



ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

PROYECTO DE DETECCIÓN DE PATRONES DE PARTICIPACIÓN EMPLEANDO MINERÍA DE DATOS EN EL ENTORNO VIRTUAL DE APLICACIONES WEB DE LA ESPOCH, PARA PREDECIR ESTUDIANTES EXITOSOS.

AUTOR: GUSTAVO XAVIER HIDALGO SOLÓRZANO

TUTOR: ING. FRANKLIN MORENO

Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo, presentado ante el Instituto de Postgrado y Educación Continua de la ESPOCH, como requisito parcial para la obtención del grado de:

MAGÍSTER EN FORMULACIÓN, EVALUACIÓN Y GERENCIA DE PROYECTOS PARA EL DESARROLLO

RIOBAMBA – ECUADOR

Junio 2016

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

CERTIFICACIÓN

EL TRIBUNAL DE TESIS CERTIFICA QUE:

El trabajo de titulación, titulado “PROYECTO DE DETECCIÓN DE PATRONES DE PARTICIPACIÓN EMPLEANDO MINERÍA DE DATOS EN EL ENTORNO VIRTUAL DE APLICACIONES WEB DE LA ESPOCH, PARA PREDECIR ESTUDIANTES EXITOSOS.”, de responsabilidad del Sr Gustavo Xavier Hidalgo Solórzano ha sido prolijamente revisado y se autoriza su presentación.

Tribunal de Tesis:

Ing. Verónica Mora. Msc.

PRESIDENTE

Ing. Franklin Geovanni Moreno Montenegro. Msc.

TUTOR DE TESIS

Ing. Gonzalo Nicolay Samaniego Erazo, Ph.D.

MIEMBRO

Ing. Jorge Ernesto Huilca Palacios. Msc.

MIEMBRO

COORDINADOR SISBIB ESPOCH

Riobamba, Junio del 2016

DERECHOS INTELECTUALES

Yo, GUSTAVO XAVIER HIDALGO SOLÓRZANO, declaro que soy responsable de las ideas, doctrinas y resultados expuestos en el **Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo**, y que el patrimonio intelectual generado por la misma pertenece exclusivamente a la Escuela Superior Politécnica de Chimborazo.

GUSTAVO XAVIER HIDALGO SOLÓRZANO
No. 080211622-8

DECLARACIÓN DE AUTENTICIDAD

Yo, GUSTAVO XAVIER HIDALGO SOLÓRZANO, declaro que el presente **Trabajo de Titulación modalidad Proyectos de Investigación y Desarrollo**, es de mi autoría y que los resultados del mismo son auténticos y originales. Los textos constantes en el documento que provienen de otra fuente están debidamente citados y referenciados.

Como autor/a, asumo la responsabilidad legal y académica de los contenidos de este proyecto de investigación de maestría.

Riobamba, 1 de Junio de 2016

GUSTAVO XAVIER HIDALGO SOLÓRZANO
No. 080211622-8

DEDICATORIA

Dedico este trabajo con mucho orgullo y cariño a mi esposa Vanessa a mis Hijos Gustavo y Diego.

Gustavo Xavier

AGRADECIMIENTO

A Dios por permitirme con salud culminar esta etapa.

A mi esposa por ser un valioso e incondicional apoyo.

A mis padres por motivarme a seguir cumpliendo metas y trazando objetivos.

Al Ing. Franklin Moreno, Dr. Nicolay Samaniego, Ing. Jorge Huilca y por su apoyo incondicional, asesoramiento y ayuda acertada en el desarrollo de la presente investigación, por su tiempo, dedicación y paciencia; por orientarme para la presentación de la misma.

Gustavo Xavier

ÍNDICE GENERAL

ÍNDICE DE TABLAS.....	i
ÍNDICE DE GRÁFICOS.....	ii
ÍNDICE DE IMÁGENES.....	iii
RESUMEN.....	iv
SUMARY.....	v
CAPÍTULO I	
1.1 Introducción.....	1
1.2 Planteamiento del problema.	2
1.3 Formulación del problema.	4
1.4 Sistematización del problema.	5
1.5 Objetivos de la investigación.	6
1.5.1 <i>Objetivo general.</i>	6
1.5.2 <i>Objetivos específicos.</i>	6
1.6 Justificación de la investigación.	6
1.7 Hipótesis.	8
CAPITULO II	
2.1 Antecedentes y estudios previos.....	9
2.2 Fundamentación Teórica de las técnicas de Minería de Datos.....	10
2.3 Generalidades.....	12
2.3.1 <i>Datos</i>	12
2.3.2 <i>La información</i>	13
2.3.3 <i>El conocimiento</i>	14
2.3.4 <i>Data WareHouse</i>	15
2.3.4.1 <i>Proceso de Extracción, Transformación y Carga</i>	16
2.3.4.2 <i>Modelo dimensional</i>	17
2.3.5 <i>Minería de datos</i>	20
2.3.5.1 <i>Proceso de Minería de Datos (DataMining)</i>	21
2.3.5.2 <i>Técnicas de Minería de Datos (DataMining)</i>	23
2.3.5.2 <i>Trabajos relacionados</i>	33
CAPITULO III	
3. MATERIALES Y MÉTODOS.....	36
3.1 Diseño de la investigación.....	36
3.2 Tipo de Investigación.....	37
3.3 Población.....	37
3.4 Muestra.....	38
3.5 Método.....	39
3.6 Técnica e instrumentos.....	40
3.7 Procesamiento y Análisis de datos.....	41
3.7.1 <i>Metodología</i>	42
3.7.1.1 <i>Determinar los patrones de participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web de la ESPOCH</i>	43
3.7.1.2 <i>Identificar el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web, compararlas con las calificaciones obtenidas en las evaluaciones acumulativas.</i>	76

3.7.1.3	<i>Determinar si el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web ayudan a que los estudiantes sea exitosos.....</i>	77
3.8	Variables e indicadores.....	79
3.9	Instrumentos.....	80
CAPITULO IV		
4.	RESULTADOS Y DISCUSIÓN.....	81
4.1	Análisis e interpretación de resultados.....	81
4.1.1	Indicadores de la variable dependiente.....	81
4.1.1.1	<i>Número de participaciones por estudiantes.....</i>	81
4.1.1.2	<i>Porcentaje de calificaciones de los estudiantes.....</i>	83
4.1.1.3	<i>Número de actividades o módulos.....</i>	85
4.1.1.4	<i>Número de participación en las actividades o módulos.....</i>	85
4.1.1.5	<i>Porcentaje de participación en las actividades o módulos.....</i>	87
4.1.2	Indicadores de la variable independiente.....	89
4.1.2.1	<i>Porcentaje de estudiantes que logran terminar el semestre con éxito.</i>	89
4.1.2.2	<i>Porcentaje de Falsos Positivos del algoritmo seleccionado.....</i>	89
4.1.2.3	<i>Porcentaje de Verdaderos Positivos del modelo seleccionado.....</i>	90
4.1.2.4	<i>Porcentaje de aprendizaje del modelo.....</i>	90
4.2	Presentación de Resultados.....	91
4.2.1	<i>Análisis de la trama.....</i>	93
4.3	Predicción.....	95
4.3.1	<i>Presentación de individuos nuevos.....</i>	96
4.3.2	<i>Creación del modelo predictivo.....</i>	98
4.4	Prueba de la hipótesis de investigación.....	100
4.4.1	<i>Planteamiento de la hipótesis.....</i>	101
4.4.2	<i>Nivel de significancia.....</i>	102
4.4.3	<i>Criterios de validación de la hipótesis.....</i>	102
4.4.4	<i>Toma de decisiones.....</i>	103
CONCLUSIONES		104
RECOMENDACIONES.....		105
BIBLIOGRAFÍA.....		
ANEXOS.....		

ÍNDICE DE TABLAS

Tabla 1-3:	Modelo de evaluación por asignatura.....	38
Tabla 2-3:	Cursos Virtuales que utilizan el modelo de evaluación.....	39
Tabla 3-3:	Técnicas e instrumentos.....	41
Tabla 4-3:	Entidades.....	45
Tabla 5-3:	Descripción de las fuentes de datos.....	46
Tabla 6-3:	Descripción de la dimensión dim_assign.....	54
Tabla 7-3:	Descripción de la dimensión dim_quiz.....	55
Tabla 8-3:	Descripción de la dimensión dim_notas_participacion.....	56
Tabla 9-3:	Descripción de la dimensión dim_roleassignment.....	56
Tabla 10-3:	Descripción de la dimensión hecho_calificaciones.....	57
Tabla 11-3:	Descripción de calificaciones de tareas.....	61
Tabla 12-3:	Descripción de calificaciones de los cuestionarios.....	62
Tabla 13-3:	Participación consolidada.....	76
Tabla 14-3:	Variables dependientes.....	79
Tabla 15-3:	Variables independientes.....	79
Tabla 1-4:	Participación por individuo.....	82
Tabla 2-4:	Porcentaje de calificaciones.....	84
Tabla 3-4:	Participaciones en módulos.....	86
Tabla 4-4:	Consolidado de participación.....	87
Tabla 5-4:	Entradas y salidas de la red neuronal.....	94
Tabla 6-4:	Estudio de EVAs en periodo actual.....	96
Tabla 7-4:	Individuos nuevos.....	100
Tabla 8-4:	Análisis y Decisiones	104

ÍNDICE DE FIGURAS

Figura 1-2: Pirámide del conocimiento.....	14
Figura 2-2: Arquitectura de un DataWare House.....	15
Figura 3-2: Proceso ETL.....	17
Figura 4-2: Estructura estrella.....	18
Figura 5-2: Estructura copo de nieve.....	18
Figura 6-2: Tabla de hechos calificaciones.....	19
Figura 7-2: Dimensiones.....	19
Figura 8-2: Metodología ETL.....	22
Figura 9-2: Árbol de decisión.....	27
Figura 10-2: Esquema redes neuronales.....	29
Figura 1-3: Proceso de Minería de Datos.....	44
Figura 2-3: Modelo Seleccionado para el análisis.....	47
Figura 3-3: Modelo estrella para el Data Warehouse.....	53
Figura 4-3: Análisis de los modelos en la curva ROC, Weka	71
Figura 5-3: Análisis de los modelos en la curva ROC, código R.....	72
Figura 6-3: Participación por individuo.....	78
Figura 1-4: Participación por individuos.....	83
Figura 2-4: Porcentaje de calificaciones.....	85
Figura 3-4: Participaciones en módulos.....	86
Figura 4-4: Participación por módulos.....	88
Figura 5-4: Total de participación.....	88
Figura 6-4: Estado de estudiantes.....	89
Figura 7-4: Datos SPSS.....	102
Figura 8-4: Campana de Gauss.....	103
Figura 9-4: Participación vs, Predicción.....	103

ÍNDICE DE CUADROS

Cuadro 1-3: Comparación de capacidades ETL.....	48
Cuadro 2-3: Comparación de capacidades analíticas.....	48
Cuadro 3-3: Comparación de capacidades de reportaría.....	48
Cuadro 4-3: Comparación de capacidades tableros de mando.....	49
Cuadro 5-3: Comparación de entidad de información provista.....	49
Cuadro 6-3: Evaluación final.....	49

ÍNDICE DE IMÁGENES

Imagen 1-3: Inicio.....	51
Imagen 2-3: Variables de entorno.....	52
Imagen 3-3: Variables de entorno.....	52
Imagen 4-3: Configuración Variables de entorno.....	52
Imagen 5-3: Crear Base de datos.....	58
Imagen 6-3: Nombre de Base de datos.....	58
Imagen 7-3: Ejecutar PDI.....	59
Imagen 8-3: Entorno de trabajo PDI.....	59
Imagen 9-3: Dim_assign.....	60
Imagen 10-3: Dim_quiz.....	60
Imagen 11-3: Dim_roleassignment.....	61
Imagen 12-3: dim_tmp_assign.....	63
Imagen 13-3: hec_calificaciones2.....	63
Imagen 14-3: hec_notas1.....	64
Imagen 15-3: dim_notas_assign.....	65
Imagen 16-3: Dim_notas_quiz.....	65
Imagen 17-3: dim_notas_participación.....	66
Imagen 18-3: hecho_calificaciones.....	67
Imagen 19-3: Conexión weka con tabla de hechos.....	73
Imagen 20-3: Detección de patrones.....	74
Imagen 21-3: Detección de patrones en foros.....	75

RESUMEN

Esta investigación define un modelo de predicción basado en el algoritmo de Minería de Datos Perceptron Multica, fue seleccionado mediante la curva ROC; de esta manera se utiliza datos históricos del Entorno Virtual de Aprendizaje (EVA) de la materia de Aplicaciones Web de noveno semestre de la Escuela de Ingeniería en Sistemas. Se aplicó en la obtención de las fuentes de datos, pre procesamiento de datos, extracción y limpieza, creación del modelo, interpretación de resultados. Se utilizó los datos fuentes Las fuentes de la Plataforma Virtual MOODLE de la ESPOCH y el Sistema Académico; de la misma forma el pre-procesamiento se realizó en el motor PostgreSQL, se crearon dos Data Mart; utilizando un modelo dimensional relacional tipo estrella, donde se crearon las tablas de dimensiones y hechos con sus respectivas medidas, el resultado del análisis es la información contenida en la tabla de hechos, sometido al weka y R para identificar los patrones y su posterior predicción. Los datos de Didáctica informática la Escuela de Ingeniería en Sistemas del actual semestre, fue sometido al modelo de predicción para evaluar el modelo. El algoritmo proporciona es de 90% de certeza, con esta predicción se aceptó la hipótesis nula.

Palabras Clave: <MOODLE><ALGORITMOS PREDICTIVOS><MINERÍA DE DATOS><PERCEPTRON MULTICAPA><APRENDIZAJE NO SUPERVIZADO>

SUMMARY

This research defines a prediction model based on the algorithm for, data mining Multilayer Perceptron was selected by the ROC curve; in this way it is used, historical data of Virtual Learning Environment (VLE) of the subject matter: Web Application in ninth semester at the school of Virtual Platform MOODLE of ESPOCH and the Academic System; in the same way the pre-processing was performed in the PostgreSQL engine two Data Mart were created; using a star-relational dimensional model, where the dimension tables and made with their respective measures were created, the result of the analysis is the information contained in the fact table, submitted to weka and R to identify patterns and its subsequent prediction. Teaching computer data School of Systems Engineering of the current semester, it was submitted to the prediction model to evaluate the model. The algorithm provides of 90% certainty, with this prediction the null hypothesis was accepted.

Keywords: <MOODLE> <PREDICTIVE ALGORITHMS> <DATA MINING> <MULTILAYER PERCEPTRON> <UNSUPERVISED LEARNING>

CAPÍTULO I

1.1 Introducción

Esta investigación tiene el objetivo de proponer un modelo en algoritmos predictivos no supervisados, basado en minería de datos para que el profesor de enseñanza superior, pueda detectar patrones de participación y notas de los estudiantes con el fin de detectar los estudiantes que van a tener éxito en una cátedra de un determinado semestre.

En el primer capítulo se describe el problema del desconocimiento de patrones que los estudiantes dejan cuando trabajan con los Entornos Virtuales de Aprendizaje, así como también los objetivos que se pretenden alcanzar con este trabajo de tesis. De esta manera se plantea la hipótesis que se pretende comprobar en este trabajo de investigación, trabajo que se lleva a cabo con los estudiantes de la Facultad de Informática y Electrónica de la Escuela Superior Politécnica de Chimborazo.

En el segundo capítulo muestra los conceptos y teoría relacionados con la herramienta de minería de datos, los modelos y algoritmos de predicción existente, para poder entender la forma como se estructura dicho modelo en base al algoritmos propuesto. De la misma manera se describe Moodle como plataforma virtual que soporta la aplicación de algunos procesos educativos, entre ellos la evaluación, además de ser la plataforma utilizada en la Escuela Superior Politécnica de Chimborazo para la gestión y administración de material educativo y a la cual tienen acceso todos los estudiantes y profesores.

El capítulo tres muestra el proceso investigativo de este trabajo, indicando el tipo de investigación aplicada, la descripción de la muestra tomada, los ambientes utilizados para la ejecución del trabajo, en el proceso llevado a cabo para ejecutar la investigación. En este capítulo se propone los algoritmos de predicción que será seleccionado por intermedio de la curva ROC y el área bajo la curva ROC, ya que con esta herramienta podremos tomar el algoritmo que mejor se adapte al modelo en base a los datos que se presenta y así poder responder a la hipótesis.

En el capítulo se exponen los resultados obtenidos en el proceso investigativo, mediante tablas y gráficos se realiza una interpretación de dichos resultados; además, con la aplicación de procesos estadísticos se explica cómo se llegó a la comprobación de la hipótesis.

1.2 Planteamiento del problema.

Diariamente se genera gran cantidad de datos en las tareas cotidianas que se realiza, por ejemplo; cuando caminan, corren o practican algún deporte, se está generando datos como: km recorridos, número de pulsaciones de nuestro corazón, cantidad de fluidos corporales, tiempos de recorridos por km, entre otros.

El mundo digital es el más grande repositorio de datos que existe, ya que constantemente se genera información gracias a herramienta de comunicación como el chat en el telefonía móvil, navegar por el Internet, sistemas de procesamiento de información, entornos virtuales de educación, entre otros.

“Estos datos están dispersos por todas partes ya sean en documentos, archivos o bases de datos; el crecimiento de los sistemas de bases de datos se debe en gran medida, gracias a la gran cantidad de datos que se generan por segundo; en estos tiempos ya no se habla de almacenamientos en Megabytes ni Gigabytes sino en Terabyte, dentro de 10 años estaremos hablando de Petabyte y en 5 años más de Exabyte, En 2002, hubieron unos cinco Exabytes de datos online”. (Hagggar, 2011, P.8).

En el 2009 el total se incrementó a 281 Exabytes, una tasa de crecimiento de 56 veces en siete años. La cantidad total de datos almacenados por empresas se está duplicando cada tres años. (Hagggar, 2011, P.8).

Los datos que se originan diariamente por la interacción de las personas; cuando se someten a una la interpretación son transformados en información, luego cuando un especialista interpreta la información se convierte en conocimiento, y es ahí donde se agrega valor a los datos.(Elearn Training Company, 2007).

La educación no puede estar alejada de este nuevo paradigma de evolución del conocimiento, con la utilización de redes sociales, plataformas de educación en línea o virtual, e-commerce, e-goberment, entre otros.

La Escuela Superior Politécnica de Chimborazo (ESPOCH), genera una gran cantidad de información de estudiantes y docentes gracias a las aplicaciones o sistemas que poseen, entre ellos están: el Sistema Académico Institucional OASIs (academico.espoch.edu.ec), la Plataforma de Educación Virtual EVIRTUAL-ESPOCH (evirtual.espoch.edu.ec), el Sistema de Recursos Humanos (recursos.espoch.edu.ec), el Sistema de Evaluación Institucional (evaluacion.espoch.edu.ec), el

Sistema Médico (medicina.espoch.edu.ec), Sistema de Biblioteca (biblioteca.espoch.edu.ec), entre otros.

La Institución carece de un medio, técnica o método de minería de datos para el análisis de estos datos, limitando en gran parte el conocimiento de este análisis y por ende impidiendo que la toma de decisiones sea efectiva para la gestión de recursos.

“La Minería de Datos educacionales es una disciplina emergente dedicada a desarrollar métodos para analizar una gran cantidad de datos provenientes de ambientes relacionados a la educación, con el objetivo de entender los datos que generan los estudiantes, profesores y otros actores relacionados a sus entornos educacionales”. (Molen. 2013,p.10).

La Minería de datos se clasifica en 2 categorías; métodos supervisados o predictivos y métodos no supervisados o descubrimiento del conocimiento. Entre los supervisados tenemos: árbol de decisión, inducción neuronal, regresión, series temporales; mientras que en los no supervisados tenemos: detecciones de desviaciones, segmentación, agrupamiento o cluster, reglas de asociación, patrones secuenciales, a estos dos últimos también se los conoce como análisis de asociación.

En el ámbito académico de la ESPOCH, tenemos: el Sistema Académico Institucional y la Plataforma de Educación Virtual; estos sistemas tiene en sus bases de datos una extensa y valiosa información de cada uno de los estudiantes, que no se ha estudiado y/o analizado en conjunto.

La cátedra de Aplicaciones Web, forma parte del quinto semestre de la Escuela de Ingeniería en Sistemas Informáticos de la Facultad de Informática y Electrónica, es una de las materias profesionalizantes de la carrera y cuenta con una amplia información y datos en las bases de datos históricas tanto en la Plataforma Virtual de Educación como en el Sistema Académico Institucional para realizar esta investigación.

La carencia de patrones de participación; se convierte en un problema para la Institución, ya que no existe un medio, método o modelo para identificar a los estudiantes NO exitosos; de esta manera los Docentes no tienen los fundamentos reales necesarios para tomar decisiones con respecto a este grupo de estudiantes y por ende el índice de deserción estudiantil se incrementa gradualmente cada semestre.

1.3 Formulación del problema.

La Escuela Superior Politécnica de Chimborazo no cuenta con un estudio adecuado para la detección de patrones de participación que permita a los docentes e investigadores definir indicadores de los estudiantes exitosos, para de esta manera identificar a los estudiantes no exitosos y poder tomar decisiones antes de que fracasen en el semestre en curso.

La ESPOCH, cuenta con un reglamento aprobado desde el año 2009, indica que existen 3 tipos de evaluaciones que son: acumulativas, finales y suspensión. La evaluación acumulativa corresponderá a pruebas parciales teóricas y/o prácticas, lecciones, trabajos de investigación y más parámetros de evaluación edumétrica, establecidos en el cronograma de actividades del docente.

La evaluación será procesual, sistemática y continua e implicará la valoración de conocimientos, destrezas, habilidades y actitudes, por lo que la calificación global se determinará de la siguiente manera:

La calificación de evaluación acumulativa, 70%, constituida por pruebas, lecciones, trabajos de investigación y más parámetros de evaluación edumétrica. Que no deberán ser menores de tres (3) componentes;

La calificación de evaluación final, 30%;

El estudiante que en la evaluación acumulativa reuniera el 90% de la calificación, veinte y cinco (25) puntos, será exonerado de rendir la evaluación final y se considerará aprobado; para lo cual se sumará al valor de la evaluación final, doce (12) puntos.

Las calificaciones se contabilizarán en la escala de cero (0) a veintiocho puntos (28), en las evaluaciones acumulativas y de 0 a 12 puntos en la evaluación final; serán siempre en cifras correspondientes a números enteros; la fracción cero punto cinco (0.5) ó más, se aproximará a la cifra inmediata superior.

Una de las actividades edumétricas que los docentes de la ESPOCH han adoptado es la participación en los Entornos virtuales de Aprendizaje, estas herramientas facilitan la labor de aprendizaje tanto para docentes como para los estudiantes, tanto así que los docentes destinan un porcentaje de la nota total del curso a la participación de los estudiantes en los EVAs.

Además de ser una de las herramientas más utilizadas, captan una gran cantidad de información tanto del docente como de los estudiantes; esta información en los actuales momento no brinda el valor agregado que debería, ya que la ESPOCH carece de un modelo para extraer esa información estudiarla y compararla con otras fuentes de datos como el Sistema académico para convertirla en conocimiento a beneficio de la institución y de los mismos estudiantes.

La Minería de Datos (Data Mining) y sus métodos supervisados junto con las técnicas de minería como: árbol de decisión, agrupamientos o Clustering entre otros; ayudan a modelar esta información para detectar los patrones de participación de los estudiantes exitosos en los Entornos Virtuales de Aprendizaje (EVA) de la Plataforma de Educación Virtual de la Escuela Superior Politécnica de Chimborazo; así los especialistas puedan analizar y evaluar esta información para de esta manera generar conocimiento y poder tomar decisiones que en el área de la academia podría ser denominada Analítica de Aprendizaje. (Moreno. 2009,p.16).

Pero ¿Cómo la Minería de Datos podría ser utilizada para detectar patrones de participación para la predicción de estudiantes exitosos.?. La respuesta a estas preguntas invita a plantearnos una hipótesis que permita conseguir resultados que ayuden a entender y manipular los datos, información y convertirlos en conocimiento.

1.4 Sistematización del problema.

La institución está interesada en conocer y entender la información que puede brindar sus Aplicaciones o Sistemas Académicos con los que cuenta; en tal virtud se acude a las técnicas de minería de datos para entender el comportamiento participativo de los estudiantes y conocer si estos patrones influyen en el éxito de los estudiantes. Por ello es importante identificar las variables que presenta este problema.

Partiendo del problema planteado surgen las siguientes interrogantes:

¿Qué impacto tendrá descubrir los patrones de éxito de un estudiante en el desarrollo académico de los estudiantes de Aplicaciones Web de la Escuela de Sistemas?,

¿Cuáles han sido los resultados de la detección de patrones de participación de estudiantes exitosos?.

1.5 Objetivos de la investigación.

1.5.1 *Objetivo general.*

Aplicar un modelo de detección de patrones de participación empleando minería de datos en el Entorno Virtual de Aprendizaje de Aplicaciones Web de la ESPOCH para predecir estudiantes exitosos.

1.5.2 *Objetivos específicos.*

Determinar los patrones de participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web de la ESPOCH.

Identificar el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web, compararlas con las calificaciones obtenidas en las evaluaciones acumulativas.

Determinar si el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web ayudan a que los estudiantes sea exitosos.

1.6 Justificación de la investigación.

La Escuela Superior Politécnica de Chimborazo es una de las Instituciones de Educación Superior (IES), más importantes del centro del país; actualmente se encuentra en una etapa de continua evolución y desarrollo, al punto que sus estudiantes están rompiendo las barreras de la brecha digital y se perfilan en grandes ideas y proyectos en la era del conocimiento.

Estos estudiantes generan una gran cantidad de información diariamente: cuando navegan por el Internet, cuando utilizan las plataformas de educación virtuales, en el momento de comunicarse por WhatsApp, Hangouts, Messenger, Facebook, Instagram, Google plus, entre otros; Esta información correctamente analizada y organizada generará conocimiento valioso para la toma de decisiones en la Institución.

La detección de patrones hace visible para los profesores, investigadores, y la propia institución, la información invisible contenida en las bases de datos; con el descubrimiento de los patrones de participación se podrá conocer indicadores de frecuencia y de impacto; de tal forma que el docente

tendrá más argumentos para la toma de decisiones en su gestión de cátedra con respecto a los patrones de estudiantes que tienen bajas probabilidad de éxito en el semestre.

Proporcionar al docente datos (por ejemplo: participación en actividades, recursos, interacción entre estudiantes, calificaciones y consultas desde la plataforma).

Si los estudiantes exitosos en cursos pasados mostraron tendencias hacia una participación de foros en los cursos o tareas, chat, entre otras; entonces el modelo de predicción podría pronosticar que los estudiantes que participan activamente en los foros, tareas y chat; y el porcentaje de calificación de la participación en el Entorno Virtual de Aprendizaje comparada con la nota del curso es alto, tienen una alta probabilidad de éxito en el curso que aquellos que no participan activamente en las actividades antes mencionadas.

Esta información se convierte en conocimiento ya que el profesor la utilizará para tomar decisiones efectivas sobre los estudiantes que tienen una alta probabilidad de fracaso en el curso.

Esta investigación pretende encontrar patrones de participación de los estudiantes en la cátedra de Aplicaciones Web de la Escuela de Ingeniería en Sistemas de la Facultad de Informática y Electrónica de la ESPOCH mediante las técnicas de minería de datos y empleando modelos de detección de patrones; para descubrir las tendencias de participación de los estudiantes exitosos de semestres pasados, con el fin de compararlos con los comportamientos de los estudiantes actuales y así saber si hay influencia directa en el éxito de los estudiantes, de esta manera también predeciremos cuales son los estudiantes que se encontrarán en el nivel de no exitoso.

Además esta investigación está enmarcada en los criterios técnicos de Aplicación e Innovación, ya que es una propuesta de solución a un problema determinado “carencia de un modelo de detección de patrones de participación” e innovación, porque propone una solución diferente a las existentes “empleando minería de datos en el Entorno Virtual de Aprendizaje de Aplicaciones Web”.

Es importante tener en cuenta que el proyecto está directamente relacionado con los objetivos de la maestría porque se enmarca en la formulación de un proyecto para el desarrollo académico de los estudiantes con el propósito de impulsar cambios en la Escuela de Ingeniería en Sistemas de la Facultad de Informática y Electrónica, en base a la toma de decisiones de parte de las autoridades.

Por último el proyecto propuesto está estrechamente enmarcado en las línea de investigación de la maestría, “Formulación de proyectos de investigación y desarrollo en áreas específicas del conocimiento: salud, ambiente, economía, administración, cultura, educación, ciencia y tecnología, entre otros”, ya que es la formulación de un proyecto de desarrollo en el área de la educación y la tecnología.

1.7 Hipótesis.

La detección de patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje, influye en el éxito de los estudiantes del semestre actual.

CAPÍTULO II

2. REVISIÓN DE LITERATURA

2.1 Antecedentes y estudios previos

Las investigaciones científicas pueden ser realizadas a partir de metodologías cuantitativas y cualitativas. La primera consiste en el contraste de teoría ya existente a partir de una serie de hipótesis surgidas de la misma, siendo necesario obtener una muestra, ya sea en forma aleatoria o discriminada, pero representativa de una población o fenómeno objeto de estudio. Por lo tanto, para realizar estudios cuantitativos es indispensable contar con una teoría ya construida, dado que el método científico utilizado en la misma es el deductivo; mientras que la segunda (metodología cualitativa) consiste en la construcción o generación de una teoría a partir de una serie de proposiciones extraídas de un cuerpo teórico que servirá de punto de partida al investigador, para lo cual no es necesario extraer una muestra representativa, sino una muestra teórica conformada por uno o más casos (Martinez.2006, p.23).

Pese a que las metodologías cualitativas están reservadas a la construcción o generación de teorías, a partir de una serie de observaciones de la realidad objeto de estudio, haciendo uso del método inductivo, según el cual se debe partir de un estado nulo de teoría, (Glasser y Strauss .1987,p.253, citado en Perry, 1998, p.788) aseguran que “en la práctica es difícil ignorar la teoría acumulada, ya que ésta es importante antes de comenzar el proceso de investigación; es decir, el primer conocimiento común ganado a través del proceso de socialización, inevitablemente. Influirá en la formulación de las hipótesis por parte del investigador...el investigador debe abstenerse de la apropiación no crítica de ésta reserva de ideas”.

Por consiguiente, “comenzar sin nada o con una absoluta limpieza del estado teórico no es ni práctico, ni preferido”. De esta manera, el marco teórico se constituye en una parte importante de una investigación –independientemente del tipo de metodología utilizado–, toda vez que ésta se beneficiará de sus aportaciones científicas.

Por consiguiente, (Sarabia 1999:55) indica que en lo metodológico, la investigación científica actual es una espiral inductivo - hipotético – deductivo con dos pasos procesales esenciales:

Fase heurística o de descubrimiento: fase hecha de observación, descripción, reflexión y generalización inductiva, con miras a generar hipótesis (lo que podría ser verdadero como solución al problema, respuesta a la cuestión o explicación del fenómeno).

Fase de justificación-confirmación: proceso de comprobación del fundamento de una hipótesis por medio de un procedimiento o dispositivo previsto al efecto (y susceptible de ser reproducido).

De acuerdo con lo anterior, este autor considera que algunas de las actividades relevantes en el proceso de investigación científica son:

La observación-descripción del fenómeno,

La exploración de la realidad para la generación de hipótesis explicativas sobre el comportamiento, las causas y los efectos del fenómeno, y

El contraste-justificación de la hipótesis propuesta en la idea de garantizar su verdadera capacidad de explicación.

Respecto a su propósito, las investigaciones realizadas a través del método de estudio de caso pueden ser: descriptivas, si lo que se pretende es identificar y describir los distintos factores que ejercen influencia en el fenómeno estudiado, y exploratorias, si a través de las mismas se pretende conseguir un acercamiento entre las teorías inscritas en el marco teórico y la realidad objeto de estudio.

2.2 Fundamentación Teórica de las técnicas de Minería de Datos.

A mediados del siglo XX, las empresas empezaron a preocuparse por los datos que generaban sus clientes entorno a los bienes y servicios que adquirían, estos datos se encontraba dispersos en bases de datos o archivos por sus diferentes puntos de ventas, bajo esas circunstancias para los directivos les era casi imposible tomar decisiones reales basados en datos reales.

Estas empresas empezaron a desarrollar aplicaciones o sistemas que pudieran acceder de mejor manera y así le permitiera tener una mejor información de sus datos; pero esta tarea se fue

convirtiéndose en un problema ya que su tecnología no era suficiente para la cantidad de datos que tenían a su alrededor, convirtiéndose en un rompecabezas infinito.

En base a esas necesidades surge una nueva tendencia que es la minería de datos; esta técnica permite mejorar en gran medida la toma de decisiones en las empresas ya que construyen mediante interfaces amigables un sinnúmero de estructuras donde fluyen los datos de manera organizada mismos que son convertidos en información gracias a algoritmos supervisados y no supervisado, para luego tener el conocimiento y así tomar decisiones más precisas en beneficio de sus empresas.

En estos tiempos las empresas utilizan estas técnicas para crear iniciativas y así alcanzar sus objetivos.

Un individuo puede generar millones de datos a lo largo de su vida, estos datos son de vital importancia cuando se convierten en información la misma que luego de estudiarla se convertirá en conocimiento, este concepto ha sido utilizado hace millones de años en el pasado y el presente no es la excepción ya que hoy más que nunca se generan una gran e importante gama de datos que generan a diario conocimiento, el mismo que es utilizado para la industria, medicina, educación, cultura entre otros.

Las personas se comunican entre sí, por la necesidad de transmitir ideas, pensamientos, noticias, eventos, proyectos. La comunicación de los seres humanos se tiene su origen hace millones de años atrás, desde la existencia misma de seres humanos vieron la necesidad de comunicarse para poder sobrevivir, ya sea para canjear artículos, alimentos, animales o para simplemente conversar sea de ciencia, astronomía, cultura, entre otros . (Vallejos. 2006, P.18).

El medio más importante para la generación de datos es la comunicación ya que cada vez que nos comunicamos generamos este recurso importante para la vida del ser humano. (Vallejos. 2006, p.18); ejemplo de ello fue cuando aparecieron las primeras enfermedades humanas, los datos fueron de vital importancia para que los médicos pudieran encontrar la cura por medio de diagnósticos en base a síntomas, es decir los datos de las enfermedades se convirtieron en información (Síntomas) y estos a su vez en conocimiento (diagnóstico o tratamiento).

Cuando los seres humanos descubrieron la rueda esos datos fueron importante para llegar a construir maquinarias que hoy en día nos permiten tener un sin número de oportunidades de desarrollo y progreso.

Los datos se encuentran en cualquier parte del entorno del individuo, cuando corren, caminan, practican deporte, comen, hablan o simplemente ven televisión. Con ello podemos tener mucha información que ayuda a investigadores a realizar grandes descubrimientos y hasta llegar a predecir algunos aspectos que podría realizar dicho individuo, pero ¿Qué es un dato?.

2.3 Generalidades

Antes de hablar de minería de datos primero debemos mencionar los aspectos que forman parte del concepto.

2.3.1 Datos

Es la representación simbólica (numérica, alfabética, algorítmica, espacial, entre otros) de un atributo o variable cuantitativa o cualitativa, describen hechos empíricos, sucesos y entidades. Es un valor o referente que recibe el computador por diferentes medios, los datos representan la materia prima que se utiliza para la construcción de la información. (Haggar, 2011, P.,8).

Informativamente hablando los datos son objetos, condiciones o situaciones. Son el conjunto básico de hechos referentes a un individuo, transacción de interés para distintos objetivos, entre los cuales se encuentra la toma de decisiones (Sinnexus, 1999). Los datos pueden ser:

Alfabéticos (A a la Z), numéricos (0 al 9), simbólicos o de caracteres especiales (% , \$, # , @ , & , entre otros).

La ciencia de la computación ha ocupado al datos como la unidad mas pequeña de la información un datos corresponde a un bit que desde en el computador es utilizado para brindar 2 clases de información como son: si – no, encendido- apagado, 0 – 1. Luego al agrupar estos datos son convertidos en información gracias a la interpretación interna de la computadora. (Moreno. 2009, p.23).

Los datos de forma aislada no pueden contener información humanamente entendible. Sólo cuando un conjunto de datos se examina conjuntamente a la luz de un enfoque, hipótesis o teoría se puede apreciar la información contenida en dichos datos. Los datos pueden consistir en números, estadísticas o proposiciones descriptivas.

Los datos convenientemente agrupados, estructurados e interpretados se consideran que son la base de la información humanamente relevante que se pueden utilizar en la toma de decisiones, la reducción de la incertidumbre o la realización de cálculos. Es de empleo muy común en el ámbito informático y, en general, prácticamente en cualquier investigación científica.

2.3.2 La información

Al conjunto organizado de los datos se lo denomina información y constituye un mensaje que cambia la perspectiva del sujeto que recibe dicha información; de esta manera las personas pueden comprender que nos está proporcionando los datos. (Torres, 2015).

Informativamente podemos decir que la información es una medida de la complejidad de los datos. Estos datos pueden ser extraídos de cualquier entorno ámbito o ubicación.

En las sociedades humanas y en parte en algunas sociedades animales, la información tiene un impacto en las relaciones entre diferentes individuos. En una sociedad la conducta de cada individuo frente a algunos otros individuos se puede ver alterada en función de qué información disponible posee el primer individuo. Por esa razón, el estudio social de la información se refiere a los aspectos relacionados con la variación de la conducta en posesión de diferentes informaciones.

La información es importante en el desarrollo de las sociedades, ya que sin ella no podrías haber sobrevivido a muchas etapas de la historia humana, gracias a la información las sociedades han trascendido a través del tiempo y el espacio, ya que con toda esa gama de datos que construyeron la información se pudo conocer la forma de vivir en sociedad o individualmente.

Gracias a la gran gama de información recolectada a través del tiempo se pueden construir las profesiones y con ellos las escuelas, colegios y universidades, ya que en cualquier campo de la ciencia la información fomentó una evolución al punto que hoy en día existen gran variedad de profesiones y profesionales que ayudan a construir un mundo mejor para las nuevas generaciones.

La información es la puerta de entrada para el dominio del conocimiento, mientras más informados estemos más conoceremos y mientras más conozcamos más poder tendremos, pero, ¿Qué es el conocimiento?.

2.3.3 El conocimiento

El conocimiento no es un concepto formalizado o establecido por algún autor, filósofo, pensador o institución académica, más bien es una perspectiva en base a su función y fundamento.

El conocimiento es la comprensión del análisis de la información que genera la ciencia. Un conjunto de la obtención del verdadero conocimiento para apoyar a la toma de decisiones es una secuencia dirigida de la siguiente manera.

Cuando él un usuario interpreta los datos, este se convierte en información; luego cuando un especialista interpreta la información presenta un valor agregado entonces se convierte en conocimiento. Este paradigma se cumple estrictamente en cualquier ciencia.

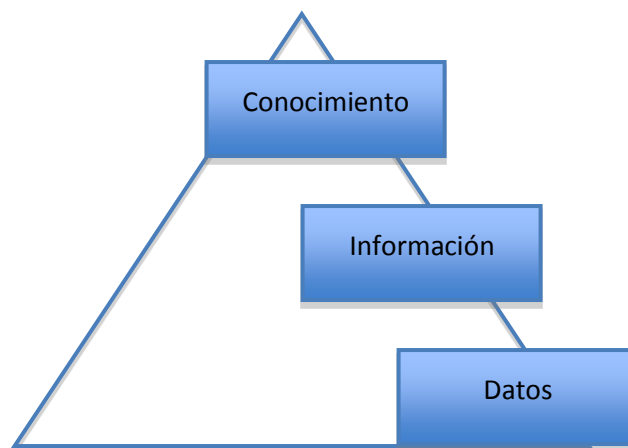


Figura. 1-2: Pirámide del conocimiento
Realizado por: Gustavo Hidalgo.2016

Una actividad esencial de todo individuo en su relación con su entorno es captar o procesar información acerca de lo que lo rodea. Este principio fundamental sitúa la actividad humana del conocer en un ámbito general propio de todos los seres de la naturaleza. El conocimiento, para el caso del hombre, consiste en su actividad relativa al entorno que le permite existir y mantenerse y desarrollarse en su existencia. (Vallejos. 2006, p.18).

Tan fundamental es esta actividad en la vida que todos "sabemos" lo que es el conocer y el conocimiento, con tal de que no tengamos que explicarlo. Tal es la situación que ocurre con casi todos los conceptos verdaderamente importantes: la palabra es perfectamente conocida y su uso

perfectamente dominado. Pero la palabra tiene una amplitud tan grande y su uso unos contextos tan variados que el concepto, tan rico y lleno de matices, resulta muy difícil de comprender y explicar.

Hoy día la ciencia habla de cognición o actividades cognitivas como un conjunto de acciones y relaciones complejas dentro de un sistema complejo cuyo resultado es lo que consideramos conocimiento. El conocimiento se logra gracias a la interacción de los elementos antes mencionados pero como lograr el conocimiento sin antes tener el elemento fundamental para el objetivo de conocer que es los datos, la gran gama de datos que se encuentran en todos los ámbitos del ser humano tiene que ser extraído de alguna manera o técnica, una de estas técnicas es la Minería de Datos (Datamining).

2.3.4 Data Warehouse

Es una base de datos accesible por los usuarios el cual contiene registros históricos y actuales de las entidades importantes que tiene la institución; Este organiza y aloja los datos necesarios, para ser utilizados en el procesamiento analítico dentro de una perspectiva de tiempo. El desarrollo de un Data Warehouse facilita información útil, entre sus características principales tenemos:

Orientado al tema: Se clasifica de acuerdo a aspectos de interés.

Integrado: Está integrada, cualquier tipo de datos será estandarizado de manera general y así será alojado en el almacén. De tiempo variante: ES el horizonte de tiempo en el que funciona el DW.

No volátil: No puede ser modificada.

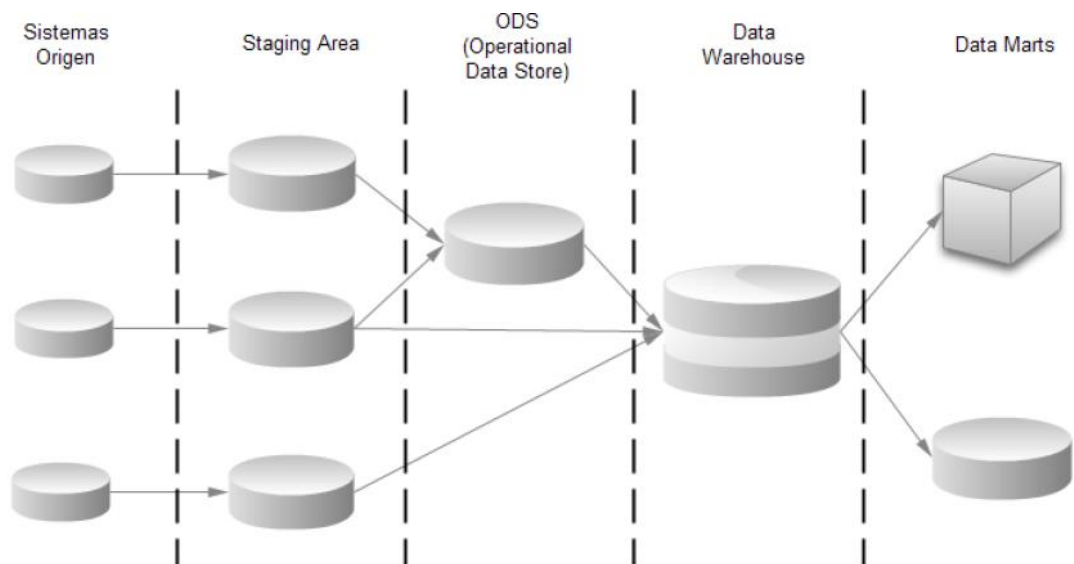


Figura 2-2. Arquitectura de un Data Warehouse
Fuente: Andrés Duque

Los sistemas Origen es donde se encuentran los datos de interés, mismos que serán objeto de la extracción; estos orígenes pueden provenir de distintos lugares como son: archivos, bases de datos, hojas de cálculo, otros sistemas, entre otros.

El área temporal o Staging Area, es un espacio de memoria donde se almacena temporalmente los datos que provienen de los orígenes, su principal objetivo es minimizar el impacto de afectación de los sistemas orígenes, de esta manera cuando se realice la extracción los orígenes no intervienen hasta que se realice la próxima carga.

Los ODS es un almacén de datos operativos, estos siguen el mismo esquema de una base de datos relacional, el DW aprovecha los datos que se encuentran aquí permitiendo dar soporte a las tareas operacionales que se ejecuten.

El Data Warehouse; es la base de datos que va a contener los datos que provienen del área temporal o del ODS.

Los DataMart se consideran como pequeños Data Warehouse, ya que contienen información específica del negocio. Este tipo de estructura es más personalizada ya que puede llegar a construirse a partir de una necesidad en particular o tema específico.

2.3.4.1 *Proceso de Extracción, Transformación y Carga*

Es un proceso de suma importancia ya que garantiza la calidad de los datos que van a ser cargados en el Data Warehouse; Este proceso también es conocido como ETL, su objetivo principal es de organizar e integrar los datos desde múltiples fuentes hasta un destino que en este caso es el Data Warehouse.

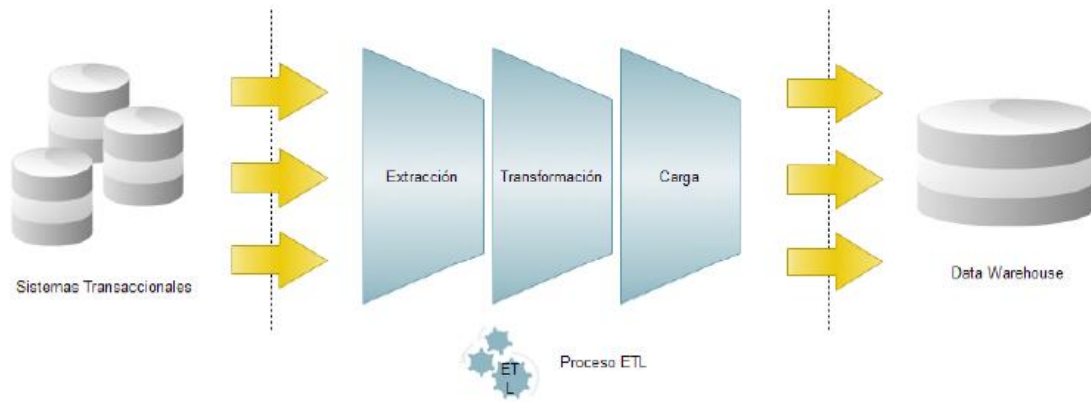


Figura. 3-2: Proceso ETL

Fuente: Andrés Duque

La Extracción se refiere a la adquisición de los datos los cuales pueden ser recogidos de diferentes fuentes; esta se almacena en un área temporal aquí solo se extraen los datos necesarios para el estudio ya que en el ambiente transaccional existe una gran cantidad de información en algunos casos innecesaria.

La transformación consiste en el refinamiento de los datos que han sido extraídos, este proceso incluye, corrección de errores, decodificaciones, borrado de datos que no son de interés, generación de claves, agregación de información, entre otras cosas.

Por último la Carga, consiste en almacenar los datos extraídos y transformados a un nuevo espacio de almacenamiento que por lo general es el Data Warehouse.

2.3.4.2 *Modelo dimensional*

Es un modelo de estructura de almacenamiento, este se caracteriza por ser tipo estrella o copo de nieve, con esto se gana maximizar el rendimiento de las consultas.

El modelo estrella es la técnica de diseño más popular en este tipo de herramientas, este tiene la característica de tener una tabla en el centro conocida como tabla de hechos, misma que se encuentra conectada radialmente con otro objeto o entidad llamada dimensión.

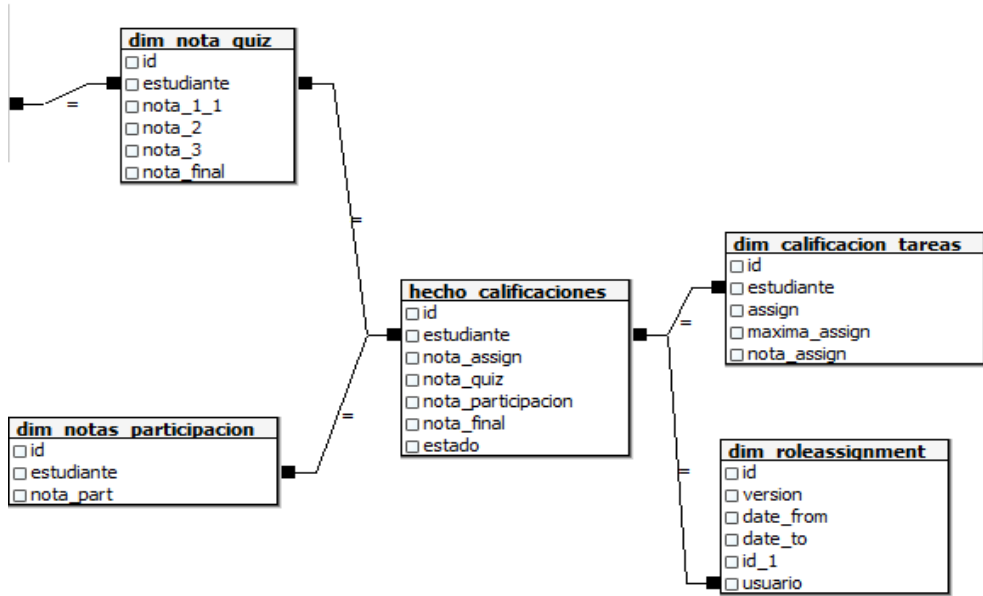


Figura. 4-2: Estructura estrella
Realizado por: Gustavo Hidalgo.2016

En la estructura copo de nieve es una extensión de la estrella, donde cada punta se explota en más puntas; la diferencia es que en esta estructura se normaliza las dimensiones para eliminar redundancia permitiendo que estos se agrupan en múltiples tablas.

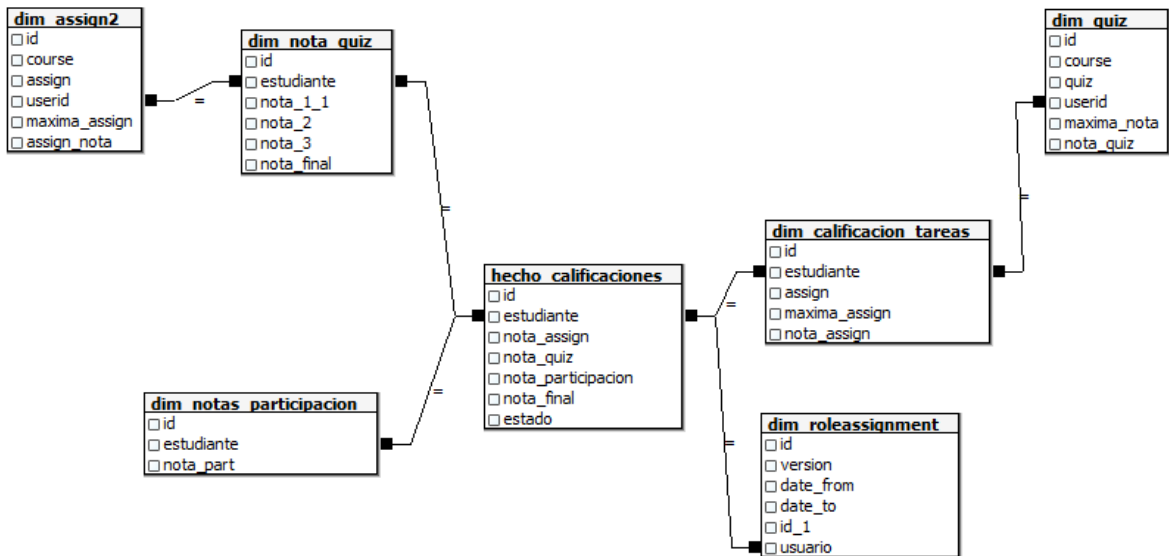


Figura. 5-2: Estructura copo de nieve
Realizado por: Gustavo Hidalgo.2016

Tabla de hechos

Es la tabla central del esquema dimensional, este contiene las medidas del negocio a ser analizado, con este tipo de tablas se representa el hecho o actividad del negocio. Por lo general este tipo de datos son numéricos y pueden agruparse.

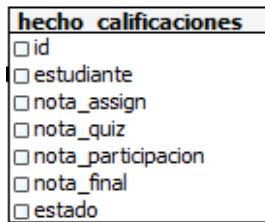


Figura. 6-2: Tabla de hechos calificaciones
Realizado por: Gustavo Hidalgo.2016

Tablas de Dimensiones

Son tablas que describen a la tabla de hechos, mediante atributos descriptores, este tipo de tablas tienen 2 tipos de atributos que son: la clave primaria, y de los atributos descriptores.

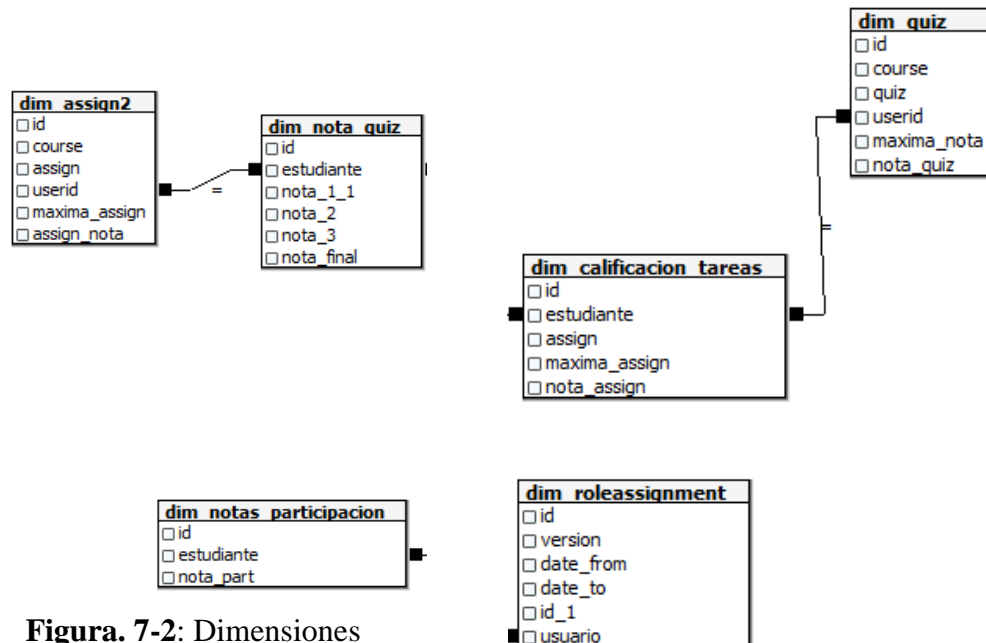


Figura. 7-2: Dimensiones
Realizado por: Gustavo Hidalgo.2016

2.3.5 Minería de datos

Existen muchas definiciones de minería de datos a continuación alguna de ellas:

La empresa Sinnexus lo define como un conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Alfonso Cutro dice: “Es el proceso de extraer información válida, auténtica y que se pueda procesar de las bases de datos de gran tamaño”.

Sin embargo la minería de datos es un conjunto de técnicas de Inteligencia Artificial, aprendizaje automático, estadísticas y sistemas de bases de datos, que tiene como objetivo extraer información de un conjunto de datos.

La minería de datos está compuesta por un conjunto de técnicas que están sujetas a continua evolución gracias a investigaciones tales como: base de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, entre otros.

El objetivo primordial de la minería de datos es la de encontrar patrones repetitivos, tendencias, reglas que identifique el comportamiento de los datos en un determinado contexto, es decir comprender el contenido de un conjunto de datos. Los algoritmos en la minería de datos se clasifican en: algoritmos supervisados o predictivos y no supervisados o descubrimiento de conocimiento.

Supervisados o predictivos: predicen el valor de un atributo clase a partir de un conjunto de datos conocidos como atributos predictores. Entre estos tenemos; Árbol de decisión, inducción neuronal, regresión, series temporales. Entre otros.

No supervisados o descubrimiento del conocimiento: descubren patrones y tendencias de datos actuales, el descubrimiento de esa información sirve para llevar a cabo acciones y obtener beneficios de ellos. Entre ellos tenemos; Detección de desviaciones, segmentación, agrupamiento o cluster, reglas de asociación, entre otros.

La aplicación de la minería de datos requiere de la realización de un conjunto de actividades previas encaminadas a preparar los datos de entrada, ya que algunas veces estos datos provienen de repositorios heterogéneos y no tienen el formato adecuado para el experimento.

La minería de datos es un proceso que interviene la dinámica del método científico ya que primero se formula la hipótesis y luego se diseña el experimento para coleccionar los datos que confirmen o refuten dicha hipótesis. Para aplicar estas técnicas avanzadas estas deben de estar integradas a un DataWarehouse y a herramienta para el análisis de datos o inteligencia de negocios.

2.3.5.1 *Proceso de Minería de Datos (DataMining)*

Esta tendencia ha siendo utilizada fuertemente en las investigaciones de todas las ciencias donde involucre gran cantidad de datos y muchas fuentes de información, esta técnica utiliza la inteligencia artificial, aprendizaje autónomo, estadísticas y sistema de base de datos para logra el objetivo de extraer información de un conjunto de datos también llamados fuentes de información que pueden ser redes sociales, entornos virtuales de aprendizaje, bases de datos distribuidas, entre otros A esta tendencia algunos autores lo denominan máquina de aprendizaje práctico. (Molina, 2010).

La minería de datos realiza un proceso secuencial el cual se expone en el siguiente apartado.

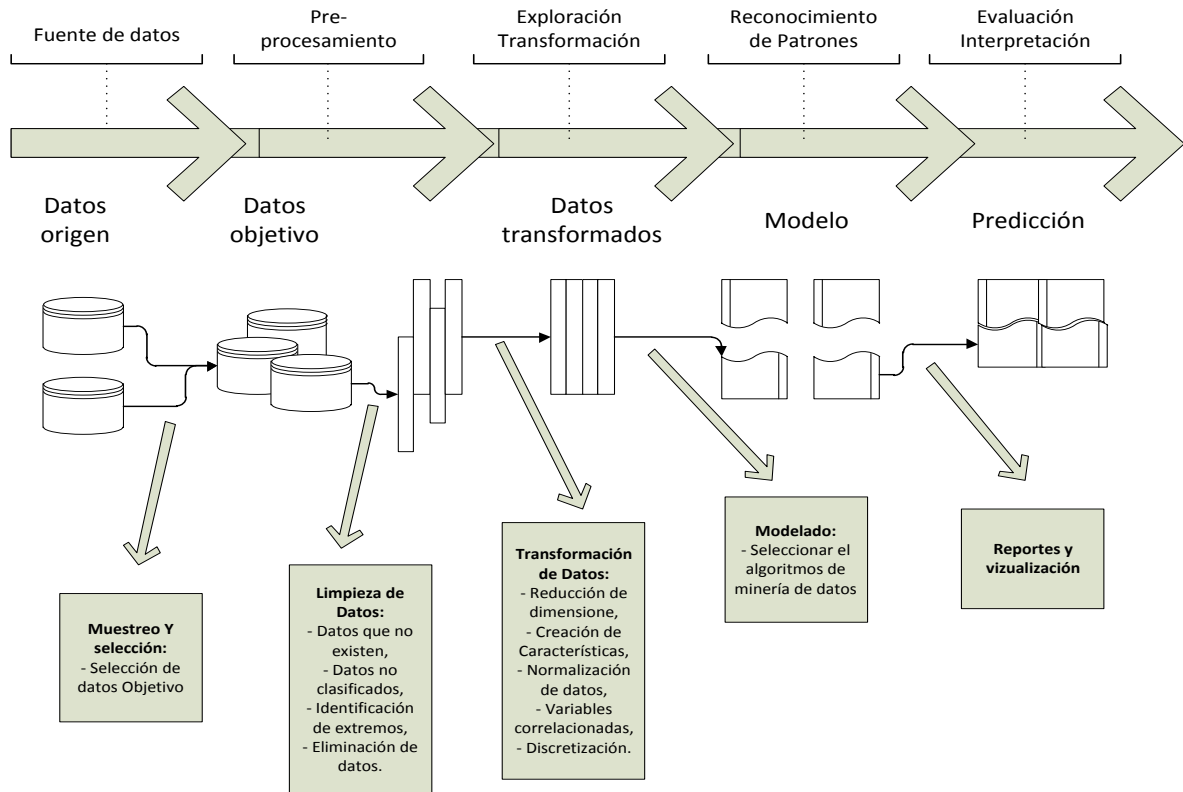


Figura 8-2: Metodología ETL
 Realizado por: Gustavo Hidalgo.2016

- 1.- Selección del conjunto de datos (aquellos que se quieren predecir o inferir) variables independientes.
- 2.- Análisis de propiedades de los datos esto se lo puede hacer mediante histogramas, diagramas de dispersión, presencia de valores atípicos.
- 3.- Transformaciones del conjunto de datos de entrada.
- 4.- Seleccionar y aplicar técnicas de minería de datos (construcción de un modelo predictivo, de clasificación o segmentación).
- 5.- Extracción del conocimiento luego de las técnicas de minería de datos se obtiene un modelo de conocimiento representan patrones de comportamiento observados en valores de variables del problema o relaciones de asociaciones entre dichas variables.
- 6.- Interoperabilidad y evaluación de datos. Conclusión.

Como habíamos dicho anteriormente el datamining (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto. (Vallejos. 2006, p.18).

Básicamente, el datamining surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales. (Cabero. 2012, p.4).

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces nos referimos al conocimiento. Vea más diferencias entre datos, información y conocimiento.

Aunque desde un punto de vista académico el término datamining es una etapa dentro de un proceso mayor llamado extracción de conocimiento en bases de datos (Knowledge Discovery in Databases o KDD) en el entorno comercial, así como en este trabajo, ambos términos se usan de manera indistinta.

Lo que en verdad hace el datamining es reunir las ventajas de varias áreas como la Estadística, la Inteligencia Artificial, la Computación Gráfica, las Bases de Datos y el Procesamiento Masivo, principalmente usando como materia prima las bases de datos. Una definición tradicional es la siguiente: "Un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos". (Fayyad y otros. 1996, p.5).

Desde nuestro punto de vista, lo definimos como "la integración de un conjunto de áreas que tienen como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten un sesgo hacia la toma de decisión". (Molina y otros. 2001, p.23).

2.3.5.2 Técnicas de Minería de Datos (DataMining)

Las técnicas de minería de datos están sujetas a continua evaluación gracias a las investigaciones tales como base de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadísticas, recuperación de información. Los algoritmos de la minería de datos se clasifican en dos: Supervisados y No supervisados.

Supervisados o predictivos: Predicen el valor de un atributo de un conjunto de datos, conociendo otros atributos. La clasificación en esta parte tenemos Árboles de decisión, Inducción neural, Regresión, Series temporales. (Molina. 2010, p.23).

No supervisados o Descubrimiento del conocimiento: Descubren patrones y tendencias en datos actuales (No utiliza datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener beneficios en ellos. Entre los no supervisados tenemos; Detección de desviaciones, segmentación, Agrupamiento, Reglas de asociación, patrones secuenciales. (Molina. 2010, p.23).

El análisis de asociación persigue el establecimiento de relaciones entre los registros individuales o grupos de registros de la base de datos. La aplicación de la Minería de Datos requiere la realización de un conjunto de actividades previas encadenadas a preparar los datos de entrada ya que algunas veces estos datos provienen de repositorios heterogéneos y no tienen el formato adecuado o contienen ruidos.

El nombre de Data Mining deriva de las similitudes entre buscar valiosa información de negocios en grandes bases de datos. (Molina. 2010, p.23), por ejemplo: encontrar información de la venta de un producto entre grandes montos de Gigabytes almacenados - y minar una montaña para encontrar una veta de metales valiosos.

Ambos procesos requieren examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente donde residen los valores. Dadas bases de datos de suficiente tamaño y calidad, la tecnología de Data Mining puede generar nuevas oportunidades de negocios al proveer estas capacidades:

Predicción automatizada de tendencias y comportamientos. DataMining automatiza el proceso de encontrar información predecible en grandes bases de datos. Preguntas que tradicionalmente requerían un intenso análisis manual, ahora pueden ser contestadas directa y rápidamente desde los datos.

Descubrimiento automatizado de modelos previamente desconocidos. Las herramientas de DataMining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Las técnicas de DataMining pueden redituar los beneficios de automatización en las plataformas de

hardware y software existentes además de ser implementada en sistemas nuevos a medida que las plataformas existentes se actualicen y nuevos productos sean desarrollados.

Cuando las herramientas de DataMining son implementadas en sistemas de procesamiento paralelo de alta performance, pueden analizar bases de datos masivas en minutos.

Procesamiento más rápido significa que los usuarios pueden automáticamente experimentar con más modelos para entender datos complejos. Alta velocidad hace que sea práctico para los usuarios analizar inmensas cantidades de datos. Grandes bases de datos, a su vez, producen mejores predicciones.

Esta técnica nos permite tener los datos limpios y listo a para realizar el análisis en el área de la educación este análisis se convierte en una herramienta importante para saber cómo mejorar el aprendizaje prediciendo el comportamiento académico de los estudiantes. A esta tendencia se la denomina Analítica de conocimiento (Learning Analytics). (Molina. 2010, p.23).

Árbol de Decisión

El algoritmo de árboles de decisión genera un modelo de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al modelo cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta.

El algoritmo de árboles de decisión utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos utilizan la selección de características para mejorar el rendimiento y la calidad del análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si utiliza demasiados atributos de predicción o de entrada al diseñar un modelo de minería de datos, el modelo puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Entre los métodos que se usan para determinar si hay que dividir el árbol figuran métricas estándar del sector para la entropía y las redes Bayesianas.

Un problema común de los modelos de minería de datos es que el modelo se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobreajustado o sobreentrenado. Un modelo sobreajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobreajustar un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol.

Cuando se prepara los datos para su uso en un modelo de árboles de decisión, conviene que comprenda qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos se utilizan.

Los requisitos para un modelo de árboles de decisión son los siguientes:

Una columna key: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una columna de predicción. Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.

Columnas de entrada. Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

Los árboles de decisión son uno de los métodos de aprendizaje inductivo más usados.

Hipótesis de aprendizaje inductivo: cualquier hipótesis encontrada que clasifique un número suficientemente grande de ejemplos de entrenamiento clasificará otros ejemplos no observados.

Razonamiento deductivo: partiendo de unas premisas se llega necesariamente a una conclusión. No aporta información nueva.

Razonamiento abductivo: partiendo del conocimiento de unos efectos (síntomas) se llega a la causa (enfermedad)

Se trata de aproximar una función desconocida a partir de ejemplos positivos y negativos de esa función. Esos ejemplos serán en realidad pares $\langle x, f(x) \rangle$, donde x es el valor de entrada y $f(x)$ el valor de la función aplicada a x . Dado un conjunto de ejemplos de f , la inducción consiste en obtener una función h que aproxime f . A esta función h se la denomina hipótesis.

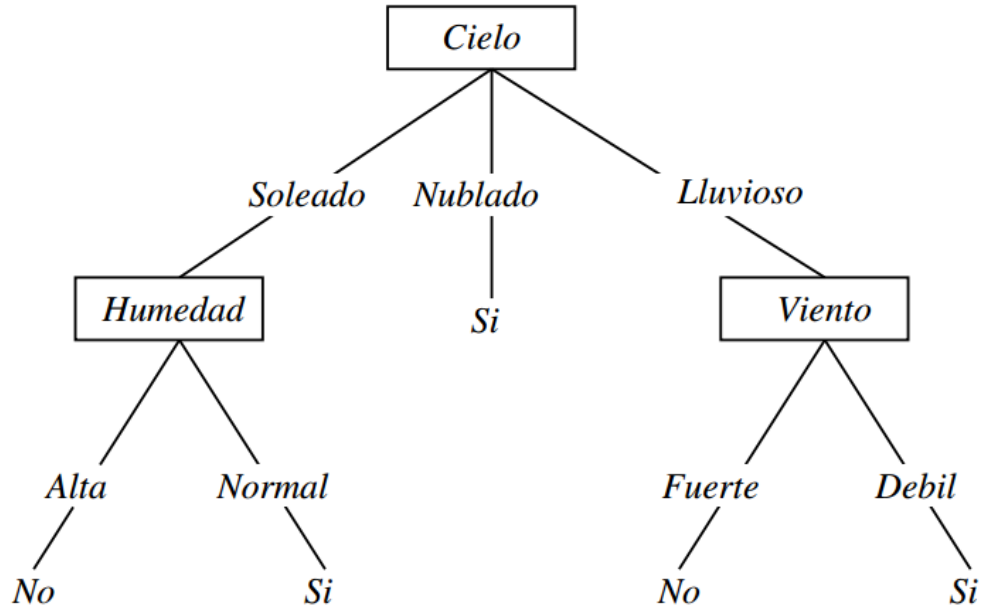


Figura. 9-2: Árbol de decisión
Realizado por: Gustavo Hidalgo.2016

Redes Neuronales

El perceptrón multicapa es una red neuronal artificial (RNA) formada por múltiples capas, esto le permite resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón (también llamado perceptrón simple). El perceptrón multicapa puede ser totalmente o localmente conectado. En el primer caso cada salida de una neurona de la capa "i" es entrada de todas las neuronas de la capa "i+1", mientras que en el segundo cada neurona de la capa "i" es entrada de una serie de neuronas (región) de la capa "i+1".

Las capas pueden clasificarse en tres tipos:

Capa de entrada: Constituida por aquellas neuronas que introducen los patrones de entrada en la red. En estas neuronas no se produce procesamiento.

Capas ocultas: Formada por aquellas neuronas cuyas entradas provienen de capas anteriores y cuyas salidas pasan a neuronas de capas posteriores.

Capa de salida: Neuronas cuyos valores de salida se corresponden con las salidas de toda la red.

La propagación hacia atrás (también conocido como retropropagación del error o regla delta generalizada), es un algoritmo utilizado en el entrenamiento de estas redes, por ello, el perceptrón multicapa también es conocido como red de retropropagación (no confundir con la red de contrapropagación).

Entre sus limitaciones tenemos no extrapola bien, es decir, si la red se entrena mal o de manera insuficiente, las salidas pueden ser imprecisas. La existencia de mínimos locales en la función de error dificulta considerablemente el entrenamiento, pues una vez alcanzado un mínimo el entrenamiento se detiene aunque no se haya alcanzado la tasa de convergencia fijada.

Cuando caemos en un mínimo local sin satisfacer el porcentaje de error permitido se puede considerar: cambiar la topología de la red (número de capas y número de neuronas), comenzar el entrenamiento con unos pesos iniciales diferentes, modificar los parámetros de aprendizaje, modificar el conjunto de entrenamiento o presentar los patrones en otro orden.

El Perceptrón multicapa es una red de alimentación hacia adelante (feedforward) compuesta por una capa de unidades de entrada (sensores), otra capa de unidades de salida y un número determinado de capas intermedias de unidades de proceso, también llamadas capas ocultas porque no se ven las salidas de dichas neuronas y no tienen conexiones con el exterior. Cada sensor de entrada está conectado con las unidades de la segunda capa, y cada unidad de proceso de la segunda capa está conectada con las unidades de la primera capa y con las unidades de la tercera capa, así sucesivamente.

Las unidades de salida están conectadas solamente con las unidades de la última capa oculta, como se muestra en la figura 10-2. Con esta red se pretende establecer una correspondencia entre un conjunto de entrada y un conjunto de salidas deseadas, de manera que:

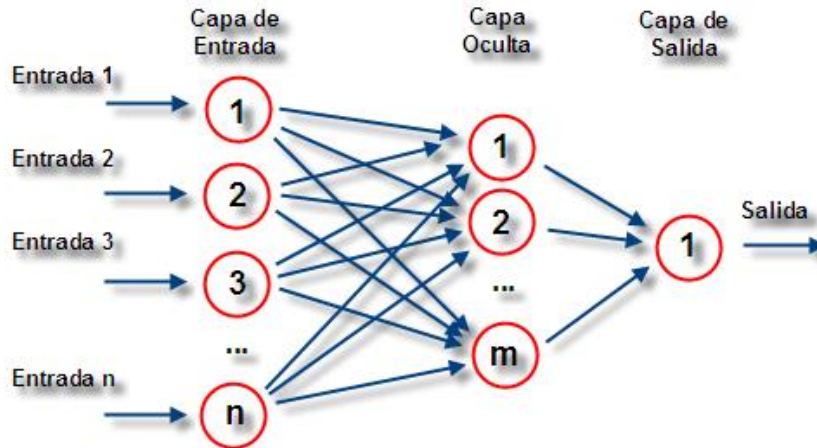


Figura: 10-2: Esquema redes neuronales
Fuente: Wikipedia

Para implementar dicha relación, la primera capa (sensores de entrada) tendrá tantos sensores como componentes tenga el patrón de entrada, es decir, N; la capa de salida tendrá tantas unidades de proceso como componentes tengan las salidas deseadas, es decir, M, y el número de capas ocultas y su tamaño dependerán de la dificultad de la correspondencia a implementar.

$$a_i = F(u_i^k + \sum_{j=1}^{n^{(k-1)}} a_j^{(k-1)} w_{ji}^{(k-1)})$$

Donde:

a: valor de la salida de una neurona

u: umbral de cada neurona

w: peso de los signoides

Bosques Aleatorios

Los bosques aleatorios (o random forests) es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.

El algoritmo para inducir un random forest fue desarrollado por Leo Breiman y Adele Cutler y Random forests es su marca de fábrica. El método combina la idea de bagging de Breiman y la

selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de decisión con variación controlada.

La selección de un subconjunto aleatorio de atributos es un ejemplo del método random subspace, el que, según la formulación de Ho, es una manera de llevar a cabo la discriminación estocástica⁵ propuesta por Eugenio Kleinberg.

En muchos problemas el rendimiento del algoritmo random forest es muy similar a la del boosting, y es más simple de entrenar y ajustar. Como una consecuencia el random forests es popular y es ampliamente utilizado.

La idea esencial del bagging es promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la variación. Los árboles son los candidatos ideales para el bagging, dado que ellos pueden registrar estructuras de interacción compleja en los datos, y si crecen suficientemente profundo, tienen relativamente baja parcialidad. Producto de que los árboles son notoriamente ruidosos, ellos se benefician grandemente al promediar.

Cada árbol es construido usando el siguiente algoritmo:

Sea N el número de casos de prueba, M es el número de variables en el clasificador.

Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado; m debe ser mucho menor que M

Elegir un conjunto de entrenamiento para este árbol y usar el resto de los casos de prueba para estimar el error.

Para cada nodo del árbol, elegir aleatoriamente m variables en las cuales basar la decisión. Calcular la mejor partición a partir de las m variables del conjunto de entrenamiento.

Para la predicción un nuevo caso es empujado hacia abajo por el árbol. Luego se le asigna la etiqueta del nodo terminal donde termina. Este proceso es iterado por todos los árboles en el ensamblado, y la etiqueta que obtenga la mayor cantidad de incidencias es reportada como la predicción.

Las ventajas del random forests son:

Es uno de los algoritmos de aprendizaje más certeros que hay disponible. Para un set de datos lo suficientemente grande produce un clasificador muy certero.

Corre eficientemente en bases de datos grandes.

Puede manejar cientos de variables de entrada sin excluir ninguna.

Da estimados de qué variables son importantes en la clasificación.

Tiene un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.

Computa los prototipos que dan información sobre la relación entre las variables y la clasificación.

Computa las proximidades entre los pares de casos que pueden usarse en los grupos, localizando valores atípicos, o (ascendiendo) dando vistas interesantes de los datos.

Ofrece un método experimental para detectar las interacciones de las variables.

El Random Forest comienza con una técnica de aprendizaje automático estándar llamada "árbol de decisiones", que, en cuanto al conjunto, corresponde a un aprendizaje. En un árbol de decisión, una entrada se introduce en la parte superior y hacia abajo a medida que atraviesa el árbol de los datos se acumulan en conjuntos más pequeños y más pequeños (sobra).

Algoritmo:

Así es como se formó este sistema, para un número determinado de árboles T:

Muestra N casos al azar con el reemplazo para crear un subconjunto de los datos. El subconjunto debe ser aproximadamente 66% del conjunto total.

En cada nodo:

Para un número m, las variables predictoras m son seleccionados al azar entre todas las variables predictoras.

La variable de predicción que proporciona la mejor división, de acuerdo con una función objetiva, se utiliza para hacer una división binaria en ese nodo.

En el siguiente nodo, elige otras m variables al azar entre todas las variables predictoras y hace lo mismo.

Dependiendo del valor de m, hay tres sistemas ligeramente diferentes:

Selección aleatoria divisor: $m = 1$

Empaquetadores de Breiman: $m = \text{número total de variables de predictor}$

Random forest: $m \ll \text{número de variables predictoras}$. Breiman sugiere tres posibles valores de m:

$\frac{1}{2} m$, \sqrt{m} y $2 \sqrt{m}$

Teorema de Bayes

En términos más generales y menos matemáticos, el teorema de Bayes es de enorme relevancia puesto que vincula la probabilidad de A dado B con la probabilidad de B dado A. Es decir, que sabiendo la probabilidad de tener un dolor de cabeza dado que se tiene gripe, se podría saber (si se tiene algún dato más), la probabilidad de tener gripe si se tiene un dolor de cabeza. Muestra este sencillo ejemplo la alta relevancia del teorema en cuestión para la ciencia en todas sus ramas, puesto que tiene vinculación íntima con la comprensión de la probabilidad de aspectos causales dados los efectos observados.

Los algoritmos de aprendizaje bayesiano pueden calcular probabilidades explícitas para cada hipótesis. También nos proporcionan un marco para estudiar otros algoritmos de aprendizaje. El aprendizaje se puede ver como el proceso de encontrar la hipótesis más probable, dado un conjunto de ejemplos de entrenamiento D y un conocimiento a priori sobre la probabilidad de cada hipótesis.

Cada ejemplo de entrenamiento afecta a la probabilidad de las hipótesis. Esto es más efectivo que descartar directamente las hipótesis incompatibles. Se puede incluir conocimiento a priori: probabilidad de cada hipótesis; y la distribución de probabilidades de los ejemplos. Es sencillo asociar un porcentaje de confianza a las predicciones, y combinar predicciones en base a su confianza.

Una nueva instancia es clasificada como función de la predicción de múltiples hipótesis, ponderadas por sus probabilidades. Incluso en algunos casos en los que el uso de estos métodos se ha mostrado imposible, pueden darnos una aproximación de la solución óptima.

$$P(h|D) = \frac{P(D|h) P(h)}{P(D)}$$

Donde:

$P(h)$ es la probabilidad a priori de la hipótesis h.

$P(D)$ es la probabilidad de observar el conjunto de entrenamiento D.

$P(D|h)$ es la probabilidad de observar el conjunto de entrenamiento D en un universo donde se verifica la hipótesis h.

$P(h|D)$ es la probabilidad a posteriori de h, cuando se ha observado el conjunto de entrenamiento D.

2.3.5.3 Trabajos relacionados

Tema: Minería de datos para descubrir estilos de aprendizaje

Autores: Elena B. Durán, Rosanna N. Costaguta

Resumen

Los estudiantes aprenden de muchas maneras: viendo y escuchando, reflexionando y actuando, razonando lógicamente e intuitivamente, memorizando y visualizando, construyendo analogías y modelos matemáticos. También los métodos de enseñanza son variados. Cuanto aprende un estudiante depende de su habilidad innata y de su preparación previa, pero también de la compatibilidad entre su estilo de aprendizaje y el estilo de enseñanza del instructor. Como docentes de la carrera Licenciatura en Sistemas de Información de la Universidad Nacional de Santiago del Estero, consideramos que resultaría útil determinar las características del perfil de aprendizaje de nuestros estudiantes para adecuar las estrategias de enseñanza.

Con este propósito encuestamos a ochenta y nueve estudiantes, aplicando el Test de Felder y Soloman. La información recabada originó una base de datos a la que se le aplicó el Proceso de Descubrimiento de Conocimiento (Knowledge Discovery from Database - KDD).

Así determinamos que existe un alto grado de homogeneidad en el estilo de aprendizaje de los alumnos. La identificación del cluster dominante permitió conocer además las características del estilo de aprendizaje compartido por la mayoría de los alumnos. Teniendo en cuenta la información descubierta se sugieren estrategias de intervención didáctica y de presentación de material educativo conforme al estilo dominante.

Tema: Evaluación visual de las relaciones entre participación de los estudiantes y sus resultados en entornos de e-learning

Autores: Gómez Aguilar, Diego Alonso, García Peñalvo, Francisco J., Therón, Roberto

Resumen:

Investigaciones recientes han confirmado el impacto del uso de los medios electrónicos en los resultados de los estudiantes. Además en anteriores trabajos se ha estudiado la relación existente entre

del acceso a la lectura y publicación en un foro y el acceso a la lectura de los recursos con el desempeño de los estudiantes, sin embargo, poco se sabe de la relación de la periodicidad temporal de estas actividades con los resultados de los alumnos.

Además algunos estudios han realizado diferentes categorizaciones de las actividades de los estudiantes demostrando su utilidad en la predicción y entendimiento del aprendizaje del alumno. En este estudio se ha implementado una visualización para encontrar estos aspectos mediante técnicas de analítica visual. Los resultados de este estudio mostraron que no sólo la participación en foros de discusión, como la lectura o la publicación, sino también el acceso a la lectura de los recursos, y la periodicidad con que estas actividades se realizan pueden alertar sobre el desempeño del estudiante.

Esto se produjo después de comprobarse la existencia de una fuerte relación de influencia entre la frecuencia de acceso a la lectura y publicación a foros de discusión, así como entre la frecuencia de acceso a la lectura a los recursos con los resultados de los estudiantes. Sin embargo, y además se ha encontrado que la periodicidad con que estas actividades se realizan a lo largo del curso es también un factor influyente y puede analizarse visualmente, gracias a que se obtiene una mejor visualización de los patrones de rendimiento de los estudiantes.

Tema: Análisis de patrones en la participación ciudadana en procesos electorales aplicando algoritmos de minería de datos.

Autor: José Luis Gutiérrez Villanueva

Resumen.

En Baja California Sur se aplica un sistema democrático para el proceso de elección a gobernador constitucional. Las elecciones se llevan a cabo con la participación de los ciudadanos y del Instituto Estatal Electoral de Baja California Sur (IEEBCS). Esto genera información muy importante: los resultados electorales y la participación ciudadana.

Los resultados electorales son el conteo final de todos los votos agrupados por candidato. La participación ciudadana es el porcentaje del total de personas que votaron entre el total de personas que podían votar en la elección.

Una de las métricas para la evaluación del proceso electoral es la participación ciudadana. Por lo tanto, el IEEBCS dentro de sus planes estratégicos para asegurar unas elecciones exitosas ejecuta una

promoción del voto antes y durante el proceso electoral para motivar a los ciudadanos a votar el día de la elección. Sin embargo esta promoción del voto se realiza de forma muy general.

En el presente trabajo se diseñó un modelo de minería de datos para el análisis de la participación ciudadana en procesos electorales de Baja California Sur. El cual permite determinar las variables y patrones de comportamiento presentados en las elecciones. Se realizó la construcción de una bodega de datos alimentada por datos de la participación ciudadana y de fuentes externas con información demográfica de la entidad. Para el análisis de la información se aplicaron técnicas de Minería de

Datos para la obtención de los patrones de comportamiento. Las técnicas utilizadas son: el método de Árbol de Decisión (dentro del algoritmo de Clasificación) y el método de K-Means (dentro de los algoritmos de Clúster o Agrupamiento). La investigación se llevó a cabo mediante la aplicación de la metodología CRISP-DM, ya que es una de las metodologías de minería de datos más completa y de las más utilizadas.

El análisis de la participación ciudadana permite llevar a cabo una planificación estratégica y focalizada a las diferentes regiones demográficas del estado, además de lograr una promoción más eficiente y un mayor porcentaje de participación.

CAPÍTULO III

3. MATERIALES Y MÉTODOS

3.1 Diseño de la investigación

Para la presente investigación, se utilizó el diseño de investigación tipo cuasi experimental; por medio de este tipo de investigación podemos aproximarnos a los resultados de una investigación experimental en situaciones en las que no es posible el control y manipulación absoluta de las variables, es decir, existe una 'exposición', una 'respuesta' y una hipótesis para contrastar, pero no hay aleatorización de los sujetos a los grupos de tratamiento y control, o bien no existe grupo control propiamente dicho.

Tal como afirma (Campbell. 1988, p. 191), "podemos distinguir los cuasiexperimentos de los experimentos verdaderos por la ausencia de asignación aleatoria de las unidades a los tratamientos". Los cuasi-experimentos son como experimentos de asignación aleatoria en todos los aspectos, excepto en que no se puede presumir que los diversos grupos de tratamiento sean inicialmente equivalentes dentro de los límites del error muestral (Campbell. 1988, p. 142).

¿Qué es un cuasi-experimento? Es una investigación que posee todos los elementos de un experimento, excepto que los sujetos no se asignan aleatoriamente a los grupos. En ausencia de aleatorización, el investigador se enfrenta con la tarea de identificar y separar los efectos de los tratamientos del resto de factores que afectan a la variable dependiente (Pedhazur y Schmelkin. 199: p. 277).

En las definiciones anteriores se observa que la característica principal de las investigaciones cuasi-experimentales es la ausencia de aleatorización de los tratamientos y, por lo tanto, la carencia de un control total sobre la situación. Al interpretar los resultados de un cuasi-experimento, hay que considerar la posibilidad de que se deban a otros factores no tenidos en cuenta (Cook y Campbell. 1986, p.156). En efecto, en un diseño cuasi-experimental, en comparación con los experimentales, hay más hipótesis alternativas que pueden ajustarse a los datos. Por ello, es imprescindible que el investigador tenga, en la medida de lo posible, un conocimiento de las variables específicas que el diseño cuasi-experimental utilizado no sea capaz de controlar.

El trabajo se enfoca en el grupo al que fue aplicado el experimento, seleccionado de forma intencional y dirigida, basada en la estructura del Reglamento de Calificaciones de la Escuela Superior Politécnica de Chimborazo y la aplicación de esta estructura en las aulas virtuales de la Plataforma de Educación Virtual de la ESCPOCH. Además, en la presente investigación se utilizan dos variables, las cuales van a ser analizadas y procesadas mediante una prueba estadística para comprobar la hipótesis. Estas variables son: Patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje y Estudiantes exitosos.

3.2 Tipo de Investigación

El tipo de investigación utilizada en este trabajo es descriptiva (permite la observación si afectar el comportamiento normal) ya que se pretende identificar patrones de participación de estudiantes en la plataforma virtual de aprendizaje de la ESPOCH, para luego predecir a los estudiantes que van a tener éxito en el semestre. De esta manera se detalla las características del modelo en base al algoritmo seleccionado, así como las características de los atributos aplicados en la muestra experimental. Adicionalmente, se van aplicar medios estadísticos para establecer la relación entre los diferentes atributos de calificación del grupo de estudiantes seleccionado.

Esta investigación también es aplicativa, ya que este modelo se someterá a un ensayo real con individuos similares nuevos donde deberá predecir a los estudiantes que van a tener éxito en el semestre.

3.3 Población

La población utilizada en la investigación es finita, el muestreo a utilizarse es intencionado, ya que se utilizará la base de datos histórica de los estudiantes de la Facultad de Informática y Electrónica, Escuela de Ingeniería en Sistemas, Carrera de Ingeniería en sistemas informáticos; que estén matriculados en las asignaturas virtuales de la plataforma virtual de educación, cuyos profesores utilicen dicha herramienta. De esta manera representaremos al Universo (U) misma que representa la cantidad de cursos virtuales creados en el espacio destinado para la carrera de sistemas informáticos ene le periodo Septiembre 2014 – Febrero 2015 y la población (P) que representa la cantidad de cursos virtuales utilizados por los docentes de la carrera de sistemas informáticos ene le periodo Septiembre 2014 – Febrero 2015.

P = 240 estudiantes en 8 Cursos virtuales en uso

3.4 Muestra

Como se había mencionado anteriormente, el muestreo a utilizarse es un MUESTREO DE JUICIO ya que se toma la muestra a partir de los estudiantes matriculados en los cursos virtuales cuyos docentes hayan realizado actividades como exámenes en línea, tareas, foros y recursos como documentos, paper, etc. De esta manera gracias a la exploración de la base de datos histórica tenemos que existen 8 cursos virtuales que tuvieron interacción profesor-estudiantes, de los cuales 3 de ellos se acoplaron en distinto porcentaje al Reglamento de Régimen Académico Art. 59, de la siguiente manera.

La **Tabla 1**, muestra la estructura de calificaciones vigente en el Reglamento de Régimen Académico, en el que se muestra las calificaciones por puntos de cada una de las evaluaciones que contempla la institución.

Tabla 1-3: Modelo de evaluación por asignatura

ACTIVIDADES A EVALUAR	PRIMER PARCIAL	SEGUNDO PARCIAL	TERCER PARCIAL	EVALUACIÓN FINAL	SUSPENSIÓN
Trabajo colaborativo	8 pts	10 pts	10 pts	12 pts	20 pts
Prácticas de aplicación					
Trabajo autónomo					
Pruebas parciales teóricas y/o prácticas					
Lecciones					
Consultas electrónicas					
Investigaciones					

Fuente: Régimen Académico Espoch

Basándonos en el modelo de evaluaciones acumulativas del Reglamento de Régimen Académico, procedemos a proporcionar pesos a los componentes o actividades a evaluar que fueron utilizadas desde la plataforma virtual en las 3 asignaturas descritas anteriormente.

En la **Tabla 2**, se describe los recursos (publicación de documentos como .doc, .docx, .pdf, .xlsx, .pptx) y actividades (Tareas, pruebas en línea, participación en foros, participación en wikis, participación en chat) que cada uno de las aulas virtuales han utilizado en el semestre, colocando un

peso de 1 si utilizo y 0 si no lo utilizo (propuesta por el investigador); de esta forma nos damos cuenta que el curso que más se adaptó al modelo de evaluación es el 652.

Esta información es muy importante al momento de escoger la muestra del presente estudio, ya que usando la técnica no probabilística con la estrategia sujeto-tipo permitirá tener una gran riqueza de datos puros, profundidad de datos y calidad de información. Por tal motivo se escoge el curso 652 que corresponde a la cátedra de Aplicaciones Web del quinto semestre de la carrera de Ingeniería en Sistemas de la Facultad de Informática y Electrónica en el periodo académico Septiembre 2014 – Febrero 2015, que ha tenido continuidad en la utilización de foros, chat, tareas, recursos, exámenes en línea, e historial de visitas en la Plataforma de Educación Virtual de la ESPOCH; además esta cátedra ajusta en un 48,5% del 80% de las evaluaciones acumulativas como lo especifica el modelo de evaluación del Régimen Académico de la ESPOCH; dejando fuera del estudio el 31,5% de la calificación del total acumulativa por considerar que la obtención de estas calificaciones las realizaron mediante otros medios que no necesariamente fue la plataforma virtual de la ESPOCH.

Tabla 2-3: Cursos Virtuales que utilizan el modelo de evaluación

ID CURSO	PRUEBAS			TAREAS			PARTICIPACIÓN	RECURSOS	PESO TOTAL
	P1	P2	P3	T1	T2	T3			
674	0	0	0	1	1	1	1	1	5/8
652	1	1	1	1	1	1	1	1	8/8
679	1	0	0	1	1	0	1	1	5/8

Fuente: Evirtual Espoch

Además el curso contiene un grupo de 29 estudiantes que registran múltiples transacciones en la plataforma de educación virtual de la ESPOCH.

3.5 Método

El método científico – Hipotetico-Deductivo es considerado como un camino lógico para buscar la solución a los problemas (Sánchez, 2012: p82); este trabajo de investigación aplica este método ya que por medio de la observación de un conjunto datos definidos como fuente de información con el propósito de obtener los resultados y deducir la hipótesis que se plantea.

De esta manera y siguiendo un procedimiento formal y ordenado para el análisis apropiado de estos datos; para lograr este objetivo se sigue los siguientes pasos planteamiento del problema, la

formulación de la hipótesis, el levantamiento de la información, el análisis e interpretación de datos, la comprobación de la hipótesis y la difusión de los resultados de la investigación.

3.6 Técnica e instrumentos

Mediante las técnicas de recopilación de información, como es la inspección de registros, entrevista, análisis documental de fuentes primarias y secundarias, y observación.

La técnica inspección de registros, permitirá identificar los elementos de estudio que en este caso son las aulas virtuales contenidas en la base de datos histórica de la plataforma de Educación Virtual de la ESPOCH.

La técnica de entrevista nos proporcionara la información completa de como el docente lleva el proceso de aprendizaje y el porcentaje de calificación de los estudiantes en la plataforma de educación virtual de la ESPOCH con respecto al 80% de la evaluación acumulativa establecida en el modelo de evaluación de Régimen Académico de la ESPOCH Art. 59.

La técnica de investigación de análisis documental de fuentes primarias y secundarias, proporcionará información precisa de los conceptos necesarios para aplicar los distintos descriptores que la investigación requiera. La técnica de observación, permitirá interpretar y descubrir los patrones que se descubran al momento de aplicar la técnica de predicción en el modelo de minería de datos.

Finalmente, la técnica de análisis comparativo entre las variables para establecer una relación entre las variables, que permita afirmar o negar la hipótesis propuesta en este trabajo. De esta manera en este trabajo de investigación se utiliza una serie de instrumento de acuerdo a como se aplique las técnicas antes mencionada, en ese sentido los instrumentos a utilizarse son los siguientes.

En la **Tabla 3-3**, se describe las técnicas, Instrumentos e instrumento de registro que se utilizará en este estudio.

Tabla 3-3: Técnicas e instrumentos

Técnica	Instrumento	Instrumento de registro
Inspección de registros	Lenguaje SQL	Registro de datos o tablas almacenadas en el disco duro.
Entrevista	Cuestionario	Papel e impresora
Análisis documental de fuentes primarias y secundarias	Internet, Libros, Reglamentos, artículos científicos, videos	Cuaderno de apuntes y lápiz
Observación	- Guía de Observación (Elemento a evaluar, fecha, nombre del evaluador, título de la tarea, Observaciones, escala) - Herramientas de extracción de datos	Cuaderno de apuntes y lápiz Registro de datos o tablas almacenadas en el disco duro.
Análisis comparativo	Herramientas estadísticas	Registro de datos o tablas almacenadas en el disco duro.

Realizado por: Gustavo Hidalgo.2016

3.7 Procesamiento y Análisis de datos

En la Escuela Superior Politécnica de Chimborazo (ESPOCH), han planteado muchos estudios respecto al impacto que tiene los Sistemas Manejadores de Contenido (SMC o CMS en inglés) en el desarrollo académico del estudiante universitario; algunos se centran en el estudio proponiendo estándares de evaluaciones, estudiando el comportamiento de los estudiantes en la plataforma virtual, etc. Estos estudios han ayudado mucho al proceso de formación académica de los estudiantes de la ESPOCH.

La inclusión de las plataformas de educación virtual en la Ley Orgánica de Educación Superior (LOES), como complemento a la formación del estudiante, ha influido en gran parte al uso y configuración de este tipo de herramientas; muchos docentes y profesores de las universidades del país utilizan estas herramientas como parte de su estrategia o metodología de enseñanza publicando artículos, libros, videos, wikis, foros, tareas, exámenes en línea entre otros. Esta tendencia tecnológica ha ayudado en gran medida a los estudiantes a mejora en su rendimiento académico ya que cuentan con una amplia información a cualquier hora y lugar.

Toda esta gama de recursos y actividades que los estudiantes generan por su participación en los SMC, son almacenados en forma de datos en una base de datos que están alojados en los diferentes servidores de los Centros de Datos (NOC) de las Instituciones de Educación Superior (IES); estos

datos si bien es cierto están organizados en las tablas de las bases de datos no nos brindan una información coherente para la toma de decisión por parte del docente.

La Escuela Superior Politécnica de Chimborazo viene utilizando este tipo de sistemas desde el 2005, misma que al estar integrado con el sistema académico Institucional, permite crear aulas virtuales, malla curricular, carga horaria de docentes, matriculación de estudiantes de una forma rápida y eficiente; esto ayuda al docente a tener su herramienta virtual al momento de empezar el periodo académica.

La ESPOCH cuenta con una gran cantidad de datos distribuidas en todo sus servidores como: Sistema Académico, Plataforma de Educación Virtual, Sistema de Evaluación, Sistema Financiero, Sistemas de Redes Inalámbricas, entre otros. En la actualidad se ha realizado poco a o nada con esos datos, es decir no hay un instrumento o herramienta que pueda interpretar esos datos y brindar información que ayude a adquirí conocimiento y así poder tomar decisiones con respecto a esos indicadores.

3.7.1 Metodología

Este trabajo de investigación se llevó a cabo mediante la recolección de información primaria o fuentes primarias (contienen información nueva y original, resultado de un trabajo intelectual. Son documentos primarios: libros, revistas científicas y de entretenimiento, periódicos, diarios, base de datos, documentos oficiales de instituciones públicas, informes técnicos y de investigación de instituciones públicas o privadas, patentes, normas técnicas) junto con el Departamento de Tecnologías de la Información y Comunicación de la ESPOCH, el administrador de la Plataforma Virtual de la ESPOCH, el docente de la cátedra de Aplicaciones Web; así como información secundaria o fuentes secundarias (contienen información organizada, elaborada, producto de análisis, extracción o reorganización que refiere a documentos primarios originales. Son fuentes secundarias: enciclopedias, antologías, directorios, libros o artículos que interpretan otros trabajos o investigaciones.). Utilizando técnicas de minería de datos se logró interpretar los datos para convertirlos en información y luego en conocimiento para la toma de decisiones por parte de los docentes.

De esta manera se procede a explicar los pasos que se siguieron para dar cumplimiento a los objetivos planteados.

3.7.1.1 *Determinar los patrones de participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web de la ESPOCH.*

Diariamente los usuarios de la tecnología generan una gran cantidad de datos inconscientemente, cada vez que utilizamos las redes sociales, servicios informáticos públicos o privados, páginas informativas, sistemas bancarios en línea, estamos dejando un camino importante de datos en aquellos contenedores de datos también llamados bases de datos. Muchas instituciones gubernamentales y empresas privadas aprovechan estos datos para su beneficio, ya sea social o económico.

Los datos que generan las personas nos brindan una gran gama de información, mismos que interpretados son convertidos en conocimiento que luego son usados para la toma de decisiones.

Una de las aristas de este trabajo de investigación es la detección de patrones de participación de los estudiantes dentro de la plataforma de educación virtual de la ESPOCH, para poder cumplir con el primer objetivo de esta investigación se tiene que empezar respondiendo las siguientes preguntas.

¿Qué es un patrón?, ¿Para qué se requiere la identificación de patrones? Las respuestas a estas preguntas ayudan a entender de mejor manera el alcance de este primer objetivo.

“Los patrones son series de variables constantes, identificables dentro de un conjunto mayor de datos” (poner a alguien). La identificación de patrones ayudará al docente a identificar el perfil del estudiante exitoso, de esta manera podrá motivar a sus nuevos estudiante a seguir este perfil para que también puedan tener éxito en el semestre.

Aspectos iniciales

La Plataforma de Educación Virtual (PEV) es un Sistema de Manejo de Contenidos (SMC o CMS siglas en inglés) conocido como MOODLE, mismo que está diseñado para almacenar en su base de datos todas las interacciones que los usuarios realizan en su aplicación (página web o interfaz web).

Este tipo de sistemas es conocido por manejar muchas formas de participación entre ellas tenemos: archivos, tareas, exámenes en línea o cuestionarios, foros, wikis, chat, entre otros. Este registro que dejan los estudiantes en las base de datos del MOODLE, nos permite tener una infinidad de datos se someterán a una transformación y limpieza para luego para procesarlos y convertirlos en información, luego se aplicará un algoritmo que permitirá crear conocimiento.

Procedimiento

Para poder aplicar Minería de Datos sobre un conjunto de datos hay que seguir el siguiente proceso. La **Figura 1-3**, describe el proceso de mineía de datos aplicado en este estudio.

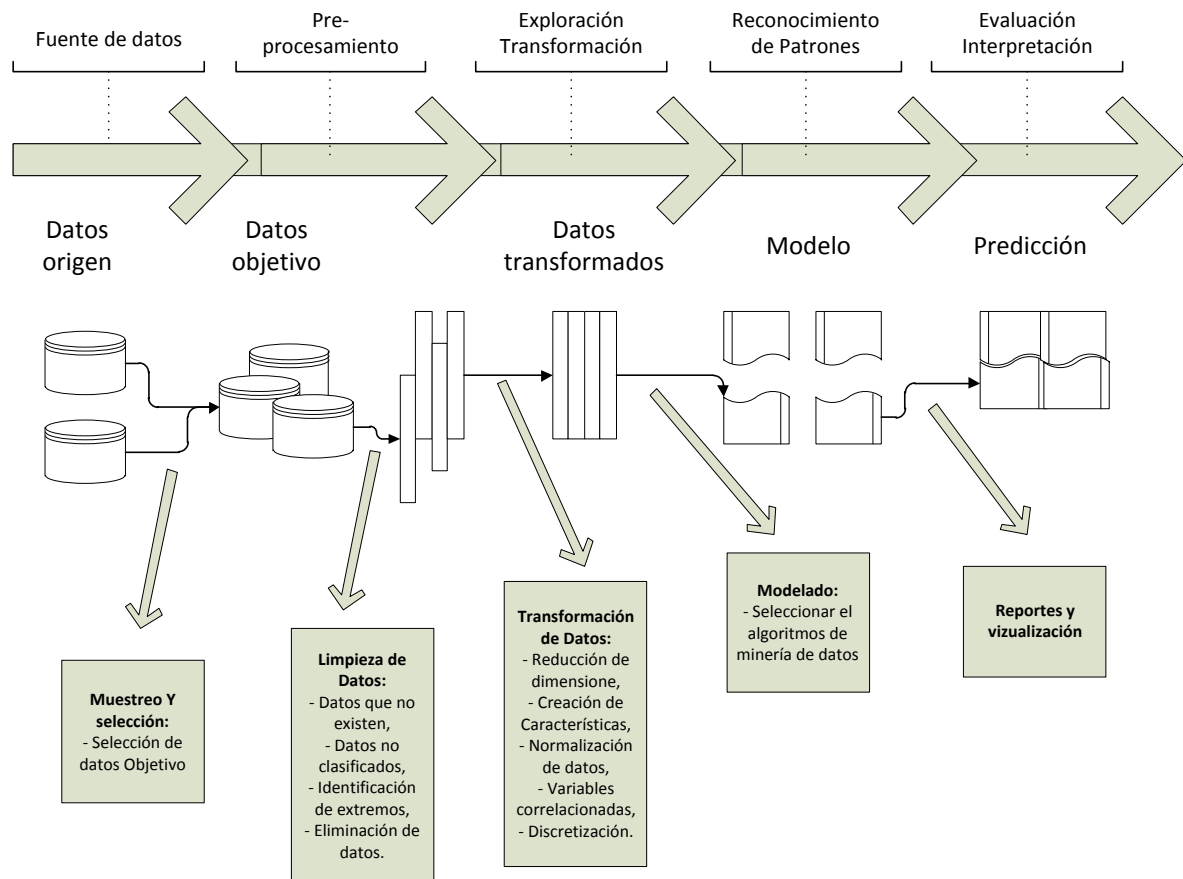


Figura 1-3: Proceso de Minería de Datos

Realizado por: Gustavo Hidalgo.2016

Fuente de datos: Son aquellas bases de datos donde se encuentran los datos origen o datos nativos, en el presente estudio se recaban datos de dos fuentes externas o de origen como son: base de datos del MOODLE y la base de datos del sistema académico.

La base de datos MOODLE está conformada por 396 tablas entre principales y secundarias, mismas que fueron restauradas en el motor de base de datos PostgreSQL con sus respectivos datos; las tablas están relacionadas y organizadas mediante el diseño conceptual y el modelo Entidad-Relación.

De esta manera tenemos que todas las actividades tienen su representación en una tabla en la base de datos, ejemplo: una tarea o asignación tiene su tabla llamada mdl_assign misma que está relacionada

directamente con la entidad principal mdl_course por medio de su id (clave principal de la tabla). Para el proceso de selección de las entidades o tablas que serán escogidas para el análisis se establecieron los siguientes parámetros según el criterio del investigador.

Usabilidad de la actividad o recurso en el aula virtual (C1)

Registro de calificación en las entidades utilizadas (C2)

Estrategia de calificación del docente aplicada en el aula virtual y que se adapte al modelo del calificación del Régimen Académico de la ESPOCH. (C3)

Registro de participación. (C4)

Basado en estos criterios se procede a seleccionar las entidades tipo secundarios, la **Tabla 4-3** presenta el análisis realizado en este estudio.

Tabla 4-3: Entidades

Entidad o Tabla	C1	C2	C3	C4
mdl_assign	SI	SI	SI	SI
mdl_assignment	NO	NO	SI	NO
mdl_badge	NO	NO	SI	NO
mdl_chat	NO	NO	SI	NO
mdl_cohort	NO	NO	SI	NO
mdl_data	NO	NO	NO	NO
mdl_event	NO	NO	SI	NO
mdl_feedback	NO	NO	NO	NO
mdl_files	NO	NO	SI	SI
mdl_forum	SI	NO	SI	SI
mdl_game	NO	NO	NO	NO
mdl_lesson	NO	NO	SI	NO
mdl_log	SI	SI	SI	SI
mdl_message	NO	NO	SI	NO
mdl_quiz	SI	SI	SI	SI
mdl_scorm	NO	NO	SI	NO
mdl_wiki	NO	NO	SI	NO
mdl_workshop	NO	NO	SI	NO
mdl_glossary	NO	NO	SI	NO
mdl_event	NO	NO	SI	NO

Realizado por: Gustavo Hidalgo.2016

Vale la pena señalar que cada una de estas entidades tiene otra entidad relacionadas de tercer nivel que contienen las participaciones y calificaciones insumo importante para el estudio.

Entidades
mdl_assign_grades
mdl_quiz_grades

Luego tenemos un conjunto de entidades que son necesariamente importantes para el análisis es decir no se puede prescindir de ellas por considerarse de nivel uno o principales.

Entidades
mdl_role_assignments
mdl_course
mdl_course_categories
mdl_context
mdl_user
mdl_role

En la Tabla 5-3 se muestra las entidades, su ubicación física, cantidad de registros activos y una descripción individual de sus características; son estas entidades las que se utilizarán para el análisis de datos en este proyecto.

Tabla 5-3: Descripción de las fuentes de datos.

Fuente	Nombre de la base de datos	Ubicación	Entidades	Número de registros	Descripción
Base de datos MOODLE	evirtual	http://127.0.0.1:5432/	mdl_assign	8878	Tareas creadas por los docentes
			mdl_assign_grades	62285	Tareas completadas por los estudiantes
			mdl_quiz	1281	Exámenes creadas por los docentes
			mdl_quiz_grades	18472	Exámenes completadas por los estudiantes
			mdl_role_assignments	174482	Asignación de usuarios a un curso determinado.
			mdl_course	6484	Conjunto de cursos creados en la plataforma.
			mdl_course_categoríe s	1932	Categoría de los cursos.
			mdl_context	110719	Contexto de los cursos.
			mdl_user	19626	Descripción completa de los estudiantes y profesores.
			mdl_role	10	Lista de roles que pueden asignar a los usuarios
mdl_log	9543110	Lista de participaciones de los usuarios			
Base de datos académico	Estados_estudiantes.xlsx	D:\Tesis_GusX\Info_Danilo	Lista de aprobados	29	Lista de aprobados

Realizado por: Gustavo Hidalgo.2016

En la **Figura 2-3**, se presenta el esquema relacional que tiene las entidades antes mencionadas, en la fuente de datos Plataforma Virtual de Educación.

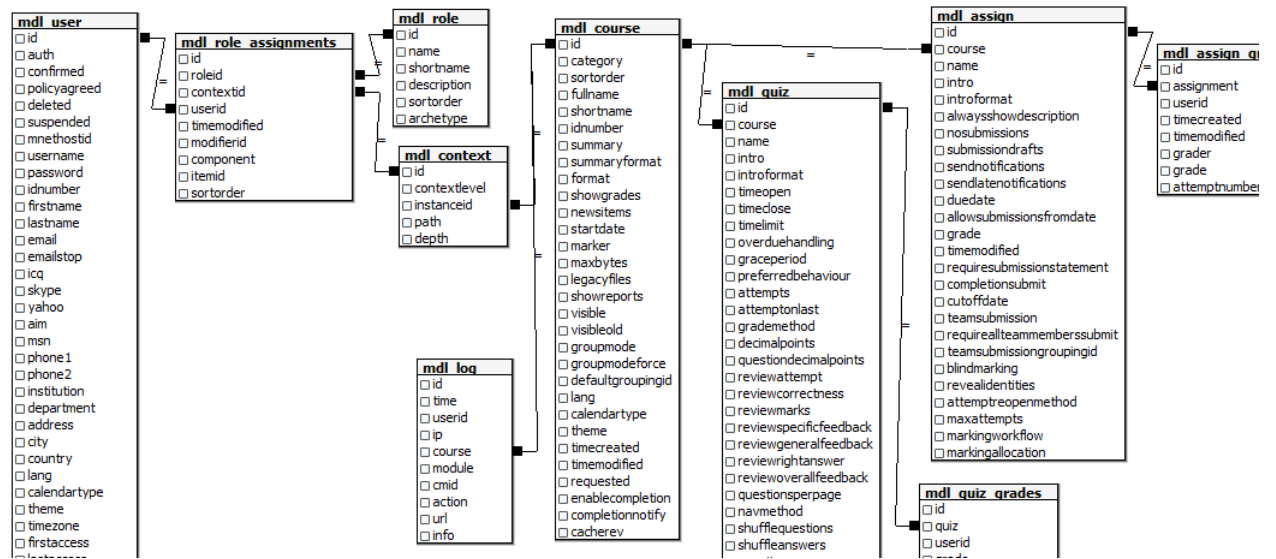


Figura 2-3: Modelo Seleccionado para el análisis.

Realizado por: Gustavo Hidalgo.2016

Pre-procesamiento: Este paso es uno de los más importantes y complejos ya que aquí se requiere moldear los datos hasta conseguir los datos objetivos, es decir se descarta lo que no nos va a servir para el análisis.

Como ya se mencionó antes, la muestra establecida intencionalmente son los estudiantes de la cátedra de Aplicaciones Web de 5^o semestre de la carrera de Ingeniería en Sistemas Informática. Para lograr identificar esta cátedra en el origen de datos “evirtual” o Plataforma de educación Virtual de la ESPOCH, tenemos que realizar una exploración y reprocesamiento de la información.

Para realizar la fase de pre-procesamiento tomamos como materia prima el origen de datos (base de datos) “evirtual” y como instrumento la herramienta ETL Pentaho. Luego los datos pre-procesados serán almacenados temporalmente en un DataWareHouse de postgresql.

En la actualidad se han realizado muchos estudios comparativos de herramientas ETL; en esta ocasión se optó por elegir la herramienta ETL basándonos en el estudio “**Implantación de Data Warehouse Open Free (13 de diciembre del 2011)**” realizado por la Universidad de la República en Montevideo Uruguay por Nicolas Gerolandi, Esteban Revelo, Germania Venzal.

A continuación se exponen un extracto de la evaluación.

Se evaluaron y compararon cada uno de los componentes por separado tomando el siguiente rango de evaluación según sus funcionalidades:

Nivel A: El componente existe y posee una cantidad de funcionalidades superior a la media.

Nivel B: El componente existe pero posee una cantidad media de funcionalidades.

Nivel C: El componente existe pero posee escasa o pobre cantidad de funcionalidades con respecto a la media.

Nivel D: El componente no se encuentra como una característica de la plataforma pero puede ser reemplazado por otro o implementado fácilmente.

Nivel E: El componente no se encuentra como una característica en la plataforma y se desconoce si puede ser reemplazado por otro.

Plataforma	Herramienta	Evaluación
Pentaho	Pentaho Data Integration (Kettle)	B
JasperSoft	Jasper ETL (Basado en TOS)	A
SpagoBI	TOS (Talend Open Studio)	A
OpenI	N/A	D
Palo	Palo ETL Server	C

Cuadro 1-3: Comparación de capacidades ETL.

Fuente: Nicolas Gerolandi

Plataforma	Tablas Pivot/Motor	Evaluación
Pentaho	JPivot/Mondrian	B
JasperSoft	JPivot/Mondrian	B
SpagoBI	JPivot/Mondrian - JPalo/Mondrian - JPivot/XMLA Server	A
OpenI	JPivot/Mondrian - JPivot/XMLA Server	A
Palo	Palo (MOLAP)	B

Cuadro 2-3: Comparación de capacidades analíticas.

Fuente: Nicolas Gerolandi

Plataforma	Creación de Reportes	Evaluación
Pentaho	Pentaho Report Designer, JasperReport, BIRT	A
JasperSoft	JasperReport	B
SpagoBI	JasperReport, BIRT	B
OpenI	N/A	D
Palo	Palo Report Manager	C

Cuadro 3-3: Comparación de capacidades de reportaría

Fuente: Nicolas Gerolandi

Plataforma	Creación/Visualización de Dashboards	Evaluación
Pentaho	CDF, JfreeChart, CCC Charts, Google Maps	A
JasperSoft	Dashboard Designer	A
SpagoBI	OpenLazlo	A
OpenI	OpenI	B
Palo	Palo	B

Cuadro 4-3: Comparación de capacidades tableros de mando.

Fuente: Nicolas Gerolandi

Plataforma	Documentación	Evaluación
Pentaho	Documentación completa, Website, Foros, Papers, wiki, blogs externos, Bibliografía, tutoriales.	A
JasperSoft	Documentación, Website, blogs externos	B
SpagoBI	Documentación completa, Website, Foros, Papers, wiki, blogs externos, tutoriales	A
OpenI	Documentación parcial, Mail Advisors, blogs externos	C
Palo	Documentación escasa	C

Cuadro 5-3: Comparación de entidad de información provista.

Fuente: Nicolas Gerolandi

Plataforma	Evaluación
Pentaho	A
JasperSoft	B
SpagoBI	A
OpenI	C
Palo	B

Cuadro 6-3: Evaluación final.

Fuente: Nicolas Gerolandi

Luego concluyen lo siguiente:

“A partir de los datos reflejados en la tabla luego de las evaluaciones se visualiza que Pentaho se diferencia bastante con respecto a SpagoBI básicamente en dos puntos principales, el primero es la cantidad de documentación e información que se encuentra disponible para consultar y el segundo la amigabilidad y facilidad de uso que brinda a usuario. Con respecto a la documentación, pensamos que es un punto a favor muy importante de Pentaho. El hecho de no tener disponible soporte técnico sobre las herramientas obliga a contar con una amplia base de conocimientos y/o una comunidad de expertos activa al momento de realizar los desarrollos y mantenimientos sobre la plataforma. También se considera importante cuan amigables sean las herramientas para los usuarios, a mayor

facilidad de uso mayor aceptabilidad tendrá sobre estos y mejor será su utilización. Por estos motivos, la plataforma que se seleccionó para la implantación de este proyecto fue Pentaho.”

Por tal motivo se escoge la herramienta ETL Pentaho para realizar el pre-procesamiento, la Extracción y Transformación de datos.

Instalación de Pentaho

Pentaho es un conjunto de herramientas de inteligencia de negocios que tiene dos versiones, la versión comercial y la versión de código abierto.

En este estudio necesitamos crear un Data Warehouse, por tal motivo se utiliza la herramienta PDI (Pentaho Data Integration) el cual es una ETL que nos permitirá extraer los datos de las base de datos origen, transformar la información a través de un modelo dimensional y cargar los resultados de la transformación en una base de datos destino tipo Data Warehouse.

Requisitos previos a la instalación del PDI

Requisitos mínimos de hardware

Procesador de arquitectura Pentium de 2.0 GHZ

768 MB de memoria RAM

Disco Duro con al menos 2 GB libres

Requisitos de software

Java Run Time Enviroment 7 (JRE)

Java Development Kit 7 (JDK)

PostgreSQL 9.4

Pasos para la instalación

Los siguientes son los pasos para instalar el PDI en un computador:

Descargar el archivo .zip del sitio Web de Pentaho que contiene el PDI: `pdi-ce-5.0.1-stable.zip` (<http://community.pentaho.com/projects/data-integration/>)

Descomprimir el archivo en cualquier ubicación dentro de C:

Instalar el JDK 7 y JRE 7

Configurar las variables de entorno. Siguiendo los siguientes pasos:

Sobre el items “Equipo”, pulsar clic derecho y buscar Propiedades.

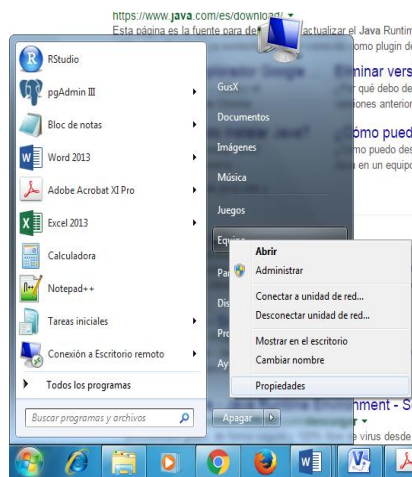


Imagen 1-3: Inicio

Realizado por: Gustavo Hidalgo.2016

Buscar “Configuración avanzada del sistema” y dar clic sobre él.



En propiedades del sistema ir a variables de entorno y dar clic.

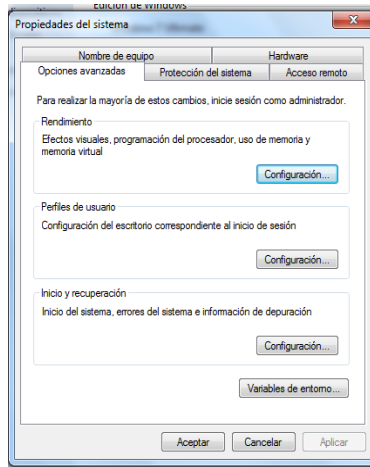


Imagen 2-3: Variables de entorno.
Realizado por: Gustavo Hidalgo.2016

En la ventana “Variables de entorno” dar clic en el botón “Nueva” y luego escribir el nombre de la variable y la dirección donde se instaló; en el caso de JDK es “JAVA_HOME” y “C:\Program Files\Java\jdk1.7.0_79”; para el JRE es “JRE_HOME” y “C:\Program Files\Java\jre7”, en ambos casos hay que pulsar el botón “Aceptar” para que se guarden los cambios.

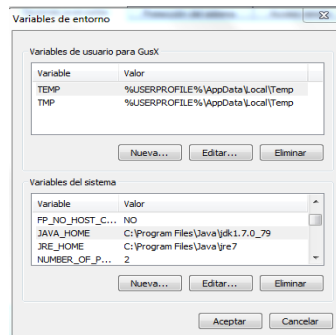


Imagen 3-3: Variables de entorno
Realizado por: Gustavo Hidalgo.2016

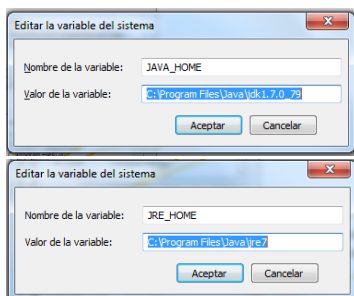


Imagen 4-3: Configuración Variables de entorno.
Elaborado: Por el autor

En el apartado “Fuentes de Datos” se definieron las entidades que se seleccionaron para el experimento, de esta manera procedemos a realizar un diagrama relacional con las entidades que formaran parte del Data Warehouse como dimensiones y tabla de hechos.

La **Figura 3-3**, describe el modelo dimensional tipo estrella copo de nieve que será implementado en el DataWare House que servirá para realizar este estudio.

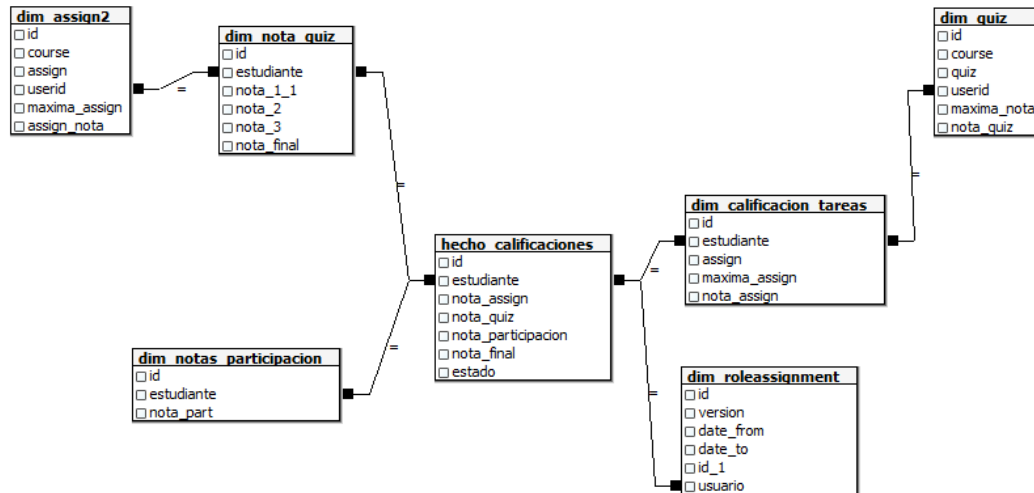


Figura 3-3: Modelo estrella para el Data Warehouse.
Realizado por: Gustavo Hidalgo.2016

Exploración Transformación: En esta parte se procede a escoger los atributos que se utilizaran para el análisis, es decir se explora los datos y se va eliminando aquellos que no considera el docente que influyan en la calificación final, luego se transforma para su análisis.

Como ya se mencionó en capítulos anteriores un Data Warehouse; es una base de datos que contiene una estructura propia de acuerdo al estudio que se quiera realizar, si bien es cierto esta base de datos tiene los mismo principios y reglas que una base de datos común no se las puede confundir ya que el Data Warehouse a diferencia de la base de datos contiene datos específicos y transformados para la investigación o estudio a realizarse. Además en este almacén de datos tenemos una estructura basada en dimensiones, medidas y hechos, en adelante se presentara la definición de estos términos aplicándolos en la solución del proyecto investigativo.

Identificación de las dimensiones

En la **Tabla 6-3**, se describe la Matriz de observación de atributos y fuentes, esta matriz nos permite saber de dónde proviene y cómo se construyen la dimensión tareas o assign.

Tabla 6-3: Descripción de la dimensión dim_assign.

Nombre de la Dimensión = dim_assign						
Tipo de tabla = dimensión						
Esquema Origen = Public/evirtual						
Tabla origen = mdl_assign, mdl_assign_grades.						
Destino				Origen		
Atributos	Descripción	Tipo dato	Clave	Esquema	Atributo	Tipo dato
mdl_assign				dim_assign		
id	indice	bigserial	Primaria	Public/ dwh_calificacione s_v2	assign	Numeric
course	indice	bigint	Secundaria	Public/ dwh_calificacione s_v2	course	Numeric
name	-	character varying	-	-	-	-
intro	-	text	-	-	-	-
introformat	-	smallint	-	-	-	-
nosubmissions	-	smallint	-	-	-	-
allowsubmissions fromdate	-	bigint	-	-	-	-
grade	Calificación	bigint	-	-	-	-
timemodified	-	bigint	-	-	-	-
requiresubmission statement	-	smallint	-	-	-	-
completionsubmit	-	smallint	-	-	-	-
cutoffdate	-	bigint	-	-	-	-
teamsubmission	-	smallint	-	-	-	-
requireallteamme mberssubmit	-	smallint	-	-	-	-
maxattempts	-	integer	-	-	-	-
markingworkflow	-	smallint	-	-	-	-
markingallocation	-	smallint	-	-	-	-
assignment	indice	bigint	-	Public/ dwh_calificacione s_v2	assign	Numeric
userid	Estudiante	bigint	-	Public/ dwh_calificacione s_v2	userid	Numeric
timecreated	-	bigint	-	-	-	-
timemodified	-	bigint	-	-	-	-
grader	Calificación	bigint		Public/ dwh_calificacione s_v2	maxima _assign	Numeric
grade	Calificación	numeric		Public/ dwh_calificacione s_v2	assign_n ota	Numeric
attemptnumber		bigint				

Fuente: Evirtual Epoch

En la **Tabla 7-3**, se describe la Matriz de observación de atributos y fuentes, esta matriz nos permite saber de dónde proviene y cómo se construyen la dimensión cuestionarios o quiz.

Tabla 7-3: Descripción de la dimensión dim_quiz.

Nombre de la Dimensión = dim_quiz						
Tipo de tabla = dimensión						
Esquema Origen = Public/evirtual						
Tabla origen = mdl_quiz, mdl_quiz_grades.						
Destino				Origen		
Atributos	Descripción	Tipo dato	Clave	Esquema	Atributo	Tipo dato
mdl_quiz				dim_quiz		
id	indice	bigserial	Primaria	Public/ dwh_calificaciones _v2	assign	Numeric
course	indice	bigint	Secundaria	Public/ dwh_calificaciones _v2	course	Numeric
name	-	character varying	-	-	-	-
intro	-	text	-	-	-	-
introformat	-	smallint	-	-	-	-
timeopen	-	bigint	-	-	-	-
timeclose	-	bigint	-	-	-	-
timelimit	-	bigint	-	-	-	-
preferredbehaviour	-	character varying	-	-	-	-
attemptonlast	-	smallint	-	-	-	-
navmethod	-	character varying	-	-	-	-
shufflequestions	-	smallint	-	-	-	-
shuffleanswers	-	smallint	-	-	-	-
questions	-	text	-	-	-	-
sumgrades	Calificación	numeric	-	-	-	-
grade	Calificación	numeric	-	-	-	-
timecreated	-	bigint	-	-	-	-
quiz	indice	bigint	-	Public/ dwh_calificaciones _v2	quiz	Numeric
userid	Estudiante	bigint	-	Public/ dwh_calificaciones _v2	userid	Numeric
timecreated	-	bigint	-	-	-	-
timemodified	-	bigint	-	-	-	-
grader	Calificación	bigint	-	Public/ dwh_calificaciones _v2	nota_m axima	Numeric
grade	Calificación	numeric	-	Public/ dwh_calificaciones _v2	Nota_q uiz	Numeric
timemodified	-	bigint	-	-	-	-

Fuente: Evirtual Epoch

En la **Tabla 8-3**, se describe la Matriz de observación de atributos y fuentes, esta matriz nos permite saber de dónde proviene y cómo se construyen la dimensión participación.

Tabla 8-3: Descripción de la dimensión dim_notas_participacion.

Nombre de la Dimensión = dim_notas_participacion						
Tipo de tabla = dimensión						
Esquema Origen = Public/evirtual						
Tabla origen = mdl_log						
Destino				Origen		
Atributos	Descripción	Tipo dato	Clave	Esquema	Atributo	Tipo dato
mdl_log				dim_notas_participacion		
id	indice	bigserial	Primaria	-	-	-
time	indice	bigint	Secundaria	Public/ dwh_calificaciones_v2	id	numeric
userid	indice	bigint	-	Public/ dwh_calificaciones_v2	estudiante	Numeric
ip	-	character varying	-	-	-	-
course	-	bigint	-	-	-	-
module	-	character varying	-	-	-	-
cmid	-	bigint	-	-	-	-
action	-	character varying	-	-	-	-
url	-	character varying	-	-	-	-
info	-	character varying	-	-	-	-

Fuente: Evirtual Epoch

En la **Tabla 9-3**, se describe la Matriz de observación de atributos y fuentes, esta matriz nos permite saber de dónde proviene y cómo se construyen la dimensión role o roles de los participantes (docentes y estudiantes).

Tabla 9-3: Descripción de la dimensión dim_roleassignment.

Nombre de la Dimensión = dim_roleassignment						
Tipo de tabla = dimensión						
Esquema Origen = Public/evirtual						
Tabla origen = mdl_user, mdl_context, mdl_role, mdl_role_assignment						
Destino				Origen		
Atributos	Descripción	Tipo dato	Clave	Esquema	Atributo	Tipo dato
mdl_role_assignment				dim_roleassignment		
id	indice	bigserial	Primaria	-	-	-
roleid	indice	bigint	Secundaria	Public/ dwh_calificaciones_v2	id	numeric
userid	indice	bigint	-	Public/ dwh_calificaciones_v2	usuario	Numeric
contextid	-	bigint	-	-	-	-
timemodified	-	bigint	-	-	-	-
modifierid	-	bigint	-	-	-	-
component	-	character varying	-	-	-	-
itemid	-	bigint	-	-	-	-
sortorder	-	bigint	-	-	-	-

Fuente: Evirtual Epoch

En la **Tabla 10-3**, se describe la Matriz de observación de atributos y fuentes, esta matriz nos permite saber de dónde proviene y cómo se construyen el hecho calificaciones que es la tabla principal para el estudio.

Tabla 10-3: Descripción de la dimensión hecho_calificaciones.

Nombre de la Dimensión = hecho_calificaciones						
Tipo de tabla = hechos						
Esquema Origen = Public/dwh_calificaciones_v2						
Tabla origen = dim_roleassignment, dim_notas_participacion, dim_quiz, dim_assign						
Destino				Origen		
Atributos	Descripción	Tipo dato	Clave	Esquema	Atributo	Tipo dato
mdl_assign				Hecho_calificaciones		
id	índice	bigserial	Primaria	-	-	-
userid	índice	bigint	Secundaria	Public/ dwh_calificaciones_v2	estudiante	Numeric
assign	índice	bigint	Secundaria			
maxima_assign	Calificación	bigint	-	-	-	-
assign_notas	Calificación	numeric	-	Public/ dwh_calificaciones_v2	nota_assign	Numeric
mdl_quiz				Hecho_calificaciones		
id	índice	bigserial	Primaria	-	-	-
userid	índice	bigint	Secundaria	Public/ dwh_calificaciones_v2	estudiante	Numeric
quiz	índice	bigint	Secundaria	-	-	-
maxima_quiz	Calificación	numeric				
nota_quiz	Calificación	numeric		Public/ dwh_calificaciones_v2	nota_quiz	Numeric
dim_notas_participacion				Hecho_calificaciones		
id	índice	bigserial	Primaria			
estudiante	índice	bigint	Secundaria	Public/ dwh_calificaciones_v2	estudiante	Numeric
nota_part	Calificación	double precision		Public/ dwh_calificaciones_v2	nota_participacion	Numeric
Estado_estudiantes.xlsx				Hecho_calificaciones		
estudiante	índice	numeric		Public/ dwh_calificaciones_v2	estudiante	Numeric
estado		String		Public/ dwh_calificaciones_v2	estado	{si,no}

Fuente: Evirtual Espoch

Implementación de la Extracción

De manera resumida los pasos para la implementación son:

Creación de la base de datos que contendrá el repositorio, en el servidor de base de datos PostgreSQL se crea una base de datos llamada “dwh_calificaciones_v2” de la siguiente manera. En la consola de administración del motor de base de datos PostgreSQL visor pgadmin III, damos clic derecho con el botón derecho del mouse en el ítems **Database**. Luego damos clic en New Database.

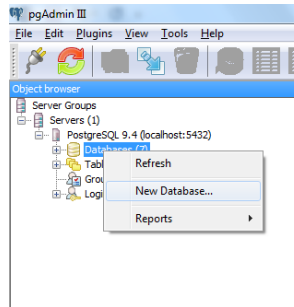


Imagen 5-3: Crear Base de datos.
Realizado por: Gustavo Hidalgo.2016

En “Name” escribimos el nombre de la base de datos tipo Data Warehouse.
“dwh_calificaciones_v2”.

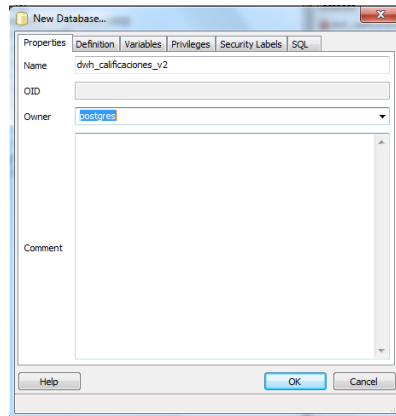


Imagen 6-3: Nombre de Base de datos.
Realizado por: Gustavo Hidalgo.2016

Creación del repositorio temporal en Pentaho. En este momento del proyecto vamos a trabajar con la herramienta ETL Pentaho.

Para iniciar la herramienta ETL, nos ubicamos en D:\Tesis_GusX\Herramientas_Mineria\pdi-ce-5.0.1-stable\data-integration, seleccionamos el archivo “Spoon” y damos clic derechos sobre este, luego seleccionamos “Ejecutar como administrador”.

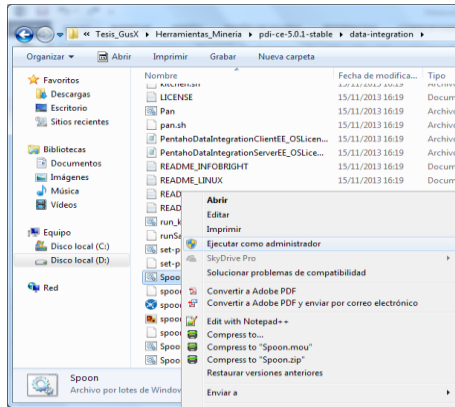


Imagen 7-3: Ejecutar PDI.
Realizado por: Gustavo Hidalgo.2016



Imagen 8-3: Entorno de trabajo PDI.
Realizado por: Gustavo Hidalgo.2016

La herramienta ETL, muestra su entorno de trabajo, donde tenemos que ir definiendo todos los elementos que nos ayudarán al pre-procesamiento de los datos al nuevo almacén de datos. Antes es necesario explicar cómo se realizó el análisis de las entidades que se utilizaran.

Reconocimiento de patrones: Este paso requiere de un estudio previo, ya que aquí se debe escoger el algoritmo que mejor se adapte al reconocimiento de patrones y su posterior uso.

Para el reconocimiento de patrones, se requiere almacenar en la tabla de hechos todos los datos necesarios para la minería de datos que se quiere aplicar dependiendo de la técnica que se elija, esta técnica será escogida gracias a el análisis de la Curva ROC. Mientras tanto tenemos que realizar la construcción del Data Warehouse como se especificó en el apartado anterior.

Se crearan 2 almacenes de datos, uno destinado a las calificaciones y el otro a las participaciones:

Data Warehouse “calificaiones”: El modelo que se pretende construir es el especificado en la **Imagen 3**. En el espacio de trabajo del Spoon diagramamos los siguientes Steps o pasos.

Dim_assign2: Para crear esta tabla en el almacén de datos diseñamos una transformación llamada, “**Trans_tareas**”, misma que contiene los siguientes elementos. Esta transformación permitirá crear una tabla en el repositorio de datos con el nombre de dim_assign, misma que contendrá todos los datos migrados desde la fuente principal “eviartual”. La Imagen 9-3, muestra los elementos que se utilizan en el Pentaho para realizar la migración antes mencionada. Vale la pena indicar que en este proceso de transformación también se ejecuta la limpieza de datos que consiste en cambiar valores nulos por el número cero (0).

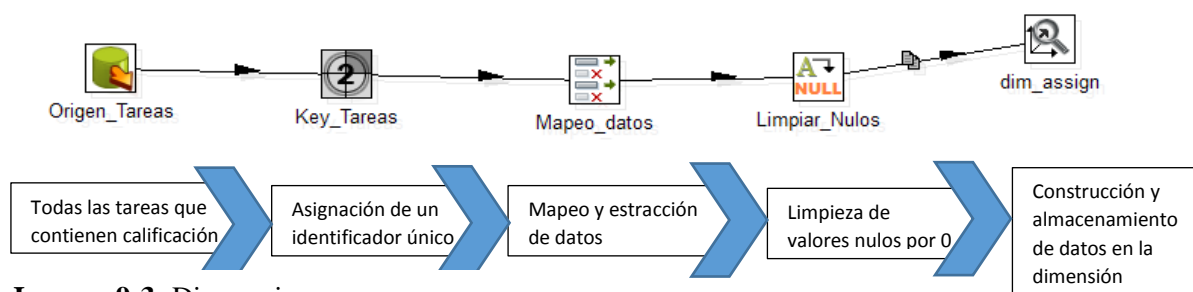


Imagen 9-3: Dim_assign.
Realizado por: Gustavo Hidalgo.2016

Dim_quiz: Para crear esta tabla en el almacén de datos creamos una transformación llamada, “**Trans_cuestionario**”, misma que contiene los siguientes elementos. Esta transformación permitirá crear una tabla en el repositorio de datos con el nombre de dim_quiz, misma que contendrá todos los datos migrados desde la fuente principal “eviartual”.

La **Imagen 10-3**, muestra los elementos que se utilizan en el Pentaho para realizar la migración antes mencionada. Vale la pena indicar que en este proceso de transformación también se ejecuta la limpieza de datos; que consiste en cambiar valores nulos por el número cero (0).

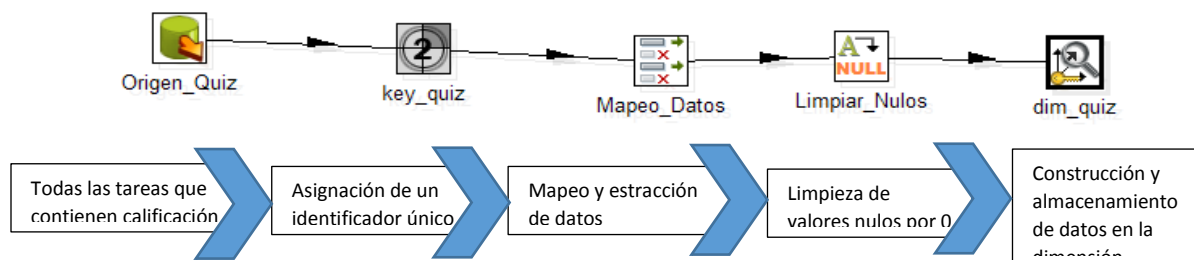


Imagen 10-3: Dim_quiz.
Realizado por: Gustavo Hidalgo.2016

Dim_roleassignment: Para crear esta tabla en el almacén de datos creamos una transformación llamada, “**Trans_roles**”, misma que contiene los siguientes elementos. Esta transformación permitirá crear una tabla en el repositorio de datos con el nombre de dim_roleassignment, misma que contendrá todos los datos migrados desde la fuente principal “eviartual”. La **Imagen 11-3**, muestra los elementos que se utilizan en el Pentaho para realizar la migración antes mencionada. Vale la pena indicar que en este proceso de transformación también se ejecuta la limpieza de datos; que consiste en cambiar valores nulos por el número cero (0).

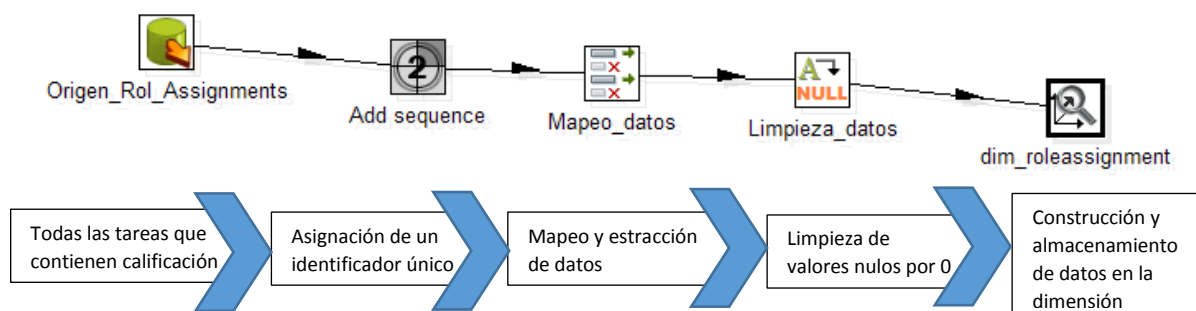


Imagen 11-3: Dim_roleassignment.

Realizado por: Gustavo Hidalgo.2016

Hay que tomar en cuenta que no todas las tareas contempladas en el Entorno Virtual de Aprendizaje (EVA), fueron utilizadas para la nota final, por tal motivo se tiene que seleccionar cuales son las que el docente utilizó para formar parte de la nota total del curso. La siguiente tabla expone el criterio de evaluación del docente respecto a esta actividad (Assign o tarea).

Tabla 11-3: Descripción de calificaciones de tareas.

Tareas - Assign						
	Total (Pts)	%	Con EVA (Pts)	%	Sin EVA (Pts)	%
Lecciones	1	3.6	1	3.6	0	0
Tareas	1	3.6	1	3.6	0	0
Informes	2	7.1	2	7.1	0	0
Total	4	14.3	4	14.3	0	0

Realizado por: Gustavo Hidalgo.2016

De la misma manera se exponen el criterio de evaluación del docente en lo que respecta a los cuestionarios o también conocidos como exámenes en línea.

Tabla 12-3: Descripción de calificaciones de los cuestionarios.

Cuestionario - Quiz						
	Total (Pts)	%	Con EVA (Pts)	%	Sin EVA (Pts)	%
Prueba 1	6	21.5	6	21.5	0	0
Prueba 2	6	21.5	3	10.7	3	10.7
Prueba 3	4	14.2	2	7.1	2	7.1
Total	16	57.2	11	39.3	5	17.8

Realizado por: Gustavo Hidalgo.2016

En esta parte de la Extracción y transformación se procede a crear una tabla temporal de dimensiones llamada “dim_tmp_assign” en la transformación “**Trans_Totales_NotasAssign**”; de esta manera se extrae mediante u SQL los datos de todas las lecciones, tareas e informes que se encuentran en las tablas origen de la fuente “evirtual”.

Luego se limpia todos los datos nulos que se encuentran en los campos, colocando el valor “0” a aquellos valores NULL, luego se realiza un extracción de datos para colocar los datos en otra tabla de dimensión temporal llamada “**dim_calificacion_tareas**”; todos estos datos son sometidos a un filtro con el objetivo de separar una tarea que es usadas como complemento de un Quiz o cuestionario (luego este dato tiene que ser sumado en la primera nota de las pruebas), esto se realiza ya que el docente en su primera prueba utilizo un cuestionario y una tarea para su primera nota.

De esta manera por un lado fluye el resultado de la tarea 1539 por ser parte de la prueba práctica de la primera nota, y por el otro lado las demás tareas con sus respectivas notas.

De esta manera se procede a cambiar los puntos en porcentaje según la **tabla 12-3**, al realizar la unión de los resultados, limpieza y ordenamiento de información para finalmente guardar estos datos transformados en la tabla “dim_tmp_assign”, luego se convertirá en un insumo importante para el cálculo al final del almacén de datos.

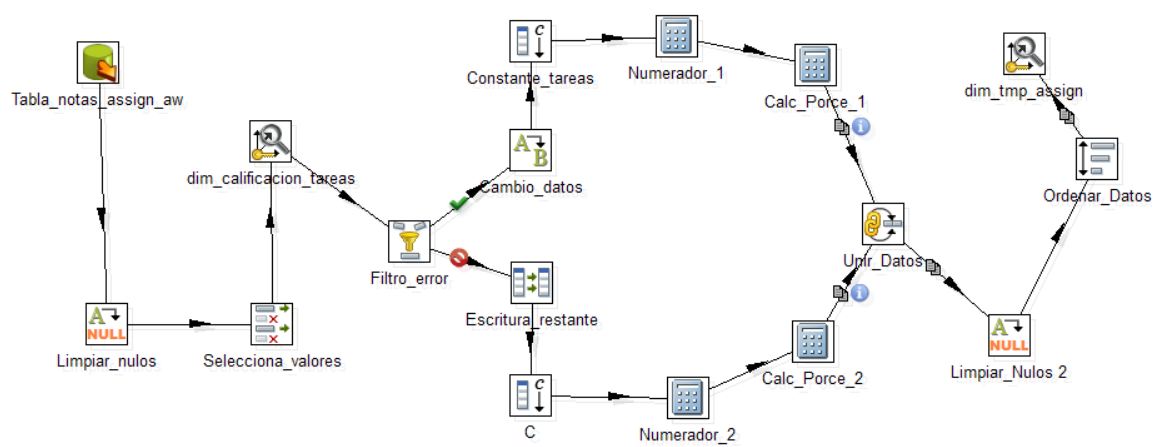


Imagen 12-3: dim_tmp_assign.
Realizado por: Gustavo Hidalgo.2016

Con la siguiente transformación almacenamos los datos de las calificaciones de las tareas y los cuestionarios en una tabla de hechos temporal llamada “**Trans_HechoCalificaciones**”. De la misma forma la transformación realiza y ejecuta la limpieza de los nulos a ceros.

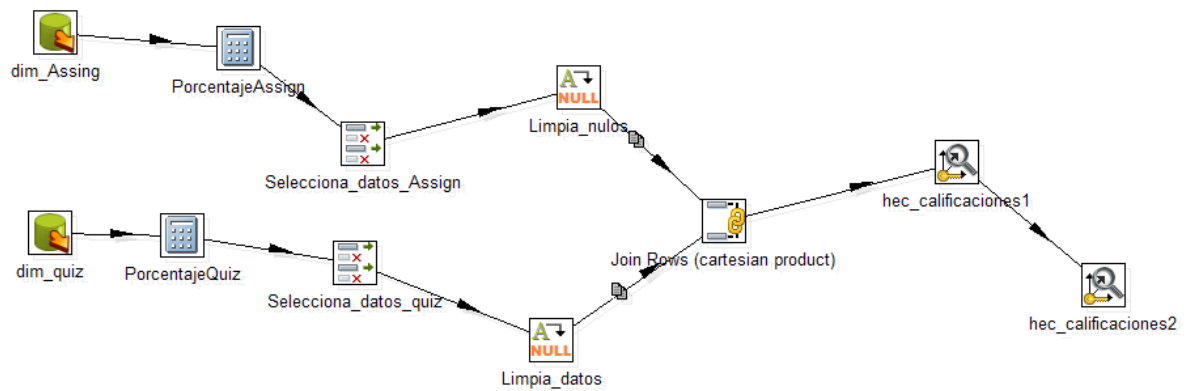


Imagen 13-3: hec_calificaciones2.
Realizado por: Gustavo Hidalgo.2016

De esta manera tendremos dos tablas temporales “dim_tmp_assign” (especifica la tarea que es tratada como cuestionario) y “hec_calificaciones” (compendio de todas las calificaciones clasificadas según las tablas de calificaciones del docente).

Ya tenemos un contenido de todas las notas tanto tareas como cuestionarios en una tabla de hechos temporal, recordemos que el docente en la nota 1 utilizó una tarea y un cuestionario para construir la primera nota por tal motivo utilizamos la transformación llamada “**Trans_Nota1**”.

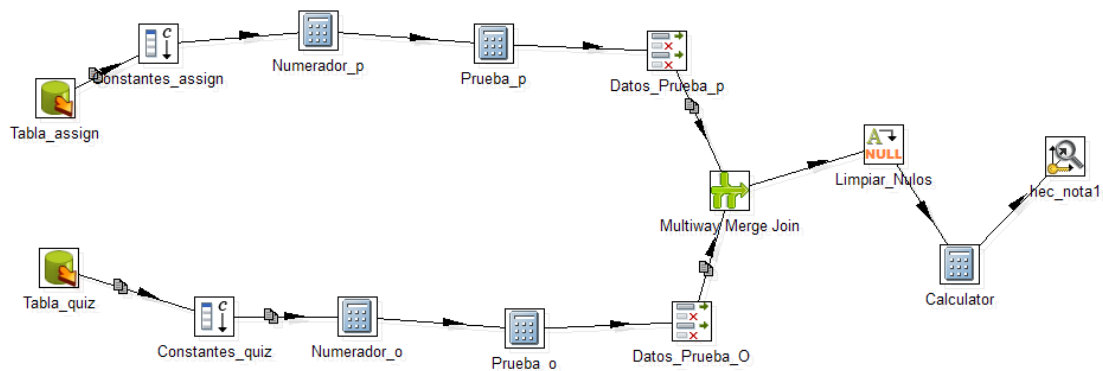


Imagen 14-3: hec_notas1.

Realizado por: Gustavo Hidalgo.2016

Esta transformación tiene el objetivo de ingresar a la tabla de hechos temporal “**Trans_HechoCalificaciones**”, identificar la tarea y el cuestionario de la primera nota para primero ordenarlo de forma ascendente, transformar la calificación a porcentaje, extraer los datos, luego unir en una sola tabla estos datos; por último limpiar los nulos si los hubiese, sumar los porcentajes de calificación de la tarea y el cuestionario para luego almacenarlos en la tabla denominada “**hec_notas1**”; esta tabla se hace necesaria por las características propias de la forma de calificación del docente.

Es importante tener en cuenta que cada docente tiene su estrategia de calificación dependiendo de los recursos y actividades que utilice en el EVA.

Antes de empezar la extracción de la tabla de hecho tenemos que organizar y categorizar los atributos que van a formar parte de esta tabla de hechos, los atributos son: nota_assign, nota_quiz, nota_participación, nota_final, estado.

Para la nota_assign se construye la dimensión dim_notas_assign. Esta tabla contenida en el repositorio contendrá todas las notas de tareas en porcentaje de los estudiantes sumadas y posteriormente almacenadas en la tabla según el parcial a quien corresponda.

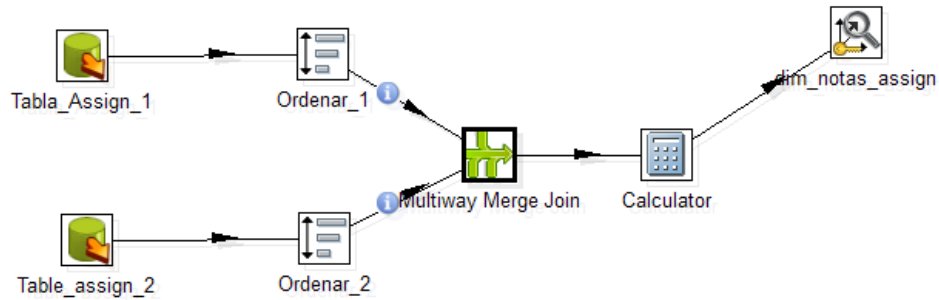


Imagen 15-3: dim_notas_assign.
Realizado por: Gustavo Hidalgo.2016

Para la nota_quiz se construye la dimensión dim_notas_quiz; Esta tabla contenida en el repositorio contendrá todas las notas de tareas en porcentaje de los estudiantes sumadas y posteriormente almacenadas en la tabla según el parcial a quien corresponda.

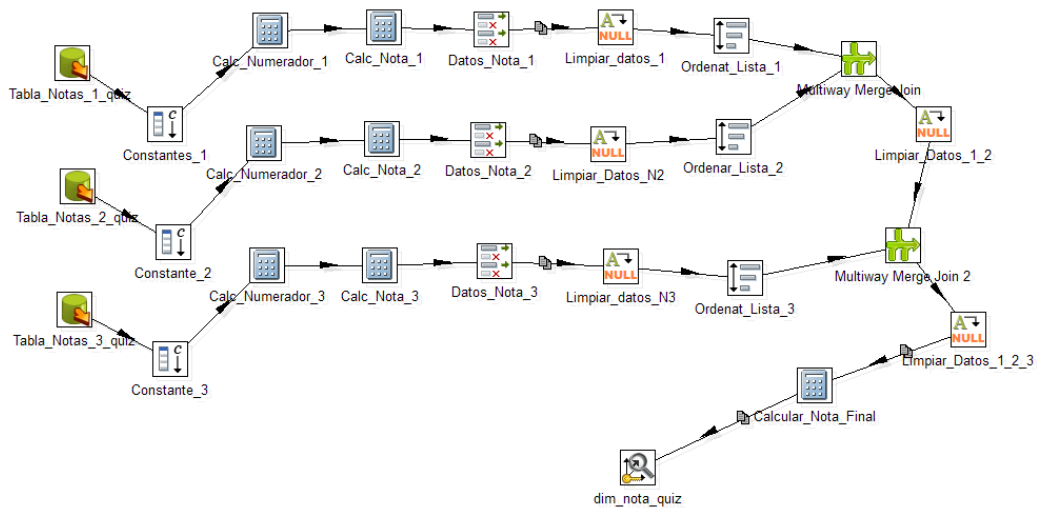


Imagen 16-3: Dim_nota_quiz.
Realizado por: Gustavo Hidalgo.2016

Para la nota de participación se construye la dimensión dim_notas_participacion; Esta tabla contenida en el repositorio contendrá todas las notas de tareas en porcentaje de los estudiantes sumadas y posteriormente almacenadas en la tabla según el parcial a quien corresponda.

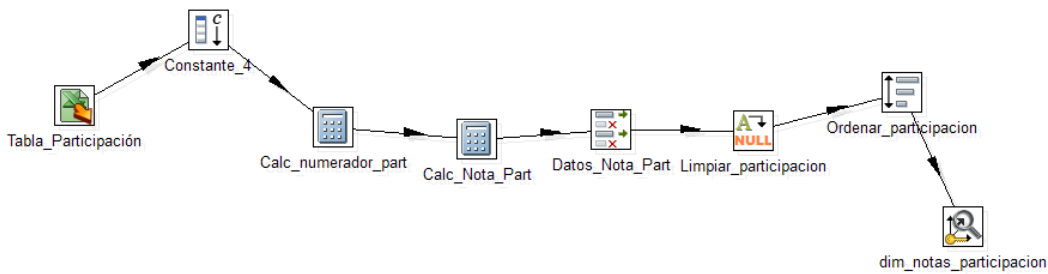


Imagen 17-3: dim_notas_participación.
Realizado por: Gustavo Hidalgo.2016

Creando tabla de hechos

Finalmente construimos la tabla de hechos “**hecho_calificaciones**”, misma que será sometida a los estudios de minería de datos correspondientes. Luego de tener los porcentajes de las calificaciones de cada uno de los argumentos almacenados en tablas según su valor por parcial; se procede a la extracción y transformación de la tabla de hechos, esta tabla consume información del resto de dimensiones extraídas anteriormente.

La transformación primero ordena los datos según el código del estudiante y luego une los conjuntos de datos de las tablas de dimensiones; luego calcula los promedios de los atributos según la estrategia de calificación estipulada en el silabo de la materia.

En este caso es la suma de las notas de tareas, cuestionarios y participación.

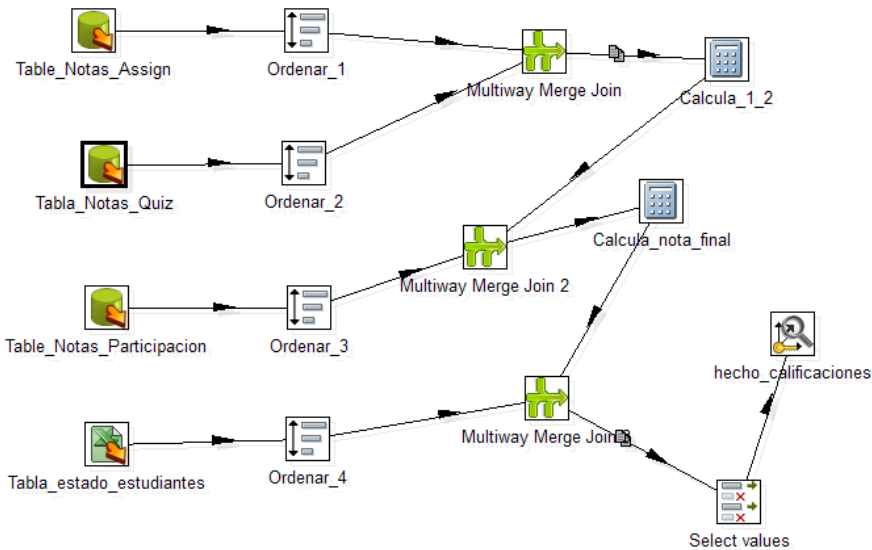


Imagen 18-3: hecho_calificaciones.

Realizado por: Gustavo Hidalgo.2016

Evaluación e Interpretación: Finalmente los resultados obtenidos se someterán a su evaluación y posterior interpretación.

Como ya se estableció en capítulos anteriores la minería de datos utiliza técnicas de Inteligencia Artificial para el aprendizaje automático, cuyo objetivo es estrictamente extraer información de un conjunto de datos para aprender automáticamente de ellos.

Las técnicas de minería de datos están sujetas a continuas evoluciones gracias a las investigaciones en el campo de: bases de datos, reconocimiento de patrones, sistemas expertos, estadísticas, recuperación de información, entre otras.

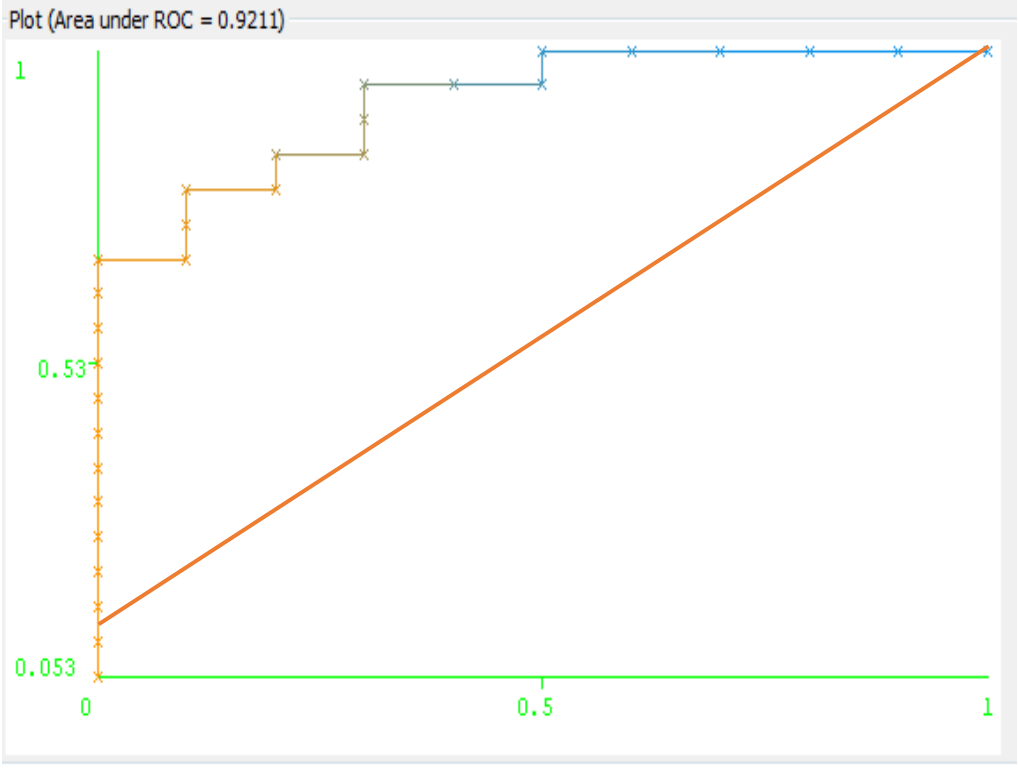
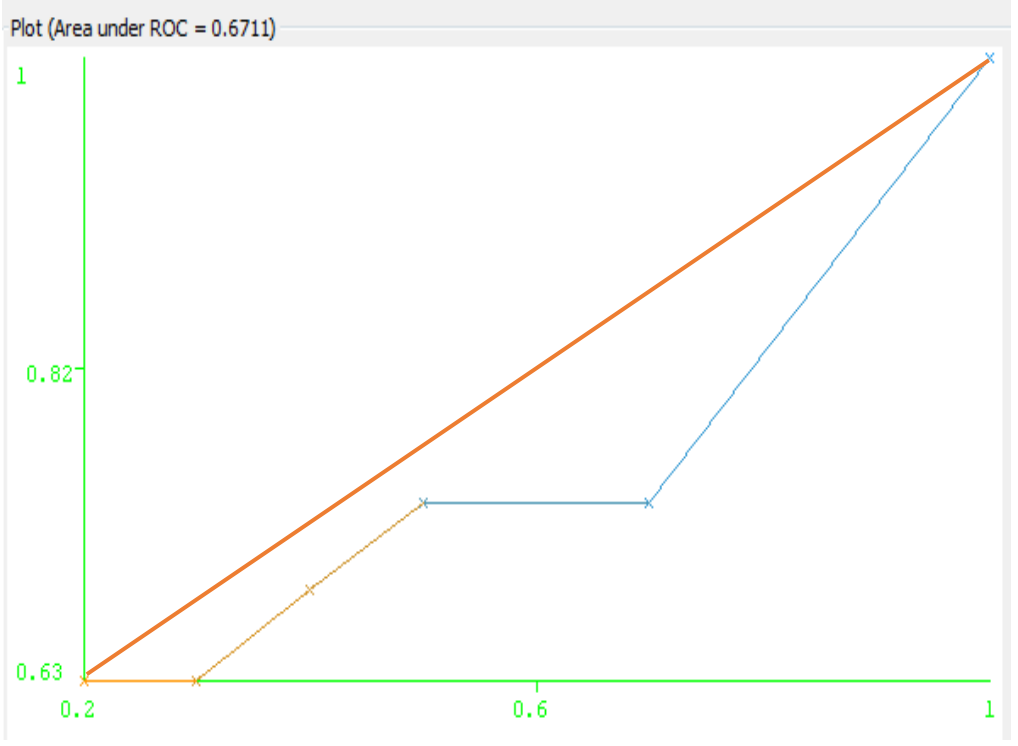
Estas técnicas son los conocidos como algoritmos los mismos que se clasifican en supervisados y no supervisados. Para este trabajo por su carácter deductivo y además que su variable discriminante es cualitativa se requiere utilizar algoritmos supervisados o predictivos.

Los algoritmos que se someterán a pruebas para encontrar el que mejor se adapte al problema que se plantea son: Redes neuronales, árbol de decisión, bosques aleatorios, bayes, vecino más cercano, Ada Boost M1.

Para conocer cuál de estos algoritmos predictivos supervisados es el que mejor se adapta a l problema presente en este trabajo se debe utilizar la curva ROC (Receiver Operating Characteristic, o Característica Operativa del Receptor) y el área bajo la curva ROC.

Esta técnica consiste en reconocer primero gráficamente cómo se comporta la curva sobre la pendiente creciente, si la curva del modelo esta sobre esta pendiente el modelo tienen una gran probabilidad de ser eficiente, mientras más alejada esta de la pendiente el modelo será más eficiente.

De la misma forma si el área bajo la curva del modelo se acerca al 100% el modelo será más eficiente.

Algoritmo	Área ROC	Gráfica (Weka)
Redes neuronales	0.9211	 <p>Plot (Area under ROC = 0.9211)</p>
Árboles de decisión	0.6711	 <p>Plot (Area under ROC = 0.6711)</p>

<p>Bosques aleatorios</p>	<p>0.7947</p>	<p>Plot (Area under ROC = 0.7947)</p>
<p>Bayes</p>	<p>0.8789</p>	<p>Plot (Area under ROC = 0.8789)</p>

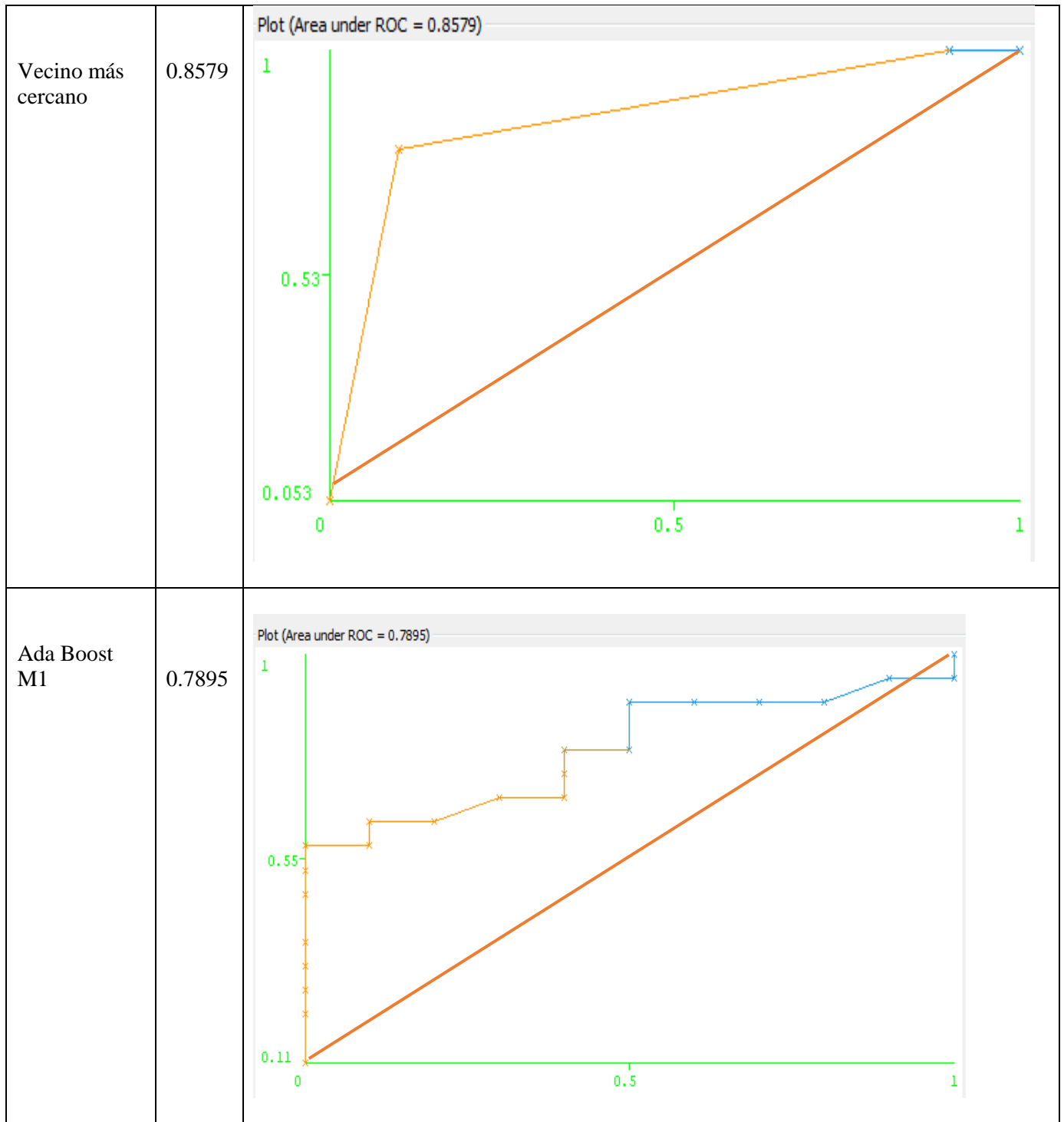


Figura 4-3: Análisis de los modelos en la curva ROC, weka.
 Realizado por: Gustavo Hidalgo.2016

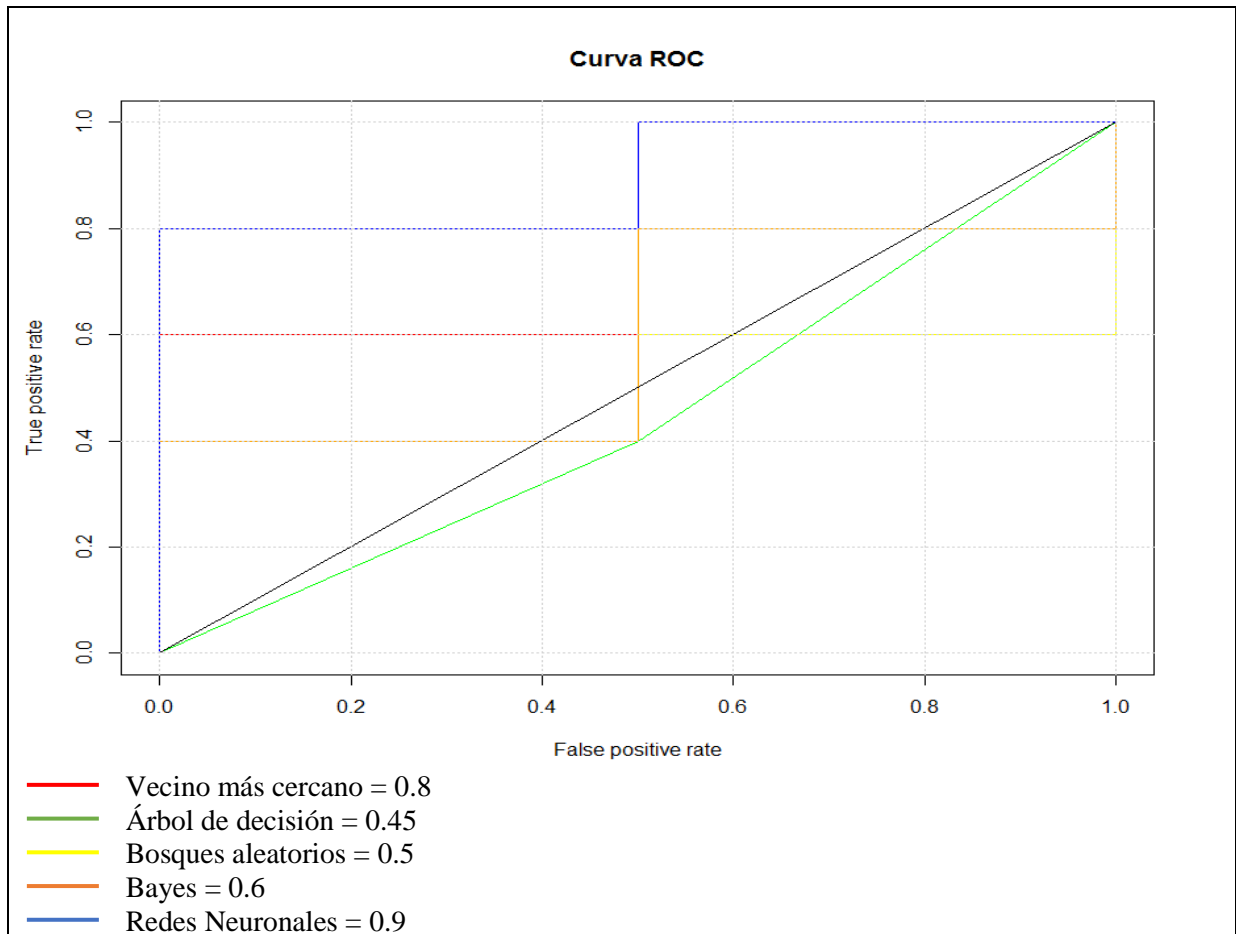


Figura 5-3: Análisis de los modelos en la curva ROC, código R.

Realizado por: Gustavo Hidalgo.2016

Bajo estas 2 premisas que se presenta en la **tabla 14 y tabla 15** el algoritmo que mejor se adapta con un 90% al problema planteado es el de **Redes Neuronales Multicapa**. Luego de haber definido el algoritmo que utilizaremos en este estudio se procede a utilizar los paquetes de minería de datos para descubrir patrones como el Rstudio y el Weka.

Patrones

Atributos más representativos

En esta sección identificaremos los patrones que según el investigador se consideran importantes para que un estudiante sea exitoso en el semestre actual.

Desde el software weka podemos encontrar aquellos argumentos mas significativos o que influyen más para que un estudiante valla por el rumbo del éxito en el semestre.

Primero nos conectamos a la tabla de hechos del Data Warehouse que construimos anteriormente, “hecho_calificaciones”.

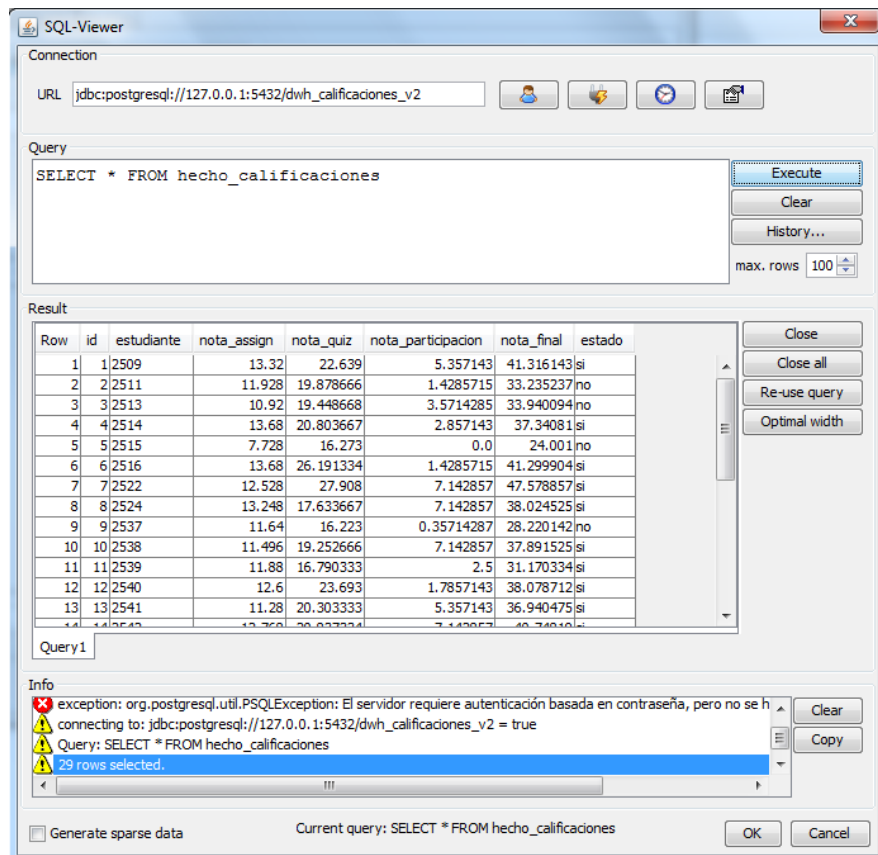


Imagen 19-3: Conexión weka con tabla de hechos.
Realizado por: Gustavo Hidalgo.2016

Este procedimiento no permitirá conocer los atributos más significativos, esto no quiere decir que los otros atributos no son importantes, ya que el sistema utiliza todos los atributos para realizar las predicciones.

En Weka, la selección de atributos se puede hacer de varias maneras. La más directa es usando la pestaña de attribute selection. Tenemos que seleccionar un método de búsqueda y un método de evaluación. En este caso se escoge:

Evaluación de atributos.

Método de búsqueda = Ranker

Método de evaluación = InfoGainAttributeEval

El método de test para evaluar los atributos será de crossvalidation de 5 hojas. Esto hará que la evaluación sea 5 veces más lenta, pero más precisa.

De esta manera podemos darnos cuenta que los atributos más representativos son **nota_assign** y **nota_participación**. Es decir, que el atributo **nota_assign** está altamente relacionado con el atributo **nota_participación**.

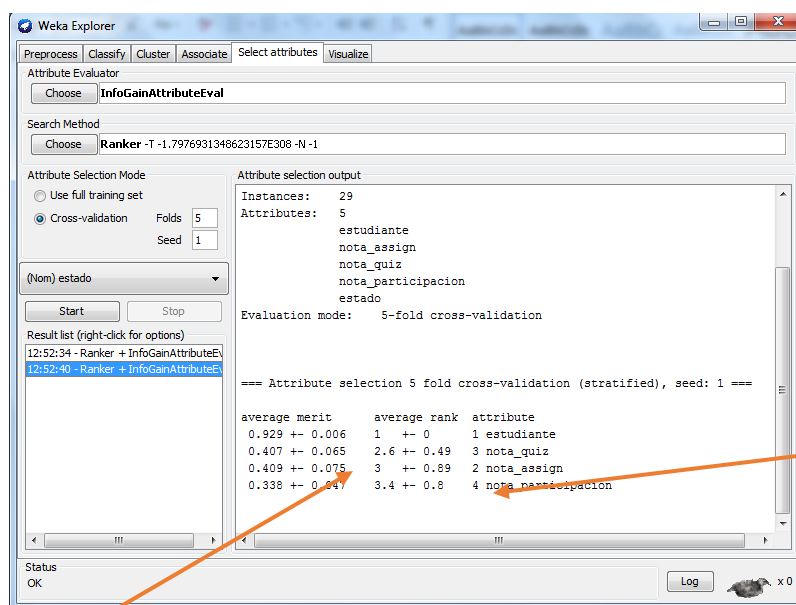
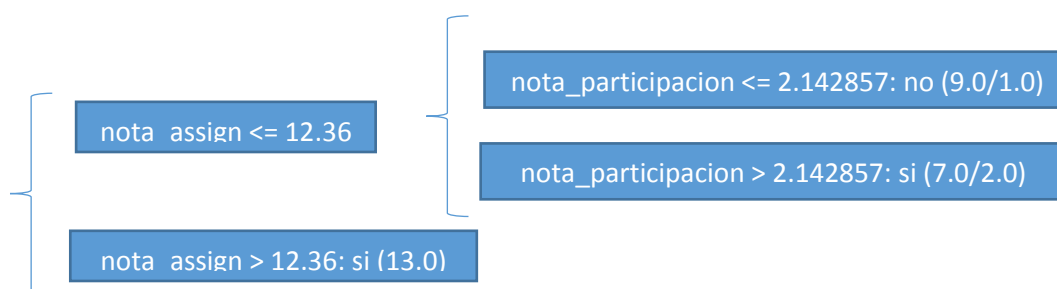


Imagen 20-3: Detección de patrones.

Realizado por: Gustavo Hidalgo.2016

Reglas de peso

El patrón anterior hacíamos referencia de los atributos más representativos, enfocados en ese principio nos damos cuenta que los estudiantes deben seguir los siguientes patrones.



Si un estudiante o grupo de estudiantes tiene valores menores e iguales que $\frac{12.36\%}{14.3\%}$ en el atributo **nota_assign** debe tener por menos un valor mayor a $\frac{2.14\%}{7.1\%}$ en el atributo **nota_participacion**, para que 7 de 9 estudiantes **APRUEBEN** el curso, es decir una probabilidad de que un 24,1% de los estudiantes del curso **APRUEBEN**.

$$Pob. Aprobación = \frac{7}{29} = 0,241$$

Si un estudiante o grupo de estudiantes tiene que tener por lo menos una valor mayor a $\frac{12.36\%}{14.3\%}$ en el atributo **nota_assign**, para que al menos 13 de cada 29 estudiantes **APROBARAN** el curso, es decir una probabilidad de que un 44,8% de los estudiantes del curso **APRUEBEN**.

$$Pob. Aprobación = \frac{13}{29} = 0,448$$

Relación entre forum y resource

La participación en los foros y la interactividad de los recursos tienen una estrecha relación ya que los participantes basan sus criterios en los foros según lo que leyeron o investigaron en los archivos que el docente publicó en el EVA. Bajo esa premisa tenemos el siguiente gráfico.

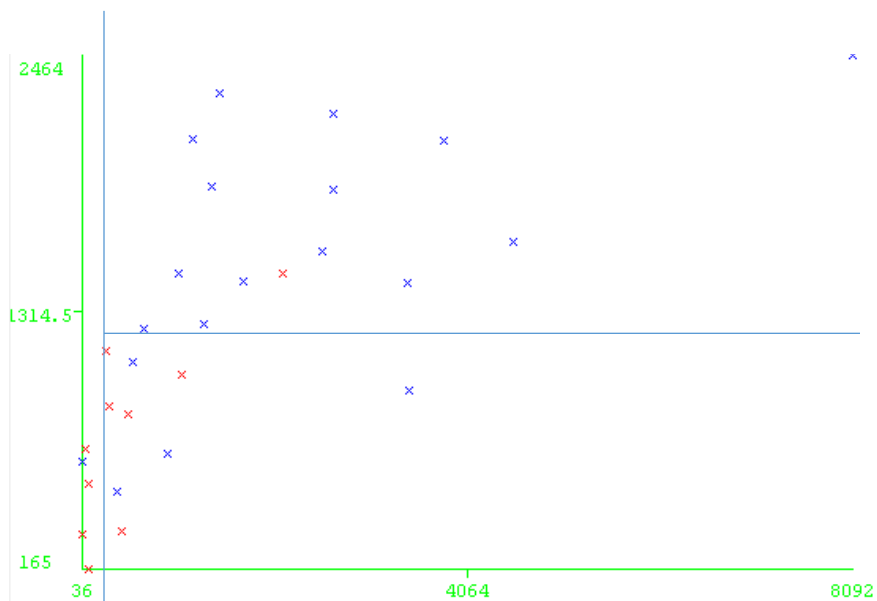


Imagen 21-3: Detección de patrones en foros.
Realizado por: Gustavo Hidalgo.2016

Los puntos de color azul representa el estado de Aprobado, los puntos rojos por lo contrario representan los reprobados. El gráfico presenta una notable característica de aprobados por encima de la línea, que se encuentra entre [X: 696; Y: 1314.5]

3.7.1.2 Identificar el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web, compararlas con las calificaciones obtenidas en las evaluaciones acumulativas.

Luego de la construcción de DataMart denominado “dwh_participación” y con ello la construcción de la tabla de hechos “hecho_participacion”, se genera la participación individual por cada módulo que está instalados en el EVA.

La tabla que a continuación se presenta nos indica el número de participaciones con su respectiva valoración en porcentajes y al final el campo participación, su pondera de acuerdo a la planificación del docente.

Tabla 13-3: Participación consolidada.

Estudiantes	Forum	course	quiz	Resource	assign	choice	page	glossary	book	folder	wiki	blog	Participación												
2509	1045	2,21	2849	4,04	2134	4,16	1485	4,10	2310	3,65	319	4,64	286	4,41	99	1,78	33	5,58	77	4,77	22	5,18	0	0,00	5,4
2511	105	0,22	250	0,35	1025	2,00	165	0,46	530	0,84	65	0,95	15	0,23	0	0,00	0	0,00	10	0,62	0	0,00	0	0,00	1,4
2513	1092	2,31	1740	2,47	2376	4,63	1032	2,85	1884	2,98	180	2,62	180	2,78	36	0,65	12	2,03	60	3,72	12	2,82	0	0,00	3,6
2514	400	0,85	1400	1,99	1740	3,39	510	1,41	1800	2,84	140	2,04	40	0,62	0	0,00	0	0,00	30	1,86	0	0,00	0	0,00	2,9
2515	81	0,17	1053	1,49	1323	2,58	702	1,94	1251	1,98	117	1,70	54	0,83	0	0,00	0	0,00	9	0,56	0	0,00	0	0,00	0,0
2516	2676	5,66	3828	5,43	2844	5,54	1860	5,13	3744	5,91	420	6,11	372	5,74	144	2,59	12	2,03	48	2,98	12	2,82	0	0,00	1,4
2522	3822	8,09	2860	4,06	2704	5,27	2080	5,74	3575	5,65	325	4,73	689	10,63	325	5,86	0	0,00	65	4,03	26	6,12	0	0,00	7,1
2524	3456	7,32	2760	3,92	1008	1,96	960	2,65	1788	2,82	228	3,32	216	3,33	228	4,11	0	0,00	48	2,98	0	0,00	0	0,00	7,1
2537	288	0,61	1680	2,38	1112	2,17	1136	3,13	1192	1,88	128	1,86	152	2,34	152	2,74	0	0,00	32	1,98	8	1,88	8	100,00	0,4
2538	2665	5,64	4524	6,42	2392	4,66	2197	6,06	3926	6,20	598	8,70	468	7,22	351	6,32	13	2,20	78	4,84	13	3,06	0	0,00	7,1
2539	696	1,47	3132	4,45	1848	3,60	1236	3,41	2388	3,77	228	3,32	144	2,22	144	2,59	0	0,00	24	1,49	0	0,00	0	0,00	2,5
2540	1200	2,54	3828	5,43	1524	2,97	2088	5,76	3456	5,46	396	5,76	636	9,81	552	9,95	312	52,79	396	24,55	180	42,35	0	0,00	1,8
2541	1313	2,78	2223	3,16	1222	2,38	1261	3,48	2431	3,84	273	3,97	52	0,80	0	0,00	0	0,00	13	0,81	13	3,06	0	0,00	5,4
2542	4550	9,63	4667	6,62	2613	5,09	1625	4,48	4368	6,90	390	5,68	351	5,41	468	8,43	26	4,40	52	3,22	0	0,00	0	0,00	7,1
2543	8092	17,13	5180	7,35	1876	3,65	2464	6,79	2702	4,27	224	3,26	378	5,83	406	7,32	0	0,00	70	4,34	42	9,88	0	0,00	7,1
2544	3444	7,29	4214	5,98	2380	4,63	1442	3,98	2352	3,72	266	3,87	280	4,32	224	4,04	0	0,00	70	4,34	0	0,00	0	0,00	7,1
2545	1482	3,14	3913	5,55	1833	3,57	2288	6,31	2678	4,23	286	4,16	416	6,42	325	5,86	0	0,00	78	4,84	26	6,12	0	0,00	3,6
2547	2145	4,54	3289	4,67	2457	4,78	1482	4,09	3471	5,48	273	3,97	208	3,21	143	2,58	0	0,00	26	1,61	0	0,00	0	0,00	3,6
2550	1404	2,97	2327	3,30	2080	4,05	1872	5,16	2717	4,29	234	3,41	247	3,81	273	4,92	117	19,80	143	8,87	0	0,00	0	0,00	1,8
2551	50	0,11	320	0,45	825	1,61	315	0,87	340	0,54	45	0,66	35	0,54	15	0,27	0	0,00	10	0,62	0	0,00	0	0,00	0,0
2562	930	1,97	1440	2,04	2325	4,53	675	1,86	2250	3,55	135	1,97	105	1,62	0	0,00	0	0,00	15	0,93	0	0,00	0	0,00	3,6
2564	330	0,70	1330	1,89	1930	3,76	890	2,45	2040	3,22	130	1,89	150	2,31	100	1,80	0	0,00	40	2,48	0	0,00	0	0,00	0,0
2565	2556	5,41	2580	3,66	1992	3,88	1584	4,37	2808	4,44	564	8,21	288	4,44	444	8,00	12	2,03	48	2,98	12	2,82	0	0,00	7,1
2573	105	0,22	1169	1,66	1022	1,99	546	1,51	1127	1,78	63	0,92	91	1,40	0	0,00	0	0,00	7	0,43	0	0,00	0	0,00	0,4
2634	1727	3,66	4059	5,76	1870	3,64	1452	4,00	2332	3,68	308	4,48	275	4,24	341	6,14	0	0,00	44	2,73	22	5,18	0	0,00	5,4
2635	570	1,21	1370	1,94	3380	6,58	1090	3,01	1700	2,69	200	2,91	110	1,70	0	0,00	10	1,69	40	2,48	10	2,35	0	0,00	3,2
2648	36	0,08	624	0,89	456	0,89	642	1,77	594	0,94	90	1,31	48	0,74	78	1,41	0	0,00	18	1,12	0	0,00	0	0,00	0,0
2693	522	1,11	1053	1,49	675	1,31	855	2,36	1026	1,62	117	1,70	126	1,94	126	2,27	36	6,09	54	3,35	27	6,35	0	0,00	1,4
2707	456	0,97	792	1,12	384	0,75	328	0,90	528	0,83	128	1,86	126	1,11	576	10,38	8	1,35	8	0,50	0	0,00	0	0,00	2,1
	47238	100	70454	100	51350	100	36262	100	63308	100	6870	100	6484	100	5550	100	591	100	1613	100	425	100	8	100	

Fuente: Evirtual Epoch

3.7.1.3 *Determinar si el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje de la cátedra de Aplicaciones Web ayudan a que los estudiantes sea exitosos.*

Como se ha mencionado en este trabajo de investigación, se utiliza un modelo de minería de datos basado en los datos históricos de la cátedra de Aplicaciones web de la Escuela de Ingeniería en Sistemas. Este modelo está relacionado con el peso de calificación de los distintos atributos que ha utilizado, en la tabla 10 y tabla 11, se describió los atributos de Assign (tareas) y Quiz (cuestionarios), mientras que el puntaje de participación según el Silabo de la materia tiene un correspondencia de 2 puntos o lo que es lo mismo el 7,1% de la nota total. Con los datos obtenidos del EVA tenemos la siguiente tabla. (El color verde representa los aprobados)

Estudiantes	Participación
2509	5,4
2511	1,4
2513	3,6
2514	2,9
2515	0,0
2516	1,4
2522	7,1
2524	7,1
2537	0,4
2538	7,1
2539	2,5
2540	1,8
2541	5,4
2542	7,1
2543	7,1
2544	7,1
2545	3,6
2547	3,6
2550	1,8
2551	0,0
2562	3,6
2564	0,0
2565	7,1
2573	0,4
2634	5,4
2635	3,2
2648	0,0
2693	1,4
2707	2,1

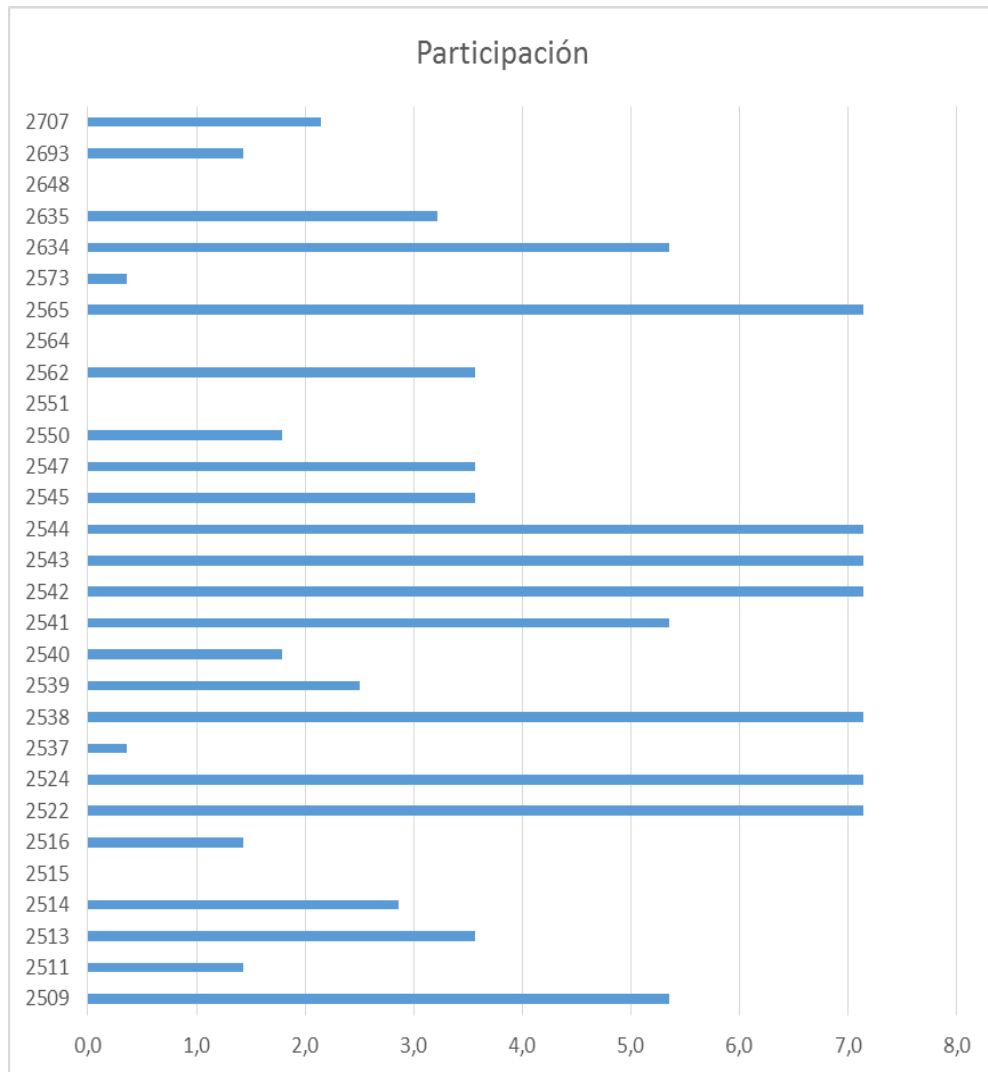


Figura 6-3: Participación por individuo.

Fuente: Evirtual Epoch

Indicadores.

De los 29 estudiantes de la materia de Aplicaciones Web, 7 tiene una calificación igual a 7,1%, que representa el 24,1% del total, de este porcentaje todo lograron aprobar el curso. 3 estudiantes (10,4%), encuentran en el rango entre [4;7,1], de este porcentaje todo lograron aprobar el curso. 8 estudiantes (27,6%), encuentran en el rango entre [2;4], de este porcentaje apenas el 17,2% aprobaron el curso, el 10,4% restante reprobaron. 11 estudiantes (37,9%), encuentran en el rango entre [0;2], de este porcentaje apenas el 13,8% aprobaron el curso, el 24,1% restante reprobaron. El 65,5% de los estudiantes que utilizaron el EVA aprobaron el curso.

3.8 Variables e indicadores

De acuerdo a la hipótesis planteada, fueron definidas las siguientes variables:

Operacionalización de la variable dependiente:

Variable Independiente: Patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje.

Operacionalización conceptual

Tabla 14-3: Variables dependientes.

Variable	Indicadores	Tipo	Concepto
V ₀ : Patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje.	I ₁ : # de participaciones por estudiantes I ₂ : % de calificaciones de los estudiantes I ₃ : # de actividades o módulos. I ₄ : # de participación en las actividades o módulos. I ₅ : % de participación en las actividades o módulos.	Variable independiente Variable cuantitativa	Conjunto estructurado de rastros que sirven para obtener evidencias o respuestas sobre el objeto a evaluar.

Realizado por: Gustavo Hidalgo.2016

Operacionalización de la variable independiente:

Variable Dependiente: Estudiantes exitosos

Operacionalización conceptual

Tabla 15-3: Variables independientes.

Variable	Indicadores	Tipo	Concepto
V ₁ : Éxito de los estudiantes	I ₆ : % de estudiantes que logran terminar el semestre con éxito. I ₇ : % de Falsos Positivos del algoritmo seleccionado. I ₈ : % de Verdaderos Positivos del modelo seleccionado. I ₉ : % de aprendizaje del modelo.	Variable dependiente	Describe lo que previsiblemente los estudiantes deberán saber, comprender y ser capaces de hacer al finalizar con éxito una asignatura.

Realizado por: Gustavo Hidalgo.2016

3.9 Instrumentos

Para la variable independiente “Patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje”, se procederá con la técnica de observación directa de los datos contenidos en las bases de datos de Plataforma virtual de Educación de la Epoch, correspondiente a los periodos de Septiembre 2014 – Febrero 2015 y la aplicación del modelo en base al algoritmo seleccionado (Redes neuronales Perceptron) con individuos nuevos mismos que saldrán de la materia de Didáctica informática del periodo actual.

La aplicación de este método nos revelara los patrones con sus respectivas calificaciones que pueden ser cuantitativas o cualitativa dependiendo del patrón, estos resultados contrastados con las calificaciones generales del curso, que fueron ingresadas en el Sistema Académico en los periodos antes mencionados no mostrará la información necesaria para resolver la hipótesis.

Esta investigación utiliza el método de recolección de datos y como instrumento la observación de datos en base a la información que genera la técnica de minería aplicada para descubrir los patrones y su comparación con los datos de las calificaciones generales de la cátedra de Aplicaciones web en el sistema académico, además de la entrevista al docente principal de la cátedra.

CAPITULO IV

4. RESULTADOS Y DISCUSIÓN

4.1 Análisis e interpretación de resultados

En el presente trabajo de investigación, se pretende demostrar la influencia que existe entre los patrones de participación y el éxito de los estudiantes en el semestre. Por lo que, fueron definidas dos variables para el análisis: los patrones de participación en un Entorno Virtual de Aprendizaje, y el éxito de los estudiantes.

4.1.1 *Indicadores de la variable dependiente*

La variable patrones de participación en los Entornos Virtuales de Aprendizaje, es medida mediante la extracción y transformación de datos a un Data Warehouse que contiene dos Data Mart, *dwh_calificaciones* y *dwh_participación*; además de la utilización de herramientas estadísticas para la detección de los mismos.

El *dwh_calificaciones* nos permite tener una tabla limpia de datos para las calificaciones de los módulos que el docente configuro en el EVA. Mientras que el *dwh_participación* nos permite tener las participaciones de los estudiantes en cada uno de los módulos que ofrece el EVA.

4.1.1.1 *Número de participaciones por estudiantes*

En la **Tabla 1-4** se muestra la cantidad total de interacciones o conteo de participaciones que el estudiante ha tenido durante todo el semestre en los distintos módulos o actividades del EVA.

Tabla 1-4: Participación por individuo

Estudiantes	Número Participación
2509	10868
2511	2190
2513	8748
2514	6160
2515	4644
2516	16260
2522	16666
2524	10740
2537	5992
2538	17745
2539	9912
2540	15804
2541	8827
2542	19526
2543	21686
2544	14728
2545	13481
2547	13741
2550	11856
2551	1970
2562	7905
2564	6960
2565	13176
2573	4179
2634	12815
2635	8520
2648	2622
2693	4806
2707	3384

.Fuente: Evirtual Epoch

La **Figura 16-4**, muestra la participación total de los estudiantes en las diferentes actividades y recursos que están vigentes en el EVA de la materia de Aplicaciones Web. El 75% de los estudiantes sobrepasan las 6000 participaciones en los diferentes módulos y actividades evidenciando que prestan mucho interés a este tipo de herramientas.



Figura 1-4: Participación por individuos.
Fuente: Evirtual Epoch

4.1.1.2 Porcentaje de calificaciones de los estudiantes

La **Tabla 17-4**, muestra el porcentaje de calificaciones de los estudiantes por actividad según la planificación del docente en el EVA y en el sílabo de la materia.

Tabla 2-4: Porcentaje de calificaciones.

Estudiantes	Assign	Quiz	Participación	Total
2509	13,3	22,6	5,4	41,3
2511	11,9	19,9	1,4	33,2
2513	10,9	19,4	3,6	33,9
2514	13,7	20,8	2,9	37,3
2515	7,7	16,3	0,0	24,0
2516	13,7	26,2	1,4	41,3
2522	12,5	27,9	7,1	47,6
2524	13,2	17,6	7,1	38,0
2537	11,6	16,2	0,4	28,2
2538	11,5	19,3	7,1	37,9
2539	11,9	16,8	2,5	31,2
2540	12,6	23,7	1,8	38,1
2541	11,3	20,3	5,4	36,9
2542	12,8	20,8	7,1	40,7
2543	13,7	28,7	7,1	49,5
2544	12,6	21,2	7,1	41,0
2545	13,2	22,1	3,6	39,0
2547	11,2	20,1	3,6	34,8
2550	14,4	19,8	1,8	36,0
2551	4,1	19,8	0,0	23,8
2562	10,8	16,6	3,6	31,0
2564	12,4	20,3	0,0	32,6
2565	13,7	20,7	7,1	41,5
2573	12,0	13,2	0,4	25,6
2634	13,1	26,6	5,4	45,0
2635	10,1	19,7	3,2	33,0
2648	10,8	21,5	0,0	32,3
2693	7,1	10,2	1,4	18,7
2707	3,6	18,1	2,1	23,8

Fuente: Evirtual Espoch

Este indicador nos muestra el 100% de los estudiantes participan en este tipo de actividad, esto se debe a que es una actividad obligatoria y de más peso en la evaluación sumativa.

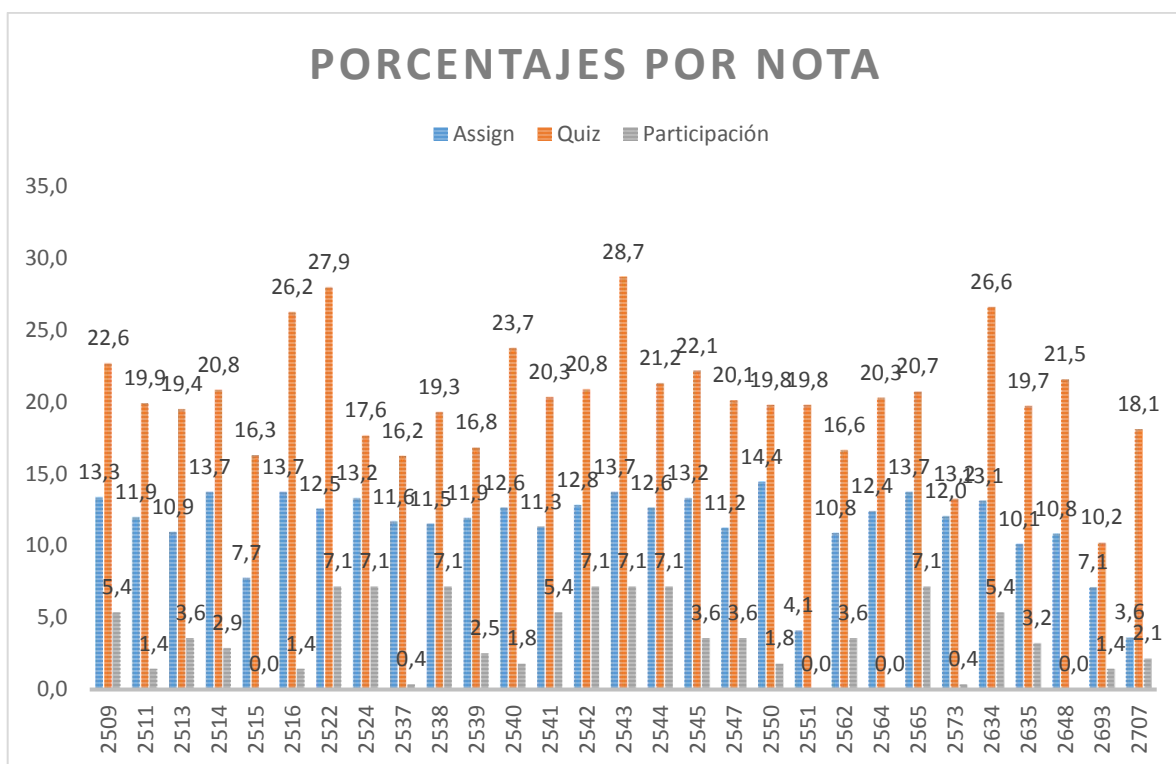


Figura 2-4: Porcentaje de calificaciones.

Fuente: Evirtual Espoch.

4.1.1.3 Número de actividades o módulos.

El número de actividades o módulos configurados son 13 en el ambiente de trabajo del EVA de cátedra de Aplicaciones Web, de los cuales los más representativos son 3: quiz, assign y log.

4.1.1.4 Número de participación en las actividades o módulos.

Los estudiantes de la cátedra de Aplicaciones Web han tenido un total de 295911 participaciones en todas las actividades o módulos configurados, mismos que se detallan a continuación.

Tabla 3-4: Participaciones en módulos.

Modulo	Total
course	70454
quiz	51350
resource	36262
forum	47238
choice	6870
user	2675
page	6484
assign	63308
glossary	5564
url	1887
book	591
folder	1613
imscp	591
label	591
wiki	425
blog	8

Fuente: Evirtual Epoch.

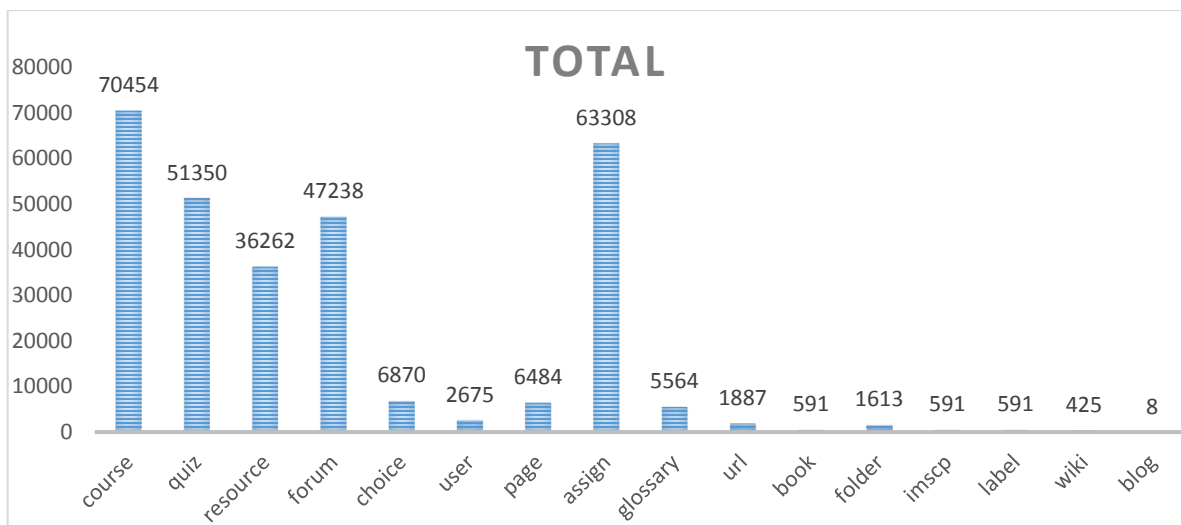


Figura 3-4: Participaciones en módulos

Fuente: Evirtual Epoch

4.1.1.5 Porcentaje de participación en las actividades o módulos.

El 65,5% de los estudiantes que interaccionaron con los módulos o actividades del EVA de aplicaciones web aprobaron el semestre. Este indicador es clave al momento de tomar decisiones por parte del docente.

Tabla 4-4: Consolidado de participación.

Estudiantes	Forum	course	quiz	Resource	assign	choice	page	glossary	book	folder	wiki	blog	Participación												
2509	1045	2,21	2849	4,04	2134	4,16	1485	4,10	2310	3,65	319	4,64	286	4,41	99	1,78	33	5,58	77	4,77	22	5,18	0	0,00	5,4
2511	105	0,22	250	0,35	1025	2,00	165	0,46	530	0,84	65	0,95	15	0,23	0	0,00	0	0,00	10	0,62	0	0,00	0	0,00	1,4
2513	1092	2,31	1740	2,47	2376	4,63	1032	2,85	1884	2,98	180	2,62	180	2,78	36	0,65	12	2,03	60	3,72	12	2,82	0	0,00	3,6
2514	400	0,85	1400	1,99	1740	3,39	510	1,41	1800	2,84	140	2,04	40	0,62	0	0,00	0	0,00	30	1,86	0	0,00	0	0,00	2,9
2515	81	0,17	1053	1,49	1323	2,58	702	1,94	1251	1,98	117	1,70	54	0,83	0	0,00	0	0,00	9	0,56	0	0,00	0	0,00	0,0
2516	2676	5,66	3828	5,43	2844	5,54	1860	5,13	3744	5,91	420	6,11	372	5,74	144	2,59	12	2,03	48	2,98	12	2,82	0	0,00	1,4
2522	3822	8,09	2860	4,06	2704	5,27	2080	5,74	3575	5,65	325	4,73	689	10,63	325	5,86	0	0,00	65	4,03	26	6,12	0	0,00	7,1
2524	3456	7,32	2760	3,92	1008	1,96	960	2,65	1788	2,82	228	3,32	216	3,33	228	4,11	0	0,00	48	2,98	0	0,00	0	0,00	7,1
2537	288	0,61	1680	2,38	1112	2,17	1136	3,13	1192	1,88	128	1,86	152	2,34	152	2,74	0	0,00	32	1,98	8	1,88	8	100,00	0,4
2538	2665	5,64	4524	6,42	2392	4,66	2197	6,06	3926	6,20	598	8,70	468	7,22	351	6,32	13	2,20	78	4,84	13	3,06	0	0,00	7,1
2539	696	1,47	3132	4,45	1848	3,60	1236	3,41	2388	3,77	228	3,32	144	2,22	144	2,59	0	0,00	24	1,49	0	0,00	0	0,00	2,5
2540	1200	2,54	3828	5,43	1524	2,97	2088	5,76	3456	5,46	396	5,76	636	9,81	552	9,95	312	52,79	396	24,55	180	42,35	0	0,00	1,8
2541	1313	2,78	2223	3,16	1222	2,38	1261	3,48	2431	3,84	273	3,97	52	0,80	0	0,00	0	0,00	13	0,81	13	3,06	0	0,00	5,4
2542	4550	9,63	4667	6,62	2613	5,09	1625	4,48	4368	6,90	390	5,68	351	5,41	468	8,43	26	4,40	52	3,22	0	0,00	0	0,00	7,1
2543	8092	17,13	5180	7,35	1876	3,65	2464	6,79	2702	4,27	224	3,26	378	5,83	406	7,32	0	0,00	70	4,34	42	9,88	0	0,00	7,1
2544	3444	7,29	4214	5,98	2380	4,63	1442	3,98	2352	3,72	266	3,87	280	4,32	224	4,04	0	0,00	70	4,34	0	0,00	0	0,00	7,1
2545	1482	3,14	3913	5,55	1833	3,57	2288	6,31	2678	4,23	286	4,16	416	6,42	325	5,86	0	0,00	78	4,84	26	6,12	0	0,00	3,6
2547	2145	4,54	3289	4,67	2457	4,78	1482	4,09	3471	5,48	273	3,97	208	3,21	143	2,58	0	0,00	26	1,61	0	0,00	0	0,00	3,6
2550	1404	2,97	2327	3,30	2080	4,05	1872	5,16	2717	4,29	234	3,41	247	3,81	273	4,92	117	19,80	143	8,87	0	0,00	0	0,00	1,8
2551	50	0,11	320	0,45	825	1,61	315	0,87	340	0,54	45	0,66	35	0,54	15	0,27	0	0,00	10	0,62	0	0,00	0	0,00	0,0
2562	930	1,97	1440	2,04	2325	4,53	675	1,86	2250	3,55	135	1,97	105	1,62	0	0,00	0	0,00	15	0,93	0	0,00	0	0,00	3,6
2564	330	0,70	1330	1,89	1930	3,76	890	2,45	2040	3,22	130	1,89	150	2,31	100	1,80	0	0,00	40	2,48	0	0,00	0	0,00	0,0
2565	2556	5,41	2580	3,66	1992	3,88	1584	4,37	2808	4,44	564	8,21	288	4,44	444	8,00	12	2,03	48	2,98	12	2,82	0	0,00	7,1
2573	105	0,22	1169	1,66	1022	1,99	546	1,51	1127	1,78	63	0,92	91	1,40	0	0,00	0	0,00	7	0,43	0	0,00	0	0,00	0,4
2634	1727	3,66	4059	5,76	1870	3,64	1452	4,00	2332	3,68	308	4,48	275	4,24	341	6,14	0	0,00	44	2,73	22	5,18	0	0,00	5,4
2635	570	1,21	1370	1,94	3380	6,58	1090	3,01	1700	2,69	200	2,91	110	1,70	0	0,00	10	1,69	40	2,48	10	2,35	0	0,00	3,2
2648	36	0,08	624	0,89	456	0,89	642	1,77	594	0,94	90	1,31	48	0,74	78	1,41	0	0,00	18	1,12	0	0,00	0	0,00	0,0
2693	522	1,11	1053	1,49	675	1,31	855	2,36	1026	1,62	117	1,70	126	1,94	126	2,27	36	6,09	54	3,35	27	6,35	0	0,00	1,4
2707	456	0,97	792	1,12	384	0,75	328	0,90	528	0,83	128	1,86	72	1,11	576	10,38	8	1,35	8	0,50	0	0,00	0	0,00	2,1
	47238	100	70454	100	51350	100	36262	100	63308	100	6870	100	6484	100	5550	100	591	100	1613	100	425	100	8	100	

Fuente: Evirtual Espoch

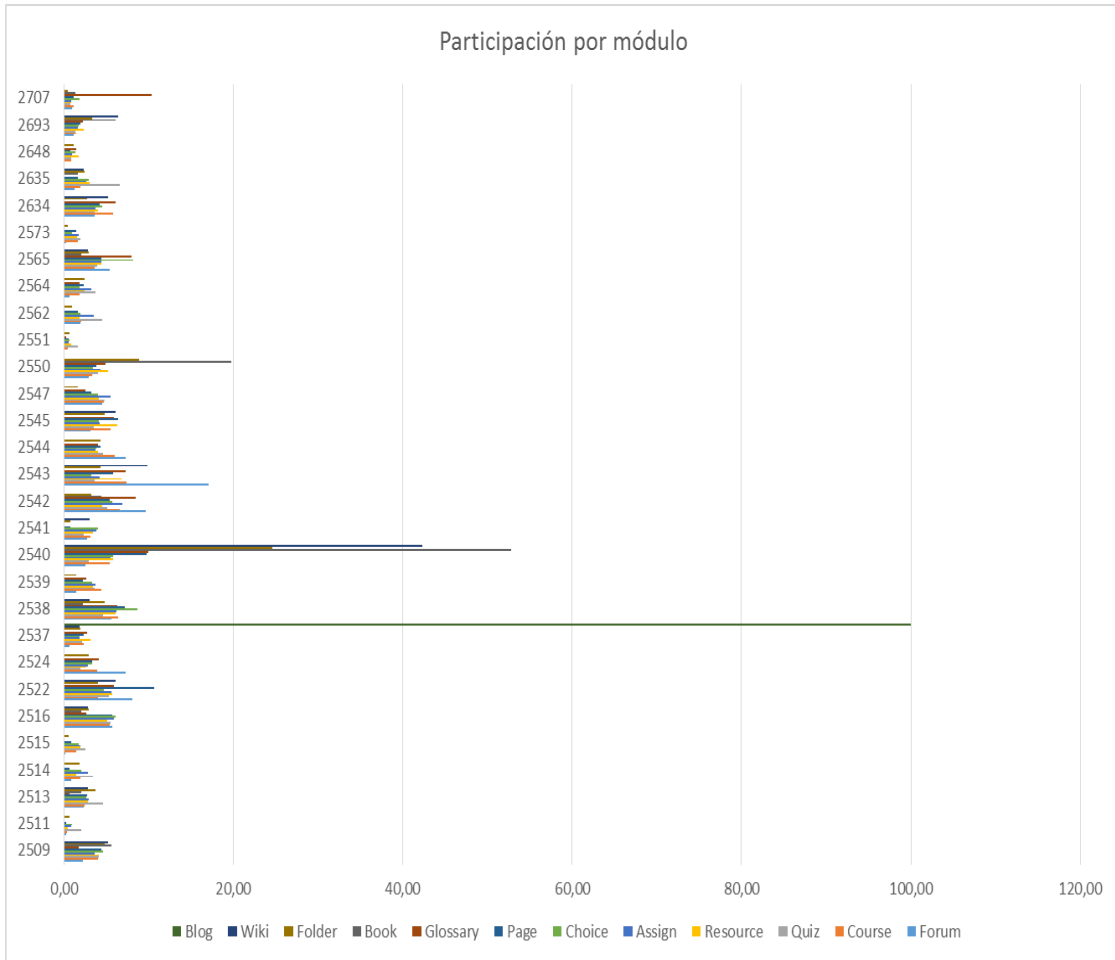


Figura 4-4: Participación por módulos.
Fuente: Evirtual Espoch

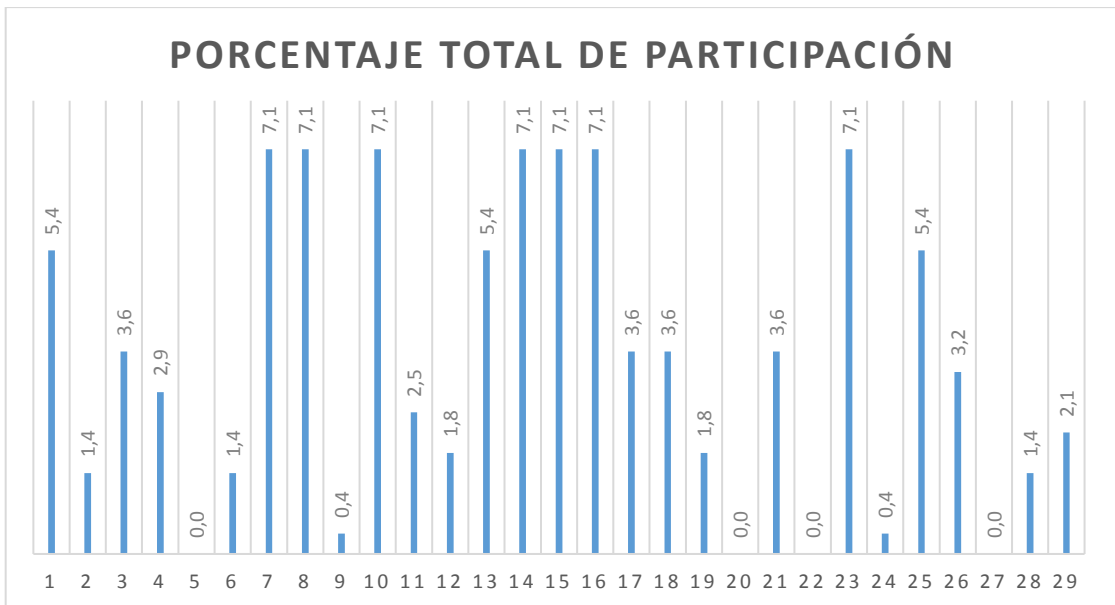


Figura 5-4: Total de participación.
Fuente: Evirtual Espoch

4.1.2 Indicadores de la variable independiente.

La variable estudiantes exitosos, es el resultado de calcular notas acumulativas (80%), el examen final (20%) y suspensión. Las notas acumulativas están enfocadas en las primeras tres notas, mismas que se las construyen a partir de tareas, consultas, exámenes parciales, lecciones, entre otras.

En este trabajo de investigación nos enfocamos a las notas acumulativas ya que es aquí donde se aprovecha el EVA, de esta manera del 80% del valor total de las notas acumulativas, se planifica en el EVA el 60% al 75% de la misma. En el caso particular del estudio el porcentaje es de 60,7%.

4.1.2.1 Porcentaje de estudiantes que logran terminar el semestre con éxito.

De 29 estudiantes que tomaron el curso el 66% de los estudiantes aprobaron la cátedra. De esta manera se puede identificar el 66% de estudiantes de participaron en los módulos del EVA aprobaron el semestre; este indicador se convierte en una herramienta importante al momento de realizar un pre-análisis de la influencia positiva que nos da los EVAs en el éxito de los estudiantes.

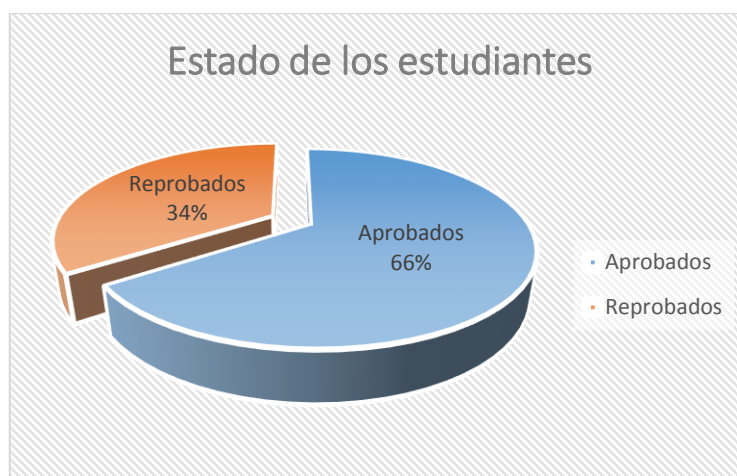


Figura 6-4: Estado de estudiantes.
Fuente: OASIs

4.1.2.2 Porcentaje de Falsos Positivos del algoritmo seleccionado.

Es la capacidad del modelo para predecir con un bajo nivel de exactitud o acierto en el éxito de los estudiantes, cuando se ingresa un grupo de nuevos individuos los cuales aún no se conoce su estado (Aprobado o reprobado). El actual trabajo luego de una serie de pruebas para conocer cuál de los algoritmos predictivos resuelve de mejor manera el problema planteado, se escogió el algoritmo de

Redes Neuronales Multicapa Perceptron mismo que tiene un porcentaje de Verdaderos positivos (TP) del 14,9%, esto es con un cross-validation del 10 folds.

4.1.2.3 Porcentaje de Verdaderos Positivos del modelo seleccionado.

Es la capacidad del modelo para predecir con mayor exactitud el éxito de los estudiantes, cuando se ingresa un grupo de nuevos individuos los cuales aún no se conoce su estado (Aprobado o reprobado). El actual trabajo luego de una serie de pruebas para conocer cuál de los algoritmos predictivos resuelve de mejor manera el problema planteado, se escogió el algoritmo de Redes Neuronales Multicapa Perceptron mismo que tiene un porcentaje de Verdaderos positivos (TP) del 89,7%, esto es con un cross-validation del 10 folds.

4.1.2.4 Porcentaje de aprendizaje del modelo.

El modelo según los datos de ensayo o experimento ofrece un 89,7 % de aprendizaje, este dato son importante ya que permite saber que tan exacto será la predicción de nuevos individuos.

```
=== Detailed Accuracy By Class ===
              TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
              0.947    0.2      0.9         0.947   0.923      0.947    si
              0.8      0.053   0.889      0.8     0.842      0.947    no
Weighted Avg.  0.897    0.149   0.896      0.897   0.895      0.947

=== Confusion Matrix ===
  a  b  <-- classified as
18  1  |  a = si
 2  8  |  b = no
```

La matriz de confusión es el resultado que brinda los algoritmos de clasificación para evaluar la predicción que realiza, de esta manera nos permite tener un medio más para la toma de decisiones.

Cuando la matriz de confusión muestra su diagonal de derecha a izquierda mayor a su diagonal inversa esto es una señal que el modelo es bueno en este caso la diagonal 18 – 8 es mayor a la diagonal 2 – 1, ya que $18 > 2$ y $8 > 1$.

Esto quiere decir que de 20 datos que indican “si” el modelo fallo o no acertó solamente en 2; de la misma manera de 9 datos que indican “no” solo fallo en 1.

Este modelo fue creado con los datos de prueba de la materia de Aplicaciones Web misma que tenía 20 aprobados y 9 reprobados. Un datos que nos indica también que el modelo va a predecir bien los datos es el Precision este nos muestra un 0.896 es decir un 89,6%, un alto margen de aceptabilidad.

De la misma forma los Verdaderos positivos, es de 89,7% es un alto margen de verdad o predicción acertadas sin equivocación, reduciendo los Falsos positivos a 14.9% un margen aceptable de error. Por último el patrón Recall o también conocido como porcentaje de aprendizaje, es del 89,7%, indicando que el modelo ha aprendido en gran medida de los datos ingresados.

4.2 Presentación de Resultados

Como ya se mencionó anteriormente el algoritmo que se escogió para este trabajo de investigación es el de Redes Neuronales Perceptron multicapa, mismo que según la curva ROC y el área bajo la curva ROC, es el que mejor resuelve la predicción con los datos que se plantea. Lugo de someter los datos a pruebas con el algoritmo del perceptron multicapa desde el paquete estadístico de minería de datos Weka, este arroja la siguiente información.

```
==== Run information ====
```

```
Scheme: weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S  
0 -E 20 -H a
```

```
Relation: estudiante_exitoso-weka.filters.unsupervised.attribute.Remove-R1
```

```
Instances: 29
```

```
Attributes: 4
```

```
nota_assign
```

```
nota_quiz
```

```
nota_participacion
```

```
estado
```

```
Test mode: evaluate on training data
```

```
==== Classifier model (full training set) ====
```

```
Sigmoid Node 0
```

```
Inputs Weights
```

```
Threshold -2.7213033150171233
```

```
Node 2 3.2882257644919406
```

```
Node 3 3.3557452613142504
```

```
Sigmoid Node 1
```

Inputs Weights
Threshold 2.7188129300582182
Node 2 -3.281701109047159
Node 3 -3.3647568390195617

Sigmoid Node 2

Inputs Weights
Threshold 1.8583864378860313
Attrib nota_assign 1.7471176767530547
Attrib nota_quiz 1.4437534365520666
Attrib nota_participacion 3.6951765388367335

Sigmoid Node 3

Inputs Weights
Threshold -2.5554184690013795
Attrib nota_assign 3.145466091177445
Attrib nota_quiz 1.0698793718416386
Attrib nota_participacion 1.522010636192594

Class si

Input
Node 0

Class no

Input
Node 1

Time taken to build model: 0.03 seconds

==== Evaluation on training set ====

==== Summary ====

Correctly Classified Instances	26	89.6552 %
Incorrectly Classified Instances	3	10.3448 %
Kappa statistic	0.7655	
Mean absolute error	0.2053	
Root mean squared error	0.3058	
Relative absolute error	45.1264 %	
Root relative squared error	64.3291 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	86.2069 %	

Total Number of Instances 29

==== Detailed Accuracy By Class ====

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.947	0.2	0.9	0.947	0.923	0.947	si
	0.8	0.053	0.889	0.8	0.842	0.947	no
Weighted Avg.	0.897	0.149	0.896	0.897	0.895	0.947	

==== Confusion Matrix ====

a b <-- classified as

18 1 | a = si

2 8 | b = no

4.2.1 Análisis de la trama

La red neuronal nos indica que el algoritmo perceptron multicapa al momento de crear el modelo para poder hacer la predicción, ha recurrido a la creación de 2 niveles ocultos con 2 nodos cada uno, mismo que tiene 4 entradas (atributos o variables predictores) y 2 variables a predecir o variables discriminantes.

Además nos está indicando que utilizo una función semoidal [0,1]. El siguiente gráfico y la tabla describen el comportamiento de la red neuronal.

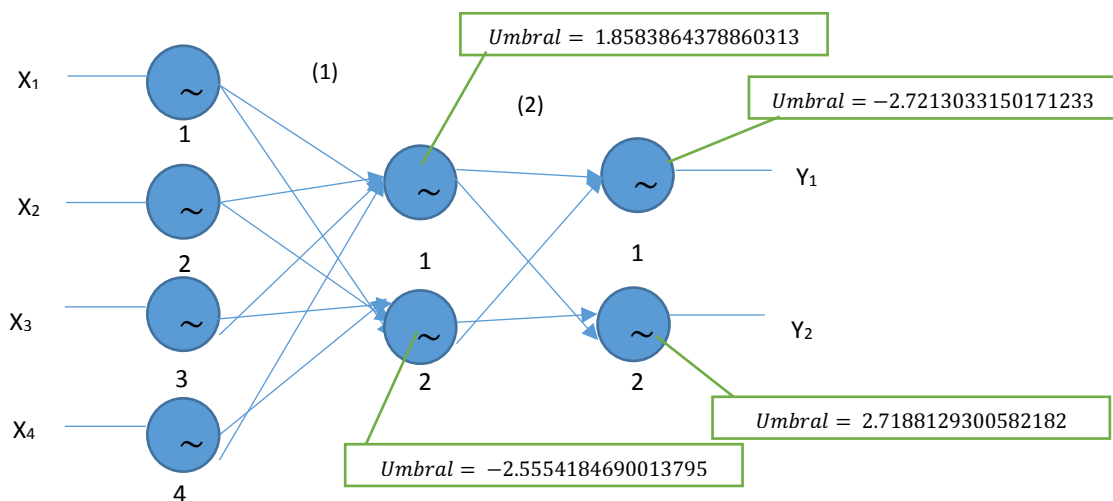


Tabla 5-4: Entradas y salidas de la red neuronal.

Peso (1)	Peso (2)	Salida
$W_{1,1}^{(1)}$ = 1.7471176767530547	$W_{1,1}^{(2)}$ = 3.2882257644919406 $W_{1,2}^{(2)}$ = -3.281701109047159	$Y_1 = 1$
$W_{1,2}^{(1)}$ = 3.145466091177445		
$W_{2,1}^{(1)}$ = 1.4437534365520666		
$W_{2,2}^{(1)}$ = 1.0698793718416386	$W_{2,1}^{(2)}$ = 3.3557452613142504 $W_{2,2}^{(2)}$ = -3.3647568390195617	$Y_2 = 1$
$W_{3,1}^{(1)}$ = 3.6951765388367335		
$W_{3,2}^{(1)}$ = 1.522010636192594		

Realizado por: Gustavo Hidalgo.2016.

De esta manera la fórmula que se utiliza en el presente trabajo y con la ayuda del modelo antes mencionado es:

$$a_i = u_i^{(k)} + \left[\sum_{j=1}^{n^{(k-1)}} z_j^{(k-1)} w_{ji}^{(k-1)} \right]$$

a = Salida del perceptron multicapa [0,1]

u = Umbral

z = Datos de entrada al perceptron.

w = Peso del sigmoide.

De esta manera si a_j es $>$ al umbral (u) entonces el valor que viene de z_j va a tener más influencia en la neurona, y es por donde seguirá el pulso eléctrico. La red neuronal reacciona mediante impulsos eléctricos, dependiendo del peso que tenga cada sigmoide, es decir; si la neurona tiene 3 entrada y el peso de cada entrada es 1, 2 ,3; gracias a la formula antes mencionada, obedecerá a la entrada 3 ya que su peso es mayor.

De esta manera va encaminando los pulsos eléctricos, hasta llegar a la salida en una de sus dos clases “sí o no”.

4.3 Predicción

Luego de detectar los patrones que influyen en que un estudiante sea exitoso, gracias al modelo que nos proporcionó el algoritmo del perceptron multicapa, el siguiente paso es someter el modelo encontrado cuyo porcentaje de aprendizaje fue de 89,7 %; a la predicción de los individuos nuevos.

Es importantes establecer que lo individuos nuevos son los datos de estudiantes que están cruzando actualmente el semestre y cuyos valores aprobación y reprobación no se conocen aún; De esta manera se ha escogido a los estudiantes de la cátedra de Didáctica informática del noveno semestre de la escuela de Ingeniería en Sistemas Informáticos de la Escuela de Sistemas de la facultad de Informática y Electrónica.

Se escogió a estos individuos nuevos por iniciativa del investigador basado en las siguientes condiciones.

Regla 1 (R1).- La cátedra en mención debe utilizar la plataforma virtual de aprendizaje de la ESPOCH.

Regla 2 (R2).- El EVA de esta catedra de contener actividades como assign, quiz y log de participación en forum, chat, course, resource, choice, entre otros.

Regla 3 (R3).- El porcentaje de calificación que se toma en cuenta desde el EVA para la nota final debe ser mayor o igual al 60,7% del 80% de las notas acumulativas.

Regla 4 (R4).- Debe pertenecer a la Escuela de sistemas.

Regla 5 (R5).- El número de estudiantes que toma la materia tiene que ser mayor o igual que 29.

Tomado en cuenta estas reglas se tiene el siguiente análisis.

Tabla 6-4: Estudio de EVAs en periodo actual.

Cátedra	R1	R2	R3	R4	R5
Interfaz y multimedia	Cumple	No Cumple	No Cumple	Cumple	Cumple
Proyecto de tesis	Cumple	No Cumple	No Cumple	Cumple	Cumple
Arquitectura de computadores	Cumple	No Cumple	No Cumple	Cumple	Cumple
Didáctica Informática	Cumple	Cumple	Cumple	Cumple	Cumple
Redes de Computadores	Cumple	No Cumple	No Cumple	Cumple	Cumple
Sistemas de información geográfica	Cumple	No Cumple	No Cumple	Cumple	Cumple

Realizado por: Gustavo Hidalgo.2016

De esta manera la única cátedra que cumple con todas las condiciones antes mencionada es Didáctica informática.

4.3.1 Presentación de individuos nuevos.

El apartado anterior permite identificar la cátedra con la que se aplicará el modelo predictivo creado anteriormente. De esta manera los nuevos individuos son los estudiantes de Didáctica Informática del noveno semestre de la Escuela de Ingeniería en Sistemas de la Facultad de Informática y Electrónica.

El archivo de individuos nuevos para la predicción sería el siguiente:

```

@relation QueryResult-weka.filters.unsupervised.attribute.Remove-R1
@attribute codigo
{4262,4320,4349,4722,4917,4924,4977,5125,5138,5172,5199,5208,5228,5295,5339,5449,5454,5476,5480,548
4,5489,5506,5546,5548,5565,5569,5580,5634}
@attribute assign numeric
@attribute quiz numeric
@attribute participación numeric
@attribute estado {si,no}
@data
4262,18.428572,24.285715,7.2,?
4320,19.428572,30.464285,3.6,?
4349,13.571428,27.607143,7.2,?
4722,17.464285,26.178572,3.6,?
4917,16.214285,37.607143,4.6,?
4924,10.678572,22.857143,3.6,?
4977,13.892858,27.857143,7.2,?
5125,16.321428,35.25,4.3,?
5138,17.5,28.107143,4.3,?
5172,16.142857,26.678572,7.2,?
5199,17.25,27.857143,5.2,?
5208,18.714285,35.6,4,?
5228,17.5,29.035715,5.2,?
5295,16.071428,30.714285,7.2,?
5339,17.714285,34.75,7.2,?
5449,18.142857,29.535715,6.4,?
5454,5.464286,14.285714,5.2,?
5476,18.607143,31.321428,6.4,?
5480,18.321428,28.464285,5.4,?
5484,18.428572,35.607143,5.4,?
5489,18.178572,29.535715,7.2,?
5506,15.571428,14.285714,5.2,?
5546,19.571428,30.607143,5.2,?
5548,19.392857,36.32143,7.2,?
5565,18.214285,37.035713,5.4,?
5569,19.428572,32.035713,5.4,?
5580,18.5,29.178572,5.4,?
5634,9.964286,31.892857,7.2,?

```

4.3.2 Creación del modelo predictivo.

Para la creación del modelo predictivo basado en el algoritmo del perceptrón, utilizamos el siguiente código en el framework R Studio.

```
setwd("D:/Tesis_GusX/Fuente_Datos_Moodle")
rm(list=ls(all=TRUE))
datos<-read.csv("aplicacionesweb_rocsn.arff",sep=";",dec=".",header = F,comment.char = "@")
taprendizaje<-datos
ttesting<-read.csv("individuosnuevos_di.arff",sep=";",dec=".",header = F,comment.char = "@")
library(nnet)
# Matriz de predictores
mpre<-cbind(datos$V1,datos$V2,datos$V3,datos$V4)
summary(mpre)
#Matriz de target:
target<-as.matrix(datos$V5)
select<-sample(1:29,8)
modal1<-nnet(V5~.,data = taprendizaje, size=100, rang = 0.1, decay =5e-4,
maxit=200,trace=FALSE)
prediccion=predict(modal1,ttesting)
names(prediccion)=c("prediccion")
datos.mpre=cbind(ttesting,prediccion)
summary(datos.mpre)
tapply(datos.mpre$prediccion, list(pvi=datos.mpre$V5),mean, na.rm=TRUE)
```

En la variable taprendizaje se le asignó los valores comprendidos en el Data Mart “dwh_calificaciones_v1”, misma que contiene todos los atributos de las calificaciones de los estudiantes incluido las dos clases de predicción en el atributo “estado”.

```
@relation estudiante_exitoso
@attribute estudiante
{2509,2511,2513,2514,2515,2516,2522,2524,2537,2538,2539,2540,2541,2542,2543,2544,2545,2547,2550,2551,2562,
,2564,2565,2573,2634,2635,2648,2693,2707}
@attribute nota_assign numeric
@attribute nota_quiz numeric
@attribute nota_participacion numeric
```


@attribute estado {si,no}

@data

2509,13.32,22.639,5.35714,si
2511,11.928,19.8787,1.42857,no
2513,10.92,19.4487,3.57143,no
2514,13.68,20.8037,2.85714,si
2515,7.728,16.273,0,no
2516,13.68,26.1913,1.42857,si
2522,12.528,27.908,7.14286,si
2524,13.248,17.6337,7.14286,si
2537,11.64,16.223,0.357143,no
2538,11.496,19.2527,7.14286,si
2539,11.88,16.7903,2.5,si
2540,12.6,23.693,1.78571,si
2541,11.28,20.3033,5.35714,si
2542,12.768,20.8373,7.14286,si
2543,13.68,28.6517,7.14286,si
2544,12.6,21.2403,7.14286,si
2545,13.248,22.131,3.57143,si
2547,11.208,20.0613,3.57143,no
2550,14.4,19.7713,1.78571,si
2551,4.08,19.75,0,no
2562,10.848,16.6157,3.57143,si
2564,12.36,20.255,0,no
2565,13.68,20.6737,7.14286,si
2573,12,13.2217,0.357143,no
2634,13.08,26.552,5.35714,si
2635,10.08,19.6963,3.21429,si
2648,10.8,21.5137,0,si
2693,7.08,10.196,1.42857,no
2707,3.6,18.06,2.14286,no

Con la ejecución del anterior scripts donde se asigna en la variable ttesting los datos de los individuos nuevos, tenemos los siguientes resultados:

Tabla 7-4: Individuos nuevos.
Predicción

id	Código	Predicción	Estado
1	4262	0,99997870	A
2	4320	0,97833270	A
3	4349	0,96385450	A
4	4722	0,97833270	A
5	4917	0,76382880	A
6	4924	0,76382890	A
7	4977	0,96385450	A
8	5125	0,00066350	R
9	5138	0,03805050	R
10	5172	0,99986370	A
11	5199	0,95268240	A
12	5208	0,96385430	A
13	5228	0,95274430	A
14	5295	0,96385450	A
15	5339	0,96385450	A
16	5449	0,98106630	A
17	5454	0,00029870	R
18	5476	0,96409100	A
19	5480	0,99030570	A
20	5484	0,95455910	A
21	5489	0,99991540	A
22	5506	0,09049020	R
23	5546	0,96381210	A
24	5548	0,96385450	A
25	5565	0,92562250	A
26	5569	0,95761340	A
27	5580	0,93133330	A
28	5634	0,81501110	A

Fuente: OASIS

Datos Reales

ID	Código	Porcentaje	Estado
1	4262	0,7	A
2	4320	0,875	A
3	4349	0,7	A
4	4722	0,7	A
5	4917	0,875	A
6	4924	0,725	A
7	4977	0,7	A
8	5125	0,7	A
9	5138	0,7	A
10	5172	0,7	A
11	5199	0,7	A
12	5208	0,9	A
13	5228	0,7	A
14	5295	0,7	A
15	5339	0,7	A
16	5449	0,7	A
17	5454	0,7	A
18	5476	0,875	A
19	5480	0,75	A
20	5484	0,7	A
21	5489	0,7	A
22	5506	0,7	A
23	5546	0,875	A
24	5548	0,9	A
25	5565	0,925	A
26	5569	0,875	A
27	5580	0,7	A
28	5634	0,7	A

4.4 Prueba de la hipótesis de investigación

La estadística es fundamental para las ciencias, para esto primero se debe plantear una hipótesis sobre una muestra que en nuestro caso será los estudiantes del noveno semestre de la materia Didáctica Informática de la carrera de Sistemas de la ESPOCH.

La detección de patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje, influye en el éxito de los estudiantes. Su verificación debe ser a través de una hipótesis estadística. Una hipótesis es una conjetura que nos ayuda a verificar si la hipótesis planteada es válida o se rechaza, para este estudio se ha planteado la hipótesis nula y la hipótesis alternativa.

Hipótesis Nula (H_0): Hace referencia a una población más no a una muestra. Siempre tiene el signo de la igualdad.

Hipótesis Alternativa (H_1): Es la hipótesis contraria a la hipótesis nula. No contiene el signo de igualdad.

4.4.1 Planteamiento de la hipótesis

H_0 = La detección de patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje, influye en el éxito de los estudiantes.

\bar{X}_A = Media aritmética de la predicción de individuos nuevos de la materia de Didáctica Informática.

\bar{X}_B =Media aritmética de las notas reales de los individuos nuevos de la materia de Didáctica Informática

$$H_0: \bar{X}_A = \bar{X}_B$$

H_1 = La detección de patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje, NO influye en el éxito de los estudiantes.

\bar{X}_A = Media aritmética de la predicción de individuos nuevos de la materia de Didáctica Informática.

\bar{X}_B =Media aritmética de las notas reales de los individuos nuevos de la materia de Didáctica Informática.

$$H_1: \bar{X}_A \neq \bar{X}_B$$

4.4.2 Nivel de significancia

También conocido como el nivel de riesgo porque se toma el riesgo de una probabilidad nula cuando esta pudo haber sido verdadera.

Se utilizará el nivel 0,05 (5%) ya que este proyecto es de consumo es decir se lo puede aplicar a otros Entornos Virtuales de Aprendizaje.

4.4.3 Criterios de validación de la hipótesis

Se lo realizará a través de una distribución normal por ser la probabilidad más útil y continua. También conocida como campana de Gauss.

Debido a que se dispone de los resultados de las evaluaciones de los mismos estudiantes para los que se hizo la predicción, además de no conocer la desviación estándar de la población, de conocer que los datos provienen de una distribución normal y además que el tamaño de la población es menor a 30; se ha procedido a aplicar la prueba de comparación de medias de muestras emparejadas basada en el cómputo de del estadístico t-student para muestras relacionadas con la finalidad de comprobar si se acepta o rechaza la hipótesis nula.

	Media	N	Desviación estándar	Media de error estándar
Par 1 PREDICCIÓN	,813769707143	28	,3308646072213	,0625275334529
PORCENTAJE	,75625000	28	,085695683	,016194962

	Diferencias emparejadas					t	gl	Sig. (bilateral)
	Media	Desviación estándar	Media de error estándar	95% de intervalo de confianza de la diferencia				
				Inferior	Superior			
Par 1 PREDICCIÓN - PORCENTAJE	,0575197071429	,3204661160754	,0605624033399	-,0667440801813	,1817834944670	,950	27	,351

Figura 7-4: Datos SPSS.

Fuente: OASIS y SPSS

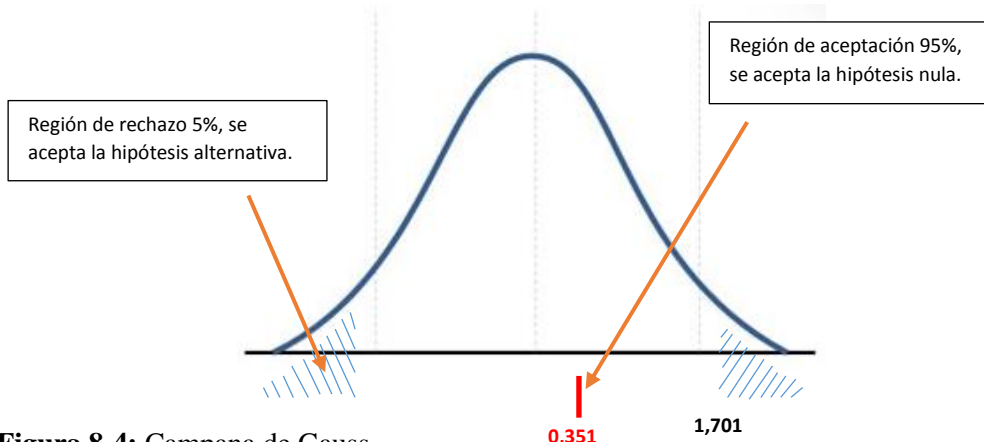


Figura 8-4: Campana de Gauss.
Realizado por: Gustavo Hidalgo.2016

Se ha obtenido un valor del estadístico t-student de 0,95 con un nivel de significación de 0,351; por lo que NO se puede rechazar la hipótesis nula donde ambas medias son iguales.

Por lo tanto la detección de patrones de participación empleando Minería de Datos en un Entorno Virtual de Aprendizaje, influye en el éxito de los estudiantes.

4.4.4 Toma de Decisiones

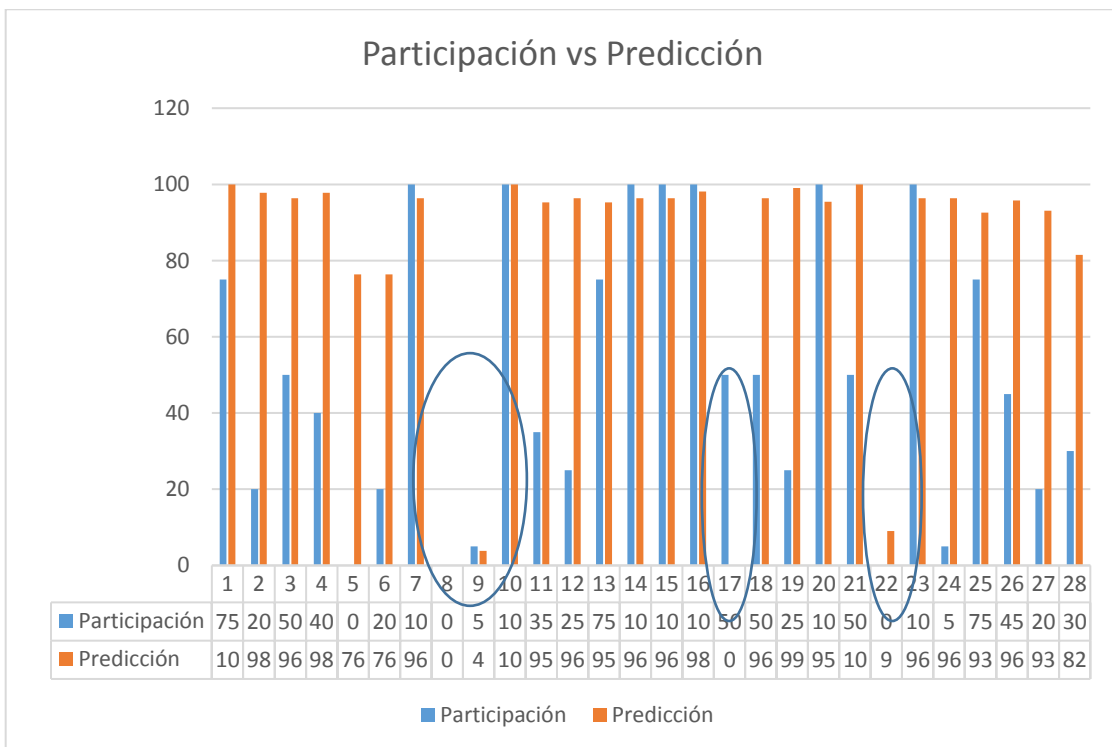


Figura 9-4: Participación vs Predicción.
Fuente: OASIS

Tabla 8-4: Análisis y Decisiones .

Análisis	Decisiones
La predicción con respecto a la participación nos indica que peligrosamente varios estudiantes no van a tener éxito en la cátedra.	El docente debe identificar a los estudiantes con dificultades.
La baja participación en el aula es la causante de esta condición. Posiblemente su participación ha sido exclusivamente en evaluaciones en línea, dejando a un lado las actividades y recursos.	El docente deberá dialogar con los estudiantes para encontrar la causa de su falta de participación.
Falta de incentivos y pérdida de interés.	El docente debe decidir e incentivar a que los estudiantes en la participación en tareas, foros, recursos, posiblemente aumentar el tiempo de gestión de aula.
Discriminación de actividades importantes.	Posiblemente el docente deberá nivelar las calificaciones de las distintas actividades, reforzar temas, tutorías extras.

Realizado por: Gustavo Hidalgo.2016

CONCLUSIONES

- Existe evidencia estadística que la predicción arroja resultados similares a los obtenidos en la evaluación real de los estudiantes.
- Los modelos predictivos basados en el algoritmo Perceptrón Multicapa es una herramienta poderosa que detección de patrones que permiten predecir con un 90% aproximadamente de exactitud.
- El uso de los Entornos virtuales de aprendizaje influye positivamente en el éxito de los estudiantes.
- Con el conocimiento adquirido gracias a los modelos predictivos se puede tomar decisiones para el futuro, basados en el aprendizaje del modelo.
- Las predicciones basadas en patrones de participación en los Entornos virtuales ayudara a los docentes a tomar las decisiones correctivas y preventivas para el éxito de sus estudiantes semestral, anual o periódicamente.
- Los patrones de participación pueden ayudar al estudiante para que siguiéndolos se encaminen en la ruta del éxito en sus estudios.
- La interpretación de la información contenida en los rastros que dejan los estudiantes en los distintos medios digitales, podría generar oportunidades para los docentes e instituciones de mejoras más eficientes y eficaces los procesos académicos; gracias a las predicciones que se realicen sobre esa información.
- Aquellos estudiantes que tuvieron participaciones mayores o iguales a un 75% en los Entornos virtuales estudiados, fueron los que tuvieron éxito en el semestre.
- Los estudiantes que mostraron tendencias en el uso de foros, exámenes en líneas, tareas, visitas a recursos didácticos, el modelo pronostica que aquellos estudiantes tienen más posibilidad de éxito en el curso, que aquellos que no participan activamente en las actividades antes mencionadas.

RECOMENDACIONES

- La institución se encuentra constantemente levantando información e indicadores para el CES, es importante tener herramientas de minería de datos que permita construir Data WareHouse para hacer predicciones que permita tomar decisiones de los procesos académicos, financieros y administrativos.
- Las herramientas de educación virtual un complemento interesante para la formación académica de los estudiantes; se recomienda que los docentes utilicen en sus EVAs actividades como foros, chat, recursos, exámenes en línea y tareas.
- El siguiente paso que en este tipo de estudios es la Analítica del Conocimiento, por tal motivo se recomienda que docentes y estudiantes se involucren en el estudio del tema.

BIBLIOGRAFÍA

- (1) **CABERO, J.** (17 de Marzo de 2012). Learning Analytics [Mensaje en un Blog]. Obtenido de <http://ibero.wiki.nmc.org/Lerning+Analytics>
- (2) **CAPTERRA. TOP LEARNING MANAGEMENT SYSTEM SOFTWARE PRODUCTS. CAPTERRA.** [En línea] Capterra, 2013. <http://www.capterra.com/learning-management-system-software/#infographic>.
- (3) **COMMUNITY, MOODLE.** Moodle. [En línea] Moodle.org, 2013. <https://moodle.org/>.
- (4) **DELIA C.,** Maestra. Qué es un modelo científico. Maestra Delia. [En línea] 2008. <http://maestradelia.wordpress.com/2008/01/20/que-es-un-modelo-cientifico-2/>.
- (5) **HAGGAR, P.** (8 de Agosto de 2011). Crecimiento de datos y estándares [Mensaje en un blog]. Recuperado el 9 de Febrero de 2015, de IBM DeveloperWorks: <http://www.ibm.com/developerworks/ssa/xml/library/x-datagrowth/>
- (6) **MOLINA, L.** (24 de Marzo de 2010). Torturando a los datos hasta que confiesen [Mensaje en un blog]. Obtenido de <http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html>
- (7) **MORENO, M.** (15 de Marzo de 2009). Aplicación de técnicas de minería de datos en la construcción y validación de modelos predictivos y asociativos a partir de especificaciones de requerimientos de software. Salamanca, España. <http://www.ibm.com/developerworks/ssa/xml/library/x-datagrowth/>
- (8) **TORRES, V.** (5 de Abril de 2015). Información [Mensaje en un Blog]: Obtenido de <https://es.wikipedia.org/wiki/informacion>
- (9) **SANTAMARÍA, F.** (13 de Septiembre de 2012). Análisis de Aprendizaje [Mensaje en un Blog]. Obtenido de <http://fernandosantamaria.com/blog/2012/09/learning-analytics-analisis-del-aprendizaje-2/>

ANEXOS

Anexo A: . PROPUESTA DE PROYECTO DE DESARROLLO ACADÉMICO.

1. DATOS GENERALES DEL PROYECTO

1.1 NOMBRE DEL PROYECTO:

Proyecto de detección de patrones de participación empleando minería de datos en entornos virtuales, para predecir estudiantes exitosos.

1.2 ENTIDAD EJECUTORA:

Institución: ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO.

Unidad Académica: INSTITUTO DE POSTGRADO Y EDUCACIÓN CONTINUA.

Participantes: GUSTAVO XAVIER HIDALGO SOLÓRZANO, ING. FRANKLIN MORENO, ING. NICOLAS SAMANIEGO, ING. JORGE HUILCA.

1.3 COBERTURA Y LOCALIZACIÓN:

Provincia: Chimborazo

Cantón: Riobamba

Parroquia: Veloz

Dirección: Panamericana sur km 1 $\frac{1}{2}$, vía a Guayaquil.

1.4 PLAZO DE EJECUCIÓN (EN MESES):

Fecha de inicio tentativo: 2 de Mayo del 2016

Fecha de finalización tentativa: 29 de Julio del 2016

Horas de trabajo: 240 horas.

Tiempo de duración en meses: 3 meses.

1.5 SECTOR Y TIPO DEL PROYECTO:

Tipo de Proyecto: Proyecto Social.

Línea de vinculación: Aplicación de las tecnologías de la información y comunicaciones.

Sector: Educación

Subsectores: Infraestructura y servicios

1.5.1 Beneficiarios

Al ser un proyecto social dirigido a la toma de decisiones en beneficio de los estudiantes; con la detección de patrones o rastros que dejan en los archivos históricos, se puede tener información importante para la toma de decisiones por parte de ellos y así logara el éxito en cada semestre.

Los docentes es otro grupo de beneficiados en el proyecto, teniendo a la mano el perfil del estudiante basado en sus patrones de participación pueden planificar su catedra basados en esas preferencias, de esta manera aportara en la disminución de la deserción y/o fracaso estudiantil cada semestre.

De la misma manera, los directores de escuela pueden planificar su gestión administrativa, tomado datos de las predicciones y planificando su recurso humano (docentes) e infraestructura física, para los semestres posteriores.

1.6 CONTRIBUCIÓN DEL PROYECTO A LAS METAS DEL PLAN NACIONAL DEL BUEN VIVIR

Estrategia 6.5: “Transformación de la educación superior y transferencia de conocimiento a través de ciencia, tecnología e innovación.”

Estrategia 6.6: “Conectividad y telecomunicaciones para la sociedad de la información y el conocimiento.”

Objetivo 2: “Mejorar las capacidades y potencialidades de la ciudadanía”.

2. DIAGNÓSTICO DEL PROBLEMA

2.1 DESCRIPCIÓN DE LA SITUACIÓN ACTUAL

La Escuela Superior Politécnica de Chimborazo (ESPOCH), está ubicada en la provincia de Chimborazo, en la ciudad de Riobamba en la Panamericana Sur Km 1 $\frac{1}{2}$, vía a Guayaquil. Es una de las Instituciones de Educación Superior (IES) reconocidas por el Consejo de Educación Superior (CES) y el SENESYT.

La ESPOCH, está comprometida en mejorar continuamente sus procesos académicos, es por eso que junto a sus docentes, empleados y estudiantes se plantea una serie de proyectos enfocados a ese fin.

Uno de los proyectos que más éxitos cosechó fue el de la plataforma virtual de Educación en línea; este sistema permite tener un conjunto de recursos y actividades disponibles para los estudiantes a cualquier instante del tiempo, esta herramienta es utilizada por los docentes y estudiantes como complemento al aprendizaje formal.

Muchos estudios se ha realizado sobre los datos que ofrece esta plataforma virtual, estándares de educación, metodologías de educación e-learning, estándares de evaluación a estudiantes, entre otros.

En la actualidad hay una gran gama de información dispersa en las bases de datos de esta aplicación web, misma que bajo un manejo especial se podría convertir en conocimiento y este a su vez ser utilizado para la toma de decisiones por parte de los docentes y autoridades de la institución.

Actualmente el 15% de los docentes de la ESPOCH utilizan las aulas virtuales para interactuar con los estudiantes luego de las clases formales, a esto se le denomina metodología B-learning o Blended learning, de los cuales se benefician alrededor del 25% de la población estudiantil.

Estudiantes y docentes, concuerdan que la aparición de este tipo de sistemas ha mejorado en gran medida el proceso de enseñanza-aprendizaje, pero ¿realmente el uso de esta plataforma y sus recursos influye en el éxito de un estudiante? ¿Existen patrones de participación en este tipo de plataforma que influyen en el éxito de los estudiantes?

Es una pregunta que aún no tiene respuesta ya que se requiere de factores internos y externos para conocer el perfil del estudiante éxitos, este proyecto intenta definir cuáles son los patrones internos desde

el punto de vista de la plataforma que influyen en que un estudiante tenga éxito en un determinado semestre.

Tipo de conexión a internet: El Instituto de postgrado y educación continua dispone del servicio de Internet que es proporcionado por la Dirección de Tecnología de la información y Comunicación (DTIC), distribuido por la empresa de telecomunicaciones Telconet, el mismo que presta servicio de internet al edificio de la IPEC.

2.2 IDENTIFICACIÓN, DESCRIPCIÓN Y DIAGNÓSTICO DEL PROBLEMA

Diariamente se genera gran cantidad de datos en las tareas cotidianas que se realiza, por ejemplo; cuando caminan, corren o practican algún deporte, se está generando datos como: km recorridos, número de pulsaciones de nuestro corazón, cantidad de fluidos corporales, tiempos de recorridos por km, entre otros.

El mundo digital es el más grande repositorio de datos que existe, ya que constantemente se genera información gracias a herramienta de comunicación como el chat en el telefonía móvil, navegar por el Internet, sistemas de procesamiento de información, entornos virtuales de educación, entre otros.

“Estos datos están dispersos por todas partes ya sean en documentos, archivos o bases de datos; el crecimiento de los sistemas de bases de datos se debe en gran medida, gracias a la gran cantidad de datos que se generan por segundo; en estos tiempos ya no se habla de almacenamientos en Megabytes ni Gigabytes sino en Terabyte, dentro de 10 años estaremos hablando de Petabyte y en 5 años más de Exabyte, En 2002, hubieron unos cinco Exabytes de datos online”. (Haggar, 2011, P.,8).

En el 2009 el total se incrementó a 281 Exabytes, una tasa de crecimiento de 56 veces en siete años. La cantidad total de datos almacenados por empresas se está duplicando cada tres años. (Haggar, 2011, P.,8).

Los datos que se originan diariamente por la interacción de las personas; cuando se someten a una la interpretación son transformados en información, luego cuando un especialista interpreta la información se convierte en conocimiento, y es ahí donde se agrega valor a los datos.(Elearn Training Company, 2007).

La educación no puede estar alejada de este nuevo paradigma de evolución del conocimiento, con la utilización de redes sociales, plataformas de educación en línea o virtual, e-commerce, e-goberment, entre otros.

La Escuela Superior Politécnica de Chimborazo (ESPOCH), genera una gran cantidad de información de estudiantes y docentes gracias a las aplicaciones o sistemas que poseen, entre ellos están: el Sistema Académico Institucional OASIS (academico.espoch.edu.ec), la Plataforma de Educación Virtual EVIRTUAL-ESPOCH (evirtual.espoch.edu.ec), el Sistema de Recursos Humanos (recursos.espoch.edu.ec), el Sistema de Evaluación Institucional (evaluacion.espoch.edu.ec), el Sistema Médico (medicina.espoch.edu.ec), Sistema de Biblioteca (biblioteca.espoch.edu.ec), entre otros. La Institución carece de un medio, técnica o método de minería de datos para el análisis de estos datos, limitando en gran parte el conocimiento de este análisis y por ende impidiendo que la toma de decisiones sea efectiva para la gestión de recursos.

“La Minería de Datos educacionales es una disciplina emergente dedicada a desarrollar métodos para analizar una gran cantidad de datos provenientes de ambientes relacionados a la educación, con el objetivo de entender los datos que generan los estudiantes, profesores y otros actores relacionados a sus entornos educacionales”. (Molen, 2013).

La Minería de datos se clasifica en 2 categorías; métodos supervisados o predictivos y métodos no supervisados o descubrimiento del conocimiento. Entre los supervisados tenemos: árbol de decisión, inducción neuronal, regresión, series temporales; mientras que en los no supervisados tenemos: detecciones de desviaciones, segmentación, agrupamiento o cluster, reglas de asociación, patrones secuenciales, a estos dos últimos también se los conoce como análisis de asociación.

En el ámbito académico de la ESPOCH, tenemos: el Sistema Académico Institucional y la Plataforma de Educación Virtual; estos sistemas tienen en sus bases de datos una extensa y valiosa información de cada uno de los estudiantes, que no se ha estudiado y/o analizado en conjunto.

La cátedra de Aplicaciones Web, forma parte del quinto semestre de la Escuela de Ingeniería en Sistemas Informáticos de la Facultad de Informática y Electrónica, es una de las materias profesionalizantes de la carrera y cuenta con una amplia información y datos en las bases de datos históricas tanto en la Plataforma Virtual de Educación como en el Sistema Académico Institucional para realizar esta investigación.

La carencia de patrones de participación; se convierte en un problema para la Institución, ya que no existe un medio, método o modelo para identificar a los estudiantes NO exitosos; de esta manera los Docentes no tienen los fundamentos reales necesarios para tomar decisiones con respecto a este grupo de estudiantes y por ende el índice de deserción estudiantil se incrementa gradualmente cada semestre.

2.2.1 Descripción del problema.

La Escuela Superior Politécnica de Chimborazo no cuenta con un estudio adecuado para la detección de patrones de participación que permita a los docentes e investigadores definir indicadores de los estudiantes exitosos, para de esta manera identificar a los estudiantes no exitosos y poder tomar decisiones antes de que fracasen en el semestre en curso.

La ESPOCH, cuenta con un reglamento aprobado desde el año 2009, indica que existen 3 tipos de evaluaciones que son: acumulativas, finales y suspensión. La evaluación acumulativa corresponderá a pruebas parciales teóricas y/o prácticas, lecciones, trabajos de investigación y más parámetros de evaluación edumétrica, establecidos en el cronograma de actividades del docente.

La evaluación será procesual, sistemática y continua e implicará la valoración de conocimientos, destrezas, habilidades y actitudes, por lo que la calificación global se determinará de la siguiente manera:

- a) La calificación de evaluación acumulativa, 70%, constituida por pruebas, lecciones, trabajos de investigación y más parámetros de evaluación edumétrica. Que no deberán ser menores de tres (3) componentes;
- b) La calificación de evaluación final, 30%;
- d) El estudiante que en la evaluación acumulativa reuniera el 90% de la calificación, veinte y cinco (25) puntos, será exonerado de rendir la evaluación final y se considerará aprobado; para lo cual se sumará al valor de la evaluación final, doce (12) puntos.

Las calificaciones se contabilizarán en la escala de cero (0) a veintiocho puntos (28), en las evaluaciones acumulativas y de 0 a 12 puntos en la evaluación final; serán siempre en cifras correspondientes a números enteros; la fracción cero punto cinco (0.5) ó más, se aproximará a la cifra inmediata superior.

Una de las actividades edumétricas que los docentes de la ESPOCH han adoptado es la participación en los Entornos virtuales de Aprendizaje, estas herramientas facilitan la labor de aprendizaje tanto para docentes como para los estudiantes, tanto así que los docentes destinan un porcentaje de la nota total del curso a la participación de los estudiantes en los EVAs.

Además de ser una de las herramientas más utilizadas, captan una gran cantidad de información tanto del docente como de los estudiantes; esta información en los actuales momento no brinda el valor agregado que debería, ya que la ESPOCH carece de un modelo para extraer esa información estudiarla y compararla con otras fuentes de datos como el Sistema académico para convertirla en conocimiento a beneficio de la institución y de los mismos estudiantes.

La Minería de Datos (Data Mining) y sus métodos supervisados junto con las técnicas de minería como: árbol de decisión, agrupamientos o Clustering entre otros; ayudan a modelar esta información para detectar los patrones de participación de los estudiantes exitosos en los Entornos Virtuales de Aprendizaje (EVA) de la Plataforma de Educación Virtual de la Escuela Superior Politécnica de Chimborazo; así los especialistas puedan analizar y evaluar esta información para de esta manera generar conocimiento y poder tomar decisiones que en el área de la academia podría ser denominada Analítica de Aprendizaje. (Moreno, 2009).

Pero ¿Cómo la Minería de Datos podría ser utilizada para detectar patrones de participación para la predicción de estudiantes exitosos?. La respuesta a estas preguntas invita a plantearnos una hipótesis que permita conseguir resultados que ayuden a entender y manipular los datos, información y convertirlos en conocimiento.

2.2.2 Sistematización del problema.

La institución está interesada en conocer y entender la información que puede brindar sus Aplicaciones o Sistemas Académicos con los que cuenta; en tal virtud se acude a las técnicas de minería de datos para entender el comportamiento participativo de los estudiantes y conocer si estos patrones influyen en el éxito de los estudiantes. Por ello es importante identificar las variables que presenta este problema.

Partiendo del problema planteado surgen las siguientes interrogantes:

¿Qué impacto tendrá descubrir los patrones de éxito de un estudiante en el desarrollo académico de los estudiantes de Aplicaciones Web de la Escuela de Sistemas?,

¿Cuáles han sido los resultados de la detección de patrones de participación de estudiantes exitosos?

2.3 Líneas del proyecto

El descubrimiento de patrones de participación de los estudiantes en la plataforma de educación virtual se ha convertido en un importante insumo de información para que docentes y estudiantes puedan tomar decisiones con respecto a sus comportamientos dentro del proceso de enseñanza aprendizaje.

Por tal motivo este proyecto se enmarca en la línea de vinculación de aplicación de las tecnologías de la información y comunicación en el sector de la educación subsectores infraestructura y servicios.

2.3.1 Descripción

Mediante la utilización de la tecnología informática se realizara ensayos de experimentos con datos de las bases de datos históricas de la plataforma virtual, de esta manera se tomará los datos de la cátedra de Aplicaciones Web del quinto semestre de la escuela de ingeniería en sistema de la facultad de informática y electrónica.

Se escogió esta cátedra ya que en el periodo comprendido de septiembre 2014 a febrero 2015, cumplió con la mayor parte de los parámetros de estudio que se requieren como es:

- 1.- Uso de actividades como quiz, assign, fórum, glossary, course, entre otros.
- 2.- Porcentaje de calificación de las actividades de EVA son mayores o iguales a 60,7%
- 3.- Era la única que cumplía los requisitos completos como se muestra en la tabla# del capítulo IV.

Además, se utiliza las herramientas y algoritmos de minería de datos para descubrir los patrones que se encuentran escondidos en esos datos.

3. Objetivos de la investigación.

3.1 Objetivo general.

Realizar un modelo de detección de patrones de participación empleando minería de datos en el Entorno Virtual de Aprendizaje de Aplicaciones Web de la ESPOCH para predecir estudiantes exitosos.

3.2 Objetivos específicos.

Determinar los patrones de participación de los estudiantes en el Entorno Virtual de Aprendizaje de la ESPOCH.

Identificar el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje, compararlas con las calificaciones obtenidas en las evaluaciones acumulativas.

Determinar si el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje influyen a que los estudiantes sea exitosos.

4. Matriz de marco lógico

Resumen narrativo de objetivos	Indicadores Verificables Objetivamente	Medios de Verificación	Supuestos
<p>FIN</p> <p>Identificar patrones de participación en el entorno virtual de educación y predecir estudiantes exitosos.</p>	<p>Numero de patrones encontrados.</p> <p>Porcentaje de verdaderos positivos mayor al 89,7%</p>	<p>Registros de la ejecución del scripts en R studio.</p>	<p>Implementando el proyecto y difundiendo sus ventajas reales.</p> <p>Detección permanente de patrones.</p>
<p>PROPÓSITO:</p> <p>Realizar un modelo de detección de patrones de participación empleando minería de datos en el Entorno Virtual de Aprendizaje de la ESPOCH para predecir estudiantes exitosos.</p>	<p>Porcentaje mayor al 89,7% de aprendizaje del modelo</p>	<p>Registro de la predicción</p>	<p>Mejoramiento del modelos al 100%.</p>
<p>COMPONENTES</p> <p>1. Determinar los patrones de participación de los estudiantes en el</p>			

<p>Entorno Virtual de Aprendizaje de la ESPOCH.</p> <p>2. Identificar el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje, compararlas con las calificaciones obtenidas en las evaluaciones acumulativas.</p> <p>3. Determinar si el porcentaje de calificación de la participación de los estudiantes en el Entorno Virtual de Aprendizaje ayudan a que los estudiantes sea exitosos.</p>	<p>Número de participaciones por estudiantes</p> <p>Porcentaje de calificaciones de los estudiantes</p> <p>Número de actividades o módulos.</p> <p>Porcentaje de participación en las actividades o módulos.</p> <p>Porcentaje de estudiantes que logran terminar el semestre con éxito.</p>	<p>Registro de participaciones de los estudiantes en la tabla hechos_participacion</p> <p>Registro de participaciones de los estudiantes en la tabla hechos_participacion</p> <p>Registro de participaciones de los estudiantes en la tabla hechos_participacion y hechos_calificaciones</p>	
<p>ACTIVIDADES</p> <p>1. Identificar las fuentes de datos origen.</p>		<p>url1: http://localhost/dwh_calificaciones_v2</p>	<p>Contar con las herramientas e implementos necesarios para la predicción de éxito</p>

1.1 Restaurar la base de datos de la plataforma virtual en un servidor local.	dos data mart en el motor de base de datos Postgresql	url2: http://localhost/dwh_participacion Programa instalado en el PC	o fracaso de distribución libre.
1.2 Instalar el weka.			Contar con computadoras sofisticadas para realizar experimentos con datos mas grandes.
1.3 Instalar el R studio.	Una instancia de weka instalada		
2. Realizar el pre procesamiento de datos		Descripción de los pasos (steps) tabla de entrada en las transacciones del pentaho.	
2.1 Creación de los scripts SQL para pre procesar los datos que se necesitan en las dimensiones.	Seis script SQL creados y ejecutados		
2.2 Supervisar la ejecución de los scripts SQL.		Previsualización de la ejecución en la consola del motor de base de datos.	
2.3 Instalar Pentaho			
2.4 Diseñar y probar las transformaciones para el proceso de extracción.	Un programa instalado en la PC.	Estacio de trabajo (Spoon) del pentaho. Usuario admin, contraseña: admin.	
3. Realizar la limpieza de datos nulos.	6 transformaciones		
3.1 Extraer los datos al data mart creado anteriormente.	realizadas en pentaho	url1: http://localhost/dwh_calificaciones_v2 url2: http://localhost/dwh_participacion	
3.2 Realizar la conexión de la tabla de hechos con el weka.			

<p>3.3 Someter los datos a los algoritmos de clasificación predictivos.</p> <p>3.4 Escoger el algoritmo mediante la curva ROC.</p> <p>3.5 Crear el modelo con el algoritmo escogido.</p> <p>3.6 Realizar la predicción.</p>	<p>19 tablas creadas para extracción de datos</p> <p>cuatro algoritmos predictivo</p> <p>El algoritmo me mas se acerque a 1 es el que mejor predecirá los aprobados.</p> <p>Los estudiantes con más del 90% de puntos serán los que aprueben el semestre.</p>	<p>Registro de experimentos en el software weka.</p> <p>Reporte del grafico de la curva ROC, en weka y en R studio.</p> <p>Lista de estudiantes con su respectivo porcentaje de aprobación.</p>	
---	---	---	--

5. Viabilidad del plan de sostenibilidad

5.1 Viabilidad Técnica

Para la ejecución del presente proyecto no es necesario adquirir computadores, licencias de software, materiales de red para el trabajo de mantenimiento hardware/software y para la capacitación, ya que se utilizará software libre y el equipamiento tecnológico, energía eléctrica, internet provee la ESPOCH.

El recurso humano para el proyecto es de 4 profesionales en el área de informática y bases de datos; este recurso puede ser proporcionado por la misma institución, misma que puede cubrir la inversión inicial ya que el monto es bajo. De esta manera el proyecto es viable técnicamente ya que la institución cuenta con la tecnología, los recursos económicos y los profesionales con los conocimientos suficientes para ejecutar el proyecto.

5.2 Especificaciones técnicas

Se utilizará la herramienta Pentaho Data Integration (Spoon) para la extracción, transformación y limpieza de datos, además de la construcción de los data mart en el motor de base de datos PostgreSQL. Luego se utilizara el Weka y el entorno de desarrollo R-Studio. Este conjunto de herramientas tiene la característica de que todas son de licencia libre, de esta manera la Institución no tiene que pagar ningún valor por licencias.

Los profesionales responsables para ejecutar este proyecto son: un coordinador y tres analistas de información y comunicación con los detalles económicos que se muestran en la factibilidad económica.

5.3 Implementación

Para la implementación del proyecto la ESPOCH deberá aportar con toda la infraestructura y recursos para su ejecución; bajo esta premisa se utilizará los siguientes recursos:

Recurso	Características	Cantidad
Computadora	Procesador i7. 2GHz Pantalla LED 15.6". 8GB en RAM DDR3. 1 TB SATA 6gb/s 7200 rpm Windows 8.1 PD 64 bits.	4
Internet	Banda ancha mínimo 3 Mbps	4

5.4 VIABILIDAD ECONÓMICA Y FINANCIERA

COSTOS Y GASTOS							
Denominación	Cant.	V. U.	Unidad	Duración.	V. T	Aporte	Metodología
OPERACIÓN Y ADMINISTRACIÓN							
Coordinación extensionis	1	20	\$/ h	240	4800	ESPOCH	Ingreso promedio, calculado a nivel hora (Valor estimado)
Operativos	3	17,5	\$/ h	240	12600	ESPOCH	Salario mínimo, referente legal, calculado a nivel hora (Valor estimado para tres estudiantes)
MATERIALES							
Papel bond 75 g	40	0,1	\$/ resma	1	4	ESPOCH	Papel calculado para elaborar informes y proyecto
Fotocopiado	80	0,02	\$/ copia	1	1,6	ESPOCH	Calculado para evidencias y respaldos
Carpetas	4	0,25	\$/ folder	1	1	ESPOCH	Calculado para proyectos e informes
HERRAMIENTAS							
Software	4	0	\$/ licencia	0	0	Autogestión	Calculado para la extracción, transformación y limpieza de datos.
INSUMOS							
Electricidad	1	40	\$/mes cons	2	80	ESPOCH	Calculado para el mantenimiento de los computadores, el valor se considera tomando en cuenta un valor estimado por alquiler un laboratorio por los cuatro meses.
Servicio internet	1	40	\$/mes cons	2	180	ESPOCH	Calculado para consultas e investigaciones posteriores, el valor se considera tomando en cuenta un valor estimado por alquiler un laboratorio por los cuatro meses.
Total					17666,6		

5.5 COSTO / BENEFICIO

El costo del proyecto según el estudio económico es de 17666.6, este valor será asumido en su totalidad por la Institución; de la misma forma el beneficio que representa esta inversión se retornara en los siguientes aspectos:

El proyecto está pensado en el mejoramiento académico de los estudiantes, de esta manera el beneficio es Institucional ya que los indicadores de gestión académico mejorarán, en cada una de sus dependencias o unidades académicas esto provocará que se cumpla con los indicadores institucionales que solicita el Consejo de Educación Superior (CES) para trazar el camino y aspirar a la recategorización a la categoría A.

Además se tendrá información y el conocimiento para tomar decisiones en beneficio de los involucrados en el sistema educativo de la institución; la información que generara el proyecto tendrá un impacto positivo ya que se podrá planificar los recursos (docentes, empleados, estudiantes, infraestructura física, recursos económicos) institucionales gracias a las predicciones que se obtengan.

6. ANÁLISIS DE SOSTENIBILIDAD

Para efectos de sostenibilidad y sustentabilidad del proyecto es necesaria la participación de las autoridades institucionales en la ejecución del proyecto y más aún en la ejecución del proyecto desde las facultades que deseen implementarlo para que se pueda replicar el contenido en futuras ocasiones tanto a los estudiantes, docentes y administrativos de la Institución, contribuyendo al mantenimiento e incremento del capital social.

Con la participación de las autoridades de las facultades con la información que se genere semestralmente que será analizada mediante la herramienta de Minería de Datos Pentaho, se garantiza la sustentabilidad del proyecto. Además el proyecto es sostenible ya que la institución no volverá a invertir un solo dólar más mientras el proyecto se ejecute.

7. PRESUPUESTO

ACTIVIDADES	INSTITUCIÓN	TIPOS DE RECURSO	DESCRIPCIÓN	VALOR TOTAL
1. Identificar las fuentes de datos origen.	ESPOCH	FISCAL	Coordinador	\$ 4.800,00
1.1 Restaurar la base de datos de la plataforma virtual en un servidor local.	ESPOCH	FISCAL	Operativo	\$ 12.600,00
1.2 Instalar el weka.	ESPOCH	FISCAL	Internet	\$ 180,00
1.3 Instalar el R studio.	ESPOCH	FISCAL	Electricidad	\$ 80,00
2. Realizar el pre procesamiento de datos	ESPOCH	FISCAL	Papel bond 75 g	\$ 4,00
2.1 Creación de los scripts SQL para pre procesar los datos que se necesitan en las dimensiones.	ESPOCH	FISCAL	Fotocopiado	\$ 1,60
2.2 Supervisar la ejecución de los scripts SQL.	ESPOCH	FISCAL	Carpetas	\$ 1,00
2.3 Instalar Pentaho				
2.4 Diseñar y probar las transformaciones para el proceso de extracción.				
3. Realizar la limpieza de datos nulos.				
3.1 Extraer los datos al data mart creado anteriormente.				
3.2 Realizar la conexión de la tabla de hechos con el weka.				
3.3 Someter los datos a los algoritmos de clasificación predictivos.				
3.4 Escoger el algoritmo mediante la curva ROC.				
3.5 Crear el modelo con el algoritmo escogido.				
3.6 Realizar la predicción.				
Total				\$ 17.666,60

