UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TÉCNICA

TITULACIÓN DE INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Aplicación de técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL.

TRABAJO DE FIN DE TITULACIÓN.

AUTOR: Ordoñez Briceño, Karla Fernanda

DIRECTOR: Valdiviezo Díaz, Priscila Marisela, Mgs

LOJA - ECUADOR

2013

CERTIFICACIÓN

Magister.

Priscila Marisela Valdiviezo Díaz

DIRECTORA DEL TRABAJO DE FIN DE TITULACIÓN

C E R T I F I C A:

Que el presente proyecto de fin de carrera, denominado "*Aplicación de Técnicas de minería de datos para predecir la deserción de los estudiantes de primer ciclo de la MAD-UTPL*", realizado por la estudiante Karla Fernanda Ordoñez Briceño, ha cumplido con los requisitos estipulados en el Reglamento General de la Universidad Técnica Particular de Loja, el mismo que ha sido coordinado y revisado durante todo el proceso de desarrollo, desde su inicio hasta la culminación, por lo cual autorizo su presentación.

Loja, Septiembre 18 del 2013

f)

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

"Yo, Karla Fernanda Ordoñez Briceño declaro ser autor (a) del presente trabajo y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales.

Adicionalmente declaro conocer y aceptar la disposición del Art. 67 del Estatuto Orgánico de la Universidad Técnica Particular de Loja que en su parte pertinente textualmente dice: "Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se real+icen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad"

f.

Autor: Karla Fernanda Ordoñez Briceño
Cédula: 0705031003

DEDICATORIA

Dedico la presente tesis principalmente a mi abuelita Margarita Riofrío a quien quiero como una madre, es la persona a quien debo muchos de mis logros, ella es y será siempre mi ejemplo de lucha y esfuerzo.

A mis padres por ser mi pilar fundamental en todo lo que he conseguido hasta estas instancias de mi vida, por su apoyo incondicional, por demostrarme siempre que el que persevera alcanza. Me han dado todo lo que soy como persona, mis valores, mis principios, mi carácter, mi empeño, mi coraje para conseguir mis sueños y no desmayar en el intento.

A mi pequeña hija Melany, que aun sin conocerla, ha sido desde 6 meses atrás mi motivación principal para conseguir este logro tan anhelado, para con ello poder ser su ejemplo a seguir.

A mi amado Diego Ronald, por estar dispuesto a escucharme en los buenos y malos momentos, dándome siempre las fuerzas necesarias para no decaer en los problemas que se me presentaban durante mi formación profesional.

AGRADECIMIENTO

La presente tesis es el resultado del esfuerzo y dedicación constante, que sin la participación de algunas personas no hubiese sido posible llevarla a feliz término. Por ello, es para mí un verdadero placer utilizar este espacio para ser justo y consecuente con ellas, expresándoles mis respectivos agradecimientos.

En primera instancia agradezco a Dios, por haberme brindado, la salud, fortaleza y la sabiduría necesaria, para llevar a cabo este deseo tan anhelado, que hoy en día se vuelve una realidad.

A mi familia les agradezco de manera especial, quienes nunca dudaron que alcanzaría este logro; gracias abuelitos, papi, mami, por brindarme en todo instante su apoyo incondicional y motivación para seguir adelante con mis estudios, siempre estuvieron presentes, dándome su palabra de aliento en los momentos más difíciles de mi formación profesional.

Quiero expresar además mis más sinceros agradecimientos a mi directora de tesis la Mgs. Priscila Valdiviezo, quien me ha brindado su orientación continua en el desarrollo del presente proyecto; gracias por su participación activa, por su disponibilidad de tiempo, y por haberme facilitado siempre los medios necesarios para llevar a cabo todas las actividades propuestas durante la realización de la presente tesis.

Son muchas las personas que han formado parte de mi vida profesional, por ello les agradezco infinitamente por haberme brindarme su amistad, consejos, apoyo, y ánimo en todos estos años de estudio.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

FIGURAS CAPíTULO 1

FIGURAS CAPíTULO 2

FIGURAS CAPITULO 3

ÍNDICE DE TABLAS

TABLAS CAPíTULO 1

TABLAS CAPíTULO 3

RESUMEN

En el presente trabajo de fin de titulación se ha obtenido un modelo de minería de datos aplicando la metodología CRISM-DM, que con la ayuda del análisis de la información, que los diferentes estudiantes proporcionan a las bases de datos del sistema académico (Syllabus) y al entorno virtual de aprendizaje (Eva) de la Universidad, se ha podido obtener patrones de comportamiento, para con ello conocer cuáles son las posibles causas por las que un alumno que cursa las asignaturas de primer ciclo de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, decide abandonar sus estudios universitarios. El presente modelo permitirá a la Institución educativa obtener beneficios económicos, ya que podrá determinar las estrategias necesarias para evitar que un estudiante deserte la carrera.

PALABRAS CLAVES: deserción estudiantil, minería de datos, sistemas informáticos, bases de datos, análisis del aprendizaje, metodología crisp-dm, clustering, asociación, clasificación.

ABSTRACT

In this thesis project has been obtained data mining model using CRISM-DM methodology, with the help of data analysis, which give students the different databases of the academic system (Syllabus) and the virtual learning environment (Eva) University has been able to obtain behavioral patterns, thereby to know what are the possible reasons why a student who attends the subjects of junior Open and Distance mode of Technical University of Loja, decides to leave college. This model will enable the educational institution to monetize, and you can determine the strategies needed to prevent a student's career deserts.

KEYWORDS: dropout, data mining, computer systems, databases, analysis of learning, methodology crisp-dm, clustering, association, classification.

INTRODUCCIÓN

La deserción estudiantil se ha convertido en un problema social que afecta a muchas Universidades en todo el mundo, reducir el número de estudiantes desertores es un tema que tienen muy presente cada uno de las Instituciones educativas, una de ellas es la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, la misma que plana implementar un plan estratégico para reducir el índice de estudiantes que deciden abandonar sus estudios, por diferentes circunstancias.

Con la ayuda del análisis de la información, que los diferentes estudiantes proporcionan a los sistemas informáticos de la Universidad, se podrá crear un modelo de análisis de datos que permita obtener patrones de comportamiento de un estudiante desertor. La creación del presente modelo se lo realiza a través del análisis de la información: personal, académica del estudiante y de la interacción en el entorno virtual del curso tanto de los estudiantes como de los docentes que integran la asignatura.

Para contribuir con la solución al problema de la deserción estudiantil se plantea la aplicación de técnicas de minería de datos, con el objeto de "Comprender cuáles son las posibles causas por lo que un alumno decide abandonar sus estudios universitarios, a través del análisis de las características de los estudiantes". De acuerdo a Hand, Mannila & Smyth (2011) "la Minería de datos es un proceso que reúne un conjunto de herramientas de diversas ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras)" que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos en las empresas.

CRISM-DM fue la metodología utilizada para la creación del modelo, la misma que es una de las más usadas en la actualidad para la generación de proyectos de Minería de datos, con ella se pretende obtener un modelo de análisis de datos, que con la ayuda de la implementación de algoritmos de Inteligencia Artificial, ya incorporados en la herramienta de preprocesamiento de datos Weka, se pueda conocer cuáles son las posibles causas por las que un alumno que cursa las asignaturas de primer ciclo de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, decide abandonar sus estudios universitarios, y así la Institución educativa pueda generar las mejores estrategias para evitar dicho problema, con el presente modelo obtendrán beneficios económicos tanto los estudiantes como la institución.

El presente proyecto de fin de titulación está formado por 5 capítulos; en el capítulo 1 se describe el estado del arte del presente proyecto definiendo en el mismo el análisis del aprendizaje y algunos conceptos que engloban la minería de datos como son las técnicas y tareas de la minería de datos para extraer conocimiento, además de analizar las diferentes herramientas y librerías que se utilizan para poder analizar el conocimiento extraído. Del mismo modo se describen en el estado de arte las diferentes áreas en la actualidad en donde se están empleando técnicas de minería de datos; se empleó además un estudio de proyectos similares acerca de la aplicación de técnicas de minería de datos para predecir la deserción de estudiantes de carreras universitarias. En el capítulo 2 se describe el análisis del problema que se requiere modelar, así mismo se establece el diseño de la solución para el mismo. En el capítulo 3 se detalla, la implementación de la Metodología CRISP-DM, para la elaboración del modelo de minería de datos. Por ultimo en el capítulo 4 se presentan las conclusiones y recomendaciones del proyecto, respecto a los resultados encontrados en la minería de datos.

CAPÍTULO 1

ESTADO DEL ARTE

1.1. Minería de datos.

En la actualidad se dispone de grandes cantidades de información las mismas que están alojadas en bases de datos, archivos, documentos impresos, páginas web que se crean por una tarea cotidiana específica, dicha información no se analiza ni se integra con el resto de conocimiento de un determinado dominio. Para lo cual existe el área de la Minería de Datos que nace de la necesidad de explicar el porqué de unos sucesos, de unos comportamientos, los cuales están ocultos en datos históricos.

Hand et al. (2011) refiere que "la Minería de Datos es un proceso que reúne un conjunto de herramientas de diversas ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras)" que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos en empresas determinadas.

La minería de datos engloba un proceso para la obtención de conocimiento a partir de datos como primeramente se encuentra: la fase de selección de un conjunto, el análisis de propiedades de los datos, la transformación de conjunto de datos de entrada, seleccionar y aplicar técnica de minería de datos, seguidamente está el proceso de extracción de datos y por último la interpretación y evaluación de los datos. Referente a las fases mencionadas es sin duda la más compleja la que corresponde a la aplicación de técnicas de minerías de datos, esto debido a muchas características que pueden influir a la hora de la extracción de conocimiento para la toma de decisiones como son los tipos de variables que influyen en el conjunto de datos, interpretación y limpieza de los mismos.

Entre algunas de las técnicas de minería de datos que existen en la actualidad se encuentran: la modelización estadística paramétrica y no paramétrica, reglas de asociación y dependencia, métodos bayesianos, árboles de decisiones y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales, máquinas de vectores soporte, clustering, algoritmos evolutivos y reglas difusas, métodos basados en casos y en vecindad. De las cuales las que han sido utilizadas para la generación del presente modelo han sido: árboles de decisión, clustering, y reglas de asociación.

1.2. Análisis del aprendizaje

El Análisis del Aprendizaje es el uso que se le brinda a los datos inteligentes y modelos de análisis para descubrir la información y conexiones sociales, para con ello predecir y asesorar en el aprendizaje. EDUCAUSE iniciativa de aprendizaje ofrece una definición ligeramente distinta, refiere que es el uso de datos y modelos para predecir el progreso del estudiante y su rendimiento. Una serie de informes y artículos de la fundación con sede en EDUCAUSE sugirieron que el análisis del aprendizaje podría ofrecer una solución que ayuda a aumentar la retención escolar (no deserción) y las tasas de éxito.

En *la 1ª Conferencia Internacional del Análisis del Aprendizaje y el Conocimiento 2011,* el comité directivo estableció que "el análisis del aprendizaje es la medición, recolección, análisis y presentación de datos sobre los alumnos y sus contextos, a fines de comprender y optimizar el aprendizaje y los entornos en los que se produce", (Dron *et al.*, 2011*)* para lo cual es utilizada la minería de datos, ya que permite la extracción de dicho conocimiento a partir de un conjunto de datos, que tiene como objetivo descubrir patrones que ayuden a mejorar las técnicas de aprendizaje en el ámbito de la educación.

Existen algunos campos relacionados con el análisis del aprendizaje tales como: Minería de datos para la educación (EDM), análisis de redes sociales e inteligencia de negocios. El EDM se centra más específicamente en herramientas y métodos para la exploración de los datos procedentes de contextos educativos (Dyckhoff, Dennis, Bültmann, Chatti Ulrik, & Schroeder, 2012); generalmente el EDM realiza un análisis de un conjunto de datos reducido, a diferencia del análisis del aprendizaje que no hace hincapié en la reducción de los datos relacionados con el aprendizaje sino que busca comprender los sistemas enteros, para de esa manera apoyar la toma de decisiones referente al modelo de aprendizaje y enseñanza en las instituciones educativas. El análisis de redes sociales busca a su vez, mediante la aplicación de modelos extraídos de la Teoría de redes sociales, predecir el comportamiento de una persona que utiliza una red social como Blogs, Wiki, etc; la presente disciplina es por ejemplo, el análisis de las relaciones estudiante a estudiante, y de estudiante-profesor, con el fin de identificar a los estudiantes con una característica o comportamiento especifico (Bienkowski, Feng, & Means, 2012); es sustancial señalar que las redes sociales son particularmente importantes para el aprendizaje desde el punto de vista de la tecnología, ya que permite mejorar las técnicas de aprendizaje en las personas, por el conocimiento que circulan en las mismas. Por su parte la Inteligencia de Negocios (BI) se enfoca no solo en el ámbito de la educación dentro de la institución educativa, sino también incluye cuestiones organizativas y financieras

1.3. Tareas de minería de datos

Se clasifican en dos grupos las tareas que se realiza en la Minería de Datos para poder extraer conocimiento oculto, estas son las predictivas que permiten predecir uno o más valores para uno o más ejemplos. Y el otro grupo es de las tareas descriptivas su objetivo no es predecir nuevos datos sino describir los existentes, a continuación se describen las tareas de ambos grupos:

1.3.1. Tareas predictivas.

1.3.1.1. Clasificación.

El objetivo de la tarea es poder clasificar un dato dentro de las clases definidas del dominio que se está modelando (Riquelme, Ruiz & Gilbert, 2006).

Ejemplos: Clasificar un mensaje de correo electrónico como spam o no; clasificar entre varios medicamentos cuál es el mejor para una determinada enfermedad, clasificar los clientes que pagan y los que no pagan los préstamos.

1.3.1.2. Regresión.

El objetivo de la tarea es poder encontrar la similitud entre valores de atributos de una determinada clase de un dominio dado (Riquelme *et al.,* 2006).

Ejemplos: Predecir el número de unidades defectuosas de una partida de productos, predecir la presión de una válvula a partir de las entradas. Determinar el stock de cada producto de una tienda a través del análisis de ventas anteriores para que el número de productos en bodega sea suficiente para poder servir rápidamente los pedidos de los clientes.

1.3.2. Tareas descriptivas.

1.3.2.1. Agrupamiento (clustering).

El objetivo de la presente tarea es obtener grupos o conjuntos en donde se incorpore elementos similares extraídos de las clases del dominio dado (Riquelme *et al.,* 2006).

Ejemplos: Se pueden agrupar los clientes por segmentos según sus perfiles, estudiar que grupos se comportan mejor ante determinados productos, y después orientar ciertos productos a ciertos grupos; además se pueden definir grupos diferencias de los empleados para poder establecer políticas sociales en la empresa.

1.3.2.2. Asociación.

El objetivo de la tarea es poder describir las relaciones que existen entre los valores de los atributos de un determinado ejemplo de un dominio establecido.

Ejemplos: Se puede realizar un análisis de la cesta de compra en un supermercado para poder obtener información acerca de los productos que los clientes compran en conjunto, con el objetivo de mejorar la ubicación de los productos en el dicho establecimiento.

1.3.2.3. Correlación.

El objetivo de la presente técnica es ver, dados los ejemplos del conjunto E= A_1, A_2, A_3-------- A_n, si dos o más atributos numéricos A_i y A_j están correlacionados linealmente o relacionados de algún otro modo mediante un análisis de varianza Coeficiente de correlación lineal de los datos (Hernández, Ramírez & Ferri, 2004).

Ejemplos: Se puede analizar en un centro de salud los factores que influyen para que un paciente pueda asistir al dicho establecimiento.

1.4. Técnicas de minería de datos.

A continuación se describen algunas técnicas de minería de datos para llevar a cabo las tareas anteriormente mencionadas.

1.4.1. Modelización estadística paramétrica.

Esta técnica consiste en explicar el comportamiento de una variable a partir del conocimiento que se está extrayendo mediante fórmulas algebraicas, por ejemplo, el saldo total bancario de las personas de una cierta edad, con un mismo nivel profesional y residiendo en una misma ciudad no es igual para todos ellos sino que sigue una cierta distribución, pero sin embargo si sabemos la edad de una persona, su nivel profesional y su ciudad de residencia podremos dar una aproximación de su saldo bancario. En el presente ejemplo la variable que queremos analizar son: saldo bancario, denominada variable de salida (output), conocida también como de respuesta, mientras que las variables edad, nivel profesional, ciudad, se denominan variables de entrada(input), (Hernández *et al.,* 2004).

La modelización lineal paramétrica en un contexto de minería de datos, tanto se utilizan métodos matemáticos ya existentes como son regresión lineal múltiple (MLR), mínimos cuadrados generalizados con una estructura de correlación no nula (GLS), modelo lineal de efectos mixtos (LME) y mínimos cuadrados parciales (PLS).

La regresión y la clasificación son dos de las tareas más utilizadas para este tipo de técnica.

Algunos algoritmos conocidos para este tipo de técnica son la regresión lineal, regresión logarítmica, y la regresión logística.

Una de las ventajas de los métodos paramétricos está la de que pueden ser ajustados con una cantidad pequeña de datos (por ejemplo, puede usarse el método de mínimos cuadrados o el de máxima verosimilitud) (Hernández *et al.,* 2004).

Una de las desventajas de la presente técnica es que su estructura es tan rigurosa que no pueden adaptarse a grandes conjuntos de datos (Hernández *et al.,* 2004), porque necesitaría un gran nivel de procesamiento para dicha actividad.

1.4.2. Modelización estadística no paramétrica.

Esta técnica permite construir modelos más flexibles, a comparación de la modelización estadística paramétrica, porque son capaces de modelar fenómenos complejos que involucran el análisis de un gran volumen de datos (Hernández *et al.,* 2004).

En la presente técnica igual que la anterior se utilizan tareas de regresión y clasificación pero en este caso son métodos no paramétricos, como k-ésimo vecino más cercano (k-NN) y vecino más similar (MSN).

1.4.3. Reglas de Asociación y Dependencia.

Esta técnica consiste en que mediante reglas se expresan patrones de comportamiento entre los datos de las clases del dominio en función de la aparición conjunta de valores de dos o más atributos (Hernández *et al.,* 2004). La característica principal de estas reglas es que tratan con atributos nominales es decir que un atributo puede tener un valor de un conjunto de valores establecidos, por ejemplo el atributo género (masculino, femenino).

En la presente técnica las reglas expresan las combinaciones de valores de los atributos que ocurren con mayor frecuencia, por lo cual utiliza la tarea de asociación.

La presente técnica puede trabajar con grandes volúmenes de datos el uso de esta técnica se lo hace por ejemplo en el análisis de la cesta de la compra de un supermercado para poder determinar qué productos se compran conjuntamente, además también se las utiliza para hacer el estudio de textos, y búsquedas de patrones en páginas web.

Las reglas de asociación hacen uso del algoritmo de aprendizaje A priori, para extraer los patrones de comportamiento.

1.4.4. Métodos Bayesianos.

Una de las características primordiales de los métodos bayesianos es el uso de distribuciones de probabilidad para cuantificar incertidumbre de los datos que se desea modelar. Estos métodos proporcionan una metodología práctica para la inferencia y predicción y, en última instancia, para tomar decisiones que involucran cantidades inciertas (Hernández *et al.,* 2004).

Una de las desventajas de los métodos bayesianos es que no pueden realizar predicciones con pocos datos, ya que no podría proporcionar un modelo correcto con poca cantidad de información proporcionada.

Hernández *et al. (*2004) señala que la presente técnica "*es una de las que más se han utilizado en problemas de inteligencia artificial, con ello en el aprendizaje automático y minería de datos, ya que es un método práctico para realizar inferencias a partir de los datos, la misma que se basa en estimar la probabilidad de pertenecía (a una clase o grupo) mediante la estimación de las probabilidades, utilizando para ello el teorema de Bayes*". Por lo tanto se la podría utilizar para modelar cualquier tipo de conocimiento.

Los métodos bayesianos utilizan la tarea de clasificación para extraer los patrones de comportamiento. Algunos algoritmos que se utilizan frecuentemente para este tipo de técnica son el clasificador bayesiano Naive, Bayes Net, los métodos basados en máxima verisimilitud y el algoritmo EM.

Por ejemplo de la ventaja que supone poder dar la probabilidad asociada a la clasificación, en un sistema de recomendación para invertir en bolsa, en que a partir de unos datos de entrada sobre un determinad producto, el sistema nos recomienda si invertir o no, nos puede proporcionar la probabilidad de inversión.

Una de las desventajas de los métodos bayesianos es que requieren de un alto costo computacional por el nivel de procesamiento que necesitan para ejecutar sus algoritmos.

1.4.5. Árboles de decisión y sistemas de reglas.

Hernández *et al. (*2004) señala que la "técnica basada en árboles de decisión es quizás el método más fácil de utilizar y de entender". Un árbol decisión es una estructura jerárquica que está formado por un conjunto de nodos, en donde cada nodo establece una condición o regla la misma que puede retornar verdadero o falso según los valores de los atributos que se desean analizar, de tal manera que la decisión final a tomar se puede determinar si-

guiendo las condiciones que se cumplen desde el nodo raíz (superior) del árbol hasta alguno de sus nodos hojas (inferior).

Las tareas que utilizan este tipo de técnica son la: clasificación, regresión y agrupamiento.

Este tipo de técnica se basan en dos tipos de algoritmos: los denominados “divide y vencerás”, como el ID3/C4.5 o el CART, y los algoritmos denominados “separa y vencerás”, como el CN2.

Esta técnica es utilizada para expresar procedimientos médicos, legales, comerciales, estratégicos. Por ejemplo en un hospital público se utiliza la presente técnica para poder determinar si es recomendable o no que un paciente se realice una determina operación según las condiciones médicas del mismo.

Una de las ventajas de los árboles de decisión es que las opciones resultantes posibles a partir de una determinada condición son precisas, es decir que con la ayuda de los árboles de decisión se puede llegar a una sola acción o decisión a tomar.

1.4.5.1. Ejemplo: Implementación de Árbol de Decisión en Weka bajo el algoritmo J48.

- Generación de un Árbol de Decisión para determinar una de las razones porque los estudiantes deciden abandonar sus estudios universitarios.

```
=== Confusion Matrix ===

  a  b   <-- classified as
  5  0 |  a = SiAbandono
  0 21 |  b = NoAbandono
```

FIGURA 1. 1. Matriz de confusión

La [*Figura. 1.1*] ilustra la matriz de confusión generada en weka la misma que señala el tipo de errores cometidos en la generación del modelo, en el presente caso existe 0 errores en la predicción.

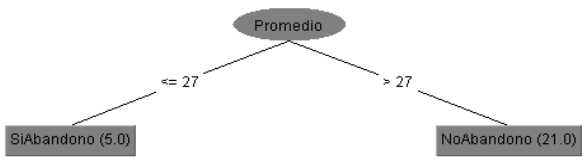

FIGURA 1. 2. Ejemplo de árbol de decisión en Weka con la Variable Promedio

La [*Figura. 1.2*] ilustra el árbol generado con la variable promedio, deduciendo que mientras el promedio del alumno del ciclo sea <=27 el alumno esté decidirá abandonar la carrera.

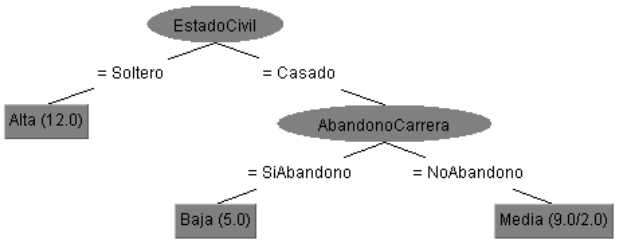

FIGURA 1. 3. Ejemplo de árbol de decisión en weka con la variable estado civil.

La [*Figura.1.3*] ilustra el árbol generado con la variable promedio, deduciendo que mientras el estado civil de un alumno sea casado esté decidirá abandonar la carrera, además se puede observar que la calidad del servicio académico fue baja, y los que no abandonaron la carrera han recibiendo una calidad de servicio Media.

1.4.6. Métodos relacionales y estructurales.

En la presente técnica se hace uso de un lenguaje de representación relacional, mucho más ponente expresivamente hablando que los tradicionales lenguajes de representación típicos de la mayoría de métodos de minería de datos ya que analizan los datos de las tablas relacionadas (Hernández *et al.,* 2004). Los métodos relacionales nos permiten descubrir patrones de comportamiento más complejos, haciendo uso de la estructura de los propios datos y las relaciones entre ellos, sin necesidad de unir todos los datos en un solo conjunto. Este tipo de técnicas son utilizadas en las áreas de bioinformática o la farmacología.

Los métodos utilizados en esta técnica son: Programación lógica inductiva (ILP por sus siglas en inglés), Aprendizaje basado en grafos, Modelos probabilísticos relacionales, Aproximaciones relacionales basadas en distancia, árboles de decisión relacionales, reglas de asociación relacionales, inducción de programas lógico-funcionales.

Una de las desventajas de la presente técnica es que son mucho menos eficientes, con respecto a la velocidad de respuesta que proporciona, comparadas con otras técnicas.

1.4.7. Redes neuronales artificiales.

Se trata de una técnica que está inspirada del funcionamiento del sistema nervioso en los seres humanos el mismo que aprenden un modelo mediante el entrenamiento de los pesos (valores) que conectan un conjunto de nodos o neuronas de la red, la presente técnica recibe como entrada un conjunto de datos de entrenamiento. La neurona se activará si el resultado es superior a un determinado límite u umbral con la finalidad de comunicarse con otras neuronas.

Una de las ventajas que ofrecen las redes neuronales es que son capaces de trabajar con información incompleta o inconsistente, ya que existen conjuntos de datos que se desea modelar con este tipo de características.

Hernández *et al. (*2004) señala que la presente técnica posee dos tipos de aprendizaje uno es el supervisado, en el mismo que se le proporciona un conjunto de datos de entrada y la respuesta correcta es útil en tareas de regresión y clasificación. Y el aprendizaje no supervisado solo se le da a la red un conjunto de datos de entrada y la red debe auto-enseñarse para proporcionar una respuesta, este aprendizaje es útil para las tareas de agrupamiento.

Las redes neuronales han sido utilizados en diversos campos por ejemplo: como la predicción de mercados financieros, control de robots, la teledetección que se refiere a recoger información a través de diferentes dispositivos electrónicos de un objeto (barco, avión, satélite).

Una de las desventajas de la presente técnica es que el modelo aprendido es difícilmente comprensible.

El algoritmo más común utilizado para este tipo de técnica es el de retropropagación (backpropation).

1.4.8. Máquinas de vectores soporte.

Se tratan de técnicas que intentan maximizar el margen entre los grupos o las clases formales.

Los campos en donde las SVM han sido aplicadas con éxito incluyen, entre otros, la visión por computador, la bioinformática, la recuperación de información, el procesamiento de lenguaje natural y el análisis de series temporales (*Pérez, 2010).*

Con las máquinas de vectores soporte se ha logrado realizar un pronóstico semanal del tipo de cambio australiano frente a cinco monedas extranjeras.

La presente técnica utiliza tareas de clasificación, regresión y agrupamiento.

Está técnica posee una desventaja porque no son adecuadas para la clasificación con grandes conjuntos de datos, ya que su procesamiento se torna lento.

1.4.9. Extracción de conocimiento con algoritmos evolutivos y reglas difusas.

Hernández *et al. (*2004) señala que a través de “los algoritmos evolutivos se puede utilizar un procedimiento determinístico para alcanzar una solución óptima: comienzan desde un punto aleatorio y se basan en una regla de transición especificada previamente para determinar la dirección de la búsqueda”. Los presentes algoritmos utilizan un algoritmo de búsqueda efectivo para poder encontrar las mejores reglas de un determinado patrón de datos y así poder brindar las mejores recomendaciones.

Los algoritmos evolutivos se han utilizado con éxito para resolver el problema de selección de instancias, entendiendo como tal la determinación del subconjunto de ejemplos más significativos dentro de un conjunto total.

Los algoritmos evolutivos se los puede utilizar para poder analizar las compras que frecuentemente realiza un cliente en una tienda específica y así poder realizar recomendaciones de productos relacionados. Además puede ser utilizada la presente técnica para poder agrupar los diferentes intereses de los alumnos en una clase y con ello poder brindar recomendaciones de los textos que pueden ser de su interés.

La lógica difusa permite modelar conocimiento impreciso y cuantitativo así como permitir manejar la incertidumbre. Por ejemplo: Si radio es pequeño y el nivel de simetría es alto entonces el nivel de concavidad es muy bajo.

La lógica difusa son útiles para construir un modelo basado en reglas que son fáciles de interpretar por los usuarios normales y que permita incorporar la información proveniente de: la experiencia de un experto, y la información proveniente de modelos matemáticos o medidas empíricas con lo que permite determinar una clase para cualquier patrón de datos admisible que llegue al sistema (Riquelme, Ruiz, & Gilbert, 2006).

La utilización de la lógica difusa para extraer conocimiento es muy útil para tratar problemas de clasificación.

La lógica difusa se puede agrupar en función del tipo de tarea que se realice como Agrupamiento, Clasificación, Reglas de Asociación, Dependencias funcionales, Sumarización de datos.

1.4.10. Métodos basados en casos y en vecindad

En la presente técnica se inicia a procesar un conjunto de ejemplos existentes para poder hacer comparaciones con los nuevos casos y los casos pasados, para aprender del ellos (Hernández *et al.,* 2004). Estos métodos determinan la similitud que puede existir entre los casos y así poder obtener los patrones de datos.

Se trata de técnicas que se basan en medir las distancias entre los valores de los atributos esta tarea la realizan con la ayuda de los métodos de los vecinos más próximos (los casos más similares) o mediante la estimación de funciones de densidad. Para este tipo de técnica se utiliza algoritmos de vecinos más próximos (K-NN), algoritmos jerárquicos (Two-step, COBWED), algoritmos no jerárquicos (K-means).

1.4.11. Algoritmos de minería de datos

El algoritmo de minería de datos es el mecanismo que crea modelos de minería de datos. Para crear un modelo, un algoritmo analiza primero un conjunto de datos, buscando patrones y tendencias específicos. Después, el algoritmo utiliza los resultados de este análisis para definir los parámetros del modelo de minería de datos (Thearling, 2007).

Los algoritmos de Minería de datos están clasificados dentro de dos grupos, los mismos que se detallan a continuación: (FACENA – UNNE, 2003).

- **Supervisados o predictivos:** predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción de datos cuya etiqueta es desconocida.

- **No supervisados o del descubrimiento del conocimiento:** con estos algoritmos se descubren patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

En la [*Tabla 1.1*], se puede observar las técnicas, que se encuentran dentro de los grupos antes descritos:

TABLA 1. 1. Clasificación de las técnicas de minería de datos

Supervisados o Predictivos	**No Supervisados o Descriptivos**
Arboles de Decisión	Detección de desvíos
Inducción Neuronal	Segmentación
Regresión	Agrupamiento ("clustering")
Series Temporales	Reglas de Asociación
	Patrones Secuenciales

Fuente: [GUTIÉRREZ. (2008)]

1.4.12. Algoritmos de clusteing o agrupamiento

Los presentes algoritmos son utilizados para crear grupos de datos, con características similares.

1.4.12.1. K-Means

Uno de los algoritmos más utilizados para el agrupamiento de datos, es el K-Medias o K-Means, por ser uno de los más veloces y eficaces. Dicho algoritmo trabaja con un método de agrupamiento por vecindad, en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar sin etiquetar.

El propósito de K-Means es ubicar a los prototipos o centros en el espacio, de forma que los datos pertenecientes al mismo prototipo tengan características similares (Moody & Draken, 1989). Todo ejemplo nuevo, una vez que los prototipos han sido correctamente situados, es comparado con estos y asociado a aquel que sea el más próximo, en los términos de una distancia previamente elegida. Normalmente, se utiliza la distancia euclidiana.

El objetivo que se busca mediante el algoritmo K-Means es minimizar la varianza total intra-grupo o la función de error cuadrático, para que el algoritmo pueda generar los mejores resultados. [*ver Figura. 1.4*]

$$V = \sum_{i=0}^{K} \sum_{j \in S_i} \left| \chi_j - \mu_i \right|^2$$

FIGURA 1. 4. Error cuadrático k-means, [gutiérrez. (2008)].

K-Means comienza particionado los datos en k subconjuntos no vacíos, aleatoriamente o usando alguna heurística. Luego calcula el centroide de cada partición como el punto medio del cluster y asigna cada dato al cluster cuyo centroide sea el más próximo. Luego los centroides son recalculados para los grupos nuevos y el algoritmo se repite hasta la convergencia, la cual es obtenida cuando no haya más datos que cambien de grupo de una iteración a otra. (Gutiérrez, 2008).

Con la finalidad de calcular el centroide más cercano a cada punto, se debe utilizar una función de distancia. Para datos reales se suele utilizar la distancia euclidiana. En el caso que se utilice datos categóricos se debe establecer una función específica de distancia para dicho conjunto de datos. Algunas de las opciones son utilizar una matriz de distancias predefinidas o una función heurística.

Dado k (No de grupos), el algoritmo K-Means se implementa en los siguientes 4 pasos (Ale, 2005):

1) Particionar los objetos en k subconjuntos no vacíos.
2) Computar los centroides de los cluster de la partición corriente. El centroide es el centro (punto medio) del cluster.
3) Asignar cada objeto al cluster cuyo centroide sea más cercano.
4) Volver al paso 2, y parar cuando no existan más reasignaciones.

Para evaluar los resultados generados por simple K-means en caso de utilizar la herramienta Weka, se recomienda aplicar Use tranining set que permite utilizar el propio conjunto de entrenamiento, que indica que porcentaje de instancias se van a cada grupo.
El mencionado modo de evaluación genera en la ventana de texto en Weka, el número de interacciones que ha realizado el algoritmo para crear el modelo, y el error cuadrático en los clusters, mientras menor sea el error cuadrático más homogéneos serán los clusters generados, tomando en cuenta que el valor del error depende de la semilla que se haya establecido en los parámetros del algoritmo.

1.4.13. Algoritmos de clasificación

Los presentes algoritmos son utilizados para clasificar un conjunto de datos, dentro de una clase especifica.

1.4.13.1. J48

J48 es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta weka, el presente algoritmo permite generar un árbol de decisión, a través de los datos ingresados, seleccionando el mejor atributo que clasifique a los datos. El presente algoritmo trabaja tanto con atributos numéricos, como nominales, además permite realizar la clasificación con ausencia de datos, para la generación del modelo.

El algoritmo J48 es uno de los más utilizados en minería de datos, permite trabajar con valores continuos para los atributos, separando los posibles resultados en las ramas respectivas. Además, los árboles son menos frondosos porque cada hoja no cubre una clase en particular sino una distribución de clases, lo cual los hace menos profundos y frondosos.

Wilford et al. (2008) menciona que el algoritmo J48 amplía las funcionalidades del C4.5, tales como permitir la realización del proceso de post-poda del árbol mediante un método basado en la reducción del error (reducedErrorPruning) o que las divisiones sobre las variables discretas (contiene número finito de valores) sean siempre binarias (binarySplits). Algunas propiedades concretas de la implementación son las siguientes:

- Admite atributos nominales y numéricos.
- Se permiten ejemplos con valores desconocidos.

El presente algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información (Hernández & Ferri, 2006) es decir, elige al atributo que mejor clasifica a los datos dentro de una categoría definida. El árbol que se genera está formado por: [*ver Figura. 1.5*]

- Nodos: Nombres de los Atributos seleccionados.
- Ramas: Valores de los determinados atributos.
- Hojas: Conjuntos de datos clasificados y etiquetados con el nombre de la clase.

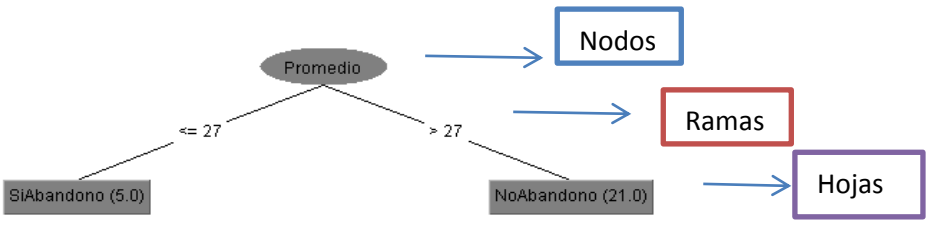

FIGURA 1. 5. Estructura de un árbol de decisión en weka

El proceso de construcción del árbol de decisión comienza por el nodo raíz, el que tiene asociados todos los ejemplos o casos de entrenamiento. Lo primero es seleccionar la variable o atributo a partir de la cual se va a dividir la muestra de entrenamiento original (nodo raíz), buscando que en los subconjuntos generados haya una mínima relación respecto a la clase. Este proceso es recursivo, es decir, una vez que se haya determinado la variable con la que se obtiene la mayor semejanza respecto a la clase en los nodos hijos, se vuelve a realizar el análisis para cada uno de los nodos hijos. Aunque en el límite este proceso se detendría cuando todos los nodos hojas contuvieran casos de una misma clase, no siempre se desea llegar a este extremo, para lo cual se implementan métodos de pre-poda y post-poda de los árboles (Wilford, Rosete & Rodríguez, 2008).

Para evaluar el modelo generado por la presente técnica en la herramienta Weka se suele utilizar: Cross-validation: evaluación con validación cruzada, la misma que es la más elaborada y costosa. Se realizan tantas evaluaciones como se indica en el parámetro Folds en la herramienta. Se dividen las instancias en tantas carpetas como indica este parámetro y en cada evaluación se toman las instancias de cada carpeta como datos de test, y el resto como datos de entrenamiento para construir el modelo. Los errores calculados son el promedio de todas las ejecuciones.

En el caso de utilizar la herramienta Weka, al momento de ejecutar el clasificador sobre los datos, en la ventana de texto de la aplicación, aparece información de ejecución, el modelo generado con todos los datos de entrenamiento y los resultados de la evaluación.

Referente a los resultados de evaluación que devuelve el clasificador se destacan los siguientes:

- Resumen (Summary): es el porcentaje global de errores cometidos en la evaluación

- Precisión detallada por clase: para cada uno de los valores que puede tomar el atributo de clase: el porcentaje de instancias con ese valor que son correctamente predichas (TP: true positives), y el porcentaje de instancias con otros valores que son incorrectamente predichas a ese valor aunque tenían otro (FP: false positives). Las otras columnas, precision, recall, F- measure, se relacionan con estas dos anteriores.

- Matriz de confusión: aquí aparece la información detallada de cuantas instancias de cada clase son predichas a cada uno de los valores posibles.

1.4.14. Algoritmos de Asociación

Son algoritmos que surgieron inicialmente para afrontar el análisis de las cestas de la compra de los comercios. Permiten expresar patrones de comportamiento entre los datos, en función de la aparición conjunta de valores de dos o más atributos.

1.4.14.1. Apriori

El algoritmo utiliza recursividad por niveles, el mismo que trabaja solamente con atributos nominales. En un primer paso genera los candidatos y seguidamente los pone a prueba para descartar los itemsets no frecuentes.

El algoritmo Apriori empieza generando todos los ítems-sets con un elemento. Usa éstos para generar los de dos elementos y así sucesivamente. Se toman todos los posibles pares que cumplen con las medidas mínimas del soporte. Esto permite ir eliminando posibles combinaciones ya que no todas se tienen que considerar. Genera las reglas revisando que cumplan con el criterio mínimo de confianza [Hernández *et al. (*2004)].

Para evaluar las reglas se emplean la medida del soporte [support] o cobertura, que indica el número de casos, ejemplos, que cubre la regla y la confianza [confidence], que indica el número de casos que predice la regla correctamente.

1.4.15. Algoritmo para la Selección de atributos.

A continuación se establecen algunas razones, para realizar la selección de atributos (Hernández Orallo et al, 2004).

- Reducir el tamaño de los datos eliminando los irrelevantes o redundantes.
- Eliminar atributos que contienen numerosos datos erróneos o faltantes.
- Mejorar la calidad del modelo centrándose en atributos relevante.
- Expresar el modelo resultante en función de menos variables mejorando la comprensión.
- Reducir la dimensionalidad a fin de representar los datos visualmente.

Surge un inconveniente, cuando se hace uso de todos los atributos disponibles, puesto que la mayoría de las técnicas de minería de datos pueden actuar incorrectamente cuando se posee tantos campos en el dataset, siendo estos: irrelevantes, redundantes o con valores erróneos. Se busca obtener modelos que se ajusten a particularidades de los datos de entrenamiento, para que brinden resultados eficaces, ya que si se trabaja con todos los datos se podría obtener deducciones poco satisfactorias y complejas. La Selección de atributos

para un conjunto de atributos, se realiza haciendo una búsqueda en el espacio del conjunto y evaluando cada uno de sus atributos.

Para realizar la evaluación del conjunto de atributos, se requiere combinar uno de los 4 evaluadores de conjuntos con alguno de los 7 métodos de búsqueda implementados en las herramientas de minería, los mismos que se detallan a continuación: (Cubero, Berzal, & Herrera, 2006)

- **Evaluador de Conjunto de Atributos**

    - WEKACfsSubsetEval: Considera el valor predictivoindividual de cada atributo
    - ClassifierSubsetEval: Usar un clasificador para evaluar
    - ConsistencySubsetEval: Mide la consistencia en términos de las clases
    - WrapperSubsetEval: Usa un clasificador + validación cruzada

- **Método de Búsqueda**

    - BestFirst: GreedyIncremental con backtracking
    - ExhaustiveSearch: Fuerza bruta
    - GeneticSearch: Algoritmo genético de búsqueda
    - GreedyStepWise: Greedyincremental sin backtracking
    - RaceSearch: Metodología RaceSearch
    - RandomSearch: Búsqueda Aleatoria
    - RankSearch: Ordena los atributos y crea un ranking de subconjuntos prometedores.

Un método más rápido pero menos preciso consiste en evaluar los atributos individualmente y ordenarlos, descartando atributos que caen debajo de un determinado umbral, los métodos de búsqueda utilizados para realizar la presente actividad son los mismos que han sido utilizados para evaluar un conjunto de atributos, y los evaluadores que se deben utilizar para evaluar los atributos individuales, se describen a continuación:

- **Evaluador Individual de Atributos**

    - WEKAChiSquaredAttributeEval: Calcula la estadística chi-cuadrado de cada atributo con respecto a la clase.
    - GainRatioAttributeEval: Evaluación por tasa de ganancias
    - InfoGainAttributeEval: Evaluación por ganancia de información
    - OneRAttributeEval: Metodología OneR

- PrincipalComponents: Análisis de principales componentes y transformación
- ReliefFAttributeEval: Evaluados basado en instancias
- SVMAttributeEval: Usar máquinas de soporte vectorial para calcular los atributos
- SymmetricalUncertAttributeEval: Evalúa atributos basándose en incertidumbre simétrica.

1.5. Correspondencia entre tareas, técnicas y algoritmos.

En la Tabla 1.2, se muestra la correspondencia que existe entre las técnicas de minería, con las tareas y los algoritmos, el texto entre paréntesis del campo 'técnica', son los algoritmos que se pueden utilizar en dicha técnica.

TABLA 1. 2. Correspondencia entre técnicas, algoritmos y las tareas

	Predictivas		**Descriptivas**		
Técnica (algoritmo)	**Clasificación**	**Regresión**	**Agrupamiento**	**Asociación**	**Correlación**
Redes Neuronales	X	X	X		
Árboles de decisión (ID.3, C4.5, C5.0)	X				
Árboles de decisión (CART)	X	X			
Árboles de decisión y sistemas de reglas(CN2)	X			X	
Redes de Kohonen			X		
Modelización Estadística (Regresión lineal), (Regresión Logarítmica		X			X
Modelización Estadística (Regresión Logística)	X			X	
Métodos basados en casos y en vecindad (K-means)			X		
Reglas de Asociación y Dependencia (A priori)				X	
Métodos Bayesianos(Naive Bayes)	X				
Métodos basados en casos y en vecindad (vecinos más próximos)	X	X	X		

Métodos basados en casos y en vecindad (Two-step, COBWED),			X		
Algoritmos genéticos y evolutivos	X	X	X	X	X
Máquinas de vectores soporte	X	X	X		

1.6. Herramientas de minerías de datos

Las herramientas de minería de datos nos facilitan el desarrollo de los modelos para la extracción de conocimiento de un dominio establecido, dichas herramientas contienen los algoritmos específicos para la aplicación de técnicas de minería de datos, se los puede utilizar mediante la interfaz gráfica que brindan algunas aplicaciones.

Pérez, González (citado por Galán, 2009) menciona algunas herramientas tanto comerciales como de libres distribuciones útiles para el desarrollo de modelos de minería de datos:

1.6.1. Spss clementine.

Es uno de los sistemas de Minería de Datos más conocidos, el mismo se caracteriza por: Acceso a datos (fuentes de datos archivos ASCII); procesamiento de datos; aplicación de técnicas de aprendizaje como (redes neuronales, reglas de asociación), incorpora técnicas de evaluación de modelos visualización de resultados como (histogramas, diagramas de dispersión).

1.6.2. Weka (Waikato environment for knowledge analysis).

Es una herramienta visual de libre distribución desarrollada por los investigadores de la Universidad de waikato en Nueva Zelanda.

Sus principales característica son: Acceso de los datos desde un archivo en formato ARFF (es un archivo de texto plano organizado en filas y columnas); preprocesador de datos (selección, transformación de atributos) visualización del entorno; aplicación de técnicas de aprendizaje como (redes neuronales, reglas de asociación, arboles de decisión).

1.6.3. Kepler.

Es una herramienta comercial distribuida por Dialogis. Posee múltiples modelos de análisis como: Redes neuronales, Regresión no lineal, Aplicaciones estadísticas. Así mismo permite el preprocesado de datos, la elección de un modelo o la manipulación de la representación gráfica de los modelos obtenidos.

1.6.4. Odms (oracle data mining suite).

Es una herramienta comercial que está diseñado sobre una arquitectura cliente servidor; ofrece una gran versatilidad en cuanto al acceso a grandes volúmenes de información.

Se caracteriza principalmente por: Acceso a datos en diversos formatos: almacenes de datos, bases de datos relacionales como SQL, Oracle, archivos planos; preprocesador de datos: muestreo de datos, patrones de datos; posee modelos de aprendizaje como: redes neuronales, regresión lineal. Además brinda herramientas de visualización para resultados estadísticos, importación de datos en Excel, Word o Power Point.

1.6.5. Dbminer.

Sistema interactivo desarrollado por la Universidad de Simon Fraser de Canadá. Su licencia es pública a excepción de la empresarial (que es comercializada). Está concebido para la extracción del conocimiento de bases de datos relacionales, almacenes de datos y Web con la utilización de técnicas como: Reglas de Asociación, Reglas difusas e incorpora tareas de clasificación.

Dentro de su arquitectura de diseño es importante destacar:

- OLAP (online analytic processing)
- OLAM (online analytic mining)

La herramienta de DBMiner posee dos modos de trabajo:

- Vía interfaz gráfica.
- Vía interfaz de script.

1.6.6. Rapid miner (yale).

Es una herramienta de aprendizaje automático implementado en Java por la Universidad de Dormund de libre distribución. El sistema incluye operaciones para Importación y preprocesamiento de datos, aprendizaje automático, validación de modelos, permite la aplicación de técnica como (redes neuronales, reglas de asociación, arboles de decisión).

1.6.7. Db2 intelligent miner.

Herramienta comercial distribuida según la arquitectura cliente / servidor y distribuida por IBM que permite procesar grandes cantidades de datos. Posee una serie de paquetes de implementación para Minería de Datos entre los que destacan:

- Db 2 intelligent miner for data: destinado para aplicar tareas de minería en bases de datos. Soporta tareas de agrupamiento, asociaciones, patrones, clasificación.

- Db 2 intelligent miner scoring: utiliza la funcionalidad de la base de datos para aplicar las técnicas de Minería de Datos.
- Db 2 intelligent miner modelling: orientado al descubrimiento de las relaciones entre los datos, como las asociaciones o el agrupamiento de dichos datos.

- Db 2 intelligent miner visualization: para visualizar los resultados procedentes de los modelos de análisis.

1.6.8. Sas enterprise miner.

Es una herramienta comercial que se centra en la Minería de Datos (de forma tradicional) y no en su funcionalidad (es el caso de SAS Text Miner)

Posee una arquitectura distribuida, es decir, tiene una potente interfaz gráfica de usuario. Las tareas que realiza esta herramienta son:

- Preprocesado de datos: tratamiento estadístico, filtros, tareas de muestreo. Modelos: árboles de decisión, regresión lineal, redes neuronales, construcción de métodos de ensamblaje.
- Visualización de resultados: a través de gráficos, diagramas, informes en formato html.

1.6.9. Statistica data miner.

Es una potente herramienta con un sistema visual desarrollado y comercializado, en la que destaca:

- Base de datos: permite trabajar con un gran volumen de información, así como importar los datos en formatos Excel, Oracle, SQL.

- Preprocesado de datos: a través del cual seleccionamos las características, el muestreo de datos, realizamos operaciones de filtrado, tratamiento de datos.

- Modelos de análisis: como reglas de asociación, redes neuronales, modelos lineales de regresión, modelos no lineales de regresión, regresión múltiple.

- Visualización: desarrollada a través de una interfaz gráfica que facilita las diversas tareas que pueda realizar el usuario.

- Destacan los diagramas de barras, diagramas de sectores, árboles de asociación, redes neuronales.

1.6.10. Cart.

Herramienta desarrollada y comercializada por Salford System y orientada a tareas de clasificación o regresión de Minería de Datos. Se destaca principalmente por su accesibilidad, capacidad de visualización o información estadística relativa al modelo.

1.7. Áreas de aplicación de la minería de datos.

1.7.1. Educación.

La minería de datos en la educación ayuda a determinar qué factores influyen para que un estudiante decida abandonar sus estudios en un determinado establecimiento, además puede contribuir para poder analizar los factores del porque los estudiantes reprueban sus materias con ello poder obtener patrones de comportamiento para proponer mejoras en las técnicas de aprendizaje que se están implementando en la institución educativa.

1.7.2. Negocio.

Referente a los negocios se puede determinar que clientes van a ser rentables durante un determinado periodo de tiempo con la finalidad de enviar ofertas a las personas que es probable que sean rentables.

1.7.3. Hábitos de compra en supermercado.

Un supermercado podría incrementar sus ventas si analizara los productos que sus clientes compran en conjunto para poder organizar las ubicaciones de cada producto en el mismo.

1.7.4. Patrones de fuga.

En una empresa es importante poder detectar cuando un cliente está pensando en cambiarse a la competencia, gracias a la minería de datos se pueden analizar los patrones de comportamiento de dichos clientes para poder predecir cuándo sucederá el presente evento, y poder tomar medidas de prevención para atraer a los clientes.

1.7.5. Fraudes.

Se pueden analizar las operaciones fraudulentas o ilegales que se realizan ya sea por ejemplo con el fraude en el uso de tarjetas de crédito o de servicios de telefonía móvil, dichas operaciones suelen seguir patrones característicos que permiten conocer en cierto grado de probabilidad si está ocurriendo realmente un fraude para con ello tomar las medidas necesarias para prevenir dichos eventos mal intencionados.

1.7.6. Seguros.

Con la ayuda de la minería de datos se puede realizar un análisis para identificar los clientes que pueden contratar nuevas pólizas. Además de la identificación de clientes con comportamiento fraudulento.

1.7.7. Medicina.

Con la ayuda de la minería de datos se puede a través de los patrones de comportamiento de un paciente poder dar un diagnóstico de la enfermedad que padece, para con ello poder realizar un agente recomendador de las medicinas útiles para este tipo de pacientes d. Además se podría analizar los grupos de riesgo para distintas patologías.

1.8. Metodología para proyectos de minería de datos (crisp-dm).

CRISP–DM (*Cross Industry Standard Process for Data Mining*), es la guía de referencia creada en el año 1996 es en la actualidad una de las más utilizada para el desarrollo de proyectos de Minería de Datos en los ambientes académico e industriales, Chapman *et al.* (2000).

La metodología se describe en términos de un proceso jerárquico consistente en un grupo de tareas descritas en cuatro niveles de abstracción (de general a específico): [*ver Figura. 1.6*]

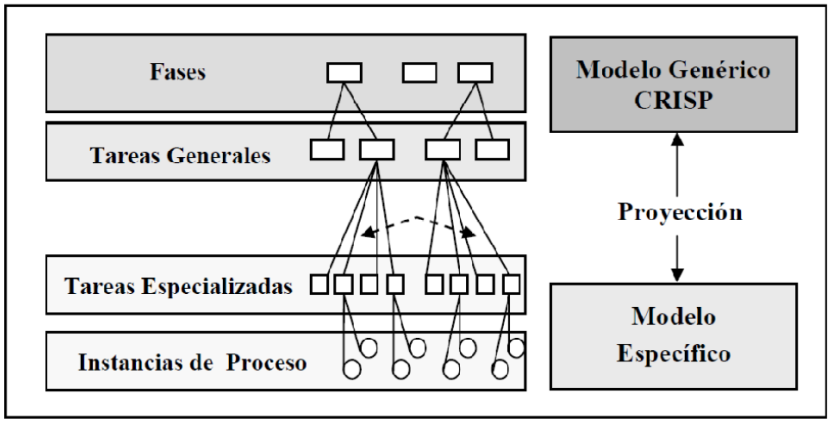

FIGURA 1. 6. Los 4 niveles del crisp–dm [chapman *et al.* (2000)].

La metodología CRISP-DM provee una representación completa del ciclo de vida de un proyecto de DM, que se divide en seis fases, sus tareas y relaciones entre ellas. [*ver Figura. 1.7*]

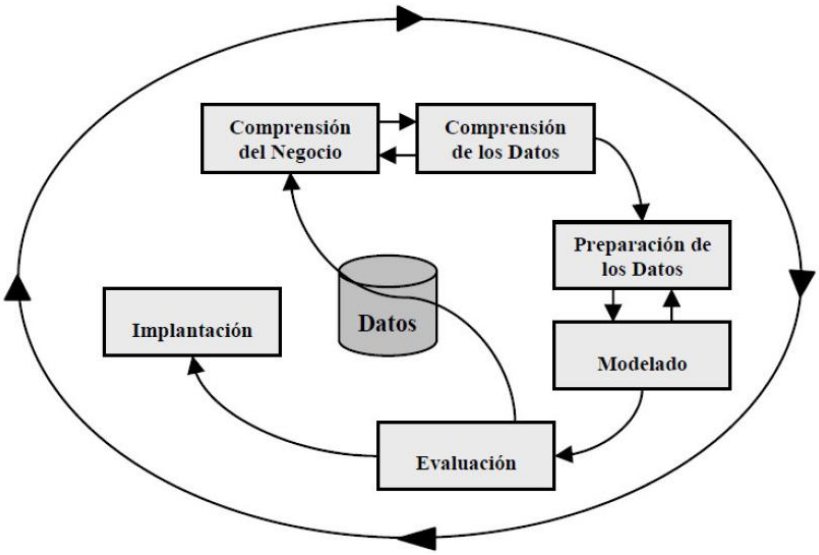

FIGURA 1. 7. Ciclo de vida de crisp–dm [chapman *et al.* (2000)].

Las Metodología CRISP–DM consta de 6 fases y cada una de ellas establecen tareas, las mismas que se describen a continuación: [*ver Figura. 1.8*]

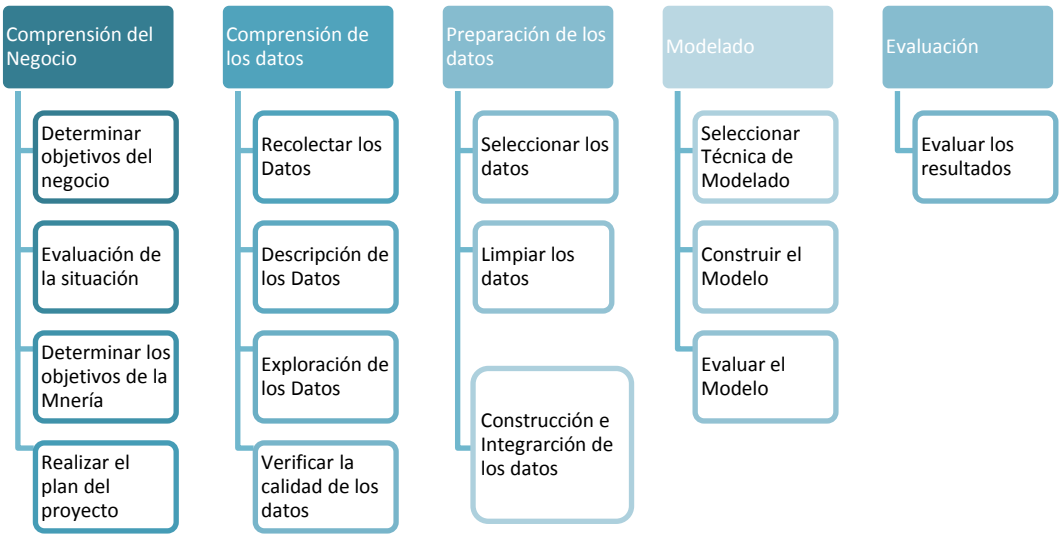

FIGURA 1. 8. Fases de crisp–dm [chapman *et al.* (2000)].

1.9. Descripción de fases de CRISP–DM, Chapman et al. (2000).

1.9.1. Comprensión del negocio.

En la presente fase se realiza un estudio completo del negocio para poder conocer el ámbito del problema y los objetivos del proyecto, lo que realmente se intenta resolver con el modelo.

1.9.2. Comprensión de los datos.

En esta fase se realiza una recolección inicial de los datos relacionados con el problema, se debe realizar un análisis de los datos con el fin de identificar problemas de calidad y detectar relaciones interesantes entre los mismos que permitan generar conocimiento sobre información oculta.

1.9.3. Preparación de los datos.

En esta fase se realiza una selección de los datos más relevantes para la creación de los dataset (conjunto de datos) a partir de los datos recopilados al inicio, a partir de los cuales se deberá realizar tareas de limpieza y transformación de tablas, registros y atributos para así tener información de calidad y poder ingresarlas en la herramienta de modelado.

1.9.4. Modelado.

En la presente fase se utilizará el conjunto de datos determinado, para procesarlo con la ayuda de una Herramienta de Minería de Datos que implemente las técnicas necesarias para la construcción del modelo.

1.9.5. Evaluación.

En la presente fase se debe comprobar la eficacia del modelo generado, evaluando si los resultados que devuelve son los correctos, con ello si está realizando las predicciones necesarias. Revisar los pasos realizados para la construcción del modelo y verificar que estos sean apropiados en función de los objetivos del negocio.

1.10. Proyectos relacionados.

1.10.1. Proyecto: Aplicando minería de datos al marketing educativo (Pinzón, 2011).

Descripción*:* El presente proyecto trata la importancia de la inteligencia de negocios y en especial una de sus técnicas, la minería de datos en el sector educativo, aplicado especialmente en los registros del estudiante, desde que ingresa en la universidad y las posibles causas de deserción en cada periodo académico, por medio de llamadas a los estudiantes que no volvieron a matricularse en la institución. Se presenta la caracterización del perfil del estudiante desertor de la Escuela de Marketing y Publicidad de la Universidad Sergio Arboleda de Colombia, bajo el estudio de variables demográficas del alumno con el registro de última matrícula del mismo semestre de abandono y las causas que lo generaron (Pinzón, 2011).

Para aplicar la técnica de minería de datos, se transformó la base de datos, incluyendo datos categóricos a numéricos, codificando cada uno de los metadatos. Se tuvieron en cuenta las siguientes variables:

- DNI: documento de identificación
- Semestre
- Ciudad de domicilio
- Ciudad de domicilio del acudiente
- Departamento de domicilio del acudiente
- Sexo o género
- Edad
- Estado civil
- País de nacimiento
- Estrato
- Medio por el cual se enteró del programa y de la universidad
- Idioma

Técnica de Minería de Datos Aplicada: Técnica de Agrupamiento bajo un método No-Jerárquico con el algoritmo K-means.

Se utilizó la presente técnica debido a que la investigación partía de extraer conocimiento, que en algunos casos la dirección de la escuela suponía, por lo cual se utilizó el método no supervisado; es decir que no se tiene variable objetivo, para primero tratar de comprender su base datos en busca de descubrir patrones y tendencias.

Herramientas utilizadas: Rapid Miner (anteriormente, YALE**):** RapidMiner ocupó el segundo lugar en herramientas de analítica y de minería de datos utilizadas para proyectos reales en 20091 y fue el primero en 2010.

1.10.2. Proyecto: Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil (Spositto, 2008).

Descripción: En el presente proyecto se realizó una evaluación del rendimiento académico y de la deserción estudiantil de los estudiantes del Departamento de Ingeniería e Investigaciones Tecnológicas (DIIT) de la Universidad Nacional de La Matanza (UNLaM). La investigación se realizó aplicando el proceso de descubrimiento de conocimiento sobre los datos de alumnos del período 2003-2008 (Spositto, 2008).

En el presente proyecto se tomaron en cuenta las siguientes variables:

- Datos del estudiante.
- Datos de las carreras del DIIT.
- Datos de los planes de estudio, vigentes y no vigentes, de las carreras.
- Datos de las materias de los planes de estudio.
- Datos de las notas, por carrera, plan de estudio y materia, de los estudiantes.
- Datos de los censos realizados a los estudiantes.

Técnica de Minería de Datos Aplicada: Se utilizaron los arboles de decisión implementando la tarea de clasificación bajo el algoritmo J48 (implementación en Weka del algoritmo C4.5) y el FT como algoritmos de minería.

Herramientas utilizadas: MS SQL Server para la generación de un almacén de datos, el software SPSS para realizar un preprocesamiento de los datos y el software Weka (Waikato Environment for Knowledge Analysis) para encontrar un clasificador del rendimiento académico y para detectar los patrones determinantes de la deserción estudiantil.

1.10.3. Proyecto: Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación (Pautsch, 2008).

Descripción: El objetivo principal del presente proyecto es maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos, que han desertado de la Carrera Analista en Sistemas de Computación de la Facultad de Ciencias Exactas, Químicas y Naturales de la Universidad Nacional de Misiones (Pautsch, 2008)..

Técnica de Minería de Datos Aplicada: Se utilizaron los arboles de decisión con el uso de la tarea de clasificación y además se utilizó la técnica de agrupamiento a través de la generación de clusteres.

Herramientas utilizadas: Se utilizó el software comercial para minería de datos *IBM DB2 Warehouse (versión 9.5).*

1.10.4. Proyecto: Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado (Domínguez, 2008)

Descripción: En el presente proyecto se elaboró un sistema de predicción para la detección de factores que influyen para el abandono escolar de alumnos que estudian en instituciones privadas de nivel superior. Se utilizó la minería de datos (CRISP-DM) y la lógica finalmente, el sistema de predicción se aplicó a una institución difusa, como técnicas de análisis. El sistema se aplicó en el Instituto de universitaria privada en un período comprendiendo un ciclo Estudios Superiores del Centro de Chiapas, México para corroborar su funcionamiento (Domínguez, 2008).

Técnica de Minería de Datos Aplicada: Reglas Difusas con la tarea de Clasificación.

Herramientas utilizadas: El sistema de inferencia difuso fue construido en Matlab.

Proyecto: Deserción universitaria. Un caso de estudio: variables que influyen y tiempo que demanda la toma de decisión (Vaira et. Al, 2010).

Descripción: El objetivo del presente trabajo es investigar cuándo es probable que ocurra un evento determinado como el de abandonar los estudios universitarios, el tiempo que lleva tomar la decisión y cuáles son las variables que más influyen en el cumplimiento de este evento. Han realizado algunos estudios sobre las causas que pueden influyen para que un estudiante tome la decisión de abandonar sus estudios universitarios que van desde: el abandono por la escasa formación previa, los reiterados fracasos en los exámenes finales, el origen social, la elección inadecuada de estudios, características familiares o circunstancias de la vida, problemas de organización de las diferentes unidades académicas, entre otras (Vaira et. Al, 2010).

Para el presente proyecto se tomaron en cuenta las siguientes variables, para determinar la deserción:

- Género.
- Tipo de escuela de la cual provenían los alumnos (pública/privada).
- Estudios alcanzados por los padres.
- Localidad de procedencia.
- Proporción de materias aprobadas según plan de estudio y carrera.
- Cantidad de atrasos totales registrados.
- Situación socio-ocupacional de los padres
- Nivel cultural de los padres
- Ingreso económico del hogar
- Tipo y tamaño del hogar.

CAPíTULO 2

ANÁLISIS DE LA PROBLEMÁTICA Y DISEÑO DE LA SOLUCIÒN

2.1. Análisis de la problemática.

La deserción de los estudiantes universitarios, es uno de los principales problemas que enfrentan las Universidades del Ecuador, ya que según investigaciones realizadas por la Senescyt (Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación), solo cuatro de cada 10 estudiantes que empiezan una carrera superior se gradúan (http://www.elcomercio.ec/sociedad/desercion-mayor-publica_0_624537721.html). La Unesco menciona que el presente problema involucra mucha pérdida de dinero, ya que señanala que (entre el 2000 y 2005) América Latina pierde hasta $420 millones por la deserción.

Cabrera *et al. (2006)* determinaron algunas categorías que se pueden entender como deserción estudiantil:

- Interrumpir la formación con la intención de retomarla en el futuro, ya sea por situaciones personales o económicas.

- Renunciar a los estudios universitarios definitivamente sin intención a retornarlos por situaciones personales.

- Dejar la carrera para iniciar otra en la misma institución ya que la que eligió en un inicio no está acorde a su perfil académico.

- Dejar la carrera para iniciar otra de su interés en una nueva institución.

- Dejar la universidad e irse a otra para completar estudios iniciados porque la universidad que escogió en un inicio no está acorde con su perfil, o no cumple con sus expectativas.

- Renunciar a la formación universitaria para incorporarse al mundo laboral.

Una publicación realizada por la Universidad Agraria del Ecuador (2012) señala que una considerable cantidad de alumnos abandonan sus estudios universitarios en el primer año de la carrera lo mismo que lo hacen porque no tuvieron la suficiente orientación e información antes de iniciar sus estudios de nivel superior y, como consecuencia, optaron por una titulación que no se ajustaba a sus expectativas. Por lo cual es importante que un estudiante razone adecuadamente la carrera que elegirá la misma que debe acoplarse a su perfil que incorpora algunas variables que pueden ser: la formación académica previa, las aptitudes, posibilidades económicas, disponibilidad de tiempo, motivación y vocación.

De algunos estudios realizados se ha podido determinar que mientras el estudiante mantenga un nivel académico adecuado, menor será la probabilidad de que el mismo decida abandonar la carrera, ya que si el estudiante obtiene notas insuficientes esté con mayor motivo decidirá en abandonar la carrera.

Además según los estudios realizados se determinaron que de los estudiantes que ingresan a primer año de universidad existen un porcentaje considerable de desertores ya sea porque la eligieron por tradición familiar, o por la influencia de los amigos.

El presente problema lo enfrentan varias instituciones universitarias a nivel nacional y mundial; por lo cual la Universidad Técnica Particular de Loja Modalidad Abierta y a Distancia (MAD), hace algunos años atrás ha creído conveniente desarrollar algunos estudios para determinar cuáles son las posibles causas por la que un estudiante decide abandonar sus estudios universitarios.

Dentro de los estudios realizados por los Docentes Investigadores de la UTPL (MAD) están la investigación realizada a "814 estudiantes que han abandonado sus estudios en el periodo abril – agosto 2008 de los centros universitarios de Quito, Guayaquil, Cuenca y otros del país, entre las razones principales de ese abandono, se pudo observar que el 31% lo hizo por dificultades económicas, el 25% por falta de tiempo para el estudio; el 15% por el sistema de estudio como: problemas académicos, de matriculación y de evaluación; y, el 10% por problemas familiares" (Arévalo & Maldonado, 2010).

Arévalo & Maldonado. (2010) mencionan además que un 21% de los estudiantes de primer ciclo decidieron abandonar la carrera de la Modalidad Abierta y a Distancia realizando un análisis del periodo Octubre 2009-Febrero 2010.

Arévalo & Maldonado. (2010) refiere además que a través del análisis de reportes del sistema de gestión académica se pudo determinar que entre las razones por las que los estudiantes de primer ciclo deciden abandonar temporalmente los estudios superiores están: el trabajo de los docentes principales y auxiliares; la calidad de material educativo (texto básico, guía didáctica, evaluaciones a distancia, evaluaciones presenciales); el nivel de interacción con los recursos tecnológicos; y, el sistema de tutorías telefónicas.

Ante este problema los Docentes Investigadores de la Universidad Técnica Particular de Loja de la Modalidad Abierta y a Distancia están trabajando sobre las posibles estrategias que se podrían adoptar para evitar que un estudiante decida abandonar sus estudios dichas estrategias se aplicaran cuando sucedan los siguientes eventos: cuando el estudiante no envié las evaluaciones a distancia en la fecha establecida, cuando el estudiante no apruebe una materia en el bimestre, y cuando el estudiante no se presente al supletorio de la determinada materia.

Para contribuir con la solución del problema de la deserción estudiantil se pretende: "*Crear un modelo predictivo que permita conocer cuáles son las posibles causas por lo que un alumno decide abandonar sus estudios universitarios, a través del análisis de las características de los estudiantes desertores de primer ciclo de la Modalidad Abierta y a Distancia de la UTPL".*

2.2. Diseño de la solución.

Para el presente proyecto se pretende aplicar técnicas de minería de datos para lograr un modelo predictivo, que permita predecir la deserción de los estudiantes de primer ciclo de la Modalidad Abierta y a distancia de la UTPL.

El modelo usará el conocimiento adquirido de los datos, que se vayan a extraer de las entidades, las mismas que han propuesto la información relacionada con la deserción de los estudiantes, y así tener como resultado un modelo útil, que ayude a resolver el problema planteado.

Para realizar el modelo de forma adecuada y ordenada, se utilizará la Metodología (CRISP-DM), detallada en el [*CAPÍTULO 1, sección 1.9*], la misma que es hoy en día, una de las más utilizadas en proyectos académicos e industriales.

Como parte del diseño de la solución del problema se planea seguir el siguiente diagrama, en el mismo que se detallan los elementos necesarios para la generación del modelo predictivo, como son: las variables que influyen en el problema para determinar la predicción; la herramienta de minería de datos que implementa las técnicas necesarias para analizar los datos de las variables establecidas para la predicción. [*ver Figura.2.1*]

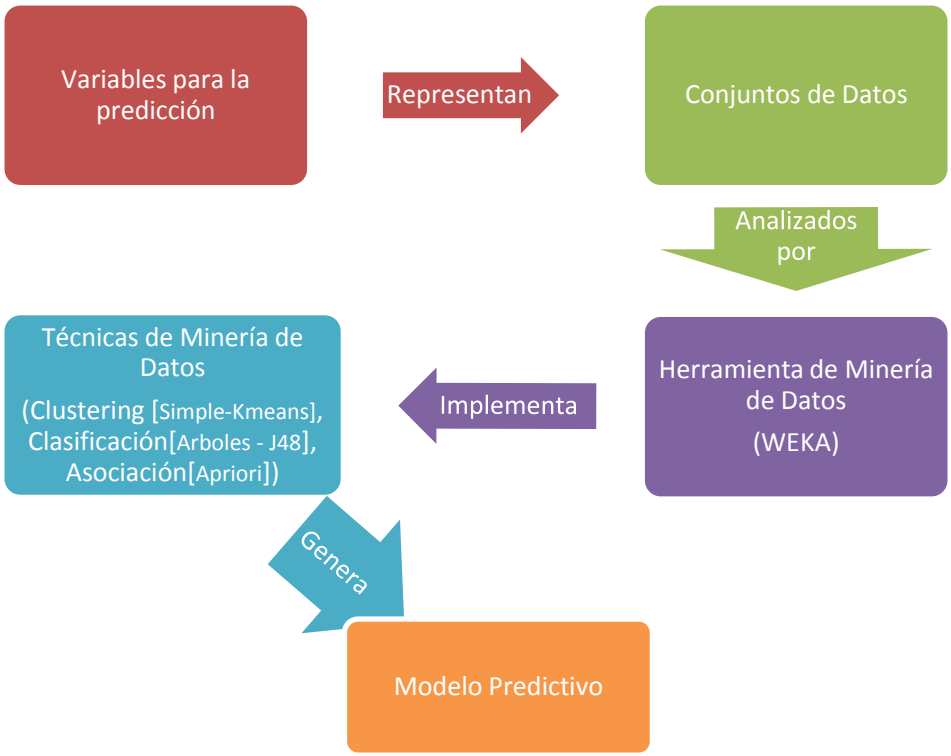

FIGURA 2. 1. Elementos para la generación del modelo predictivo

2.2.1. Variables para la predicción.

Una vez realizado el análisis del problema, se ha podido determinar algunas variables que están relacionadas, con las características tanto de los estudiantes, como en el entorno en el que se desenvuelven, para poder efectuar sus estudios universitarios; las mismas que se describen a continuación: [*ver Figura. 2.2*]

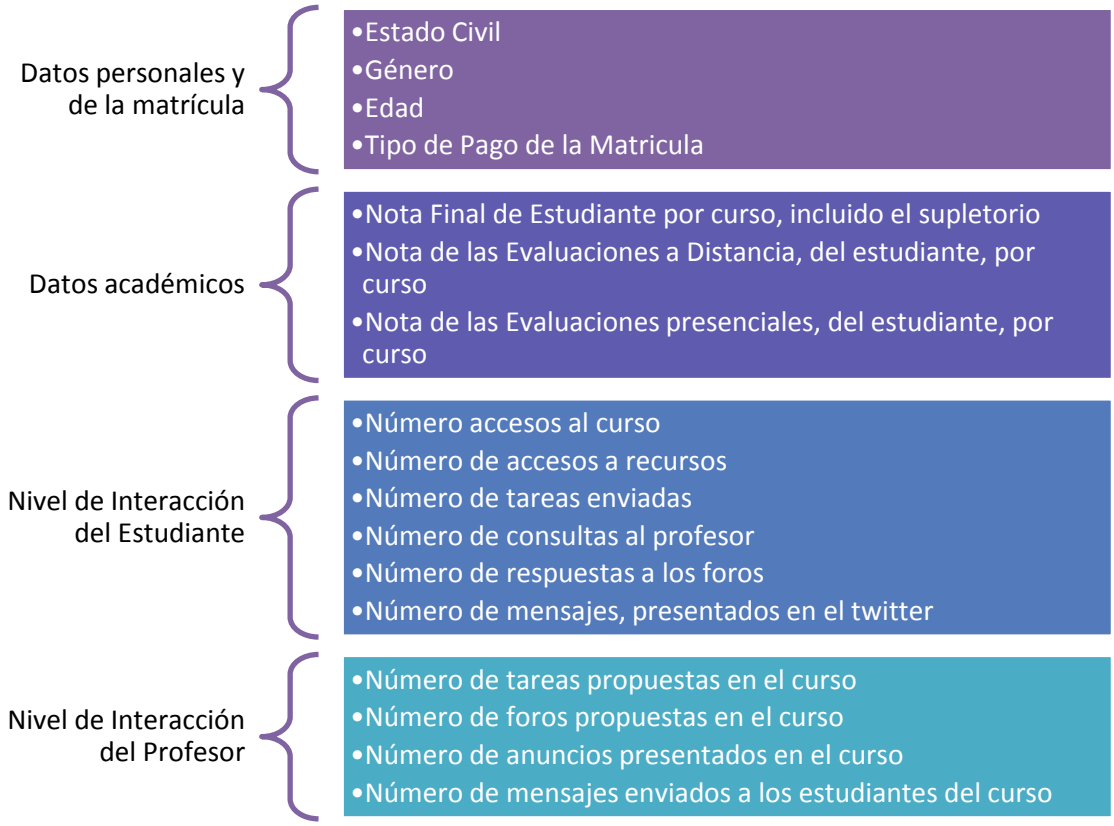

FIGURA 2. 2. Variables para la predicción

2.2.2. Herramienta de minería de datos a utilizar.

La herramienta que se utilizará para implementar la técnica de minería de datos será Weka.

WEKA (Waikato Environment for Knowledge Analysis) es una herramienta de libre distribución multiplataforma que permite la experimentación de análisis de datos mediante la aplicación, análisis y evaluación de las técnicas más relevantes de análisis de datos, principalmente las provenientes del aprendizaje automático, sobre cualquier conjunto de datos del usuario (García & Álvarez, 2008).

Se ha considerado la herramienta Weka por las siguientes razones:

- Está disponible libremente bajo la licencia pública general de GNU.

- Es portable porque está implementada en Java y puede correr en casi cualquier plataforma como Linux, Windows y Mac OS.

- Es una herramienta que es fácil de utilizar, ya que dispone de una interfaz gráfica de usuario.

- Weka tiene implementados varios algoritmos, con los cuales se pueden realizar tareas de minería de datos, permitiendo el preprocesamiento de datos, como clustering, clasificación, asociación, regresión, visualización.

- Permite realizar manipulaciones en los datos aplicando filtros ya sea a los atributos o a las instancias de los mismos, entre algunos de los filtros que nos permite aplicar Weka están los filtros “Remove” y “Discretize”, que eliminan atributos y discretizan atributos numéricos, respectivamente.

- La presente herramienta además proporciona acceso a bases de datos vía SQL gracias a la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que ya puede ser procesada con Weka.

- En otras herramientas de minería de datos, se puede trabajar con bases de datos o incluso, archivos Excel, o ficheros csv, de igual manera como lo hace Weka; sin embargo suele existir problemas, al querer cargar datos en los formatos descritos, ya que deben tener una estructura específica, para que las herramientas lo puedan reconocer; por lo cual Weka posee una ventaja adicional, ya que dispone de un formato propio .arff , por lo que se pueden crear los dataset necesarios de forma rápida y sencilla, en un bloc de notas.

- Weka además, proporciona información adicional, que ayudan a evaluar el nivel de efectividad del modelo obtenido.

2.2.3. Técnicas de minería de datos a utilizar.

Para obtener un modelo eficaz de predicción, se aplicaran algunas técnicas de minería de datos, de las cuales se escogerá la que ofrezca los mejores resultados, las presentes técnicas se describen a continuación:

2.2.3.1. Clustering.

Una de las técnicas que se utilizará será clustering bajo el algoritmo Simple-Kmeans.

Clustering es una técnica más de Aprendizaje Automático, en la que el aprendizaje realizado es no supervisado, es decir que no necesita introducir alguna clase para agrupar los datos.

Según el análisis de los trabajos relacionados, detallada en el [*Capitulo 1, sección 1.10*], la presente técnica es la más comúnmente utilizada para problemas de deserción estudiantil ya que permite obtener grupos de estudiantes según sus características.

El proceso de clustering consiste en la división de los datos en grupos de objetos similares. Para medir la similitud entre objetos se suelen utilizar diferentes formas de distancia: distancia euclídea, de Manhatan, de Mahalanobis, etc. El representar los datos por una serie de clusters, conlleva la pérdida de detalles, pero consigue la simplificación de los mismos, por ende una mayor comprensión de los datos.

2.2.3.2. Arboles de Decisión.

Una de las técnicas que se utilizará para poder predecir la deserción del estudiante será la técnica de árboles de decisión, implementando la tarea de clasificación bajo el algoritmo J48.

Se seleccionó la presente técnica ya que según el análisis realizado en el estado del arte se ha podido determinar, que los árboles de decisión mediante el uso de la tarea de clasificación es muy utilizada, para poder predecir cuándo un estudiante decide abandonar sus estudios; es una técnica muy útil puesto que brinda resultados precisos, además que es una de las más sencillas de entender he implementar.

El algoritmo J48 es uno de los más utilizados en minería de datos, en Weka es una implementación del algoritmo C4.5; permite trabajar con atributos tanto nominales, como numéricos.

2.2.3.3. Reglas de Asociación.

Dentro de las técnicas, que también se utilizaran serán las reglas de asociación, implementando, el algoritmo A priori en Weka.

Es una técnica fácil de implementar y comprender, permite generar todos los conjuntos válidos para poder crear parámetros que ayuden a resolver problemas determinados.

Representan una de las técnicas más valiosas e interesantes para la extracción de conocimiento oculto en grandes volúmenes de datos (bases de datos). Es una manera muy popular de expresar patrones de datos de una base de datos, los mismos que pueden servir para conocer el comportamiento general del problema que se requiere resolver, y poder así tener la información suficiente para la toma de decisiones.

Aunque los árboles de decisión pueden producir un conjunto de reglas, las reglas de asociación generan un conjunto de reglas independientes que no necesariamente formarán un árbol.

CAPÍTULO 3

DESARROLLO DEL PROYECTO

Para el desarrollo del presente proyecto de Tesis de pregrado se ha creído conveniente implementar la metodología CRISP-DM propuesta por Chapman *et al.* (2000) la misma que se detalla en el Estado del Arte [*Capítulo 1, sección 1.9*].

3.1. Fase I. Comprensión del negocio

3.1.1. Objetivos del negocio.

La Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja tiene actualmente los siguientes objetivos:

- Generar profesionales eficientes, a través de la ciencia para que sirvan a la sociedad.
- Desarrollar una universidad como alma máter para el siglo XXI.
- Desarrollo de una docencia pertinente y de alto nivel.
- Promover la Educación Superior a Distancia, en Ecuador y el Mundo.
- Promover la Investigación, el Desarrollo e Innovación en los estudiantes y docentes.
- Realizar proyectos de investigación científica y tecnológica vinculados con las diferentes carreras que existen en la institución educativa.

3.1.2. Evaluación de la situación.

Después de un análisis del problema descrito en el Capítulo 2.1: Análisis de la Problemática se ha podido evaluar la situación actual para el desarrollo del proyecto detallando los recursos, requerimientos, supuestos y restricciones que se muestran a continuación:

3.1.2.1. Recursos.

- *Personal* [*ver Tabla 3.1*]

TABLA 3. 1. Recursos – personal

Recurso	**Cargo**
Ing. Juan Carlos Torres	Experto del Negocio
Ing. Priscila Valdiviezo	Experto en Minería de Datos

- *Datos* [*ver Tabla 3.2*]

TABLA 3. 2. Recursos - datos

Recurso
- Información personal, académica y socioeconómica de los estudiantes de 1er Ciclo de las carreras que poseen la mayor cantidad de población en cada una de las Áreas de la MAD de la UTPL ofertadas en el período Octubre 2011- Febrero 2012.
- Información de las actividades que se realicen en los cursos de 1er ciclo de las carreras seleccionas de la MAD de la UTPL ofertadas en el período Octubre 2011- Febrero 2012.
- Lista de los Estudiantes matriculados en segundo ciclo de las carreras seleccionadas para el análisis de la MAD de la UTPL ofertadas en el período Abril 2012- Agosto 2012.

- *Hardware* [*ver Tabla 3.3*]

TABLA 3. 3. Recursos - hardware

Marca	**Disco**	**Memoria RAM**	**Procesador**
Hp	600 GB	8 Gb	Intel core i7

- *Software* [*ver Tabla 3.4*]

TABLA 3. 4. Recursos – software

Herramienta	**Descripción**
Microsoft Excel 2010	Herramienta utilizada para realizar los cálculos Matemáticos sobre los datos
Navicat Premium	Herramienta utilizada para realizar la manipulación y consultas a las bases de datos
XAMPP	Herramienta utilizada para levantar el servicio del Gestor de Bases de Datos Mysql.
Weka 3.6	Herramienta utilizada para implementar las técnicas de Minería de Datos necesarias.
Bloc de Notas	Herramienta que se utilizará para crear los DataSet en formato de los archivos de Weka .arff.

3.1.3. Requerimientos.

A continuación se detallan los requerimientos necesarios para el desarrollo exitoso del proyecto:

- Contar con la información necesaria y certera para la realización de un modelo de calidad.
- Aprender el uso de la Herramienta de minería de datos planteada para realizar el modelo.
- Contar con la participación continua de los expertos del negocio con la finalidad de evaluar los resultados obtenidos en cada fase.

3.1.4. Suposiciones

- Se supone que un estudiante pertenecerá a 1er ciclo si por lo menos está cursando una materia del mismo correspondiente a las troncales y de formación básica.

- Para determinar que un estudiante de 1er ciclo ha desertado la carrera se supone que este no está matriculado en el 2do ciclo del siguiente periodo de su respectiva carrera, es decir del periodo Abril 2012 – Agosto 2012.

- Se supone que un estudiante será posible desertor si no asiste a realizar la Evaluación Presencial del primer o la del segundo bimestre, además se lo considerará posible desertor si no envía la Evaluación a Distancia del primer o segundo bimestre.

- Para determinar cuál es el nivel de interacción del profesor en el curso, se ha tomado algunos atributos entre ellos las respuestas que ha proporcionado el profesor a los estudiantes por medio del envío de mensajes, como dicho campo no se encuentra establecido en la base de datos del EVA, se lo determino mediante los mensajes que han enviado los alumnos al profesor del curso, por lo que se supone que los mensajes que envía el alumno al profesor son preguntas, y los mensajes que envía el profesor al alumno son respuestas.

- Se supone que un curso puede tener más de un profesor titular, los cuales pueden realizar cualquier tipo de actividades en el mismo.

- Se supone que los alumnos y los profesores pueden enviar mensajes en el Entorno Virtual de Aprendizaje (EVA) encontrándose en cualquier curso, de los que le aparezcan en el entorno, o inclusive puede enviar mensajes localizándose fuera de los cursos es decir en la ventana principal del EVA.

- Si no existe nota en algunas de las evaluaciones efectuadas por el estudiante, se asume que dicho alumno no presento la evaluación, por lo tanto tendrá la puntuación de *0*.

3.1.5. Restricciones.

- Solo se dispone del campo tipo de pago de la matrícula del estudiante como información socioeconómica, ya que para el período “Octubre 2011- Febrero 2012”, con el que se está trabajando para el análisis de los datos, no cuenta con dicha información; estos datos hubiesen sido importante incluirlos en el modelo, para conocer si el nivel socioeconómico del estudiante interviene, dentro de las razones para que decida desertar la carrera.

- En las bases de datos recopilada existen valores faltantes en los campos de género, estado civil y tipo de pago de la matrícula del alumno los mismos que podría en un mínimo porcentaje afectar a la generación del modelo predictivo.

- Para la generación del modelo de minería se analizan dos bases de datos como son la del Entorno Virtual de Aprendizaje y la del Sistema Académico (Syllabus), las mismas que en comparación presentan una restricción ya que en la primera almacena códigos de identificación del curso según su paralelo, y los datos que maneja el Sistema Académico no son por paralelos, por lo tanto para la creación de los dataSet no se los realizará por paralelos, si no por cursos.

- Al momento de recopilar la información de la interacción del Profesor en el curso localizada en el EVA, se realizará un promedio de los datos recogidos, por cada paralelo del curso determinado según corresponda el caso; como se muestra en el [*Anexo 2 - B*].

- Existen datos faltantes en los códigos del profesor de algunos paralelos de los cursos del Entorno Virtual de Aprendizaje (EVA), lo que dificulta extraer la variable del Porcentaje de Respuesta que ha proporcionado el profesor del curso a sus estudiantes.

3.1.6. Terminología.

3.1.6.1. Terminología del Negocio.

- *Materias Troncales*.- Son materias que proporcionan los contenidos específicos y propios a través de los cuales se alcanzarán las competencias establecidas. [Utpl (2012)]

- *Materias de Formación Básica*.- Son materias con temáticas fundamentales que reflejan la dinámica de la universidad. [Utpl (2012)]

- *EVA*.- Entorno Virtual de Aprendizaje.

- *MAD*.- Modalidad Abierta y a Distancia.

- UTPL.- Universidad Técnica Particular de Loja.

3.1.7. Terminología de Minería de Datos.

- *Data Set (Almacén de Datos).*- Es un conjunto de datos históricos, internos o externos y descriptivos de un contexto o área de estudio, que están integrados y organizados de tal forma que permiten aplicar eficientemente herramientas para resumir, describir y analizar los datos con el fin de ayudar en la toma de decisiones estratégicas. [Hernández, Ramírez, & Ferri (*2004*)]

- *Int*.- Tipo de Dato que acepta valores numéricos, utilizados en las bases de datos.

- *Real*.- Tipo de Dato que valores numéricos, utilizados en los dataset .arff.

- *String.*- Tipo de Dato que acepta cadena de caracteres, comprendidos entre: letras, números, símbolos.

- *Nominal*.- Tipo de Dato que puede tomar un conjunto de valores especificados previamente, utilizados en los dataset.

- *Discretización*.- Es la conversión de un valor numérico en un valor nominal ordenado (que representa un intervalo).

3.1.8. Objetivos de la Minería.

- Encontrar patrones de comportamiento de los estudiantes desertores.

- Identificar grupos de estudiantes, según sus características comunes.

- Establecer que variables influyen, con mayor frecuencia para que un estudiante decida desertar la carrera en la MAD de la UTPL.

3.1.9. Plan de Trabajo.

El tiempo estimado para la realización del presente proyecto de tesis es de 8 meses, para lo cual se ha realizado un plan de trabajo en donde se detallan las tareas que se realizaran según la Metodología CRISP-DM adoptada para la generación del modelo, es importante tomar en cuenta que algunas de las siguientes actividades se las realizaba en paralelo: [*ver Tabla 3.5*]

TABLA 3. 5. Plan de trabajo

Fase	Tareas	Tiempo (semanas)
Comprensión del Negocio	Realizar Estado del Arte	2 semanas
	Realizar Análisis de la Problemática y Diseño de la Solución	2 semana
	Plantear Objetivos, Requerimientos, Supuestos Restricciones.	1 semana
Comprensión de los Datos	Recolección de datos iniciales.	4 semanas
	Descripción de los datos.	1 semana
	Exploración de datos.	2 semanas
	Verificación de la calidad de los datos.	1 semanas
Preparación de los Datos	Seleccionar los Datos	1 semana
	Limpiar los Datos	1 semana
	Estructurar los Datos	1 semana
	Integrar los Datos	1 semana
	Formateo de los Datos	1 semana
Modelado	Selección de la técnica de modelado	1 semana
	Construcción del Modelo.	5 semanas
Evaluación	Evaluar los Resultados	2 semanas

3.2. FASE II: Comprensión de los Datos.

Se realizó una recolección inicial de los datos relacionados con el problema, además se procedió a realizar un análisis de los mismos con el fin de identificar las relaciones entre ellos, y así generar conocimiento sobre alguna información oculta.

Los datos obtenidos corresponden a una muestra de estudiantes que cursan las cinco materias de primer ciclo de la modalidad abierta y a distancia de la UTPL, que son de formación básica y troncales. Estas materias corresponden a las carreras que poseen la mayor población de estudiantes de las cuatro áreas académicas de la UTPL: área técnica, administrativa, biológica, y humanística. La muestra corresponde al período académico Octubre 2012 – Febrero 2013, y se detalla a continuación:

TABLA 3. 6. Muestra poblacional

AREA	CARRERA	Materia	Número de estudiantes
ADMINISTRATIVA	ADMINISTRACIÓN DE EMPRESAS	ADMINISTRACIÓN I	**988**
		CONTABILIDAD GENERAL	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	
BIOLÓGICA	GESTIÓN AMBIENTAL	INTRODUCCION A LAS CIENCIAS AMBIENTALES	**714**
		BIOLOGIA GENERAL	

		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	
SOCIO HUMANÍSTICA	JURISPRUDENCIA	DERECHO CONSTITUCIONAL	**1304**
		INTRODUCCION AL DERECHO	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	
TÉCNICA	INFORMATICA	FUNDAMENTOS INFORMATICOS	**449**
		LOGICA DE LA PROGRAMACION	
		METODOLOGÍA DE ESTUDIO	

		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	

Cabe señalar que los estudiantes que conforman la muestra, para el proceso de aprendizaje cuentan con recursos como guías didácticas, libros base y acceso al entorno virtual de aprendizaje. En base a la lectura de los recursos disponibles para el estudiante se deben desarrollar dos evaluaciones a distancia, una por cada bimestre, que luego deben ser subidas al entorno virtual. Además los estudiantes deben rendir dos evaluaciones presenciales parciales, una por cada bimestre, y si no logran obtener el puntaje mínimo de aprobación (28/40) tienen opción a una tercera evaluación (supletoria).

En base a esto, las fuentes de información de dónde se extrajo conocimiento fueron las bases de datos internas de la institución de los siguientes sistemas: Entorno Virtual de Aprendizaje (EVA) y el Sistema Académico (Syllabus), los mismos que se describen a detallen en la siguientes sección.

3.2.1. Recolección de Datos.

Las fuentes de información útiles que se creyeron convenientes para extraer conocimiento de alta calidad fueron las bases de datos internas de la institución como son los datos del: Entorno Virtual de Aprendizaje (EVA) y del Sistema Académico (Syllabus); cuyas aplicaciones están basadas en el procesamiento tradicional de datos, que se conoce como procesamiento transaccional en línea (OLTP), las mismas que son suficientes para cubrir necesidades diarias dentro de la institución educativa y se describen a continuación:

- *Entorno Virtual de Aprendizaje (EVA)*.- Es un sistema web que ofrece diferentes servicios al estudiante como correo electrónico, interacción con los cursos matriculados de la carrera con lo cual se podrá obtener información sobre las actividades que realizan cada uno de los estudiantes de 1er ciclo que cursan las principales materias de las carreras seleccionadas para el análisis ofertadas en el periodo Octubre 2011/Febrero 2012 de la MAD de la UTPL.

- *Sistema Académico (Syllabus)*.- Sistema web que facilita los procesos académicos que realiza un estudiante como son: matricula, pago en línea, acceso al expediente estudiantil, acceso a notas. En el presente sistema se podrá obtener información sobre las notas, información personal y socioeconómica de los estudiantes.

3.2.1.1. Descripción de los datos

a. Descripción de los datos iniciales

Los datos inicialmente recopilados se encuentran almacenados en tablas relacionadas, y en Hojas de Cálculo en Microsoft Excel.

A continuación se realiza la descripción de la información recopilada en el inicio de las bases de datos de los Sistemas antes mencionados:

- *Tablas del Entorno Virtual de Aprendizaje (EVA)*
  - *Tabla: mdl_user_utpl*.- Registra los datos de los usuarios (alumnos y profesores). La presente tabla se la puede observar en el [*Anexo 4-A*].
  - *Tabla: mdl_enrol_utpl*.- Registra el número de rol que desempeñan los usuarios (5 = alumno, 3= profesor). La presente tabla se la puede observar en el [*Anexo 4-B*].
  - *Tabla: mdl_course_utpl.*- Registra información referente a los cursos, es aquí donde se guarda toda la información, que le fue suministrada al crearlo. La presente tabla se la puede observar en el [*Anexo 4-C*].
  - *Tabla: mdl_course_sections*.- Registra las secciones o módulos de un determinado curso, es decir si un curso tiene 4 secciones esta tabla contendrá un código para cada sección con el id del curso al que pertenecen dichas secciones, es aquí donde se registra toda la información correspondiente a esa sección (datos de esa sección). La presente tabla se la puede observar en el [*Anexo 4-D*].
  - *Tabla: mdl_assignment*.- Registra los datos de las tareas, toda la información que se le suministra al momento de crearla y además guarda el id del curso con el que está relacionada o sea a que curso pertenece dicha tarea. La presente tabla se la puede observar en el [*Anexo 4-E*].

- *Tabla: mdl_forum*.- Registra información sobre los foros propuestos en un determinado curso. La presente tabla se la puede observar en el [*Anexo 4-F*].

- *Tabla: mdl_message*.- Registran todos los mensajes que son enviados desde el estudiante al profesor y viceversa los mismos que no han sido leídos. La presente tabla se la puede observar en el [*Anexo 4-G*].

- *Tabla: mdl_message_read*.- Registran todos los mensajes que son enviados desde el estudiante al profesor y viceversa los mismos que han sido leídos. La presente tabla se la puede observar en el [*Anexo 4-H*].

- *Tabla: mdl_message_answered*.- Registran todos los identificadores de los mensajes creados con su respectivo estado (0= no leído, 1= leído, 2=respondido), la presente tabla está relacionada por medio del messageid con las tablas: mdl_message_read y mdl_message. La presente tabla se la puede observar en el [*Anexo 4-I*].

- *Tabla: mdl_periodo_utpl*.-Registran los códigos de los periodos académicos desde el más antiguo hasta el actual. La presente tabla se la puede observar en el [*Anexo 4-J*].

En él [*Anexo 3 – A]*, se muestra la relación que existe entre las tablas descritas anteriormente del Entorno Virtual de Aprendizaje, por medio de un diagrama conceptual.

- *Tablas del Sistema Académico (Syllabus)*

La información inicial recolectada de la Base de Datos del Sistema Académico se encuentra almacenada en hojas de cálculo de Microsoft Excel, la misma que se detalla a continuación:

- *Tabla: Notas Esquema 1*.- Se encuentra la información relacionada con los datos: personales, socioeconómicos y académicos de todos los estudiantes matriculados en las carreras seleccionadas del periodo Octubre/2011 - Febrero/2012. La presente tabla se la puede observar en el [*Anexo 5-A*].

- *Tabla: Identificaciones Abril2012 – Agosto2012*.- Se registran los números de cedulas de identidad de los estudiantes matriculados en 2do ciclo del periodo Abril2012 – Agosto2012. La presente tabla se la puede observar en el [*Anexo 5-B*].

- *Tabla: categorías_cursos*.- Se registran las categorías de las carreras con sus respectivos cursos ofertados. La presente tabla se la puede observar en el [*Anexo 5-C*].

En él *[Anexo 3 – B]* se muestra la relación que existe entre las tablas descritas anteriormente del Sistema Académico (Syllabus), por medio de un diagrama conceptual.

b. Descripción de los DataSet seleccionados

Los DataSet que se detallan en la presente sección contienen información referente a cada asignatura de 1er Ciclo del periodo Octubre 2011 – Febrero 2012 ofertadas en las carreras que poseen una mayor población en cada área de la MAD de la UTPL. [*ver Tabla 3.7*]

TABLA 3. 7. Asignaturas de 1er ciclo seleccionadas para el dataset

ÁREA	CARRERA	ASIGNATURA
ÁREA ADMINISTRATIVA	ADMINISTRACIÓN DE EMPRESAS UTPL-ECTS-1ª	ADMINISTRACIÓN I
		CONTABILIDAD GENERAL
		METODOLOGÍA DE ESTUDIO
		REALIDAD NACIONAL Y AMBIENTAL
		EXPRESIÓN ORAL Y ESCRITA
ÁREA BIOLÓGICA	GESTIÓN AMBIENTAL UTPL-ECTS -1ª	INTRODUCCION A LAS CIENCIAS AMBIENTALES
		BIOLOGIA GENERAL
		METODOLOGÍA DE ESTUDIO
		REALIDAD NACIONAL Y AMBIENTAL
		EXPRESIÓN ORAL Y ESCRITA
ÁREA SOCIO HUMANISTICA	JURISPRUDENCIA UTPL-ECTS-1ª	DERECHO CONSTITUCIONAL
		INTRODUCCION AL DERECHO
		METODOLOGÍA DE ESTUDIO
		REALIDAD NACIONAL Y AMBIENTAL

		EXPRESIÓN ORAL Y ESCRITA
ÁREA TÉCNICA	INFORMATICA UTPL-ECTS-1ª	FUNDAMENTOS INFORMATICOS
		LOGICA DE LA PROGRAMACION
		METODOLOGÍA DE ESTUDIO
		REALIDAD NACIONAL Y AMBIENTAL
		EXPRESIÓN ORAL Y ESCRITA

La información de los siguientes DataSet descritos se encuentra almacenada en ficheros .arff que es el formato oficial de la Herramienta Weka.

A continuación se detallan los atributos seleccionados para los DataSet (Almacenes de Datos) relacionados con la información de los alumnos de 1er ciclo del periodo Octubre 2011-Febrero 2012, los mismos que serán utilizados para la generación del modelo:

- *DataSet: Variables de Interacción del Profesor en el Curso*
  Contiene los datos referentes a la interacción que tiene el profesor dentro del curso relacionada con la propuesta de foros, tareas, contestación de preguntas a los estudiantes a través de la respuesta de los mensajes, y presentación de anuncios. Los campos del presente DataSet serán utilizados para discretizar los valores de los atributos por carreras y así generar un nuevo campo nominal llamado 'Nivel Interacción del Profesor' que permitirá valores (Alto, Medio, Bajo), el mismo que será utilizado en el Data Set: Deserción Estudiantil seleccionado para la generación del modelo.

  Los valores de los campos que se especifican a continuación se seleccionarán para aplicar el filtro no supervisado 'Discretize' de la herramienta Weka que será útil para convertir los campos de numérico a nominal. [*ver Tabla 3.8*]

TABLA 3. 8. Dataset: Variables de interacción del profesor en el curso.

Campo	Descripción	Tipo de Dato
NUM_TAREAS_PROPUESTAS	Número de Tareas propuestas por el profesor de la asignatura.	Real
NUM_FOROS_PROPUESTOS	Número de Tareas propuestas por el profesor de la asignatura.	Real
NUM_ANUNCIOS	Número de Anuncios presentados en el curso por parte del profesor de la asignatura.	Real
PORC_RESP_PROF_AL_EST	Porcentaje de respuesta que el profesor ha dado al estudiante, dicho campo se lo calcula, según el número de mensajes que el alumno envía al profesor de la asignatura, comparado con el número de mensaje que ha enviado el profesor a los estudiantes de la determinada asignatura.	Real

Al aplicar la discretización de los valores de los campos antes mencionados de la [*Tabla 3.8*] en la herramienta weka se obtuvieron algunos rangos numéricos clasificados como Bajo = 1, Medio = 2 y Alto = 3, detallados en el [*Anexo 3 - A*], luego los presentes valores fueron utilizados para obtener una sola variable del 'Nivel Interacción del Profesor' a través de la Moda, la misma que se explica en el [*Anexo 3 – B*].

- *DataSet Preliminar: Deserción Estudiantil*
  Contiene información referente a datos personales, socioeconómicos, académicos del estudiante, además que almacena información referente a la interacción que tiene el profesor y el alumno en el curso. [*ver Tabla 3.9*]

TABLA 3. 9. Data set deserción estudiantil

Campo	Descripción	Tipo de Dato
CURSO	Asignatura que está cursando el estudiante	Nominal
EDAD	Edad del estudiante	real
GENERO	Genero del Estudiante	nominal(Masculino, Femenino)
ESTADO_CIVIL	Estado Civil del Estudiante	nominal(Soltero, Casado)
TIPO_PAGO_MATRICULA	Tipo de pago de matrícula que ha efectuado el estudiante.	nominal (Contado, Crédito)
PRIM_EVAL_DISTANCIA	Nota de la Primera evaluación a distancia realizada por el estudiante.	Real
PRIM_EVAL_PRESENCIAL	Nota de la Primera evaluación presencial dada por el estudiante.	Real
NOTA_1BIM	Nota del Primer Bimestre tomando en cuenta las evaluaciones presentadas por el estudiante.	Real
SEG_EVAL_DISTANCIA	Nota de la Segunda evaluación a distancia realizada por el estudiante.	Real
SEG_EVAL_PRESENCIAL	Nota de la Segunda evaluación presencial dada por el estudiante.	Real
NOTA_2BIM	Nota del Segundo Bimestre tomando en cuenta las evaluaciones presentadas por el estudiante.	Real

NOTA_FINAL	Nota Final de la asignatura tomada por el estudiante tomando en cuenta las notas del supletorio según corresponda el caso.	Real
NOTA_SUP1	Nota obtenida por los estudiantes que han tenido que rendir el examen supletorio del 1er bimestre.	Real
NOTA_SUP2	Nota obtenida por los estudiantes que han tenido que rendir el examen supletorio del 2do bimestre.	Real
ESTADO_APROBACION	Estado de Aprobación de la asignatura que está cursando el estudiante tomando en cuenta el supletorio según corresponda el caso.	nominal (Aprobado, Reprobado)
NIVEL_INTER_EST	Nivel de interacción dentro del curso en todo el periodo académico de cada estudiante.	nominal (SI, NO)
NIVEL_INTER_PROF	Nivel de interacción dentro del curso que está impartiendo el profesor. [Anexo 3 - B]	nominal (Alto, Medio, Bajo)
DESERTOR	Se registra un 'SI' en el caso de que el alumno no se haya matriculado en el 2do Ciclo del siguiente periodo es decir del Abril 2012 – Agosto 2012; caso contrario se registrará 'NO'	nominal (SI, NO)

- *Decodificación de los campos*

Para decodificar los campos tanto para los numéricos y los de tipo string del dataset definido, se han establecido algunos conjuntos de valores que se detallan a las siguientes tablas: [*ver Tabla 3.10*]

TABLA 3. 10. Decodificación del campo curso

CARRERA	Ítem	ASIGNATURA
ADMINISTRACIÓN DE EMPRESAS UTPL-ECTS-1ª	AdministracionI	ADMINISTRACIÓN I
	ContabilidadG	CONTABILIDAD GENERAL
	MetodologiaE	METODOLOGÍA DE ESTUDIO
	RealidadN	REALIDAD NACIONAL Y AMBIENTAL
	ExpreO	EXPRESIÓN ORAL Y ESCRITA
GESTIÓN AMBIENTAL UTPL-ECTS -1ª	IntroduccionCA	INTRODUCCION A LAS CIENCIAS AMBIENTALES
	BiologiaG	BIOLOGIA GENERAL
	MetodologiaE	METODOLOGÍA DE ESTUDIO
	RealidadN	REALIDAD NACIONAL Y AMBIENTAL
	ExpreO	EXPRESIÓN ORAL Y ESCRITA
JURISPRUDENCIA UTPL-ECTS-1ª	DerechoC	DERECHO CONSTITUCIONAL
	IntroduccionD	INTRODUCCION AL DERECHO
	MetodologiaE	METODOLOGÍA DE ESTUDIO
	RealidadN	REALIDAD NACIONAL Y AMBIENTAL
	ExpreO	EXPRESIÓN ORAL Y ESCRITA

INFORMATICA UTPL-ECTS-1ª	FundamentosI	FUNDAMENTOS INFORMATICOS
	LogicaP	LOGICA DE LA PROGRAMACION
	MetodologiaE	METODOLOGÍA DE ESTUDIO
	RealidadN	REALIDAD NACIONAL Y AMBIENTAL
	ExpreO	EXPRESIÓN ORAL Y ESCRITA

La [*Tabla 3.11*.] del campo edad especifica algunos rangos de valores, los mismos que fueron determinados por medio, de la aplicación de distribución de frecuencias de la información recolectada.

TABLA 3. 11. Decodificación del campo edad

Campo	Ítem	Significado
Edad	16 – 26	16a26
Edad	27 – 37	27a37
Edad	38 – 48	38a48
Edad	49 – 59	49a59
Edad	60 – 70	60a70
Edad	71 -81	71a81

TABLA 3. 12. Decodificación del campo género

Campo	Ítem
Género	Femenino
Género	Masculino

TABLA 3. 13. Decodificación del campo estado civil

Campo	Ítem
Estado Civil	Soltero
Estado Civil	Casado

TABLA 3. 14. Decodificación del campo tipo de pago de matricula

Campo	Ítem
Tipo_Pago_Matricula	Contado
Tipo_Pago_Matricula	Crédito

Los rangos especificados en la siguiente tabla, han sido determinados por la dirección académica de la Universidad Técnica Particular de Loja. [*ver Tabla 3.15*]

TABLA 3. 15. Decodificación del campo nota final

ESCALA CUALITATIVA DE LA NOTA FINAL			
Campo	**Valores**	**Ítem**	**Descripción**
NotaFinal	40-39	A	Sobresaliente
NotaFinal	38-36	B	Notable
NotaFinal	35-33	C	Bien
NotaFinal	32-30	D	Satisfactorio
NotaFinal	29-28	E	Suficiente
NotaFinal	27-14	FX	Insuficiente
NotaFinal	13 o menos	F	Deficiente

TABLA 3. 16. Decodificación del campo nivel_inter_est.

Campo	Ítem
Nivel_inter_est	Alto
Nivel_inter_est	Bajo

TABLA 3. 17. Decodificación del campo nivel_inter_prof

Campo	Ítem
Nivel_inter_prof	Alto
Nivel_inter_prof	Bajo

TABLA 3. 18. Decodificación del campo present_todas_las_eval

Campo	Ítem
Present_todas_las_eval	SI
Present_todas_las_eval	NO

TABLA 3. 19. Decodificación del campo desertor

Campo	Ítem
Deserto	SI
Deserto	NO

3.2.1.2. Exploración de los datos.

Luego de establecer cuál será el dataset preliminar, que se utilizará para la generación del modelo se continúa con la exploración de los datos, la misma que contiene cada una de las variables seleccionadas en el almacén de datos, para con ello analizar la calidad que existe en los mismos, y además conocer la distribución que existen en los datos por cada variable. Tomando en cuenta que la exploración que ha realizado, se incluyeron los datos de todos los estudiantes matriculados, en las asignaturas de 1er ciclo, ofertadas en las 19 carreras de la MAD de la UTPL, y la generación del modelo predictivo se lo realizará, considerando, solamente 1 carrera por área, ya que según lo expresado por los expertos, y de lo analizado, las carreras de una determinada área poseen un comportamiento similar.

a. *Distribución según la Edad.-* La [*Figura.3.1*] se puede visualizar a través de una gráfica de pastel que de un total de 8289 estudiantes de 1er ciclo matriculados la mayoría figuran entre los 16 a 26 años.

TABLA 3. 20. Frecuencias de la edad

Edad	Estudiantes
16 – 26	4417
27 – 37	2720
38 – 48	863
49 – 59	181
60 – 61	19
71 – 81	87
Datos Erróneos	2
TOTAL	8289

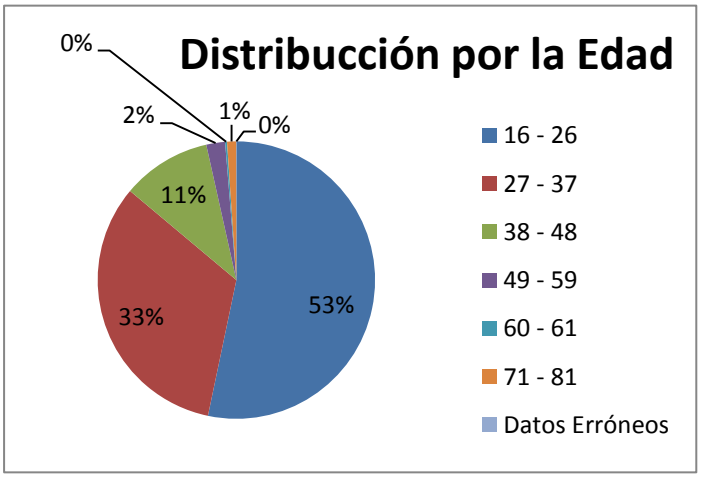

FIGURA 3. 1. Frecuencias del género

b. *Distribución según el Género.-* En la [*Figura.3.2*] se puede observar que de un total de 8289 estudiantes matriculados en 1er ciclo, la mayoría de la población pertenece al género femenino.

TABLA 3. 21. Frecuencias del género.

GENERO	ESTUDIANTES
F	4814
M	3466
Datos faltan-tes	9
TOTAL	8289

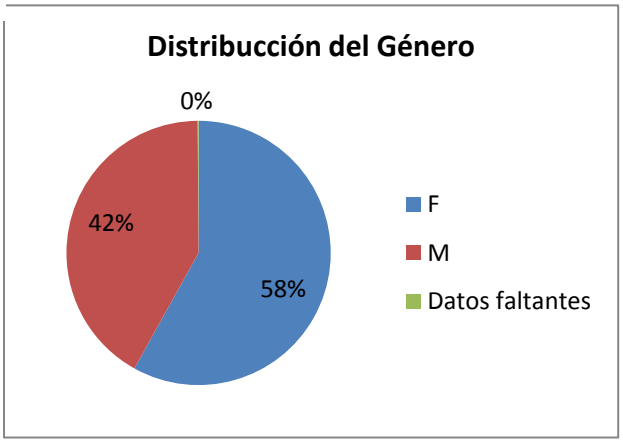

FIGURA 3. 2. Distribución por el género

c. *Distribución según el Estado Civil.-* En la [*Figura.3.3*] se puede observar que de un total de 8289 estudiantes de 1er ciclo matriculados, la mayoría de la población posee un estado civil 'soltero'.

TABLA 3. 22. Frecuencia del estado civil

ESTADO CIVIL	ESTUDIANTES
Soltero	5056
Casado	2119
Divorciado	18
Unión Libre	112
Viudo	12
Religioso	33
Otro	2
Datos faltantes	937
TOTAL	8289

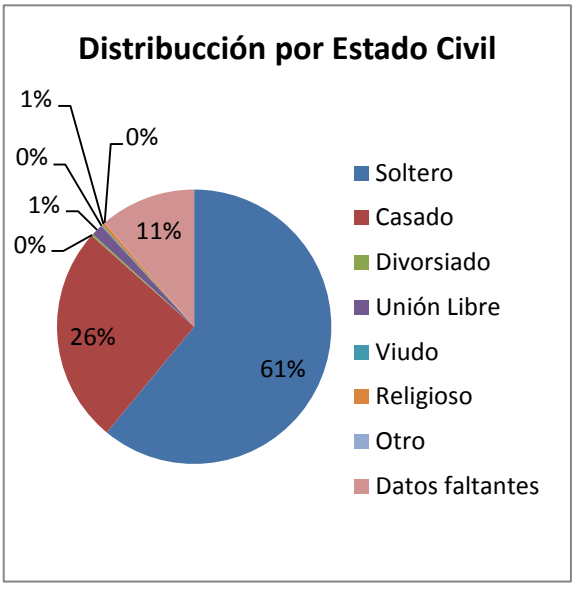

FIGURA 3. 3. Distribución por el estado civil

d. *Distribución según el Tipo de Pago de Matricula.-* En la [*Figura. 3.4*] se puede observar que de un total de 8289 estudiantes de 1er ciclo matriculados, en su mayoría han cancelado la matricula al contado.

TABLA 3. 23. Frecuencia del Tipo de pago de matricula

TIPO DE PAGO DE MATRICULA	ESTUDIANTES
CONTADO	6516
CREDCONVENIO	532
CU6040	1210
DEBITO_ROL	1
CUESPECIAL	16
Datos Faltantes	14
TOTAL	8289

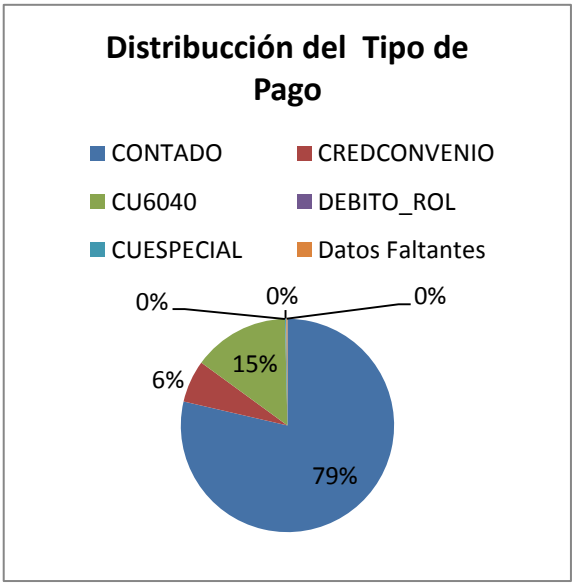

FIGURA 3. 4. Distribución del tipo de pago De matricula

e. *Distribución según el Estado de Aprobación-* En la [*Figura. 3.5*] se puede observar que de un total de 8289 estudiantes de 1er ciclo matriculados, la mayoría a reprobado en al menos una asignatura.

TABLA 3. 24. Frecuencia del tipo De pago de matricula

ESTADO DE APROBACIÓN	ESTUDIANTES
APROBADO	3652
REPROBADO	4618
ANULADO	15
Datos Faltantes	4
TOTAL	8289

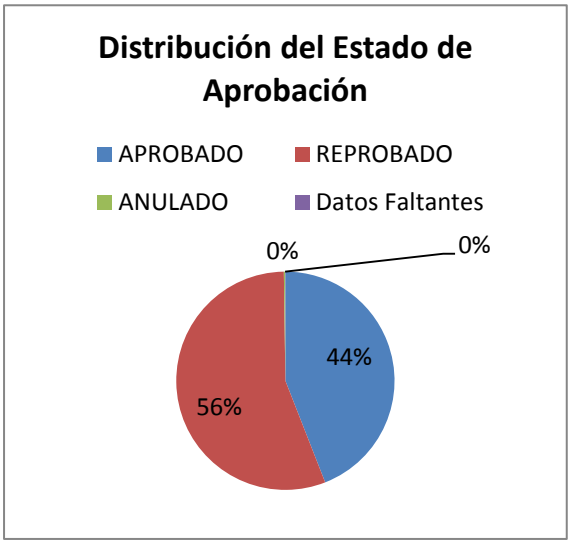

FIGURA 3. 5. Distribución del estado De aprobación

f. *Distribución de la Deserción por Carrera***-** En la [*Figura. 3.6.*] se puede observar en forma estadística el porcentaje de deserción por carrera, tomando en cuenta, que se considera un estudiante desertor, cuando este no se ha matriculado en el siguiente periodo, del que se está analizado actualmente. En la gráfica se visualiza que un total de 8289 estudiantes matriculados en 1er ciclo, la carrera que posee un mayor nivel de deserción estudiantil, es la carrera de INFORMÁTICA UTPL-ECTS-1A, seguida con una mínima diferencia, por la carrera de Ingeniería Administración de Empresas Turísticas y Hoteleras UTPL-ECTS-1A. En la [*Tabla 3.25]* muestra que la carrera de Jurisprudencia UTPL-ECTS-1A posee un mayor número de estudiantes matriculados.

TABLA 3. 25. Distribución de la deserción por carrera

CARRERA	NUMERO DE DESERTORES	POBLACIÓN	PORCENTAJE DE DESERCIÓN
Administración de Empresas UTPL-ECTS-1ª	798	**988**	80,77%
ADMINISTRACION EN BANCA Y FINANZAS UTPL-ECTS-1ª	220	**284**	77,46%
Administración en Gestión Pública UTPL-ECTS-1ª	118	**155**	76,13%
Asistencia Gerencial y Relaciones Publicas UTPL-ECTS-1ª	255	**333**	76,58%
CCEE-Educación Básica Ocho Ciclos UTPL-ECTS-1ª	415	**624**	66,51%
CCEE-Mención Físico Matemáticas Ocho Ciclos UTPL-ECTS-1ª	118	**159**	74,21%
CCEE-Mención Lengua y Literatura UTPL-ECTS-1ª	77	**119**	64,71%
CCEE-Mención Químico Biológicas UTPL-ECTS-1ª	72	**91**	79,12%
Ciencias Humanas y Religiosas UTPL-ECTS-1ª	27	**47**	57,45%

Comunicación Social UTPL-ECTS-1ª	242	**308**	78,57%
Economía UTPL-ECTS-1ª	155	**201**	77,11%
Educación Infantil UTPL-ECTS-1ª	156	**310**	50,32%
Gestión Ambiental UTPL-ECTS -1ª	541	**714**	75,77%
Informática UTPL-ECTS-1ª	366	**449**	81,51%
Ingeniería Administración de Empresas Turísticas y Hoteleras UTPL-ECTS-1ª	231	**285**	81,05%
Ingeniería en Contabilidad y Auditoría UTPL-ECTS-1ª	525	**731**	71,82%
INGLES UTPL-ECTS-1ª	238	**375**	63,47%
Jurisprudencia UTPL-ECTS-1ª	990	**1304**	75,92%
PSICOLOGIA UTPL-ECTS-1ª	624	**812**	76,85%
Deserción MAD UTPL	6168	**8289**	74,41%

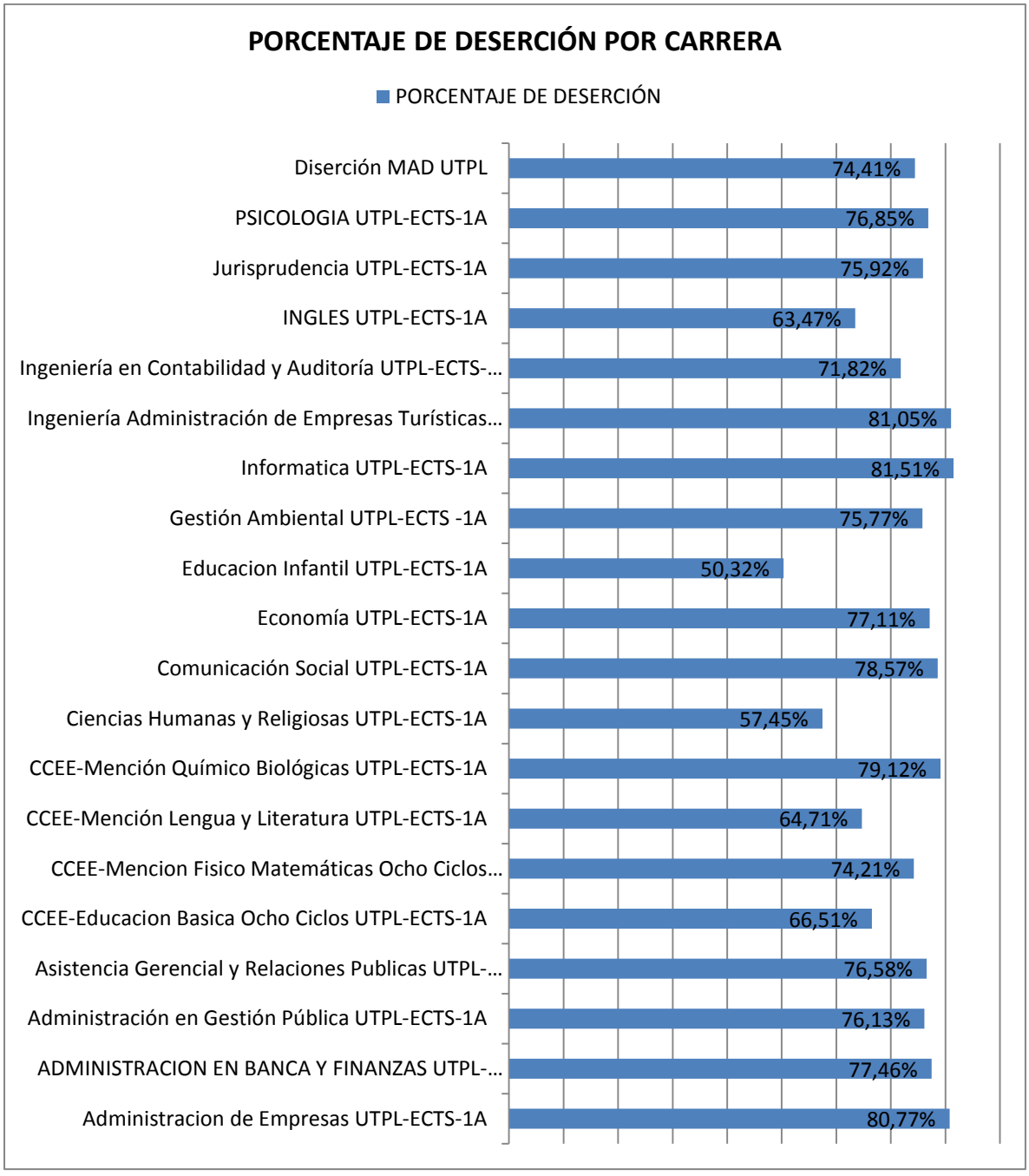

FIGURA 3. 6. Distribución de deserción por carreras

g. *Distribución del Rendimiento Académico por Áreas:* En la [*Figura.3.7*] se puede observar que todos los promedios de las áreas están dentro del rango cualitativo 'insuficiente' siendo el más alto el del Área Socio Humanística, con un total de 26 puntos sobre 40.

TABLA 3. 26. Frecuencias del rendimiento académico por áreas

ÁREAS	PROMEDIO
AREA ADMINISTRATIVA	23
AREA BIOLOGICA	24
AREA SOCIO HUMANISTICA	26
ÁREA TÉCNICA	23

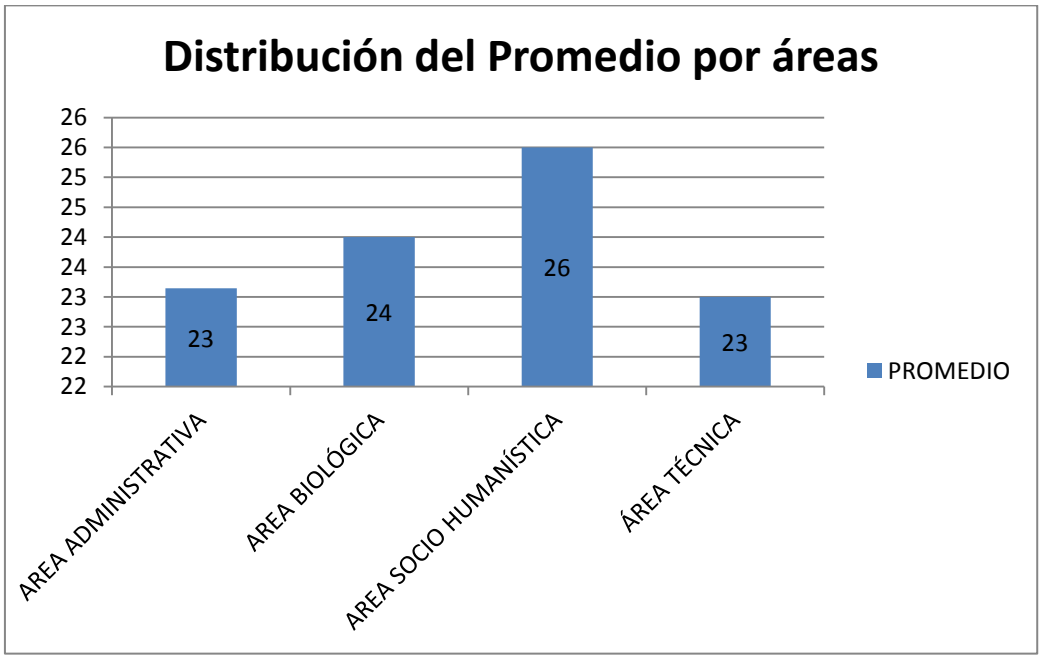

FIGURA 3. 7. Distribución rendimiento académico por áreas

- *Distribución de la Interacción del Profesor por Áreas*.- En la [*Figura.3.8*] se puede observar que el Área Administrativa posee un mayor nivel de interacción en tareas, foros y presentación de anuncios en el EVA.
  En la [*Figura.3.9*] se puede observar que el Área Administrativa y Técnica poseen un mayor nivel de interacción en el porcentaje de respuestas que ha dado el profesor a los estudiantes, a través del envío de mensajes en el EVA.

TABLA 3. 27. Frecuencias de la interacción del profesor

ÁREA	Número de Tareas Propuestas	Número de Foros Propuestos	Número de Anuncios Presentados	Respuestas del Profesor %
ADMINISTRATIVA	1	2	14	72%
BIOLÓGICA	1	1	12	63%
SOCIO HUMANÍSTICA	1	2	17	62%
TÉCNICA	3	3	19	71%

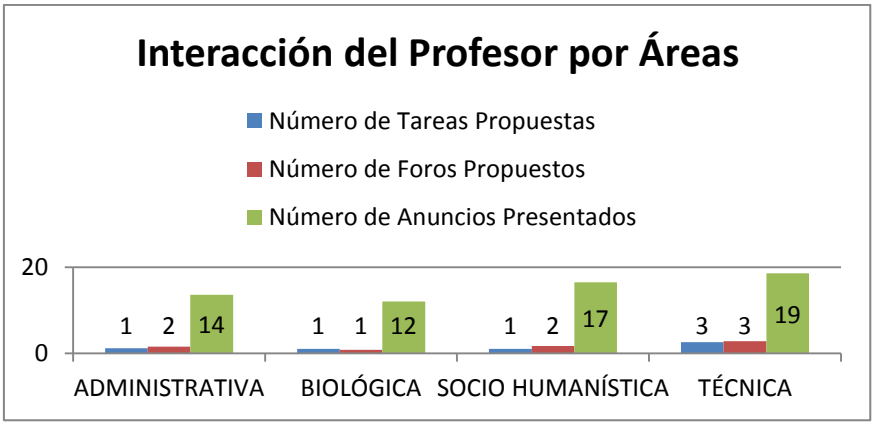

FIGURA 3. 8. Distribución de la interacción del profesor

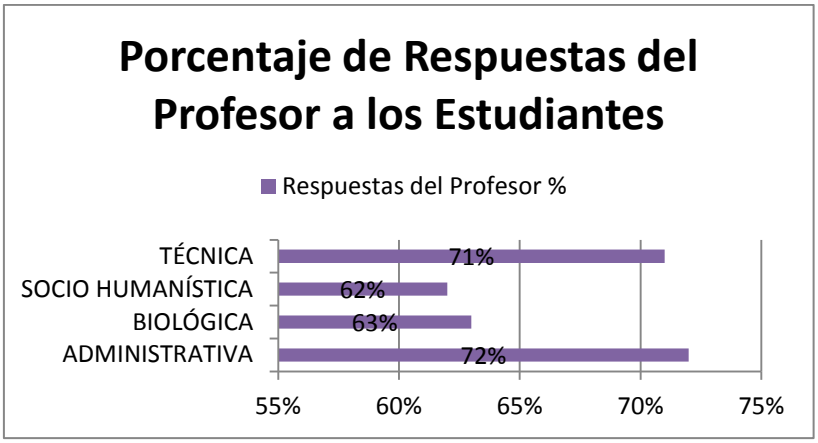

FIGURA 3. 9. Distribución de la interacción del profesor – respuestas

3.2.1.3. Verificación de la Calidad de los Datos

Luego de realizar la exploración de los datos seleccionados, se encontraron algunos campos que contienen valores con inconvenientes de calidad, como son:

- *GENERO*.- Contiene valores faltantes.

- *ESTADO_CIVIL*.- Contiene valores que pertenecen a un grupo poco frecuentes, como son: 'UNION LIBRE', 'DIVORSIADO', 'RELIGIOSO' y 'OTRO'.

- *TIPO_PAGO_MATRICULA*.- Contiene diferentes tipos de pago en el grupo de cancelaciones a crédito.

- *ESTADO_APROBACION*.- Contiene dentro del grupo valores como 'ANULADO', estado que se registra a los estudiantes que por algún motivo ha decido ya no seguir la asignatura, lo que significa que ya no consta como alumno de dicha materia.

A continuación se muestra en la [*Tabla 3.28*] un resumen de los campos que se han analizado en la exploración de los datos, de todas las carreras de la MAD-UTPL.

TABLA 3. 28. Resumen de atributos

Campo	Tipo de Dato	Total	Nulos y va-cíos	Media	Desv.e.	Moda	Min	Max
CURSO	nominal	95	0	-	-	-	-	-
EDAD	real	8289	2	26	9,8	20	16	81
GENERO	nominal	8289	9	-	-	-	-	-
ESTADO_CIVIL	nominal	8289	937	-	-	-	-	-
TIPO_PAGO_MATRICULA	nominal	8289	14	-	-	-	-	-
PRIM_EVAL_DISTANCIA	Real	8289	0	5,4	0,9	5,9	0	6
PRIM_EVAL_PRESENCIAL	Real	8289	0	8,2	2,8	9,4	0	6
NOTA_1BIM	Real	8289	0	13	3,91	14	0	20
SEG_EVAL_DISTANCIA	Real	8289	0	5,2	1,1	6	0	6
SEG_EVAL_PRESENCIAL	Real	8289	0	9,23	2,64	9,4	0	14
NOTA_2BIM	Real	8289	0	14	3,99	16	0	20
NOTA_FINAL	Real	8289	0	27	9,14	31	0	40
Nota_Sup1	Real	8289	0	10,85	4,78	12	0	20
Nota_Sup2	Real	8289	0	11,14	4,07	12	0	20
ESTADO_APROBACION	nominal	8289	0	-	-	-	-	-
NIVEL_INTER_ALUM	nominal	8289	0	-	-	-	-	-

NIVEL_INTER_PROF	nominal	8289	0	-	-	-	-	-
PRESENT_TODAS_LAS_EVAL	nominal	8289	0	-	-	-	-	-
SUPLETORIO	nominal	8289	0	-	-	-	-	-
ASISTIO_SUPLETORIO	nominal	8289	0	-	-	-	-	-
DESERTOR	nominal	8289	0	-	-	-	-	-

La calidad de los datos se analizará más detalladamente en la Fase 3 *[sección 3.1.3.2]*

3.2.2. FASE III: Preparación de Datos.

En la presente fase se construirá el DataSet definitivo, en el cual se procederá a ingresar los datos íntegros, sin errores y faltantes. Luego de construir los dataset correspondientes se procederá a ingresarlos a la herramienta Weka escogida para el modelado.

3.2.2.1. Seleccionar los Datos

En el data set construido, constan los estudiantes matriculados en 1er ciclo del periodo Octubre 2011 – Febrero 2012, de las carreras de Administración de Empresas (Área Administrativa), Jurisprudencia (Área Socio Humanística), Informática (Área Técnica), y Gestión Ambiental (Área Biológica), en el data set construido se le han aplicado diferentes filtros en Weka con la finalidad de obtener un dataset de calidad. Los campos que se han tomado en cuenta para la generación del modelo predictivo, se refiere a datos personales, información socioeconómica (tipo de pago), datos académicos y la participación que tiene el profesor como los estudiantes en el curso.

Los campos que se han considerado para establecer en el data set definitivo, para el presente estudio, de la deserción estudiantil son los siguientes:

CURSO.- Campo que contiene el nombre de la asignatura ofertadas en 1er ciclo, de las carreras de Administración de Empresas (Área Administrativa), Jurisprudencia (Área Socio Humanística), Informática (Área Técnica), y Gestión Ambiental (Área Biológica).

EDAD.- Campo nominal que contendrá los rangos de valores de las edades que se han especificado anteriormente, se han tomado en cuenta todas las edades de los estudiantes de las carreras seleccionadas, para el análisis.

GENERO.- Campo nominal que contendrá los valores de masculino y femenino.

ESTADO_CIVIL.- Campo nominal que contendrá los valores de soltero y casado.

TIPO_PAGO_MATRICULA.- Campo que contendrá los tipos de pago al contado y a crédito.

NOTA_FINAL.- Campo nominal que contendrá letras según la nota, los valores que se establecieron para cada letra, se especificaron anteriormente.

ESTADO_APROBACION.- Campo nominal que especifica el estado de aprobación ya sea Aprobado a Reprobado de la asignatura que ha cursado el estudiante.

NIVEL_ INTER_PROF.- Campo nominal que especifica el nivel de interacción que tiene el profesor en el curso, ya sea alto, medio o bajo.

NIVEL_INTER_EST.- Campo nominal que especifica el nivel de interacción que tiene el estudiante en el curso, ya sea alto, medio o bajo.

SUPLETORIO.- Campo nominal que especifica si el estudiante, debe rendir o no el examen supletorio, de la asignatura correspondiente. Para establecer el presente campo se analizaron las variables: NOTA_1BIM y NOTA_2BIM, por cada asignatura que ha cursado un respectivo estudiante. Estableciendo un SI, en el caso que el estudiante haya obtenido un puntaje menor a 14 por cada bimestre, caso contrario se establece un NO.

ASISTIO_SUPLETORIO.- Campo nominal que especifica si el estudiante, presenta o no la evaluación de supletorio, de la asignatura correspondiente. La presente variable se determinó analizando, las variables de: NOTA_SUP1, NOTA_SUP2 y SUPLETORIO. Estableciendo un SI en la presente variable, en el caso de que un estudiante teniendo que rendir el examen SUPLETORIO, si tiene calificación en las variables antes descritas, en caso contrario que dicho estudiante haya tenido que rendir el examen, y si por algún motivo no asistió se ha establecido un NO.

PRESENTARON_TODAS_LAS_EVALUACIONES.- Campo nominal que especifica si los estudiantes han presentado las evaluaciones de primer y segundo bimestre. La presente variable fue determinada, a través del análisis, de las variables: PRIM_EVAL_DISTANCIA, PRIM_EVAL_PRESENCIAL, SEG_EVAL_DISTANCIA, SEG_EVAL_PRESENCIAL. Estableciendo un SI, en el caso de que el estudiante tenga calificación en las 4 variables anteriormente descritas; caso que no haya presentado todas las evaluaciones se ha establecido un NO.

DESERTOR.- Campo nominal que especifica si el estudiante se ha matriculado en el 2do ciclo del siguiente periodo del que estamos analizando es decir 'Abril 2012 – Agosto 2012'.

3.2.2.2. Limpieza de los Datos

En la presente sección se ha realizado una limpieza de datos, con la finalidad de dar tratamiento a las inconsistencias encontradas, y así poder generar un modelo de calidad.

En la [*sección 3.1.2.4*], donde se realizó la verificación de la calidad de los datos, se encontraron algunas inconsistencias, por lo cual se requiere, establecer una solución para los registros que posee valores erróneos y faltantes. A continuación se detalla la limpieza de los datos que ha realizado:

El atributo ***GENERO***, posee valores faltantes, lo que se realizó, es reemplazar los valores faltantes, por el valor que tenga la mayor cantidad de instancias, en el presente caso, el género *Soltero* es el que se repite con mayor frecuencia, por lo tanto se reemplazaron los valores faltantes dicho valor.

El atributo ***ESTADO_CIVIL***, contiene valores que pertenecen a un grupo poco frecuentes, como son: 'UNION LIBRE', 'DIVORSIADO', 'RELIGIOSO' y 'OTRO', lo que se realizó es agrupar los valores en solo dos grupos, como son SOLTERO Y CASADO, por lo tanto los valores de Divorciado, Religioso y Otro, fueron reemplazados por 'Soltero', y los valores de Unión Libre fueron reemplazados por 'Casado'.

El atributo ***TIPO_PAGO_MATRICULA***, contiene diferentes tipos de pago en el grupo de cancelaciones a crédito, por tanto se procedió a agrupar los valores de CREDCONVENIO, CU6040, DEBITO_ROL, CUESPECIAL, reemplazando dichos valores por el tipo de pago a 'Crédito'.

El atributo ***ESTADO_APROBACION***.- contiene dentro del grupo valores como 'ANULADO', estado que se registra a los estudiantes que por algún motivo ha decido ya no seguir la asignatura, lo que significa que ya no consta como alumno de dicha materia; lo que se realizó fue eliminar los registros que poseen dicho valor. Hay que tomar en cuenta que dichos registros son pocos frecuentes en la información recolectada.

3.2.2.3. Construcción e Integración de los Datos

En la presente sección, se ha realizado la construcción e integración de los datasets definitivos, siendo estos creados, por cada curso de las carreras seleccionadas para el análisis. Luego se procede al ingreso de cada uno de los dataset en la herramienta Weka.

Una vez realizada la limpieza de los datos en la sección anterior, con los valores que debieron ser agrupados, reemplazados y eliminados, se procede a establecer el dataset definitivo, el mismo que se detalla a continuación: [*ver Tabla 3.29*]

TABLA 3. 29. Dataset definitivo

Campo	Descripción	Tipo de Dato
CURSO	Curso del Estudiante	Nominal
EDAD	Edad del Estudiante	Nominal
GENERO	Genero del Estudiante	Nominal
ESTADO_CIVIL	Estado Civil del Estudiante	Nominal
TIPO_PAGO_MATRICULA	Tipo de pago con la que ha cancelado la matricula el estudiante	Nominal
NOTA_FINAL	Nota final de estudiante	Nominal
ESTADO_APROBACION	Estado de Aprobación de la asignatura que ha cursado el estudiante	Nominal
NIVEL_ INTER_PROF	Nivel de Interacción del profesor en el curso	Nominal
NIVEL_INTER_EST	Nivel de interacción del estudiante en el curso	Nominal
PRESENTARON_TODAS_LAS _EVAL	Registra un estado de no en el caso que un estudiante no haya presentado al menos un de las 4 evaluaciones del curso.	Nominal (SI, NO)

SUPLETORIO	Registra 'SI', en el caso de que el estudiante, tenga que rendir el examen supletorio, caso contrario registra un 'NO'.	Nominal (SI, NO)
ASISTIO_SUPLETORIO	Registra un 'SI' en el caso de que el estudiante no se presente a emitir la respectiva evaluación supletoria, caso contrario se registrara un 'NO'	Nominal (SI, NO)
DESERTOR	Registra un 'no' si el estudiante no deserto la carrera caso contrario registra 'si'	Nominal (SI, NO)

3.2.3. FASE IV: Modelado

Una vez establecido el DataSet definitivo, libre de inconsistencias, se procede a, crear el modelo, seleccionando las técnicas de minería de datos, más apropiadas para el problema, las mismas que se han descrito en la [*sección 1.1 Diseño de la Solución*].

3.2.3.1. Seleccionar Técnica de Modelado

Las técnicas que se utilizaron para la creación del presente modelo, se eligieron en función de los siguientes criterios:

- Apropiada al problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

A continuación se establecen las técnicas utilizadas para el modelo, con la correspondiente tarea y algoritmo implementado, las mismas que se describen en el Diseño de la Solución [*capítulo 2, sección 2.3.1*].

TABLA 3. 30. Técnicas utilizadas para la generación del modelo

Técnica	Tarea	Algoritmo
Clusters	Agrupamiento - Clustering	Simple-Kmeans
Árboles de Decisión	Clasificación	J48
Reglas de Asociación	Asociación	A priori

3.2.3.2. Construcción del modelo y resultados experimentales

A continuación se procede a explicar los resultados de cada uno de los experimentos realizados, a través de las técnicas de agrupamiento (clustering), clasificación y asociación.

a. Agrupamiento (Clustering)

Para realizar el agrupamiento de los datos se ha utilizado el algoritmo SimpleK-*Means,* el mismo que es el más utilizado para este tipo de tareas*,* otras ventajas que ofrece la presente técnica, se encuentran detalladas en el Diseño de la Solución [*capítulo 2, sección 2.3.1*].

Al momento de aplicar el algoritmo SimpleK-*Means*, se realiza algunas configuraciones, de los valores que están establecidos por defecto en la herramienta Weka; en el primer parámetro, llamado *displayStdDevs* se especifica el valor de *True*(verdadero) con el fin de mostrar la desviación estándar de los atributos numéricos y contar los atributos nominales del cluster, el parámetro de distanceFunción es para establecer que método de los disponibles será utilizado para calcular las distancias entre los valores, en el mismo se deja la opción que viene por defecto *EuclideanDistance* que es una de las más utilizadas por su efectividad, en el parámetro dontReplaceMissingValues se deja la que viene por defecto que es *False*(Falso), en el parámetro del numCluster se establece el número de cluster que deseamos por lo tanto se ha colocado 3 para todos los dataset, se ha colocado dicho número porque existen tantas instancias en cada uno de los dataset de los cursos, y por último se elige el número de seed(semilla) para realizar una distribución aleatoria inicial a partir de la cual el algoritmo comienza las sucesivas iteraciones; el número elegido para llevar a cabo el proceso, se lo elige tomando en cuenta el valor, que muestra el menor error cuadrático. Al ser el

SimpleK-Means un algoritmo que se basa en minimizar la suma del error cuadrático, podría tomarse a ese error como punto de partida para la elección del número óptimo de la semilla.

Una vez realizada la conFigurauración pertinente, en los parámetros del algoritmo SimpleK-Means, se procedió a analizar los datos, por cada una de las asignaturas de 1er ciclo de las carreras seleccionadas. A continuación se muestran los resultados, obtenidos al momento de ejecutar el algoritmo de agrupamiento.

- *Resultados del Agrupamiento (Clustering)*

Se ha creído conveniente aplicar el algoritmo Simple K-Means, generando 3 clusters con los atributos de cada asignatura, las mismas que corresponden a 1er ciclo de la carrera de Jurisprudencia, a continuación se muestran los resultados obtenidos en Full Data, por cada asignatura, tomando en cuenta que dicho campo muestra el atributo que posee el mayor número de instancias: [*ver Tabla 3.31]*

- ***CARRERA: JURISPRUDENCIA***

TABLA 3. 31. Clusters generados – Carrera de Jurisprudencia

	Derecho Constitucional ***Seed = 10***	**Introducción al Derecho** ***Seed = 10***	**Metodología de Estudio** ***Seed = 10***	**Realidad Nacional** ***Seed = 10***	**Expresión Oral** ***Seed = 10***
Atributo	**Full Data** 1129 (100% de la población)	**Full Data** 1129 (100% de la población)	**Full Data** 1143 (100%)	**Full Data** 1093 (100% de la población)	**Full Data** 1128(100% de la población)
Edad	16a26 652 (57%)	16a26 652 (57%)	16a26 446 (39%)	16a26 440 (40%)	16a26 436 (38%)
Género	Masculino 652 (57%)	Masculino 652 (57%)	Masculino 666 (58%)	Masculino 630 (57%)	Masculino 654 (57%)
Estado Civil	Soltero 772 (68%)	Soltero 772 (68%)	Soltero 781 (68%)	Soltero 749 (68%)	Soltero 358 (31%)
Tipo de Pago de Matricula	Contado 886 (78%)	Contado 886 (78%)	Contado 888 (78%)	Contado 852 (77%)	Contado 884 (78%)
Nota Final	FX 486 (43%)	FX 486 (43%)	FX 312 (27%)	FX 345 (31%)	FX 392 (34%)

Estado de Aprobación	REPROBADO 753 (66%)	REPROBADO 753 (66%)	APROBADO 646 (56%)	REPROBADO 562 (51%)	REPROBADO 632 (56%)
Nivel Interacción del Profesor	Bajo 1129 (100%)	Bajo 1129 (100%)	Alto 1135 (99%)	Medio 1093 (100%)	Alto 1128 (100%)
Nivel de Interacción del Estudiante	Bajo 384 (34%)	Bajo 384 (34%)	Medio 392 (34%)	Medio 381 (34%)	Bajo 918 (81%)
Presento todas las Evaluaciones	SI 835 (73%)	SI 835 (73% del Cluster)	SI 820 (72%)	SI 798 (73%)	SI 814 (72%)
Supletorio	SI 911 (80%)	SI 911 (80% del Cluster)	SI 665 (58%)	SI 754 (68%)	SI 752 (66%)
Asistió al Supletorio	SI 574 (50%)	SI 574 (50% del Cluster)	NO_LE_CORRESPONDE 471 (41%)	SI 442 (40%)	SI 424 (37%)
Deserto	SI 841 (74%)	SI 841 (74% del Cluster)	SI 859 (75%)	SI 811 (74%)	SI 854 (75%)

En la *[Tabla 3.31],* podemos observar de manera general, que en todos los Full Data generados, predominan con un mayor número de instancias, la edad de 16 a 26 años, el género masculino, el estado civil soltero, el tipo de pago al contado. La mayoría de estudiantes que están cursando las materias de 1er ciclo, han reprobado con una nota de FX = 14 a 26 puntos sobre 40, a excepción de los estudiantes que han cursado Metodología de Estudio que en su mayoría han aprobado la misma. Se observa además que la mayoría de estudiantes que han cursado las presentes materias, han presentado todas las evaluaciones correspondientes.

Interpretación de los Clusters por curso: *Carrera de Jurisprudencia*

En base a la investigación realizada y la consulta con los expertos en el tema de la deserción estudiantil en la MAD de la UTPL, se procedió a realizar el análisis de los clusters de cada materia de la carrera de Jurisprudencia; con lo cual se pudo percibir que, hubiese sido importante analizar también la información socioeconómica del estudiante, además de conocer antecedentes académicos del estudiante, es decir saber la nota con la que obtuvo su Bachillerato y saber en qué colegio lo culmino; dicha información no fue posible analizarla, puesto que las bases de datos estudiada, no contaban con la misma.

A continuación se detalla el análisis realizado, en los clusters, los mismos que han sido generados por cada curso de 1er ciclo de la carrera de Jurisprudencia:

- ***Clusters de Derecho Constitucional***

Tamaño de la Población: 1129 instancias

Instancias de los Clusters:

Cluster 0 → 282 (25%)

Cluster 1 → 549 (49%)

Cluster 2 → 298 (26%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 3686.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                             (1129)        (282)         (549)         (298)
=============================================================================================
CURSO                      DerechoC      DerechoC      DerechoC      DerechoC
EDAD                         16a26         16a26         27a37         16a26
GENERO                    Masculino     Masculino     Masculino     Masculino
ESTADO_CIVIL                Soltero       Soltero       Soltero       Casado
TIPO_PAGO_MATRICULA         Contado       Contado       Contado       Contado
NOTA_FINAL                       FX             D            FX            FX
ESTADO_APROBACION        REPROBADO      APROBADO     REPROBADO     REPROBADO
NIVEL_INTER_PROF              Bajo          Bajo          Bajo          Bajo
NIVEL_INTER_EST               Bajo          Alto         Medio          Alto
PRESENT_TODAS_LAS_EVAL          SI            SI            SI            SI
SUPLETORIO                      SI            NO            SI            SI
ASISTIO_SUPLETORIO              SI NO_LE_CORRESPONDE        NO            SI
DESERTOR                        SI            NO            SI            SI



Time taken to build model (full training data) : 0.2 seconds
```

FIGURA 3. 10. Resultados – Simple k-means – Derecho constitucional – Jurisprudencia.

En la [*Figura 3.10*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la carrera de Jurisprudencia. Se puede visualizar que existe un total de 3 interacciones, y una suma de error cuadrático de 3686.0 para un semilla de 10, una vez realizada diferentes pruebas con varias semillas en los parámetros del algoritmo, se determinó que la semilla antes mencionada es la que presenta una menor cantidad en el error cuadrático, por lo tanto es una de las mejores distribuciones de datos que se puede obtener en los clusters. Hay que tomar en cuenta que el algoritmo, toma más tiempo para su ejecución al momento de que aumenta el número de interacciones, considerando que el tiempo de incremento es mínimo, comparado con los resultados propuestos para las otras asignaturas analizadas.

A continuación se detallan las características de los 3 clusters obtenidos, con una semilla de 10, del curso de Derecho Constitucional:

TABLA 3. 32. Resultados Simplek-Means – Derecho Constitucional – Jurisprudencia

Atributo	**Full Data 1129 (100% de la población)**	**Cluster0 282 (25% de la población)**	**Cluster1 549 (49% de la población)**	**Cluster2 298 (26% de la población)**
Edad	16a26 652 (57%)	16a26 102 (36% del Cluster)	27a37 271 (49% del Cluster)	16a26 151 (50% del Cluster)
Género	Masculino 652 (57%)	Masculino 157 (55% del Cluster)	Masculino 321 (58% del Cluster)	Masculino 174 (58% del Cluster)
Estado Civil	Soltero 772 (68%)	Soltero 180 (63% del Cluster)	Soltero 464 (84% del Cluster)	Casado 170 (57% del Cluster)
Tipo de Pago de Matricula	Contado 886 (78%)	Contado 230 (81% del Cluster)	Contado 421 (76% del Cluster)	Contado 235 (78% del Cluster)
Nota Final	FX 486 (43%)	D 149 (52% del Cluster)	FX 292 (53% del Cluster)	FX 194 (65% del Cluster)

Estado de Aprobación	Reprobado 753 (66%)	Aprobado 282 (100% del Cluster)	Reprobado 508 (92% del Cluster)	Reprobado 245 (82% del Cluster)
Nivel Interacción del Profesor	Bajo 1129 (100%)	Bajo 282 (100% del Cluster)	Bajo 549 (100% del Cluster)	Bajo 298 (100% del Cluster)
Nivel de Interacción del Estudiante	Bajo 384 (34%)	Alto 98 (34% del Cluster)	Medio 253 (46% del Cluster)	Alto 144 (48% del Cluster)
Presento todas las Evaluaciones	SI 835 (73%)	SI 280 (99% del Cluster)	SI 298 (54% del Cluster)	SI 257 (86% del Cluster)
Supletorio	SI 911 (80%)	NO 64 (22% del Cluster)	SI 549 (100% del Cluster)	SI 298 (100% del Cluster)
Asistió al Supletorio	SI 574 (50% del Cluster)	NO_LE_CORRESPONDE 218 (77% del Cluster)	NO 315 (57% del Cluster)	SI 276 (92% del Cluster)
Deserto	SI 841 (74% del Cluster)	NO 212 (75% del Cluster)	SI 32 (5% del Cluster)	SI 44 (14% del Cluster)

En la [*Tabla 3.32*], se puede visualizar, que la edad que predomina en el clusters 0 y 2 es de 16 a 26 años, siendo masculino el género que se repite con mayor frecuencia en los 3 clusters. El estado civil de soltero, prevalece en los clusters 0 y 1, a diferencia del cluster 2 que constan los estudiantes casados. En los 3 clusters, predomina el Tipo de Pago al Contado. En los clusters 1,2 predomina el estado de Reprobado, con una nota de FX=14 a 27 puntos sobre 40; a diferencia del cluster 0, que constan los estudiantes que Aprobaron con una nota D=30 a 32 puntos. El nivel de interacción Bajo del profesor es el que predomina en todos los clusters, tomando en cuenta que existe un solo docente quien dicta el curso; además se observa que predomina el nivel de interacción Alto del estudiante, en los clusters 0 y 1. Los estudiantes de todos los clusters SI han presentado todas las evaluaciones. En el cluster 0 los estudiantes no han tenido que rendir la evaluación supletoria, al contrario del cluster 1 y 2 que si han tenido que rendir la evaluación supletoria, tomando en cuenta que los estudian-

tes del cluster 2 SI asistieron a dar la evaluación supletoria, a diferencia del cluster 1 que NO asistieron. Los estudiantes que constan como desertores, y han cursado la materia de Derecho Constitucional, son los del cluster 1 y 2, a diferencia del cluster 0 que son los que han aprobado la asignatura y no constan como desertores.

Los estudiantes que han cursado la asignatura de Derecho constitucional, y constan como desertores, poseen las siguientes características comunes:

- Son estudiantes de género masculino, que han pagado la matricula al contado.
- Han reprobado la asignatura con una nota de FX= 14 a 27 puntos sobre 40,
- Han presentado todas las evaluaciones de la asignatura, tanto las presenciales como las a distancia.
- Dichos estudiantes han tenido una interacción considerable en el curso de Media y Alta, es decir que acceden con frecuencia al curso, realizan consultas al profesor por medio de los mensajes; sin embargo el profesor ha obtenido una interacción Baja, por lo tanto no propone tareas suficientes en el curso, no da contestación a todas las preguntas de los estudiantes.

Hay que tomar en cuenta que tanto el género del estudiante, el estado civil y el tipo de pago de matrícula no poseen demasiada influencia para que el estudiante decida desertar, ya que en los 3 clusters existen estudiantes solteros, de los cuales, también constan como no desertores en el clusters 0; además el tipo de pago al contado no influye como una variable socioeconómica para que el estudiante decida desertar por falta de pago de la matrícula, ya que la mayoría de estudiantes han pagado al contado.

- ***Clusters de Introducción al Derecho***

Tamaño de la Población: 1143 instancias

Instancias de los Clusters:

Cluster 0 → 257 (22%)

Cluster 1 → 595 (52%)

Cluster 2 → 291 (25%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 3503.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                 Full Data            0            1            2
                             (1143)       (257)        (595)        (291)
=============================================================================================
CURSO                  IntroduccionD  IntroduccionD  IntroduccionD  IntroduccionD
EDAD                           16a26          27a37          16a26          27a37
GENERO                     Masculino      Masculino       Femenino      Masculino
ESTADO_CIVIL                 Soltero        Soltero        Soltero          Casado
TIPO_PAGO_MATRICULA          Contado        Contado        Contado        Contado
NOTA_FINAL                        FX              D             FX             FX
ESTADO_APROBACION         REPROBADO       APROBADO      REPROBADO      REPROBADO
NIVEL_INTER_PROF               Alto           Alto           Alto           Alto
NIVEL_INTER_EST              Medio           Bajo           Alto          Medio
PRESENT_TODAS_LAS_EVAL          SI             SI             SI             SI
SUPLETORIO                      SI             NO             SI             SI
ASISTIO_SUPLETORIO              SI NO_LE_CORRESPONDE          SI             SI
DESERTOR                        SI             NO             SI             SI



Time taken to build model (full training data) : 0.05 seconds
```

FIGURA 3. 11. Resultados – Simple K-Means- – Introducción Al Derecho - Jurisprudencia

Referente a los resultados propuestos para la asignatura de Introducción al Derecho, ilustrada en la [*Figura 3.11*], se puede visualizar que existe un total de 4 interacciones, y una suma de error cuadrático de 3503.0 para un semilla de 10, dicha semilla es la que muestra uno de los menores errores cuadráticos.

A continuación se detallan las características de los 3 clusters, obtenidos con una semilla de 10, del curso de Introducción al Derecho:

TABLA 3. 33. Resultados simplek-means – introducción al derecho – jurisprudencia

Atributo	Full Data 1143 (100% de la población)	Cluster0 257 (22% de la población)	Cluster1 595 (52% de la población)	Cluster2 291 (25% de la población)
Edad	16a 26 449 (39%)	27a37 106 (41% del Cluster)	16a26 366 (61% del Cluster)	27a37 165 (56% del Cluster)
Género	Masculino 660 (57%)	Masculino 149 (57% del Cluster)	Femenino 342 (57% del Cluster)	Masculino 258 (88% del Cluster)
Estado Civil	Soltero 781 (68%)	Soltero 157 (61% del Cluster)	Soltero 524 (88% del Cluster)	Casado 100 (34% del Cluster)
Tipo de Pago de Matricula	Contado 889 (77%)	Contado 207 (80% del Cluster)	Contado 453 (76% del Cluster)	Contado 229 (78% del Cluster)
Nota Final	FX 546 (47%)	D 125 (48% del Cluster)	FX 364 (61% del Cluster)	FX 182 (62% del Cluster)
Estado de Aprobación	REPROBADO 811 (70%)	APROBADO 257 (100% del Cluster)	REPROBADO 549 (92% del Cluster)	REPROBADO 262 (90% del Cluster)
Nivel Interacción del Profesor	Alto 1143 (100%)	Alto 257 (100% del Cluster)	Alto 595 (100% del Cluster)	Alto 291 (100% del Cluster)
Nivel de Interacción del Estudiante	Medio 386 (33%)	Bajo 111 (43% del Cluster)	Alto 260 (43% del Cluster)	Medio 148 (50% del Cluster)

Presento todas las Evaluaciones	SI 832 (72%)	SI 255 (99% del Cluster)	SI 386 (64% del Cluster)	SI 191 (65% del Cluster)
Supletorio	SI 953 (83%)	NO 188 (73% del Cluster)	SI 593 (99% del Cluster)	SI 291 (100% del Cluster)
Asistió al Supletorio	SI 607 (53%)	NO_LE_CORRESPONDE 188 (73% del Cluster)	SI 349 (58% del Cluster)	SI 189 (64% del Cluster)
Deserto	SI 852 (74%)	NO 220 (85% del Cluster)	SI 549 (92%)	SI 266 (91%)

En la [*Tabla 3. 33*], se puede visualizar que la edad que predomina en el clusters 0 y 2 es de 27 a 37 años, siendo masculino el género que se repite con mayor frecuencia en ambos clusters. El estado civil de soltero, prevalece en los clusters 0 y 1, a diferencia del cluster 2 que constan los estudiantes casados. En los 3 clusters, predomina el Tipo de Pago al Contado. En los clusters 1 y 2 predomina el estado de Reprobado, con una nota de FX=14 a 27 puntos sobre 40; a diferencia del cluster 0, que constan los estudiantes que Aprobaron con una nota D=30 a 32 puntos. A diferencia de los clusters generados de la materia de Derecho Constitucional, en la presente materia, el nivel de interacción del profesor es Alto. En los clusters 0, 1 y 2 están predominando respectivamente el nivel Bajo, Alto, Medio en la interacción del estudiante. Los estudiantes de todos los clusters SI han presentado todas las evaluaciones. En el cluster 0 los estudiantes no han tenido que rendir la evaluación supletoria, al contrario del cluster 1 y 2 que si han tenido que rendir la evaluación supletoria, tomando en cuenta que los estudiantes del cluster 1 y 2 si asistieron a dar la respectiva evaluación supletoria. Los estudiantes que constan como desertores, y han cursado la materia de Introducción al Derecho, son los del cluster 1 y 2, a diferencia del cluster 0, que son los estudiantes que han aprobado la asignatura y no son desertores.

Los estudiantes que han cursado la asignatura de Introducción al Derecho, y constan como desertores, poseen las siguientes características comunes:

- Son estudiantes de género masculino, que han pagado la matricula al contado, y además han reprobado la asignatura.

- Los estudiantes de dicha asignatura SI han presentado todas las evaluaciones de la misma, sin embargo no han obtenido un buen puntaje, por tanto tuvieron que rendir un evaluación supletoria, y aunque asistiendo a dar la respectiva prueba, obtuvieron una nota de FX= 14 a 27 puntos sobre 40.

- Dichos estudiantes además han tenido una interacción considerable en el curso de Media y Alta, es decir que acceden con frecuencia al curso, realizan consultas al profesor por medio de los mensajes; y el profesor ha obtenido una interacción Alta, por lo tanto ha propuesto las suficientes tareas en el curso y ha dado contestación a la mayoría de las preguntas de los estudiantes.

Luego de analizar los clusters generados de la materia de Derecho Constitucional, y de la materia de Introducción al Derecho, se ha podido verificar que los estudiantes de las dos materias poseen comportamientos similares, siendo estas materias troncales de la carrera de Jurisprudencia. A diferencia que la edad que predomina en los clusters de Derecho Constitucional es de 16 a 26 años, y la edad predominante en los clusters de Introducción al Derecho es 27 a 37 años, siendo estas las edades que poseen la mayoría de estudiantes de 1er Ciclo de Jurisprudencia. Además se ha podido notar, en los clusters de ambas materias, que el nivel de interacción del profesor no ha influido en gran escala, para que un estudiante no repruebe la asignatura y por ende no deserte la carrera; ya que existen estudiantes que a pesar, de que el profesor ha obtenido una Alta interacción en el curso, los mismos han reprobado la materia, y son los que constan como desertores. Se ha podido constatar, que en los clusters generados en ambas materias troncales, los estudiantes que han cursado las mismas, y las han aprobado, no constan como desertores.

- ***Clusters de Metodología de Estudio***

Tamaño de la Población: 1143 instancias

Instancias de los Clusters:

Cluster 0 → 528 (46%)

Cluster 1 → 346 (30%)

Cluster 2 → 262 (23%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 3629.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                     Full Data          0          1          2
                                (1136)     (528)      (346)      (262)
=============================================================================================
CURSO                      MetodologiaE MetodologiaE MetodologiaE MetodologiaE
EDAD                              16a26      16a26      27a37      16a26
GENERO                        Masculino  Masculino  Masculino   Femenino
ESTADO_CIVIL                    Soltero    Soltero     Casado    Soltero
TIPO_PAGO_MATRICULA             Contado    Contado    Contado    Contado
NOTA_FINAL                           FX         FX          D          D
ESTADO_APROBACION             APROBADO   REPROBADO  APROBADO   APROBADO
NIVEL_INTER_PROF                  Alto       Alto       Alto       Alto
NIVEL_INTER_EST                 Medio      Medio       Bajo      Medio
PRESENT_TODAS_LAS_EVAL             SI         NO         SI         SI
SUPLETORIO                         SI         SI         NO         NO
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE         NO NO_LE_CORRESPONDE NO_LE_CORRESPONDE
DESERTOR                           SI         SI         SI         SI



Time taken to build model (full training data) : 0.05 seconds
```

FIGURA 3. 12. Resultados – Simple K-Means – Metodología De Estudio - Jurisprudencia

Referente a los resultados propuestos para la asignatura de Metodología de Estudio, ilustrada en la [*Figura 3.12*], se puede observar que existe un total de 4 interacciones, y una suma de error cuadrático de 3629.0 para un semilla de 10, dicha semilla es la que muestra una de las mejores distribuciones de los datos entre los clusters.

A continuación se detallan las características de los 3 clusters obtenidos con una semilla de 10, del curso de Metodología de Estudio:

TABLA 3. 34. Resultados Simplek-Means – Metodología De Estudio – Jurisprudencia

Atributo	**Full Data 1143 (100%)**	**Cluster0 257 (22% de la población)**	**Cluster1 595 (52% de la población)**	**Cluster2 291 (25% de la población)**
Edad	16a26 446 (39%)	16a26 238 (45% del cluster)	27a37 189 (54% del cluster)	16a26 163 (62% del cluster)
Género	Masculino 666 (58%)	Masculino 336 (63% del cluster)	Masculino 261 (75% del cluster)	Femenino 193 (73% del cluster)

Estado Civil	Soltero 781 (68%)	Soltero 402 (76% del cluster)	Casado 216 (62% del cluster)	Soltero 249 (95% del cluster)
Tipo de Pago de Matricula	Contado 888 (78%)	Contado 390 (73% del cluster)	Contado 284 (82% del cluster)	Contado 214 (81% del cluster)
Nota Final	FX 312 (27%)	FX 309 (58% del cluster)	D 168 (48% del cluster)	D 122 (46% del cluster)
Estado de Aprobación	APROBADO 646 (56%)	REPROBADO 486 (92% del cluster)	APROBADO 342 (98% del cluster)	APROBADO 262 (100% del cluster)
Nivel Interacción del Profesor	Alto 1135 (99%)	Alto 527 (99% del cluster)	Alto 346 (100% del cluster)	Alto 262 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 392 (34%)	Medio 198 (37% del cluster)	Bajo 166 (47% del cluster)	Medio 119 (45% del cluster)
Presento todas las Evaluaciones	SI 820 (72%)	NO 309 (58% del cluster)	SI 340 (98% del cluster)	SI 261 (99% del cluster)
Supletorio	SI 665 (58%)	SI 528 (100% del cluster)	NO 272 (78% del cluster)	NO 199 (75% del cluster)
Asistió al Supletorio	NO_LE_CORRESPONDE 471 (41%)	NO 322 (60% del cluster)	NO_LE_CORRESPONDE 272 (78% del cluster)	NO_LE_CORRESPONDE 199 (75% del cluster)

Deserto	SI 859 (75%)	SI 512 (96% del cluster)	SI 191 (55% del cluster)	SI 156 (59% del cluster)

En la [*Tabla 3.34*], se puede visualizar que la edad que predomina en los clusters 0 y 2 es de 16 a 26 años, siendo el género masculino y femenino que predominan respectivamente en ambos clusters, y son los estudiantes solteros que sobresalen en dichos clusters; discrepando dichos valores en el cluster 1, que es donde predomina la edad de 27 a 37 años, el género masculino, y el estado civil casado. El estado civil de soltero, prevalece en los clusters 0 y 2. En los 3 clusters, predomina el Tipo de Pago al Contado. En los clusters 1 y 2 predomina el estado de Aprobado, con una nota de D= 30 a 32 puntos sobre 40; a diferencia del cluster 0, que constan los estudiantes que Reprobaron con una nota de FX=14 a 27 puntos. A diferencia de los clusters generados de la materia de Derecho Constitucional, en los presentes clusters, el nivel de interacción del profesor en el presente curso es Alto. En los clusters 0, 1 y 2 están predominando respectivamente el nivel Medio, Bajo, Medio en la interacción del estudiante. Los estudiantes de los clusters 1 y 2 SI han presentado todas las evaluaciones, a diferencia del cluster 0 que predominan los estudiantes que no presentaron todas las evaluaciones. En el cluster 0 los estudiantes SI han tenido que rendir la evaluación supletoria, y predominan en el mismo los que no se presentaron a rendir la respectiva evaluación; al contrario del cluster 1 y 2 que NO han tenido que rendir la evaluación supletoria ya que han aprobado la asignatura directamente. En todos los 3 clusters generados predominan los estudiantes que han cursado Metodología de Estudio, y son desertores, a pesar que en los clusters 1 y 2, constan los estudiantes que aprobaron la asignatura.

Los estudiantes que han cursado la asignatura de Metodología de Estudio que forma parte de las asignaturas de formación básica en la carrera de Jurisprudencia, y constan como desertores, poseen las siguientes características comunes:

- Son estudiantes que en su mayoría son hombres, que han pagado la matricula al contado.

- Son estudiantes que si han Aprobado la asignatura, y en su mayoría SI han presentado todas las evaluaciones.

- Además son estudiantes que han tenido en su mayoría una interacción Media en el curso, es decir que si realiza actividades en el curso; y el profesor ha obtenido una interacción Alta, por lo tanto ha propuesto las suficientes tareas en el curso, ha dado respuesta a la mayoría de las preguntas que realizan los estudiantes.

Se puede observar en los clusters formados, que el nivel de interacción del estudiante no tiene una considerable influencia para que los alumnos puedan aprobar la asignatura, ya que en el cluster 1, a pesar que los estudiantes han obtenido un nivel de interacción bajo, de igual manera predominan en el mismo los estudiantes que han aprobado la asignatura, esto suceso podría ser porque dichos estudiantes, han empleado tiempo suficiente en estudiar el contenido de la materia, para con ello obtener buenas calificaciones en las evaluaciones tanto presencial como a distancia. Además existen estudiantes que han obtenido un nivel de interacción medio, sin embargo han reprobado la asignatura.

Luego de analizar los clusters formados de la materia de Derecho Constitucional (troncal), Introducción al Derecho (troncal), y ahora Metodología de Estudio (F. básica); se ha determinado que en la mayoría de clusters formados en las primeras 2 materias, predominan los estudiantes que constan como desertores y que SI han reprobado las respectivas materias. Al contrario de los clusters generados en la materia de Metodología de Estudio en donde predominan los estudiantes que SI aprobaron la asignatura, y SI desertaron la carrera, tomando en cuenta que la misma forma parte de las materias de Formación Básica de la carrera, siendo estas las que aprueban con mayor frecuencia los estudiantes de 1er ciclo, según lo analizado en experimentos que se detallan posteriormente.

- ***Clusters de Realidad Nacional***

Tamaño de la Población: 1093 instancias

Instancias de los clusters:

Cluster 0 → 442 (40%)

Cluster 1 → 409 (37%)

Cluster 2 → 242 (22%)

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 3426.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                             (1093)        (442)         (409)         (242)
=============================================================================================
CURSO                     RealidadN     RealidadN     RealidadN     RealidadN
EDAD                          16a26         27a37         16a26         16a26
GENERO                    Masculino     Masculino     Masculino     Masculino
ESTADO_CIVIL                Soltero       Soltero       Soltero       Soltero
TIPO_PAGO_MATRICULA         Contado       Contado       Contado       Contado
NOTA_FINAL                       FX             D            FX             F
ESTADO_APROBACION         REPROBADO     APROBADO      REPROBADO     REPROBADO
NIVEL_INTER_PROF             Medio         Medio         Medio         Medio
NIVEL_INTER_EST              Medio          Alto         Medio         Medio
PRESENT_TODAS_LAS_EVAL          SI            SI            SI            NO
SUPLETORIO                      SI            NO            SI            SI
ASISTIO_SUPLETORIO              SI NO_LE_CORRESPONDE        SI            NO
DESERTOR                        SI            SI            SI            SI



Time taken to build model (full training data) : 0.05 seconds
```

FIGURA 3.13. Resultados – Simple K-Means – Realidad Nacional - Jurisprudencia

Referente a los resultados propuestos para la asignatura de Realidad Nacional, ilustrada en la [*Figura 3.13*], se puede observar que existe un total de 5 interacciones, y una suma de error cuadrático de 3426.0 para un semilla de 10, siendo dicha semilla la que presenta una de las mejores distribuciones de los datos en los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con una semilla de 10, del curso de Realidad Nacional:

TABLA 3. 35. Resultados Simplek-Means - Realidad Nacional - Jurisprudencia

Atributo	**Full Data 1093 (100% de la población)**	**Cluster0 442 (40% de la población)**	**Cluster1 409 (37% de la población)**	**Cluster2 242 (22% de la población)**
Edad	16a26 440 (40%)	27a37 202 (45% del cluster)	16a26 228 (55% del cluster)	16a26 111 (45% del cluster)
Género	Masculino 630 (57%)	Masculino 269 (60% del cluster)	Masculino 218 (53% del cluster)	Masculino 143 (59% del cluster)

Estado Civil	Soltero 749 (68%)	Soltero 252 (57% del cluster)	Soltero 312 (76% del cluster)	Soltero 185 (76% del cluster)
Tipo de Pago de Matricula	Contado 852 (77%)	Contado 367 (83% del cluster)	Contado 316 (77% del cluster)	Contado 169 (69% del cluster)
Nota Final	FX 345 (31%)	D 210 (47% del cluster)	FX 293 (71% del cluster)	F 190 (78% del cluster)
Estado de Aprobación	REPROBADO 562 (51%)	APROBADO 437 (98% del cluster)	REPROBADO 315 (77% del cluster)	REPROBADO 242 (100% del cluster)
Nivel Interacción del Profesor	Medio 1093 (100%)	Medio 442 (100% del cluster)	Medio 409 (100% del cluster)	Medio 242 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 381 (34%)	Alto 165 (37% del cluster)	Medio 160 (39% del cluster)	Medio 89 (36% del cluster)
Presento todas las Evaluaciones	SI 798 (73%)	SI 434 (98% del cluster)	SI 362 (88% del cluster)	NO 2 (0% del cluster)
Supletorio	SI 754 (68%)	NO 339 (76% del cluster)	SI 409 (100% del cluster)	SI 242 (100% del cluster)
Asistió al Supletorio	SI 442 (40%)	NO_LE_CORRESPONDE 339 (76% del cluster)	SI 325 (79% del cluster)	NO 223 (92% del cluster)
Deserto	SI 811 (74%)	SI 223 (50% del cluster)	SI 347 (84% del cluster)	SI 241 (99% del cluster)

En la [*Tabla 3.35*], se puede visualizar que la edad que predomina en los clusters 1 y 2 es de 16 a 26 años, a diferencia del cluster 0, que predomina la edad de 27 a 37 años; siendo el género masculino y soltero los que predominan en todos los clusters. En los 3 clusters, predomina el Tipo de Pago al Contado. En los clusters 1 y 2 predomina el estado de Reprobado, a diferencia del cluster 0 que predominan los estudiantes que han Aprobado con una

nota de D=30 a 32 puntos. El nivel de interacción del profesor en el curso es Medio, a diferencia de materia de Derecho Constitucional, que mostraba un nivel de interacción Bajo. En los clusters 0, 1 y 2 están predominando respectivamente el nivel Alto, Medio, Medio en la interacción del estudiante. Los estudiantes de los clusters 0 y 1 SI han presentado todas las evaluaciones, a diferencia del cluster 2 que predominan los estudiantes que no presentaron todas las evaluaciones. En el cluster 1 y 2 los estudiantes SI han tenido que rendir la evaluación supletoria, sin embargo, en el cluster 2 predominan los que NO se presentaron a rendir la evaluación respectiva; a diferencia del cluster 1, que SI se presentaron. En el cluster 0 predominan los estudiantes que han aprobado la asignatura, directamente sin dar el respectivo supletorio. En todos los 3 clusters generados predominan los estudiantes que han cursado Realidad Nacional, y son desertores, a pesar que en el cluster 0, constan los estudiantes que aprobaron la asignatura.

Los clusters generados en la presente materia, poseen un comportamiento similar a los clusters generados de la materia de Metodología de Estudio, siendo ambas asignaturas las que corresponden a las de Formación Básica de la carrera de Jurisprudencia. Puesto que, en los clusters formados, de dichas materias, predominan los estudiantes que son desertores. En los clusters generados de ambas materias descritas, aproximadamente la mitad de los estudiantes aprueban, sin embargo predominan en los clusters los estudiantes desertores. Esto se debe, a que existen estudiantes, que a pesar de que estos pueden aprobar alguna de las asignaturas que pertenecen a las de formación básica, dichos estudiantes pueden reprobar en alguna de las asignaturas que forman parte de las troncales de la carrera, y es ahí donde se verifica, que la principal característica de un posible desertor es reprobar en alguna o ambas asignaturas que corresponden a las troncales de la carrera; dicho suceso se analizará en profundidad en los experimentos realizados con árboles de decisión, los mismos que se revisan posteriormente.

- ***Clusters de Expresión Oral***

Tamaño de la Población: 1128 instancias

Instancias de los Clusters:

Cluster 0 → 442 (39%)

Cluster 1 → 399 (35%)

Cluster 2 → 287 (25%)

```
kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 2935.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0                1                2
                             (1128)        (442)            (399)            (287)
=============================================================================================
CURSO                         Expre0       Expre0           Expre0           Expre0
EDAD                           16a26        27a37            16a26            16a26
GENERO                     Masculino     Femenino        Masculino        Masculino
ESTADO_CIVIL                 Soltero      Soltero          Soltero          Soltero
TIPO_PAGO_MATRICULA          Contado      Contado          Contado          Contado
NOTA_FINAL                        FX           FX                D                F
ESTADO_APROBACION         REPROBADO    REPROBADO        APROBADO        REPROBADO
NIVEL_INTER_PROF               Alto         Alto             Alto             Alto
NIVEL_INTER_EST                Bajo         Bajo             Bajo             Bajo
PRESENT_TODAS_LAS_EVAL           SI           SI               SI               NO
SUPLETORIO                       SI           SI               NO               SI
ASISTIO_SUPLETORIO               SI           SI NO_LE_CORRESPONDE               NO
DESERTOR                         SI           SI               NO               SI



Time taken to build model (full training data) : 0.05 seconds
```

FIGURA 3. 14. Resultados – Simple K-Means – Expresión Oral - Jurisprudencia

En la [*Figura 3.14*], se puede observar los resultados propuestos para la asignatura de Expresión Oral, en donde el algoritmo a retornado 5 interacciones, y una suma de error cuadrático de 2935 para un semilla de 10, siendo dicha semilla la que ofrece una de las mejores distribuciones de datos en los clusters, de igual manera como sucedió con las anteriores asignaturas de la carrera de Jurisprudencia.

A continuación se muestran los 3 clusters generados con sus respectivos porcentajes de población utilizada, para la asignatura de Expresión Oral de la carrera de Jurisprudencia:

TABLA 3. 36. Resultados Simplek-Means – Expresión Oral – Jurisprudencia

Atributo	**Full Data 1128(100%)**	**Cluster0 442 (39% de la población)**	**Cluster1 399 (35% de la población)**	**Cluster2 287 (25% de la población)**
Edad	16a26 436 (38%)	27a37 206 (46% del cluster)	16a26 142 (35% del cluster)	16a26 159 (55% del cluster)
Género	Masculino 654 (57%)	Femenino 238 (53% del cluster)	Masculino 241 (60% del cluster)	Masculino 209 (72%) del cluster
Estado Civil	Soltero 358 (31%)	Soltero 143 (32% del cluster)	Soltero 157 (39% del cluster)	Soltero 58 (20% del cluster)

Tipo de Pago de Matricula	Contado 884 (78%)	Contado 346 (78% del cluster)	Contado 335 (83% del cluster)	Contado 203 (70% del cluster)
Nota Final	FX 392 (34%)	FX 318 (71% del cluster)	D 171 (42% del cluster)	F 203 (70% del cluster)
Estado de Aprobación	REPROBADO 632 (56%)	REPROBADO 350 (79% del cluster)	APROBADO 399 (100% del cluster)	REPROBADO 282 (98% del cluster)
Nivel Interacción del Profesor	Alto 1128 (100%)	Alto 442 (100% del cluster)	Alto 399 (100% del cluster)	Alto 287 (100% del cluster)
Nivel de Interacción del Estudiante	Bajo 918 (81%)	Bajo 371 (83% del cluster)	Bajo 309 (77% del cluster)	Bajo 238 (82% del cluster)
Presento todas las Evaluaciones	SI 814 (72%)	SI 383 (86% del cluster)	SI 396 (99% del cluster)	NO 35 (12% del cluster)
Supletorio	SI 752 (66%)	SI 430 (97% del cluster)	NO 364 (91% del cluster)	SI 287 (100% del cluster)
Asistió al Supletorio	SI 424 (37%)	SI 352 (79% del cluster)	NO_LE_CORRESPONDE 364 (91% del cluster)	NO 250 (87% del cluster)
Deserto	SI 854 (75%)	SI 396 (89% del cluster)	NO 225 (56% del cluster)	SI 284 (98% del cluster)

En la [*Tabla 3.36*], se puede visualizar que la edad que predomina en los clusters 1 y 2 es de 16 a 26 años, a diferencia del cluster 0, que predomina la edad de 27 a 37 años; siendo el género masculino, los que predominan en los clusters 1 y 2. En los 3 clusters, predomina el tipo de pago al Contado y el estado civil Soltero. En los clusters 0 y 3 predomina el estado de Reprobado, a diferencia del cluster 1 que predominan los estudiantes que han Aprobado con una nota de D=30 a 32 puntos. El nivel de interacción del profesor en el curso es Alto, por ello es el nivel que sobresale en todos los cluster; y el nivel de interacción del estudiante que predomina es Bajo. Los estudiantes de los clusters 0 y 1 SI han presentado todas las evaluaciones, a diferencia del cluster 2 que predominan los estudiantes que no presentaron

todas las evaluaciones. En el cluster 0 y 2 los estudiantes SI han tenido que rendir la evaluación supletoria, sin embargo, en el cluster 0 predominan los que SI se presentaron a rendir la evaluación respectiva; a diferencia del cluster 2 que NO se presentaron. En el cluster 1 predominan los estudiantes que han aprobado la asignatura, directamente sin dar el respectivo supletorio. En los clusters 0 y 2 predominan los estudiantes que han cursado Expresión Oral, y son desertores, constando en dichos clusters los que reprobaron la asignatura.

Los clusters generados, en la materia de Expresión Oral, muestran un grupo de estudiantes que son desertores, los mismos que poseen las siguientes características:

- Son estudiantes con genero civil soltero, los mismos que han pagado la matricula al Contado.

- Dichos estudiantes han tenido un Nivel de Interacción Bajo en el curso, los cuales no han accedido un considerable número de veces al entorno virtual del curso, y el profesor de la asignatura ha tenido un nivel de interacción Alto en el curso, por tanto dicho docente ha brindado el apoyo respectivo a los estudiantes por medio del entorno virtual.

Existe un cierto grupo de estudiantes que si han aprobado la presente asignatura, y no constan como desertores, hay que tomar en cuenta que cierto grupo de estudiantes, podrían también haber aprobado otras materias de las anteriormente analizadas, tomando en cuenta que la presente asignatura corresponde a las de formación básica de la carrera de Jurisprudencia.

Se observa en los presentes clusters, lo que se analizó en los clusters de la materia de Metodología de Estudio, que el nivel de interacción del profesor no influye, en gran escala para que un estudiante aprueba la materia, ya que se visualiza en los clusters 0 y 2, que a pesar de que el profesor ha obtenido un nivel de interacción alta, los estudiantes de dichos clusters han reprobado, considerando además que los estudiantes han obtenido un nivel bajo de interacción en todos los clusters, sin embargo en el cluster 0, han aprobado la asignatura los estudiantes que igualmente han obtenido un nivel de interacción Bajo. Por tanto no necesariamente los estudiantes que tengan una Alta interacción en el curso, aprobaran la asignatura, sino que además depende de las buenas calificaciones que obtenga en las evaluaciones.

- ***CARRERA: ADMINISTRACIÓN DE EMPRESAS***

Se ha creído conveniente aplicar el algoritmo SimpleK-Means, generando 3 clusters con los atributos de cada asignatura, las mismas que corresponden a 1er ciclo de la carrera de Administración de Empresas, a continuación se muestran los resultados obtenidos en Full Data, por cada asignatura. [*ver Tabla 3.37]*

TABLA 3. 37. Resultados Del Clustering – Carrera De Administración De Empresas

	Administración I ***Seed = 10***	**Contabilidad General** ***Seed = 20***	**Metodología de Estudio** ***Seed = 20***	**Realidad Nacional** ***Seed = 20***	**Expresión Oral** ***Seed = 20***
Atributo	**Full Data** 901 (100% de la población)	**Full Data** 812 (100% de la población)	**Full Data** 851 (100% de la población)	**Full Data** 837 (100% de la población)	**Full Data** 842 (100%)
Edad	16a26 472 (52%)	16a26 427 (52%)	16a26 453 (53%)	16a26 445 (53%)	16a26 450 (53%)
Género	Femenino 464 (51%)	Femenino 415 (51%)	Femenino 439 (51%)	Femenino 429 (51%)	Femenino 432 (51%)
Estado Civil	Soltero 642 (71%)	Soltero 574 (70%)	Soltero 605 (71%)	Soltero 589 (70%)	Soltero 596 (70%)
Tipo de Pago de Matricula	Contado 728 (80%)	Contado 666 (82%)	Contado 697 (81%)	Contado 683 (81%)	Contado 683 (81%)
Nota Final	FX 486 (43%)	F 230 (28%)	FX 243 (28%)	FX 243 (29%)	FX 257 (30%)
Estado de Aprobación	REPROBADO 684 (75%)	REPROBADO 437 (53%)	REPROBADO 428 (50%)	REPROBADO 486 (58%)	REPROBADO 472 (56%)
Nivel Interacción del Profesor	Alto 901 (100%)	Alto 812 (100%)	Medio 851 (100%)	Alto 837 (100%)	Alto 842 (100%)

Nivel de Interacción del Estudiante	Alto 323 (35%)	Alto 412 (50%)	Alto 296 (34%)	Bajo 286 (34%)	Alto 294 (34%)
Presento todas las Evaluaciones	SI 615 (68%)	SI 499 (61%)	SI 568 (66%)	SI 559 (66%)	SI 556 (66%)
Supletorio	SI 784 (87%)	SI 559 (68%)	SI 532 (62%)	SI 608 (72%)	SI 546 (64%)
Asistió al Supletorio	SI 458 (50%)	NO 308 (37%)	NO 260 (90%)	SI 311 (37%)	NO_LE_CORRESPONDE 296 (35%)
Deserto	SI 720 (79%)	SI 642 (79%)	SI 283 (98%)	SI 674 (80%)	SI 674 (80%)

En la *[Tabla 3.37],* podemos observar de manera general, que en todos los Full Data generados, predominan con un mayor número de instancias, la edad de 16 a 26 años, el género femenino, el estado civil soltero, el tipo de pago al contado. La mayoría de estudiantes que están cursando las materias de 1er ciclo, han reprobado con una nota de FX = 14 a 27 puntos sobre 40, de los cuales gran segmento constan como Desertores. Se observa además que la mayoría de profesores han obtenido una interacción Alta en el curso a diferencia del profesor de Metodología de la Programación que ha obtenido una interacción Media. La mayoría de estudiantes de las asignaturas de Contabilidad General y Metodología de Estudio, no se han presentado a dar la respectiva evaluación supletoria.

Interpretación de los Clusters por curso: *Carrera de Administración de Empresas*

De igual manera a lo analizado en los clusters de la carrera de Jurisprudencia, se pudo determinar que sería importante analizar datos académicos y socioeconómicos del estudiante, para con ello conocer si dichas variables son influyentes para que un estudiante decida desertar la carrera.

A continuación se detalla el análisis realizado, en cada uno de los clusters de la carrera de Administración de Empresas:

- ***Clusters de Administración I***

Tamaño de la Población: 901 instancias

Instancias de los Clusters:

Cluster 0 → 161 (18%)

Cluster 1 → 541 (60%)

Cluster 2 → 199 (22%)

```
kMeans
======

Number of iterations: 7
Within cluster sum of squared errors: 2778.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                 Full Data             0             1             2
                              (901)         (161)         (541)         (199)
=============================================================================================
CURSO                  AdministracionI AdministracionI AdministracionI AdministracionI
EDAD                         16a26         27a37         16a26         27a37
GENERO                     Femenino      Masculino      Femenino      Femenino
ESTADO_CIVIL               Soltero       Soltero       Soltero       Casado
TIPO_PAGO_MATRICULA        Contado       Contado       Contado       Contado
NOTA_FINAL                    FX            D             FX            FX
ESTADO_APROBACION          REPROBADO     APROBADO      REPROBADO     REPROBADO
NIVEL_INTER_PROF           Alto          Alto          Alto          Alto
NIVEL_INTER_EST            Alto          Medio         Alto          Medio
PRESENT_TODAS_LAS_EVAL     SI            SI            SI            SI
SUPLETORIO                 SI            NO            SI            SI
ASISTIO_SUPLETORIO         SI  NO_LE_CORRESPONDE       SI            SI
DESERTOR                   SI            NO            SI            SI



Time taken to build model (full training data) : 0.06 seconds
```

FIGURA 3. 15. Resultados – Simple K-Means- Administración De Empresas – Administración I

En la [*Figura 3.15*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la asignatura de Administración I de la carrera de Administración de Empresas. Se puede visualizar que existe un total de 7 interacciones, y una suma de error cuadrático de 2778,0 para un semilla de 20, siendo dicha semilla la que ofrece los mejores resultados respecto a la distribución de los datos en los clusters generados.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 20, del curso de Administración I.

TABLA 3. 38. Resultados Simplek-Means – Administración I – Administración De Empresas

Atributo	Full Data 901 (100% de la población)	Cluster0 443 (49% de la población)	Cluster1 261 (29% de la población)	Cluster2 197 (22% de la población)
Edad	16a26 472 (52%)	16a26 245 (55% del Cluster)	27a37 120 (45% del Cluster)	16a26 114 (57% del Cluster)
Género	Femenino 464 (51%)	Masculino 234 (52% del Cluster)	Femenino 171 (65% del Cluster)	Masculino 113 (57%) del Cluster
Estado Civil	Soltero 642 (71%)	Soltero 322 (72% del Cluster)	Soltero 184 (70% del Cluster)	Soltero 136 (69% del Cluster)
Tipo de Pago de Matricula	Contado 728 (80%)	Contado 364 (82% del Cluster)	Contado 203 (77% del Cluster)	Contado 161 (81% del Cluster)
Nota Final	FX 486 (43%)	FX 230 (51% del Cluster)	F 203 (77% del Cluster)	FX 109 (55% del Cluster)
Estado de Aprobación	Reprobado 684 (75%)	Reprobado 310 (69% del Cluster)	Reprobado 236 (90% del Cluster)	Reprobado 138 (70% del Cluster)
Nivel Interacción del Profesor	Alto 901 (100%)	Alto 443 (100% del Cluster)	Alto 261 (100% del Cluster)	Alto 197 (100% del Cluster)
Nivel de Interacción del Estudiante	Alto 323 (35%)	Medio 274 (61% del Cluster)	Alto 126 (48% del Cluster)	Alto 197 (100% del Cluster)
Presento todas las Evaluaciones	SI 615 (68%)	SI 388 (87% del Cluster)	NO 221 (84% del Cluster)	SI 187 (94% del Cluster)

Supletorio	SI 784 (87%)	SI 372 (83% del Cluster)	SI 242 (92% del Cluster)	SI 170 (86% del Cluster)
Asistió al Supletorio	SI 458 (50%)	SI 291 (65% del Cluster)	NO 220 (84% del Cluster)	SI 145 (73% del Cluster)
Deserto	SI 720 (79%)	SI 334 (75% del Cluster)	SI 237 (90% del Cluster)	SI 149 (75% del Cluster)

En la [*Tabla 3.38*], se puede visualizar, que la edad predominante en los clusters 0 y 2 es de 16 a 26 años, siendo masculino el género que se repite con mayor frecuencia en dichos clusters. El estado civil de soltero, prevalece en todos los clusters, tomando en cuenta que este posee el mayor número de población. En los 3 clusters, predomina el Tipo de Pago al Contado. En todos los clusters sobresale el estado Reprobado, predominando en ellos una calificación insuficiente de F= 13 o menos y FX = 14 a 27. El nivel de interacción Alto es el que tiene el profesor en el curso; siendo el nivel Alto en la interacción del estudiante, también el que prepondera en los clusters 1 y 2. Solamente los estudiantes del cluster 0 y 2 SI han presentado todas las evaluaciones; sin embargo la mayoría de estudiantes de los 3 clusters han tenido que rendir el examen supletorio; presentándose a dar el respectivo supletorio solo los del cluster 1 y 2. Los estudiantes que constan como desertores, y han cursado la materia de Administración I, predominan en todos los clusters.

Los estudiantes que han cursado la asignatura de Derecho constitucional, y constan como desertores, poseen las siguientes características comunes:

- Son estudiantes de género masculino, que han pagado la matricula al contado.

- Son estudiantes que en su mayoría han obtenido la una nota de FX= 14 a 27 puntos sobre 40 en la presente asignatura, a pesar que si presentaron todas las evaluaciones de la misma.
- En el curso virtual, tanto el profesor como los estudiantes han obtenido una interacción Alta, es decir que interactúan con frecuencia en dicho entorno virtual del curso.

Como observamos en el análisis de los clusters de la carrera de Jurisprudencia, se ha podido determinar lo mismo con la carrera de Administración de Empresas, ya que los atributos: genero, estado civil y el tipo de pago de matrícula no poseen demasiada influencia para que el estudiante decida desertar, ya que la mayoría de desertores son hombres solteros, y pagan la matricula al contado, es decir que no necesariamente los estudiantes que pagan a crédito la matrícula y son casados van a desertar la carrera, ya que se podría asumir que dichas personas no poseen el tiempo necesario para aplicarlo al aprendizaje de la asignatura.

- ***Clusters de Contabilidad General***

Tamaño de la Población: 812 instancias

Instancias de los Clusters:

Cluster 0 → 271 (33%)

Cluster 1 → 321 (40%)

Cluster 2 → 220 (27%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2286.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0              1              2
                              (812)        (271)          (321)          (220)
=============================================================================================
CURSO                   ContabilidadG  ContabilidadG  ContabilidadG  ContabilidadG
EDAD                            16a26          16a26          16a26          27a37
GENERO                       Femenino       Femenino       Masculino      Masculino
ESTADO_CIVIL                  Soltero        Soltero        Soltero        Soltero
TIPO_PAGO_MATRICULA           Contado        Contado        Contado        Contado
NOTA_FINAL                          F             FX              F              C
ESTADO_APROBACION          REPROBADO      APROBADO       REPROBADO      APROBADO
NIVEL_INTER_PROF                Alto           Alto           Alto           Alto
NIVEL_INTER_EST                 Alto           Alto           Alto           Alto
PRESENT_TODAS_LAS_EVAL            SI             SI             NO             SI
SUPLETORIO                        SI             SI             SI             NO
ASISTIO_SUPLETORIO                NO             SI             NO NO_LE_CORRESPONDE
DESERTOR                          SI             SI             SI             NO



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 16. Resultados – Simple K-Means – Contabilidad General - Administración De Empresas.

En la [*Figura 3.16*], se visualiza los resultados propuestos para la asignatura de Contabilidad General, con, los mismos que muestran un total de 3 interacciones, y una suma de error cuadrático de 2286 para un semilla de 20, igual al caso anterior, la presente semilla es la que brinda un menor valor en el error cuadrático.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 20, del curso de Contabilidad General.

TABLA 3. 39. Resultados Simplek-Means – Contabilidad General – Administración De Empresas

Atributo	**Full Data** **812 (100% de la población)**	**Cluster0** **271 (33% de la población)**	**Cluster1** **321 (40% de la población)**	**Cluster2** **220 (27% de la población)**
Edad	16a26 427 (52%)	16a26 169 (62% del cluster)	16a26 190 (59% del cluster)	27 a 37 117 (53% del cluster)
Género	Femenino 415 (51%)	Femenino 176 (64% del cluster)	Masculino 180 (56% del cluster)	Masculino 122 (55% del cluster)
Estado Civil	Soltero 574 (70%)	Soltero 199 (73% del cluster)	Soltero 246 (76% del cluster)	Soltero 129 (58% del cluster)
Tipo de Pago de Matricula	Contado 666 (82%)	Contado 225 (83% del cluster)	Contado 253 (78% del cluster)	Contado 188 (85% del cluster)
Nota Final	F 230 (28%)	FX 106 (39% del cluster)	F 227 (70% del cluster)	C 72 (32% del cluster)
Estado de Aprobación	Reprobado 437 (53%)	Aprobado 155 (57% del cluster)	Reprobado 321 (100% del cluster)	Aprobado 220 (100% del cluster)
Nivel Interacción del Profesor	Alto 812 (100%)	Alto 271 (100% del cluster)	Alto 321 (100% del cluster)	Alto 220 (100% del cluster)

Nivel de Interacción del Estudiante	Alto 412 (50%)	Alto 134 (49% del cluster)	Alto 159 (49% del cluster)	Alto 119 (54% del cluster)
Presento todas las Evaluaciones	SI 499 (61%)	SI 249 (91% del cluster)	NO 291 (90% del cluster)	SI 220 (100% del cluster)
Supletorio	SI 559 (68%)	SI 229 (84% del cluster)	SI 321 (100% del cluster)	NO 211 (95% del cluster)
Asistió al Supletorio	NO 308 (37%)	SI 212 (78% del cluster)	NO 291 (90% del cluster)	NO_LE_CORRESPONDE 211 (95% del cluster)
Deserto	SI 642 (79%)	SI 242 (89% del cluster)	SI 321 (100% del cluster)	NO 141 (64% del cluster)

En la [*Tabla 3.39*], se puede visualizar, que de igual manera a los clusters anteriores, en los presentes, la edad que se repite con mayor frecuencia en los clusters es de 16 a 26 años, a diferencia del cluster 2 que sobresalen los estudiantes de 27 a 37 años. En los clusters 0 y 1 sobresale una nota insuficiente para aprobar la asignatura como son FX = 14 a 27 y F = 13 o menos respectivamente; al contrario del cluster 2 que predominan los estudiantes que han aprobado la asignatura obteniendo una nota de C = 33 a 35. El profesor de la asignatura posee una interacción alta en el curso, por lo tanto en todos los clusters sobresale el nivel Alto; además también predomina en todos los clusters el nivel de interacción del estudiante Alto, sin embargo solo los estudiantes del cluster 2, que han cursado la presente asignatura constan como No Desertores.

En los presentes clusters podemos observar que en el cluster 2, se encuentran los estudiantes No desertores, de los cuales poseen las siguientes características:

- En su mayoría han aprobado la presente asignatura obteniendo un puntaje considerable, tomando en cuenta que la presente materia pertenece al grupo de las troncales de la carrera.

- La mayoría de estudiantes han obtenido una interacción alta en el curso.

- Han presentado todas las evaluaciones correspondientes de la asignatura.

Después de haber interpretado los presentes clusters, se pudo constatar lo que se analizó en los clusters de la carrera de Jurisprudencia, que la mayoría de estudiantes que constan como aprobados en una determinada asignatura, son No desertores, hecho que será analizado posteriormente, a profundidad con la ayuda de los árboles de decisión.

- ***Clusters de Metodología de Estudio***

Total de la Población: 851

Clustered Instances:

Cluster 0 → 286 (34%)

Cluster 1 → 237 (28%)

Cluster 2 → 328 (39%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2479.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                    Full Data           0           1           2
                                (851)       (286)       (237)       (328)
=============================================================================
CURSO                    MetodologiaE MetodologiaE MetodologiaE MetodologiaE
EDAD                            16a26       16a26       16a26       27a37
GENERO                       Femenino    Femenino    Masculino    Femenino
ESTADO_CIVIL                  Soltero     Soltero     Soltero     Soltero
TIPO_PAGO_MATRICULA           Contado     Contado     Contado     Contado
NOTA_FINAL                         FX           F          FX           D
ESTADO_APROBACION          REPROBADO   REPROBADO   REPROBADO   APROBADO
NIVEL_INTER_PROF              Medio       Medio       Medio       Medio
NIVEL_INTER_EST               Alto        Bajo        Medio       Alto
PRESENT_TODAS_LAS_EVAL          SI          NO          SI          SI
SUPLETORIO                      SI          SI          SI          NO
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE       NO          SI NO_LE_CORRESPONDE
DESERTOR                        SI          SI          SI          SI



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 17. Resultados – Simple K-Means – Metodología De Estudio - Administración De Empresas

Con respecto a los resultados proporcionados para la materia de Metodología de Estudio ilustrados en la [*Figura 3.17*], se puede observar que el algoritmo a realizado 3 interacciones, obteniendo un error cuadrático de 2479,0 para un semilla de 20; con dicha semilla se obtienen los mejores resultados en la distribución de datos, de igual manera como sucedió con las anteriores materias de la presente carrera.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 20, del curso de Metodología de Estudio.

TABLA 3. 40. Resultados Simplek-Means – Metodología De Estudio – Administración De Empresas

Atributo	**Full Data 851 (100%)**	**Cluster0 286 (34% de la población)**	**Cluster1 237 (28% de la población)**	**Cluster2 328 (39% de la población)**
Edad	16a26 453 (53%)	16a26 173 (60% del cluster)	16a26 151 (63% del cluster)	27 a 37 148 (45% del cluster)
Género	Femenino 439 (51%)	Femenino 171 (59% del cluster)	Masculino 147 (62% del cluster)	Femenino 150 (45% del cluster)
Estado Civil	Soltero 605 (71%)	Soltero 222 (77% del cluster)	Soltero 188 (79% del cluster)	Soltero 195 (59% del cluster)
Tipo de Pago de Matricula	Contado 697 (81%)	Contado 227 (79% del cluster)	Contado 195 (82% del cluster)	Contado 275 (83% del cluster)
Nota Final	FX 243 (28%)	F 164 (57% del cluster)	FX 148 (62% del cluster)	D 162 (49% del cluster)
Estado de Aprobación	Reprobado 428 (50%)	Reprobado 268 (93% del cluster)	Reprobado 160 (67% del cluster)	Aprobado 328 (100% del cluster)
Nivel Interacción del Profesor	Medio 851 (100%)	Medio 286 (100% del cluster)	Medio 237 (100%)	Medio 328 (100% del cluster)

Nivel de Interacción del Estudiante	Alto 296 (34%)	Bajo 121 (42% del cluster)	Medio 111 (46% del cluster)	Alto 126 (38% del cluster)
Presento todas las Evaluaciones	SI 568 (66%)	NO 249 (87% del cluster)	SI 206 (86% del cluster)	SI 325 (99% del cluster)
Supletorio	SI 532 (62%)	SI 286 (100% del cluster)	SI 224 (94% del cluster)	NO 306 (93% del cluster)
Asistió al Supletorio	NO 260 (90%)	NO 260 (90% del cluster)	SI 183 (77% del cluster)	NO_LE_CORRESPONDE 306 (93% del cluster)
Deserto	SI 283 (98%)	SI 283 (98% del cluster)	SI 205 (86% del cluster)	SI 191 (58% del cluster)

En la [*Tabla 3.40*], se puede observar, que de igual manera, a los clusters de la asignatura de Administración I, en todos los presentes clusters predominan los estudiantes que SI desertaron la carrera, que en su mayoría son mujeres solteras, las mismas que además han reprobado la asignatura, porque han obtenido un puntaje insuficiente. Hay que tomar en cuenta que en cluster 2 sobresalen los estudiantes que Si han aprobado la asignatura sin embargo constan como desertores el 58% del clusters, considerando que el resto de estudiantes que poseen las mismas características NO constan como desertores, es decir el 42% del clusters.

En el presente análisis, se continua estableciendo lo deducido anteriormente, que cuando los estudiantes aprueban la asignatura, tienen mayor posibilidad a NO desertar la carrera; aunque en los presente clusters se presentó una interrogante al momento de analizar el cluster 2, ya que en este la mayoría de estudiantes que han aprobado la asignatura de Metodología de Estudio constan como desertores, siendo esta una materia de Formación Básica; por lo tanto en las siguientes secciones se aplicarán experimentos específicos, con árboles de decisión para conocer, la influencia que tiene, cuando un estudiante decide desertar, al momento de reprobar tanto las materias correspondientes a las troncales y de formación básica de la carrera.

- ***Clusters de Realidad Nacional***

Total de la Población: 837

Instancias del cluster:

Cluster 0 → 282 (34%)

Cluster 1 → 311 (37%)

Cluster 2 → 244 (29%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 2330.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                     Full Data          0          1          2
                                (837)      (282)      (311)      (244)
=============================================================================================
CURSO                         RealidadN  RealidadN  RealidadN  RealidadN
EDAD                              16a26      16a26      16a26      27a37
GENERO                         Femenino  Masculino   Femenino  Masculino
ESTADO_CIVIL                    Soltero    Soltero    Soltero    Soltero
TIPO_PAGO_MATRICULA             Contado    Contado    Contado    Contado
NOTA_FINAL                           FX          F         FX          C
ESTADO_APROBACION             REPROBADO  REPROBADO  REPROBADO  APROBADO
NIVEL_INTER_PROF                   Alto       Alto       Alto       Alto
NIVEL_INTER_EST                    Bajo       Alto       Bajo       Alto
PRESENT_TODAS_LAS_EVAL               SI         NO         SI         SI
SUPLETORIO                           SI         SI         SI         NO
ASISTIO_SUPLETORIO                   SI         NO         SI NO_LE_CORRESPONDE
DESERTOR                             SI         SI         SI         NO



Time taken to build model (full training data) : 0.05 seconds
```

FIGURA 3. 18. Resultados – Simple K-Means – Realidad Nacional - Administración De Empresas.

En la [Figura 3.18], se muestra los resultados propuestos para la carrera de Realidad Nacional, en donde el algoritmo ha obtenido un error cuadrático de 2330,0 para un semilla de 20; con dicha semilla se obtienen los mejores resultados en la distribución de datos.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 20, del curso de Realidad Nacional.

TABLA 3. 41. Resultados Simplek-Means – Realidad Nacional – Administración De Empresas

Atributo	**Full Data 837 (100% de la población)**	**Cluster0 282 (34% de la población)**	**Cluster1 311 (37% de la población)**	**Cluster2 244 (29% de la población)**
Edad	16a26 445 (53%)	16a26 171 (60% del cluster)	16a26 210 (67% del cluster)	27 a 37 131 (53% del cluster)
Género	Femenino 429 (51%)	Masculino 165 (58%)	Femenino 211 (67% del cluster)	Masculino 143 (58% del cluster)
Estado Civil	Soltero 589 (70%)	Soltero 215 (76% del cluster)	Soltero 239 (76% del cluster)	Soltero 135 (55% del cluster)
Tipo de Pago de Matricula	Contado 683 (81%)	Contado 221 (78% del cluster)	Contado 260 (83% del cluster)	Contado 202 (82% del cluster)
Nota Final	FX 243 (29%)	F 208 (73% del cluster)	FX 186 (59% del cluster)	C 92 (37% del cluster)
Estado de Aprobación	Reprobado 486 (58%)	Reprobado 272 (96% del cluster)	Reprobado 214 (68% del cluster)	Aprobado 244 (100% del cluster)
Nivel Interacción del Profesor	Alto 837 (100%)	Alto 282 (100%)	Alto 311 (100%)	Alto 244 (100% del cluster)
Nivel de Interacción del Estudiante	Bajo 286 (34%)	Alto 108 (38% del cluster)	Bajo 126 (40% del cluster)	Alto 92 (37% del cluster)

Presento todas las Evaluaciones	SI 559 (66%)	NO 259 (91% del cluster)	SI 292 (93% del cluster)	SI 244 (100% del cluster)
Supletorio	SI 608 (72%)	SI 282 (100% del cluster)	SI 302 (97% del cluster)	NO 220 (90% del cluster)
Asistió al Supletorio	SI 311 (37%)	NO 256 (90% del cluster)	SI 261 (83% del cluster)	NO_LE_CORRESPONDE 220 (90% del cluster)
Deserto	SI 674 (80%)	SI 281 (99% del cluster)	SI 282 (90% del cluster)	NO 133 (54% del cluster)

En la [*Tabla 3.41*], se puede observar, que la edad predominante en la mayoría de los clusters es de 16 a 26 años, como en la mayoría de clusters analizados. Además se puede observar que en los clusters 0 y1, preponderan los estudiantes que han reprobado la asignatura, de los cuales la mayoría constan como desertores en los clusters mencionados; a excepción del cluster 2 donde predominan los estudiantes que han aprobado la asignatura con una Nota de C = 33 a 35, de los cuales la mayoría constan como No Desertores. En los clusters 0, se observa que predominan los estudiantes que han obtenido una interacción Alta en el curso, que a pesar de ello, en su mayoría han reprobado la asignatura y constan como desertores, es decir que se vuelve a recalcar lo encontrado en análisis de los clusters de la carrera de Jurisprudencia, que no necesariamente los estudiantes que poseen una Alta interacción en el curso son los más propensos a aprobar la asignatura.

Las características comunes de los estudiantes desertores que se ha podido encontrar en el análisis de los presentes clusters son las siguientes:

- Son estudiantes solteros, que han pagado la matricula al contado, los mismos que en su mayoría poseen de 16 a 26 años de edad.

- Son estudiantes que han reprobado la asignatura, mayoría de los cuales han tenido que proporcionar la evaluación correspondiente de la asignatura y son han asistido a presentarla.

Se ha podido constatar en el análisis de la presente asignatura, que no necesariamente los estudiantes que poseen un nivel de interacción alto en el curso, son los que aprueban la asignatura.

- ***Clusters de Expresión Oral***

Total de la Población: 842

Instancias de los clusters:

Cluster 0 → 285 (34%)

Cluster 1 → 323 (38%)

Cluster 2 → 234 (28%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2355.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                  Full Data             0              1              2
                               (842)         (285)          (323)          (234)
=============================================================================================
CURSO                         Expre0        Expre0         Expre0         Expre0
EDAD                          16a26         16a26          27a37          16a26
GENERO                      Femenino      Femenino       Femenino      Masculino
ESTADO_CIVIL                 Soltero       Soltero        Soltero        Soltero
TIPO_PAGO_MATRICULA          Contado       Contado        Contado        Contado
NOTA_FINAL                        FX             F              D             FX
ESTADO_APROBACION         REPROBADO     REPROBADO      APROBADO      REPROBADO
NIVEL_INTER_PROF               Alto          Alto           Alto           Alto
NIVEL_INTER_EST                Alto         Medio           Bajo           Bajo
PRESENT_TODAS_LAS_EVAL           SI            NO             SI             SI
SUPLETORIO                       SI            SI             NO             SI
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE          NO NO_LE_CORRESPONDE          SI
DESERTOR                         SI            SI             SI             SI



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 19. Resultados – Simple K-Means – Expresión Oral - Administración De Empresas

En la [*Figura 3.19*], se muestra los resultados propuestos por el algoritmo Simple K-means, en donde se observa un error cuadrático de 2355,0 para un semilla de 20; con dicha semilla se obtienen los mejores resultados en la distribución de datos.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 20, del curso de Expresión Oral.

TABLA 3. 42. Resultados Simplek-Means – Expresión Oral – Administración De Empresas

Atributo	Full Data 842 (100%)	Cluster0 285 (34% de la población)	Cluster1 323 (38% de la población)	Cluster2 234 (28% de la población)
Edad	16a26 450 (53%)	16a26 181 (63% del cluster)	27 a 37 151 (46% del cluster)	16a26 144 (61% del cluster)
Género	Femenino 432 (51%)	Femenino 156 (54% del cluster)	Femenino 174 (53% del cluster)	Masculino 132 (56% del cluster)
Estado Civil	Soltero 596 (70%)	Soltero 221 (77% del cluster)	Soltero 194 (60% del cluster)	Soltero 181 (77% del cluster)
Tipo de Pago de Matricula	Contado 683 (81%)	Contado 225 (78% del cluster)	Contado 276 (85% del cluster)	Contado 182 (77% del cluster)
Nota Final	FX 257 (30%)	F 192 (67% del cluster)	D 157 (48% del cluster)	FX 185 (79% del cluster)
Estado de Aprobación	Reprobado 472 (56%)	Reprobado 274 (96% del cluster)	Aprobado 323 (100% del cluster)	Reprobado 198 (84% del cluster)
Nivel Interacción del Profesor	Alto 842 (100%)	Alto 285 (100% del cluster)	Alto 323 (100% del cluster)	Alto 234 (100% del cluster)
Nivel de Interacción del Estudiante	Alto 294 (34%)	Medio 113 (39% del cluster)	Bajo 116 (35% del cluster)	Bajo 98 (41% del cluster)
Presento todas las Evaluaciones	SI 556 (66%)	NO 258 (90% del cluster)	SI 322 (99% del cluster)	SI 207 (88% del cluster)

Supletorio	SI 546 (64%)	SI 285 (100% del cluster)	NO 296 (91% del cluster)	SI 234 (100% del cluster)
Asistió al Supletorio	NO_LE_CORRESPONDE 296 (35%)	NO 260 (91% del cluster)	NO_LE CORRESPONDE 296 (91% del cluster)	SI 202 (86% del cluster)
Deserto	SI 674 (80%)	SI 279 (97% del cluster)	SI 181 (56% del cluster)	SI 214 (91% del cluster)

En la [*Tabla 3.42*], se puede observar, que en todos los clusters predominan los estudiantes que constan como Desertores, y la mayoría de los mismos han reprobado la asignatura; sin embargo solamente los del cluster 1, predominan los estudiantes que han Aprobaron la asignatura, que en su mayoría constan como desertores de la carrera.

Se puede constatar que en el cluster 1 predominan los estudiantes desertores, de los cuales en su mayoría han aprobado la asignatura directamente sin estar en el supletorio, y han obtenido una interacción baja en el curso; por lo tanto en los presentes clusters se ha podido determinar, lo mismo que se estableció en el análisis de Metodología de Estudio de la presente carrera, que a pesar, de que los estudiantes aprueben en las materias que son de formación básica de igual manera la mayoría constan como desertores de la carrera; para lo cual se analizará el número de materias troncales reprobadas en experimentos posteriores con árboles de decisión, para con ello comprobar si es esta una de las razones por las que un estudiante decide desertar la carrera de Administración de Empresas.

- ***CARRERA: GESTIÓN AMBIENTAL***

De igual manera como se ha realizado con las anteriores carreras, en la presente se continúa aplicando el algoritmo SimpleK-Means, generando 3 clusters con los atributos de cada asignatura, las mismas que corresponden a 1er ciclo de la carrera de Gestión Ambiental, a continuación se muestran los resultados obtenidos en Full Data, por cada asignatura. [*ver Tabla 3.43]*

TABLA 3. 43. Resultados Del Clustering – Carrera Gestión Ambiental

	Introducción a la Ciencias Ambientales ***Seed = 50***	**Biología General** ***Seed = 50***	**Metodología de Estudio** ***Seed = 500***	**Realidad Nacional** ***Seed = 500***	**Expresión Oral** ***Seed = 10***
Atributo	**Full Data** 653 (100% de la población)	**Full Data** 636 (100% de la población)	**Full Data** 635 (100% de la población)	**Full Data** 617 (100% de la población)	**Full Data** 600 (100%)
Edad	16a26 408 (62%)	16a26 401 (63%)	16a26 155 (24%)	16a26 396 (64%)	16a26 375 (62%)
Género	Masculino 412 (63%)	Masculino 394 (61%)	Masculino 397 (62%)	Masculino 383 (62%)	Masculino 374 (62%)
Estado Civil	Soltero 525 (80%)	Soltero 515 (80%)	Soltero 512 (80%)	Soltero 503 (81%)	Soltero 478 (79%)
Tipo de Pago de Matricula	Contado 512 (78%)	Contado 499 (78%)	Contado 499 (78%)	Contado 484 (78%)	Contado 470 (78%)
Nota Final	FX 204 (31%)	F 262 (41%)	FX 195 (30%)	FX 210 (34%)	FX 218 (36%)
Estado de Aprobación	APROBADO 343 (52%)	REPROBADO 520 (81%)	APROBADO 359 (56%)	REPROBADO 341 (55%)	REPROBADO 353 (58%)
Nivel Interacción del Profesor	Alto 653 (100%)	Alto 636 (100%)	Alto 851 (100%)	Bajo 617 (100%)	Alto 600 (100%)
Nivel de Interacción del Estudiante	Alto 219 (33%)	Medio 214 (33%)	Medio 211 (33%)	Medio 206 (33%)	Alto 200 (33%)
Presento todas las Evaluaciones	SI 490 (75%)	SI 472 (74%)	SI 475 (74%)	SI 465 (75%)	SI 447 (74%)

Supletorio	SI 386 (59%)	SI 588 (92%)	SI 374 (58%)	SI 446 (72%)	SI 412 (68%)
Asistió al Supletorio	NO_LE_CORRESPONDE 267 (40%)	SI 375 (58%)	NO_LE_CORRESPONDE 261 (41%)	SI 260 (42%)	NO 371 (61%)
Deserto	SI 165 (25%)	SI 473 (74%)	SI 474 (74%)	SI 459 (74%)	SI 440 (73%)

En la *[Tabla 3.43],* podemos observar de manera general, que en todos los Full Data generados, predominan con un mayor número de instancias los siguientes valores: la edad de 16 a 26 años, el género masculino, el estado civil soltero, el tipo de pago al contado. La mayoría de estudiantes han aprobado la asignatura, sin embargo la nota que posee mayor cantidad de instancias es FX = 14 a 27 puntos sobre 40. La mayoría de profesores han obtenido un nivel de interacción Alto en el curso, a diferencia de la Realidad Nacional que ha obtenido un nivel Bajo. Los estudiantes han obtenido un nivel de interacción entre Medio y Alto en todas las asignaturas analizadas. Se puede observar además, que la mayoría de estudiantes, si han presentado tanto las evaluaciones presenciales, como las a distancias de todas las asignaturas. Finalmente es importante mencionar, que la mayoría de estudiantes de la presente carrera han desertado.

Interpretación de los Clusters de los cursos: *Carrera de Gestión Ambiental*

De igual manera a lo analizado en los clusters de las anteriores carreras, se pudo determinar que sería importante analizar datos académicos y socioeconómicos del estudiante, y de esa manera conocer, si dichas variables influyen para que un estudiante deserte la carrera.

A continuación se detalla el análisis realizado, en cada uno de los clusters de la carrera de Gestión Ambiental:

- ***Clusters de Introducción a las Ciencias Ambientales***

Tamaño de la Población: 653 instancias

Instancias de los Cluster:

Cluster 0 → 251 (38%)

Cluster 1 → 277 (42%)

Cluster 2 → 125 (19%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1805.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                  Full Data           0           1           2
                               (653)       (251)       (277)       (125)
=============================================================================
CURSO                  IntroduccionCA  IntroduccionCA  IntroduccionCA  IntroduccionCA
EDAD                           16a26       16a26       16a26       16a26
GENERO                      Masculino   Masculino   Masculino   Masculino
ESTADO_CIVIL                  Soltero     Soltero     Soltero     Soltero
TIPO_PAGO_MATRICULA           Contado     Contado     Contado     Contado
NOTA_FINAL                         FX          FX           D           F
ESTADO_APROBACION           APROBADO    REPROBADO   APROBADO    REPROBADO
NIVEL_INTER_PROF               Alto        Alto        Alto        Alto
NIVEL_INTER_EST                Alto       Medio        Alto        Bajo
PRESENT_TODAS_LAS_EVAL           SI          SI          SI          NO
SUPLETORIO                       SI          SI          NO          SI
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE        SI NO_LE_CORRESPONDE    NO
DESERTOR                         SI          SI          SI          SI


Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 20. Resultados – Simple K-Means- Introducción a las Ciencias Ambientales– Gestión Ambiental.

Respecto a los resultados generados por el algoritmo Simple K-menas, para la asignatura de Introducción a las Ciencias ambientales de la carrera de Gestión ambiental, se puede observar en la [*Figura 3.20*] un total de 4 interacciones, y 1805,0 como suma del error cuadrático, con una semilla establecida de 50, la misma que se comprobó según pruebas realizadas, ser la que muestran un menor error cuadrático.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Introducción a las Ciencias Ambientales.

TABLA 3. 44. Resultados Simplek-Means – Introducción A Las Ciencias Ambientales – Gestión Ambiental

Atributo	**Full Data 653 (100% de la población)**	**Cluster0 251 (38% de la población)**	**Cluster1 277 (42% de la población)**	**Cluster2 125 (19% de la población)**
Edad	16a26 408 (62%)	16a26 181 (72% del Cluster)	16a26 151 (54% del Cluster)	16a26 76 (60% del Cluster)
Género	Masculino 412 (63%)	Masculino 150 (59% del Cluster)	Masculino 171 (61% del Cluster)	Masculino 91 (72% del Cluster)
Estado Civil	Soltero 525 (80%)	Soltero 207 (82% del Cluster)	Soltero 217 (78% del Cluster)	Soltero 101 (80% del Cluster)
Tipo de Pago de Matricula	Contado 512 (78%)	Contado 184 (73% del Cluster)	Contado 230 (83% del Cluster)	Contado 98 (78% del Cluster)
Nota Final	FX 204 (31%)	FX 166 (66% del Cluster)	D 101 (36% del Cluster)	F 85 (68% del Cluster)
Estado de Aprobación	APROBADO 343 (52%)	REPROBADO 186 (74% del Cluster)	APROBADO 277 (100% del Cluster)	REPROBADO 124 (99% del Cluster)
Nivel Interacción del Profesor	Alto 653 (100%)	Alto 251 (100% del Cluster)	Alto 103 (37% del Cluster)	Alto 125 (100% del Cluster)
Nivel de Interacción del Estudiante	Alto 219 (33%)	Medio 104 (41% del Cluster)	Alto 126 (48% del Cluster)	Bajo 54 (43% del Cluster)

Presento todas las Evaluaciones	SI 490 (75%)	SI 251 (100% del Cluster)	SI 277 (100% del Cluster)	NO 124 (99% del Cluster)
Supletorio	SI 386 (59%)	SI 372 (83% del Cluster)	NO 267 (96% del Cluster)	SI 125 (100% del Cluster)
Asistió al Supletorio	NO_LE_CORRESPONDE 267 (40%)	SI 190 (75% del Cluster)	NO_LE_CORRESPONDE 267 (96% del Cluster)	NO 120 (96% del Cluster)
Deserto	SI 165 (25%)	SI 218 (86% del Cluster)	SI 145 (52% del Cluster)	SI 125 (100% del Cluster)

En la [*Tabla 3.44*], se puede visualizar, que la edad que predomina en todos los clusters es de 16 a 26 años, siendo masculino el género, el estado civil soltero, y el tipo de pago al contado el que se repite con mayor frecuencia en dichos clusters. En todos los cluster sobresale una nota diferente, siendo FX = 16 a 27, la que predomina en el cluster 0, D la que predomina, y F = 13 o menos la que prevalece en el cluster 2, tomando en cuenta que el estado de Reprobado sobresale en el cluster 0 y 2, siendo Aprobado el estado que predomina en el cluster 1. El docente la presente asignatura ha obtenido un nivel de interacción Alto en el curso, por lo cual dicho valor es el que sobresale en todos los clusters; siendo el nivel Medio, Alto, Bajo el que sobresale en los clusters 0, 1 y 2 respectivamente en la interacción del estudiante. Solamente los estudiantes del cluster 0 y 1 SI han presentado todas las evaluaciones, a diferencia de los estudiantes del cluster 2 que no han presentado; sin embargo la mayoría de estudiantes de los clusters 1 y 2, si han tenido que rendir el examen supletorio; presentándose a dar el respectivo supletorio solo los del cluster 0, a diferencia de los que predominan en el 2, que no han asistido a rendir la correspondiente evaluación supletoria. Los estudiantes que constan como desertores, y han cursado la materia de Introducción a las Ciencias Ambientales, predominan en todos los clusters.

Los estudiantes que han cursado la asignatura de Introducción a las Ciencias Ambientales, y constan como desertores, poseen las siguientes características comunes, las mismas que son similares a las establecidas en el análisis de las carreras anteriores:

- Son estudiantes solteros de género masculino, que han pagado la matricula al contado.

- Son estudiantes que en su mayoría han reprobado la asignatura, a pesar que si presentaron todas las evaluaciones de la misma.

- En el curso virtual, los estudiantes han obtenido diferentes niveles de interacción, tanto bajo, medio y alto.

- El profesor ha obtenido un nivel de interacción Alto en el curso, por lo tanto si interactúa con frecuencia en el curso.

En los presentes clusters se ha podido observar un comportamiento diferente a los analizados en las anteriores carreras, ya que en el cluster 1, sobresalen los estudiantes que han aprobado la asignatura, y constan igualmente como desertores, tomando en cuenta que esta, es una materia troncal de la carrera; a diferencia de lo que se determinó en las carrera de Jurisprudencia y Administración de Empresas, donde los estudiantes desertan en su mayoría cuando han reprobado al menos una materia troncal; sin embargo en la presente carrera se encontró que un poco más de la mitad de estudiantes, es decir el 52% del clúster que han aprobado de igual manera han desertado, por lo cual probablemente dichos alumnos reprobaron en otras materias de la carrera, por ende tomaron dicha decisión.

De igual manera como lo que hemos determinado en el análisis de las carreras anteriores, en la presente se puedo constatar lo mismo, que los atributos: género, estado civil y el tipo de pago de matrícula no poseen demasiada influencia para que el estudiante decida desertar, como además se pudo determinar que la mayoría de estudiantes que obtuvieron un nivel de interacción alto en el curso aprobaron la asignatura, sin embargo un poco más de la mitad de dichos estudiantes, constan como desertores en la presente asignatura.

Es importante observar que el nivel de interacción del profesor en el curso, no posee un nivel de influencia considerable para que un estudiante no decida desertar la carrera, ya que a pesar de que el docente de la presente asignatura obtuvo un nivel de interacción Alto de igual manera la mayoría de estudiantes de la materia constan como desertores.

- ***Clusters de Biología General***

Tamaño de la Población: 636 instancias

Instancias de los Clusters:

Cluster 0 → 388 (61%)

Cluster 1 → 100 (16%)

Cluster 2 → 148 (23%)

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 1639.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute                 Full Data         0         1         2
                              (636)     (388)     (100)     (148)
=============================================================================
CURSO                     BiologiaG BiologiaG BiologiaG BiologiaG
EDAD                          16a26     16a26     16a26     16a26
GENERO                    Masculino Masculino  Femenino Masculino
ESTADO_CIVIL                Soltero   Soltero   Soltero   Soltero
TIPO_PAGO_MATRICULA         Contado   Contado   Contado   Contado
NOTA_FINAL                        F        FX         D         F
ESTADO_APROBACION         REPROBADO REPROBADO  APROBADO REPROBADO
NIVEL_INTER_PROF               Alto      Alto      Alto      Alto
NIVEL_INTER_EST               Medio      Bajo     Medio      Alto
PRESENT_TODAS_LAS_EVAL           SI        SI        SI        NO
SUPLETORIO                       SI        SI        SI        SI
ASISTIO_SUPLETORIO               SI        SI        SI        NO
DESERTOR                         SI        SI        NO        SI



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 21. Resultados Simple K-Means- Biología General – Gestión Ambiental

En la [*Figura 3.21*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la asignatura de Biología General de la carrera de Gestión Ambiental. Se puede visualizar que existe un total de 6 interacciones, y una suma de error cuadrático de 1639.0 para un semilla de 50, una vez realizada diferentes pruebas con varias semillas en los parámetros del algoritmo, se determinó que la semilla antes mencionada es la que presenta una menor cantidad en el error cuadrático, por lo tanto es una de las mejores distribuciones de datos que se puede obtener en los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Biología General.

TABLA 3. 45. Resultados Simplek-Means – Biología General – Gestión Ambiental

Atributo	Full Data 636 (100% de la población)	Cluster0 388 (61% de la población)	Cluster1 100 (16% de la población)	Cluster2 148 (23% de la población)
Edad	16a26 401 (63%)	16a26 254 (65% del cluster)	16a26 56 (56% del cluster)	16a26 91 (61% del cluster)
Género	Masculino 394 (61%)	Masculino 249 (64% del cluster)	Masculino 56 (56% del cluster)	Masculino 101 (68% del cluster)
Estado Civil	Soltero 515 (80%)	Soltero 328 (84% del cluster)	Soltero 66 (66% del cluster)	Soltero 121 (81% del cluster)
Tipo de Pago de Matricula	Contado 499 (78%)	Contado 297 (76% del cluster)	Contado 88 (88% del cluster)	Contado 114 (77% del cluster)
Nota Final	F 262 (41%)	FX 243 (62% del cluster)	D 47 (47% del cluster)	F 143 (96% del cluster)
Estado de Aprobación	REPROBADO 520 (81%)	Reprobado 361 (93% del cluster)	Aprobado 89 (89% del cluster)	Reprobado 148 (100% del cluster)
Nivel Interacción del Profesor	Alto 636 (100%)	Alto 388 (100% del cluster)	Alto 100 (100% del cluster)	Alto 148 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 214 (33%)	Bajo 157 (40% del cluster)	Medio 46 (46% del cluster)	Alto 58 (39% del cluster)
Presento todas las Evaluaciones	SI 472 (74%)	SI 368 (94% del cluster)	SI 99 (99% del cluster)	NO 143 (96% del cluster)

Supletorio	SI 588 (92%)	SI 373 (96% del cluster)	SI 67 (67% del cluster)	SI 148 (100% del cluster)
Asistió al Supletorio	SI 375 (58%)	SI 299 (77% del cluster)	SI 65 (65% del cluster)	NO 137 (92% del cluster)
Deserto	SI 473 (74%)	SI 312 (80% del cluster)	NO 86 (86% del cluster)	SI 147 (99% del cluster)

En la [*Tabla 3.45*], se puede observar, que igualmente a los anteriores clusters, la edad que prevalece en los presentes, es de 16 a 26 años, siendo esta la que posee la mayor cantidad de matriculados en 1er curso. En los clusters 0 y 2 predomina el estado Reprobado, de los cuales en su mayoría han desertado la carrera, a pesar que en el cluster 2 poseen una interacción alta en el curso; el cluster 1 hace contraste con los otros clusters, ya que es donde sobresalen los estudiantes que han Aprobado y no constan como desertores, además que si han presentado todas las evaluaciones, y han obtenido una interacción Media en el curso.

Podemos observar que en el cluster 0 y 2, se encuentran los estudiantes desertores, los cuales poseen las siguientes características comunes:

- Son estudiantes que en su mayoría son hombres, solteros, que poseen una edad de 16 a 26 años, y que además han pagado la matricula al contado.

- Han reprobado la asignatura, a pesar que los del cluster 0, si se presentaron todos las evaluaciones y si asistieron a dar la correspondiente evaluación supletoria.

En los presentes clusters, se ha podido determinar lo que se encontró, en el análisis de las carreras anteriores, ya que la presente materia corresponde a las troncales de la carrera, se obtuvo un cluster en donde sobresalen los estudiantes que han aprobado y en su mayoría no constan como desertores; por lo cual se vuelve a establecer lo mencionado en las conclusiones anteriores, que mientras los estudiantes aprueben al menos una materia troncal son más propensos a no desertar la carrera.

- ***Clusters de Metodología de Estudio***

Total de la Población: 635

Instancias de los Cluster:

Cluster 0 → 193 (30%)

Cluster 1 → 261 (41%)

Cluster 2 → 181 (29%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1768.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                     Full Data          0          1          2
                                 (635)      (193)      (261)      (181)
=============================================================================================
CURSO                      MetodologiaE MetodologiaE MetodologiaE MetodologiaE
EDAD                             16a26      16a26      16a26      16a26
GENERO                        Masculino  Masculino  Masculino  Masculino
ESTADO_CIVIL                    Soltero    Soltero    Soltero    Soltero
TIPO_PAGO_MATRICULA             Contado    Contado    Contado    Contado
NOTA_FINAL                           FX         FX          D         FX
ESTADO_APROBACION             APROBADO   APROBADO   APROBADO  REPROBADO
NIVEL_INTER_PROF                 Alto       Alto       Alto       Alto
NIVEL_INTER_EST                 Medio       Bajo      Medio      Medio
PRESENT_TODAS_LAS_EVAL             SI         SI         SI         NO
SUPLETORIO                         SI         SI         NO         SI
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE         SI NO_LE_CORRESPONDE         NO
DESERTOR                           SI         SI         SI         SI



Time taken to build model (full training data) : 0.06 seconds
```

FIGURA 3. 22. Resultados – Simple K-Means- Metodología De Estudio– Gestión Ambiental.

Respecto a los resultados propuestos para la asignatura de Metodología de estudio de la carrera de Gestión ambiental, se puede observar en la [*Figura 3.22*] un total de 4 interacciones, y 1768,0 como suma del error cuadrático, con una semilla establecida de 500, la misma que se comprobó según pruebas realizadas, que es la que muestra los mejores resultados referente a la distribución de datos en los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 500, del curso de Metodología de Estudio.

TABLA 3. 46. Resultados Simplek-Means – Metodología De Estudio – Gestión Ambiental

Atributo	**Full Data 635 (100% de la población)**	**Cluster0 193 (30% de la población)**	**Cluster1 261 (41% de la población)**	**Cluster2 181 (29% de la población)**
Edad	16a26 155 (24%)	16a26 127 (65% del cluster)	16a26 156 (59% del cluster)	27 a 37 119 (65% del cluster)
Género	Masculino 397 (62%)	Masculino 119 (61% del cluster)	Masculino 149 (57% del cluster)	Masculino 129 (71% del cluster)
Estado Civil	Soltero 512 (80%)	Soltero 154 (79% del cluster)	Soltero 206 (78% del cluster)	Soltero 152 (83% del cluster)
Tipo de Pago de Matricula	Contado 499 (78%)	Contado 152 (78% del cluster)	Contado 214 (81% del cluster)	Contado 133 (73% del cluster)
Nota Final	FX 195 (30%)	FX 88 (45% del cluster)	D 118 (45% del cluster)	FX 107 (59% del cluster)
Estado de Aprobación	APROBADO 359 (56%)	Aprobado 95 (49% del cluster)	Aprobado 261 (100% del cluster)	Reprobado 181 (100%)
Nivel Interacción del Profesor	Alto 851 (100%)	Alto 193 (100% del cluster)	Alto 261 (100% del cluster)	Alto 181 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 211 (33%)	Bajo 81 (41% del cluster)	Medio 92 (35% del cluster)	Medio 78 (43% del cluster)
Presento todas las Evaluaciones	SI 475 (74%)	SI 178 (92% del cluster)	SI 261 (100% del cluster)	NO 145 (80% del cluster)

Supletorio	SI 374 (58%)	SI 193 (100% del cluster)	NO 261 (100% del cluster)	SI 181 (100% del cluster)
Asistió al Supletorio	NO_LE_CORRESPONDE 261 (41%)	SI 178 (92% del cluster)	NO_LE_CORRESPONDE 261 (100% del cluster)	NO 171 (94% del cluster)
Deserto	SI 474 (74%)	SI 159 (82% del cluster)	SI 135 (51% del cluster)	SI 180 (99% del cluster)

En la [*Tabla 3.46*], se puede visualizar, que en la mayoría de los clusters, sobresalen los hombres solteros, de edad entre los 16 a 26 años, los mismos que han pagado la matricula al contado. En el cluster 0, prevalecen los estudiantes que han aprobado la asignatura sin embargo, la mayoría de estudiantes han obtenido la calificación de FX = 14 a 27, en el presente cluster; en el cluster 1 sobresalen los estudiantes que han aprobado la asignatura con una nota de D, y en el cluster 2 predominan los estudiantes que han reprobado la asignatura con una nota de FX = 14 a 27. El profesor de la presente asignatura ha obtenido un nivel de interacción Alto en el curso, sin embargo los estudiantes han presentado un nivel de interacción Bajo, Medio, Medio en los clusters 0, 1, y 2 respectivamente. La mayoría de estuantes de todos los clusters 0 y 1 si han presentado todas las evaluaciones correspondientes a la asignatura, discrepando con el cluster 2, que no han presentado todas las evaluaciones; además se puede visualizar que la mayoría de estudiantes del cluster 2, donde constan los reprobados, no se han presentado a rendir la respectiva evaluación supletoria de la asignatura. Se puede observar que en todos los cluster, la mayoría de estudiantes constan como desertores.

Dentro de las características comunes que se han encontrado en los 3 clusters, sobre los estudiantes desertores, constan las siguientes:

- Son estudiantes, soltero de género masculino, que en su mayoría han aprobado la asignatura, presentando todas las evaluaciones correspondientes, tomado en cuenta que Metodología de Estudio pertenece al grupo de las materias de Formación Básica de la carrera de Gestión Ambiental.

Realizando el análisis de los presentes clusters, se ha podido determinar, lo ya encontrado en análisis anteriores, que a pesar que la mayoría de estudiantes han aprobado la presente asignatura, perteneciendo está a las materias de formación básica de la carrera, gran parte de dicha población ha desertado la carrera, por lo que se puede decir que no necesariamente aprobar una asignatura de formación básica asegura, que un estudiante no deserte la carrera.

Se puede observar en los clusters, que a pesar que algunos estudiantes han obtenido un nivel de interacción Bajo, de igual manera han aprobado la asignatura, es decir que dicha variable no posee demasiada influencia para que un estudiante puede aprobar la asignatura.

- ***Clusters de Realidad Nacional***

Total de la Población: 617

Instancias de los clusters:

Cluster 0 → 207 (34%)

Cluster 1 → 283 (46%)

Cluster 2 → 127 (21%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 1671.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                              (617)         (207)         (283)         (127)
=============================================================================================
CURSO                     RealidadN     RealidadN     RealidadN     RealidadN
EDAD                          16a26         16a26         16a26         16a26
GENERO                    Masculino     Masculino     Masculino     Masculino
ESTADO_CIVIL                Soltero       Soltero       Soltero       Soltero
TIPO_PAGO_MATRICULA         Contado       Contado       Contado       Contado
NOTA_FINAL                       FX             C            FX             F
ESTADO_APROBACION         REPROBADO     APROBADO      REPROBADO     REPROBADO
NIVEL_INTER_PROF               Bajo          Bajo          Bajo          Bajo
NIVEL_INTER_EST              Medio         Bajo          Alto          Medio
PRESENT_TODAS_LAS_EVAL          SI            SI            SI            NO
SUPLETORIO                      SI            NO            SI            SI
ASISTIO_SUPLETORIO              SI NO_LE_CORRESPONDE        SI            NO
DESERTOR                        SI            NO            SI            SI



Time taken to build model (full training data) : 0.08 seconds
```

FIGURA 3. 23. Resultados – simple k-means- realidad nacional– gestión ambiental.

Respecto a los resultados propuestos para la asignatura de Realidad Nacional de estudio de la carrera de Gestión ambiental, se puede observar un total de 3 interacciones, y 1671,0 como suma del error cuadrático, con una semilla establecida de 500, la misma de igual manera al caso anterior presenta los mejores resultados referente a la distribución de datos.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 500, del curso de Realidad Nacional.

TABLA 3. 47. Resultados Simplek-Means – Realidad Nacional – Gestión Ambiental

Atributo	**Full Data 617 (100% de la población)**	**Cluster0 207 (34% de la población)**	**Cluster1 283 (46% de la población)**	**Cluster2 127 (21% de la población)**
Edad	16a26 396 (64%)	16a26 103 (49% del cluster)	16a26 211 (74% del cluster)	16a26 82 (64% del cluster)
Género	Masculino 383 (62%)	Masculino 130 (62% del cluster)	Masculino 166 (58% del cluster)	Masculino 87 (68% del cluster)
Estado Civil	Soltero 503 (81%)	Soltero 154 (74% del cluster)	Soltero 243 (85% del cluster)	Soltero 106 (83% del cluster)
Tipo de Pago de Matricula	Contado 484 (78%)	Contado 174 (84% del cluster)	Contado 213 (75% del cluster)	Contado 97 (76% del cluster)
Nota Final	FX 210 (34%)	C 77 (37% del cluster)	FX 191 (67% del cluster)	F 103 (81% del cluster)
Estado de Aprobación	REPROBADO 341 (55%)	Aprobado 206 (99% del cluster)	Reprobado 216 (76% del cluster)	Reprobado 124 (97% del cluster)
Nivel Interacción del Profesor	Bajo 617 (100%)	Bajo 207 (100% del cluster)	Bajo 283 (100% del cluster)	Bajo 127 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 206 (33%)	Bajo 83 (40% del cluster)	Alto 121 (42% del cluster)	Medio 50 (39% del cluster)

Presento todas las Evaluaciones	SI 465 (75%)	SI 206 (99% del cluster)	SI 255 (90% del cluster)	NO 123 (96% del cluster)
Supletorio	SI 446 (72%)	NO 171 (82% del cluster)	SI 283 (100% del cluster)	SI 127 (100% del cluster)
Asistió al Supletorio	SI 260 (42%)	NO_LE_CORRESPONDE 171 (82% del cluster)	SI 219 (77% del cluster)	NO 122 (96% del cluster)
Deserto	SI 459 (74%)	NO 125 (60% del cluster)	SI 252 (89% del cluster)	SI 125 (98% del cluster)

En la [*Tabla 3.47*], se puede observar, que los estudiantes que predominan en los clusters, son hombres solteros, los mismos que poseen una edad entre los 16 a los 26 años. En los clusters 1 y 2, se puede visualizar que sobresalen los estudiantes que han reprobado con una nota de FX = 14 a 27 y F = 13 o menos respectivamente, a diferencia del cluster 0, que han aprobado con una nota de C = 33 a 35. El profesor posee en la presente asignatura un nivel de interacción Bajo, es decir que el profesor no interactúa con frecuencia en el EVA del curso. En todos los clusters, sobresalen los estudiantes que si han presentado todas las evaluaciones, sin embargo la mayoría se ha quedado en supletorio y ha reprobado la asignatura. En los clusters 1 y 2 sobresalen los estudiantes que constan como desertores en la presente asignatura siendo estos los que en su mayoría han reprobado; a diferencia del cluster 0, en donde constan los estudiantes que no han desertado y si han aprobado la asignatura.

Una vez analizados los clusters, se ha podido determinar que la mayoría de estudiantes que constan como desertores, han reprobado la presente asignatura, tomando en cuenta que así como la anterior, la presente forma parte del grupo de las materias troncales de la carrera de Gestión Ambiental.

Comparando, el presente análisis, con el de los clusters de Metodología de Estudio, donde a pesar que la mayoría de estudiantes desertaban habiendo aprobado la asignatura, siendo esta una de Formación Básica; en el presente análisis discrepa con lo determinado en el anterior, ya que la mayoría de estudiantes que han aprobado Realidad Nacional, no constan como desertores, a pesar que existe un 40%, que si han desertado, por lo cual, posterior-

mente se analizará dicha interrogante, con la ayuda de los árboles de decisión, para con ello comprobar si el número de materias troncales aprobadas influye en la presente carrera, para que un estudiante decida desertar.

- ***Clusters de Expresión Oral***

Total de la Población: 600

Instancias de los Clusters:

Cluster 0 → 215 (36%)

Cluster 1 → 160 (27%)

Cluster 2 → 225 (38%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 1515.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                Cluster#
Attribute               Full Data          0          1          2
                          (600)      (215)      (160)      (225)
=============================================================================
CURSO                      ExpreO     ExpreO     ExpreO     ExpreO
EDAD                        16a26      16a26      16a26      16a26
GENERO                  Masculino  Masculino  Masculino  Masculino
ESTADO_CIVIL              Soltero    Soltero    Soltero    Soltero
TIPO_PAGO_MATRICULA       Contado    Contado    Contado    Contado
NOTA_FINAL                     FX          D          F         FX
ESTADO_APROBACION       REPROBADO  APROBADO REPROBADO REPROBADO
NIVEL_INTER_PROF             Alto       Alto       Alto       Alto
NIVEL_INTER_EST              Alto      Medio      Medio       Alto
PRESENT_TODAS_LAS_EVAL         SI         SI         NO         SI
SUPLETORIO                     SI         NO         SI         SI
ASISTIO_SUPLETORIO             NO         NO         NO         SI
DESERTOR                       SI         NO         SI         SI


Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 24. Resultados – Simple K-Means- Expresión Oral – Gestión Ambiental

En la [*Figura 3.24*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la asignatura de Expresión Oral de la carrera de Gestión Ambiental. Se puede visualizar que existe un total de 4 interacciones, y una suma de error cuadrático de 1515.0 para un semilla de 10, una vez realizada diferentes pruebas con varias semillas en los parámetros del algoritmo, se determinó que la semilla antes mencionada es la que presenta una menor en el error cuadrático, por lo tanto es una de las mejores distribuciones de datos que se puede obtener en los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 10, del curso de Expresión Oral.

TABLA 3. 48. Resultados Simplek-Means – Expresión Oral – Gestión Ambiental

Atributo	Full Data 600 (100%)	Cluster0 215 (36% de la población)	Cluster1 160 (27% de la población)	Cluster2 225 (38% de la población)
Edad	16a26 375 (62%)	16a26 119 (55% del cluster)	16a26 101 (63% del cluster)	16a26 155 (68% del cluster)
Género	Masculino 374 (62%)	Masculino 127 (59% del cluster)	Masculino 112 (70% del cluster)	Masculino 135 (60% del cluster)
Estado Civil	Soltero 478 (79%)	Soltero 162 (75% del cluster)	Soltero 132 (82% del cluster)	Soltero 184 (81% del cluster)
Tipo de Pago de Matricula	Contado 470 (78%)	Contado 179 (83% del cluster)	Contado 128 (80% del cluster)	Contado 163 (72% del cluster)
Nota Final	FX 218 (36%)	D 113 (52% del cluster)	F 107 (66% del cluster)	FX 171 (76% del cluster)
Estado de Aprobación	REPROBADO 353 (58%)	Aprobado 215 (100% del cluster)	Reprobado 160 (100% del cluster)	Reprobado 193 (85% del cluster)
Nivel Interacción del Profesor	Alto 600 (100%)	Alto 215 (100% del cluster)	Alto 160 (100% del cluster)	Alto 225 (100% del cluster)
Nivel de Interacción del Estudiante	Alto 200 (33%)	Medio 84 (39% del cluster)	Medio 73 (45% del cluster)	Alto 99 (44% del cluster)
Presento todas las Evaluaciones	SI 447 (74%)	SI 213 (99% del cluster)	NO 140 (87% del cluster)	SI 214 (95% del cluster)

Supletorio	SI 412 (68%)	NO 188 (87% del cluster)	SI 160 (100% del cluster)	SI 225 (100% del cluster)
Asistió al Supletorio	NO 371 (61%)	NO 188 (87% del cluster)	NO 145 (90% del cluster)	SI 187 (83% del cluster)
Deserto	SI 440 (73%)	NO 144 (66% del cluster)	SI 160 (100% del cluster)	SI 209 (92% del cluster)

En la [*Tabla 3.48*], se puede observar, que en los presentes clusters predominan la edad de 16 a 26 años, el género masculino, el estado civil soltero, y el tipo de pago al contado, de igual manera como sucedía en alguna de las materias antes analizadas de la presente carrera. Además se observa que el estado que sobresale en los clusters 1 y 2 es Reprobado, en donde predominan los estudiantes que han tenido que rendir una evaluación supletoria, sin embargo solo la mayoría de estudiantes del cluster 2 asistieron; y además en dichos clusters predominan los estudiantes que constan como desertores en la carrera. En el cluster 0 se observa que, sobresalen los estudiantes que han aprobado la asignatura y no constan como desertores.

Se pudo constatar en los presentes clusters que a pesar, de que existen estudiantes que han obtenido un nivel de interacción alto en el curso, de igual manera han reprobado la asignatura y han desertado. Además se comprobó nuevamente lo encontrado en análisis anteriores que existen estudiantes, que como han reprobado la presente asignatura constan como desertores, y la mayoría que han aprobado, no han desertado, tomando en cuenta que Realidad Nacional forma parte del grupo de las materias de Formación Básica de la carrera; aunque existe en dicho grupo un 34% de estudiantes que de igual manera han desertado, por lo tanto se debería considerar también si el mencionado grupo ha reprobado materias troncales.

- ***CARRERA: INFORMÁTICA***

Por último procedemos a realizar la aplicación de la técnica de agrupamiento o clustering para la carrera de Informática, la misma que pertenece al área técnica, de igual manera como se ha realizado con las anteriores carreras, se aplicará el algoritmo SimpleK-Means, generando 3 clusters con los atributos de cada asignatura ofertada en 1er ciclo de la carrera, a continuación se muestran los resultados obtenidos en Full Data, por cada asignatura. [*ver Tabla 3.49]*

TABLA 3. 49. Resultados Del Clustering – Carrera Informática

	Fundamentos Informáticos ***Seed = 50***	**Lógica de la Programación** ***Seed = 50***	**Metodología de Estudio** ***Seed = 50***	**Realidad Nacional** ***Seed = 500***	**Expresión Oral** ***Seed = 50***
Atributo	**Full Data** 378 (100% de la población)	**Full Data** 378 (100% de la población)	**Full Data** 393 (100% de la población)	**Full Data** 380 (100% de la población)	**Full Data** 374 (100% de la población)
Edad	16a26 228 (60%)	16a26 229 (60%)	16a26 239 (60%)	16a26 230 (60%)	16a26 225 (60%)
Género	Masculino 287 (75%)	Masculino 289 (76%)	Masculino 298 (75%)	Masculino 292 (76%)	Masculino 285 (76%)
Estado Civil	Soltero 294 (77%)	Soltero 298 (78%)	Soltero 309 (78%)	Soltero 297 (78%)	Soltero 289 (77%)
Tipo de Pago de Matricula	Contado 301 (79%)	Contado 299 (79%)	Contado 312 (79%)	Contado 304 (80%)	Contado 299 (79%)
Nota Final	FX 116 (30%)	F 201 (53%)	FX 119 (30%)	FX 115 (30%)	FX 116 (31%)
Estado de Aprobación	REPROBADO 221 (58%)	REPROBADO 304 (80%)	REPROBADO 203 (51%)	REPROBADO 221 (58%)	REPROBADO 221 (59%)
Nivel Interacción del Profesor	Alto 378 (100%)	Alto 378 (100%)	Medio 393 (100%)	Bajo 380 (100%)	Alto 374 (100%)

Nivel de Interacción del Estudiante	Medio 154 (40%)	Medio 151 (39%)	Medio 154 (39%)	Medio 151 (39%)	Medio 151 (40%)
Presento todas las Evaluaciones	SI 231 (61%)	SI 228 (60%)	SI 249 (63%)	SI 238 (62%)	SI 232 (62%)
Supletorio	SI 281 (74%)	SI 348 (92%)	SI 144 (36%)	SI 286 (75%)	SI 250 (66%)
Asistió al Supletorio	NO 146 (38%)	SI 190 (50%)	NO_LE_CORRESPONDE 144 (36%)	SI 131 (34%)	NO 135 (36%)
Deserto	SI 303 (80%)	SI 301 (79%)	SI 323 (82%)	SI 310 (81%)	SI 305 (81%)

En la *[Tabla 3.49],* podemos observar de manera general, que en todos los Full Data generados, predominan con un mayor número de instancias los siguientes valores: la edad de 16 a 26 años, el género masculino, el estado civil soltero, el tipo de pago al contado. La mayoría de estudiantes han reprobado cada una de las asignaturas, siendo 14 a 26 puntos sobre 40, la nota que más han obtenido los estudiantes con mayor frecuencia. La mayoría de estudiantes han desertado, además se observa que los profesores de Metodología de Estudio y Realidad Nacional han obtenido un nivel de interacción Medio y Bajo, a diferencia de los profesores de las demás asignaturas que han obtenido un nivel Alto; también se visualiza que la mayoría de estudiantes de todas la carreras han obtenido un nivel de interacción Medio en el curso, por lo tanto no interactúan con frecuencia en los cursos.

Interpretación de los Clusters de los cursos: *Carrera de Informática*

A continuación se detalla el análisis realizado, en cada uno de los clusters de la carrera de Informática:

- ***Clusters de Fundamentos Informáticos***

Tamaño de la Población: 378 instancias

Instancias de los clustes:

Cluster 0 → 165 (44%)

Cluster 1 → 101 (27%)

Cluster 2 → 112 (30%)

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 956.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0                1                2
                              (378)           (165)            (101)            (112)
=============================================================================================
CURSO                   FundamentosI    FundamentosI     FundamentosI     FundamentosI
EDAD                          16a26           16a26            16a26            16a26
GENERO                    Masculino       Masculino        Masculino        Masculino
ESTADO_CIVIL                Soltero         Soltero          Soltero          Soltero
TIPO_PAGO_MATRICULA         Contado         Contado          Contado          Contado
NOTA_FINAL                       FX              FX                C                F
ESTADO_APROBACION         REPROBADO       REPROBADO        APROBADO        REPROBADO
NIVEL_INTER_PROF              Alto            Alto             Alto             Alto
NIVEL_INTER_EST               Medio           Medio            Bajo             Bajo
PRESENT_TODAS_LAS_EVAL           SI              SI               SI               NO
SUPLETORIO                       SI              SI               NO               SI
ASISTIO_SUPLETORIO               NO              SI NO_LE_CORRESPONDE              NO
DESERTOR                         SI              SI               NO               SI



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 25. Resultados – Simple K-Means- Fundamentos Informáticos – Informática

En la [*Figura 3.25*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la asignatura de Fundamentos Informáticos de la carrera de Informática. Se puede visualizar que existe un total de 6 interacciones, y una suma de error cuadrático de 956.0 para un semilla de 50, una vez realizada diferentes pruebas con varias semillas en los parámetros del algoritmo, se determinó que la semilla antes mencionada es la que presenta una menor cantidad en el error cuadrático, por lo tanto es una de las mejores distribuciones de datos que se puede obtener en los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Fundamentos Informáticos.

TABLA 3. 50. Resultados Simplek-Means – Fundamentos Informáticos – Informáticos

Atributo	Full Data 378 (100% de la población)	Cluster0 165 (44% de la población)	Cluster1 101 (27% de la población)	Cluster2 112 (30% de la población)
Edad	16a26 228 (60%)	16a26 118 (71% del cluster)	16a26 48 (47% del cluster)	16a26 62 (55% del cluster)
Género	Masculino 287 (75%)	Masculino 126 (76% del cluster)	Masculino 75 (74% del cluster)	Masculino 86 (76% del cluster)
Estado Civil	Soltero 294 (77%)	Soltero 134 (81% del cluster)	Soltero 70 (69% del cluster)	Soltero 90 (80% del cluster)
Tipo de Pago de Matricula	Contado 301 (79%)	Contado 136 (82% del cluster)	Contado 78 (77% del cluster)	Contado 87 (77% del cluster)
Nota Final	FX 116 (30%)	FX 90 (54% del cluster)	C 49 (48% del cluster)	F 84 (75% del cluster)
Estado de Aprobación	REPROBADO 221 (58%)	REPROBADO 110 (66% del cluster)	APROBADO 101 (100% del cluster)	REPROBADO 111 (99% del cluster)
Nivel Interacción del Profesor	Alto 378 (100%)	Alto 165 (100% del cluster)	Alto 101 (100% del cluster)	Alto 112 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 154 (40%)	Medio 90 (54% del cluster)	Bajo 40 (39% del cluster)	Bajo 44 (39% del cluster)
Presento todas las Evaluaciones	SI 231 (61%)	SI 128 (77% del cluster)	SI 100 (99% del cluster)	NO 109 (97% del cluster)
Supletorio	SI 281 (74%)	SI 165 (100% del cluster)	NO 97 (96% del cluster)	SI 112 (100% del cluster)

Asistió al Supletorio	NO 146 (38%)	SI 128 (77% del cluster)	NO_LE_CORRESPONDE 97 (96% del cluster)	NO 109 (97% del cluster)
Deserto	SI 303 (80%)	SI 153 (92% del cluster)	NO 63 (62% del cluster)	SI 112 (100% del cluster)

En la [*Tabla 3.50*], se puede visualizar, que en todos los clusters predominan los estudiantes solteros de género masculino, los mismos que en su mayoría poseen una edad de 16 a 26 años, y han pagado la matricula al contado. En los clusters 0 y 2 sobresalen los estudiantes que han Reprobado la asignatura, a diferencia del cluster 1 donde predominan los que han Aprobado. En los presentes clusters se puede observar también, que el profesor ha obtenido una interacción Alta en el curso, por lo tanto, si ha interactuado con frecuencia en el EVA del curso, sin embargo, sucede lo contrario, con los estudiantes ya que en su mayoría han obtenido un nivel de interacción Bajo en el curso, según lo que se visualiza, en los clusters 1 y 2. Además se puede observar que en los clusters 0 y 1 predominan los estudiantes que Si han presentado todas las evaluaciones, a diferencia del cluster 2 que se encuentran los estudiantes que No han presentado todas las evaluaciones. En los clusters 0 y 2 constan los estudiantes desertores, al contrario del cluster 1, donde sobresalen los No desertores.

Los estudiantes que han cursado la asignatura de Fundamentos Informáticos, y constan como desertores, poseen las siguientes características comunes, las mismas que son similares a las establecidas en el análisis de las carreras anteriores:

- Son estudiantes solteros de género masculino, que han pagado la matricula al contado, y además poseen una edad entre los 16 a 26 años. Es importante tomar en cuenta que dichos atributos poseen el mayor número de instancias en la presente asignatura.

- Son estudiantes que en su mayoría han reprobado la asignatura.

- En el curso virtual, los estudiantes han obtenido, un nivel de interacción entre Medio y Bajo.
- El profesor ha obtenido un nivel de interacción Alto en el curso, por lo tanto si interactúa con frecuencia en el curso.

En el análisis de los presentes clusters, se ha podido constatar lo establecido en anteriores carreras, ya que al momento, de que el estudiante reprueba una materia y esta pertenece al grupo de las troncales, es más propensos a desertar la carrera, como podemos observar en los clusters, que los estudiantes que reprueban la asignatura en su mayoría constan como desertores, y los estudiantes que han aprobado, en su mayoría no han decido desertar.

Además se ha podido determinar, que no necesariamente los estudiantes que presenten un nivel de interacción Bajo en el entorno del curso, va a reprobar la materia; como también la presentación de todas las evaluaciones a distancia, no asegura que el estudiante apruebe o repruebe la materia.

- ***Clusters de Lógica de la Programación***

Tamaño de la Población: 378 instancias

Instancias de los Clusters:

Cluster 0 → 180 (48%)

Cluster 1 → 133 (35%)

Cluster 2 → 65 (17%)

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 857.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                     Cluster#
Attribute                Full Data          0          1          2
                            (378)      (180)      (133)       (65)
=============================================================================
CURSO                      LogicaP    LogicaP    LogicaP    LogicaP
EDAD                         16a26      16a26      16a26      27a37
GENERO                    Masculino  Masculino  Masculino  Masculino
ESTADO_CIVIL                Soltero    Soltero    Soltero    Soltero
TIPO_PAGO_MATRICULA         Contado    Contado    Contado    Contado
NOTA_FINAL                        F          F         FX          D
ESTADO_APROBACION        REPROBADO  REPROBADO  REPROBADO  APROBADO
NIVEL_INTER_PROF              Alto       Alto       Alto       Alto
NIVEL_INTER_EST               Medio      Alto       Medio      Bajo
PRESENT_TODAS_LAS_EVAL           SI         NO         SI         SI
SUPLETORIO                       SI         SI         SI         SI
ASISTIO_SUPLETORIO               SI         NO         SI         SI
DESERTOR                         SI         SI         SI         NO



Time taken to build model (full training data) : 0.02 seconds
```

FIGURA 3. 26. Resultados – Simple K-Means- Lógica De La Programación – Informática

En la [*Figura 3.26*], se observan los resultados proporcionados para la asignatura de Lógica de la Programación de la carrera de Informática, en donde indica, un total de 3 interacciones realizadas por el algoritmo, y se visualiza un valor de 857,0 en la suma del error cuadrático, con una semilla de 50, tomando en cuenta que se realizaron algunas pruebas de igual manera como los casos anteriores, y se comprobó que con dicha semilla se obtiene uno de los más bajos valores en el error cuadrático, para obtener una adecuada distribución de los datos.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Lógica de la Programación.

TABLA 3. 51. Resultados Simplek-Means – Fundamentos De La Programación – Informática.

Atributo	**Full Data 378 (100% de la población)**	**Cluster0 180 (48% de la población)**	**Cluster1 133 (35% de la población)**	**Cluster2 65 (17% de la población)**
Edad	16a26 229 (60%)	16a26 116 (64% del cluster)	16a26 89 (66% del cluster)	27a37 36 (55% del cluster)
Género	Masculino 289 (76%)	Masculino 141 (78% del cluster)	Masculino 99 (74% del cluster)	Masculino 49 (75% del cluster)
Estado Civil	Soltero 298 (78%)	Soltero 144 (80% del cluster)	Soltero 106 (79% del cluster)	Soltero 48 (73% del cluster)
Tipo de Pago de Matricula	Contado 299 (79%)	Contado 135 (75% del cluster)	Contado 113 (84% del cluster)	Contado 51 (78% del cluster)
Nota Final	F 201 (53%)	F 166 (92% del cluster)	FX 89 (66% del cluster)	D 33 (50% del cluster)
Estado de Aprobación	REPROBADO 304 (80%)	Reprobado 179 (99% del cluster)	Reprobado 124 (93% del cluster)	Aprobado 64 (98% del cluster)

Nivel Interacción del Profesor	Alto 378 (100%)	Alto 180 (100% del cluster)	Alto 133 (100% del cluster)	Alto 65 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 151 (39%)	Alto 66 (36% del cluster)	Medio 77 (57% del cluster)	Bajo 27 (41% del cluster)
Presento todas las Evaluaciones	SI 228 (60%)	NO 144 (80% del cluster)	SI 128 (96% del cluster)	SI 64 (98% del cluster)
Supletorio	SI 348 (92%)	SI 179 (99% del cluster)	SI 131 (98% del cluster)	SI 38 (58% del cluster)
Asistió al Supletorio	SI 190 (50%)	NO 148 (82% del cluster)	SI 122 (91% del cluster)	SI 37 (56% del cluster)
Deserto	SI 301 (79%)	SI 179 (99% del cluster)	SI 118 (88% del cluster)	NO 61 (93% del cluster)

En la [*Tabla 3.51*], se puede observar, que igualmente a los anteriores clusters, en los presentes predominan, los estudiantes solteros de género masculino, que han pagado la matricula al contado, los mismos que en su mayoría poseen una edad de 16 a 26 años, a diferencia del cluster 2, en donde predominan los estudiantes que poseen una edad de 27 a 37 años. En los clusters 0 y 1 prevalecen los estudiantes que han reprobado la asignatura y constan como desertores, al contrario del cluster 2 que constan los estudiantes no desertores, que han aprobado la presente asignatura, perteneciendo esta al grupo de las troncales de la carrera, igual a la anterior materia. Podemos observar que en el cluster 0, sobresalen los estudiantes que han No han presentado todas las evaluaciones, además que teniendo que dar la evaluación supletoria no han asistido a presentarla, a diferencia del cluster 1 y 2, que predominan los estudiantes que Si han presentado todas las evaluaciones y Si han asistido a rendir la correspondiente evaluación supletoria.

Entre las características similares principales de un posible desertor, se encontraron las siguientes, las mismas que ya han sido descubiertas en análisis anteriores:

- Los estudiantes poseen una edad entre los 16 a 26 años.

- Los estudiantes han reprobado la asignatura.

En el presente análisis de los clusters generados, se encontró un comportamiento similar a carreras anteriores y a la asignatura de Fundamentos Informáticos analizada anteriormente, ya que se observa que mientras el estudiante reprueba la asignatura, son mayormente propensos a desertar la carrera, y cuando el estudiante aprueba la asignatura, tiene una gran posibilidad de no desertar la carrera, tomando en cuenta que la materia de Fundamentos Informáticos y Lógica de la Programación, pertenecen al grupo de las troncales de la carrera.

En análisis posteriores se generaran los clusters de las materias que forman parte del grupo de Formación Básica de la carrera, para con ello poder observar si, su comportamiento es similar al encontrado hasta el momento o ha cambiado.

- ***Clusters de Metodología de Estudio***

Tamaño de la Población: 393 instancias

Instancias de los Clusters:

Cluster 0 → 145 (37%)

Cluster 1 → 99 (25%)

Cluster 2 → 149 (38%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 978.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                        Cluster#
Attribute                 Full Data            0            1            2
                             (393)        (145)         (99)        (149)
=============================================================================================
CURSO                   MetodologiaE MetodologiaE MetodologiaE MetodologiaE
EDAD                           16a26        16a26        16a26        16a26
GENERO                      Masculino    Masculino    Masculino    Masculino
ESTADO_CIVIL                  Soltero      Soltero      Soltero      Soltero
TIPO_PAGO_MATRICULA           Contado      Contado      Contado      Contado
NOTA_FINAL                         FX            F           FX            D
ESTADO_APROBACION          REPROBADO    REPROBADO    REPROBADO    APROBADO
NIVEL_INTER_PROF              Medio        Medio        Medio        Medio
NIVEL_INTER_EST               Medio        Medio         Alto         Bajo
PRESENT_TODAS_LAS_EVAL           SI           NO           SI           SI
SUPLETORIO                       SI           SI           SI           NO
ASISTIO_SUPLETORIO  NO_LE_CORRESPONDE        NO           SI NO_LE_CORRESPONDE
DESERTOR                         SI           SI           SI           SI



Time taken to build model (full training data) : 0.14 seconds
```

FIGURA 3. 27. Resultados – Simple K-Means- Metodología De Estudio – Informática

En la [*Figura 3.27*], se observan los resultados proporcionados para la carrera de Lógica de la Programación de la carrera de Informática, en donde indica, un total de 4 interacciones realizadas por el algoritmo, y se visualiza un valor de 978,0 en la suma del error cuadrático, con una semilla de 50, tomando en cuenta que se realizaron algunas pruebas de igual manera como los casos anteriores, y se comprobó que con dicha semilla se obtiene uno de los valores más bajos en el error cuadrático, para obtener una adecuada distribución de los datos, entre los clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Metodología de Estudio.

TABLA 3. 52. Resultados Simplek-Means – Metodología De Estudio – Informática

Atributo	**Full Data 393 (100% de la población)**	**Cluster0 145 (37% de la población)**	**Cluster1 99 (25% de la población)**	**Cluster2 149 (38% de la población)**
Edad	16a26 239 (60%)	16a26 87 (60% del cluster)	16a26 67 (67% del cluster)	16a26 85 (57% del cluster)

Género	Masculino 298 (75%)	Masculino 113 (77% del cluster)	Masculino 73 (73% del cluster)	Masculino 112 (75% del cluster)
Estado Civil	Soltero 309 (78%)	Soltero 115 (79% del cluster)	Soltero 83 (83% del cluster)	Soltero 111 (74% del cluster)
Tipo de Pago de Matricula	Contado 312 (79%)	Contado 112 (77% del cluster)	Contado 80 (80% del cluster)	Contado 120 (80% del cluster)
Nota Final	FX 119 (30%)	F 75 (51% del cluster)	FX 59 (59% del cluster)	D 73 (48% del cluster)
Estado de Aprobación	REPROBADO 203 (51%)	REPROBADO 138 (95% del cluster)	REPROBADO 65 (65% del cluster)	APROBADO 149 (100% del cluster)
Nivel Interacción del Profesor	Medio 393 (100%)	Medio 145 (100% del cluster)	Medio 99 (100% del cluster)	Medio 149 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 154 (39%)	Medio 76 (52% del cluster)	Alto 45 (45% del cluster)	Bajo 58 (38% del cluster)
Presento todas las Evaluaciones	SI 249 (63%)	NO 136 (93% del cluster)	SI 85 (85% del cluster)	SI 149 (100% del cluster)
Supletorio	SI 144 (36%)	SI 145 (100% del cluster)	SI 99 (100% del cluster)	NO 144 (96% del cluster)
Asistió al Supletorio	NO_LE_CORRESPONDE 144 (36%)	NO 132 (91% del cluster)	SI 92 (92% del cluster)	NO_LE_CORRESPONDE 144 (96% del cluster)
Deserto	SI 323 (82%)	SI 144 (99% del cluster)	SI 88 (88% del cluster)	SI 91 (61% del cluster)

En la [*Tabla 3.52*], se puede visualizar, que de igual manera a los anteriores clusters, en los presentes, sobresalen los estudiantes hombres solteros que poseen una edad entre los 16 a 26 años, los mismos que han pagado la matricula al contado, se puede observar además que el profesor ha obtenido un nivel de interacción Medio en el curso por lo tanto, dicho nivel predomina en todos los clusters, a diferencia de las materias anteriores, ya que los profesores tenían un nivel Alto de interacción. Se visualiza también que en los clusters 0 y 1 predominan los estudiantes que han Reprobado y constan como desertores, los cuales han obtenido un nivel de interacción Medio y Alto, respectivamente; contrastando el cluster 2, en donde predominan los estudiantes que constan como desertores y han Aprobado la asignatura, los mismos que en un considerable cantidad han obtenido un nivel de interacción Bajo en el curso.

Dentro de las características comunes que se han encontrado en los 3 clusters, sobre los estudiantes desertores, constan las siguientes:

- Son estudiantes soltero de género masculino, que han pagado la matricula al contado, y que además poseen una edad entre los 16 a 26 años, la presente característica, también fue encontrada en el análisis de las anteriores asignaturas.

- Son estudiantes que en cierta cantidad han aprobado y reprobado la asignatura, considerando que la presente materia forma parte del grupo de Formación Básica de la carrera.

En los clusters de la presente asignatura, se ha podido determinar, lo ya descubierto en análisis anteriores, que a de que la mayoría de estudiantes han aprobado la presente asignatura, perteneciendo está al grupo de formación básica de la carrera; la mayoría de dicho grupo constan como desertores, por lo que se puede señalar que no necesariamente aprobar una asignatura de formación básica asegura, que un estudiante no deserte la carrera.

Se puede observar además, un comportamiento también encontrado en análisis anteriores, que el nivel de interacción del estudiante, no influye en gran escala para que el estudiante apruebe o no la materia, ya que como se visualiza en el cluster 2, que cierta cantidad de estudiantes han obtenido un nivel de interacción Bajo en el curso, de los cuales todos han aprobado la asignatura.

- ***Clusters de Realidad Nacional***

Tamaño de la Población: 380 instancias

Instancias de los clustes:

Cluster 0 → 166 (44%)

Cluster 1 → 117 (31%)

Cluster 2 → 97 (26%)

```
kMeans
======

Number of iterations: 6
Within cluster sum of squared errors: 950.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                              (380)         (166)         (117)          (97)
=============================================================================================
CURSO                     RealidadN     RealidadN     RealidadN     RealidadN
EDAD                          16a26         16a26         16a26         27a37
GENERO                    Masculino     Masculino     Masculino     Masculino
ESTADO_CIVIL                Soltero       Soltero       Soltero       Soltero
TIPO_PAGO_MATRICULA         Contado       Contado       Contado       Contado
NOTA_FINAL                       FX            FX             F             C
ESTADO_APROBACION         REPROBADO     REPROBADO     REPROBADO     APROBADO
NIVEL_INTER_PROF               Bajo          Bajo          Bajo          Bajo
NIVEL_INTER_EST               Medio          Bajo         Medio         Medio
PRESENT_TODAS_LAS_EVAL           SI            SI            NO            SI
SUPLETORIO                       SI            SI            SI            NO
ASISTIO_SUPLETORIO               SI            SI            NO NO_LE_CORRESPONDE
DESERTOR                         SI            SI            SI            NO



Time taken to build model (full training data) : 0.03 seconds
```

FIGURA 3. 28. Resultados – Simple K-Means- Realidad Nacional – Informática

En la [*Figura 3.28*], se observan los resultados generados, al momento de ejecutar el algoritmo Simple K-means, para la asignatura de Fundamentos Informáticos de la carrera de Informática. Se puede visualizar que el modelo se ha demorado 0.03 segundos en construirse, que existe un total de 6 interacciones, y una suma de error cuadrático de 950.0 para un semilla de 500, una vez realizada diferentes pruebas con varias semillas en los parámetros del algoritmo, se determinó que la semilla antes mencionada es la que presenta una menor cantidad en el error cuadrático, por lo cual se ha obtenido una adecuada distribución de los datos entre los 3 clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 500, del curso de Realidad Nacional.

TABLA 3. 53. Resultados Simplek-Means – Realidad Nacional – Informática

Atributo	Full Data 380 (100% de la población)	Cluster0 166 (44% de la población)	Cluster1 117 (31% de la población)	Cluster2 97 (26% de la población)
Edad	16a26 230 (60%)	16a26 124 (74% del cluster)	16a26 69 (58% del cluster)	27a37 48 (49% del cluster)
Género	Masculino 292 (76%)	Masculino 117 (70% del cluster)	Masculino 96 (82% del cluster)	Masculino 79 (81% del cluster)
Estado Civil	Soltero 297 (78%)	Soltero 135 (81% del cluster)	Soltero 93 (79% del cluster)	Soltero 69 (71% del cluster)
Tipo de Pago de Matricula	Contado 304 (80%)	Contado 144 (86% del cluster)	Contado 88 (75% del cluster)	Contado 72 (74% del cluster)
Nota Final	FX 115 (30%)	FX 96 (57% del cluster)	F 93 (79% del cluster)	C 44 (45% del cluster)
Estado de Aprobación	REPROBADO 221 (58%)	Reprobado 108 (65% del cluster)	Reprobado 113 (96% del cluster)	Aprobado 97 (100% del cluster)
Nivel Interacción del Profesor	Bajo 380 (100%)	Bajo 166 (100% del cluster)	Bajo 117 (100% del cluster)	Bajo 97 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 151 (39%)	Bajo 72 (43% del cluster)	Medio 55 (47% del cluster)	Medio 44 (45% del cluster)
Presento todas las Evaluaciones	SI 238 (62%)	SI 140 (84% del cluster)	NO 116 (99% del cluster)	SI 97 (100% del cluster)

Supletorio	SI 286 (75%)	SI 162 (97% del cluster)	SI 117 (100% del cluster)	SI 127 (100% del cluster)
Asistió al Supletorio	SI 131 (34%)	SI 139 (83% del cluster)	NO 108 (92% del cluster)	NO_LE_CORRESPONDE 90 (92% del cluster)
Deserto	SI 310 (81%)	SI 149 (89% del cluster)	SI 115 (98% del cluster)	NO 51 (52% del cluster)

En la [*Tabla 3.53*], se puede observar, que los estudiantes hombres solteros, predominan en los clusters, los mismos que en su mayoría poseen una edad de 16 a 26 años, de igual manera como sucedió en los clusters de las anteriores materias. A diferencia de las materias anteriores, en la presente el profesor ha obtenido un nivel de interacción Bajo en el curso, por lo cual en los 3 clusters predomina el nivel Bajo. En los clusters 0 y 1, se puede visualizar que sobresalen los estudiantes que han reprobado con una nota de FX = 14 a 27 y F = 13 o menos respectivamente, a diferencia del cluster 2, que han aprobado con una nota de C. En el cluster 1 se puede observar que sobresalen los estudiantes, que No han presentado todas las evaluaciones, y teniendo que dar la evaluación supletoria no se han presentado a rendirla, por lo tanto reprobaron la misma. Podemos observar que en los clusters 0 y 1 constan los estudiantes que han cursado la presente asignatura y han desertado, y en el cluster 2 constan los no desertores.

Luego de analizar cada uno de los clusters se ha podido determinar lo también encontrado en análisis anteriores, que mientras un estudiante reprueba una asignatura es más propenso a desertar la carrera, y cuando la aprueba, existen más posibilidades que no deserte la carrera; es importante tomar en cuenta además, que no es seguro, que un estudiante no deserte la carrera al momento de que este aprueba la asignatura de Formación Básica de la carrera, ya que como se puede observar en el cluster 2, donde constan 97 estudiantes que han aprobado Realidad Nacional, considerando que dicha materia pertenece al grupo de Formación Básica, solamente un total de 51 estudiantes, no han desertado la carrera, el resto de dicho población si ha desertado.

Se puede observar que de igual, a las anteriores materias, al momento que un estudiante no presente todas las evaluaciones de la asignatura es más propenso a no presentarse a dar la evaluación supletoria y por ende reprobar la asignatura.

Se ha podido constatar en el presente análisis, que una considerable cantidad de estudiantes ha aprobado la asignatura, a pesar que el profesor a obtenido un nivel de interacción Bajo en el curso, por lo tanto dicha variable no influyen en gran escala para que un estudiante apruebe o repruebe la asignatura.

- ***Clusters de Expresión Oral***

Total de la Población: 374 instancias

Instancias de los clustes:

Cluster 0 → 136 (36%)

Cluster 1 → 137 (37%)

Cluster 2 → 101 (27%)

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 893.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                 Full Data             0             1             2
                              (374)         (123)         (122)         (129)
=============================================================================================
CURSO                        ExpreO        ExpreO        ExpreO        ExpreO
EDAD                          16a26         16a26         16a26         16a26
GENERO                    Masculino     Masculino     Masculino     Masculino
ESTADO_CIVIL                Soltero       Soltero       Soltero       Soltero
TIPO_PAGO_MATRICULA         Contado       Contado       Contado       Contado
NOTA_FINAL                       FX            FX             F             D
ESTADO_APROBACION        REPROBADO     REPROBADO     REPROBADO     APROBADO
NIVEL_INTER_PROF              Alto          Alto          Alto          Alto
NIVEL_INTER_EST               Medio         Medio         Medio         Bajo
PRESENT_TODAS_LAS_EVAL           SI            SI            NO            SI
SUPLETORIO                       SI            SI            SI            NO
ASISTIO_SUPLETORIO               NO            SI            NO NO_LE_CORRESPONDE
DESERTOR                         SI            SI            SI            SI



Time taken to build model (full training data) : 0.02 seconds
```

FIGURA 3. 29. Resultados – Simple K-Means- Expresión Oral Y Escrita – Informática

Respecto a los resultados propuestos para la asignatura de Expresión Oral de la carrera de Informática, en la [*Figura 3.29*] se puede observar, que el modelo se ha demorado en construirse 0.02 segundos, ha realizado el algoritmo un total de 4 interacciones para generar el modelo, y ha obtenido un total de 893,0 como suma del error cuadrático, con una semilla establecida de 50, la misma que presenta la mejor distribución de los datos entre los 3 clusters.

A continuación se detallan las características de los 3 clusters, obtenidos con un semilla de 50, del curso de Expresión Oral.

TABLA 3. 54. Resultados Simplek-Means – Expresión Oral – Informática

Atributo	Full Data 374 (100% de la población)	Cluster0 123 (33% de la población)	Cluster1 122 (33% de la población)	Cluster2 129 (34% de la población)
Edad	16a26 225 (60%)	16a26 93 (75% del cluster)	16a26 70 (57% del cluster)	16a26 62 (48% del cluster)
Género	Masculino 285 (76%)	Masculino 87 (70% del cluster)	Masculino 100 (81% del cluster)	Masculino 98 (75% del cluster)
Estado Civil	Soltero 289 (77%)	Soltero 105 (85% del cluster)	Soltero 99 (81% del cluster)	Soltero 85 (65% del cluster)
Tipo de Pago de Matricula	Contado 299 (79%)	Contado 106 (86% del cluster)	Contado 96 (78% del cluster)	Contado 97 (75% del cluster)
Nota Final	FX 116 (31%)	FX 87 (70% del cluster)	F 93 (76% del cluster)	D 56 (43% del cluster)
Estado de Aprobación	REPROBADO 221 (59%)	Reprobado 99 (80% del cluster)	Reprobado 122 (100% del cluster)	Aprobado 129 (100% del cluster)
Nivel Interacción del Profesor	Alto 374 (100%)	Alto 123 (100%)	Alto 122 (100% del cluster)	Alto 129 (100% del cluster)
Nivel de Interacción del Estudiante	Medio 151 (40%)	Medio 61 (49% del cluster)	Medio 49 (40% del cluster)	Bajo 57 (44% del cluster)

Presento todas las Evaluaciones	SI 232 (62%)	SI 101 (82% del cluster)	NO 120 (98% del cluster)	SI 129 (100% del cluster)
Supletorio	SI 250 (66%)	SI 123 (100% del cluster)	SI 122 (100% del cluster)	NO 124 (96% del cluster)
Asistió al Supletorio	NO 135 (36%)	SI 105 (85% del cluster)	NO 117 (95% del cluster)	NO_LE_CORRESPONDE 124 (96% del cluster)
Deserto	SI 305 (81%)	SI 113 (91% del cluster)	SI 121 (99% del cluster)	SI 71 (55% del cluster)

En la [*Tabla 3.54*], se puede observar, que en los presentes clusters predominan: la edad de 16 a 26 años, el género masculino, el estado civil soltero, y el tipo de pago al contado, de igual manera como sucedía en alguna de las materias antes analizadas de la presente carrera. En los cluster 0 y 1 se observar que sobresalen los estudiantes que ha reprobado la asignatura con una nota de FX = 14 a 27 y F = 13 o menos, respectivamente, en donde el nivel de interacción del estudiante que predomina es Medio, a diferencia del cluster 2, donde predominan los estudiantes que aprobaron la asignatura con una nota D, en donde el estudiante ha obtenido un nivel Bajo de interacción. Se puede visualizar, que en todos los clusters predominan los estudiantes que Si han desertado la asignatura.

Se pudo constatar en los presentes clusters Una vez analizado los clusters se ha podido confirmar lo determinado en el análisis de los anteriores clusters, ya que al momento, de que un estudiante reprueba la asignatura, posee más posibilidades a desertar la carrera, aunque en los presentes clusters se observa que existen estudiantes, que si han aprobado la asignatura sin embargo de igual manera la mayoría han desertado, considerando que Expresión Oral forma parte del grupo de Formación Básica de la carrera, ya que el comportamiento de desertores y no desertores, es diferente en materias troncales; por lo tanto se analizará con árboles de decisión, el número de materias troncales que reprueba el estudiante tanto desertor como no desertor, como ya se analizó con las anteriores carreras.

A pesar, de que existen estudiantes que han obtenido un nivel de interacción Bajo en el curso, de igual manera han aprobado la asignatura, por lo cual se vuelve a recalcar lo establecido en análisis anteriores, que la presente variable no posee una alta influencia para que el estudiante apruebe o repruebe la asignatura, ya que la nota final del estudiante también depende del rendimiento que tenga en cada una de las evaluaciones tanto presencial como a distancia correspondientes a la materia.

Como ha sucedido en las anteriores materias, en la presente también, la mayoría de estudiantes que no presentan las evaluaciones de la asignatura, teniendo que rendir la evaluación supletoria no se presentan a efectuarla.

b. Selección de Atributos

Para evaluar el nivel de calidad de los atributos del dataset, se creyó conveniente aplicar métodos de selección de atributos, que propone la herramienta Weka.

Se ha realizado una selección de los atributos, para con certeza conocer cuáles son los atributos que poseen un mayor nivel de relevancia, con ello poder elegir dichos atributo en los algoritmos de clasificación y asociación, los mismos que se implementaran en las siguientes secciones, y así poder conocer las razones más específicas de la deserción estudiantil.

Weka propone una serie de métodos eficaces para realizar la selección de atributos, de los cuales aplicaran: filtros.

- ***Resultados de la Selección de Atributos***

  - ***CARRERA: JURISPRUDENCIA***

A continuación se muestran los resultados de los filtros aplicados para el dataset de la carrera de Jurisprudencia, que han ayudado a evaluar los atributos, proporcionando con ello un ranking de los mismos:

TABLA 3. 55. Ranking De Atributos – **Chisquaredattributeeval - Jurisprudencia**

ChiSquaredAttributeEval		
Atributo	**Ranked**	**Clase**
NOTA_FINAL	1886.2783	DESERTOR
ESTADO_APROBACION	1758.1897	
ASISTIO_SUPLETORIO	1484.9648	
SUPLETORIO	1304.035	
PRESENT_TODAS_LAS_EVAL	581.0545	
EDAD	99.6954	
ESTADO_CIVIL	99.1503	
TIPO_PAGO_MATRICULA	10.4938	
GENERO	3.5506	
CURSO	1.167	
NIVEL_INTER_PROF	0.6269	
NIVEL_INTER_EST	0.3568	

La [*Tabla 3.55*] muestra los resultados del algoritmo evaluador **ChiSquaredAttributeEval**: el mismo que calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo. Considerando que la correlación es comúnmente utilizada en estadística, para indicar la fuerza y la dirección de una relación lineal entre dos variables aleatorias.

En el presente caso la clase seleccionada es desertor, para con ello mostrar la influencia que tienen los atributos establecidos con respecto a dicha clase. Se observa en la tabla el orden de los atributos según la importancia que establece el Ranked, el mismo que se refiere al orden medio en el que quedó cada atributo en cada uno de los 10 ciclos. El algoritmo establece como mejor atributos a *Nota_Final, Estado_Aprobacion, Asistio_Supletorio, Supletorio*.

Posteriormente se ha realizado una evaluación de los atributos antes descritos, con otros algoritmos de filtros evaluadores, para con ello analizar cuáles son los atributos más importantes y menos importantes del dataset seleccionado. Las deducciones de estos algoritmos se muestran en las [*Tablas 3.56, 3.57, 3.58*].

TABLA 3. 56. Ranking De Atributos – **Gainratioattributeeval – Jurisprudencia**

GainRatioAttributeEval	
Atributo	**Weight (Peso)**
SUPLETORIO	0.74261
ESTADO_APROBACION	0.58282
ASISTIO_SUPLETORIO	0.55666
PRESENT_TODAS_LAS_EVAL	0.50613
NOTA_FINAL	0.32306
DESERTOR	0.30485
NIVEL_INTER_PROF	0.16627
CURSO	0.12643
ESTADO_CIVIL	0.07359
EDAD	0.06315
GENERO	0.06024
NIVEL_INTER_EST	0.05642
TIPO_PAGO_MATRICULA	0.00588

TABLA 3. 57. Ranking De Atributos – **Infogainattributeeval – Jurisprudencia**

InfoGainAttributeEval	
Atributo	**Weight (Peso)**
ASISTIO_SUPLETORIO	0.8678
NOTA_FINAL	0.74768
SUPLETORIO	0.63845
ESTADO_APROBACION	0.57281
PRESENT_TODAS_LAS_EVAL	0.4272
CURSO	0.29353
DESERTOR	0.24773
NIVEL_INTER_PROF	0.22658
EDAD	0.11648
NIVEL_INTER_EST	0.08763
ESTADO_CIVIL	0.06619
GENERO	0.05914
TIPO_PAGO_MATRICULA	0.00445

TABLA 3. 58. Ranking De Atributos – **Relieffattributeeval - Jurisprudencia**

ReliefFAttributeEval	
Atributo	**Weight (Peso)**
ASISTIO_SUPLETORIO	0.40302
NOTA_FINAL	0.35232
SUPLETORIO	0.29895
ESTADO_APROBACION	0.28903
PRESENT_TODAS_LAS_EVAL	0.22734
CURSO	0.20032
NIVEL_INTER_EST	0.1934
NIVEL_INTER_PROF	0.17249
GENERO	0.16861
DESERTOR	0.16734
EDAD	0.16408
ESTADO_CIVIL	0.08337
TIPO_PAGO_MATRICULA	-0.00452

A través de los resultados propuestos por los 3 algoritmos evaluadores, descritos en las [*Tablas 3.56, 3.57, 3.58*], se pudo verificar lo que se estableció cuando se crearon los clusters, encontrando que el atributo *TIPO_PAGO_MATRICULA*, es irrelevante en el dataset, ya que observando en las tablas descritas, dicho campo presenta el menor peso comparado con los otros atributos. Solo los algoritmos *InfoGainAttributeEval* y *ReliefFAttributeEval*, concuerdan que los atributos más relevantes son ASISTIO_SUPLETORIO, el algoritmo *GainRatioAttributeEval* discrepa, ya que el mismo establece que el atributo Supletorio es el más relevante. A pesar de la diferencia que tiene cada atributo en colocar el peso del ranking de cada atributo, concuerdan en que ASISTIO_SUPLETORIO, NOTA_FINAL, SUPLETORIO, ESTADO_APROBACION y PRESENT_TODAS_LAS_EVAL, son los atributos que ocupan los 5 primeros puestos de importancia en el dataset.

- ***CARRERA: ADMINISTRACIÓN DE EMPRESAS***

A continuación se muestran los resultados que han proporcionado los algoritmos evaluadores, los mismos que ayudan a evaluar la calidad de los atributos y con ello conocer el ranking de los mismos:

TABLA 3. 59. Ranking De Atributos – **Chisquaredattributeeval – Administración De Empresas**

ChiSquaredAttributeEval		
Atributo	**Ranked**	**Clase**
NOTA_FINAL	1311.70539	DESERTOR
ESTADO_APROBACION	1177.21583	
ASISTIO_SUPLETORIO	1086.49124	
SUPLETORIO	975.46538	
PRESENT_TODAS_LAS_EVAL	493.65276	
EDAD	113.53136	
ESTADO_CIVIL	82.82951	
TIPO_PAGO_MATRICULA	11.36131	
CURSO	0.57289	
NIVEL_INTER_EST	0.23568	
GENERO	0.17283	
NIVEL_INTER_PROF	0.0047	

En la [*Tabla. 3.59*], se visualiza el orden de los atributos según el nivel de correlación que existe entre ellos, con respecto a la clase Desertor. El algoritmo establece que el atributo que posee una mayor relación respecto a la clases son: *Nota_Final, Estado_Aprobacion, Asistio_Supletorio*, *Supletorio,* el atributo Nota_Final es el que posee una mayor influencia, de igual manera como sucedió en la carrera de Jursiprudencia, por lo tanto las variables que forman parte del rendimiento académico del estudiante son las que poseen una alta influencia para que un estudiante deserte la carrera.

TABLA 3. 60. Ranking De Atributos – **Gainratioattributeeval – Administración De Empresas**

GainRatioAttributeEval	
Atributo	**Weight (Peso)**
ESTADO_APROBACION	0.22171863
SUPLETORIO	0.17640996
ASISTIO_SUPLETORIO	0.12974951
PRESENT_TODAS_LAS_EVAL	0.12308031
NOTA_FINAL	0.10044501
ESTADO_CIVIL	0.01539987
EDAD	0.01281425
TIPO_PAGO_MATRICULA	0.0029266
CURSO	0.00004186
GENERO	0.0000294
NIVEL_INTER_EST	0.00002542
NIVEL_INTER_PROF	0.0000011

En la [*Tabla 3.60*] se puede observar, los resultados del ranking según su nivel de importancia, los mismos que fueron propuestos por el algoritmo GainRationAtributeEval. Se pudo verificar a través de los presentes resultados lo que se estableció cuando se crearon los clusters, encontrando que el atributo *NIVEL_INTER_PROF*, es irrelevante en el dataset, ya que observando en las tablas descritas, dicho campo presenta el menor peso comparado con los otros atributos, de igual manera al ejecutar otros algoritmos evaluadores en Weka, suministraron como resultado, en la última posición el *NIVEL_INTER_PROF*.

- ***CARRERA: GESTIÓN AMBIENTAL***

A continuación se muestran los resultados que han proporcionado los algoritmos evaluadores, los mismos que ayudan a evaluar la calidad de los atributos y con ello conocer el ranking de los mismos:

TABLA 3. 61. Ranking De Atributos – **Chisquaredattributeeval –Gestión Ambiental**

ChiSquaredAttributeEval		
Atributo	**Ranked**	**Clase**
NOTA_FINAL	808.27983	DESERTOR
ESTADO_APROBACION	757.92552	
SUPLETORIO	521.83158	
ASISTIO_SUPLETORIO	378.51546	
PRESENT_TODAS_LAS_EVAL	304.25327	
EDAD	84.45013	
ESTADO_CIVIL	27.55783	
TIPO_PAGO_MATRICULA	23.62061	
GENERO	10.03355	
NIVEL_INTER_EST	0.97687	
CURSO	0.40159	
NIVEL_INTER_PROF	0.00288	

En la [*Tabla. 3.61*], se visualiza el orden de los atributos según el nivel de correlación que existe entre ellos, con respecto a la clase Desertor. El algoritmo establece que el atributo que posee una mayor relación respecto a la clases son: *Nota_Final, Estado_Aprobacion, Supletorio, Asistio_Supletorio*, el atributo Nota_Final es el que posee una mayor influencia, de igual manera como sucedió en las carreras de Jursiprudencia y Administración de Empresas, por lo tanto las variables que forman parte del rendimiento académico del estudiante son las que poseen una alta influencia para que un estudiante deserte la carrera.

TABLA 3. 62. Ranking De Atributos – **Gainratioattributeeval – Gestión Ambiental**

GainRatioAttributeEval	
Atributo	**Weight (Peso)**
ESTADO_APROBACION	0.184785878
SUPLETORIO	0.128718759
PRESENT_TODAS_LAS_EVAL	0.115345353
NOTA_FINAL	0.083079114
ASISTIO_SUPLETORIO	0.05502133
EDAD	0.012365436
ESTADO_CIVIL	0.008531536
TIPO_PAGO_MATRICULA	0.007605599
GENERO	0.002389994
NIVEL_INTER_EST	0.000142278
CURSO	0.000039525
NIVEL_INTER_PROF	0.000000927

En la [*Tabla 3.62*] se puede observar, los resultados del ranking según su nivel de importancia, los mismos que fueron propuestos por el algoritmo GainRationAtributeEval. El algoritmo retornó la variable NIVEL_INTER_PROF, como la menos relevante del dataset, dicha variable, también ha sido selecciona como la menos importante para la carrera de Administración de Empresas.

- ***CARRERA: INFORMÁTICA***

A continuación se muestran los resultados que han proporcionado los algoritmos evaluadores, los mismos que ayudan a medir la calidad de los atributos y con ello conocer el ranking de los atributos:

TABLA 3. 63. RANKING DE ATRIBUTOS – **CHISQUAREDATTRIBUTEEVAL – INFORMÁTICA**

ChiSquaredAttributeEval		
Atributo	**Ranked**	**Clase**
NOTA_FINAL	622.2835	DESERTOR
ESTADO_APROBACION	543.0159	
ASISTIO_SUPLETORIO	455.7962	
SUPLETORIO	398.7645	
PRESENT_TODAS_LAS_EVAL	231.8646	
EDAD	43.1183	
ESTADO_CIVIL	3.9702	
NIVEL_INTER_EST	1.7444	
CURSO	1.1524	
NIVEL_INTER_PROF	0.6713	
GENERO	0.0575	
TIPO_PAGO_MATRICULA	0.0411	

En la [*Tabla. 3.63*], se visualiza el orden de los atributos según el nivel de correlación que existe entre ellos, con respecto a la clase Desertor. El algoritmo establece que el orden de los atributos que poseen una mayor relación con respecto a la clases son: *Nota_Final, Estado_Aprobacion, Asistio_Supletorio*, *Supletorio,* siendo el atributo Nota_Final el que posee una mayor influencia, ya que se encuentra en primer lugar en el ranking, de igual manera como sucedió en las 3 carreras antes analizadas, por lo tanto las variables que forman parte del rendimiento académico del estudiante son las que poseen una alta influencia para que un estudiante deserte la carrera.

c. Clasificación

Para realizar la clasificación de los datos se ha utilizado Arboles de decisión, aplicando el algoritmo **J48**, el mismo que se basa en obtener medidas derivadas de las frecuencias relativas de las clases, y con ello pueda conseguir más nodos puros. El algoritmo J48 utiliza el proceso de poda, que significa eliminar condiciones de las ramas del árbol o de algunas reglas; los nodos que están por debajo del límite de poda se eliminan, ya que se consideran demasiado específicos. Las ventajas que ofrece la presente técnica, se encuentran detalladas en el Diseño de la Solución [*capítulo 3, sección 3.3.2*].

El algoritmo J48 posee algunas propiedades, en las mismas se estableció un 25% como nivel de confianza para la poda del árbol de decisión, y se fijó en 2 el mínimo de instancias por hoja, con cuyos valores se obtienen los mejores resultados con el presente algoritmo.

- ***Resultados de la Clasificación***
  - ***CARRERA: JURISPRUDENCIA***

Se ha creído conveniente aplicar arboles de decisión, seleccionando los atributos más relevantes respecto a la deserción, relacionados con la información académica que obtenido un estudiante en cada asignatura. Dichos atributos fueron seleccionados previo un análisis, con la ayuda del algoritmo evaluador **ChiSquaredAttributeEval** descrito en la [*Tabla 3.64*].

TABLA 3. 64. Resultados- Árboles De Decisión (J48) – Carrera De Jurisprudencia

Experimentos	Atributos Seleccionados	Atributos Clasificados
Experimentos 1	Curso	Supletorio,
	Nota_Final	Nota_Final
	Estado_Aprobación	Estado_Aprobación
	Nivel_Inter_Prof	Asistio_Supletorio.
	Nivel_Inter_Est	
	Present_todas_las_Eval	
	Supletorio	
	Asistio_Supletorio	
Experimentos 2	Curso	Curso
	Estado de Aprobación	Estado_Aprobación
Experimentos 3	Asistio_Supletorio	Asistio_Supletorio
	Estado_Aprobacion	Estado_Aprobacion
Experimentos 4	No_Materias_Troncales_Rep	No_Materias_Troncales_Rep
	Deserto	Deserto

Interpretación de los Arboles de Decisión: *Carrera de Jurisprudencia*

A continuación se detalla el análisis realizado, en cada uno de los experimentos con árboles de decisión, además se realiza una interpretación de la matriz de confusión generada en cada experimento:

- ***Árbol de Decisión:*** *Experimento 1*

Atributos del Data Set: Nota_Final, Estado_Aprobacion, Supletorio y Asistio_Supletorio.

Atributos Clasificados: Supletorio (Nodo Principal), Nota_Final, Estado_Aprobación y Asistio_Supletorio.

Tamaño de la Población: 5629 instancias.

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4646               82.5369 %
Incorrectly Classified Instances       983               17.4631 %
Kappa statistic                          0.7284
Mean absolute error                      0.1592
Root mean squared error                  0.2823
Relative absolute error                 36.5002 %
Root relative squared error             60.4529 %
Total Number of Instances             5629

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.226      0.745     0.895     0.813      0.892    SI
                 0.555     0.063      0.785     0.555     0.65       0.872    NO
                 1         0          1         1         1          1        NO_LE_CORRESPONDE
Weighted Avg.    0.825     0.114      0.829     0.825     0.818      0.917

=== Confusion Matrix ===

    a    b    c   <-- classified as
 2138  251    0 |    a = SI
  732  914    0 |    b = NO
    0    0 1594 |    c = NO_LE_CORRESPONDE
```

FIGURA 3. 30. Resultados J48 – Experimento 1 - Carrera De Jurisprudencia

En la [*Figura. 3.30*], ilustra que existe un total de 4646(82.53%) instancias correctamente clasificadas, y un total de 983(17.46%) clasificadas incorrectamente. Si la clasificación hubiera sido perfecta se esperaría encontrar únicamente elementos en la diagonal, en la matriz de confusión, por lo cual se puede observar que se tiene un total del 15.92% en promedio de error absoluto, al intentar clasificar todas las instancias.

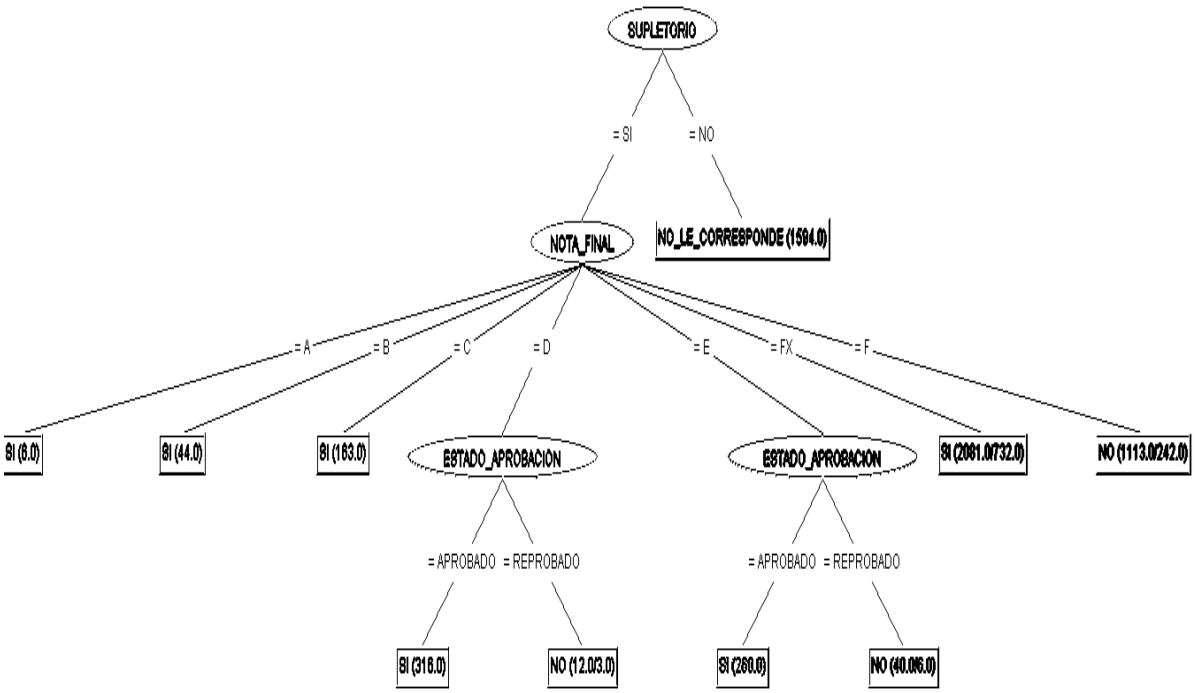

FIGURA 3. 31. Gráfica Del Árbol De Decisión – Carrera De Jurisprudencia

En la [*Figura. 3.31*] se visualiza el árbol de decisión, generado con 3 atributos del dataset principal, siendo estos Supletorio (Nodo Principal), Nota_Final, Estado_Aprobación y Asistio_Supletorio.

El árbol expresa que 6 estudiantes que han obtenido una Nota de A = 39 - 40 puntos, 44 estudiantes que han obtenido B = 36 – 38, y 163 estudiantes que han obtenido C = 33 – 35, todos ellos han Aprobado, obtenido dichas calificaciones dando el respectivo supletorio. Indica además el árbol que de los estudiantes que han obtenido una Nota D = 30 – 32, un total de 316 estudiantes han aprobado dando el respectivo supletorio; sin embargo existen 9 alumnos que han obtenido D = 30 – 32, teniendo que dar el respectivo supletorio pero no asistieron a dar la evaluación correspondiente por ende reprobaron. También existen 260 estudiantes que dando el supletorio han aprobado, obtenido E = 28 – 29 como calificación. Un total de 34 estudiantes que tienen E = 28 – 29, no se presentaron a dar el respectivo supletorio por ende han reprobaron la asignatura. Un total de 1349 estudiantes han obtenido FX = 14 – 27, a pesar que si se presentaron a dar el supletorio respectivo reprobaron. Un total de 871 estudiantes reprobaron teniendo F= 13 o menos como calificación, a pesar que si presentaron la evaluación correspondiente. En el árbol también se puede visualizar un total de 1594 estudiantes que si aprobaron sin necesidad de dar una evaluación supletoria.

El árbol nos indica que existen estudiantes que a pesar de haber obtenido D=30 – 32 como calificación, la misma que es suficiente para aprobar la asignatura; sin embargo dichos estudiantes, tuvieron que asistir al supletorio, ya que no han obtenido el puntaje suficiente en cada bimestre, es decir mínimo 14; dicha nota la han obtenido al sumar las calificaciones de los dos bimestres, por lo tanto han tenido que rendir el examen supletorio del bimestre que no han obtenido el puntaje adecuado, con todo no lograron aprobar la asignatura porque no asistieron a presentar la evaluación supletoria.

Existen estudiantes que a pesar, de obtener un buen puntaje en algunos de los dos bimestres, pueden reprobar la asignatura por no asistir a dar la evaluación supletoria del bimestre correspondiente.

También en el árbol se puede analizar que existe un porcentaje considerable de estudiantes que no se presentan a dar la evaluación supletorio, ya que poseen una nota sumamente baja como es F=13 o menos, por lo que ya no se sienten capacitados para poderla rendir, y deciden simplemente ya no estudiar y reprobarla.

- ***Árbol de Decisión:*** *Experimento 2*

Atributos del Data Set: Curso, Estado de Aprobación

Atributos Clasificados: Curso (Nodo Principal), Estado de Aprobación

Tamaño de la Población: 5629 instancias.

```
Number of Leaves  :     5

Size of the tree :      6


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3404               60.4726 %
Incorrectly Classified Instances      2225               39.5274 %
Kappa statistic                          0.1295
Mean absolute error                      0.4678
Root mean squared error                  0.4838
Relative absolute error                 95.8399 %
Root relative squared error             97.936  %
Total Number of Instances             5629

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.271     0.151      0.569     0.271     0.367      0.609    APROBADO
                 0.849     0.729      0.614     0.849     0.713      0.609    REPROBADO
Weighted Avg.    0.605     0.484      0.595     0.605     0.567      0.609

=== Confusion Matrix ===

    a    b   <-- classified as
  646 1735 |    a = APROBADO
  490 2758 |    b = REPROBADO
```

FIGURA 3. 32. Resultados Algoritmo J48 – Experimento 2 - Carrera De Jurisprudencia

En la [*Figura. 3.32*], ilustra que existe un total de 3404(60.47%) instancias correctamente clasificadas, y un total de 2225(39.52%) clasificadas incorrectamente. En el presente experimento se tiene 46.78% en el promedio de error absoluto, al intentar clasificar todas las instancias. La matriz de confusión nos muestra la distribución de los ejemplos por clase, en la cual podemos observar que el campo 'Estado de Aprobación' no es la mejor la clase para clasificar correctamente todos los datos del campo 'Curso'.

FIGURA 3. 33. Gráfica Del Árbol De Decisión – Experimento 2 - Carrera De Jurisprudencia

En el árbol se puede observar que la materia de Derecho Constitucional posee un total de 753 estudiantes que han 'Reprobado' de 1129 matriculados. La materia de Introducción al Derecho posee un total de 811 estudiantes que han 'Reprobado' de 1143. La materia de Metodología de Estudio posee un total de 646 estudiantes que han 'Aprobado', de un total de 1136 matriculados. La materia de Realidad Nacional posee un total de 532 estudiantes que han 'Reprobado' de un total de 1093. La materia de Expresión Oral posee un total de 632 estudiantes que han 'Reprobado', de un total de 1128 matriculados. [*Ver Figura. 4.33*]

La mayoría de estudiantes de 1er Ciclo de la carrera de Jurisprudencia, han aprobado con mayor frecuencia, las materias que forman parte de las asignaturas de Formación Básica, siendo estás las que implican menor complejidad para su aprendizaje, ya que el contenido de las mismas es abordado en el colegio.

- ***Árbol de Decisión:*** *Experimento 3*

Atributos del Data Set: Asistio_Supletorio, Estado_Aprobacion

Atributos Clasificados: Asistio_Supletorio (Nodo Principal), Estado_Aprobacion

Tamaño de la Población: 5629 instancias.

```
Size of the tree :      4


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        4842               86.0188 %
Incorrectly Classified Instances       787               13.9812 %
Kappa statistic                          0.7004
Mean absolute error                      0.1876
Root mean squared error                  0.3063
Relative absolute error                 38.4271 %
Root relative squared error             62.0075 %
Total Number of Instances             5629

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.669     0          1         0.669     0.802      0.912    APROBADO
                 1         0.331      0.805     1         0.892      0.912    REPROBADO
Weighted Avg.    0.86      0.191      0.887     0.86      0.854      0.912

=== Confusion Matrix ===

    a    b   <-- classified as
 1594  787 |    a = APROBADO
    0 3248 |    b = REPROBADO
```

FIGURA 3. 34. Resultados Algoritmo J48 – Experimento 3 - Carrera De Jurisprudencia

En la [*Figura. 3.34*], ilustra que existe un total de 4842(86.01%) instancias correctamente clasificadas, y un total de 787(13.98%) clasificadas incorrectamente. En el presente experimento se tiene un 18.76% en el promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar que las 787 instancias que debieron ser clasificadas en la clase 'APROBADO', han sido clasificadas en la clase de 'REPROBADO'.

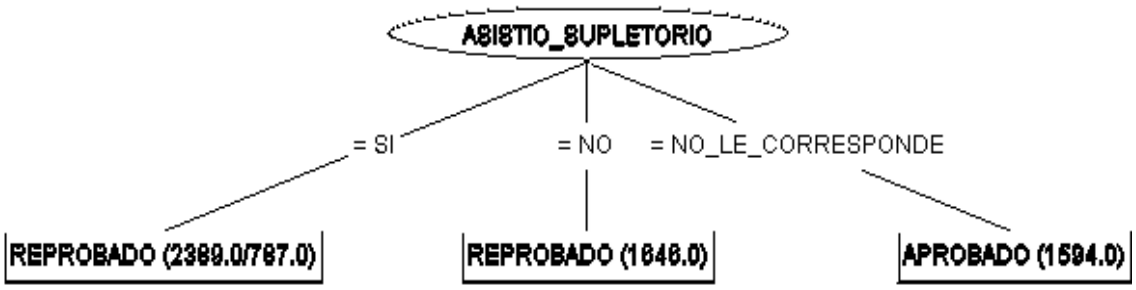

FIGURA 3. 35. Gráfica Del Árbol De Decisión – Experimento 3 - Carrera De Jurisprudencia

En el árbol se puede observar que de un total de 3248 estudiantes que reprobaron los cursos, tomando en cuenta todas las asignaturas de la carrera, un total de 1602 estudiantes Reprobaron la asignatura, si asistiendo a dar la respectiva evaluación supletoria, simplemente un total de 787 estudiantes Aprobaron la asignatura, si asistiendo a dar la respectiva evaluación supletoria. Un total de 1646 estudiantes, teniendo que rendir la respectiva evaluación supletoria, no han asistido a presentarla, por lo tanto han reprobado la asignatura. Un total de 1594 estudiantes han aprobado la asignatura directamente, es decir sin estar en supletorio. [*Ver Figura. 3.35*]

La mayoría de estudiantes que han cursado las asignaturas de 1er ciclo de la carrera de Jurisprudencia, no asisten a dar la evaluación supletoria, por ende han reprobado la asignatura, ya sea porque los estudiantes no están lo suficientemente preparados para dicha evaluación por tanto no asisten a presentarla y reprueban.

- ***Árbol de Decisión:*** *Experimento 4*

Atributos del Data Set: NoMateriasTroncalesRep, Deserto.

Atributos Clasificados: NoMateriasTroncalesRep, Deserto (Nodo Principal).

Tamaño de la Población: 1302 instancias.

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1033               79.3395 %
Incorrectly Classified Instances       269               20.6605 %
Kappa statistic                          0.53
Mean absolute error                      0.2514
Root mean squared error                  0.3547
Relative absolute error                 68.6339 %
Root relative squared error             82.9021 %
Total Number of Instances             1302

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.77      0.134      0.948     0.77      0.85       0.801    SI
                 0.866     0.23       0.545     0.866     0.669      0.801    NO
Weighted Avg.    0.793     0.157      0.851     0.793     0.806      0.801

=== Confusion Matrix ===

   a   b   <-- classified as
 761 227 |   a = SI
  42 272 |   b = NO
```

FIGURA 3. 36. Resultados Algoritmo J48 – Experimento 4 - Carrera De Jurisprudencia

En la [*Figura. 3.36*], ilustra que existe un total de 1033(79.33%) instancias correctamente clasificadas, y un total de 269(20.66%) clasificadas incorrectamente. En el presente experimento se tiene un 25.14% de promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar la distribución de los ejemplos, por cada clase.

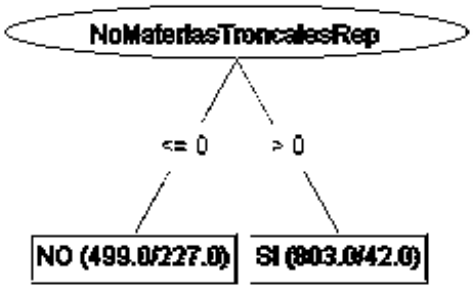

FIGURA 3. 37. Gráfica Del Árbol De Decisión – Experimento 4 - Carrera De Jurisprudencia

Para el presente experimento se ha creído conveniente analizar información del número materias troncales reprobadas; ya que se pudo constatar en el experimento anterior que los estudiantes reprueban con mayor frecuencias las materias troncales, y cuando se construyeron los clusters se comprobó que la mayoría de estudiantes que reprueban en alguna asignatura, desertan la carrera, por lo cual ahora se analizaron dichas variables, para con ello analizar de mejor manera la deserción estudiantil. [*Ver Figura. 3.37*]

El árbol nos muestra que un total de 272 estudiantes, que no reprobaron ninguna materia troncal de 1er ciclo, como resultado de ello no desertaron la carrera, solo un total de 227 que no reprobaron en alguna materia troncal, desertaron la carrera. Se observa en el árbol además que, un total de 761 estudiantes que si reprobaron en alguna o ambas, de las materias troncales correspondientes, como resultado de ello desertaron la carrera, y solo un total de 42 estudiantes que reprobaron en una o más materias troncales, no decidieron desertar la carrera. [*Ver Figura. 3.37*]

La mayoría de estudiantes que han desertado la carrera han reprobado en alguna o ambas materias troncales ofertadas en primer ciclo de la carrera de Jurisprudencia. Tomando en cuenta que las Materias Troncales ofertadas en cada carrera, implican aplicarle mayor tiempo para su comprensión; ya que el contenido de las mismas es fundamental en la carrera. Existen estudiantes que además no poseen las bases necesarias para comprender con mayor facilidad dicho contenido de la asignatura, puesto que no han elegido la carrera idónea, según su perfil profesional.

- ***CARRERA: ADMINISTRACIÓN DE EMPRESAS***

A continuación se muestran los atributos que se han seleccionado para cada experimento con árboles de decisión, para con ello analizar, de forma específica la deserción estudiantil Dichos atributos fueron seleccionados previo un análisis y con la ayuda del algoritmo evaluador **ChiSquaredAttributeEval** descrito en la [*Tabla 3.65*].

TABLA 3. 65. Resultados- Árboles De Decisión (J48) – Carrera De Administración De Empresas.

Experimentos	Atributos Seleccionados	Atributos Clasificados
Experimentos 1	Nota_Final	Nota_Final
		Estado_Aprobación
	Estado_Aprobación	Present_todas_las_Eval
		Supletorio
	Present_todas_las_Eval	
	Supletorio	
Experimentos 2	Curso	Curso
	Estado de Aprobación	Estado_Aprobación
	Asistio_Supletorio	Asistio_Supletorio
Experimentos 4	No_Materias_Troncales_Rep	No_Materias_Troncales_Rep
	Deserto	Deserto

Interpretación de los Arboles de Decisión: *Carrera de Administración de Empresas*

A continuación se detalla el análisis realizado, en cada uno de los experimentos con árboles de decisión, además se realiza una interpretación de la matriz de confusión generada en cada experimento:

- ***Árbol de Decisión:*** *Experimento 1*

Atributos del Data Set: Nota_Final, Estado_Aprobación, Present_todas_las_Eval, Supletorio.

Atributos Clasificados: Estado_Aprobacion (Nodo Principal) Nota_Final, Estado_Aprobación.

Tamaño de la Población: 4243 instancias.

```
Correctly Classified Instances        3783               89.1586 %
Incorrectly Classified Instances       460               10.8414 %
Kappa statistic                          0.7426
Mean absolute error                      0.1499
Root mean squared error                  0.2741
Relative absolute error                 36.6871 %
Root relative squared error             60.6532 %
Total Number of Instances             4243

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0.903     0.138      0.942     0.903     0.922      0.943
                 0.862     0.097      0.781     0.862     0.82       0.943
Weighted Avg.    0.892     0.126      0.896     0.892     0.893      0.943

=== Confusion Matrix ===

    a    b   <-- classified as
 2736  293 |    a = SI
  167 1047 |    b = NO
```

FIGURA 3. 38. Resultados Algoritmo J48 – Experimento 1 - Carrera De Administración De Empresas

En la [*Figura. 3.38*], ilustra que existe un total de 3783(89.15%) instancias correctamente clasificadas, y un total de 460(10.84%) clasificadas incorrectamente. En el presente experimento se tiene un 14.99% de promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar la distribución de los ejemplos, por cada clase, los que han sido clasificados en cada clase de forma correcta he incorrecta.

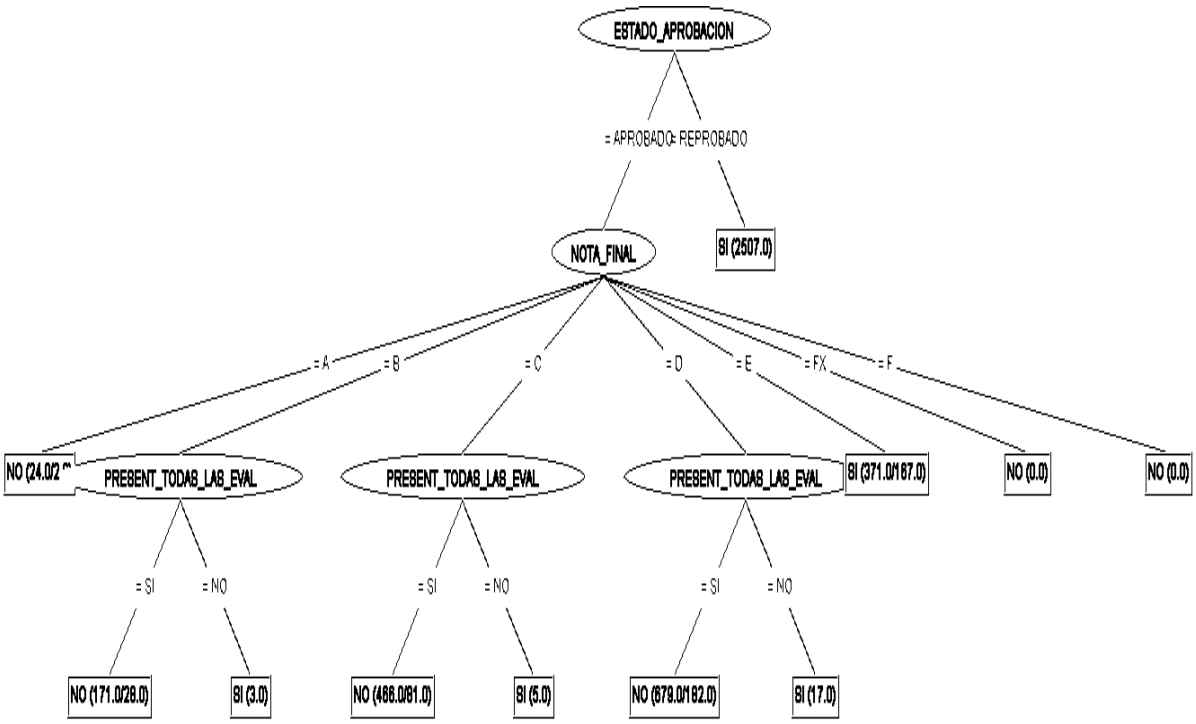

FIGURA 3. 39. Gráfica Del Árbol De Decisión – Experimento 1 - Carrera De Administración De Empresas

El árbol expresa que de 4243 estudiantes matriculados en la carrera de Administración de Empresas un total de 22 han aprobado, obtenido una Nota de A = 39 - 40 puntos, dichos estudiantes han obtenido la mencionada calificación sin tener que rendir un supletorio, tomando en cuenta que A es la nota máxima que puede obtener un estudiante. Un total de 143 estudiantes han aprobado la asignatura, obtenido una Nota de B = 36 – 38, presentado todas las evaluaciones y sin tener que tener que dar la correspondiente evaluación supletoria, solo un total de 3 estudiantes han obtenido el mencionado puntaje, asistiendo a dar el correspondiente supletorio, ya que no han presentado las respectivas evaluaciones de la materia. Un total de 385 estudiantes que han aprobado la asignatura han obtenido C = 33 – 35 como puntaje, todos ellos poseen dichas calificaciones sin tener que rendir la evaluación supletoria, ya que si presentaron todas las evaluaciones correspondientes, solo un total de 5 estudiantes obtuvieron la calificación de C, dando el respectivo supletorio, ya que no presentaron todas las evaluaciones en el periodo académico. [*ver Figura. 3.39]*

Indica además el árbol que de los estudiantes que han obtenido una Nota D = 30 – 32, un total de 497 estudiantes han aprobado dando el respectivo supletorio; sin embargo existen 17 alumnos que han obtenido D = 30 – 32, teniendo que dar el respectivo supletorio, de los cuales no han asistido a dar las correspondientes evaluaciones de la asignatura. También existen 204 estudiantes que han aprobado, obteniendo E = 28 – 29 como calificación. Indica además el árbol que existen 2407 estudiantes que han reprobado la asignatura de la carrera de Administración de Empresas. [*ver Figura. 3.39***]**

Los estudiantes que han aprobado la asignatura directamente, han presentado todas las evaluaciones que correspondientes a las asignaturas que cursan, tomando en cuenta que han obtenido puntajes suficientes en dichas evaluaciones para no tener que asistir a un supletorio. Además existen en su minoría estudiantes que han aprobado la asignatura teniendo que rendir una evaluación supletoria, obteniendo en ella una nota suficiente para poder aprobar la materia.

- ***Árbol de Decisión:*** *Experimento 2*

Atributos del Data Set: Curso, Estado de Aprobación

Atributos Clasificados: Curso (Nodo Principal), Estado de Aprobación

Tamaño de la Población: 4243 instancias

```
Correctly Classified Instances        2758               65.0012 %
Incorrectly Classified Instances      1485               34.9988 %
Kappa statistic                          0.4787
Mean absolute error                      0.2975
Root mean squared error                  0.3862
Relative absolute error                 67.2689 %
Root relative squared error             82.1278 %
Total Number of Instances             4243

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.209     0.107      0.516     0.209      0.297      0.559    SI
                 0.807     0.246      0.648     0.807      0.719      0.844    NO
                 1         0.172      0.699     1          0.823      0.909    NO_LE_CORRESPONDE
Weighted Avg.    0.65      0.176      0.616     0.65       0.599      0.762

=== Confusion Matrix ===

   a    b    c   <-- classified as
 314  669  522 |    a = SI
 294 1230    0 |    b = NO
   0    0 1214 |    c = NO_LE_CORRESPONDE
```

FIGURA 3. 40. Resultados Algoritmo J48 – Experimento 2 - Carrera De Administración De Empresas

En la [*Figura. 3.40*], se observa que existe un total de 2758(89.15%) instancias correctamente clasificadas, y un total de 1485(10.84%) clasificadas incorrectamente. En el presente experimento se tiene un 29.75% de promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar la distribución de los ejemplos, por cada clase, se puede observar, que las instancias que representan a la variable a y b, están en una cierta cantidad clasificadas incorrectamente en clases que no corresponden.

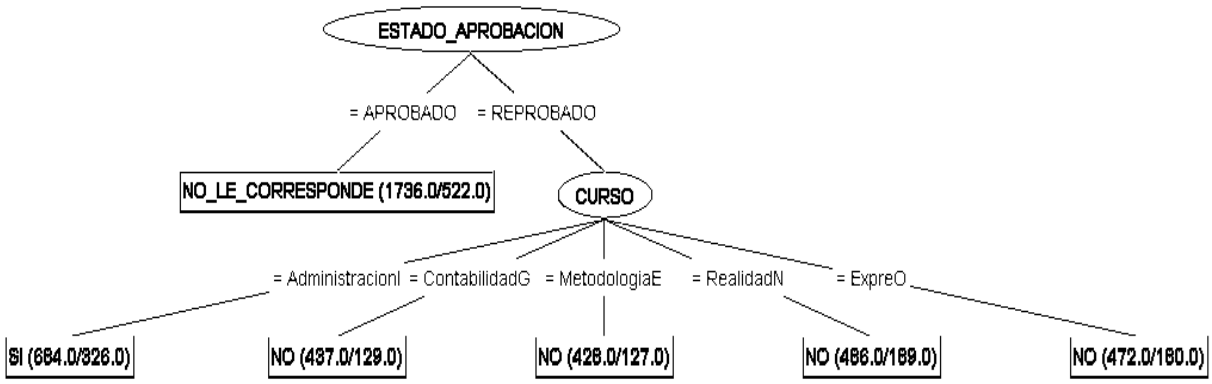

FIGURA 3. 41. Gráfica Del Árbol De Decisión – Experimento 2 - Carrera De Administración De Empresas

El árbol ilustrado en la [*Figura. 3.41*] indica, que de 1736 estudiantes, que han aprobado las asignaturas de primer ciclo de la carrera de Administración de empresas, un total de 1214 han aprobado las asignaturas directamente sin tener que rendir una evaluación supletoria; solo un total de 522 han aprobado las asignaturas, teniendo que dar el respectivo supletorio. Se puede observar además en el árbol que de 2507 estudiantes que han reprobado las asignaturas de 1er ciclo de la carrera de Administración de Empresas, 358 son de la materia

de Administración I, 308 son de la materia de Contabilidad General, 301 son de la materia de Metodología de Estudio, 297 pertenecen a la asignatura de Realidad Nacional, y 292 son de Expresión Oral; dichos estudiantes han reprobado las asignaturas, no asistiendo a dar la correspondiente evaluación supletoria.

El árbol nos indica, que la mayoría de estudiantes reprobados se encuentran en la materia de Administración I, que corresponden a las materias troncales de la carrera, y en la materia que reprueban con menor frecuencia los estudiantes es Metodología de Estudio, tomando en cuenta que dicha materia corresponde al grupo de asignaturas de Formación Básica de la Carrera.

- ***Árbol de Decisión:*** *Experimento 3*

Atributos del Data Set: NoMateriasTroncalesRep , Deserto.

Atributos Clasificados: NoMateriasTroncalesRep (Nodo Principal), Deserto.

Tamaño de la Población: 988 instancias.

```
Correctly Classified Instances         855               86.5385 %
Incorrectly Classified Instances       133               13.4615 %
Kappa statistic                          0.6335
Mean absolute error                      0.178
Root mean squared error                  0.2988
Relative absolute error                 57.2196 %
Root relative squared error             75.8048 %
Total Number of Instances              988

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.86      0.111      0.97      0.86      0.912      0.847    SI
                 0.889     0.14       0.601     0.889     0.718      0.847    NO
Weighted Avg.    0.865     0.116      0.899     0.865     0.874      0.847

=== Confusion Matrix ===

   a   b   <-- classified as
 686 112 |   a = SI
  21 169 |   b = NO
```

FIGURA 3. 42. Resultados Algoritmo J48 – Experimento 3 - Carrera De Administración De Empresas

En la [*Figura. 3.42*], ilustra que existe un total de 855(86.53%) instancias correctamente clasificadas, y un total de 133(13.46%) clasificadas incorrectamente. En el presente experimento se tiene un 17.8% de promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar que existen instancias incorrectamente clasificadas en las variables SI y NO, de la clase Desertor.

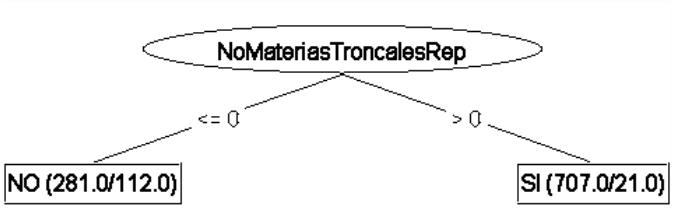

FIGURA 3. 43. Gráfica Del Árbol De Decisión – Experimento 3 - Carrera De Administración De Empresas

De igual manera como se realizó el experimento, del número de materias troncales reprobadas en la carrera de Jurisprudencia, se procede a ejecutar lo mismo en la carrera de Administración de Empresas, ya que según lo analizado en el experimento anterior, se pudo constatar que los estudiantes reprueban con mayor frecuencias las materias troncales, y cuando se construyeron los clusters se comprobó que la mayoría de estudiantes que reprueban en alguna asignatura troncal, desertan la carrera, por lo cual ahora se creyó conveniente realizar un análisis dichas variables, para con ello observar el comportamiento de los estudiantes desertores. [*Ver Figura. 3.43*]

El árbol nos muestra que un total de 190 estudiantes, que no reprobaron ninguna materia troncal de 1er ciclo, como resultado de ello un total de 169 estudiantes no desertaron la carrera. Se observa en el árbol además que, un total de 798 estudiantes que si reprobaron en alguna o ambas materias troncales, como resultado de ello un total de 686 estudiantes desertaron la carrera, y solo un total de 21 estudiantes que reprobaron en una o más materias troncales, no decidieron desertar la carrera. [*Ver Figura. 3.43*]

La mayoría de estudiantes que han decidido desertar la carrera de Administración de Empresas, han reprobado una o ambas asignaturas troncales; como de igual manera existe gran parte de estudiantes que al momento de no reprobar ninguna materia, correspondientes a las troncales de la carrera, no constan como desertores. Es decir que al momento que un estudiante de 1er ciclo repruebe en alguna de las materias troncales sería un posible desertor de la carrera de Administración de Empresas.

- ***CARRERA: GESTIÓN AMBIENTAL***

A continuación se muestran los atributos seleccionados para cada experimento, con árboles de decisión, para con ello analizar, de forma específica la deserción estudiantil. Dichos atributos forman parte del rendimiento académico del estudiante, los mismos que fueron seleccionados con la ayuda del algoritmo evaluador **ChiSquaredAttributeEval** descrito en la [*Tabla 3.66*].

TABLA 3. 66. Resultados- Árboles De Decisión (J48) – Carrera De Gestión Ambiental

Experimentos	Atributos Seleccionados	Atributos Clasificados
Experimentos 1	Curso	Curso
	Estado_Aprobacion	Estado_Aprobacion
Experimentos 2	Estado de Aprobación	Estado_Aprobacion
	Supletorio	Supletorio
	Asistio_Supletorio	Asistio_Supletorio
Experimentos 4	No_Materias_Troncales_Rep	No_Materias_Troncales_Rep
	Deserto	Deserto
	Estado_Civil	Estado_Civil

Interpretación de los Arboles de Decisión: *Carrera de Gestión Ambiental*

A continuación se detalla el análisis realizado, en cada uno de los experimentos con árboles de decisión, además se realiza una interpretación de la matriz de confusión generada en cada experimento:

- ***Árbol de Decisión:*** *Experimento 1*

Atributos del Data Set: Curso, Estado_Aprobacion

Atributos Clasificados: Curso (Nodo Principal), Estado_Aprobación.

Tamaño de la Población: 3141 instancias.

```
Correctly Classified Instances        1916               60.9997 %
Incorrectly Classified Instances      1225               39.0003 %
Kappa statistic                          0.1989
Mean absolute error                      0.454
Root mean squared error                  0.4769
Relative absolute error                 92.7698 %
Root relative squared error             96.4197 %
Total Number of Instances             3141

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.523     0.326      0.545     0.523     0.534      0.631    APROBADO
                 0.674     0.477      0.655     0.674     0.665      0.631    REPROBADO
Weighted Avg.    0.61      0.412      0.608     0.61      0.609      0.631

=== Confusion Matrix ===

   a    b   <-- classified as
 702  639 |    a = APROBADO
 586 1214 |    b = REPROBADO
```

FIGURA 3. 44. Resultados Algoritmo J48 – Experimento 1 - Carrera De Gestión Ambiental

En la [*Figura. 3.44*], ilustra que existe un total de 1916(60.99%) instancias correctamente clasificadas, y un total de 1225(39.00%) clasificadas incorrectamente. En el presente experimento se tiene un 45.40% de promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar la distribución de los ejemplos, por cada clase, los que han sido clasificados en las clases que representan 'a' y 'b', de forma correcta he incorrecta.

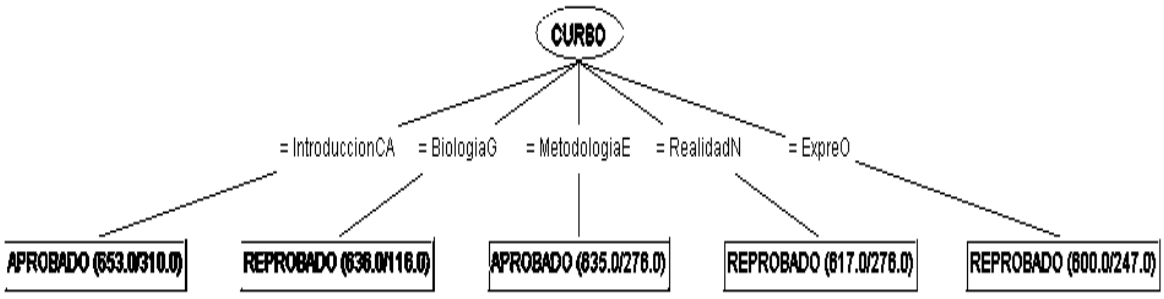

FIGURA 3. 45. Gráfica Del Árbol De Decisión – Experimento 1 - Carrera De Administración De Empresas

El árbol expresa, que de 1341 aprobados, 343 han sido de la materia de Introducción a las Ciencias Ambientales de un total de 653 matriculados, 116 de Biología General de 636 matriculados, 359 de Metodología de Estudio de un total de 635 matriculados, 276 de Realidad

Nacional de un total de 617 matriculados, y 247 de Expresión Oral de un total de 600 matriculados. [*ver Figura. 3.45*]

Podemos observar que una considerable cantidad de estudiantes han aprobado Introducción al Derecho, considerando que esta es una materia troncal, y suele ser más complicado el aprendizaje de la misma. En la materia de Biología General existen pocos estudiantes, que han aprobado, ya que esta es también una materia troncal.

Además se puede constatar que la mayoría de estudiantes, que han cursado Realidad Nacional y Expresión Oral, han reprobado dichas asignaturas, a diferencia de los que han cursado Metodología de Estudio que en su mayoría han aprobado.

- ***Árbol de Decisión:*** *Experimento 2*

Atributos del Data Set: Asistio_Supletorio, Supletorio, Estado_Aprobacion

Atributos Clasificados: Supletorio (Nodo Principal)**,** Asistio_Supletorio (Nodo Principal), Estado_Aprobacion

Tamaño de la Población: 3141 instancias.

```
Correctly Classified Instances        2290               72.9067 %
Incorrectly Classified Instances       851               27.0933 %
Kappa statistic                          0.6054
Mean absolute error                      0.1905
Root mean squared error                  0.3086
Relative absolute error                 43.2926 %
Root relative squared error             65.8016 %
Total Number of Instances             3141

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.323     0          1         0.323     0.488      0.824    SI
                 1         0.388      0.527     1         0.69       0.799    NO
                 1         0          1         1         1          1        NO_LE_CORRESPONDE
Weighted Avg.    0.729     0.117      0.857     0.729     0.702      0.869

=== Confusion Matrix ===

   a   b   c   <-- classified as
 406 851   0 |   a = SI
   0 949   0 |   b = NO
   0   0 935 |   c = NO_LE_CORRESPONDE
```

FIGURA 3. 46. Resultados Algoritmo J48 – Experimento 2 – Carrera De Gestión Ambiental

En la [*Figura. 3.46*], ilustra que existe un total de 2290(72.90%) instancias correctamente clasificadas, y un total de 851(27.09%) clasificadas incorrectamente. En el presente experimento se tiene un 19.05% en el promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar que las instancias que clasifican a las variables representadas por ‘a’ y ‘b’, poseen valores que no han sido clasificados en las clases que corresponden.

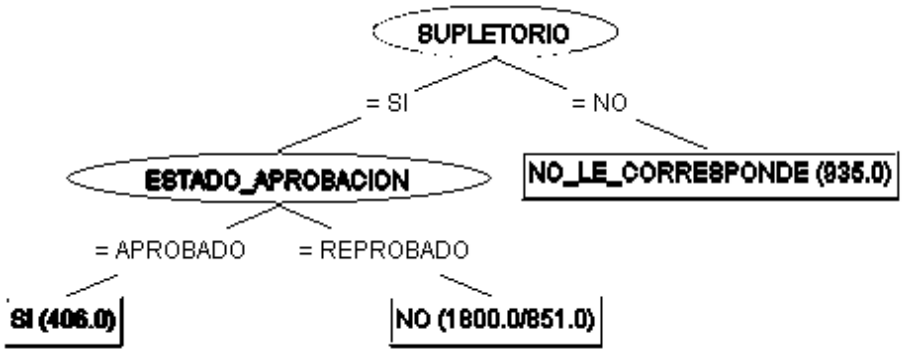

FIGURA 3. 47. Gráfica Del Árbol De Decisión – Experimento 2 - Carrera De Gestión Ambiental

En el árbol se puede observar que de un total de 2206 estudiantes que se han quedado en supletorio, tomando en cuenta todas las 5 asignaturas de 1er ciclo de la carrera de Gestión Ambiental, un total de 406 han aprobado la asignatura, presentado la correspondiente evaluación supletoria, y un total de 949 no se han presentado a rendir la correspondiente evaluación por ende han reprobado la asignatura, además un total de 851 estudiantes, que si se presentaron a rendir la evaluación, de igual manera reprobaron, ya que no obtuvieron el puntaje requerido. Se observa además que un total de 935 estudiantes, han aprobado la asignatura directamente. [*Ver Figura. 3.47*]

Se ha podido constatar en el presente árbol, que la mayoría de estudiantes que si se presentaron a rendir la evaluación supletoria, no aprobaron, ya sea por la falta de preparación para la prueba, o por desinterés en la materia. Existe además un número considerable de estudiantes, que sabiendo que están en supletorio, de la asignatura, no asisten a dar la correspondiente evaluación, por lo cual reprueban.

- ***Árbol de Decisión:*** *Experimento 3*

Atributos del Data Set: NoMateriasTroncalesRep , Deserto, Estado Civil.

Atributos Clasificados: NoMateriasTroncalesRep (Nodo Principal), Deserto, Estado Civil.

Tamaño de la Población: 714 instancias.

```
Correctly Classified Instances         542               75.9104 %
Incorrectly Classified Instances       172               24.0896 %
Kappa statistic                          0.1544
Mean absolute error                      0.2978
Root mean squared error                  0.3869
Relative absolute error                 81.0008 %
Root relative squared error             90.2877 %
Total Number of Instances              714

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.946     0.827      0.782     0.946     0.856      0.741    SI
                 0.173     0.054      0.508     0.173     0.259      0.741    NO
Weighted Avg.    0.759     0.639      0.715     0.759     0.711      0.741

=== Confusion Matrix ===

   a   b   <-- classified as
 512  29 |   a = SI
 143  30 |   b = NO
```

FIGURA 3. 48. Resultados Algoritmo J48 – Experimento 3 - Carrera De Gestión Ambiental

En la [*Figura. 3.48*], ilustra que existe un total de 542(75.91%) instancias correctamente clasificadas, y un total de 172(24.08%) clasificadas incorrectamente. En el presente experimento se tiene un 29.78% de promedio de error absoluto, es decir, que si se han clasificado, sin errores la mayoría de instancias. En la matriz de confusión se puede observar que existen instancias incorrectamente clasificadas en los valores de SI y NO, de la clase Desertor.

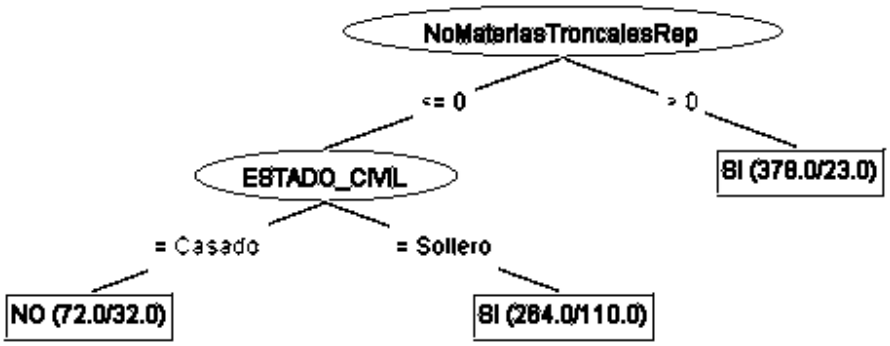

FIGURA 3. 49. Gráfica Del Árbol De Decisión – Experimento 3 - Carrera De Gestión Ambiental

En el árbol de decisión que se ilustra en la *[Figura. 3.49]* se puede observar, que de un total de 173 estudiantes no desertores, 150 no han desertado la carrera, ya que no han reprobado, ninguna materia troncal de 1er ciclo, de los cuales la mayoría son solteros.

Se puede visualizar en el árbol también que, de un total de 541 estudiantes desertores, 355 han reprobado al menos una materia troncal, los cuales constan como desertores en la carrera. [*Ver Figura. 3.49*]

En el presente análisis se ha podido constatar, el mismo suceso, que se verifico en las carreras anteriores, ya que la mayoría de estudiantes de Gestión Ambiental, han decidido desertar la carrera, porque han reprobado al menos una materia troncal; y como también existe estudiantes, que como han aprobado al menos una materia troncal, los mismos, no han decidido desertar la carrera; por lo tanto los estudiantes que reprueban al menos una materia troncal son más propensos a desertar la carrera. [*Ver Figura. 3.49*]

- ***CARRERA: INFORMÁTICA***

A continuación se muestran los atributos seleccionados para cada experimento, con árboles de decisión, para con ello analizar, de forma específica la deserción estudiantil. Dichos atributos forman parte del rendimiento académico del estudiante, los mismos que fueron seleccionados con la ayuda del algoritmo evaluador **ChiSquaredAttributeEval** descrito en la [*Tabla 3.67*].

TABLA 3. 67. Resultados- Árboles De Decisión (J48) – Carrera De Informática

Experimentos	**Atributos Seleccionados**	**Atributos Clasificados**
Experimentos 1	Estado de Aprobación	Estado_Aprobacion
	Supletorio	Supletorio
	Asistio_Supletorio	Asistio_Supletorio
Experimentos 2	No_Materias_Troncales_Rep	No_Materias_Troncales_Rep
	Deserto	Deserto

Interpretación de los Arboles de Decisión: *Carrera de Informática*

A continuación se detalla el análisis realizado, en cada uno de los experimentos con árboles de decisión, además se realiza una interpretación de la matriz de confusión generada en cada uno de los experimentos:

- ***Árbol de Decisión:*** *Experimento 1*

Atributos del Data Set: Asistio_Supletorio, Supletorio, Estado_Aprobacion

Atributos Clasificados: Asistio_Supletorio (Nodo Principal), Estado_Aprobacion

Tamaño de la Población: 1903 instancias.

```
Correctly Classified Instances        1659               87.1781 %
Incorrectly Classified Instances       244               12.8219 %
Kappa statistic                          0.7113
Mean absolute error                      0.1678
Root mean squared error                  0.2897
Relative absolute error                 35.4159 %
Root relative squared error             59.5287 %
Total Number of Instances             1903

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.667     0          1         0.667     0.8        0.93     APROBADO
                 1         0.333      0.827     1         0.906      0.93     REPROBADO
Weighted Avg.    0.872     0.205      0.894     0.872     0.865      0.93

=== Confusion Matrix ===

   a    b   <-- classified as
 489  244 |    a = APROBADO
   0 1170 |    b = REPROBADO
```

FIGURA 3. 50. Resultados Algoritmo J48 – Experimento 1 – Carrera De Informática

En la [*Figura. 3.50*], ilustra que existe un total de 1659(87.17%) instancias correctamente clasificadas, y un total de 244(27.09%) clasificadas incorrectamente. En el presente experimento se tiene un 16.78% en el promedio de error absoluto, al intentar clasificar todas las instancias. En la matriz de confusión se puede observar que las instancias que clasifican a las variables representadas por 'a' y 'b', poseen valores que no han sido clasificados en las clases correspondientes.

FIGURA 3. 51. Gráfica Del Árbol De Decisión – Experimento 1 - Carrera De Informática

En árbol que se muestra en la [*Figura. 3.51]*, se puede observar que de 1414 estudiantes, que se han quedado en supletorio, un total de 461 estudiantes, ha reprobado la asignatura, si asistiendo a rendir la respectiva evaluación; además se visualiza que 709 estudiantes han reprobado no asistiendo a rendir la correspondiente evaluación supletoria, se visualiza que solamente un total de 489 estudiantes han aprobado la asignatura, directamente, es decir sin tener que rendir un examen supletorio.

Luego de un previo análisis del presente árbol de decisión, se ha podido constatar, que la mayoría de estudiantes no asisten a rendir la evaluación supletoria que les corresponde, por lo tanto reprueban la asignatura. Como también se observó, al aplicar la técnica de cluster, que gran parte de los estudiantes que no presentan las evaluaciones tanto presencial como a distancia, de igual manera no asisten a rendir la evaluación supletoria correspondiente.

- ***Árbol de Decisión:*** *Experimento 2*

Atributos del Data Set: NoMateriasTroncalesRep, Deserto.

Atributos Clasificados: NoMateriasTroncalesRep (Nodo Principal), Deserto.

Tamaño de la Población: 446 instancias.

```
Correctly Classified Instances         376               84.3049 %
Incorrectly Classified Instances        70               15.6951 %
Kappa statistic                          0.5764
Mean absolute error                      0.1936
Root mean squared error                  0.3118
Relative absolute error                 63.6987 %
Root relative squared error             80.0986 %
Total Number of Instances              446

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.837     0.133      0.965     0.837     0.897      0.819    SI
                 0.867     0.163      0.55      0.867     0.673      0.819    NO
Weighted Avg.    0.843     0.138      0.888     0.843     0.855      0.819

=== Confusion Matrix ===

   a   b   <-- classified as
 304  59 |   a = SI
  11  72 |   b = NO
```

FIGURA 3. 52. Resultados Algoritmo J48 – Experimento 2 - Carrera De Informática

En la [*Figura. 3.52*], ilustra que existe un total de 376(84.30%) instancias correctamente clasificadas, y un total de 70(15.69%) clasificadas incorrectamente. En el presente experimento se tiene un 19.36% de promedio de error absoluto, es decir, que si se han clasificado, sin errores la mayoría de instancias. Se observa en la matriz de confusión, que existen instancias incorrectamente clasificadas en los valores de SI y NO, de la clase Desertor.

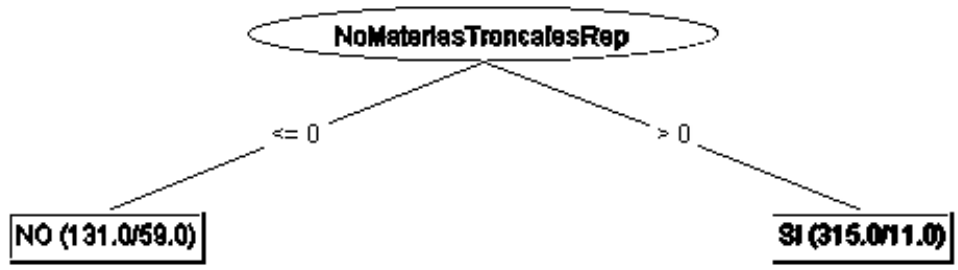

FIGURA 3. 53. Gráfica Del Árbol De Decisión – Experimento 2 - Carrera De Informática

Se puede visualizar en el árbol que, de un total de 83 estudiantes no desertores, 72 no han reprobado ninguna materia del grupo de las troncales de 1er ciclo de la carrera. Además se puede observar que de 363 estudiantes, que han desertado la carrera, un total de 304, han reprobado al menos 1 materia troncal perteneciente a la carrera. [*Ver Figura. 3.53*]

Se ha podido constatar, con el análisis del presente árbol, el mismo comportamiento, que se verifico en las carreras anteriores, ya que la mayoría de estudiantes de 1er ciclo de Informática, han decidido desertar la carrera, porque han reprobado al menos una materia troncal; y los estudiantes, que han aprobado al menos una materia troncal, no han desertado; por lo tanto existe mayor posibilidad de deserción, en los estudiantes que reprueban al menos una materia troncal.

d. Asociación

Para generar las reglas de asociación se ha creído conveniente aplicar el algoritmo, A priori, el mismo que es el más utilizado en la presente técnica, con el fin de obtener las 10 mejores reglas con un mínimo de confianza de 0.9.

El algoritmo A priori utiliza recursividad por niveles. En un primer paso genera los candidatos y seguidamente los pone a prueba para descartar los itemsets no frecuentes.

Al aplicar el algoritmo A priori, se establecen algunas propiedades como son el número de reglas (numRules), en donde se especificó 10; se estableció un mínimo de confianza (minMetric) de 0.9, que indica el número de casos que predice la regla correctamente; y en soporte [support] o cobertura se especificó el valor de 0.4, que indica el número de casos, ejemplos que cubre la regla.

Algunas de las ventajas que ofrece la presente técnica, se encuentran detalladas en el Diseño de la Solución [*capítulo 3, sección 2.3.3*].

- ***Resultados de la Técnica de Asociación***
    - ***CARRERA: JURISPRUDENCIA***

Para generar las mejores reglas de asociación se creyó conveniente, seleccionar ciertos atributos para el experimento que se ha realizado. Al ejecutar el algoritmo A priori, se obtuvieron las 10 mejores reglas, cuyos resultados se muestran a continuación: [*ver Tabla 3.68*]

TABLA 3. 68. Resultados- Reglas De Asociación– Carrera De Jurisprudencia

Experimentos	Atributos Seleccionados	Atributos Asociados	Mejores Reglas Seleccionadas
Experimento 1	Curso Nota_Final Estado_Aprobación Nivel_Inter_Prof Nivel_Inter_Est Pre- sent_todas_las_Eval Supletorio Asistio_Supletorio	Estado_Aprobacion Asistio_Supletorio Supletorio Nota_Final Present_todas_las_Eval	Regla 3 Regla 4 Regla 5 Regla 10

A continuación se muestran los resultados del modelo de reglas de asociación para la presente carrera:

Tamaño de la Población: 5629 instancias.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.4 (2252 instances)
Minimum metric <confidence>: 0.5
Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

 1. ESTADO_APROBACION=REPROBADO 3248 ==> SUPLETORIO=SI 3248    conf:(1)
 2. ASISTIO_SUPLETORIO=SI 2389 ==> SUPLETORIO=SI 2389    conf:(1)
 3. ESTADO_APROBACION=APROBADO 2381 ==> PRESENT_TODAS_LAS_EVAL=SI 2304    conf:(0.97)
 4. SUPLETORIO=SI 4035 ==> ESTADO_APROBACION=REPROBADO 3248    conf:(0.8)
 5. NIVEL_INTER_PROF=Alto 3406 ==> PRESENT_TODAS_LAS_EVAL=SI 2466    conf:(0.72)
 6. NIVEL_INTER_PROF=Alto 3406 ==> SUPLETORIO=SI 2369    conf:(0.7)
 7. SUPLETORIO=SI 4035 ==> PRESENT_TODAS_LAS_EVAL=SI 2505    conf:(0.62)
 8. PRESENT_TODAS_LAS_EVAL=SI 4099 ==> SUPLETORIO=SI 2505    conf:(0.61)
 9. PRESENT_TODAS_LAS_EVAL=SI 4099 ==> NIVEL_INTER_PROF=Alto 2466    conf:(0.6)
10. SUPLETORIO=SI 4035 ==> ASISTIO_SUPLETORIO=SI 2389    conf:(0.59)
```

FIGURA 3. 54. Resultados Reglas De Asociación – Experimento 1 - Carrera De Jurisprudencia

En la [*Figura 3.54*], se muestran los resultados obtenidos mediante el algoritmo Apriori, para la generación del modelo de la carrera de Jurisprudencia; podemos observar que el modelo generado ha tenido un soporte del 0.4, con lo cual ha predicho un total de 2252 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.5. El algoritmo a priori ha utilizado un total de 12 ciclos para la generación del modelo. Se visualiza en la imagen además que las 10 primeras reglas, son las mejores ya que poseen un nivel de confianza mayor que 0.5.

A continuación se interpretan las reglas que se consideran las más relevantes, de las 10 generadas por el modelo antes descrito:

Regla 3.- De 2381 estudiantes que han Aprobado las asignaturas, un total de 2304 estudiantes SI presentaron todas las evaluaciones. [*Ver Figura. 3.54*]

Regla 4.- De 4035 estudiantes que han tenido que dar supletorio de las asignaturas, un total de 3248 estudiantes han Reprobado. [*Ver Figura. 3.54*]

Regla 5.- De 3406 instancias que tienen un Nivel Alto en la Interacción del Profesor, un total de 2466 estudiantes SI han presentado todas las evaluaciones de las asignaturas. [*Ver Figura. 3.54*]

Reglas 10.- De 4035 estudiantes que han tenido que dar supletorio de las asignaturas, un total de 2389 estudiantes, han asistido a presentar la evaluación supletoria. [*Ver Figura. 3.54*]

La mayoría de los estudiantes que han presentado todas las evaluaciones de las asignaturas que cursa, son más propensos a aprobar la asignatura; sin embargo existen estudiantes que a pesar de haber presentado todas las evaluaciones respectivas de las asignaturas han reprobado por no haber obtenido buenas calificaciones en dichas pruebas; por lo tanto no es seguro que los estudiantes que presenten todas las evaluaciones podrán aprobar la asignatura. El nivel de Interacción del Profesor es influyente para que los estudiantes puedan presentar sus evaluaciones, ya que si el docente no responde a las inquietudes de los estudiantes, y no habilita el enlace respectivo para que los estudiantes puedan subir las evaluaciones; los alumnos tendrán inconvenientes para presentar dichas evaluaciones. Existen estudiantes que estando en supletorio, no asisten a dar la respectiva evaluación, por lo cual han reprobado la asignatura.

Las 10 reglas de clasificación generadas, no muestran los mejores resultados, para poder establecer las principales razones de la deserción, al contrario de las técnicas de clustering y clasificación que se aplicaron anteriormente, las mismas que mostraron resultados mucho más relevantes.

- ***CARRERA: ADMINISTRACIÓN DE EMPRESAS***

Se ha realizado un experimento para generar las mejores reglas de asociación, se creyó conveniente, seleccionar solo los atributos más relevantes con respecto a la deserción, para la generación de las reglas. Al ejecutar el algoritmo A priori, se obtuvieron las 10 mejores reglas, cuyos resultados se muestran a continuación:

TABLA 3. 69. Resultados- Reglas De Asociación– Carrera De Administración De Empresas

Experimentos	**Atributos Seleccio-nados**	**Atributos Asociados**	**Mejores Reglas Seleccionadas**
Experimento 1	Curso Edad Nota_Final Estado_Aprobación Nivel_Inter_Prof Nivel_Inter_Est Pre- sent_todas_las_Eval Supletorio Asistio_Supletorio	Estado_Aprobacion Asistio_Supletorio Supletorio Nota_Final Present_todas_las_Eval	Regla 2 Regla 3 Regla 9 Regla 10

A continuación se muestran los resultados del modelo de reglas de asociación para la pre-sente carrera:

Tamaño de la Población: 4243 instancias.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (1485 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 12

Size of set of large itemsets L(3): 3

Best rules found:

 1. ESTADO_APROBACION=REPROBADO 2507 ==> SUPLETORIO=SI 2507    conf:(1)
 2. ESTADO_APROBACION=REPROBADO NIVEL_INTER_PROF=Alto 2079 ==> SUPLETORIO=SI 2079    conf:(1)
 3. ASISTIO_SUPLETORIO=NO 1524 ==> ESTADO_APROBACION=REPROBADO 1524    conf:(1)
 4. ASISTIO_SUPLETORIO=NO 1524 ==> SUPLETORIO=SI 1524    conf:(1)
 5. SUPLETORIO=SI ASISTIO_SUPLETORIO=NO 1524 ==> ESTADO_APROBACION=REPROBADO 1524    conf:(1)
 6. ESTADO_APROBACION=REPROBADO ASISTIO_SUPLETORIO=NO 1524 ==> SUPLETORIO=SI 1524    conf:(1)
 7. ASISTIO_SUPLETORIO=NO 1524 ==> ESTADO_APROBACION=REPROBADO SUPLETORIO=SI 1524    conf:(1)
 8. EDAD=16a26 ESTADO_APROBACION=REPROBADO 1522 ==> SUPLETORIO=SI 1522    conf:(1)
 9. ASISTIO_SUPLETORIO=SI 1505 ==> SUPLETORIO=SI 1505    conf:(1)
10. ESTADO_APROBACION=APROBADO 1736 ==> PRESENT_TODAS_LAS_EVAL=SI 1702    conf:(0.98)
```

FIGURA 3. 55. Resultados Reglas De Asociación – Experimento 1 – Carrera Administración De Empresas

Con respecto a los resultados obtenidos mediante el algoritmo Apriori, para la generación del modelo de la carrera de Administración de Empresas; podemos observar en la [ver Figura 3.55] que el modelo generado ha tenido un soporte del 0.35, con lo cual ha predicho un total de 1485 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.9. El algoritmo a priori ha utilizado un total de 13 ciclos para la generación del modelo. Se visualiza en la imagen además que las 10 primeras reglas, son las mejores ya que poseen un nivel de confianza de 1, el mismo que significa el 100% de cumplimiento de la regla.

A continuación se interpretan las reglas que se consideran las más relevantes, de las 10 generadas por el modelo antes descrito:

Regla 2.- Un total de 2979 estudiantes han Reprobado las asignaturas, teniendo una interacción Alta el profesor en dichos cursos. [*Ver Figura. 3.35*]

Regla 3.- Un total 1524 estudiantes, les correspondía presentarse a dar la respectiva evaluación supletoria, sin embargo no asistieron a rendirla, por ende reprobaron la asignatura. [*Ver Figura. 3.35*]

Regla 9.- De 3929 estudiantes que tenían que rendir la evaluación supletoria de las materias correspondientes, solo un total de 1505 se presentaron a rendirla. [*Ver Figura. 3.35*]

Reglas 10.- De 1736 estudiantes que han aprobado las asignaturas, un total de 1702 estudiantes, han presentado todas las evaluaciones, tanto las presenciales como a distancia, de las asignaturas correspondientes. [*Ver Figura. 3.35*]

La mayoría de estudiantes que han reprobado las asignaturas de primer ciclo de la carrera de Administración de Empresas, han sido guiados por un tutor, el mismo que ha obtenido una interacción Alta en el curso, por lo tanto la interacción del profesor no poseen una alta influencia para que el alumno pueda aprobar la asignatura, en la presente carrera. Además la mayoría de estudiantes han aprobado las asignaturas, si presentando todas las evaluaciones de las asignaturas correspondientes.

De igual manera a lo que se analizó en la carrera de Jurisprudencia, se pudo constatar en la presente carrera, ya que las 10 reglas de clasificación generadas, no muestran los mejores resultados, para poder establecer las principales razones de la deserción, al contrario de las técnicas de clustering y clasificación que se aplicaron, las mismas que muestran resultados mucho más representativos.

- ***CARRERA: GESTIÓN AMBIENTAL***

Se ha realizado un experimento para generar las mejores reglas de asociación, se creyó conveniente, seleccionar solo los atributos más relevantes con respecto a la deserción, para la generación de las reglas. Al ejecutar el algoritmo A priori, se obtuvieron las 10 mejores reglas, cuyos resultados se muestran a continuación: [*ver Tabla 3.70*]

Experimentos	Atributos Seleccio-nados	Atributos Asociados	Mejores Reglas Seleccionadas
Experimento 1	Curso Nota_Final Estado_Aprobación Nivel_Inter_Prof Nivel_Inter_Est Pre- sent_todas_las_Eval Supletorio Asistio_Supletorio	Estado_Aprobacion Asistio_Supletorio Supletorio Nota_Final Present_todas_las_Eval	Regla 2 Regla 10

TABLA 3. 70. Resultados- Reglas De Asociación– Carrera De Gestión Ambiental

A continuación se muestran los resultados del modelo de reglas de asociación para la presente carrera:

Tamaño de la Población: 3141 instancias.

```
Apriori
=======

Minimum support: 0.3 (942 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 11

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 8

Best rules found:

 1. ESTADO_APROBACION=REPROBADO 1800 ==> SUPLETORIO=SI 1800    conf:(1)
 2. ESTADO_APROBACION=REPROBADO NIVEL_INTER_PROF=Alto 1459 ==> SUPLETORIO=SI 1459    conf:(1)
 3. ASISTIO_SUPLETORIO=SI 1257 ==> SUPLETORIO=SI 1257    conf:(1)
 4. PRESENT_TODAS_LAS_EVAL=SI ASISTIO_SUPLETORIO=SI 1132 ==> SUPLETORIO=SI 1132    conf:(1)
 5. NOTA_FINAL=FX 1083 ==> ESTADO_APROBACION=REPROBADO 1083    conf:(1)
 6. NOTA_FINAL=FX 1083 ==> SUPLETORIO=SI 1083    conf:(1)
 7. NOTA_FINAL=FX SUPLETORIO=SI 1083 ==> ESTADO_APROBACION=REPROBADO 1083    conf:(1)
 8. NOTA_FINAL=FX ESTADO_APROBACION=REPROBADO 1083 ==> SUPLETORIO=SI 1083    conf:(1)
 9. NOTA_FINAL=FX 1083 ==> ESTADO_APROBACION=REPROBADO SUPLETORIO=SI 1083    conf:(1)
10. ESTADO_APROBACION=REPROBADO PRESENT_TODAS_LAS_EVAL=SI 1043 ==> SUPLETORIO=SI 1043    conf:(1)
```

FIGURA 3. 56. Resultados Reglas De Asociación – Experimento 1 - Carrera De Gestión Ambiental

En la [*ver Figura 3.56*], se muestran los resultados obtenidos con reglas de asociación para la carrera de Gestión Ambiental, con la ayuda del algoritmo A priori, podemos observar en la imagen que el modelo generado ha tenido un soporte del 0.3, con lo cual ha predicho un total de 942 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.9. El algoritmo a priori ha utilizado un total de 14 ciclos para la generación del modelo.. Se visualiza en la imagen además que las 10 primeras reglas, son las mejores ya que poseen un nivel de confianza de 1, el mismo que significa el 100% de cumplimiento de la regla.

A continuación se interpretan las reglas que se consideran las más relevantes, de las 10 generadas por el modelo antes descrito: [*ver Figura 3.56*]:

Regla 2.- Un total de 1459 estudiantes, han Reprobado las asignaturas, teniendo una interacción Alta el profesor en dichos cursos.

Reglas 10.- Un total de 1043 estudiantes, han reprobado, presentado todas las evaluaciones, tanto las presenciales como a distancia, de las asignaturas correspondientes.

Se ha podido constatar en las reglas descritas, lo mismo que se ha encontrado en análisis de las anteriores carreras, ya que existen estudiantes, que aunque presentado todas las evaluaciones de la asignatura aun así, ha reprobado la misma. Además los docentes que han dictado la asignatura, han obtenido un nivel Alto de interacción en el entorno virtual del curso, aun así, los estudiantes han reprobado la asignatura. Es decir que las variables antes mencionadas no poseen una alta influencia para que un estudiante repruebe o no la asignatura.

Además se ha logrado determinar, que las 10 reglas de clasificación generadas, no muestran los mejores resultados, para predecir la deserción estudiantil; lo mismo sucedió en las carreras anteriores, ya que también la presente técnica no brindaba buenos resultados; al contrario de las técnicas de clustering y clasificación, que son mucho más efectivas para este tipo de problemas.

- ***CARRERA:INFORMÁTICA***

Se ha creído conveniente, seleccionar los atributos más relevantes con respecto a la deserción, para la generación de las mejores reglas. Al ejecutar el algoritmo A priori, se obtuvieron las 10 mejores reglas, es decir con el máximo nivel de confianza; cuyos resultados se muestran a continuación:

TABLA 3. 71. Resultados- Reglas De Asociación– Carrera De Informática

Experimentos	Atributos Seleccionados	Atributos Asociados	Mejores Reglas Seleccionadas
Experimento 1	Curso Nota_Final Estado_Aprobación Nivel_Inter_Prof Nivel_Inter_Est Present_todas_las_Eval Supletorio Asistio_Supletorio	Estado_Aprobacion Asistio_Supletorio Supletorio Nota_Final	Regla 2 Regla 10

A continuación se muestran los resultados del modelo de reglas de asociación para la presente carrera:

Tamaño de la Población: 1903 instancias.

```
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.35 (666 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 13

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 15

Size of set of large itemsets L(3): 4

Best rules found:

 1. ESTADO_APROBACION=REPROBADO 1170 ==> SUPLETORIO=SI 1170    conf:(1)
 2. EDAD=16a26 ESTADO_APROBACION=REPROBADO 771 ==> SUPLETORIO=SI 771    conf:(1)
 3. ESTADO_APROBACION=REPROBADO NIVEL_INTER_PROF=Alto 746 ==> SUPLETORIO=SI 746    conf:(1)
 4. PRESENT_TODAS_LAS_EVAL=NO 731 ==> SUPLETORIO=SI 731    conf:(1)
 5. ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO 709    conf:(1)
 6. ASISTIO_SUPLETORIO=NO 709 ==> SUPLETORIO=SI 709    conf:(1)
 7. SUPLETORIO=SI ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO 709    conf:(1)
 8. ESTADO_APROBACION=REPROBADO ASISTIO_SUPLETORIO=NO 709 ==> SUPLETORIO=SI 709    conf:(1)
 9. ASISTIO_SUPLETORIO=NO 709 ==> ESTADO_APROBACION=REPROBADO SUPLETORIO=SI 709    conf:(1)
10. ASISTIO_SUPLETORIO=SI 705 ==> SUPLETORIO=SI 705    conf:(1)
```

FIGURA 3. 57. Resultados Reglas De Asociación – Experimento 1 – Carrera De Informática

En la [*Figura 3.57*], se muestran los resultados obtenidos con reglas de asociación para la carrera de Informática, con la ayuda del algoritmo A priori, podemos observar en la imagen que el modelo generado ha tenido un soporte del 0.35, con lo cual ha predicho un total de 666 instancias de forma correcta, con un mínimo de confianza previamente establecido de 0.9. Se visualiza además que el algoritmo a priori ha utilizado un total de 13 ciclos para la generación del modelo.

A continuación se interpretan las reglas que se consideran las más relevantes, de las 10 generadas por el modelo antes descrito:

Regla 3.- Un total de 746 estudiantes, han Reprobado las asignaturas, teniendo el profesor un nivel Alto de interacción en dichos cursos.

Reglas 4.- Un total de 731 estudiantes, se han quedado en supletorio, presentado las evaluaciones de la asignatura correspondiente.

Lo más relevante que se ha encontrado analizado las presentes reglas generadas para la carrera de Informática es que, a pesar que los docentes han obtenido en su mayoría un nivel alto de interacción en el curso de igual manera los estudiantes han reprobado la asignatura. Es decir que las variables antes mencionadas no poseen una alta influencia para que un estudiante repruebe o no la asignatura.

Es importante recalcar, lo mismo que se estableció en las carreras antes analizadas, ya que aplicando la presente técnica de reglas de asociación, no se han obtenido los mejores resultados, para determinar las posibles razones de la deserción en los estudiantes, pues las reglas generadas son sucesos redundantes, los mismos que en su mayoría, ya se conocen en la institución.

e. Gráficos de Dispersión para el análisis de los datos

- ***CARRERA: JURISPRUDENCIA***

Distribución de la Deserción por Curso.- En la [*Figura. 3.58*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza que la mayoría de estudiantes que han cursado las asignaturas de la carrera de Jurisprudencia, constan como desertores. Siendo la materia de Introducción al Derecho, la que posee la mayoría de desertores.

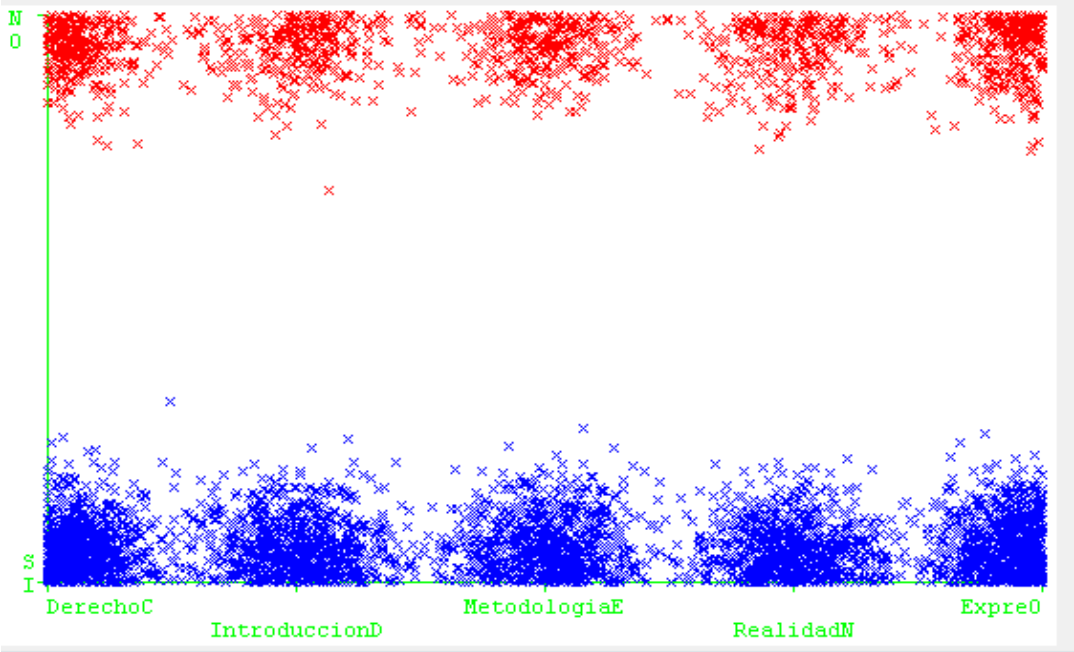

FIGURA 3. 58. Distribución de la deserción por *curso*

A continuación se realiza un análisis de la interrelación que existen entre las variables que se consideran las más relevantes según lo analizado en los clusters, con respecto a la deserción:

Distribución de Interrelación de Estado Aprobación – Curso - Desertor.- En la [*Figura. 3.59*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). En la imagen se visualiza, que el curso Introducción al Derecho posee mayor cantidad de Reprobados, y el curso de Metodología de Estudio es el que posee mayor cantidad de Aprobados. Podemos observar, que los estudiantes que han cursado las materias de Derecho Constitucional e Introducción al Derecho, y han reprobado, son en su mayoría desertores.

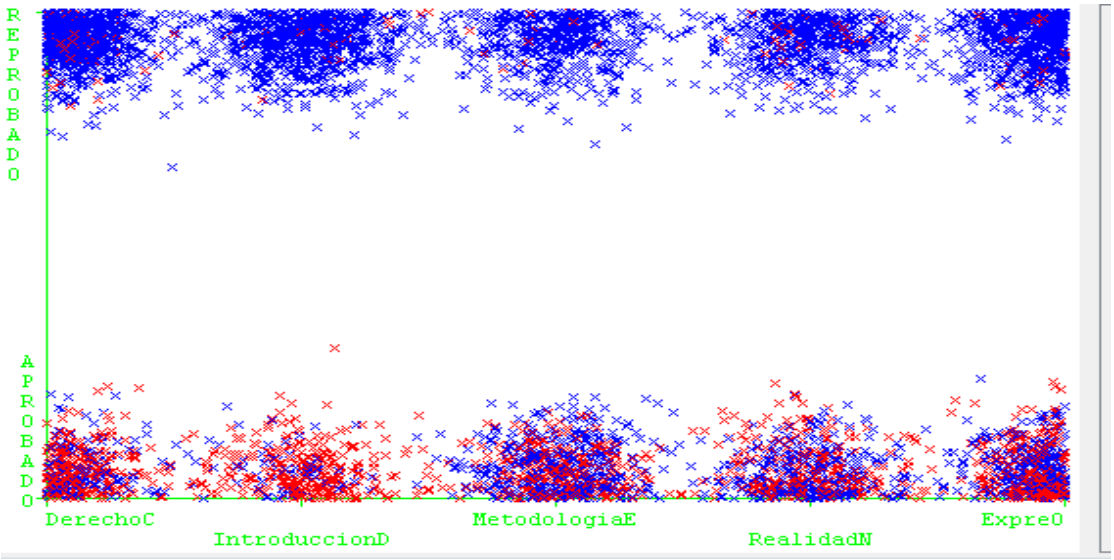

FIGURA 3. 59. Interrelación Estado Aprobación – Curso – Desertor

Distribución de Interrelación de Estado Aprobación – Edad – Desertor.- En la [*Figura. 3.60*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la edad de 16 a 26 años posee la mayor población. Se puede observar además a través de la gráfica, que los estudiantes que poseen una edad de 16 a 26 reprueban con mayor frecuencia y de igual manera son los que desertan con mayor frecuencia. Los estudiantes que poseen una edad de 27 a 37 años han aprobado con mayor frecuencia, y es la edad que posee en su mayoría los estudiantes no desertores, tomando en cuenta que la mayoría de estudiantes matriculados en primer ciclo en la carrera de Jurisprudencia, se encuentran entre los 16 a los 37 años.

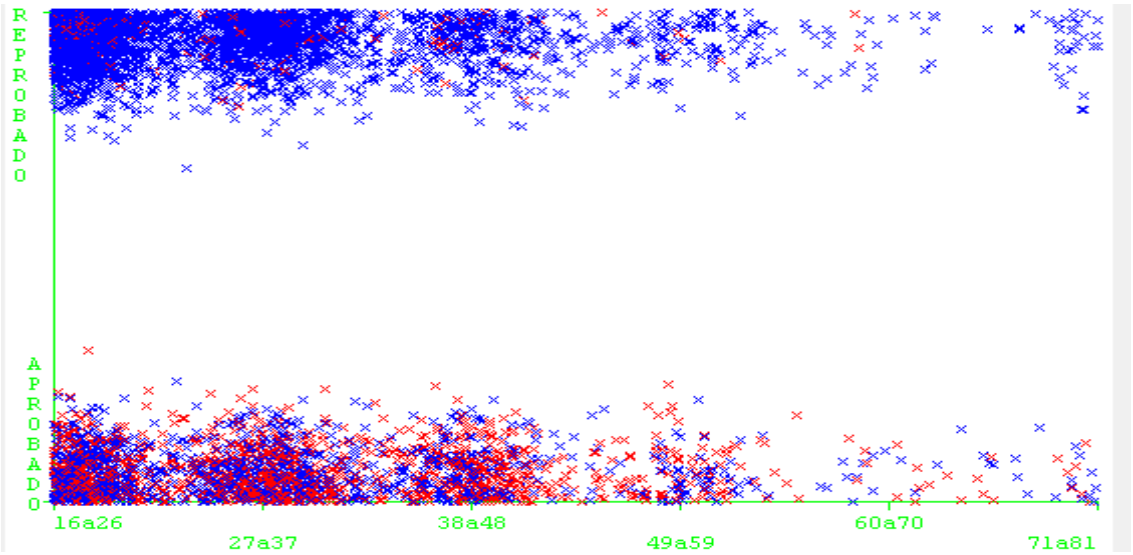

FIGURA 3. 60. Interrelación Estado Aprobación – Edad – Desertor

Distribución de Interrelación de Supletorio – Nota Final– Desertor.- En la [*Figura. 3.61*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la Nota Final que posee menor cantidad de instancias es A = 39 a 40, y la nota que posee mayor cantidad de instancias es FX = 14 a 28 y F = 13 o menos, siendo estas las menores calificaciones, que puede obtener un alumno. En la gráfica también se puede observar que la mayoría de estudiantes les ha correspondido, proporcionar el supletorio, obteniendo con ello, la nota FX=14 a 28 y F = 13 o menos puntos, siendo dichas notas de las asignaturas, las que poseen con mayor frecuencia los estudiantes desertores.

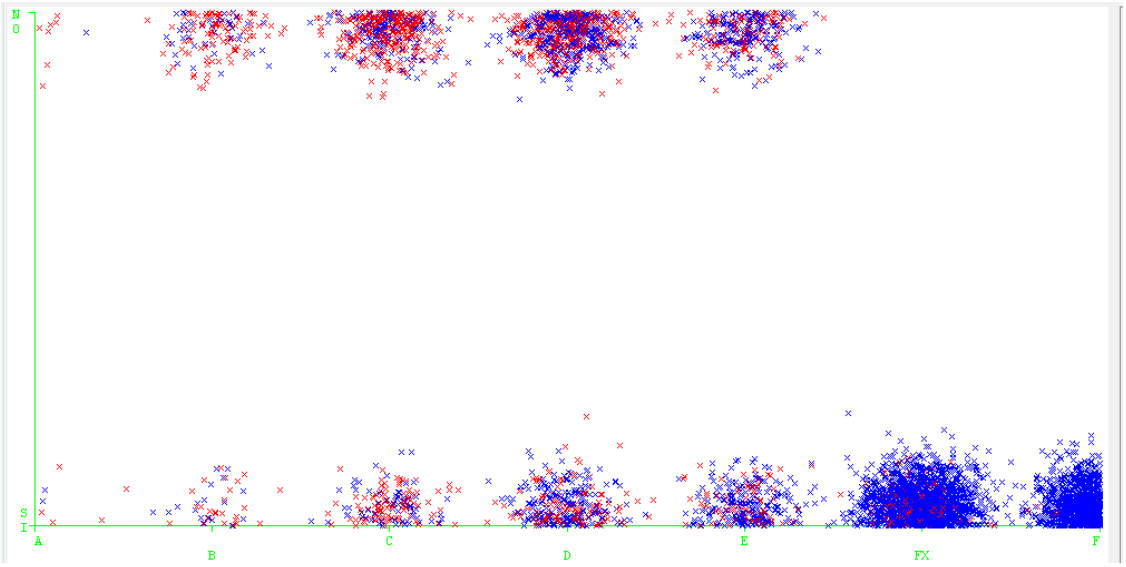

FIGURA 3. 61. Interrelación Supletorio – Nota Final – Desertor

Distribución de Interrelación de Estado de Aprobación – Nivel Interacción del Estudiante – Desertor.- En la [*Figura. 3.62*] se observa una serie de colores en pequeños grupos de la variable desertor, que representan (SI = azul, NO =rojo). Se observa en la imagen además, que la mayoría de estudiantes, que han obtenido un Nivel de Interacción Alto han Reprobado, siendo dichos alumnos los que constan en su mayoría como desertores, considerando que gran parte de los estudiantes han obtenido un nivel Alto de Interacción. Se visualiza también que existe mayor cantidad de desertores, de los cuales, la mayor parte son estudiantes que han reprobado las asignaturas.

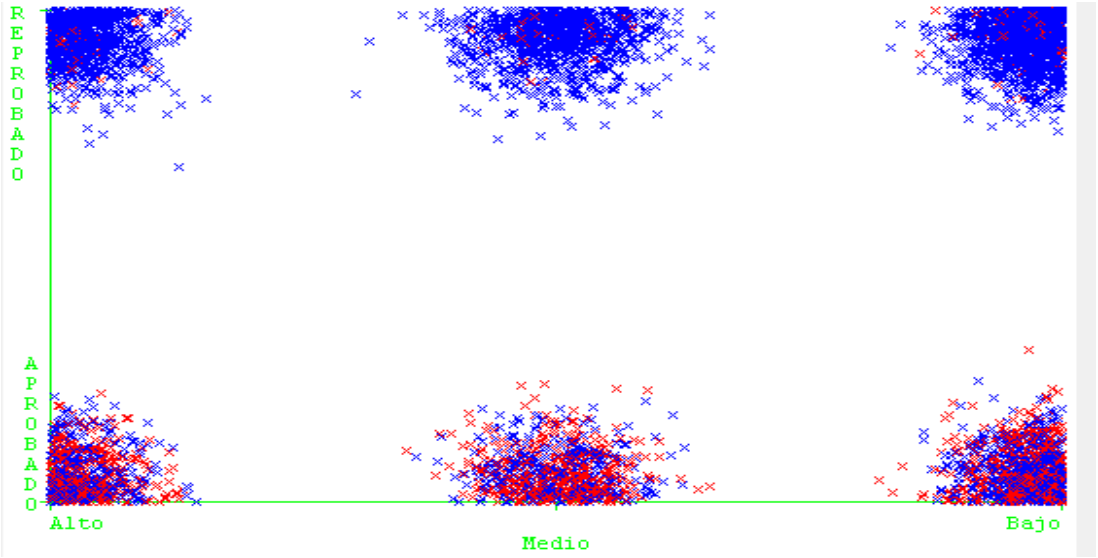

FIGURA 3. 62. Interrelación Estado Aprobación – Nivel De Interacción Estudiante – Desertor

- ***CARRERA: ADMINISTRACIÓN DE EMPRESAS***

Distribución de la Deserción por Curso.- En la [*Figura. 4.63*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza que la mayoría de estudiantes, que han cursado las asignaturas de la carrera de Administración de Empresas, constan como desertores. Siendo la materia Administración I, la que posee la mayor cantidad de estudiantes que constan como desertores, dicha asignatura pertenece a las troncales de la carrera.

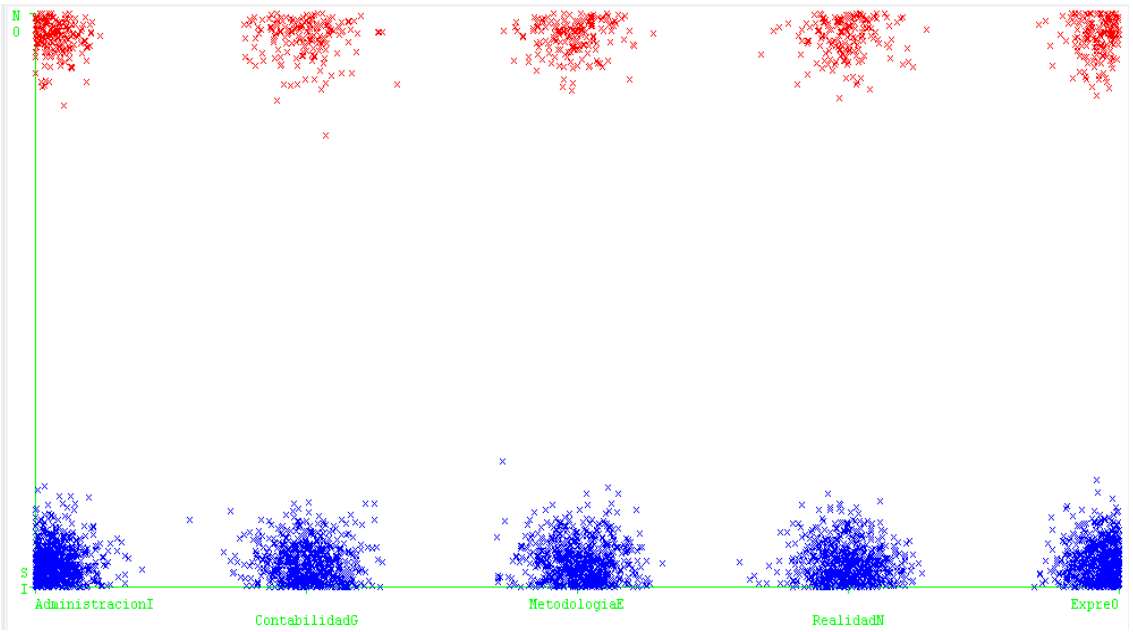

FIGURA 3. 63. Distribución De La Deserción Por *Curso*

Distribución de Interrelación de Estado Aprobación - Curso - Desertor.- En la [*Figura. 3.44*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). En la imagen puede visualizar, que el curso Administración I posee mayor cantidad de Reprobados, y el curso de Metodología de Estudio es el que posee mayor cantidad de Aprobados. Podemos observar, que los estudiantes que han cursado las materias de Administración I Contabilidad General y han reprobado, son en su mayoría desertores.

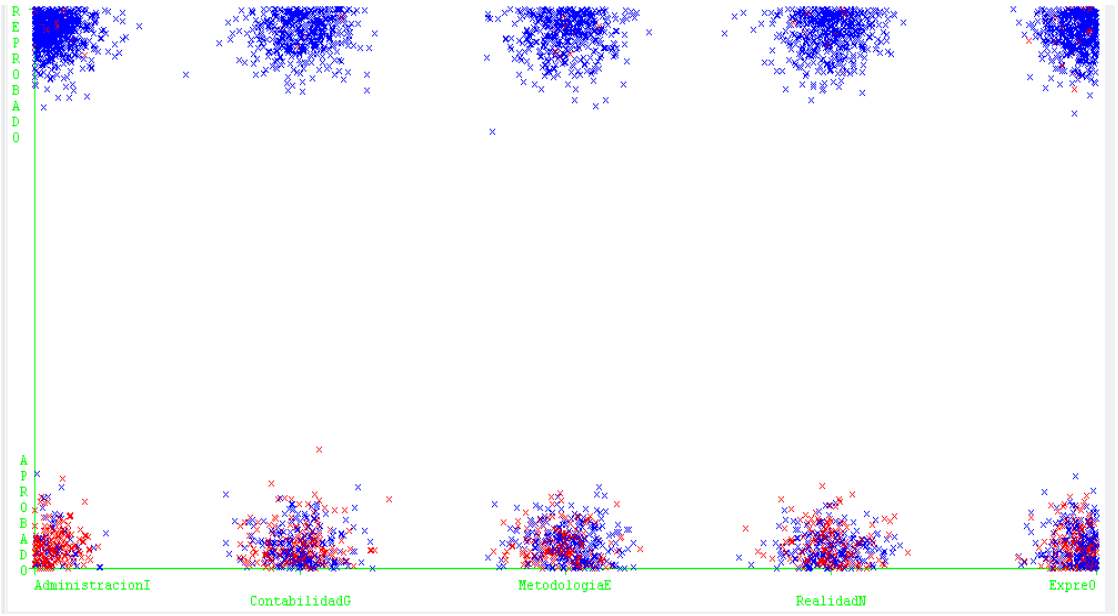

FIGURA 3. 64. Interrelación Estado Aprobación – Curso – Desertor

- ***CARRERA: GESTIÓN AMBIENTAL***

Distribución de la Deserción por Curso.- En la [*Figura. 4.65*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza que la mayoría de estudiantes, que han cursado las asignaturas de la carrera de Gestión Ambiental, constan como desertores. Siendo la materia Introducción a las Ciencias Ambientales, la que posee la mayor cantidad de estudiantes que constan como desertores, ya que se presenta una mayor aglomeración de valores, en dicha asignatura, la misma que pertenece a las troncales de la carrera.

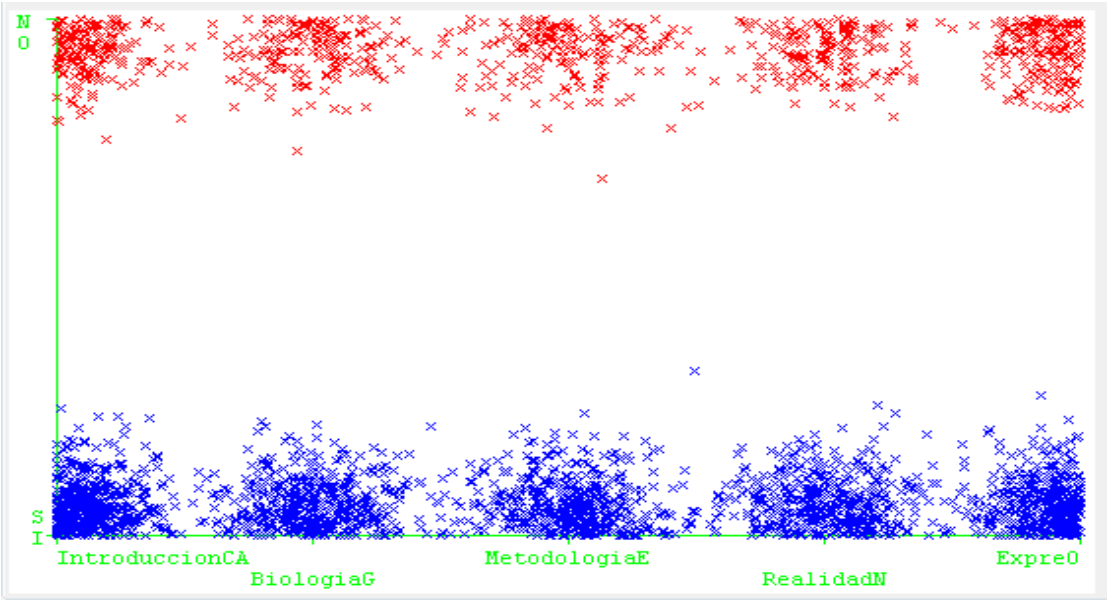

FIGURA 3. 65. Distribución De La Deserción Por *Curso*

Distribución de Interrelación de Estado Aprobación – Edad – Desertor.- En la [*Figura. 3.66*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la edad de 16 a 26 años posee la mayor población, seguida con una considerable diferencia, la edad de 27 a 37 años. Se puede observar además, en la gráfica, que los estudiantes que poseen una edad de 16 a 26 reprueban con mayor frecuencia y de igual manera son los que desertan con mayor frecuencia. Es importante tomar en cuenta que la mayoría de estudiantes matriculados en 1er ciclo de la carrera de Gestión Ambiental posee entre los 16 a 26 años de edad.

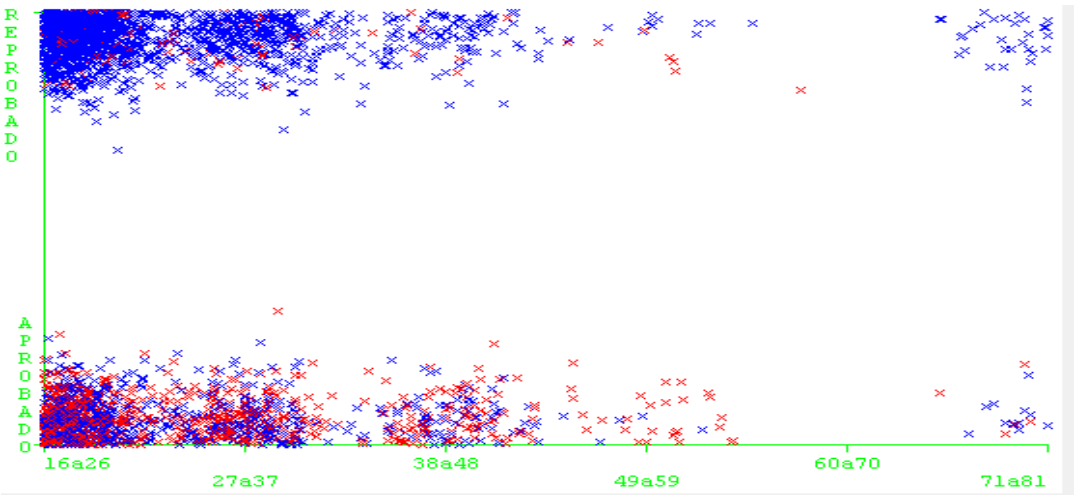

FIGURA 3. 66. Interrelación Estado Aprobación – Edad – Desertor

- ***CARRERA: INFORMÁTICA***

Distribución de la Deserción por Curso.- En la [*Figura. 3.67*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza que la mayoría de estudiantes, que han cursado las asignaturas de la carrera de Informática, constan como desertores. Siendo la materia de Metodología de Estudio, la que posee la mayor cantidad de estudiantes que figuran como desertores, ya que se presenta una mayor aglomeración de valores, en dicha asignatura, tomando en cuenta que en todas las asignaturas, se observa una similar conjunto de valores.

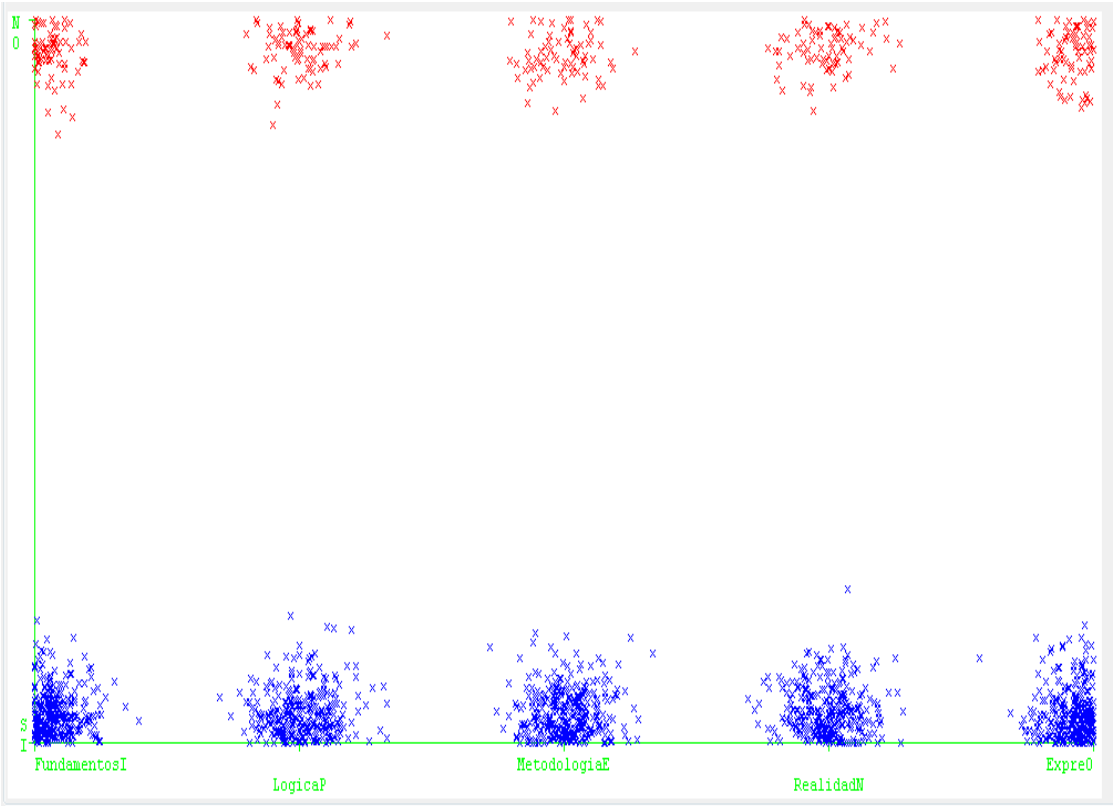

FIGURA 3. 67. Distribución De La Deserción Por *Curso*

Distribución de Interrelación de Supletorio – Nota Final– Desertor.- En la [*Figura. 3.68*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la Nota Final que posee menor cantidad de instancias es A = 39 a 40, y la nota que posee mayor cantidad de instancias es FX = 14 a 28 y F = 13 o menos, siendo estas las menores calificaciones, que puede obtener un alumno. En la gráfica también se observa que la mayoría de estudiantes se han quedado en supletorio en las asignaturas de 1er ciclo de la presente carrera.

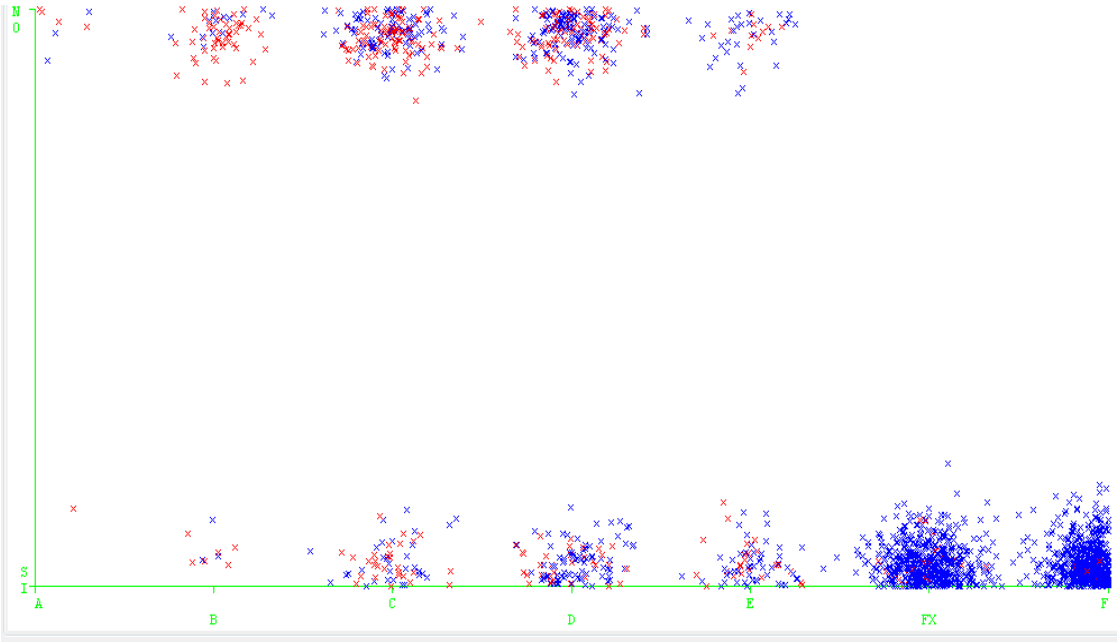

FIGURA 3. 68. Interrelación Supletorio – Nota Final – Desertor

Distribución de Interrelación de Estado de Aprobación – Nivel Interacción del Estudiante – Desertor.- En la [*Figura. 3.69*] se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen además, que la mayoría de estudiantes, que han obtenido un Nivel Medio de interacción han Reprobado, siendo dichos alumnos los que constan en su mayoría como desertores, seguido están los que han obtenido un nivel Bajo de interacción con una pequeña diferencia de valores. Se observa también en la gráfica, que en la carrera existe mayor cantidad de desertores, de los cuales, la mayoría son estudiantes que han reprobado las asignaturas, y el pequeño grupo de los no desertores en su mayoría, han aprobado las materias.

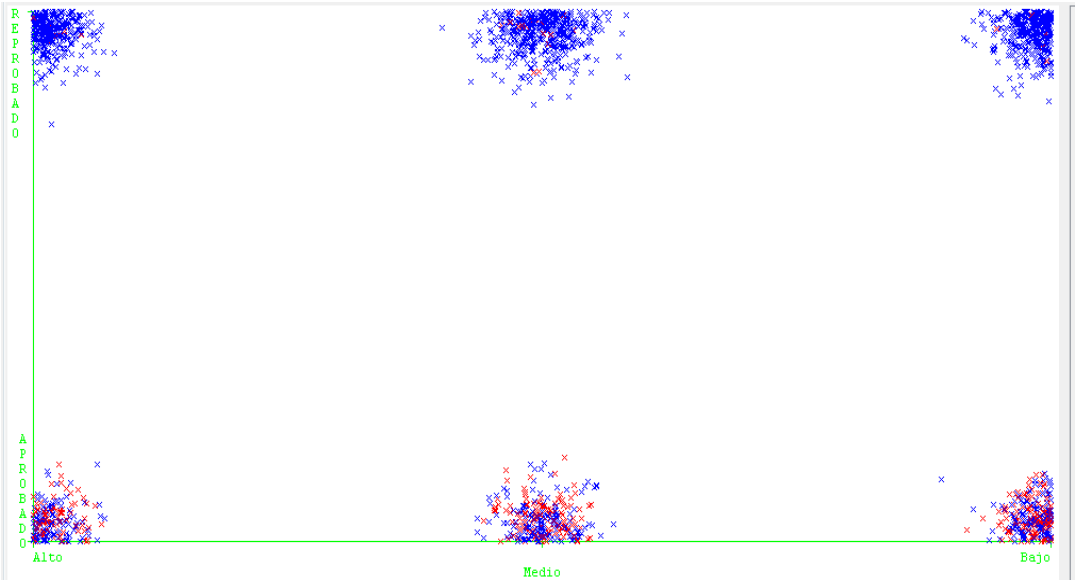

FIGURA 3. 69. Interrelación Estado Aprobación – Nivel De Interacción Estudiante – Desertor

f. Conclusiones de experimentos

CARRERA: Jurisprudencia – Área Socio Humanística

- Los estudiantes que reprueban en una o ambas materias troncales, han desertado la carrera con mayor frecuencia.

- El tipo de pago de matrícula, no es una variable que influye directamente como factor socioeconómico, para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes desertores han cancelado la matricula al contado.

- Existen estudiantes que a pesar de haber aprobado en alguna de las materias de formación básica, constan como desertores.

- La mayoría de estudiantes desertores de 1er ciclo de la carrera de Jurisprudencia, poseen una edad de 16 a 26 años, siendo estos las personas que no poseen la suficiente seguridad de haber elegido la carrera idónea.

- La mayoría de estudiantes desertores han obtenido una nota de menos 27 puntos en las asignaturas de 1er ciclo de la carrera de Jurisprudencia, de los cuales gran parte no se han presentado a dar el correspondiente examen supletorio, por ende han reprobado la asignatura, y la mayoría de los mismos son desertores.

- El género y el estado civil son datos personales que no influyen, para que un estudiante decida desertar la carrera.

- El nivel de interacción del profesor no tiene una influencia importante, para que los estudiantes deserten la carrera, ya que la mayoría de docentes de las materias de 1er ciclo han obtenido una interacción Alta en los cursos, a pesar de ello la mayoría de estudiantes han desertado.

- El nivel de interacción del estudiante no posee una alta influencia para que los estudiantes decidan desertar, ya que existen estudiantes que han obtenido un nivel alto y medio en la interacción, sin embargo han reprobado la asignatura y constan como desertores.

- Si el estudiante presenta o no presenta todas las evaluaciones tanto presencial, como a distancia de los dos bimestres de la asignatura, no influye para que el estudiante decida desertar la carrera, ya que existen estudiantes que a pesar de haber presentado todas las evaluaciones, han reprobado, y han desertado la carrera, tomando en cuenta que las notas de dichas evaluaciones son bajas.

- Existen un considerado número de estudiantes que sabiendo que están en supletorio, por un bajo rendimiento académico, deciden no presentarse al supletorio, de los cuales la mayoría decide desertar.
- El nivel de Interacción del Profesor es influyente para que los estudiantes puedan presentar sus evaluaciones, ya que si el docente no responde a las inquietudes de los estudiantes, y no habilita el enlace respectivo para que los estudiantes puedan subir las evaluaciones; los alumnos tendrán inconvenientes para presentar dichas evaluaciones.

- La mayoría de estudiantes que aprueban las asignaturas, han presentado todas las evaluaciones tanto presencial y a distancia de la misma, tomando en cuenta que dichos estudiantes han obtenido buenas calificaciones para poder aprobarla.

CARRERA: Administración de Empresas – Área Administrativa

- El género, estado civil, son atributos personales que no influyen para que un estudiante deserte la carrera.

- El tipo de pago de matrícula, no influye como factor socioeconómico para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes matriculados en las materias de 1er ciclo de la carrera de Administración de Empresas, son desertores y han cancelado la matricula al contado.

- La mayoría de estudiantes que constan como desertores, han reprobado en al menos una asignatura troncal de la carrera de Administración de Empresas.

- No necesariamente los estudiantes que poseen una Alta interacción en el curso son los más propensos a aprobar la asignatura, es decir que el nivel de interacción del estudiante no es una variable que posee una alta influencia, para que un estudiante pueda aprobar la asignatura, ya que también depende del puntaje que obtenga en las evaluaciones de la asignatura.

- Las variables con mayor influencia con respecto a la deserción son: la nota final, el estado de aprobación, conocer si está en el supletorio de la asignatura, conocer si asistió a dar la correspondiente evaluación supletoria, y saber si ha presentado todas las evaluaciones de la materia.

- Se pudo constatar que las variables que poseen un menor nivel de importancia en el dataset son el Nivel de Interacción del Estudiante y del Profesor, ya que existen estudiantes que a pesar de haber obtenido un nivel alto en la interacción en el curso, y a pesar de haber tenido un docente que ha obtenido también un nivel de interacción alto en el curso, de igual manera han reprobado la asignatura.

- Existen estudiantes que a pesar de haber aprobado en alguna de las materias de formación básica, constan igualmente como desertores.

- Una considerable cantidad de la población de matriculados en asignaturas de 1er ciclo de la carrera de Administración de Empresas no asiste a rendir la respectiva evaluación supletoria, de los cuales la mayoría constan como desertores.

CARRERA: Gestión Ambiental– Área Biológica

- La mayoría de estudiantes que han desertado la carrera, son hombres solteros y poseen una edad entre los 16 a 26 años, tomando en cuenta que la mayor parte de estudiantes matriculados en 1er ciclo de la presente carrera, poseen las características antes mencionadas.

- La variable tipo de pago de matrícula no posee demasiada influencia, como factor socioeconómico, para que el estudiante deserte la carrera.

- El nivel de interacción del profesor en el curso, no posee una influencia considerable para que un estudiante no decida desertar la carrera, ya que a pesar de que el docente de la presente asignatura obtuvo un nivel de interacción Alto de igual manera la mayoría de estudiantes de la materia constan como desertores.

- Existen estudiantes, que a pesar de haber obtenido un nivel de interacción alto en el curso, de igual manera han reprobado y desertado la carrera, además que también existen estudiantes que han obtenido un nivel de interacción Bajo, que de igual manera han aprobado; por lo tanto no es una variable que posee una influencia importante para que un estudiante pueda aprobar o reprobar la asignatura, ya que también depende de las notas que obtenga en las evaluaciones presenciales y a distancia de la asignatura.

- La presentación de todas las evaluaciones de la materia no asegura que el estudiante pueda aprobar la asignatura, ya que también depende del puntaje que obtenga en cada una de ellas.

- La mayoría de estudiantes que desertan, es porque han reprobado en alguna asignatura, y son aún más propensos a desertar, si dicha materia forma parte del grupo de las troncales de la carrera.

- La mayoría de estudiantes que se presentan a dar la evaluación supletoria, no aprueban la asignatura, y son los más propensos a desertar la carrera.

CARRERA: Informática – Área Técnica

- La mayoría de estudiantes desertores, son solteros, de estado civil masculino, y poseen una edad entre los 16 a 26 años, tomando en cuenta que dichos valores contienen la mayor cantidad de instancias.

- El tipo de pago de matrícula no es un atributo que posee demasiada influencia, como factor socioeconómico, para que el estudiante deserte la carrera, ya que la mayoría de estudiantes que desertaron la carrera, han pagado la matricula al contado.

- Los estudiantes son más propensos a desertar la carrera, al momento de reprobar una materia que pertenece al grupo de las troncales de la carrera.

- La presentación de todas las evaluaciones a distancia, no asegura que el estudiante apruebe o repruebe la materia, ya que también depende del puntaje que obtenga en cada una de ellas.

- El nivel de interacción del estudiante y del profesor no son variables, que influyen a gran escala para que un estudiante pueda aprobar o reprobar una asignatura, ya que también depende de las notas que obtenga las evaluaciones tanto presencial como a distancia correspondientes a la asignatura que estén cursando.

- La mayoría de estudiantes que no presentan todas las evaluaciones correspondientes a la asignatura, teniendo que rendir la evaluación supletoria, ya no se presentan a efectuarla.

- Los atributos que forman parte del rendimiento académico del estudiante son los que poseen una mayor influencia para que el estudiante decida desertar o no la carrera, como son la Nota_Final, Estado_Aprobacion, Supletorio, Asistio_Supletorio, Present_todas_las_eval.

- La mayoría de estudiantes que se han quedado en supletorio de una determinada asignatura, han reprobado la misma.

3.2.4. FASE V: Evaluación.

En la presente fase, se realiza una evaluación de los resultados obtenidos en el modelo, tomando en cuenta el cumplimiento de los objetivos de minería, los mismos que fueron descritos en la [*sección 3.1.1.3*].

El modelo predictivo obtenido, a través de la técnica de clustering ha ofrecido resultados satisfactorios, ya que nos ha permitido conocer las características principales de un posible desertor, las mismas que principalmente están relacionadas con el nivel académico del estudiante. Sin embargo según, las investigaciones realizadas y a la consulta con los expertos del tema, hubiese sido importante analizar también la información socioeconómica del estudiante ya que por ser la MAD de la UTPL, una institución pagada, puede influir también para que el estudiante deserte la carrera.

Los resultados obtenidos con árboles de decisión, nos permitieron conocer con mayor exactitud, la principal razón por la que un estudiante deserta la carrera, la misma que esta relaciona con el número de materias troncales reprobadas por un estudiante desertor; aunque dicha conclusión y otras más ya se las había obtenido con la técnica de clustering.

Los modelos obtenidos con reglas de asociación, mostraron, resultados poco satisfactorios, ya que la mayoría de deducciones que proporciono la menciona técnica, ya se las conocía, al momento de realizar el análisis del problema.

CONCLUSIONES

A través de los resultados obtenidos con los modelos creados, se ha podido concluir lo siguiente:

- La minería de datos posee importantes ventajas, para poder descubrir patrones de comportamiento de un estudiante desertor, ya que brinda un alto valor agregado para el análisis y la generación del nuevo conocimiento.

- La carreras analizadas han tenido un similar comportamiento respecto a que: la mayoría de desertores de dichas carreras han reprobado en al menos una materia troncal, ya que en las carreras analizadas: Jurisprudencia, Administración de Empresas, Informática y Gestión Ambiental existen estudiantes, que a pesar de haber aprobado en alguna de las materias de formación básica, de igual manera han desertado.

- La carrera de Informática perteneciente al Área Técnica es la que posee, un mayor porcentaje de deserción estudiantil, siendo este el 81,51 %, puesto que es una carrera que involucra un alto nivel de esfuerzo y dedicación, puesto que los contenidos que se indican en la misma son complejos de analizar y comprender, por ende la mayoría de estudiantes abandonan la carrera, porque presentan un bajo rendimiento académico.

- La mayoría de estudiantes desertores en las 4 carreras analizadas, poseen una edad, entre los 16 a 26 años, en mucho de los casos analizados se puedo determinar que mientras menor sea la edad que empiece una persona sus estudios superiores, mayor es la posibilidad de que el estudiante opte por desertar la carrera, ya sea porque dichos estudiantes, no tuvieron la suficiente orientación e información antes de iniciar sus estudios universitarios, o porque simplemente siguieron la carrera por influencia de los padres o amigos, por ende optaron por una titulación que no se ajustaba a sus perspectivas futuras o a su perfil académico.

- Se ha logrado determinar, según el análisis realizado en las 4 carreras, que una considerable cantidad de estudiantes, que No han presentado todas las evaluaciones correspondientes a las asignaturas que cursan, teniendo que presentarse a rendir la correspondiente evaluación supletoria, no han asistido a proporcionar dicho examen, por lo tanto han reprobado, y dichas personas son en su mayoría desertores.

- Según los resultados obtenidos en las 4 carreras analizadas, se puedo constatar que las variables: género, estado, civil; no influyen significativamente para que un estudiante deserte la carrera. Además se determinó que el tipo de pago de matrícula, no es una variable que influye directamente como factor socioeconómico, para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes desertores han cancelado la matricula al contado.

- La variable del nivel de interacción del profesor utilizada para la generación del modelo predictivo, no posee una influencia importante para que los estudiantes, aprueben la asignatura, y por ende no decidan desertar, ya que a pesar que la mayoría de profesores de las carreras analizadas, han obtenido un nivel de interacción Alto en el entorno del curso, la mayoría de estudiantes han reprobado; y además existen estudiantes que han aprobado la asignatura a pesar que el profesor ha obtenido un nivel de interacción Bajo. De igual manera se pudo constatar que la interacción del estudiante en el EVA no posee una alta influencia para que los estudiantes deserten la carrera, ya que existen estudiantes que han obtenido un nivel alto y medio en la interacción, sin embargo han reprobado la asignatura y constan como desertores. Por lo tanto la dedicación y el desempeño que aplique el estudiante en las evaluaciones tanto presencial como a distancia de las asignaturas que cursa, son las que influirán en mayor porcentaje para que el estudiante pueda aprobar la materia.

- Se ha podido constatar, según el análisis realizado en las 4 carreras, que si el estudiante presenta o no todas las evaluaciones tanto presencial, como a distancia de los dos bimestres de la asignatura, no influye en gran escala para que el estudiante decida desertar la carrera, ya que existen estudiantes que a pesar de haber presentado todas las evaluaciones, han reprobado, y han desertado la carrera, por lo tanto para que el estudiante apruebe la materia depende del puntaje que obtenga en dichas evaluaciones, ya que una considerable cantidad de estudiantes que si presentan todas las evaluaciones mencionadas, han obtenido una baja calificación en las mismas.

- Con la ayuda del análisis realizado con las técnicas de minería de datos y con los algoritmos evaluadores de atributos, se ha logrado determinar que: los atributos que forman parte del rendimiento académico del estudiante, son los que poseen una mayor influencia en la deserción del estudiante.

- Referente a las técnicas empleadas para la generación de los modelos, se pudo constatar que la técnica de clustering, brinda resultados eficaces referentes a los problemas de deserción estudiantil, ya que, con la mencionada técnica se pudo conocer las características principales de un posible desertor. Además se encontró que los árboles de decisión es una técnica, que en cierta medida facilita, realizar experimentos específicos, para con ello conocer con mayor exactitud las razones de la deserción estudiantil en la MAD de la UTPL. Al contrario de las anteriores técnicas las reglas de asociación que se aplicaron, no brindaron los mejores resultados, por lo tanto no es una técnica útil para ser aplicada, para problemas de deserción estudiantil.

RECOMENDACIONES

- Para poder crear un modelo de forma correcta y ordenada es importante, hacer uso de una metodología para el desarrollo de proyectos de minería de datos, siendo en la actualidad CRISP – DM una de las más utilizadas, ya que la misma propone las fases necesarias, para generar un modelo de calidad.

- Cuando se realiza el análisis de información, es necesario, implementar más de una técnica de minería de datos, para con ello comparar los resultados, y constatar cual técnica ha resultado ser más eficaz para el problema que se está analizando.

- Es necesario contar con personas que conozcan la temática que se requiere analizar, para con la experiencia que ellos poseen, poder determinar cuáles serían las variables idóneas para la creación del modelo predictivo.

- Según el análisis realizado y a la consulta con los expertos del tema, es recomendable también para la presente temática, analizar la información socioeconómica del estudiante, ya que por ser la universidad una institución educativa con fines de lucro, puede la mencionada variable tener una considerable influencia para el estudiante deserte la carrera.

- Al momento de generar un modelo para analizar la deserción estudiantil, es recomendable aplicar la técnica de clustering, ya que según lo evaluado, dicha técnica es la que brinda los mejores resultados, para el análisis de la deserción.

FUTURAS LINEAS DE INVESTIGACIÓN

- Para trabajos futuros, sería importante incorporar variables que formen parte de la ficha socioeconómica del estudiante, las mismas que no pudieron ser tomadas en cuenta, para el presente proyecto, ya que las bases de datos analizadas no contaban con dicha información. Por ser la MAD de la UTPL una institución educativa con fines de lucro, se sugiere analizar los campos relacionados con el nivel socioeconómico del estudiante, para con ello analizar en qué porcentaje influye la situación económica, para que un estudiante decida desertar la carrera. Los campos que se deberían tomar en cuenta para el análisis serian: el Tipo de Financiamiento para los estudios, Vivienda que habita, Tarjeta de Crédito, Rango correspondiente a los ingresos económicos; dichos campos en la actualidad son solicitados de forma obligatoria en el Sistema Académico (Syllabus), de la MAD de la UTPL.

- Se propone además, el análisis de la deserción, en todas las carreras de la MAD de la UTPL, tomando en cuenta muestras de estudiantes de todos los ciclos, puesto que en el presente estudio, solo se consideraron los estudiantes matriculados en asignaturas de 1er ciclo, de las carreras que poseen la mayor población de cada una de las 4 áreas, que existen en la institución educativa.

BIBLIOGRAFÍA

- Arévalo, Flora. & Maldonado, Judith (2010*). Estrategias para promover la retención estudiantil en un sistema de educación a distancia*. UTPL. Recuperada de: http://memorias.utpl.edu.ec/sites/default/files/documentacion/cread-andes/cc/utpl-cread-andes-2010-judith-maldonado-flora-arevalo.pdf

- Bouckaert, R. (2008). Bayesian Network Clasifiers in Weka for Version 3-5-7. University of Waikato. .Recuperada de: http://www.cs.waikato.ac.nz/~remco/weka.bn.pdf.

- Bienkowski, M., Feng, M., & Means, B. (2012). *Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief*. *Educational Technology*. Washington.

- Cabrera, Lidia., Bethencourt, José., Álvarez, Pedro., & González, Miriam. (2006). *El problema del abandono de los estudios universitarios*. Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE), *v. 12*, n. 2. Http://www.uv.es/RELIEVE/v12n2/relievev12n2_1.htm

- Chapman, Pete., Clinton, Julian., Kerber, Randy., Khabaza, Thomas., Reinartz, Thomas., Shearer, Colin., & Wirth, Rüdiger., (Traducción realizada por Gutiérrez, Daniel). (2000*). CRISP-DM 1.0- Guía paso a paso de Minería de Datos*. Disponible en: http://www.dataprix.com/es/metodolog-crisp-dm-para-miner-datos

- Cubero, J. Berzal, F. & Herrera, F. (2006). *Fundamentos de minería de datos*. Universidad de Granada. España. Recuperada de : http://sci2s.ugr.es/docencia/m1/Preprocesamiento-Weka-MD.pdf

- Domínguez, M. (2008). *Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado* (Tesis de Posgrado). Universidad Valle Del Grijalva, México. Recuperada de: http://pcti.mx/tesis-de-posgradoen-mexico?task=callelement&format=raw&item_id=382&element=5832706c-3ae3-408b-93e3-bca7418d0376&method=download

- Dron J., Duval E., Wiley D. *et al.* (2011*). 1st International Conference on Learning Analytics and Knowledge 2011.*Banff, Alberta, Canada.

- Dyckhoff, A. L., Dennis, Z., Bültmann, M., Chatti Ulrik, M. A., & Schroeder, U. (2012). Design and Implementation of a Learning Analytics Toolkit for Teachers. *Educational Technology & Society*, *15*, 58–76. Retrieved from http://www.ifets.info/journals/15_3/5.pdf

- *Duda*, Richard . *Hart*, Peter., Stork, David.(1997). *Pattern classification and scene analysis*. New York: John Wiley and Sons.

- Ecuadorinmediato.com (2012). Ecuador busca frenar alta deserción de universitarios con nuevo sistema de admisión ecuadorinmediato(Periódico).Recuperada de: http://www.ecuadorinmediato.com/index.php?Module=Noticias&func=news_user_view&id=163722&umt=ecuador_busca_frenar_alta_desercion_universitarios_con_nuevo_sistema_admision

- educause. (2007). Academic Analytics(6101) .Recuperado de http://net.educause.edu/ir/library/pdf/PUB6101.pdf.

- FACENA – UNNE. (2003). *Minería de Datos.* Universidad Nacional del Nordeste. Argentina. Recuperado de http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/SDataMining.pdf.

- Galán, M. (2009). *Sistemas_herramientas_mineria_datos.*España. Recuperado de http://www.oocities.org/es/mineria.datos/sistemas_herramientas_mineria_datos.pdf

- García, María & Álvarez, Aránzazu. (2008). *Análisis de Datos en WEKA – Pruebas de Selectividad.* Recuperada de: http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf.

- Gutiérrez, R. (2008). *Aplicación de minería de datos para la prevención de hechos delictivos* (Tesis Pregrado).Instituto Tecnológico de Buenos Aires, Argentina.

- Hand, D.J., Mannila, H. & Smyth, P. (2011). *Principles of Data Mining.* Cambridge, MA. USA: MIT Press.

- Hernández, J. & Ferri, C. (2006). *Introducción al Weka*. Curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software. Universitat Politècnica de València.

- Hernández, J., Ramírez, M. & Ferri, C. (2004).*Introducción a la Minería de Datos.* España: Pearson.

- López, Jorge., García, Juan., Sánchez, Leticia., & Solana, Emilia. (2007).*Las redes bayesianas como herramientas de modelado en psicología*. Servicio de Publicaciones de la Universidad de Murcia. Murcia (España). vol. 23, nº 2. 307-316.

- Moodle.org (2006) Recuperado de: https://moodle.org/.../Descripcion_Fisica_de_la_Base_de_Datos_Moo.dle.doc

- Moody, J. & Darken, C. (1989). *Fast Learning in networks of locally tuned processing units*. Neural Computation.

- *Pautsch, J. (2008). Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación*. (Tesis de Pregrado). Universidad nacional de misiones, Argentina, Recuperado de: http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

- Pérez, C., González, D.(2007). *Minería de datos técnicas y herramientas.* España: *Paraninfo Cengage Learning.*

- *Pérez, D. (2010).Estudio del efecto de las medidas de similaridad intervalares difusas en sistemas de clasificación* (Tesis de Pregrado). Universidad Pública Navarrensis, Pamplona.

- Pinzón, L. (2011). Aplicando minería de datos al marketing educativo*. Notas D Marketing*. Vol.1 (1), 45-61. Recuperado de: http://www.usergioarboleda.edu.co/investigacionmarketing/marketing/articulo5MineriaDatos.pdf

- Riquelme, J., Ruiz, R. & Gilbert K. (2006). Minería de Datos: Conceptos y Tendencias. *Revista Iberoamericana de Inteligencia Artificial, vol.10 (029), 11-18.*

- Romero, C., Ventura, S. & Castro C. (2006). Aplicación de Algoritmos Evolutivos como Técnica de Minería de Datos para la Mejora de Cursos Hipermedia Adaptativos basados en Web. *Revista iberoamericana de educación a distancia*, VOL.6 (2), 1-23.

- Ryan S.J.d. Baker. (2010). *Encyclopedia Chapter Draft v10* [versión electrónica]. Pennsylvania., USA: International Encyclopedia of Education (3rd edition)., http://users.wpi.edu/~rsbaker/Encyclopedia%20Chapter%20Draft%20v10%20-fw.pdf

- Spositto, O., Etcheverry, M., Ryckeboer, H., Bossero, J. (2008). *Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil*. Argentina. Recuperado de http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

- Thearling, K. (2007). *An Overview of Data Mining Techniques*. Recuperada de: http://www.thearling.com/text/dmtechniques/dmtechniques.htm.

- Universidad Agraria del Ecuador (2011). *¿Por qué abandonan los estudios los universitarios?*. El Misionero - Universidad Agraria del Ecuador. Recuperada de: http://www.elmisionero.com.ec/index.php?option=com_content&view=category&layout=blog&id=504&Itemid=23

- Utpl(2012). Modelo Académico de la Utpl. Recuperado de: http://www.utpl.edu.ec/casaabierta/sites/default/files/carreras-utpl-abrilagosto2012.pdf

- Vaira, S. Avila, O. Ricardi, P. Bergesio, A. (2010).Deserción universitaria. Un caso de estudio: variables que influyen y tiempo que demanda la toma de decisión. *Revista FABICIB*. Volumen 14. PÁGS. 107 a 115. Recuperado de: http://bibliotecavirtual.unl.edu.ar:8180/publicaciones/bitstream/1/2946/1/FABICIB_14_2010_pag_107_115.pdf

- Valero, Sergio., Salvador, Alejandro., García, Marcela. (2009). *Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos.* Universidad Tecnológica de Izúcar de Matamoros Recuperado de: http://www.utim.edu.mx/~svalero/docs/e1.pdf

- Wilford, Ingrid., Rosete, Alejandro., Rodríguez, Alfredo. (2008). *Aplicación de la Minería de Datos para el análisis de información clínica.* Revista Cubana de Informática Médica. Recuperada de: http://www.rcim.sld.cu/revista_18/articulos_htm/mineriadatos.htm#ibl

ANEXOS

ANEXO 1: SENTENCIAS SQL.

ANEXO 1 – A: Código sql utilizado para consultar las tareas propuestas en el curso.

La siguiente consulta muestra las tareas propuestas en un determinado curso previamente establecido.

```
SELECT * FROM `mdl_assignment` Where course='37237';
```

ANEXO 1 – B: Código SQL utilizado para consultar los foros propuestos en el curso.

La siguiente consulta muestra los foros propuestos en un determinado curso previamente establecido.

```
SELECT * FROM `mdl_forum` Where course = '37237';
```

ANEXO 1 – C: Código SQL utilizado para consultar los anuncios presentados en el curso.

La siguiente consulta muestra los anuncios presentados en un determinado curso previamente establecido.

```
SELECT * FROM `mdl_course_sections` Where course='37237';
```

ANEXO 1 – D: Código SQL utilizado para consultar el número de mensajes enviados del profesor al estudiante de un determinado curso.

La siguiente consulta muestra todos los mensajes no repetidos enviados del profesor a los estudiantes del curso especificado con un código previamente establecido, además se especifican los códigos ‘0’, ‘1’ y ‘2’ para extraer los mensajes que ha enviado el mismo profesor localizándose fuera del curso que está dictando, es decir en la ventana principal del Entorno Virtual de Aprendizaje (EVA).

```
SELECT mr.useridfrom, mr.useridto, mr.message, mr.messageid, ma.messageid,
FROM_UNIXTIME (timecreated, '%d/%m/%Y') as Fecha
FROM mdl_message_read mr, mdl_message_answered ma
Where ma.courseid IN ('37237', '0', '1', '2' )
AND ma.messageid=mr.messageid
AND mr.useridfrom IN ('3231')
AND mr.useridto IN (Select userid From mdl_enrol_utpl Where courseid='37237')

GROUP BY mr.message
```

ANEXO 1 – E: Código SQL utilizado para consultar el número de mensajes enviados del estudiante al profesor de un determinado curso.

La siguiente consulta muestra todos los mensajes no repetidos enviados de los estudiantes al profesor del curso especificado con un código previamente establecido, además se especifican los códigos ‘0’, ‘1’ y ‘2’ para extraer los mensajes que ha enviado el mismo profesor localizándose fuera del curso que está dictando, es decir en la ventana principal del Entorno Virtual de Aprendizaje (EVA).

```
SELECT mr.useridfrom, mr.useridto, mr.message, mr.messageid, ma.messageid,
FROM_UNIXTIME (timecreated, '%d/%m/%Y') as Fecha
FROM mdl_message_read mr, mdl_message_answered ma
Where ma.courseid IN ('37237', '0', '1', '2')
AND ma.messageid=mr.messageid
AND mr.useridto IN ('3231')
AND mr.useridfrom IN (Select userid From mdl_enrol_utpl Where courseid='37237')

GROUP BY mr.message
```

ANEXO 2: OBTENCIÓN DE LA VARIABLE 'NIVEL DE INTERACCIÓN DEL PROFESOR EN EL CURSO', Y ATRIBUTOS RELACIONADOS.

ANEXO 2 – A: Obtención del Campo: Porcentaje de Respuesta del Profesor al Estudiante.

El campo de porcentaje de respuestas del profesor es obtenido realizando una regla de 3, dicho cálculo se lo realiza mediante los atributos de: número de mensajes enviados del profesor a los estudiantes y viceversa **[ver Anexo 1]**; ya que según el número de preguntas que realice el alumno al profesor por medio del envío de mensajes el profesor proporcionará las respectivas respuestas. A continuación se detalla el cálculo realizado en la herramienta Microsoft Excel:

REGLA DE TRES

Var2 = 100%
Var1 = X

X= 100. $(\frac{Var1}{Var2})$

X = 100. $(\frac{39}{49})$

X = 79, 59
X= 80 R//

Var1-> MENSAJES ENVIADOS DE EL PROFESOR A LOS ESTUDIANTES
Var2-> MENSAJES ENVIADOS DE LOS ESTUDIANTES AL PROFESOR
X -> PORCENTAJE DE RESPUESTA DEL PROFESOR A LOS ESTUDIANTES

CARRERA	CURSO	CODIGO CURSO	PARALELOS	ID PROFESOR	MENSAJES ENVIADOS DE EL PROFESOR AL ESTUDIANTE	MENSAJES ENVIADOS DE LOS ESTUDIANTES AL PROFESOR	PORCENTAJE DE RESPUESTA DEL PROFESOR A LOS ESTUDIANTES
Asistencia Gerencial y Relaciones Publicas UTPL-ECTS-1A	ADMINISTRACIÓN I	36027	A	42613	39	49	=PRODUCT(100;(K3/L3))

80%

FIGURA 2A. Obtención del campo Porcentaje de Respuesta del profesor al alumno

ANEXO 2 – B: Promedio de las variables relacionadas con la Interacción del Profesor en caso de que existan varios paralelos en un curso.

A continuación se muestra el cálculo del promedio redondeado que se realiza en el caso de que exista más de un paralelo de un determinado curso:

CODIGOS CURSO	idPROFESOR	PARALELOS	NUMERO_DE_TAREAS_PROPUESTAS_EN_EL CURSO	NUMERO DE FOROS PROPUESTOS	NUMERO_ANUNCIOS_EN_EL_CURSO	MENSAJES ENVIADOS DEL PROFESOR AL ESTUDIANTE	MENSAJES ENVIADOS DEL ESTUDIANTE AL PROFESOR	PORCENTAJE DE RESPUESTA DEL PROFESOR
36045	27571	A	2	1	17	49	86	57
36913	3089	B	2	0	0	13	72	18
		Promedio Interaccion	2	1	9	31	79	39

FIGURA 2B. Promedio de las variables relacionadas con la Interacción del Profesor en el curso

ANEXO 2 – C: Discretización de los valores relacionados con la Interacción del Profesor en el curso.

A continuación se detallan los rangos de valores obtenidos por las asignaturas de las carreras de una determinada área al momento de discretizar los campos de: número de tareas propuestas, número de foros propuestos, número de anuncios presentados y porcentaje de respuestas del profesor de un determinado curso; el último campo es obtenido realizando una regla de tres con el número de mensajes que envía el profesor al estudiante y viceversa detallada en el **[Anexo 2 - A]**.

TABLA 2A. Discretización de los valores relacionados con la Interacción del Profesor en el Curso

ÁREA	Variable	Rango		
		Bajo	**Medio**	**Alto**
ADMINISTRATIVA (ADMINISTRACCIÓN DE EMPRESAS)	Numero de Tareas	-inf – 1.33	1.33 – 1.66	1.66 – inf
	Numero de Foros	-inf – 0.66	0.66 – 1.33	1.33 – inf
	Numero de Anuncios	-inf – 7	7– 13	13 – inf
	Respuestas del Profesor	-inf – 50.33	50.33– 61.66	61.66 – inf
BIOLÓGICA (GESTIÓN AMBIENTAL)	Numero de Tareas			1
	Numero de Foros	-inf – 0.33	0.33 – 0.66	0.66 – inf
	Numero de Anuncios	-inf – 9	9 – 13	13 – inf
	Respuestas del Profesor	-inf – 44.66	44.66 – 64.33	64.33 – inf

SOCIO (HUMANÍSTICA JURISPRUDENCIA)	Numero de Tareas			1
	Numero de Foros	-inf – 3	3 – 5	5 – inf
	Numero de Anuncios	-inf – 15	15 – 20	20 – inf
	Respuestas del Profesor	-inf – 51.33	51.33 – 59.66	59.66 – inf
TÉCNICA (INFORMÁTICA)	Numero de Tareas	-inf – 2.33	2.33 – 3.66	3.66 – inf
	Numero de Foros	-inf – 2	2 – 4	4 – inf
	Numero de Anuncios	-inf – 14.33	14.33 – 23.66	23.66 – inf
	Respuestas del Profesor	-inf – 30	30 – 60	60 – inf

Los rangos de valores detallados anteriormente fueron establecidos ya redondeados a su inmediato superior en cada uno de los archivos de Excel de las 4 carreras seleccionadas de las áreas correspondientes de la MAD de la UTPL. a continuación se muestra una captura que se realizó del Curso Administración I de la carrera ADMINISTRACCIÓN DE EMPRESAS UTPL-ECTS-1A del Área Administrativa; en el mismo se demuestra que en 'número de tareas propuestas' tiene un nivel *Medio* de interacción, en 'número de foros propuestos' tiene un nivel *Alto*, en 'número de anuncios presentados en el curso' tiene un nivel *Alto,* y el porcentaje de respuesta del profesor a los estudiantes tiene un nivel de interacción *Alto*.

CURSO		NUMERO DE TAREAS PROPUESTAS	NUMERO DE FOROS PROPUESTOS	NUMERO ANUNCIOS PRESENTADOS	PORCENTAJE DE RESPUESTA DEL PROFESOR AL ESTUDIANTE (%)
ADMINISTRACIÓN I	INTERACCIONES	1	2	19	64
	NIVEL DE INTERACCIÓN	Medio	Alto	Alto	Alto

FIGURA 2C. Ejemplo de la obtención de la variable de Nivel de Interacción del Profesor en el Curso.

ANEXO 2 – D: Discretización para obtener el campo de Nivel de Interacción del Profesor.

Para determinar la variable de Nivel de Interacción del Profesor en el Curso se agruparon las variables descritas en el *Anexo 3 – A*, las mismas que a partir de los niveles ya encontrados se logró obtener con la ayuda de la *Moda,* se escoge entre los valores que se tengan en las variables el que más se repita, en el caso que se tengan valores que se repitan el mismo número de veces se escogerá el más optimista.

CURSO		NUMERO DE TAREAS PROPUESTAS	NUMERO DE FOROS PROPUESTOS	NUMERO ANUNCIOS PRESENTADOS	PORCENTAJE DE RESPUESTA DEL PROFESOR AL ESTUDIANTE (%)
ADMINISTRACIÓN I	INTERACCIONES	1	2	19	64
	NIVEL DE INTERACCIÓN	Medio	Alto	Alto	Alto
	NIVEL DE INTERACCIÓN DEL PROFESOR EN EL CURSO (MODA)	*Alto*			

FIGURA 2D. Discretización para obtener el campo de Nivel de Interacción del Profesor

ANEXO 3: MODELOS FÍSICOS DE LAS BASES DE DATOS UTILIZADAS.

ANEXO 3 – A: Modelo Físico del Entorno Virtual de Aprendizaje.

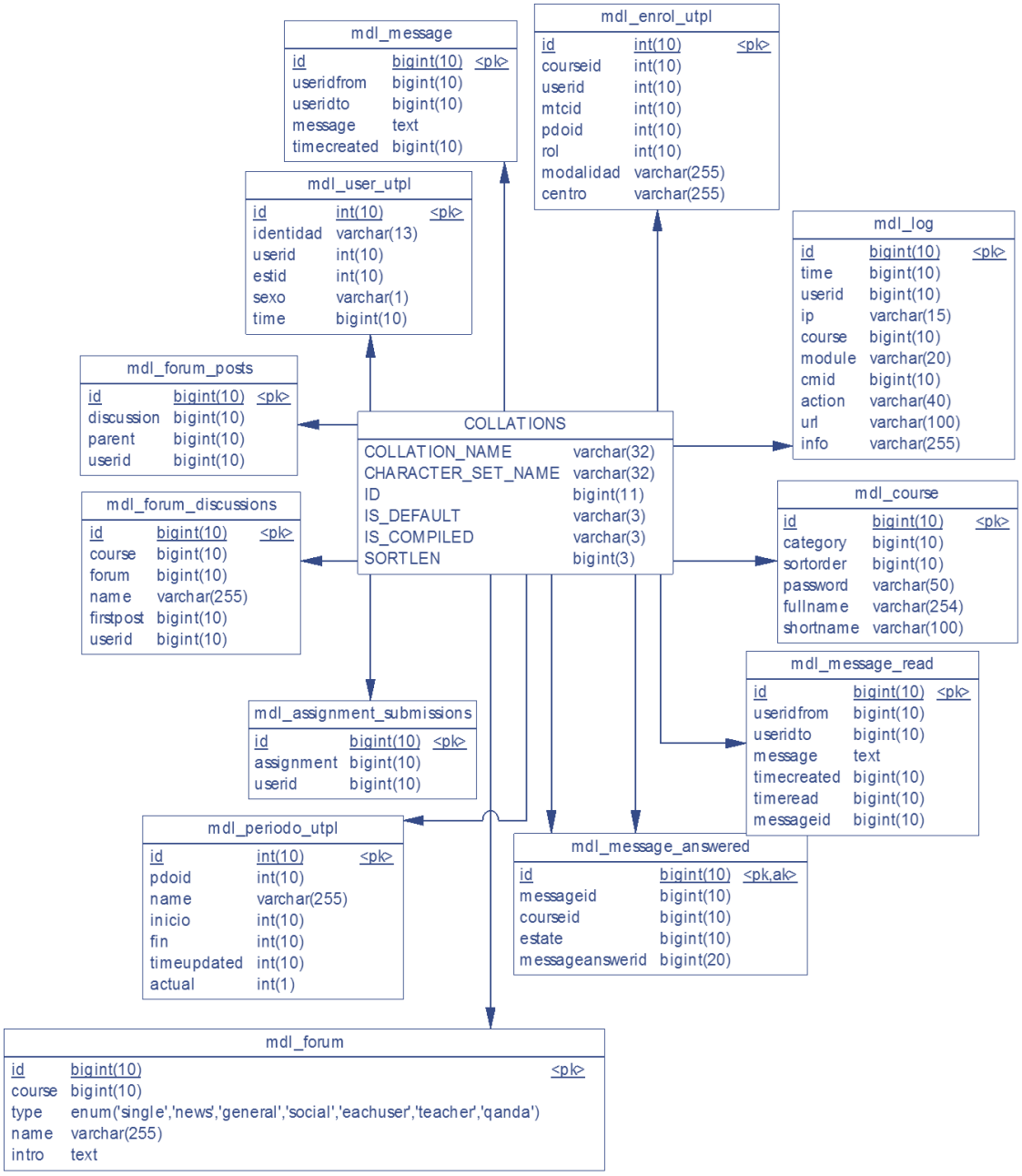

FIGURA 3A. Modelo Físico del Entorno Virtual de Aprendizaje

ANEXO 3 – B: Modelo Físico del Sistema Académico.

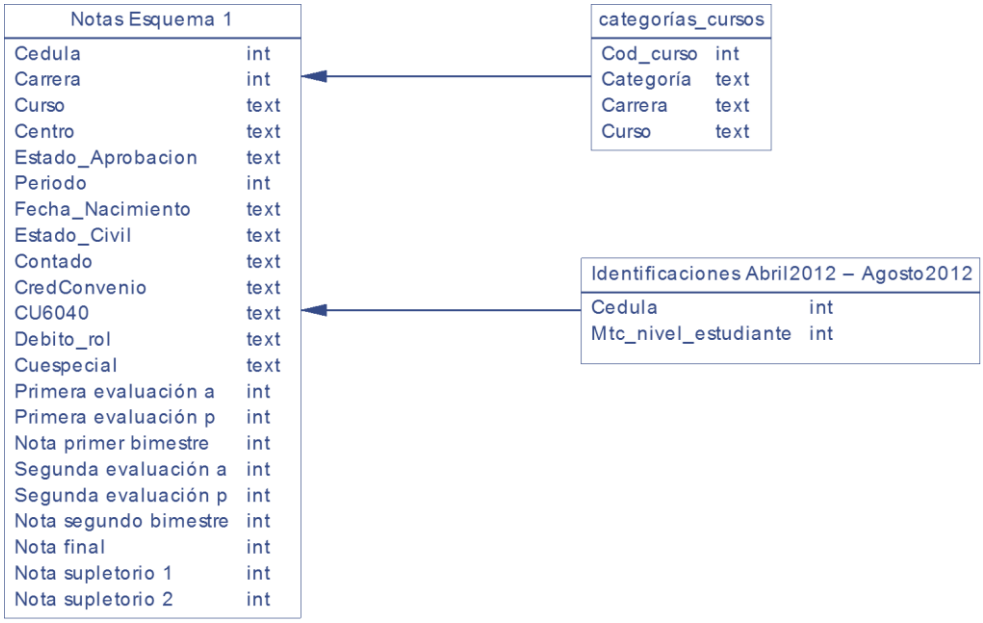

FIGURA 3B. Modelo Físico del Sistema Académico

ANEXO 4: TABLAS DEL ENTORNO VIRTUAL DE APRENDIZAJE (EVA).

ANEXO 4 – A: Tabla: mdl_user_utpl.

TABLA 4A. mdl_user_utpl

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tabla mdl_user_utpl	Int auto increment Primary key	10	Not null
Identidad	Número de Cedula del usuario	Varchar	13	Not null
Userid	Id del usuario	Int	10	Not null
Sexo	Sexo del usuario	Varchar	1	Not null

ANEXO 4– B: Tabla: mdl_enrol_utpl.

TABLA 4B. mdl_enrol_utpl

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tabla mdl_user_utpl	Int auto increment Primary key	10	Not null
Courseid	Id del curso al que pertenece el usuario	Int	10	Not null
Userid	Id del usuario	Int	10	Not null
Pdoid	Periodo lectivo en el que se dicta el curso.	Int	10	Not null
Rol	Número del rol que realiza el usuario.	Int	10	Not null
Modalidad	Modalidad a la que pertenece el usuario ya sea Presencial o Abierta/Distancia	Varchar	255	Not null
Centro	Ciudad donde se encuentra el centro de estudios al que pertenece el usuario.	Varchar	255	Not null

ANEXO 4 – C: Tabla: mdl_course_utpl.

TABLA 4C. mdl_course_utpl

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tabla mdl_course_utpl	Int auto increment Primary key	10	Not null
Category	Id de la categoría a la que pertenece el curso, se encuentran establecidos según la carrera que concierne al curso.	Bigint	10	0
Fullname	Nombre completo del curso	Varchar	254	0
Shortname	Nombre corto del curso	Varchar	100	0

ANEXO 4 – D: Tabla: mdl_course_sections.

TABLA 4D. mdl_course_sections

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tabla mdl_course_sections	Int auto increment Primary key	10	Not null
Course	Id del curso al que pertenece la sección	Int	10	Not null

Section	Es el número de sección que describe, por ejemplo: si un curso tiene 5 secciones en este campo se registrara un 0 en la primera sección, un 1 en la segunda sección, un 2 en la tercera sección y así sucesivamente.	Int	10	Not null
Summary	Registra la descripción del módulo, por ejemplo. *Módulo I: Educación a Distancia y Tecnologías de la Información y Comunicación*	Text		Not null
Visible	Si el modulo es visible o no. (1= visible, 0 = no visible)	Int	1	Not null

ANEXO 4 – E: Tabla: mdl_assignment.

TABLA 4E. mdl_assignment

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tarea que se está creando	Int Auto Increment Primary key	10	Not null
Course	Id del curso al que pertenece esa tarea	Int	10	Not null
Name	Nombre de la tarea	Var char	255	Not null
Description	Descripción de la tarea	Text		Not null

ANEXO 4 – F: Tabla: mdl_forum.

TABLA 4F. mdl_forum

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id del fórum	Int Auto Increment Primary key	10	Not null
Course	Id del curso al que pertenece el foro	Int	10	Not null
Type	Tipo de foro (eachuser= Cada persona plantea un tema, single= Debate sencillo, general= Foro para uso general)	Enum		Not null
Name	Nombre del foro	Var char	255	Not null
Intro	Introducción o descripción del foro	Text		Not null

ANEXO 4 – G: Tabla: mdl_message.

TABLA 4G. mdl_message

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id del mensaje	Int Auto Increment Primary key	10	Not null
Useridfrom	Id del usuario que envía el mensaje	Bigint	10	Not null
Useridto	Id del usuario que recibe el mensaje	Bigint	10	Not null
Message	Contenido del mensaje	Text		Not null

Timecreated	Registra la fecha que se creó el mensaje en formato Unix.	Date		Not null

ANEXO 4 – H: Tabla: mdl_message_read.

TABLA 4H. mdl_read

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id del mensaje	Int Auto Increment Primary key	10	Not null
Useridfrom	Id del usuario que envía el mensaje	Bigint	10	Not null
Useridto	Id del usuario que recibe el mensaje	Bigint	10	Not null
Message	Contenido del mensaje	Text		Not null
Timecreated	Registra la fecha que se creó el mensaje en formato Unix.	Date		Not null
Messageid	Registra el Id que está relacionado con la tabla:mdl_message_answered	Bigint	10	Not null

ANEXO 4 – I: Tabla: mdl_message_answered

TABLA 4I. mdl_message_answered

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Id de la tabla: mdl_message_answered	Int Auto Increment Primary key	10	Not null
Messageid	Id del mensaje principal al cual ha dado respuesta un determinado usuario.	Bigint	10	Not null

Courseid	Id del curso de donde se creó el mensaje; existen ids 0, 1, 2 para identificar los mensajes que se enviaron fuera de los cursos.	Bigint	10	Not null
Estate	Estado del mensaje creado. (0=no leído, 1=leído, 2= respondido)	Bigint	10	Not null
messageanswerid	Id del mensaje de respuesta.	Bigint	20	Not null

ANEXO 4 – J: Tabla: mdl_periodo_utpl

TABLA 4J. mdl_periodo_utpl

Campo	Descripción	Tipo de Carácter	Longitud	Restricción
Id	Identificador de la tabla	Int	10	Not null
Pdoid	Código para representar el estado periodo	Int	10	Not null
Name	Nombre del periodo académico.	Varchar	255	Not null
Inicio	Fecha de inicio del periodo	Int	10	Not null
Fin	Fecha final del periodo académico.	Int	10	Not null

En el [*Anexo 3 – A]*, se muestra la relación que existe entre las tablas descritas anteriormente del Entorno Virtual de Aprendizaje, por medio de un diagrama conceptual.

ANEXO 5: TABLAS DEL SISTEMA ACADÉMICO (SYLLABUS).

TABLA 5A. Notas Esquema 1.

Campo	Descripción
Cedula	Número del Cedula del estudiante
Carrera	Carrera que está siguiendo el estudiante
Asignatura	Asignatura que cursa el estudiante
Centro	Ciudad donde se encuentra el centro universitario.
Estado_aprobacion	Estado de Aprobación de la asignatura que está cursando el estudiante, considerando la calificación del supletorio según corresponda el caso. (Aprobado, Reprobado, Anulado)
Periodo	Periodo del ciclo académico en que se haya matriculado el estudiante.
Fecha_nacimiento	Fecha de nacimiento del estudiante
Estado_civil	Estado Civil de estudiante
Contado	Registra una x si el tipo de pago de matrícula es al Contado
Credconvenio	Registra una x si el tipo de pago de matrícula es por medio de un Convenio a Crédito.
Cu6040	Registra una x si el tipo de pago de matrícula es a Crédito por medio de una cuenta bancaria.
Debito_rol	Registra una x si el tipo de pago de matrícula es a Crédito por medio del Debito al Rol de Pagos del Trabajo.
Cuespecial	Registra una x si el tipo de pago de matrícula es a Crédito por medio de una Cuenta Especial.
Primera evaluación a	Nota de la Primera Evaluación a Distancia (Primer Bimestre)
Primera evaluación p	Nota de la Primera Evaluación Presencial (Primer Bimestre)
Nota primer bimestre	Promedio del Primer Bimestre tomando en cuenta las notas de las evaluaciones presentadas por el estudiante.
Segunda evaluación a	Nota de la Segunda Evaluación a Distancia (Segundo Bimestre)

Segunda evaluación p	Nota de la Segunda Evaluación Presencial (Segundo Bimestre)
Nota segundo bimestre	Promedio del Segundo Bimestre tomando en cuenta las notas de las evaluaciones presentadas por el estudiante.
Nota final	Nota Final de la asignatura que está cursando el estudiante, considerando la calificación del supletorio según corresponda el caso.
Nota supletorio 1	Nota del Supletorio del Primer Bimestre de la asignatura que está cursando el estudiante.
Nota supletorio 2	Nota del Supletorio del Segundo Bimestre de la asignatura que está cursando el estudiante.

ANEXO 5 – B: Tabla: Identificaciones Abril2012 – Agosto2012.

TABLA 5B. Identificaciones de los estudiantes matriculados *Abril2012 – Agosto2012*

Campo	Descripción
Cedula	Número de Cedula del Estudiante Matriculado
Mtc_nivel_estudiante	Nivel en que está matriculado el estudiante

ANEXO 5 – C: Tabla: categorías_cursos.

TABLA 5C. Categorías de los cursos.

Campo	Descripción
Cod_curso	Código de identificación del curso
Categoría	Número de categoría de la carrera
Carrera	Nombre de la Carrera.
Curso	Nombre del Curso.

En él *[Anexo 3 – B]* se muestra la relación que existe entre las tablas descritas anteriormente del Sistema Académico (Syllabus), por medio de un diagrama conceptual.

ANEXO 6: PAPER.

PREDICCIÓN DE LA DESERCIÓN DE LOS ESTUDIANTES DE 1ER CICLO DE LA MODALIDAD ABIERTA Y A DISTANCIA DE LA UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA (MAD-UTPL)

*[*DROPOUT PREDICTION OF STUDENTS OF 1ST STAGE OF THE MAD-UTPL*]*

por / by

Karla Ordoñez
Cargo: Becaria de Investigación
Dep. Ciencias de la Computación y Electrónica
(Sección Dep. Inteligencia Artificial)
(kfordonez@utpl.edu.ec)

Mg. Priscila Valdiviezo
Cargo: Directora de la Sección
Departamental de Inteligencia Artificial
Dep. Ciencias de la Computación y Electrónica
(Sección Dep. Inteligencia Artificial)
(pmvaldiviezo@utpl.edu.ec)

Ms. Juan Carlos Torres
Cargo: Docente Investigador de la UTPL
Dep. Ciencias de la Computación y Electrónica
(Sección Dep. Inteligencia Artificial)
(pmvaldiviezo@utpl.edu.ec)

Sistemas Informáticos y Computación, Universidad Técnica Particular de Loja

Ecuador, Loja

RESUMEN

El problema de la deserción estudiantil lo enfrentan varias instituciones universitarias a nivel nacional y mundial; por lo cual la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, hace algunos años atrás ha creído conveniente desarrollar algunos estudios para determinar cuáles son las posibles causas por la que un estudiante decide abandonar sus estudios universitarios.

Para contribuir con la solución del problema, en el presente artículo se describe cada uno de los pasos que propone la metodología CRISM-DM, la misma que es una de las más utilizadas en la actualidad para la generación de proyectos de Minería de datos, con ella se pretende obtener un modelo basado en Minería de Datos, que con la ayuda de la implementación de algoritmos de Inteligencia Artificial, ya incorporados en la herramienta de preprocesamiento de datos Weka, se pueda conocer cuáles son las posibles causas por las que un alumno que cursa las asignaturas de primer ciclo de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja, decide abandonar sus estudios universitarios, esto se lo ha realizado a través del análisis de la información: personal, académica, del estudiante y de la interacción en el entorno virtual del curso tanto de los estudiantes como de los docentes que dictan la asignatura.

Palabras Clave: Minería de datos, deserción estudiantil, clustering, clasificación, asociación, moodle.

ABSTRACT

The dropout problem I faced several universities nationally and globally, for which the Open and Distance mode Technical University of Loja, a few years ago has seen fit to develop studies to determine the possible causes why a student decides to leave college.

To contribute to solving the problem, this article describes each of the steps proposed CRISM-DM methodology, the same that is one of the most commonly used today for generating data mining projects, with it is intended to obtain a model with the help of implementing AI algorithms, and incorporated into the data preprocessing tool Weka, you can know what are the possible reasons why a student who attends the subjects first cycle of Open and Distance mode Technical University of Loja, decides to leave college, this is what has made through the analysis of information: personal, academic, student and interaction in the virtual environment of the course both students and teachers that dictate the course.

Key words: *Data Mining, dropout, clustering, classification, association, moodle.*

1. INTRODUCCIÓN

La deserción universitaria es un problema que lo enfrentan las instituciones de nivel superior por diferentes factores que afectan tanto el desarrollo personal y académico del profesional en formación. Esta deserción se ve mayormente acentuada en aquellos sistemas de estudios a distancia donde el estudiante es el responsable de su propio aprendizaje, en este caso de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja.

Para contribuir con la solución a este problema de deserción se plantea la aplicación de técnicas de minería de datos, con el objeto de "Comprender cuáles son las posibles causas por lo que un alumno decide abandonar sus estudios universitarios, a través del análisis de las características de los estudiantes". De acuerdo a Hand, Mannila & Smyth (2011) "la Minería de datos es un proceso que reúne un conjunto de herramientas de diversas ciencias (Estadística, Informática, Matemáticas, Ingeniería, entre otras)" que persigue extraer conocimiento oculto o información no trivial de grandes volúmenes de datos, con la finalidad de dar soluciones a problemas específicos en empresas determinadas.

Algunos trabajos realizados en torno a este tema están por ejemplo el presentado por (Pinzón, 2011), en la que se aplica minería de datos a los registros del estudiante, desde que ingresa en la universidad y se determinan las posibles causas de deserción en cada periodo académico. Se presenta además la caracterización del perfil del estudiante desertor. Para aplicar las técnicas de minería de datos en este trabajo se utilizaron las siguientes variables: DNI, semestre, ciudad de domicilio, sexo o género, edad, estado civil, país de nacimiento, medio por el cual se enteró del programa y de la universidad, idioma, entre otras.

Así mismo en (Spositto, 2008), se aplica técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil, en este proyecto se tomaron en cuenta las siguientes variables: datos del estudiante, datos de las carreras, datos de los planes de estudio, vigentes y no vigentes de las carreras, datos de las materias de los planes de estudio, datos de las notas y datos de los censos realizados a los estudiantes. En (Pautsch, 2008) se utiliza también minería de datos para el análisis de la deserción en la Carrera de Analista en Sistemas de Computación, con el objetivo de maximizar la calidad que los modelos tienen para clasificar y agrupar a los estudiantes, de acuerdo a sus características académicas, factores sociales y demográficos. Un estudio similar es el presentado por (Domínguez, 2008) en la que se usa minería de datos y lógica difusa para elaborar un sistema de predicción para la detección de factores que influyen para el abandono escolar de alumnos que estudian en instituciones privadas de nivel superior.

Otro caso de deserción universitaria se presenta en (Vaira et. al, 2010), donde se investiga cuándo es probable que ocurra un evento determinado como el de abandonar los estudios universitarios, el tiempo que lleva tomar la decisión y cuáles son las variables que más influyen en el cumplimiento de este evento. Las causas que mencionan pueden influir en el abandono de los estudios universitarios van desde: el abandono por la escasa formación previa, los reiterados fracasos en los exámenes finales, el origen social, la elección inadecuada de estudios, características familiares o circunstancias de la vida, problemas de organización de las diferentes unidades académicas, entre otras.

Estos y otros trabajos previos son los que han alimentado la experiencia realizada en esta investigación y toman como base la necesidad de determinar este modelo predictivo en la que se analiza una serie de variables de índole personal y académico de los estudiantes. Además se analizan diferentes patrones de comportamiento del estudiante durante su interacción en las asignaturas disponibles en un entorno virtual de aprendizaje al cual los estudiantes están enrolados, a fin de conocer si dicha interacción influye o no en la deserción.

Considerando que, entre las técnicas de minería de datos que existen en la actualidad se encuentran: la modelización estadística paramétrica y no paramétrica, reglas de asociación y dependencia, métodos bayesianos, árboles de decisiones y sistemas de reglas, métodos relacionales y estructurales, redes neuronales artificiales, máquinas de soporte de vectores, clustering, algoritmos evolutivos y reglas difusas, métodos basados en casos y en vecindad; en esta investigación se utilizan: técnicas de clustering para comprender los comportamientos de estudiantes en busca de descubrir patrones y tendencias de posibles desertores y árboles de decisión para la clasificación y comprensión de qué variables son las más importantes dentro del proceso de predicción.

2. METODOLOGÍA APLICADA

La metodología utilizada para la aplicación de las técnicas de minería de datos, fue: CRISP–DM (Cross Industry Standard Process for Data Mining, por sus siglas en inglés), una de las más utilizadas para el desarrollo de proyectos de Minería de Datos en entornos académicos (Chapman et al., 2000). Esta metodología provee una representación completa del ciclo de vida de un proyecto de minería de datos. En base a esto las fases que se siguieron en esta investigación fueron:

a. Comprensión del Negocio.- En esta fase se revisó información sobre la modalidad abierta y a distancia de la UTPL, sus objetivos, características de la población estudiantil, modelo educativo (Rubio, 2009), etc., para poder determinar el ámbito del problema. En base a esto se terminaron los objetivos de este proyecto, a fin de establecer lo que se intenta resolver con la aplicación de minería de datos, los cuales son:

- Encontrar patrones de comportamiento de los estudiantes desertores.
- Identificar grupos de estudiantes, según sus características comunes.
- identificar que variables influyen en la deserción de los estudiantes de primer ciclo de la modalidad abierta y a distancia.

En esta fase también se revisaron las herramientas que serían las más adecuadas para obtener resultados para el análisis de la deserción, de ahí que después de un estudio de las diferentes herramientas de minería de datos que existen en la actualidad, se creyó conveniente utilizar Weka, la misma que es de libre distribución, fácil de manejar, además que permite analizar grandes volúmenes de datos, y tiene implementados varios técnicas de aprendizaje y minería de datos como: clustering, clasificación, asociación, regresión y visualización, etc., que podían ser aplicadas en este contexto.

b. Comprensión de los Datos.- Se realizó una recolección inicial de los datos relacionados con el problema, además se procedió a realizar un análisis de los mismos con el fin de identificar las relaciones entre ellos, y así generar conocimiento sobre alguna información oculta.

Los datos obtenidos corresponden a una muestra de estudiantes que cursan las cinco materias de primer ciclo de la modalidad abierta y a distancia de la UTPL, que son de formación básica y troncales. Estas materias corresponden a las carreras que poseen la mayor población de estudiantes de las cuatro áreas académicas de la UTPL: área técnica, administrativa, biológica, y humanística. La muestra corresponde al período académico Octubre 2012 – Febrero 2013, y se detalla a continuación:

Tabla. 1. *Muestra poblacional*

AREA	CARRERA	Materia	Número de estudiantes
ADMINISTRATIVA	ADMINISTRACIÓN DE EMPRESAS	ADMINISTRACIÓN I	**988**
		CONTABILIDAD GENERAL	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	
BIOLÓGICA	GESTIÓN AMBIENTAL	INTRODUCCION A LAS CIENCIAS AMBIENTALES	**714**
		BIOLOGIA GENERAL	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	

SOCIO HUMANÍSTICA	JURISPRUDENCIA	DERECHO CONSTITUCIONAL	**1304**
		INTRODUCCION AL DERECHO	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	
TÉCNICA	INFORMATICA	FUNDAMENTOS INFORMATICOS	**449**
		LOGICA DE LA PROGRAMACION	
		METODOLOGÍA DE ESTUDIO	
		REALIDAD NACIONAL Y AMBIENTAL	
		EXPRESIÓN ORAL Y ESCRITA	

Cabe señalar que los estudiantes que conforman la muestra, para el proceso de aprendizaje cuentan con recursos como guías didácticas, libros base y acceso al entorno virtual de aprendizaje. En base a la lectura de los recursos disponibles para el estudiante se deben desarrollar dos evaluaciones a distancia, una por cada bimestre, que luego deben ser subidas al entorno virtual. Además los estudiantes deben rendir dos evaluaciones presencia-

les parciales, una por cada bimestre, y si no logran obtener el puntaje mínimo de aprobación (28/40) tienen opción a una tercera evaluación (supletoria).

En base a esto, las fuentes de información de dónde se extrajo conocimiento fueron las bases de datos internas de la institución de los siguientes sistemas: Entorno Virtual de Aprendizaje (EVA) y el Sistema Académico (Syllabus). El Entorno Virtual de Aprendizaje es un sistema web que ofrece diferentes servicios al estudiante, como: acceso a los cursos matriculados en la carrera, acceso a recursos educativos, actividades de aprendizaje, consultas al profesor, correo electrónico, etc. Esta fuente de información permitió determinar los comportamientos de los usuarios en base a las interacciones que realizan en el curso (materia). Por otro lado, el Sistema Académico (Syllabus) es un sistema web que facilita los procesos académicos que realiza un estudiante como son: matricula, pago en línea, acceso al expediente estudiantil, acceso a notas, entre otros. De este sistema se obtuvo información sobre las notas, información personal y socioeconómica de los estudiantes.

Una vez obtenidos los datos se procedió a realizar una limpieza de los mismos, con la finalidad de dar tratamiento a las inconsistencias encontradas en algunas de las variables. En esta caso se encontró que la variable *ESTADO_CIVIL*, contenía valores que pertenecían a un grupo poco frecuente, como: 'UNION LIBRE', 'DIVORSIADO', 'RELIGIOSO' y 'OTRO', lo que se realizó fue agrupar los valores en solo dos grupos, como son SOLTERO Y CASADO, por lo tanto los valores de Divorciado, Religioso y Otro, fueron reemplazados por 'Soltero', y los valores de Unión Libre fueron reemplazados por 'Casado'.

La variable *TIPO_PAGO_MATRICULA*, contenía diferentes tipos de pago en el grupo de cancelaciones a crédito, como: CREDCONVENIO, CU6040, DEBITO_ROL, CUESPECIAL, por tanto se procedió a agrupar estos valores por tipo de pago a 'Crédito'.

La variable *ESTADO_APROBACION*. Contenía algunos valores como 'ANULADO', estado que se registra a los estudiantes que por algún motivo ha decido ya no tomar la materia, lo que significaba que ya no constaba como estudiante de dicha asignatura; por lo que se procedió a eliminar los registros que poseían dicho valor. Estos registros fueron pocos frecuentes en la información recolectada.

Una vez realizada la limpieza de los datos se procedió a elaborar el conjunto de datos definitivo, en base a las siguientes variables.

Tabla. 2. *Variables utilizadas*

Campo	Descripción	Tipo de Dato
CURSO	Curso del estudiante	Nominal
EDAD	Edad del estudiante	Nominal
GENERO	Género del estudiante	Nominal
ESTADO_CIVIL	Estado civil del estudiante	Nominal
TIPO_PAGO_MATRICULA	Tipo de pago con la que ha cancelado la matricula	Nominal
NOTA_FINAL	Nota final de estudiante	Nominal
ESTADO_APROBACION	Estado de aprobación de la asignatura que ha cursado el estudiante	Nominal
NIVEL_ INTER_PROF	Nivel de interacción del profesor en el curso	Nominal
NIVEL_INTER_EST	Nivel de interacción del estudiante en el curso	Nominal
PRESENTARON_TODAS_LAS_EVAL	Registra un estado de NO en el caso que un estudiante no haya presentado al menos un de las 4 evaluaciones del curso.	Nominal (SI, NO)
SUPLETORIO	Registra ‘SI’, en el caso de que el estudiante, tenga que rendir el examen supletorio, caso contrario registra un ‘NO’.	Nominal (SI, NO)

ASISTIO_SUPLETORIO	Registra un 'SI' en el caso de que el estudiante no se presente a emitir la respectiva evaluación supletoria, caso contrario se registrara un 'NO'	Nominal (SI, NO)
DESERTOR	Registra un 'no' si el estudiante no deserto la carrera caso contrario registra 'si'	Nominal (SI, NO)

En el caso de la última variable, si el estudiante no registre una matrícula en el siguiente ciclo o período académico (segundo ciclo), en este caso se considera como desertor.

c. Modelado

En esta fase se procedió a aplicar las técnicas de minería de datos seleccionadas, las cuales se muestra en la siguiente tabla.

Tabla 3. *Técnicas utilizadas para la experimentación*

Técnica	Tarea	Algoritmo
Clustering	Agrupamiento	Simple-Kmeans
Árboles de Decisión	Clasificación	J48

En base a los objetivos propuestos en esta investigación y considerando algunos estudios realizados en esta línea (Pinzón, 2011; Spositto, 2008; Pautsch, 2008) se ha creído conveniente trabajar con la técnica de clustering ya que permiten agrupar estudiantes en subclases de acuerdo a su nivel de participación y semejanza de acceso a la plataforma virtual (Mejía et al, 2008). Por lo que con esta técnica se procedió a agrupar a los estudiantes de las materias en diferentes grupos relacionados con las actividades realizadas en las mismas, y así descubrir patrones que reflejen comportamientos similares en los estudiantes. Los árboles de decisión fueron utilizados para en base a unas entradas descritas por medio del conjunto de variables, comprender cuáles son más importantes en la deserción del estudiante.

3. RESULTADOS OBTENIDOS

Una vez seleccionadas las técnicas se procede a aplicar los algoritmos seleccionados a los datos obtenidos de los estudiantes, para luego interpretar y evaluar los resultados.

En los siguientes apartados se presentan los resultados de la aplicación de las técnicas de agrupamiento (clustering) y clasificación, y además se procede analizar la calidad de los variables del conjunto de datos con la ayuda del algoritmo ChiSquaredAttributeEval.

Resultados de la aplicación de la técnica de agrupamiento

Se creyó conveniente aplicar el algoritmo Simple-KMeans, para generar los grupos por carrera con las variables consideradas para el análisis de la deserción en cada una de las materias de las carreras seleccionadas, obteniendo en cada una los siguientes resultados:

En cuanto a la carrera de Jurisprudencia correspondiente al área Sociohumanística se ha podido determinar que:

- Los estudiantes que reprueban en una o ambas materias troncales, han desertado la carrera con mayor frecuencia.

- El tipo de pago de matrícula, no es una variable que influye directamente como factor socioeconómico, para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes desertores han cancelado la matricula al contado.

- Existen estudiantes que a pesar de haber aprobado en alguna de las materias de formación básica, constan como desertores.
- La mayoría de estudiantes desertores de 1er ciclo de la carrera de Jurisprudencia, poseen una edad de 16 a 26 años, siendo estos las personas que no poseen la suficiente seguridad de haber elegido la carrera idónea.

- La mayoría de estudiantes desertores han obtenido una nota de menos 27 puntos en las asignaturas de 1er ciclo de la carrera de Jurisprudencia, de los cuales gran parte no se han presentado a dar el correspondiente examen supletorio, por ende han reprobado la asignatura, y la mayoría de los mismos son desertores.

- El género y el estado civil son datos personales que no influyen, para que un estudiante decida desertar la carrera.

- El nivel de interacción del profesor no tiene una influencia importante, para que los estudiantes deserten la carrera, ya que la mayoría de docentes de las materias de 1er ciclo han obtenido una interacción Alta en los cursos, a pesar de ello la mayoría de estudiantes han desertado.

- El nivel de interacción del estudiante no posee una alta influencia para que los estudiantes decidan desertar, ya que existen estudiantes que han obtenido un nivel alto y medio en la interacción, sin embargo han reprobado la asignatura y constan como desertores.

- Si el estudiante presenta o no presenta todas las evaluaciones tanto presencial, como a distancia de los dos bimestres de la asignatura, no influye para que el estudiante decida desertar la carrera, ya que existen estudiantes que a pesar de haber presentado todas las evaluaciones, han reprobado, y han desertado la carrera, tomando en cuenta que las notas de dichas evaluaciones son bajas.

- Existen un considerado número de estudiantes que sabiendo que están en supletorio, por un bajo rendimiento académico, deciden no presentarse al supletorio, de los cuales la mayoría decide desertar.

- El nivel de Interacción del Profesor es influyente para que los estudiantes puedan pre-sentar sus evaluaciones, ya que si el docente no responde a las inquietudes de los estudiantes, y no habilita el enlace respectivo para que los estudiantes puedan subir las evaluaciones; los alumnos tendrán inconvenientes para presentar dichas evaluaciones.

- La mayoría de estudiantes que aprueban las asignaturas, han presentado todas las evaluaciones tanto presencial y a distancia de la misma, tomando en cuenta que dichos estudiantes han obtenido buenas calificaciones para poder aprobarla.

En cuanto a la carrera de Administración de Empresas correspondiente al área se ha podido determinar que:

- El género, estado civil, son atributos personales que no influyen para que un estudiante deserte la carrera.

- El tipo de pago de matrícula, no influye como factor socioeconómico para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes matriculados en las materias de 1er ciclo de la carrera de Administración de Empresas, son desertores y han cancelado la matricula al contado.

- La mayoría de estudiantes que constan como desertores, han reprobado en al menos una asignatura troncal de la carrera de Administración de Empresas.

- No necesariamente los estudiantes que poseen una Alta interacción en el curso son los más propensos a aprobar la asignatura, es decir que el nivel de interacción del estudiante no es una variable que posee una alta influencia, para que un estudiante pueda aprobar la asignatura, ya que también depende del puntaje que obtenga en las evaluaciones de la asignatura.

- Las variables con mayor influencia con respecto a la deserción son: la nota final, el estado de aprobación, conocer si está en el supletorio de la asignatura, conocer si asistió a dar la correspondiente evaluación supletoria, y saber si ha presentado todas las evaluaciones de la materia.

- Se pudo constatar que las variables que poseen un menor nivel de importancia en el dataset son el Nivel de Interacción del Estudiante y del Profesor, ya que existen estudiantes que a pesar de haber obtenido un nivel alto en la interacción en el curso, y a pesar de haber tenido un docente que ha obtenido también un nivel de interacción alto en el curso, de igual manera han reprobado la asignatura.

- Existen estudiantes que a pesar de haber aprobado en alguna de las materias de formación básica, constan igualmente como desertores.

- Una considerable cantidad de la población de matriculados en asignaturas de 1er ciclo de la carrera de Administración de Empresas no asiste a rendir la respectiva evaluación supletoria, de los cuales la mayoría constan como desertores.

En cuanto a la carrera de Gestión Ambiental correspondiente al área Biológica se ha podido determinar que:

- La mayoría de estudiantes que han desertado la carrera, son hombres solteros y poseen una edad entre los 16 a 26 años, tomando en cuenta que la mayor parte de estudiantes matriculados en 1er ciclo de la presente carrera, poseen las características antes mencionadas.

- La variable tipo de pago de matrícula no posee demasiada influencia, como factor socioeconómico, para que el estudiante deserte la carrera.

- El nivel de interacción del profesor en el curso, no posee una influencia considerable para que un estudiante no decida desertar la carrera, ya que a pesar de que el docente de la presente asignatura obtuvo un nivel de interacción Alto de igual manera la mayoría de estudiantes de la materia constan como desertores.

- Existen estudiantes, que a pesar de haber obtenido un nivel de interacción alto en el curso, de igual manera han reprobado y desertado la carrera, además que también existen estudiantes que han obtenido un nivel de interacción Bajo, que de igual manera han aprobado; por lo tanto no es una variable que posee una influencia importante para que un estudiante pueda aprobar o reprobar la asignatura, ya que también depende de las notas que obtenga en las evaluaciones presenciales y a distancia de la asignatura.

- La presentación de todas las evaluaciones de la materia no asegura que el estudiante pueda aprobar la asignatura, ya que también depende del puntaje que obtenga en cada una de ellas.

- La mayoría de estudiantes que desertan, es porque han reprobado en alguna asignatura, y son aún más propensos a desertar, si dicha materia forma parte del grupo de las troncales de la carrera.

- La mayoría de estudiantes que se presentan a dar la evaluación supletoria, no aprueban la asignatura, y son los más propensos a desertar la carrera.

Finalmente en la carrera de Informática correspondiente al área técnica se ha podido determinar que:

- La mayoría de estudiantes desertores, son solteros, de estado civil masculino, y poseen una edad entre los 16 a 26 años, tomando en cuenta que dichos valores contienen la mayor cantidad de instancias.

- El tipo de pago de matrícula no es un atributo que posee demasiada influencia, como factor socioeconómico, para que el estudiante deserte la carrera, ya que la mayoría de estudiantes que desertaron la carrera, han pagado la matricula al contado.

- Los estudiantes son más propensos a desertar la carrera, al momento de reprobar una materia que pertenece al grupo de las troncales de la carrera.

- La presentación de todas las evaluaciones a distancia, no asegura que el estudiante apruebe o repruebe la materia, ya que también depende del puntaje que obtenga en cada una de ellas.

- El nivel de interacción del estudiante y del profesor no son variables, que influyen a gran escala para que un estudiante pueda aprobar o reprobar una asignatura, ya que también depende de las notas que obtenga las evaluaciones tanto presencial como a distancia correspondientes a la asignatura que estén cursando.

- La mayoría de estudiantes que no presentan todas las evaluaciones correspondientes a la asignatura, teniendo que rendir la evaluación supletoria, ya no se presentan a efectuarla.

- Los atributos que forman parte del rendimiento académico del estudiante son los que poseen una mayor influencia para que el estudiante decida desertar o no la carrera, como son la Nota_Final, Estado_Aprobacion, Supletorio, Asistio_Supletorio, Present_todas_las_eval.

- La mayoría de estudiantes que se han quedado en supletorio de una determinada asigna-tura, han reprobado la misma.

De estos resultados podemos deducir que: la mayoría de estudiantes desertores de todas las carreras poseen una edad entre los 16 a 26 años, que el tipo de pago de matrícula no ha influido, como factor socioeconómico, para que el estudiante deserte la carrera. Se determinó también que el nivel de interacción del estudiante y del profesor en el entorno vir-

tual del curso no son variables, que influyen a gran escala para que un estudiante pueda deserte la carrera, ya que existen estudiantes que a pesar de haber obtenido un nivel de interacción alto o medio, de igual manera han abandonado la carrera. Además se puedo observar que la mayoría de estudiantes que reprueban alguna asignatura correspondiente a las troncales de la carrera, son posibles desertores de la carrera que están siguiendo.

Resultados de la aplicación de la técnica de clasificación

Para realizar la clasificación de los datos se utilizó árboles de decisión, aplicando el algoritmo J48. Para las cuatro carreras seleccionadas para el análisis se ha creído conveniente realizar la clasificación de los estudiantes que han desertado, considerando también el número de materias troncales reprobadas, el presente experimento se lo realizó para comprobar lo obtenido con la técnica de clustering, puesto que cuando se construyeron los grupos se comprobó que la mayoría de estudiantes que reprueban alguna asignatura, desertan la carrera, por lo cual en este caso se consideró esta variable (NoMateriasTroncalesRep), para con ello analizar de mejor manera la deserción estudiantil.

Con respecto a los resultados de la clasificación obtenidos en la carrera de Jurisprudencia, se ha podido determinar que: un total de 272 estudiantes, que no reprobaron ninguna materia troncal de 1er ciclo, como resultado de ello no desertaron la carrera, solo un total de 227 que no reprobaron en alguna materia troncal, desertaron la carrera. Se observa en el árbol además que, un total de 761 estudiantes que si reprobaron en alguna o ambas, de las materias troncales correspondientes, como resultado de ello desertaron la carrera, y solo un total de 42 estudiantes que reprobaron en una o más materias troncales, no decidieron desertar la carrera. Concluyendo con ello que: la mayoría de estudiantes que han desertado la carrera han reprobado en alguna o ambas materias troncales ofertadas en primer ciclo de la carrera de Jurisprudencia. Tomando en cuenta que las Materias Troncales ofertadas en cada carrera, implican aplicarle mayor tiempo para su comprensión; ya que el contenido de las mismas es fundamental en la carrera. Existen estudiantes que además no poseen las bases necesarias para comprender con mayor facilidad dicho contenido de la asignatura, puesto que no han elegido la carrera idónea, según su perfil profesional.

En la carrera de Administración de Empresas, se ha podido determinar que: que un total de 190 estudiantes, que no reprobaron ninguna materia troncal de 1er ciclo, como resultado de ello un total de 169 estudiantes no desertaron la carrera. Se observa en el árbol además que, un total de 798 estudiantes que si reprobaron en alguna o ambas materias troncales, como resultado de ello un total de 686 estudiantes desertaron la carrera, y solo un total de

21 estudiantes que reprobaron en una o más materias troncales, no decidieron desertar la carrera. Deduciendo con ello que: la mayoría de estudiantes que han decidido desertar la carrera de Administración de Empresas, han reprobado una o ambas asignaturas troncales; como de igual manera existe gran parte de estudiantes que al momento de no reprobar ninguna materia, correspondientes a las troncales de la carrera, no constan como desertores. Es decir que al momento que un estudiante de 1er ciclo repruebe en alguna de las materias troncales sería un posible desertor de la carrera de Administración de Empresas.

Con respecto a los resultados de la clasificación obtenidos en la carrera de Gestión Ambiental, se ha podido determinar que: de un total de 173 estudiantes no desertores, 150 no han desertado la carrera, ya que no han reprobado, ninguna materia troncal de 1er ciclo, además la mayoría de dichos estudiantes están casados. Se puede visualizar además, que de un total de 541 estudiantes desertores, 355 han reprobado al menos una materia troncal, los cuales constan como desertores en la carrera. Con la presente interpretación se ha podido constatar, el mismo suceso, que se verifico en las carreras anteriores, ya que la mayoría de estudiantes de Gestión Ambiental, han decidido desertar la carrera, porque han reprobado al menos una materia troncal; y como también existe estudiantes, que como han aprobado al menos una materia troncal, los mismos, no han decidido desertar la carrera; por lo tanto los estudiantes que reprueban al menos una materia troncal son más propensos a desertar la carrera. [*Ver Figura. 1*]

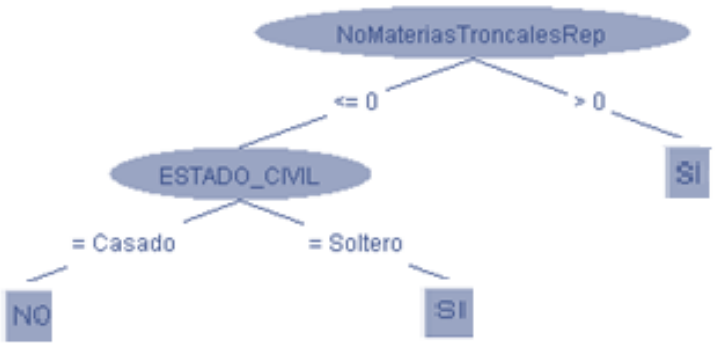

Figura. 1. Gráfica del Árbol de Decisión – Deserción según el Número de materia troncales reprobadas - Carrera de Gestión Ambiental

En la carrera de Informática, se ha podido determinar que: de un total de 83 estudiantes no desertores, 72 no han reprobado ninguna materia del grupo de las troncales de 1er ciclo de la carrera. Además se puede observar que de 363 estudiantes, que han desertado la carrera, un total de 304, han reprobado al menos 1 materia troncal perteneciente a la carrera. Con la presente interpretación se ha podido constatar, el mismo comportamiento, que se

verifico en las carreras anteriores, ya que la mayoría de estudiantes de 1er ciclo de Informática, han decidido desertar la carrera, porque han reprobado al menos una materia troncal; y los estudiantes, que han aprobado al menos una materia troncal, no han desertado; por lo tanto existe mayor posibilidad de deserción, en los estudiantes que reprueban al menos una materia troncal.

De estos resultados propuestos por la técnica de clasificación, podemos deducir, que la mayoría de estudiantes que reprueban uno o ambas materias troncales ofertadas en primer ciclo de la carrera que siguen, son más propensos a desertar la misma.

Resultados de la influencia de las variables

Para evaluar el nivel de calidad de las variables utilizadas, se creyó conveniente aplicar métodos de selección de atributos, que propone la herramienta Weka.

En la tabla 4, se muestra los resultados del algoritmo evaluador ChiSquaredAttributeEval, aplicado a los datos de la de la carrera de: Jurisprudencia, Administración de Empresas, Gestión Ambiental e Informática, el mismo que calcula el valor estadístico Chi-cuadrado de cada variable con respecto a la clase (variable Desertor) y obtiene el nivel de correlación entre la clase y cada variable, brindando con ello un ranking del nivel de influencia entre los mismos.

En la tabla siguiente se puede visualizar el orden de las variables según el nivel de correlación que existe entre ellas, con respecto a la clase Desertor. El algoritmo establece que en todas las carreras analizadas, las variable que poseen una mayor relación respecto a la clase Desertor son: Nota_Final, Estado_Aprobacion, Asistio_Supletorio, Supletorio, siendo la Nota_Final la que posee mayor influencia; por lo tanto, analizando dichos resultados, se puede deducir que las variables que forman parte del rendimiento académico (Notas) del estudiante son las que influyen con mayor frecuencia en la deserción o no del estudiante.

Tabla 4. *Ranking de Variables – ChiSquaredAttributeEval*

ChiSquaredAttributeEval					
Variable	**Ranked de Jurisprudencia**	**Ranked de Administración de Empresas**	**Ranked de Biología**	**Ranked de Informática**	**Clase**
NOTA_FINAL	1886.2783	1311.70539	808.27983	622.2835	DESERTOR
ESTADO_APROBACION	1758.1897	1177.21583	757.92552	543.0159	
ASISTIO_SUPLETORIO	1484.9648	1086.49124	378.51546	455.7962	
SUPLETORIO	1304.035	975.46538	521.83158	398.7645	
PRESENT_TODAS_LAS_EVAL	581.0545	493.65276	304.25327	231.8646	
EDAD	99.6954	113.53136	84.45013	43.1183	
ESTADO_CIVIL	99.1503	82.82951	27.55783	3.9702	
TIPO_PAGO_MATRICULA	10.4938	11.36131	23.62061	0.0411	
GENERO	3.5506	0.17283	10.03355	0.0575	
CURSO	1.167	0.57289	0.40159	1.1524	
NIVEL_INTER_PROF	0.6269	0.0047	0.00288	0.6713	
NIVEL_INTER_EST	0.3568	0.23568	0.97687	1.7444	

3.1. Gráficos de Dispersión

Respecto a la carrera de Jurisprudencia, se ha obtenido la gráfica de la ***Distribución de la Deserción por Curso,*** la misma que se ilustra en la [*Figura. 2*] en donde se observa una serie de colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza que la mayoría de estudiantes que han cursado las asignaturas de la carrera de Jurisprudencia, constan como desertores. Siendo la materia de Introducción al Derecho, la que posee la mayoría de desertores.

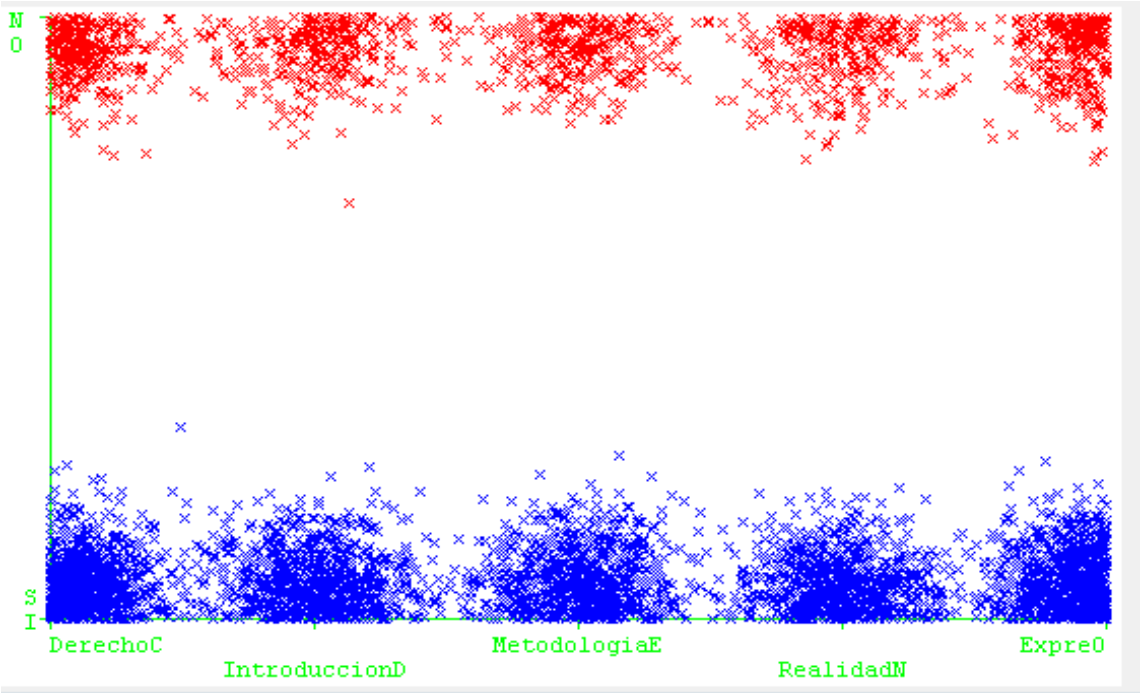

Figura. 2. Distribución de la Deserción por *Curso*

Respecto a la carrera de Administración de empresas, se ha obtenido la gráfica de la ***Distribución de Interrelación de las variables: Estado de Aprobación - Curso - Desertor,*** la misma que se ilustra en la [*Figura. 3*] en donde se observan dos colores de la variable desertor, que representan (SI = azul, NO =rojo). En la imagen puede visualizar, que el curso Administración I posee mayor cantidad de Reprobados, y el curso de Metodología de Estudio es el que posee mayor cantidad de Aprobados. Podemos observar, que los estudiantes que han cursado las materias de Administración I Contabilidad General y han reprobado, son en su mayoría desertores.

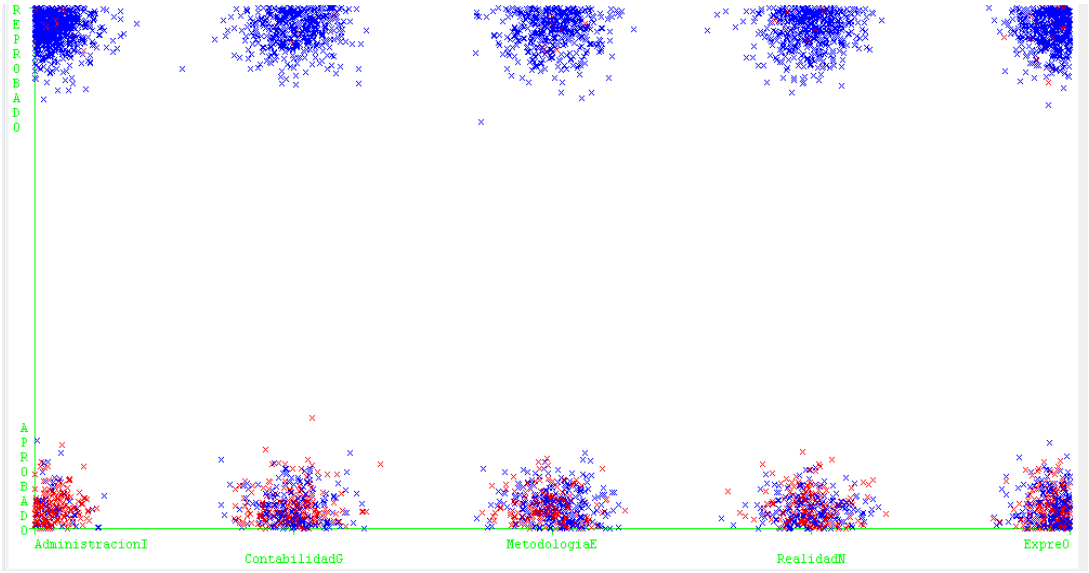

Figura. 3. Interrelación Estado Aprobación – Curso – Desertor

Respecto a la carrera de Gestión Ambiental, se ha obtenido la gráfica de la observan ***Distribución de Interrelación entre las variables de: Estado de Aprobación – Edad – Desertor,*** la misma que se ilustra en la [*Figura. 4*] en donde se observan dos colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la edad de 16 a 26 años posee la mayor población, seguida con una considerable diferencia, la edad de 27 a 37 años. Se puede observar además, en la gráfica, que los estudiantes que poseen una edad de 16 a 26 reprueban con mayor frecuencia y de igual manera son los que desertan con mayor frecuencia. Es importante tomar en cuenta que la mayoría de estudiantes matriculados en 1er ciclo de la carrera de Gestión Ambiental posee entre los 16 a 26 años de edad.

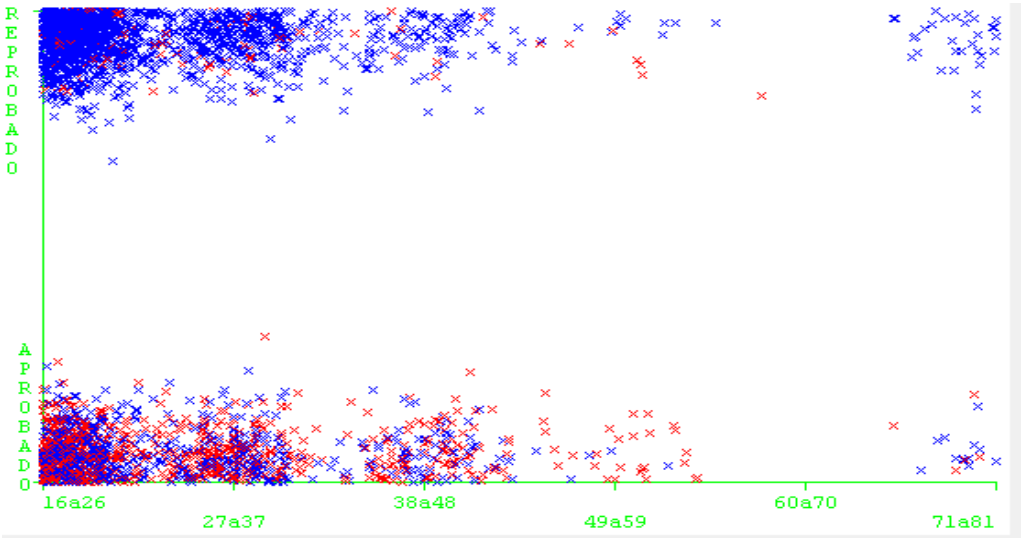

Figura. 4. Interrelación Estado Aprobación – Edad – Desertor

Respecto a la carrera de Gestión Ambiental, se ha obtenido la gráfica de la observan ***Distribución de Interrelación de las variables: Supletorio – Nota Final– Desertor,*** la misma que se ilustra en la [*Figura. 5*] en donde se observan dos colores de la variable desertor, que representan (SI = azul, NO =rojo). Se visualiza en la imagen que la Nota Final que posee menor cantidad de instancias es A = 39 a 40, y la nota que posee mayor cantidad de instancias es FX = 14 a 28 y F = 13 o menos, siendo estas las menores calificaciones, que puede obtener un alumno. En la gráfica también se observa que la mayoría de estudiantes se han quedado en supletorio en las asignaturas de 1er ciclo de la presente carrera.

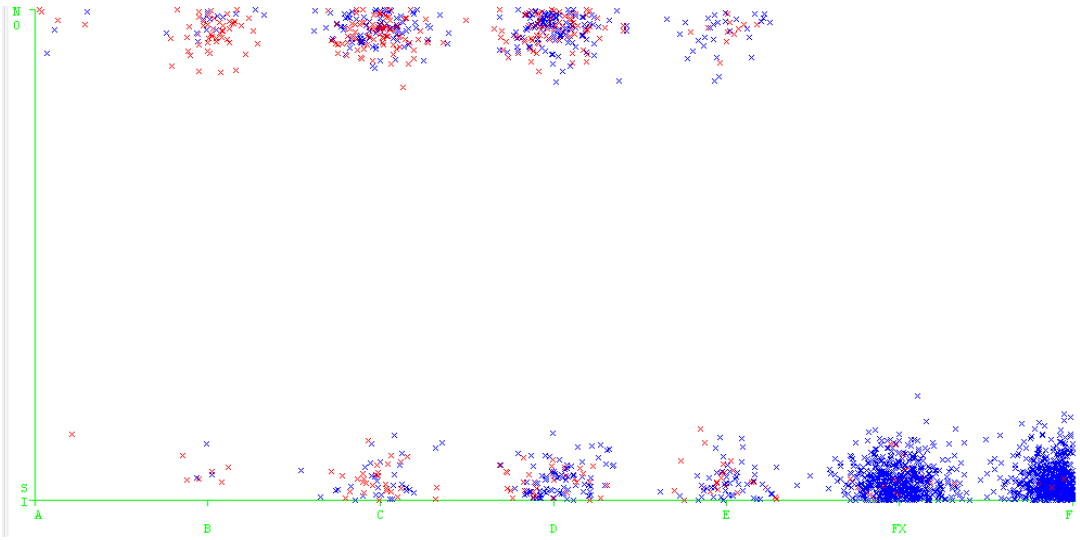

Figura. 5. Interrelación Supletorio – Nota Final – Desertor

4. DISCUSIÓN

A través de los resultados obtenidos, se puede concluir que las técnicas de minería de datos poseen importantes ventajas, para poder descubrir patrones de comportamiento de un estudiante que deserta una carrera, ya que brinda un alto valor agregado para el análisis y la generación del nuevo conocimiento.

Referente a las técnicas empleadas para la generación de los modelos, se pudo constatar que la técnica de agrupamiento (clustering), brinda resultados eficaces referentes a los problemas de deserción estudiantil, ya que, con la mencionada técnica se pudo conocer las características principales de un posible desertor. Además se encontró que los árboles de decisión es una técnica, que en cierta medida facilita, realizar experimentos específicos, para con ello conocer con mayor exactitud las razones de la deserción estudiantil en la modalidad abierta y a distancia de la UTPL.

La carreras analizadas han tenido un similar comportamiento ya que la mayoría de estudiantes desertores han reprobado en al menos una materia troncal, aunque en las carreras analizadas (Jurisprudencia, Administración de Empresas, Informática y Gestión Ambiental) existen estudiantes, que a pesar de haber aprobado alguna de las materias de formación básica, de igual manera han desertado. De los resultados obtenidos se pudo además observar que la carrera de Informática perteneciente al Área Técnica es la que posee, un mayor porcentaje de deserción, siendo este el 81,51%, diferencia de las otras carreras.
Así mismo, podemos concluir que la mayoría de estudiantes desertores en las 4 carreras analizadas, una considerable cantidad de estudiantes, que no han presentado todas las evaluaciones de las asignaturas que cursan, y que no se han presentado a la evaluación supletoria, han reprobado las asignaturas y por tanto en su mayoría son desertores.

Según los resultados obtenidos en las cuatro carreras analizadas, se puedo constatar además que las variables: género, estado, civil; no influyen significativamente para que un estudiante deserte la carrera. Además se determinó que el tipo de pago de matrícula, no es una variable que influye directamente, para que un estudiante decida desertar la carrera, ya que la mayoría de estudiantes desertores han cancelado la matricula al contado.

De igual manera se pudo constatar que la interacción del estudiante en el EVA no posee una alta influencia para que los estudiantes deserten la carrera, ya que existen estudiantes que han obtenido un nivel alto y medio en la interacción, sin embargo han reprobado la asignatura y constan como desertores. Por lo tanto la dedicación y el desempeño que aplique el estudiante en las evaluaciones tanto presencial como a distancia de las asignaturas que cursa, son las que influirán en mayor porcentaje en la deserción de un estudiante.

Se pudo constatar además, según el análisis realizado en las cuatro carreras, que si el estudiante presenta o no todas las evaluaciones tanto presencial, como a distancia de los dos bimestres de la asignatura, no influye en gran escala para que el estudiante decida desertar la carrera, ya que existen estudiantes que a pesar de haber presentado todas las evaluaciones, han reprobado, y han desertado la carrera, por lo tanto para que el estudiante apruebe la materia depende del puntaje que obtenga en dichas evaluaciones, ya que una considerable cantidad de estudiantes que si presentan todas las evaluaciones mencionadas, han obtenido una baja calificación en las mismas.

Finalmente, con la ayuda de las técnicas de minería de datos y los algoritmos evaluadores de variables, se ha logrado determinar que las variables que forman parte del rendimiento académico del estudiante, son los que poseen una mayor influencia en la deserción del estudiante.

REFERENCIAS BIBLIOGRÁFICAS

Hand, D.J., Mannila, H. & Smyth, P. (2011). *Principles of Data Mining*. Cambridge, USA: MIT Press.

Pinzón, L. (2011). Aplicando minería de datos al marketing educativo. *Notas de Marketing*. Vol(1), n. 1, 45-61. Recuperado de: http://www.usergioarboleda.edu.co/investigacionmarketing/marketing/articulo5MineriaDatos.pdf

Spositto, O., Etcheverry, M., Ryckeboer, H., Bossero, J. (2008). Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil. Argentina. Recuperado de http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

Pautsch, J. (2008). Minería de Datos aplicada al análisis de la deserción en la Carrera de Analista en Sistemas de Computación. Argentina. Recuperado de: http://www.iiis.org/CDs2010/CD2010CSC/CISCI_2010/PapersPdf/CA156FK.pdf

Domínguez, M. (2008). Minería de datos y lógica difusa como método para la predicción del abandono escolar de alumnos de institutos de nivel superior privado. México. Recuperado de: http://pcti.mx/tesis-de-posgradoen-mexico?task=callelement&format=raw&item_id=382&element=5832706c-3ae3-408b-93e3-bca7418d0376&method=download

Vaira, S. Avila, O. Ricardi, P. Bergesio, A. (2010).Deserción universitaria. Un caso de estudio: variables que influyen y tiempo que demanda la toma de decisión. *Revista FABICIB*. Vol(14), 107-115. Recuperado de: http://bibliotecavirtual.unl.edu.ar:8180/publicaciones/bitstream/1/2946/1/FABICIB_14_2010_pag_107_115.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R., (2000). CRISP-DM 1.0- Guía paso a paso de Minería de Datos. Recuperado de: http://www.dataprix.com/files/Metodologia_CRISP_DM.pdf

Rubio, M. (2009). *Nuevas Orientaciones y metodología para la educación a distancia*. Loja: Universidad Técnica Particular de Loja.

Mejía, C., Mancera, L., Gómez, S., Baldiris, S., Fabregat, R. (2008). Supporting Competence upon DotLRN through Personalization. 7th OpenACS / .LRN conference. Valencia (Spain). 18-19. November 2008.