



UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

MODALIDAD PRESENCIAL

ESCUELA DE CIENCIAS DE LA COMPUTACIÓN

Técnicas de minería de datos para identificar patrones de colaboración de los estudiantes que hacen uso del EVA de la UTPL.

Trabajo de fin de carrera previa a la obtención del título de ingeniero en sistemas informáticos y computación.

Autora:

Pulla Elizalde, Cinthia Elizabeth

Director:

Ing. Cueva Carrión Samanta Patricia

Coodirector:

Ing. Valdiviezo Díaz Priscila Marisela

LOJA - ECUADOR

2011

CERTIFICACIÓN

Ing. Samanta Cueva

DIRECTOR DE TESIS

CERTIFICA:

*Haber dirigido y supervisado el desarrollo del presente proyecto de tesis previo a la obtención del título de **INGENIERA EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN**, y una vez que este cumple con todas las exigencias y los requisitos legales establecidos por la Universidad Técnica Particular de Loja, autoriza su presentación para los fines legales pertinentes.*

Loja, Diciembre del 2011

Ing. Samanta Cueva Carrión

CESIÓN DE DERECHOS

Yo, Cinthia Elizabeth Pulla Elizalde declaro ser autora del presente trabajo y eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones legales.

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

Cinthia Pulla Elizalde

AUTORÍA

El presente proyecto de tesis con cada una de sus observaciones, análisis, evaluaciones, conclusiones y recomendaciones emitidas, es de absoluta responsabilidad del autor.

Además, es necesario indicar que la información de otros autores empleada en el presente trabajo está debidamente especificada en fuentes de referencia y apartados bibliográficos.

Cinthia Pulla Elizalde

DEDICATORIA

Con un profundo cariño quiero dedicar esta tesis:

A Dios por ser la luz que guía mi existencia, por las bendiciones que le ha dado a mi vida, por las personas maravillosas que están en ella y por tomarme en sus brazos y no dejarme caer.

A mis padres Blanca y Wilson por todo el esfuerzo que han hecho siempre por mí, por sus consejos, por el apoyo que me han dado a lo largo de mi vida, por los valores que me han inculcado que me hacen ser la mujer que hoy soy.

A mis hermanas Gaby y Hillary que de una u otra forma me han brindado su apoyo a veces sin siquiera saberlo y por tolerar a la hermana que les ha tocado.

A Wilsito que desde que naciste hermanito has llenado de alegría mi corazón.

A mis abuelitas Ulvia y Rosa en las que siempre he encontrado una palabra de aliento, y por tenerme siempre presente en sus oraciones.

A Pablo por estar siempre ahí con una sonrisa para mí en este largo camino, por brindarme su amor y por su dulzura que llena mi vida.

A mis amigos que puedo contarlos con la palma de mi mano pero sé que puedo contar con ellos siempre sin condición.

Gracias a todos, que Dios los cuide siempre y me de la bendición de tenerlos siempre a mi lado.

Cinthia.

AGRADECIMIENTO

Al finalizar mi carrera tengo sentimientos encontrados, alegría al ver una de mis aspiraciones cristalizadas y tristeza porque aquí quedan cinco años de mi vida, finalizando una etapa muy importante de ella.

Son varias las personas a las que agradezco profundamente, personas que intervinieron tanto en mi proyecto de tesis como a lo largo de mi vida universitaria.

Agradezco a la ingeniera Samanta Cueva por el apoyo y dirección que he recibido de ella a lo largo de este proyecto.

A la ingeniera Priscila Valdiviezo que creyó en mis capacidades cuando opté por su tema propuesto y quien junto a mi directora me supieron guiar.

Al ingeniero Guido Riofrío por sus indicaciones y predisposición para orientarme en este campo.

A los ingenieros: Jorge López, Marta Agila, Inesita Jara, Juan Carlos Morocho, Nelson Piedra, Daniel Guamán, Rommel Torres, Julia Pineda, referentes en cada uno de los campos en los que se han especializado y de quienes desde el inicio de mi carrera me han brindado grandes enseñanzas.

Índice de contenidos

CERTIFICACIÓN.....	ii
CESIÓN DE DERECHOS.....	iii
AUTORÍA.....	iv
DEDICATORIA.....	v
AGRADECIMIENTO.....	vi
ÍNDICE DE GRÁFICAS.....	ix
ÍNDICE DE TABLAS.....	xiii
ECUACIONES.....	xiv
OBJETIVOS Y RESUMEN DE LA TESIS.....	xv
CAPÍTULO 1: SITUACIÓN ACTUAL DE LOS ENTORNOS COLABORATIVOS..	1
1.1. INTRODUCCIÓN.....	2
1.2. MOODLE.....	2
1.3. MODELADO DE USUARIO.....	3
1.3.1. Representaciones del Modelo.....	4
1.3.2. Soporte al aprendizaje colaborativo.....	5
1.4. TRABAJOS RELACIONADOS – SOPORTE A LA COLABORACIÓN.....	6
1.4.1. Sistema Hipermedia De Aprendizaje Colaborativo Adaptativo (SHACA) ...	6
1.4.2. Modelo de Estudiante Colaborativo (MEC).....	7
1.4.3. Soporte Adaptativo Al Aprendizaje Colaborativo e Individual (ASCIL).....	11
1.4.4. Sistema de Soporte a la Colaboración (CSCL).....	13
1.5. CONSTRUCCIÓN DE UN MODELO COLABORATIVO.....	13
1.6. MINERÍA DE DATOS EN LA EDUCACIÓN.....	14
1.7. MÉTODOS Y HERRAMIENTAS PARA EL ANÁLISIS COLABORATIVO...21	
1.8. ESTUDIO DE LOS ALGORITMOS DE AGRUPAMIENTO.....	26
1.9. HERRAMIENTAS PARA LA MINERÍA DE DATOS.....	31
CAPÍTULO II:	
MINERÍA DE DATOS APLICADA A ENTORNOS COLABORATIVOS DEL	
EVA.....	38

2.1.	Introducción.....	39
2.2.	Metodología.....	39
2.3.	Integración y Recopilación.....	41
2.3.1.	Selección de Tablas.....	41
2.3.2.	Selección de Materias.....	43
2.4.	PREPROCESAMIENTO.....	47
2.4.1.	Selección, Limpieza y transformación.....	47
2.5.	MINERÍA DE DATOS Y EXTRACCIÓN DE PATRONES.....	50
2.5.1.	Prueba de Carga de Datos en WEKA.....	50
2.5.2.1.1.	Tercera Experimentación.....	52
2.5.2.1.1.1.	Descripción del Procedimiento.....	52
2.5.2.1.1.1.1.	Fundamentos de la Programación.....	52
2.5.2.1.1.1.2.	Algoritmo K-Means.....	55
2.5.2.1.1.1.3.	Algoritmo Expectation Maximation	59
2.5.2.1.1.1.4.	Algoritmo Cluster Jerárquico.....	61
2.5.2.1.1.2.	Lógica de la Programación	63
2.5.2.1.1.2.1.	Algoritmo K-Means.....	64
2.5.2.1.1.2.2.	Algoritmo Expectation Maximation	67
2.5.2.1.1.2.3.	Algoritmo Cluster Jerárquico.....	68
2.5.2.1.1.3.	Fundamentos Informáticos	70
2.5.2.1.1.3.1.	Algoritmo K-Means.....	72
2.5.2.1.1.3.2.	Algoritmo Expectation Maximation	75
2.5.2.1.1.3.3.	Algoritmo Cluster Jerárquico.....	76
2.6.	EVALUACIÓN E INTERPRETACIÓN.....	78
2.6.1.	Fundamentos de la Programación.....	78
2.6.2.	Lógica de la Programación.....	81
2.6.3.	Fundamentos Informáticos.....	84
	DISCUSIÓN.....	88
	CONCLUSIONES Y RECOMENDACIONES.....	92
	BIBLIOGRAFÍA.....	96
	ANEXOS.....	102

ÍNDICE DE GRÁFICAS

Capítulo I

Figura 1. 1 Arquitectura del Sistema SHACA, (Arteaga & Fabregat, 2002)	7
Figura 1. 2 Componentes del Modelo de Estudiante Colaborativo (MEC). (Duran, 2006)	9
Figura 1. 3 Arquitectura de ASCIL. (Arteaga & Fabregat, 2002)	12
Figura 1. 4 Tareas del EDM. (Corso & Alfaro, 2010)	20
Figura 1. 5 Comparativa de cálculo de distancias (Cruz, 2010)	28
Figura 1. 6 Algoritmo K-Means (Saharkhiz, 2009)	29
Figura 1. 7 Ventana de Creación de Foros en moodle	33
Figura 1. 8 Tipos de Foros existentes en moodle	34
Figura 1. 9 Tipo de calificaciones para Foros	35

Capítulo II

Figura 2. 1 Fases del Proceso de Extracción de Conocimiento (Hernández, Ramírez, & Ferri, 2004)	39
Figura 2. 2 Diagrama Entidad Relación Foros y logs	42
Figura 2. 3 Nodos de Interacción del EVA	43
Figura 2. 4 Búsqueda de periodo en tabla prefix_periodo_utpl	46
Figura 2. 5 Planes Académicos correspondientes al periodo Octubre-Febrero 2011	47
Figura 2. 6 Foros de Aprendizaje curso "Fundamentos de la Programación"	52
Figura 2. 7 Tabla prefix_forum del curso Fundamentos de la Programación	53
Figura 2. 8 Simulación de Creación de un nuevo foro	53
Figura 2. 9 Vista General del curso Fundamentos de la Programación	54
Figura 2. 10 Vista General II del curso Fundamentos de la Programación	55
Figura 2. 11 Relación sexo_usr – num_respuestas_post	58
Figura 2. 12 Relación número de acceso - promedio_foros	58
Figura 2. 13 Resultados Ejecución Algoritmo EM	60
Figura 2. 14 Esquema Algoritmo de Clustering Jerárquico	61
Figura 2. 15 Resultados ejecución Algoritmo Cluster Jerárquico	62
Figura 2. 16 Dendograma - Algoritmo de Clustering Jerárquico	62
Figura 2. 17 Foros de Aprendizaje curso "Lógica de la Programación"	63
Figura 2. 18 Tabla prefix_forum del curso Lógica de la Programación	64
Figura 2. 19 Vista General del curso Lógica de la Programación	64
Figura 2. 21 Relación sexo_usr – Cluster	66
Figura 2. 22 Relación Número de acceso a foros- Número de posts	67
Figura 2. 23 Resultados del Algoritmo EM	68
Figura 2. 24 Resultados ejecución Algoritmo Cluster Jerárquico	69
Figura 2. 25 Dendograma –Algoritmo de Clustering Jerárquico	70
Figura 2. 26 Foros de Aprendizaje curso "Fundamentos de la Programación"	70
Figura 2. 27 Tabla prefix_forum del curso Fundamentos Informáticos	71
Figura 2. 28 Vista General del curso Fundamentos Informáticos	71
Figura 2. 30 Relación sexo_usr - Cluster	74
Figura 2. 31 Relación num_acceso_foros - num_respuestas_post	74

Figura 2. 32 Resultados de Ejecución Algoritmo EM	75
Figura 2. 33 Esquema de Aplicación Clustering Jerárquico	76
Figura 2. 34 Dendograma - Algoritmo Clustering Jerárquico	77
Figura 2. 35 Gráfica por criterios de colaboración en Foros del curso "Fundamentos de la Programación"	81
Figura 2. 36 Gráfica por criterios de colaboración en Foros del curso "Lógica de la Programación"	84
Figura 2. 37 Gráfica por criterios de colaboración en Foros del curso "Fundamentos Informáticos"	86

Anexo A

Figura A.1 Directorio del Servidor XAMPP	103
Figura A.2 Path Archivo de Configuración MOODLE	103
Figura A.3 Archivo conFigura.php	104
Figura A.4 Tabla prefix_conFigura en PHPMyAdmin	105
Figura A.5 Modificación campo "alternateloginurl"	105
Figura A.6 Interfaz Moodle "Categorías"	106

Anexo C

Figura C.1 Creación de tabla prefix_log_fundProgramacion a partir de prefix_log	119
Figura C.2 Vista Previa de la tabla Generada	119
Figura C.3 Exportación de la tabla prefix_log_fundProgramacion en formato CSV	120
Figura C.4 Archivo CSV con los datos filtrados, tabla prefix_log_fundProgramacion	120
Figura C.5 Procedimiento creación de tabla prefix_log_logProgramacion y exportación de archivo .CSV	121
Figura C.6 Tabla prefix_log_fundProgrmacion_test	122
Figura C.7 Vista Previa de la tabla Generada	122

Anexo D

Figura D.1 Consulta para la Obtención de todos los alumnos pertenecientes a un curso.	124
Figura D.2 Consulta para obtener el número de acceso a FOROS.	125
Figura D.3 Consulta para obtener los subtemas leídos por un estudiante.	125
Figura D.4 Consulta para obtener el número de mensajes que un usuario ha agregado.	126
Figura D.5 Recuperación del número de debates agregados.	126
Figura D.6 Consulta para obtener el número de mensajes actualizado por un usuario	127
Figura D.7 Consulta para obtener el número de archivos adjuntos a un post	127
Figura D. 8 Consulta para obtener el número de suscripciones realizadas por un usuario.	128
Figura D.9 Tiempo Promedio de Interacción	128
Figura D.10 Panel de Administración del Curso - "Calificaciones "	129
Figura D.11 Exportación de Calificaciones a Excel	129
Figura D.12 Selección de Items para el reporte	130
Figura D.13 Descarga de Archivo.	130
Figura D.14 Resultado Generación de Reporte	131
Figura D.15 Cálculo de Promedios de Foros.	131

Figura D.16 Herramienta DREAMCODER for MYSQL.....	132
Figura D.17 Creación de atributos.....	132
Figura D.18 Importación de tabla desde Excel.....	133

Anexo E

Figura E.1 Matriz en EXCEL, recopilación de atributos, Fundamentos de la Programación.	135
Figura E.2 Matriz en EXCEL, recopilación de atributos, Lógica de la Programación.	135
Figura E.3 Matriz en EXCEL recopilación de atributos, Fundamentos Informáticos.	136

Anexo F

Figura F. 1 Path MySQL Connector.....	138
Figura F. 3 Parámetros de Conexión con la Base de Datos.....	138
Figura F. 2 URL Base de Datos.....	138
Figura F. 4 Informe de Conexión.....	139
Figura F. 5 Búsqueda General del Registro de Logs.....	139
Figura F. 6 Mensaje de error por tamaño de Memoria.....	139
Figura F. 7 Archivo RunWeka, edición campo MaxHeap.....	140
Figura F. 8 Consulta tabla Logs con Filtros.....	140
Figura F. 9 Ruta DatabaseUtils.props.....	141
Figura F. 10 Edición archivo DataBaseUtils.props.....	141

Anexo G

Figura G.1 Selección de archivos CSV.....	143
Figura G.2 Vista WEKA- Selección de Pestaña Cluster.....	144
Figura G.3 Selección de Algoritmo K-MEANS (Clustering).....	144
Figura G.4 Propiedades del algoritmo, Selección Número de Clusters.....	145

Anexo H

Figura H.1 Fórmula Excel para la Conversión de horas en formato hh:mm:ss a horas.....	146
Figura H.2 Vista Previa Resultado de Conversión.....	146

Anexo I

Figura I.1 Origen de Datos Herramienta Clementine.....	148
Figura I.2 Filtro de Variables.....	148

Figura 1.3 Verificación del tipo de dato del Origen	149
Figura 1.4 Ajuste y Conexión del ícono del Algoritmo K-MEANS.....	149
Figura 1.5 Verificación de Resultados	150
Figura 1.6 Pestaña "Model ", desviación estándar.....	150
Figura 1.7 Pestaña "Viewer", gráfica de atributos y Clusters.....	151
Figura 1.8 Sumario del Proyecto.....	151
Figura 1.9 Resultados diagrama proceso de Clustering K-MEANS.....	152
Figura 1.10 Resultados de la experimentación y Agrupamiento.....	152

Anexo J

Figura J.1 Vista General de la Asignatura "Fundamentos de la Programación"	155
Figura J.2 Salida de Información "Fundamentos de la Programación".....	155
Figura J.3 Relación número de interacciones, nota final para Fundamentos de la Programación	156
Figura J.4 Vista General de la Asignatura "Lógica de la Programación"	157
Figura J.5 Salida de Información "Lógica de la Programación"	157
Figura J.6 Relación número de interacciones, nota final para Lógica de la Programación	158
Figura J.7 Vista General de la Asignatura "Fundamentos Informáticos"	159
Figura J.8 Salida de Información "Fundamentos Informáticos"	159
Figura J.9 Relación número de interacciones, nota final para Fundamentos Informáticos.	160
Figura J.10 Diagrama Genérico para la extracción de Conocimiento en Clementine	161
Figura J.11 Resultado Gráfico curso "Fundamentos de la Programación".....	163
Figura J.12 Iteraciones y errores para el curso "Fundamentos de la Programación"	163
Figura J.13 Resultado Gráfico curso "Lógica de la Programación"	164
Figura J.14 Iteraciones y errores para el curso "Lógica de la Programación"	164
Figura J.15 Resultado Gráfico curso "Fundamentos Informáticos".....	165
Figura J.16 Iteraciones y errores para el curso "Fundamentos Informáticos"	166
Figura J.17 Filtro "Add Expression" para la Obtención de la mejora en rendimiento del primer al Segundo Bimestre.....	166
Figura J. 18 Salida de Información "Fundamentos de la Programación" - Experimento 2 Sin Discretizar.....	168
Figura J.19 Vista General del Curso "Fundamentos de la Programación" sin discretizar	168
FiguraJ. 20 Gráfica Relación Número de Interacción/Nota Final del curso Fundamentos de la Programación.	169
Figura J. 21 Categorización de Variable nota_final.....	170
Figura J.22 Vista General de la Asignatura "Fundamentos de la Programación" – Experimento 2 Discretización.....	171
Figura J.23 Salida de Información "Fundamentos de la Programación" - Experimento 2 Discretización	171
Figura J. 24 Gráfica Relación Número de Interacción/Nota_Final del curso Fundamentos de la Programación Discretizando.....	173

Figura J. 25 Vista General de la Asignatura "Lógica de la Programación" –Experimento 2	174
Figura J. 26 Salida de Información "Lógica de la Programación" - Experimento 2 ...	174
Figura J. 27 Relación número de interacciones, nota final para "Fundamentos de la Programación"- Experimento 2	175
Figura J. 28 Vista General de la Asignatura "Fundamentos Informáticos" – Experimento 2.....	176
Figura J. 29 Salida de Información "Fundamentos Informáticos" - Experimento 2 ...	177
Figura J. 30 Relación número de interacciones, clusters para "Fundamentos Informáticos"- Experimento 2	178
Figura J.31 Experimentación del curso Fundamentos de la Programación con el Algoritmo EM - Sin Discretizar	180
Figura J.32 Experimentación del curso Fundamentos de la Programación con el Algoritmo EM - Discretizado.....	181
Figura J.33 Experimentación del curso Lógica de la Programación	182
Figura J. 34 Experimentación del curso Fundamentos Informáticos	183

ÍNDICE DE TABLAS

Capítulo I

Tabla 1. 1 EDM Usuarios/Objetivos. Adaptation of (Romero & Ventura, Educational Data Mining: A Review of the State of the Art, 2010)	16
Tabla 1. 2 Resumen de Métodos de Análisis Colaborativo.	25

Capítulo II

Tabla 2. 1 Herramientas para el descubrimiento de información	40
Tabla 2. 2 Descripción de tablas	41
Tabla 2. 3 Descripción atributos principales de Foros y Logs.	42
Tabla 2. 4 Recopilación de nodos con mayor conexión.....	44
Tabla 2. 5 Materias y Docentes con mayor nivel de Interacción.	45
Tabla 2. 6 Resumen Colaboración Foros en materias.....	46
Tabla 2. 7 Descripción de atributos usados para la recopilación de información en foros.	49
Tabla 2. 8 Descripción atributos adicionales curso "Fundamentos de la Programación"	49
Tabla 2. 9 Opciones de configuración Algoritmo K-Means en WEKA	55
Tabla 2. 10 Resultados de la ejecución del Algoritmo K-Means	56
Tabla 2. 11 Opciones de configuración Algoritmo EM	59
Tabla 2. 12 Resultados ejecución del algoritmo K-Means	65
Tabla 2. 13 Resultado del Algoritmo K-Means	72
Tabla 2. 14 Recopilación de resultados Fundamentos de la Programación.....	79
Tabla 2. 15 Recopilación de resultados Fundamentos de la Programación	82
Tabla 2. 16 Recopilación de resultados Fundamentos Informáticos.....	85

ECUACIONES

<i>Ecuación 1 Fórmula de la Distancia Euclidiana.....</i>	<i>27</i>
<i>Ecuación 2 Ecuación Distancia de Manhattan</i>	<i>27</i>
<i>Ecuación 3 Ecuación Distancia de Mahalanobis</i>	<i>27</i>

RESUMEN

El presente trabajo abordó el nivel de colaboración estudiando el entorno colaborativo con mayor número de usuarios con el que cuenta la UTPL, el Entorno Virtual de Aprendizaje (EVA). Se utilizó la metodología inductiva como técnica de inferencia, seleccionándose las técnicas más aptas de MINERÍA DE DATOS para la identificación de patrones de comportamiento colaborativo en los estudiantes de modalidad abierta mediante la búsqueda de elementos colaborativos en los FOROS y la relación con las calificaciones obtenidas mediante el análisis de sus registros. En la etapa de minería se usó los algoritmos de AGRUPAMIENTO K-MEANS, EM y Clustering Jerárquico. Para el efecto se seleccionaron mediante la herramienta GEPHI los cursos: Fundamentos de la Programación, Lógica de la Programación y Fundamentos Informáticos pertenecientes al periodo Octubre-Febrero 2011 de la Carrera Informática UTPL-ECTS. Se etiquetó a cada grupo de estudiantes como: Alumnos con nivel de Colaboración Alto, Medio y Bajo. Los resultados demostraron la no utilización de todos los recursos de la plataforma educativa por parte de los docentes y de forma global un bajo interés colaborativo de parte de los estudiantes.

CAPÍTULO I

SITUACIÓN ACTUAL DE LOS ENTORNOS COLABORATIVOS

1.1. INTRODUCCIÓN

En esta primera parte se desarrollará el estado del arte del aprendizaje colaborativo y lo que esto implica.

Se realizará una breve descripción de la plataforma tecnológica MOODLE, base del EVA. Además se abordarán los temas de modelado de usuario colaborativo, los modelos de estudiante colaborativo, los comportamientos que se pueden determinar, el proceso de construcción de modelos de estudiante colaborativo, el soporte al aprendizaje colaborativo en modelos de estudiante más un compendio de los sistemas construidos bajo el modelo de colaboración y técnicas de Educational Data Mining más representativas.

1.2. MOODLE

MOODLE¹, que significa entorno de aprendizaje dinámico orientado a objetos y modular, se define como un sistema de gestión de cursos, un paquete de software diseñado para ayudar al profesor a crear fácilmente cursos de calidad en línea, se encuentra constituido por los siguientes módulos: Foros, Diarios, Cuestionarios, Consultas, Tareas, Recursos, Wiki, que pretenden facilitar el aprendizaje desde una posición participativa, fue creado por Martin Dougiamas en el 2003.

Las características de administración que ofrece MOODLE son:

- ◆ La administración general, se define el usuario administrador durante la instalación.
- ◆ La capacidad de añadir nuevos módulos.
- ◆ Personalización del sitio, cambio de apariencia con la ayuda de temas.
- ◆ La existencia de un total de 35 paquetes de idiomas.
- ◆ El código está escrito en PHP bajo GNU GPL.

El EVA en la Universidad constituye uno de los elementos más importantes para el logro de las actividades académicas, actualmente es utilizado tanto para las carreras de la modalidad presencial como abierta. Aquí el docente tiene la capacidad de subir recursos de wikis o de otras fuentes acorde a los temas que desea tratar, proponer

¹ <http://docs.moodle.org/>

una discusión abierta sobre algún tema en particular, comunicarse a través del chat con un alumno o viceversa además tanto facilitador como estudiante disponen de la Red Social de Aprendizaje (RSA) propias del curso en común.

El uso de esta plataforma educativa nos ofrece un potencial colaborativo bastante significativo medido en función de la usabilidad e interacción que los estudiantes generen.

1.3. MODELADO DE USUARIO

Se puede definir al modelado de usuario como: “El proceso de presentar al usuario ciertas recomendaciones, contenidos o cualquier otro recurso adaptándolo a las características que el sistema guarda de él” (Gaudioso Vásquez, 2002)

Sin duda alguna el modelo de usuario es uno de los componentes más importantes sino el principal para el funcionamiento de un sistema adaptativo ya que permite conocer las características propias del estudiante que son relevantes para el proceso de adaptación.

Según (González G, Duque M, & Ovalle C, 2008) manifiestan que para la construcción de un modelo de estudiante se deben tener los siguientes procesos:

- ◆ Definir Características
- ◆ Captura Inicial (usando formularios, test, etc.)
- ◆ Actualización (Alta, Media o Baja fidelidad)
- ◆ Implementación (Agente, Minería, Redes, etc.)

Cuando se habla de la definición de características describimos al nivel de comprensión de un tema, nivel de aprendizaje, características psicológicas, estado de ánimo, propósitos, patrones de comportamiento, destrezas, todo esto sirve como base para el establecimiento de su perfil.

La captura inicial indica la manera en que se obtendrá la información que va a ser procesada, esto estará sujeto a constantes actualizaciones las metodologías de la fidelidad se podría definir como aquellas que usan la observación directa para obtener datos del alumno mientras que la baja fidelidad lo hace tomando datos de interacción (debe haber interacción para que se realice la actividad colaborativa) luego de esto estaría listo para la implementación bien sea en forma de agentes o para el minado de datos, este último con el que se trabajaría.

El proceso de modelado de usuario según (Gaudioso Vásquez, 2002) sigue las siguientes pautas:

- ◆ Recogida de datos
- ◆ Identificación de tareas de adaptación adecuadas
- ◆ Construcción del modelo
- ◆ Mejora de la respuesta de los componentes de aprendizaje
- ◆ Validación del modelo construido.

Como ya se mencionó el modelo de usuario es uno de los requerimientos básicos para la construcción de un sistema adaptativo más se requiere tener en claro el ámbito que se le quiere dar al sistema. Es decir si necesitamos un sistema abierto es necesario que los modelos de usuario sean abiertos y flexibles. (Gaudioso Vásquez, 2002). Para ello se necesita usar técnicas de IA².

Un punto válido a señalar dentro del diseño de sistemas adaptativos es la importancia de las técnicas para la detección de errores de los alumnos que nos permitirán abordar de forma clara el problema de aprendizaje que podrían tener los mismos.

1.3.1. Representaciones del Modelo

Ya definidas las características del modelo de usuario, el siguiente paso consiste en representarlo, esto puede darse de tres maneras bien sea de forma explícita, implícita o híbrida, todo dependerá de la dimensión y estructura que se desee alcanzar.

Modelos Explícitos: Se presentan mediante reglas en base a un conocimiento. Difíciles de construir.

Modelos Implícitos: Se basan en atributos que recogen los datos de interacción en forma de aprendizaje automático.

Modelos Híbridos: Es una combinación de ambas técnicas con el fin de obtener los máximos beneficios de ellas.

² I.A : Inteligencia Artificial: f. Inform. Desarrollo y utilización de ordenadores con los que se intenta reproducir los procesos de la inteligencia humana. (Fuente: Diccionario de la Lengua Española).

Dentro de la educación virtual los sistemas adaptativos con mayor aceptación son los Sistemas Tutoriales Inteligentes que consisten en tres módulos que se comunican entre sí estos módulos son: El modelo de dominio, tutor y el de alumno.

La complejidad del modelo del alumno lo expresa (Boeira, 2001) al manifestar que: "Los modelos del alumno tradicionales son duramente criticados ya que su modelamiento es apenas eficiente en dominios limitados, ya que el modelo del alumno está basado en hipótesis predefinidas que giran en torno de reglas también predefinidas. Esto quiere decir que esta dependencia al dominio no puede describir o predecir toda la variedad del comportamiento humano".

1.3.2. Soporte al aprendizaje colaborativo

El aprendizaje colaborativo tiene como objetivo primordial el promover los procesos de colaboración en los estudiantes, detectando para ello perfiles de colaboración mediante la identificación de estereotipos

Según (Gaudioso Vásquez, 2002) existen ciertas tareas que facilitarían el entorno de cooperación y a las que se le debe dar especial impulso para alcanzar este fin, estas tareas son:

Avisar al tutor de la posibilidad de que un alumno tenga dificultades en el uso de algún servicio. Para ello necesitamos conocer el nivel de fracaso de un usuario en una determinada actividad.

Aconsejar a un usuario que interactúe más con un determinado servicio. Para esto se necesita conocer el nivel de actividad de cada estudiante.

Agrupación automática de usuarios dentro de una misma comunidad. Sugiere la conformación de grupos de acuerdo a ciertas condiciones pre-establecidas. El sistema sugiere u organiza los grupos de acuerdo a similitudes y/o diferencias entre los estudiantes.

Aconsejar al tutor una clasificación de mensajes en el foro. Esta tarea puede ser de gran utilidad para el tutor a la hora de administrar un determinado grupo, identificando temas de interés que sirvan para añadir categorías al foro, una mejor gestión de los servicios de comunicación mejora sin lugar a dudas el acceso a la información y por tanto la colaboración.

Según (Jermann, Soller, & Muehlenbrock, 2001) (y otros autores más. Ver [Sección 1.3](#)) se distinguen cuatro fases en todo sistema colaborativo:

- ◆ **Recolectar información de la Interacción:** Se monitorean las acciones colaborativas que realizan los alumnos y se guardan para poder procesarlas posteriormente.
- ◆ **Construir el Modelo de la Interacción:** Utilizando los datos capturados se construye el modelo de la interacción que incluyen indicadores relacionados con las actividades colaborativas.
- ◆ **Comparar el Modelo de Interacción con el modelo ideal.** El modelo ideal de la interacción es un conjunto de indicadores que describen estados deseables y no deseables de la interacción. En una situación ideal, un estado deseable de la interacción sería que todos los alumnos de un grupo alcancen un grado de participación similar y que consigan ponerse de acuerdo frecuentemente. (Arteaga & Fabregat, 2002)
- ◆ **Sugerir acciones a tomar.** Si existe una marcada diferencia del estado actual con el ideal se sugieren ciertas acciones para remediar esta situación.

1.4. TRABAJOS RELACIONADOS – SOPORTE A LA COLABORACIÓN

1.4.1. Sistema Hipermedia De Aprendizaje Colaborativo Adaptativo (SHACA)

SHACA es un sistema que integra el aprendizaje individual con el aprendizaje colaborativo, y en el que ambos aprendizajes son adaptativos (Arteaga & Fabregat, 2002). Una característica importante es que el modelo utilizado para la adaptación del aprendizaje individual es el mismo para el aprendizaje colaborativo, así que cualquier cambio en el modelo del estudiante afecta el comportamiento adaptativo individual como colaborativo. La definición del Modelo de la Colaboración, es sin duda alguna uno de los componentes más complejos y críticos en esta propuesta. Estos espacios deben permitir que los participantes realmente se encuentren en ambientes donde se construya conocimiento colaborativamente y no exclusivamente en un ambiente para la difusión superficial de conocimiento o intercambio de opiniones personales. Este es el reto principal que se sostiene.

Este modelo incluye las reglas para la conformación de los Espacios de Colaboración y está orientado a resolver cuestiones como: ¿Qué condiciones debe cumplir un estudiante para pertenecer a un grupo particular?, ¿Las características de cada grupo

serán previamente definidas o serán dinámicas?, ¿Los grupos permitirán el uso de roles de los estudiantes?, ¿El cumplimiento de los objetivos del grupo cambia el estado del Modelo del Estudiante?, ¿Cuáles son las condiciones para que un estudiante salga de un grupo e ingrese en otro?

En la Figura 1.1 se observa la arquitectura de SHACA donde el estudiante y el profesor son los actores principales en este proceso.

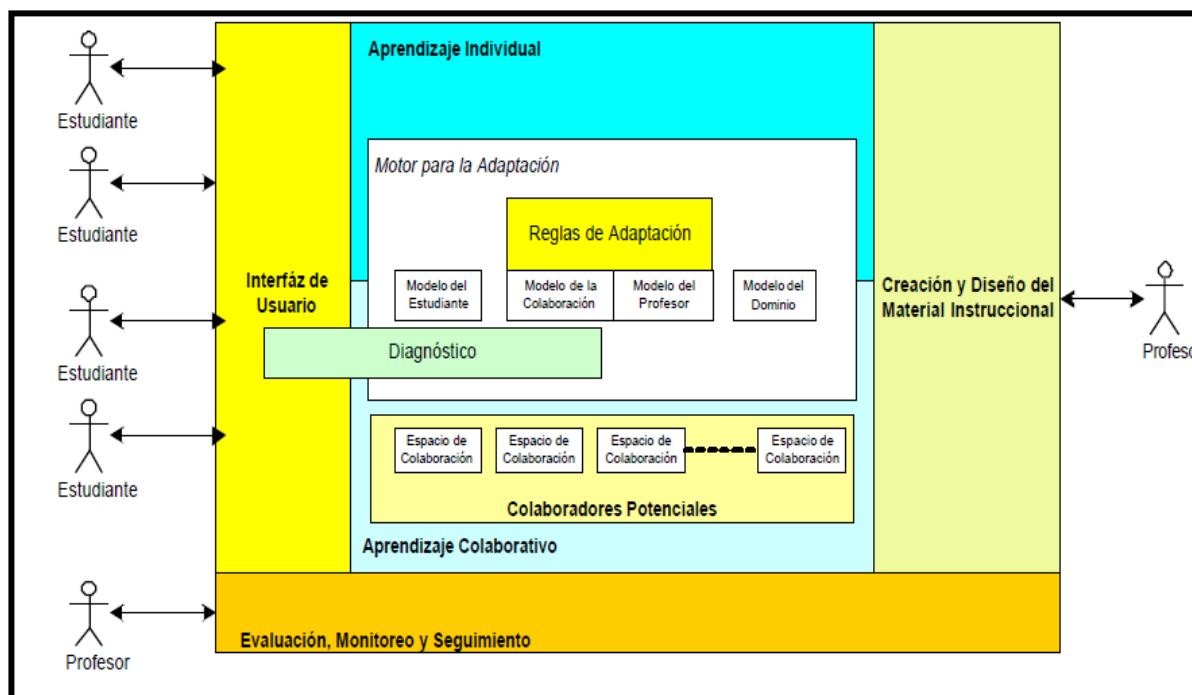


Figura 1.1 Arquitectura del Sistema SHACA, (Arteaga & Fabregat, 2002)

1.4.2. Modelo de Estudiante Colaborativo (MEC)

(Duran, 2006) Definió un modelo que incluye rasgos personales y destrezas colaborativas del estudiante que posee y que fueron identificadas a partir del análisis de las actividades colaborativas previas. Para ello se utilizó técnicas de Minería de Usos de la Web o como sus siglas en inglés WUM (WEB USAGE MINING).

A este modelo lo nombró como MODELO DE ESTUDIANTE COLABORATIVO (MEC) consistente en un proceso de: Selección de la estructura, Inicialización del modelo y Diagnóstico.

Como propuesta en la selección de la estructura se define la información sobre el alumno que será registrada, dentro de esta se detalla el perfil del alumno, sus

características personales, tanto en datos demográficos (nombre, apellido, fecha de nacimiento) como de dominio, estos incluyen los errores cometidos por el alumno al realizar una acción en el sistema.

Se analiza el Perfil de Colaboración de los estudiantes en el que se incluyen: Habilidades Colaborativas, Taxonomía de Destrezas donde cada habilidad se divide en sub habilidades definiéndose atributos.

Ya para el perfil de grupo se establecen las características que identifican al grupo y describen al grupo como un todo. Se considerarán aquí: identificación del grupo, conjunto de alumnos que integran el grupo, objetivo del grupo, creencias compartidas por los miembros, errores cometidos por el grupo, y rol que desempeña cada integrante.

En la representación del modelo se trabaja con una representación híbrida, los datos de usuario y de grupo se representan en base a un conocimiento previamente declarado, los datos colaborativos por otra parte serán representados mediante una serie de atributos que se calculan mediante la ejecución de tareas de aprendizaje automático.

INICIALIZACIÓN: *La primera etapa es la etapa de pre procesamiento:* En esta etapa se obtienen los logs generados por el sistema SAVER (sistema estudiado en este caso), se le realiza la correspondiente limpieza y unificación de los datos.

A continuación viene la etapa de *Descubrimiento del conocimiento* donde se definen los Métodos y algoritmos de descubrimiento, se evaluará la aplicación de reglas de asociación para descubrir conocimiento en las actividades colaborativas.

En la etapa de análisis de conocimiento descubierto se realiza una selección del total de conocimiento descubierto, aquel que resulte más significativo para retroalimentar el modelo, con entradas que permitan luego soportar recomendaciones personalizadas para mejorar la colaboración, se definen algunos criterios que convendría considerar para seleccionar sólo aquel conocimiento más significativo, la representación de este modelo se puede observar en la Figura 1.3.

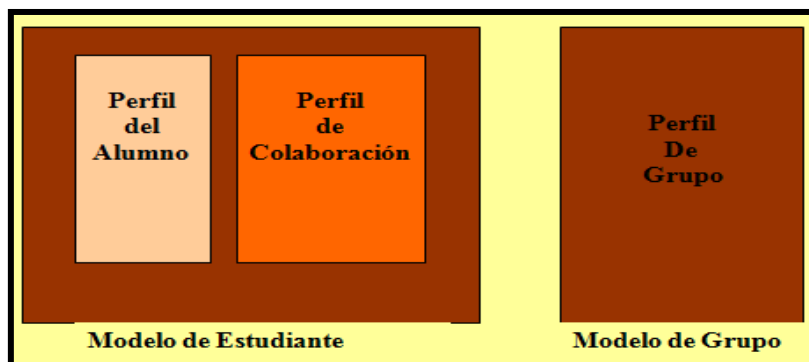


Figura 1. 2 Componentes del Modelo de Estudiante Colaborativo (MEC). (Duran, 2006)

Finalmente en el diagnóstico se realiza una actualización del modelo del alumno con los datos recogidos en cuanto a su nivel de colaboración.

Según (Gaudioso Vásquez, 2002) los objetivos principales para la adaptación del soporte adaptativo a la colaboración son:

- ◆ Formar grupos de trabajo cuyo objetivo es facilitar la cooperación entre alumnos con perfiles similares.
- ◆ Dar soporte en el uso de servicios de comunicación para fomentar su uso en el grupo.
- ◆ Dar soporte en el uso de servicios para compartir información y para aconsejar al usuario que documentos pueden ser de su interés, bien porque el usuario que los mandó era afín o porque la temática puede ser de su interés.
- ◆ Encontrar un usuario afín para colaborar. Seleccionando entre todos los usuarios del grupo, aquellos con los que el sistema piensa que la colaboración puede ser efectiva.
- ◆ Detectar perfiles de colaboración. Identificar estereotipos que ayuden a gestionar los grupos de trabajo y las relaciones personales.

El aprendizaje cooperativo o colaborativo³ es aquel que se consigue mientras se establecen comunicaciones con otros compañeros de aprendizaje (Rodríguez Anaya, 2009). Con esta definición podemos decir que no es necesario colaborar para aprender cooperativamente.

³ Las palabras cooperativo y colaborativo serán usados de forma similar en esta tesis. Mas es importante mencionar que ciertos autores definen las tareas cooperativas como la asignación de sub tareas para llegar a una global, mientras que tareas colaborativas aquellas que mediante la comunicación y negociación se llegan a un acuerdo.. (Gaudioso Vásquez, 2002)

Debemos reconocer que los estudiantes de la modalidad a distancia que es a los cuales nos vamos a dirigir tienen la responsabilidad de controlar tanto sus procesos de aprendizaje como de colaboración. Sin lugar a duda el desarrollo de las TIC's⁴ hacen que la colaboración sea posible en escenarios donde la personalización se ve de cierta forma afectada por la distancia.

Los objetivos que plantea (Rodríguez Anaya, 2009) en su modelo de colaboración son: que sea utilizado por los estudiantes y tutores, fácil de entender por personas, y que contenga información relativa a la colaboración.

Así pues autores como (Gaudioso Vásquez, 2002), menciona a los sistemas CSCL (Computer Support Collaborative) como un apoyo al trabajo/aprendizaje cooperativo y colaborativo, donde cada alumno posee una especie de entrenador que lo dirigirá a lo largo de sus sesión, en la [Sección 1.4.4](#) se ahondará en este tema.

A demás de tomar en cuenta las características colaborativas individuales es preponderante describir y establecer características de agrupación.

Como se ha visto anteriormente en la propuesta de (Duran, 2006) se dan pautas en el proceso de colaboración, (Jermann *et al.*, 2001) han planteado ciertos cambios, estas son las principales directrices a tomar en cuenta.

- ◆ **Recogida de Datos:** Salvado de datos de las acciones colaborativas.
- ◆ **Seleccionar indicadores:** Representación del estado actual de la colaboración en base a experiencias anteriores.
- ◆ **Diagnóstico de la Interacción:** Compara el estado actual de la colaboración con un estado ideal, aquel en el que se obtenga el promedio más alto de cooperación.
- ◆ **Otras acciones:** Acciones para mejorar el estado actual de colaboración, derivadas del análisis de un único indicador, por decirlo de otra forma prestarle una mayor atención de forma individual o grupal a aquellos estudiantes con un menor grado de colaboración.

⁴ TICS: Tecnologías de Información y Comunicación. se encargan del estudio, desarrollo, implementación, almacenamiento y distribución de la información mediante la utilización de hardware y software como medio de sistema informático. (Fuente: <http://tics.org.ar>)

1.4.3. Soporte Adaptativo Al Aprendizaje Colaborativo e Individual (ASCIL)

ASCIL tiene como objetivo la colaboración entre estudiantes a partir de su disponibilidad para colaborar. Posee tres componentes:

AHA: Para la creación de cursos en línea

CLAROLINE: Administración de cursos en línea

MODELO DE COLABORACIÓN: Guarda información para el inicio de las actividades de aprendizaje.

MOTOR ADAPTATIVO COLABORATIVO: Implementa las reglas de adaptación básicas, construyendo el conjunto de colaboradores potenciales para cada estudiante este motor utiliza la información contenida en el Modelo del Usuario (estudiante) para conseguir el comportamiento adaptativo. A partir de ello, el motor adaptativo sugiere para cada estudiante un conjunto de Colaboradores Potenciales que se crea dinámicamente y que es diferente para cada estudiante. (Arteaga & Fabregat, 2002)

INTERFAZ INTEGRADO: Se ha fusionado dos enlaces uno para cursos adaptativos y otro para el conjunto de colaboradores.

Todos estos componentes se encuentran diagramados en la Figura 1.3

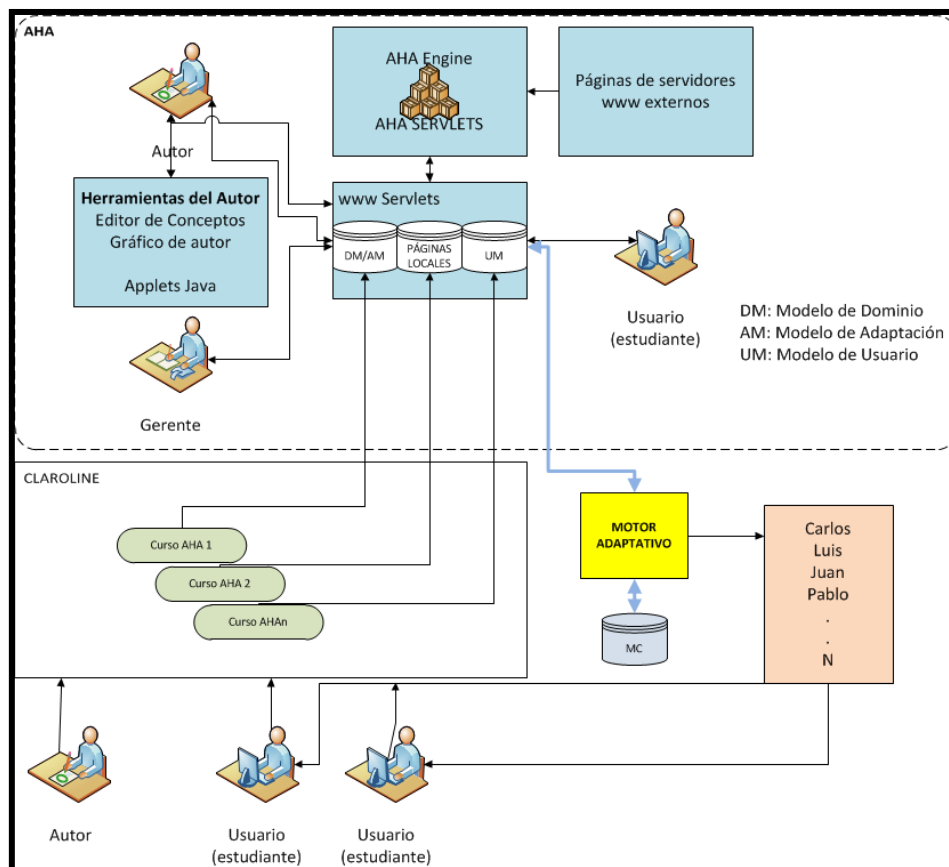


Figura 1. 3 Arquitectura de ASCIL. Adaptación de (Arteaga & Fabregat, 2002)

ASCIL posee un proceso de interacción que se conforma de la siguiente manera:

1. El estudiante se registra en un curso a través de la pantalla de registro de CLAROLINE, y automáticamente queda registrado como estudiante del curso de AHA.
2. En cada uno de los cursos que el estudiante inicie, se incluye un enlace al material del curso desarrollado en AHA y una vez que el estudiante inicia el estudio del material, el motor adaptativo de ASCIL crea el primer conjunto de Colaboradores Potenciales que se va actualizando a lo largo del proceso de aprendizaje.
3. Cuando el estudiante requiere apoyo puede dirigirse al enlace "Lista de Colaboradores" y le aparecerán los estudiantes que cumplen las condiciones establecidas en las reglas de adaptación. (Arteaga & Fabregat, 2002)

1.4.4. Sistema de Soporte a la Colaboración (CSCL)

Según (Rodríguez Anaya, 2009) Los entornos colaborativos vía Web están basados en investigaciones en torno a los sistemas colaborativos soportados por ordenador (CSCL es sus siglas en inglés), que es un área de innovación pedagógica que desarrolla un modelo educativo que consiste en aumentar la interacción entre los miembros de un grupo cuyo interés común es el aprendizaje y que cuenta con el ordenador como el elemento para interactuar y el medio para comunicarse.

El nivel de este soporte tecnológico puede abarcar desde tareas meramente administrativas o de gestión, hasta el soporte indirecto de las interacciones colaborativas mediante aplicaciones tales como sistemas de videoconferencia (De la Fuente Valentín *et al.*, 2009) .

En el trabajo realizado por (Rodríguez Anaya, 2009) se describe el proceso de análisis de la colaboración estos pasos son:

- ◆ **Recogida de datos.** Esta fase cubre la observación y recogida de las interacciones.
- ◆ **Construcción de un patrón o plantilla de la interacción.** Se realiza una selección e inferencia de uno o más indicadores de alto nivel con los que representar el actual estado de interacción.
- ◆ **Comparación del actual estado de interacción con uno deseado.** El estado deseado se define como un conjunto de valores de los indicadores con los que se puede discriminar la productividad de las interacciones.
- ◆ **Guía/consejo.** Análisis del estado deseado con el actual.
- ◆ **Evaluación y diagnóstico.** Después de la fase de evaluación/consejo, se evalúa dicha actuación

1.5. CONSTRUCCIÓN DE UN MODELO COLABORATIVO

Ahora que ya se tiene conocimiento acerca de lo que es un modelo de usuario y específicamente de un modelo de estudiante colaborativo y que se analizó los diferentes procedimientos que debemos tener en cuenta para su construcción, se

Hay que tomar en cuenta que la mayoría de los autores plantea dos técnicas a seguir: La primera consiste en el desarrollo de modelos de usuario en base al conocimiento

de expertos y el segundo define técnicas de aprendizaje automático. Ambas técnicas tienen la capacidad de ser combinadas, según los requerimientos del usuario.

Los pasos a seguir tomando en cuenta de las diferentes investigaciones serían:

1. Estudio de campo
2. Recogimiento de Información
3. Diseño Modelo de datos
4. Construcción del modelo
5. Pruebas del Modelo
6. Implementación del Modelo

En el estudio de campo se verificará el estado actual de las herramientas y sistemas actuales capaces de ayudarnos en la definición de nuestro modelo.

En la segunda etapa se recogerá la información de los alumnos bien sea en forma de encuesta o usando los logs del sistema.

Como tercera etapa en base a la información seleccionada se diseñará un modelo de usuario colaborativo.

Para la cuarta etapa estaremos en disposición de construir el modelo de usuario colaborativo.

Luego de esto (Quinta etapa) se realizarán las pruebas de rigor de tal manera que el modelo pueda validarse en el entorno, finalmente se implementará el modelo.

1.6. MINERÍA DE DATOS EN LA EDUCACIÓN

El desarrollo tradicional de los cursos e-learning es una actividad ardua en el que el profesor del curso tiene que elegir el contenido que se mostrará, decidir sobre la estructura de los contenidos, y determinar los elementos de contenido más apropiado para cada tipo de usuario potencial del curso. (Romero, Ventura, & García, Data mining in course management systems: Moodle case study and tutorial).

El sitio web de la comunidad de Educational Data Mining⁵ (EDM), define la minería de datos educativos de la siguiente manera: "La minería de datos en la educación es una disciplina emergente, cuyo interés radica en la elaboración de métodos para explorar los tipos de información que proceden de los centros educativos y el uso de los métodos para comprender mejor a los estudiantes y la manera en que aprenden".

El Argumento de (Corso & Alfaro, 2010) Data Mining tiene como objetivo reunir los beneficios de varias áreas como la estadística, inteligencia artificial, las bases de datos y el pre procesamiento masivo, usando las bases de datos como materia prima.

En el ámbito educativo la minería de datos proporciona entre otras características, criterios y pautas para personalizar el sistema de enseñanza estableciendo cambios estructurales en el mismo.

Existen diversos contextos donde se podría implementar EDM, (Baker & Yacef, 2009) manifiesta la existencia de cuatro áreas claves.

La primera: En los últimos años los investigadores han utilizado EDM para deducir el desenvolvimiento del estudiante dentro del sistema y lo aburrido o frustrado que podría sentirse, han podido ampliar además el modelo de estudiante para la determinación de posibles fracasos o falta de retentiva del alumno.

La segunda: Para el descubrimiento o mejora de la estructura de los modelos de conocimiento de dominio, algunos investigadores han sido capaces de desarrollar enfoques automatizados que se pueden descubrir modelos precisos de estructura de dominio, directamente de los datos.

La tercera: Un tercer aspecto clave de la aplicación de métodos de EDM ha sido en el estudio pedagógico de apoyo (tanto en software para el aprendizaje y el aprendizaje en otros dominios, como los comportamientos), para descubrir qué tipos de apoyo pedagógico son más eficaces, ya sea de forma general o por grupos de estudiantes o en situaciones diferentes.

La Cuarta: La búsqueda empírica de pruebas para perfeccionar y ampliar las teorías educativas y fenómenos educativos conocidos, para una comprensión más profunda de los factores clave que afectan el aprendizaje.

⁵ <http://www.educationaldatamining.org>

Según (Romero & Ventura, Educational Data Mining: A Review of the State of the Art, 2010) una consideración inicial, parece implicar sólo dos grandes grupos, los alumnos y los instructores, en realidad hay más grupos que participan con muchos más objetivos, como puede verse en la Tabla 1.1

Usuarios/actores	Objetivos de uso- Data Mining
Alumnos/Estudiantes/Pupilos	Personalizar el e-learning, realizando una recomendación de actividades y tareas, forjando un aprendizaje basado en experiencia.
Educadores/Profesores/ Instructores/tutores	Obtener una retroalimentación acerca de la enseñanza, analizando el comportamiento del estudiante detectando que estudiantes necesitan un mayor soporte y que errores son los más comunes que se puede llegar a tener personalizando, adaptando los cursos para un mejor aprendizaje y usabilidad.
Desarrolladores de Cursos/Investigadores Educativos	Para la evaluación y mantenimiento de cursos, valorando la estructura de contenido de los cursos, comparando técnicas de minería de datos con el fin de recomendar las más útiles para cada tarea.
Organizadores/Proveedores de aprendizaje/Universidades/ Empresas de formación privada	Para la toma de decisiones en instituciones de nivel superior. Encontrando la mejor relación costo/eficiencia. Seleccionando los candidatos más calificados para la admisión en sus universidades.
Administradores/Administradore s de centro educativo/Administradores de red/Administradores de Sistema	Para desarrollar la mejor manera de organizar los recursos institucionales humanos y materiales y su oferta educativa, para establecer parámetros de eficiencia del sitio, determinando el enfoque y eficiencia de la educación a distancia.

Tabla 1. 1 EDM Usuarios/Objetivos. Adaptation of (Romero & Ventura, Educational Data Mining: A Review of the State of the Art, 2010)

Para facilitar esta tarea, necesitamos métodos de análisis de datos y herramientas para observar el comportamiento de los estudiantes y maestros para ayudar en la detección de posibles errores, deficiencias y posibles mejoras. El análisis de datos tradicional en el e-learning es la hipótesis que mas impulso tiene en el sentido de que el usuario parte de una pregunta y explora los datos para confirmar su intuición.

Si bien esto es útil cuando se maneja una cantidad de datos pequeña puede ser muy difícil para el usuario buscar patrones más complejos que se relacionan con diferentes aspectos de los datos. Una alternativa al análisis de datos tradicional es el uso de

minería de datos como un método inductivo para descubrir de forma automática la información oculta en los datos.

Según (Trcka & Pechenizkiy, 2009) el Proceso de minería de la Educación (EPM) tiene por objeto la construcción completa y compacta de los modelos de procesos educativos que son capaces de reproducir todo el comportamiento observado, la verificación de si el comportamiento del modelo (ya sea pre-escrito o descubierto a partir de datos) coincide con el comportamiento observado (control de la conformidad), y que información se proyecta en los registros en el modelo, para una mejor comprensión del proceso.

En la minería de datos se extraen automáticamente los datos, la información que se recopilará se basará únicamente en los datos que se obtengan de este análisis en lugar de basarse en la investigación o impresión humana, construyendo modelos de análisis que descubre patrones y tendencias interesantes de información sobre el uso del estudiante que pueden ser utilizadas por el profesor para mejorar el aprendizaje del estudiante y el mantenimiento del sistema.

(Rodríguez Anaya, 2009) Afirma que aunque hay una falta de metodología y estándares en el análisis de la colaboración en entornos educativos, se han realizado distintos experimentos para medir o identificar la colaboración que se realizaba entre usuarios de un sistema.

De estos experimentos se deduce, en primer lugar, que hay que tener en cuenta el método de adquisición de la información, lo que se corresponde con el pre proceso de la técnica de minería de datos utilizado.

Se pueden identificar tres métodos:

Cualitativo: Realizando de forma directa preguntas a los individuos participantes en la investigación, o expertos evaluando las actividades de los participantes.

Cuantitativo: Recogiendo información de estadísticas de las actividades de los participantes.

Mixta: Usando ambos métodos a la vez.

Dentro del contexto del análisis de la colaboración, hay que mencionar un conjunto de investigaciones las mismas que se agrupan dentro de la red de excelencia llamada Kaleidoscope⁶.

Kaleidoscope es la Red Europea para la innovación científica en materia de Tecnologías para la Educación. Con esta visión Kaleidoscope se apoya en equipos cooperativos de investigadores, líderes en campos clave: ciencias de la educación, tecnologías de la información, ciencias sociales.

La técnica del análisis de las interacciones se centra en la obtención de indicadores de nivel medio obtenidos mediante procesos estadísticos principalmente del conjunto de interacciones, en bruto, que un usuario de un sistema realiza.

El objetivo de estas investigaciones ha sido monitorizar y hacer seguimiento de las interacciones de los estudiantes en un entorno de aprendizaje colaborativo de Técnicas de Minado.

Según (Romero Morales, Ventura Soto, & Hervás Martínez, 2005) La aplicación de técnicas de minería de datos en la educación se puede ver desde dos puntos de vista u orientaciones distintas:

Orientado hacia los autores. Con el objetivo de ayudar a los profesores y/o autores de los sistemas de e-learning para que puedan mejorar el funcionamiento o rendimiento de estos sistemas a partir de la información de utilización de los alumnos.

Sus principales aplicaciones son: obtener una mayor realimentación de la enseñanza, conocer más sobre como los estudiantes aprenden en el web, evaluar a los estudiantes por sus patrones de navegación, reestructurar los contenidos el sitio web para personalizar los cursos, clasificar a los estudiantes en grupos, etc.

Orientado hacia los estudiantes. Con el objetivo de ayudar o realizar recomendaciones a los alumnos durante su interacción con el sistema de e-learning para poder mejorar su aprendizaje. Sus principales aplicaciones son: sugerir buenas experiencias de aprendizaje a los estudiantes, adaptación del curso según el progreso del aprendiz, ayudar a los estudiantes dando sugerencias y atajos, recomendar caminos más cortos y personalizados, etc.

⁶ (<http://www.no-kaleidoscope.org/pub/>)

Las técnicas de aprendizaje automático son una alternativa para clasificar y predecir acciones futuras de los estudiantes en el sistema.

En algunos enfoques de sistemas recomendadores, durante el proceso de recomendación se opta por modelar las preferencias de los usuarios mediante técnicas de aprendizaje automático, tales como: redes neuronales, árboles de decisión, redes bayesianas, etc. (Valdiviezo, Santos, & Boticario, 2010)

Según (Romero Morales, Ventura Soto, & Hervás Martínez, 2005) las etapas de minado se corresponderían de la siguiente manera:

Pre procesamiento. Consiste en la recogida o extracción de los datos, limpieza de datos, división de las partes, selección de los atributos e integración de datos.

Minería de datos. Consiste en la selección de los algoritmos de minería de datos a utilizar y la aplicación de dichos algoritmos sobre los datos.

Post procesamiento. Consiste en la interpretación, evaluación de los resultados obtenidos y la utilización del conocimiento descubierto. (Romero, Ventura, & García, Data mining in course management systems: Moodle case study and tutorial) Proponen el siguiente proceso:

Recopilar datos. El sistema LMS⁷ es empleado por los estudiantes y la utilización de la información y la interacción se almacena en la base de datos.

Pre procesamiento de los datos. Los datos se limpian y se transforman en un formato adecuado para ser explotado. Con el fin de pre-procesar los datos, se puede utilizar una herramienta de administración de base de datos.

Aplicar minería de datos: Se aplican los algoritmos de minería de datos y se construye el modelo usando datos específicos y herramientas de minería de datos.

Interpretación, evaluación y despliegue de los resultados. Los resultados del modelo son interpretados y utilizados para la adopción de nuevas medidas. El profesor puede utilizar la información descubierta para tomar decisiones sobre los estudiantes y las actividades del curso de MOODLE con el fin de mejorar el

⁷ LMS: Learning Management System en Español se traduciría como Sistema de Manejo de Aprendizaje: Se emplea para administrar, distribuir y controlar las actividades de formación no presencial (o aprendizaje electrónico) de una institución u organización. (Fuente: <http://es.wikipedia.org>)

aprendizaje de los estudiantes. Una ampliación de este proceso se puede observar en la Figura 1.4

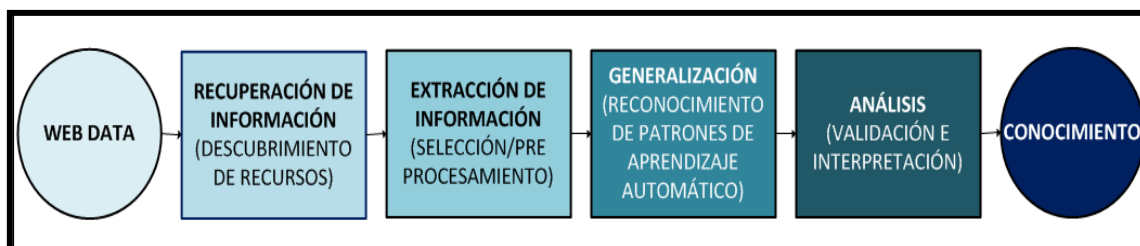


Figura 1. 4 Tareas del EDM. (Corso & Alfaro, 2010)

Como se puede observar los procesos son similares para ambos autores lo que se manifiesta en el proceso ideal para realización del presente trabajo. Las técnicas más utilizadas que describe (Romero, Ventura, & García, Data mining in course management systems: Moodle case study and tutorial) son:

La Clasificación y Agrupamiento: Las técnicas de clasificación y agrupamiento o clustering consisten en la habilidad intelectual para ordenar o dividir fenómenos complejos. Su aplicación a sistemas de e-learning permite agrupar a los usuarios por su comportamiento de navegación, agrupar a las páginas por su contenido, tipo o acceso y agrupar los comportamientos de navegación similares.

Reglas de Asociación: Las reglas de asociación descubren relaciones entre atributos de un conjunto de datos que superan unos determinados umbrales. Su aplicación más típica ha sido en los sistemas de comercio electrónico para informar sobre las preferencias de compra de los clientes. En sistemas de e-learning permite descubrir relaciones o asociaciones entre distintas páginas Web visitadas.

Análisis de secuencias o secuencia de patrones. Es una técnica de minería de datos que descubre secuencias dentro de un conjunto de datos. Al igual que las reglas de asociación, también se han aplicado en sistemas de comercio electrónico para descubrir secuencias de acciones de los clientes. Su aplicación a sistemas de e-learning permite analizar secuencias de páginas visitadas durante una sesión o en distintas sesiones de un mismo usuario.

Para nuestro proyecto sin duda alguna la técnica a utilizar será la de Agrupación para verificar la forma como se relacionan los alumnos y el nivel de colaboración existente entre ellos y las actividades que por voluntad realizan.

1.7. Métodos y Herramientas para el Análisis colaborativo

Como ya se ha visto el proceso de extracción de conocimiento se compone de un preproceso, minería de datos como tal y un postproceso según (Romero & Ventura, Educational Data Mining: A Review of the State of the Art, 2010).

El preprocesamiento es la etapa fundamental, pues en ella se selecciona la información que va a ser analizada, si no hay un adecuado manejo de esta los datos serán inexactos o erróneos.

(Rodríguez Anaya, 2009), identifica que los indicadores de la actividad del estudiante y la actividad que es producida por un estudiante en particular en otro son características claves de colaboración, en este estudio se identifica como la respuesta que da un estudiante a otro en los foros.

Este autor además habla señala algunas técnicas para adquirir la información sea de forma *Cualitativa* (Meier et al., 2007) , *Cuantitativo* (Redondo et al, 2003); (Hong, 2001); (Bratitits et al., 2008)) y *Mixta* ((Collazos et al., 2002); (Collazos et al ., 2007); (Daradoumis et al., 2006); (Martinez et al, 2006); (Perera et al., 2009); (De Pedro, 2007)).

Así mismo existen diferentes métodos de Inferencia tales como:

- ◆ **Análisis de un experto** (Meier et al., 2007); donde un especialista analiza los datos y realiza un juicio.
- ◆ **Comparación con un modelo preexistente** (Redondo et al, 2003).
- ◆ **Diferentes técnicas estadísticas o de aprendizaje automático, como árboles de decisión, cluster, minería de patrones, etc.**, ((Talavera & Gaudioso, 2004); (Redondo et al, 2003); (Hong, 2001); (Perera et al., 2009); (De Pedro, 2007)),
- ◆ **Técnicas de análisis de las interacciones**, las cuales obtienen indicadores estadísticos de las interacciones de los usuarios del sistema ((Daradoumis et al., 2006); (Martinez et al, 2006); (Bratitits et al., 2008)),
- ◆ Incluso se pueden caracterizar los sistemas por **no utilizar ningún sistema de inferencia** ((Collazos et al ., 2007); (Collazos et al., 2002)).

En este mismo trabajo se mencionan algunas herramientas para el análisis colaborativo (Rodríguez Anaya, 2009), a continuación un extracto de las más relevantes en función del entorno en el que se ejecutan las actividades y el tipo de dato de entrada.

DomoSim-TPC (Redondo et al, 2003)

- ◆ Se basa en datos cuantitativos de las interacciones de un usuario en un curso.
- ◆ Las interacciones se etiquetan según la actividad.
- ◆ El modelo es comparado mediante un algoritmo de lógica borrosa.
- ◆ Resultado: El grado de pertenencia con el nivel de colaboración.

Talavera y Gaudioso (2004)

- ◆ Estudiantes con iniciativa, claves para la colaboración.
- ◆ Usa clustering
- ◆ Agrupa a los estudiantes de acuerdo a la semejanza entre ellos.
- ◆ Una vez que se tienen los grupos hechos, los comparan con la colaboración que conocen por parte de los alumnos, y que el tutor o tutores del curso han dado.
- ◆ De la comparación hayan que los grupos encontrados identifican diferentes comportamientos de los alumnos respecto a la colaboración, por lo que encuentran que el método utilizado identifica en los alumnos la colaboración realizada.
- ◆ Cuantitativo

(Daradoumis et al., 2006) & (Martinez et al, 2006)

- ◆ Proponen un marco teórico para el análisis de la colaboración dividido en tres capas.
- ◆ El modelo de capas se divide en indicadores de alto nivel, que son evaluaciones cualitativas de los tutores.
- ◆ En el nivel medio realizan un análisis de redes sociales, y utilizan datos de dicho análisis y del log de comunicaciones para establecer unos indicadores cuantitativos
- ◆ En el nivel inferior (densidad de la red, número de comunicaciones). No proponen un método de inferencia.
- ◆ Su objetivo es mostrar el análisis en capas que han especificado.

(De Pedro, 2007)

- ◆ Realiza un estudio de la colaboración en un entorno de aprendizaje colaborativo basado en la técnica de Wiki.
- ◆ Los estudiantes escriben un documento de forma colaborativa al estilo de las wikis.
- ◆ Solicita a los estudiantes que etiqueten sus contribuciones al documento con una serie de etiquetas establecidas o que propongan nuevas.
- ◆ De este modo, las contribuciones están etiquetadas e identificado el autor. A partir de ahí, realiza un análisis estadístico del número, tamaño, tipo y corrección de las contribuciones de un estudiante.
- ◆ El tutor también revisa las contribuciones aportando el análisis.
- ◆ Evalúa la metodología propuesta en tres casos de estudio y concluye advirtiendo de las ventajas del método al dar una herramienta que hace posible la autorregulación y de las desventajas de un análisis cuantitativo de las contribuciones al no considerar la información semántica y estructural de los documentos.

(Meier et al., 2007)

- ◆ Se propone una metodología de análisis de la colaboración llamado “plan de evaluación” (rating scheme en inglés).
- ◆ El plan de evaluación es un método cualitativo de análisis de la colaboración.
- ◆ Su objetivo es identificar la validez de los atributos cualitativos propuestos.
- ◆ Este método consiste en hacer estimaciones de la colaboración entre una pareja en una tarea colaborativa, que se realiza vía web y que se graba en vídeo para un análisis posterior.
- ◆ Proponen varios indicadores para identificar la colaboración y unos expertos evalúan las escenas grabadas en video según los indicadores propuestos.
- ◆ Estas investigaciones no tienen como finalidad desvelar la colaboración sino evaluar el propio método de análisis de la colaboración.

(Collazos et al ., 2007)

- ◆ Método cuantitativo de detección de la colaboración.
- ◆ Utilizan ciertos atributos obtenidos de un análisis estadístico de las interacciones y un análisis de los contenidos.

- ◆ El primer objetivo de la investigación fue identificar la validez de los atributos utilizados (Collazos et al., 2002).
- ◆ Al continuar la investigación el objetivo que siguieron fue el de monitorizar al estudiante según los indicadores que habían propuesto.
- ◆ No proponen ningún juicio o inferencia.

(Bratit et al., 2008)

- ◆ En sus trabajos utilizan dos capas obteniendo los datos de los ficheros log y de la base de datos, donde está incluido un análisis de las redes sociales.
- ◆ Tampoco realizan ninguna inferencia con los datos obtenidos pero sí los preparan para su visualización en una herramienta de monitorización.

(Perera et al., 2009)

- ◆ Estudios sobre el trabajo colaborativo de una serie de estudiantes en un curso de informática, el cual consistía en programar en equipo.
- ◆ La comunicación se realizó mediante “tickers” (servicio que consiste en la asignación de tareas a un alumno) y la wiki, que permiten la creación y mantenimiento de un documento por varios autores.
- ◆ De los dos servicios obtienen datos cuantitativos para realizar un análisis estadístico y otro aplicando técnicas de clustering.
- ◆ Para validar ambas técnicas utilizan los resultados finales en el curso de los equipos y alumnos.
- ◆ Además de las dos técnicas de análisis, utilizan minería de secuencia de patrones.

En la Tabla2 se expone una matriz comparativa de las diferentes técnicas para el Análisis Colaborativo.

Método/ Autor(es)	Tipo	Algoritmo/ Método	Característica principal	Resultado
DomoSim-TPC (Redondo et al, 2003)	Cuantitativo	Lógica Borrosa	Mide interacción de los alumnos.	Grado de pertenencia Colaborativo
Talavera y Gaudioso (2004)	Cuantitativo	Clustering (Agrupamiento)	Comparación de grupos en base a la colaboración conocida y la	Identificación de comportamientos en los grupos.

			proporcionada por el tutor.	
(Daradoumis et al., 2006) & (Martinez et al, 2006)	-Cualitativo y Cuantitativo	Modelo de tres capas	-No proponen de método de inferencia.	-Mostrar el análisis en capas.
(De Pedro, 2007)	Cualitativo	Usa técnica de Wiki.	Etiquetación de contribución.	-Análisis y Evaluación de herramientas. -Ventajas y desventajas del método empleado.
(Meier et al., 2007)	Cualitativo	Estimación	-Plan de Evaluación. -Propone indicadores de colaboración -Escenas se graban en video.	Identificar la validez de los atributos cualitativos propuestos.
(Collazos et al., 2007)	Cuantitativo y Cualitativo	Análisis estadístico de interacción y de contenido.	-No proponen juicio ni inferencia.	-Identificar validez de atributos. -Monitorizar al estudiante.
Bratits et al., 2008)	Cuantitativo	Modelo de dos capas	-No realizan ninguna inferencia.	Visualización de los datos en una herramienta de monitorización.
(Perera et al., 2009)	Cuantitativo	Clustering	-Asignación de tareas para el desarrollo de documentos en equipo. -Utiliza secuencia de patrones.	Análisis Estadístico del trabajo colaborativo.

Tabla 1. 2 Resumen de Métodos de Análisis Colaborativo.

Sin importar el método, el ciclo del análisis de la minería de datos se cierra al utilizar el conocimiento extraído para la toma de decisiones. El método de adquisición de datos nos asegura que el análisis se puede realizar de forma regular y frecuente en cualquier circunstancia.

Tanto los modelos cualitativos como cuantitativos proponen una característica que los diferencia del resto, sean por la utilización o no de métodos de inferencia, el uso de patrones, indicadores, atributos, la técnica utilizada o el algoritmo empleado.

Para la selección de las herramientas y metodologías a tomar en cuenta se debe tomar en consideración dos aspectos:

1. El modelo a construir se basará en el alumno, pero el usuario final será el docente quien tomará de referencia las salidas para la evaluación de los alumnos.
2. El algoritmo empleado debe ser capaz de reflejar el comportamiento grupal de los estudiantes.

Es por ello que se ha seleccionado el modelo de (Talavera & Gaudio, 2004) , donde tanto estudiante como facilitador están ampliamente involucrados en el ambiente de cooperación, no podría ser de otra manera pues es el docente es quien de forma continua evalúa a su educando, detectando dificultades y pondera sus habilidades colaborativas.

1.8. Estudio de los Algoritmos de Agrupamiento

Según (García & Álvarez, 2003) Los algoritmos de agrupamiento buscan grupos de instancias con características similares, bajo una comparación entre valores de atributos de las instancias definidos en los algoritmos. El proceso de agrupar un conjunto de objetos físicos o abstractos dentro de clases con objetos similares se denomina clustering.

El clustering es una de las principales tareas en el proceso de minería de datos para descubrir grupos e identificar distribuciones y características interesantes en los datos. Consiste en agrupar una colección de datos en un conjunto de grupos de tal manera que los objetos que pertenecen a un grupo sean homogéneos entre sí, buscando que la heterogeneidad entre los distintos grupos sea lo más elevada posible.

En el proceso de clustering, no hay clases predefinidas ni registros muestra que permitan conocer las relaciones existentes entre los datos, los clusters o grupos se van creando de acuerdo a las características de los datos, no a una asignación de clases ya predefinidas, por lo que el clustering es también conocido como clasificación no supervisada. (Hernández Valadez, 2006)

Las formulas de distancia más usadas son:

Distancia euclidiana Es llamada también distancia clásica, definida como la longitud de la recta que une dos puntos en el espacio, se deduce a partir del Teorema de Pitágoras. Donde X y Y son las distancias origen y destino en una recta. Ecuación 1.

$$Eucl(x, y) = \sqrt{\sum_{i=1}^d (x_i - y_i)^2}$$

Ecuación 1 Fórmula de la Distancia Euclidiana

Distancia de Manhattan: Es llamada también distancia por cuadras (city block), está hace referencia a recorrer un camino no en diagonal (por el camino más corto) si no zigzagueando. Ecuación 2. El nombre viene precisamente como alusión del arreglo en rejilla de la mayor parte de las calles en la isla de Manhattan, y el concepto forma parte de la Geometría de Taxi, propuesta por Hermann Minkowski en el siglo XIX (IOS, 2006).

$$Manh(x, y) = \sum_{i=1}^d |x_i - y_i|$$

Ecuación 2 Ecuación Distancia de Manhattan

Distancia de Mahalanobis: Fue introducida por Mahalanobis en 1936. Es una distancia más robusta que utiliza la matriz de covarianzas S para incorporar la dependencia entre las dos variables. Formalmente, la distancia de Mahalanobis entre dos variables aleatorias con la misma distribución de probabilidad y con matriz de covarianza Σ se define como Ecuación 3.

$$Mahalanobis(x, y) = \sqrt{(x - y)Cov(D)^{-1}(x - y)}$$

Ecuación 3 Ecuación Distancia de Mahalanobis

La comparación de las distancias se muestra en la Figura 1.5

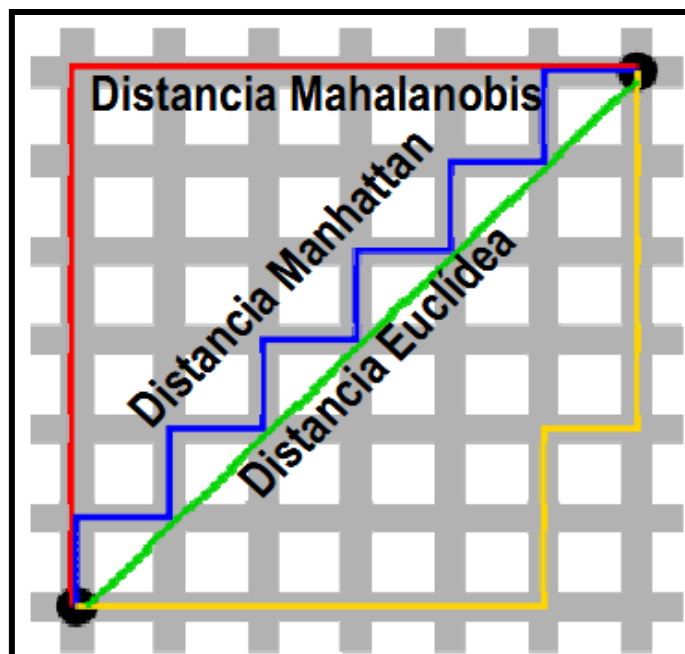


Figura 1. 5 Comparativa de cálculo de distancias (Cruz, 2010)

Los algoritmos de clustering se dividen en dos grandes grupos: Los *Agrupamientos Numéricos* y los *Simbólicos* entre los primeros están K-Medias, el Algoritmo Expectation Maximation, entre los Simbólicos, El algoritmo Clustering Jerárquico.

ALGORITMO K-MEDIAS: Se trata de un algoritmo clasificado como Método de Particionado y Recolocación. Este método es hasta ahora el más utilizado en aplicaciones científicas e industriales. El nombre le viene porque representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método únicamente se puede aplicar a atributos numéricos, y los outliers⁸ le pueden afectar muy negativamente.

Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. La suma de las discrepancias entre un punto y su centroide, expresado a través de la distancia apropiada, se usa como función objetivo. La función objetivo, suma de los cuadrados de los errores entre los puntos y sus centroides respectivos, es igual a la varianza total dentro del propio cluster.

El proceso con el que se desenvuelve el algoritmo consiste según (Molina & García, 2004) consiste en:

⁸ Outlier: Dato erróneo o Extremo que pertenece a una muestra de otra población que la estudiada.

1. Elegir k ejemplos que actúan como semillas (k número de clusters).
2. Para cada ejemplo, añadir ejemplo a la clase más similar.
3. Calcular el centroide de cada clase, que pasan a ser las nuevas semillas
4. Si no se llega a un criterio de convergencia (por ejemplo, dos iteraciones no cambian las clasificaciones de los ejemplos), volver

En la Figura 1.6 se explica de una forma gráfica el procedimiento que se lleva a cabo.

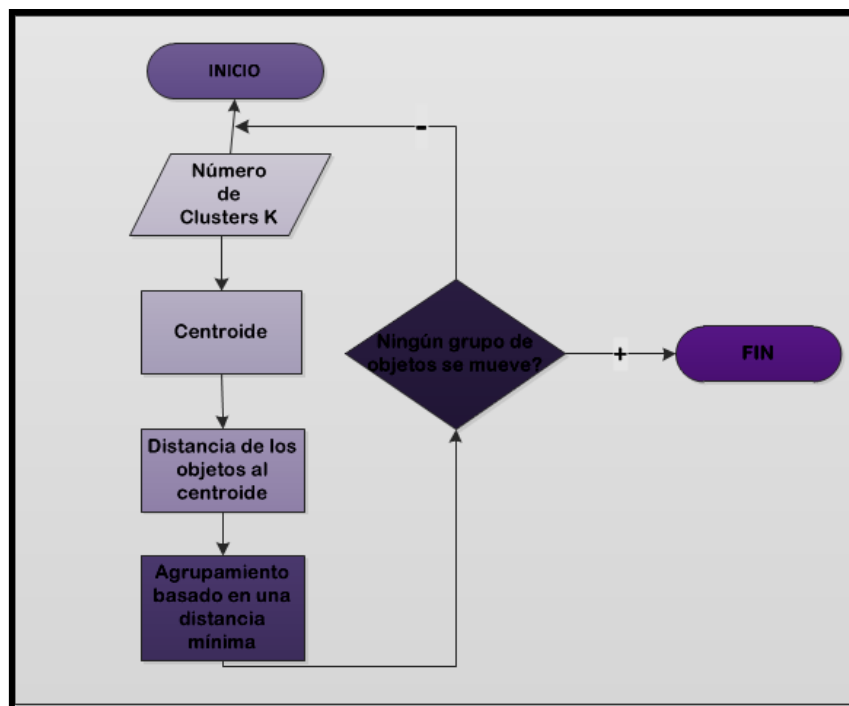


Figura 1.6 Algoritmo K-Means Adaptación de (Saharkhiz, 2009)

Para la realización de este trabajo se aplicará el algoritmo de agrupamiento K-medias, por ser uno de los más veloces y eficientes, aunque también hay que decir que es uno de los más limitados. Este algoritmo precisa únicamente del número de categorías similares en las que queremos dividir el conjunto de datos.

ALGORITMO EXPECTATION-MAXIMATION: El algoritmo EM asigna a cada instancia una distribución de probabilidad de pertenencia a cada cluster. El algoritmo puede decidir cuántos clusters crear basado en validación cruzada o se le puede especificar a priori cuántos debe generar.

Utiliza el modelo Gaussiano finito de mezclas, asumiendo que todos los atributos son variables aleatorias independientes. Este algoritmo es bastante más elaborado que el K-Medias, ya que requiere muchas más operaciones.

El ajuste de los parámetros del modelo requiere alguna medida de su bondad, es decir, cómo de bien encajan los datos sobre la distribución que los representa. Este valor de bondad se conoce como el likelihood o verisimilitud de los datos. (Garre, Cuadrado, & Sicilia, 2005). Se conforma principalmente de dos pasos:

- ◆ **(Expectation)** Para cada instancia calcular la probabilidad de pertenecer a un cluster.
- ◆ **(Maximization)** Estimar los parámetros que caracterizan las distribuciones a partir de las nuevas probabilidades (hacer más probable esa observación).

Donde las probabilidades de los clusters se almacenan como pesos asociadas a las instancias.

En el algoritmo k-medias se finalizaba cuando ningún ejemplo de entrenamiento cambiaba de cluster en una iteración, alcanzándose así un “punto fijo”, el algoritmo EM es un poco más complicado, dado que tiende a converger pero nunca se llega a ningún punto fijo. (Molina & García, 2004)

CLUSTERING JERÁRQUICO: “Se caracterizan porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo. Pueden ser, a su vez de dos tipos: aglomerativos y divisivos.

Los métodos aglomerativos comienzan con n clusters de un objeto cada uno. En cada paso del algoritmo se recalculan las distancias entre los grupos existentes y se unen los 2 grupos más similares o menos disimilares. El algoritmo acaba con 1 cluster conteniendo todos los elementos. Los métodos divisivos comienzan con 1 cluster que engloba a todos los elementos. En cada paso del algoritmo se divide el grupo más heterogéneo. El algoritmo acaba con n clusters de un elemento cada uno”. (Figueras, 2001).

Según (Rodríguez Anaya, 2009) un planteamiento extremo poco operativo sería la obtención de un gran número de clusters o grupos de instancias y que cada grupo significara un nivel de colaboración distinto. Al dar una etiqueta del lenguaje natural a cada grupo, podría ser difícil de entender o de establecer la diferencia entre distintos grupos.

Este trabajo se delimitará al estudio de los Algoritmos K-Medias, EM y Clustering Jerárquico. Se agrupará a los alumnos usando tres etiquetas:

- ◆ Alumnos con Colaboración Alta
- ◆ Alumnos con Colaboración Media
- ◆ Alumnos con Colaboración Baja

Se ha utilizado esta clasificación pues permite definir categóricamente el comportamiento colaborativo de los estudiantes sin caer en ambigüedades.

1.9. Herramientas para la Minería de Datos

Existen un sinnúmero de herramientas tanto libres como comerciales igual de poderosas para la manipulación de grandes cantidades de datos.

En esta tesis se ha seleccionado una de cada tipo, WEKA, Clementine SPSS y KNIME, por la versatilidad y potencia que ofrecen.

WEKA: Es un conjunto de librerías JAVA para la extracción de conocimientos desde bases de datos, soporta varias tareas típicas de minería de datos, especialmente pre procesamiento de datos, agrupamiento, clasificación, regresión, visualización y características de selección.

Es un software ha sido desarrollado bajo licencia GPL lo cual ha impulsado que sea una de las suites más utilizadas en el área en los últimos años.

Incluye las siguientes características:

- ◆ Diversas fuentes de datos (ASCII, JDBC).
 - Interfaz visual basado en procesos/flujos de datos (rutas).
 - Distintas herramientas de minería de datos: reglas de asociación (a priori, Tertius, ...).
- ◆ Agrupación/segmentación/conglomerado (Cobweb, EM y k-medias), clasificación (redes neuronales, reglas y árboles de decisión, aprendizaje Bayesiana) y regresión (Regresión lineal, SVM..).
- ◆ Manipulación de datos (pick & mix, muestreo, combinación y separación).
 - Combinación de modelos (Bagging, Boosting ...)
- ◆ Visualización anterior (datos en múltiples gráficas) y posterior (árboles, curvas ROC, curvas de coste..).
- ◆ Entorno de experimentos, con la posibilidad de realizar pruebas estadísticas (t-test). Sus técnicas se basan en la hipótesis de que los datos están disponibles en

un único archivo plano o relación, donde cada punto marcado es etiquetado por un número fijo de atributos. (Orallo & Ferri, 2006)

CLEMENTINE SPSS: Es una herramienta de Data Mining que permite desarrollar modelos predictivos y desplegarlos para mejorar la toma de decisiones. Está diseñada teniendo en cuenta a los usuarios empresariales.

Es la solución líder en minería de datos que ayuda a las organizaciones a comprender el comportamiento de las personas y a predecir qué es lo que harán. Al utilizar Clementine, los analistas y usuarios de negocios podrán acceder datos de varias fuentes para producir, evaluar, y desplegar modelos analíticos rápida y fácilmente. La arquitectura abierta y escalable del producto permite obtener el máximo provecho de la infraestructura actual, haciendo de la minería de datos un proceso efectivo. Ahora llamado SPSS Modeler. (IBM®, 2011)

KNIME: Knime es una plataforma modular de exploración de datos, que permite a un usuario crear flujos de datos, o pipelines, de forma visual e intuitiva. Además permite ejecutar de forma selectiva algunos de los pasos creados, así como ejecutar todo el flujo desarrollado. Tras la ejecución, los resultados se pueden investigar mediante vistas interactivas tanto de los datos como de los modelos. (Guerra, 2008)

En este capítulo se ha analizado la situación actual de los entornos colaborativos, las adaptaciones de los autores en cada uno de sus estudios, el proceso de modelado del usuario, el de minería de datos, los principales algoritmos de agrupamiento y las herramientas especializadas en la detección de comportamientos colaborativos, esto ha permitido efectuar la identificación de todos los interesados, orientar esta tesis al siguiente paso de la misma y a definir el modelo que más se ajusta al escenario planteado, se ha seleccionado para el efecto las propuestas de Cristóbal Romero y de Talavera & Gaudioso, estudiadas a lo largo de este apartado, las cuales en caso de ser necesario en el transcurso de este estudio se realizarán las debidas modificaciones.

1.10. Descripción de los Foros en MOODLE

La presente investigación ha delimitado su campo de estudio a la actividad de FOROS en MOODLE, por lo que es preciso conocer su funcionamiento y las alternativas de configuración que posee.

Esta actividad tal vez sea la más importante siendo a través de los foros donde se da la mayor parte de los debates y discusión de los temas del curso. Se dice que esta actividad es asincrónica ya que los participantes no tienen que acceder al sistema al mismo tiempo. Su icono estándar es: 🗨️

En todas las asignaturas del EVA por lo menos existe un foro por cada bimestre, el tipo que se utilice dependerá de la configuración que le haya dado el profesor y de la forma en cómo se desea emitir y captar la información.

Como muestra de las funcionalidades que se pueden agregar a los foros se exponen las siguientes cabeceras configurables al momento de su creación. Figura 1.7

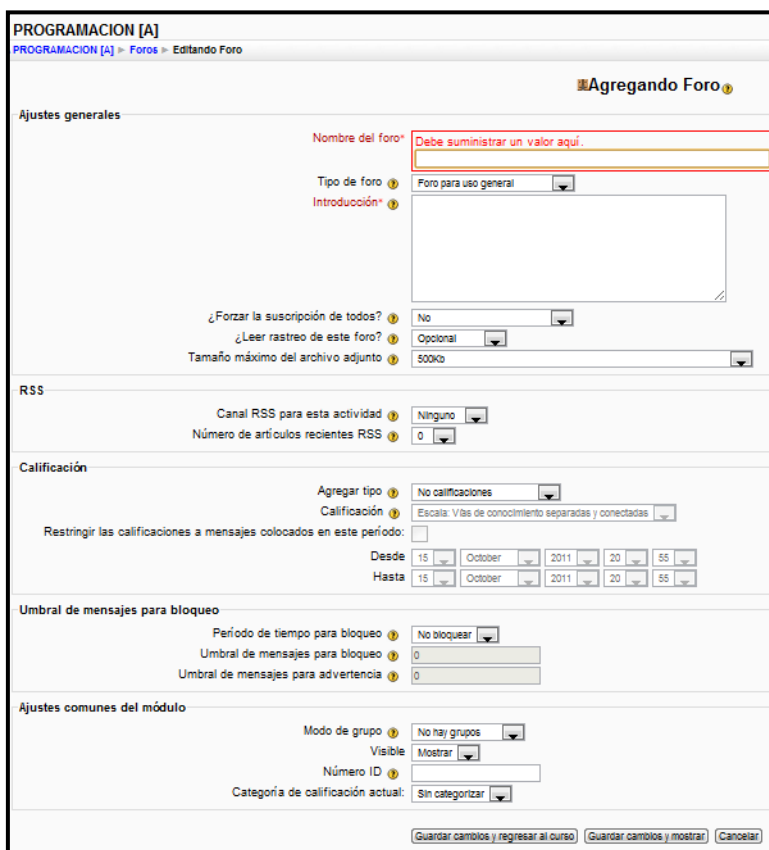


Figura 1. 7 Ventana de Creación de Foros en moodle

- ◆ **Nombre del Foro**
- ◆ **Tipo de Foro:** En la página oficial de MOODLE (MOODLE, 2009) se presentan dos categorías:
 - **Foro de Uso General:** Ubicado en la Sección 0 del Curso.
 - **Foro de Aprendizaje:** Foros de alguna sección específica de un curso.

Además de esta clasificación los foros técnicamente pueden dividirse en: Figura 1.8

- **Foro Normal para uso general:** Es un foro abierto donde cualquiera puede empezar un nuevo tema de debate cuando quiera.
- **Foros de debate sencillo:** Simplemente un intercambio de ideas sobre un solo tema, todo en un página, respondiendo a un único planteamiento inicial. Útil para debates cortos y muy concretos.
- **Foro un Debate por Persona:** Donde cada persona puede plantear un nuevo tema de debate y todos pueden responder.
- **Foro P y R:** Cuando se desea que una pregunta en particular sea contestada. En un foro P & R, los tutores lanzan la pregunta y los estudiantes contestan con posibles respuestas. Por defecto, un foro P & R requiere que un estudiante conteste una vez antes de ver las respuestas de los otros estudiantes. Esta característica permite una igualdad de oportunidades para la respuesta inicial entre todos los estudiantes, fomentando el pensamiento original e independiente.

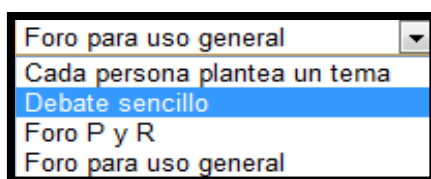


Figura 1. 8 Tipos de Foros existentes en moodle

Introducción: Descripción general del foro y su intención.

Suscripción: Nos indica la aprobación de recibir o no los mensajes en el correo electrónico. Las personas generalmente pueden elegir si quieren estar suscritas a cada foro o no. Sin embargo, el profesor puede elegir forzar la suscripción a un foro en particular donde todos los usuarios del curso se suscribirán automáticamente, incluso los que se incorporen más tarde.

Rastreo: Permite a los usuarios hacer un seguimiento de los mensajes leídos y no leídos del foro. Hay tres opciones para este parámetro:

- **Opcional** (por defecto) - Los estudiantes pueden activar o desactivar el seguimiento en el foro a discreción.

- **Conectado** - el seguimiento siempre está activado en este foro para todos los miembros.
- **Desconectado** - el seguimiento siempre está desactivado en este foro para todos los miembros.

Tamaño máximo del Adjunto: Es posible adjuntar archivos binarios a los mensajes de los foros. Se puede limitar el tamaño de esos archivos (desde el máximo permitido por el Campus virtual) o deshabilitar la posibilidad de adjuntarlos. (CES. S. RAMÓN Y CAJAL, 2008)

RSS: Cuando el sitio moodle tiene habilitados los Canales RSS, en la configuración de los foros aparecen las siguientes opciones:

- **Canales RSS en esta actividad:** Esta opción nos permite habilitar los canales RSS en este foro. Podemos escoger entre dos tipos de foros:
- **Número de artículos RSS recientes:** Esta opción permite seleccionar el número de artículos a incluir en el canal RSS. Si este número se ajusta a 5, los 5 artículos más recientes serán enviados a los suscriptores. Cuando haya nuevos temas (o discusiones) los más antiguos serán reemplazados en el canal RSS. Un número comprendido entre 5 y 20 puede ser apropiado para la mayoría de los foros. Auméntelo sólo si se trata de un foro muy utilizado.

Calificación: En moodle los foros son calificables mediante una escala. Figura 1.9 Por defecto, solamente los profesores pueden calificar mensajes de foros, la opción “Anular Permisos” posibilita que los estudiantes también califiquen los mensajes de los demás. Esta herramienta es útil para dar niveles de participación a los estudiantes, cualquier calificación dada en el foro se graba en el Libro de Calificaciones. (MOODLE, 2009).

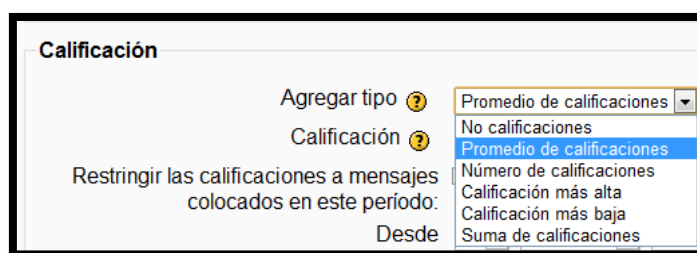


Figura 1.9 Tipo de calificaciones para Foros

Umbral de mensajes para bloqueo: Especifica el número de mensajes que puede enviar un estudiante en el período definido.

Período de tiempo para bloqueo: Es el período dentro del cual el estudiante no puede enviar más que un número determinado de mensajes.

Umbral de mensajes para advertencia: Esta opción es para indicar el número de mensajes que un estudiante puede hacer antes de recibir un aviso. Si se pone 0, se deshabilitan los avisos.

Modo de Grupo: La configuración del modo de grupo tiene tres opciones:

No hay grupos

Grupos separados: Cada grupo puede ver solamente a su propio grupo, los otros son invisibles.

Grupos visibles: Cada grupo trabaja dentro de su propio grupo, pero puede ver también a los otros.

Si el modo es de grupos separados:

- ◆ Los profesores tienen la opción de agregar un nuevo tema de debate a todos los participantes o a un grupo seleccionado.
- ◆ Los estudiantes solamente pueden empezar discusiones dentro de su grupo.
- ◆ Los estudiantes solamente pueden contestar a discusiones iniciadas por otros miembros de su grupo o discusiones iniciadas por un profesor para su propio grupo.

Si el modo es de grupos visibles:

- ◆ Los profesores tienen la opción de agregar un nuevo tema de debate a todos los participantes o a un grupo seleccionado. Si un profesor añade un nuevo tema de debate a un grupo, entonces solamente los miembros del grupo pueden responder a él.
- ◆ Los estudiantes solamente pueden empezar discusiones dentro de su grupo.
- ◆ Los estudiantes solamente pueden contestar a discusiones iniciadas por otros miembros de su grupo o profesores.

Visible para los estudiantes: Se puede ocultar la actividad a los estudiantes seleccionando "Ocultar".

Número ID: Especificar un número ID proporciona una forma de identificar el foro para propósitos de cálculo de calificaciones. Si la actividad no está incluida en ningún cálculo de calificaciones, entonces se puede dejar en blanco. (MOODLE, 2009)

CAPÍTULO II

MINERÍA DE DATOS APLICADA A ENTORNOS COLABORATIVOS DEL EVA

2.1. Introducción

En este capítulo se trabajará con el proceso de minería de datos estudiados en la primera parte de la tesis y la implementación de las técnicas de (Talavera & Gaudioso, 2004).

Para ello se comenzará con la etapa de Pre-Procesamiento que incluye la selección y limpieza de datos, luego se empleará los algoritmos K-MEANS y EM como complemento para la determinación de patrones de comportamiento similares entre los estudiantes.

Luego de terminada la sección de minería de datos se procederá a la obtención de patrones y análisis de resultados

2.2. Metodología

La base de datos con la que se va a trabajar en esta primera parte de la minería corresponde a la del Periodo Octubre-Febrero2011.

El proceso que se va a seguir pertenece al de (Hernandez, Ramirez, & Ferri, 2004) Figura 2.1, con cada una de las tareas que esta estructura plantea.

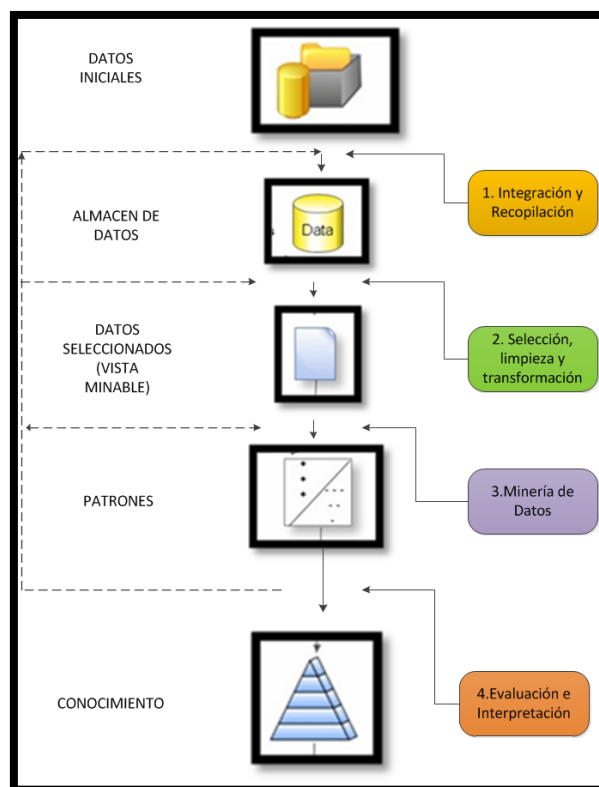


Figura 2.1 Fases del Proceso de Extracción de Conocimiento (Hernández, Ramírez, & Ferri, 2004)

Las herramientas que se han seleccionado, se basan unas en las características propias del proyecto, mientras que otras en las características de la aplicación, en las propiedades de complemento a herramientas similares y a su facilidad de uso. En la Tabla 2.1 se especifican las herramientas a utilizar.

Representación	Herramienta	Función
SOFTWARE BASE	MOODLE	Herramienta educativa para la interacción.
MANIPULACIÓN DE SOFTWARE	GEPHI	Para la selección de las materias con mayor número de interacciones en base a los docentes con mayor número de conexiones.
SERVIDOR	XAMPP	Software libre que integra la base de datos MySQL, el servidor Apache, Perl y PHP.
DBA MANAGER	PHPMyAdmin(incluido en XAMPP), NAVICAT Enterprise, DreamCoder for MySQL.	Permiten la manipulación de la base de datos, cada uno con características de complemento al otro.
VALIDACIÓN Y EJECUCIÓN DE ALGORITMOS	WEKA 3.6 Clementine SPSS KNIME	Soporte a la minería de datos.

Tabla 2. 1 Herramientas para el descubrimiento de información

Se ha considerado una etapa previa al preprocesamiento de los datos para su mejor comprensión haciendo una integración de la base de datos del EVA a MOODLE, lo que puede apreciar en el **ANEXO A**.

2.3. Integración y Recopilación

2.3.1. Selección de Tablas

Como ya se ha mencionado el objetivo de esta tesis es determinar el nivel de Colaboración en el entorno EVA, específicamente se analizará la interacción existente en los foros.

A nivel de base de datos las principales entidades involucradas en la investigación son las que se muestran en la Tabla 2.2

Entidad	Descripción
prefix_forum	Información acerca de todos los foros.
prefix_forum_post	Almacena todos los mensajes de los foros.
prefix_forum_discussion	Almacena todos los foros de discusión
prefix_log	Registro de acciones de todos los usuarios

Tabla 2. 2 Descripción de tablas

Las tres primeras tablas se utilizarán más adelante para la cuantificación de la colaboración en determinadas asignaturas mientras que prefix_logs es quizás la máxima referencia para la identificación de eventos colaborativos.

Tanto prefix_forum, prefix_forum_discussions como prefix_forum_posts corresponden a los tipos de foros existentes. El interés de este trabajo no es tratar a cada uno de ellos por separado, sumando a esto la correspondencia entre estas entidades como se muestra en el diagrama entidad-relación de la Figura 2.2 se ha creído conveniente tratar el módulo Foros sin diferencia de forma.

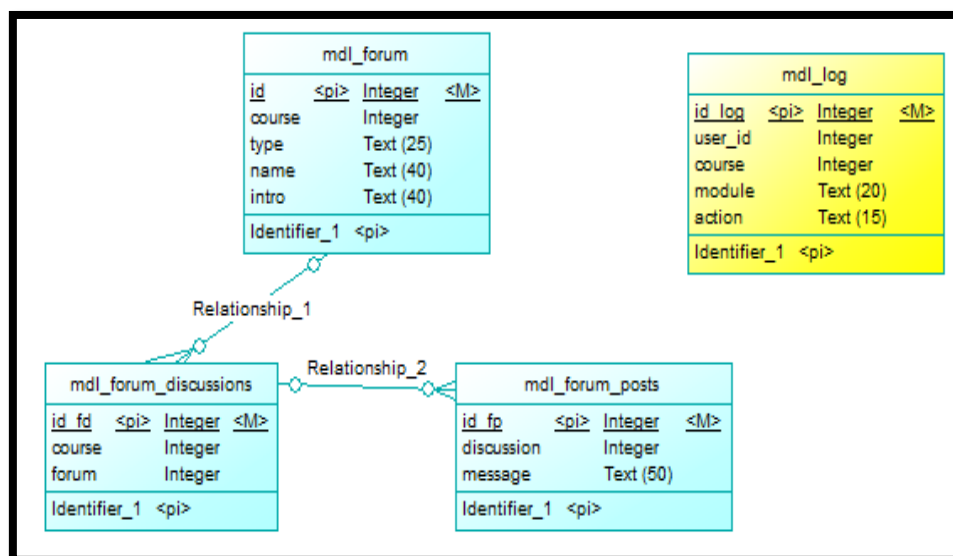


Figura 2. 2 Diagrama Entidad Relación Foros y logs

En la Tabla 2.3 se presenta una descripción de los principales atributos de cada entidad.

Tabla	Atributo	Descripción
prefix_forum	Id (todas las entidades)	Identificador único de cada entidad.
	course	Identificación del número de curso.
	type	Índole del foro: news, single, teacher, general. (novedades, foro único, foros de profesor y foros de tema general respectivamente)
	name	Título o tema del foro
	intro	Mensaje de inicio del foro
prefix_forum_discussions	fórum	Identificador del foro al que pertenece la discusión
prefix_forum_posts	discussion	Referencia al identificador de la tabla prefix_forum_discussions
	message	Mensaje de participación del for
prefix_logs	userid	Identificador de la tabla de usuarios.
	module	Tipo de actividad realizada en este caso el que evaluará será: fórum.
	action	Tipo de acción que realiza el usuario

Tabla 2. 3 Descripción atributos principales de Foros y Logs.

prefix_forum, prefix_forum_discussions, prefix_forum_posts cada una de ellas están relacionadas de tal manera que brinda información complementaria de los foros como los temas de discusiones, mensajes enviados, a que curso se realizó el aporte, etc.

Cabe recalcar que el atributo *action* será únicamente como identificativo de la que la acción “add discussion” y “add post” se estén dando.

Id_usuario	Número de Interacciones
5	151
33	111
2879	73
44	42
453	35
2912	27
3	26
3087	26
32100	21
3523	20

Tabla 2. 4 Recopilación de nodos con mayor conexión⁹.

Los 10 facilitadores laboran en la Carrera de informática. A continuación se seleccionará aleatoriamente dos materias de las que dirigen los docentes mencionados, se puede observar esto en la Tabla 2.5

Id Docentes	Materias que imparten	Plan de Estudio	Id_plan	Id_curso
5	Fundamentos de la Programación[A]	Informática UTPL-ECTS-1A	2251	28741
	Lenguaje de Alto Nivel [A]	Informática	1684	17323
33	Lógica de la programación [B]	Informática UTPL-ECTS-1A	2251	28737
	Lógica de la programación[A]	Informática UTPL-ECTS-1A	1859	21221
2879	Fundamentos Informáticos [A]	Informática UTPL-ECTS-1A	2251	28739
	Fundamentos Informáticos [A]	Informática UTPL-ECTS-1A	1859	21222
44	Base de datos II [A]	Informática	437	2132
	Base de datos II [A]	Informática	1886	22065
453	Bases de Datos I [A]	Informática	437	2131
	Sistemas III	Informática	2250	29012

⁹ Cabe recalcar que en el puesto 9no se encontraba jgochoa un estudiante, debido a los objetivos el proyecto no se lo contempló.

2912	Sistemas basados en el conocimiento [A]	Informática	437	2170
	Lógica de la Programación [C]	Informática UTPL-ECTS-1A	2251	30055
3	Lógica Matemática [A]	Informática	1684	17317
	Lógica Matemática [A]	Informática	1189	12574
3087	Sistemas basados en el conocimiento [A]	Informática	2250	28991
	Sistemas basados en el conocimiento [A]	Informática	2072	25287
32100	Redes y Sistemas Distribuidos [A]	Informática	2250	28989
	Redes y Sistemas Distribuidos [A]	Informática	1684	17354
3523	Estadística [A]	Informática UTPL-ECTS-1A	2251	28731
	Estadística Analítica[A]	Informática	1886	22114

Tabla 2. 5 Materias y Docentes con mayor nivel de Interacción.

Las consultas que se van hacer contra la base de datos son de tipo Cuantitativo, se utilizará para facilitar la misma los identificadores de la Tabla 2.6.

Se comprueba cada una de las materias explicadas anteriormente agrupándolas por docente.

MATERIAS												
<i>Id Doce nte</i>	<i>Fund amen tos de la progr amación</i>	<i>Len guaje de alto nive l</i>	<i>Lógica de la Progra mación</i>	<i>Funda mento s Inform áticos</i>	<i>Ba se de Da tos II</i>	<i>Bas e de dat os I</i>	<i>Si ste mas III</i>	<i>Siste mas basad os en el conoc imient o</i>	<i>Lógic a Mate mática</i>	<i>Rede s y Siste mas Distri buido s</i>	<i>Esta dística</i>	<i>Esta dística Analí tica</i>
5	798	538										
33			237 227									
2879				102 304								
44					71 99							
453						44	66					
2912			109					34				

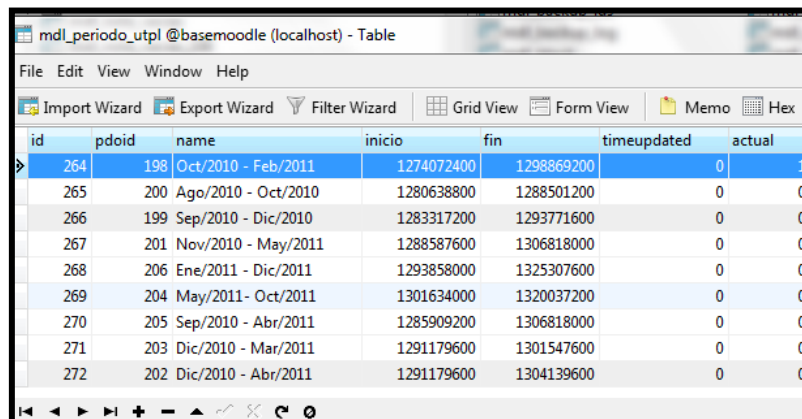
3								31 29				
3087									648 470			
3210 0										191 118		
3523											80	99

Tabla 2. 6 Resumen Colaboración Foros en materias.

En el **ANEXO B** se encuentra el procedimiento para encontrar el número de colaboraciones de forma general en cada una de las materias, un resumen de esta práctica se puede observar en la Tabla 8.

En base al resultado de este análisis se ha considerado tomar los cursos de: Fundamentos de la Programación [A], Lógica de la Programación [B], Fundamentos Informáticos [A] de los facilitadores con el id: 5, 33, 2879 respectivamente.

Como se ha manifestado a lo largo de esta tesis los cursos evaluados deben corresponder al periodo: Octubre/Febrero 2011, para mayor seguridad de que esto se cumpla se consulta la tabla “prefix_periodo_utpl”, Figura 2.4.



id	pdoid	name	inicio	fin	timeupdated	actual
264	198	Oct/2010 - Feb/2011	1274072400	1298869200	0	1
265	200	Ago/2010 - Oct/2010	1280638800	1288501200	0	0
266	199	Sep/2010 - Dic/2010	1283317200	1293771600	0	0
267	201	Nov/2010 - May/2011	1288587600	1306818000	0	0
268	206	Ene/2011 - Dic/2011	1293858000	1325307600	0	0
269	204	May/2011 - Oct/2011	1301634000	1320037200	0	0
270	205	Sep/2010 - Abr/2011	1285909200	1306818000	0	0
271	203	Dic/2010 - Mar/2011	1291179600	1301547600	0	0
272	202	Dic/2010 - Abr/2011	1291179600	1304139600	0	0

Figura 2. 4 Búsqueda de periodo en tabla prefix_periodo_utpl

Una vez identificado el periodo, en la tabla “prefix_plan_utpl” se realizará una nueva búsqueda, esta vez de todos los planes de estudio dentro del periodo referido, basándonos en los cursos descritos en la Tabla 2.5. Los resultados obtenidos se muestran en la Figura 2.5.

pdoid is equal to 198
[<Click here to add>](#) <Move Up> <Move Down> <C

id	category	plnid	pdoid
2053		2245	1022
2054		2246	1009
2055		2247	781
2056		2248	623
2057		2249	1014
2058		2250	1
> 2059		2251	1012
2060		2252	536
2061		2253	910
2062		2254	1002
2063		2255	1001
2064		2256	1023
2065		2257	10
2066		2258	109
2067		2259	1007
2068		2260	1113
2069		2261	1021
2070		2262	1005
2071		2263	14
2072		2264	1020
2073		2265	487
2074		2266	0
2076		2268	319
2077		2269	637
2078		2270	632
2079		2271	634

SELECT * FROM 'mdl_plan_utpl' WHERE 'pdoid' = '1

Figura 2. 5 Planes Académicos correspondientes al periodo Octubre-Febrero 2011

De las 20 materias listadas en la Tabla 2.5 se seleccionaron las 3 primeras que cumplían con los requisitos de:

- ◆ Estar dentro del periodo estipulado y que
- ◆ El plan de estudios correspondiera a la malla de créditos “UTPL-ECTS-1A”.

2.4. PREPROCESAMIENTO

Según (S. Zhang, 2003) el Preprocesamiento de Datos engloba a todas aquellas técnicas de análisis de datos que permite mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento/minería de datos puedan obtener mayor y mejor información. Si queremos buenos resultados indudablemente necesitamos buenos datos.

2.4.1. Selección, Limpieza y transformación

Para este fin se consideraron algunas opciones: **Anexo C**, más estas no contenían atributos factibles para la extracción del conocimiento, puesto que en la primera opción se toma tan solo la información contenida en la tabla *logs* mientras que la segunda

opción si bien se tomaron atributos de otras tablas no se proporcionaba información específica relevante al tópico foros.

Entonces se consideró tomar en cuenta los siguientes campos, unos propios de la tabla que las contiene y otros nuevos con cálculos, o lo que se conoce como agregación¹⁰.

Para el proceso de clustering entre mayores atributos se especifiquen, mayor será la similitud con la que se formen los grupos (siempre y cuando estos provean de información efectiva) es por esto que en la Tabla 2.7 se presenta una recopilación de atributos capaces de proporcionar información específica de los miembros de una asignatura.

Atributo	Origen	Descripción
user_id	prefix_Log	Identificación del Usuario.
sexo_usr	prefix_user_utpl	Sexo de los individuos, M si es Masculino, F si es Femenino
num_acceso_foros	prefix_log	Número de veces que un usuario ha accedido a un curso, se utiliza la acción "view forum"
subtemas_leidos	prefix_log	Número de veces en las que un estudiante ha leído un hilo de mensajes.
num_respuestas_post	prefix_log	Número de veces que un estudiante ha respondido o agregado un hilo en el foro.
num_respuestas_debates	prefix_forum_discussions	Número de veces que un estudiante ha agregado un mensaje en una discusión como respuesta a otro mensaje.
num_mens_act	prefix_log	Número de veces que un usuario ha actualizado un mensaje en el foro.
arch_adjuntos	prefix_forum_posts prefix_forum_discussions	Archivos adjuntos a los mensajes en foros.
numForos_subscr	prefix_forum	Foros en los que se ha

¹⁰ Agregación: Consiste en crear nuevos atributos para mejorar la calidad, visualización o comprensibilidad del conocimiento extraído, sin sustituir los valores iniciales (Hernández, Ramírez y Ferri, 2004)

	prefix_forum_subscriptions	subscrito un usuario.
prom_horas	prefix_log	Número de horas promedio que un usuario ha usado para la gestión de foros.
nota_final	Reporte de Moodle (ver Anexo D).	Nota final sobre 100, calificación de todas las actividades y evaluación vía EVA.
course	prefix_log prefix_course	Identificación de la asignatura

Tabla 2. 7 Descripción de atributos usados para la recopilación de información en foros.

En una etapa inicial se contemplaron los atributos `user_id`, `total_num_interaccion_foros`, `num_debates`, `total_num_mensajes_foros`, `num_veces_foros_add`, `segundos`, `course`, pero como se verá más adelante **Sección 2.5.2** no reunían las características necesarias para la determinación de comportamientos colaborativos en Foros.

Para la asignatura de Fundamentos de la Programación [A] se tomaron en tres atributos adicionales, debido a que este curso contenía la calificación de foros (`rating`)¹¹ dentro del sistema. Tabla 2.8

Atributo	Descripción
calif_prom_foro_1bim	Promedio de calificaciones en foros correspondientes al primer bimestre.
calif_prom_foro_2bim	Promedio de calificaciones en foros correspondientes al segundo bimestre.
prom_foros	Promedio de foros 1er y 2do Bimestre.

Tabla 2. 8 Descripción atributos adicionales curso "Fundamentos de la Programación"

La metodología para obtener estos valores incluyendo "*nota_final*" (que se incluye en los tres cursos) se ha basado unos en consultas individuales para cada uno de los campos y otros los últimos mencionados en reportes desde MOODLE. Este proceso

¹¹ El rating fue configurado por los profesores de tal manera que sólo ellos tuviera la capacidad de asignarlo.

está disponible en el **ANEXO D**. Cada uno de los atributos obtenidos se colocó a modo de matriz en una hoja de EXCEL **ANEXO E**.

Luego de esto se realizó su exportación a la base de Datos (**Anexo D. Sección D.7**) y de la base de datos a formato CSV (**Anexo D. Sección D.8**) con el que se trabajará en la próxima etapa.

2.5. MINERÍA DE DATOS Y EXTRACCIÓN DE PATRONES

Esta etapa comprende la ejecución de los algoritmos de agrupamiento. Se ha trabajado con: K-Means, EM y Clustering Jerárquico. Para el tratamiento automatizado y para la ejecución de las pruebas de rigor de los datos se ha seleccionado la Herramientas WEKA.

Lo que se busca con la aplicación de estas técnicas es la generación de modelos que describan patrones y relaciones entre los datos.

Un modelo es una descripción global del conjunto de datos. Toma una perspectiva completa y total. En contraste un patrón es una propiedad local de los datos, tal vez sólo la tienen ciertas instancias o atributos. (Nevárez).

2.5.1. Prueba de Carga de Datos en WEKA

En WEKA existen algunas formas de cargar la data para ser analizada, en forma de archivos en formato .csv, artff propia de WEKA, por medio de base de datos, o si se encuentra alojada en la WEB por medio de su URL.

En primera instancia se ha decidido realizarlo por medio de la base de datos con ayuda del **mysql-connector-java-5.0.8-bin** pero por la dificultad que representaba básicamente por el tipo de dato admitido **Anexo F** se optó trabajar con archivos de formato .CSV. Se procede como en el **Anexo G**.

2.5.2. Resultados de los Algoritmos de Agrupación

Lo que sigue es la ejecución de los algoritmos de agrupación mencionados en la **Sección 1.8** para cada uno de los cursos y la visualización de sus resultados.

En la primera prueba se observó que la variable “tiempo” ejercía una enorme influencia en el incremento de la suma de error cuadrático¹² **Anexo J** (Sección J.1), afectando la distribución de grupos en los tres cursos seleccionados, el tipo de dato al que correspondía esta instancia (Date) alteró la conformación de los conglomerados, por lo que la experimentación fue descartada.

Partiendo de esta premisa se hizo cambios dentro de la etapa de pre procesamiento. En el segundo experimento **Anexo J** (Sección J.3) se decidió convertir a segundos el valor dado en el formato “hh:mm:ss”, fundamentándose en que en primer lugar “los datos se someten a un proceso de estandarización” (Molina & García, 2004).

Existe la posibilidad de eliminar el atributo time_promedio del primer experimento con lo que se reduciría, incluso sería menor que en el segundo experimento la suma del error cuadrático, pero esta es una variable imprescindible para la determinación del comportamiento colaborativo, por lo que se decidió transformarla.

La segunda prueba si bien arrojó resultados con un menor índice de error cuadrático. (Algoritmo K-Means) **Anexo J** (J.3.1) no proporcionaba los elementos suficientes para un análisis efectivo del comportamiento colaborativo en foros, a esto se le sumó la baja probabilidad (Algoritmo EM) **Anexo J** (J.3.2) de que un estudiante pertenezca a un cluster.

En esta misma experimentación se abordó el uso de filtros para discretizar los atributos, resultando ser una técnica efectiva en el ordenamiento e identificación de tendencias, pero que debe ser usado de forma no recurrente pues su uso tiende a incrementar en el algoritmo K-Means la suma de errores cuadráticos.

Otro aspecto relevante que se comprobó con EM es que en el cambio en los parámetros de configuración tanto de iteraciones como número de semillas se mostró invariable la verisimilitud.

Como resultante de las dos primeras experiencias se efectuó una tercera tomando en cuenta las deficiencias de sus antecesoras.

¹² Suma de error cuadrático: Suma de errores en la formación de grupos, comportamiento de la función objetivo.

2.5.2.1.1. Tercera Experimentación

Producto de una retroalimentación se consideró el incremento de las instancias “sexo_usr”, “número_acceso_foros”, subtemas_leídos, num_mensajes_act, arch_adjuntos, numForos_subscr, ”. Descritos en la **Tabla 2.7**, con el fin de obtener una visión más amplia de la colaboración realizada por el estudiante. El tiempo empleado por el educando ya no se contempló en segundos sino en horas su participación. **Anexo H**.

2.5.2.1.1.1. Descripción del Procedimiento

Si bien se utilizaron en las primeras prácticas los Algoritmos K-Means y EM se advertirá un tercer algoritmo, el de Clustering Jerárquico para la experimentación con cada una de las materias, los modelos se evaluarán, finalmente se seleccionará aquel cuyos resultados reflejen una asignación de grupos más compacta y con un rango de error mínimo.

Se realizará una descripción del curso para luego proceder con la ejecución de cada uno de los algoritmos.

2.5.2.1.1.1.1. Fundamentos de la Programación

Fundamentos de la Programación cuenta con foros tanto para la exposición de consultas realizadas por el docente como para las resoluciones de ejercicios.

En total existen 10 secciones de las cuales 3 anidan dos foros en lugar de uno. En total de 13 foros, 8 de los cuales reflejan el criterio de los estudiantes los restantes el desarrollo de algoritmos. De igual forma se observa que todos los mensajes han sido leídos (al menos una vez) . **Figura 2.6**

Sección	Foro	Descripción	Temas	Mensajes no leídos	Rastrear	Suscrito	RSS
3	Foro 1 - 1er Bim	Consulte y comparta en el foro sobre una herramienta que ayuda a aprender a programar en Java que se denominada GreenFoot y trate de responder a la pregunta, según su criterio, ¿Cuáles son sus ventajas?	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	Solución ecuaciones de segundo grado	Suba la solución al problema propuesto	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	Foro 2 - 1er Bim	Una de las características de Java se que es un lenguaje Multihilo (Multithread) ¿Qué significa esto? ¿Qué tipos de aplicaciones podríamos construir aprovechando dicha característica?	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6	Foro - Solución Validación de Datos	Elabore un programa Java que permita validar el ingreso de datos hecho por el usuario. El programa debe tener el siguiente comportamiento: El programa solicita que se ingrese la edad de una persona, se lee el valor ingresado y si el valor se ingresado es menor que 17 o mayor que 70, el programa ...	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Foro 3 - 1er Bim	Existe un lenguaje denominado JavaScript, determine, en caso de existir ¿Cuál es la relación entre JavaScript y Java?	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7	Foro - Solución Fibonaci	La participación en este foro es OPCIONAL, pero le ayudará en su preparación para el examen presencial.	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Foro 4 - 1er Bim	¿Qué significan las siglas IDE dentro de programación? Liste algunos IDE para desarrollar programas en Java	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8	Solución Números primos	Muestre la solución al problema planteado	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11	Foro 1 - 2do Bim	Consulte si es posible o no incluir programas Java dentro del motor de base de datos Oracle. Si es posible explique brevemente y en sus propios términos lo que es necesario.	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Foro - Solución Histograma de valores	Realice un programa en Java que utilizando arreglos visualice un histograma de valores, para las respuestas a 5 preguntas que se plantearon a 15 personas. Respuestas: 1, 3, 5, 4, 2, 1, 1, 3, 5, 5, 2, 2, 4, 5, 1. La salida del programa sería la siguiente: Pregunta 1: **** Pregunta 2: ****	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12	Foro 2 - 2do Bim	Realice una consulta sobre las bases de datos embebidas y liste algunas base de datos embebidas construidas usando Java	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13	Foro 3 - 2do Bim	En Java existen unas clases que se denominan POJOs consulte cuál es su estructura general y para que sirven.	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14	Foro 4 - 2do Bim	Google ha creado un lenguaje de programación denominada "G" busque información sobre el mismo y trate de determinar si es un lenguaje orientado a objetos	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figura 2. 6 Foros de Aprendizaje curso "Fundamentos de la Programación"

Para obtener un conjunto de datos más amplio acerca de los foros en este curso se ha mapeado los datos encontrados en su tabla principal (prefix_forum) Figura 2.7 con las características de configuración Figura 2.8 que se presentan al simular la creación de un nuevo foro. Estos son:

- ◆ Para este curso solo se han evaluado los foros de consultas.
- ◆ La calificación ha sido realizada únicamente por el docente “id: 5”
- ◆ El tamaño máximo de los archivos adjuntos es de : 2MB
- ◆ No se ha forzado la **subscripción** a foros.
- ◆ El **rastreo** se ha desconectado.
- ◆ El canal **rss** para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- ◆ No se ha bloqueado el uso de los foros.

	id	course	type	name	Intro	assessed	assessedstart	assessedfinish	scale	maxbytes	forcesubscribe	trackingtype	rsstype	rssarticles	timemodified	warnafter	blockafter	blockperiod
<input checked="" type="checkbox"/>	12765	28741	single	Foro 1 - 1er Bim	Consulte y comparta en el foro sobre una herramienta...	5	1287731800	1290833700	2	1	0	1	2	10	1287732243	0	0	0
<input checked="" type="checkbox"/>	12877	28741	single	Solución ecuaciones de segundo grado	Solita la solución al problema propuesto	0	0	0	0	1	0	1	2	10	1287335423	0	0	0
<input checked="" type="checkbox"/>	12976	28741	single	Foro 2 - 1er Bim	Una de las características de Java es que es un...	5	1287949800	1290833700	2	1	0	1	2	10	1287949843	0	0	0
<input checked="" type="checkbox"/>	13073	28741	single	Foro 3 - 1er Bim	Existe un lenguaje denominado JavaScript, ¿determin...	5	1288711500	1290833700	2	1	0	1	2	10	1288711579	0	0	0
<input checked="" type="checkbox"/>	13072	28741	single	Foro - Solución Validación de Datos	<div style="text-align: justify;">Elabore un progr...	0	0	0	0	1	0	1	2	10	1288711246	0	0	0
<input checked="" type="checkbox"/>	13142	28741	single	Foro - Solución Fibonacci	La participación en este foro es OPCIONAL, pero le...	0	0	0	0	1	0	1	2	10	1288224585	0	0	0
<input checked="" type="checkbox"/>	13150	28741	single	Foro 4 - 1er Bim	¿Qué significan las siglas IDE dentro de programac...	5	1288229000	1290833700	2	1	0	1	2	10	1288229237	0	0	0
<input checked="" type="checkbox"/>	13275	28741	single	Solución Números primos	Muestre la solución al problema planteado	0	0	0	0	1	0	1	2	10	1288762011	0	0	0
<input checked="" type="checkbox"/>	13670	28741	single	Foro 1 - 2do Bim	Consulte si es posible o no incluir programas Java...	5	1292812500	1298955300	2	1	0	1	2	10	1294708216	0	0	0
<input checked="" type="checkbox"/>	13971	28741	single	Foro - Solución Histograma de valores	Realice un programa en Java que utilizando...	0	0	0	0	1	0	1	2	10	1292815180	0	0	0
<input checked="" type="checkbox"/>	14116	28741	single	Foro 2 - 2do Bim	Realice una consulta sobre las bases de datos entre...	5	1294065000	1298961100	2	1	0	1	2	10	1294065071	0	0	0
<input checked="" type="checkbox"/>	14291	28741	single	Foro 3 - 2do Bim	En Java existen unas clases que se denominan POJOs...	5	1294688600	1296276900	2	1	0	1	2	10	1294688620	0	0	0
<input checked="" type="checkbox"/>	14411	28741	single	Foro 4 - 2do Bim	Google ha creado un lenguaje de programación llamado...	5	1296222100	1296276900	2	1	0	1	2	10	1296222308	0	0	0

Figura 2. 7 Tabla prefix_forum del curso Fundamentos de la Programación

¿Forzar la suscripción de todos?

¿Leer rastreo de este foro?

Tamaño máximo del archivo adjunto

RSS

Canal RSS para esta actividad

Número de artículos recientes RSS

Calificación

Agregar tipo

Calificación

Restringir las calificaciones a mensajes colocados en este período:

Desde

Hasta

Figura 2. 8 Simulación de Creación de un nuevo foro

En WEKA, la vista general de la asignatura Figura 2.9 muestra la ausencia de discusiones creadas y consecuentemente la lectura de estas, sucede lo mismo con los

archivos adjuntos. Estos atributos no harán diferencia entre un grupo y otro por lo que finalmente se descartan, mas adelante serán usados en la fase de interpretación.

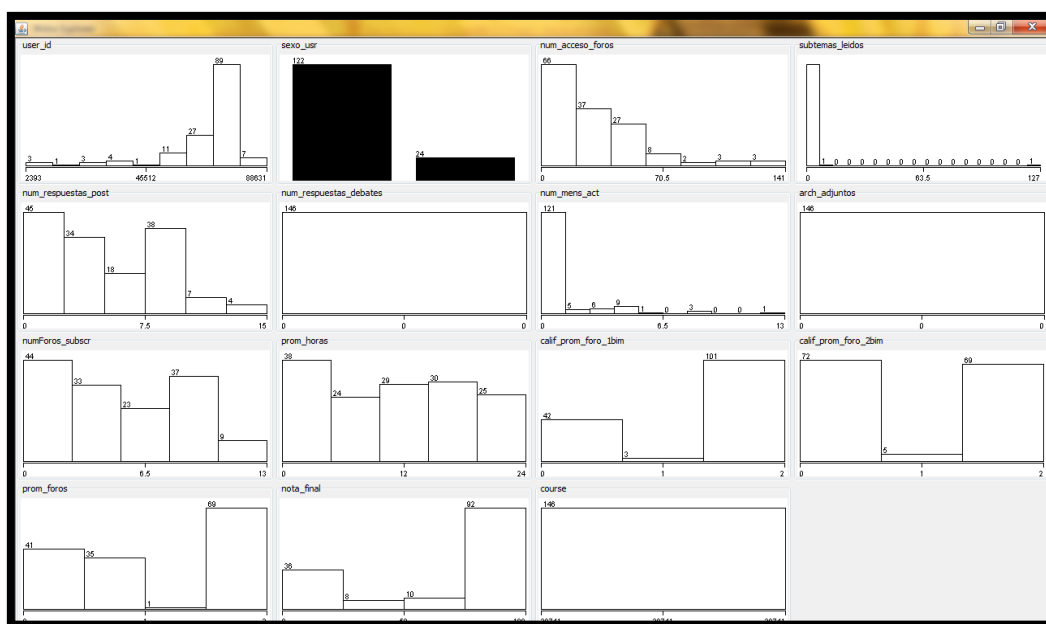


Figura 2. 9 Vista General del curso Fundamentos de la Programación

Previa a la etapa de pruebas se preparan los datos en WEKA. Utilizando los filtros: “Numeric to Nominal” para la conversión del atributo “sexo_usr” siendo los estudiantes del sexo masculino los que representan un 84% el total de los estudiantes de Fundamentos de la Programación, “AddExpression” para determinar la mejora de un bimestre a otro lo que resulta en “mejoraBimestre” y el filtro “Discretize” para “nota_final” debido al rango de calificaciones amplio que posee y tomando en cuenta que un uso constante de este filtro incrementaría la suma de errores cuadráticos (Tabla J.6). Con lo que la vista General de la asignatura con la clase “sexo_usr” sería la expuesta en la Figura 2.10

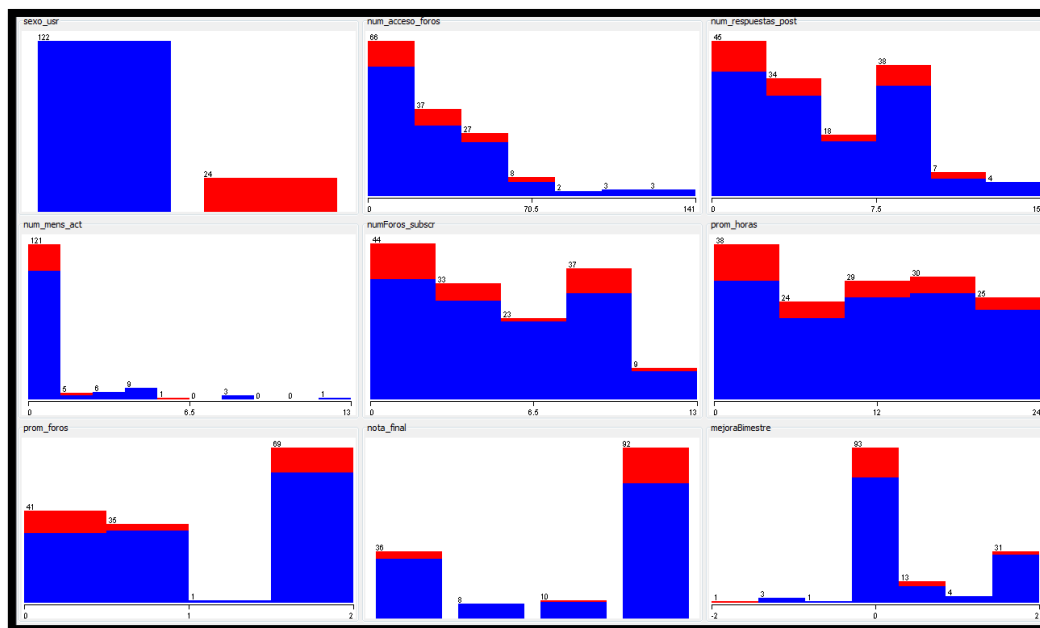


Figura 2. 10 Vista General II del curso Fundamentos de la Programación

Luego de esto es posible guardar un nuevo archivo con los cambios realizados, este tomará el formato “.arff”, que podrá ser usado en cualquier otra herramienta de Data Mining.

2.5.2.1.1.1.2. Algoritmo K-Means

El criterio de comparación de este algoritmo es la distancia. Representa cada uno de los clusters por la media de sus puntos. La configuración utilizada para este curso será la predeterminada en WEKA. Tabla 2.9

Parámetros	Valor
Máximo de Iteraciones	500
Número de Clusters	3
Semilla	10

Tabla 2. 9 Opciones de configuración Algoritmo K-Means en WEKA

Tras la ejecución de K-MEANS los resultados son los de la Tabla 2.10

Número de Iteraciones: 4

Suma de errores cuadráticos: 93.1006598855943

Atributo	Full Data (146)	0	1	2
sexo_usr	M (122) F (24)	M(32) F(10)	M(60) F(10)	M(30) F (4)
num_acceso_foros	30.2397	5.4524	37.6286	45.6471
num_respuestas_post	4.9452	0.7381	6.2714	7.4118
num_mens_act	0.863	0	0.7571	2.1471
numForos_subscr	4.863	0.5	6.3857	7.1176
prom_horas	10.9932	7.4762	12.2429	12.7647
prom_foros	1.1445	0.0952	1.5303	1.6465
nota_final	(75-inf)'	(75-inf)'	(75-inf)'	(-inf-25]'
mejoraBimestre	0.461	0.0319	0.7986	0.2962

Tabla2. 10 Resultados de la ejecución del Algoritmo K-Means

Los grupos se han dividido: 42 para el cluster0, 70 para el cluster1, 34 para el cluster2 lo que se traduce en un 29%, 48% y 23% respectivamente.

La suma de errores cuadráticos disminuyó de 179,561936636949 en la primera experimentación a 93,10065988559433 en relación un 51.84% lo que implica un 48,16% de confiabilidad, menor que el 64,67% del segundo experimento pero considerándose los atributos adicionales que hacen del origen de datos una fuente más completa.

Las características de los grupos son:

Grupo 0: Formado mayormente por hombres, el número de acceso a foros es mínimo, de la misma forma el número de mensajes creados, no han actualizado ninguno de sus mensajes, se han comprometido (suscrito) a un número mínimo de foros, el tiempo que ocupan es bastante corto en relación al resto de grupos lo que se refleja en el promedio de la actividad en ambos bimestres aunque la mejora en estos no sea muy significativa. Al contrario del nivel de participación bajo la “nota final” de estos alumnos es mayor a 75 sobre 100.

Grupo 1: Mayormente constituido por hombres, el número de acceso a los foros es menor al de Grupo 2 pero mayor que el del Grupo 0, sucede lo mismo con el número de mensajes creados, actualizados y suscritos. El número de horas que emplean los alumnos en esta actividad es levemente menor que el del Grupo 2 igual que el promedio en foros; a diferencia del resto de grupos se denota una mejora entre el

primero y segundo bimestre, en cuanto a la nota final se ubica en un intervalo mayor a 75.

Grupo 2: Número superior de estudiantes de sexo masculino, en este grupo se encuentran los estudiantes que mayor acceso tienen a los foros, mayor número de mensajes en los foros, actualizados y a quienes los alumnos se han suscrito mayormente, ocupan un tiempo importante en el cumplimiento de esta actividad, lo que se ve reflejado en su promedio pero no en su nota final, la mejoría en la calificación entre un bimestre y otro se ubica en un punto medio.

La conformación de grupos ha sido compacta, diferenciándose únicamente en la nota final.

Cualitativamente el nivel de colaboración se daría de esta manera:

Grupo 2: Alumnos con nivel de colaboración Alta

Grupo 1: Alumnos con nivel de colaboración Media

Grupo 0: Alumnos con nivel de colaboración Baja

La relación X: sexo_usr, Y: num_respuestas_post , eligiendo como color sex_usr
Figura. 2.11. Esta Figura indica que a pesar de que el número de mujeres es menor que el de hombres su número de respuestas no distan mucho de las realizadas por los hombres.

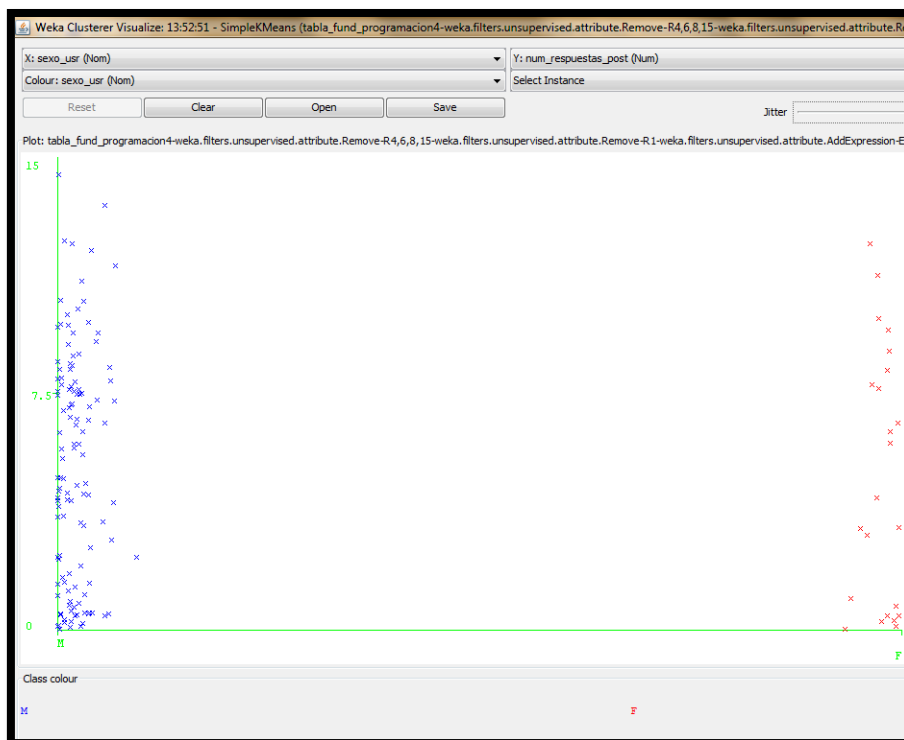


Figura 2. 11 Relación sexo_usr – num_respuestas_post

Otra relación fruto de esta experiencia es la de X: número de acceso y Y: prom_foros, esta vez se ha coloreado por clusters. Figura 2.12

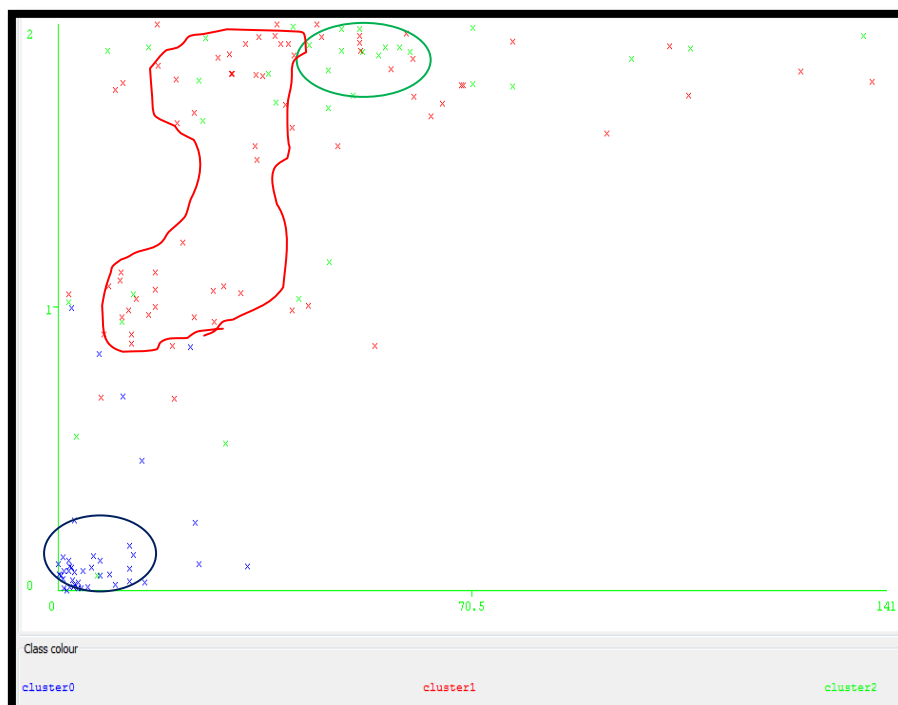


Figura 2. 12 Relación número de acceso - promedio_foros

Los educandos del cluster 2 (color verde) son quienes destacan (ratificándose el análisis previo) de entre el resto de conglomerados con el mayor número de acceso

proporcional a la calificación bimestral de la actividad estudiada; seguidos por el Cluster1 (rojo) y el Cluster 0 (color azul) en ese orden.

Cabe mencionar que en los resultados gráficos el Jitter¹³ alejará las muestras que están físicamente muy próximas, esto tiene utilidad cuando se concentran tanto los puntos que no es posible discernir la cantidad de éstos en un área.

2.5.2.1.1.1.3. Algoritmo Expectation Maximation

EM trabaja mediante un criterio de probabilidad, es capaz de determinar dinámicamente el número de clusters (-1) mediante validación cruzada¹⁴, además de poder personalizarse. Se seguirá utilizando 3 clusters para la conformación de grupos, de tal manera que se tenga un escenario lo más parecido posible al del resto de algoritmos.

La configuración que se ha dado para esta prueba es la precargada en la herramienta WEKA. Tabla 2.11

Parámetro	Descripción	Valor
numClusters	Número de clusters, si es -1 el algoritmo determinará automáticamente el número de clusters.	3
maxIteration	Número máximo de iteraciones del algoritmo si esto no convergió antes.	100
Debug	Muestra información sobre el proceso de clustering.	False
Seed	Muestra información sobre el proceso de clustering.	100
minSrdDev	Desviación típica mínima admisible en las distribuciones de densidad.	1e-6

Tabla 2. 11 Opciones de configuración Algoritmo EM

Los resultados tras la aplicación del algoritmo EM se muestran en la Figura 2.13.

¹³ Jitter: Ruido aleatorio en las muestras. (García D. , 2005)

¹⁴ Validación Cruzada: Es la práctica estadística de partir una muestra de datos en subconjuntos de tal modo que el análisis es inicialmente realizado en uno de ellos, mientras los otros subconjuntos son retenidos para su uso posterior en la confirmación y validación del análisis inicial. (Wikipedia, 2011)

```

EM
==
Number of clusters: 3

Attribute          Cluster
                   0      1      2
                   (0.1) (0.41) (0.49)
-----
sexo_usr
M                  14.8121 51.4375 58.7504
F                   1.0588 10.9304 15.0108
[total]            15.8709 62.3679 73.7612
num_acceso_foros
mean               44.3166 51.3864  9.7295
std. dev.          25.7602 28.0124  9.1772

num_respuestas_post
mean                6.2966  8.0419  2.079
std. dev.           2.9304  2.731  2.5441

num_mens_act
mean                4.252  1.1102  0
std. dev.           3.8555  1.5796  1.9883

numForos_subscr
mean                5.6696  8.2973  1.8181
std. dev.           2.9232  1.6742  2.1482

prom_foros
mean                1.3009  1.901  0.478
std. dev.           0.4552  0.103  0.5533

nota_final
*(-inf-25]'         9.0077 15.9552 14.0371
*(25-50]'           1.5969  3.2595  6.1436
*(50-75]'           1.004  5.996  6
*(75-inf]'          6.2623 39.1573 49.5804
[total]            17.8709 64.3679 75.7612

mejoraBimestre
mean                1.0682  0.0844  0.6605
std. dev.           1.0218  0.199  1.0046
  
```

Figura 2. 13 Resultados Ejecución Algoritmo EM

A diferencia del algoritmo K-Means, Expectation Maximation arroja resultados para los campos del atributo nominal "sexo_usr".

Los grupos se han formado de la siguiente manera:

Grupo 0: 13 (9%)

Grupo 1: 62 (42%)

Grupo 2: 71 (49%)

Su registro de verisimilitud es de: -16.96638 mayor que la experimentación previa - 21.3299.

El Cluster 2 denota un perfil con un mayor nivel colaborativo, le sigue el Cluster 1, finalmente el Cluster 0; tal cual sucedió con el K-Means.

El algoritmo EM se le conoce como K-Means Probabilístico, por cuanto se corresponde con los resultados iniciales de esta experimentación.

2.5.2.1.1.4. Algoritmo Cluster Jerárquico

WEKA si bien cuenta con “HierarchicalClusterer” dentro de sus algoritmos de Clustering, no proporciona un soporte efectivo visual del dendograma¹⁵ resultante, por lo que se utilizará una herramienta adicional para la experimentación. Se ha seleccionado a KNIME para el efecto.

KNIME cuenta con extensiones de otras poderosas herramientas de Datamining como SPSS , R y el mismo WEKA, su interfaz y creación de escenarios es similar a Clementine SPSS.

Se parte del origen de datos “.arff” que se creó en la etapa de pre-procesamiento (Sección 2.5.2.1.1), se prueba con el 33% de los datos, lo que sigue es la conexión al algoritmo, finalmente los resultados se escriben a modo de archivo .csv Figura 2.14

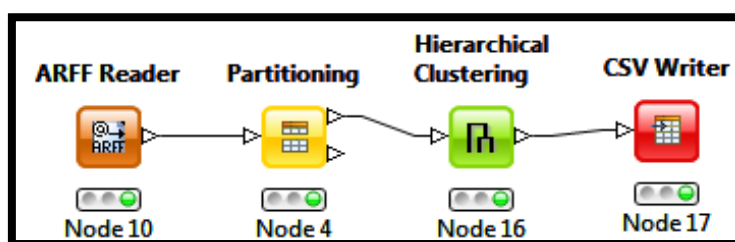


Figura 2. 14 Esquema Algoritmo de Clustering Jerárquico

Los resultados que muestra este procedimiento indican cómo se han dividido los conglomerados. Figura 2.15

¹⁵ Dendograma: Es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. (Vicente, 2006)

Figura 2. 15 Resultados ejecución Algoritmo Cluster Jerárquico

Este algoritmo es especialmente útil si se requiere conocer el valor óptimo por el cual se debería agrupar los elementos, más a priori se ha definido a 3 el número de clusters Figura 2.16 , distribuyéndose así:

Cluster 0: 2, Cluster 1: 44, Cluster 2: 2

Si tenemos en cuenta que esta información solo representa el 33% de los datos reales, los mismos tendrían una representación de: 4,17% - 91,67% - 4,17% respectivamente.

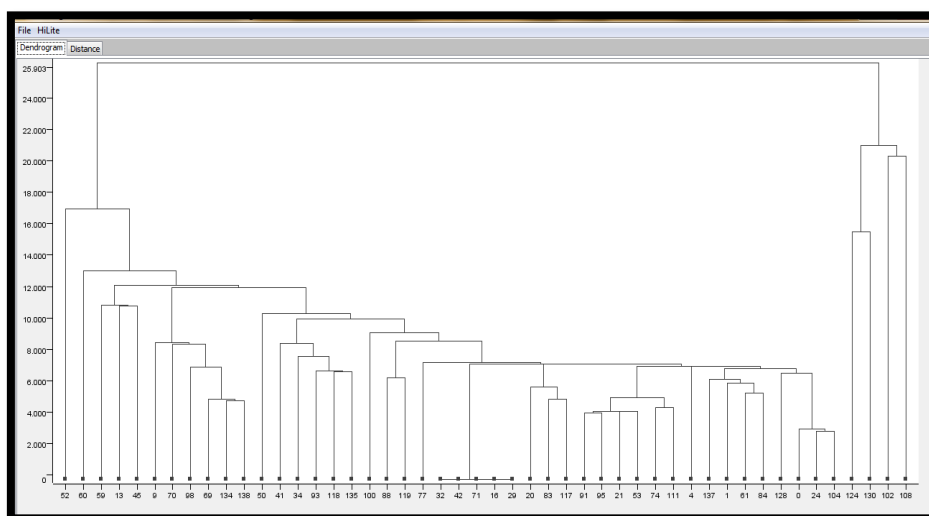


Figura 2. 16 Dendrograma - Algoritmo de Clustering Jerárquico

2.5.2.1.1.2. Lógica de la Programación

Lógica de la Programación [B] posee foros tanto de tipo Preguntas y Respuestas (3) como de discusiones (2), 2 de los primeros son ejercicios prácticos el restante de consultas; de las discusiones una es de consulta, la otra práctica.

Todos los mensajes han sido leídos (al menos una vez). Figura 2.17

Foros de aprendizaje							
Sección	Foro	Descripción	Temas	Mensajes no leídos	Rastrear	Suscrito	RSS
2	Foro 1: Primer Bimestre	Para participar en este foro, es preciso haber estudiado el capítulo de la guía didáctica, consiste en responder a la pregunta siguiente ¿Qué principios de la lógica sustentan el desarrollo de los programas de computadora? y como consecuencia de la respuesta anterior responder a la ...	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	foro practico	Para el problema siguiente realice lo siguiente: 1. Identifica salidas, procesos y entradas. 2. representelos en formato de enunciados consider los criterios y tipos del apartado Principios en la elaboración de enunciados del texto básico. Problema: Se desea desarrollar un programa que calcule ...	1	0	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
6	Foro de soluciones a los ejercicios de lógica	Resolver y colocar los resultados de los ejercicios indicados en el documento de esta semana. No es necesario resolver todos, la idea es ir aportando de a poco las soluciones en el foro, de modo que se genere un espacio de discusión con las posibles soluciones.	9	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11	Dudas generales para la evaluación presencial	Estimados: En este foro sírvanse colocar sus inquietudes generales en relación a los contenidos de la asignatura de Lógica de la programación, estas preguntas serán respondidas de manera inmediata.	1	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13	Foro temático segundo bimestre	Las miniespecificaciones y los diagramas de flujo son técnicas especializadas de representación de algoritmos, cada una de ellas con su características que las hacen útiles en diferentes situaciones. Para responder a este foro resuelva los siguiente s ejercicios tanto en miniespecificación ...	23	0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figura 2. 17 Foros de Aprendizaje curso "Lógica de la Programación"

Utilizando el recurso de la tabla prefix_forum Figura 2.18 integrándolo con la información del curso integrado en el EVA, se ha recopilado la siguiente información:

- ◆ Para este curso no existió calificación en los foros, según lo observado en el libro de calificaciones.
- ◆ El tamaño máximo permitido en archivos adjuntos es de : 5MB
- ◆ No se ha forzado la **subscripción** a foros.
- ◆ El **rastreo** se ha desconectado.
- ◆ El canal **rss** para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- ◆ No se ha bloqueado el uso de los foros.

A diferencia de la Minería del curso Fundamentos de la Programación, Lógica de la Programación si registra actividad para los atributos: subtemas_leidos, num_respuestas_debates y arch_adjuntos.

Se muestran los resultados en la Tabla 2.12

Número de iteraciones: 5
Suma de errores cuadráticos: 77.402302380404

Atributo	Full Data (154)	0	1	2
sexo_usr	M (143) F (11)	M (39) F (2)	M (54) F (6)	M (50) F(3)
num_acceso_foros	12.6753	3.5366	23.5167	7.4717
subtemas_leidos	5.8506	0.6829	12.3667	2.4717
num_respuestas_post	0.5649	0.0488	1	0.4717
num_respuestas_debates	0.2078	0.0488	0.4333	0.0755
num_mens_act	0.2338	0.0488	0.4667	0.1132
arch_adjuntos	0.0844	0	0.2167	0
numForos_subscr	0.7532	0.1707	1.3833	0.4906
prom_horas	11.2078	1.122	12.5	17.5472
nota_final	(-inf-25]'	(-inf-25]'	(75-inf)'	(-inf-25]'

Tabla 2. 12 Resultados ejecución del algoritmo K-Means

La suma de errores cuadráticos bajó de 148,7633869348921 a 34,6864615892839 de la primera a la segunda experimentación, en la actual llega a 77,402302380404 debido al cambio de escenario al realizar la agregación de atributos y la aplicación de filtros.

Los grupos se han dividido:

Cluster 0: 41, **Cluster 1:** 60, **Cluster 2:** 53 representando un 27%, 39% y 34% respectivamente.

Las características de este grupo son:

Grupo 0: Mayormente formado por hombres, menor número de acceso a foros, número menor de subtemas leídos así como creados, menor número de mensajes creados y actualizados, no existen archivos adjuntos, bajo número de foros suscritos, el promedio en horas es ínfimo.

Grupo 1: Formado en su mayoría por hombres, cuentan con mayor acceso a foros, es superior en cuanto a discusiones creadas y leídas, los mensajes que se han creado, actualizado y en los que se ha adjuntado un archivo es considerable respecto a los otros grupos, la mayoría de los estudiantes que se han suscrito a foros pertenecen a este grupo, aún su nivel de colaboración alto el tiempo que tardan en realizar actividades concernientes a Foros es menor que el Grupo 2 y mayor que el Grupo 1.

Grupo 2: Como en los otros grupos predominan los hombres, el número de veces que los alumnos acceden a foros es menor que el Grupo1 y mayor que el Grupo 0, situación que se repite en las discusiones leídas y en el número de mensajes creados, aunque la creación de debates se ubica en un punto medio así como los que actualizan, los estudiantes no han subido archivos como complemento a su participación pero si se han suscrito a la mayoría de foros y han ocupado un tiempo importante en el cumplimiento de esta actividad.

Cualitativamente los grupos se denominaran:

Grupo 1: Alumnos con nivel de colaboración Alta.

Grupo 2: Alumnos con nivel de colaboración Media.

Grupo 0: Alumnos con nivel de colaboración Baja.

Aunque en primera instancia no se pudo visualizar el número de mujeres pertenecientes a cada cluster, mediante la relación X: sexo_usr - Y: Cluster, eligiendo como color sex_usr Figura. 2.20 se obtuvo que: El cluster 0 está conformado por 2 personas del sexo femenino en el Cluster 1 por 6 y 3 en el Cluster 2.

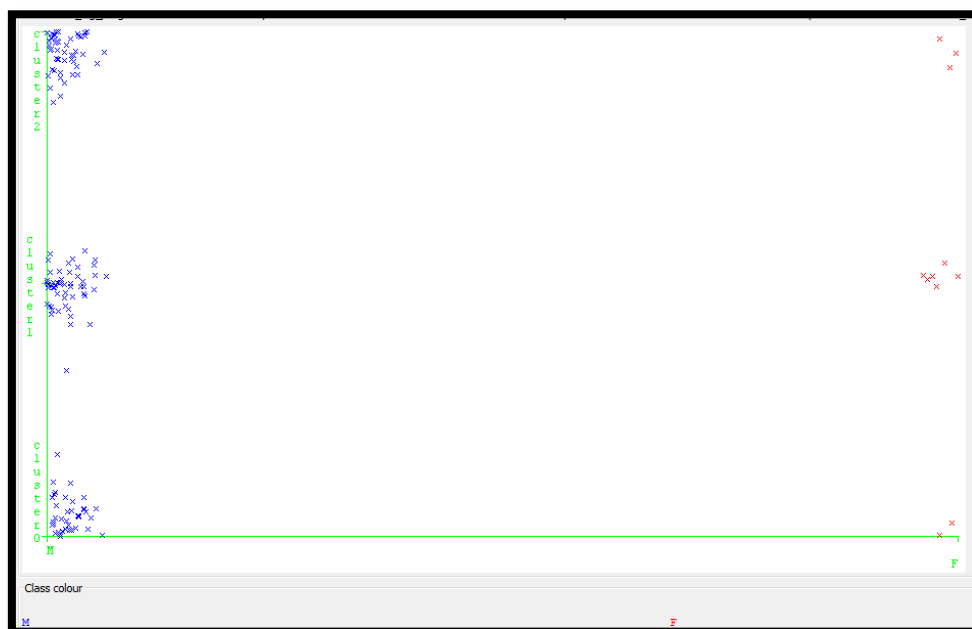


Figura2. 20 Relación sexo_usr – Cluster

Otra relación X: num_acceso_foros Y: num_respuestas_post indica la poca afluencia a nivel general de los estudiantes en los foros, lo que se ve reflejado en una

colaboración mínima. Los educandos del Cluster 1 reflejan una mayor predisposición para participar en este tipo de actividad. Figura 2.21

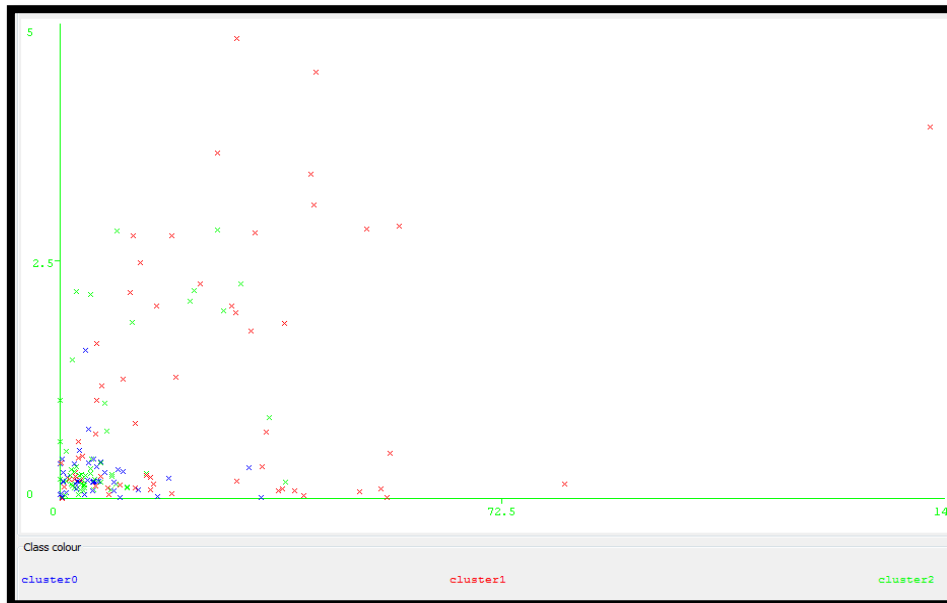


Figura 2. 21 Relación Número de acceso a foros- Número de posts

2.5.2.1.1.2.2. Algoritmo Expectation Maximation

Se realizan las pruebas del algoritmo con la configuración precargada de la herramienta de aprendizaje automático con la que se ha venido trabajando. Los resultados para Lógica de la Programación se muestran en la Figura 2.22

Attribute	Cluster		
	0 (0.31)	1 (0.05)	2 (0.64)
sexo_usr			
M	45.325	7.6014	93.0736
F	4.4247	2.5753	7
[total]	49.7498	10.1766	100.0736
num_acceso_foros			
mean	24.1139	60.4576	3.1224
std. dev.	12.8808	34.8486	4.3286
subtemas_leidos			
mean	8.9196	50.3443	0.6469
std. dev.	9.0185	50.1406	1.2061
num_respuestas_post			
mean	1.5993	0.6888	0.051
std. dev.	1.2531	1.4421	0.22
num_respuestas_debates			
mean	0.4797	1.1124	0
std. dev.	0.6797	0.5888	0.5071
num_mens_act			
mean	0.5199	1.3668	0
std. dev.	0.8699	1.4708	0.6935
arch_adjuntos			
mean	0.0105	1.5285	0
std. dev.	0.1023	1.54	0.4977
numForos_subscr			
mean	1.7288	2.3612	0.1442
std. dev.	0.9764	0.4818	0.407
prom_horas			
mean	12.881	10.6176	10.4423
std. dev.	6.9272	6.5444	8.6957
nota_final			
'(-inf-25]'	13.9504	1.0496	65
'(25-50]'	3.9901	2.0099	6
'(50-75]'	8.4959	4.5041	9.9999
'(75-inf)'	25.3134	4.613	21.0736
[total]	51.7498	12.1766	102.0736

Figura 2. 22 Resultados del Algoritmo EM

Los grupos se han dividido así:

Grupo 0: 54 (35%)

Grupo 1: 9 (6%)

Grupo 2: 91 (59%)

Con un registro de verisimilitud: -13.55598

Luego de revisar la probabilidad por conglomerado y atributo, se denominará a los grupos de la siguiente manera:

Grupo 2: Alumnos con nivel de colaboración Alta

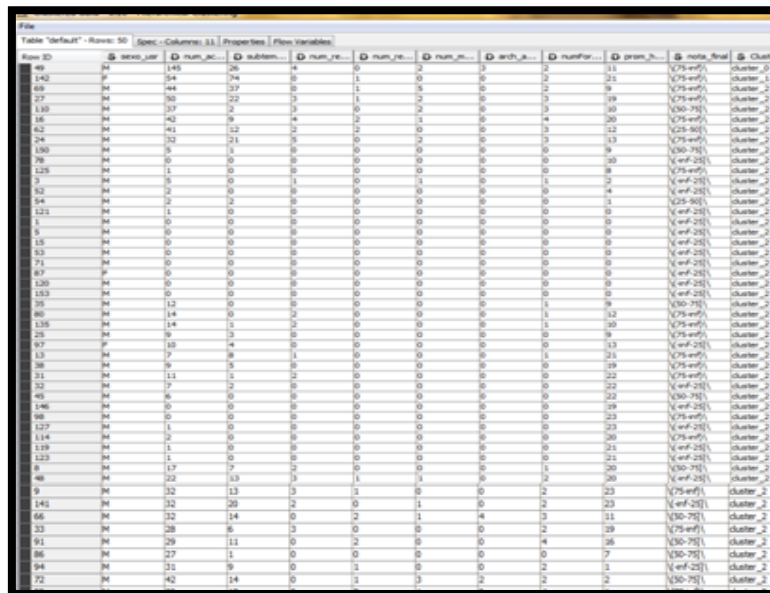
Grupo 1: Alumnos con nivel de colaboración Media

Grupo 0: Alumnos con nivel de colaboración Baja.

2.5.2.1.1.2.3. Algoritmo Cluster Jerárquico

Como ya se explicó en la asignatura “Fundamentos de la Programación” se trabajará con la herramienta KNIME para la aplicación de este algoritmo.

La salida tras su ejecución se muestra en la Figura 2.23



ID	Nombre	Cluster
39	M	cluster_0
142	F	cluster_1
69	M	cluster_2
217	M	cluster_2
135	M	cluster_2
56	M	cluster_2
62	M	cluster_2
24	M	cluster_2
180	M	cluster_2
78	M	cluster_2
125	M	cluster_2
7	M	cluster_2
52	M	cluster_2
54	M	cluster_2
121	M	cluster_2
1	M	cluster_2
15	M	cluster_2
53	M	cluster_2
71	M	cluster_2
87	F	cluster_2
120	M	cluster_2
153	M	cluster_2
35	M	cluster_2
80	M	cluster_2
135	M	cluster_2
215	M	cluster_2
107	F	cluster_2
13	M	cluster_2
38	M	cluster_2
11	M	cluster_2
52	M	cluster_2
45	M	cluster_2
146	M	cluster_2
98	M	cluster_2
127	M	cluster_2
114	M	cluster_2
119	M	cluster_2
133	M	cluster_2
8	M	cluster_2
48	M	cluster_2
9	M	cluster_2
143	M	cluster_2
66	M	cluster_2
33	M	cluster_2
91	M	cluster_2
86	M	cluster_2
94	M	cluster_2
72	M	cluster_2

Figura 2. 23 Resultados ejecución Algoritmo Cluster Jerárquico

La distribución de los alumnos en función del 33% de los datos destinados para la prueba es:

Grupo 0: 1 alumno

Grupo 1: 1 alumno

Grupo 2: 48 alumnos

El dendograma resultante es el que se muestra en la Figura 2.25. Si se hace un corte sobre el mismo se aprecia que a pesar de las numerosas gráficas se ha logrado agrupar estos elementos en 3 conglomerados.

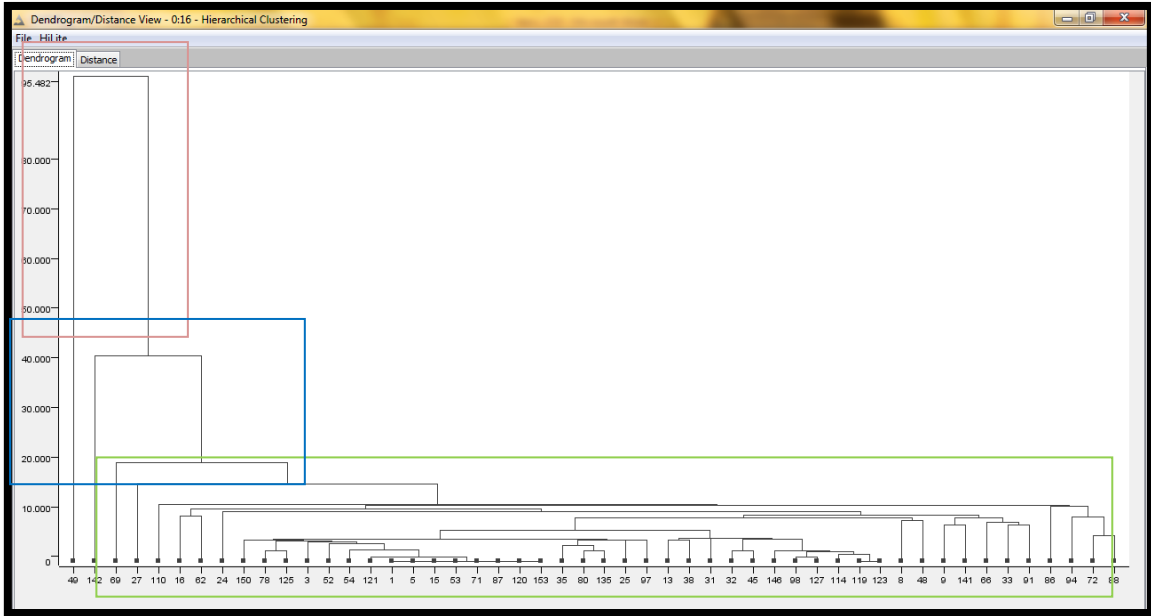


Figura 2. 24 Dendograma –Algoritmo de Clustering Jerárquico

2.5.2.1.1.3. Fundamentos Informáticos

Fundamentos Informáticos cuenta con 3 secciones, una de ellas contiene dos foros anidados, en total posee 4 foros, 1 de ellos es de tipo discusión el resto utiliza el formato Preguntas y Respuestas. Figura 2.25





Foros de aprendizaje							
Sección	Foro	Descripción	Temas	Mensajes no leídos	Rastrear	Suscrito	RSS
7	Foros Fundamentos Informáticos	En este espacio, los estudiantes de la asignatura de Fundamentos Informáticos debatirán sobre los temas planteados. Seguramente esta actividad será muy fructífera. A continuación, haga click en el foro que le corresponda, más abajo en las temáticas creadas por Ruth Reátegui, quien figura ...	7	0	<input type="checkbox"/>	<input type="checkbox"/>	
17	Preguntas 2do bimestre	Por favor indique aquí cuáles son sus dudas respecto a lo que ha estudiado de los contenidos propuestos para el segundo bimestre.	1	0	<input type="checkbox"/>	<input type="checkbox"/>	
19	Foro 3: Ética Informática	Para esta actividad es necesario que realice la lectura "Ética informática", que se encuentra en el apartado métodos prácticos del texto base, para luego emitir opinión respecto a lo siguiente: <ul style="list-style-type: none"> ¿Cree que las reglas citadas de ética se practican en nuestro medio? ¿Hacen falta ... 	1	0	<input type="checkbox"/>	<input type="checkbox"/>	
	Foro 4: Importancia de la IA	<ul style="list-style-type: none"> ¿Cuáles son las ventajas que brinda la IA? ¿Cree usted que en el futuro la IA podrá sustituir a los trabajadores de una empresa u organización? 	1	0	<input type="checkbox"/>	<input type="checkbox"/>	

Figura 2. 25 Foros de Aprendizaje curso "Fundamentos de la Programación"

Se ha recurrido a la tabla prefix_forum Figura 2.26, para una visión más amplia de las características activadas para este curso.

Estas son:

	id	course	type	name	intro	assessed	assesstimestart	assesstimefinish	scale	maxbytes	forcesubscribe	trackingtype	rsstype	rssarticles	timemodified	warnafter	blockafter	blockperiod
<input type="checkbox"/>	13038	28739	general	Foros Fundamentos Informáticos	<>>En este espacio, los estudiantes de la asigna...	0	0	0	0	512000	0	1	2	10	1289407050	0	0	0
<input type="checkbox"/>	13905	28739	single	Preguntas 2do bimestre	Por favor indique aquí cuáles son sus dudas respe...	0	0	0	0	512000	0	1	2	10	1292337133	0	0	0
<input type="checkbox"/>	14089	28739	single	Foro 3: Ética Informática	Para esta actividad es necesario que realice la l...	0	0	0	0	512000	0	1	2	10	1293803391	0	0	0
<input type="checkbox"/>	14092	28739	single	Foro 4: Importancia de la IA	<u><u>Cuales son las ventajas que brinda ...	0	0	0	0	512000	0	1	2	10	1293803348	0	0	0

Figura 2. 26 Tabla prefix_forum del curso Fundamentos Informáticos

- ◆ Para este curso no se ha realizado evaluación alguna de foros.
- ◆ El tamaño máximo de los archivos adjuntos es de : 5MB
- ◆ No se ha forzado la **subscripción** a foros.
- ◆ El **rastreo** se ha desconectado.
- ◆ El canal **rss** para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- ◆ El uso de foros no está bloqueado.

Los filtros que se han aplicado son: “Discretize” dividido en 4 rangos de igual dimensión y “Numeric to Nominal”, este último ha permitido conocer que los estudiantes de este curso en un 80% son hombres.

La vista general de la asignatura Figura 2.27 refleja una participación minúscula.

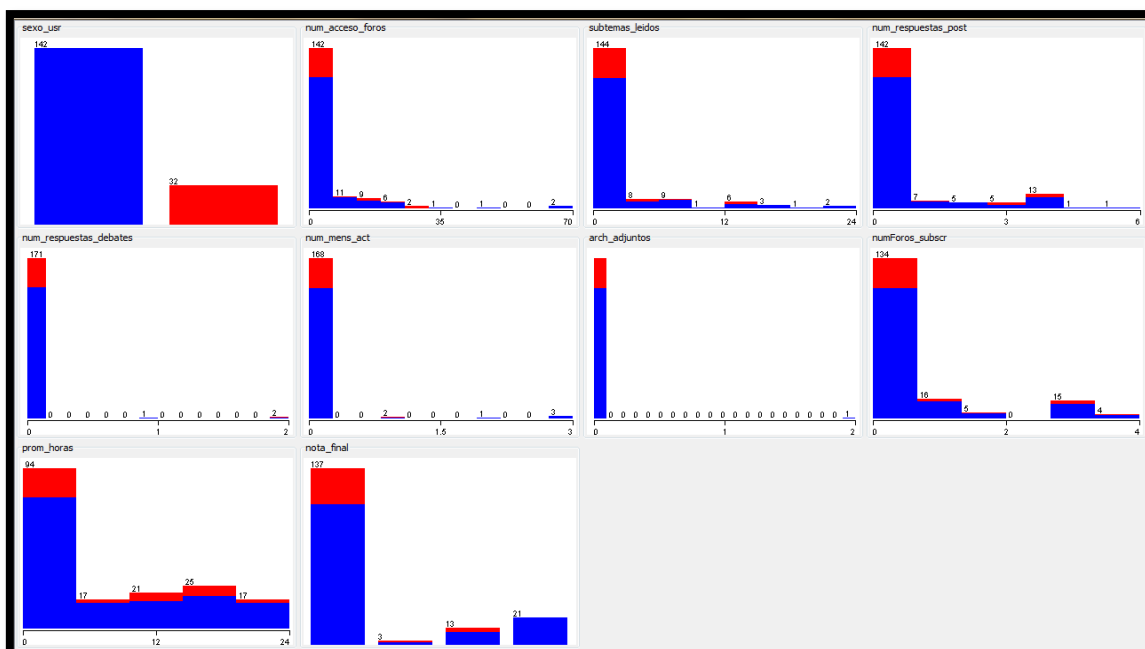


Figura 2. 27 Vista General del curso Fundamentos Informáticos

2.5.2.1.1.3.1. Algoritmo K-Means

Se utilizan los datos precargados en WEKA que consisten en:

- ◆ Máximo de Iteraciones: 500
- ◆ Número de Clústers: 3
- ◆ Semilla: 10

Luego de la ejecución del algoritmo se obtuvieron los resultados mostrados en la Tabla 2.13

Número de iteraciones: 5
Suma de errores cuadráticos: 87.64590519277633

Atributo	Full Data (174)	0	1	2
sexo_usr	M (142) F (32)	M(49) F(9)	M(74) F(18)	M(19) F (5)
num_acceso_foros	4.5287	3.5172	0.2065	23.5417
subtemas_leidos	2.0632	1.6552	0.1087	10.5417
num_respuestas_post	0.546	0.2069	0.0109	3.4167
num_respuestas_debates	0.0287	0	0	0.2083
num_mens_act	0.0747	0	0	0.5417
arch_adjuntos	0.0115	0	0	0.0833
numForos_subscr	0.5	0.2931	0.0217	2.8333
prom_horas	6.8621	15.7759	0.5	9.7083
nota_final	(-inf-25]'	(-inf-25]'	(-inf-25]'	(-inf-25]'

Tabla 2. 13 Resultado del Algoritmo K-Means

Los estudiantes han sido divididos en tres grupos: el Cluster 0 cuenta con 58 estudiantes, el Cluster 1 con 92 y el Cluster 2 con 24, lo que equivale a un 33%, 53% y 14% respectivamente de 174 alumnos.

De 103,78037289349791 en la suma de errores cuadráticos de la primera experimentación se paso a 22,661696453425375 en la segunda a 87,64590519277633 en esta última.

Es notable la diferencia con la prueba inicial, mientras que con la siguiente si bien es menor el error cuadrático hay que tomar en cuenta el incremento de las variables y el uso de filtros que pudieran haber generado tal dilatación. Calificación final es menor a 25.

Las características de estos grupos son:

Grupo 0: Formada mayormente por hombres, número de acceso a foros bastante menor que el Grupo 2 pero ligeramente mayor que el Grupo 1, el número de discusiones leídas es mayor que el Grupo 1 y bastante menor con respecto al Grupo 2, lo que se repite para el número de mensajes creados y al de foros suscritos, tanto la participación en debates, mensajes actualizados y archivos adjuntos es Nula, el tiempo utilizado en este grupo es mayor que el del resto.

Grupo 1: Compuesto en su mayoría por hombres, Bajo nivel colaborativo tanto en el número de acceso en foros, así como en las discusiones leídas , el número de mensajes creados, los foros suscritos y el promedio de horas en las que se realiza la actividad de foros. Las discusiones creadas, los mensajes actualizados y los archivos adjuntos tienen un estado nulo la nota final es menor a 25 sobre 100.

Grupo 2: Usuarios masculinos en una proporción superior, los estudiantes muestran de este conglomerado muestran un mayor número de acceso a los foros, son los únicos que crean y revisan discusiones, los estudiantes se han suscrito a un mayor número de foros y han participado efectivamente en la creación de estos. A diferencia de los otros grupos estos alumnos si han adjuntado archivos a sus respuestas pero no han dedicado mayor tiempo que el Grupo 0 en esta actividad. La calificación se muestra en un rango menor a 25 sobre 100.

Cualitativamente los grupos se expresan así:

Grupo 2: Alumnos con nivel de interacción Alta

Grupo 0: Alumnos con nivel de interacción Media

Grupo 1: Alumnos con nivel de interacción Baja

La relación X: sexo_usr, Y: Cluster- Figura 2.28 eligiendo como color Cluster Figura. 2.30 muestra que la distribución de mujeres si bien es mínima, en el cluster 1 se encuentra la mayor parte con 20 educandos seguida por el cluster 0 con 9 y el cluster 2, con 5 alumnos.

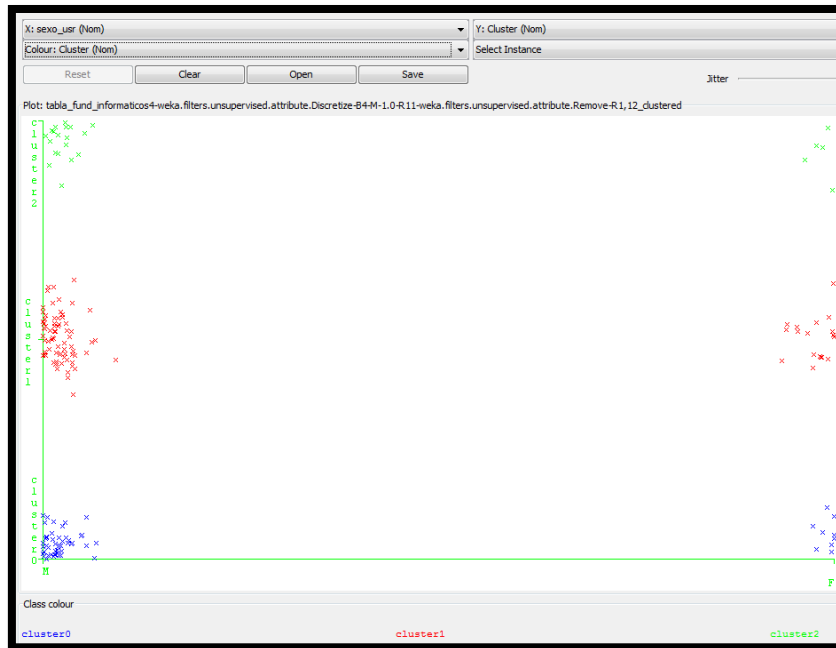


Figura 2. 28 Relación sexo_usr - Cluster

Otra relación a exponer es la de X: num_acceso_foros, Y: num_respuestas_post coloreado con Cluster. Figura 2.29

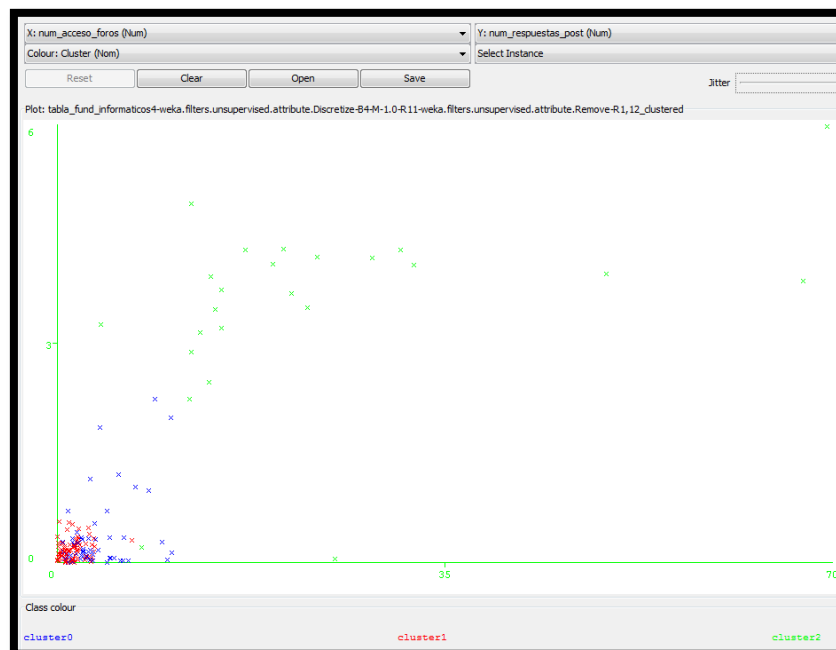


Figura 2. 29 Relación num_acceso_foros - num_respuestas_post

El número de acceso de forma general es limitado, no así el número de mensajes creados, que el cluster 2 ha superado por mucho.

2.5.2.1.1.3.2. Algoritmo Expectation Maximation

Se hará uso de la configuración predeterminada en WEKA para EM, la que consiste en un número de cluster de 3, máximo número de iteraciones 100, semilla 100 y desviación 1e-6.

La salida de este algoritmo se muestra en la Figura 2.30

Attribute	Cluster		
	0 (0)	1 (0.11)	2 (0.89)
sexo_usr			
M	1.3552	16.5088	127.136
F	1.0003	4.0001	29.9996
[total]	2.3554	20.5089	157.1356
num_acceso_foros			
mean	0.7246	26.9314	1.8646
std. dev.	1.4787	17.2458	3.3105
subtemas_leidos			
mean	0.2501	13.9213	0.6526
std. dev.	0.9575	5.1443	1.6171
num_respuestas_post			
mean	0.0015	3.5946	0.1835
std. dev.	0.0383	1.393	0.6255
num_respuestas_debates			
mean	0	0.2701	0
std. dev.	0.2263	0.6429	0.2263
num_mens_act			
mean	0	0.7024	0
std. dev.	0.4303	1.1356	0.4303
arch_adjuntos			
mean	0	0.1081	0
std. dev.	0.1516	0.4521	0.1516
numForos_subscr			
mean	0.0081	2.7362	0.2343
std. dev.	0.0896	1.0221	0.6649
prom_horas			
mean	4.6677	9.8544	6.5101
std. dev.	7.2934	7.4553	7.9619
nota_final			
(-inf-25]	1.2702	11.8767	126.853
(25-50]	1.0025	1	3.9975
(50-75]	1.0112	2.6189	12.3699
(75-inf)	1.0715	7.0132	15.9152
[total]	4.3554	22.5089	159.1356

Figura 2. 30 Resultados de Ejecución Algoritmo EM

Los grupos se han formado así:

- Grupo 0:** 29 (17%)
- Grupo 1:** 27 (16%)
- Grupo 2 :** 118 (68%)

Con un registro de verisimilitud de : -10.19145

La revisión de las probabilidades obtenidas nos muestra que cualitativamente los resultados estarían dados de esta forma:

Cluster 2: Alumnos con nivel de colaboración Alto

Cluster 1: Alumnos con nivel de colaboración Medio

Cluster 0: Alumnos con nivel de colaboración Bajo

2.5.2.1.1.3.3. Algoritmo Cluster Jerárquico

Con KNIME se carga el archivo “.arff” que se obtuvo del preprocesamiento con WEKA, para posteriormente actualizarse en el esquema inicialmente planteado. Figura 2.31

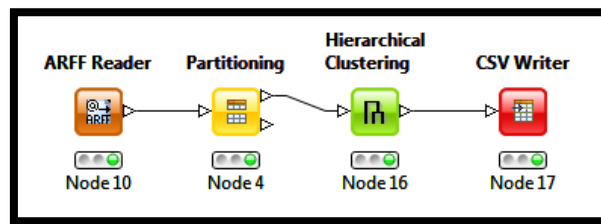


Figura 2. 31 Esquema de Aplicación Clustering Jerárquico

El 33% de los datos utilizados para esta prueba se han distribuido así:

Cluster 0: 1

Cluster 1: 1

Cluster 2: 55

El dendograma resultante Figura 2.32 si se corta a un nivel de distancia dado, se obtiene una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

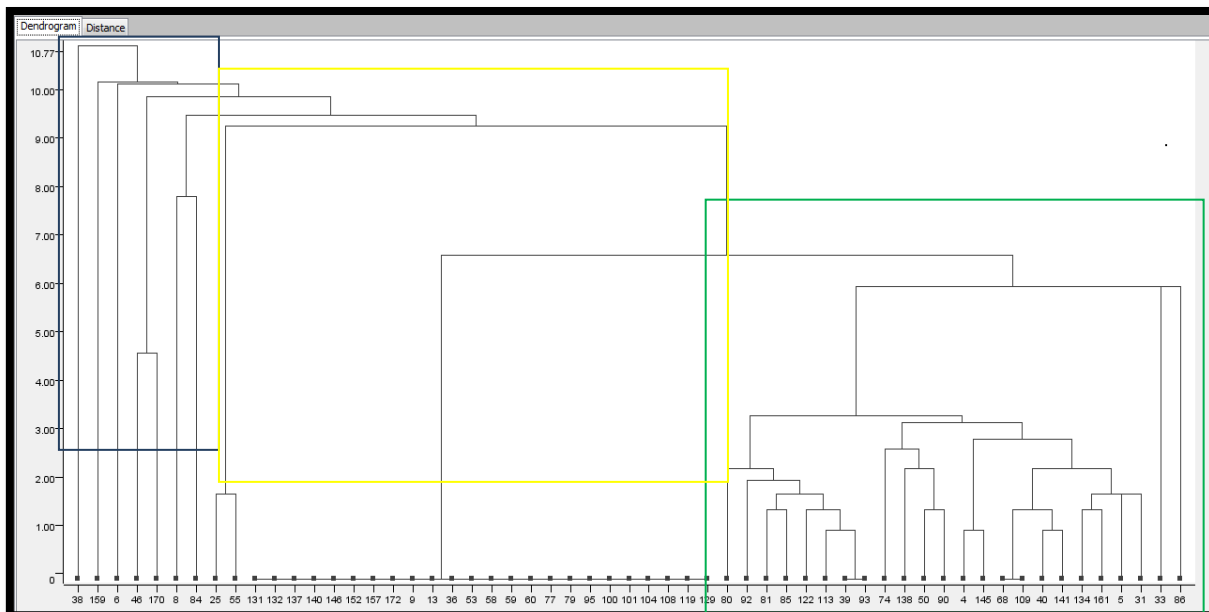


Figura 2. 32 Dendrograma - Algoritmo Clustering Jerárquico

La ejecución de los Algoritmos K-Means, EM y de Clustering Jerárquico han permitido la consolidación de resultados, es decir tanto el algoritmo K-Means como el EM poseen un grado de similitud y correspondencia entre la distancia y la probabilidad de pertenecer o no a un conglomerado, mientras que Hierarchical Clustering la distribución de estos en un espacio separado por niveles de similitud/disimilitud edido en distancias.

Tras la realización de estas experimentaciones se concluye que:

- ◆ Los algoritmos de particionado K-Means y EM funcionan mejor con datos numéricos.
- ◆ Clustering Jerárquico funciona mejor con una cantidad pequeña de datos
- ◆ En K-Means la suma de errores cuadráticos disminuye al incrementar la semilla, debido a la dependencia del arranque cuando no se establece una solución que mejor divide el conjunto de instancias.
- ◆ Obtener una menor suma de errores cuadráticos certifica que los estudiantes han sido agrupados de una forma óptima, siendo la variación mínima entre estos.
- ◆ Se debe experimentar cambiando los parámetros de configuración de la herramienta para encontrar una solución que se adapte a los objetivos planteados.
- ◆ Aunque existe una gran cantidad de alumnos, generalmente mayor a 100 son pocos los que realmente participan.

2.6. EVALUACIÓN E INTERPRETACIÓN

En esta fase se analizarán los resultados de cada una de las experimentaciones seleccionadas y se evaluará las posibles soluciones y recomendaciones para todos los casos.

2.6.1. Fundamentos de la Programación

Para el uso de foros en el curso Fundamentos de la Programación el docente ha activado solamente las características básicas de los mismos, entiéndase con esto el no bloqueo, la suscripción no obligatoria, el canal rss y la capacidad de responder un mensaje. Aunque no exista restricción de archivos adjuntos son inexistentes, de igual forma solo se ha hecho uso de foros tipo Preguntas y Respuestas, no se han planteado discusiones.

El rastreo de mensajes es un punto fuerte para la determinación de comportamientos colaborativos mas no fue activado, esta información fue recogida desde el *log* del sistema.

Aunque las aportaciones en foros de los estudiantes fueron calificadas, esa potestad fue exclusiva del facilitador, no se habilitó la cooperación de los estudiantes para la asignación de notas entre sus compañeros.

Los conglomerados se han calificado tomando en cuenta siete aspectos

- Sexo
- Colaboración directa (num_respuestas_post, num_mens_act, numForos_subscr).
- Colaboración indirecta (num_acceso_foros, subtemas_leidos)
- Tiempo empleado en toda acción que implique Foros.
- Calificación promedio en Foros
- Mejora de Calificación del Primer al Segundo Bimestre.
- Nota final

En este curso las participaciones en el foro fueron indirectamente proporcionales a las calificaciones finales, los alumnos del primero grupo (0) no han ingresado de forma recurrente, pocos son los mensajes que se han agregado y actualizado. El tiempo que han utilizado es mínimo en comparación con el resto de clusters, lo que podría

reflejarse en un bajo promedio de actividad, diferente a su nota final superior (> 75 sobre 100).

Los estudiantes del segundo grupo (1) que poseen un nivel medio de colaboración tienden a tener una calificación final alta (>75) en contraste con los que poseen mayor nivel de colaboración, registran un tiempo ligeramente menor que el del tercer grupo en sus prácticas, estos estudiantes denotan una mejoría en el segundo bimestre en sus participaciones en foros.

Los del tercer 3 grupo (2) obtuvieron una alta participación en agregado y lectura de mensajes y cualquier acción que involucre foros, dedican más tiempo al cumplimiento de esta actividad, registran las más altas calificaciones en foros, alcanzaron notas finales mínimas pero si registran un incremento en sus calificaciones de un bimestre a otro.

En la Tabla 2.14 se muestra los patrones con los que se desenvuelven los estudiantes de este curso.

PARÁMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directo	Bajo	Medio	Alto
Grado de Colaboración Indirecto	Bajo	Medio	Alto
Tiempo Empleado	Bajo	Medio	Alto
Promedio en Foros	Bajo	Medio	Alto
Mejora de Bimestre	Bajo	Alta	Medio
Calificación Final	Alta	Alta	Baja

Tabla 2. 14 Recopilación de resultados Fundamentos de la Programación

El Grado de colaboración “Bajo” es realmente preocupante si se presta atención al resto de parámetros que están en las mismas condiciones.

Así pues los patrones obtenidos de los estudiantes demuestran que:

- ◆ En la medida que acceden los estudiantes colaboran, por ejemplo los estudiantes que accede un mayor número de veces tienden a crear un mayor número de mensajes tanto como respuestas al moderador como a las de sus compañeros.

- ◆ El factor “tiempo” indica en este caso que ningún estudiante ha ocupado más tiempo del que realmente ha necesitado para cumplir con sus actividades, es decir los alumnos que más tiempo han empleado mejores resultados presentan contrariamente a los que dedican un tiempo mínimo.
- ◆ Los alumnos que poco o en forma nula participan actualizando y agregando nuevos mensajes se ubican en un rango alto de calificaciones finales debido a que se promedian con las notas de otras actividades.
- ◆ Los integrantes del Grupo 1 están en un “proceso de crecimiento académico” pues como ya se indicó se ubican en un punto medio en cada una de las interacciones que realizan, en el tiempo utilizado y en el promedio final de foros pero se distinguen por la mejoría en su desempeño del primer al segundo bimestre y la más alta calificación final. Los alumnos de este grupo muestran tener una alta motivación
- ◆ Los alumnos del Grupo 2 representan un caso especial, en primera instancia son los que mayores valores presenta, sobresalen en el grado de colaboración en foros y en su calificación bimestral aunque el rango en que se ha superado sea bajo no es mayor problema, lo que sí lo es la escasa calificación final.
- ◆ Este curso posee un alto número de estudiantes del sexo masculino 122 a diferencia de 24 del sexo Femenino, en los tres grupos predominan los hombres siendo en el Cluster 1 donde se concentran en su mayoría en otras palabras en promedio los hombres tienden a tener un grado de colaboración media, las mujeres se ubican en igual número en el Cluster 0 y 1 lo que indican que su colaboración tiende a ser de media a baja.

“Promedio en Foros” es una instancia que únicamente (de los cursos estudiados) se encuentra en este curso, el docente ha tomado en cuenta cada una de las participaciones de los estudiantes y les ha dado una calificación o rating de acuerdo a su contribución.

Se observa un patrón de comportamiento homogéneo de los estudiantes diferenciándose únicamente de la mejora bimestral y la calificación final.

En función de la colaboración realizada los diferentes grupos se calificarían así:

Grupo 2: NIVEL DE COLABORACIÓN ALTA

Grupo 1: NIVEL DE COLABORACIÓN MEDIA

Grupo 0: NIVEL DE COLABORACIÓN BAJO

De un 100% de estudiantes de la materia de Fundamentos de la programación los alumnos con un nivel de colaboración alta representa un 29% (42), los de nivel de colaboración media un 48%(70) y los de bajo nivel colaborativo un 23%(34) de una población total de 146 alumnos.

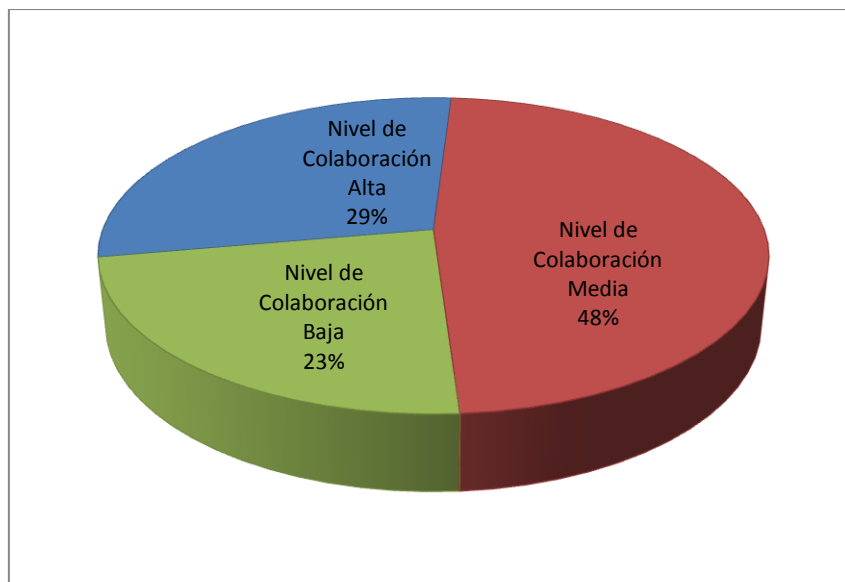


Figura 2. 33 Gráfica por criterios de colaboración en Foros del curso "Fundamentos de la Programación"

Se observa en la Figura 2.33 como los estudiantes con nivel de colaboración media dominan el escenario, seguidos de los de baja y por último los que cuentan con un mayor número de colaboraciones. Que la mitad de una población se encuentre en un punto medio de colaboración significa que existen alumnos entusiastas por contribuir y aprender pero que deben ser orientados de forma que optaran por participar un poco más de lo acostumbrado.

2.6.2. Lógica de la Programación

Para este curso no existió calificación en los foros, según lo observado en el libro de calificaciones.

El tamaño máximo que se ha permitido para los archivos adjuntos es de: 5MB, al igual que Fundamentos de la Programación no se ha forzado la suscripción a foros, así como el rastreo que se ha desconectado.

Lógica de la Programación si registra actividad para los atributos: subtemas_leídos, num_respuestas_debates y arch_adjuntos.

Una vez más esta asignatura está formada por hombres que representan un 93% del total de alumnos (154).

Los grupos de este curso se calificaron tomando en cuenta cuatro aspectos: El Grado de colaboración directa, el grado de colaboración indirecta, el tiempo ocupado en esta actividad y la calificación final, la instancia “prom_foros” no se ha utiliza para este curso puesto que los foros no fueron valorados.

El primer grupo (0) el número de acceso a foros es mínimo, número mínimo de debates, pocos foros en los que se ha participado, número de mensajes mínimos agregados y actualizados, tiempo menor ocupado en foros, finalmente los miembros de este grupo son los que junto al tercer grupo menor calificación poseen (< 25). Este grupo corresponde al de la minoría de estudiantes.

El segundo grupo (1) por su parte cuenta con un mayor número de acceso a foros, número mayor de debates, mayor número de foros en los que se ha participado, actualizado y en los que se ha adjuntado un archivo, número menor que el grupo 2 pero mayor que el 0 de mensajes agregados, tiempo promedio en actividad de foros, nota final mucho mayor que el primer grupo (0). Son el grupo de mayor representación.

El tercer grupo (2) poseen un acceso promedio a foros, así mismo de discusiones leídas y mensajes creados, aunque los debates se ubican en un punto medio, los estudiantes no han subido archivos pero si se han suscrito a la mayoría de foros y han ocupado un tiempo importante en el cumplimiento de esta actividad. Su nota final se ubica en el rango de menor a 25.

Una recopilación de estos resultados se presenta en la Tabla 2.15

PARÁMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directa	Bajo	Alto	Medio
Grado de colaboración indirecta	Bajo	Alto	Bajo/Nulo
Tiempo Empleado	Bajo	Medio	Alto
Calificación Final	Bajo	Alto	Bajo

Tabla 2. 15 Recopilación de resultados Fundamentos de la Programación

Los patrones de comportamiento que se han recopilado de este curso son:

- ◆ Los alumnos dedican buena parte de su tiempo a las prácticas colaborativas (agregar mensajes, actualizar, responder mensajes). En este caso las interacciones si se ven reflejadas en su nota final.
- ◆ Para este curso existen debates aunque pocos estudiantes hacen uso de estos, una vez más por la poca práctica que poseen (esta materia se da en Primer Ciclo).
- ◆ En Lógica de la programación aunque los alumnos poco ingresen a la plataforma educativa si participan de forma mediana en la ejecución de las actividades que impliquen foros.
- ◆ El tiempo no es un indicador que se dé de forma homogénea, pues aquellos que registran mayor tiempo no necesariamente están generando una colaboración.
- ◆ El grupo de quienes colaboran en superior medida son la mayoría en este curso.
- ◆ El número de estudiantes del sexo masculino es notablemente mayor (143) en comparación con los del femenino (11) en ambos casos el mayor número de estudiantes cuentan con un alto grado de colaboración esto es proporcional a su calificación final.

La escala descriptiva de los grupos en cuanto a colaboración en foros es:

Grupo 1: NIVEL DE COLABORACIÓN ALTA

Grupo2: NIVEL DE COLABORACIÓN MEDIA

Grupo 0: NIVEL DE COLABORACIÓN BAJO

El total de alumnos para este curso es 154. Luego del análisis realizado se puede determinar qué: El porcentaje de estudiantes con nivel de colaboración alta es del 39% lo que resulta en 60 individuos, los de colaboración media con una representatividad del 34% es decir 53 y el de colaboración bajo un 27% significando 41 estudiantes.

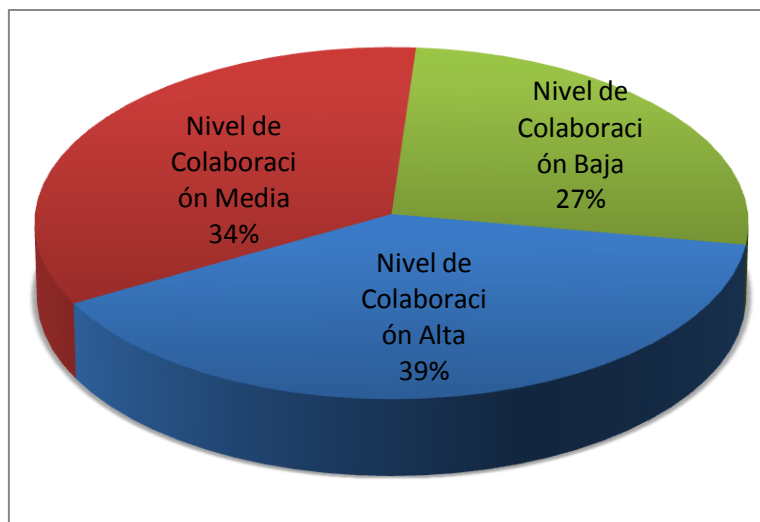


Figura 2. 34 Gráfica por criterios de colaboración en Foros del curso "Lógica de la Programación"

La Figura 2.34 denota claramente una diferencia no tan pronunciada entre quienes han colaborado activamente y quienes se mantienen en una posición intermedia y baja.

Esta imagen da fe de que los estudiantes realmente están contribuyendo, receptando información y generando conocimiento a modo de ideas y pensamientos que es como se suele hacer en los foros.

2.6.3. Fundamentos Informáticos

Para este curso no se ha realizado evaluación alguna de foros, en cuanto a los mensajes adjuntos el tamaño máximo es de: 5MB. Una vez más no se ha forzado a los estudiantes a suscribirse a los foros, ni a ejecutar el rastreo de estos. Se prefiere utilizar los canales rss para los mensajes.

Al igual que Lógica de la Programación los parámetros que se revisaran serán: Grado de colaboración directo, Grado de colaboración indirecto, tiempo empleado en cualquier actividad que implique foros y calificación final, a excepción de la instancia "debates" que si bien existen entradas estas son mínimas.

Los grupos cumplen con las siguientes características:

El primer grupo (0) cuenta con un número de acceso a foros menor que el Grupo 2 pero mayor que el Grupo 1, el número de discusiones leídas es mayor que el Grupo 1 y bastante menor con respecto al Grupo 2, lo que se repite para el número de mensajes creados y al de foros suscritos, tanto la participación en debates, mensajes

actualizados y archivos adjuntos es Nula, el tiempo utilizado en este grupo es mayor que el del resto.

El segundo grupo (1) cuenta con un bajo nivel colaborativo tanto en el número de acceso en foros, como en las discusiones leídas, el número de mensajes creados, los foros a los que se han suscrito y el promedio de horas que ocupan en los foros. Las discusiones creadas, los mensajes actualizados y los archivos adjuntos poseen un estado nulo la nota final es menor a 25 sobre 100.

El tercer grupo (2) registra un mayor número de acceso a los foros, son los únicos que crean y revisan discusiones estos estudiantes se han suscrito a un mayor número de foros y han participado positivamente en la creación de estos. A diferencia de los otros grupos estos alumnos han adjuntado archivos a sus respuestas pero no han dedicado mayor tiempo que el Grupo 0 en esta actividad. Cuentan con una calificación menor a 25 sobre 100.

La Tabla 2.16 recopila estos resultados.

PARÁMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directa	Medio	Bajo	Alto
Grado de colaboración indirecta	Medio	Bajo	Alto
Tiempo Empleado	Alto	Bajo	Medio
Calificación Final	Alto	Bajo	Medio

Tabla 2. 16 Recopilación de resultados Fundamentos Informáticos

El Cluster 1 aunque es homogéneo y representa el grupo con mayor número de integrantes 92 (53%) también constituye el de menor nivel colaborativo con una amplia diferencia de medias en sus centroides en comparación con el resto de segmentos.

La mayoría de los estudiantes de este curso poseen un nivel de participación mínima pero en la nota final tienden a recuperarse notablemente.

El tiempo empleado se corresponde de forma directa en cada uno de los clusters con lo que concluye que para estos profesionales la dedicación es un factor importante para la obtención de un rendimiento sobresaliente.

El grado de colaboración se basa en la siguiente escala cualitativa resultado de la experimentación:

Grupo 2: NIVEL DE COLABORACIÓN ALTA

Grupo 0: NIVEL DE COLABORACIÓN MEDIA

Grupo 1: NIVEL DE COLABORACIÓN BAJO

El total de la población en este curso es de: 174

Quienes cuenta con un nivel de colaboración alta representan un 14% (24) total de la población, los de mediano nivel colaborativo un 33% (58) y aquellos con niveles mínimos de colaboración representan una mayoría con un 53% (92).

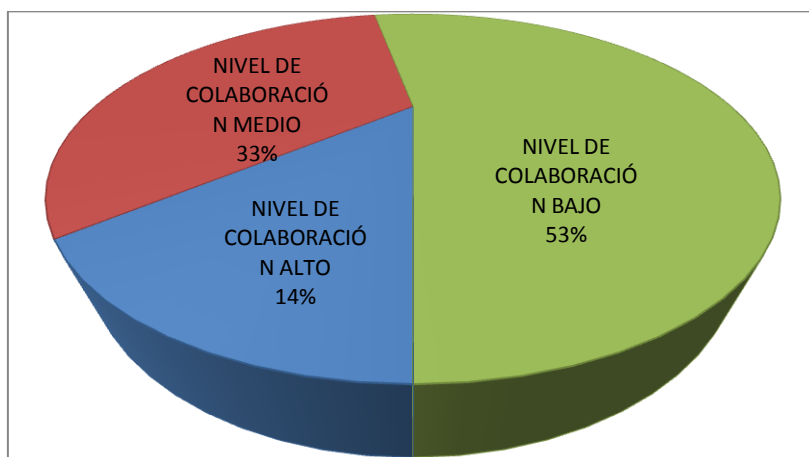


Figura 2. 35 Gráfica por criterios de colaboración en Foros del curso "Fundamentos Informáticos"

En la Figura 2.35 se observa que el nivel de colaboración bajo es el que predomina, lo cual genera un patrón de comportamiento colaborativo con resultados poco alentadores, pero que pueden mejorarse con la capacitación y motivación correspondiente.

Los patrones de colaboración en Fundamentos de la Programación reflejan que:

- ◆ Mientras mayor número de veces accedan los estudiantes mas colaboran.
- ◆ Quienes mayormente participan en foros no ocupan el factor tiempo en la misma cantidad.
- ◆ Existen 142 alumnos del sexo Masculino diferenciándose con 32 del sexo Femenino. Tanto hombres como mujeres en su mayor número se ubican en el cluster 1 con nivel colaborativo bajo.
- ◆ Los estudiantes prefieren o creen conveniente no adjuntar archivos en sus respuestas.

El comportamiento colaborativo en los tres cursos pudiera diferenciarse debido a factores como: El conocimiento de las herramientas con las que cuenta en el EVA, La dificultad propia de la materia, motivación del docente y al tiempo disponible de cada estudiante.

No se pueden tomar más atributos ni acciones colaborativas ya que no han sido habilitadas todos los recursos con los que cuenta esta actividad

DISCUSIÓN

La presente tesis abordó la investigación, análisis y uso de las técnicas de minería de datos, específicamente de los algoritmos de agrupamiento K-MEANS, EXPECTATION MAXIMATION y CLUSTER JERÁRQUICO para la realización de experimentos que permitieron extraer patrones de comportamiento colaborativo en Foros de los estudiantes de los cursos de: Fundamentos de la Programación, Lógica de la Programación y Fundamentos Informáticos de la UTPL, modalidad ABIERTA dentro del periodo Octubre 2010-Febrero 2011.

Estado de la Situación Actual de los Entornos Colaborativos

Se inició con el estudio de la Minería de Datos en la Educación, la cual mediante criterios fundamentados dan pautas para la personalización del proceso de enseñanza que se adapte al usuario de modo que pueda aprender de una forma más óptima.

El estudio de los algoritmos de agrupamiento K-MEANS, EM y Clustering Jerárquico también formó parte de este estudio, se seleccionaron estos dos debido a su naturaleza numérica y de particionado adaptable a todo tipo de datos de entrada.

Las herramientas que se utilizaron fueron WEKA y SPSS Clementine si bien la segunda proporcionaba una interfaz gráfica intuitiva estaba orientada más a un nivel de usuario final se realizó un experimento con esta más WEKA fue la elegida para la experimentación general en primera instancia, como complemento a estas se utilizó KNIME para la ejecución de pruebas de Clustering Jerárquico.

Minería de datos en Entornos Colaborativos del EVA

En esta segunda etapa se establecieron las fases para el proceso de Minería de datos, tomando como base el propuesto por (Hernández, Ramírez, & Ferri, 2004).

En la etapa de Integración y recolección se adquirió la base de datos de los estudiantes de la UTPL la que fue luego integrada a MOODLE para una construcción del escenario más cercano a la realidad.

El preprocesamiento (Limpieza, Selección y Transformación) fue una etapa que constituyó el 50% de este trabajo y consistió en la elección de los cursos con los que se trabajaría utilizándose la herramienta GEPHI para una selección fundamentada en el número de contribuciones que poseía el docente y luego la materia que impartía, se realizó el tratamiento de los datos, la selección de aquellos que proporcionarían información relevante a la participación de foros y su relación con el aspecto académico del educando. Luego de esto se crearon las consultas que traerían esta

data y fueron llenados a modo de matriz, para luego ser cargadas en la base de datos y transformada a formato .CSV se procedió así para cada una de las materias.

La etapa de Minería fue la siguiente en ser desarrollada se dividieron las experimentaciones por algoritmo. Se trató el Algoritmo K-MEANS en primera instancia, para Fundamentos de la Programación a los atributos ya definidos se agregó un indicador más: “mejoría de un estudiante de un bimestre a otro”. Este experimento se lo realizó de dos formas: 1) Con variables numéricas en su totalidad y 2) con la variable “nota_final” discretizada. Dicha variable se escogió porque el rango de calificaciones era amplio.

De las pruebas realizadas se obtuvo que el conjunto de datos sin discretizar presentaba un menor valor en la suma cuadrática de errores a diferencia del que si fue discretizado, además de esto se probó con diferentes iteraciones y semillas hasta encontrar un error mínimo.

En los cursos de: Lógica de la Programación y Fundamentos Informático no se ha realizado la calificación de los foros por lo que las calificaciones de ambos bimestres “prom_foros” no pudieron utilizarse para estas materias, de igual forma se procedió con la experimentación y con la prueba de diferentes semillas e iteraciones.

En todos los experimentos el cambio de valor en iteraciones no provocó ninguna diferencia en los resultados.

Luego de esto se realizaron las pruebas con el Algoritmo EM, donde el indicador esta vez es la verisimilitud que es la probabilidad de que un elemento pertenezca o no a un cluster, además se cambió el valor de semillas e iteraciones para identificar la variación que podrían tener pero no se observó diferencia alguna.

Los resultados de este algoritmo sirvieron de complemento al de K-Means pues la probabilidad de que estos existan en un cluster se correspondía.

Llegando a este punto se contempló la posibilidad de incrementar el número de atributos de tal forma que se obtuvieran características más amplias que permitieran realizar un juicio más centrado en el comportamiento colaborativo de los estudiantes en foros.

Lo que llevó a una reestructuración en el pre procesamiento y a partir de los experimentos realizados la generación de uno nuevo que contempla las deficiencias de sus antecesores, uniéndose a K-Means y a EM , el algoritmo de Cluster jerárquico.

Finalmente en la etapa de evaluación e interpretación se analizaron los resultados de la etapa anterior concluyendo que en Fundamentos de la Programación la interacción no estaba relacionada con la calificación final; el tiempo empleado fue usado en la misma medida con la que se detectó su interacción es decir los que mayor colaboración representaban mayor tiempo ocupaban en la realización de acciones concernientes a foros, otro punto a tomar en cuenta es la existencia de un grupo de estudiantes que se ubicaban en un punto medio en interacción pero que se ha superado en sus participaciones de un bimestre a otro. El tipo de foro predominante en todos los cursos fue el sencillo.

Para Lógica de la Programación se han utilizado a parte de los foros sencillos, las discusiones pero con un mínimo de participaciones, el grupo en forma general cuenta con un nivel alto de colaboración en su mayoría. Este curso ocupa un tiempo importante en el cumplimiento de las actividades concernientes a foros, aunque esto no siempre implica una participación activa, en este curso si ha sido un factor importante.

En el curso de Fundamentos Informáticos el mayor grupo de estudiantes son los que menor colaboración tienen, el tiempo empleado está ligado a la calificación final recibida lo que hace pensar que los miembros del curso dedican el tiempo preciso para cada una de las actividades en foros, estas comprenden el agregado, actualización y revisión de mensajes.

Cuando los estudiantes no se ven forzados a participar en los foros por lo general no muestran motivación para realizarlo. Ninguno de los foros han sido bloqueados es decir no se ha delimitado el tiempo en el que estarán disponibles lo que implica el descuido por parte de los educandos.

Otro patrón importante que se ha observado es que solo en el curso Fundamentos de la Programación los foros han sido evaluados implicando que los estudiantes se esfuercen por participar de una forma más contundente.

Al menos debe existir 1 foro por bimestre cuando es así los resultados de la colaboración se incrementan cuando son más de 1 la colaboración se reparte entre todos los existentes.

Finalmente el comportamiento colaborativo de los estudiantes esta íntimamente ligado a factores como el tiempo disponible, la predisposición y los recursos utilizados de la plataforma.

CONCLUSIONES Y RECOMENDACIONES

A más de las conclusiones expuestas a lo largo de este trabajo se indican las siguientes:

- ◆ El nivel colaborativo de los estudiantes no es proporcional a su calificación final, influyen otros factores como: calificaciones de otras actividades y exámenes.
- ◆ El tiempo que registran en la actividad foros pudiera no utilizarse con fines colaborativos.
- ◆ Aunque la cantidad de estudiantes del sexo femenino sea minoritaria, sus colaboraciones se encuentran en el mismo nivel que los del masculino.
- ◆ Las capacidades tecnológicas y actitudes del docente en un entorno a distancia son preponderantes en el rendimiento académico de sus educandos es así que se debe contar con capacitaciones periódicas de tal forma que se aproveche al máximo los recursos con los que cuenta la plataforma educativa.
- ◆ Categorizar a los estudiantes por su nivel de colaboración permite a los docentes centrarse en aquellos alumnos que necesitan mayor atención y soporte. La retroalimentación que se realice no solo debe señalar puntos negativos de una conducta sino también reforzar la actitud comprometida de los estudiantes y las mejoras que puedan tener dentro de un periodo de tiempo.
- ◆ El trabajo colaborativo es sin duda el mayor apoyo con el que puede contar un estudiante, puntos de vista diferentes o similares permiten la existencia de debates que enriquecen el pensamiento analítico y crítico de los alumnos.
- ◆ La etapa de Pre Procesamiento en Minería de Datos constituye un 50% mínimo del total de un proyecto, esta fase aún terminada si no cumple con las expectativas de la Minería deberá ser revisada y cambiada cuantas veces sea necesario.
- ◆ La búsqueda de nuevas instancias incrementaría la eficiencia en la formación de grupos al maximizarse el número de similitudes a evaluarse.
- ◆ Generalmente las participaciones que se ubican en un punto medio son las que tienen mayor predisposición para obtener una calificación alta.
- ◆ El clustering al categorizarse como descriptivo fue la técnica que mejor se adapta para la realización de esta investigación, el reunir grupos por características colaborativas similares es el punto focal de este trabajo.
- ◆ K-Means fue la técnica que mejor se adaptó a los objetivos de este trabajo por la celeridad en la conformación de grupos especialmente cuando la población no es de gran tamaño como fue en este caso.

- ◆ Tanto EM como el Clustering Jerárquico sirvieron de complemento a K-Means para el análisis de los grupos debido a su naturaleza probabilística y subjetiva en ese orden.
- ◆ Los estudiantes mostraron de forma global un bajo interés colaborativo en gran parte de ellos fue nulo, dando a entrever falta de motivación o habilidad para la ejecución de esta actividad.

RECOMENDACIONES

- ◆ Se recomienda la capacitación de los estudiantes de los primeros ciclos acerca del uso de los foros en el EVA de tal manera que se le dé el seguimiento pertinente, identificándose los errores más comunes que pudiera cometer y que obstruyeran la ejecución plena de sus actividades.
- ◆ Ampliar el campo de investigación de las habilidades sociales en el EVA a fin de optimizar su uso y generar nuevos conocimientos.
- ◆ Se podría experimentar además con la formación de grupos de trabajo con la capacidad de calificarse entre sí cada una de sus contribuciones generando un ambiente colaborativo activo.
- ◆ Así también para la experimentación se recomienda hacer todas las variaciones posibles de modo que se obtenga los resultados más fiables.
- ◆ El uso de atributos numéricos como nominales para una representación de los datos más cercana a la realidad.
- ◆ Se recomienda la evaluación de las herramientas de Data Mining que mejor se adapten a los objetivos planteados en la investigación.

TRABAJOS FUTUROS

Se sugieren los siguientes trabajos que en un futuro podrían realizarse en el ámbito de comportamientos colaborativos en foros.

- ◆ La realización de la minería de datos en diferentes periodos de tal forma que se compruebe si los patrones se repiten.
- ◆ Minería de Texto aplicada a los mensajes valorándolos cualitativamente bajo criterios de: conocimiento del tema y relevancia.
- ◆ La creación de grupos de trabajo formados aleatoriamente a quienes se le asigne una tarea específica con la capacidad de que sus miembros puedan calificarse entre sí, el objetivo de esto es medir el grado de colaboración de los estudiantes y su capacidad para trabajar en equipo.

BIBLIOGRAFÍA

- ◆ Arteaga, & Fabregat. (2002). *Integración del aprendizaje individual y del colaborativo en un sistema hipermedia adaptativo*. (I. d. (IliA)., Ed.) España: Universitat de Girona.
- ◆ Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. 1-14.
- ◆ Boeira, A. (2001). *Um Modelo do Aluno Adaptativo para Sistemas na Web*. Porto Alegre, Rio de Janeiro.
- ◆ Bratits et al. (2008). Supporting members of a learning community using interaction analysis tools: the example of the Kaleidoscope NoE scientific network Proceedings of the IEEE International Conference on Advanced Learning Technologies. *ICALT 2008*, (págs. 809-813). Santander, España.
- ◆ Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted* , 6, 87-129.
- ◆ CES. S. RAMÓN Y CAJAL. (2008). *Manual Moodle*. Obtenido de <http://www.campuscajal.com>.
- ◆ Collazos et al . (2007). *Evaluating Collaborative Learning Processes using System-based Measurement*. 10(3).
- ◆ Collazos et al. (2002). *Evaluating Collaborative Learning Processes*. Universidad de Chile, Department of Computer Science. Springer-Verlag Berlin Heidelberg 2002.
- ◆ Corso, C., & Alfaro, S. (2010). *Algoritmos de Data Mining aplicados en la enseñanza basada en la Web*. Universidad Tecnológica Nacional, Departamento de Sistemas de Información, Córdoba.
- ◆ Cruz, L. (2010). *Minería de datos con Aplicaciones*. Universidad NacionalAutonomade Mexico.
- ◆ Daradoumis et al. (2006). A Layered Framework for Evaluating OnLine Collaborative Learning Interactions. *International Journal of Human-Computer Studies*, 64, págs. 622-635.
- ◆ De la Fuente Valentín, e. a. (Mayo de 2009). Modelos de Aprendizaje Colaborativo en Entornos a Distancia con Learning Design: Un Caso de Estudio. *IEEE-RITA* , 4 (2) .

- ◆ De Pedro, e. a. (2007). New Method Using Wikis and Forums to Evaluate Individual Contributions in Cooperative Work while Promoting Experiential Learning: Results from Preliminary experience. *Symposium On Wikis (WikiSym)*. Montreal.
- ◆ Duran, E. (2006). Modelo del Alumno para Sistemas de Aprendizaje Colaborativo. *Workshop de Inteligencia Artificial en Educación*. Mendoza: Universidad Nacional de Santiago del Estero.
- ◆ Felder M, R., & Silverman K, L. (1988). *Learning and Teaching Styles In Engineering Education* (Vol. 78(7)). Engr. Education.
- ◆ Figueras, S. (2001). *Análisis de conglomerados o cluster*.
- ◆ Galbiate, J. (2011). *Material de Apoyo al Aprendizaje de la Estadística*. Recuperado el 27 de 09 de 2011, de Homepage de Jorge Galbiati Riesco: http://www.jorgegalbiati.cl/ejercicios_4/ConceptosBasicos.pdf
- ◆ García, D. (2005). *Manual de WEKA*.
- ◆ García, M., & Álvarez, A. (2003). *Análisis de Datos en WEKA – Pruebas de Selectividad*.
- ◆ García, M., & Quintales, L. M. (2002). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN LA CONSTRUCCIÓN Y VALIDACIÓN DE MODELOS PREDICTIVOS Y ASOCIATIVOS A PARTIR DE ESPECIFICACIONES DE REQUISITOS DE SOFTWARE*. Salamanca.
- ◆ Garre, M., Cuadrado, J. C., & Sicilia, M. (2005). *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*. Universidad de Alcalá, Departamento de Ciencias de la Computación, Alcalá de Henares, Madrid.
- ◆ Gaudioso Vásquez, E. (2002). *Contribuciones al Modelado del Usuario en Entornos Adaptativos de Aprendizaje y Colaboración a través de Internet mediante técnicas de Aprendizaje Automático*. Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial, Madrid.
- ◆ Gephi. (5 de Marzo de 2010). Gephi Tutorial Quick Start. *0.7alpha2*.
- ◆ González G, H. M., Duque M, N. D., & Ovalle C, D. A. (2008). Modelo del Estudiante para Sistemas Adaptativos de Educación Virtual. *Revista Avances en Sistemas e Informática* , 5 (1), 199-206.
- ◆ Guerra, L. (07 de Mayo de 2008). Primeros pasos con KNIME.

- ◆ Hernández Valadez, E. (2006). *Algoritmo de clustering basado en entropía para descubrir grupos en atributos mixtos*. Instituto Politécnico Nacional, Departamento de Ingeniería Eléctrica. México: Centro de Investigación y de Estudios Avanzados.
- ◆ Hernandez, J., Ramirez, M., & Ferri, C. (2004). *Introducción a la Minería de Datos*. Pearson.
- ◆ Hong, W. (2001). Spinning Your Course Into A Web Classroom - Advantages And Challenges. *International Conference on Engineering Education* Augusto 6-10. Oslo, Norway.
- ◆ IBM®. (Agosto de 2011). *IBM*. Recuperado el Agosto de 2011, de <http://www-01.ibm.com/software/analytics/spss/products/modeler/>
- ◆ IOS. (12 de 03 de 2006). Recuperado el 01 de 11 de 2011, de Sitio Web de Improved Outcomes Software: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Manhattan_Distance_Metric.htm
- ◆ J. Pérez¹, e. (2007). Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer. *2º Taller Latino Iberoamericano de Investigación de Operaciones*. Acapulco, Guerrero, México: México.
- ◆ Jermann, P., Soller, A., & Muehlenbrock, M. (2001). From Mirroring to Guiding: A Review of State of the Art Technology for Supporting Collaborative Learning. *First European Conference on Computer-Supported Collaborative Learning* (págs. 324–331). Kai Hakkarainen Pierre Dillenbourg, Anneke: European Perspectives on Computer-Supported Collaborative Learning - Maastricht McLuhan Institute.
- ◆ Martinez et al, .. (2006). Studying participation networks in collaboration using mixed methods. *International Journal of Computer-Supported Collaborative Learning*, 1, págs. 383-408.
- ◆ Meier et al. (2007). *A rating scheme for assessing the quality of computer-supported collaboration processes*. Freiburg, Alemania: Computer-Supported Collaborative Learning 2:63–86.
- ◆ Molina, J., & García, J. (2004). *Técnicas de Análisis de Datos*. Universidad Carlos III de Madrid, Madrid.

- ◆ MOODLE. (12 de Junio de 2009). *Moodle.org*. Recuperado el 10 de 10 de 2011, de Foros: <http://docs.moodle.org/19/es/Foros>
- ◆ Nevárez, L. (s.f.). *Minería de Datos*. Recuperado el 06 de 09 de 2011, de <http://leonardonevarez.host56.com/archivos/MineriaDatos.ppt>
- ◆ Orallo, J., & Ferri, C. (2006). *Curso de Doctorado Extracción Automática de Conocimiento*. Universitat Politècnica de València.
- ◆ Perera et al. (2009). Clustering and Sequential Pattern Mining of Online Collaborative Learning Data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 21, págs. 759-772.
- ◆ Redondo et al, .. (2003). *Applying Fuzzy Logic to Analyze Collaborative Learning Experiences in an e-Learning Environment*. USDLA Journal. (United States Distance Learning Association).
- ◆ Rodríguez Anaya, A. (2009). *Prospección de la colaboración utilizando herramientas de minería de datos en ambiente abiertos de aprendizaje colaborativo con el objetivo de mejorar la gestión del proceso de colaboración*. Universidad Nacional de Educación a Distancia, Departamento de Inteligencia Artificial, Madrid.
- ◆ Romero Morales, C., Ventura Soto, S., & Hervás Martínez, C. (2005). *Estado actual de la aplicación de la minería de datos a los sistemas de enseñanza basada en web*. Escuela Politécnica Superior. Universidad de Córdoba, Departamento de Informática y Análisis Numérico.
- ◆ Romero, C., & Ventura, S. (2010). *Educational Data Mining: A Review of the State of the Art*. *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS*.
- ◆ Romero, C., Ventura, S., & García, E. *Data mining in course management systems: Moodle case study and tutorial*. University of Córdoba, aDepartment of Computer Sciences and Numerical Analysis, Córdoba.
- ◆ S. Zhang, C. Z. (2003). *Data preparation for data mining*. *Applied Artificial Intelligence*.
- ◆ Saharkhiz, A. (03 de 01 de 2009). *K-Means Clustering Used in Intention Based Scoring Projects*. Recuperado el 20 de 09 de 2011, de [codeproject.com: http://www.codeproject.com/KB/recipes/K-Mean_Clustering.aspx](http://www.codeproject.com/KB/recipes/K-Mean_Clustering.aspx)
- ◆ Talavera, L., & Gaudioso, E. (2004). Mining Student Data To Characterize Similar Behavior Groups In Unstructured Collaboration Spaces. *In: Proceedings*

of the Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence, (págs. 17-23). Valencia, España.

- ◆ Trcka, N., & Pechenizkiy, M. (2009). *From Local Patterns to Global Models: Towards Domain Driven Educational Process Mining*. Eindhoven University of Technology, Department of Computer Science.
- ◆ Valdiviezo, P. M., Santos, O. C., & Boticario, J. G. (2010). Aplicación de Métodos de diseño centrado en el usuario y minería de datos para definir recomendaciones que promuevan el uso del foro en una experiencia virtual de aprendizaje. *13:2*, 237-264.
- ◆ Vicente, J. (2006). *Introducción al Análisis del Clusters*. Universidad de Salamanca, Departamento de Estadística.
- ◆ Wikipedia. (25 de Octubre de 2011). Recuperado el 25 de Junio de 2011, de Validación Cruzada: http://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada
- ◆ Zhunio, F. (02 de 08 de 2011). *Ahora haciendo data mining con microsoft sql server*. Recuperado el 20 de 09 de 2011, de Métodos de discretización (minería de datos) en SQL Server Analysis Services

Ing. Samanta P. Cueva

DIRECTORA DE TESIS

Ing. Priscila M. Valdiviezo

CO- DIRECTORA DE TESIS

Cinthia E. Pulla E

TESISTA

ANEXOS

ANEXO A

Integración de base de datos en MOODLE.

En primera instancia se contó con la base de datos en extensión .sql sin interfaz gráfica lo que de cierta forma dificultaba el escenario en el que se desenvolvían las interacciones. Por ello se optó por su integración con el entorno MOODLE.

Para lo cual hubo que dirigirse al servidor que contiene la carpeta XAMPP como se muestra en la Figura A.1

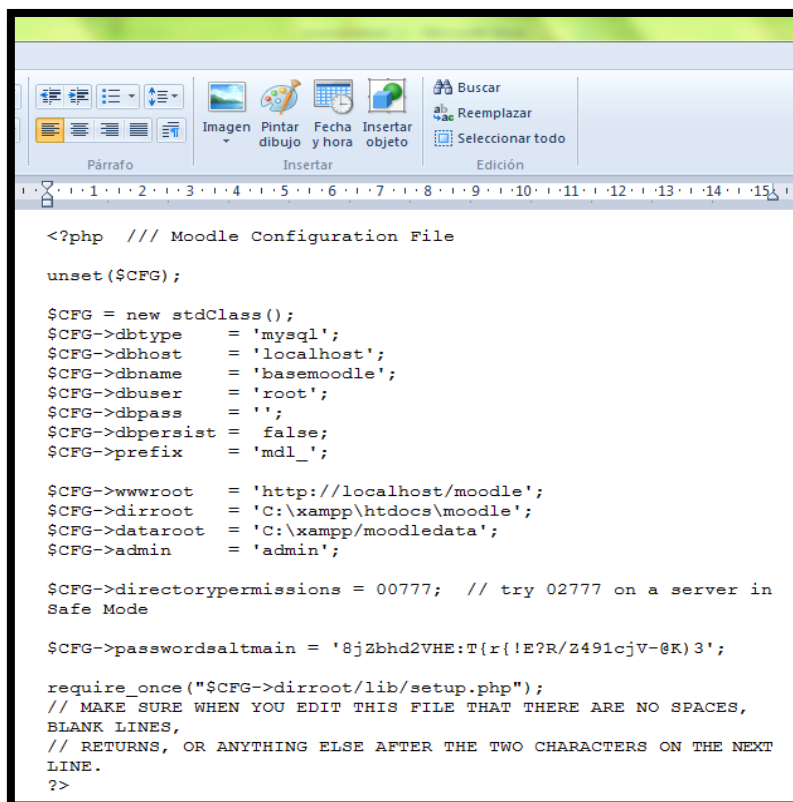
Nombre	Fecha de modifica...	Tipo	Tamaño
Joomla	13/11/2010 22:09	Carpeta de archivos	
moodle	22/03/2011 13:52	Carpeta de archivos	
phd_2_05	13/11/2010 10:14	Carpeta de archivos	
PHPMailer	13/11/2010 20:23	Carpeta de archivos	
server	01/03/2011 8:44	Carpeta de archivos	
xampp	20/12/2009 0:00	Carpeta de archivos	

Figura A.1 Directorio del Servidor XAMPP

Ingresa a “moodle”, ubicarse en el archivo “config.php” como se muestra en la Figura A.3.

Nombre	Fecha de modifica...	Tipo	Tamaño
lib	22/03/2011 13:31	Carpeta de archivos	
login	22/03/2011 13:32	Carpeta de archivos	
message	22/03/2011 13:32	Carpeta de archivos	
mnet	22/03/2011 13:32	Carpeta de archivos	
mod	22/03/2011 13:33	Carpeta de archivos	
my	22/03/2011 13:33	Carpeta de archivos	
notes	22/03/2011 13:33	Carpeta de archivos	
pix	22/03/2011 13:33	Carpeta de archivos	
question	22/03/2011 13:33	Carpeta de archivos	
rss	22/03/2011 13:33	Carpeta de archivos	
search	22/03/2011 13:33	Carpeta de archivos	
sso	22/03/2011 13:33	Carpeta de archivos	
tag	22/03/2011 13:33	Carpeta de archivos	
theme	22/03/2011 13:33	Carpeta de archivos	
user	22/03/2011 13:33	Carpeta de archivos	
userpix	22/03/2011 13:33	Carpeta de archivos	
config	30/03/2011 10:23	Archivo PHP	1 KB
config-dist	03/04/2010 8:03	Archivo PHP	20 KB
COPYING	07/05/2009 11:03	Documento de tex...	18 KB
file	10/04/2009 8:05	Archivo PHP	8 KB
help	03/03/2009 8:05	Archivo PHP	9 KB
index	26/04/2009 8:07	Archivo PHP	12 KB
install	22/02/2010 8:03	Archivo PHP	49 KB

Figura A.2 Path Archivo de Configuración MOODLE



```
<?php /// Moodle Configuration File

unset($CFG);

$CFG = new stdClass();
$CFG->dbtype      = 'mysql';
$CFG->dbhost      = 'localhost';
$CFG->dbname      = 'basemoodle';
$CFG->dbuser      = 'root';
$CFG->dbpass      = '';
$CFG->dbpersist  = false;
$CFG->prefix      = 'mdl_';

$CFG->wwwroot     = 'http://localhost/moodle';
$CFG->dirroot     = 'C:\xampp\htdocs\moodle';
$CFG->dataroot    = 'C:\xampp\moodledata';
$CFG->admin       = 'admin';

$CFG->directorypermissions = 00777; // try 02777 on a server in
Safe Mode

$CFG->passwordsaltmain = '8jZbhd2VHE:T{r{!E?R/Z491cjV-@K)3';

require_once("$CFG->dirroot/lib/setup.php");
// MAKE SURE WHEN YOU EDIT THIS FILE THAT THERE ARE NO SPACES,
BLANK LINES,
// RETURNS, OR ANYTHING ELSE AFTER THE TWO CHARACTERS ON THE NEXT
LINE.
?>
```

Figura A.3 Archivo conFigura php

Se abre el archivo conFigura php con el editor de texto como se observa en la Figura A.3 en "*dbname*" se especifica el nombre de la base de datos y de preferencia se podría especificar un nuevo "*wwwroot*" para el ingreso por el explorador a MOODLE.

A continuación con ayuda de phpMyAdmin se realiza una consulta sencilla para ubicar la tabla *prefix_conFigura* , Figura A.4

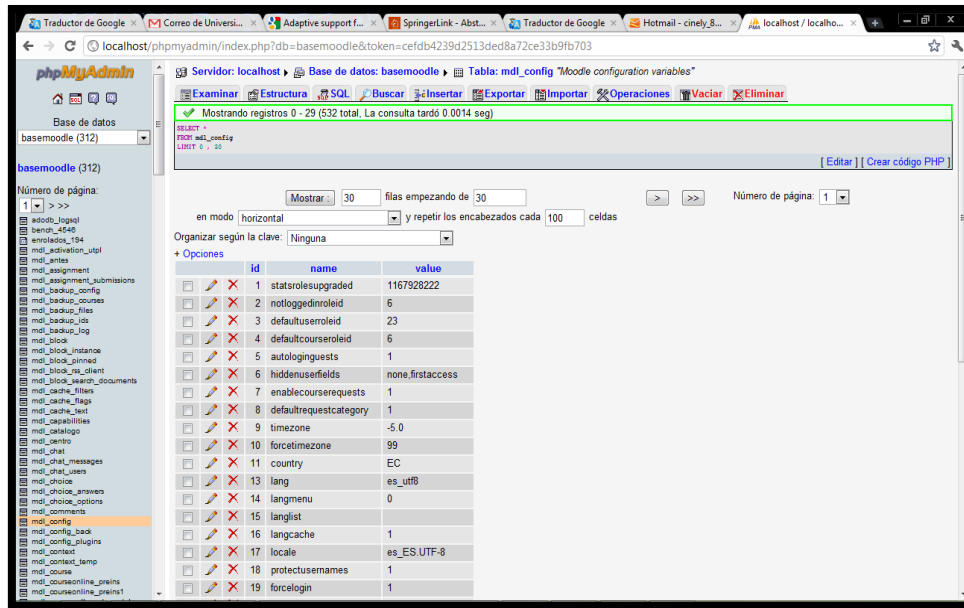


Figura A.4 Tabla prefix_conFigura en PHPMyAdmin

El objetivo es borrar del campo "alternateloginurl" el valor existente, Figura A.5, de este modo se permanecerá en la base de datos instalada y el explorador no se re direccionará hacia el portal de la UTP en línea.

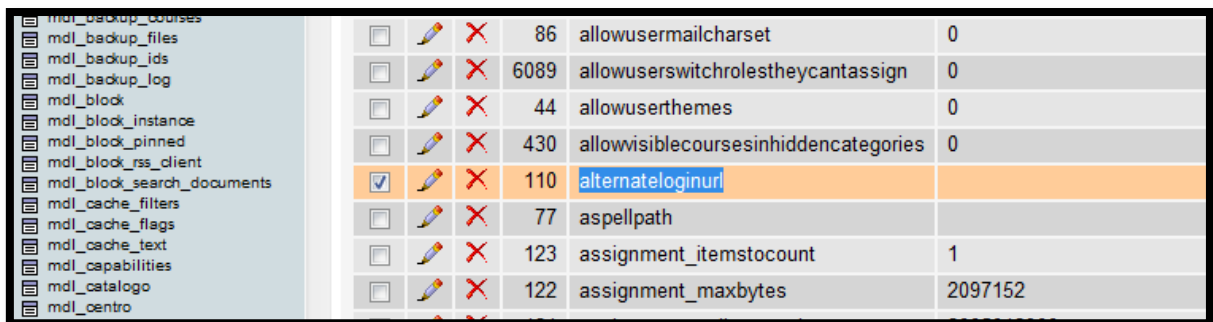


Figura A.5 Modificación campo “alternateloginurl”

Se Ingresa al entorno con el nombre de usuario y contraseña del “admin”, la categoría que nos compete “ABIERTA Y A DISTANCIA”. Figura A.6.



Figura A.6 Interfaz Moodle "Categorías"

Clic en la categoría mencionada, enseguida se presentará la ventana con las subcategorías correspondientes a esa modalidad. Esta clasificación representa a todos las carreras existentes.

ANEXO B

Consulta SQL de interacción en Foros

Fundamentos de la Programación [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course= 28741
```

798 mensajes en
8 foros tanto del
primer como
segundo bimestre.

Lenguaje de Alto Nivel [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=17323
```

538 mensajes en
10 foros.

Lógica de la Programación [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=21221
```

237mensajes en
2 foros.

Lógica de la Programación [B]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=28737
```

227mensajes en
5 foros.

2879

Fundamentos Informáticos [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=28739
```

102 mensajes en
4 foros.

Fundamentos Informáticos [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=21222
```

304 mensajes en
3 foros.

Base de Datos II [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=2132
```

71 mensajes en 9
foros.

Base de Datos II [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=22065
```

99 mensajes en 9
foros.

Base de Datos I [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=2131
```

44 mensajes en 3
foros.

Sistemas III [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=29012
```

66 mensajes en 5
foros.

2912

Sistemas basados en el conocimiento [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=2170
```

34 mensajes en 2
foros.

Lógica de la programación [C]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=30055
```

109 mensajes en
4 foros.

Sistemas basados en el conocimiento [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=28991
```

31 mensajes en 2 foros.

Sistemas basados en el conocimiento [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=25287
```

29 mensajes en 2 foros.

Lógica Matemática[A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=17317
```

648 mensajes en
2 foros.

Lógica Matemática[A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
prefix_forum_discussions.id
WHERE prefix_forum.course=12574
```

470 mensajes en
2 foros.

32100

Redes y Sistemas Distribuidos[A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
    prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
    prefix_forum_discussions.id
WHERE prefix_forum.course=28989
```

191 mensajes en
10 foros.

Redes y Sistemas Distribuidos[A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
    prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
    prefix_forum_discussions.id
WHERE prefix_forum.course=17354
```

118 mensajes en
9 foros.

3523

Estadística [A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
    prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
    prefix_forum_discussions.id
WHERE prefix_forum.course=28731
```

80 mensajes en
22 foros.

Estadística Analítica[A]

```
SELECT DISTINCT
prefix_forum.id,
prefix_forum.course,
prefix_forum.type,
prefix_forum.`name`,
prefix_forum.intro,
prefix_forum_discussions.forum,
prefix_forum_discussions.course,
prefix_forum_posts.discussion,
prefix_forum_posts.message
FROM
prefix_forum
INNER JOIN prefix_forum_discussions ON prefix_forum_discussions.forum =
    prefix_forum.id
INNER JOIN prefix_forum_posts ON prefix_forum_posts.discussion =
    prefix_forum_discussions.id
WHERE prefix_forum.course=22114
```

99 mensajes en 2
foros.

ANEXO C

Intentos de Selección de atributos

Opción 1

Una opción que al principio se consideró podría ser más efectiva es la de crear una nueva tabla de logs a partir de prefix_log filtrando el resultado por curso, módulo y la acción de agregar discusiones o posts Figura C.1.

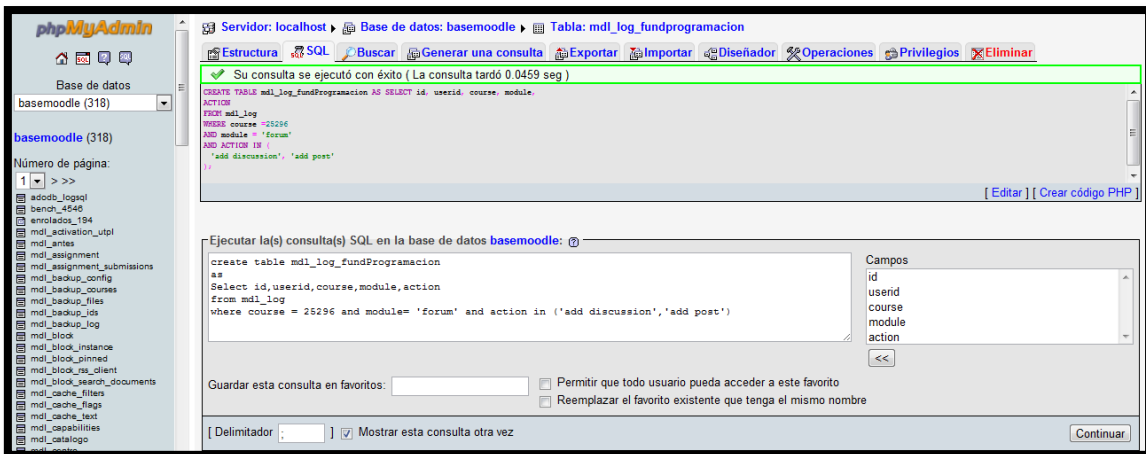


Figura C.1 Creación de tabla prefix_log_fundProgramacion a partir de prefix_log

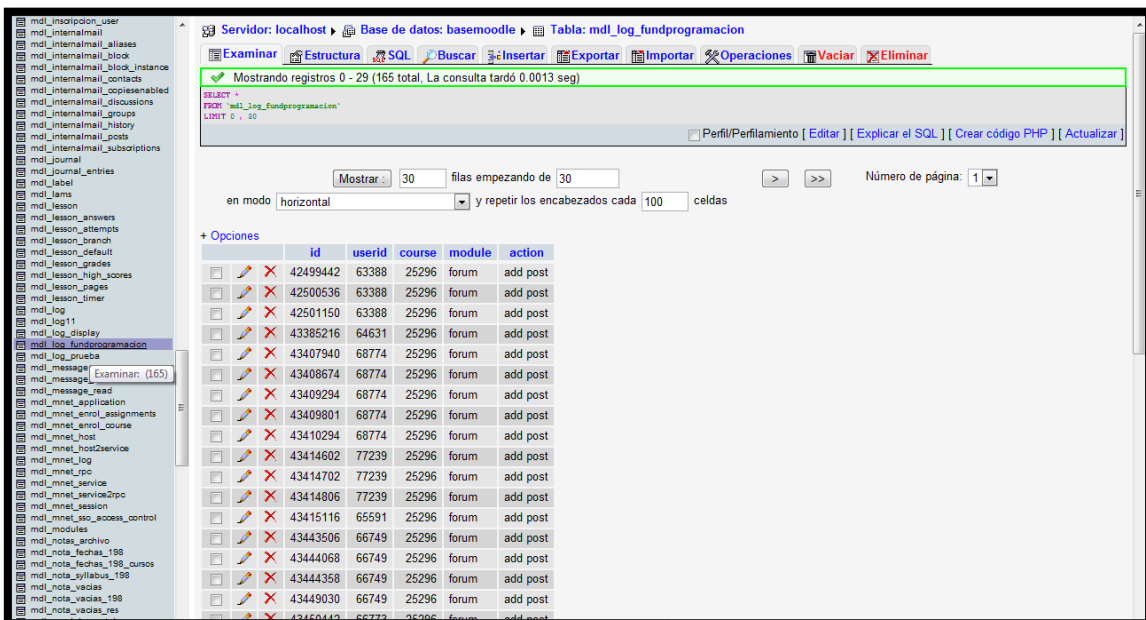


Figura C.2 Vista Previa de la tabla Generada

Ingresamos a la tabla ya creada Figura C.2 y en la tabla de Exportar seleccionamos “CSV para datos de MS Excel” Figura C.3 ¿Porqué este ítem y no “Datos CSV”?,

pues la primera genera este formato de archivo con la sintaxis compatible con WEKA los campos separados por puntos y coma“;” Figura C.4.

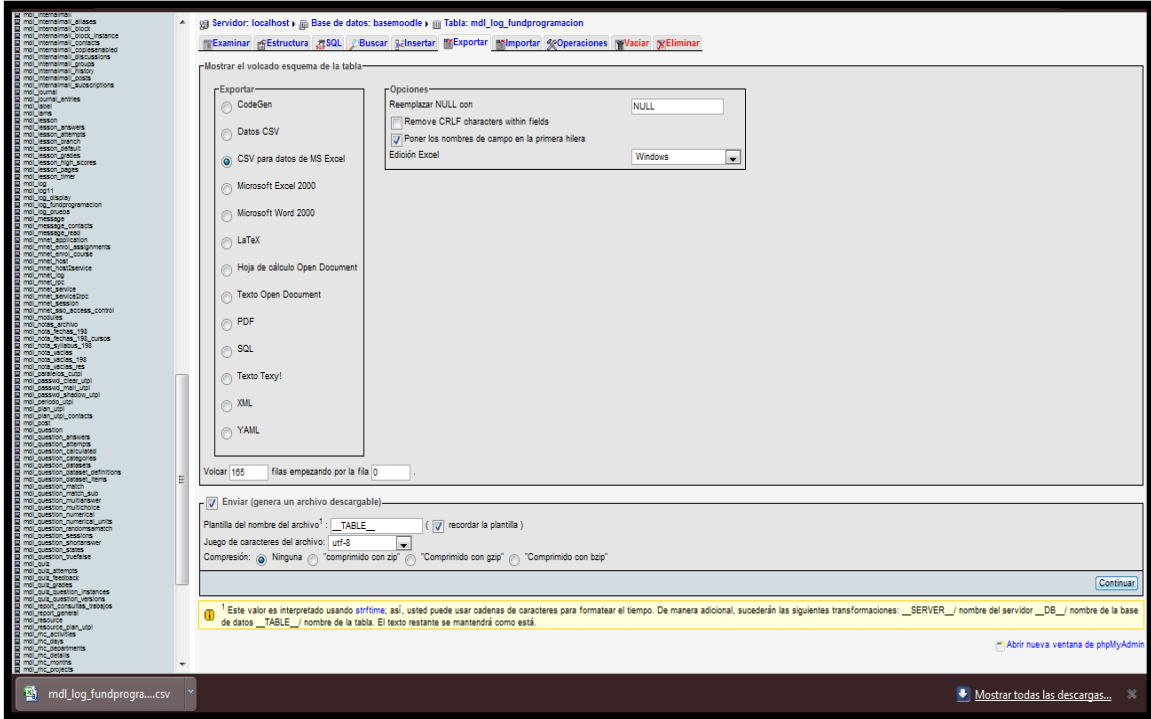


Figura C.3 Exportación de la tabla prefix_log_fundProgramacion en formato CSV

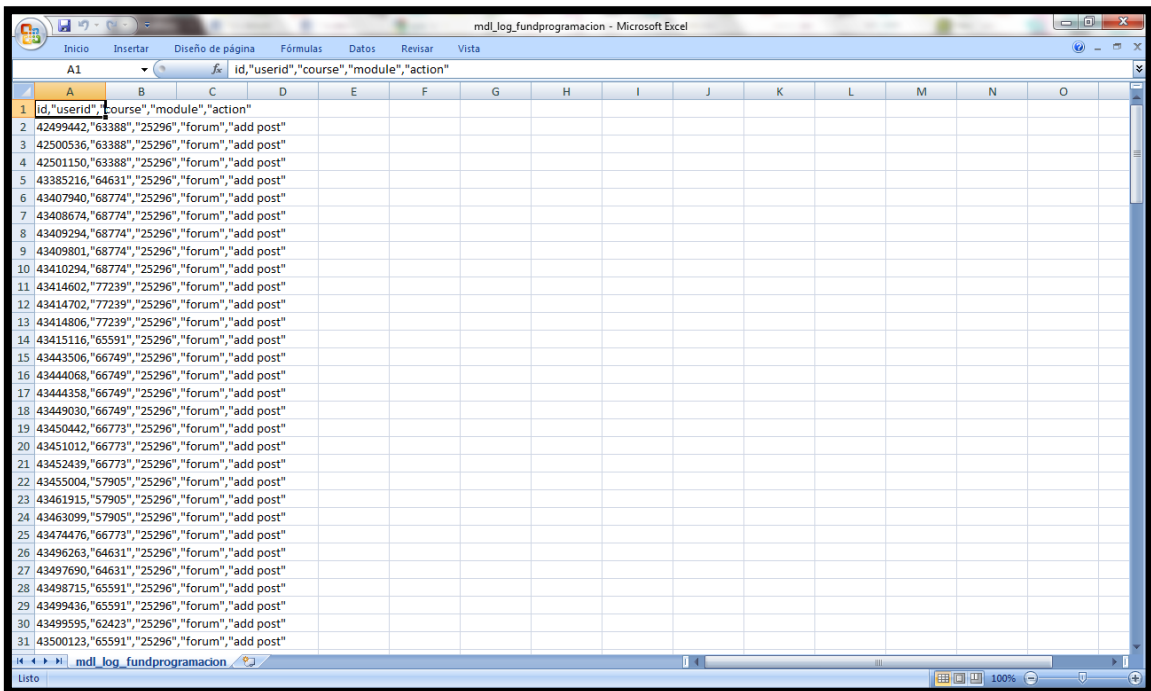
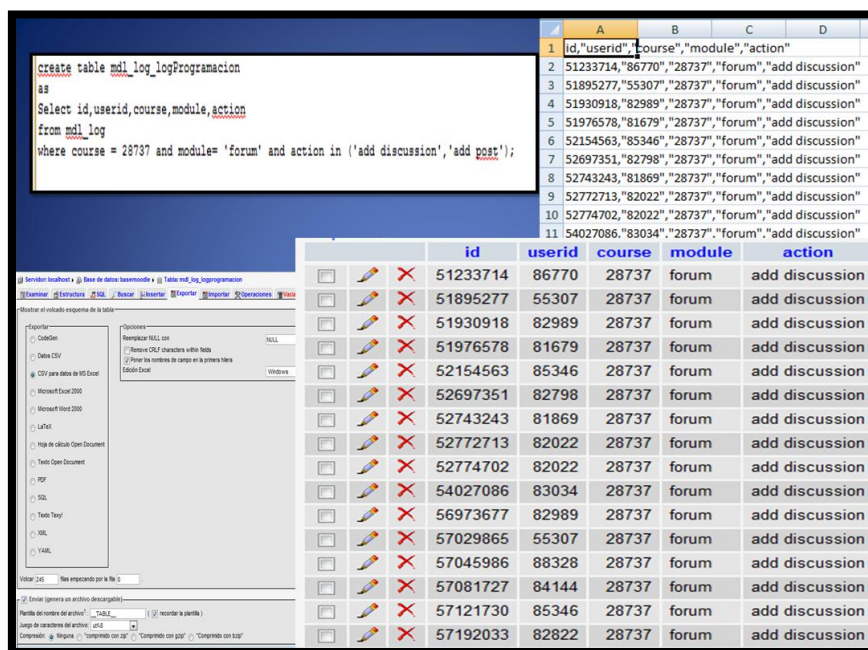


Figura C.4 Archivo CSV con los datos filtrados, tabla prefix_log_fundProgramacion

Se realiza el mismo procedimiento para la materia de Lógica de la Programación Figura C.5 y Fundamentos informáticos.



The screenshot shows a database management interface. On the left, a SQL query is entered in a text area:

```
create table mdl_log_logProgramacion
as
Select id,userid,course,module,action
from mdl_log
where course = 28737 and module= 'forum' and action in ('add discussion','add post');
```

On the right, a table of data is displayed with the following columns: id, userid, course, module, and action. The table contains 11 rows of data, all with 'add discussion' as the action.

id	userid	course	module	action
51233714	86770	28737	forum	add discussion
51895277	55307	28737	forum	add discussion
51930918	82989	28737	forum	add discussion
51976578	81679	28737	forum	add discussion
52154563	85346	28737	forum	add discussion
52697351	82798	28737	forum	add discussion
52743243	81869	28737	forum	add discussion
52772713	82022	28737	forum	add discussion
52774702	82022	28737	forum	add discussion
54027086	83034	28737	forum	add discussion
57029865	55307	28737	forum	add discussion
57045986	88328	28737	forum	add discussion
57081727	84144	28737	forum	add discussion
57121730	85346	28737	forum	add discussion
57192033	82822	28737	forum	add discussion

Figura C.5 Procedimiento creación de tabla prefix_log_logProgramacion y exportación de archivo .CSV

Estos atributos no reunían la suficiente información, no ofrecían una vista minable nos referimos con ello a atributos capaces de proporcionar información relevante, descriptiva de cómo están sucediendo los eventos, así que se tomó en cuenta una segunda opción.

Opción 2

Otra opción es la de reunir algunas tablas y escoger los atributos que provean mayor información para la minería.

La consulta base que se ha utilizado es:

```
create table prefix_log_fundProgmacion_test
as
SELECT log.id, log.userid, log.course, log.module, log.action,
foro.type,discusiones.name, foro.intro, post.message
FROM prefix_log AS log, prefix_forum AS foro, prefix_forum_posts AS post,
prefix_forum_discussions AS discusiones
WHERE log.course =28741
```

AND module = 'forum'
 AND ACTION IN (
 'add discussion', 'add post'
)
 AND post.discussion = discusiones.id
 AND discusiones.forum = foro.id
 AND log.course = foro.course

Donde log.course tomará los valores dependiendo del curso: 28741, 28737, 28739 para Fundamentos de la programación, Lógica de la Programación y Fundamentos Informáticos respectivamente. Figura C.7.

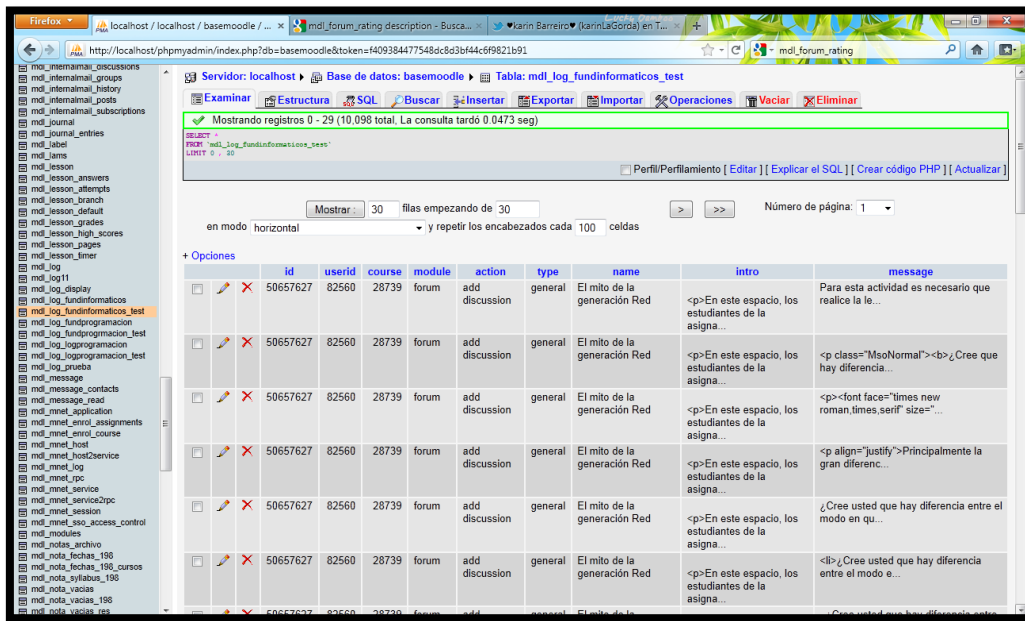


Figura C.6 Tabla prefix_log_fundProgrmacion_test

	id	userid	course	module	action	type	name	intro	message
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	Para esta actividad es necesario que realice la le...
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	<p class="MsoNormal">¿Cree que hay diferencia...
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	<p><font face="times new roman,times,serif" size="...
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	<p align="justify">Principalmente la gran diferenc...
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	¿Cree usted que hay diferencia entre el modo en qu...
<input type="checkbox"/>	50657627	82560	28739	forum	add discussion	general	El mito de la generación Red	<p>En este espacio, los estudiantes de la asigna...	¿Cree usted que hay diferencia entre el modo e...

Figura C.7 Vista Previa de la tabla Generada

ANEXO D

Obtención de atributos con características Minables y Exportación de Tablas

Se seleccionaron ocho atributos para Lógica de la Programación y Fundamentos Informáticos y once para Fundamentos de la programación (por carencia de los tres restantes en las primeras materias) solo dos se tomaron directamente de la tabla (userid, course) que los contenía, para el resto se requirió de consultas individuales para cada uno de los usuarios que forman parte de una asignatura.

Lo primero que se realiza es una consulta SQL para la obtención de todos los estudiantes inscritos en el curso. Figura D.1

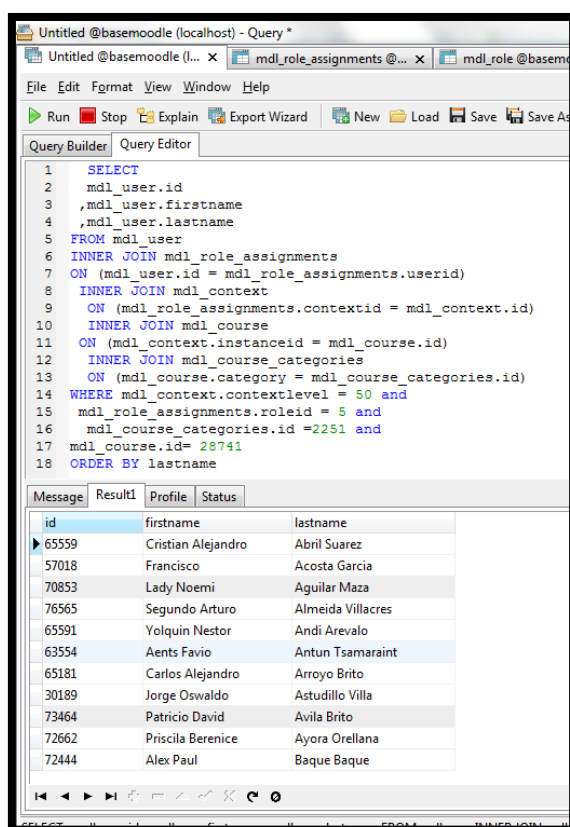


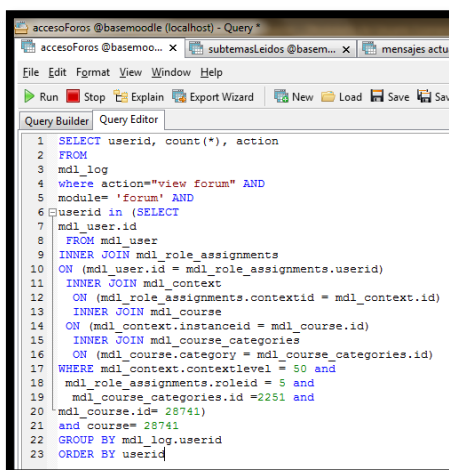
Figura D.1 Consulta para la Obtención de todos los alumnos pertenecientes a un curso.

Donde prefix_context.contextlevel = "50" hace referencia al contexto de **curso**, prefix_role_assignments.roleid= "5" indica que el rol que se enlistará será el de **profesional en formación**, en otras palabras los estudiantes.

Prefix_course.categories.id= "2251" el periodo como se ha manifestado "**Octubre-Febrero 2011**" y prefix_course.id la materia.

Se procede de la misma forma para el resto. Luego se realizará la respectiva consulta para cada uno de los atributos.

D.1. NÚMERO ACCESO A FOROS (num_acceso_foros)



```

1 SELECT userid, count(*), action
2 FROM
3 mdl_log
4 where action="view_forum" AND
5 module= 'forum' AND
6 userid in (SELECT
7 mdl_user.id
8 FROM mdl_user
9 INNER JOIN mdl_role_assignments
10 ON (mdl_user.id = mdl_role_assignments.userid)
11 INNER JOIN mdl_context
12 ON (mdl_role_assignments.contextid = mdl_context.id)
13 INNER JOIN mdl_course
14 ON (mdl_context.instanceid = mdl_course.id)
15 INNER JOIN mdl_course_categories
16 ON (mdl_course.category = mdl_course_categories.id)
17 WHERE mdl_context.contextlevel = 50 and
18 mdl_role_assignments.roleid = 5 and
19 mdl_course_categories.id =2251 and
20 mdl_course.id= 28741)
21 and course= 28741
22 GROUP BY mdl_log.userid
23 ORDER BY userid

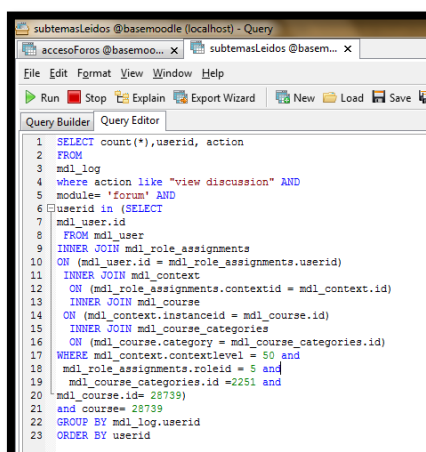
```

Figura D.2 Consulta para obtener el número de acceso a FOROS

En la Figura D.2 se presenta el valor correspondiente al número de veces que un estudiante ha accedido a los foros, se utiliza la acción “view forum” , se filtra esta consulta por curso y por id de usuario, se ha utilizado una subconsulta para obtener el número de acceso de todos los miembros de la materia.

D.3. SUBTEMAS LEÍDOS (subtemas_leidos)

Representa el número de veces que se ha leído un hilo de conversación Figura D.3



```

1 SELECT count(*),userid, action
2 FROM
3 mdl_log
4 where action like "view discussion" AND
5 module= 'forum' AND
6 userid in (SELECT
7 mdl_user.id
8 FROM mdl_user
9 INNER JOIN mdl_role_assignments
10 ON (mdl_user.id = mdl_role_assignments.userid)
11 INNER JOIN mdl_context
12 ON (mdl_role_assignments.contextid = mdl_context.id)
13 INNER JOIN mdl_course
14 ON (mdl_context.instanceid = mdl_course.id)
15 INNER JOIN mdl_course_categories
16 ON (mdl_course.category = mdl_course_categories.id)
17 WHERE mdl_context.contextlevel = 50 and
18 mdl_role_assignments.roleid = 5 and
19 mdl_course_categories.id =2251 and
20 mdl_course.id= 28739)
21 and course= 28739
22 GROUP BY mdl_log.userid
23 ORDER BY userid

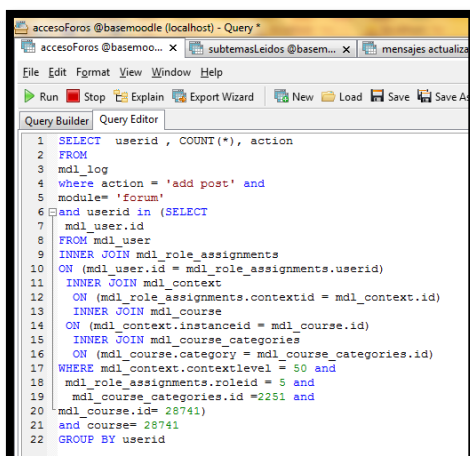
```

Figura D.3 Consulta para obtener los subtemas leídos por un estudiante.

Donde el modulo es igual a “forum” y action a “view discussion”.

D.4. MENSAJES AGREGADOS (num_respuestas_post)

Número de mensajes que un usuario ha agregado Figura D.4



```

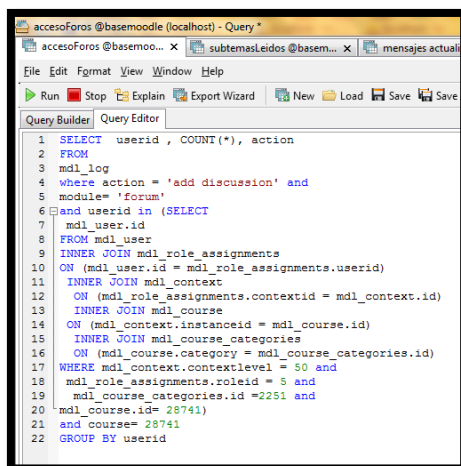
1 SELECT userid , COUNT(*) , action
2 FROM
3 mdl_log
4 where action = 'add post' and
5 module= 'forum'
6 and userid in (SELECT
7 mdl_user.id
8 FROM mdl_user
9 INNER JOIN mdl_role_assignments
10 ON (mdl_user.id = mdl_role_assignments.userid)
11 INNER JOIN mdl_context
12 ON (mdl_role_assignments.contextid = mdl_context.id)
13 INNER JOIN mdl_course
14 ON (mdl_context.instanceid = mdl_course.id)
15 INNER JOIN mdl_course_categories
16 ON (mdl_course.category = mdl_course_categories.id)
17 WHERE mdl_context.contextlevel = 50 and
18 mdl_role_assignments.roleid = 5 and
19 mdl_course_categories.id =2251 and
20 mdl_course.id= 28741)
21 and course= 28741
22 GROUP BY userid
  
```

Figura D.4 Consulta para obtener el número de mensajes que un usuario ha agregado.

Se filtra acción con “add post” y modulo igual a “forum”.

D.5. DISCUSIONES (num_respuestas_debates)

Las discusiones son hilos de conversaciones iniciados generalmente por el docente, representando también la respuesta a un mensaje de otro usuario. Figura D.5



```

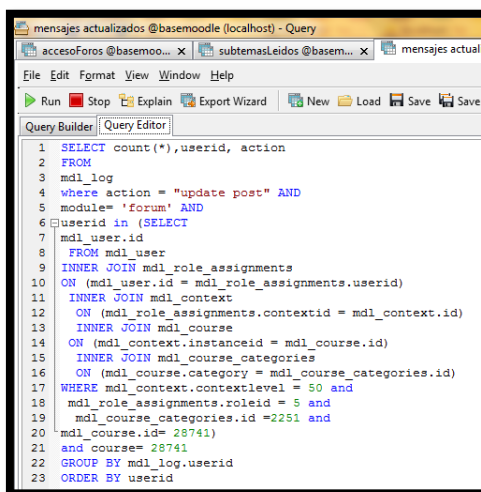
1 SELECT userid , COUNT(*) , action
2 FROM
3 mdl_log
4 where action = 'add discussion' and
5 module= 'forum'
6 and userid in (SELECT
7 mdl_user.id
8 FROM mdl_user
9 INNER JOIN mdl_role_assignments
10 ON (mdl_user.id = mdl_role_assignments.userid)
11 INNER JOIN mdl_context
12 ON (mdl_role_assignments.contextid = mdl_context.id)
13 INNER JOIN mdl_course
14 ON (mdl_context.instanceid = mdl_course.id)
15 INNER JOIN mdl_course_categories
16 ON (mdl_course.category = mdl_course_categories.id)
17 WHERE mdl_context.contextlevel = 50 and
18 mdl_role_assignments.roleid = 5 and
19 mdl_course_categories.id =2251 and
20 mdl_course.id= 28741)
21 and course= 28741
22 GROUP BY userid
  
```

Figura D.5 Recuperación del número de debates agregados.

Como se puede observar en este caso (curso: Fundamentos de la Programación) los estudiantes no presentan ninguna discusión a su haber, más si exploramos únicamente al docente, este si cuenta con 13, pero no se utilizará pues no es válido como aporte del estudiante.

D.6. NÚMERO DE MENSAJES ACTUALIZADOS (num_mens_act)

Número de mensajes que un usuario ha actualizado en un curso Figura D.5



```

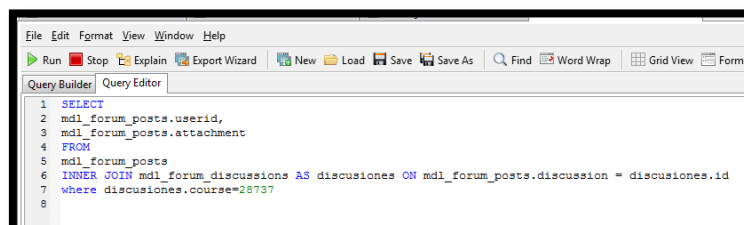
1 SELECT count(*),userid, action
2 FROM
3 mdl_log
4 where action = "update post" AND
5 module= 'forum' AND
6 userid in (SELECT
7 mdl_user.id
8 FROM mdl_user
9 INNER JOIN mdl_role_assignments
10 ON (mdl_user.id = mdl_role_assignments.userid)
11 INNER JOIN mdl_context
12 ON (mdl_role_assignments.contextid = mdl_context.id)
13 INNER JOIN mdl_course
14 ON (mdl_context.instanceid = mdl_course.id)
15 INNER JOIN mdl_course_categories
16 ON (mdl_course.category = mdl_course_categories.id)
17 WHERE mdl_context.contextlevel = 50 and
18 mdl_role_assignments.roleid = 5 and
19 mdl_course_categories.id =2251 and
20 mdl_course.id= 28741)
21 and course= 28741
22 GROUP BY mdl_log.userid
23 ORDER BY userid
  
```

Figura D.6 Consulta para obtener el número de mensajes actualizado por un usuario

Donde en action es igual a “update post”.

D.7. DATOS ADJUNTOS (arch_adjuntos)

Archivos adjuntos dentro de un mensaje. Figura D.7



```

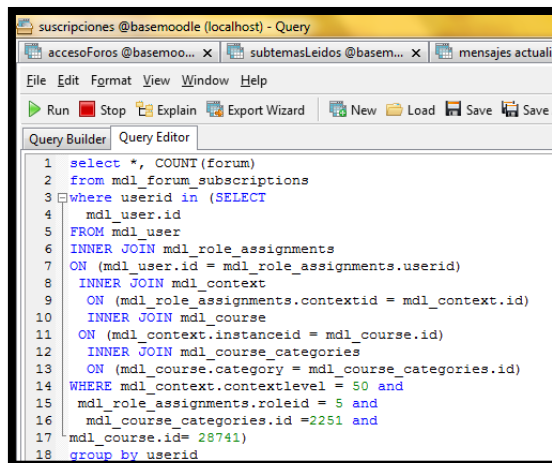
1 SELECT
2 mdl_forum_posts.userid,
3 mdl_forum_posts.attachment
4 FROM
5 mdl_forum_posts
6 INNER JOIN mdl_forum_discussions AS discusiones ON mdl_forum_posts.discussion = discusiones.id
7 where discusiones.course=28737
8
  
```

Figura D.7 Consulta para obtener el número de archivos adjuntos a un post

Discusiones.course = 28737 representa al curso, el campo mdl_forum.attachment contiene los archivos que se han agregado junto al post.

D.8. FOROS SUBSCRITOS (numForos_subscr)

Número de mensajes en los que un estudiante se ha registrado y por ende ha aceptado recibir mensajes a su correo electrónico del seguimiento de la temática planteada. Figura D.8



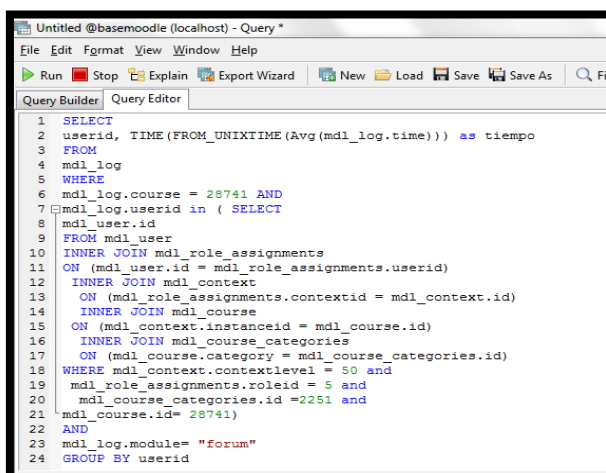
```

1 select *, COUNT(forum)
2 from mdl_forum_subscriptions
3 where userid in (SELECT
4   mdl_user.id
5 FROM mdl_user
6 INNER JOIN mdl_role_assignments
7 ON (mdl_user.id = mdl_role_assignments.userid)
8 INNER JOIN mdl_context
9 ON (mdl_role_assignments.contextid = mdl_context.id)
10 INNER JOIN mdl_course
11 ON (mdl_context.instanceid = mdl_course.id)
12 INNER JOIN mdl_course_categories
13 ON (mdl_course.category = mdl_course_categories.id)
14 WHERE mdl_context.contextlevel = 50 and
15   mdl_role_assignments.roleid = 5 and
16   mdl_course_categories.id =2251 and
17   mdl_course.id= 28741)
18 group by userid
  
```

Figura D. 8 Consulta para obtener el número de suscripciones realizadas por un usuario

D.9. PROMEDIO DE INTERACCIÓN (time_promedio)

Esta consulta requirió del registro y acceso al foro de moodle¹⁶ donde se explicaba el formato en el que se manejan las fechas este es timestamp , con FROM_UNIXTIME, obtendríamos la fecha en formato “dd:mm:yy hh:mm:ss”, en vista de que lo que se necesitaba era el número de horas se utilizó la función *TIME*, para obtener solo la hora y la función *AVG* para el promedio de las mismas. Figura D.9



```

1 SELECT
2   userid, TIME (FROM_UNIXTIME (Avg (mdl_log.time))) as tiempo
3 FROM
4   mdl_log
5 WHERE
6   mdl_log.course = 28741 AND
7   mdl_log.userid in ( SELECT
8     mdl_user.id
9 FROM mdl_user
10 INNER JOIN mdl_role_assignments
11 ON (mdl_user.id = mdl_role_assignments.userid)
12 INNER JOIN mdl_context
13 ON (mdl_role_assignments.contextid = mdl_context.id)
14 INNER JOIN mdl_course
15 ON (mdl_context.instanceid = mdl_course.id)
16 INNER JOIN mdl_course_categories
17 ON (mdl_course.category = mdl_course_categories.id)
18 WHERE mdl_context.contextlevel = 50 and
19   mdl_role_assignments.roleid = 5 and
20   mdl_course_categories.id =2251 and
21   mdl_course.id= 28741)
22 AND
23   mdl_log.module= "forum"
24 GROUP BY userid
  
```

Figura D.9 Tiempo Promedio de Interacción

¹⁶ <http://moodle.org/mod/forum>

“calif_prom_foro_1bim”, “calif_prom_foro_2bim”, “prom_foros” y “nota_final”, se obtuvieron a partir de un reporte de calificaciones en moodle, los tres primeros atributos estuvieron disponibles únicamente para la asignatura de: “Fundamentos de la Programación”.

D.6. GENERACIÓN DE REPORTE DE CALIFICACIÓN DE FOROS (PROMEDIOS) Y NOTA FINAL

Para la generación del reporte se procedió de la siguiente forma:

1. Se accede al curso en mención
2. En el panel de Administración, se ingresa a “Calificaciones”. Figura D.10
3. Se selecciona “Exportar”, en este caso a una “Hoja de Cálculo de Excel” (por su facilidad de manipulación, frente a las otras opciones) Figura D.11
4. Se señala los ítems que se requieren en el informe Figura D.12
5. Clic en “Enviar”
6. Finalmente Clic en “Descargar” Figura D.13
7. El resultado es el que se muestra en la Figura D.14



Figura D.10 Panel de Administración del Curso - "Calificaciones "

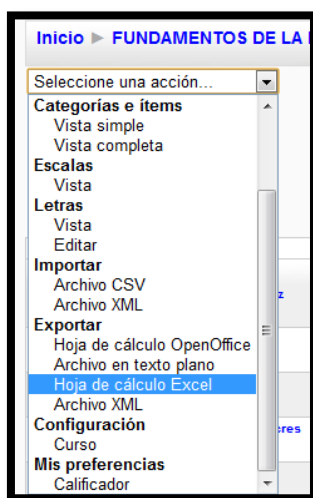


Figura D.11 Exportación de Calificaciones a Excel

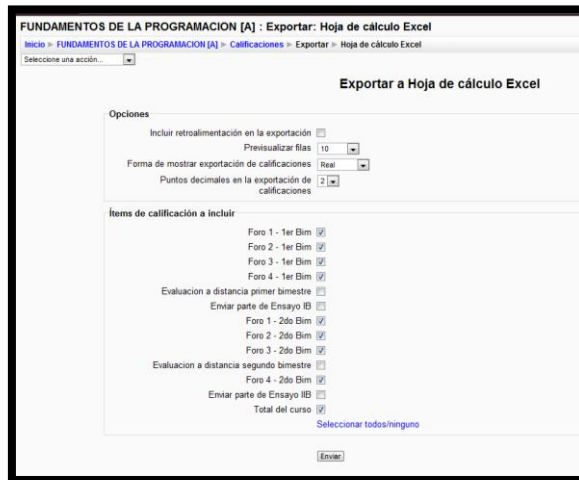
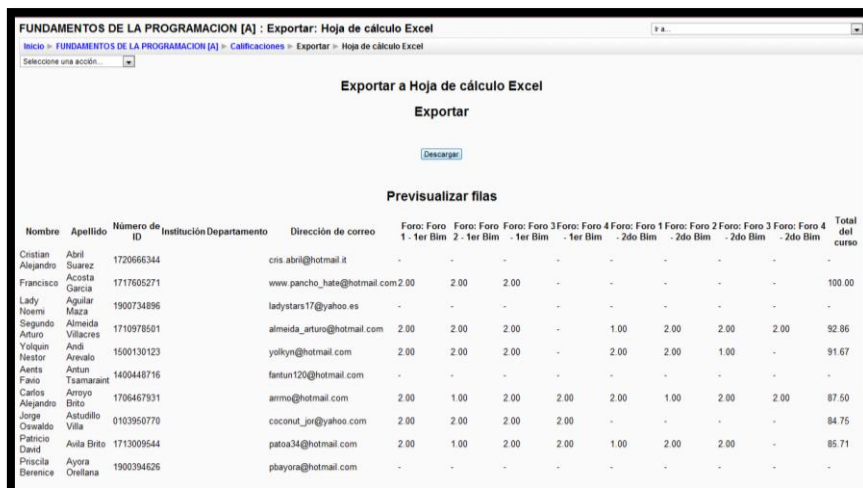


Figura D.12 Selección de Ítems para el reporte



FUNDAMENTOS DE LA PROGRAMACION [A] : Exportar: Hoja de cálculo Excel

Inicio > FUNDAMENTOS DE LA PROGRAMACION [A] > Calificaciones > Exportar > Hoja de cálculo Excel

Seleccione una acción: [v]

Exportar a Hoja de cálculo Excel

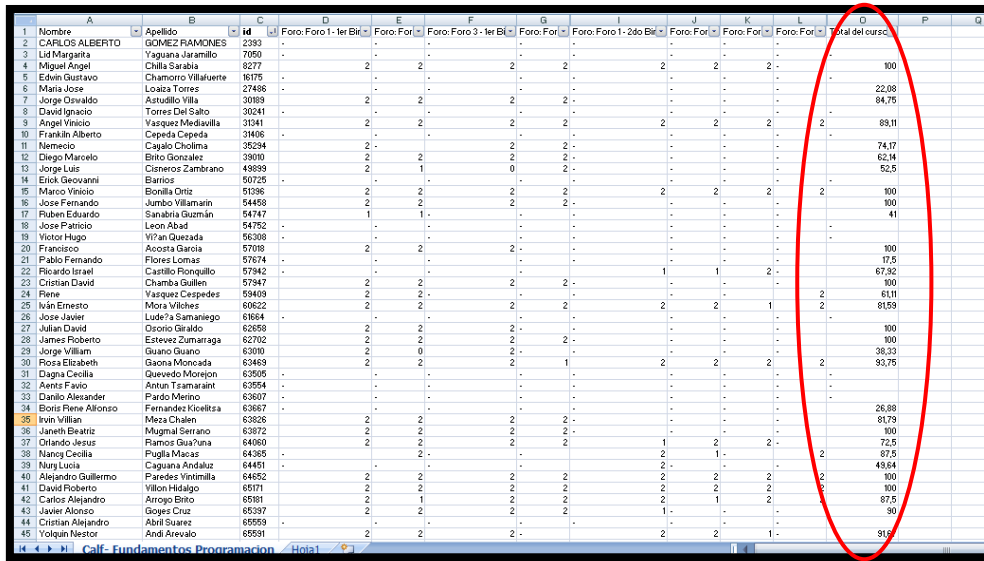
Exportar

[Descargar]

Previsualizar filas

Nombre	Apellido	Número de ID	Institución	Departamento	Dirección de correo	Foro 1 - 1er Bim	Foro 2 - 1er Bim	Foro 3 - 1er Bim	Foro 4 - 1er Bim	Foro 1 - 2do Bim	Foro 2 - 2do Bim	Foro 3 - 2do Bim	Foro 4 - 2do Bim	Total del curso
Cristian	Abri	1720666344			cris.abri@hotmail.it	-	-	-	-	-	-	-	-	-
Alejandro	Suarez	1717605271			www.pancho_hate@hotmail.com	2.00	2.00	2.00	-	-	-	-	-	100.00
Francisco	Acosta	1900734896			ladystars17@yahoo.es	-	-	-	-	-	-	-	-	-
Lady	Garcia	1710978501			almeida_arturo@hotmail.com	2.00	2.00	2.00	-	1.00	2.00	2.00	2.00	92.86
Noemi	Maza	1500130123			yolkyn@hotmail.com	2.00	2.00	2.00	-	2.00	2.00	1.00	-	91.67
Segundo	Almeida	1400448716			fantun120@hotmail.com	-	-	-	-	-	-	-	-	-
Arturo	Villacres	1706467931			armoc@hotmail.com	2.00	1.00	2.00	2.00	2.00	1.00	2.00	2.00	87.50
Yolquin	Andr	0103950770			cocconut_jor@yahoo.com	2.00	2.00	2.00	2.00	-	-	-	-	84.75
Nestor	Arevalo	1713009544			patoa34@hotmail.com	2.00	1.00	2.00	2.00	1.00	2.00	2.00	-	85.71
Alejandro	Brito	1900394626			playora@hotmail.com	-	-	-	-	-	-	-	-	-
Jorge	Astudillo													
Oswaldo	Villa													
Patricio	David													
Priscila	Ayora													
Berenice	Orellana													

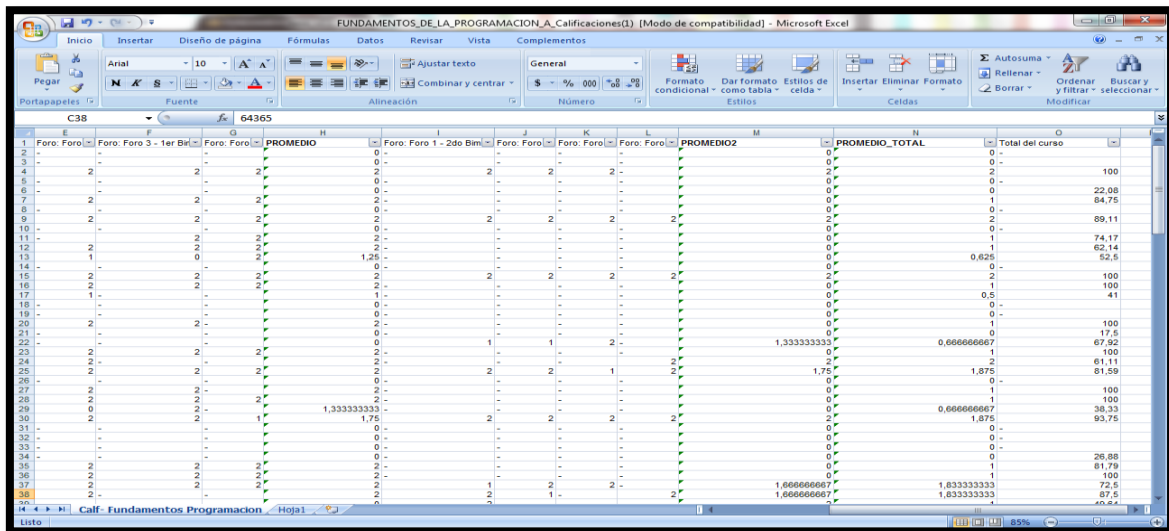
Figura D.13 Descarga de Archivo



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Nombre	Apellido	id	Foro: Foro 1- 1er Bli	Foro: For	Foro: Foro 3- 1er Bli	Foro: For	Foro: Foro 1- 2do Bli	Foro: For	Foro: For	Foro: For	Foro: For	Foro: For	Foro: For	Total del curso		
1	CARLOS ALBERTO	GOMEZ RAMONES	2393													
2	Lid Margarita	Chila Saebla	9277													
3	Miguel Angel	Yaguana Jaramillo	7090		2									100		
4	Edwin Gustavo	Chamorro Villaluz	16175							2						
5	María Jose	Loaiza Torres	27486											22,08		
6	Jorge Osvaldo	Antuñillo Villa	30989		2									84,75		
7	David Ignacio	Torres Del Salto	30241													
8	Angel Vinicio	Vasquez Mediasvilla	31341		2	2				2				89,11		
9	Franklin Alberto	Cepeda Cepeda	34006													
10	Nemesio	Cajado Ovima	35294													
11	Diego Marcelo	Brizo Gonzalez	39890		2	2				2				74,17		
12	Jorge Luis	Cisneros Zambrano	49899		2	1				2				62,14		
13	Erik Giovanni	Baticos	50125											52,5		
14	Marco Vinicio	Bonilla Ortiz	51298		2	2				2				100		
15	Jose Fernando	Jumbo Villamarin	54458		2	2				2				100		
16	Ruben Eduardo	Sanabria Guzman	54747		1	1								41		
17	Jose Francisco	Leon Abad	54792													
18	Victor Hugo	Vizán Quezada	56308													
19	Francisco	Acosta Garcia	57018		2	2								100		
20	Fabio Fernando	Flores Lomas	57674											17,5		
21	Ricardo Israel	Castillo Flonquillo	57942							1		1	2	67,92		
22	Cristian David	Chamba Guillen	57947		2	2				2				100		
23	Rene	Vasquez Cespedes	59409		2	2								61,11		
24	Ivan Ernesto	Mora Vilches	60622							2			1	81,59		
25	Jose Javier	Lude? Samaniego	61684													
26	Julian David	Osorio Giraldo	62658		2	2								100		
27	James Roberto	Estevez Zamarraga	62702		2	2				2				100		
28	Jorge William	Guano Guano	63030		2	0								38,33		
29	Rosa Elizabeth	Gaona Moncada	63469		2	2				2		2	2	93,75		
30	Diagna Cecilia	Quevedo Morejon	63505													
31	Antes Fayo	Antun Tamaraant	63594													
32	Daniel Alexander	Pardo Merino	63607													
33	Boris Rene Alfonso	Fernandez Kiehliza	63667											26,88		
34	Ivan Villan	Mesa Chaban	63626		2	2				2				81,79		
35	Janeth Beatriz	Mugmal Serrano	63972		2	2				2				100		
36	Otilando Jesus	Famos Guayana	64060		2	2				2				72,5		
37	Nancy Cecilia	Puglia Macias	64365		2	2				1		2	2	87,5		
38	Nury Lucia	Cajuna Andujar	64451											49,64		
39	Alejandro Guillermo	Paredes Vintimilla	64652		2	2				2			2	100		
40	David Roberto	Villon Hidalgo	65171		2	2				2		2	2	100		
41	Carlos Alejandro	Arango Eiro	65891		1	2				2		1	2	87,5		
42	Javier Alonso	Gomez Cruz	66397		2	2				2		1	-	90		
43	Cristian Alejandro	Abril Suarez	66569													
44	Yoliqui Nestor	Andi Arevalo	66591		2	2				2		2	1	91,6		

Figura D.14 Resultado Generación de Reporte

Como se aprecia en la Figura D.14 la columna “Total del curso” que se traduciría como “nota_final” si existe, para los atributos faltantes (promedios) se realizan cálculos sencillos, el resultado es el que se muestra en la Figura D.15



	Foro: Foro	Foro: Foro 3- 1er Bli	Foro: For	PROMEDIO	Foro: Foro 1- 2do Bli	Foro: For	Foro: For	Foro: For	PROMEDIO2	PROMEDIO_TOTAL	Total del curso
1				0					0		0
2				0					0		0
3				0					0		0
4		2		2					2		100
5				0					0		0
6		2		2					2		22,08
7				0					0		84,75
8		2		2					2		0
9		2		2					2		89,11
10				0					0		0
11		2		2					2		74,17
12		2		2					2		1
13		1		1,25					0,625		62,14
14				0					0		52,5
15		2		2					2		0
16		2		2					2		100
17		1		1					0,5		100
18				0					0		41
19		2		2					2		0
20				0					0		100
21				0		1			1,333333333		17,5
22		2		2					2		67,92
23		2		2					2		100
24		2		2					2		61,11
25		2		2					1,75		81,59
26				0					0		0
27		2		2					2		100
28		2		2					2		100
29		0		1,333333333					0,666666667		38,33
30		2		1,75					1,875		72,5
31				0					0		93,75
32				0					0		0
33				0					0		0
34				0					0		0
35		2		2					2		26,88
36		2		2					2		100
37		2		2					1,666666667		81,79
38		2		2					1,666666667		100
39				0					0		72,5
40				0					0		87,5
41				0					0		90
42				0					0		91,6

Figura D.15 Cálculo de Promedios de Foros.

D.7. De Excel a la Base de datos

Un punto aparte merece la herramienta “DREAMCODER for MYSQL” con la que se trabajó para la exportación del archivo de Excel, donde se colocaba la información obtenida de las consultas a la base de datos. Figura D.16

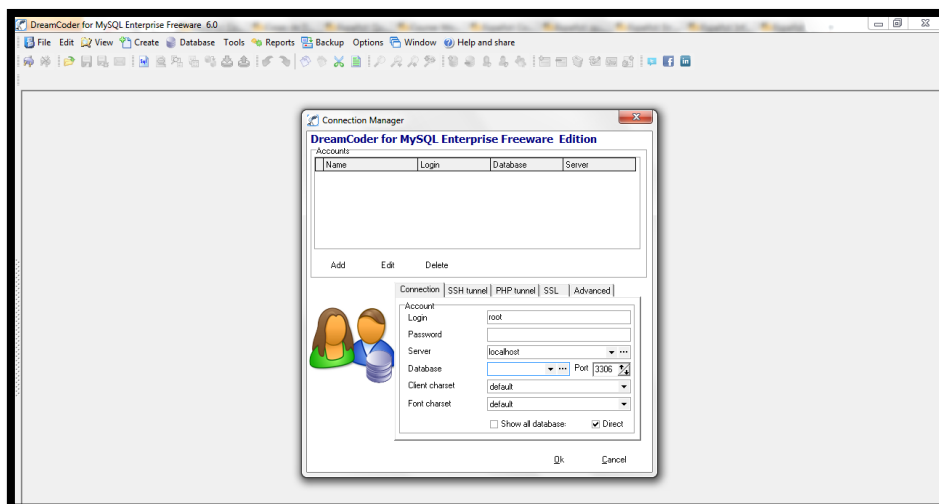


Figura D.16 Herramienta DREAMCODER for MYSQL

Se debía crear una nueva tabla con sus atributos correspondientes Figura D.17

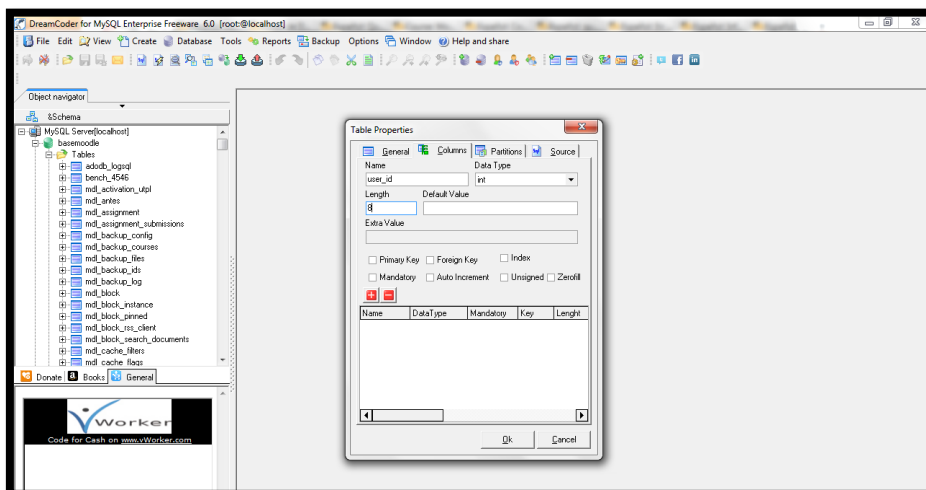


Figura D.17 Creación de atributos

Finalmente se selecciona todas las filas se copia de Excel y se pega a la herramienta Figura D.18.

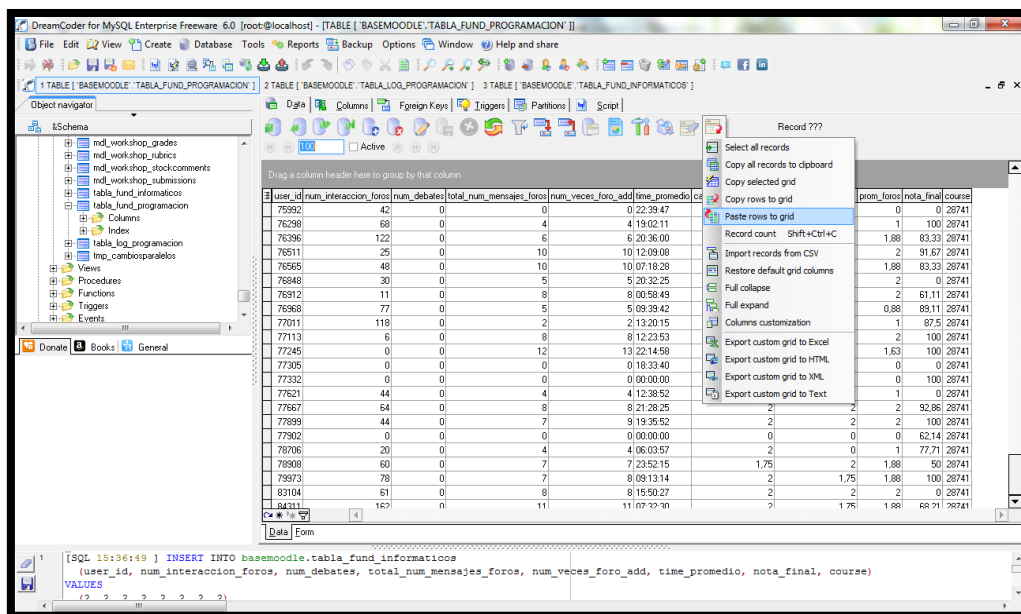


Figura D.18 Importación de tabla desde Excel

Con esto tendríamos ya a las tablas en la base de datos, el siguiente paso es “Exportarlas”

D.8. Exportación de MYSQL a formato “CSV” para WEKA

Existen algunas opciones dentro de la exportación de tablas al formato “CSV”, mas las herramientas que se han utilizado Navicat y DreamCoder no cuentan con soporte en esta extensión pero PHPMyAdmin sí, así que se trabajará con esta aunque el proceso será ligeramente más largo.

Se procede de la siguiente manera.

Se ingresa a la tabla ya creada, en la pestaña de *Exportar* seleccionamos “CSV para datos de MS Excel” ¿Porqué este ítem y no “Datos CSV”? , pues la primera genera este formato de archivo con la sintaxis compatible con WEKA los campos separados por puntos y coma “;” Figura D.19.

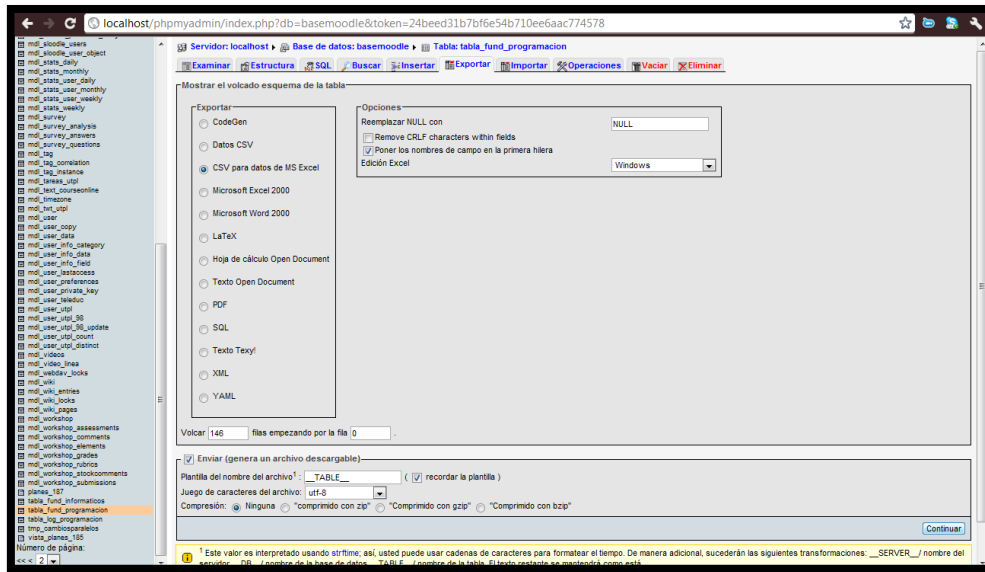


Figura D.19 Exportación de la tabla prefix_log_fund_programación en formato CSV

Se realiza el mismo procedimiento para la materia de Lógica de la Programación y Fundamentos informáticos.

Gráficas de recopilación de Atributos en una Matriz de Excel

1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
2	user_id	sexo_usi	num_acceso	subtemas_leid	num_respuestas_pc	num_respuestas	num_mens	arch_adjun	numForos_sul	time_promed	prom_hor	calif_prom	calif_pr	prom_forc	nota_final	course
2	2292	1	2	0	0	0	0	0	0	9:34:39	10	0,00	1,00	0,00	0,00	28741
3	7050	2	13	0	0	0	0	0	0	4:32:39	5	0,00	0,00	0,00	100,00	28741
4	8277	1	52	0	7	0	1	0	7	4:44:47	5	2,00	2,00	2,00	0,00	28741
5	16175	1	3	0	0	0	0	0	0	11:55:56	11	0,00	0,00	0,00	92,86	28741
6	27456	2	6	0	0	0	0	0	0	4:11:25	4	0,00	0,00	0,00	31,67	28741
7	30189	1	41	0	5	0	3	0	1	19:25:21	19	2,00	0,00	1,00	0,00	28741
8	30241	1	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	87,50	28741
9	31541	1	48	0	8	0	0	0	8	23:02:58	23	2,00	2,00	2,00	84,75	28741
10	31606	1	6	0	0	0	0	0	0	5:55:45	6	0,00	0,00	0,00	85,71	28741
11	35234	1	45	0	3	0	1	0	3	20:54:47	21	2,00	0,00	1,00	0,00	28741
12	39010	1	21	0	4	0	0	0	4	5:18:20	5	2,00	0,00	1,00	33,75	28741
13	49833	1	8	0	4	0	0	0	4	16:52:07	17	1,25	0,00	0,63	85,36	28741
14	50125	1	3	0	0	0	0	0	0	16:09:08	16	0,00	0,00	0,00	17,50	28741
15	51536	1	68	1	15	0	3	0	11	10:45:46	11	2,00	2,00	2,00	0,00	28741
16	54458	1	3	0	4	0	0	0	4	0:22:30	0	2,00	0,00	1,00	0,00	28741
17	54747	1	10	0	2	0	0	0	2	2:45:21	3	1,00	0,00	0,50	100,00	28741
18	54752	1	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	62,14	28741
19	56306	1	6	2	0	0	0	0	0	6:51:53	7	0,00	0,00	0,00	80,36	28741
20	57018	1	22	6	5	0	0	0	4	1:31:54	2	2,00	0,00	1,00	86,92	28741
21	57674	1	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	43,64	28741
22	57842	1	10	4	3	0	0	0	4	18:11:16	18	0,00	1,33	0,67	100,00	28741
23	57847	1	20	5	4	0	0	0	4	12:53:21	13	2,00	0,00	1,00	35,75	28741
24	58419	1	10	0	3	0	0	0	3	14:52:33	15	2,00	2,00	2,00	0,00	28741
25	60622	1	62	127	13	0	0	0	11	0:06:07	0	2,00	1,75	1,88	92,86	28741
26	61664	1	1	0	0	0	0	0	0	16:08:32	16	0,00	0,00	0,00	78,85	28741
27	62658	1	8	0	3	0	0	0	3	10:37:30	11	2,00	0,00	1,00	84,29	28741
28	62702	1	23	0	7	0	0	0	6	14:33:05	15	2,00	0,00	1,00	77,12	28741
29	65010	1	23	0	3	0	0	0	3	22:41:38	23	1,33	0,00	0,67	91,67	28741
30	63469	2	55	1	9	0	2	0	9	20:22:55	20	1,75	2,00	1,88	93,75	28741
31	63505	2	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	67,92	28741
32	63554	1	1	0	0	0	0	0	0	16:50:01	11	0,00	0,00	0,00	74,17	28741
33	63607	1	8	0	0	0	0	0	0	9:18:18	9	0,00	0,00	0,00	0,00	28741
34	63667	1	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	100,00	28741
35	63826	1	21	0	4	0	3	0	4	1:03:07	1	2,00	0,00	1,00	100,00	28741
36	63872	2	34	0	5	0	0	0	4	14:51:20	15	2,00	0,00	1,00	100,00	28741
37	64060	1	24	0	7	0	0	0	7	11:33:01	17	2,00	1,57	1,83	0,00	28741
38	64365	2	42	0	11	0	6	0	11	16:28:37	16	2,00	1,67	1,83	83,33	28741
39	64451	2	10	0	1	0	0	0	1	8:54:35	9	0,00	2,00	1,00	100,00	28741
40	64652	1	43	0	8	0	0	0	8	9:12:39	9	2,00	2,00	2,00	52,50	28741
41	65171	1	39	0	7	0	0	0	8	20:43:50	21	2,00	2,00	2,00	83,62	28741
42	65181	1	57	2	10	0	8	0	3	11:18:40	11	1,75	1,75	1,75	0,00	28741
43	65397	1	29	1	9	0	1	0	6	8:53:40	10	2,00	1,00	1,50	100,00	28741
44	65553	1	0	0	0	0	0	0	0	0:00:00	0	0,00	0,00	0,00	0,00	28741
45	65591	1	40	0	0	0	1	0	8	19:30:39	20	2,00	1,67	1,83	0,00	28741
46	65929	1	32	6	0	0	0	0	7	0:38:20	1	2,00	1,67	1,83	83,33	28741
47	66392	2	76	3	9	0	0	0	3	12:38:11	13	2,00	1,50	1,75	86,82	28741
48	66438	1	0	0	0	0	0	0	0	2:43:47	3	0,00	0,00	0,00	0,00	28741
49	67545	1	17	0	5	0	0	0	5	3:28:14	3	1,75	0,00	0,88	100,00	28741

Figura E.1 Matriz en EXCEL, recopilación de atributos, Fundamentos de la Programación.

1	A	B	C	D	E	F	G	H	I	J	K	L	M
2	user_id	sexo_usi	num_acceso	subtemas_leid	num_respuestas	num_respuestas	num_mens_act	arch_adj	numForos_sul	time_promed	prom_horas	nota_final	course
2	251	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
3	3706	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
4	14750	0	0	0	0	0	0	0	0	0:00:00	0	0	28737
5	16881	1	5	0	1	0	0	1	0	1:45:13	2	0	28737
6	26120	1	1	0	0	0	0	0	0	23:03:12	23	0	28737
7	30241	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
8	30877	1	4	2	0	0	0	0	0	23:55:41	24	7,39	28737
9	31408	1	10	0	1	0	0	0	0	11:23:23	11	0,00	28737
10	43859	1	17	0	7	2	0	0	0	20:08:01	20	54,25	28737
11	46479	1	32	13	3	1	0	0	2	23:26:14	23	94,6	28737
12	46987	1	1	0	0	0	0	0	0	18:05:18	18	0	28737
13	47686	1	6	3	1	0	0	0	1	12:13:53	12	85,19	28737
14	47732	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
15	48747	1	7	8	1	0	0	0	1	21:22:12	21	93,5	28737
16	49839	1	0	0	0	0	0	0	0	0:00:00	0	24,67	28737
17	50229	1	39	0	0	0	0	0	0	0:00:00	0	0	28737
18	55307	1	42	9	4	2	1	0	4	20:28:31	20	89,85	28737
19	57949	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
20	58465	1	6	7	1	1	0	0	1	8:39:14	9	0	28737
21	62149	1	0	0	0	0	0	0	0	0:00:00	0	17,83	28737
22	63808	1	12	6	3	0	0	0	1	0:16:29	0	91,5	28737
23	63826	1	3	0	0	0	0	0	0	19:14:41	19	45,33	28737
24	65781	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
25	67625	2	1	1	0	0	0	0	0	18:35:02	19	0	28737
26	68357	1	32	21	5	0	2	0	3	12:46:22	13	99,6	28737
27	71477	1	9	3	0	0	0	0	0	9:17:33	9	96,3	28737
28	71526	1	3	1	0	0	0	0	0	17:33:57	18	100	28737
29	71597	1	50	22	3	1	2	0	3	18:31:49	19	94,85	28737
30	72123	2	5	2	2	0	0	0	2	21:44:29	22	0	28737
31	72444	1	22	7	2	0	0	0	1	11:00:13	11	95,8	28737
32	72473	1	0	0	0	0	0	0	0	0:00:00	0	77,06	28737
33	72745	1	11	1	2	0	0	0	0	21:49:46	22	92,5	28737
34	73045	1	7	2	0	0	0	0	0	21:56:12	22	0	28737
35	73059	1	28	6	3	0	0	0	2	19:06:20	19	93,75	28737
36	73609	1	9	3	1	0	0	0	1	1:14:20	1	68,75	28737
37	73994	1	12	0	0	0	0	0	1	9:01:28	9	67,64	28737
38	74006	1	0	0	0	0	0	0	0	0:00:00	0	0	28737
39	74051	1	8	2	0	0	0	0	0	20:26:21	20	100	28737
40	75155	1	9	5	0	0	0	0	0	19:02:14	19	94,5	28737

Figura E.2 Matriz en EXCEL, recopilación de atributos, Lógica de la Programación.

	A	B	C	D	E	F	G	H	I	J	K	L	M
	user_ic	sexo	num_acceso_for	subtemas_leidos	num_respuestas	num_respuestas	num_mensajes	arch_adj	numForos	time_promedio	prom_horas	nota_final	courses
2	28137	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
3	34419	1	0	0	0	0	0	0	0	0.00.00	0	70	28739
4	38330	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
5	45478	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
6	49912	1	0	0	0	0	0	0	0	6:36:16	7	0	28739
7	54938	1	1	1	0	0	0	0	0	11:36:13	12	90	28739
9	58534	1	2	0	0	0	0	0	0	3:12:40	3	0	28739
10	61184	1	21	14	4	0	0	2	4	7:30:25	8	87.5	28739
11	61487	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
12	64628	1	4	2	1	0	0	0	1	17:13:29	17	0	28739
13	66354	2	0	0	0	0	0	0	0	0.00.00	0	0	28739
14	68384	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
15	68572	1	0	0	0	0	0	0	0	0.00.00	0	49.17	28739
16	72254	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
17	72463	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
18	72628	2	4	1	0	0	0	0	0	12:10:50	12	0	28739
19	73998	2	0	0	0	0	0	0	0	0.00.00	0	0	28739
20	74244	1	5	1	0	0	0	0	0	18:58:07	19	0	28739
21	74817	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
22	75205	2	1	0	0	0	0	0	0	18:12:11	18	0	28739
23	75768	1	0	0	0	0	0	0	0	0.00.00	0	63.75	28739
24	77184	1	4	4	3	0	0	0	2	16:10:09	16	0	28739
25	77620	1	3	0	0	0	0	0	0	17:07:41	17	72.5	28739
26	79973	1	2	3	1	0	0	0	1	13:00:04	13	87.5	28739
27	80549	1	14	0	3	0	0	0	3	4:40:54	5	0	28739
29	80640	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
30	80652	1	10	7	1	0	0	0	1	19:30:53	20	0	28739
31	80793	2	0	0	0	0	0	0	0	0.00.00	0	0	28739
32	80848	2	0	0	0	0	0	0	0	0.00.00	0	0	28739
33	80915	2	0	0	0	0	0	0	0	12:45:11	13	0	28739
34	80918	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
35	80931	1	8	6	0	0	0	0	0	1:40:10	2	0	28739
37	81014	1	9	8	1	0	0	0	1	22:51:00	23	0	28739
38	81055	1	0	0	0	0	0	0	0	0.00.00	0	0	28739
40	81244	1	16	21	5	0	0	0	3	1:39:47	2	0	28739
41	81256	1	1	1	0	0	0	0	0	20:04:53	20	0	28739
42	81260	1	2	0	0	0	0	0	0	8:48:24	9	0	28739
43	81330	1	10	8	2	0	0	0	1	18:08:32	18	80	28739

Figura E.3 Matriz en EXCEL recopilación de atributos, Fundamentos Informáticos.

ANEXO F

Carga de datos en WEKA desde la Base de datos

Se descarga y copia **mysql-connector-java-5.0.8-bin** a la carpeta raíz de WEKA. Figura F.1 y establecer la ruta de conexión en la aplicación Figura F.2, en este caso la base de prueba se encuentra alojada localmente y tiene por nombre “basemoodle”

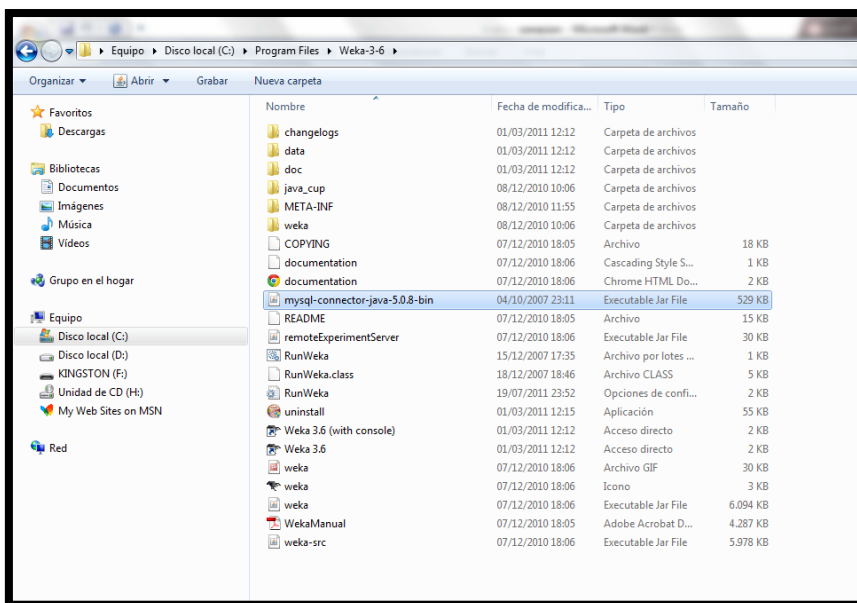


Figura F. 1 Path MySQL Connector

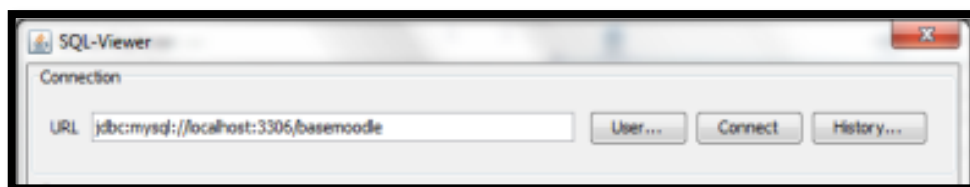


Figura F. 2 URL Base de Datos

Se Ingresa el usuario y password y Clic en “OK”. Figura F.3.

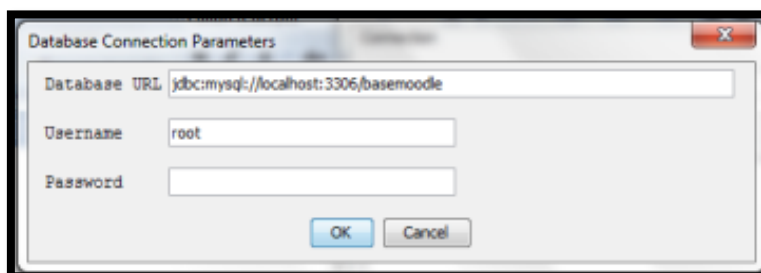


Figura F. 3 Parámetros de Conexión con la Base de Datos

Clic en *Connect* para realizar la conexión, se observará en la sección Info, que efectivamente se ha realizado. Figura F.4

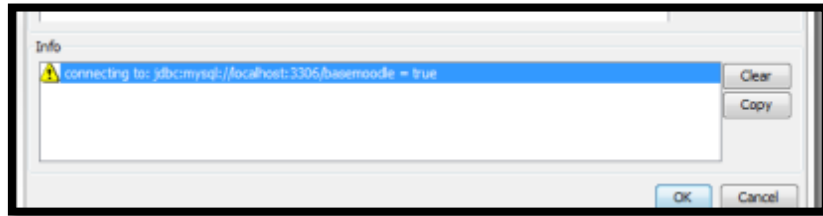


Figura F. 4 Informe de Conexión

En aquella ventana se puede colocar la consulta de acuerdo a los datos que vayamos a utilizar. Se ha decidido colocar a modo de prueba, una consulta que me traiga todos los valores de la tabla de *logs*. Figura F.5

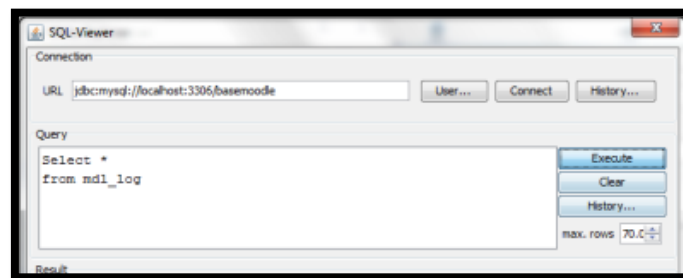


Figura F. 5 Búsqueda General del Registro de Logs.

Lamentablemente como era de esperarse el tamaño de la tabla de logs, excede el permitido. Figura F.6. Se tienen dos opciones: Cambiar el archivo *runWeka.jar* y asignarle un rango mayor en el campo *maxheap* Figura F.7 (no muy factible) o realizar una consulta con filtros que permitan obtener solamente la información que se necesitará. Figura F.8

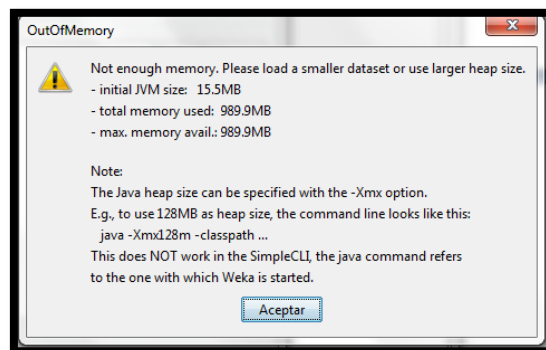


Figura F. 6 Mensaje de error por tamaño de Memoria

```

# Version $Revision: 1.3 $

# setups (prefixed with "cmd ")
cmd_default=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap# -
classpath "#wekajar#;#cp#" #mainclass#
cmd_console=cmd.exe /K start cmd.exe /K "java -
Dfile.encoding=#fileEncoding# -Xmx#maxheap# -classpath
\"#wekajar#;#cp#\" #mainclass#"
cmd_explorer=java -Dfile.encoding=#fileEncoding# -Xmx#maxheap#
-classpath "#wekajar#;#cp#" weka.gui.explorer.Explorer

# placeholders ("bla" in command gets replaced with content of
key "bla")
# Note: "#wekajar#" gets replaced by the launcher class, since
that jar gets
# provided as parameter
#maxheap=1024#
# The MDI GUI
#mainclass=weka.gui.Main
# The GUIChooser
mainclass=weka.gui.GUIChooser
# The file encoding; use "utf-8" instead of "Cp1252" to display
UTF-8 characters in the
# GUI, e.g., the Explorer
fileEncoding=Cp1252

# The classpath placeholder. Add any environment variables or
jars to it that
# you need for your Weka environment.
# Example with an environment variable (e.g., THIRD_PARTY_LIBS):
# cp=%CLASSPATH%;%THIRD_PARTY_LIBS%
# Example with an extra jar (located at D:\libraries

```

Figura F. 7 Archivo RunWeka, edición campo MaxHeap

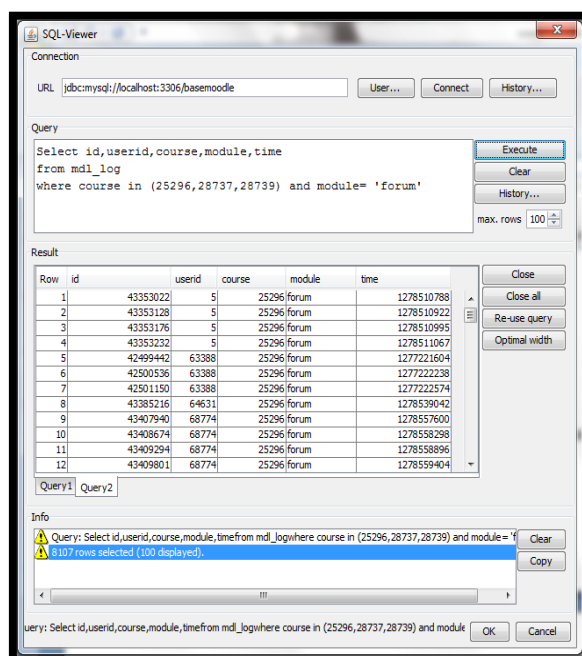


Figura F. 8 Consulta tabla Logs con Filtros

Rescatando de la consulta de la Figura F.8 los cursos con los que se trabajará y el módulo de Foros con el cual se realizará el análisis.

Al parecer esto era todo lo que se requería para la carga de datos pero los atributos dispuestos pertenecían a un tipo de dato que no se podía cargar Figura F.9, se dieron problemas con el tipo de valor INT/INTEGER, aparentemente se debían realizar cambios en el archivo DatabaseUtils.props, gestión que se realizó tanto en la zona de tipos de archivos como el restablecimiento de jdbcDriver como jdbcURL Figura F.10, pero aun así persistió el error, lamentablemente no existe mucha información acerca

de este error en particular, por lo que finalmente se ha decidido realizar el estudio con archivos CSV.

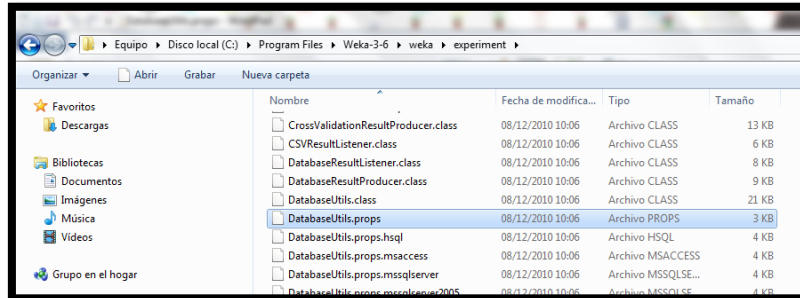


Figura F. 9 Ruta DatabaseUtils.props

```
# General information on database access can be found here:
# http://weka.wikispaces.com/Databases
#
# Version: $Revision: 5836 $

# The comma-separated list of jdbc drivers to use
#jdbcDriver=RmiJdbc.RJDriver,jdbc.idbDriver
#jdbcDriver=jdbc.idbDriver
#jdbcDriver=RmiJdbc.RJDriver,jdbc.idbDriver,org.gjt.mm.mysql.Dri
#jdbcDriver=com.mckoi.JDBCdriver,org.hsqldb.jdbcDriver
#jdbcDriver=com.mysql.jdbc.Driver
#jdbcDriver=org.gjt.mm.mysql.Driver

# The url to the experiment database
#jdbcURL=jdbc:rdmi://expserver/jdbc:idb=experiments.prp
#jdbcURL=jdbc:idb=experiments.prp
#jdbcURL=jdbc:mysql://localhost:3306/basemoodle

# the method that is used to retrieve values from the db
# (java datatype + ResultSet.<method>)
string, getString() = 0; --> nominal
boolean, getBoolean() = 1; --> nominal
double, getDouble() = 2; --> numeric
byte, getByte() = 3; --> numeric
short, getShort() = 4; --> numeric
int, getInteger() = 5; --> numeric
long, getLong() = 6; --> numeric
float, getFloat() = 7; --> numeric
date, getDate() = 8; --> date
text, getString() = 9; --> string
time, getTime() = 10; --> date
# the original conversion: <column type>=<conversion>
```

Figura F. 10 Edición archivo DataBaseUtils.props

ANEXO G

Carga de datos en WEKA definitiva

En la fase de minería lo esencial es la prueba de los datos ya obtenidos con un algoritmo de Inteligencia artificial como se ha mencionado será el K-MEANS (clustering).

Los pasos a seguir son los siguientes.

1. Consiste en seleccionar los archivos .CSV que se exportaron desde la base de datos, uno a uno. Figura G3.
2. Se selecciona la pestaña “Cluster” Figura G3.
3. Clic en “SimpleKMeans”, EM ,
4. Seleccionamos el número de cluster, en este caso 3 (Alto, Medio y Bajo) Figura G3.
5. Clic en Start.

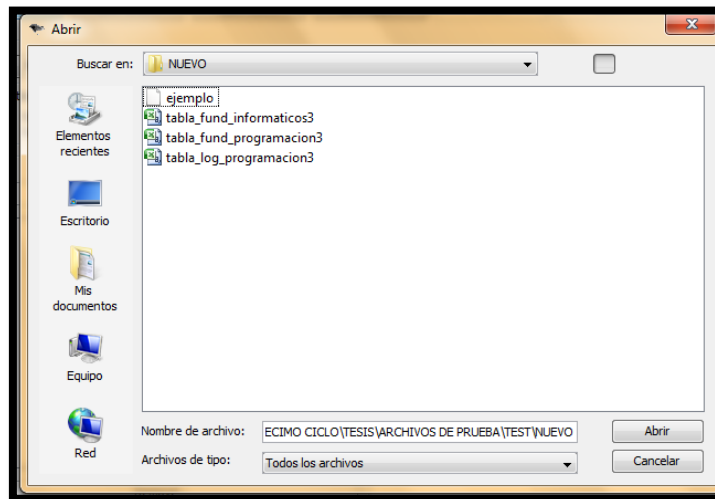


Figura G.1 Selección de archivos CSV

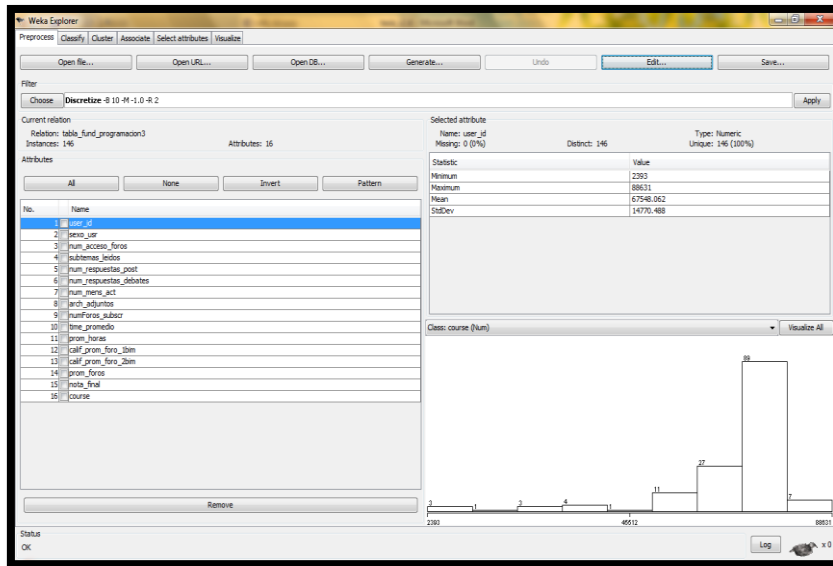


Figura G.2 Vista WEKA- Selección de Pestaña Cluster

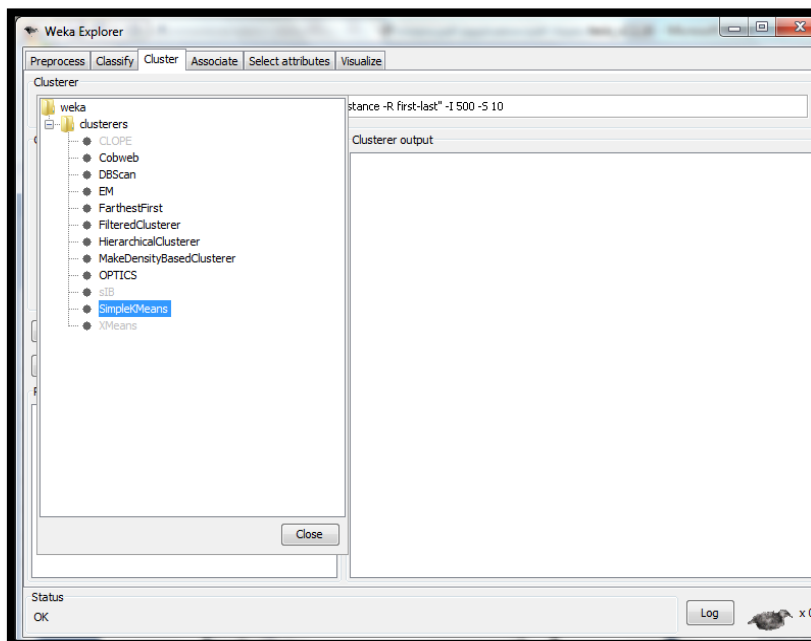


Figura G.3 Selección de Algoritmo K-MEANS (Clustering)

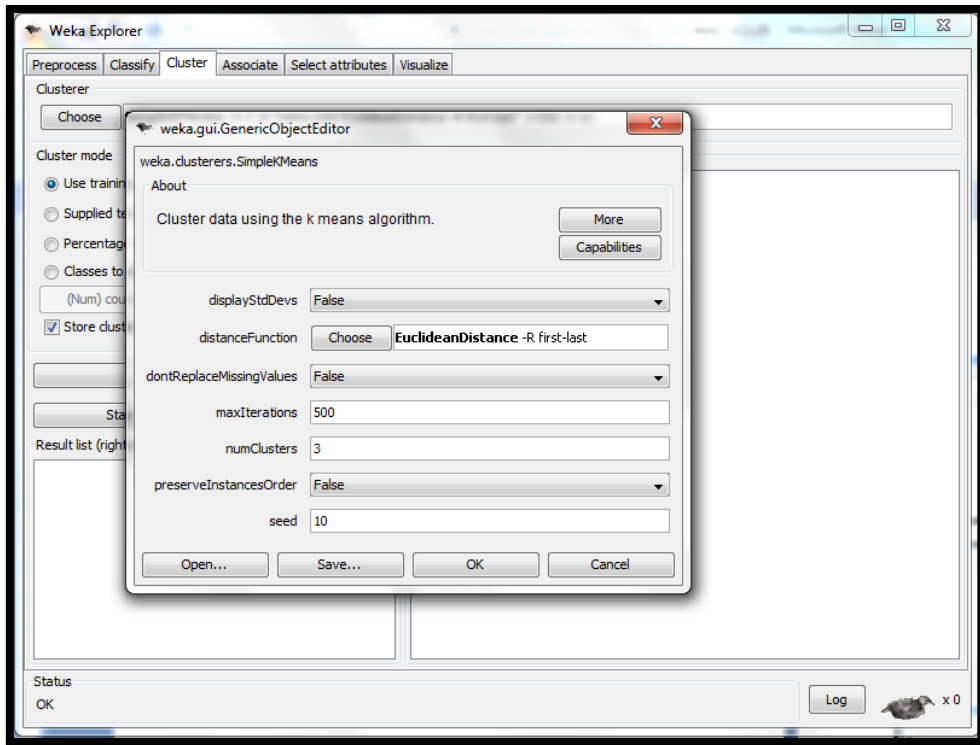


Figura G.4 Propiedades del algoritmo, Selección Número de Clusters

ANEXO H

Conversión del atributo time de “hh:mm:ss” a segundos y luego a horas

Se agregó para el efecto una nueva columna “segundos”, para la segunda experimentación se transformó los segundos a horas usando la fórmula de la Figura H.1

`=(HORA(J2)*1)+(MINUTO(J2)/60)+SEGUNDO(J2)/3600`

Figura H.1 Fórmula Excel para la Conversión de horas formato HH:MM:SS a horas

Resultando tal como se muestra en la Figura H.2

J	K
0:00:00	0
18:12:11	18
0:00:00	0
16:10:09	16
17:07:41	17
13:00:04	13
4:40:54	5
0:00:00	0
19:30:53	20
0:00:00	0
0:00:00	0
12:45:11	13
0:00:00	0
1:40:10	2
22:51:00	23
0:00:00	0
1:39:47	2
20:04:53	20
8:48:24	9
18:08:32	18
20:16:53	20
9:25:05	9
0:00:00	0
0:00:00	0
18:12:11	18

Figura H.2 Vista Previa Resultado de Conversión

Luego de esto se procede a cargar en la base de datos tal como se ha indicado en los anteriores anexos.

ANEXO I

Clementine 13.0 , creación de Diagramas de Flujo o Stream para ejecución del algoritmo K-MEANS

1. Se escoge el medio por el cual se va a cargar la data, bien sea desde la base de datos o por medio de un archivo, se seleccionó la segunda opción.
2. Doble clic en el ícono de archivo, se abrirá una ventana donde escogeremos el origen del mismo. Figura I.1

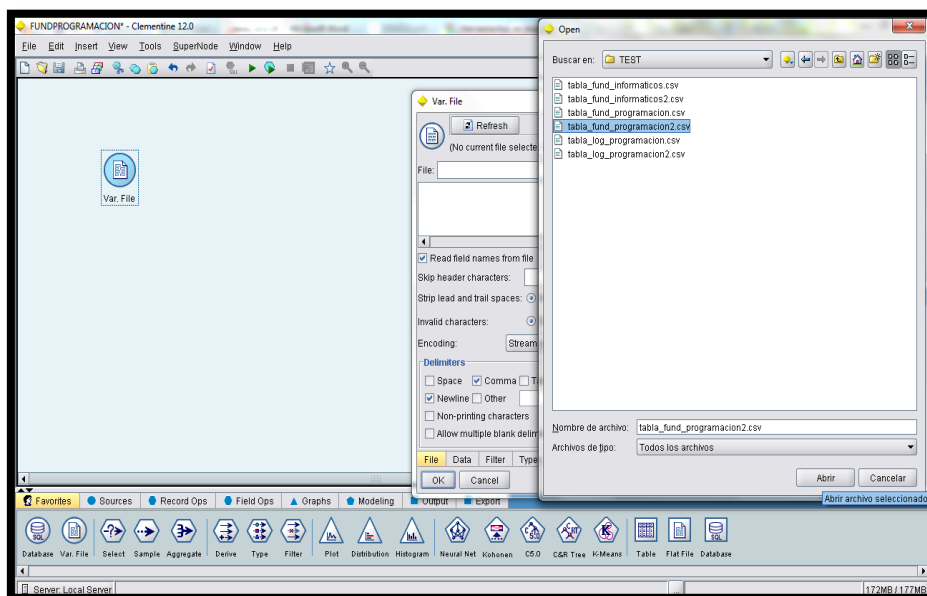


Figura I.1 Origen de Datos Herramienta Clementine.

3. En la pestaña “Filter” de la misma ventana se seleccionaran los atributos con los que se han trabajado en los anteriores experimentos (igualdad de condiciones). Figura I.2

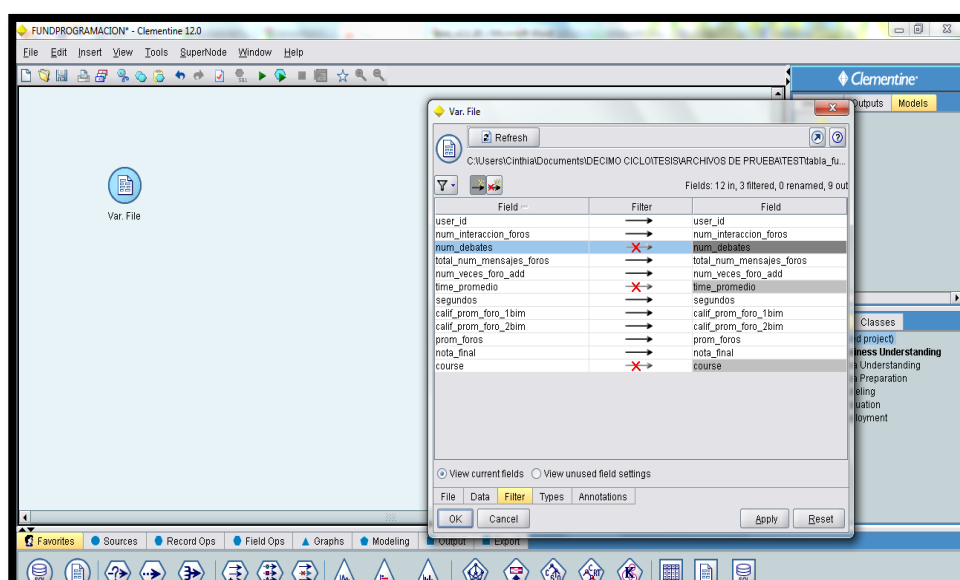


Figura I.2 Filtro de Variables

4. Aplicar y Ok.

5. El siguiente paso consiste en verificar el tipo de dato que se va a manejar, se arrastra para ello el ícono “type” y se conecta el ícono de origen con este.
6. Observamos que la data contenida este en el correcto tipo de dato y de ser necesario se harán cambios, sino aceptamos. Figura I.3

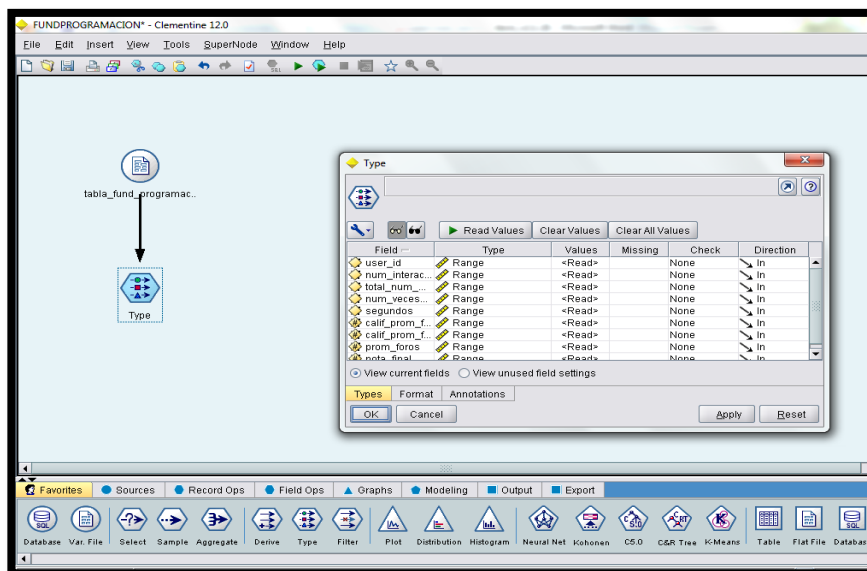


Figura I.3 Verificación del tipo de dato del Origen

7. Se agrega el ícono K-MEANS, “type” se conecta a este.
8. Doble Clic para ver sus propiedades, en esta ventana se escogerán el número de cluster que deseamos que se divida el conjunto, Figura I.4

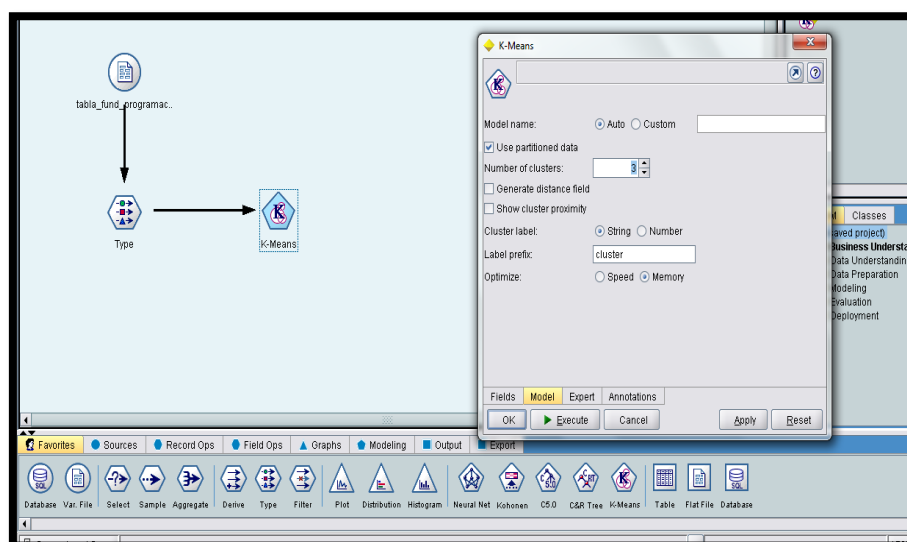


Figura I.4 Ajuste y Conexión del ícono del Algoritmo K-MEANS

9. Clic en Ejecutar.
10. Inmediatamente en el lado derecho de la ventana de trabajo aparecerán los resultados de este proceso en forma de ícono.
11. Clic derecho “Browse” para su revisión Figura I.5

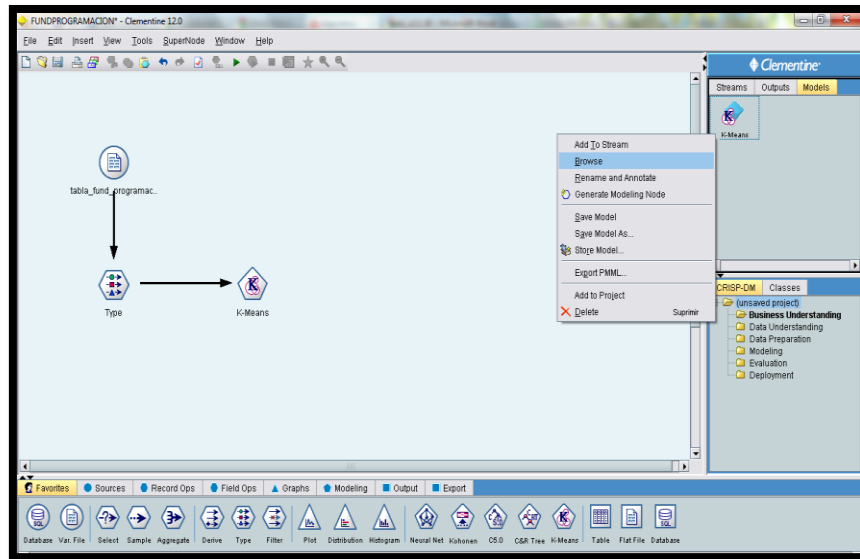


Figura I.5 Verificación de Resultados

12. Existen 3 vistas en la nueva ventana, “Model” Figura I.6 que muestra la desviación estándar de cada atributo de cada cluster, “Viewer” Figura I.7 que a modo gráfico da una vista general de la formación de cluster y sus atributos y “Summary” Figura I.8 información del proyecto, número de iteraciones y errores encontrados en la experimentación.



Figura I.6 Pestaña "Model", desviación estándar

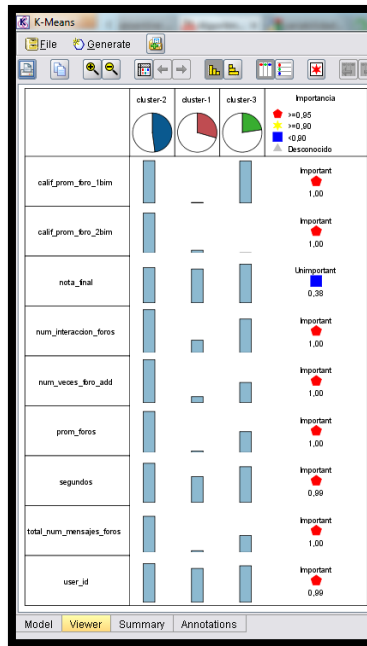


Figura I.7 Pestaña "Viewer", gráfica de atributos y Clusters

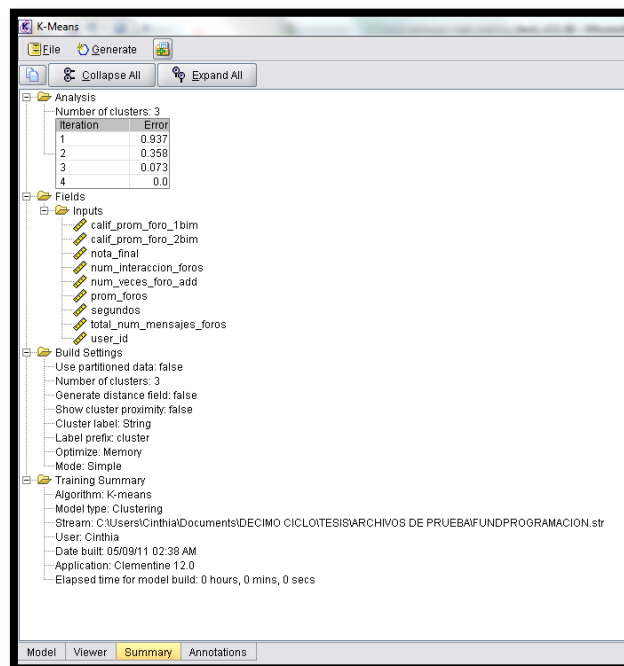


Figura I.8 Sumario del Proyecto.

13. Se agrega el resultado de la experimentación al marco de trabajo y se conecta con "type", a esta se agrega una tabla para ver los resultados. Figura I.9
14. Clic en ejecutar y se analiza la tabla Figura I.10

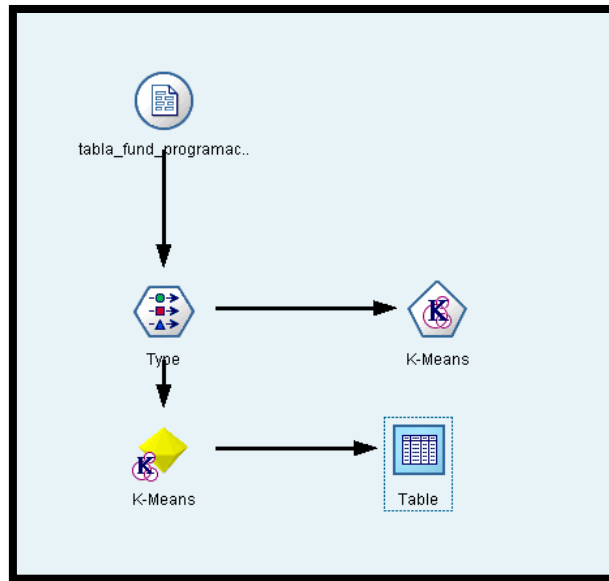


Figura I.9 Resultados diagrama proceso de Clustering K-MEANS

	user_id	num_interaccion_foros	total_num_mensajes_foros	num_veces_foro_add	segundos	calif_prom_foro_1bim	calif_prom_foro_2bim	prom_foros	nota_final	\$K-M-K-Means
1	2393	2	0	0	34479	0.000	0.000	0.000	0.000	cluster-1
2	7050	13	0	0	16359	0.000	0.000	0.000	100.000	cluster-1
3	8277	61	0	7	17087	2.000	2.000	2.000	0.000	cluster-2
4	15175	3	0	0	40556	0.000	0.000	0.000	93.800	cluster-1
5	27486	8	0	0	15095	0.000	0.000	0.000	91.670	cluster-1
6	30189	83	5	5	69921	2.000	0.000	1.000	0.000	cluster-3
7	30241	0	0	0	0	0.000	0.000	0.000	87.500	cluster-1
8	31341	56	8	8	82978	2.000	2.000	2.000	84.750	cluster-2
9	31406	13	0	0	21315	0.000	0.000	0.000	85.710	cluster-1
10	35294	54	3	3	75287	2.000	0.000	1.000	0.000	cluster-3
11	39010	25	4	4	19100	2.000	0.000	1.000	93.750	cluster-3
12	48899	12	4	4	60727	1.250	0.000	0.630	85.360	cluster-3
13	50725	6	0	0	59098	0.000	0.000	0.000	17.500	cluster-1
14	51386	89	13	15	38906	2.000	2.000	2.000	0.000	cluster-2
15	54458	21	4	4	1350	2.000	0.000	1.000	0.000	cluster-3
16	54747	15	2	2	9921	1.000	0.000	0.500	100.000	cluster-1
17	54752	0	0	0	0	0.000	0.000	0.000	62.140	cluster-1
18	56308	8	0	0	24693	0.000	0.000	0.000	80.360	cluster-1
19	57018	43	5	5	5514	2.000	0.000	1.000	86.920	cluster-3
20	57674	0	0	0	0	0.000	0.000	0.000	49.640	cluster-1
21	57942	30	3	3	65476	0.000	1.330	0.670	100.000	cluster-1
22	57947	37	4	4	46401	2.000	0.000	1.000	93.750	cluster-3
23	59409	13	3	3	53553	2.000	2.000	2.000	93.750	cluster-2
24	60622	206	12	13	367	2.000	1.750	1.880	92.880	cluster-2
25	61664	1	0	0	58112	0.000	0.000	0.000	78.850	cluster-1
26	62658	11	3	3	38250	2.000	0.000	1.000	84.290	cluster-3
27	62702	53	7	7	52385	2.000	0.000	1.000	77.120	cluster-3
28	63010	30	3	3	81698	1.330	0.000	0.670	91.670	cluster-3
29	63469	79	9	9	73375	1.750	2.000	1.880	93.750	cluster-2
30	63505	0	0	0	0	0.000	0.000	0.000	67.920	cluster-1
31	63554	1	0	0	60607	0.000	0.000	0.000	74.170	cluster-1
32	63607	8	0	0	33499	0.000	0.000	0.000	0.000	cluster-1
33	63667	0	0	0	0	0.000	0.000	0.000	100.000	cluster-1
34	63826	35	4	4	3787	2.000	0.000	1.000	100.000	cluster-3
35	63872	39	5	5	53480	2.000	0.000	1.000	100.000	cluster-3
36	64060	37	7	7	81981	2.000	1.670	1.830	0.000	cluster-2
37	64365	76	11	11	59377	2.000	1.670	1.830	83.330	cluster-2
38	64451	11	1	1	32075	0.000	2.000	1.000	100.000	cluster-1
39	64652	61	8	8	33159	2.000	2.000	2.000	52.500	cluster-2

Figura I.10 Resultados de la experimentación y Agrupamiento

En la Figura I.10 se puede observar a detalle que usuario pertenece a determinado Cluster.

Se repite este proceso para cada curso.

ANEXO J

Experimentaciones con el Algoritmo K-Means en WEKA y Clementine SPSS

J.1. Primer Experimentación

Parámetros	Descripción	Valor
Función de Distancia	Función de distancia a utilizar para la comparación de casos.	Euclídea
Máximo de Iteraciones	Número máximo de repeticiones	500
Número de Cluster	Número de agrupaciones	3
Seed/Semilla	Semilla a partir de la cual se genera el número aleatorio para inicializar los centros de los clusters.	10

Tabla J. 1 Configuración del Algoritmo K-means en WEKA- 1er Experimento

Para la realización de las pruebas se utilizaron las opciones de configuración descritas en la Tabla J.1; las mismas que se encuentran precargadas en la herramienta WEKA.

La primera experimentación consistió en el uso de todos los atributos mencionados en la sección 3.4.1. El detalle de este experimento se puede observar en el **Anexo J**, donde también se utilizó la Herramienta Clementine SPSS que si bien en cada iteración devolvía un menor grado de error no ofrecía un valor agregado a las prestaciones de WEKA, Clementine se orienta más a un usuario final.

En esta sección se presenta la recopilación de los resultados de las pruebas por cada curso.

J.1.1. Curso Fundamentos de la Programación

La vista general del curso es la que se muestra en la Figura J.1, donde es notable las calificaciones sobresalientes en contraste con la interacción en foros que son relativamente bajas

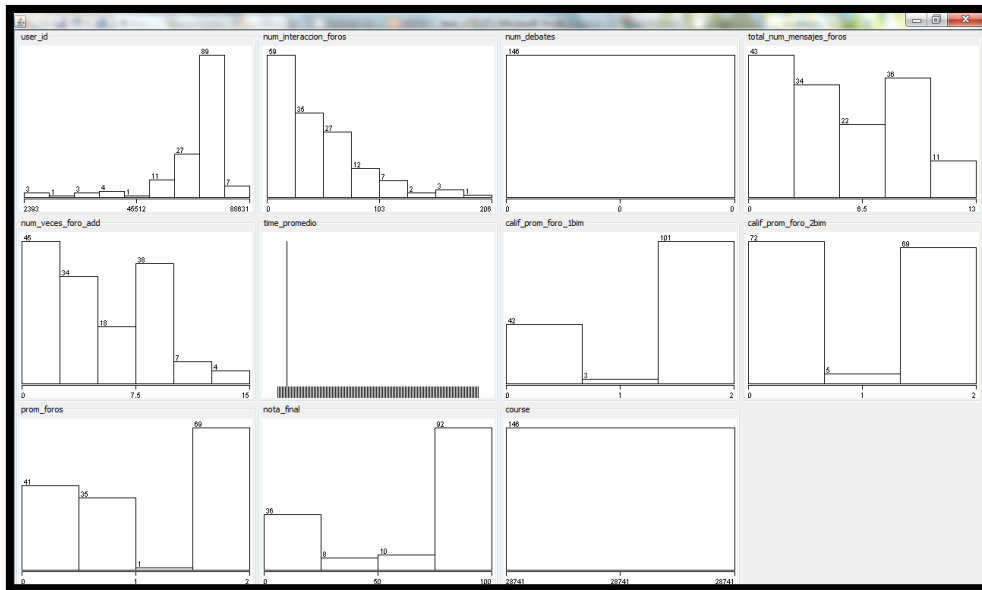


Figura J.1 Vista General de la Asignatura "Fundamentos de la Programación"

Se ha realizado el proceso dividiendo a los alumnos en 3 clusters el primero de 48 correspondiendo un 33%, el segundo de 70 equivalente a un 48% y el tercero de 28 a 19% Figura J.2

```

kMeans
=====

Number of iterations: 4
Within cluster sum of squared errors: 179.56193663694927
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute          Full Data      Cluster#
                   (146)         (48)         (70)         (28)
-----
user_id            67548.0616    62660.6458    70431.9429    68716.7857
num_interaccion_foros  44.0616      19.0833      51.9          67.2857
num_debates        0             0            0            0
total_num_mensajes_foros  4.9589      0.6458      6.8857      7.5357
num_veces_foros_add  0            0            0            0
time_promedio      00:00:00     00:00:00     04:44:47     19:25:21
calif_prom_foro_1bim  1.3745      0.1875      1.9524      1.9643
calif_prom_foro_2bim  0.9134      0.111       1.2324      1.4914
prom_foros         1.1445      0.1496      1.5933      1.7282
nota_final         65.3432     63.8573     89.2024     8.2425
course             28741       28741       28741       28741

Clustered Instances

0      48 ( 33%)
1      70 ( 48%)
2      28 ( 19%)
  
```

Figura J.2 Salida de Información "Fundamentos de la Programación"

El algoritmo arrojó los siguientes resultados:

grupo 0: menor número de interacciones en foros, menor número de mensajes en foros, menor calificación en foros, nota final media

grupo 1: interacciones medianas en foro, número de mensajes mayor que el grupo 0 y menor que el grupo 2, calificación media en foros, nota final alta

grupo 2: mayor interacción en foros, mayor número de mensajes en foro, calificación en foros alta, nota final baja.

En la Figura J.3 se muestra como se encuentran dispersos los clusters, bastante alejados de un centroide, definido, las interacciones son bastante bajas en función de las calificaciones finales.

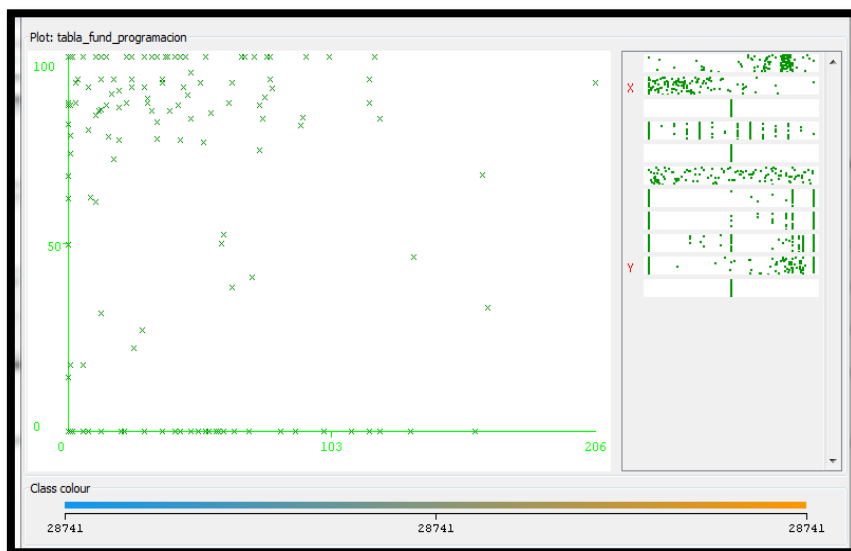


Figura J.3 Relación número de interacciones, nota final para Fundamentos de la Programación

La suma de errores cuadráticos ascendió a: 179,561936636949, las dimensiones de cada uno de los conglomerados se muestran en la Tabla J.2

PARÁMETROS	VALORES		
Clusters	0	1	2
Número de Alumnos	48	70	28
Porcentaje Representativo	33%	48%	19%

Tabla J. 2 Resultados de la primera experimentación curso Fundamentos de la Programación

La distribución no fue homogénea, los resultados que arrojó el algoritmo indican que el Grupo 0 pertenece a los estudiantes con menor colaboración, los del Grupo 1 a los de colaboración media y finalmente los del Grupo 2 los que tienen mayor colaboración.

J.1.3. Curso Lógica de la Programación

La vista general del Curso se muestra en la Figura J.4, en este caso la nota final como las interacciones son bastante bajas.

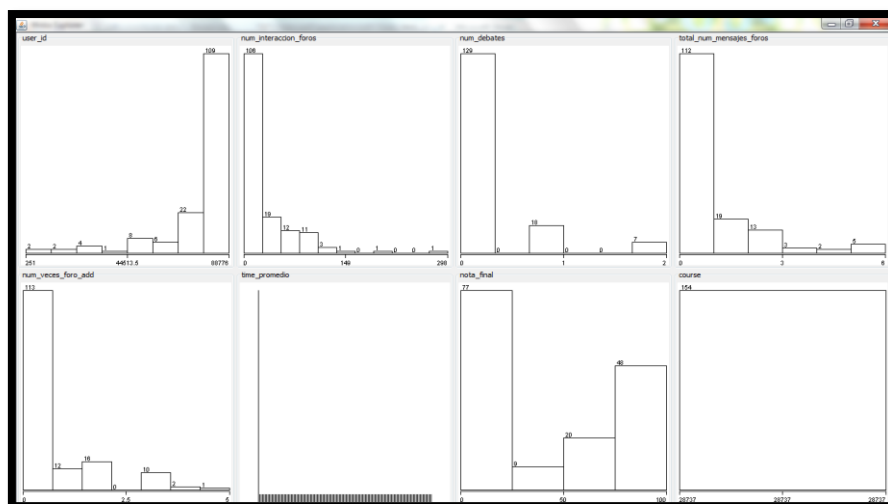


Figura J.4 Vista General de la Asignatura "Lógica de la Programación"

En la Figura J.5 se muestra como se han dividido los estudiantes formándose como se estipuló 3 grupos el primero de 82 el segundo de 23 y el tercero de 49 estudiantes correspondiendo en un 53%, 15% y 32%.

```

kMeans
=====
Number of iterations: 9
Within cluster sum of squared errors: 148.7633869348921
Missing values globally replaced with mean/mode

Cluster centroids:
Attribute          Full Data      Cluster#
                   (154)         (82)          (23)          (49)
-----
user_id            76143.0519    73975.6341    78715.6522    78562.6122
num_interaccion_foros  27.0195      7.7927       86.4348      31.3061
num_debates        0.2078       0.0366       1.2609       0
total_num_mensajes_foros  0.9351      0.1829       3.7391       0.8776
num_veces_foro_add  0.5649       0.1829       0.913        1.0408
time_promedio      00:00:00     00:00:00     20:08:01     11:23:23
nota_final         40.5201      7.6199       67.4683      82.9286
course             28737        28737        28737        28737

Clustered Instances

0      82 ( 53%)
1      23 ( 15%)
2      49 ( 32%)
  
```

Figura J.5 Salida de Información "Lógica de la Programación"

Se han formado los grupos de la siguiente manera:

grupo 0: número de interacción en foros baja, número de mensajes agregados bajo, tiempo nulo realizando acciones en foros, nota final baja.

grupo 1: número de interacción en foros alta, número de mensajes agregados alto, mayor tiempo realizando acciones en foros, nota final media.

grupo 2: número de interacción en foros media, número de mensajes agregados medio, tiempo medio realizando acciones en foros, nota final alta.

En la Figura J.6 se puede visualizar un incremento en las interacciones pero con la calificación final con tendencia baja.

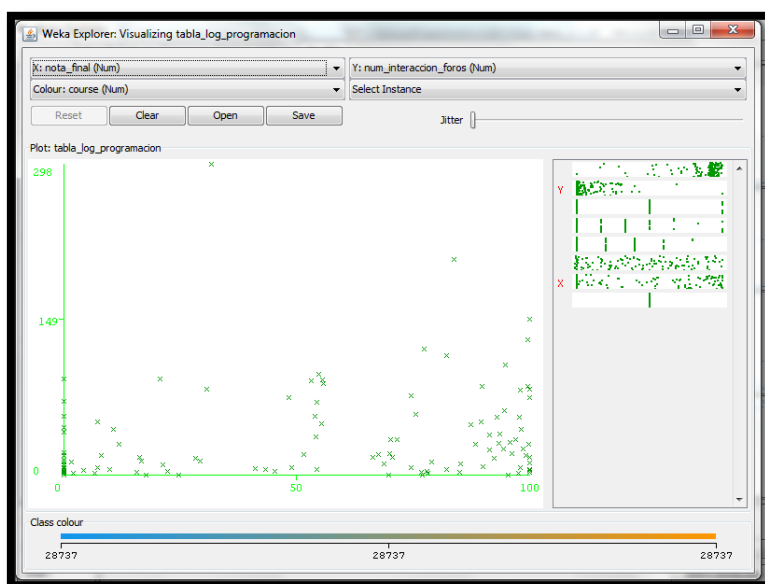


Figura J.6 Relación número de interacciones, nota final para Lógica de la Programación

Para **Lógica de la Programación** la suma de Errores Cuadráticos es de: 148,763386934892, en la Tabla J.3 se encuentra una descripción de cómo se han distribuido los alumnos en los 3 clusters.

PARÁMETROS	VALORES		
Clusters	0	1	2
Número de Alumnos	82	23	42
Porcentaje Representativo	53%	15%	32%

Tabla J. 3 Resultados de la primera experimentación curso Lógica Programación

En el Grupo 0 se ubican los alumnos cuya interacción en foros es mínima, los del Grupo 1 corresponden a los que mayor colaboración tienen, finalmente los del Grupo 2 que representan el punto medio colaborativo en este curso.

J.1.3. Curso Fundamentos Informáticos

En la Figura J.7 se muestra la vista general para la materia de “Fundamentos Informáticos”.

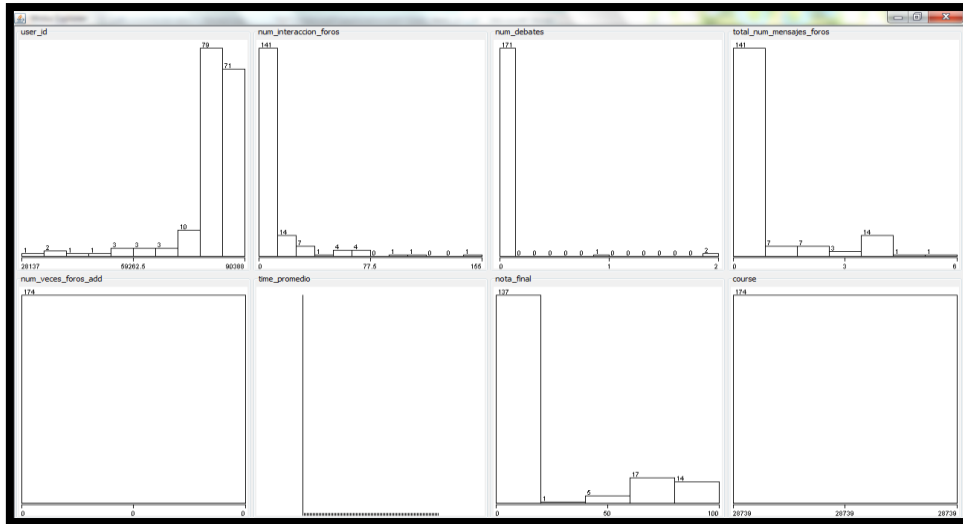


Figura J.7 Vista General de la Asignatura "Fundamentos Informáticos"

La salida de información luego de ejecutado el algoritmo, nos proporcionará datos más precisos, estos se observan en la Figura J.8. Los clusters se formaron con 29, 130 y 15 alumnos, equivalente a un 17%, 75% y 15% respectivamente.

```

kMeans
=====

Number of iterations: 9
Within cluster sum of squared errors: 103.78037289349791
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                Full Data      Cluster#
                          (174)         0           1           2
                          (174)         (29)        (130)        (15)
-----
user_id                   81297.8793  81949.7931  81421.7692   78963.8
num_interaccion_foros    9.5517      16          2.4231      58.8667
num_debates              0.0287     0.0345     0           0.2667
total_num_mensajes_foros 0.5575     1.069      0.0846     3.6667
num_veces_foros_add      0           0           0           0
time_promedio            00:00:00   11:36:13   00:00:00   06:36:16
nota_final               15.8213    76.1207    3.5352     5.722
course                   28739      28739     28739      28739

Clustered Instances

0      29 ( 17%)
1     130 ( 75%)
2      15 (  9%)
  
```

Figura J.8 Salida de Información "Fundamentos Informáticos"

Los grupos formados poseen las siguientes características:

grupo 0: Número de interacción media en foros, pocos debates creados, número medio de mensajes creados, alto tiempo empleado en cualquier acción que implique foros en el curso, nota final alta

grupo 1: Número de interacción en foros baja, debates creados nulos, número bajo de mensajes creados, tiempo nulo empleado en interacciones en foros, calificación baja.

grupo 2: Mayor número de Interacción en foros, mayor número de debates, mayor número de mensajes en foros, tiempo promedio usado en acciones que impliquen foros en el curso, nota final mucho menor que la del grupo 0 pero mayor que la del grupo 1.

La relación nota final, número de interacciones relativamente alto pero aun con calificaciones bajas se muestra en la Figura J.9.

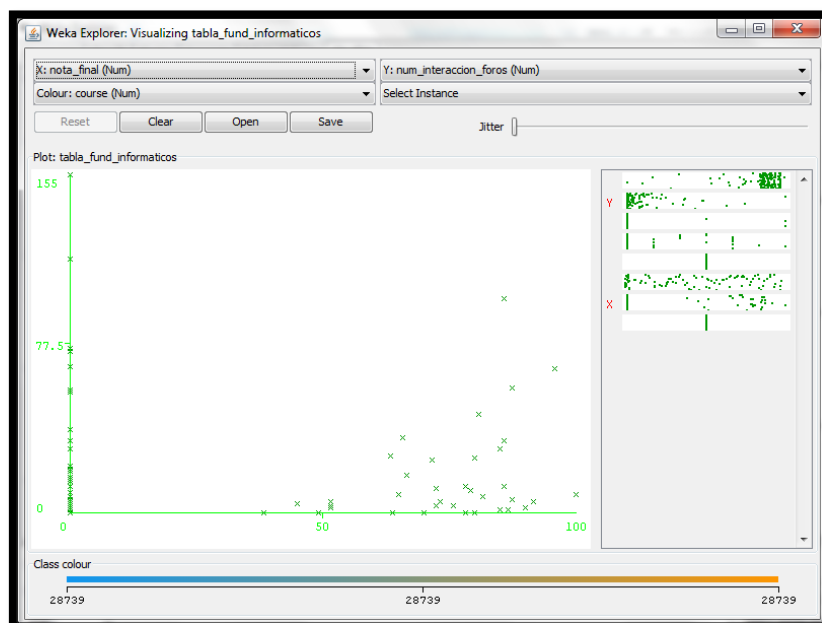


Figura J.9 Relación número de interacciones, nota final para Fundamentos Informáticos.

Fundamentos Informáticos aunque presenta un menor número de errores cuadráticos ($103,780372893497$) con respecto al de las primeras materias no deja de ser elevado. Cada clúster se ha distribuido de forma desigual (Tabla J.4) predominando los alumnos del Grupo 1 precisamente los que menor colaboración registran, el Grupo 0 los de colaboración promedio y los del Grupo 2 con el nivel de colaboración más alto.

PARÁMETROS	VALORES		
Clusters	0	1	2
Número de Alumnos	29	130	15
Porcentajes	17%	75%	9%

Tabla J. 4 Resultados de la primera experimentación curso Fundamentos Informáticos.

J.2. Resultados Herramienta Clementine

Se realizó además la experimentación con la herramienta Clementine SPSS 13.0, donde por medio de un procedimiento estructurado **Anexo H** se logró la obtención de un diagrama sencillo Figura J.10 pero a la vez poderoso para la formación y análisis de conglomerados o clusters.

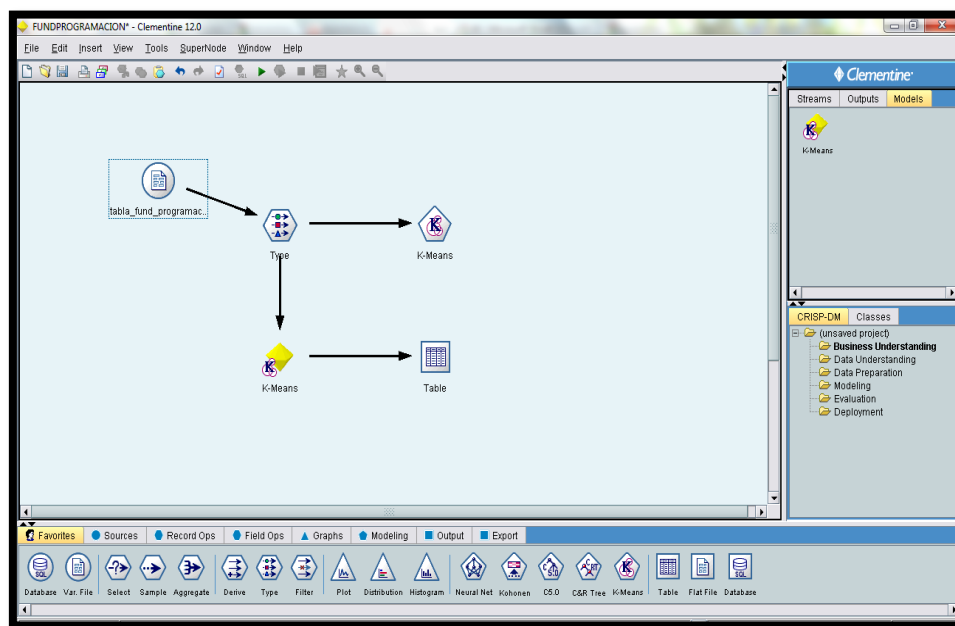


Figura J.10 Diagrama Genérico para la extracción de Conocimiento en Clementine

Los resultados que se obtienen de esta experimentación se reflejarán en 3 vistas para cada asignatura.

J.2.1. Fundamentos de la Programación

La primera vista es la que se da información acerca de la desviación estándar¹⁷, esta es mínima a excepción de “segundos” (se eliminó de la lista de variables “user_id”) debido a su dimensión, diferente una cantidad de otra.

Los estudiantes se han dividido de la siguiente forma:

cluster 1: 43

cluster 2: 70

cluster 3: 33

La segunda vista supone un análisis más detallado de cada cluster, así pues el cluster 2 denota una mayor cantidad de aportes e interacciones, siguiendo del Cluster 3 y Finalmente el 1. La “nota_final” no es un argumento relevante en esta experimentación. Figura J.11

Lo que supondría la siguiente escala:

Cluster 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Cluster 3: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Cluster 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

¹⁷ Desviación estándar: Es una medida del grado de dispersión de los datos con respecto al valor promedio

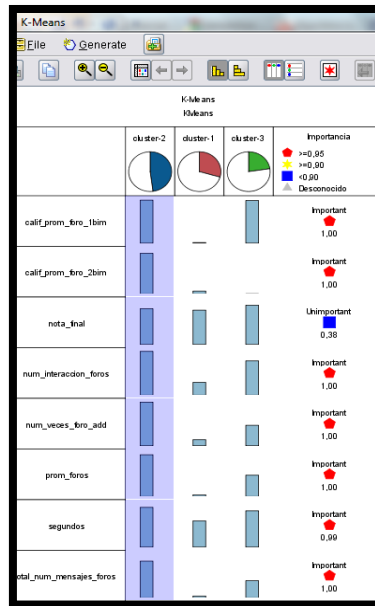


Figura J.11 Resultado Gráfico curso "Fundamentos de la Programación"

En la pestaña "Summary" se observa el número de repeticiones y el error en cada una de estas, como se puede notar en la quinta repetición esta sube, quizás porque "la condición de convergencia no toma en cuenta el error cuadrático" (J. Pérez1, 2007) y tiende a bajar para finalmente llegar a 0. Figura J.12

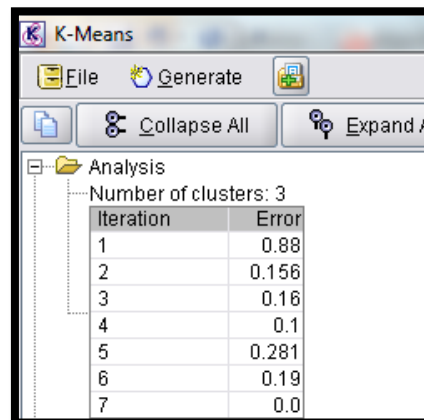


Figura J.12 Iteraciones y errores para el curso "Fundamentos de la Programación"

J.2.2. Lógica de la Programación

Sucede lo mismo con respecto a la desviación estándar del curso anterior, "segundos" posee la más elevada.

Para este curso los estudiantes se han dividido así:

Cluster 1: 47

Cluster 2: 67

Cluster 3: 40

La pestaña “Viewer” arrojó resultados alentadores para los alumnos pertenecientes al cluster 2, siguiendo el cluster 3 y finalmente el 1 Figura J.13.

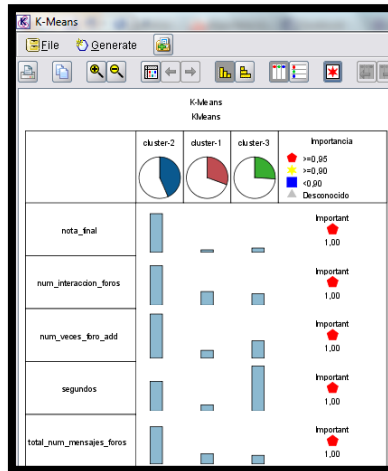


Figura J.13 Resultado Gráfico curso "Lógica de la Programación"

La escala cualitativa para este curso es:

Cluster 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Cluster 3: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Cluster 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

El número de repeticiones y errores tendió siempre a bajar en cada iteración. Figura J.14

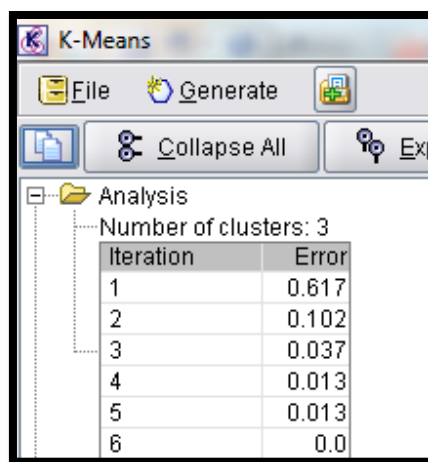


Figura J.14 Iteraciones y errores para el curso "Lógica de la Programación"

J.2.3. Fundamentos Informáticos

Se repite lo que con el resto de cursos, segundos posee una desviación estándar elevada, el porqué de esto se explica en el primer curso trabajado.

Los clusters se han dividido de la siguiente manera:

Cluster 1: 94

Cluster 2: 35

Cluster 3: 45

El cluster 2 si bien es el grupo más pequeño de estudiantes, es el que mayor interacción denota, seguido del cluster 3 y finalmente el cluster 1. Figura J.15

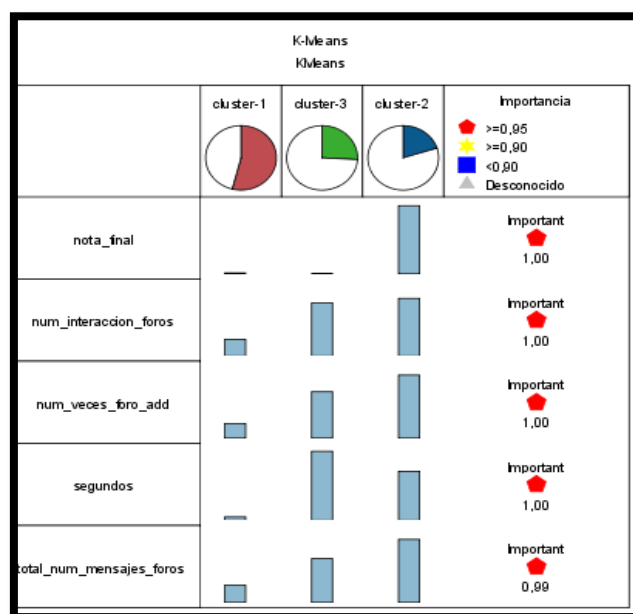


Figura J.15 Resultado Gráfico curso "Fundamentos Informáticos"

La escala cualitativa para Fundamentos Informáticos es:

Cluster 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Cluster 3: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Cluster 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJA

Mientras que las iteraciones y errores tendieron siempre a bajar hasta llegar a 0. Figura J.16

Number of clusters: 3	
Iteration	Error
1	0.561
2	0.237
3	0.046
4	0.016
5	0.03
6	0.007
7	0.0

Figura J.16 Iteraciones y errores para el curso "Fundamentos Informáticos"

J.3. Segunda Experimentación

J.3.1. Algoritmo K-MEANS

Lo que se realizó en primera instancia fue incrementar una variable al conjunto de las ya existentes por medio del filtro "add expression". Figura J.17

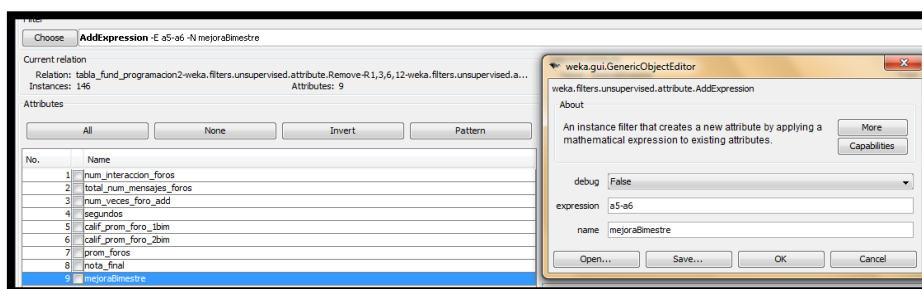


Figura J.17 Filtro "Add Expression" para la Obtención de la mejora en rendimiento del primer al Segundo Bimestre

Las opciones de configuración que se utilizó fueron las precargadas en la Herramienta WEKA siendo el máximo de iteraciones "500", Número de Cluster "3", y semilla "10"

J.3.1.1. Fundamentos de la Programación

Para este curso se presentan dos resultados el primero sin discretizar y el segundo discretizando la variable "nota_final", se utilizó esta variable pues su rango de calificaciones era bastante amplio, además de que devolvía un menor valor en la suma de errores cuadráticos a diferencia de "prom_foros" y del resto de variables numéricas.

Se discretizó partiendo de la premisa de que los valores se incluyen en depósitos para que haya un número limitado de estados posibles. (Zhunio, 2011) y así obtener una tendencia clara de cómo ha sido el progreso de la calificaciones de los estudiantes en base a las interacciones realizadas.

Para el efecto se crearon por separados experimentaciones con variables discretas y numéricas **Anexo K**.

Además se decidió experimentar con diferentes valores de semillas tanto para los valores discretizados como para los que no, estos valores fueron asignados en rangos 10 veces mayores cada vez ya que los incrementos pequeños no hacen una diferencia significativa en los resultados. Tabla J.5

curso	Iteraciones	seed	Suma de errores cuadráticos	
			Sin Discretizar	Discretizando
Fundamentos de la Programación	500	10	45,6675349271115	65,96890697427813
		100	48,37936920820893	63,440045344404794
		1000	46,53461491667241	79,20934099740465
		10000	45,66753492711149	63,44004534440478
		100000	49,91548162381365	79,19960173215362
		5000	45,6675349271115	65,96890697427813
	50000	45,6675349271115	65,96890697427813	
	500000	45,6675349271115	65,96890697427813	

Tabla J. 5 Matriz de número de Iteraciones/semilla – Suma de errores Cuadráticos para Fundamentos de la Programación

No existen cambios al incrementar el número de iteraciones pero sí al cambiar el tamaño de la “semilla” pues representa el número de inicialización de los centros del cluster y por lo tanto como van a ser distribuidos.

Se presentan los resultados con mejor valoración tanto los que no se han sometido a un proceso de discretizado como las que sí.

J.3.1.1.1. Sin Discretizar

Se procede con la ejecución del algoritmo Figura J .18

```

=== Model and evaluation on training set ===

kMeans
=====

Number of iterations: 5
Within cluster sum of squared errors: 45.6675349271115
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                Full Data      Cluster#
                        (146)         (44)         (73)         (29)
=====
num_interaccion_foros    44.0616      17.1136     51.9863      65
total_num_mensajes_foros 4.9589       0.2955     6.7671      7.4828
num_veces_foro_add       4.9452       0.9773     6.4932      7.069
segundos                 39579.0137   27394.4318 42933.5479 49621.7931
prom_foros               1.1445       0.1023     1.5577      1.6859
nota_final               65.3432     63.3664    90.0507     6.1479
mejoraBimestre           0.461        0.0077     0.7497     0.4221

Clustered Instances

0      44 ( 30%)
1      73 ( 50%)
2      29 ( 20%)

```

Figura J. 18 Salida de Información "Fundamentos de la Programación" - Experimento 2 Sin Discretizar

La visualización general de esta experimentación es la de la Figura J.19, donde se aprecia que el contexto calificaciones (entiéndase calificaciones de ambos bimestres y nota final) es bastante alta en relación a la interacción realizada por los estudiantes.

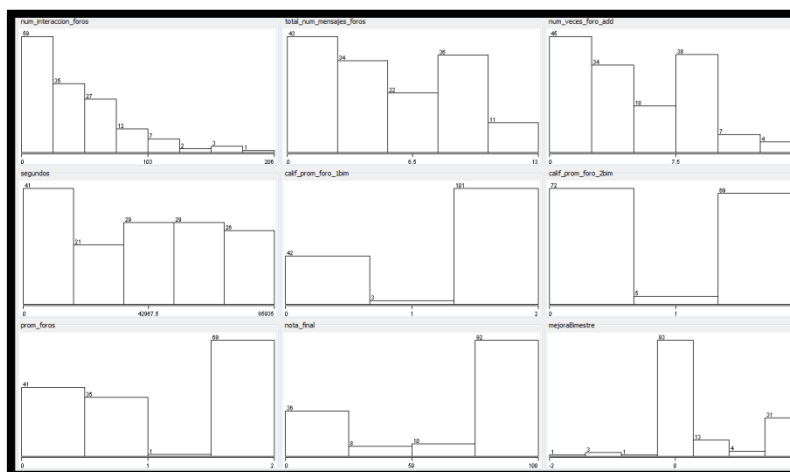


Figura J.19 Vista General del Curso "Fundamentos de la Programación" sin discretizar

La suma de errores cuadráticos bajó visiblemente de 179,561936636949 a 45,6675349271115 un 25,43%, lo que implica un mayor grado de confiabilidad (74,57%) en la agrupación realizada.

Se formaron 3 grupos de alumnos en base lo dispuesto en la configuración del algoritmo, el primero grupo de 44, el segundo de 73 y el tercero de 29 de un total de 146 alumnos, lo que equivale a un 30%, 50% y 20% respectivamente.

Los grupos que se formaron poseen las siguientes características.

Grupo 0: Bajo número de interacción en foros, bajo nivel de foros participado, registran pocos mensajes agregados, menor tiempo invertido en foros, calificación mínima en promedio de actividades concernientes en foros, calificación final media, baja mejora mínima en promedio de un bimestre a otro .

Grupo 1: Interacción en foros mayor que la del grupo 0 pero ligeramente menor que el grupo 2, número de foros participado ligeramente menor que el grupo 2, número de veces promedio en los que un usuario ha agregado un foro, tiempo medio invertido en foros, calificación promedio en foros , nota final más alta del resto de grupos, mejora más alta de calificación de un bimestre a otro.

Grupo 2: mayor número de interacciones en foro, mayor número de mensajes en los que se ha participado, mayor número de foros agregados, mayor tiempo invertido en los foros, mayor promedio en calificación de foros, nota final inferior al resto de grupos, mejora de calificación promedio de un bimestre a otro

Estos grupos se calificarán de la siguiente manera:

Grupo 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Grupo1: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Grupo 0: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

Como en el primer experimento se ha detectado que el nivel de colaboración en foros no está ligado de forma proporcional a la nota final del educando.

Los grupos quedarían dispersos estando X: num_interaccion_foros y Y: nota_final como se muestra en la Figura J.20



FiguraJ. 20 Gráfica Relación Número de Interacción/Nota Final del curso Fundamentos de la Programación.

Con estos cambios se distingue con mayor similitud y consistencia se han conformado los diferentes grupos, siendo los puntos azules el cluster 0, los rojos el cluster 1 y los verdes el 2, además de una mayor consistencia en la formación de grupos.

Como se analizó en un punto anterior el cluster 0 presentan una interacción sumamente baja indirectamente proporcional a la calificación bastante aceptable que obtienen los miembros de éste, los del Cluster 1 presentan un nivel de colaboración media y nota final alta, finalmente los del Cluster 2 denotan una mayor interacción más esta no es proporcional a la nota final recibida.

J.3.1.1.2. Discretizando

Se discretizó la calificación de los estudiantes en los foros dividiéndolos en 4 rangos con un intervalo de 0.5 cada uno, en particiones iguales (25%) lo que correspondería el 100%. Las calificaciones de notas finales están valoradas sobre “100”, mediante esta transformación se obtuvo que 36 estudiantes tienen una calificación menor a 25, 8 estudiantes entre 25 y 50, 10 estudiantes obtuvieron un promedio en foros de 50 a 75 y finalmente un gran número de estos (92) calificaciones mayores a 75. Figura J.21, por lo que se estaría hablando de un promedio sobresaliente del curso en general.

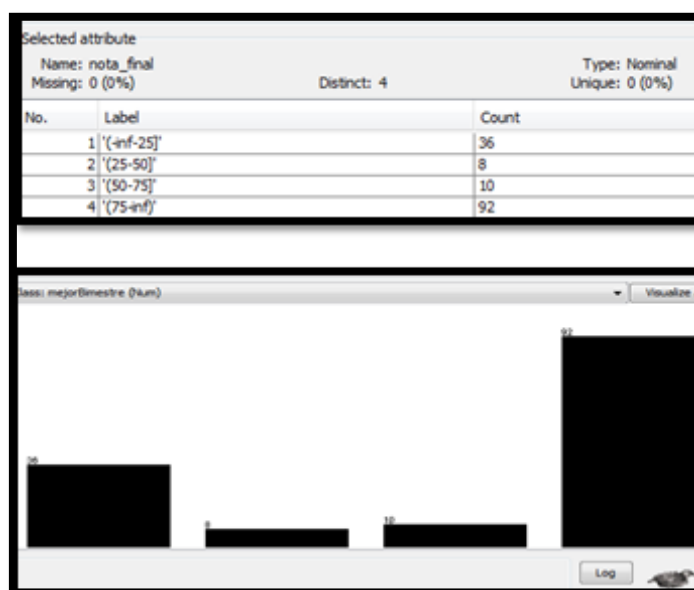


Figura J. 21 Categorización de Variable nota_final

En la Figura J.22 se muestra una vista general de la asignatura. Se registra una calificación final alta inversamente proporcional al número de interacciones.

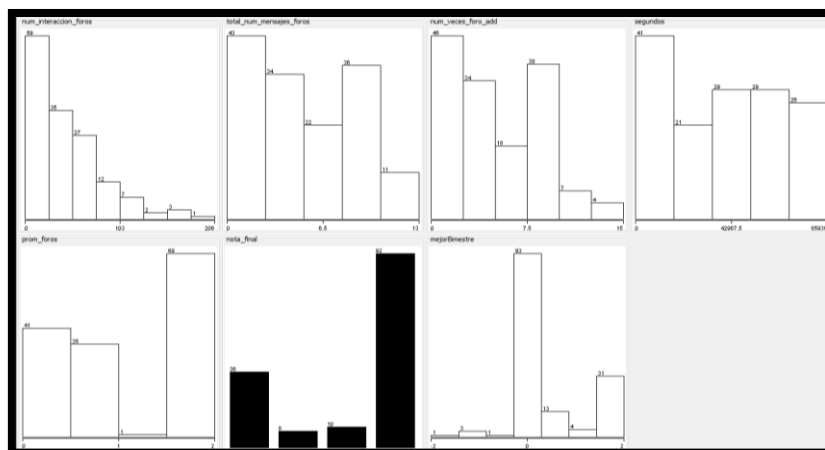


Figura J.22 Vista General de la Asignatura "Fundamentos de la Programación" –Experimento 2 Discretización

En este caso también se removerá el atributo “num_debates”, pues en este curso no existían discusiones creadas por los estudiantes además de calif_prom_foro_1bim y calif_prom_foro_2bim puesto que estos valores no son necesarios si ya se cuenta con el promedio de los mismos.

Se utiliza la semilla en un valor de 100 y el número de iteraciones predeterminado. La salida del proceso de Clustering luego de su ejecución se puede observar en la Figura J.2.3, hay que acotar que este resultado es el más óptimo en función de Suma cuadrada de errores dentro de los experimentos de discretización.

```

kMeans
=====

Number of iterations: 3
Within cluster sum of squared errors: 63.440045344404794
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                Full Data      Cluster#
                        (146)         0           1           2
=====
num_interaccion_foros    44.0616       31.3448     51.3611     53.1923
total_num_mensajes_foros 4.9589        1.8793      5.3056      8.1538
num_veces_foro_add       4.9452        2.1207      5.3889      7.7885
segundos                 39579.0137    30328.8276  47264.0833  44576.0962
prom_foros               1.1445        0.4817      1.1842      1.8563
nota_final               '(75-inf)'    '(75-inf)'  '(-inf-25]' '(75-inf)'
a5-a6                   0.461         0.8478      0.2706      0.1615

Clustered Instances

0      58 ( 40%)
1      36 ( 25%)
2      52 ( 36%)

```

Figura J.23 Salida de Información "Fundamentos de la Programación" - Experimento 2 Discretización

El grupo de alumnos se ha segmentado de la siguiente manera: El cluster 0 en 58 (40%), el cluster 1 en 36 (25%) y el cluster 2 en 52 (36%). Sus características son:

Grupo 0: Menor número de interacciones , menor número de foros participados, así mismo menor número de mensajes agregados, menor tiempo realizando alguna actividad que implique el uso de foros, promedio final en foros bajo, nota final mayor a

75, mayor rango de diferencia entre la calificación del primer y segundo Bimestre (incremento de calificación).

Grupo 1: Mayor número de interacciones en foros que el grupo 0 menor que el 2, mayor número de foros participados que el grupo 0 menor que el grupo 2, mayor número de foros agregados que el grupo 0 menor que el grupo 1, mayor tiempo invertido en foros , promedio en foros mayor que el grupo 0 y menor que el grupo 1, nota final menor a 25 puntos, diferencia mínima entre las calificaciones del primer a segundo bimestre.

Grupo 2: Mayor número de interacciones realizadas en foros, número de foros participados y mensajes agregados, tiempo mayor que el grupo 0 pero menor que el grupo 1 utilizado para actividades en foros, mayores calificaciones en foros, nota final mayor a 75, menor rango de diferencia entre las calificaciones del primer al segundo bimestre lo que implica menor esfuerzo de un bimestre a otro.

Estas segmentaciones pasaran a tomar las siguientes denominaciones:

Grupo 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Grupo 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Grupo 0: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

De 179,561936636949 en la primera experimentación se llegó a 63,440045344404794 en la segunda lo que implica un margen de error del 35.33%, por ende la confiabilidad en esta prueba se incrementó en un 64,67%.

La gráfica número de interacción en foros en relación a la nota final (Figura J.24) es una fiel demostración de cómo se distribuyen los grupos y de cómo el número de interacciones va ligado a la calificación final recibida, en el caso de los estudiantes con nivel de colaboración alta también cuenta con una nota final del mismo tipo, pero también da muestra de que los que menor interacción en foros tienen tienden a tener mayor calificación también.

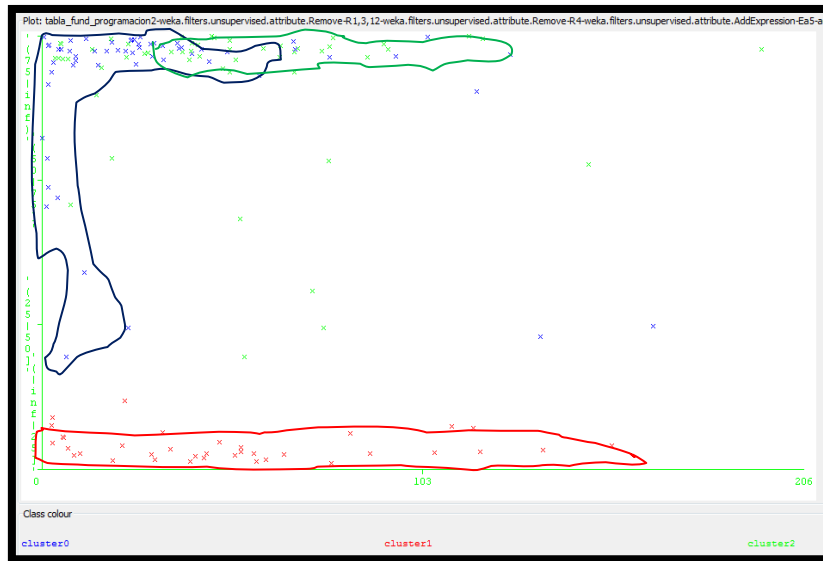


Figura J. 24 Gráfica Relación Número de Interacción/Nota_Final del curso Fundamentos de la Programación Discretizando

Haciendo una comparación en base a un indicador de confiabilidad del algoritmo se obtiene que es con el modo no discretizado con el que se trabajará para la descripción de las recomendaciones. Tabla J.6

	Sin Discretizar	Discretizando
confiabilidad	74,57%	64,67%
% de error	25,43%	35,33%

Tabla J. 6 Recopilación de resultados curso Fundamentos de la Programación

J.3.1.2. Lógica de la Programación

Para este curso se ha decidido no discretizar las instancias debido a que el rango de valores hallado es relativamente bajo y a lo poco necesario de esta labor en la actual experimentación. La vista general es la que se muestra en la Figura J.25.

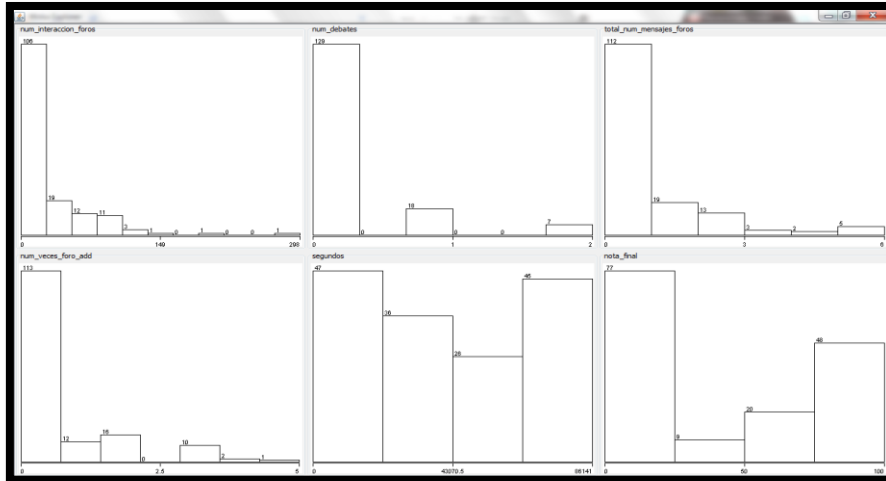


Figura J. 25 Vista General de la Asignatura "Lógica de la Programación" –Experimento 2

A simple vista los datos iniciales muestran que los estudiantes en este curso dedican un buen tiempo a la práctica colaborativa en foros y a diferencia de “Fundamentos de la Programación”, si se han creado discusiones aunque muy pocas. Ya en la ejecución con el algoritmo los resultados son los que se muestran en la Figura J.26.

```

=== Model and evaluation on training set ===

kMeans
=====
Number of iterations: 9
Within cluster sum of squared errors: 34.6864615892839
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute          Full Data      Cluster#
                   (154)         0          1          2
                   (154)         (79)       (21)       (54)
-----
num_interaccion_foros  27.0195    8.6203    90.8095    29.1296
num_debates           0.2078    0.0506    1.3333     0
total_num_mensajes_foros  0.9351    0.2405    3.8095    0.8333
num_veces_foro_add    0.5649    0.1899    0.9048    0.9815
segundos              40268.6494 34102.8228 44134.619 47785.5926
nota_final            40.5201    4.2663    71.3105    81.5841

Clustered Instances

0      79 ( 51%)
1      21 ( 14%)
2      54 ( 35%)

```

Figura J. 26 Salida de Información "Lógica de la Programación" - Experimento 2

En este caso la suma de errores cuadráticos bajó de 148,7633869348921 a 34,6864615892839 implicando una reducción de error del 23.32% y por lo tanto un grado de confiabilidad del 76,68%. Los tres clusters se dividieron 79 para el primer grupo que representa el 51% de todos los datos, 21 para el segundo (14%) y 54 con 35% de representación para el tercer cluster de un total de 154 alumnos.

Cada cluster formado se interpreta de la siguiente manera:

grupo 0: interacción mínima en foros, número medio de debates, pocos foros en los que se ha participado, número de mensajes mínimos agregados, tiempo menor ocupado en foros., nota final baja.

grupo 1: mayor número de interacción en foros, número mayor de debates, mayor número de foros en los que se ha participado, número menor que el grupo 2 pero mayor que el 0 de mensajes agregados, tiempo promedio en actividad de foros, nota final media

grupo 2: interacción en foros media, número de debates nulo, pocos foros en los que se ha participado, mayor número de mensajes agregados, mayor tiempo ocupado en actividad de foros, nota final mayor.

La escala cualitativa quedaría así:

Grupo 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Grupo2: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Grupo 0: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

Las interacciones para este curso si se dan de forma proporcional con la calificación final obtenida.



Figura J. 27 Relación número de interacciones, nota final para "Fundamentos de la Programación"- Experimento 2

Estando X: num_interaccion_foros y Y: nota_final, el resultado es el que se muestra en la Figura J.27. Se denota homogeneidad en cada una de las agrupaciones, pues como se explicó el margen de error es mínimo.

Se ha tomado en consideración la experimentación del anterior curso como una forma de verificar la existencia errores mínimos en los valores, los resultados se presentan en la Tabla J.7

Curso	Iteraciones	Seed	Suma de errores cuadráticos
Lógica de la Programación	500	10	34.6864615892839
		100	34.6864615892839
		1000	34.702221137433014
		10000	36.26338955550469
		100000	34.686461589283894
	5000	10	34.6864615892839
	50000	10	34.6864615892839
	500000	10	34.6864615892839

Tabla J. 7 Matriz de número de Iteraciones/semilla – Suma de errores Cuadráticos para Lógica de la Programación.

La prueba realizada en primera instancia ha resultado ser la más óptima solución al obtenerse una mínima suma de errores cuadráticos. No se registran cambios al probar con diferentes valores en el número de iteraciones por lo que se concluye que estas no ejercen ninguna influencia al menos en esta experimentación.

J.3.1.3. Fundamentos Informáticos

Las variables que se utilizan para este curso son las mismas que la del curso “Lógica de la Programación” se ha creído conveniente proceder de la misma forma, obviando además la variable “num_debates” ya que los outliers afectarían negativamente esta prueba.

La vista general de cursos Figura J.28, denota en forma general una menor interacción en comparación con el resto de materias.

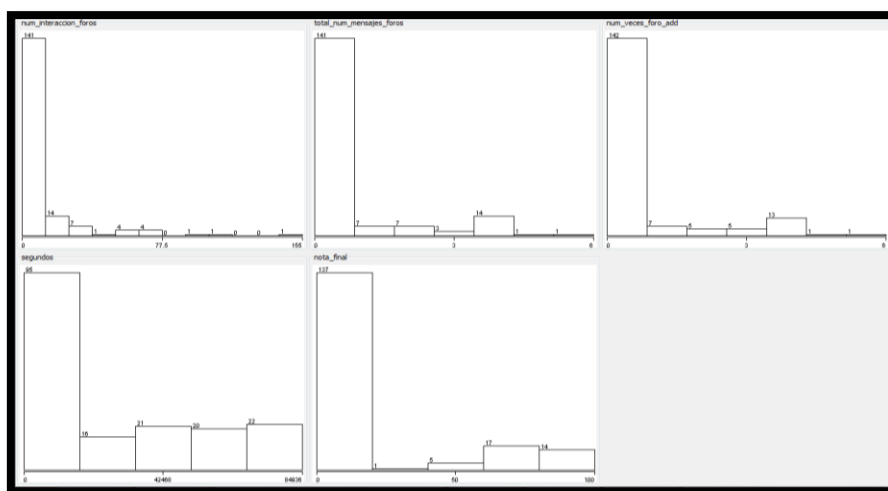


Figura J. 28 Vista General de la Asignatura "Fundamentos Informáticos" –Experimento 2

Ejecutando el algoritmo se obtienen los siguientes resultados Figura J.29 donde de 103,78037289349791 se paso a 22,661696453425375 la suma de errores cuadráticos, representando un 21,84% el grado de error traduciendo una efectividad del 78.16%

```

kMeans
=====

Number of iterations: 5
Within cluster sum of squared errors: 22.661696453425375
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute          Full Data          Cluster#
                   (174)             0          1          2
                   (174)             (52)       (100)      (22)
-----
num_interaccion_foros    9.5517    7.4423    1.01    53.3636
total_num_mensajes_foros 0.5575    0.2308    0.03    3.7273
num_veces_foro_add       0.546     0.2308    0.03    3.6364
segundos                 24721.3736 59555.3462 4267.31 35359.5455
nota_final               15.8213   31.4425   4.5958  29.9236

Clustered Instances

0      52 ( 30%)
1     100 ( 57%)
2      22 ( 13%)

```

Figura J. 29 Salida de Información "Fundamentos Informáticos" - Experimento 2

Los tres clusters se han formado el primero con 52 el segundo 100 y el tercero 22 representando un 30%, 57% y 13%.

Los grupos cumplen con las siguientes características:

Grupo 0: número medio de interacciones, de foros participados y mensajes agregados, mayor tiempo empleado en actividades concernientes en foros, nota final mayor que el resto de grupos.

Grupo 1: número de interacción menor en foros, mínimo número de foros participados, menor número de mensajes agregados, tiempo reducido en actividades de foro, nota final mínima.

Grupo 2: mayor número de interacción en foros, mayor número de foros participado, mayor número de foros agregados, tiempo menor que el del grupo 0 y mayor que el grupo 1 empleado en foros, nota final media.

La escala cualitativa quedaría así:

Grupo 2: ESTUDIANTES CON NIVEL DE COLABORACIÓN ALTA

Grupo 0: ESTUDIANTES CON NIVEL DE COLABORACIÓN MEDIA

Grupo 1: ESTUDIANTES CON NIVEL DE COLABORACIÓN BAJO

La relación Siendo X: interaccion_foros y Y: Clusters, quedaría de la siguiente forma, identificándose claramente su distribución, como se indica en la Figura J.30.



Figura J. 30 Relación número de interacciones, clusters para "Fundamentos Informáticos"- Experimento 2

La homogeneidad en la formación de grupos corresponde con la reducida suma de errores cuadráticos para este curso. Es necesario experimentar con los distintos valores de semilla/iteraciones para este curso, el resultado se observa en la Tabla J.8.

Curso	Iteraciones	Seed	Suma de errores cuadráticos	
Lógica de la Programación	500	10	23.661696453425375	
		100	24.14320313761428	
		1000	23.661696453425378	
		10000	24.137523668484363	
		100000	24.17059005574087	
	5000	10	23.661696453425375	
		50000	10	23.661696453425375
		500000	10	23.661696453425375

Tabla J. 8 Matriz de número de Iteraciones/semilla – Suma de errores Cuadráticos para Fundamentos Informáticos

El valor más óptimo fue el probado en la experimentación, se ratifica una vez más que el valor de las iteraciones en nada influyen en el crecimiento de la suma de errores cuadráticos por poco variante que sea esta.

J.3.2. Algoritmo EM

El algoritmo EM proviene de la estadística y es bastante más elaborado que el K-medias, con el coste de que requiere muchas más operaciones y es apropiado cuando

sabemos que los datos tienen una variabilidad estadística¹⁸ de modelo conocido. (García & Álvarez, 2003). Si bien para este algoritmo se puede buscar el número de grupos más apropiado se predeterminaran a 3 clusters, además se partirá para cada curso del procesado ya definido anteriormente, de tal forma que se encuentre en “igualdad” de condiciones con respecto a los anteriores experimentos

Un argumento válido en el análisis de este segundo algoritmo de agrupamiento es que se basa en criterios estadísticos y no mediante distancias entre vectores de atributos

Se tomará como base para la realización de las pruebas los parámetros predeterminados en WEKA.

Los experimentos que se realizaron con este algoritmo **Anexo L** revelaron que al cambiar el número de semillas 10 veces cada vez a partir de 10 hasta 10000, se muestra invariable el grado de verisimilitud, esto ocurrió para las tres materias.

J.3.2.1. Fundamentos de la Programación

Los resultados obtenidos con los datos sin **DISCRETIZAR** de esta experimentación son los mostrados en la Figura J.31

¹⁸ Variabilidad Estadística: Nombre que se da a las diferencias en el comportamiento de todo fenómeno observable que se repite bajo iguales condiciones. (Galbiate, 2011)

Attribute	Cluster		
	0 (0.25)	1 (0.49)	2 (0.27)
=====			
num_interaccion_foros			
mean	13.415	43.7744	72.8174
std. dev.	20.2451	29.7727	49.0813
total_num_mensajes_foros			
mean	0	6.2878	7.1153
std. dev.	3.7708	2.7428	2.9322
num_veces_foro_add			
mean	0.6389	5.9933	7.01
std. dev.	2.2749	3.1061	3.3742
segundos			
mean	26725.3604	43278.4846	44705.86
std. dev.	25868.0703	25026.9576	25616.4133
prom_foros			
mean	0	1.4756	1.598
std. dev.	0.7976	0.5027	0.5392
nota_final			
mean	62.6287	92.2794	18.9514
std. dev.	37.8561	7.9218	28.6307
mejoraBimestre			
mean	13.415	37.4866	65.702
std. dev.	20.2451	29.3268	47.9128

Figura J.31 Experimentación del curso Fundamentos de la Programación con el Algoritmo EM - Sin Discretizar

Se formaron 3 grupos: El grupo 0 de 44, el segundo de 68 y el tercero de 34 equivalentes en un 30%, 47% y 23% respectivamente.

Con un Log likelihood o registro de verisimilitud de : -31.88892

Los datos **DISCRETIZADOS** presentan el siguiente resultado (Figura J.32)

Attribute	Cluster		
	0 (0.32)	1 (0.37)	2 (0.32)
num_interaccion_foros			
mean	46.0032	12.5475	78.3405
std. dev.	22.4138	11.7873	46.2796
total_num_mensajes_foros			
mean	7.9247	1.2553	6.2669
std. dev.	0.9328	1.7166	3.7307
num_veces_foro_add			
mean	8.0039	1.3092	6.0831
std. dev.	1.0349	1.9059	4.1146
segundos			
mean	44590.8425	33164.1342	41968.7963
std. dev.	24161.85	26972.3544	26528.5992
prom_foros			
mean	1.9068	0.3822	1.2629
std. dev.	0.1112	0.5292	0.6266
nota_final			
'(-inf-25]'	11.0409	11.6985	16.2606
'(25-50]'	3.9901	3.957	3.0528
'(50-75]'	4.9895	5.989	2.0216
'(75-inf)'	30.1467	35.7319	29.1214
[total]	50.1672	57.3763	50.4565
mejoraBimestre			
mean	38.0785	11.2923	72.0735
std. dev.	22.3754	11.0869	45.3482

Figura J.32 Experimentación del curso Fundamentos de la Programación con el Algoritmo EM - Discretizado

Se segmentaron los grupos de la siguiente manera:

Grupo 0: 46 (32%)

Grupo 1: 54 (37%)

Grupo 2 46 (32%)

Para **Fundamentos de la Programación** el nivel de verisimilitud fue de : -27.05513. En las pruebas realizadas sobre la instancia discretizada no existe cambio alguno, en las que no se discretizó con semilla 1000 hubo una ligera casi imperceptible disminución de la verisimilitud. Tabla J.9

Numero de Iteraciones	Semilla	Discretizando	Sin Discretizar
100	100	-27.05512	-31.88892
	1000	-27.05512	-31.77027
	10000	-27.05512	-31.88892
	100000	-27.05512	-31.88892
1000	100	-27.05512	-31.88892
10000	100	-27.05512	-31.88892
100000	100	-27.05512	-31.88892

Tabla J. 9 Evaluación en el cambio de semillas/iteraciones para el curso Fundamentos de la Programación

J.3.2.2. Lógica de la Programación

Para lógica de la Programación los resultados fueron los que se muestran en la Figura J.33

Attribute	Cluster		
	0 (0.36)	1 (0.47)	2 (0.17)
=====			
num_interaccion_foros			
mean	3.0024	24.8952	84.652
std. dev.	5.1993	26.0282	60.608
num_debates			
mean	0	0	1.2307
std. dev.	0.5071	0.5071	0.5045
total_num_mensajes_foros			
mean	0.0097	0.7427	3.4613
std. dev.	0.0981	0.9703	1.5989
num_veces_foro_add			
mean	0.0143	0.8642	0.9231
std. dev.	0.1187	1.1456	1.4121
segundos			
mean	33344.7423	44980.3151	42148.1843
std. dev.	33231.936	25971.6706	26398.0403
nota_final			
mean	0.9562	63.7385	61.513
std. dev.	3.0771	33.6875	35.9486

Figura J.33 Experimentación del curso Lógica de la Programación

Los estudiantes se agruparon así:

Grupo 0 : 55 (36%)

Grupo 1: 76 (49%)

Grupo 2: 23 (15%)

Grado de verisimilitud de: -23.2495.

Se muestra un mínimo casi imperceptible cambio en la verisimilitud al experimentar con diferentes semillas e iteraciones. Tabla J.10

Numero de Iteraciones	Semilla	Verosimilitud
100	100	-23.2495
	1000	-23.6451
	10000	-23.6451
	100000	-23.6451
1000	100	-23.2495
10000	100	-23.2495
100000	100	-23.2495

Tabla J. 10 Evaluación en el cambio de semillas/iteraciones para el curso Lógica de la Programación

J.3.2.3. Fundamentos Informáticos

Los resultados para este curso son los que se indican en la Figura J.34.

```
EM
==

Number of clusters: 3

Attribute          Cluster
                   0          1          2
                   (0.65)   (0.15)   (0.2)
=====
num_interaccion_foros
  mean              0.7247   48.0775   9.2306
  std. dev.         1.5714    33.841    6.2308

total_num_mensajes_foros
  mean              0          3.4333   0.1962
  std. dev.         1.3012    1.0728    0.3971

num_veces_foro_add
  mean              0          3.3572   0.1962
  std. dev.         1.2928    1.1996    0.3971

segundos
  mean             10735.049 39367.4609 59953.3348
  std. dev.        19450.6302 26332.9155 19297.942

nota_final
  mean              7.0004   31.2494   33.2725
  std. dev.         20.8797   40.1223   38.944
```

Figura J. 34 Experimentación del curso Fundamentos Informáticos

Los grupos se han formado de la siguiente manera:

Grupo 0: 97 (56%)

Grupo 1: 26 (15%)

Grupo 2: 51 (29%)

Su registro de verisimilitud es de: -21.3299

En las pruebas realizadas con diferentes semillas el resultado de verisimilitud es constante Tabla J.11

Número de Iteraciones	Semilla	Verosimilitud
100	100	-21.3299
	1000	-21.3299
	10000	-21.3299
	100000	-21.3299
1000	100	-21.3299
10000	100	-21.3299
100000	100	-21.3299

Tabla J. 11 Evaluación en el cambio de semillas/iteraciones

ANEXO K

PAPER

TÉCNICAS DE MINERÍA DE DATOS PARA IDENTIFICAR PATRONES DE COLABORACIÓN DE LOS ESTUDIANTES QUE HACEN USO DEL EVA DE LA UTPL

Ing. Samanta Cueva

Cinthia Pulla Elizalde

Ing. Priscila Valdiviezo

IICC IC-IS

Universidad Técnica Particular de Loja

Unidad de Virtualización

spcueva@utpl.edu.ec

cepulla@utpl.edu.ec

pmvaldiviezo@utpl.edu.ec

RESUMEN: En el presente trabajo se abordará el nivel de colaboración estudiando el entorno colaborativo con mayor número de usuarios con el que cuenta la UTPL. Este espacio se denomina, Entorno Virtual de Aprendizaje (EVA), utilizado como herramienta fundamental para el desenvolvimiento académico de sus educandos tanto de las modalidades Abierta como Clásica. Se utilizará la metodología inductiva como técnica de inferencia, seleccionándose las características y técnicas más aptas de MINERÍA DE DATOS para la identificación de patrones de comportamiento colaborativo en los estudiantes de la UTPL modalidad abierta mediante la búsqueda de elementos colaborativos dentro del entorno MOODLE específicamente FOROS y su relación con las calificaciones obtenidas, mediante el análisis de sus tablas usando para ello los ALGORITMOS DE AGRUPAMIENTO K-MEANS, EM y Clustering Jerárquico para el descubrimiento de grupos potenciales de estudiantes hacia el establecimiento de estrategias pedagógicas.

Este análisis se concentra dentro del Período Octubre-Febrero 2011. Finalmente se interpretarán y validarán los resultados, verificando si el nivel de colaboración ejerce de forma directa en el ámbito académico del educando.

PALABRAS CLAVE: minería, e-learning, clustering, algoritmos, K-Means, EM, Jerárquico, Foro, MOODLE, experimentación, UTPL, EVA, comportamiento, colaboración.

1. MINERIA DE DATOS EN LA EDUCACIÓN

El desarrollo tradicional de los cursos e-learning es una actividad ardua en el que el profesor del curso tiene que elegir el contenido que se mostrará, decidir sobre la estructura de los contenidos, y determinar los elementos de contenido más apropiado para cada tipo de usuario potencial del curso. [1]

El sitio web de la comunidad de Educational Data Mining (EDM)¹⁹, define la minería de datos educativos de la siguiente manera: "La minería de datos en la educación es una disciplina emergente, cuyo interés radica en la elaboración de métodos para explorar los tipos de información que proceden de los centros educativos y el uso de los métodos para comprender mejor a los estudiantes y la manera en que aprenden".

El argumento de [2] es que Data Mining tiene como objetivo reunir los beneficios de varias áreas como la

¹⁹ <http://www.educationaldatamining.org>

estadística, inteligencia artificial, las bases de datos y el pre procesamiento masivo, usando las bases de datos como materia prima.

En el ámbito educativo la minería de datos proporciona entre otras características, criterios y pautas para personalizar el sistema de enseñanza estableciendo cambios estructurales en el mismo. Existen diversos contextos donde se podría implementar EDM, [3] manifiesta la existencia de cuatro áreas claves:

La primera: Para deducir el desenvolvimiento del estudiante dentro del sistema y lo aburrido o frustrado que podría sentirse.

La segunda: Para el descubrimiento o mejora de la estructura de los modelos de conocimiento de dominio.

La tercera: Un tercer aspecto clave de la aplicación de métodos de EDM ha sido en el estudio pedagógico de apoyo (tanto en software para el aprendizaje y el aprendizaje en otros dominios, como los comportamientos), para descubrir qué tipos de apoyo pedagógico son más eficaces, ya sea de forma general o por grupos de estudiantes o en situaciones diferentes.

La Cuarta: La búsqueda empírica de pruebas para perfeccionar y ampliar las teorías educativas y fenómenos educativos conocidos, para una comprensión más profunda de los factores clave que afectan el aprendizaje.

Según [4] una consideración inicial, parece implicar sólo dos grandes grupos, los alumnos y los instructores, en realidad hay más grupos que participan con muchos más objetivos, como puede verse en la Tabla 1.1

Usuarios/actores	Objetivos de uso- Data Mining
Alumnos/Estudiantes /Pupilos	Personalizar el e-learning, realizando una recomendación de actividades y tareas, forjando un aprendizaje basado en experiencia.
Educadores/Profesores/Instructores/tutores	<ul style="list-style-type: none"> • Retroalimentación • Analizar el comportamiento del estudiante • Brindar mayor soporte • Detección de los errores más comunes que puede llegar a tener un estudiante
Desarrolladores de Cursos/Investigadores Educativos	Para la evaluación y mantenimiento de cursos, valorando la estructura de estos.
Organizadores/Proveedores de aprendizaje/Universidades/ Empresas de formación privada	<ul style="list-style-type: none"> • Para la toma de decisiones en instituciones de nivel superior. • Para encontrar la mejor relación costo/eficiencia. • Seleccionar los candidatos más calificados para la admisión en sus universidades.
Administradores/Administradores de centro educativo/Administradores de red/Administradores de Sistema	<ul style="list-style-type: none"> • Para desarrollar la mejor manera de organizar los recursos institucionales humanos y materiales y su oferta educativa. • Establecer parámetros de eficiencia del sitio, determinando el enfoque y eficiencia de la educación a distancia.

Tabla 1.1 EDM Usuarios/Objetivos. Adaptation of [4]

2. PROCESO DE DESCUBRIMIENTO DE CONOCIMIENTO DE

[1] Proponen el siguiente proceso:

Recopilar datos. El sistema LMS²⁰ es utilizado por los estudiantes y la utilización de la información y la interacción se almacena en la base de datos

²⁰ LMS: Learning Management System en Español se traduciría como Sistema de Manejo de Aprendizaje: Se emplea para administrar, distribuir y controlar las actividades de formación no presencial (o aprendizaje electrónico) de una institución u organización. (Fuente: <http://es.wikipedia.org>)

Pre procesamiento de los datos.

Los datos se limpian y se transforman en un formato adecuado para ser explotados. Con el fin de pre-procesar los datos de MOODLE, se puede utilizar una herramienta de administración de base de datos

Aplicar minería de datos: Se aplican los algoritmos de minería de datos y se construye el modelo usando datos específicos y herramientas de minería de datos.

Interpretación, evaluación y despliegue de los resultados: Los resultados del modelo son interpretados y utilizados para la adopción de nuevas medidas. El profesor puede utilizar la información descubierta para tomar decisiones sobre los estudiantes y las actividades del curso de MOODLE con el fin de mejorar el aprendizaje de los estudiantes. Una ampliación de este proceso se puede observar en la Figura 2.1

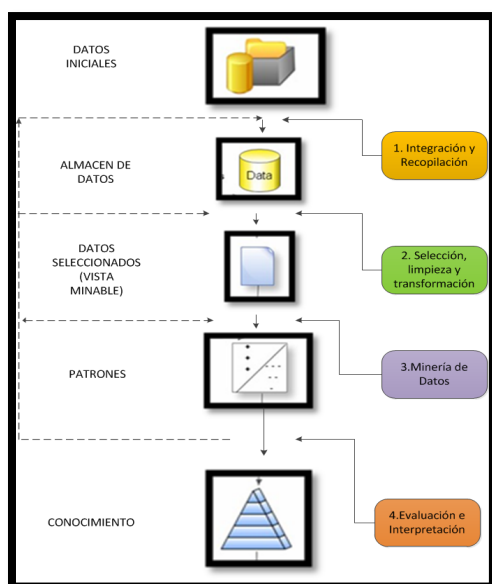


Figura 2. 1 Adaptación de las Fases del Proceso de Extracción de Conocimiento [5]

3. ESTUDIO DE LOS ALGORITMOS DE AGRUPAMIENTO

K-Medias: Se trata de un algoritmo clasificado como Método de Particionado y Recolocación. Este método es hasta ahora el más utilizado en aplicaciones científicas e industriales. El nombre le viene porque representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su centroide. Este método únicamente se puede aplicar a atributos numéricos, y los outliers⁸ le pueden afectar muy negativamente.

Sin embargo, la representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. La suma de las discrepancias entre un punto y su centroide, expresado a través de la distancia apropiada, se usa como función objetivo. La función objetivo, suma de los cuadrados de los errores entre los puntos y sus centroides respectivos, es igual a la varianza total dentro del propio cluster. [6]

Expectation-Maximation: El algoritmo EM asigna a cada instancia una distribución de probabilidad de pertenencia a cada cluster. El algoritmo puede decidir cuántos clusters crear basado en validación cruzada o se le puede especificar a priori cuántos debe generar.[7]

Clustering Jerárquico: “Se caracterizan porque en cada paso del algoritmo sólo un objeto cambia de grupo. Si un objeto ha sido asignado a un grupo ya no vuelve a cambiar de grupo. [8]

4. Descripción de los foros en MOODLE

Esta actividad tal vez sea la más importante siendo a través de los foros donde se da a mayor parte de los debates y discusión de los temas del curso. Se dice que esta actividad es asincrónica ya que los participantes no tienen que acceder al sistema al mismo tiempo. Su icono estándar es:

En todas las asignaturas del EVA por lo menos existe un foro por cada bimestre, el tipo que se utilice dependerá de la configuración que le haya dado el profesor y de la forma en cómo se desea emitir y captar la información.

Como muestra de las funcionalidades que se pueden agregar a los foros se exponen las siguientes cabeceras configurables al momento de su creación.

- Nombre del foro
- Tipo del foro {Normal, debate sencillo, debate por persona, Preguntas y respuestas}
- Introducción
- Subscripción
- Rastreo
- Adjunto
- RSS
- Calificación
- Bloqueo
- Grupos

5. DESARROLLO

Se inicia la minería de datos que incluye cuatro fases: Integración y Recopilación, Selección, Limpieza y transformación, Minería de datos y Evaluación e interpretación.

Luego de terminada la sección de minería de datos se procederá a la obtención de patrones y análisis de resultados

5.1. INTEGRACIÓN Y RECOPIACIÓN

En la etapa de Integración y recopilación se adquirió la base de datos de los estudiantes de la UTPL la que fue luego integrada a MOODLE para una construcción del escenario más cercana a la realidad.

5.2. SELECCIÓN, LIMPIEZA Y TRANSFORMACIÓN

El preprocesamiento (Limpieza, Selección y Transformación) fue una etapa que constituyó el 50% de este trabajo y consistió en la elección de los cursos con los que se trabajaría utilizándose la herramienta GEPHI Figura 5.1 para una selección fundamentada en el número de contribuciones Tabla 5.1 que poseía el docente y luego la materia que impartía.

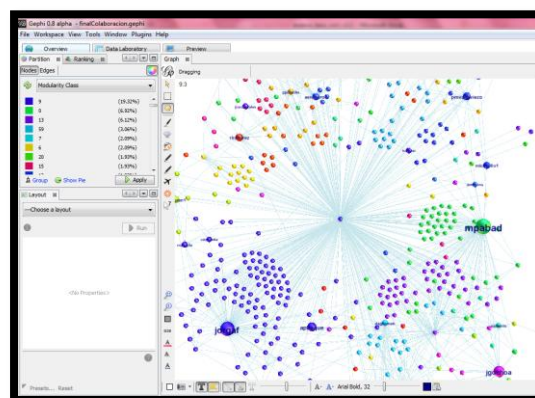


Figura 5. 1 Nodos de Interacción en el EVA

Id_usuario	# de Interacciones
5	151
33	111
2879	73
44	42
453	35
2912	27
3	26
3087	26
32100	21

Tabla 5. 1 Recopilación de nodos con mayor colaboración

Se realizó el tratamiento de los datos, la selección de aquellos que proporcionarían información relevante a la participación de foros y su relación con el aspecto académico del educando. Tabla 5.2

Atributo	Origen	Descripción
user_id	prefix_log	Identificación del Usuario.
sexo_usr	prefix_user_utpl	Sexo de los individuos, M si es Masculino, F si es Femenino
num_acceso_foros	prefix_log	Número de veces que un usuario ha accedido a un curso, se utiliza la acción "view_forum"
subtemas_leidos	prefix_log	Número de veces en las que un estudiante ha leído un hilo de mensajes.
num_respuestas_post	prefix_log	Número de veces que un estudiante ha respondido o agregado un hilo en el foro.
num_respuestas_debates	prefix_forum_discussions	Número de veces que un estudiante ha agregado un mensaje en una discusión como respuesta a otro mensaje.
num_mens_act	prefix_log	Número de veces que un usuario ha actualizado un mensaje en el foro.
arch_adjuntos	prefix_forum_posts prefix_forum_discussions	Archivos adjuntos a los mensajes en foros.
numForos_subscr	prefix_forum_subscriptions	Foros en los que se ha suscrito un usuario.
prom_horas	prefix_log	Número de horas promedio que un usuario ha usado para la gestión de foros.
nota_final	Reporte de Moodle (ver Anexo D).	Nota final sobre 100, calificación de todas las actividades y evaluación vía EVA.
course	prefix_log prefix_course	Identificación de la asignatura

Tabla 5. 2 Descripción de atributos usados para la recopilación de información en foros.

Luego de esto se crearon las consultas de cada uno de los atributos siendo llenados a modo de matriz, para luego ser cargadas en la base de datos

y transformada a formato .CSV se procedió así para cada una de las materias.

6. MINERIA DE DATOS

La etapa de Minería fue la siguiente en ser tratada, se dividieron las experimentaciones por Algoritmo. Se trató el Algoritmo K-MEANS en primera instancia, seguido EM, finalmente Clustering Jerárquico.

Se realizaron dos experimentaciones a modo de evaluación y entrenamiento en el manejo de la herramienta.

Aquí una descripción de las mismas.

En la primera prueba se observó que la variable tiempo ejercía una enorme influencia en el incremento de la suma de error cuadrático afectando la distribución de grupos en los tres cursos seleccionados, el tipo de dato al que correspondía esta instancia (Date) alteró la conformación de los conglomerados, por lo que la experimentación fue descartada. Partiendo de esta premisa se hizo cambios dentro de la etapa de pre procesamiento.

En el segundo experimento se decidió convertir a segundos el valor dado en el formato —hh:mm:ss, fundamentándose en que en primer lugar “los datos se someten a un proceso de estandarización” [6].

La segunda prueba si bien arrojó resultados con un menor índice de error cuadrático no proporcionaba los elementos suficientes para un análisis efectivo del comportamiento colaborativo en foros, a esto se le sumó la baja probabilidad (Algoritmo EM) .

En esta misma experimentación se abordó el uso de filtros para discretizar los atributos, resultando ser una técnica efectiva en el ordenamiento e identificación de tendencias, pero que debe ser usado de forma no recurrente pues su uso tiende a incrementar en el algoritmo K-Means la suma de errores cuadráticos.

Otro aspecto relevante que se comprobó con EM es que en el cambio en los parámetros de configuración tanto de iteraciones como número de semillas se mostró invariable la verisimilitud.

Como resultante de las dos primeras experiencias se efectuó una tercera tomando en cuenta las deficiencias de sus antecesoras

6.1. TERCERA EXPERIMENTACIÓN

Previa a la etapa de pruebas se preparan los datos en WEKA. Utilizando los filtros:

Numeric to Nominal: Para la conversión del atributo `sexo_usr`.

AddExpression: para determinar la mejora de un bimestre a otro que resulta en "mejoraBimestre" y el filtro

Discretize para nota_final: debido al rango de calificaciones amplio que posee y tomando en cuenta que un uso constante de este filtro incrementaría la suma de errores cuadráticos.

6.1.1. Fundamentos de la Programación

- Para este curso solo se han evaluado los foros de consultas.
- La calificación ha sido realizada únicamente por el docente —id: 5
- El tamaño máximo de los archivos adjuntos es de : 2MB
- No se ha forzado la suscripción a foros.
- El rastreo se ha desconectado.
- El canal rss para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- No se ha bloqueado el uso de los foros.

Los resultados tras la ejecución del algoritmo K-MEANS son los que se muestran en la Tabla 6.1

Atributo	Full Data (146)	0	1	2
<code>sexo_usr</code>	M (122) F (24)	M(32) F(10)	M(60) F(10)	M(30) F (4)
<code>num_acceso_foros</code>	30.2397	5.4524	37.6286	45.6471
<code>num_respuestas_post</code>	4.9452	0.7381	6.2714	7.4118
<code>num_resp_ect</code>	0.863	0	0.7571	2.1471
<code>numForos_subsc</code>	4.863	0.5	6.3857	7.1176
<code>prom_horas</code>	10.9932	7.4762	12.2429	12.7647
<code>prom_foros</code>	1.1445	0.0952	1.5303	1.6465
<code>nota_final</code>	(75-inf)*	(75-inf)*	(75-inf)*	(-inf-25]*
<code>mejoraBimestre</code>	0.461	0.0319	0.7986	0.2962

Tabla 6. 1 Salida Algoritmo K-Means curso Fundamentos de la Programación

Los grupos se han dividido: 42 para el cluster0, 70 para el cluster1, 34 para el cluster2 lo que se traduce en un 29%, 48% y 23% respectivamente.

La suma de errores cuadráticos disminuyó de 179,561936636949 en la primera experimentación a 93,10065988559433 en relación un 51.84% lo que implica un 48,16% de confiabilidad, menor que el 64,67% del segundo experimento pero considerándose los atributos adicionales que hacen del origen de datos una fuente más completa.

Para el Algoritmo EM los resultados se muestran en la Figura 6.1

```

EM
==
Number of clusters: 3

Attribute          Cluster
                   0      1      2
                   (0.1) (0.41) (0.49)
-----
sexo_usr
M                  14.8121 51.4375 58.7504
F                  1.0588 10.9304 15.0108
[total]           15.8709 62.3679 73.7612
num_acceso_foros
mean              44.3166 51.3864  9.7295
std. dev.        25.7602 28.0124  9.1772
num_respuestas_post
mean              6.2966  8.0419  2.079
std. dev.        2.9304  2.731  2.5441
num_mens_act
mean              4.252  1.1102  0
std. dev.        3.8555  1.5796  1.9883
numForos_subscr
mean              5.6696  2.2973  1.2101
std. dev.        2.9232  1.6742  2.1482
prom_foros
mean              1.3009  1.901  0.478
std. dev.        0.4952  0.103  0.5533
nota_final
'(-inf-25]'      9.0077 15.9552 14.0371
'(25-50]'       1.5969  3.2595  6.1436
'(50-75]'       1.004  5.996  6
'(75-inf]'      6.2623 39.1573 49.5804
[total]         17.8709 64.3679 75.7612
mejoraSemestre
mean              1.0682  0.0844  0.6605
std. dev.        1.0218  0.199  1.0046
  
```

Figura 6.1 Salida Algoritmo EM para Fundamentos de la Programación

El Cluster 2 denota un perfil con un mayor nivel colaborativo, le sigue el Cluster 1, finalmente el Cluster 0; tal cual sucedió con el K-Means.

El algoritmo EM se le conoce como K-Means Probabilístico, por cuanto se corresponde con los resultados iniciales de esta experimentación.

Finalmente para este curso se decidió ejecutar el algoritmo de Clustering Jerárquico. Figura 6.2

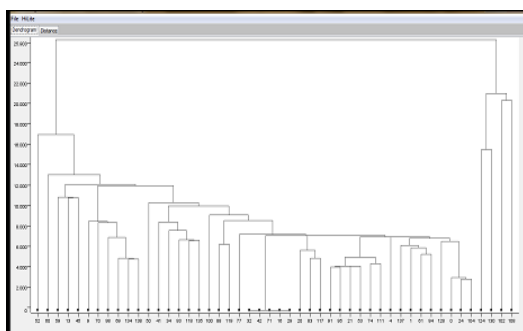


Figura 6.2 Salida Clustering Jerárquico curso Fundamentos de la Programación

Este algoritmo es especialmente útil si se requiere conocer el valor óptimo por el cual se debería agrupar los elementos, más a priori se ha definido a 3 el número de clústeres, distribuyéndose así:

Cluster 0: 2, Cluster 1: 44, Cluster 2: 2

6.1.2. Lógica de la Programación

Lógica de la Programación [B] posee foros tanto de tipo Preguntas y Respuestas (3) como de discusiones (2), 2 de los primeros son ejercicios prácticos el restante de consultas; de las discusiones una es de consulta, la otra práctica.

- Para este curso no existió calificación en los foros, según lo observado en el libro de calificaciones.
- El tamaño máximo permitido en archivos adjuntos es de: 5MB
- No se ha forzado la suscripción a foros.
- El rastreo se ha desconectado.
- El canal rss para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- No se ha bloqueado el uso de los foros.

La ejecución del algoritmo K-Means dio como resultado la Tabla 6.2

Atributo	Full Data (154)	0	1	2
sexo_usr	M (143) F (11)	M (39) F (2)	M (54) F (6)	M (60) F (3)
num_acceso_foros	12.6753	3.5366	23.5167	7.4717
subtemas_leidos	5.8506	0.6829	12.3667	2.4717
num_respuestas_post	0.5649	0.0488	1	0.4717
num_respuestas_debates	0.2078	0.0488	0.4333	0.0755
num_mens_act	0.2338	0.0488	0.4667	0.1132
arch_adjuntos	0.0844	0	0.2167	0
numForos_subscr	0.7532	0.1707	1.3833	0.4906
prom_horas	11.2078	1.122	12.5	17.5472
nota_final	(-inf-25]'	(-inf-25]'	(75-inf)'	(-inf-25]'

Tabla 6. 2 Salida del Algoritmo K-Means curso Lógica de la Programación

La suma de errores cuadráticos bajó de 148,7633869348921 a 34,6864615892839 de la primera a la segunda experimentación, en la actual llega a 77,402302380404 debido al cambio de escenario al realizar la agregación de atributos y la aplicación de filtros.

Los grupos se han dividido:

Cluster 0: 41, **Cluster 1:** 60, **Cluster 2:** 53 representando un 27%, 60% y 53% respectivamente.

El algoritmo EM produjo el resultado mostrado en la Figura 6.3

Attribute	Cluster		
	0	1	2
	(0.31)	(0.05)	(0.64)

sexo_usr			
M	45.325	7.6014	93.0736
F	4.4247	2.5753	7
[total]	49.7498	10.1766	100.0736
num_acceso_foros			
mean	24.1139	60.4576	3.1224
std. dev.	12.8808	34.8486	4.3286
subtemas_leidos			
mean	8.9196	50.3443	0.6469
std. dev.	9.0185	50.1406	1.2061
num_respuestas_post			
mean	1.5993	0.6555	0.651
std. dev.	1.2531	1.4421	0.22
num_respuestas_debates			
mean	0.4797	1.1124	0
std. dev.	0.6797	0.5888	0.5071
num_mens_act			
mean	0.5199	1.3668	0
std. dev.	0.8699	1.4708	0.6935
arch_adjuntos			
mean	0.0105	1.5285	0
std. dev.	0.1023	1.54	0.4977
numForos_subscr			
mean	1.7288	2.3612	0.1442
std. dev.	0.9764	0.4818	0.407
prom_horas			
mean	12.881	10.6176	10.4423
std. dev.	6.9272	6.5444	8.6957
nota_final			
* (-inf-25]'	13.9504	1.0496	65
* (25-50]'	3.9901	2.0099	6
* (50-75]'	8.4959	4.5041	9.9999
* (75-inf)'	25.3134	4.613	21.0736
[total]	51.7498	12.1766	102.0736

Figura 6. 3 Resultados algoritmo EM para el curso "Lógica de la Programación"

En este caso los grupos se han dividido así:

Grupo 0: 54 (35%)

Grupo 1: 9 (6%)

Grupo 2: 91 (59%)

Para Clustering Jerárquico la distribución de los alumnos en función del 33% de los datos destinados para la prueba es:

Grupo 0: 1 alumno

Grupo 1: 1 alumno

Grupo 2: 48 alumnos

El dendograma²¹ resultante es el que se muestra en la Figura 6.4. Si se hace un corte sobre el mismo se aprecia que a pesar de las numerosas gráficas se ha logrado agrupar estos elementos en 3 conglomerados.

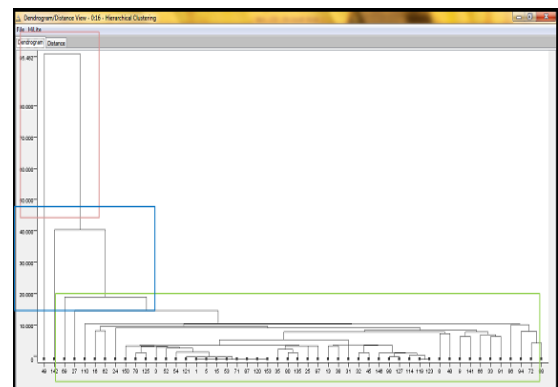


Figura 6. 4 Dendograma curso Lógica de la Programación

²¹ Es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis de clusters. [9]

6.1.3. Fundamentos Informáticos

Fundamentos Informáticos cuenta con 3 secciones, una de ellas contiene dos foros anidados, en total posee 4 foros, 1 de ellos es de tipo discusión el resto utiliza el formato Preguntas y Respuestas.

- Para este curso no se ha realizado evaluación alguna de foros.
- El tamaño máximo de los archivos adjuntos es de : 5MB
- No se ha forzado la subscripción a foros.
- El rastreo se ha desconectado.
- El canal rss para esta actividad son los mensajes donde se listaran los 10 artículos más recientes.
- El uso de foros no está bloqueado.

La salida luego de la ejecución del algoritmo K-Means es la mostrada en la Tabla 6.3

Atributo	Full Data (174)	0	1	2
sexo_usr	M (142) F (32)	M(49) F(9)	M(74) F(18)	M(19) F(5)
num_acceso_foros	4.5287	3.5172	0.2065	23.5417
subtemas_leidos	2.0632	1.6552	0.1087	10.5417
num_respuestas_post	0.546	0.2069	0.0109	3.4167
num_respuestas_debates	0.0287	0	0	0.2083
num_mens_act	0.0747	0	0	0.5417
arch_adjuntos	0.0115	0	0	0.0833
numForos_subscr	0.5	0.2931	0.0217	2.8333
prom_horas	6.8621	15.7759	0.5	9.7083
nota_final	(-inf-25]'	(-inf-25]'	(-inf-25]'	(-inf-25]'

Tabla 6. 3 Resultado del algoritmo K-Means para el curso Fundamentos Informáticos

Los estudiantes han sido divididos en tres grupos: el Cluster 0 cuenta con 58 estudiantes, el Cluster 1 con 92 y el Cluster 2 con 24, lo que equivale a un 33%, 92% y 24% respectivamente de 174 alumnos.

De 103,78037289349791 en la suma de errores cuadráticos de la primera experimentación se paso a 22,661696453425375 en la segunda a 87,64590519277633 en esta última.

Es notable la diferencia con la prueba inicial, mientras que con la siguiente si bien es menor el error cuadrático hay que tomar en cuenta el incremento de las variables y el uso de filtros que pudieran haber generado tal dilatación. Calificación final es menor a 25.

El resultado luego de la aplicación del algoritmo EM se indica en la Figura 6.5

```

Attribute          %cluster
                   0      1      2
                   (0)  (0.11) (0.89)
-----
SEXO_USR
M                   1.3552  14.9088  127.136
F                   1.0003  4.0003  29.9996
[total]            2.3554  20.9091  157.1356
num_acceso_foros
mean               0.7244  24.9314  3.9444
std. dev.          1.4787  17.2458  3.3105
subtemas_leidos
mean               0.2501  13.9213  0.4526
std. dev.          0.9575  5.1443  1.4171
num_respuestas_post
mean               0.0015  3.5946  0.1835
std. dev.          0.0383  1.393  0.4255
num_respuestas_debates
mean               0          0.2701  0
std. dev.          0.2263  0.6429  0.2263
num_mens_act
mean               0          0.7024  0
std. dev.          0.4303  1.1356  0.4303
arch_adjuntos
mean               0          0.1081  0
std. dev.          0.1516  0.4521  0.1516
numForos_subscr
mean               0.0081  2.7362  0.2343
std. dev.          0.0896  1.0221  0.4649
prom_horas
mean               4.6677  9.8544  4.5101
std. dev.          7.2934  7.4953  7.9419
nota_final
*(-inf-25]*        1.2702  11.8767  126.653
*(25-50]*          1.0025  1  3.9975
*(50-75]*          1.0112  2.6189  12.3499
*(75-100]*         1.0719  7.0132  15.9152
[total]            4.3554  22.5098  159.1356
  
```

Figura 6. 5 Salida Algoritmo EM para Fundamentos Informáticos

Los grupos se han formado así:

- Grupo 0: 29 (17%)
- Grupo 1: 27 (16%)
- Grupo 2 : 118 (68%)

Con un registro de verisimilitud de : - 10.19145

La revisión de las probabilidades obtenidas nos muestra que cualitativamente los resultados estarían dados de esta forma:

Cluster 2: Alumnos con nivel de colaboración Alto

Cluster 1: Alumnos con nivel de colaboración Medio

Cluster 0: Alumnos con nivel de colaboración Bajo

Finalmente el algoritmo de Clustering Jerárquico Figura 6.6

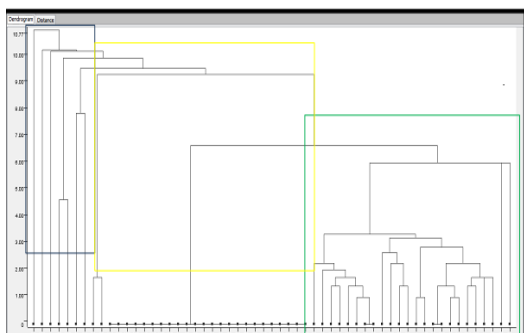


Figura 6. 6 Dendograma Curso Fundamentos Informáticos

El 33% de los datos utilizados para esta prueba se han distribuido así:

Cluster 0: 1

Cluster 1: 1

Cluster 2: 55

El dendograma resultante si se corta a un nivel de distancia dado, se obtiene una clasificación del número de grupos existentes a ese nivel y los elementos que los forman.

7. EVALUACION E INTERPRETACION

7.1. FUNDAMENTOS DE LA PROGRAMACIÓN

Los conglomerados se han calificado tomando en cuenta siete aspectos:

- Sexo
- Colaboración directa (num_respuestas_post, num_mens_act, numForos_subscr).
- Colaboración indirecta (num_acceso_foros, subtemas_leidos)
- Tiempo empleado en toda acción que implique Foros.
- Calificación promedio en Foros
- Mejora de Calificación del Primer al Segundo Bimestre.
- Nota final

En este curso las participaciones en el foro fueron indirectamente proporcionales a las calificaciones finales, los alumnos del primero grupo (0) no han ingresado de forma recurrente, pocos son los mensajes que se han agregado y actualizado.

El tiempo que han utilizado es mínimo en comparación con el resto de clusters, lo que podría reflejarse en un bajo promedio, diferente a su nota final superior (> 75 sobre 100).

Los estudiantes del segundo grupo (1) poseen un nivel medio de colaboración tienden a tener una calificación final alta (>75) en contraste con los que posee mayor nivel de colaboración, registran un tiempo ligeramente

menor que el del tercer grupo en sus

prácticas, estos estudiantes denotan una mejoría en el segundo bimestre en sus participaciones en foros.

Los del tercer 3 grupo (2) obtuvieron una alta participación en agregado y lectura de mensajes y cualquier acción que involucre foros, dedican más tiempo al cumplimiento de esta actividad, registran las más altas calificaciones en foros, alcanzaron notas

finales mínimas pero si registran un incremento en sus calificaciones de un bimestre a otro.

En la Tabla 7.1 se muestra los patrones con los que se desenvuelven los estudiantes de este curso.

PARÁMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directo	Bajo	Medio	Alto
Grado de Colaboración Indirecto	Bajo	Medio	Alto
Tiempo Empleado	Bajo	Medio	Alto
Promedio en Foros	Bajo	Medio	Alto
Mejora de Bimestre	Bajo	Alta	Medio
Calificación Final	Alta	Alta	Baja

Tabla 7. 1 Recopilación de resultados Fundamentos de la Programación

En función de la colaboración realizada los diferentes grupos se calificarían así:

Grupo 2: NIVEL DE COLABORACIÓN ALTA

Grupo 1: NIVEL DE COLABORACIÓN MEDIA

Grupo 0: NIVEL DE COLABORACIÓN BAJO

De un 100% de estudiantes de la materia de Fundamentos de la programación los alumnos con un nivel de colaboración alta representa un 29% (42), los de nivel de colaboración media un 48%(70) y los de bajo nivel colaborativo un 23%(34) de una población total de 146 alumnos. Figura 7.1



Figura 7.1 Gráfica por criterios de colaboración- Fundamentos de la Programación

7.2. LÓGICA DE LA PROGRAMACIÓN

Los grupos de este curso se calificaron tomando en cuenta cuatro aspectos: El Grado de colaboración directa, el grado de colaboración indirecta, el tiempo ocupado en esta actividad y la calificación final, la instancia prom_foros no se ha utiliza para este curso puesto que los foros no fueron valorados.

El primer grupo (0) el número de acceso a foros es mínimo, número mínimo de debates, pocos foros en los que se ha participado, número de mensajes mínimos agregados y actualizados, tiempo menor ocupado en foros, finalmente los miembros de este grupo son los que junto al tercer

grupo menor calificación poseen (< 25). Este grupo corresponde al de la minoría de estudiantes.

El segundo grupo (1) por su parte cuenta con un mayor número de acceso a foros, número mayor de debates, mayor número de foros en los que se ha participado, actualizado y en los que se ha adjuntado un archivo, número menor que el grupo 2

pero mayor que el 0 de mensajes agregados, tiempo promedio en actividad de foros, nota final mucho mayor que el primer grupo (0). Son el grupo de mayor representación.

El tercer grupo (2) poseen un acceso promedio a foros, así mismo de discusiones

leídas y mensajes creados, aunque los debates se ubican en un punto medio, los estudiantes no han subido archivos pero si se han suscrito a la mayoría de foros y han ocupado un tiempo importante en el cumplimiento de esta actividad. Su nota final se ubica en el rango de menor a 25. Tabla 7.2

PARAMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directa	Bajo	Alto	Medio
Grado de colaboración indirecta	Bajo	Alto	Bajo/Nulo
Tiempo Empleado	Bajo	Medio	Alto
Calificación Final	Bajo	Alto	Bajo

Tabla 7. 2 Recopilación de resultados Lógica de la Programación

El total de alumnos para este curso es 154. Luego del análisis realizado se puede determinar qué: El

porcentaje de estudiantes con nivel de colaboración alta es del 39% lo que resulta en 60 individuos, los de colaboración media con una representatividad del 34% es decir 53 y el de colaboración bajo un 27% significando 41 estudiantes. Figura 7.2



Figura 7.2 Gráfica por criterios de colaboración- Lógica de la Programación

7.3. FUNDAMENTOS INFORMÁTICOS

Al igual que Lógica de la Programación los parámetros que se revisaran serán: Grado de colaboración directa, Grado de colaboración indirecta, tiempo empleado en cualquier actividad que implique foros y calificación final, a excepción de la instancia debates que si bien existen entradas estas son mínimas.

Los grupos cumplen con las siguientes características:

El primer grupo (0) cuenta con un número de acceso a foros menor que el Grupo 2 pero mayor que el Grupo 1, el número de discusiones leídas es mayor que el Grupo 1 y bastante menor con respecto al

Grupo 2, lo que se repite para el número de mensajes creados y al de foros suscritos, tanto la participación en debates, mensajes actualizados y archivos adjuntos es Nula, el tiempo utilizado en este grupo es mayor que el del resto.

El segundo grupo (1) cuenta con un bajo nivel colaborativo tanto en el número de acceso en foros, como en las discusiones leídas, el número de mensajes creados, los foros a los que se han suscrito y el promedio de horas que ocupan en los foros. Las discusiones creadas, los mensajes actualizados y los archivos adjuntos poseen un estado nulo la nota final es menor a 25 sobre 100.

El tercer grupo (2) registra un mayor número de acceso a los foros, son los únicos que crean y revisan discusiones estos estudiantes se han suscrito a un mayor número de foros y han participado positivamente en la creación de estos. A diferencia de los otros grupos estos alumnos han adjuntado archivos a sus respuestas pero no han dedicado mayor tiempo que el Grupo 0 en esta actividad. Cuentan con una calificación menor a 25 sobre 100.

La Tabla 7.3 recopila estos resultados:

PARAMETRO	CLUSTERS		
	0	1	2
Grado de colaboración directa	Medio	Bajo	Alto
Grado de colaboración indirecta	Medio	Bajo	Alto
Tiempo Empleado	Alto	Bajo	Medio
Calificación Final	Alto	Bajo	Medio

Tabla 7. 3 Recopilación de resultados Fundamentos Informáticos

El grado de colaboración se basa en la siguiente escala cualitativa resultado de la experimentación:

Grupo 2: NIVEL DE COLABORACIÓN ALTA

Grupo 0: NIVEL DE COLABORACIÓN MEDIA

Grupo 1: NIVEL DE COLABORACIÓN BAJO

El total de la población en este curso es de: 174

Quienes cuenta con un nivel de colaboración alta representan un 14% (24) total de la población, los de mediano nivel colaborativo un 33% (58) y aquellos con niveles mínimos de colaboración representan una mayoría con un 53% (92). Figura 7.3



Figura 7.3 Gráfica por criterios de colaboración- Fundamentos Informáticos

Cuando los estudiantes no se ven forzados a participar en los foros por lo general no muestran motivación para realizarlo. Ninguno de los foros han sido bloqueados es decir no se ha delimitado el tiempo en el que estarán disponibles lo que implica el descuido por parte de los educandos.

Otro patrón importante que se ha observado es que solo en el curso Fundamentos de la Programación los foros han sido evaluados implicando que los estudiantes se esfuercen por participar de una forma más contundente.

Al menos debe existir 1 foro por bimestre cuando es así los resultados de la colaboración se incrementan cuando son más de 1 la colaboración se reparte entre todos los existentes.

Finalmente el comportamiento colaborativo de los estudiantes está íntimamente ligado a factores como el tiempo disponible, la predisposición y los recursos utilizados de la plataforma.

8. CONCLUSIONES

- El nivel colaborativo de los estudiantes no es proporcional a su calificación final, influyen otros factores como: calificaciones de otras actividades y exámenes.
- El tiempo que registran en la actividad foros pudiera no utilizarse con fines colaborativos.
- Aunque la cantidad de estudiantes del sexo femenino sea minoritaria, sus colaboraciones se encuentran en el mismo nivel que los del masculino
- Las capacidades tecnológicas y actitudes del docente en un entorno a distancia son preponderantes en el rendimiento académico de sus educandos es así que se debe contar con capacitaciones periódicas de tal forma que se aproveche al máximo los recursos con los que cuenta la plataforma educativa.
- Categorizar a los estudiantes por su nivel de colaboración permite a los docentes centrarse en aquellos alumnos que necesitan mayor atención y soporte. La retroalimentación que se realice no solo debe señalar puntos negativos de una conducta sino también reforzar la actitud comprometida de los estudiantes y las mejoras que puedan tener dentro de un periodo de tiempo.
- El trabajo colaborativo es sin duda el mayor apoyo con el que puede contar un estudiante, puntos de vista diferentes o similares permiten la existencia de debates que enriquecen el pensamiento analítico y crítico de los alumnos.
- La etapa de Pre Procesamiento en Minería de Datos constituye un 50% mínimo del total de un proyecto, esta fase aún terminada si no cumple con las expectativas de la Minería deberá ser revisada y cambiada cuantas veces sea necesario.
- La búsqueda de nuevas instancias incrementaría la eficiencia en la formación de grupos al maximizarse el número de similitudes a evaluarse.
- Generalmente las participaciones que se ubican en un punto medio son las que tienen mayor predisposición para obtener una calificación alta.
- El clustering al categorizarse como descriptivo fue la técnica que mejor se adapta para la realización de esta investigación, el reunir grupos por características colaborativas similares es el punto focal de este trabajo.
- K-Means fue la técnica que mejor se adaptó a los objetivos de este trabajo por la celeridad en la conformación de grupos especialmente cuando la población no es de gran tamaño como fue en este caso.

- Tanto EM como el Clustering Jerárquico sirvieron de complemento a K-Means para el análisis de los grupos debido a su naturaleza probabilística y subjetiva en ese orden.
- Los estudiantes mostraron de forma global un bajo interés colaborativo en gran parte de ellos fue nulo, dando a entrever falta de motivación o habilidad para la ejecución de esta actividad.

9. RECOMENDACIONES

- Se recomienda la capacitación de los estudiantes de los primeros ciclos acerca del uso de los foros en el EVA de tal manera que se le dé el seguimiento pertinente, identificándose los errores más comunes que pudiera cometer y que obstruyeran la ejecución plena de sus actividades.
- Ampliar el campo de investigación de las habilidades sociales en el EVA a fin de optimizar su uso y generar nuevos conocimientos.
- Se podría experimentar además con la formación de grupos de trabajo con la capacidad de calificarse entre sí cada una de sus contribuciones generando un ambiente colaborativo activo.
- Así también para la experimentación se recomienda hacer todas las variaciones posibles de modo que se obtenga los resultados más fiables.
- El uso de atributos numéricos como nominales para una representación de los datos más cercana a la realidad.
- Se recomienda la evaluación de las herramientas de Data Mining que mejor se adapten a los objetivos planteados en la investigación.

10. TRABAJOS FUTUROS

- Se sugieren los siguientes trabajos que en un futuro podrían realizarse en el ámbito de comportamientos colaborativos en foros.
- La realización de la minería de datos en diferentes periodos de tal forma que se compruebe si los patrones se repiten.
- Minería de Texto aplicada a los mensajes valorándolos cualitativamente bajo criterios de: conocimiento del tema y relevancia.
- La creación de grupos de trabajo formados aleatoriamente a quienes se le asigne una tarea específica con la capacidad de que sus miembros puedan calificarse entre sí, el objetivo de esto es medir el grado de colaboración de los estudiantes y su capacidad para trabajar en equipo.

11. REFERENCIAS

- [1] Romero, C., Ventura, S., & García, E. Data mining in course management systems: Moodle case study and tutorial. University of Córdoba, Department
- [2] Corso, C., & Alfaro, S. (2010). Algoritmos de Data Mining aplicados en la enseñanza basada en la Web. Universidad Tecnológica Nacional, Departamento de Sistemas de Información, Córdoba.
- [3] Baker, R., & Yacef, K. (2009). The State of Educational Data Mining in

2009: A Review and Future Visions. 1-14.

[4] Romero, C., & Ventura, S. (2010). Educational Data Mining: A Review of the State of the Art. IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS.

[5] Hernandez, J., Ramirez, M., & Ferri, C. (2004). Introducción a la Minería de Datos. Pearson.

[6] Molina, J., & García, J. (2004). Técnicas de Análisis de Datos. Universidad Carlos III de Madrid, Madrid.

[7] Garre, M., Cuadrado, J. C., & Sicilia, M. (2005). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. Universidad de Alcalá, Departamento de Ciencias de la Computación, Alcalá de Henares, Madrid.

[8] Figueras, S. (2001). Análisis de conglomerados o cluster.

[9] Vicente, J. (2006). Introducción al Análisis del Clusters. Universidad de Salamanca, Departamento de Estadística.