UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TÉCNICA

TITULO DE INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

Aplicación de técnicas de minería de datos para diseñar un modelo descriptivo de la labor tutorial en las asignaturas de formación básica de modalidad a distancia.

TRABAJO DE TITULACIÓN.

AUTORA: Maza Quezada, Mariuxi Mabel.

DIRECTOR: Riofrio Calderon, Guido Eduardo, Ing.

LOJA – ECUADOR

2017

Septiembre, 2017

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN

Ingeniero.

Guido Eduardo Riofrio Calderon.

DOCENTE DE TITULACIÓN

De mi consideración:

El presente trabajo de titulación: Aplicación de técnicas de minería de datos para diseñar un modelo descriptivo de la labor tutorial en las asignaturas de formación básica de modalidad a distancia por Mariuxi Mabel Maza Quezada, ha sido orientado y revisado durante su ejecución, por cuanto se aprueba la presentación del mismo.

Loja, Febrero de 2017

f)……

DECLARACIÓN DE AUDITORÍA Y CESIÓN DE DERECHOS

"Yo Mariuxi Mabel Maza Quezada declaro ser autor (a) del presente trabajo de titulación: Aplicación de técnicas de minería de datos para diseñar un modelo descriptivo de la labor tutorial en las asignaturas de formación básica de modalidad a distancia, de la Titulación de Sistemas Informáticos y Computación, siendo Guido Eduardo Riofrio Calderon director del presente trabajo; y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales. Además, certifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo, son de mi exclusiva responsabilidad.

Adicionalmente declaro conocer y aceptar la disposición del Art. 88 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 o trabajos de titulación que se realicen con el apoyo financiero, académico o institucional (operativo) de la Universidad"**.**

.....
Mariuxi Mabel Maza Quezada
1104872849

DEDICATORIA

Este proyecto está dedicado a Dios, quien sin lugar a duda ha estado presente en cada instante minúsculo de mi vida, en la presencia de mis padres, hermano, familia, amigos, docentes y desconocidos permitiéndome sentir su presencia y su amor incondicional.

A mis queridos padres Sara Quezada y Vicente Maza, quienes sembraron en mi la perseverancia y valentía ante los obstáculos que se presentaron en el camino para poder obtener este logro universitario, siendo así mi motivación para cada día ser no solo mejor profesional si no también mejor persona.

A mi hermano Javier Maza que sin dudarlo me brindó su apoyo incondicional, haciéndome sentir que nada es imposible de alcanzar, cuando amas lo que haces.

Con humildad y cariño este título es dedicado para ustedes mi amada familia.

Mariuxi Maza

AGRADECIMIENTO

Todo lo que conlleve sacrificios y mucha constancia va acompañado del apoyo infinito de quienes nos quieren y desean lo mejor para nosotros, es por ello que por medio de este espacio quiero agradecer primeramente a Dios por guiar mi camino, y permitirme llenar este periodo de mi vida con experiencias muy enriquecedoras.

A mis padres, hermano y familia en general que siempre me brindaron su apoyo incondicional en todo momento, su paciencia, su confianza y sus sacrificios para poder cumplir este logro profesional. Mil gracias familia, todo esto sin duda les será multiplicado.

A mi universidad que gracias a su gran misión de formar profesionales dispuestos al servicio de los demás, brindan a los estudiantes el apoyo necesario para que cumpla con sus metas, ya sea por medio de becas o reconocimientos académicos, permitiéndonos alcanzar una profesión universitaria.

A mi tutor de Tesis Ing. Guido Riofrio quien guio mi trabajo sin estimar tiempo, transmitiendo sus conocimientos, compartiendo experiencias y poniendo a mi disposición los recursos que demandaba el desarrollo de este trabajo de fin de titulación. Gracias.

A la titulación de Sistemas Informáticos y Computación, a mis docentes quienes enriquecieron mis conocimientos a lo largo de esta formación profesional.

Y como no agradecer a mi amigos/as que sin duda fueron un gran regalo de Dios en esta etapa universitaria gracias por su apoyo y consideración.

Un gracias muy grande para todos ustedes.

Mariuxi Maza

ÍNDICE DE CONTENIDOS

APROBACIÓN DEL DIRECTOR DEL TRABAJO DE TITULACIÓN i
DECLARACIÓN DE AUDITORÍA Y CESIÓN DE DERECHOS ii
DEDICATORIA iii
AGRADECIMIENTO iv
ÍNDICE DE CONTENIDOS v
ÍNDICE DE ILUSTRACIONES vii
ÍNDICE DE TABLAS viii
RESUMEN 1
ABSTRACT 2
INTRODUCCION 3
CAPÍTULO I: MARCO TEORICO 5
1.1 Minería de datos (data mining). 6
1.1.1 Definiciones. 6
1.1.2 Criterios para aplicar la minería de datos. 7
1.1.3 Técnicas data mining. 8
1.1.4 Herramientas de data mining. 13
1.1.5 Áreas de aplicación de data mining. 17
1.2 Metodologías para el desarrollo de proyectos de minería de datos. 22
1.2.1 Metodología SEMMA. 22
1.2.2 Metodología CRISP-DM. 24
1.2.3 KDD. 27
1.2.4 Estudio comparativo. 29
1.3 Educación A Distancia. 30
1.3.1 Importancia de la Educación a distancia. 31
1.3.2 Teorías de modelos de educación a distancia. 31
1.3.3 Proceso instruccional. 34
1.4 E-learning. 36
1.4.1 Evolución de E-learning. 36
1.4.2 Recursos de aprendizaje. 38
1.5 Materias básicas impartidas en las universidades. 38
1.6 Perfiles de los docentes de educación a distancia. 38
1.7 Casos de deserción a inicios de una carrera universitaria. 43
1.8 Programa de formación básica en la UTPL. 45
CAPÍTULO II: DEFINICIÓN Y PLANTEAMIENTO DE LA SOLUCIÓN DEL PROBLEMA 46
2.1 Definición del problema. 47

2.2 Planteamiento de la solución del problema. 48
CAPÍTULO III: PROPUESTA DEL MODELO 52
3.1 Fase 1. Comprensión del negocio. 53
3.1.1 Objetivos del negocio. 53
3.1.2 Evaluación de la situación. 53
3.1.2.1 Recursos. 53
3.1.2.2 Requerimientos. 54
3.1.2.3 Supuestos. 54
3.1.2.4 Restricciones. 54
3.1.2.5 Terminología. 55
3.1.3 Objetivos de la minería. 56
3.1.4 Plan del proyecto. 56
3.2 Fase 2. Comprensión de los datos. 57
3.2.1 Recolección de datos. 58
3.2.2 Descripción de los datos. 60
3.2.3 Exploración de los datos. 61
3.2.4 Verificación de la calidad de los datos. 63
3.3 Fase 3. Preparación de los datos. 64
3.3.1 Selección de datos. 64
3.3.2 Limpieza de datos 65
3.3.3 Construcción e Integración de datos. 66
3.4 Fase 4. Modelo. 66
3.4.1 K-means. 67
1. Selección de técnica de modelado 67
2. Optimización de la técnica de agrupación k-means. 67
3. Descripción del modelado con técnicas k-means. 69
3.4.2 Análisis de componentes principales. 71
1. Construcción del modelo con técnica de ACP. 73
2. Descripción del modelo a través de la técnica de ACP. 77
3.5 Fase 5. Evaluación. 78
3.5.1 Discusión de resultados. 79
CONCLUSIONES 85
RECOMENDACIONES 86
BIBLIOGRAFIA 87
ANEXOS 95

ÍNDICE DE ILUSTRACIONES

Ilustración 1. Técnicas de minería de datos 8
Ilustración 2. Herramientas con más usabilidad 16
Ilustración 3. Fases de la metodología de SEMMA 22
Ilustración 4. Dinámica de la metodología SEMMA 24
Ilustración 5. Niveles de abstracción de la metodología CRISP-DM 24
Ilustración 6. Ciclo de vida del proceso de modelado de CRISP-DM 25
Ilustración 7. Proceso de KDD 27
Ilustración 8. Metodologías más utilizadas en minería de datos 29
Ilustración 9. Modelos instruccionales de EaD 35
Ilustración 10. Evolución de E-Learning 37
Ilustración 11. Los cuatro pilares de la educación 39
Ilustración 12. Perfil de docente EaD 42
Ilustración 13. Pasos de K-mean 50
Ilustración 14. Data Seleccionada 57
Ilustración 15. Datos a correlacionar 61
Ilustración 16. Correlación de campos 62
Ilustración 17. Regla de asociación 63
Ilustración 18. Resultado de reglas de asociación 63
Ilustración 19. Data final 66
Ilustración 20. Selección de campos y data numérica 67
Ilustración 21. Inicio para la validación de k-means 68
Ilustración 22. Codo de JAMBU, estabilidad de métodos 68
Ilustración 23. Inercias inter-clases 69
Ilustración 24. Aplicación de k-means con algoritmo Hartigan y 4K 70
Ilustración 25. Sentencia plot 71
Ilustración 26. Análisis de componentes en 2D, clustplot de 3 clusters. 72
Ilustración 27. plotcluster de 3 clusters 72
Ilustración 28. plotcluster de 4 clusters 73
Ilustración 29. Desviación estándar 74
Ilustración 30. Carga de variable en PC1 74
Ilustración 31. Importancia de componentes 75
Ilustración 32. Proporción de varianza 76
Ilustración 33. Valores propios con 3 CP 76
Ilustración 34. Variables seleccionadas 77
Ilustración 35. Representación gráfica de los PC 78
Ilustración 36. Presentación de la unión de datos 81

ÍNDICE DE TABLAS

Tabla 1. Herramientas de data mining 14
Tabla 2. Metodologías aplicadas a data mining 26
Tabla 3. Teorías de modelos de EaD 33
Tabla 4. Programa tutorial de educación a distancia 43
Tabla 5. Descripción de muestra 58
Tabla 6. Muestra de atributos 59
Tabla 7. Nombre de las tablas de la BD del SYLLABUS 60
Tabla 8. Resultado de análisis de correlación 62
Tabla 9. Campos y atributos seleccionados 64
Tabla 10. Limpieza de datos 65
Tabla 11. Cluster de 4 71
Tabla 12. Cluster de 3 71
Tabla 13. Nuevas variables 77

RESUMEN

Con el fin de desarrollar un modelo descriptivo acerca de la labor tutorial del docente de educación abierta y a distancia de la UTPL, en este proyecto se aplican técnicas de minería de datos como: k-means y análisis de componentes principales dentro de la herramienta R, utilizando la metodología CRISP-MD. Los datos utilizados para el análisis son en base a resultados cuantitativos obtenidos de las actividades en el entorno virtual de aprendizaje (EVA) como: foros, chats, video conferencias, y quiz; de las asignaturas de formación básica de metodología del aprendizaje, desarrollo espiritual y expresión oral y escrita dentro de los periodos oct-2014/feb-2015 y abr-2015/ago-2015.

Presentando como resultado que la técnica de análisis de componentes principales es la más óptima para el trabajo de los datos, y con la que son definidos 3 componentes principales, nombrados con las etiquetas de: debate, consultas y comunicación, conteniendo cada una las acciones en las que los estudiantes realizan mayor actividad dentro del EVA.

PALABRAS CLAVE: EVA, minería de datos, metodología CRISP-DM, clustering, componentes principales.

ABSTRACT

In order to develop a descriptive model about the tutorial work of the open and distance education teacher of the UTPL, this project applies data mining techniques such as: k-means and analysis of main components within the tool R, Using the CRISP-MD methodology; The data used for the analysis are based on quantitative results obtained from activities in the virtual learning environment (EVA) such as: forums, chats, video conferences, and quiz; Of the basic training subjects of methodology of learning, spiritual development and oral and written expression within the periods of October-2014 / Feb-2015 and Apr-2015 / Aug-2015.

As a result, the main component analysis technique is the most optimal for data work, which allows defining 3 main components with discussion, query and communication tags, which contain the actions in which the students perform Greater activity within the EVA.

KEY WORDS: EVA, data mining, CRISP-DM methodology, clustering, main components.

INTRODUCCION

El presente trabajo nace de la problemática sobre la deserción de estudiantes, el cual se presenta en universidades de todo el mundo y entre ellas la UTPL. (Arévalo & Maldonado, 2010) han ubicado que, dificultades económicas, disponibilidad de tiempo del estudiante, problemas académicos, matriculación, dificultades familiares, es lo que promueve esta deserción, sumándose un incorrecto diseño de planes de estudio o mallas curriculares, debilidad en metodologías de enseñanza y aprendizaje como indica (Centro Microdatos, 2008). Por tal razón se han desarrollado proyectos de minería de datos que orienten a las instituciones, para disminuir este problema de deserción, entre ellos está el de (Malbernat, Clemens, Varela, & Urrizaga, 2015) que permitió ubicar los tipos de docentes con los que cuenta su institución, clasificándolos en grupos como: innovadores, flemático, refractarios o indiferentes.

Con el fin de brindar un aporte de ayuda a este problema, se desarrolló un modelo descriptivo de la labor tutorial en las asignaturas de formación básicas comprendidas por: Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, del docente de modalidad abierta y a distancia de la UTPL, tomando en cuenta las acciones más realizadas por parte de los estudiantes, dentro del entorno virtual de aprendizaje en los periodos de oct-2014/feb-2015 y abr-2015/ago-2015, y de esta manera, sugerir al docente la revisión de dichas acciones, en cuanto a la metodología o recursos que comparten, para aplicarlos en aquellas que no tienen mayor actividad o a su vez tomar las acciones como un medio, para llegar de mejor manera al conocimiento que adquieren sus estudiantes.

El proceso para conseguir el resultado de las acciones se lo realizo con la metodología CRISP-DM la cual según (Moine, Haedo, & Gordillo, 2011) es la más utilizada y hasta la actualidad no ha tenido reemplazo. Dentro de esta metodología hacemos uso de técnicas descriptivas de minería de datos, tales como: kmeans y análisis de componentes principales, ejecutadas en de la herramienta RStudio que en base a la encuesta de KDnuggets Software anual 16ª, realizada por (Piatetsky, 2015) es la más utilizada por expertos de minería de datos.

La estructura del desarrollo de este trabajo consta de 3 capítulos descritos a continuación:

En el primer capítulo se abordarán los temas de investigación con respecto a definiciones y uso de herramientas minería de datos, trabajos con datos masivos y problemas presentados con este tipo de datos.

El segundo capítulo se define el problema y el planteamiento de la solución, que se empleará en el proyecto.

En el tercer capítulo se realizará todo el proceso de minería de datos conlleva el proyecto, junto a esto se determinará una discusión sobre los resultados, conclusiones y recomendaciones en base a los análisis realizados, con las técnicas seleccionadas de minería de datos, sobre los datos seleccionados.

CAPÍTULO I: MARCO TEORICO

1.1 Minería de datos (data mining).

1.1.1 Definiciones.

Para el análisis de grandes datos se podría aplicar tabulaciones en los resultados de encuestas y operaciones estadísticas, obteniendo data para concluir o determinar patrones, utilizando una muestra si la cantidad de datos era muy grande; pero en la actualidad contamos con muchas herramientas de análisis estadístico que facilitan este proceso, de ahí que data mining nos permite trabajar con herramientas que enfocan el conocimiento más relevante de datos en gran tamaño, tomando en cuenta que ara nuestro proyecto la cantidad aproximada de estudiantes a ser analizados varía entre 7000 a 11000 alumnos, la minería de datos a través de los pasos y herramientas que cuenta para trabajar con masividad de datos nos servirá para desarrollar nuestro análisis y conclusiones.

Para respaldar lo dicho anteriormente citamos algunos conceptos:

- (Microsoft, 2016) señala que la minería de datos o data mining es el proceso de detectar la información en acción de grandes conjuntos de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en la data. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiada data.

- (Sinnexus, 2016) comparte el concepto con Microsoft sobre la minería de datos, resumiendo que es el proceso de trabajar con gran cantidad de datos para adquirir conocimiento, pero además menciona que data mining inicia para ayudar a comprender el contenido de un repositorio de datos, mediante el uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales; de forma general, los datos son la materia prima, pasando a ser información útil en el momento que el usuario les atribuye algún significado especial.

Davenport y Prusak (1999 citados en Sinnexus, 2016) permiten definir los términos de datos, información y conocimiento de la siguiente manera:

- Datos es el conjunto discreto de valores que no dicen nada sobre el porqué de las cosas y no son orientativos para la acción.

- Información es el proceso que se da a este conjunto de datos proporcionando un significado siendo de utilidad para una toma de decisiones

- Conocimiento se deriva de la información a través de la comparación con otros elementos, predicción de consecuencias, búsqueda de conexiones y la conversación con otros portadores de conocimiento, pero demanda de la experiencia de quien esté realizando el análisis de los datos ya hechos información.

(Sinnexus, 2016) además describe las cuatro etapas a seguir en una minería de datos, independientemente de la técnica de extracción elegida:

1. **Determinación de los objetivos:** Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.
2. **Pre procesamiento de los datos:** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.
3. **Determinación del modelo:** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
4. **Análisis de los resultados:** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

1.1.2 Criterios para aplicar la minería de datos.

(Febles Rodríguez & González Pérez, 2001) consideran que para aplicar los métodos de la minería de datos en la bioinformática (fusión del estudio sobre la genética molecular con la informática), existen dos criterios importantes que también se los puede considerar de forma general para un análisis y una toma de decisiones en otro campo. Estos criterios con la factibilidad económica y la factibilidad técnica los mismos que se definen de la siguiente manera:

- Factibilidad económica - organizativa: existe potencialmente un impacto significativo, no se conocen métodos alternativos, se dispone de personal calificado, no existen problemas de legalidad o violación de la información.

- Factibilidad técnica: se dispone de suficientes datos, los cuales contienen rasgos relevantes, existe poco ruido en los datos y se domina la aplicación de los métodos.

1.1.3 Técnicas data mining.

(Palomo Miñambres Oscar, 2011) describe a las técnicas de minería de datos como algoritmos más o menos sofisticados, aplicados sobre un conjunto de datos para obtener resultados, tomando en cuenta que la utilización de estos algoritmos o técnicas se eligen según la solución que requiera el problema a ser resuelto.

(López, 2007) al mencionar que independientemente de las técnicas existentes, lo que se pretende es descubrir el conocimiento embebido en los datos, muestra estar de acuerdo con el concepto de (Palomo Miñambres Oscar, 2011) y mediante la Ilustración 1 nos permite visualizar los dos grupos de clasificación de las técnicas de minería de datos, predictiva y descriptiva cada una con sub-clasificación y su respectiva descripción dada por (López, 2007) al final de la imagen, pero también presenta las técnicas auxiliares, que son más bien las herramientas de refuerzo aplicadas en la verificación del trabajo de la minería, y de las que se hablaran más adelante en la sección 1.1.4**.**

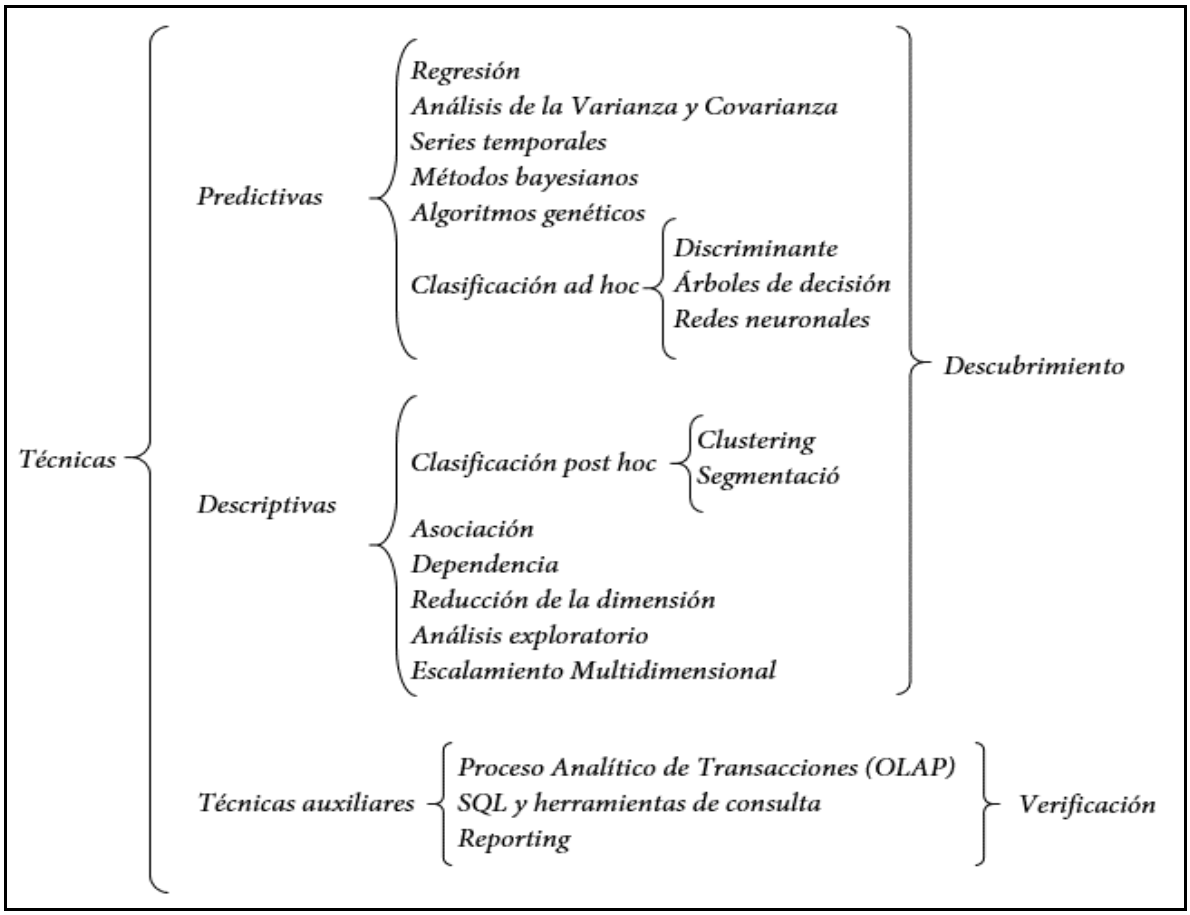


Ilustración 1. Técnicas de minería de datos
Fuente. (López, 2007)

Las técnicas **predicativas** especifican el modelo de los datos en base a un previo conocimiento teórico, y comprenden tipos de **regresión**, **análisis de varianza y covarianza, series temporales, métodos bayesianos, algoritmos genéricos, análisis discriminante, árboles de decisión y redes neuronales**, tanto los arboles de decisión como las redes neuronales y análisis discriminante pueden ser técnicas de clasificación que permiten extraer perfiles de comportamiento o clases, enfocados a construir un modelo para clasificar cada nuevo dato.

En las técnicas **descriptivas** el punto de partida nace de modelos creados automáticamente, mediante el reconocimiento de patrones. En este grupo se incluyen técnicas de **clustering** y **segmentación** que en cierto modo son también técnicas de clasificación, las técnicas de **asociación** y **dependencia**, las técnicas de **análisis exploratorio de datos** y las técnicas de **reducción de dimensión** (factorial, componentes principales, correspondencia, entre otros) y el **escalamiento multidimensional** ya vistas en la fase de transformación.

En los siguientes párrafos se describirá a cada método según la técnica predictiva o descriptiva a la que pertenezcan.

Dentro de las **Predictivas** tenemos:

- **Regresión:** El análisis de regresión es una metodología estadística que es utilizada para la predicción numérica. Existe la regresión lineal que se basa en encontrar la "mejor" línea para encajar dos atributos (o variables), de modo que un atributo se puede utilizar para predecir el otra, y tenemos también la regresión lineal múltiple que es una extensión de la regresión lineal, donde más de dos atributos son
  Implicados y los datos se ajustan a una superficie multidimensional. (Closas, Arriola, Kuc Zening, Amarilla, & Jovanovich, 2013)
- **Análisis de la Varianza y covarianza:** El grado en que los datos numéricos tienden a extenderse se llama la dispersión, o *varianza* de los datos. La varianza es una de las medidas más comunes de la dispersión, así como rango.
  Además, trabaja con variables dependientes métricas, buscando la existencia de diferencias significativas entre los grupos creados a partir de una división de la muestra total. (Closas et al., 2013), concepto respaldado por (Arriaza, 2006) cuando lo resume en que el objetivo del análisis de la varianza es la visualización del efecto que dan dos factores sobre una variable métrica.
  La *covarianza* permite analizar el impacto de la variable nominal u ordinal y otra métrica denominada covariante sobre una variable métrica, combinando análisis de

varianza con análisis de regresión y de esta manera poder separar la variabilidad de la variable dependiente relacionada con la covariable reforzando las diferencias entre grupos (Arriaza, 2006)

- **Análisis canónico:** (Härdle & Simar, 2015) aplica esta técnica en un proyecto de marketing teniendo en claro que considera que el análisis de correlación canónica es una herramienta estándar de análisis estadístico multivariado aplicada en el descubrimiento y la cuantificación de las asociaciones entre dos conjuntos de variables.
- **Redes neuronales:** (Han, Kamber, & Pei, 2011) mencionan que las redes neuronales tienen un tiempo de larga formación, requieren de parámetros empíricos como lo es la topología de red o estructura, y su interpretación hace que dentro de la minería de datos hayan sido en su inicio las más aptas para el análisis de un proyecto.

Dentro de una clasificación de red ad hoc de técnicas predictivas están:

- **Discriminante:** (Han, Kamber, & Pei, 2011) describe el análisis discriminante o discriminación de datos como la comparación de las características generales de los objetos de datos de la clase objetivo, con las características generales de los objetos de una o un conjunto de clases diferentes, pudiendo ser tanto el objetivo y las clases diferentes especificados por el usuario, y los correspondientes objetos de datos recuperado a través de consultas de bases de datos. Por ejemplo, el usuario puede tener gusto de comparar según características generales de los productos de software cuyas ventas aumentaron en un 10% en el último año las cuales se reducen en al menos un 30% durante el mismo período. Los métodos utilizados para los datos discriminación son similares a los utilizados para la caracterización de datos.

  (Arriaza, 2006) diferencia esta técnica multivariante del análisis de regresión en la forma que se encuentre categorizada la variable dependiente, es decir, si es de tipo nominal puede ser Consumidor A, Consumidor B, Consumidor C, esto significa que es una variable con valor informativo; en el caso de ser la variable de tipo ordinal contendrá valor numérico que puede ser ubicado en un nivel alto, medio o bajo de un nivel de ingresos y por ultimo si la variable dependiente cuenta tan solo con dos categorías como lo presenta la variable género (Masculino y Femenino), indica la utilización de un modelo logístico como alternativa al análisis discriminante para tratar el problema.
- **Arboles de decisión:** (Han, Kamber, & Pei, 2011) indica que los arboles de decisión pueden ser convertidos en reglas de clasificación ya que presentan una estructura de árbol similar al diagrama de flujo, en donde cada nodo indica una prueba en un valor

de atributo, cada rama representa un resultado de la de prueba, y las hojas representan las clases o distribuciones de clase, así como el uso de una red neuronal, cuando se utiliza para la clasificación, pues es una colección de unidades de procesamiento similares a neuronas con conexiones ponderadas entre el unidades. Hay muchos otros métodos para la construcción de modelos de clasificación, tales como clasificación bayesiana, máquinas de vectores soporte, y k-nearest-neighbor que es un método de clasificación supervisada.

Y en las técnicas **Descriptivas** se detallan de la siguiente manera:

En la clasificación de post ad hoc:

- **Clustering:** una técnica no supervisada. Organización de valores similares denominados como clústeres.
  (Han, Kamber, & Pei, 2011) lo describe como una tarea en donde por la agrupación permite detectar valores ubicados fuera del conjunto de conglomerados denominados como valores atípicos.
  Teniendo en cuenta que, las etiquetas de clase no están presentes en los datos de entrenamiento, simplemente porque no se conoce, y Clustering se puede utilizar para generar tales etiquetas. Los objetos son agrupados en base a maximizar la similitud dentro de las clases y minimizar la similitud fuera de las clases. Es decir, las agrupaciones de objetos están formadas de manera que los objetos dentro de un grupo tienen una alta similitud en comparación con los otros, pero son muy diferentes a los objetos en otros grupos. La agrupación también puede facilitar la formación de la taxonomía que organiza las observaciones en una jerarquía de clases que agrupan similares eventos.
- **Segmentación:** es otro tipo de agrupación, pero se la utiliza para dividir un grande grupo de datos. (Han, Kamber, & Pei, 2011) menciona que la segmentación puede ser utilizado para la detección de los valores atípicos que están muy lejos de cualquier grupo, siendo así más interesante que los casos comunes. Las aplicaciones de detección de valores atípicos incluyen la detección de fraude de tarjetas de crédito y el control de las actividades delictivas en el comercio electrónico. Por ejemplo, los casos excepcionales en las transacciones con tarjeta de crédito, que son muy frecuentes, pueden correr el riesgo de una actividad fraudulenta.
  El análisis de conglomerados se puede utilizar como una herramienta independiente para comprender mejor la distribución de los datos, para observar las características de cada grupo, y centrarse en un conjunto particular de grupos para su posterior análisis. Alternativamente, puede servir como un pre procesamiento paso para otros

algoritmos, como la caracterización, selección de subconjuntos de atributos, y clasificación, que entonces operar sobre los grupos detectados y los atributos seleccionados o características.

Saliendo de la sub clasificación de post hoc de las técnicas descriptivas continuamos con el resto.

- **Asociación:** (Han, Kamber, & Pei Jian, 2011) señala que las reglas de asociación se descartan si no se satisfacen tanto un acceso de mínimo apoyo como el acceso de mínima confianza. Aplicar este análisis puede llevar a descubrir correlaciones estadísticas interesantes entre los asociados, por ejemplo, usando representaciones de datos en aplicaciones informáticas como el llamado attribute–value pair.
  (Perez, 2014) concuerdan con (Han, Kamber, & Pei, 2011) al decir que las reglas de asociación expresan patrones de comportamiento entre los datos en función de su aparición conjunta, y teniendo resultados de combinaciones de valores de los atributos que ocurren más veces, pudiendo ser aplicadas en un análisis de patrones en compra de supermercados, para adquirir una mejora en la distribución de sus productos.
  Existen algunos algoritmos utilizados para determinar reglas de asociación entre ellos Apriori, Partition, Eclat.
- **Dependencia:** (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) en el artículo del descubrimiento del conocimiento indican que el modelado de dependencia, permite encontrar un modelo con las dependencias más importantes entre variables. Además, la existencia de dos niveles de dependencia como son el nivel estructural y el nivel cuantitativo, el primero especifica las variables que son dependientes de forma local y el segundo especifica los puntos fuertes de las dependencias utilizando una escala numérica.
- **Reducción de la dimensión:** en la reducción de dimensión (Han & Kamber, 2006), menciona que la codificación de datos o transformaciones se aplican a fin de obtener un comprimido de la representación de los datos originales. Si los datos originales pueden ser reconstruidos a partir de los datos comprimidos sin ninguna pérdida de información, es considerada como ganancia. Pero si solo podemos reconstruir una aproximación de los datos originales, la reducción de datos se considera una pérdida. Existen varios algoritmos bien definidos por la compresión de datos. En otras palabras, la reducción de la dimensión solo trabaja con limitados datos del volumen o cantidad inicial a ser analizado.

- **Análisis exploratorio de datos (AED):** (Figueras & Gargallo Valero, 2003) nos describe este análisis como el conjunto de técnicas estadísticas que buscan obtener una comprensión básica de la data y las relaciones entre las variables analizadas.

  Además, el AED proporciona métodos sistemáticos sencillos para recoger, organizar y preparar los datos; detectando fallos en el diseño, planteando tratamiento, brindando una posible Evaluación de datos ausentes (missing), identificando casos atípicos (outliers) y comparando los supuestos subyacentes en la mayor parte de las técnicas multivariante (normalidad, linealidad, homocedasticidad).

  Menciona además que es un error descuidar el examen previo de los datos, por el tiempo este lleva; pues las tareas implícitas en este examen son una parte esencial en cualquier análisis estadístico.
- **Escalamiento multidimensional (MDS):** se originó en el campo psicológico, pues como aduce (Arce, de Francisco, & Arce, 2010), permite tratar el problema de distancias que contienen error o distancias distorsionadas por los mecanismos perceptivos de los seres humanos.

  (Guerreo & Ramirez, 2012) describen la técnica del escalamiento multidimensional, como una técnica de representación espacial, que permite visualizar sobre un mapa un conjunto de estímulos pudiendo ser firmas, productos, candidatos políticos, ideas u otros artículos; para analizar la posición relativa, es decir, algún espacio no conocido que por lo general se la ubica con puntos cardinales.

  Y el propósito del MDS es transformar los criterios similares de un grupo de individuos que compartan un conjunto de comparación de objetos o estímulos en distancias alterables para representarlas en un espacio multidimensional, de manera que si un individuo juzga a los objetos A y B como los más similares entonces las técnicas de MDS colocarán a los objetos A y B en el gráfico en donde la distancia entre ellos sea más pequeña que la distancia entre cualquier otro par de objetos.

  Por ejemplo, tenemos un grupo de personas con las que aplicaremos la técnica de escalamiento multidimensional para analizar el consumo respecto a 5 vitaminas para niños denominándolas V1, V2, V3, V4, V5. Ubicamos las mismas en una escala de tal forma que sus posiciones nos dan una medida de preferencia en el consumo de una de ellas.

1.1.4 Herramientas de data mining.

Las herramientas son objetos que facilitan el desarrollo de los trabajos o proyectos que se deseen desarrollar en diferentes campos y dentro de la minería de datos existe una gran variedad.

Las herramientas o software de minería de datos contienen técnicas específicas para extraer conocimiento de un conjunto definido de datos.

La Tabla 1 indica herramientas, privadas y de código abierto; detallando la compañía, técnicas, plataforma e interfaz con la que trabaja cada una de ellas, según (López, 2007; B. Vargas, 2014).

Tabla 1. Herramientas de data mining

PRODUCTO	COMPAÑÍA	TECNICAS	PLATAFORMA	INTERFAZ
KNOWLEDGE SEEKER	Angoss	Árboles de decisión	Win / UNIX	ODBC
CART	Salford Systems	Árboles de decisión	Win / UNIX	ODBC
CLEMENTINE DE SPSS IBM	SPSS	Amplio abanico	UNIX	ODBC
DATA SURVEYOR	Data Distilleries	Amplio abanico	UNIX	ODBC
GAEN SMARTS	Urban Science	Gráficos-Ganancias	Win / UNIX	
INTELLIGENT MINER	IBM	Amplio Abanico	UNIX (AIX)	IBM, DB2
MICOSTRATEGY	Micostrategy	Datawarehouse	Win	Oracle
POLYANALYST	Megaputer	Simbolicas	Win	Oracle, ODBC
DARWIN	Oracle	Amplio abanico	Win / UNIX	Oracle
ENTERPRISE MINER	SAS Institute	Amplio Abanico	Win / UNIX / Mac	
SGI MINESET	Silicon Graphies	Asociación y Clasificación	UNIX	Oracle, Sybase, Informix
WIZSOFT / WIZWHY	Wizdoft	Algoritmo propio de corrección	Win	ODBC, OLE DB
MINER DE SAS	Salford Systems	Árboles de decisión	Win	ORACLE, IBM Netezza, SAP HANA
STATISTICA	StatSoft	Análisis de Agregación, multivariantes y	Win, OLE, DDE	OLE,

		modelos avanzados de regresión lineal y no lineal		
ORACLE DATA MINING	Oracle Data Mining	Análisis predictivo	Win, Solaris, SUSE 9	SQL, PL / SQL Y JAVA
MATLAB	Cleve Moler-Matrix Laboratory	Análisis exploratorio, clasificación, arboles de regresión, redes neuronales artificiales	Win, UNIX, GNU/Linux	OpenGL
CODIGO LIBRE				
R	Laboratorios Bell, Universidad Auklan	Regresión, Asociación	GNU GLP, Win, Mac, UNIX	Tcltk, LaTex, (X)Emacs, Python
RAPIDMINER ANTES YALE	SourceForge	Aprendizaje automático, redes neuronales, arboles de decisión	Multiplataforma	Acces, Oracle, IBM DB2, MySQL, Postgres
ORANGE	Universidad de Ljubljana	Exploración, validación cruzada, arboles de regresión básica	Qt multiplataforma, Mac, Linux, Win	C++, Python
WEKA	The University of Waikato	Pre procesamiento de datos, clustering, clasificación, regresión, visualización y selección	Multiplataforma	ASCII, JDBC
KNIME	KNIME.com	Cluster, Clasificación	Eclipse RCP, LINUX, Win, Mac	BioSolveIT, GPLv3

Fuente. Elaboración propia a partir de (López, 2007; B. Vargas, 2014)

Después de conocer la variedad de herramientas con que cuenta data mining, la encuesta de KDnuggets Software anual 16ª, realizada por (Piatetsky, 2015) acerca de la usabilidad que tiene los software de minería de datos, basándose en la participación de 2800 votantes, ante 93 herramientas diferentes, resultando un listado de 10 más destacadas, de donde R resalta como la herramienta más utilizada por los mineros y científicos de datos quienes participaron como votantes.

A continuación la Ilustración 2 con el porcentaje de usabilidad de las 10 herramientas más destacadas.

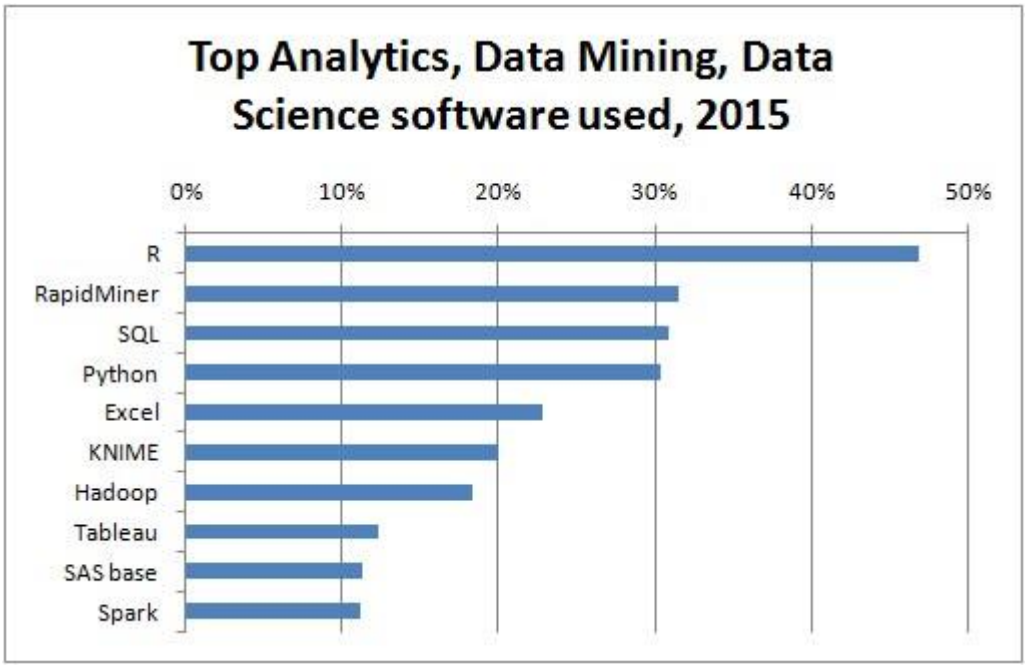


Ilustración 2. Herramientas con más usabilidad
Fuente. (Piatetsky, 2015)

En base a al resultado de usabilidad describiremos las primeras cuatro herramientas más utilizadas en la minería de datos, tales como: R, RapidMiner, SQL y Python.

R[1] cuenta con la una gran calidad de diseño para la presentación de diagramas, incluyendo símbolos y formas cuando sea necesario.

Para las tareas computacionalmente intensivas, C, C ++ y Fortran se vinculan y llaman en tiempo de ejecución. Los usuarios avanzados pueden escribir código C para manipular objetos R directamente.

(Zhao, 2015) describe a R como un entorno de software libre para el cálculo estadístico y gráficos, con importación y exportación de datos, que contiene 6,600+ paquetes disponibles como: CRAN, Bioconductor, R-Forge, GitHub, entre otros, y además menciona que es ampliamente utilizado tanto en academias como en la industria y permite realizar tareas como:

- Aprendizaje supervisado y el aprendizaje no supervisado.
- Análisis de cluster (conglomerados) y mezcla de modelos.
- Análisis de series temporales

[1] https://www.r-project.org/about.html

- Estadística multivalente
- Análisis de datos espaciales

Actualmente cuenta con editor multiplataforma muy amigable con el usuario denominada RStudio.

RapidMiner[2] herramienta de minería de datos, anteriormente conocida como YALE, es utilizada a nivel internacional en aplicaciones empresariales, de gobierno y académicas.

Algunas de las características de esta herramienta son:

- Utiliza un modelo de cliente/servidor
- Escrito en lenguaje de programación Java
- Contiene más de 500 de técnicas para el pre-procesamiento de datos
- Proporciona una interfaz gráfica para el diseño y ejecución de flujos de trabajo.

Microsoft SQL Server[3], cuenta con características y herramientas para aplicar minería de datos, pues la combinación de Integration Services, Reporting Services y SQL Server Data Mining ofrece una plataforma integrada para el análisis predictivo que abarca la limpieza y preparación de datos, aprendizaje automático y generación de informes.

SQL Server Data Mining incluye varios algoritmos estándar, incluyendo la EM y K-means clustering modelos, redes neuronales, regresión logística y regresión lineal, árboles de decisión y los clasificadores de Bayes. Todos los modelos han integrado visualizaciones para ayudarle a desarrollar, refinar y evaluar sus modelos. La integración de la minería de datos en una solución de inteligencia de negocios le ayuda a tomar decisiones inteligentes acerca de problemas complejos.

1.1.5 Áreas de aplicación de data mining.

El conocimiento que se puede adquirir a través de un determinado grupo de datos, dentro de herramientas que nos ayudan a trabajar con las distintas técnicas de la minería de datos, puede ser posible en diversas áreas, por ejemplo:

[2] http://www.postecnologia.com/2015/03/rapid-miner-software-business-intelligence.html
[3] https://technet.microsoft.com/en-us/library/bb510517(v=sql.105).aspx

- **En educación**

Rodas (2001 citado en (Molina, 2002)) menciona el estudio sobre la *inserción profesional de los recién titulados de la carrera de Ingeniería en Sistemas computacionales del Instituto Tecnológico de Chihuahua II*; cuyo objetivo era comprobar si su plan de estudios y el aprovechamiento del alumno produjeron una buena inserción laboral u optaron por otras variables.

Considerando características como: género, edad, escuela de procedencia, desempeño académico, zona económica de su vivienda y actividad profesional.

Aplicando conjuntos aproximados descubrió cuatro variables: zona económica, procedencia del colegio, nota de ingreso y promedio final de la carrera, con cuyo resultado determino que la universidad debe hacer un estudio socioeconómico sobre grupos de alumnos que pertenecían a las clases económicas bajas para dar posibles soluciones, pues tres de las cuatro variables no dependían de la universidad.

(Formia, 2012) en su trabajo sobre la *evaluación de conocimiento en base de datos y su aplicación a la deserción de los alumnos universitarios,* hace énfasis en las técnicas no supervisadas, ya que sus datos solo presentan características personales y académicas de los estudiantes. Con el uso de la herramienta RapidMiner aplica K-means, determinando 5 grupos, que no describen de forma clara los atributos, por lo que aplica algoritmos de selección para establecer los atributos y como respaldo a este resultado, aplica arboles de decisión en la extensión de weka presente en la RapinMiner, obteniendo finalmente, los atributos deseados, asignándoles valores de porcentaje y determinando que grupos tienen más incidencia de abandono y grupo posee menos incidencia de abandono.

(Azoumana, 2013) presenta un artículo con resultados del análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad de Ingeniería de Sistemas, con técnicas de minería de Datos, Utilizando la herramienta Weka, agrupando una muestra de 707 estudiantes, que presentaban casos de deserción clasificadas en 5 variables denominadas: pérdida de semestre, dificultad financiera, ingreso al mercado laboral, otros intereses atraen al estudiante e indeterminado, que comprendieron los períodos 2007-2012, determinando que la deserción es el factor indeterminado.

(Malbernat et al., 2015) presentan un trabajo sobre la *aplicación de minería de datos en la gestión de docentes de educación superior*, utilizando datos con características cuantitativas sobre la preparación y actitud de cada profesor, en donde se utilizaron técnicas como K-means y EM (esperanza-maximización) dentro de herramientas como:

weka, PSPP, RapidMiner de donde con k-means de K=4, en weka y 13 iteraciones se obtuvo un 68% de indiferentes, que absorben un 20% de refractarios y casi un 40% de innovadores. Por tal razón divide al grupo de indiferentes, en reticentes por estar cerca de los refractarios y en flemáticos por tener una actitud y preparación a la media. Al aplicar k-means con k=4 en la herramienta PSPP, obtuvo el mismo resultado.

Realizaron otro análisis con k=5 pero esta vez en weka como en RapidMiner, obteniendo en weka un 24% de refractarios y 14% de innovadores con un nuevo grupo de indiferentes denominada desorientados con alto nivel de actitud y bajo en preparación; pero en RapidMiner obtuvo un resultado más equilibrado con 23% de desorientados, 51% de reticentes, 8% de flemáticos y 5% de innovadores. Concluyendo que un 50% docentes pertenecen al grupo de reticentes y una probabilidad de 5% en el grupo de innovadores.

Y en el uso del algoritmo EM con weka para 3 grupos, utilizando una distribución normal, el resultado fue 22,5% de sujetos refractarios, 55% de Indiferentes y 22,5% de Innovadores y para 4 grupos con EM, obtuvo 17% para los refractarios y 16% los reticentes. Dejando advertido que con la implementación se adquiriría una división más homogénea.

- **En el Gobierno**

(Camana, 2014) presenta una investigación, realizada en la universidad tecnológica indométrica, utilizando técnicas de minería de datos para encontrar patrones de comportamiento o datos que incidan en una elección presidencial, teniendo como objetivo seleccionar las juntas receptoras del voto más representativas, durante elecciones presidenciales del 2009.

Teniendo mejor resultado con la técnica de agrupamiento llamada K-means, dentro de la herramienta de minería weka, se dedujo que, entre mayor sea el grupo seleccionado, menor será el número de muestra en cada uno, entonces aplicaron 20 agrupaciones, de donde en la numero 13 se ubicaron las juntas receptoras del voto más útiles para realizar cualquier tipo de sondeo.

(Molina, 2002) indica el anuncio realizado por John Aschcroft director del FBI sobre el análisis de las bases de datos comerciales para detectar terroristas, ingresando a las bases de datos comerciales inclinados a los hábitos y preferencias de compra de los consumidores, para identificar intentos terroristas antes de que sean ejecutados.

Expertos testifican que al unir todas las bases de datos probablemente mediante el número de la Seguridad Social, se obtendrá información como: si una persona fuma, qué talla y tipo de ropa usa, su registro de arrestos, su salario, las revistas a las que está suscrito, su altura y peso, sus contribuciones a la Iglesia, grupos políticos u organizaciones no gubernamentales, sus enfermedades crónicas (como diabetes o asma), los libros que lee, los productos de supermercado que compra, si tomó clases de vuelo o si tiene cuentas de banco abiertas, entre otros, la misma que permitirá llegar a tan importante localización de atentados.

Para este grande proyecto especulan una inversión inicial de setenta millones de dólares estadounidenses para consolidar los almacenes de datos, desarrollar redes de seguridad para compartir información e implementar nuevo software analítico y de visualización.

- **En investigaciones espaciales**

(Molina, 2002) menciona el *Proyecto SKYCAT*, en donde el Second Palomar Observatory Sky Survey (POSS-II) durante seis años, coleccionó tres terabytes de imágenes que contenían aproximadamente dos millones de objetos en el cielo. Tres mil fotografías fueron digitalizadas a una resolución de 16 bits por píxel con 23.040 x 23.040 píxeles por imagen. El objetivo era formar un catálogo de todos esos objetos. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) se basa en técnicas de agrupación (clustering) y árboles de decisión para poder clasificar los objetos en estrellas, planetas, sistemas, galaxias, etc. con una alta confiabilidad (Fayyad et al., 1996). Los resultados han ayudado a los astrónomos a descubrir dieciséis nuevos cuásares con corrimiento hacia el rojo, considerados entre los objetos más lejanos del universo y, por consiguiente, más antiguos. Estos cuásares son difíciles de encontrar y permiten saber más acerca de los orígenes del universo.

- **En clubes deportivos**

(Molina, 2002) señala que la Asociación de Futbol Milán, utiliza redes neuronales para prevenir lesiones y optimizar el acondicionamiento de cada atleta, ayudando de esta manera a seleccionar el fichaje de un posible jugador o a alertar al médico del equipo de una posible contusión. El sistema, creado por Computer Associates International, es sustentado por datos de cada jugador, en función a su rendimiento, alimentación, respuesta a estímulos externos, obtenidos y analizados cada 15 días, y además de actividades monitoreadas por veinticuatro sensores conectados al cuerpo, que transmiten señales de radio.

Actualmente el sistema dispone de 5.000 casos registrados que permiten predecir alguna posible lesión y con ello el club intenta ahorrar dinero evitando comprar jugadores que presenten una alta probabilidad de lesión, lo que haría incluso renegociar su contrato.

El sistema pretende encontrar las diferencias entre las lesiones de atletas de ambos sexos, así como saber si una determinada lesión se relaciona con el estilo de juego de un país concreto donde se practica el fútbol.

Otro caso que presenta Molina es el de los *equipos de la NBA*, los cuales utilizan el Advanced Scout, un software desarrollado por IBM, que aplica técnicas de minería de datos para detectar patrones estadísticos y eventos raros. Tiene una interfaz gráfica muy amigable orientado al análisis del juego de los equipos de la National Basketball Association (NBA). Con este conocimiento, los entrenadores crearon estrategias alternativas para tratar con el doble marcaje.

- **En meteorología**

(Camana, 2016) hace referencia de la aplicación de minería de datos para establecer patrones de comportamiento, con información recopilada desde el año 1995 hasta el 2005, usando los datos meteorológicos que posee el Observatorio Astronómico de Quito(OAQ) e implementando un almacén de datos, en la herramienta Clementine de minería de datos, obtuvo patrones de comportamiento meteorológico.

- **En finanzas**

(LOGICALIS, 2014) comenta como se aplica la minería de datos tanto en sector bancario como financiero, buscando proveer datos que aseguren un análisis sistemático en condiciones avanzadas y garanticen fiabilidad a través de:

- Diseño y construcción multidimensional de datos.
- Predicción de pagos sobre préstamos
- Análisis de políticas de crédito del cliente.
- Clasificación de los clientes para crear ofertas de manera personalizada.
- Detención de delitos financieros como el lavado de activos.

- **En la industria de telecomunicaciones**

(LOGICALIS, 2014) muestra que la minería de datos, dan un gran aporte en la identificación de patrones de telecomunicaciones, teniendo así las siguientes ventajas:

- Detección de actividades y patrones fraudulentos, como: hábitos y tendencias inusuales.
- Mejorar la utilización de los recursos.
- Mejorar la calidad del servicio.
- Analizar datos de telecomunicaciones de manera multidimensional.
- Asociar y analizar de manera multidimensional patrones secuenciales.

1.2 Metodologías para el desarrollo de proyectos de minería de datos.

Al aplicar la minería de datos en diferentes proyectos tenemos pocas opciones en cuanto a metodologías se trata, las mismas que según experiencias de expertos son muy buenas para obtener excelente resultado en el desarrollo del mismo.

Se debe tomar en cuenta que al elegir una de ellas los requisitos que presente la posible solución del problema del proyecto a ser ejecutado, son la guía para la elección de la metodología a trabajar.

SEMMA, CRISP-DM Y KDD son algunas de las metodologías aplicadas para la minería de datos, y a continuación las citaremos para hablar de cada una de ellas.

(Montequín et al., 2002) en su trabajo de *investigación* determina lo siguiente acerca de SEMMA Y CRISP-DM.

1.2.1 Metodología SEMMA.

SAS Institute desarrollador de esta metodología, la define como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones de negocio desconocidos.

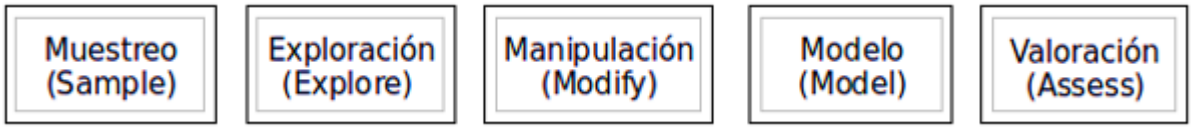


Ilustración 3. Fases de la metodología de SEMMA
Fuente. (Montequín et al., 2002)

El nombre de esta terminología es el acrónimo correspondiente a las cinco fases básicas del proceso, tal como se muestra en la Ilustración 3.

El proceso se inicia con la **extracción de la muestra** de la población sobre la que se va a aplicar el análisis, teniendo como objetivo tratar el problema en estudio, en base a la muestra tomada. Teniendo en cuenta que la muestra debe ser muy representativa, para la correcta validación y análisis de resultado del modelo a ser aplicado.

La forma más común de obtener una muestra es la selección al azar, es decir, cada uno de los individuos de una población tiene la misma posibilidad de ser elegido. Este método se denomina muestreo aleatorio simple, teniendo en cuenta que para el análisis del proceso se debe considerar la asociación de acuerdo al nivel de confianza de la muestra.

Una vez determinada una muestra o conjunto de muestras significativas de la población en estudio, seguimos con la segunda fase, en donde se procede a la **exploración** de la información disponible, con el fin de simplificar en lo posible el problema, optimizando la eficiencia del modelo. Para ello se utilizan herramientas de visualización o de técnicas estadísticas que ayuden a identificar relaciones entre variables. De esta forma se pretende determinar cuáles son las variables explicativas o también conocidas como variables independientes que van a servir como entradas al modelo.

La tercera fase de la metodología consiste en la **manipulación de los datos**, en base a la exploración realizada, de forma que se definan y tengan el formato adecuado de los datos que serán introducidos en el modelo.

Una vez definido el formato de los datos, se procede al **modelado de los datos**, con el fin de establecer una relación entre las variables independientes y las variables objeto del estudio, que posibiliten inferir el valor de las mismas con un nivel de confianza determinado. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como técnicas de redes neuronales, técnicas adaptativas, árboles de decisión, reglas de asociación y computación evolutiva, basadas en datos.

Para finalizar tenemos la cuarta fase que es la **valoración de los resultados** mediante el análisis de bondad del modelo o modelos, contrastado con otros métodos estadísticos o con nuevas poblaciones de muestra.

Dicho de forma gráfica tenemos la Ilustración 4.

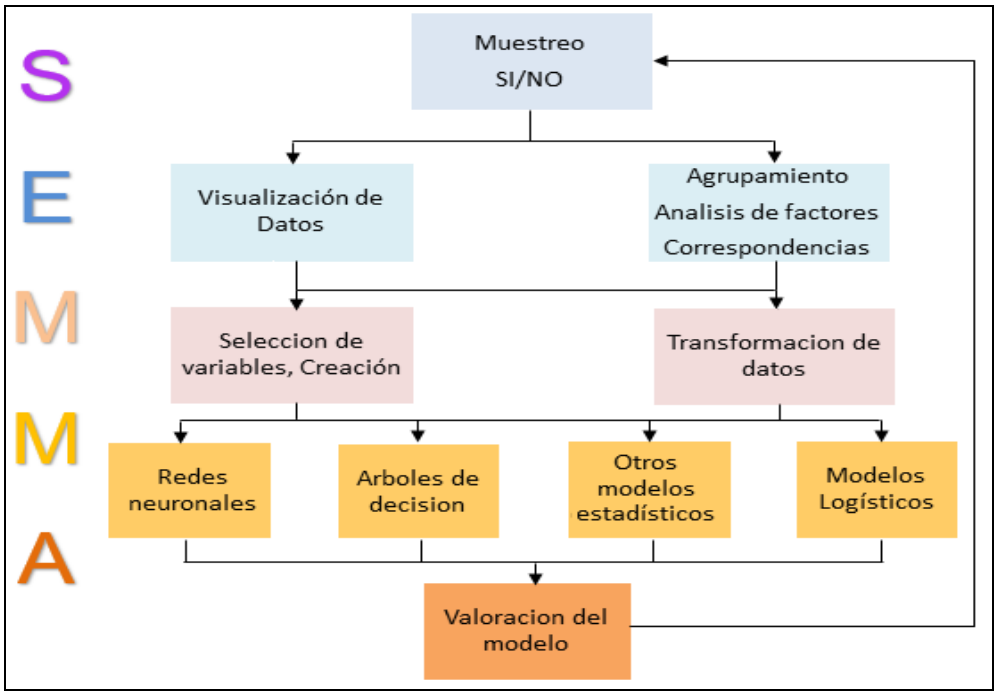


Ilustración 4. Dinámica de la metodología SEMMA
Fuente. (Montequín et al., 2002)

1.2.2 Metodología CRISP-DM.

(Montequín et al., 2002) indica que la metodología CRISP-DM, consta de cuatro niveles de abstracción y seis fases dentro de un ciclo de vida, los niveles de abstracción se están organizados de forma jerárquica en tareas que van desde el nivel más general hasta el más específicos Ilustración 5.

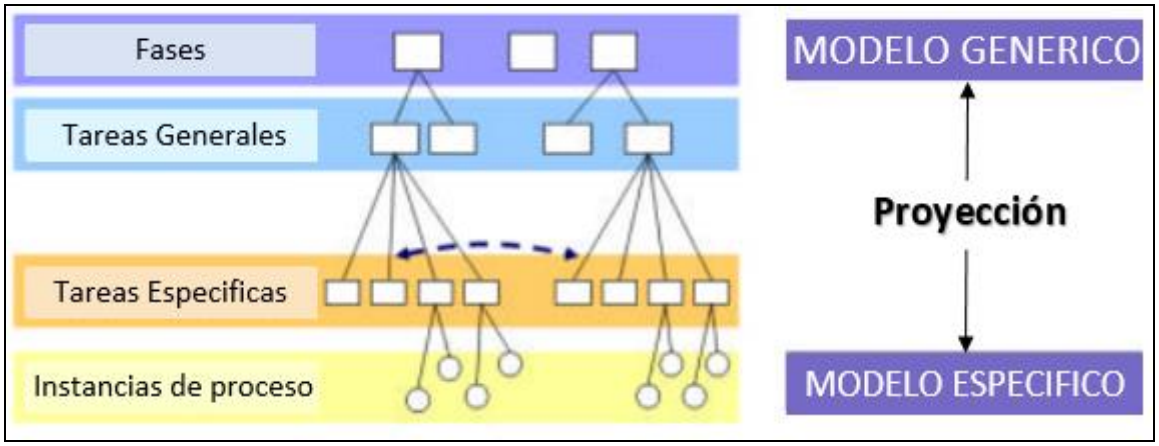


Ilustración 5. Niveles de abstracción de la metodología CRISP-DM
Fuente. (Montequín et al., 2002)

El ciclo de vida de la metodología CRISP-DM presentado en la Ilustración 6, muestra cómo trabajan de manera iterativa sus seis fases en el desarrollo del proyecto.

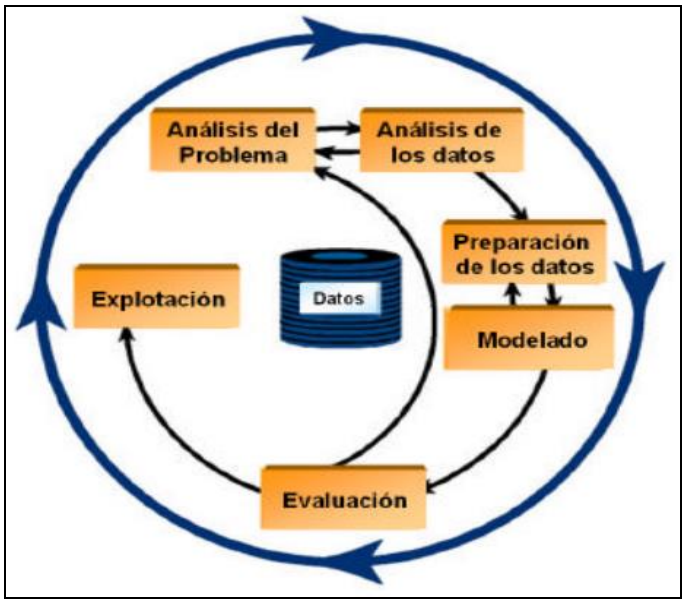


Ilustración 6. Ciclo de vida del proceso de modelado de CRISP-DM
Fuente. (Montequín et al., 2002)

(Chapman et al., 2000) explican que las flechas muestran las relaciones más habituales entre las fases, pero también se pueden establecer relaciones entre cualquiera de ellas sin importar la secuencia, los siguientes párrafos nos detallaran el trabajo de cada una de ellas:

- **Análisis del problema** comprende los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

- **Análisis de datos** entiende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

- **Preparación de datos** incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato, en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados, realizando en esta fase la interacción entre las fases de preparación y modelado de forma sistemática.

- **Modelado** selecciona las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se seleccionan en función de los siguientes criterios:

    - Ser apropiada al problema

- Disponer de datos adecuados
- Cumplir los requerimientos del problema
- Tiempo necesario para obtener un modelo
- Conocimiento de la técnica

Antes de proceder al modelado de los datos se debe establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Los parámetros utilizados en la generación del modelo dependen de las características de los datos.

- **Evaluación** aprecia el modelo, desde el cumplimiento de los criterios de éxito del problema, revisando el seguimiento del proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que, a la vista del desarrollo posterior del proceso, se hayan podido cometer errores. Si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo.

Por lo general los proyectos de Data Mining se deben documentar y presentar los resultados en orden de manera comprensible para conseguir un incremento del conocimiento.

En la fase de explotación se debe de asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

(Olson David & Delen Dursun, 2008) basado en el proyecto Australiano de desarrollo de software aplicado por Nayak y Qui nos muestra la Tabla 2 con la comparación entre las dos metodologías mencionadas anteriormente, señalando en la última columna las actividades aplicadas a las diferentes comparaciones entre la metodología CRISP Y SEMMA.

Tabla 2. Metodologías aplicadas a data mining

ACTIVIDADES DE COMPARACION DE Nayak & Qiu / METODOLOGIAS	CRISP	SEMMA
Se definieron objetivos **Desarrollar herramientas para utilizar mejor los informes de problemas**	Comprensión del negocio	Toma pregunta definida
Análisis de datos en los informes de problemas	comprensión de los datos	Muestra Explorar

Pre-procesamiento de datos **Limpieza de datos** **Transformación de datos**	Preparación de datos	Modificar los datos
Modelado de datos	Modelado	Modelo
Análisis de los resultados	Evaluación	Evaluar

Fuente: (Olson David & Delen Dursun, 2008)

1.2.3 KDD.

(Fayyad et al., 1996) apuntan que KDD es el proceso de extracción de datos, para identificar patrones válidos y potenciales dentro de grandes volúmenes de información, es de forma interactiva e iterativa más no automática.

Una vez identificado el problema y objetivo del proyecto esta metodología consta de 9 etapas, figuradas en la Ilustración 7, tomando en cuenta que no para todos los autores esta metodología se aplica a la minería de datos sino más bien data mining es parte de una de las fases de la KDD como nos recuerda Hernández (2004 citado en (Candás Romero, 2006)).

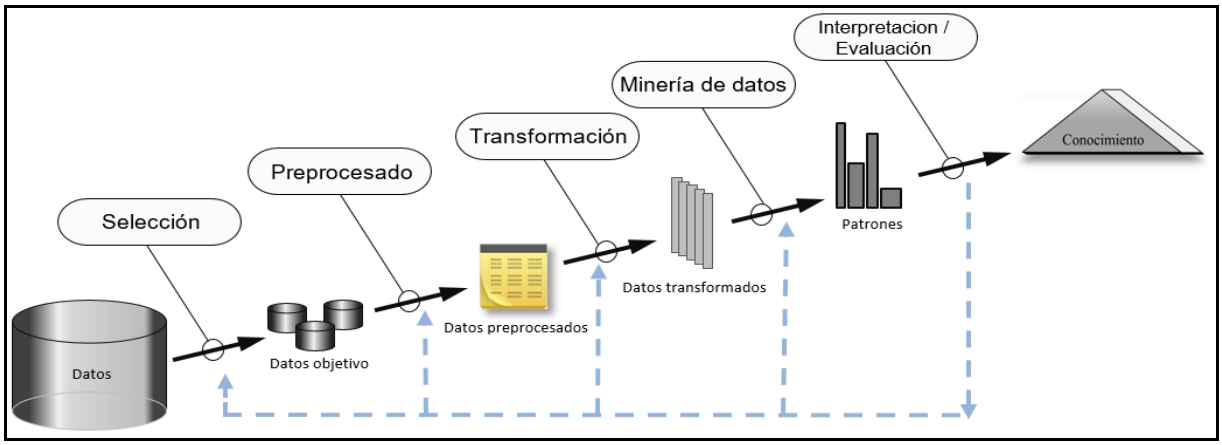


Ilustración 7. Proceso de KDD
Fuente. (Fayyad et al., 1996)

A continuación la descripción del proceso KDD según (Fayyad et al., 1996)

1. Desarrolla una **comprensión** del dominio de la aplicación relevante ante el conocimiento y la identificación de la meta del proceso de KDD desde el punto de vista del cliente.

2. Creación de un conjunto de datos, es decir **seleccionar** un conjunto de datos, o agrupar un subconjunto de variabilidad o muestras de datos.

3. **Limpieza** de datos y **pre-procesamiento**, operaciones básicas que incluyen la eliminación de ruido, recogida de la información necesaria para modelar o reconocer el ruido, decidir sobre las estrategias para el manejo de los campos de datos que faltan, y la contabilidad de tiempo-secuencia de información y los cambios conocidos.

4. **Reducción** de datos y proyección, atreves de la búsqueda de características útiles para representar los datos en función del objetivo de la tarea, con métodos de reducción o transformación.
   obtenemos el número efectivo de variables que bajo consideración del experto puede reducirse, o representarse de acuerdo a varios datos que se encuentren.

5. **Búsqueda** de patrones de interés particular en forma de representación, en función del objetivo de minería de datos. Por ejemplo, el resumen, clasificación, regresión, clustering, entre otros.

6. **Análisis** y **modelo exploratorio** y la **selección de hipótesis** no es más que la elección de los algoritmos de la minería de datos y la selección del método para ser utilizado en la búsqueda de patrones de datos, es decir que en esta etapa se definen el o los modelos de parámetros que van a ser apropiados para la toma de decisiones, por ejemplo: modelos de datos categóricos, los mismos que se diferencian entre los modelos de los vectores sobre los reales; así como también se realiza la búsqueda de un método de extracción de datos particulares, con los criterios generales del proceso de KDD por ejemplo, el interés del usuario final podría ser más en la comprensión del modelo de sus capacidades predictivas.
   Minería de datos: en esta etapa se realiza la búsqueda de patrones de interés en una representación en particular, formulario o un conjunto de tales representaciones, incluyendo reglas de clasificación, regresión o árboles, y la

agrupación. El usuario puede significativamente ayudar al método de minería de datos correctamente realizar los pasos precedentes.

7. **Interpretación** de los patrones extraídos, probablemente esta etapa implique volver a cualquiera de las etapas anteriores y con ello repetir el proceso, con otros datos, algoritmos o metas. Este paso también puede implicar visualización de los patrones extraídos y modelos o la visualización de los datos dado los modelos extraídos que nos ayudaran a borrar patrones redundantes e irrelevantes.

8. **Conocimiento** descubierto, esta etapa permite aplicar el conocimiento adquirido en una mejora del sistema con nuevas medidas o simplemente documentar e informar a las partes interesadas.
   Además, presenta la revisión y resolución de potenciales conflictos con el conocimiento extraído o mejor dicho con el conocimiento existente.

KDD es una metodología para el descubrimiento de conocimientos en bases de datos, en donde la minería de datos se considera como la etapa de mayor trabajo, aclarando otros conceptos sobre esta metodología (Fayyad et al., 1996) indica que KDD no es metodología de minería de datos, puesto que, la minería de datos es la etapa fundamental en la que la metodología del descubrimiento de conocimiento realiza su mayor desarrollo y junto a las demás etapas consigue un resultado exitoso.

1.2.4 Estudio comparativo.

(Moine et al., 2011) menciona que el crecimiento en el área de minería de datos inicia desde el año 2000 y pone en consideración la Ilustración 8, para indicar que de las tres metodologías antes mencionadas, la más destacada por su usabilidad es CRISP-DM.

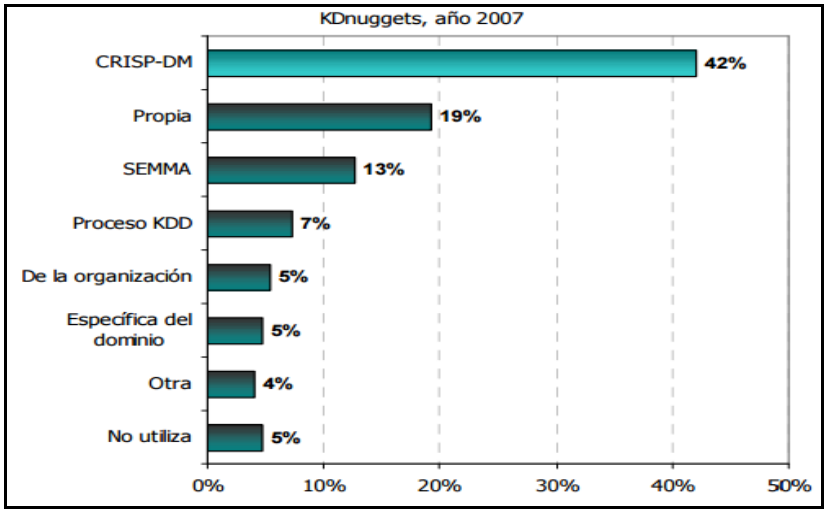


Ilustración 8. Metodologías más utilizadas en minería de datos
Fuente. (Moine et al., 2011)

Este ranquin se mantiene hasta el 2014 y sigue siendo el más recomendado, en el 2015 se presenta una tendencia de metodologías denominada ASUM-DM extendiendo la ya mencionada CRISP-DM, pero aún se esperan resultados para medir el éxito que pueda traer como una metodología de minería de datos. (singular, 2016).

1.3 Educación A Distancia.

Muchas plataformas de enseñanza se suman a la utilización de la red y herramientas tecnológicas, para compartir y obtener información como es el caso de los MOOC, OCW, entre otros.

Se suman a esta gran labor grandes universidades que brindan su educación a distancia, con calidad y gran responsabilidad, para formar grandes profesionales como la UTPL, que cuenta con la Educación Abierta y a Distancia y la educación continua.

Esta forma de educación, facilita la preparación profesional de quienes tienen un puesto de trabajo en tiempo completo; y porque no decir, ayuda en la práctica del conocimiento que adquiere en su educación.

(Aretio, 2001) Autor del libro “La educación a distancia” tras una ardua investigación para encontrar la definición de educación a distancia se atreve a decir que es un sistema tecnológico de comunicación bidireccional, que puede ser masivo, basado en la acción sistemática y conjunta de recursos didácticos y el apoyo de una organización y tutoría, que, separados físicamente de los estudiantes, propician en éstos un aprendizaje independiente y también cooperativo.

A continuación (Martínez, 2008), enuncia las características principales de la educación a distancia:

1. Mediación tecnológica, hacen posible la comunicación multimedia, atreves de audio o video, medios usados entre el profesor, el estudiante y los contenidos del curso.
2. Separación de profesor y alumno durante la mayor parte del proceso instruccional.
3. Influencia de una organización de apoyo al estudiante.
4. Provisión de una comunicación bidireccional entre el profesor, el tutor o la agencia educativa y el estudiante.

Se debe tomar en cuenta que el uso de recursos en la educación a distancia no debe ser limitado, pero si debe ser de calidad, puesto que por el hecho de no contar con un docente que imparta la clase, el estudiante precisa de textos, videos o audios que le ayuden a guiar su aprendizaje, los mismos que son sugeridos y por el docente tutor.

1.3.1 Importancia de la Educación a distancia.

La preparación o crecimiento profesional, es uno de los objetivos más importantes de cualquier persona, sin embargo, existen algunas cadenas que detienen este interés, por ello cualquier opción que reduzca el tiempo de estar en un salón de clases para aprender es muy acogido.

La idea de aprender y trabajar, ocuparse de un hogar y terminar los estudios, hacer más de una sola cosa a la vez es bastante útil para quien desee especializarse en una carrera pendiente o algo nuevo para incrementar su conocimiento, para llegar a esta facilidad de preparación han pasado muchos años y estudios de estrategias para implementarlas como la nueva manera de incursionar una carrera.

(Aretio, 2001) Considera a la educación a distancia como una mediación pedagógica capaz de promover y acompañar el aprendizaje de los interlocutores, es decir, de promover en los educandos la tarea de construir y de apropiarse del conocimiento del mundo, basada cada vez más en los avances tecnológicos.

El maestro continúa desempeñando su función de guía, sin pretensiones de sustituir la actividad creadora del alumno.

El empleo racional de los recursos tecnológicos beneficia considerablemente la localización, reconocimiento, procesamiento y utilización la información obtenida.

Las tecnologías modernas ayudan al educando a resolver problemas, a conocer mejor la realidad objetiva del entorno.

Dentro de las características que mencionan (Raúl Fernández Aedo, Pedro Mario Server García, Elianis Cepero Fadragas, & Romero, 2005) el proceso de aprendizaje a distancia es la respuesta a muchas interrogantes que surgen frente al hecho social de la educación.

Ante la continua preocupación por la necesidad y derecho de una educación permanente, la educación a distancia es una alternativa válida, ya que facilita estrategias de educación permanente.

1.3.2 Teorías de modelos de educación a distancia.

Se han realizado varias investigaciones para generar un modelo de educación a distancia, y no se ha conseguido el resultado anhelado, antes bien se han descubierto diferentes teorías. (Keegan, 1996), en su libro “Fundamentos de la educación a distancia”, nos da un enfoque de las tres más importantes, como son:

- Teoría de la Independencia y la autonomía

- Teoría de la Industrialización
- Teorías de la interacción y la comunicación

(Frías, 2005) cita a Lorenzo García Aretio, quien en base al análisis realizado por Keengan plantea su propia teoría denominándola como: teoría integradora o del diálogo didáctico mediado.

A continuación para mejor conocimiento de estas teorías, la **Tabla 3** detalla el autor y la descripción de cada una.

Tabla 3. Teorías de modelos de EaD

MODELOS DE EDUCACION A DISTANCIA		
NOMBRE DEL MODELO	AUTOR(ES)	DESCRIPCION
Teoría de la Independencia y la autonomía	Rudolf Manfred Delling, Charles A. Wedemeyer y Michael G. Moore. (1960)	El aprendizaje es independiente y autónomo por parte del estudiante pues existe la separación de contenidos de los cursos y el apoyo del tutor, es decir, no se toma en cuenta la interacción.
Teoría de la Industrialización	Otto Peters (1970)	El aprendizaje es convertido en enseñanza y los materiales físicos antes utilizados se vuelven herramientas de trabajo para interacciones en línea, como debates. Este modelo conlleva a los estudiantes tener más libertad y responsabilidad para su autoaprendizaje; tomando en cuenta que el punto de partida que el docente toma para la división del trabajo de un curso determinado es muy importante, pues el modelo se asemeja a la labor de forma industrial.
Teorías de la interacción y la comunicación	Börje Holmberg, John A Bååth, David Sewart, Kevin C. Smith y John S. Daniel (1995)	El aprendizaje ahora tiene una guía más interactiva, pues presenta actividades de colaboración, recursos de aprendizaje y tareas conjuntas, además la parte central del curso tiene lugar en la influencia recíproca entre docente y estudiante a través de comunicación telefónica o mensajería.
Teoría integradora o del diálogo didáctico mediado	Lorenzo García Aretio (2001)	El aprendizaje es de total autoría del estudiante basándose en la disponibilidad de su tiempo, espacio y ritmo de aprendizaje, integra las teorías antes mencionadas en el hecho que utiliza medios de comunicación como vía telefónica, mensajería o telemática para las tutorías que el estudiante requiera y recibe los recursos necesarios de parte de la institución para que genere el conocimiento necesario de un determinado curso.

Fuente. Elaboración propia a partir de (Frías, 2005; Keegan, 1996; Núñez, 2016)

Entonces, es correcto decir, que se puede implementar uno de estos modelos dentro de la educación a distancia de una institución, o a su vez como Lorenzo García, crear una nueva teoría en base a las necesidades o requerimientos que presente la entidad educativa, tomando como referencia las teorías existentes.

1.3.3 Proceso instruccional.

El desempeño del docente no se anula ni pasa a segundo plano en lo que se refiere a la educación a distancia antes bien estos se convierten en moderadores de los contenidos que impartan ya sea ellos mismos o sus estudiantes en un entorno virtual. Siendo así tanto alumno como profesor los principales componentes de esta forma de enseñanza.

(Raúl Fernández Aedo et al., 2005) plantean un esquema de sistemas que cuenta con los siguientes componentes:

- **Entrada** corresponde a los alumnos que se desean mejorar y todos los recursos que van a contribuir en la transformación de esos estudiantes. Aquí tienen un papel fundamental los medios que se diseñen y las características del alumno

- **Salida** es el alumno mejorado, es decir, el que ha alcanzado los niveles exigidos en logro de los objetivos de aprendizaje.

- **Procesador** son las interacciones y/o experiencias de aprendizaje que proporciona el sistema de educación a distancia. El texto debe estar elaborado de tal forma que permita al alumno resolver la mayoría de las dudas que se le pudieren presentar. Pero no sólo se pueden emplear textos, también es necesario el uso de otros medios: la televisión, las redes computacionales, la radio, entre otros.

- **Control** permite diagnosticar las conductas de entrada, verifica los resultados finales y supervisa todo el proceso. Aquí tienen especial importancia los trabajos de investigación y de aplicación que deben realizar e informar los estudiantes.

- **Ambiente** son todas las variables que influyen en el sistema, que no puede controlar, aunque sí puede influir en ellas. Dentro de esa preparación están mencionadas algunas variables propias del Ambiente: lugar de estudio, horarios adecuados, uso de recursos de la comunidad, incluso vida familiar. La influencia que se ejerce sobre esos factores se traduce en las recomendaciones que se le entregan al estudiante para que aproveche de una manera eficaz y eficiente los aprendizajes propuestos.

- **Retroalimentación** permite confirmar los resultados de los esfuerzos de enseñanza y de aprendizaje, para lo cual el centro emisor necesita un adecuado Sistema De Información Administrativa.

Pero debemos tener en cuenta que es tan solo una sugerencia pues el modelo institucional con el que el docente va a seguir para elaborar un curso dependerá de los requisitos que este tenga en su materia.

(Rodríguez, 2009) afirma que en la educación distancia, los diferentes modelos de diseño instruccional que pretenden clarificar a quien recibe la instrucción, las formas de lograr el aprendizaje, evitando que la distancia sea un impedimento.

En la Ilustración 9 nos muestra diferencias entre algunos modelos de diseño.

MODELOS INSTRUCCIONALES			
MODELO	**TEORIA DE APRENDIZAJE QUE PREDOMINA**	**CARACTERÍSTICAS BÁSICAS**	**N° DE PASOS/ELEMENTOS**
ASSURE	Cognitivismo (Gagné)	Diseño de instrucción que incorpora el uso de los medios y tecnología.	6 secuenciales
Modelo de los Procedimientos de Interservicios para el desarrollo de Sistemas Instruccionales	Conductista	Utilizado por las fuerzas armadas del país.	5 secuenciales
Jerrold Kemp	Constructivista	La forma oval del modelo da al diseñador el sentido que el diseño y el proceso de desarrollo es un ciclo continuo que requiere de planeación, y que la evaluación constante asegura una instrucción eficaz.	9 flexibles, no lineales
Dick y Carey	Conductista	Este modelo describe todas las fases de un proceso interactivo, que comienza identificando las metas instruccionales y termina con una evaluación sumaria.	10 secuenciales

Ilustración 9. Modelos instruccionales de EaD
Fuente. (Rodríguez, 2009)

1.4 E-learning.

E-Learning (aprendizaje electrónico), pertenece a uno de las modalidades de la educación a distancia más utilizado.

Para definir lo que es E-learning encontramos una exposición de numerosos conceptos, llegando a una simple definición; e-learning es un proceso innovador de las universidades, en lo que enseñanza-aprendizaje se refiere. (Schneckenberg, 2004) nos ayuda a comprender esto de mejor manera, con algunos componentes involucrados en esta nueva metodología de estudio, tales como:

- Progreso tecnológico de aplicaciones de e-learning.
- Interés económico de las empresas y los actores involucrados.
- Diferentes modelos económicos y estrategias de organización del e-learning.
- Rol cambiante de los formadores y de los alumnos en entornos virtuales de formación.
- Importancia de la pedagogía de los medios para el desarrollo futuro y una integración sostenible del e-learning en la educación superior.

Sin embargo, es preciso tomar en cuenta que e-learning posee sus propias "reglas de juego" y, en consecuencia, presenta similitudes y diferencias tanto con la formación a distancia como con la formación presencial. En realidad, el e-learning es el conjunto de características de ambas modalidades con elementos genuinamente suyos. (Seoane & García, 2010)

1.4.1 Evolución de E-learning.

El camino para tener lo que ahora llamamos E-learning ha sido de muchos obstáculos; aunque el verdadero inicio no se ha definido verazmente se ha podido recopilar su historia de evolución desde el año 1950 como podremos apreciar en la Ilustración 10, gracias a infografías de diferentes autores, dedicados al estudio de la evolución de e-learning, entre ellos citamos a (Algieri et al., 2014a; García Aretio, 1999a).

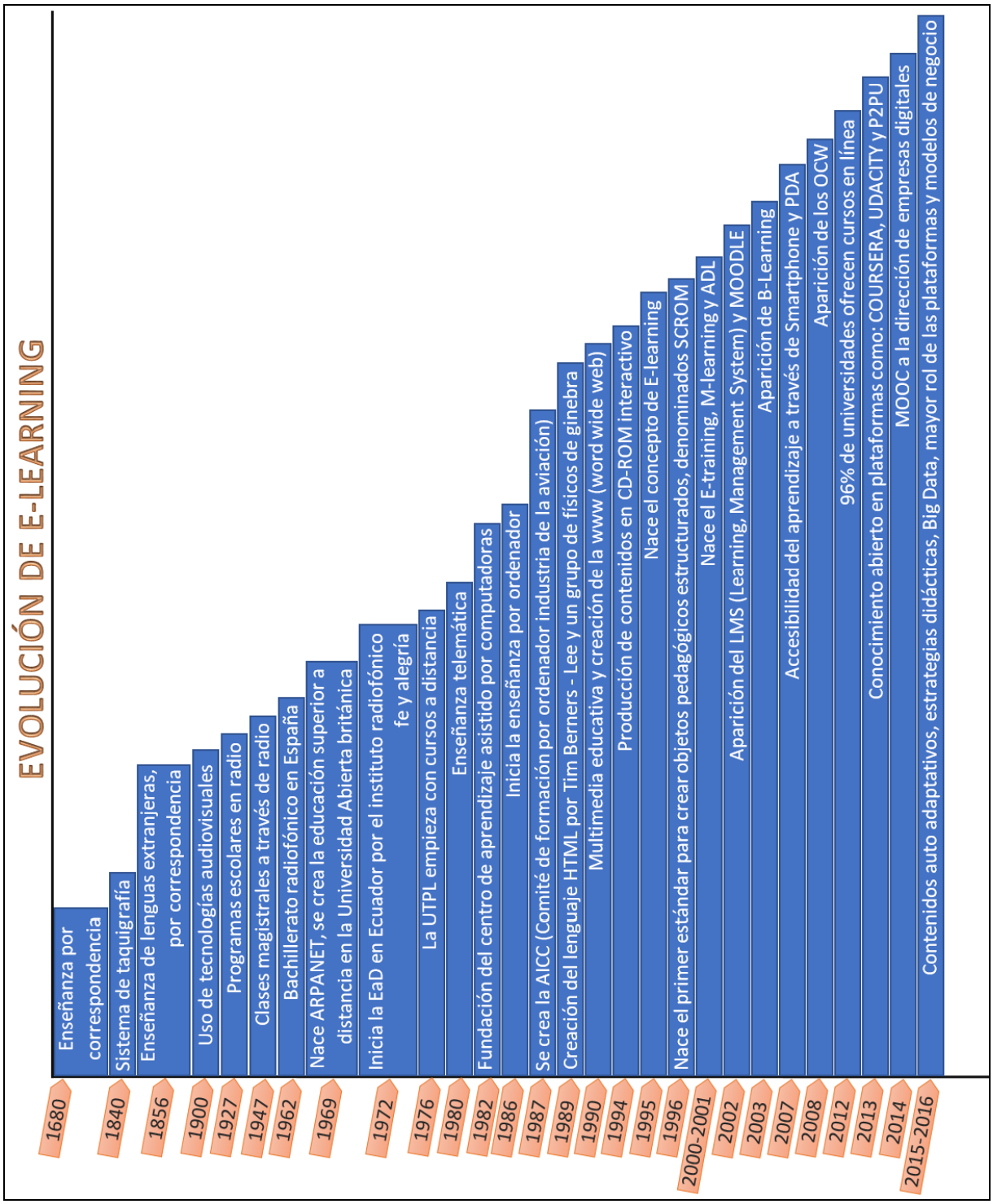


Ilustración 10. Evolución de E-Learning

Fuente. Elaboración propia a partir de (alexcouoh4, 2016; Algieri et al., 2014b; García Aretio, 1999b; WeLearning, 2015)

E-learning ha dado grandes pasos para la formación profesional, sin depender de un espacio físico, buscando la forma y el recurso para lograrlo, se inicia desde una enseñanza por correspondencia, pasando por una enseñanza radiofónica y finalmente adoptando tecnologías como video llamadas, chat o foros, que ha permitido al docente formar a alumnos a quienes no podían llegar físicamente, y a los alumnos que no contaban co un tiempo propicio para asistir a un salón de clases, organizar su horario de estudio según su disponibilidad de tiempo y finamente conseguir un título profesional.

1.4.2 Recursos de aprendizaje.

Dentro de una publicación realizada por la editorial UOC ("Evolución y retos de la educación virtual," 2011) indica que los recursos utilizados para y en el desarrollo de la actividad por parte de los estudiantes influyen y determinan el logro de los objetivos de aprendizaje propuestos, hoy en el ámbito de educación superior, nos referiremos tanto a los contenidos como a las herramientas y plataformas, en tanto que componentes posibles de los actuales escenarios de aprendizaje virtuales o parcialmente mediados por TIC.

Los recursos educativos abiertos son materiales en formato digital que se ofrecen de manera gratuita y abierta para educadores, estudiantes y autodidactas, para su uso y re-uso en la enseñanza, el aprendizaje y la investigación y estos pueden ser libros, artículos, materiales didácticos, guías y referencias de lecturas, materiales de un curso, y estos pueden ser desarrollados ya sea por profesores como por universidades, bibliotecas, organizaciones educativas o alguna persona interesada en crearlos.

1.5 Materias básicas impartidas en las universidades.

Las materias básicas que se imparten en las universidades son aquellas que pueden ser convalidadas sin ninguna restricción en el caso que el estudiante acceda o cambie de carrera en cualquier área de conocimiento (U.Alicante, 2007).

El porcentaje y número de créditos que representa aprobar estas materias, difiere entre universidades. (Universia, 2008) cita el "anexo II del Real Decreto 1393/2007", para indicar que la suma de créditos de las materias de formación básica de un estudiante universitario debe ser de 36 inclinadas a asignaturas de mínimo de 6 créditos cada una.

La UTPL sigue esta normativa y cuenta con: Metodología de Estudio, Realidad Nacional y Ambiental, Expresión Oral y Escrita, Antropología, Ética, Computación y Jornadas de Investigación Temática y Formación Espiritual como parte de las asignaturas de formación básica consideradas como el 10% del conocimiento adquirido por parte del estudiante significando una suma de 36 créditos, como lo menciona (Universia, 2008).

1.6 Perfiles de los docentes de educación a distancia.

Para apuntar al perfil de un docente de educación en cualquier modalidad ya sea esta presencial o a distancia, se debe tomar en cuenta 4 pilares fundamentales de la educación, como muestra la Ilustración 11 de manera muy sencilla y clara; elaborada dentro de un blog denominado "LA HUELLA DEL MAESTRO" (Velasco Laura, Rodenas Sonia, & Virseda Silvia, 2008) basándose en el informe Jacques Delors realizado por la Comisión Europea a través de la UNESCO sobre la educación, dejándonos muy en claro que si uno de estos falla o falta la educación tambalea hasta llegar al punto de derrumbarse.

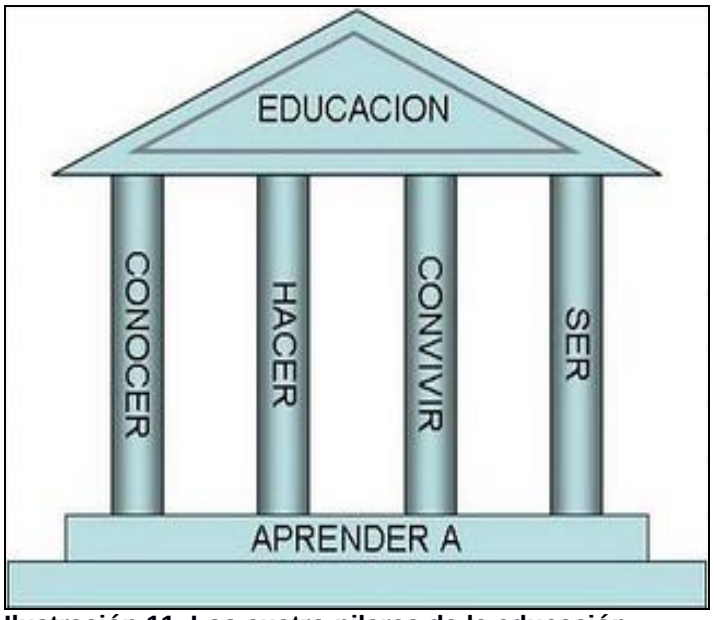


Ilustración 11. Los cuatro pilares de la educación
Fuente. (Velasco Laura et al., 2008)

(Delors Jacques, 1996) menciona que la comunicación en el siglo XXI forma parte de un recurso indispensable en el fortalecimiento de la educación, ya que a través de la misma se fomenta la transmisión de conocimientos teóricos y técnicos adecuados a la civilización cognoscente.

La educación desde cierto punto constituye la guía para el desarrollo de proyectos comunitarios e individuales, yendo más allá de la acumulación de conocimientos, adecua métodos versátiles del mundo en desarrollo, llevando a cabo la misión de la educación en base a 4 pilares de aprendizaje fundamentales, que son:

1. Aprender a conocer
2. Aprender a hacer
3. Aprender a convivir
4. Aprender a ser

A continuación (Delors Jacques, 1996) nos ayuda a definir a cada uno de ellos:

1. APRENDER A CONOCER

Este pilar se refiere al conocimiento que adquiere una persona para entender el mundo que lo rodea desarrollando sus habilidades profesionales y formas de comunicación con los demás, llegando a un fin que justifica su interés por comprender, conocer y descubrir todo esto. Pero es tan infinito y evolutivo el conocimiento, debemos tener en cuenta que no existe un saber absoluto.

2. APRENDER A HACER

Va muy vinculado con el *aprender a conocer*; busca por medio de nuestro desempeño en los trabajos o actividades, influir sobre el entorno.

Este pilar se inclina más a la formación profesional, pues se basa en la manera de enseñar al alumno a poner en práctica sus saberes y, al mismo tiempo, cómo adaptar la enseñanza a su futuro.

El aprender hacer no es solo el fin de adquirir una calificación profesional, va de la mano con el crear un espíritu competitivo que procure la capacitación del individuo para un gran número de situaciones y el trabajar en equipo, es decir, no limitarse a obtener conocimientos, para que la guía como docente llegue a incentivar al estudiante.

3. APRENDER A CONVIVIR

La palabra convivir ya nos da un indicio de lo que trata este pilar, sin embargo, como nos recuerda Delors se conoce que desde el siglo xx, hasta en tiempos antiguos la violencia y guerras han producido grandes catástrofes y pérdidas de vidas.

Es una tarea ardua llegar a que el ser humano se complemente de valores, y elimine estereotipos de competitividad nociva en el aspecto económico, ético, social de cada nación. Aún mejor la educación debe fomentar el trabajo multidisciplinario de equipo, pero en ocasiones el mal uso de la información y de la comunicación hace que se agrave la realidad conflictiva.

La experiencia demuestra que, para disminuir ese riesgo, no basta con organizar el contacto y la comunicación entre miembros de grupos diferentes, sino en organizarse con un plan específico, formulando objetivos y participando en proyectos comunes.

El fomento de esta actitud de empatía en la escuela será fecundo para los comportamientos sociales a lo largo de la vida.

4. APRENDER A SER

Este último pilar se refiere a las capacidades desarrolladas de cada individuo, como ya hemos mencionado el crecimiento profesional es continuo, teniendo en cuenta que no por tener más o menos capacidad en desarrollar cierto proyecto o actividad nos ubicamos en niveles de ser mejor o peor, cada individuo es único e importante para el desarrollo y éxito de cualquier trabajo.

“Habrá que ofrecer a niños y jóvenes todas las oportunidades posibles de descubrimiento y experimentación -estética, deportiva, científica, cultural y social”

(Delors Jacques, 1996)

La educación nos aporta conocimiento necesario para tener un pensamiento autónomo y crítico, elaborar nuestros propios juicios y tomar las decisiones correctas en diferentes situaciones de nuestra vida; de ahí la importancia de adquirirlo frecuentemente.

Claro que al hablar de la enseñanza y crecimiento profesional los dos primeros pilares son los que se orientan a estos conceptos y la armonía social, en donde se comparta responsabilidades y experiencias, corresponden a los pilares de aprender a vivir juntos o aprender a convivir como lo hemos nombrado y el de aprender a ser. Sin embargo, esto es solo una idea teórica ya que en la realidad la valoración de los pilares es de forma desigual.

Después de tener en claro lo que implica la educación, hablaremos del perfil que un docente de la modalidad abierta y distancia debe tener para su desempeño profesional.

(Perdomo Maribel, 2008) nos facilita este perfil mediante la elaboración de un artículo expuesto en el “Congreso Virtual de Calidad en Educación a Distancia”, resaltando lo ya expuesto anteriormente sobre la educación en general, como son los 4 pilares de la educación: Aprender a conocer, aprender hacer, aprender a convivir, aprender a ser.

Perdomo enfatiza que el compromiso de una educación a distancia no es tan solo del docente sino también de la institución que esta presta sus servicios. Pues el apoyo adecuado para el crecimiento educativo, como capacitaciones constantes en diferentes áreas y herramientas como una buena plataforma tecnológica, son la clave para contar con profesores que por sí solos descubran y diseñan nuevos métodos didácticos apropiados para asumir el rol de formar estudiantes de auto aprendizaje.

El modelo que propone Pedroso es basado en competencias, como son:

- **Competencias Pedagógicas:** componen habilidades, aptitudes y conocimientos básicos de la pedagogía a distancia, formando docentes que resuelvan problemas de tipo didáctico relacionados con la planificación y desarrollo de la docencia.
- **Competencias Tecnológicas:** integran habilidades, aptitudes y conocimientos básicos de los entornos virtuales de aprendizaje y recursos tecnológicos necesarios para la mediación didáctica a distancia.
- **Competencias de Orden Ético y Legal:** orientan el conocimiento de un docente de la EaD para aplicarlos en circunstancias o políticas para proteger los derechos de

autor, propiedad intelectual, entre otros aspectos, de la legalidad de la distribución de materiales por Internet.

Las mismas que al tenerlas en cuenta el desempeño del docente se extiende a facilitador de experiencias de aprendizaje que promuevan el estudio autónomo e independiente de los estudiantes.

Finalmente, el modelo del perfil por competencias del docente a Distancia que permite romper paredes, como limitaciones geográficas, físicas y métodos tradicionales, e introduce métodos de enseñanza y porque no decir también de aprendizaje y formas de comunicación, se basa en tres roles: *facilitador*, *tutor*, *mediador de tecnología*, *promotor de aspectos éticos y legales*, *e investigador*. Descritos en la Ilustración 12.

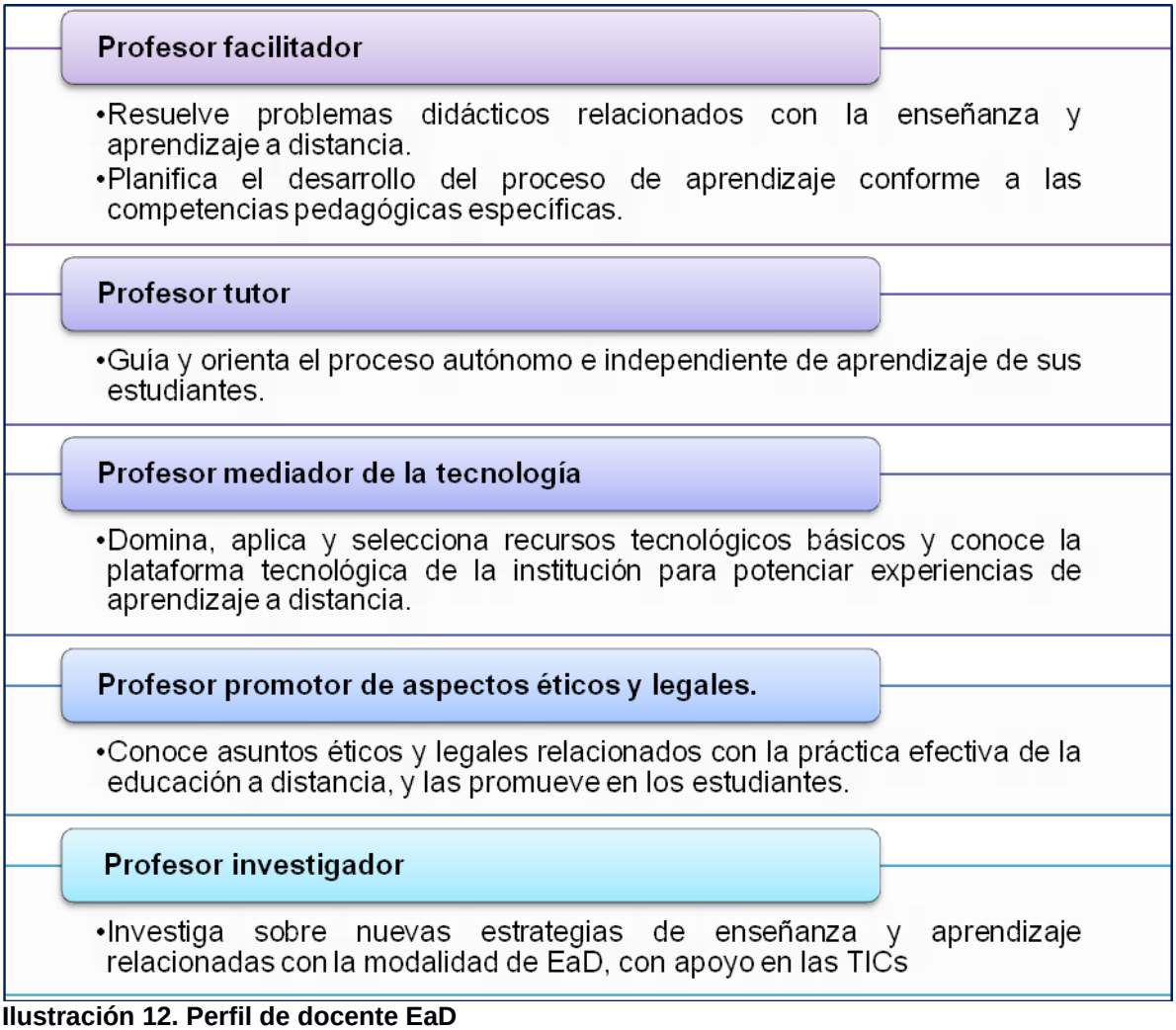


Ilustración 12. Perfil de docente EaD
Fuente. Elaboración Propia

Como esta investigación se han realizado sin duda algunas más; (Guanipa Perez & Urdaneta Marcos, 2007) autores de la investigación llamada “Perfil de competencias del

docente para la educación a distancia" sumando a los roles antes mencionados un programa tutorial aplicado a la educación universitaria a distancia apreciado en la Tabla 4

Tabla 4. Programa tutorial de educación a distancia

OBJETIVOS ESPECÍFICOS	ROLES DEL TUTOR EN LÍNEA	FUNCIONES DEL TUTOR EN LÍNEA
• Asegurarse que el estudiante entienda las ideas y argumentos presentados en las unidades y programas del curso. • Facilitar las dificultades académicas del estudiante. • Ayudar al estudiante a hacer un uso apropiado de los medios y estrategias Instruccionales disponibles en su contexto particular de aprendizaje. • Retroalimentar al sistema de evaluación que controla el proceso de enseñanza–aprendizaje.	• Orientador, motivador • Facilitador. • Mediador. • Asesor académico. • Evaluador del aprendizaje. • Evaluador	• Ayudar al estudiante a organizar patrones de estudio, familiarizarse con los métodos de enseñanza y tomar decisiones sobre cursos a escoger. • Atender consultas de estudiantes sobre materiales didácticos. • Corregir, calificar y comentar pruebas de evaluación a distancia. • Asesorar y orientar al alumno. • Asesorar a los estudiantes en relación a la conducta y hábitos de estudio más recomendados para estudiar a distancia. • Servir de intermediario

Fuente. (Guanipa Perez & Urdaneta Marcos, 2007)

1.7 Casos de deserción a inicios de una carrera universitaria.

Todos los periodos de inicios universitarios empiezan con un gran número de estudiantes, pero en todas las universidades que ofrecen ya sea en la educación a distancia como la clásica/presencial pasan por el abandono de muchos de sus estudiantes, a continuación, conoceremos algunas razones de estos desertores.

- En la universidad de Costa Rica (Abarca & Sánchez, 2005) realizaron un investigación basada en 450 entrevistas a estudiantes y 30 a expertos con

académicos de alto rango de la universidad y determinaron que sin importar el sexo influyen factores como el *ambiente educativo*, *la edad* al presentarse inmadurez en muchos estudiantes, apatía por programas curriculares y el factor económico; los mismos que van acompañados por diferentes características de dichos desertores las mismas que pueden ser:

- Encontrarse en una carrera que no es de su agrado.
- Falta de información para tomar la carrera
- Seguir una carrera a fin de la que realmente deseaba estudiar
- Cubrir necesidades económicas trabajando o trámites burocráticos

- En el caso de la educación superior en Sonora, México (Abril Valdez, Román Pérez, Cubillas Rodríguez, & Moreno Celaya, 2008) presenta un estudio basado tras una encuesta a 147 jóvenes sobre situación familiar, historia escolar, motivos de deserción y planes futuros, entre otros. Los resultados muestran que 86% de las personas participantes abandonó la escuela entre el primer y tercer semestre, con un promedio de calificación, en el último semestre cursado, de 7.49. Las principales razones para dejar de estudiar fueron los factores económicos, haber reprobado materias y la falta de interés. De los participantes, 93% no estaba satisfecho con el nivel de estudios alcanzado, sin embargo, no tenía planeado retomar estas actividades. Los resultados muestran la necesidad de un modelo de intervención basado en políticas educativas con mayores incentivos para una adherencia al sistema escolar, flexibilizar el tránsito entre subsistemas y reestructuración de las redes de comunicación entre los actores principales.

- (Garrálaga, 2014) indica que en el caso de la UTPL se aplicó un proyecto de tesis denominado “mentoría entre pares”; con estudiantes egresados como mentores y estudiantes de primer ciclo como mentorizados, cuya finalidad fue mejorar la calidad de los procesos de orientación académica, personal y sumar a ello la implementación de la cultura de acompañamiento.

  El resultado de dicha propuesta fue una disminución del 20% de deserción de los estudiantes de primer ciclo y un aporte de reforzamiento académico en los estudiantes egresados que cumplieron con el papel de mentorizadores.

1.8 Programa de formación básica en la UTPL.

Dentro de la oferta académica de la UTPL de modalidad a distancia encontramos: Metodología de estudio, Realidad Nacional y Ambiental, Expresión Oral y Escrita, Antropología, Ética, Computación y Jornadas de Investigación Temática y Formación Espiritual que forman parte del 10% de la aprobación de créditos para terminar una carrera, independientemente que la duración de la titulación sea de 8 o 10 niveles, dato verificado por la pagina informativa de la UTPL en el área de modalidad a distancia; (Rubio, 2014) de las cuales en nuestra investigación serán tomadas en cuenta Realidad Nacional, Expresión Oral y Escrita; y Metodología de Estudio, por ser las materias con mayor deserción según investigaciones anteriores.

CAPÍTULO II: DEFINICIÓN Y PLANTEAMIENTO DE LA SOLUCIÓN DEL PROBLEMA

2.1 Definición del problema.

El tema de la deserción de estudiantes que se encuentra presente en todas las universidades, es muy extenso y hasta la actualidad no se ha logrado descubrir una solución del 100%. Indicadores como dificultades económicas, disponibilidad de tiempo, problemas académicos, matriculación, o problemas familiares, mencionados por (Arévalo & Maldonado, 2010), docentes investigadoras de la UTPL, han sido algunas de las causas del abandono. Ana barrera comenta para (El comercio, 2016), que la falta de orientación vocacional es también uno de los factores que incide en la deserción de los estudiantes. Pero para el 2016 el titular del Senecyt Rene Ramírez asegura que el tema de deserción ha disminuido un 20% gracias al cambio de reforma en la educación.

En la UTPL el análisis ciclo a ciclo, por los docentes investigadores responsables, el abandono en Modalidad Abierta y a Distancia alcanza casi el 50%, por parte de alumnos que ingresan a la unidad educativa, es decir, la mitad del alumnado que ingresa a estudiar en la MAD, abandona sus estudios en el primer ciclo.

Con la incorporación del modelo educativo basado en competencias y créditos académicos ECTS desde el año 2007 surgen materias de formación básica como: Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, como algunas de las materias que están presentes en todas las mallas curriculares de la Modalidad Abierta y a Distancia.

Por el hecho de ser materias de formación básica y de primer ciclo, cuentan con un número elevado de alumnos, entre 7000 a 11000 estudiantes, esto dependiendo del periodo y la materia/componente; por ello es necesario brindar asesoría u orientación, a través de un seguimiento de las actividades elaboradas en las materias. Además de otros factores de calidad como lo es el planteamiento de evaluaciones presenciales, evaluaciones a distancia, materiales educativos, entre otros, los cuales serán vitales para el logro de competencias, resultados de aprendizaje y por consiguiente la aprobación de los componentes.

Es por ello que, la institución, ha realizado varios cambios en diferentes procesos que lleven a mejorar, la calidad de las materias, tales como la mentoría entre pares mencionado en el CAPITULO I o la reducción en el número de alumnos por docente, siendo por ello necesario un mayor seguimiento, atención y apoyo a los alumnos de primer ciclo que inician en la Modalidad Abierta y a Distancia por la susceptibilidad en razón del abandono.

Aunque se ha tomado medidas como las ya mencionadas, el abandono en los ciclos iniciales continúa, es por ello la necesidad de realizar un análisis de la evolución de estas materias dentro de los últimos periodos culminados de oct-2014/feb-2015 y abr-2015/ago-

2015, tomando las acciones que los estudiantes realicen dentro de las diferentes actividades como: video conferencias, quiz, chats y foros, dentro del entorno virtual de aprendizaje.

2.2 Planteamiento de la solución del problema.

Para la solución del problema, se plantea el desarrollo de un modelo descriptivo, acerca de la labor tutorial del docente de modalidad a abierta y a distancia, en base a las actividades que los estudiantes realicen en el entorno virtual de aprendizaje en las materias de formación básica de Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, comprendidas en los periodos oct-2014/feb-2015 y abr-2015/ago-2015.

La minería de datos ha servido como soporte técnico en decisiones empresariales como lo presentan (Aular & Pereira, 2007)en un artículo acerca del razonamiento estadístico, el enfoque de patrones, procesamiento paralelo, herramientas basadas en la teoría de la decisión y el aprendizaje automático que en su mayoría aplica reglas a partir de los datos para la extracción de conocimiento.

Por tal razón para el desarrollo del modelo, con la data proporcionada se aplicará minería de datos para su estudio, tratamiento y análisis, obteniendo finalmente la extracción de conocimiento, con el que podremos determinar conclusiones y recomendaciones del proyecto y por ende el modelo descriptivo.

Este trabajo de minería de datos se lo elaborara en la herramienta R, la misma que se acopla con el tipo de datos obtenidos de la base de datos del EVA, siendo la más recomendada en la sección 1.1.4 Herramientas de data mining, por su variabilidad en cuanto a gráficas para visualizar los resultados de las técnicas empleadas.

En la sección 1.2 estudiamos las metodologías existentes para el desarrollo de un proyecto de minería de datos y (Moine et al., 2011) cita resultados del estudio comparativo de metodologías realizado por la comunidad KDnuggets (Data Mining Community's Top Resource), del cual se destaca según la usabilidad que han tenido expertos de minería de datos, la metodología CRISP-DM, por tal razón se la ha considerado para ser aplicada en el desarrollo de este proyecto.

Entonces, los datos de las diversas acciones de las actividades que un estudiante realiza en el entorno virtual de aprendizaje, considerando las materias de formación básica de: Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, pertenecientes a los periodos de oct-2014/feb-2015 y abr-2015/ago-2015; y obtenida del resultado de consultas de la base de datos del EVA y del reporte solicitado de la base de datos del SYLLABUS, Se realizará la exploración, limpieza e integración de la misma.

Se utilizará MySQL para la integración de los datos, Excel para la limpieza y R para la exploración y la interpretación de los mismos a través de técnicas de minería cómo la de clustering o análisis de componentes principales.

Los datos del proyecto no determinan clases para realizar un análisis, pero podemos agrupar características comunes de los estudiantes con determinadas acciones, por esta razón se ha considerado la técnica de **clustering,** que como se menciona en la sección 1.1.3 se basa en agrupar objetos con la finalidad de maximizar similitud dentro de las clases y minimizar la similitud fuera de ellas.

En esta técnica, (Fernández Santana, 1991) indica una clasificación de modelos como:

- Jerárquicos aglomerativos: distancias mínimas, distancias máximas, distancias entre centroides, distancias ponderadas, Ward
- Jerárquicos divisivos: monotéticos o politéticos
- Partición iterativa: centros móviles, K-means, Hill-Climbing, Isodata
- Búsqueda de densidad: NORMIX, NORMAP
- Análisis factorial <<Q>>
- Clumping
- Métodos basados en la teoría de los grafos

En base a esta clasificación y los datos con la que se cuenta, es conveniente utilizar el modelo de **partición iterativa** con el algoritmo más conocido que es **K-means**, el cual basa su mejora en los resultados obtenidos de una participación en un grupo anterior, según (M. Vargas, 2012); he indica algunas características que se necesita tomar en cuenta para aplicar este algoritmo, tales como:

- Se especifica a priori los grupos de trabajo.
- Se trabaja directamente con la matriz de datos originales, no con la matriz de distancias, permitiéndonos el análisis de gran número de casos.
- No precisa del almacenamiento de la matriz de aproximaciones NxN para la determinación de grupos.
- Es considerado dentro de los métodos de reasignación, por contar con la asignación a un clúster en un proceso específico y posteriormente ser reasignado en otro proceso en un clúster diferente.

El aporte que la UC3M nos brinda a través de su reporte del (*Análisis de Cluster y Arboles de Clasificación*, 2010) como aplicar el algoritmo K-mean en los siguientes pasos:

1. Se toman al azar k clusters iniciales.
2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clusters y se reasignan a los que estén más próximos. Se vuelven a recalcular los centroides de los k clusters después de las reasignaciones de los elementos.
3. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo.

En la Ilustración 13 podemos entender mejor el proceso del algoritmo k-means

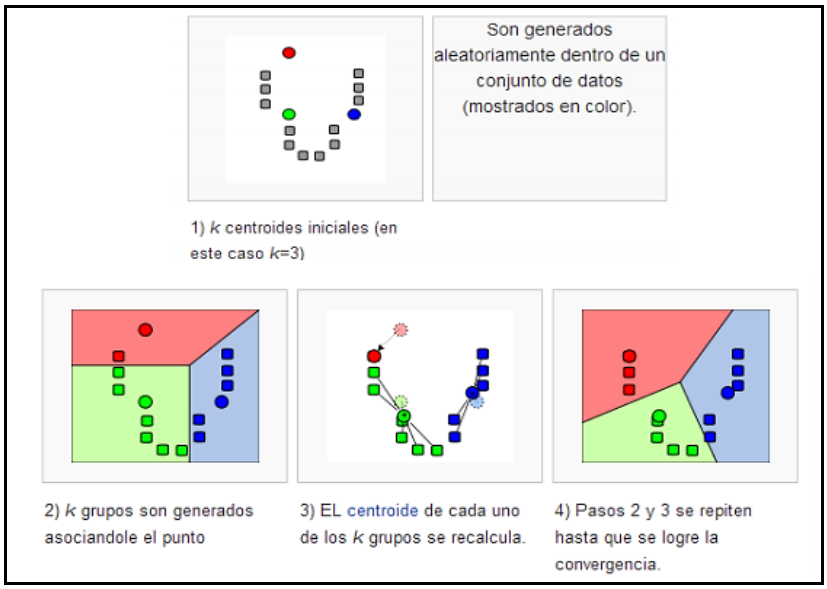


Ilustración 13. Pasos de K-mean
Fuente. (Picasso Iñaki Díaz Covián, 2014)

K-Means en R

R como se menciona anteriormente es la herramienta de minería de datos con la que se analizaran la data. Esta herramienta dispone de varios algoritmos para el procedimiento de agrupación, entre ellos el algoritmo **k-means**, con el que se procesaran los datos.

(Aluja, 2008) indica las sentencias con las que trabaja este algoritmo en la herramienta seleccionada de la siguiente manera: kmeans(x, centers, iter.max =10)

Descripción de cada sentencia:

- x: una matriz numérica de los datos o de una trama de datos con todas las columnas numéricas.
- centers: número de agrupaciones o conjunto de grupo inicial, se elige el primer conjunto aleatorio de filas de x a medida de los clusters iniciales.
- iter.max: número máximo de iteraciones permitidas

Al obtener el resultado de este algoritmo se procederá a determinar conclusiones, pero si no presenta información relevante para el modelo descriptivo del proyecto, se aplicará otra técnica, como lo es el análisis de componentes principales que también se presenta en las técnicas descriptivas de minería de datos, la misma que se describirá a detalle en su desarrollo.

CAPÍTULO III: PROPUESTA DEL MODELO

3.1 Fase 1. Comprensión del negocio.

3.1.1 Objetivos del negocio.

La UTPL como universidad “expertise” en educación a distancia ha tenido como objetivo principal llegar a las personas que no puedan acceder a la educación superior presencial, permitiendo con ello educación permanente con igualdad de oportunidades para quienes buscan mejorar cada día su nivel profesional, sumándose a ello:

- Proceso formativo, mediante la aplicación de las nuevas tecnologías.
- Crear una autonomía responsable del estudiante en el aprendizaje
- Desarrollar y aplicar proyectos de investigación tecnológica y científica en relación a las carreras que imparte la UTPL.

3.1.2 Evaluación de la situación.

Al contar con la base de datos del EVA y reporte de datos nominales del syllabus de la UTPL, se analizó el capítulo II referente a la problemática a tratar, pudiendo determinar: *recursos, requerimientos, supuestos, restricciones y terminología* con los que trabajaremos para hacer la minería de datos.

3.1.2.1 Recursos.

En cuanto a los recursos tenemos:

a) Recurso personal

1. Expertos en el negocio
2. Experto en minería de datos
3. Administrador del EVA
4. Administrador de la base de datos del syllabus

b) Recursos de datos

- Información nominal de todos los estudiantes que cursaron materias de nivel básico los periodos oct-2014/feb-2015 y abr-2015/ago-2015.
- Datos de las acciones en el EVA que un estudiante ha realizado en las materias de nivel básico.

c) Recurso de hardware

- PC LENOVO ThinkCentre
- con 16GB de RAM
- procesador Intel CORE i7

d) Recurso de software

- **MySQL Workbench:** utilizada para gestionar y consultar en la base de datos.
- **XAMPP:** usada para manipular los datos y crear vistas de la base de datos.
- **Microsoft Excel 2010:** utilizada para el proceso de cálculos matemáticos de datos
- **R:** usada para aplicar el análisis estadístico y las técnicas de minería de datos.

3.1.2.2 Requerimientos.

Los requerimientos para el desarrollo del proyecto son los siguientes:

- Acceso a los datos de la base de datos del EVA y del SYLABUS
- Conocimiento sobre el manejo de las herramientas: MySql, XAMPP Y R
- Calendario de trabajo para revisiones con los expertos.

3.1.2.3 Supuestos.

Para trabajar con los datos y aplicar la minería suponemos que:

- Los estudiantes se encuentran matriculados en materias de nivel básico independientemente del ciclo.
- Los datos de los estudiantes corresponden a los periodos oct-2014/feb-2015 y abr-2015/ago-2015.
- Los datos nominales de los estudiantes se los obtiene por solicitud de reportes a los administradores de la base de datos del syllabus.
- Según la participación en las actividades del EVA será evaluada el nivel de interactividad con su tutor.

3.1.2.4 Restricciones.

Los limitantes presentes para nuestro proyecto son:

- Datos nominales faltantes en los usuarios en la base de datos del EVA
- Al no contar con todos los nominales de los estudiantes, dependemos de una petición de reporte a los administradores de la base de datos del syllabus.
- La relación de las tablas para obtener consultas se encuentra muy confusa

3.1.2.5 Terminología.

Los términos más utilizados en el proyecto ya sea la terminología del negocio como en la minería de datos son los siguientes:

Terminología del negocio

- EVA: Entorno Virtual de Aprendizaje.
- SYLLABUS: Software que facilita los procesos académicos de los estudiantes, por ejemplo: consulta de notas, matriculas, entre otras.
- MAD: Modalidad Abierta y a Distancia.
- UTPL: Universidad Técnica Particular de Loja.
- Materias de Formación Básica: son componentes con temáticas fundamentales que reflejan la dinámica de la universidad.
- E-Learning: es la educación y capacitación a través del internet.

Terminología de la minería de datos

- Data set: almacenamiento de datos.
- Data mining: extracción de información seleccionada de grandes bases de datos.
- Datos anormales: data resultante de errores.
- Modelo analítico: estructura y proceso para analizar un conjunto de datos.
- Modelo lineal: modelo analítico que asume relaciones lineales entre una variable dependiente y sus variables independientes.
- Modelo no lineal: es un modelo analítico que no asume una relación lineal en los factores de las variables estudiadas.
- Dato nominal: pueden ser categorizado y no llevan ningún orden de presentación.
- Dato ordinal: posee una orden de presentación, podemos hacer comparaciones de mayor a menor y al mismo tiempo de igualdades y desigualdades.
- Dato atípico: se considera al dato distante del resto de datos
- Discretización: extraer de un conjunto infinito de datos una limitada cantidad, para trabajarlos.
- K-means: metodología de data mining para aplicar cauterización.
- CP: componente principal.

3.1.3 Objetivos de la minería.

Los objetivos de la minería de datos son:

- Diseñar un modelo descriptivo de la labor tutorial del docente en base a las acciones con mayor número de participación de los estudiantes en las asignaturas de formación básica de MAD.
- Analizar el modelo de base de datos con el que cuenta moodle, en relación de las actividades de los estudiantes de la MDA en materias de Formación básica en el EVA.
- Aplicar técnicas de minería de datos en la data seleccionada.

3.1.4 Plan del proyecto.

El desarrollo del proyecto se ha dividido en 5 etapas para facilitar la organización del trabajo y con ello se ha estimado un tiempo de 9 meses, es importante destacar que algunas actividades se trabajan en paralelo, a continuación, se detallan a cada una de ellas.

1. Estado del arte (6 semanas):
    - Revisión bibliográfica en torno a la definición y uso de herramientas de data mining.
    - Revisión bibliográfica de trabajos con datos masivos.
    - Revisión bibliográfica sobre los problemas presentados con datos masivos.
    - Redacción de la investigación
2. Definición y alcance del problema (7 semanas):
    - Análisis en los inconvenientes presentados antes del uso de data mining.
    - Definición del problema
    - Alcance del problema
3. Pre procesamiento de datos (11 semanas):
    - Obtención de datos
    - Selección de datos
4. Selección de herramienta (8 semanas)
    - Aplicación de Técnicas de Data Mining

- Análisis de resultado
- Interpretación

5. Conclusiones y recomendaciones (4 semanas)
    - Definiciones finales acerca de los resultados obtenidos.
    - Recomendaciones para mejorar las tutorías de docentes MAD.

3.2 Fase 2. Comprensión de los datos.

En esta fase se inicia la identificación y recolección de datos afines con el problema, comprobando la calidad y las relaciones entre ellos.

La información que la data nos proporciona es acerca de las actividades de docentes que imparten y estudiantes que reciben materias de nivel básico de Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, en la modalidad abierta y a distancia de la UTPL, correspondiente a los periodos oct-2014/feb-2015 y abr-2015/ago-2015. Ilustración 14

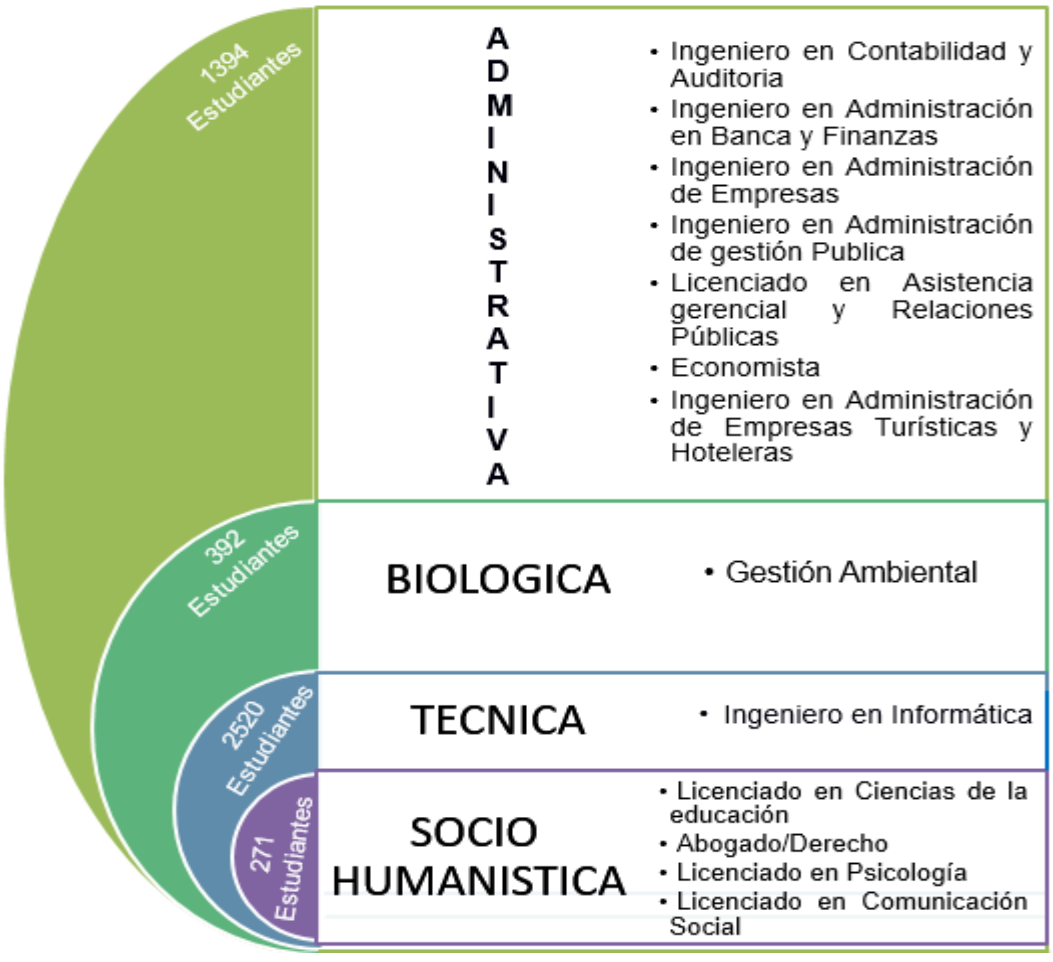


Ilustración 14. Data Seleccionada
Fuente. Elaboración Propia

3.2.1 Recolección de datos.

Para el desarrollo de este proyecto se han considerado datos nominales referentes a: edad, cedula, género y estado civil, de los estudiantes, desde la base de datos del Syllabus; y datos acerca de las actividades como chats, video conferencias, foros y quiz subidos por los docentes al EVA, y desarrolladas por los estudiantes, información que fue adquirida de la base de datos del entorno virtual de aprendizaje, tomando en cuenta que pertenezcan a los matriculados en todos o uno de los componentes de nivel básico de Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, pertenecientes a la modalidad abierta y a distancia dentro de los periodos oct-2014/feb-2015 y abr-2015/ago-2015.

A continuación tenemos la Tabla 5 con la descripción y los datos recolectados:

Tabla 5. Descripción de muestra

Dato	Descripción
Usuario	Cada usuario se identifica por un id, el mismo que considera el numeral 3 para referirse que el usuario es estudiante. Y se relaciona con el número de cedula con las tablas de categorías y cursos.
Estudiante	Cada estudiante se identifica por su número de cedula y puede relacionarse con una o varias categorías identificadas como componentes de la modalidad abierta y a distancia y con un curso o paralelo de la o las categorías previamente relacionado y a su vez con un periodo académico y actividades.
Titulación	Las titulaciones se identifican por el nombre con las que se ofertan en las 4 áreas (Administrativa, Biológica, Socio Humanística e Informática) de la UTPL dentro de la modalidad abierta y a distancia, relacionada con la cedula del estudiante.
Categorías	Cada categoría se identifica por un id y se encuentra relacionada con un curso.
Cursos	Los cursos se relacionan con la categoría por el id_categoría y cada curso pertenece a un paralelo diferente identificándose por su id_curso.
Plan académico	Contiene los periodos académicos identificados por un código y con el campo de pdoid se relaciona con el id de periodo y el campo categoría con el id de categorías.
Periodo	Cada periodo se identifica con un id y se relaciona con el id de la tabla mdl_course_categories.
Actividades	Cada actividad como chat, video conferencia, quiz o foro tiene algunas acciones y cada una de estas acciones se identifican por un id

Fuente. Elaboración propia

Tablas y documentos que contienen la data

Para la minería de datos en la Tabla 6 se indica la selección de campos de la tabla de la base de datos del EVA o documento resultante de la base de datos del Syllabus, al que corresponde cada uno.

Tabla 6. Muestra de atributos

ATRIBUTO	TABLA/DOCUMENTO
Cedula del estudiante	mdl_usuario /reporte syllabus
Genero del estudiante	mdl_usuario/reporte syllabus
Edad del estudiante	Reporte syllabus
Estado civil del estudiante	Reporte syllabus
Nivel académico del estudiante	Reporte syllabus
Titulación del estudiante	Reporte syllabus
Tipo de matrícula del estudiante	Reporte syllabus
Tipo del estudiante: nuevo, segundo en adelante, nuevo en la titulación	Reporte syllabus
Categoría	mdl_curse_categories
Curso	mdl_course
Periodo	mdl_syllabus_pdo
chat_historial	mdl_logs
chat_report	
chat_talk	
chat_viewAllChat	
chat_viewChat	
Elluminate_viewAllElluminate	
Elluminate_viewMeeting	
Elluminate_viewRecording	
Elluminate_viewElluminate	
addDiscussionForum	
deletePostForum	
subscribeForum	
subscribeallForum	
updatePostForum	
viewDiscussionForum	
viewForumForum	
viewForumsForum	
viewSubscribersForum	
Quiz_attempt	
Quiz_closeAttempt	
Quiz_evaluaciones	
Quiz_review1	
Quiz_view	
Quiz_viewAll	

Fuente. Elaboración propia

3.2.2 Descripción de los datos.

Para realizar la extracción de datos hemos trabajado con consultas en MySQL en la base de datos del EVA-UTPL, tomando de ella 8 tablas, las cuales se relacionaron por medio del atributo identidad de estudiantes con la data proporcionada por los administradores de la base de datos de SYLLABUS por falta de datos nominales en el resultado de las consultas de la base de datos del EVA.

En el **Anexo 1** encontraremos el diseño de un modelo entidad relación, de las tablas descritas en la Tabla 7, correspondientes a la base de datos del EVA.

Tabla 7. Nombre de las tablas de la BD del SYLLABUS

TABLA	DESCRIPCION
mdl_user_utpl	Contiene datos de los docentes como su cédula y el género al que pertenecen.
mdl_role_assignments	Contiene el rol al que pertenece cada usuario, en nuestro caso 3=DOCENTE
mdl_context	Es la tabla de unión entre el Docente y el Curso en donde contextlevel=50 para considerarse que es de tipo curso el contexto que se va relacionar.
mdl_course	Contiene todos los cursos de la UTPL.
mdl_syllabus_pdo	Contiene los periodos de estudio
mdl_syllabus_pln	Contiene los periodos y la categoría con la que se relacionan los cursos
mdl_course_categories	Contiene la categoría o nombre de las materias a las que corresponden los cursos.
mdl_log	Contiene las actividades que los usuario realizan en el EVA, tales como: chat, video conferencias, quiz y foros.

Fuente. Elaboración propia

Datos iniciales

Se ha modificado el formato de edad por medio de la formula **=SIFECHA()**, pasando de date a numérico utilizando la herramienta de office llamada Excel.

Dentro de la misma herramienta se limpiaron caracteres especiales como tildes y comas.

Para la presentación integral de todos los atributos se utilizó la herramienta de xamp, levantando una base de datos local, para poder relacionar la data obtenida de la BD del EVA junto con la del reporte brindado por el administrador de la BD del syllabus, vinculando el

número de identificación o cedula entre las dos BD y acoplando vistas para la presentación final de la información.

Con la nueva data obtenida como resultado de la vinculación de las dos BD se ha generado dos tipos de datos: Varchar y Numérico, en el **Anexo 2** Podremos observar a que variable corresponde cada uno.

3.2.3 Exploración de los datos.

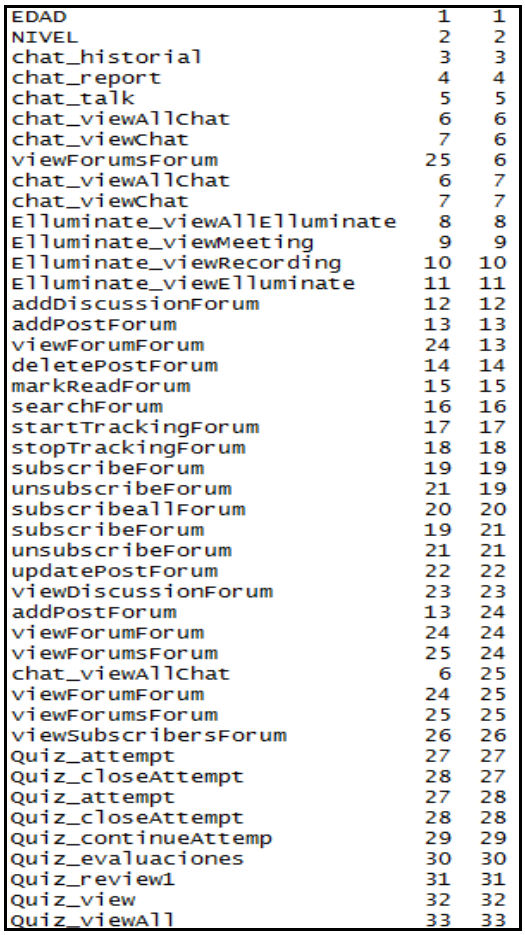
En la extracción de datos por medio de las consultas realizadas con MySQL en la base de datos del EVA, tenemos actividades de chats, foros, video llamadas y quiz; cada una de ellas presentan diferentes acciones, por ejemplo: la actividad de chat tiene acciones de chat_historial, chat_report, chat_talk.

Para explorar la data se aplica un análisis de correlación entre las acciones de estas actividades, la edad y nivel académico Ilustración 15, dentro de la herramienta R.

Numeral	Campo
1	edad
2	nivel
3	chat_historial
4	chat_report
5	chat_talk
6	chat_viewAllChat
7	chat_viewChat
8	Elluminate_viewAllElluminate
9	Elluminate_viewMeeting
10	Elluminate_viewRecording
11	Elluminate_viewElluminate
12	addDiscussionForum
13	addPostForum
14	deletePostForum
15	markReadForum
16	searchForum
17	startTrackingForum
18	stopTrackingForum
19	subscribeForum
20	subscribeallForum
21	unsubscribeForum
22	unsubscribeAllForum
23	updatePostForum
24	viewDiscussionForum
25	viewForumForum
26	viewForumsForum
27	viewSubscribersForum
28	Quiz_attempt
29	Quiz_closeAttempt
30	Quiz_continueAttemp
31	Quiz_evaluaciones
32	Quiz_review1
33	Quiz_view
34	Quiz_viewAll

Ilustración 15. Datos a correlacionar
Fuente. Elaboración propia

Como se aprecia en la Ilustración 15, se ha numerado cada dato para una adecuada interpretación del resultado presente en la Ilustración 16.



EDAD	1	1
NIVEL	2	2
chat_historial	3	3
chat_report	4	4
chat_talk	5	5
chat_viewAllChat	6	6
chat_viewChat	7	6
viewForumsForum	25	6
chat_viewAllChat	6	7
chat_viewChat	7	7
Elluminate_viewAllElluminate	8	8
Elluminate_viewMeeting	9	9
Elluminate_viewRecording	10	10
Elluminate_viewElluminate	11	11
addDiscussionForum	12	12
addPostForum	13	13
viewForumForum	24	13
deletePostForum	14	14
markReadForum	15	15
searchForum	16	16
startTrackingForum	17	17
stopTrackingForum	18	18
subscribeForum	19	19
unsubscribeForum	21	19
subscribeallForum	20	20
subscribeForum	19	21
unsubscribeForum	21	21
updatePostForum	22	22
viewDiscussionForum	23	23
addPostForum	13	24
viewForumForum	24	24
viewForumsForum	25	24
chat_viewAllChat	6	25
viewForumForum	24	25
viewForumsForum	25	25
viewSubscribersForum	26	26
Quiz_attempt	27	27
Quiz_closeAttempt	28	27
Quiz_attempt	27	28
Quiz_closeAttempt	28	28
Quiz_continueAttemp	29	29
Quiz_evaluaciones	30	30
Quiz_review1	31	31
Quiz_view	32	32
Quiz_viewAll	33	33

Ilustración 16. Correlación de campos
Fuente. Elaboración propia

Finalmente, tomando en cuenta el numeral asignado en cada dato de la Ilustración 15 el resultado del análisis de correlación expuesto en la Ilustración 16, presenta a 10 datos como mejor correlacionados, siendo estos: viewForumForum, updatePostForum, unsubscribeForum, viewDiscussionForum, Quiz_attempt, chat_viewAllChat, addPostForum, subscribeForum, unsubscribeAllForum, viewSubscribersForum.

La Tabla 8 presenta a estos datos a través de tres columnas en donde las dos primeras visualizan los datos en relación y en la tercera el numero con que fue asignado cada dato en la Ilustración 15, para identificar a cada uno según su número.

Tabla 8. Resultado de análisis de correlación

Dato 1	Dato 2	Relaciones	
		Dato 1	Dato 2
ViewForumForum	chat_viewAllChat	25	6
updatePostForum	addPostForum	24	13
unsubscribeForum	subscribeForum	21	19
viewDiscussionForum	unsubscribeAllForum	24	22
Quiz_attempt	viewSubscribersForum	28	27

Fuente. Elaboración propia

Reglas de asociación

Se aplicó reglas de asociación para evaluar la calidad de datos y con ello un mejor acierto de los resultados.

La Ilustración 17 indica como obtuvimos los valores lógicos de la regla de asociación, misma que fue aplicada a cada una de las variables consideradas como las acciones de las actividades que los estudiantes realizan en el EVA.

```
36  cl <- 5
37  table(data02[,cl])
38  (idx <- which(data02[,cl]>59))
39  data02[idx,cl] = 0
40
```

Ilustración 17. Regla de asociación
Fuente. Elaboración propia

Como ejemplo de los resultados tenemos las acciones de: chat_historial y chat_report, en la Ilustración 18.

```
> cl <- 3 chat_historial
> table(data02[,cl])
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21   26   27   36   37   64
8881 1205  494  280  149   81   43   27   17   16   12   10    8    6    5    2    4    2    3    1    3    1    1    2    1    1    1

> cl <- 4  chat_report
> table(data02[,cl])
    0     1     2     3     4     5     6     7     9    10    11    12    14
10902   228    58    36    14     3     4     2     2     2     2     2     1
```

Cantidad de acciones
Cantidad de estudiantes

Ilustración 18. Resultado de reglas de asociación
Fuente. Elaboración propia

Descripción de los resultados

chat_historial: es el historial de chat de un usuario, en donde encontramos un máximo de 64 actividades sobre la acción que es realizada por 1 solo estudiante y el mínimo de 0 actividades realizada por 8881 estudiantes, siendo de esta manera una variable de muy poca prioridad para nuestro posterior análisis.

chat_report: reportes sobre un chat obtenidos por el usuario, podemos ver que máximo se han realizado petición de 14 reportes y esto ha sido por un usuario y que su mínimo de peticiones que es de 0 ha sido por 10902 estudiantes, es decir, que esta acción al igual que la primera no es muy realizada por los estudiantes.

3.2.4 Verificación de la calidad de los datos.

La data obtenida en la exploración presento ciertos valores sin concepto como: nulos, inconsistentes y datos duplicados, presentes en las variables de Edad y Género.

3.3 Fase 3. Preparación de los datos.

En esta fase presentaremos los datos limpios, listos para ser ingresados a la herramienta de minería de datos seleccionada denominada R y aplicar las técnicas que nos ayuden al análisis de la data y con ello a una toma de decisiones.

3.3.1 Selección de datos.

La data construida constituye información personal y académica (actividades del EVA) de estudiantes que tomaron materias de formación básica (Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio), en cualquier ciclo de la modalidad abierta y distancia de la UTPL comprendida en los periodos oct-2014/feb-2015 y abr-2015/ago-2015.

Los campos y atributos seleccionados de la data se presentan en la Tabla 9.

Tabla 9. Campos y atributos seleccionados

CAMPO	ATRIBUTO
mdl_user_utpl	Id
mdl_role_assignments	contexid, userid
mdl_context	
Reporte syllabus	Genero
	Edad
	Estado civil
	Titulación
	Tipo de matricula
	Tipo de estudiante
mdl_course	id, category
mdl_syllabus_pdo	pdoid
mdl_syllabus_pln	pdoid
mdl_course_categories	Id
mdl_role_assigments	contexid, userid
mdl_log	userid, course, action

Fuente. Elaboración propia

3.3.2 Limpieza de datos.

Al verificar la data obtenida como resultado de las consultas realizadas en la base de datos del EVA y del reporte de la base de datos del SYLLABUS se encontraron algunas inconsistencias, como: campos vacíos, valores faltantes, caracteres especiales, datos duplicados, datos agrupados, mismas que fueron corregidas para poder trabajar de manera óptima en el análisis que resultara de aplicar la minería de datos.

La Tabla 10 muestra que campos presentaron estas inconsistencias y como fueron resueltas.

Tabla 10. Limpieza de datos

Campos	Corrección
Identidad	Algunos presentaban el ruc no el número de cedula, así que se eliminó el 001 correspondiente al RUC para trabajar con los dígitos correspondiente al número de cedula
Genero	Teníamos dos columnas de género una de la base de datos del EVA que describían Masculino y Femenino por sus siglas (M y F) y otra del reporte la base de datos del Syllabus que presentaban los dos géneros con toda la palabra, así nos quedamos con el campo genero del EVA
Edad	Este campo se presentaba con el nombre de fecha de nacimiento y con formato date, cambiamos el nombre del campo por edad y la presentación de la fecha de nacimiento aplicando la fórmula para calcular la edad y obtener un dato convertido a int.
Estado civil	Encontramos algunos campos vacíos, al ser un número máximo de 10 y contar con suficiente data eliminamos la fila
Categoría	Encontramos caracteres especiales como tildes, identificamos las palabras tildadas y las reemplazamos por no tildadas
Curso	Encontramos caracteres especiales como tildes y comas identificamos las palabras tildadas y las reemplazamos por no tildadas y aplicamos la fórmula de “=MAYUSC()“ para tener toda la información en mayúscula.
Actividad	Se aplicó un count en las actividades que los estudiantes realizaron para obtener información numérica y se cambió la dirección de la presentación de vertical a horizontal.

Fuente. Elaboración propia

3.3.3 Construcción e Integración de datos.

Para tener el documento final con extensión .csv con el que se trabaja en R, se han relacionado e integrado los datos de las bases de datos EVA y SYLLABUS a través del campo cedula.

A continuación el **Anexo 3** presenta los campos, descripción y tipo de dato con lo que se caracteriza la una nueva data, que resulto después del procedimiento aplicado en la fase anterior, siendo estos campos los definitivos de la data para aplicar la minería de datos.

Como ya se mencionó anteriormente los datos no fueron obtenidos de una sola base de datos, se trabajó con dos bases EVA y Syllabus, y la data resultante se la integró dentro de Excel, una herramienta de office la misma que se visualiza en la Ilustración 19.

IDENTIDAD	GENI	EDAD	ESTADO	NIVE	TITULACI	TIPOMATR	TIPO_ESTUDI	CURSO	PERIODO		chat_his	chat_req	chat_tal	chat_vi	chat_viev	Elluminat	Ellumir	Ellumi	Ellumin	addDis	subscril	subscrit	update	viewDi	view	viewF	view	Quiz_at	Quiz_cl	Quiz_r	Quiz_r	Quiz_vi	Quiz_vi
1727270538	M	23	Soltero	9	GESTION	Ordinaria	NUEVO EN L	REALIDAD NACIONAL Y AMBIEN	Abr2015	Ago2015	0	0	0	0	4	0	0	0	0	0	0	0	0	0	6	1	0	1	1	1	0	0	0
1727421834	F	23	Casado	1	INGENIEI	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO J2	Abr2015	Ago2015	2	0	0	0	0	0	0	0	8	0	0	0	0	0	11	1	0	3	3	1	0	0	0
1727464305	F	21	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO N4	Abr2015	Ago2015	0	0	17	6	9	0	0	0	0	0	0	0	0	0	10	7	0	3	3	1	1	0	0
1727511378	F	21	Soltero	1	DERECHC	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO F3	Abr2015	Ago2015	1	0	9	8	0	0	1	0	0	0	0	0	1	1	12	8	0	4	4	1	0	10	0
1727563270	F	19	Soltero	1	ECONOM	Ordinaria	SEGUNDA EN	REALIDAD NACIONAL Y AMBIEN	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	1	0	7	0	0	2	2	1	1	3	0
1727604884	M	22	Soltero	1	GESTION	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO O2	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	5	5	0	0	8	0
1727980490	F	22	Soltero	1	CIENCIAS	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO F4	Abr2015	Ago2015	1	0	11	7	29	0	8	0	0	0	0	0	0	0	13	0	0	4	4	0	1	0	0
1728012863	F	19	Soltero	1	DERECHC	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO S3	Abr2015	Ago2015	3	0	15	5	0	0	8	1	18	0	0	0	0	0	21	2	0	3	3	9	3	0	0
1728117225	F	22	Soltero	1	DERECHC	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO E3	Abr2015	Ago2015	1	0	4	8	31	0	14	2	0	0	1	0	0	0	20	6	0	5	5	0	0	0	0
1750009191	M	24	Soltero	1	INGENIEI	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO J2	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	0	4	0
1750020792	F	22	Soltero	1	PSICOLO	Ordinaria	NUEVO	REALIDAD NACIONAL Y AMBIEN	Abr2015	Ago2015	0	0	7	0	0	0	8	0	0	0	0	0	1	0	8	0	0	4	4	1	0	10	0
1750242958	F	22	Soltero	2	INFORM/	Extraordin	NUEVO EN L	EXPRESION ORAL Y ESCRITA P4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
1750245399	F	21	Soltero	1	GESTION	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO P2	Abr2015	Ago2015	6	2	20	5	0	0	5	4	28	0	1	0	1	0	31	12	0	3	3	1	7	0	0
1750257568	F	25	Soltero	1	PSICOLO	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO U4	Abr2015	Ago2015	2	0	0	8	0	0	1	0	0	0	1	0	0	0	9	11	0	6	6	0	0	16	0
1750273870	M	20	Soltero	2	INFORM/	Extraordin	SEGUNDA EN	REALIDAD NACIONAL Y AMBIEN	Abr2015	Ago2015	0	0	7	0	0	0	0	0	0	0	0	0	0	0	4	0	0	3	3	1	1	0	0
1750273870	M	20	Soltero	2	INFORM/	Extraordin	SEGUNDA EN	METODOLOGIA DE ESTUDIO H3	Oct2014	Feb2015	0	0	0	0	0	0	0	2	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1750284141	F	23	Soltero	1	CIENCIAS	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO B5	Abr2015	Ago2015	0	0	17	0	0	0	3	0	0	0	0	0	0	0	2	1	0	3	3	0	0	6	0
1750344275	F	24	Soltero	1	DERECHC	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO E3	Abr2015	Ago2015	0	0	0	2	0	0	0	1	0	0	0	0	0	0	5	0	0	5	4	0	0	14	0
1750393801	M	23	Soltero	1	DERECHC	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO O3	Abr2015	Ago2015	0	0	0	3	4	0	0	0	0	0	3	0	4	0	13	12	0	1	1	0	0	0	0
1750395798	F	20	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO N4	Abr2015	Ago2015	0	0	0	4	0	0	2	0	4	0	0	0	1	0	12	9	0	2	2	0	1	0	0
1750400937	F	22	Soltero	1	PSICOLO	Ordinaria	NUEVO EN L	METODOLOGIA DE ESTUDIO L4	Abr2015	Ago2015	0	0	8	0	0	0	0	0	0	0	0	0	1	1	13	0	0	3	3	1	0	6	0
1750405563	F	23	Soltero	2	DERECHC	Ordinaria	NUEVO EN L	METODOLOGIA DE ESTUDIO D3	Abr2015	Ago2015	0	0	0	0	0	0	7	0	2	0	0	0	0	0	5	0	0	1	1	0	0	0	0
1750423228	F	20	Soltero	7	ASISTEN	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO V1	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	1	0	0	2	0
1750429340	M	23	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO N4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	0	5	0
1750437772	M	22	Soltero	4	CIENCIAS	Extraordin	NUEVO EN L	METODOLOGIA DE ESTUDIO I4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	2	0
1750438879	F	21	Soltero	1	INGENIEI	Ordinaria	NUEVO	EXPRESION ORAL Y ESCRITA A2	Abr2015	Ago2015	1	1	30	3	21	0	0	0	0	0	2	0	0	0	7	8	0	3	3	0	3	0	0
1750665471	F	20	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO L4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	1	0	5	0
1750762047	M	23	Soltero	2	INGENIEI	Extraordin	SEGUNDA EN	REALIDAD NACIONAL Y AMBIEN	Abr2015	Ago2015	1	0	0	0	5	0	0	0	0	0	0	0	0	0	9	0	0	2	2	1	1	0	0
1750765693	F	21	Soltero	1	INFORM/	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO V4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	7	0	3	3	0	0	5	0
1750772517	F	19	Soltero	1	INGENIEI	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO I2	Abr2015	Ago2015	1	0	21	11	15	0	2	0	0	0	0	0	0	1	20	17	0	5	5	2	0	0	0
1750852103	F	21	Soltero	2	INGENIEI	Extraordin	SEGUNDA EN	METODOLOGIA DE ESTUDIO F2	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0
1750875906	M	21	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO N4	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	0	4	0
1750880377	M	21	Soltero	1	DERECHC	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO K3	Abr2015	Ago2015	0	0	9	0	0	5	0	0	0	0	0	0	0	0	9	17	0	3	3	1	5	6	0
1750961391	M	39	Casado	1	DERECHC	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO T3	Abr2015	Ago2015	0	1	0	8	18	0	1	0	0	0	0	0	0	0	16	2	0	3	3	2	0	0	0
1751028695	M	29	Soltero	1	COMUNI	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO V3	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	2	0
1751029586	F	22	Soltero	1	INGENIEI	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO D2	Abr2015	Ago2015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	5	5	0
1751040013	F	20	Soltero	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO N4	Abr2015	Ago2015	0	0	19	5	8	0	0	0	0	0	0	0	0	0	17	8	0	3	3	0	0	0	0
1751068931	F	23	Union Li	1	PSICOLO	Ordinaria	NUEVO	METODOLOGIA DE ESTUDIO M4	Abr2015	Ago2015	0	12	24	5	0	0	0	0	0	0	0	0	0	0	0	0	0	3	3	0	1	12	0
1751091057	F	21	Soltero	1	INGENIEI	Extraordin	NUEVO	METODOLOGIA DE ESTUDIO B2	Abr2015	Ago2015	1	0	0	0	0	0	0	0	0	0	0	0	0	0	21	13	0	3	3	1	0	5	5
1751107242	F	23	Union Li	1	CIENCIAS	Especial	NUEVO	EXPRESION ORAL Y ESCRITA M3	Abr2015	Ago2015	0	0	0	0	0	1	0	0	0	0	0	0	1	1	11	2	0	3	3	0	0	4	0

Ilustración 19. Data final
Fuente. Elaboración propia

3.4 Fase 4. Modelo.

En esta fase se procede a elegir las técnicas de minería de datos aplicables para el planteamiento de la solución del problema, trabajando con los datos de los campos definidos en la sección 0, teniendo en cuenta, dos posibles técnicas, la primera técnica es **k-means**, con la que a simple vista es favorable trabajar los datos, pero de no obtener resultados favorables, se aplicara una segunda técnica que es la de **análisis de componentes principales**.

3.4.1 K-means.

1. Selección de técnica de modelado.

Para aplicar técnicas de minería de datos a la data final, se ha considerado manejar la herramienta R, la cual se describe en la sección 2.2, este software/herramienta proporciona facilidad para el desarrollo de los informes finales, a través de las capacidades gráficas, manejo y gestión de los datos que contiene, siendo muy útil para la deducción de resultados.

Para el desarrollo del proyecto se cuenta con datos que a simple vista proporcionan características para clusterizar, es por ello que se considera la técnica de clúster, como la más indicada para realizar el análisis. Se aplica el algoritmo **k-means** dentro de la herramienta **R** con una validación cruzada de métodos de este algoritmo los cuales son: Hartigan-Wong, Lloyd, Forgy y MacQueen, y finalmente se utiliza la sentencia **plot** para graficar los resultados.

2. Optimización de la técnica de agrupación k-means.

Para la construcción del modelo descriptivo de la labor tutorial del docente MAD, se ha creído conveniente realizar una optimización de métodos a través de una validación cruzada.

Como primer paso tenemos la filtración de la data numérica; requisito principal, solicitado por el algoritmo kmeans y se considera solo los campos de las acciones que realizan los estudiantes dentro del EVA como lo indica la Ilustración 20.

```
setwd("C:/Users/mineria/Google Drive/tesis2016 -MINERIA/R")

dataFile <- "Datos 30.09.16.csv"
data <- read.table(file=dataFile, header=T, sep=";", stringsAsFactors=FALSE);
str(data)

colsToRemove <- which( colnames(data) %in% c("IDENTIDAD","CATEGORIA", "GENERO","EDAD","NIVEL",
                                             "ESTADO_CIVIL", "TITULACION","TIPOMATRICULA",
                                             "TIPO_ESTUDIANTE","CURSO","PERIODO"))

datos <- data[,-colsToRemove]
dim(datos)

str(datos)
```

Ilustración 20. Selección de campos y data numérica
Fuente. Elaboración propia

Posteriormente, se determina el número de clusters usados para la validación del método k-means, en cuanto a los cuatro algoritmos con los que cuenta, son: Hartigan-Wong, Lloyd, Forgy y MacQueen.

Para llegar a la validación, se aplican sentencias que comprenden 30 ejecuciones de cada método con 100 iteraciones por ejecución como se muestra en la Ilustración 21.

```
InerciaIC.Hartigan = rep(0, 30)
InerciaIC.Lloyd = rep(0, 30)
InerciaIC.Forgy = rep(0, 30)
InerciaIC.MacQueen = rep(0, 30)
for (k in 1:30) {
  grupos = kmeans(datos, k, nstart=1, iter.max = 100, algorithm = "Hartigan-Wong")
  InerciaIC.Hartigan[k] = grupos$tot.withinss

  grupos = kmeans(datos, k, nstart=1, iter.max = 100, algorithm = "Lloyd")
  InerciaIC.Lloyd[k] = grupos$tot.withinss

  grupos = kmeans(datos, k, nstart=1, iter.max = 100, algorithm = "Forgy")
  InerciaIC.Forgy[k] = grupos$tot.withinss

  grupos = kmeans(datos, k, nstart=1, iter.max = 100, algorithm = "MacQueen")
  InerciaIC.MacQueen[k] = grupos$tot.withinss
}

plot(InerciaIC.Hartigan, col = "blue", type = "b", main = "CODO DE JAMBU", xlab = "Número de inercias")
points(InerciaIC.Lloyd, col = "red", type = "b")
points(InerciaIC.Forgy, col = "green", type = "b")
points(InerciaIC.MacQueen, col = "magenta", type = "b")
legend("topright", legend = c("Hartigan", "Lloyd", "Forgy", "MacQueen"),
              col = c("blue", "red", "green", "magenta"), lty = 1, lwd = 1)
```

Ilustración 21. Inicio para la validación de k-means
Fuente. Elaboración propia

Estas sentencias finalmente retornan una gráfica, denominada codo de jambu presente en la Ilustración 22, la cual indica una estabilidad de los cuatro métodos, desde la ejecución 4; valor que será asignado a la variable que represente el número de clusters, aplicado en la optimización de los métodos para el algoritmo kmeans.

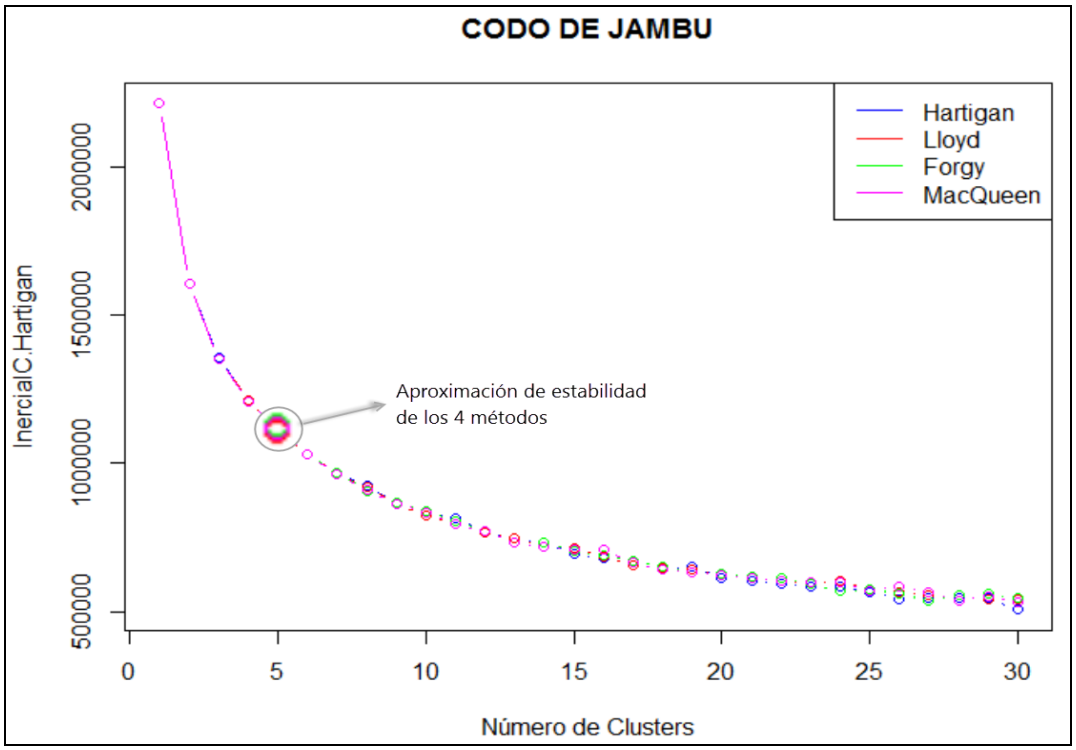


Ilustración 22. Codo de JAMBU, estabilidad de métodos
Fuente. Elaboración propia

Una vez obtenido el número de los clusters, procedemos a realizar la optimización del método, en relación a los datos con los que se cuenta, en base al resultado de la inercia inter-clases más alta, es decir, la mayor distancia entre el número de clusters; porque

cuando la inercia-interclases es más alta es en donde se presenta mejor clusterización, como menciona (Rodriguez, 2014)

Entonces para los datos del proyecto, el método de kmeans para clusterizar con **clusters** de **4**, es el mejor algoritmo es **Hartigan** con una inercia inter-clases de **1003249** sobre los otros métodos, presente en la Ilustración 23

```
> Hartigan <- 0
> Lloyd <- 0
> Forgy <- 0
> MacQueen <- 0
> for (i in 1:50) {
+   grupos <- kmeans(datos, 4, iter.max = 100, algorithm = "Hartigan-Wong")
+   Hartigan <- Hartigan + grupos$betweenss
+   grupos <- kmeans(datos, 4, iter.max = 100, algorithm = "Lloyd")
+   Lloyd <- Lloyd + grupos$betweenss
+   grupos <- kmeans(datos, 4, iter.max = 100, algorithm = "Forgy")
+   Forgy <- Forgy + grupos$betweenss
+   grupos <- kmeans(datos, 4, iter.max = 100, algorithm = "MacQueen")
+   MacQueen <- MacQueen + grupos$betweenss
+ }
>
> Hartigan/50
[1] 1003249
> Lloyd/50
[1] 1002673
> Forgy/50
[1] 1002964
> MacQueen/50
[1] 1003064
>
```

Ilustración 23. Inercias inter-clases
Fuente. Elaboración propia

El algoritmo de Hartigan con el que hemos optimizado el método kmeans, según (Maitra & Ramler, 2010) tiene como objetivo, proporcionar una implementación simple y eficiente de kmean, cuando la distancia métrica para la agrupación es euclidiana, se puede utilizar en áreas informáticas tan solo cuando dos vectores apunten a un mismo lugar como lo hace el coseno similitud o en una correlación en donde se identifiquen características similares, que es el caso que pretende identificar en nuestra data.

3. Descripción del modelado con técnicas k-means.

Una vez obtenido el número de clusters y el algoritmo aplicable al método kmeans según los datos del proyecto, procedemos a trabajar con ellos en lo que es la clusterización, agrupando los datos de la matriz asignada según características similares de los mismos, en un número de 4, clusters, pretendiendo obtener un grupo con la variación más pequeña de datos dentro de un cluster en relación con otro cluster, es decir se asigna al azar cada punto de datos a un cluster o grupo, se calcula la media dentro del cluster, y finalmente cada punto de datos se asigna iterativa a su centroide más cercano, para la reducción mínima de la variación dentro del cluster, hasta descartar diferencias importantes.

Al igual que en el enunciado 2 de esta sección, se inicia llamando la data, estableciendo la matriz con la filtración de valores numéricos, recuperando la dimensión de los datos con **dim**, aplicando un **str** para visualizar la data que se necesita sea leída y continuamos con las sentencias que nos permiten aplicar kmeans como lo muestra la Ilustración 24.

```
> setwd("C:/Users/mineria/Google Drive/tesis2016 -MINERIA/R")
> library(cluster)
> library(fpc)
>
> dataFile <- "Datos 30.09.16.csv"
> data <- read.table(file=dataFile, header=T, sep=";", stringsAsFactors=FALSE);
>
> colsToRemove <- which( colnames(data) %in% c("IDENTIDAD","CATEGORIA", "GENERO","EDAD","NIVEL",
+                                              "ESTADO_CIVIL", "TITULACION","TIPOMATRICULA",
+     I                                        "TIPO_ESTUDIANTE","CURSO","PERIODO"))
>
> data01 <- data[,-colsToRemove]
> dim(data01)
[1] 4551   23
> str(data01)
'data.frame':   4551 obs. of  23 variables:
 $ chat_historial              : int  0 1 1 0 0 0 0 2 1 0 ...
 $ chat_report                 : int  0 0 1 0 0 0 0 1 0 0 ...
 $ chat_talk                   : int  0 24 8 0 0 10 13 2 5 18 ...
 $ chat_viewAllChat            : int  0 0 2 0 0 0 0 8 0 0 ...
 $ chat_viewChat               : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Elluminate_viewAllElluminate: int  0 0 0 0 0 0 11 0 3 0 ...
 $ Elluminate_viewMeeting      : int  3 13 5 0 0 13 7 0 15 6 ...
 $ Elluminate_viewRecording    : int  0 0 1 0 0 0 0 0 0 0 ...
 $ Elluminate_viewElluminate   : int  0 17 19 0 0 24 24 0 0 17 ...
 $ addDiscussionForum          : int  0 0 0 0 0 0 0 0 0 0 ...
 $ subscribeForum              : int  0 0 0 0 0 2 0 0 0 0 ...
 $ subscribeallForum           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ updatePostForum             : int  0 0 0 0 0 0 0 0 0 1 ...
 $ viewDiscussionForum         : int  0 4 0 0 0 0 0 0 0 0 ...
 $ viewForumForum              : int  2 35 5 0 0 7 20 4 12 16 ...
 $ viewForumsForum             : int  0 1 3 0 0 5 12 3 3 0 ...
 $ viewSubscribersForum        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Quiz_attempt                : int  4 3 3 3 1 3 7 3 4 5 ...
 $ Quiz_closeAttempt           : int  4 3 3 3 0 3 7 3 4 5 ...
 $ Quiz_evaluaciones           : int  2 1 0 0 0 0 0 1 1 0 ...
 $ Quiz_review1                : int  0 3 4 0 0 1 5 2 0 5 ...
 $ Quiz_view                   : int  7 0 0 6 2 0 0 15 31 0 ...
 $ Quiz_viewAll                : int  0 1 0 0 0 0 0 0 0 0 ...
>
> km<-kmeans(data01, 4, iter.max=100, algorithm="Hartigan")
>
```

Ilustración 24. Aplicación de k-means con algoritmo Hartigan y 4K
Fuente. Elaboración propia

Ahora procedemos a graficar resultados, y para ello se utilizó la sentencia **plot**, la cual es una función genérica que nos ayuda a realizar graficas de dispersión, series de tiempo o funciones, entre otras (Kelmansky, 2006), en este proyecto se la utilizara como una gráfica de dispersión y presenta la siguiente estructura Ilustración 25:

2. Matriz o nombre de la data
3. Atributo de **cluster** dentro de la variable aplicada el método kmeans con el algoritmo Hartigan.
4. **main** que corresponde al encabezado de la gráfica
5. **pch** que es la representación los datos en la gráfica que varía de 1 a 25 en diferentes figuras nosotros aplicamos 21 que corresponde a puntos.

6. **cex** en donde se define el tamaño de la figura seleccionada por pch.

```
km<-kmeans(data01, 4, iter.max=100, algorithm="Hartigan")
plot(data01,col=km$cluster,
     main="kmeans resuelto con 4 clusters y 100 iteraciones",
     pch=21, cex=1)
```

Ilustración 25. Sentencia plot
Fuente. Elaboración propia

Esto dará como resultado un lote de clusterizacion, sobre las acciones en las los estudiantes participan dentro del EVA. **Anexo 4**

Tomando en cuenta que contamos con 23 acciones del estudiante, tras un análisis exploratorio, se buscó relaciones que pudieran brindar una adecuada clusterizacion, de las que se tomó como referencia a dos indicadas en la Tabla 11, para indicar que este cluster de número 4 según como recomendó el análisis del codo jambu no proporciono la inercia-interclases adecuada, dicho en otras palabras no fue la mejor clusterización.

Tabla 11. Cluster de 4

cluster=4		
RELACION	**ACCION**	**Num. ACCION**
1	elluminate_viewElluminate	9
	viewForumForum	15
2	Chatviewchat	5
	quiz_review1	21

Fuente. Elaboración propia

Por ello se realizó otra prueba aplicando un clúster de 3, para buscar un mejor resultado **Anexo 5**, igualmente que el clúster anterior se determinó 2 de las mejores Tabla 12, pero no se visualizó ninguna mejora.

Tabla 12. Cluster de 3

cluster=3		
RELACION	**ACCION**	**Num. ACCION**
1	Chat_talk	3
	chat_viewAllChat	4
2	Chat_viewChat	5
	elluminate_viewRecord	8

Fuente. Elaboración Propia.

Determinando finalmente que la función **plot** nos permite hacer un análisis exploratorio en donde demuestra que kmeans no es el método más indicado para realizar una clusterización de la data establecida.

3.4.2 Análisis de componentes principales.

Se aplica funciones como **clustplot** y **plotcluster** para comprobar la distribución que kmeans provoca en los datos.

De donde clustplot, de 3 clusters, presente en la Ilustración 26, permite visualizar una gráfica con etiquetas denominadas como component 1 y component 2 en el eje de las abscisas (x) y de las ordenadas (y), con la agrupación de la data a un solo sentido, aproximado al origen.

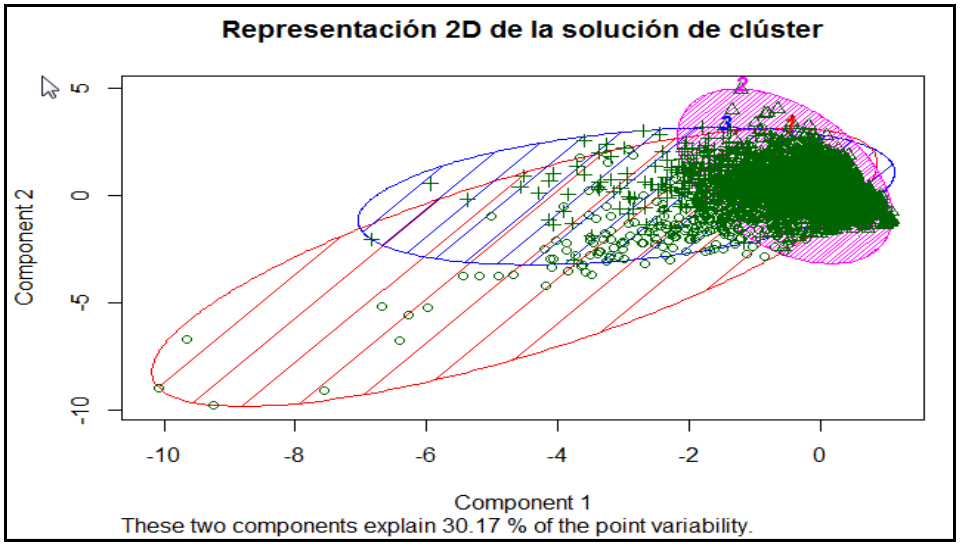


Ilustración 26. Análisis de componentes en 2D, clustplot de 3 clusters.
Fuente. Elaboración propia

Posteriormente se aplica plotcluster, con 3 cluster Ilustración 27 y 4 cluster Ilustración 28, cuyo resultado indica que la data tiende a agruparse a un solo sentido con aproximación a 0, igual que se demuestra con los resultados de la función clustplot

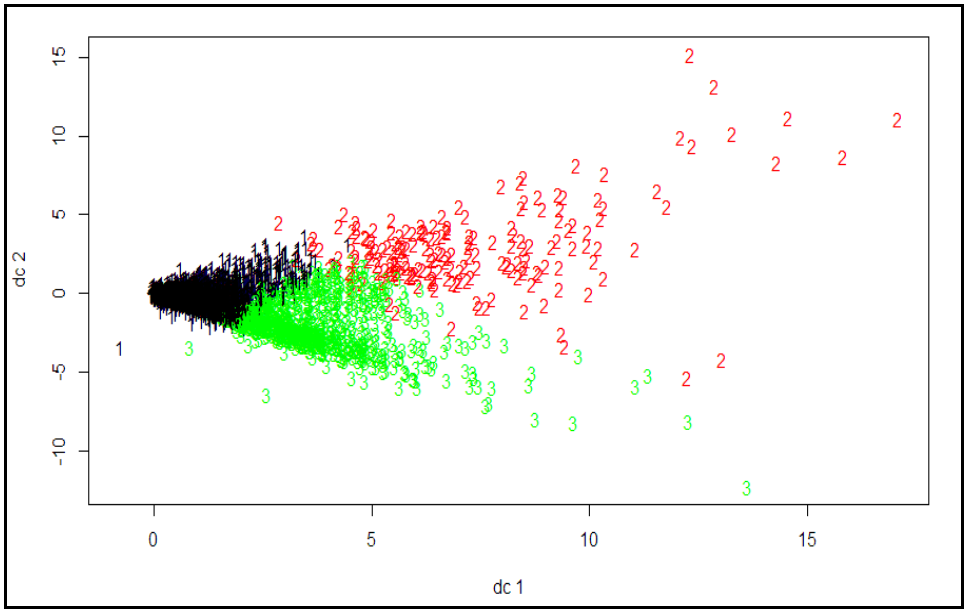


Ilustración 27. plotcluster de 3 clusters
Fuente. Elaboración propia.

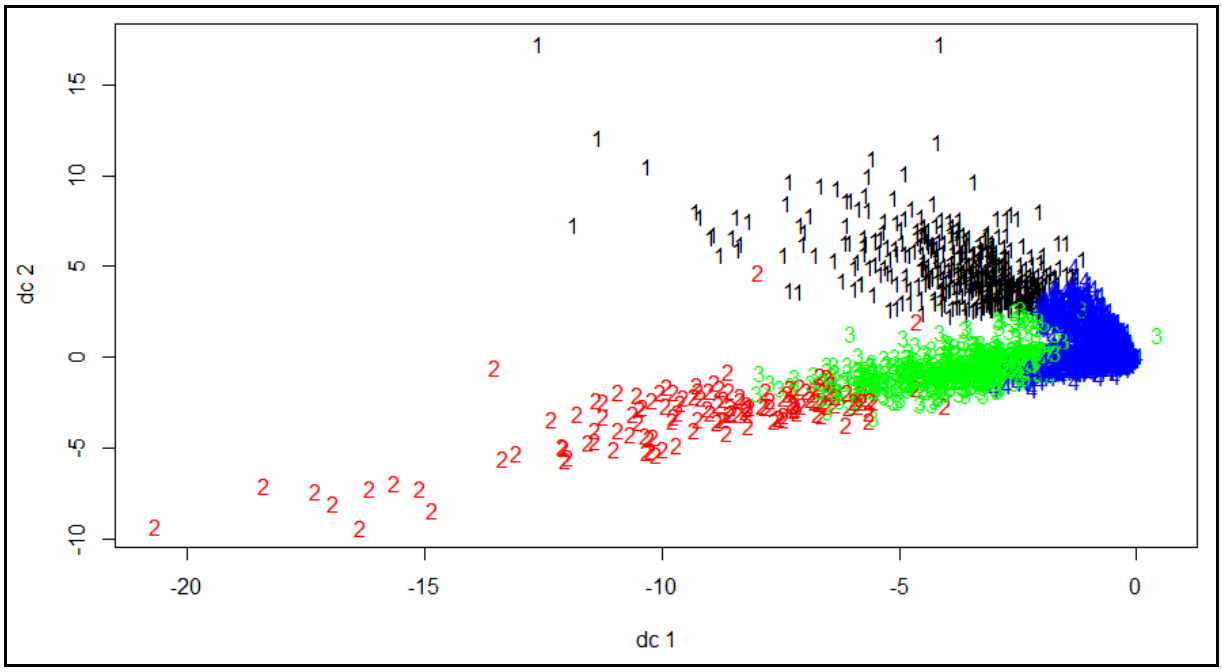


Ilustración 28. plotcluster de 4 clusters
Fuente. Elaboración propia

(de la Fuente, 2011) indica que el análisis de componentes principales, es una técnica matemática que no requiere la suposición de normalidad multivalente de los datos, y es utilizada para estudiar las relaciones que se presentan entre un determinado número de variables correlacionadas, es decir, variables que cuentan con información en común, para transformar el conjunto original de variables en un conjunto pequeño de nuevas variables, que no posean redundancia en la información, es así como las nuevas variables son combinaciones lineales de las anteriores y se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra.

Basándose en la matriz de covarianza, utilizada cuando los datos son dimensionalmente homogéneos y presentan valores medios similares, como es el caso que podemos observar en las ilustraciones anteriores de las funciones plot, plotcluster y clustplot, iniciamos el ACP (análisis de componentes principales).

Previo a la llamada de la base de trabajo que consta de 23 acciones y 4551 datos, se presenta la matriz de correlación con la sentencia **cor(). Anexo 6**, considerando la mejor información posible de las 23 acciones, se pretende obtener el componente principal, y para ello realizamos una estandarización de los datos, centrando y dividiendo para la desviación estándar, almacenando la mejor compensación de resultados.

1. **Construcción del modelo con técnica de ACP.**

Vamos a obtener un índice de referencia que nos ayude a medir el desempeño de los estudiantes en las acciones que participan dentro del EVA, tomando en cuenta según el tipo

de data con la que se trabaje es preciso aplicar matriz de correlación o matriz de covarianza; si contamos con datos que presenten características de escalas o unidades en distinta medida se deberá aplicar una matriz de correlación y de esta manera estandarizaríamos los datos, caso contrario si los datos son presentados en una misma medida se aplica la matriz de covarianza y seguimos con el procedimiento de análisis de componentes principales.

Teniendo en cuenta lo antes mencionado procedemos a determinar los componentes principales para el respectivo análisis.

1. Se ejecuta el comando **prcomp** para realizar el análisis de componentes principales, de la matriz de datos denominada **datos**:
    - acp<-prcomp(datos)

2. Se imprime el resultado de comando **prcomp**, obteniendo la desviación estándar de cada cp (componente principal), en este caso 23 componentes, y la matriz de covarianza, con su rotación o carga de cada variable, que son los coeficientes de las combinaciones lineales de las variables continuas, por ejemplo tenemos las cargas del componente 1 en la Ilustración 30 y podemos observar todas las cargas en los componentes restantes dentro del **Anexo 7**.

```
> acp
Standard deviations:
 [1] 14.30422037  9.49344767  7.73525400  6.20003329  5.29789250  4.05795757  3.43827119  3.26880122  2.83541513  2.35598779  1.68059295
[12]  1.60453899  1.51063003  1.32706970  1.12992877  0.86801363  0.64897399  0.52074120  0.41254667  0.15705106  0.15030206  0.09277866
[23]  0.03463031
```

Ilustración 29. Desviación estándar
Fuente. Elaboración propia

PC1 = 0.0599159079 + 0.0051642050 + 0.4857779006 + 0.3457185080 + 0.4010171731 + 0.0318225637 + 0.1525039064 + 0.0294136537 + 0.2042172117 + 0.0003890685 + 0.0091306397 + 0.0010434885 + 0.0285401356 + 0.0598236501 + 0.5574310394 + 0.2999965410 + 0.0002670942 + 0.0357384457 + 0.0363093139 + 0.0126610661 + 0.0461448860 + 0.0556145288 + 0.0066051544

Ilustración 30. Carga de variable en PC1
Fuente. Elaboración propia

3. Aplicamos la función **summary** a la variable que se le asignó el análisis de componentes principales con el comando **prcomp**, con el fin de obtener la importancia de los componentes, es decir, generar un resumen con el que se determine la proporción de varianza de cada cp y con ello nos oriente a elegir el número de componentes correctos para el análisis.

```
> summary(acp)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     14.304 9.4934 7.7353 6.20003 5.29789 4.05796 3.43827 3.26880
Proportion of Variance  0.421 0.1854 0.1231 0.07909 0.05775 0.03388 0.02432 0.02198
Cumulative Proportion   0.421 0.6064 0.7295 0.80861 0.86636 0.90024 0.92456 0.94654
                           PC9    PC10    PC11    PC12    PC13    PC14    PC15
Standard deviation     2.83542 2.35599 1.68059 1.6045 1.5106 1.32707 1.12993
Proportion of Variance 0.01654 0.01142 0.00581 0.0053 0.0047 0.00362 0.00263
Cumulative Proportion  0.96308 0.97450 0.98032 0.9856 0.9903 0.99393 0.99656
                          PC16    PC17    PC18    PC19    PC20    PC21    PC22
Standard deviation     0.86801 0.64897 0.52074 0.41255 0.15705 0.15030 0.09278
Proportion of Variance 0.00155 0.00087 0.00056 0.00035 0.00005 0.00005 0.00002
Cumulative Proportion  0.99811 0.99897 0.99953 0.99988 0.99993 0.99998 1.00000
                          PC23
Standard deviation     0.03463
Proportion of Variance 0.00000
Cumulative Proportion  1.00000
>
```

Ilustración 31. Importancia de componentes
Fuente. Elaboración propia

La Ilustración 31 muestra el resultado de la función **summary**, detallando lo siguiente:

- La primera fila describe la desviación estándar asociada a cada componente, en donde el PC1 se destaca con un 14.3 % de los demás componentes.
- La segunda fila indica la proporción de varianza en los datos dada por cada componente, en nuestro caso se puede considerar los dos primeros componentes como los más recomendables, ya que la suma de los mismos es de un 60,6% que supera el 50% de la información del conjunto de datos.
- La tercera fila se refiere a la proporción acumulada de la varianza explicada, es decir la suma progresiva de la proporción de varianza.

Dentro de este análisis la segunda fila es la más importante y con la que se trabaja para establecer los componentes principales. Como ya se mencionó los dos primeros suman un 60,6% así que es preciso tomar un componente más, para cumplir con la regla del 70% o mayor que debe tener la variación total de los componentes principales agrupados, como indica (León González, Llinás Solano, & Tilano, 2011).Para que la perdida de información no genere resultados inconsistentes.

En este caso la agrupación será de los tres primeros componentes con una variación estándar de 72,9%.

Expresado de manera gráfica se presenta la Ilustración 32, en donde se toma en cuenta que la varianza sea => 1 como lo indica (Peña, 2002) al calcular los componentes principales.

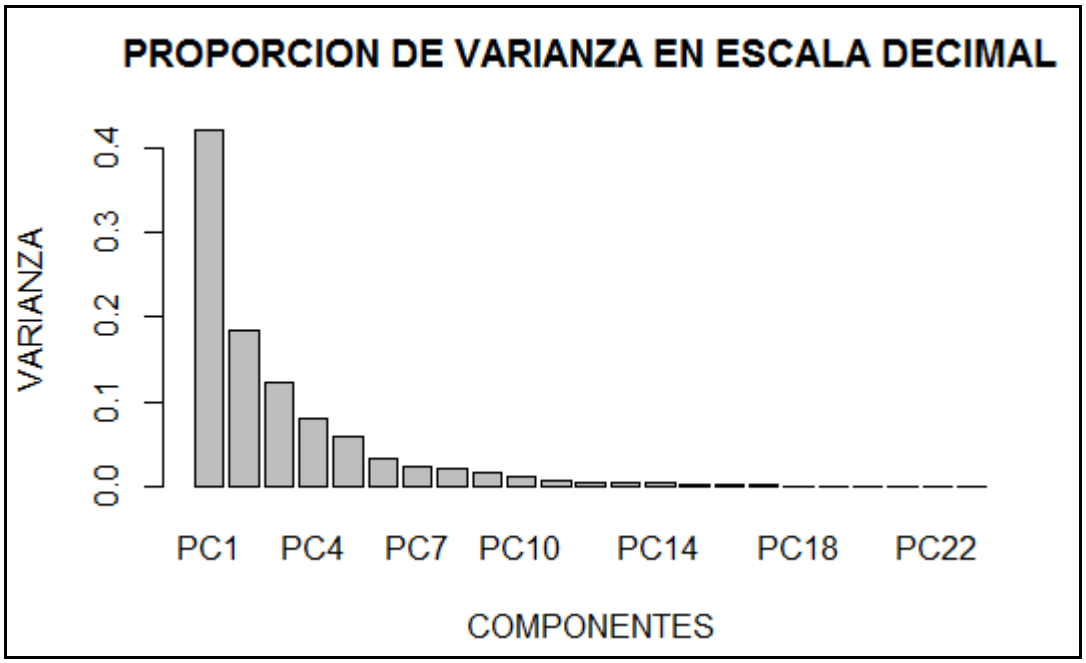


Ilustración 32. Proporción de varianza
Fuente. Elaboración propia

4. Ejecutamos la carga de los componentes principales elegidos, en donde analizaremos las variables con mayor valor absoluto, para determinar resultados finales Ilustración 33, a través del comando:
    - acpR<-acp$rotation[,1:3]

VARIABLES	PC1	PC2	PC3
chat_historial	0.0599159078682018	0.011601610812508	-0.00450756594717506
chat_report	0.00516420500951442	0.00205255419544047	0.000828171821534215
chat_talk	0.485777900594798	-0.416651457064584	0.702150630038944
chat_viewAllChat	0.34571850798649	0.382158680005205	-0.124221674632362
chat_viewChat	0.401017173148763	0.525565718485133	0.284264566201522
Elluminate_viewAllElluminate	0.0318225637429574	-0.0778019853320218	-0.0697687868021004
Elluminate_viewMeeting	0.152503906400671	-0.101854375533708	-0.00799688785238297
Elluminate_viewRecording	0.0294136537445713	0.00730276946302174	-0.0135589781769003
Elluminate_viewElluminate	0.204217211735974	-0.59034630893092	-0.342760734743069
addDiscussionForum	0.000389068527975494	0.000452348035015341	0.00007162866
subscribeForum	0.00913063971881332	0.00576328536194938	-0.00994975863847305
subscribeallForum	0.00104348854651537	0.00111758467332808	-0.00174021084840164
updatePostForum	0.0285401356399219	-0.011335214105284	-0.0198993337947761
viewDiscussionForum	0.059823650097724	0.0402911597466618	-0.0742976355699451
viewForumForum	0.557431039448255	-0.0927958207974173	-0.413985717606966
viewForumsForum	0.299996540990984	0.164643656127341	-0.330948304980297
viewSubscribersForum	0.00026709424298055	0.000006652231	-0.000427088880660197
Quiz_attempt	0.0357384456542711	-0.0233641931934086	-0.0120449645580517
Quiz_closeAttempt	0.036309313863222	-0.0232052629450453	-0.011855917116645
Quiz_evaluaciones	0.0126610660540526	-0.00750093761415401	-0.00265571251099086
Quiz_review1	0.0461448859978325	-0.00689329100087082	-0.0188963630918878
Quiz_view	-0.0556145288235807	0.00797199422421632	0.0182730968896598
Quiz_viewAll	0.00660515442437978	-0.0158673947499238	-0.0145860775590721

Ilustración 33. Valores propios con 3 CP
Fuente. Elaboración Propia

5. Selección de variables. (Jolliffe, 1972), da como criterio de selección, tomar de entre las variables de los componentes principales determinados, aquellos valores absolutos en los que su coeficiente se aproximen a 1 y si la variable coincide en dos PCs la posición de esta variable se asigna tomando en cuenta la que tenga mayor valor; aplicando este criterio obtenemos los datos de la Ilustración 34, que indica las variables con mayor dato informativo ubicadas en cada PC.

VARIABLES	PC1	PC2	PC3
chat_talk	0.485777900594798	-0.416651457064584	0.702150630038944
chat_viewAllChat	0.34571850798649	0.382158680005205	-0.124221674632362
chat_viewChat	0.401017173148763	0.525565718485133	0.284264566201522
Elluminate_viewElluminate	0.204217211735974	-0.59034630893092	-0.342760734743069
viewForumForum	0.557431039448255	-0.0927958207974173	-0.413985717606966
viewForumsForum	0.299996540990984	0.164643656127341	-0.330948304980297

Ilustración 34. Variables seleccionadas
Fuente. Elaboración propia

De donde obtenemos el siguiente detalle:

- PC1 = viewForumForum, cuya variable equivale al número de veces que el usuario ha revisado el foro dentro de otro un foro.
- PC2 = chat_viewAllChat, chat_viewChat y Elluminate_viewElluminate, cuyas variables corresponden a: todos los chats revisados por el estudiante, las veces que el estudiante a revisado un determinado chat y las veces que ha entrado a una determinada videoconferencia, respectivamente.
- PC3 = chat_talk y viewForumsForum, en donde la primera variable significa el número de veces que un chat se ha abierto y la segunda variable corresponde a número de foros revisados por el estudiante.

2. Descripción del modelo a través de la técnica de ACP.

Finalmente se etiquetará a cada componente, tomando en cuenta la característica que identifique las variables correspondientes a cada una, transformando así a las componentes principales resultantes en nuevas variables para el análisis final. Tabla 13.

Tabla 13. Nuevas variables

COMPONENTE	VARIABLES	ETIQUETA DEL COMPONENTE
PC1	viewForumForum	**Debate:** participación del estudiante dentro de un foros abierto a partir de otro foro.
PC2	chat_viewAllChat chat_viewChat Elluminate_viewElluminate	**Consultas:** revisión del estudiante de conversaciones y videoconferencias.
PC3	chat_talk viewForumsForum	**Comunicación:** conversaciones y participaciones de foros.

Fuente. Elaboración propia

3.5 Fase 5. Evaluación.

Para llegar a los objetivos planteados del proyecto, se procedió a la aplicación de dos técnicas descriptivas de minería de datos la k-means, la cual una vez establecidos los cluster de 3 y 4 para el respectivo análisis de resultados, indico que no es el mejor método para trabajar con la data, pues según las gráficas de correlación que podemos apreciar en los anexos **Anexo 2** y **Anexo 3** no proporcionan un diseño para definir que existen características similares para agrupar los datos en diferentes clusters.

Por tal razón se emplea el análisis de componentes principales, cuya finalidad es disminuir el número de variables sin afectar los resultados de un gran conjunto de datos, disminuyendo el tiempo para obtener dichos resultados, los componentes principales a ser elegidos deben cumplir un porcentaje >= al 70% de variabilidad del conjunto original, y en el caso de este proyecto al aplicar la técnica se pudo obtener una variabilidad de 72,9% con los 3 primeros componentes, mismos que se relacionan con cada una de las acciones dentro de las actividades de, chats, foros, video conferencias y quiz que los estudiantes realizaron en los periodos oct-2014/feb-2015 y abr-2015/ago-2015, dentro del EVA.

Con este resultado de componentes se registró una actividad predominante en 6 acciones, las cuales se estimaron considerando el mayor valor absoluto entre ellas. En la Ilustración 35 visualizamos la correlación de las variables originales sobre dos de sus componentes principales.

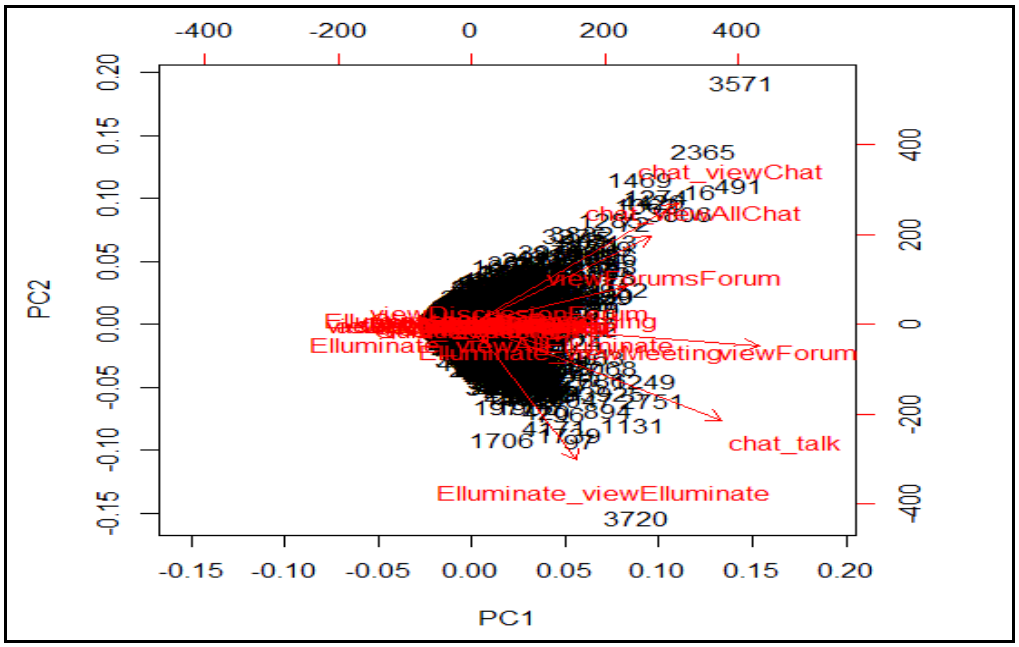


Ilustración 35. Representación gráfica de los PC
Fuente. Elaboración propia.

La correlación que indica la Ilustración 35 entre las 32 acciones listadas en la Ilustración 15 presenta un ángulo menor de 90°, y según el concepto del análisis de componentes

principales es un indicador muy acertado para aplicar esta técnica a los datos que se han obtenido.

3.5.1 Discusión de resultados.

El marco teórico y el desarrollado de este proyecto nos permite definir qué, la minería de datos es la extracción de conocimientos de una gran cantidad de datos, aplicando técnicas ya sean predictivas o descriptivas, dentro de herramientas desarrolladas para ejecutar sus algoritmos y gráficas, permitiendo de tal manera llegar a los análisis y conclusiones de un determinado objetivo, por ejemplo: permiten encontrar patrones de comportamientos en profesionales que se encuentran ya en el entorno laboral, sirve para verificar si el pensum de estudio es correcto o en que se debe mejorar, identifica como se encuentra la calidad de los docentes, permite elegir mejores juntas receptoras del voto presidencial, además ayuda localizar futuros atentados, así como predecir mejores pagos sobre préstamos, asimismo mejorar la calidad de servicios, entre otros, mencionadas en diferentes áreas en la sección 1.1.5.

En este proyecto el objetivo es aplicar minería de datos para el diseño de un modelo descriptivo de la labor tutorial en las asignaturas de formación básica de la modalidad abierta y a distancia de la UTPL, para brindar de esta manera, una posible solución al problema de deserción de estudiantes, que por parte de los docentes investigadores de la institución han localizado mayor abandono en este tipo de materias.

El proceso para trabajar con minería de datos inicia con la elección de una metodología para contar con un orden en lo que recolección, elección, análisis y obtención del conocimiento se refiere, tal como lo realiza (Formia, 2012) en su trabajo final de grado, quien opta directamente por el proceso de KDD, en donde minería de datos es una parte del proceso de esta metodología. La autora considera el uso del proceso KDD porque la secuencia de sus fases no es estricta y dependiendo del resultado de cada fase se puede realizar cambios constantemente. En el caso del presente proyecto para la selección de la metodología realizamos investigación bibliográfica, de donde se consideró a CRISP-DM, SEMMA y KDD como las más utilizadas en la minería de datos(MD), y de las cuales, tras el resultado de una encuesta sobre la usabilidad de metodologías de MD realizada a expertos en cuanto a usabilidad, presente en la Ilustración 8, se considera a la metodología de CRISP-DM como la más utilizada para la extracción de conocimiento, y en base a la recomendación de este resultado de expertos, se procede a utilizarla en este proyecto.

Además, tomando en cuenta que las fases de CRISP-DM realiza un trabajo iterativo entre sus fases, es decir, que se puede establecer relaciones entre cualquiera de sus fases sin importar la secuencia que estas sigan, tal como menciona (Chapman et al., 2000); trabajo

que también presentan las fases del proceso de KDD, pero a diferencia de CRISP-DM, KDD hace a la minería de datos como parte de una de sus fases, mientras que CRISP-DM es una metodología propia de la minería de datos.

El realizar un proyecto, no exige el uso de una metodología existente, como las antes mencionadas, pero para muchos es más eficaz trabajar con una; pues brinda una guía de trabajo para llegar al objetivo planteado de un determinado proyecto, pero existen quienes aplican o realizan sus trabajos sobre su propia metodología, así es como encontramos un artículo de (Amaya, Barrientos, & Heredia, 2014), sobre la creación de un modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos, quienes no tomaron una metodología preexistente, sino que crearon una según sus necesidades.

Una vez establecida la metodología de desarrollo, procedemos a su aplicación, como se muestra desde el primeramente analizamos el problema planteado que se menciona en el segundo párrafo de esta discusión, y según esta problemática se procede a localizar y obtener los datos necesarios, que más adelante serian analizados mediante técnicas descriptivas de minería de datos, los mismos que fueron tomamos de dos bases de datos, la primera fue la base de datos del EVA que estaban dentro de Moodle, y que por medio de permisos institucionales se pudo tener acceso y con sentencias SQL en MySQL realizar las consultas con las que se obtuvo los datos cuantitativos como: la cedula del estudiante y las acciones que los estudiantes realizaban en el entorno virtual de aprendizaje y algunos cualitativos como las materias y género, todos estos en función de las materias de formación básica de: Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio, de la modalidad abierta y a distancia de la UTPL, dentro de los periodos oct-2014/feb-2015 y abr-2015/ago-2015. Las operaciones de lenguaje relacional SQL, también son aplicadas por el proyecto antes mencionado de (Formia, 2012), lo cual fue suficiente para adquirir su data, limpiarla y seleccionar su muestra, en nuestro caso a más del resultado de las consultas en la BD, se requirió datos cualitativos como: estado civil, edad y tipo de matrícula, por lo que optamos a una solicitud de reporte, al administrador de la base de datos del Syllabus, con las características de los periodos y la modalidad de estudio que aplicamos en la BD del EVA, finalmente unificamos todos los datos obtenidos de las dos bases de datos, dentro de la herramienta de Excel tomando como id las cedulas de los estudiantes, para unirlas.

Con un análisis exploratorio de los datos resultantes, se pudo apreciar más de 90000 datos, que contenían información inconsistente, como: valores faltantes, caracteres especiales, datos duplicados y acciones sin una adecuada presentación, pues se encontraban listadas de forma vertical seguidas del tipo de actividad a la que correspondían, como podemos apreciar en la Ilustración 36, dando lugar a los datos duplicados de los estudiantes, por ello

se procede a levantar la información en la herramienta XAMP, para crear vistas y poder tener la presentación de un estudiante por cada acción de manera horizontal, disminuyendo los datos a un numero de 4551 con 33 columnas, y así obteniendo una mejor presentación, que podemos apreciar en la Ilustración 19 de la sección 0

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	IDENTIDAD	GENERO	EDAD	ESTADO_CIV	NIVEL	TITULACION	TIPOMATRIC	TIPO_ESTUDI	CURSO	PERIODO	ACCIONES	ACCION	ACTIVIDAD
2	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	1	add post	forum
3	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	1	attempt	quiz
4	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	1	close attemp	quiz
5	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	2	continue att	quiz
6	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	2	search	forum
7	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	1	subscribe	forum
8	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	2	view	quiz
9	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	7	view forum	forum
10	102584398	MASCULINO	43	Casado	2	Extraordinar	SEGUNDA EN	ADMINISTRA	EXPRESION (	Abr/2015 - A	8	view forums	forum

Ilustración 36. Presentación de la unión de datos
Fuente. Elaboración propia

Una vez obtenidos los datos, se realiza un nuevo análisis del contenido y se considera que para verificar los datos de las acciones dentro del periodo y las materias seleccionadas los datos cualitativos nos proporcionaron esa seguridad, pero el fin es trabajar con las acciones de los estudiantes, así que dentro de la herramienta RStudio, descrita más adelante y seleccionada para tratar los datos con técnicas de MD, filtramos solo las 23 acciones, con los 4551 datos, con esta matriz definitiva de datos, continuamos con la creación del modelo descriptivo, seleccionando técnicas de agrupación de minería de datos como k-means, mencionada en la sección 2.2, pues según sus características, como la identificación a priori de los grupos de trabajo y el trabajo directo con la matriz de datos originales y no con la matriz de distancias, se cree conveniente utilizarla en el trabajo de la creación de nuestro modelo, esta técnica ha sido muy utilizada en trabajos relacionados a nuestra problemática, tal es el caso de (Formia, 2012) quien utilizo k-medias para agrupar el conjunto de alumnos con características relevantes de deserción; en nuestro caso se utiliza la técnica para agrupar las acciones con mayor número de actividad por parte del estudiante.

La técnica seleccionada, por estos dos proyectos, además de sus tipos de agrupación, difieren en la herramienta donde son ejecutadas, en nuestro caso el software fue RStudio, como se menciona en el párrafo anterior, el mismo que fue elegido en previa investigación bibliográfica y al igual que la metodología seleccionada, una encuesta a expertos por parte de (Piatetsky, 2015) sobre la usabilidad de herramientas para la minería de datos, indica que el resultado favorable lo obtuvo RStudio. En el caso del proyecto de Alejandra citada en (Formia, 2012) creyó conveniente realizar su trabajo en el software denominado RapidMiner, herramienta de código abierto al igual que RStudio, en la que además normaliza los datos, que transforma de cualitativos a cuantitativos, pues como ya se mencionó, el método seleccionado k-means solo trabaja con datos numéricos.

Para aplicar k-means, se inicia asignando el número de agrupaciones, del resultado que se obtiene en la estabilidad que presentan sus cuatro algoritmos denominados: Hartigan-Wong, Lloyd, Forgy y MacQueen, en la gráfica del codo de jambu Ilustración 22, cuyo resultado es de 4 agrupaciones, y además se optimiza que el método funcionara mejor con el algoritmo Hartigan según el resultado de su alta inercia con respecto a los tres algoritmos restantes.

Entonces se aplicó kmean con con las 23 acciones de los estudiantes, en 4 agrupaciones y con el algoritmo Hartigan, pero los resultados en las gráficas no fueron favorables, pues los datos se presentaron a una sola dirección **Anexo 4**, es decir no surgieron las agrupaciones a ser interpretadas, por lo que se volvió a realizar el proceso, pero esta vez con 3 agrupaciones **Anexo 5**, sin tener mejor resultado, finalmente se intentó con 5 agrupaciones pero el resultado seguía igual.

En el caso del proyecto de (Formia, 2012) obtuvo 5 agrupaciones ya l igual que nuestro caso, no le permiten describir las agrupaciones, ni calcular los valores frecuentes en los grupos, entonces recurre a la aplicación de dos esquemas: uno es la selección de atributos de la metodología select forward, para elegir un subconjunto de los conjuntos resultantes, para mejorar el entendimiento, el desempeño predictivo y la eficiencia del modelo, aplicado a través del operador Optimize Selection en RapidMiner y el otro esquema se refiere a la selección genética de características, que es conocido como selección multiobjetivo, de propósito general y se adapta a los casos que poseen poco conocimiento en el manejo del problema, esto lo realiza a través de los operadores Optimize Selection (Evolutionary) y Performance (CFS) de RapidMiner. Los esquemas que utiliza tienen resultados similares en los subconjuntos, presentando estos atributos: estado_civil, padre_vive, rel _trab_carrera, alu _trab_remmon, cant_fami_cargo y cant_hijos_alum, Sit_laboral_padre = “DesOcupado”, pero como ultima validación utiliza la técnica de árboles de decisión, con el algoritmo C4.5 y el operador W-J48 de la extensión weka de RapidMiner, el cual le permite eliminar atributos que no son útiles y tomar el más relevante para la división de cada nodo interno, y cuyo resultado son los mismos atributos de los dos algoritmos de selección que utilizo.

Entonces toma el subconjunto resultante de la metodología select forward, validadas con las 2 técnicas de selección, para volver a ejecutar kmeans, con k=5 y con los atributos seleccionados, finalmente muestra el resultado de las características de deserción de los estudiantes dentro de las 5 agrupaciones: en la agrupación 0, contiene un 62% con mayor relación de dependencia; en la agrupación 4, presenta un 62% en aquellos estudiantes que son mayores monotributistas, en la agrupación 3 con 55% en mayores sin información, en la agrupación 1, con 56% con estudiantes que trabajan y en la agrupación 2, con 38% corresponde a estudiantes que no trabajan.

La solución que se aplicó en nuestro caso, también fue buscar otra técnica, pero a diferencia de la selección de atributos y el árbol de decisión, la que se eligió fue el análisis de componentes principales, la cual tiene como objetivo comprimir los datos de un conjunto original en un grupo más pequeño sin perder la importancia de la información, como indica (de la Fuente, 2011) en la sección 3.4.2, para determinar que el ACP era la técnica apropiada se analizaron las dos ilustraciones: a la Ilustración 27 con 3 agrupaciones y la Ilustración 28 con las 4 agrupaciones que fueron recomendadas por el codo de jambu, estas graficas fueron el resultado de la sentencia plotcluster con kmeans. Al visualizar dichas ilustraciones se observó que el agrupamiento de los datos se aproximaba hacia el origen, y uno de los requisitos de ACP es que los datos se aproximen a 0, es decir, hacia el origen.

Entonces, se inicia la utilización de la sentencia acp<-prcomp(datos), para aplicar el análisis de componentes principales a los datos, en donde las 23 acciones pasan a ser 23 variables a ser reducidas; la sentencia prcomp da como resultado la desviación estándar de cada componente y la matriz de covarianza, con su rotación o carga de cada variable, continuamos el proceso aplicando la sentencia summary, para generar un resumen con la proporción de varianza de cada uno de los 23 CP para determinar el número de componentes a ser analizados, de donde se obtuvieron 3, con un suma de 72,9%, tomando en cuenta que la suma de su varianza debería superar un porcentaje de 70%, según (León González et al., 2011), continuamos con el desarrollo de este análisis tal.
En donde al determinar los componentes principales, se ubican las rotaciones o cargas que les corresponde a cada uno, es decir, las acciones, las mismas que nos ayudaron a determinar, cuales sobresalen según el valor absoluto que se aproxime a 1. De donde, se obtuvo 6 acciones definidas como: viewForumForum, cuya variable equivale al número de veces que el usuario ha revisado el foro dentro de otro un foro; chat_viewAllChat, corresponde a todos los chats revisados por el estudiante; chat_viewChat, las veces que el estudiante a revisado un determinado chat, Elluminate_viewElluminate, son las veces que ha entrado a una determinada videoconferencia; chat_talk, número de veces que un chat se ha abierto y viewForumsForum, que es el número de foros revisados por el estudiante. Estas variables están distribuidas en los 3 componentes principales, a los cuales se los etiqueta según el conocimiento del experto, guiándose en el tipo de datos de su resultado. En nuestro caso por la orientación del tutor del proyecto. El resultado final se presenta en la Tabla 13.

Es necesario señalar que, la problemática de los dos proyectos principales citados en la discusión, es igual, pero el objetivo fue totalmente distinto, pues mientras (Formia, 2012) buscaba características de deserción en los estudiantes universitarios dentro de un solo periodo académico, el presente proyecto ya se basó en estas características similares , que

localizadas anteriormente, en investigaciones anteriores, por parte de docentes investigativos de la UTPL, por tal motivo, el objetivo fue encontrar las acciones en las que el estudiante destaque su participación, es decir, en las que ingrese a interactuar dentro del EVA, con datos de dos periodos académicos.

Finalmente se consiguió el modelo descriptivo de la labor tutorial de las asignaturas de formación básica de la modalidad abierta y a distancia de la UTPL, ya que por medio de las acciones obtenidas el tutor podrá tomarlas en cuenta, ya sea para reforzar más contenido por medio de estas acciones que muestran más interés en los estudiantes, o aplicar las mismas metodologías para aquellas que crean necesarias, en la formación de sus alumnos.

CONCLUSIONES

Con la finalidad de cumplir los objetivos planteados en el proyecto se concluye los siguientes puntos:

- El modelo descriptivo se lo obtuvo a través de la técnica de análisis de componentes principales, en donde los ángulos de correlación entre variables son menores a 90° y la variabilidad de los 3 componentes elegidos es de 72,9%, datos que confirman la aplicación de la técnica.

- Se obtuvieron 6 acciones (viewForumForum, chat_viewAllChat, chat_viewChat, Elluminate_viewElluminate, chat_talk y viewForumsForum) con mayor valor absoluto, agrupadas en tres componentes principales, determinando que tienen mayor actividad dentro del EVA, por los estudiantes.

- De las tablas que pertenecen al modelo de datos de Moodle, solo un 3% es equivalente a 8 tablas que se utilizan para realizar el análisis del proyecto.

- Se aplicaron dos técnicas descriptivas de minería de datos denominadas kmeans, y análisis de componentes principales.

RECOMENDACIONES

Al concluir este proyecto de fin de titulación, se cree conveniente tomar en cuenta las siguientes recomendaciones:

- Empezar el desarrollo de un proyecto aplicando una metodología, que este acorde al tipo de técnicas o resultados a obtener, como se realizó en este trabajo de minería de datos con la metodología CRISP-DM.

- Revisión constante de los objetivos del proyecto para no desviar el trabajo y cumplir con el cronograma planteado.

- Iniciar la minería de datos con un análisis exploratorio, para tener una visualización del comportamiento de los datos y posteriormente aplicar una técnica, para el análisis final.

- implementar más de una técnica de minería para obtener resultados satisfactorios.

- Incluir como parte del proyecto a un asesor de la base de datos del EVA, y un experto en data mining, para poder brindar una correcta interpretación de los datos y resultados.

- Adquirir herramientas de software y hardware, según los requerimientos del proyecto, evitando así posibles complicaciones que retrasen la culminación del mismo.

- Para posteriores análisis se recomienda aplicar minería de texto.

BIBLIOGRAFIA

Abarca, A., & Sánchez, M. A. (2005). La deserción estudiantil en la educación superior: El caso de la Universidad de Costa Rica. *Actualidades Investigativas en Educación*, *5*(4). http://doi.org/10.15517/aie.v5i4.9186

Abril Valdez, E., Román Pérez, R., Cubillas Rodríguez, M. J., & Moreno Celaya, I. (2008). ¿Deserción o autoexclusión? Un análisis de las causas de abandono escolar en estudiantes de educación media superior en Sonora, México. *Revista Electrónica de Investigación Educativa*, *10*(1), 1–16. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1607-40412008000100007&lng=es&nrm=iso&tlng=pt

alexcouoh4. (2016). Historia de e-lerning | Line.do - Descubre historias por medio de cronologías y ¡cuenta las tuyas también! Retrieved from https://line.do/es/historia-de-e-lerning/10r0/vertical

Algieri, R. D., Tornese, E. B., Mazzoglio, M. J., Dogliotti, C., Gazzotti, A., Jimenez, H. N., & Rey, L. M. (2014a). *EVEA en Anatomía: Usos, aplicaciones, experiencias y bases pedagógicas* (DUNKEN). Buenos Aires.

Algieri, R. D., Tornese, E. B., Mazzoglio, M. J., Dogliotti, C., Gazzotti, A., Jimenez, H. N., & Rey, L. M. (2014b). *EVEA en Anatomía: Usos, aplicaciones, experiencias y bases pedagógicas* (DUNKEN). Buenos Aires. Retrieved from https://books.google.com.ec/books?id=MfyMBAAAQBAJ&pg=PA38&lpg=PA38&dq=el+e-learning+en+1950-1960&source=bl&ots=__sU3Z5abf&sig=ilSHVR7YeBsDX5BMuk2UMqwT_S0&hl=es&sa=X&ved=0ahUKEwjmzM7YwLLPAhXFKB4KHXXXDF0Q6AEIKzAE#v=onepage&q&f=false

Aluja, T. (2008). *Clustering Data Mining course*. Retrieved from http://www.cs.upc.edu/~belanche/Docencia/mineria/English-september-2008/DM Clustering.pdf

Amaya, Y., Barrientos, E., & Heredia, D. (2014). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. Retrieved from https://pdfs.semanticscholar.org/9469/6703f9a64f284e3d0e0d16575557aeed98bf.pdf

Análisis de Cluster y Arboles de Clasificación. (2010). Retrieved from http://www.docdatabase.net/more-an225lisisde-clustery-arbolesde-clasificaci243n-598299.html

Arce, C., de Francisco, C., & Arce, I. (2010). Escalamiento multidimensional: Concepto y aplicaciones. *Papeles Del Psicologo*.

Aretio, L. G. (2001). La educación a distancia | Planeta de Libros. Retrieved December 21, 2015, from http://www.planetadelibros.com/la-educacion-a-distancia-libro-14415.html

Arévalo, F., & Maldonado, J. (2010). Estrategias para promover la retención estudiantil en un sistema de educación a distancia. Retrieved from http://reposital.cuaed.unam.mx:8080/jspui/bitstream/123456789/2757/1/judith_maldonado_flora_arevalo_estrategias_educacion_a_distancia.pdf

Arriaza, M. (2006). GUÍA PRÁCTICA DE ANÁLISIS DE DATOS. Retrieved from http://www.um.es/jmpaz/AGP1213/guia_practica_de_analisis_de_datos.pdf

Aular, Y. J. M., & Pereira, R. T. (2007). Minería de Datos como Soporte a la Toma de Decisiones Empresariales. *Opción*, *23*(52). Retrieved from http://www.produccioncientifica.luz.edu.ve/index.php/opcion/article/view/6402

Azoumana, K. (2013). Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos.

Camana, R. (2014). Una Experiencia Personal: Pico y Pala en la Explotación y Visualización de Datos Electorales. *Revista Tecnológica ESPOL – RTE*, *27*(1), 1–13. Retrieved from https://www.academia.edu/28491226/Potenciales_Aplicaciones_de_la_Minería_de_Datos_en_Ecuador

Camana, R. (2016). Potenciales Aplicaciones de la Minería de Datos en Ecuador. Retrieved from http://www.rte.espol.edu.ec/index.php/tecnologica/article/viewFile/288/199

Candás Romero, J. (2006). Minería de datos en bibliotecas: bibliominería. *BID: Textos Universitaris de Biblioteconomia I Documentació*, (17). Retrieved from http://www2.ub.edu/bid/consulta_articulos.php?fichero=17canda2.htm

Centro Microdatos. (2008). Estudio sobre causas de la desercion universitaria, 1–143. Retrieved from https://www.google.com.ec/url?sa=t&rct=j&q=&esrc=s&source=web&cd=9&ved=0ahUKEwip9O-UwvrRAhVTziYKHe0uD8wQFghZMAg&url=http://www.oei.es/historico/pdf2/causas-desercion-universitaria-chile.pdf&usg=AFQjCNE89sgEKr0mEdRzVISqBVSTfVGfxA&sig2=Uhc3gr9Zd

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM. Retrieved March 21, 2016, from https://www.the-modeling-agency.com/crisp-dm.pdf

Closas, A., Arriola, E. A., Kuc Zening, C. I., Amarilla, M. R., & Jovanovich, E. C. (2013). Análisis multivariante, conceptos y aplicaciones en Psicología Educativa y Psicometría. *Enfoques: Revista de La Universidad Adventista Del Plata, ISSN 1514-6006, Vol. 25, N^{o}. 1, 2013, Págs. 65-92*, *25*(1), 65–92. Retrieved from file:///D:/Escritorio/Dialnet-AnalisisMultivarianteConceptosYAplicacionesEnPsico-5229555.pdf

de la Fuente, S. (2011). ANÁLISIS DE COMPONENTES PRINCIPALES (ACP). Retrieved from http://www.fuenterrebollo.com/Economicas/ECONOMETRIA/MULTIVARIANTE/ACP/ACP.pdf

Delors Jacques. (1996). *LA EDUCACIÓN encierra un tesoro*. *Santillana* (UNESCO). Madrid. Retrieved from http://www.bib.ufro.cl/portalv3/files/informe-a-la-unesco-de-la-comision-internacional-sobre-la-eduacion-para-el-sigleo-xxi.pdf

El comercio. (2016). El 26% de los universitarios se retiró en los primeros años | El Comercio. Retrieved February 1, 2017, from http://www.elcomercio.com/actualidad/ecuador-universitarios-desercion-educacion-jovenes.html

Evolución y retos de la educación virtual. (2011). Retrieved January 11, 2016, from http://openaccess.uoc.edu/webapps/o2/bitstream/10609/9781/1/TRIPA__e-learning_castellano.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996, March 15). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*. Retrieved from http://www.aaai.org/ojs/index.php/aimagazine/article/view/1230

Febles Rodríguez, J. P., & González Pérez, A. (2001). Aplicación de la minería de datos en la bioinformática. *ACIMED*, *10*(2), 69–76. Retrieved from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000200003&lng=es&nrm=iso&tlng=es

Fernández Santana, O. (1991). El análisis de Cluster : aplicación, interpretación y validación. *Papers : Revista de Sociologia*, (37), 065–076.

Figueras, M. S., & Gargallo Valero, P. (2003). Análisis exploratorio de datos. Retrieved from

http://www.5campus.com/leccion/aed

Formia, A. (2012). Evaluación de técnicas de Extracción de Conocimiento en Bases de Datos y su aplicación a la deserción de alumnos universitarios. Retrieved from http://sedici.unlp.edu.ar/bitstream/handle/10915/26772/Documento_completo.pdf?sequence=1

Frías, Y. (2005). Un modelo del Proceso Educativo a Distancia para la Universidad de Pinar del Río, Cuba, 41. Retrieved from http://rc.upr.edu.cu/bitstream/DICT/78/1/2012.3.20.u1.s11.t.pdf

García Aretio, L. (1999a). HISTORIA DE LA EDUCACIÓN A DISTANCIA. *RIED. Revista Iberoamericana de Educación a Distancia*, *2*(1). http://doi.org/10.5944/ried.2.1.2084

García Aretio, L. (1999b). HISTORIA DE LA EDUCACIÓN A DISTANCIA. *RIED. Revista Iberoamericana de Educación a Distancia*, *2*(1). http://doi.org/10.5944/ried.2.1.2084

Garrálaga, I. M. F. (2014). Sistema de mentoría para nuevos estudiantes de primer ciclo. Retrieved January 13, 2016, from http://dspace.utpl.edu.ec/handle/123456789/11717

Guanipa Perez, M., & Urdaneta Marcos. (2007). Perfil de competencias del docente para la educación a distancia • GestioPolis. Retrieved from http://www.gestiopolis.com/perfil-de-competencias-del-docente-para-la-educacion-a-distancia/

Guerreo, F., & Ramirez, J. (2012). EL ANÁLISIS DE ESCALAMIENTO MULTIDIMENSIONAL: UNA ALTERNATIVA Y UN COMPLEMENTO A OTRAS TÉCNICAS MULTIVARIANTES. *La Sociología En Sus Escenarios*, *0*(25).

Han, J., & Kamber, M. (2006). Data Mining Concept and Techniques 2nd edition. Retrieved from https://archive.org/stream/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016#page/n1/mode/2up

Han, J., Kamber, M., & Pei, J. (2011). *DATA MINING: Concepts and Techniques*. Retrieved from https://books.google.com.ec/books?hl=es&lr=&id=pQws07tdpjoC&oi=fnd&pg=PP1&ots=tyKqWSozXX&sig=dNO3phmiYWjiMjAM2UwVFSgamUM#v=onepage&q&f=false

Han, J., Kamber, M., & Pei Jian. (2011). *DATA MINING*. Retrieved from http://www.cse.hcmut.edu.vn/~chauvtn/data_mining/Texts/[1] Data Mining - Concepts and Techniques (3rd Ed).pdf

Härdle, W. K., & Simar, L. (2015). Applied Multivariate Statistical Analysis - Wolfgang Karl Härdle, Léopold Simar - Google Libros. In *Applied Multivariate Statiscal Analysis* (Third). Retrieved from https://books.google.com.ec/books?id=3Wz205ve5ioC&pg=PA385&lpg=PA385&dq=Hotelling+(1935)+analisis+canonica&source=bl&ots=IAdgapCos2&sig=tv5W9okIuO5qq0BPxdOeCBAg11o&hl=es&sa=X&ved=0ahUKEwjj8dzZ4rfOAhXqJ8AKHU0LC_sQ6AEISDAG#v=onepage&q&f=false

Jolliffe, I. T. (1972). Discarding Variables in a Principal Component Analysis. I: Artificial Data. *Applied Statistics*, *21*(2), 161–165. http://doi.org/10.2307/2346488

Keegan, D. (1996). *Foundations of distance education*. book, Psychology Press. Retrieved from https://books.google.com.ec/books?hl=es&lr=&id=nYkrTWDj5twC&oi=fnd&pg=PR11&dq=Keegan,+D.+(1996)&ots=UOR7krbs7Y&sig=M-mx2GTWFnBwShirZHFsu5BVpBI#v=onepage&q&f=false

Kelmansky, D. (2006). *Introducción al lenguaje R*. Retrieved from http://www.dm.uba.ar/materias/analisis_expl_y_conf_de_datos_de_exp_de_marrays_Mae/2006/1/practicas/TP3-2006.pdf

León González, Á., Llinás Solano, H., & Tilano, J. (2011). Análisis multivariado aplicando componentes principales al caso de los desplazados. *Revista Científica Ingeniería Y Desarrollo*, *23*(23), 119–142. Retrieved from http://rcientificas.uninorte.edu.co/index.php/ingenieria/article/viewArticle/2098/4467#t7

LOGICALIS. (2014). Minería de datos: aplicaciones más populares a día de hoy. Retrieved February 3, 2017, from https://blog.es.logicalis.com/analytics/mineria-de-datos-aplicaciones-que-ya-son-una-realidad?__hstc=61804339.eeb3a29edafd5f3032740028a0ae12b0.1485740912263.1486131981441.1486164197345.8&__hssc=61804339.1.1486164197345&__hsfp=1307689764

López, C. P. (2007). *Minería de datos: técnicas y herramientas*. Retrieved from https://books.google.com.ec/books?hl=es&lr=&id=wz-D_8uPFCEC&oi=fnd&pg=PR4&dq=técnicas+y+herramientas,+César+Pérez+López&ots=Th_-vk8z8J&sig=ko_RBLhuHrf7XKXvXtIBwdBMxZI#v=onepage&q=técnicas y herramientas, César Pérez L%

Maitra, R., & Ramler, I. P. (2010). A k-mean-directions Algorithm for Fast Clustering of Data

on the Sphere. *Journal of Computational and Graphical Statistics*, *19*(2), 377–396. http://doi.org/10.1198/jcgs.2009.08155

Malbernat, L. R., Clemens, M. P., Varela, A. E., & Urrizaga, M. (2015). Aplicación de técnicas de data mining en gestión de docentes de educación superior.

Martínez, C. (2008). La educación a distancia: sus características y necesidad en la educación actual. Retrieved from https://webcache.googleusercontent.com/search?q=cache:8xTcAhJZBvwJ:https://dialnet.unirioja.es/descarga/articulo/5057022.pdf+&cd=5&hl=es&ct=clnk&gl=ec

Microsoft. (2016). Conceptos de minería de datos. Retrieved January 20, 2017, from https://msdn.microsoft.com/es-ES/library/ms174949(v=sql.120).aspx

Moine, J., Haedo, A., & Gordillo, S. (2011). Estudio comparativo de metodologías para minería de datos. Retrieved from http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1

Molina, L. (2002). Data mining: torturando a los datos hasta que confiesen. Retrieved from https://www.businessintelligence.info/assets/dss/molina-torturando-datos.pdf

Montequín, R., Teresa, M., Cabal, Á., Valeriano, J., Fernández, M., Manuel, J., & Valdés, G. (2002). Metodologías Para La Realización De Proyectos De Data Mining, pp. 257–265. Retrieved from http://aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf

Núñez, M. (2016). La virtualización de la educación superior en América Latina: entre tendencias y paradigmas Virtualization of Higher Education in Latin America: Between Trends and Paradigms. *Artic*, *48*(1), 30–2016. http://doi.org/10.6018/red/48/1

Olson David, & Delen Dursun. (2008). Advanced Data Mining Techniques. Retrieved from https://books.google.com.ec/books?hl=es&lr=&id=2vb-LZEn8uUC&oi=fnd&pg=PA2&dq=SEMMA,+the+assessment+of+the+results+by+analyzing+how+well+the+model+or+models&ots=zV8ZY31RpW&sig=OBSzL2jAy_c5RbA8iIJVf-qCyj4#v=onepage&q&f=false

Palomo Miñambres Oscar. (2011). *Minería de Datos*. Retrieved from http://www.it.uc3m.es/jvillena/irc/practicas/11-12/12mem.pdf

Peña, D. (2002). ANÁLISIS DE DATOS MULTIVARIANTES. Retrieved from

http://www.dpye.iimas.unam.mx/lety/archivos/cursoinegi/apoyos/ANAÌ• LISIS DE DATOS MULTIVARIANTES(1).pdf

Perdomo Maribel. (2008). Modelo para la formacion de los docentes que trabajan en la Educación a Disctancia, fundamentado en un enfoque por competencias. *EduQ@*. Retrieved from http://eduqa2008.eduqa.net/eduqa2008/images/ponencias/eje_tematico_4/4_46_Modelo_para_la_formacion_de_los_docentes__Perdomo_de_Vasquez.pdf

Perez, S. L. (2014). Minería de datos (Reglas de asociací on arboles de decisí on). Retrieved from http://computacion.cs.cinvestav.mx/~sperez/cursos/md/14i/ReglasAsociacionYArboles.pdf

Piatetsky, G. (2015). Encuesta: ¿Qué Analytics, minería de datos, software de Ciencia de datos / herramientas que utiliza en los últimos 12 meses? Retrieved December 1, 2015, from http://www.kdnuggets.com/polls/2015/analytics-data-mining-data-science-software-used.html

Picasso Iñaki Díaz Covián, P. (2014). Descubrimiento de conocimiento en bases de datos espaciales, 34. Retrieved from http://dspace.sheol.uniovi.es/dspace/bitstream/10651/27914/6/TFM_I%C3%B1akiDiazCovian.pdf

Raúl Fernández Aedo, M. D., Pedro Mario Server García, Elianis Cepero Fadragas, & Romero. (2005). LA UNIVERSIDAD Y LA EDUCACIÓN A DISTANCIA. Retrieved January 6, 2016, from http://sedici.unlp.edu.ar/bitstream/handle/10915/24467/Documento_completo.pdf?sequence=1

Rodríguez, A. del C. M. (2009). El diseño instruccional en la educación a distancia. Retrieved January 6, 2016, from http://www.redalyc.org/pdf/688/68812679010.pdf

Rodriguez, O. (2014). *Lección N^{o}4 Calibración y Selección de Modelos - LAPLACE - YouTube*. Retrieved from https://www.youtube.com/watch?v=UwX4Ta78J0U

Rubio, M. (2014). Modalidad Abierta y a Distancia Guía general de educación a distancia. Retrieved from http://www.utpl.edu.ec/sites/default/files/pregrado/guia-general-MAD.pdf

Schneckenberg, D. (2004). El “e-learning” transforma la educación superior. *Educar*. Servei de Publicacions. Retrieved from

http://dialnet.unirioja.es/servlet/articulo?codigo=989408&info=resumen&idioma=CAT

Seoane, A., & García, F. (2010). 1. Antecedentes históricos. De la formación a distancia al eLearning. Retrieved January 7, 2016, from http://antia.fis.usal.es/sharedir/TOL/

singular. (2016). s|ngular - CRISP-DM: La metodología para poner orden en los proyectos de Data Science. Retrieved January 30, 2017, from https://data.sngular.team/es/art/25/crisp-dm-la-metodologia-para-poner-orden-en-los-proyectos-de-data-science

Sinnexus. (2016a). Datamining (Minería de datos). Retrieved November 6, 2015, from http://www.sinnexus.com/business_intelligence/datamining.aspx

Sinnexus. (2016b). Datos, información, conocimiento. Retrieved November 6, 2015, from http://www.sinnexus.com/business_intelligence/piramide_negocio.aspx

U.Alicante. (2007). Grados. Retrieved January 11, 2016, from http://web.ua.es/es/oia/preguntas/grados.html#materias-basicas

Universia. (2008). Planes de estudio. Retrieved January 30, 2017, from http://pre.universia.es/preguntas-frecuentes/estudios-universitarios/planes-estudio/

Vargas, B. (2014). Introducción a R - con minería de datos. Retrieved from http://es.slideshare.net/BLANCAVG/INTRODUCCIN-A-R-CON-MINERA-DE-DATOS

Vargas, M. (2012). RESUMEN ANÁLISIS CLUSTER, 7–10. Retrieved from http://www.ugr.es/~mvargas/

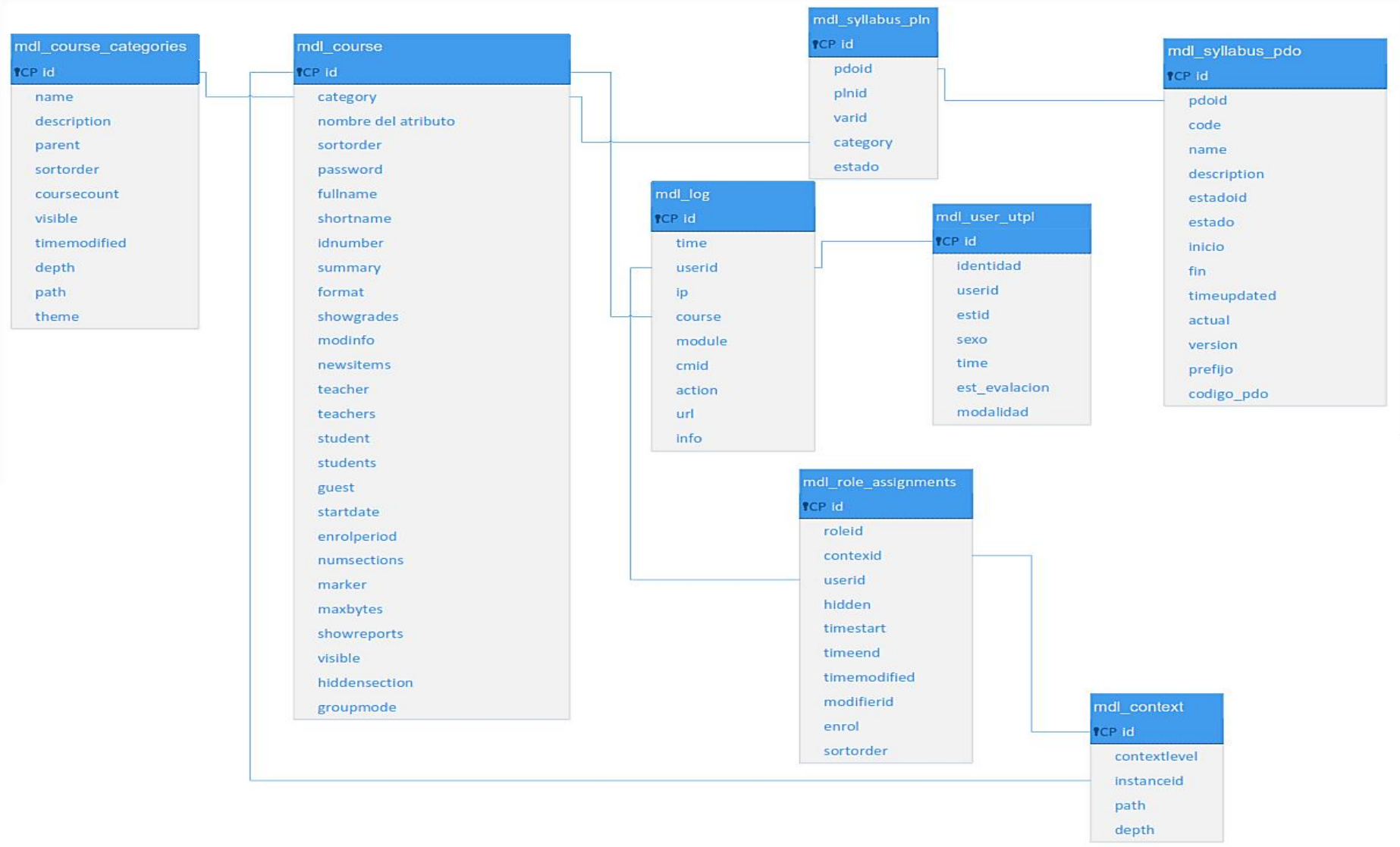
Velasco Laura, Rodenas Sonia, & Virseda Silvia. (2008). LA HUELLA DEL MAESTRO: Informe Delors: Los pilares de la educación. Retrieved from http://svirseda.blogspot.com/2013/11/informe-delors.html

WeLearning. (2015). La historia del e-Learning, en Infografía.

Zhao, Y. (2015). Introducción a la Minería de datos con R y de importación / exportación de datos en I - RDataMining.com: R y Minería de Datos. Retrieved from http://www.rdatamining.com/docs/introduction-to-data-mining-with-r-and-data-import-export-in-r

ANEXOS

Anexo 1. Modelo entidad relación de las variables de la base de datos SYLLABUS utilizadas en el proyecto



Anexo 2. Tipo de datos iniciales

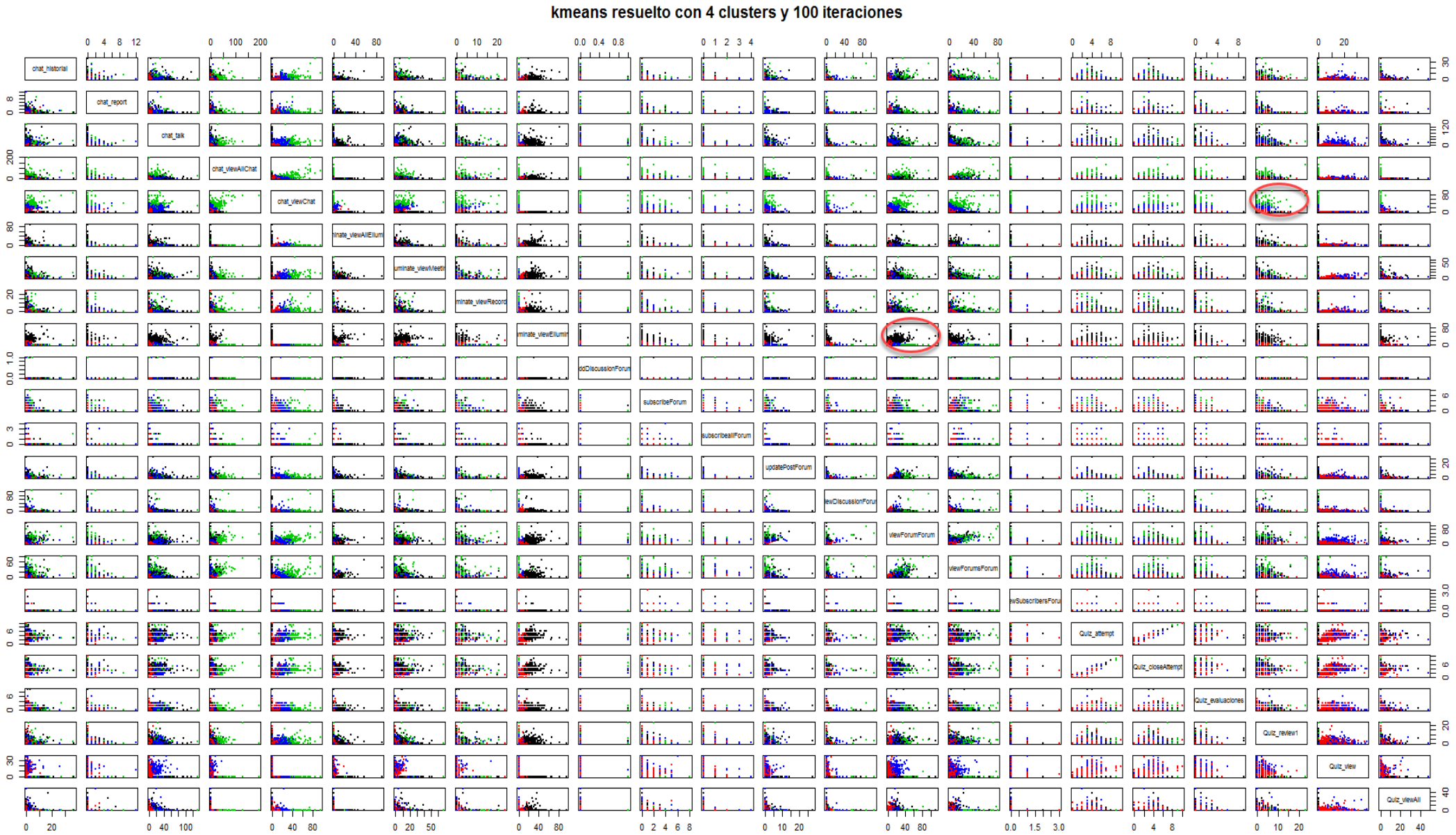
DATO	DESCRIPCION	TIPO DE DATO
Cedula del estudiante	Identificación del estudiante	Varchar
Genero del estudiante	Genero del estudiante: femenino o masculino	Varchar
Edad del estudiante	La edad del estudiante	Numérico
Estado civil del estudiante	Estado civil de un estudiante como: soltero, casado, divorciado, unión libre.	Varchar
Nivel académico del estudiante	Ciclo académico de un estudiante que va desde primero a décimo.	Numérico
Titulación del estudiante	Nombre de la titulación	Varchar
Tipo de matrícula del estudiante	Nombre del tipo de matrícula ya sea: especial, extraordinario u ordinaria	Varchar
Tipo del estudiante	Nombre del tipo de estudiante ya sea: nuevo, segundo en adelante, nuevo en la titulación	Varchar
Categoría	Nombre de las tres categorías de educación básica de MDA de la UTPL: Realidad Nacional, Expresión Oral y Escrita y Metodología de Estudio	Varchar
Curso	Nombre de la categoría con el nombre del paralelo	Varchar
Periodo	Nombre de los semestres de los estudiantes matriculados	Varchar
chat_historial	Número de chat que el usuario tiene como historial	Numérico
chat_report	Número de reportes que el usuario ha hecho	Numérico
chat_talk	Número de veces que un chat se ha abierto	Numérico
chat_viewAllChat	Total de chat vistos por el usuario	Numérico
chat_viewChat	Cuantas veces el usuario ha visto un chat	Numérico
Elluminate_viewAllElluminate	Total, de video llamadas asistidas por el estudiante	Numérico
Elluminate_viewMeeting	Número de participaciones en video llamadas	Numérico
Elluminate_viewRecording	Número de grabaciones que tiene un usuario para una video llamada	Numérico
Elluminate_viewElluminate	Número de las veces que se ha abierto una video llamada	Numérico
addDiscussionForum	Número de discusiones que el usuario a ingresado	Numérico
addPostForum	Número de anuncios que el usuario a ingresado	Numérico
deletePostForum	Número de anuncios que el usuario a borrado	Numérico

markReadForum	Número de los foros que se han leído por el usuario	Numérico
searchForum	Número de búsquedas que un usuario a echo la búsqueda de un foro.	Numérico
startTrackingForum	Número de foros a los que ha iniciado	Numérico
stopTrackingForum	Número de foros de los que ha salido	Numérico
subscribeForum	Número de foros a los que se ha inscrito un usuario	Numérico
subscribeallForum	Número de todos los foros a los q se ha inscrito el usuario	Numérico
unsubscribeForum	Número de los foros a los que no se ha inscrito el usuario	Numérico
unsubscribeAllForum	Total de foros a los que no se ha inscrito el usuario	Numérico
updatePostForum	Número de los anuncios que el usuario a actualizado	Numérico
viewDiscussionForum	Número de las veces que ha visto la discusión de un foro	Numérico
ViewForumForum	Número que el usuario se ha visto el foro de un foro	Numérico
viewForumsForum	Número que el usuario ha visto todos los foros de un determinado foro	Numérico
viewSubscribersForum	Número de suscritores a un foro	Numérico
Quiz_attempt	Número de intentos de prueba de un usuario	Numérico
Quiz_closeAttempt	Numero de cierres en un intento de prueba	Numérico
Quiz_continueAttemp	Número de intentos que el usuario a ingresado a la lección	Numérico
Quiz_evaluaciones	Número de evaluaciones en las que el usuario a participado	Numérico
Quiz_review1	Número de revisiones que el usuario a dado a una lección	Numérico
Quiz_view	Número de veces que el usuario a revisado un cuestionario	Numérico
Quiz_viewAll	Total de veces que el usuario ha visto todos los cuestionarios	Numérico

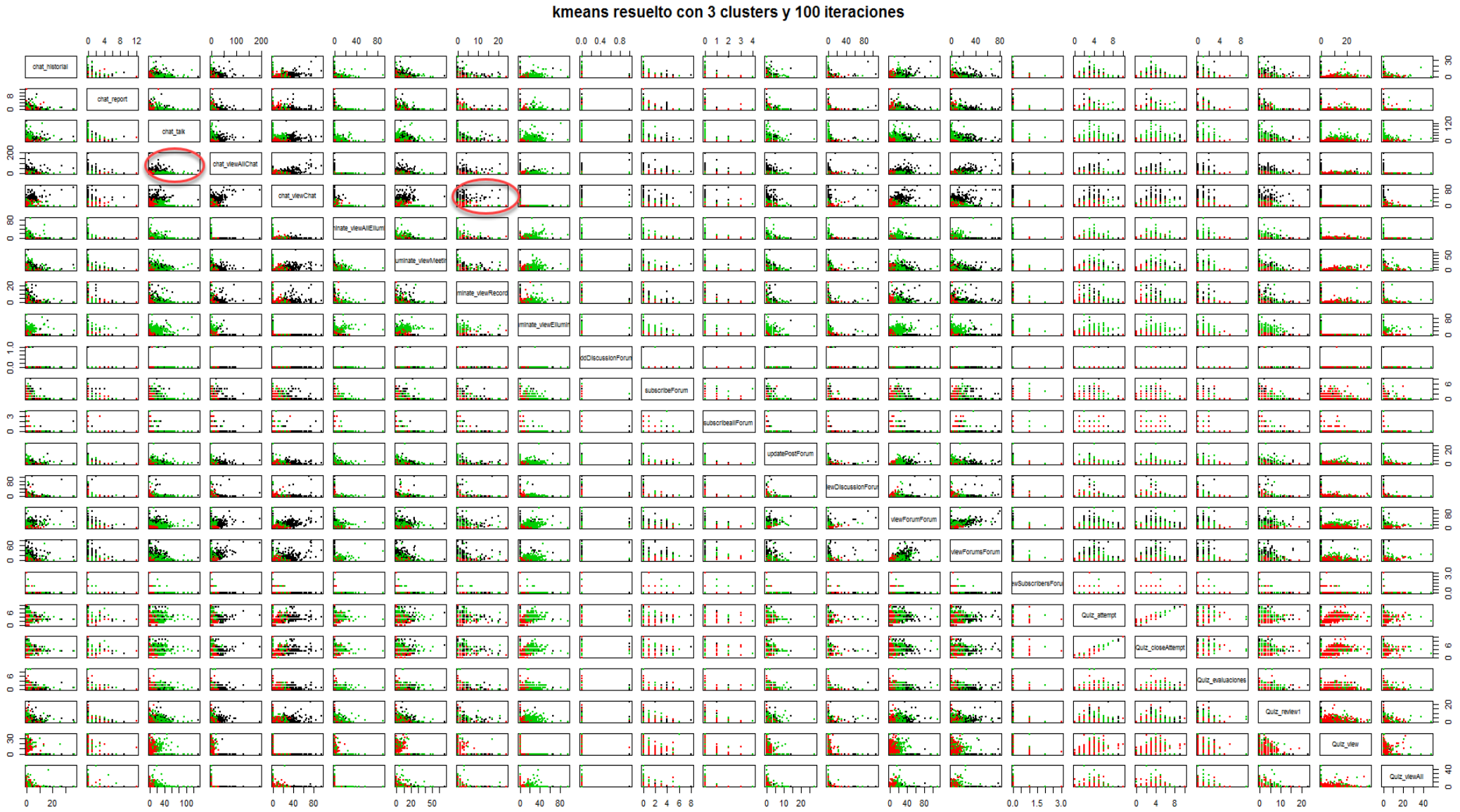
Anexo 3. Descripción de la nueva data

CAMPO	DESCRIPCION	TIPO DE DATO
IDENTIDAD	Cedula del estudiante	Varchar
GENERO	Genero del estudiante	Varchar
EDAD	Edad del estudiante	Int
ESTADO_CIVIL	Estado civil del estudiante	Varchar
NIVEL	Ciclo que curso el estudiante	Int
TITULACION	Carrera del estudiante	Varchar
TIPOMATRICULA	Nombre del tipo de matrícula del estudiante	Varchar
TIPO_ESTUDIANTE	Nombre del tipo de estudiante	Varchar
CATEGORIA	Nombre de la categoría de formación básica	Varchar
CURSO	Paralelo del curso del estudiante	Varchar
PERIODO	Nombre de los semestres	Varchar
chat_historial	Historial de chats	Int
chat_report	Reporte de un chat	Int
chat_talk	Número de veces que un chat se ha abierto	Int
chat_viewAllChat	Todos los chats vistos por el usuario	Int
chat_viewChat	Veces que el usuario ha visto un chat	Int
Elluminate_viewAllElluminate	Todos las video llamadas vistos por el usuario	Int
Elluminate_viewMeeting	Número de participaciones en video llamadas	Int
Elluminate_viewRecording	Número de grabaciones que tiene un usuario de una video llamada	Int
Elluminate_viewElluminate	Número de las veces que se ha abierto una video llamada	Int
AddDiscussionForum	Número de discusiones que el usuario a ingresado	Int
deletePostForum	Numero de anuncios que el usuario a borrado	Int
subscribeForum	Número de foros a los que se ha inscrito un usuario	Int
subscribeallForum	Número de todos los foros a los q se ha inscrito el usuario	Int
updatePostForum	Número de los anuncios que el usuario a actualizado	Int
viewDiscussionForum	Número de las veces que ha visto la discusión de un foro	Int
viewForumForum	Número de veces que el usuario ha revisado el foro dentro de otro foro	Int
viewForumsForum	Número que el usuario ha visto todos los foros de un determinado foro	Int
viewSubscribersForum	Número de suscritores a un foro	Int
Quiz_attempt	Número de intentos de prueba de un usuario	Int
Quiz_closeAttempt	Número de veces que se ha cerrado una lección	Int
Quiz_evaluaciones	Número de evaluaciones en las que el usuario a participado	Int
Quiz_review1	Número de revisiones que el usuario a dado a una lección	Int
Quiz_view	Número de veces que el usuario a revisado un cuestionario	Int
Quiz_viewAll	Total de veces que el usuario ha visto todos los cuestionarios	Int

Anexo 4. Resultado de plot con 4 cluster y 100 iteraciones



Anexo 5. Resultado de plot con 3 cluster y 100 iteraciones



Anexo 6. Correlaciones

```
> cor(datos)
                                chat_historial  chat_report    chat_talk chat_viewAllChat chat_viewChat Elluminate_viewAllElluminate Elluminate_viewMeeting Elluminate_viewRecording Elluminate_viewElluminate addDiscussionForum
chat_historial                     1.00000000  0.189086472  0.288604764      0.323671579   0.407794825                0.109877901           0.222713604             0.15384041              0.187791064        0.036487399
chat_report                        0.18908647  1.000000000  0.106664514      0.139537854   0.165306841               -0.006754155           0.072975649             0.07271843              0.056016247       -0.005212767
chat_talk                          0.28860476  0.106664514  1.000000000      0.232592423   0.323774499                0.097177810           0.372352973             0.14043056              0.330964223        0.061698121
chat_viewAllChat                   0.32367158  0.139537854  0.232592423      1.000000000   0.609389594               -0.048284834           0.206392515             0.26830268              0.045576839        0.259124884
chat_viewChat                      0.40779482  0.165306841  0.323774499      0.609389594   1.000000000               -0.009699191           0.343972588             0.28499290             -0.155973007        0.141430876
Elluminate_viewAllElluminate       0.10987790 -0.006754155  0.097177810     -0.048284834  -0.009699191                1.000000000           0.169059369             0.10280630              0.362431555        0.053619515
Elluminate_viewMeeting             0.22271360  0.072975649  0.372352973      0.206392515   0.343972588                0.169059369           1.000000000             0.19743792              0.483557641        0.047292618
Elluminate_viewRecording           0.15384041  0.072718431  0.140430556      0.268302677   0.284992897                0.102806298           0.197437922             1.00000000              0.188300268        0.029493110
Elluminate_viewElluminate          0.18779106  0.056016247  0.330964223      0.045576839  -0.155973007                0.362431555           0.483557641             0.18830027              1.000000000        0.009371639
addDiscussionForum                 0.03648740 -0.005212767  0.061698121      0.259124884   0.141430876                0.053619515           0.047292618             0.02949311              0.009371639        1.000000000
subscribeForum                     0.08256422  0.060977207  0.071770621      0.194606391   0.145674533                0.017195420           0.075870971             0.07077806              0.035557682        0.033720078
subscribeallForum                  0.06615673  0.028307253  0.005875161      0.083332869   0.067537607                0.011471876           0.004021121             0.02827463              0.008189409        0.033996341
updatePostForum                    0.18322516  0.044849849  0.237598005      0.187040474   0.159841065                0.081407415           0.185381290             0.10718255              0.214599224        0.037207826
viewDiscussionForum                0.16010926  0.051931298  0.051983480      0.188984423   0.178323150                0.011521454           0.115829717             0.14699136              0.040563046        0.072778397
viewForumForum                     0.40878149  0.124194175  0.454493230      0.427132586   0.406845804                0.161980732           0.390146111             0.26345152              0.370349009        0.077986415
viewForumsForum                    0.31649794  0.115497051  0.222026985      0.637050058   0.408887845                0.176265108           0.196023120             0.20509428              0.151794732        0.137618910
viewSubscribersForum               0.01822194  0.004591018  0.017582524      0.009472021   0.002282872               -0.005812982           0.004553173             0.02470108              0.007517974       -0.002883717
Quiz_attempt                       0.16628857  0.062653515  0.280820361      0.147777107   0.162968776                0.106368692           0.230549669             0.10099886              0.281054895        0.040071586
Quiz_closeAttempt                  0.16977677  0.067047804  0.285880669      0.151491794   0.168518494                0.102411413           0.235589976             0.10392626              0.282466995        0.041176766
Quiz_evaluaciones                  0.12399823  0.030151178  0.198334336      0.098564678   0.102574462                0.073712087           0.148857859             0.05620543              0.151710493        0.044367321
Quiz_review1                       0.18400609  0.057874143  0.238348823      0.213128916   0.207853989                0.090163795           0.148852780             0.13454723              0.171971014        0.106476540
Quiz_view                         -0.04680448 -0.003552623 -0.047642277     -0.093892838  -0.249732911               -0.066167804          -0.153875957            -0.06173711             -0.276265626       -0.008747205
Quiz_viewAll                       0.05494699  0.077896069  0.040856994     -0.051155308  -0.022231189               -0.032172760           0.053019078             0.03350995              0.087089961       -0.006707996
                                subscribeForum subscribeallForum updatePostForum viewDiscussionForum viewForumForum viewForumsForum viewSubscribersForum Quiz_attempt Quiz_closeAttempt Quiz_evaluaciones Quiz_review1    Quiz_view
chat_historial                      0.08256422       0.0661567343    0.1832251570         0.160109260     0.40878149     0.31649794          0.018221944   0.16628857        0.16977677        0.12399823  0.184006092 -0.046804483
chat_report                         0.06097721       0.0283072526    0.0448498487         0.051931298     0.12419418     0.11549705          0.004591018   0.06265351        0.06704780        0.03015118  0.057874143 -0.003552623
chat_talk                           0.07177062       0.0058751611    0.2375980054         0.051983480     0.45449323     0.22202699          0.017582524   0.28082036        0.28588067        0.19833434  0.238348823 -0.047642277
chat_viewAllChat                    0.19460639       0.0833328685    0.1870404745         0.188984423     0.42713259     0.63705006          0.009472021   0.14777711        0.15149179        0.09856468  0.213128916 -0.093892838
chat_viewChat                       0.14567453       0.0675376072    0.1598410648         0.178323150     0.40684580     0.40888785          0.002282872   0.16296878        0.16851849        0.10257446  0.207853989 -0.249732911
Elluminate_viewAllElluminate        0.01719542       0.0114718762    0.0814074146         0.011521454     0.16198073     0.17626511         -0.005812982   0.10636869        0.10241141        0.07371209  0.090163795 -0.066167804
Elluminate_viewMeeting              0.07587097       0.0040211211    0.1853812896         0.115829717     0.39014611     0.19602312          0.004553173   0.23054967        0.23558998        0.14885786  0.148852780 -0.153875957
Elluminate_viewRecording            0.07077806       0.0282746297    0.1071825497         0.146991359     0.26345152     0.20509428          0.024701081   0.10099886        0.10392626        0.05620543  0.134547234 -0.061737108
Elluminate_viewElluminate           0.03555768       0.0081894094    0.2145992243         0.040563046     0.37034901     0.15179473          0.007517974   0.28105489        0.28246700        0.15171049  0.171971014 -0.276265626
addDiscussionForum                  0.03372008       0.0339963408    0.0372078260         0.072778397     0.07798642     0.13761891         -0.002883717   0.04007159        0.04117677        0.04436732  0.106476540 -0.008747205
subscribeForum                      1.00000000       0.1731855870    0.0700925896         0.044357757     0.21278577     0.37420229          0.212234639   0.10516702        0.10727861        0.06259172  0.076765168  0.028933176
subscribeallForum                   0.17318559       1.0000000000   -0.0004965868         0.005611038     0.07865173     0.21070481          0.160374313   0.04790120        0.04948405        0.03947773  0.071036930 -0.003165015
updatePostForum                     0.07009259      -0.0004965868    1.0000000000         0.069680274     0.49250990     0.22512589          0.008627168   0.15125699        0.15600028        0.13304450  0.194996469 -0.040064814
viewDiscussionForum                 0.04435776       0.0056110379    0.0696802738         1.000000000     0.26063788     0.25339578          0.053474364   0.04458574        0.04590330        0.02065091  0.101817537 -0.024052049
viewForumForum                      0.21278577       0.0786517312    0.4925099015         0.260637881     1.00000000     0.65465325          0.061606829   0.37254145        0.37929612        0.25494900  0.380241739 -0.048349705
viewForumsForum                     0.37420229       0.2107048065    0.2251258936         0.253395785     0.65465325     1.00000000          0.052992317   0.22954949        0.23328163        0.15229482  0.278167144 -0.033170686
viewSubscribersForum                0.21223464       0.1603743126    0.0086271675         0.053474364     0.06160683     0.05299232          1.000000000   0.03977263        0.03870823        0.02633503  0.014692761  0.027561509
Quiz_attempt                        0.10516702       0.0479011981    0.1512569915         0.044585737     0.37254145     0.22954949          0.039772628   1.00000000        0.98785019        0.40584656  0.231126398  0.363873676
Quiz_closeAttempt                   0.10727861       0.0494840495    0.1560002809         0.045903301     0.37929612     0.23328163          0.038708229   0.98785019        1.00000000        0.41177384  0.233495083  0.351571683
Quiz_evaluaciones                   0.06259172       0.0394777283    0.1330445020         0.020650913     0.25494900     0.15229482          0.026335033   0.40584656        0.41177384        1.00000000  0.177654857  0.139417260
Quiz_review1                        0.07676517       0.0710369297    0.1949964694         0.101817537     0.38024174     0.27816714          0.014692761   0.23112640        0.23349508        0.17765486  1.000000000 -0.004476716
Quiz_view                           0.02893318      -0.0031650148   -0.0400648141        -0.024052049    -0.04834970    -0.03317069          0.027561509   0.36387368        0.35157168        0.13941726 -0.004476716  1.000000000
Quiz_viewAll                        0.04413209      -0.0121377564    0.0202879705         0.006890566     0.07759756     0.10609648         -0.012083510   0.09575187        0.09126796        0.04007634  0.074364736  0.109067406
                                Quiz_viewAll
chat_historial                 0.054946989
chat_report                    0.077896069
chat_talk                      0.040856994
chat_viewAllChat              -0.051155308
chat_viewChat                 -0.022231189
Elluminate_viewAllElluminate  -0.032172760
Elluminate_viewMeeting         0.053019078
Elluminate_viewRecording       0.033509951
Elluminate_viewElluminate      0.087089961
addDiscussionForum            -0.006707996
subscribeForum                 0.044132086
subscribeallForum             -0.012137756
updatePostForum                0.020287971
viewDiscussionForum            0.006890566
viewForumForum                 0.077597558
viewForumsForum                0.106096477
viewSubscribersForum          -0.012083510
Quiz_attempt                   0.095751866
Quiz_closeAttempt              0.091267960
Quiz_evaluaciones              0.040076337
Quiz_review1                   0.074364736
Quiz_view                      0.109067406
Quiz_viewAll                   1.000000000
```

Anexo 7. Matriz de covarianza

```
> acp
Standard deviations:
 [1] 14.30422037  9.49344767  7.73525400  6.20003329  5.29789250  4.05795757  3.43827119  3.26880122  2.83541513  2.35598779  1.68059295
[12]  1.60453899  1.51063003  1.32706970  1.12992877  0.86801363  0.64897399  0.52074120  0.41254667  0.15705106  0.15030206  0.09277866
[23]  0.03463031

Rotation:
                                      PC1           PC2           PC3           PC4           PC5           PC6           PC7           PC8
chat_historial                0.0599159079  1.160161e-02 -4.507566e-03  0.0125842126 -0.0241842708  0.0405787935 -0.0085042216 -0.0063347344
chat_report                   0.0051642050  2.052554e-03  8.281718e-04  0.0019107884  0.0005233009  0.0050640897 -0.0004891553 -0.0004905523
chat_talk                     0.4857779006 -4.166515e-01  7.021506e-01 -0.1674777856  0.1995157930 -0.1553275054 -0.0480425763 -0.0411195287
chat_viewAllChat              0.3457185080  3.821587e-01 -1.242217e-01  0.1991402341  0.6756706400  0.0779016359  0.3941705185 -0.0470886932
chat_viewChat                 0.4010171731  5.255657e-01  2.842646e-01  0.3345057954 -0.3946314722  0.2037850975 -0.1391293436  0.0907252392
Elluminate_viewAllElluminate  0.0318225637 -7.780199e-02 -6.976879e-02  0.0600014411 -0.0317398833  0.0284376872 -0.3850870993  0.1437026603
Elluminate_viewMeeting        0.1525039064 -1.018544e-01 -7.996888e-03  0.2379471804 -0.1557217193  0.4885006914 -0.1906157776 -0.0839208841
Elluminate_viewRecording      0.0294136537  7.302769e-03 -1.355898e-02  0.0276200973 -0.0052976685  0.0386100871  0.0059682530 -0.0208431686
Elluminate_viewElluminate     0.2042172117 -5.903463e-01 -3.427607e-01  0.5084698277  0.1184500077  0.1816266762 -0.0066871599  0.0262954074
addDiscussionForum            0.0003890685  4.523480e-04 -7.162866e-05  0.0003071512  0.0010516959  0.0002656101  0.0003243159 -0.0003698490
subscribeForum                0.0091306397  5.763285e-03 -9.949759e-03 -0.0080442582  0.0084259953 -0.0057561178 -0.0323340876  0.0119901964
subscribeallForum             0.0010434885  1.117585e-03 -1.740211e-03 -0.0008314313  0.0014037581 -0.0021984276 -0.0068901555  0.0028694192
updatePostForum               0.0285401356 -1.133521e-02 -1.989933e-02 -0.0201687404 -0.0266762063 -0.0019754354  0.0599485566  0.0137026291
viewDiscussionForum           0.0598236501  4.029116e-02 -7.429764e-02 -0.0346811748 -0.0492921501 -0.0058268746 -0.1354048979 -0.9691290846
viewForumForum                0.5574310394 -9.279582e-02 -4.139857e-01 -0.4206886412 -0.4224116105 -0.0649677889  0.3516158855  0.0487254894
viewForumsForum               0.2999965410  1.646437e-01 -3.309483e-01 -0.1773663034  0.2729581779 -0.2827744870 -0.6891146174  0.1062371784
viewSubscribersForum          0.0002670942 -6.652231e-05 -4.270889e-04 -0.0009012959 -0.0001913067 -0.0001020137 -0.0003686755 -0.0010507060
Quiz_attempt                  0.0357384457 -2.336419e-02 -1.204496e-02 -0.0432432443  0.0062076471  0.1571284342 -0.0242798863  0.0442621308
Quiz_closeAttempt             0.0363093139 -2.320526e-02 -1.185592e-02 -0.0422612089  0.0049118437  0.1535531640 -0.0225477670  0.0437113038
Quiz_evaluaciones             0.0126610661 -7.500938e-03 -2.655713e-03 -0.0137584141 -0.0011735268  0.0318856470 -0.0024097181  0.0142146315
Quiz_review1                  0.0461448860 -6.893291e-03 -1.889636e-02 -0.0236562593 -0.0128719018  0.0085485389  0.0085419462  0.0129600670
Quiz_view                    -0.0556145288  7.971994e-03  1.827310e-02 -0.5343171744  0.2272022205  0.7195035764 -0.0930642628  0.0343381690
Quiz_viewAll                  0.0066051544 -1.586739e-02 -1.458608e-02 -0.0264838995 -0.0049582188  0.0371782537 -0.0825429378  0.0192277622
                                      PC9          PC10          PC11          PC12          PC13          PC14          PC15          PC16
chat_historial                1.261227e-01 -0.0593115695 -0.1487117948  0.3827154685 -7.788607e-01  0.3769440047 -0.2354303058  0.0181675941
chat_report                   8.836597e-03 -0.0152097200 -0.0180177534  0.0179658404 -2.069244e-02  0.0019861686  0.0004245414 -0.0061089950
chat_talk                    -2.426580e-02  0.0231638599 -0.0025384131  0.0041516660  1.118814e-02 -0.0005056947 -0.0116509114  0.0003919318
chat_viewAllChat             -2.595732e-02  0.2082463294 -0.0840952337 -0.0206534694  4.049983e-02  0.0734875335 -0.0127367802 -0.0162062714
chat_viewChat                 2.693380e-01 -0.1900845048  0.0468274937  0.0360789823  3.370795e-02 -0.1264279205  0.1260924424  0.0008160043
Elluminate_viewAllElluminate  4.130538e-01  0.7736741949 -0.1590527021 -0.0329076565  8.466850e-02  0.0999457305 -0.0201198664 -0.0116834800
Elluminate_viewMeeting       -7.290111e-01  0.2344545595 -0.0276934578 -0.0471668820 -4.498652e-02  0.0841871968 -0.0686659473  0.0080496320
Elluminate_viewRecording      6.311485e-02 -0.0055407697 -0.0493741769  0.0088716429  5.204755e-02 -0.4934817067 -0.8606864412  0.0505467125
Elluminate_viewElluminate     2.723369e-01 -0.2558523073  0.0576748146  0.0607695295 -6.009448e-03 -0.1465509043  0.1326437318 -0.0066641606
addDiscussionForum            4.610780e-04  0.0013471115 -0.0008045671 -0.0018740265  5.450452e-04  0.0005289471  0.0014724848 -0.0001420247
subscribeForum               -1.361044e-02 -0.0274635938  0.0198896934  0.0045517169  5.328129e-03 -0.0176223095  0.0025450586  0.0170424612
subscribeallForum            -9.316261e-04 -0.0059508645  0.0064448825 -0.0021933958 -6.205000e-03 -0.0013569113 -0.0018480247 -0.0006653621
updatePostForum               4.449966e-03  0.0347195240 -0.0232234353 -0.0048067084 -8.777097e-04 -0.0122076352  0.0674833465  0.9927898461
viewDiscussionForum           1.559266e-01  0.0173824221  0.0014185884 -0.0206995293  3.689311e-02  0.0291561559  0.0223491467  0.0145479285
viewForumForum               -2.271288e-02  0.1161456948 -0.0260284647  0.0355483619  4.767528e-02 -0.0002875866  0.0104329479 -0.0724445457
viewForumsForum              -1.741319e-01 -0.2241515873  0.1131290177  0.0160749401 -3.856153e-02 -0.0462948363 -0.0013025015  0.0401113025
viewSubscribersForum         -1.902065e-05 -0.0006864941  0.0020073380  0.0002101243  4.287027e-05 -0.0001995390 -0.0019950950 -0.0015124894
Quiz_attempt                  1.261558e-01 -0.1530610113  0.1990207799 -0.1953442681  2.176670e-01  0.4730951310 -0.2518639834  0.0246035890
Quiz_closeAttempt             1.221273e-01 -0.1525120482  0.2016509485 -0.1963465840  2.158459e-01  0.4764761312 -0.2587124905  0.0280129780
Quiz_evaluaciones             2.242919e-02 -0.0231316501  0.0386166271 -0.0650646197  2.788134e-02  0.1228480316 -0.0674079725  0.0434309518
Quiz_review1                  6.672476e-02 -0.0436653348 -0.1609544569 -0.8672198776 -4.438555e-01 -0.1110338424  0.0426331718 -0.0155232395
Quiz_view                     1.966015e-01 -0.0810305172  0.0195458793  0.0948479827 -5.604993e-02 -0.2260537014  0.1508228099 -0.0060625869
Quiz_viewAll                 -1.389893e-02 -0.2746911098 -0.9021913355  0.0191666010  2.793287e-01  0.1470460472 -0.0149383777 -0.0045598171
                                      PC17          PC18          PC19          PC20          PC21          PC22          PC23
chat_historial               -0.0164934824 -1.062395e-02  0.0295414369  3.494565e-03 -0.0014929405  2.564835e-04 -1.130449e-03
chat_report                  -0.0071496972 -5.140279e-02 -0.9978268732  3.731286e-03 -0.0116610460 -1.268786e-03 -2.929529e-03
chat_talk                    -0.0037153135 -2.736589e-04  0.0006348910 -3.751585e-04 -0.0003836583  1.140846e-04  6.985484e-05
chat_viewAllChat             -0.0020946266 -1.169580e-02  0.0011805364 -2.893526e-03 -0.0002752823  2.710174e-04  1.694218e-03
chat_viewChat                 0.0089523212  8.212580e-03  0.0085661351  8.924391e-04  0.0016826325 -7.827326e-04 -3.603160e-05
Elluminate_viewAllElluminate -0.0071001893 -1.850575e-02 -0.0060684726 -3.084907e-03  0.0012873452  2.450076e-04  1.289556e-03
Elluminate_viewMeeting       -0.0089673024  5.550326e-04 -0.0057434007 -8.478628e-04 -0.0017123731 -4.559748e-05  6.617878e-06
Elluminate_viewRecording     -0.0011023339  5.580143e-03 -0.0008739745  1.495869e-03 -0.0026206232  1.295341e-03 -1.437652e-03
Elluminate_viewElluminate     0.0103216138  6.293210e-03  0.0073815735  8.017480e-04  0.0021237643 -5.733439e-04 -8.207951e-05
addDiscussionForum            0.0010492765 -1.404855e-04  0.0028509375 -6.062097e-03  0.0016184889  3.132087e-03 -9.999654e-01
subscribeForum                0.0093355273 -9.955730e-01  0.0511029993  3.549906e-02 -0.0003550012  3.619559e-02  1.168674e-04
subscribeallForum             0.0039165144 -3.008661e-02 -0.0018283970 -9.902653e-01 -0.0413006816  1.284899e-01  6.324153e-03
updatePostForum              -0.0373178271  1.359493e-02 -0.0066192365 -2.196093e-03 -0.0019401681 -2.023706e-03 -4.115565e-05
viewDiscussionForum           0.0043723488 -8.447129e-03  0.0010754914 -1.689390e-03  0.0001036342  1.642478e-03  5.210457e-04
viewForumForum               -0.0038473313 -6.206590e-03 -0.0002593376 -2.046533e-03 -0.0006426713  1.109840e-03 -1.274203e-04
viewForumsForum               0.0001562518  4.837237e-02 -0.0016080514  8.080326e-03  0.0002394132 -2.174278e-03 -5.135493e-04
viewSubscribersForum          0.0012959894 -4.032196e-02  0.0030548869 -1.266007e-01 -0.0179046704 -9.909571e-01 -2.355503e-03
Quiz_attempt                 -0.1241933008 -4.389881e-05  0.0026090623  3.121634e-02 -0.7071335263  9.488564e-03 -1.359533e-03
Quiz_closeAttempt            -0.1123834611 -4.104693e-03 -0.0140905479 -2.973549e-02  0.7054587349 -8.111781e-03  1.194450e-03
Quiz_evaluaciones             0.9844008226  8.560485e-03 -0.0091483764  4.377666e-03 -0.0087596307  6.337406e-04  1.010418e-03
Quiz_review1                 -0.0249756922 -7.173603e-03 -0.0016067600  3.995805e-03  0.0001641793 -7.752734e-04  1.430482e-03
Quiz_view                     0.0091321236  9.660651e-03  0.0070498150 -4.246394e-05  0.0044151876 -2.303836e-04  9.956786e-05
Quiz_viewAll                  0.0013860767 -7.805742e-03  0.0157265949 -5.522180e-03  0.0018309682 -5.337425e-04  5.634393e-04
>
```