



Universidad Popular Autónoma del Estado de Puebla
Centro Interdisciplinario de Posgrados

Investigación y Consultoría

Departamento de Ingeniería

Doctorado en Planeación Estratégica y Dirección de
Tecnología

Título de la investigación

Diseño del Modelo predictivo “PredATIS” en una
Comunidad Virtual Académica de Alerta Temprana en
indicadores Educativos (CVAATI).

Tesis que para obtener el Grado de Doctor
en Planeación Estratégica y Dirección de Tecnología

Presenta

Petra Norma Maya Pérez

Puebla, México.

Junio, 2018



UPAEP – Secretaría General

Dirección General de Apoyos Académicos

Dirección del Centro de Recursos para el Aprendizaje y la Investigación.

Biblioteca Central - **Karol Wojtyla**

Tesis Digitales Restricciones de uso:

DERECHOS RESERVADOS ©

PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de textos, imágenes, gráficas, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente de donde la obtuvo mencionando el autor o autores involucrados en el documento.

Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Carta de Aprobación de Tesis



Universidad Popular Autónoma del Estado de Puebla
Centro Interdisciplinario de Posgrados
Investigación y Consultoría
Departamento de Ingeniería
Doctorado en Planeación Estratégica
y Dirección de Tecnología

Se aprueba la Tesis:

Nombre de la tesis

Diseño del Modelo predictivo "PredATIS" en una Comunidad Virtual
Académica de Alerta Temprana en indicadores Educativos (CVAATI)

Nombre del Alumno

Petra Norma Maya Pérez

Comité Asesor

Dr. Jorge Rafael Aguilar Cisneros

Director(a) de Tesis

Dra. Rosa Angélica Zamora Ríos

Co-Director (a)

Dr. José Miguel Barrón Adame

Asesor(a)

Puebla, México.

15 de Junio de 2018

DEDICATORIAS

El presente trabajo está dedicado a toda mi familia, a mis padres y hermanos por todo el apoyo brindado en los momentos más difíciles y estresantes, que me permitieron desarrollar y lograr este proyecto de vida profesional.

A mis estimados estudiantes que por algún motivo han dejado sus estudios universitarios.

AGRADECIMIENTOS

A Dios principalmente por permitirme Vida y Salud, sin ello no sería posible realizar este trabajo.

A mi familia especialmente a mi mamá y hermana Araceli, por todo su apoyo, comprensión y paciencia que hicieron posible concluir otra etapa de mi vida profesional.

A mi amiga Adelina y sus familiares que me ofrecieron su apoyo incondicional de mi estancia en su hogar.

A la Universidad Tecnológica del Valle de Toluca por brindarme todo el apoyo para realizar mis estudios de Doctorado.

A mis asesores de dirección de Tesis doctoral Dr. Jorge Rafael Cisneros Aguilar, Dra. Rosa Angélica Zamora Ríos y Dr. José Miguel Barrón Adame por su apoyo brindado para la conclusión de este documento.

A la coordinación e instructores del programa educativo de este doctorado de la Universidad Popular Autónoma del Estado de Puebla en el apoyo de gestión de documentos y por compartir sus conocimientos y experiencia.

ÍNDICE GENERAL

Carta de Aprobación de Tesis.....	ii
DEDICATORIAS.....	iii
AGRADECIMIENTOS.....	iv
RESUMEN.....	vii
ABSTRACT.....	viii
INTRODUCCIÓN.....	ix
CAPÍTULO 1. PROPÓSITO Y ORGANIZACIÓN.....	1
1.1 Planteamiento del problema.....	1
1.2 Preguntas de investigación.....	10
1.3 Propósito de la investigación.....	10
1.4 Objetivo General.....	11
1.5 Objetivos específicos.....	12
1.6 Justificación de la investigación.....	13
1.7 Alcances y limitaciones.....	15
1.8 Organización del estudio.....	16
CAPÍTULO 2. MARCO TEÓRICO.....	19
2.1 Antecedentes relacionados con estadísticos educativos.....	19
2.1.1 Indicadores estadísticos educativos.....	19
2.1.2 Causas de deserción y reprobación escolar en los distintos niveles educativos.....	24
2.1.3 Sistemas de Alerta Temprana (SIAT) en indicadores educativos.....	27
2.2 Comunidades Virtuales para establecer una comunicación estratégica.....	28
2.2.1 Comunidad virtual.....	28
2.2.2 Aspectos relevantes para la creación de una Comunidad Virtual.....	31
2.2.3 Comunidad Virtual Académica (CVA).....	36
2.3 Estrategias Tecnológicas.....	37
2.4 Antecedentes de Minería de Datos y Redes Neuronales Artificiales.....	42
2.4.1 Descubrimiento de Conocimiento en Bases de Datos KDD.....	42
2.4.2 Minería de Datos (Data Mining DM).....	46
2.4.3 Minería de Datos Educativa (Educational Data Mining EDM).....	50
2.4.4 Redes Neuronales Artificiales (RNA).....	52
2.5 Modelos de Predicción en indicadores educativos.....	61
2.5.1 Revisión de Modelos de Predicción.....	61
2.6 Ingeniería y Arquitectura de Desarrollo de Software de la Comunidad Virtual.....	68
2.6.1 Modelos de Desarrollo de SW.....	68
2.6.2 Metodologías para el Desarrollo Ágil de SW.....	72
2.6.3 Modelado UML.....	78
2.6.4 Modelo MVC arquitectura en el Desarrollo de SW.....	85
CAPÍTULO 3. METODOLOGÍA.....	90

3.1 Fundamentos de la Investigación	90
3.2 Diseño de la Investigación	91
3.3 Hipótesis	93
3.4 Selección de la muestra.....	94
3.5 Recolección de datos	95
CAPÍTULO 4. DESARROLLO DEL MODELO PROPUESTO EN LA COMUNIDAD VIRTUAL ACADÉMICA.....	99
4.1 Diseño de la Comunidad Virtual Académica CVAATI	99
4.1.1 Modelado UML y arquitectura de Desarrollo de la CVAATI.	105
4.1.2 Construcción de la CVAATI.	112
4.2 Diseño del Modelo PredATIS	116
4.2.1 Análisis de Algoritmo de aprendizaje con DM y RNA.....	117
4.2.2 Integración y Recopilación de Datos en la CVAATI.	123
4.2.3 Selección, Limpieza y Transformación de Datos en CVAATI.	125
4.2.4 Diseño del Modelo PredATIS con DM.	137
4.2.5 Diseño del Modelo PredATIS con RNA.	147
CAPÍTULO 5. MODELADO DE PredATIS EN LA CVAATI	155
5.1 Modelado de PredATIS con algoritmos de DM	155
5.1.1 Modelado basado en algoritmos de Clasificación y Selección de DM.	155
5.2 Modelado de PredATIS con algoritmos de RNA.	162
5.2.1 Modelado de PredATIS basado en RNA de FBR.	164
CAPÍTULO 6. ANÁLISIS Y DISCUSIÓN	181
6.1 Análisis de Resultados del modelo PredATIS.....	181
6.1.1 Validación del Modelo PredATIS basado en algoritmos de Clasificación DM.....	181
6.1.2 Validación del Modelo PredATIS basado en algoritmos de RNA de FBR.	185
6.2 Discusión	188
CONCLUSIONES Y RECOMENDACIONES	189
TRABAJO FUTURO Y DIVULGACIONES	191
REFERENCIAS	193

RESUMEN

En esta tesis doctoral, se propone el diseño del Modelo Predictivo de Análisis, Transformación e Interpretación de datos educativos en el nivel Superior (PredATIS), con el propósito de abordar la problemática multifactorial de la deserción y reprobación escolar, en una Comunidad Virtual Académica de Alerta Temprana en Indicadores Educativos (CVAATI). PredATIS es un modelo predictivo y adaptativo, con aporte científico al campo de investigación de Minería de Datos (DM), donde se aplican métodos y técnicas de clasificación DM y Redes Neuronales Artificiales (RNA), a partir de algoritmos no supervisados y supervisados de aprendizaje automático para la selección y búsqueda de mejores atributos, como factores de mayor impacto sobre el fenómeno en estudio; adicionalmente de una aportación tecnológica, al generar una herramienta que permita proveer información útil, la explotación de datos que generan conocimiento de la temática de indicadores educativos, reportar alertas predictivas, con la finalidad de establecer una comunicación estratégica, que coadyuve en la toma de decisiones del entorno educativo. En el diseño de PredATIS se combinan metodologías, técnicas y algoritmos de DM e Ingeniería de Software y herramientas WEB para el diseño propio de la CVAATI en el pre procesamiento de datos, que son la base del proceso DM del Modelo. Ofreciendo una ventaja con respecto los modelos que se han analizado hasta el momento, principalmente en México.

ABSTRACT

In doctoral thesis, we propose the design Predictive Model of educational data Analysis, Transformation and Interpretation in Higher Education Institutions (PredATIS), for to study the multifactorial problem of school dropout and failing in Academic Virtual Community Early Warning in Educational Indicators called CVAATI. PredATIS is a predictive and adaptive model, with scientific contribution to Data Mining (DM) research field, where methods and techniques of DM classification and Artificial Neural Networks (ANN) are applied, with unsupervised and supervised learning algorithms automatic for the selection and search of better attributes and impact great factors on the phenomenon of study; additionally a technological contribution, by generating a tool that provides useful information, the data exploitation for generate knowledge in educational indicators, report predictive alerts, to establish a strategic communication, that contributes in the decision making at educational environment. In PredATIS design, included methodologies, techniques and algorithms of DM and CVAATI own development was through Software Engineering and WEB tools for data pre-processing, used in the Model DM process. Offering a competitive advantage with respect to the models that have currently been analyzed in Mexico mainly.

INTRODUCCIÓN

Desde su creación en 1999, las comunidades virtuales (CV) han venido evolucionando. Una CV es un grupo de personas que interactúan entre sí por un interés en común o propósito compartido y que su ciclo de vida dependerá de la importancia que le dan los miembros de la organización a las temáticas de la comunidad (Valdivia, 2009). Actualmente, las CV tienen una gran influencia en la forma de comunicarse en una organización en el sector: empresarial, educativo, social, político y cultural (Salinas, 2005). Concretamente, la investigación de esta tesis se centra en el ámbito educativo, debido a que existe un gran volumen de información a través de plataformas educativas, Sistemas Tutores Inteligentes y Sistemas de Información para la recopilación, almacenamiento y análisis de datos personales, económicos, sociales y pruebas o evaluaciones de estudiantes en Instituciones de Educación Superior (IES). Es, en estas instituciones, donde es indispensable contar con una herramienta tecnológica que ayude a prevenir y mitigar el problema de la deserción y reprobación escolar. Al respecto, algunos investigadores han abordado este fenómeno, proponiendo modelos construidos con técnicas, métodos y algoritmos no supervisados de Minería de Datos (DM), específicamente en la Minería de Datos Educativa (EDM) e Inteligencia Artificial (IA), en la mayoría de los casos; dando un aporte científico a estos campos de investigación, en los últimos cinco años (Calders & Pechenizkiy , 2011). Estos trabajos fueron desarrollados y adaptados a un software (SW) específico, entre los más citados son: WEKA (Waikato Environment for Knowledge Analysis) desarrollado en la Universidad de Waikato, RapidMiner, SOAP (Selection of attributes by projections), herramientas de Microsoft® Office, entre otros. Aplicados en todos los niveles educativos y solo pocos estudios para las IES (Amaya, Barrientos & Heredia, 2015; Lanzarini, Charnelli, Baldino & Díaz, 2015; Márquez, Romero & Ventura, 2013, 2015; Formia, 2013;

Treviño, Ibarra, Castán, Laria, & Guzmán, 2013; Fischer E., 2012; Valero, Vargas & García, 2010).

Sin embargo, por el momento estos trabajos no son capaces de contribuir con una predicción lo más cercano posible a la situación de deserción, reprobación y por ende eficiencia terminal, por lo que sigue representando un reto importante y un campo de investigación por parte de las IES como lo señalan Treviño et al. (2013). Adicionalmente, se observa que el tiempo invertido es insuficiente para la generación de estadísticos educativos de forma predictiva y que muchas veces trae como consecuencia, no aplicar acciones estratégicas en el momento preciso y la toma de decisiones efectiva, que coadyuve a este fenómeno, afectando al sector educativo, social y económico Nacional e Internacional. Es, en este sentido, se visualiza un área de oportunidad, en esta tesis se propone un Modelo de Predicción basado Minería de Datos que permita extraer y gestionar información de forma inteligente cambiando el paradigma tradicional de reportes estadísticos educativos, generando conocimiento a través de método y técnicas DM aplicando algoritmos supervisados y no supervisados, que apoye a futuros análisis e incluso hasta una predicción de situaciones particulares de estudiantes de una Comunidad Virtual Académica para mitigar el problema de la deserción y reprobación escolar en IES. El desarrollo de la tesis, se guía con la metodología adaptativa MSWD&DM por sus siglas en inglés (Methodology for Design Software and Data Mining) que combina la Ingeniería de Software (SW) utilizada en la construcción de la CVAATI y las fases del proceso de KDD por sus siglas en inglés (Knowledge Discovery from Databases) en el diseño del modelo PredATIS, donde se pretende implementar a futuro en una IES, para la emisión de alerta temprana en indicadores educativos en un software de diseño propio, que aún los modelos analizados no han cubierto; además de

contribuir en el campo de investigación de EDM en México principalmente. Obteniendo como resultado una herramienta tecnológica que coadyuve a mitigar este fenómeno, sin embargo, es importante señalar que no es una solución a la disminución de los índices de deserción o abandono escolar, ya que se ven influenciados por muchos otros factores que están fuera del alcance de esta investigación. Considerando, finalmente que la CVAATI permitirá proveer información útil y pre procesadas antes de aplicar el modelo DM y RNA, logrando emitir reportes alertas predictivas y complementarias, como puntos clave para identificar de forma temprana los factores y variables de mayor impacto, que influyen en la deserción o reprobación del estudiante, estableciendo una comunicación estratégica entre los académicos de la IES.

CAPÍTULO 1. PROPÓSITO Y ORGANIZACIÓN

1.1 Planteamiento del problema

Actualmente con el crecimiento del Internet, existe un gran número de entornos virtuales como las Comunidades Virtuales (CV), Sistemas de Información, Plataformas Educativas, Sistemas Tutores Inteligentes, video tutoriales, entre otros. Valdivia (2009) define una CV como un grupo de personas que interactúan entre sí por un interés en común o propósito compartido, y que aprenden del otro utilizando para ello los servicios disponibles en la Red. Son organizaciones temporales, cuyo ciclo de vida estará en función de la importancia que le asigne un miembro de la organización a las temáticas que se traten o se discutan. Sus miembros no están obligados a mantenerse en la Comunidad Virtual, lo que se traduce en que con el tiempo el número de participantes disminuya, y pueda por tanto, no seguir cumpliendo con las expectativas del grupo, y con una tendencia a desaparecer. Las CV se han convertido en una herramienta muy útil desde tres perspectivas: - la gestión de procesos educativos a través de plataformas educativas, aulas virtuales, simuladores, video tutoriales entre otros; -la gestión empresarial permitiendo mejorar la dinámica de trabajo interno, las relaciones con sus clientes o incrementar su eficiencia procedimental; y -la función social, que se han convertido en un lugar, en el que el individuo puede desarrollarse y relacionarse con los demás, actuando así como un instrumento de socialización y de esparcimiento como las redes sociales (Facebook, QZone, Google+, linkedIn, Instagram, Twitter, Tumblr, Sina Weibo, Vkontakte, Snapchat), y mensajería en tiempo real (Skype, Windows Live Messenger, Google Talk). Actualmente existe un gran volumen de diferentes tipos de CV que influyen en la forma de comunicarse desde una perspectiva

empresarial, educativa, social, política y cultural, actuando como un instrumento de socialización y de esparcimiento.

Analizando estos entornos virtuales, se observa que, son técnicamente especializados y profesionalmente certificados, desarrollados con tecnología vanguardista en su momento, y actualizando conforme a la dinámica de las Tecnologías de la Información y Comunicación (TIC); representando un gran abanico de cualidades positivas, como relaciones independientes de raza, religión, género o geografía, pueden ocasionar múltiples problemas.

Así como la percepción de riesgo o incerteza en participar en una comunidad, es un fenómeno bastante común en diferentes tipos de CV de aprendizaje (plataformas educativas), de práctica, de conocimiento, además de otros factores que inciden sobre su funcionalidad y desempeño tecnológico, oportunidad/tiempo, seguridad y sociabilidad. (Iriberri & Leroy, 2009: 7)

Agregando los Sistemas de Información (SI) para la administración de datos personales, económicos, sociales, pruebas o evaluaciones del ingreso económico familiar del estudiante. Que son implementados en gran volumen en IES, de forma independiente, con un entorno de inclusión y gestión de contenidos y almacenamiento, y sobre todo usados hacia un propósito en común para el cual fueron creadas; en él que, se emiten reportes estadísticos de sus miembros usuarios a los que están destinados (educandos) y solo proporcionan escasa información al personal académico o directivo que impiden dar seguimiento a la trayectoria escolar del estudiante durante un curso o tiempo de su carrera profesional, es decir son estáticos y tradicionales, no permiten de visualizar de forma automatizada a estudiantes vulnerables a la deserción, reprobación o abandono de su carrera profesional e identificando a través de un

análisis los posibles factores externos (problema familiares, económicos, problemas de salud, entre otros) de mayor impacto que causan un bajo rendimiento escolar. Además de utilizar otras herramientas que requieren de mayor tiempo e inversión de recursos humanos (directivos, administrativos y docente especializado en las temática o asignaturas que se imparten) y materiales (equipos y software adicional). Observando de todo ello, que muchas veces estos reportes son tardíos y no resultan útiles en este fenómeno de investigación. Representando con todo esto, alto impacto que se ve reflejado en los indicadores educativos de las IES y que se van sumando a las estadísticas reportadas a Nivel Nacional sobre el porcentaje global de deserción y/o reprobación, y eficiencia terminal. Como se evidencia en la Tabla 1 según, estadísticas reportadas del Sistema de Información y Gestión Educativa (SIGED), Sistema Nacional de Información Estadística Educativa (SNIEE).

Tabla 1. Estadísticas modalidad escolarizada (cifras preliminares 2013-2014)

Nivel Educativo / Indicador	2012-2013	2013-2014 ^{e/}
	%	%
Educación Superior		
Absorción	85.9	74.8
Abandono escolar	7.2	6.9
Cobertura (Incluye Posgrado) (18 a 23 años de edad) ^{1/}	25.8	26.5
Cobertura (No incluye Posgrado) (18 a 22 años de edad) ^{1/}	28.6	29.4
Cobertura (No incluye Posgrado) (18 a 22 años de edad) ^{1/ 4/}	32.1	33.1

Notas:

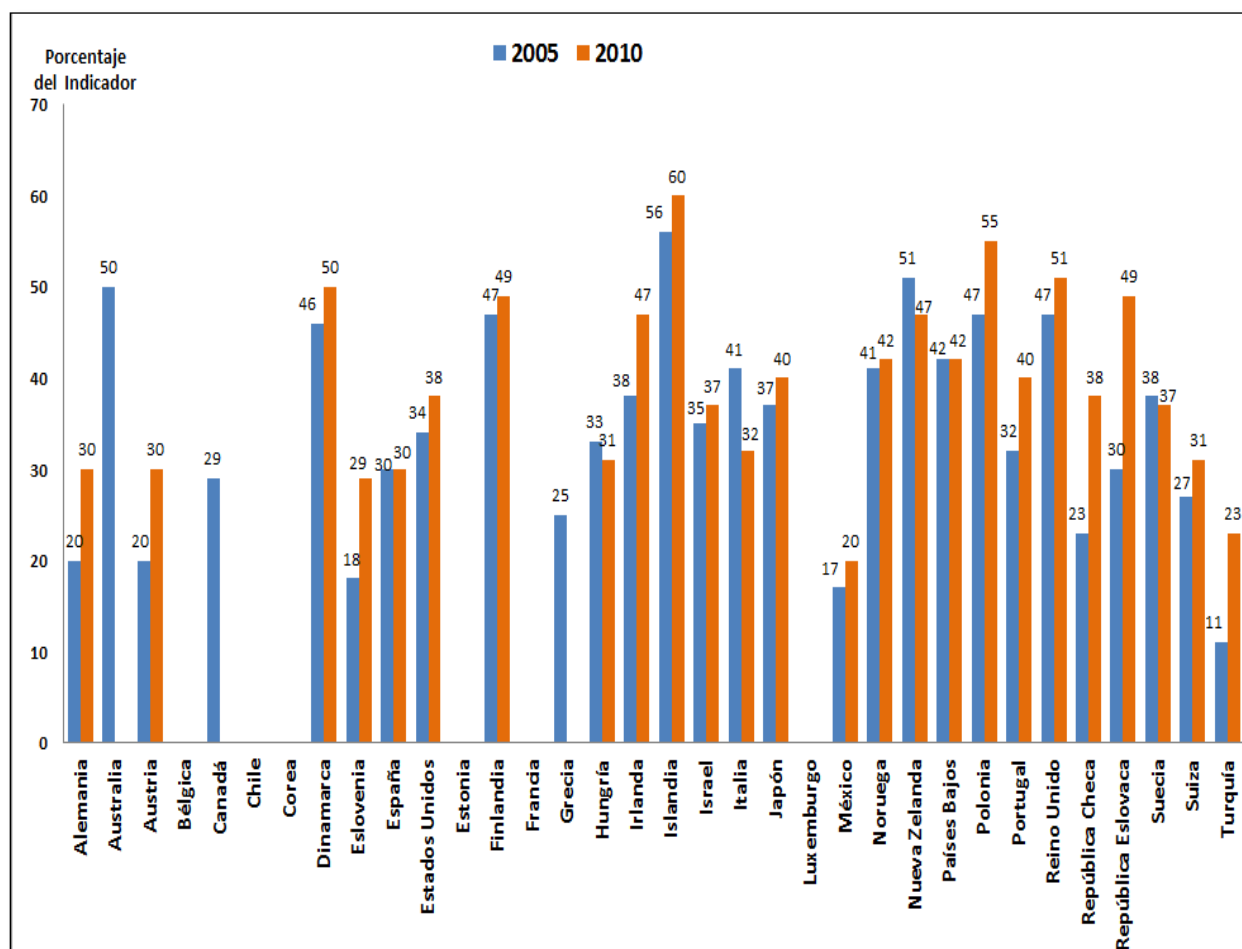
^{e/} cifras estimadas para los indicadores de abandono escolar, reprobación, eficiencia terminal y tasa de terminación.

^{1/} Para los cálculos se utilizaron, proyecciones de población a mitad de año, CONAPO

Nivel Educativo / Indicador	2012-2013	2013-2014 ^{e/}
	%	%
Educación Superior		
2013.		
^{2/} Estimaciones en base al censo de población y vivienda 2010, unidad de medida de grados.		
^{3/} Información al 31 de diciembre de cada año, INEA.		
^{4/} Incluye la modalidad no escolarizada.		

Fuente: Estadísticas (SNIEE y SIGED, 2014).

En la gráfica 1 se presentan las estadísticas e indicadores del año 2005 y 2010 del Nivel Educativo Superior de Licenciatura, de 34 países más desarrollados del mundo y ocho que forman el grupo G20, reportadas por la Organización para la Cooperación y el Desarrollo Económico (OCDE); donde el porcentaje máximo de titulación no es mayor al 60% y en la mayoría de los países están por debajo del 50%; en este sentido se visualiza como un tema de interés Nacional e Internacional, observando principalmente en México que representan tan solo el 17 y 20%, razón por la cual se hace necesaria, analizar este fenómeno de deserción y reprobación escolar, para coadyuvar esta problemática en las IES (Secretaría de Educación Pública (SEP), 2014).



Gráfica 1. Tasa de Graduación del año 2005 y 2010 en Educación Superior de Licenciatura, grupo OCDE.

Fuente: Estadísticas internacionales SIGED y OCDE (2010).

Después del estudio previo y exploratorio dentro del entorno educativo de nivel superior, la investigación se realiza con datos de una muestra representativa de una Universidad Tecnológica del Estado de México, donde se puede observar en la Tabla 2, los bajos índices de egreso de estudiantes en tres periodos escolares, visualizando un escenario con altos índices de deserción y reprobación, a pesar de la aplicación de estrategias educativas que se han venido implementando por el personal directivo, administrativos y docente.

Tabla 2. Estadísticas de indicadores de egreso por cohorte Generacional 2011-2015

Cohorte	Estudiantes que ingresaron en cohorte	Egresados por cohorte	Porcentaje de Egresados (Egresados/ingreso por cohorte)
SEP 2011- AGO 2013	241	73	30.30%
SEP2012 - AGO2014	263	95	36.10%
SEP2013 -AGO 2015	232	81	34.90%

Fuente: (UIPPE, 2015)

Todos los estudios sobre la deserción escolar difundidos hasta ahora, señalan que la deserción es un fenómeno multicausal y complejo, que no puede ser resuelto por políticas o acciones estratégicas educativas tradicionales que atiendan parcialmente el problema (Delen, 2010). Representando con ello un fenómeno a resolver, que se apoyan con otras áreas de las TIC y especialmente de Minería de Datos (DM).

La DM es un área multidisciplinar que actualmente se usa de forma exitosa en la solución de problemas en los campos de la ciencia, la salud, los juegos, los negocios y la educación, etc., donde trata diversos paradigmas de la computación tales como la construcción de árboles de decisión, la generación de reglas de inducción, las redes neuronales, el aprendizaje basado en instancias, los métodos bayesianos o los algoritmos estadísticos (Witten, Frank, & Hall, 2011), y desde hace relativamente poco tiempo también se han utilizado algoritmos de soft-computing como los algoritmos evolutivos. Específicamente existe una aplicación de DM en el campo de la educación, denominada Minería de Datos Educativa (Educational Data Mining, EDM), (Romero & Ventura, 2007, 2010), en ella se han realizado investigaciones en estos últimos años,

sobre análisis de problemas en los sistemas tradicionales de educación para predecir el rendimiento académico de los estudiantes, en los cursos basados en la web, en los sistemas de gestión para el aprendizaje de contenidos, en los sistemas inteligentes de aprendizaje, así como en la predicción del fracaso y abandono escolar en nivel básico, y medio superior (Márquez et al., 2013).

Entre las tareas principales de DM esta la predicción, descripción, segmentación y la asociación. Donde la predicción se puede realizar por la clasificación que es una de las actividades más usada, esta ocurre cuando es necesario asignar un objeto a un grupo predefinido o clase, lo cual se decide basándose en un determinado número de atributos que se observan en el objeto (Zhang, Oussena, Clark, & Kim, 2010). El objetivo de la clasificación es inducir un modelo para predecir a qué clase pertenece un objeto.

Márquez et al. (2013) comentan que la EDM se ha utilizado en estudios exploratorios, transversales o longitudinales que coadyuven el fenómeno educativo. Siendo un área emergente multidisciplinaria que ha venido madurando en estos últimos cinco años y se han presentado en las conferencias EDM, a través de revistas científicas, (Society International Educational Data Mining, 2015), considerada un campo de oportunidad para el profundizar en el desarrollo de investigación en entornos educativos.

Derivado del estudio exploratorio realizado, surge la propuesta de investigación sobre el Diseño del Modelo predictivo “PredATIS” en una Comunidad Virtual Académica de Alerta Temprana en indicadores Educativos (CVAATI), que permita la recopilación de datos desde fuentes implementadas como plataformas educativas, laboratorios virtuales, sistemas de información de estudiantes, etc. y posteriormente aplicar el proceso de DM y EDM para generar

conocimiento que permita importarla a la CVA y a su vez emitir reportes de alerta temprana a los miembros de la misma comunidad académica para aplicar acciones estratégicas oportunas que coadyuven un impacto positivo en los indicadores educativos.

La propuesta del Modelo “PredATIS” se diseñará aplicando las metodologías y técnicas de extracción del conocimiento tanto descriptivas (agrupamiento, correlaciones y reglas de asociación) como predictivas (la clasificación), además de algoritmos de selección y búsqueda de mejores atributos con DM aplicando técnicas de árboles de decisión, sistemas de aprendizaje automático supervisado y no supervisado; la investigación se apoyará con la suite de Minería de Datos WEKA (Calders & Pechenizkiy , 2011) y Matlab para el desarrollo de los algoritmos de aprendizaje con RNA, para realizar pruebas piloto y posteriormente compararlo con la ejecución en el software desarrollado de la CVAATI en una futura implementación que queda fuera del alcance de este trabajo. Esto con la finalidad de proveer información complementaria a la CVA de fuentes de datos existentes de la propia IES en estudio, para apoyar en la identificación de forma temprana las variables y factores personales, académicos, físicos (enfermedades y discapacidad), económicas, familiares, sociales, institucionales, pedagógicas, laborales, adiciones como el exceso de tiempo dedicados a redes sociales, mensajería instantánea, videojuegos, televisión, computadora, consumo de alcohol y el tiempo de dejar de estudiar cuando no alcanzan ingresar a nivel inmediato superior; donde según estudios reportados por Márquez et al. (2013) son causales de incrementar un alto valor porcentual de indicadores educativos; señalando que con estadística básica o tradicional no es fácil de obtener o predecir la situación del estudiante (Calders & Pechenizkiy , 2011). Al finalizar el diseño se pretende, implementarlo con datos de la CVA en una IES del estado de México, dándole un valor agregado

a las CV existentes que la convierte en una Comunidad Virtual dinámica e inteligente para generación del conocimiento entre sus miembros (Directivos, Administrativos, Docentes e Instructores) a través del modelo PredATIS desarrollada en software propio.

En tanto que el diseño de la CVAATI considera aspectos propuestos por (Freud, 2008), basados en una planeación estratégica que sirva de guía para un desarrollo tecnológico; integrando metodologías de la ingeniería de software como es el caso del Proceso Unificado Ágil por sus siglas en inglés Agile Unified Process (AUP), donde se va construyendo, adaptando, modificando y evaluando los componentes funcionales en cada fase e ir refinando antes de ser publicados (Pressman, R., 2005-2006; Someerville, 2005); así como del Modelo Vista Controlador (MVC) y de herramientas Web. Los contenidos propuestos para la Comunidad Virtual Académica CVAATI son:

- Datos de Ingreso del Alumno (datos iniciales personales, notas diagnósticas claves del perfil profesional al cual se proyectan).
- Recopilación de datos complementarios (Tutoría Inicial y Entrevista inicial)
- Kardex del Alumno (en cada cuatrimestre que cursa): Notas parciales y Finales
- Trayectoria académica
- Aplicación de encuestas sobre datos generales, aspectos socioeconómicos, de salud, personales, vocacionales y académicos en cada período escolar.
- Aplicación de modelo de Minería de Datos (KDD) en períodos escolares.
- Reporte de alerta de estudiantes con riesgo a deserción o reprobación de la IES.

Puntualizando que la arquitectura MVC y el framework Responsive web design, separa la lógica de negocio de la interfaz de usuario donde facilita la evolución por separado de ambos aspectos e

incrementa reutilización y flexibilidad de acceso desde cualquier plataforma y tipo de dispositivo. Además ofrece ventajas sobre la optimización del desarrollo, el uso de recursos tecnológicos y aplicación de un modelo de Minería de Datos, permitirán crear una comunidad virtual más dinámica y flexible en la inclusión y gestión de resultados sobre el tiempo de respuesta para el análisis del fenómeno en estudio (Pavón , 2008; López, Martínez, & Moreno, 2003).

1.2 Preguntas de investigación

Las interrogantes que se presentan en esta investigación son:

- ¿Cómo diseñar un modelo predictivo con DM que contribuya en la gestión del conocimiento e información útil de Bases de Datos externas de forma inteligente e implementar pruebas de entrenamiento en una CV Académica que emita avisos de alertas tempranas y seguimiento académico para coadyuvar positivamente en los índices de deserción y reprobación de estudiantes de educación superior?
- ¿La CVA diseñada permitirá generar métodos de extracción de información sobre la base de datos y exportar para su análisis con la idea de generar conocimiento sobre el fenómeno que se aborda, a través de metodologías y técnicas de tratamiento de datos con DM de forma automatizada?

1.3 Propósito de la investigación

El propósito fundamental de la investigación presentada, es diseñar un modelo predictivo “**PredATIS**” tomando como base metodologías y técnicas de DM tanto predictivas (la clasificación) como descriptivas (agrupamiento, correlaciones y reglas de asociación); así como de algoritmos de selección y búsqueda de mejores atributos, aplicando técnicas de árboles de

decisión, sistemas de aprendizaje automático supervisado y no supervisado; donde se utilizarán datos de entrenamiento (training) y datos para prueba (test).

Posteriormente implementarlo con datos de la comunidad virtual académica (CVA) diseñada con base a una planeación estratégica y tecnológica seleccionando la arquitectura de desarrollo MVC y el modelo de desarrollo basado en el Proceso Unificado Ágil (AUP) que es una versión simplificada del Proceso Unificado Racional por sus siglas en inglés Rational Unified Process (RUP). Generando así, la **CVAATI** como una comunidad virtual académica inteligente e innovadora para proveer de información útil y pre procesada, que son la base del procesamiento DM; reportando alertas predictivas, que permita identificar de forma temprana a los estudiantes en riesgo de deserción o reprobación y sus principales causas; esto, con la finalidad de comunicar estratégicamente al personal directivo y académico de la IES para implementar acciones que puedan apoyar a los estudiantes y coadyuvar en la disminución de los estadísticos educativos en estudio.

1.4 Objetivo General

Diseñar un modelo de predicción aplicando metodologías, técnicas y algoritmos de minería de datos en una comunidad virtual académica desarrollada con software propio, que permita generar información de alertas tempranas para identificar a los estudiantes en riesgo a deserción o reprobación en una muestra tomada de una Institución de Educación Superior del Estado de México.

1.5 Objetivos específicos

Para lograr el objetivo general planteado se deben alcanzar los siguientes objetivos específicos:

- Analizar las metodologías de investigación aplicables al diseño del modelo de predicción propuesto en la CVA.
- Analizar y evaluar metodologías, técnicas y algoritmos de DM y RNA aplicables para la creación del modelo de predicción de alerta temprana de deserción o reprobación de estudiantes en IES de la CVA propuesta.
- Analizar los modelos de arquitectura y desarrollo de software (SW) para el diseño y construcción de la CVAATI para el pre procesamiento de datos.
- Diseñar el modelo de predicción, con datos pre procesados de la CVA propuesta, utilizando las técnicas y algoritmos de DM seleccionados y desarrollados, con el software de Matlab y WEKA, para ir evaluando a través de tareas de RNA, clasificación y selección de atributos de mayor impacto en la deserción y reprobación de estudiantes.
- Verificar e interpretar los resultados obtenidos de los experimentos realizados del Modelo PredATIS propiamente desarrollado, en distintos conjuntos de datos de la IES seleccionada, con la finalidad de comparar la precisión calculada de los algoritmos de DM y RNA para la predicción de estudiantes con riesgo a deserción o reprobación e identificación de causas de mayor incidencia de esta investigación e ir aplicando los ajustes antes y después de la ejecución.

1.6 Justificación de la investigación

Actualmente existen Comunidades Virtuales con un potencial tecnológico e innovador, sin embargo están destinadas a un propósito específico como: *la gestión de procesos educativos* (de plataformas educativas, aulas virtuales, simuladores, video tutoriales), *la gestión empresarial*: (dinámica de trabajo interno, las relaciones con sus clientes o incrementar su eficiencia procedimental- CRM, Sistemas de Información, etc), *las redes sociales y mensajería instantánea* (Salinas, 2005). Sin embargo a pesar de contar con un gran volumen de datos con estos entornos virtuales, no es suficiente para el tratamiento de información útil y solo se quedan como un banco de datos que se consultan de forma independiente para resolver una situación educativa al momento, pero no son capaces de generar y extraer conocimiento que apoye, en la toma de decisiones de forma temprana e inteligentemente para la predicción de las trayectorias escolares de los estudiantes en las IES. Haciendo esto necesaria, una investigación en el campo de DM, buscar y aplicar técnicas y algoritmos de clasificación, agrupamiento, arboles de decisión y redes neuronales artificiales (Calders & Pechenizkiy , 2011); para la ejecución de tareas predictivas y descriptivas en el tratamiento de datos como es el estudio del fenómeno (predicción de deserción y reprobación) que tiene como objetivo principal, el diseño de un modelo de predicción y emisión de alertas tempranas utilizando la Minería de datos y RNA que sea posible de implementarlo en un futuro, en una comunidad académica en una IES del Estado de México.

Donde la CVA propuesta pretende enriquecer e innovar el desarrollo tecnológico alineados con el ámbito educativo, la gestión del seguimiento de desempeño académico y la colaboración (Salinas, 2000), a través de reportes de alertas tempranas de avisos sobre

estudiantes en riesgo de deserción o reprobación, en áreas de cursos que se imparten en instituciones de educación superior, optimizando con esto, la gestión de la comunicación entre sus integrantes (Directores, Administrativos y Docentes) para predecir el índice de egresos y tomar medidas o aplicar estrategias una vez que ya se estima un escenario futuro y no esperar hasta que suceda cuando muchas veces ya es tardío y ya no se puede hacer nada al respecto, representando con esto el valor agregado de este proyecto de investigación, en relación a los entornos virtuales existentes de forma independiente para su propósito que fueron creados o aplicados.

Además de estudios realizados a la fecha, se observa que existe investigación abundante de los modelos de DM e IA en otras áreas como la medicina, la robótica, mecatrónica, sistema de toma de decisiones, entre otras, pero no en el desarrollo de software para entornos educativos (Calders & Pechenizkiy , 2011). Así es como esta investigación pretende centrarse en esta área, contribuyendo así, con un aporte de investigación en DM y especialmente en EDM al diseñar y crear un modelo predictivo **PredATIS** aplicando técnicas y algoritmos de este campo, que sea posible aplicarlo a futuro, en una comunidad virtual académica de IES que representaría un desarrollo tecnológico e innovador, no trivial y tradicional que no solo pretende cubrir una necesidad en el ámbito educativo sino facilitar la información de forma inteligente.

El proyecto de investigación se considera viable en el diseño del modelo predictivo e implementarlo con datos de una CV académica y en un futuro aplicarlo a una IES, con la finalidad de cambiar el paradigma tradicional sobre el análisis de datos con herramientas adicionales y que realizan actualmente el personal directivo, administrativo y docente en la en la mayoría de las instituciones de educación superior en México, donde solo, son vistas como

gestores de contenidos de cursos impartidos, comunicación social, reportes de estadísticas con sistemas informático, análisis de resultados de estudiantes de ingreso (psicométricos, nivel de inglés, entre otros); representados con estructuras muy acotadas y diseñadas con la interacción directa con el personal académico, entornos cerrados, grupos limitados, orientados a la obtención de reportes para una situación emergente de la Institución Educativa Superior (IES) pero carece de generar conocimiento relacionado para futuros análisis e incluso hasta una predicción de situaciones particulares de estudiantes de las IES. Además de ser, un área de investigación actual en otras instituciones internacionales como España, Reino unido, EUA., Canadá, entre otros (Siemens, 2013).

1.7 Alcances y limitaciones

Alcances. El modelo predictivo será diseñado con las características y requerimientos para su futura implementación en la comunidad Virtual Académica, en este trabajo solo se pretende extraer datos y contenidos reales hasta donde sea posible dado el nivel de confidencialidad y manejo desde fuentes externas a la IES, sobre entornos educativos como: plataformas educativas, sistemas de información existentes de la Institución Educativa Superior (IES) seleccionada. En el diseño del modelo se utilizarán metodologías, técnicas y algoritmos de DM y RNA seleccionados, para su tratamiento y preparación, donde se pretende evaluar y seleccionar la información útil para efectos de la investigación, mediante un estudio exploratorio y longitudinal utilizando una suite (WEKA o RapidMiner)) de Minería de Datos y Matlab para entrenamiento con RNA, posteriormente realizarlo en la CVAATI desarrollada con Software propio, donde se pretende interpretar los resultados y comparar el nivel de precisión de cada algoritmo de aprendizaje. Todo ello con la finalidad de su futura aplicación en la emisión de

alertas tempranas de estudiantes con riesgo de deserción o reprobación, destinados al personal académico de la IES seleccionada y que esta fuera del alcance del trabajo propuesto en este proyecto de Tesis cuyo centro de interés es el diseño y construcción del modelo.

Limitaciones. El proyecto de investigación es extenso por lo que centra interés primero en el diseño del modelo predictivo aplicando Minería de Datos y RNA, para su posterior implementación con datos de una comunidad virtual académica diseñada y seleccionada, tomando como prototipo su sistema de gestión académica de la IES especificada de solo dos períodos escolares de una generación de un Programa Educativo (PE), que servirá de base sobre los datos de entrenamiento y pruebas en el análisis y evaluación del modelo propuesto. Dejando como trabajo futuro implementar la CVAATI en todos los grupos del programa educativo (PE) del área de Ingeniería de la misma comunidad académica, así como en todos sus PE de la IES y de ser posible en otras Instituciones perteneciente al mismo sistema educativo de nivel superior, interesadas y que está fuera del alcance de esta investigación.

1.8 Organización del estudio

El documento de tesis que se presenta, se divide en los capítulos que se describe a continuación:

- El capítulo I *Propósito y Organización* aborda la problemática, la justificación, objetivos, preguntas de investigación, alcances y limitaciones, así como la viabilidad del proyecto de investigación.
- En el Capítulo 2 *Marco Teórico* se presenta la revisión de la literatura sobre indicadores educativos, causas de deserción escolar y sistemas de alerta temprana, comunidades virtuales para establecer comunicación estratégica, así como las estrategias tecnológicas, metodologías

y modelos de ingeniería de software útil en el desarrollo de proyectos en entornos virtuales educativos, además se describen bases teóricas de metodologías y técnicas Minería de Datos(DM), de Minería de Datos aplicada a entornos educativos (EDM) y de Redes Neuronales Artificiales (RNA), adicionalmente se presenta una revisión de modelos de predicción de trabajos realizados en el campo de estudio sobre indicadores educativos.

- El capítulo 3 *Metodología* describe los fundamentos y el diseño de la investigación para el desarrollo del modelo predictivo propuesto y la metodología de desarrollo propiamente de la CVAATI como una de las aportaciones de este trabajo.
- El capítulo 4 *Desarrollo del Modelo propuesto en la CVA* presenta la metodología y arquitectura del diseño y construcción de la CVAATI y el diseño del modelo predictivo PredATIS propuesto en esta investigación a través del metodologías y técnicas de clasificación y selección DM y algoritmos de RNA.
- En el capítulo 5 *Modelado de PredATIS en la CVAATI* se describe el desarrollo e implementación del modelo PredATIS en la CVA donde se muestran los resultados obtenidos de la ejecución y aplicación de algoritmos de DM y RNA en los experimentos realizados de las muestras tomadas en la IES en estudio.
- El capítulo 6 *Análisis y Discusión* se presenta el análisis y discusión de resultados obtenidos del modelo PredATIS con el software desarrollado en Matlab y WEKA para la ejecución del entrenamiento en comparación con el modelado de la CVA, que se pretende implementar a futuro. Centrando la investigación en la precisión de resultados sobre predicción de causas de mayor impacto y la identificación de estudiantes en riesgo de deserción o reprobación.

- Finalmente se presentan las Conclusiones, Recomendaciones, Trabajo Futuro y Divulgación que resumen las aportaciones logradas en este proyecto de tesis, expone las líneas de mejora del trabajo presentado y perspectivas futuras, así como las divulgaciones de los resultados obtenidos.

CAPÍTULO 2. MARCO TEÓRICO

2.1 Antecedentes relacionados con estadísticos educativos

2.1.1 Indicadores estadísticos educativos.

La Educación es una estrategia que apoya a fomentar el desarrollo social y económico de un país, por lo que la población debe tener acceso, ya que sin ella no podrá tener un nivel educativo en crecimiento y el desarrollo personal y profesional del individuo por lo que se ve limitado, y como consecuencia impide un desarrollo de la sociedad. Comparando esta directriz con lo que sucede en México principalmente, se observa en la Tabla 3, que las estadísticas reportadas por el SIGED, SNIEE y la OCDE (SEP-SIGED, 2014; SEP-SNIEE, 2014) sobre altos indicadores de deserción, reprobación, eficiencia terminal son bajos en todos los niveles educativos, razón por la cual esta investigación se centra en el nivel de educación superior.

Tabla 3. Estadísticas de indicadores educativos en México 2012-2014

Nivel Educativo / Indicador	2012- 2013	2013-2014 ^{e/}	
	%	%	Posición
Educación Básica			
Cobertura (3 a 14 años de edad) ^{1/}	92.1	92.8	24
Tasa Neta de Escolarización (3 a 14 años de edad) ^{1/}	91.0	91.7	24
Educación Preescolar			
Atención de 3 años ^{1/}	23.6	25.6	22
Atención de 4 años ^{1/}	80.3	81.5	25
Atención de 5 años ^{1/}	83.7	83.3	18
Atención de 3, 4 y 5 años (Tasa Neta de Escolarización) ^{1/}	62.6	63.5	27
Cobertura (3 a 5 años de edad) ^{1/}	62.8	63.6	27
Educación Primaria			
Abandono escolar	0.5	0.3	11
Reprobación	0.3	0.1	03
Eficiencia Terminal	98.2	99.1	09

Nivel Educativo / Indicador	2012- 2013	2013-2014^{e/}	
	%	%	Posición
Tasa de Terminación	105.1	105.3	20
Cobertura (6 a 11 años de edad) ^{1/}	106.2	105.6	23
Tasa Neta de Escolarización (6 a 11 años de edad) ^{1/}	99.3	98.9	15
Educación Secundaria			
Absorción	96.0	96.2	23
Abandono escolar	3.2	3.1	05
Reprobación	4.4	3.7	06
Eficiencia Terminal	89.5	90.9	05
Tasa de Terminación	84.0	86.8	08
Cobertura (12 a 14 años de edad) ^{1/}	93.4	96.2	16
Tasa Neta de Escolarización (12 a 14 años de edad) ^{1/}	81.0	83.7	20
Educación Media Superior			
Absorción	92.6	93.7	29
Abandono escolar	15.0	15.5	23
Reprobación	9.8	9.0	02
Eficiencia Terminal	63.1	61.6	23
Tasa de Terminación	45.7	44.7	29
Cobertura (15 a 17 años de edad) ^{1/}	60.1	62.4	29
Tasa Neta de Escolarización (15 a 17 años de edad) ^{1/}	47.9	50.1	27
Educación Superior			
Absorción	86.5	77.0	17
Abandono escolar	8.0	7.9	14
Cobertura (Incluye Posgrado) (18 a 23 años de edad) ^{1/}	20.8	21.9	24
Cobertura (No Incluye Posgrado) (18 a 22 años de edad) ^{1/}	23.4	24.6	24
Cobertura (No Incluye Posgrado) (18 a 22 años de edad) ^{1/ 4/}	25.1	26.6	25
Otros Indicadores			
Grado Promedio de Escolaridad ^{2/}	9.3	9.3	11
Analfabetismo ^{3/}	3.9	3.8	13

Fuente: Estadísticas de SNIEE y SIGED (2014).

Indicadores. Un indicador educativo como lo define la SEP-SIGED (2014), es un “instrumento que nos permiten medir y conocer la tendencia o desviación de las acciones educativas, que determinan el éxito o grado de avance de otras metas del sistema educativo y registrando mediciones sistemáticas en cada inicio y fin de cursos” (Sección de Estadística - Indicadores, párr. 1).

Los tres indicadores más representativos según reporte de deserción escolar en México (Atlantia Search, 2014) para evaluar la eficiencia del sistema educativo son la deserción, la reprobación y la eficiencia terminal, que se encuentran relacionados, al momento de analizar las posibles causas y consecuencias del desempeño de los estudiantes y el rol del propio sistema educativo.

- *La deserción* expresa el número o porcentaje de estudiantes que abandonan las actividades escolares antes de terminar algún grado o nivel educativo. La deserción se clasifica en tres vertientes:
 - a) Deserción intracurricular: cuando el estudiante abandona durante el ciclo escolar.
 - b) Deserción intercurricular: el estudiante abandona al finalizar el ciclo escolar, independientemente de que haya aprobado o no.
 - c) Deserción total: es la combinación de ambas.
- *La reprobación* es la proporción de estudiantes que finalizaron el ciclo escolar pero que no cumplieron con los requisitos para ser promovidos del grado o nivel educativo que finaliza. Los datos necesarios para el cálculo de éste indicador son la existencia y número de alumnos aprobados por grado para cada ciclo escolar. Con la siguiente fórmula:

$$\text{Índice de reprobación} = \frac{\text{número de estudiantes reprobados en evaluación ordinaria o extraordinaria}}{\text{Matricula inicial atendida en el cuatrimestre}}$$

Matricula inicial atendida en el cuatrimestre

- *La Eficiencia terminal* permite conocer el número de alumnos que termina un nivel educativo de manera regular (dentro del tiempo ideal establecido) y el porcentaje de alumnos que lo culminan extemporáneamente. Es la relación porcentual que resulta de dividir el número de egresados de un nivel educativo determinado, entre el número de estudiantes de nuevo ingreso que entraron al primer grado de ese nivel educativo años antes.

$$\text{Índice de et} = \frac{\text{número de egresados de un nivel educativo determinado}}{\text{número de estudiantes de nuevo ingreso que entraron al primer grado de ese nivel educativo años antes}}$$

Hablar de abandono escolar es referirse a un problema presente en todos los grados escolares, lo cual conlleva al rezago de la población en esta materia y, con ello, al atraso social y económico. En nuestro país se han realizado múltiples estudios sobre el abandono escolar, por lo general, de carácter cuantitativo, que dan a conocer tasas de deserción, montos de población en rezago, etc., y se carece de investigaciones de tipo cualitativo, que expliquen las causas de la deserción escolar, con las que se puedan establecer políticas educativas más eficientes y dirigidas a grupos específicos de la población. Según investigaciones cualitativas emitidas por instancias oficiales de la SEP, el Instituto Nacional de Estadística y Geografía (INEGI), la Asociación Nacional de Universidades e Instituciones de Educación Superior (ANUIES) y el Instituto de la Juventud (INJUVE) sobre deserción escolar en los diversos grupos de edad de los jóvenes con que trabajan, además de complementar con investigaciones independientes de académicos publicadas a nivel Iberoamérica o América Latina, y que son, en su mayoría, fuentes de referencia de las mismas instancias oficiales de México. Además de publicaciones, artículos y estudios de otros organismos oficiales que brindaban información de México en materia de Educación, como la Comisión Económica para América Latina y el Caribe (CEPAL), la Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura (OEI), la Organización para la Cooperación y el Desarrollo Económicos (OCDE), la

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) y el Fondo Internacional de Emergencia de las Naciones Unidas para la Infancia UNICEF (Atlantia Search, 2014; SEP-SIGED, 2014; SEP-SNIEE, 2014).

Por otro lado, los datos cuantitativos que brindan un amplio panorama de las tasas de deserción y sus características particulares según sexo, geografía, nivel económico y otros rasgos sociales, los muestra el Instituto Nacional de Evaluación Educativa (INEE), órgano encargado de compilar información de la Dirección General de Planeación (DGP) del Sistema Nacional de Educación (a través del formato 911), de los Censos del INEGI y las proyecciones de población del Consejo Nacional de Población (CONAPO), y de esta forma evaluar la calidad, el desempeño y los resultados del Sistema Educativo Nacional en la educación preescolar, primaria, secundaria y media superior en nuestro país (SEP-SIGED, 2014).

Actualmente el fenómeno de deserción y reprobación escolar en los diferentes niveles educativos básico, medio y superior; es un tema que se está abordando en muchos países, con el objeto de determinar los múltiples factores que influyen en él (Álvarez, 2009), además del rendimiento académico de los estudiantes (Araque, Roldán, & Salguero, 2009). Valero (2009) menciona que se encuentran entre los problemas más complejos y frecuentes en las IES de México.

La Subsecretaría de Educación Media Superior (SEMS) expone que a pesar de que se han venido implementando programas o proyectos a nivel nacional en las instituciones educativas para la obtención de recursos extraordinarios, y que son empleados conforme lo especificado por la SEP en rubros para mitigar este fenómeno con la finalidad de reducir el

abandono escolar, siendo uno de estos, el programa federal PAAGES (Proyecto para el Avance en la Autonomía de la Gestión Escolar), observando que los estudios de este organismo solo está dedicado al nivel educativo medio superior y no al nivel superior, siendo una razón más que motivan a la investigación de este trabajo.

En las IES por ejemplo se han preocupado por disminuir estos índices a través de programas de tutorías, asesorías especializadas, talleres, congresos, entre otros, con la intención de involucrar directamente y aumentar el compromiso del estudiante; sin embargo esto, no es suficiente y se repite en cada ciclo escolar (Valero S. , 2009), principalmente en los primeros semestres de su estudios universitarios, donde se señala que este elevado fracaso se ve impactado en la mayoría de los casos por aspectos sociales, económicos y humanos que la sociedad no debe ignorar (Más-Estellés, y otros, 2009).

2.1.2 Causas de deserción y reprobación escolar en los distintos niveles educativos.

Márquez et al. (2013) resumen en la Tabla 4 las principales categorías en: Personales, Académicas, Físicas, Económicas, Familiares, Sociales, Institucionales, Pedagógicas, Laborales, Adiciones y otras; describiendo cada una de las variables y factores que influyen en el estudiante al abandono o fracaso escolar en diferentes niveles educativos.

Tabla 4. Variables o factores que influyen en los estudiantes de deserción o reprobación escolar

Categorías	Variables/Factores
Personales	Edad, género, población, estado, personalidad, importancia que le da a su educación, tiempo dedicado a estudiar, estado civil, actitudes, valores, reingreso a la IES, problemas personales (noviazgo, matrimonio o embarazos no deseados).

Categorías	Variables/Factores
Académicas	Bajo Promedio de bachillerato, bajo puntaje de CENEVAL, pocas habilidades en asignaturas de Razonamiento Lógico-Matemático y Programación, bajo nivel de dominio del idioma inglés, bajas calificaciones o reprobación en más del 20% de asignaturas de cada periodo escolar, bajo coeficiente intelectual, no es su primera opción de estudio seleccionado.
Físicas	Enfermedades, discapacidad, problemas visuales o auditivos.
Económicas	Bajo nivel socioeconómico o de ingreso familiar, no contar con una beca.
Familiares	Problemas familiares frecuentes, bajo apoyo afectivo de los padres al estudiante, deficiente nivel cultural de la familia, número de hermanos del estudiante, orden de nacimiento entre los hermanos, nivel de educación de los padres, actitud orientadora de los padres hacia el trabajo escolar, ausencia de material de apoyo a la educación en casa.
Sociales	Estrato social, marginación, número de amigos en el aula, pocas relaciones interpersonales y de trabajo en equipo.

Categorías	Variables/Factores
Institucionales	Nivel de satisfacción del estudiante con la institución, requisitos de ingreso, número de oportunidades para aprobar una asignatura, número permitido de asignaturas reprobadas, tipos de exámenes y número de ocasiones que se pueden presentar, plazo para concluir los estudios, grupos con más de 40 alumnos por aula, maestros aspirantes o interinos, baja calidad académica de los profesores, escuelas con población estudiantil muy grande, infraestructura en malas condiciones o insuficiente, una gran carga académica para los estudiantes.
Pedagógicas	Control de Locus académico (¿qué controla el desempeño Académico del estudiante?, ¿dónde está?), tipo de carrera a estudiar, estilo de aprendizaje, problemas específicos de aprendizaje, carencia de métodos efectivos para el aprendizaje.
Laborales	Si trabaja tiempo semanal dedicado al empleo.
Adicciones	Exceso de tiempo dedicado a la Televisión, videojuegos o juegos en computadora, acceso a redes sociales, mensajería instantánea y el consumo de alcohol y/o drogas.
Otras	Retrasar el ingreso a la escuela, permanecer un periodo de tiempo.

Fuente: Elaboración propia a partir de Márquez et al. (2013).

Magaña (2002) menciona que estos factores y variables tienen una gran influencia en el rendimiento académico del estudiante de las IES, denominando a este fenómeno como “el problema de las mil causas”, derivado de este estudio, se argumentaría, que se requiere de un análisis multifactorial, agregando que aquí es donde se hace útil y se apoya en áreas de descubrimiento del conocimiento como Minería de Datos (DM) e Inteligencia artificial, entre otras.

2.1.3 Sistemas de Alerta Temprana (SIAT) en indicadores educativos.

Un Sistema de Alerta Temprana (SIAT), según Graso (2012), es una buena solución para detectar lo más temprano posible a los estudiantes con alto riesgo de abandonar, es cualquier sistema diseñado para alertar sobre un riesgo potencial, su propósito es prevenir que se presente el problema y pueda ser peligroso (UNEP, 2012). Esta definición se deriva de los diferentes tipos de SIAT, los cuales se han usado donde la detección es muy importante, por ejemplo en ataques militares, en la prevención de conflictos, en las crisis económicas/bancarias, en ambientes de peligro/desastres, en epidemias tanto humanas como de animales, en los sismos, etc.

Esta investigación se centra en un entorno educativo, definiendo un SIAT como un conjunto de procedimientos e instrumentos que se utilizan para detectar de manera temprana las características o atributos que tienen los estudiantes en riesgo de abandonar la escuela, además involucra la implementación de intervenciones apropiadas para hacer que los estudiantes permanezcan en la escuela (Heppen & Bowles, 2008). Según la SEMS en México, un SIAT es un conjunto de procedimientos e instrumentos automatizados que permiten, por una parte, detectar oportunamente a alumnos de educación media superior que están en riesgo de abandonar

los estudios y, por otra, poner en marcha con la debida oportunidad, las intervenciones adecuadas para lograr su permanencia en la escuela (SEMS, 2013).

Las características o atributos a detectar por un SIAT son aspectos del rendimiento académico de los estudiantes, los cuales pueden reflejar de forma precisa y oportuna el riesgo de abandonar de cada estudiante. Pero detectar estos indicadores es difícil, porque no existe una única razón o motivo por la que los estudiantes abandonen, ya que se sabe que es un problema multifactorial como se ha descrito anteriormente. Por tanto es común observar todos los factores que pueden afectar en el rendimiento académico de los estudiantes hasta que estos han abandonado la escuela.

2.2 Comunidades Virtuales para establecer una comunicación estratégica

2.2.1 Comunidad virtual.

Definir una comunidad virtual no es una tarea fácil, y al paso del tiempo varios autores de diferentes disciplinas han realizado estudios y analizado este concepto, donde cada uno centra su propia definición según su entorno al cuál están destinados, en los siguientes párrafos se presentan definiciones al respecto.

Desde el año 1985 en Sausalito, California, surge la primera comunidad WELL *Whole Earth 'Lectronic Link*, creada por Stewart Brand, editor de la Revista Whole Earth Catalog (The Well Group, Inc, 1985) y formada por un grupo de personas motivadas por la Tecnología, cuyo acceso estaba limitado personajes del ámbito científico y a los expertos en informática, entre sus miembros principales destacan Kevin Kelly(Editor de la revista Wired) y Mitchell Kapor (fundador de Lotus), evolucionando desde un Sistema de Tablón de Anuncios BBS (Bulletin

Board System) hasta convertirse en uno de los primeros proveedores de servicios de Internet ISP sobre los años 90's, y es en donde las comunidades crecen de forma exponencial, gracias al surgimiento de la World Wide Web, correo electrónico, chats, mensajería instantánea entre otros; convirtiéndose en comunidades virtuales accesibles al público en general, e integrándose por grupos de personas que se reúnen para un interés común y que se comunican por medio de estos servicios, donde está ausente el espacio físico, la transcendencia de las fronteras del tiempo y de espacio.

Rheingold (1996), es el primero en definir una Comunidad Virtual, como “agregados sociales que surgen de la red, cuando una cantidad suficiente de gente lleva a cabo estas discusiones públicas durante un tiempo suficiente, con suficientes sentimientos humanos como para formar redes de relaciones personales en el espacio cibernético” en (Sección de Introducción, párr. 24). Además el autor afirma que los que participamos en las comunidades virtuales hacemos “prácticamente todo lo que hace la gente en la vida real, pero dejando atrás nuestros cuerpos” en (Sección de Introducción, párr. 26). Actualmente esta comunidad sigue en línea con 4000 miembros a través de foros, email y cuentas shell y páginas web sobre temas que dependen de la naturaleza e intereses de los participantes desde San Francisco, EUA.

Por otra parte, Meirinhos & Osorio (2009) comentan que las comunidades virtuales se convierten en un medio para compartir y construir conocimiento donde se incluyen las siguientes características:

- Compartir intereses y objetivos
- Actitud de participación activa
- Acceso a recursos
- Reciprocidad de información

- Los miembros comparten unas convenciones o reglas.

El autor argumenta que es más frecuente relacionar el concepto de comunidad dentro del contexto educativo, donde permite identificar los diferentes tipos de comunidad y sus formas de interacción para facilitar el desarrollo de metodologías e instrumentos que soporten de modo eficaz las actividades planificadas según el tipo de comunidad. Identifica los siguientes tipos de comunidad:

- a) *Comunidad Online*: Plant (2004: 54) define como “un grupo colectivo de entidades, individuos u organizaciones que permanecen juntos temporalmente o permanentemente a través de un medio electrónico para interactuar recíprocamente en un espacio común sobre algún problema o interés”.
- b) *Comunidades de interés*: sus miembros son provenientes de culturas y medios variados y persiguen objetivos que no se integran en una dinámica colaborativa, en el sentido de haber involucramiento cognitivo para la producción conjunta de conocimiento. Su dinámica se encuadra más en el campo de la ayuda para la resolución de problemas individuales. La interacción tiene mucho más a ver con la resolución de problemas personales que grupales (Fischer & Ostwald, 2005).
- c) *Comunidades de interés inteligente*: se forma en torno a un interés común, como la definición o realización de un proyecto, con base en un grupo task-force, a veces para funcionar en un tiempo determinado. Entre sus miembros predomina una heterogeneidad de conocimientos y competencias (Meirinhos & Osorio, 2009).
- d) *Comunidades de aprendizaje*: emergen en contexto institucional y son constituidas por miembros (alumnos, profesores) de una o de varias instituciones que pueden ser

geográficamente dispersas. Son más direccionadas para contextos académicos de aprendizaje y formación. Sin embargo, tienen muchos puntos comunes con las denominadas comunidades de práctica (Meirinhos & Osorio, 2009).

Valdivia (2009) considera que las CV son un grupo de personas que interactúan entre sí por un interés en común o propósito compartido, y que aprenden del otro utilizando (Millington, 2008) para ello los servicios disponibles en la Red.

2.2.2 Aspectos relevantes para la creación de una Comunidad Virtual.

Para lograr la sostenibilidad de una CV, se debe seguir un proceso de gestión hacia el crecimiento y debe partir del reconocimiento de las relaciones que se establecen dentro de la organización y solo es posible desde las acciones cotidianas, y la forma de relacionarse y comunicarse en una red (CV). Además se debe considerar el público y la intencionalidad, al momento de gestionar un entorno, en donde no todas las estrategias y decisiones se encuentran en el mismo nivel; entonces no es lo mismo la definición de una estrategia de TIC en un entorno e-learning corporativo que el diseño de un proceso de un modelo de negocios (Salinas, 2005).

Tomando como referencia los siguientes autores, se puede decir que, en la creación de una CV se requiere saber los aspectos relevantes como se indica a continuación:

- Millington (2008), argumenta que como primera instancia se debe averiguar lo que realmente importa a los miembros de la comunidad, cuestionarse sobre: ¿Qué imagen tratan de proyectar a los demás?, ¿Acerca de qué hablan en una conversación online?, ¿Quiénes y cuáles son los grandes factores de influencia?, ¿Qué tecnología utilizan?, ¿Cuáles son sus personalidades y motivaciones para participar de forma virtual?

- Además de incluir las siguientes cuestiones ¿Cuál será el enfoque de su comunidad?, ¿La comunidad va a ser activista o pacifista?, ¿Va a centrarse en asuntos concretos dentro de las regiones temáticas o geográficas?, ¿En cuánto tiempo se puede poner en marcha?

Salinas (2005) indica que debe incluir tres funciones para lograr la intencionalidad de la CV para que tenga un enfoque educativo, de aprendizaje y formativo:

- Función pedagógica: Distribución de materiales, Comunicación e interacción, Situaciones comunicativas, Gestión de los espacios de comunicación.
- Función organizativa: Marco institucional, Estrategia de implementación, Contexto.
- Tecnología apropiada: Tecnología física, Herramientas, Sistema de comunicación, Infraestructura, Infoestructura.

Estas tres funciones dan lugar a tres estrategias de gestión, que a su vez marcan las etapas de crecimiento de la comunidad.

- Estrategias de introducción y/o implementación: se incluye la gestión de los procesos de política institucional, de análisis del contexto, de implementación, dirigidos a la definición y puesta en marcha de un proyecto de e-learning, b-learning o la utilización de TIC.
- Estrategias de implantación y diseminación en la institución: aquí se sitúa la gestión del entorno virtual que hace referencia al proceso de convertir el e-learning o b-learning en parte de la cultura institucional.
- De práctica y experiencia diaria dirigida a seleccionar la más adecuada combinación de métodos, medios y técnicas que ayude al miembro a alcanzar la meta deseada del modo más sencillo y eficaz. En otras palabras, diseñar y ejecutar estrategias didácticas.

Cely (2014) señala que el crecimiento de la comunidad se debe medir en tres indicadores de éxito que son: *el crecimiento, el compromiso y el sentido de comunidad*, como lo propone Millington y que describe a continuación:

- Crecimiento: número de miembros que han hecho una contribución en el último (mes o semana).
- Compromiso: capacidad de respuesta, la participación, etc. número promedio de contribuciones por miembro activo, número promedio de las respuestas a una discusión, tiempo medio para recibir respuesta y número medio de visitas por miembro activo.
- Sentido de comunidad: Teniendo en cuenta 4 factores claves: lista de miembros, influencia, la integración y cumplimiento de las necesidades y conexión emocional compartida.

Cely (2014: 33) propone una fórmula para evaluar el impacto de la comunidad virtual, donde “recoge el impacto reproductor de buenas y malas experiencias a partir del tamaño de la red social de cada individuo y el nivel de influencia sobre sus seguidores y amigos”.

$$(CLV) = CLV * i * n$$

donde **CLV** equivale a el valor de vida del cliente, **i** es el nivel de influencia; **n** hace referencia al tamaño de la red. Argumentando que “Al valor del ciclo de vida del cliente se debe agregar el impacto de otros a través de la red social” Cely (2014: 33).

Además Cely (2014: 34) agrega las siguientes recomendaciones tomando como referencia las que señala Millington (2008):

- Programar reuniones de seguimiento como hábito desde el principio de la comunidad.

- Una vez que se tenga una base constante de miembros activos (alrededor de 100), se puede comenzar a promocionarlo.
- Pasar tiempo hablando con los miembros de forma individual, planear eventos, escribir de forma colaborativa sobre temas relevantes y discutir con las personas de la comunidad para asegurar que los usuarios hablan de los logros obtenidos en la comunidad. Poco a poco, la gente empezará a oír hablar de la comunidad y a unirse.
- Para mantener la participación alta, se deben resaltar los temas más comunes y permitir que se creen otras áreas dentro de la comunidad.
- Seleccionar las personas más activas para que se hagan cargo de estos nuevos subgrupos.
- Invitar a oradores a unirse a la comunidad para discutir temas, trabajar en una causa o logro al que la comunidad puede aspirar.

Como complemento a las sugerencias anteriores, Leimeister, Ebner, & Krcmar, (2005) mencionan que el apoyar, crear y mantener la confianza entre los miembros y los operadores de la CV, contribuyen al mantenimiento del éxito de la comunidad.

Fonseca (2008) agrega que la conformación de una CV debe apuntar a que los usuarios, donde la participación de quienes conforman la comunidad deberá ir más allá de la interacción con herramientas tecnológicas y servicios de comunicación electrónica para convertirse en procesos de comunicación activa que permita a quien conforma una comunidad virtual ser autor, coautor y par. Es decir ver al usuario no sólo como consumidor, sino que es productor y consumidor de los contenidos que se proponen en la comunidad.

Otro aspecto importante a incluir en una CV es la construcción del espacio virtual en el que se va a desenvolver.

En la actualidad existen arquitectos de la información que diseñan y crean edificios digitales, conciben espacios tridimensionales, pretenden definir, analizar y volver a definir entornos tan dinámicos como los digitales, organizados y gestionados como sistemas a ser utilizados por comunidades de usuarios en diferentes contextos sociales y económicos. (Cely, 2014: 36)

Por tal razón, al abordar cualquier estudio de los espacios electrónicos y el diseño conceptual de una comunidad virtual, es necesario tener presente los espacios, el esquema de organización, funcionamiento y navegación del sitio como lo menciona Zea & Atuesta (2007).

Finalmente se agregan las siete dimensiones señaladas por Zachman a considerar y vincular en la creación y administración de una CV:

- “Dónde: espacio – sentido de lugar.
- Quiénes: actores/miembros – sentido de comunidad.
- Para qué: interacciones – relaciones, aprendizaje, sentido de comunidad.
- Cuándo: frecuencia – sentido de pertenencia.
- Qué: funcionalidad – actualidades, noticias.
- Cómo: pedagogía/didáctica: estrategia de aprendizaje.
- Por qué: demanda social “ (Cely, 2014: 36)

2.2.3 Comunidad Virtual Académica (CVA).

Es claro que la incorporación de las TIC en la educación, como la creación e implementación de CV, SIAT, entre otros, representan un desafío que forma parte de un proceso con repercusiones que afectan a las distintas dimensiones de la sociedad, donde las Instituciones educativas de nivel superior tienen un rol importante en este impacto con el nivel de competencias y sus indicadores educativos que reflejan en cada ciclo escolar.

Después de una revisión de la literatura respecto a los diferentes tipos de comunidades virtuales, sistemas de alerta temprana y con base en la definición de Meirinhos & Osorio (2009); se procede a definir una *comunidad virtual académica de alerta temprana* como entorno virtual integrado por miembros de una IES (Directores, Administradores, Docentes e Instructores) para compartir intereses y objetivos, participación activa, acceso a recursos, interacción de datos manejados, reciprocidad de información y construir conocimiento a través de datos fuentes disponibles; con la finalidad de detectar y analizar de forma temprana los factores o variables que tienen los estudiantes en riesgo de abandonar sus estudios profesionales, aplicando acciones estratégicas implementadas en la propia Institución Educativa (sistemas de tutorías, asesorías de nivelación, académicas, de atención personalizada y padres de familia, etc.).

Así es, como la investigación centra su interés, no solo mantener los contenidos y obtener estadísticos estáticos de aplicaciones de Tecnologías de la Información (TI) desarrolladas, reportes de Excel que en la mayoría se manejan, sino ir más allá de una simple *gestión tradicional a la gestión inteligente*, es decir convertir a un escenario dinámico integrando las TIC innovadoras, aplicando modelos con métodos y técnicas de DM, IA y Sistemas Expertos (SE),

Sistemas Inteligencia de Negocios (SIN), entre otros. Esto con la finalidad de generar conocimiento y explotar la información útil para contribuir, siendo este el principal objetivo de nuestra investigación, de proveer de información relevante para su tratamiento y generación de conocimiento y establecer una comunicación estratégica entre el personal académico de la IES en estudio, al emitir alertas tempranas de estudiantes con alto riesgo de deserción o reprobación.

2.3 Estrategias Tecnológicas

Villagómez (2011) define la “Estrategia Tecnológica como el proceso de adopción y ejecución de decisiones sobre las políticas, estrategias, planes y acciones relacionadas con la creación, difusión y uso de la tecnología” (Sección 2, párr. 1). El concepto de estrategia tecnológica “es más amplio que el de investigación y desarrollo (I+D) tradicional. Comprende no sólo la investigación y desarrollo de nuevos productos y procesos, sino que su acción debe extenderse a todas las funciones o subsistemas de una organización” (Sección 2, párr. 2).

Desde hace más de dos décadas, la (I +D) especialistas en el crecimiento económico, pero la gestión de la tecnología y su inclusión en la estrategia de las organizaciones, no ha sido del todo satisfactoria en su totalidad. Actualmente en muchas organizaciones la tecnología se gestiona al más alto nivel, como una variable estratégica principal para el desarrollo de dichas organizaciones y por ende sus procesos operativos cotidianos. Considerando de vital importancia que las organizaciones elaboren una estrategia tecnológica, que sea siempre original, y por tanto, nunca podrá ser el resultado de la aplicación de una receta determinada (Cendejas, 2014). El autor sugiere incluir las experiencias empresariales tales como:

- a) “La importancia del uso de herramientas que sirvan para, suscitar ideas y promover la discusión en las etapas previas” (Cendejas, 2014:10).
- b) “La necesidad de que la estrategia tecnológica debe elaborarse conjuntamente con la estrategia global de la organización, mediante un proceso iterativo que se dirija a la formulación simultánea” (Cendejas, 2014:11).
- c) “Plasmar la estrategia tecnológica en un plan de desarrollo tecnológico que muestre con claridad las opciones efectuadas” (Cendejas, 2014:11).

La toma de decisiones organizacionales, se ven reflejadas y llevadas a cabo con base al plan estratégico de la misma, analizando y cuestionando lo siguiente como lo señala Cendejas (2014:11) y adaptandolo en el diseño de una CV desde dos perspectivas, una esta relacionada con la TI y la otra referida a la funcionalidad de la CV propuesta:

- ¿Cómo evoluciona el entorno virtual?
- ¿En qué entorno virtual se habrá de competir en el futuro?
- ¿En qué tipo de organizaciones se presentan oportunidades de éxito de una CV?
- ¿En qué estado se encuentran nuestras tecnologías propuestas?
- ¿Están cerca de sus límites físicos?
- ¿Qué nuevas tecnologías pueden tener impacto en nuestra cadena de valor, aumentando la rentabilidad en una CV en comparación con las existentes?
- ¿Qué tecnologías se están desarrollando en las CV existentes?
- ¿Qué estrategia adoptan nuestros competidores?
- ¿Cómo se relaciona nuestra estrategia tecnológica con la estrategia global de las CV actuales?

- ¿Cuáles son nuestros puntos fuertes y débiles del diseño de la CV propuesta?
- ¿Qué elementos funcionales favorecen y superan la cadena de valor en respecto a las CV existentes?

En la Figura 1 se puede visualizar el proceso descrito por Matthews (2008), donde se muestra la evidencia de la estrecha interrelación de la organización y la tecnología. Las consideraciones tecnológicas que se mencionan son las responsables de las otras áreas de la empresa que permitirán emerger la estrategia corporativa de la cual se generará de la estrategia tecnológica, que se traducirá en el plan tecnológico.

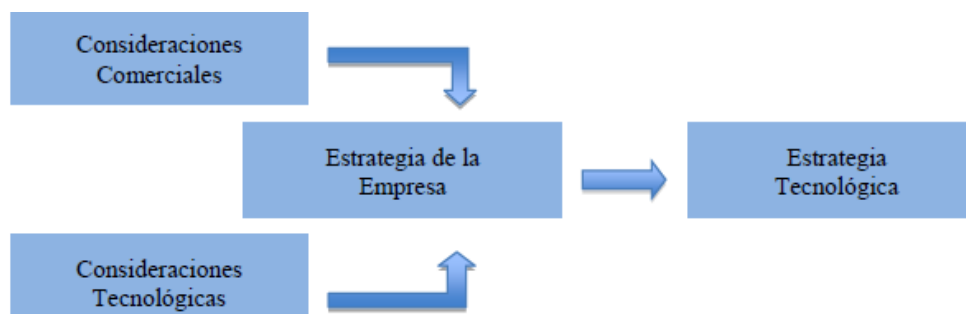


Figura 1. Plan tecnológico.

Fuente: Matthews (2008).

Porter, Anderson & Torres (2011) señalan que la tecnología permite desarrollar las ventajas de ser la primera o ser seguidora, cómo obtener o conceder licencias. Además, recomiendan examinar de manera sistemática todas las tecnologías de la organización que posibiliten una reducción de costos que favorezcan las diferenciaciones y concentrarse en ellas; para fines de la investigación se centra nuestro interés en las TIC. Basados en las referencias del autor, se resume los pasos finales para poder formular la estrategia tecnológica:

- Identificar las TIC que intervienen en las actividades administrativas y educativas de las IES.

- Identificar las TIC relevantes en otras IES potenciales o centros de investigación educativos a nivel Nacional e Internacional.
- Determinar la evolución probable de las TIC claves que participan en un desarrollo de entorno virtual.
- Determinar las TIC y los cambios tecnológicos potenciales que afecten a la ventaja competitiva y la estructura del sector.
- Valorar la capacidad de las IES en las TIC importantes y desarrollar mejoras en sus procesos educativos y administrativos.
- Seleccionar una estrategia tecnológica en consonancia con la estrategia competitiva global de la IES. Señalando que los distintos proyectos de I+D no deben aprobarse si no tienen un efecto sobre los costos o la diferenciación.

La tecnología como herramienta estratégica. La Formulación propia de estrategias tecnológicas y de experiencias tomadas con: la importancia del uso de herramientas para suscitar ideas y promover la discusión en las etapas previas, surge la necesidad de que la estrategia tecnológica se elabore conjuntamente con la estrategia global de la organización y plasmar la estrategia tecnológica en un plan de desarrollo tecnológico que muestre con claridad las opciones efectuadas (Matthews, 2008).

Una estrategia tecnológica bien diseñada puede lograr la diferenciación en el desempeño del producto, la estrategia tecnológica debe tratar tres aspectos:

- Que tecnologías debe desarrollar.
- Buscar o no el liderazgo tecnológico en esas tecnologías.

- El papel de licencias de tecnología o desarrollo tecnológico con open source o software free.

El papel de las tecnologías de la información en la inteligencia organizacional.

Cuando se habla de que la inteligencia organizacional, se relaciona con la capacidad para reunir, analizar y diseminar información interna y externa a las IES, esto no sería posible sin el uso de las TIC, ya que se requiere de redes locales, metropolitanas y de internet que garanticen el flujo de información de la institución así como las bases de datos, técnicas y herramientas para el manejo y el análisis del gran volumen de datos disponibles. “De igual forma contar con acceso libre a internet, que permita el acceso a esa enorme fuente de información que posibilite la realización de búsquedas eficaces y establecer una comunicación interpersonal y grupal” dentro de las IES (Cendejas, 2014:47).

Derivado de ello, se puede señalar que han existido grandes avances en la última década, en investigación centrada en el entorno educativo aplicando cuatro áreas de las TI; consideradas como la base para la explotación de la información y pieza fundamental en la inteligencia organizacional, las cuales se mencionan a continuación:

- Minería de datos (Data Mining - DM).
- Minería Web (Web Mining- WM).
- Minería textual (Text Mining - TM)
- Minería de datos Educativo (Educational Data Mining EDM).

2.4 Antecedentes de Minería de Datos y Redes Neuronales Artificiales

2.4.1 Descubrimiento de Conocimiento en Bases de Datos KDD.

Fayyad define KDD por sus siglas en inglés (*Knowledge Discovery from Databases*) como un proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. (Fayyad et al. 1996; Usama et al., 1996).

La diferencia es que busca el descubrimiento del conocimiento sin una hipótesis predefinida o preconcebida con respecto a sistemas tradicionales de explotación de datos basados en la existencia de una hipótesis o modelos previos. El proceso KDD define una metodología que provee una representación completa del ciclo de vida de un proyecto **DM** (Formia, 2013), descrito en la Figura 2.

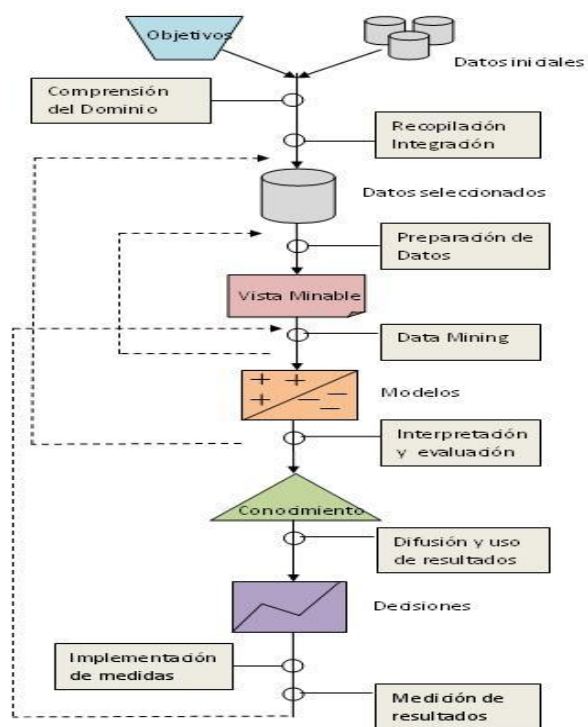


Figura 2. Proceso KDD.

Fuente (Formia, 2013: 7)

Por su parte, Vallejos y Martínez (2006) comentan que los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Entonces los especialistas del área, se refieren al conocimiento cuando elaboran o encuentran un modelo, haciendo que la interpretación de la información y el modelo representen un valor agregado; señalando que en los últimos tiempos, creció considerablemente las investigaciones para generar y almacenar información en los se ha estimado que la cantidad de datos en el mundo almacenados en bases de datos se duplica cada 20 meses. Es así que hoy las organizaciones tienen gran cantidad de datos almacenados y organizados, pero a los cuales no les pueden analizar eficientemente en su totalidad.

Con las sentencias SQL (*Structured Query Language*) se puede realizar un primer análisis, aproximadamente el 80% de la información se obtiene con estas técnicas. El 20% restante, que la mayoría de las veces, contiene la información más importante, requiere la utilización de técnicas más avanzadas como la DM, paso consistente de KDD e IA (Vallejos & Martínez, 2006).

El proceso de extracción de conocimiento KDD, se encarga de sacar conclusiones o información limpia a partir de unos datos que generalmente están organizados en una base de datos. Siendo este proceso un tanto iterativo debido a que la salida de algunas de sus fases puede requerir que se vuelva a una fase anterior, si es necesario repetir varias iteraciones hasta extraer conocimiento de los datos que se requieren para el estudio (Galán, 2015: 9).

El proceso de KDD se divide en cinco fases principales, en este proyecto se centra principalmente en una de ellas que es la DM, pero es esencial, analizar las demás ya que no se puede hablar de DM sin tener unos conocimientos previos de los datos en sí, de ¿cómo se

seleccionan, limpian, transforman y almacenan estos datos? y por supuesto una vez concluida la fase de DM será necesario saber ¿cómo evaluar e interpretar las conclusiones obtenidas del estudio. En la **Figura 3;Error! No se encuentra el origen de la referencia.** se presentan las fases del proceso de KDD de forma global (Hernández, Ramírez & Ferri, 2004, citado en Galán, 2015).

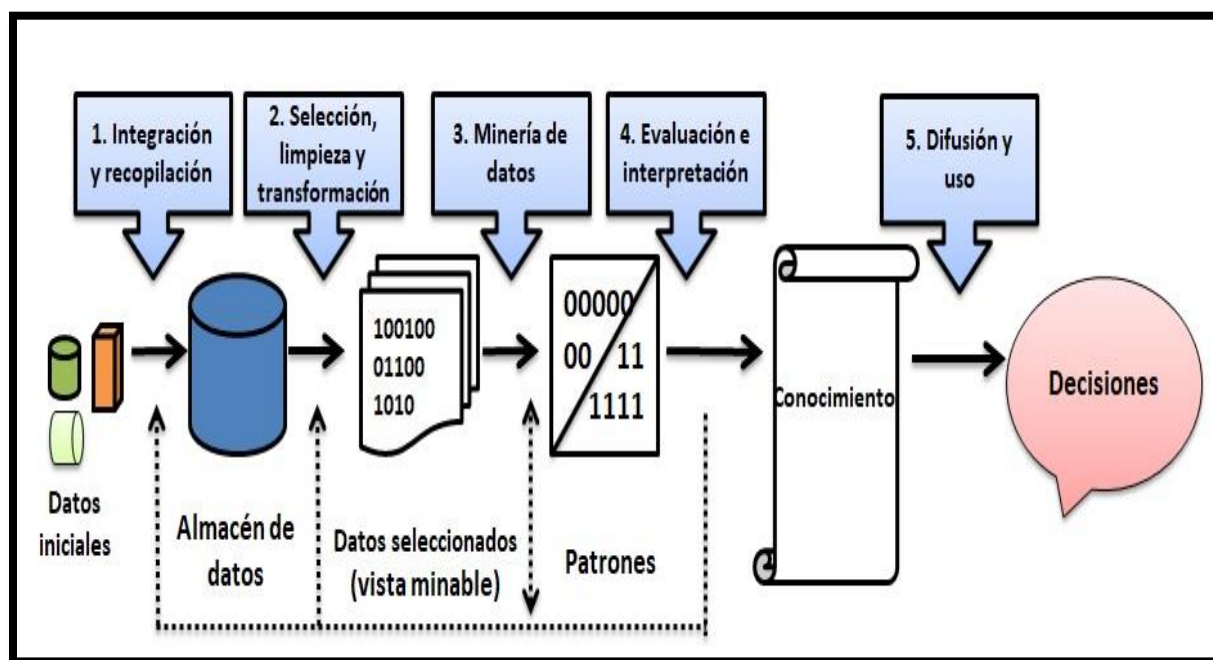


Figura 3. Fases Proceso KDD.

Fuente: Hernández et al., (2004) citado en Galán (2015: 9)

A continuación se describen las fases de este proceso.

- a) Integración y recopilación de datos: aquí se decide de dónde se van a obtener los datos que se utilizarán más adelante, es decir que fuentes de información resultan ser útiles. Después se transforman todos los datos a un formato en común, ya que como se mencionó, que estos pueden provenir de fuentes heterogéneas, generalmente esto se consigue usando un almacén de datos.

- b) Selección, limpieza y transformación: los datos recopilados en el almacén pueden contener errores en sus valores, o incluso puede que a algunos de estos les falte algún valor, que estos sean erróneos. En esta fase se trata de corregir o incluso eliminar estos datos, y se decide qué hacer con aquellos datos que estén incompletos. También se realiza una selección de aquellos datos que son relevantes para el proceso de extracción de conocimiento que se desea realizar.
- c) Minería de datos (*por sus siglas en inglés Data Mining DM*): es la fase principal que se trata en este proyecto, se debe decidir cuál es la tarea (agrupar, clasificar, etc.) que se va a realizar y se elige el método que se va a emplear para ello.
- d) Evaluación e interpretación: en la fase de minería de datos debe dar unos resultados, por ejemplo unos patrones observados en los datos. En la fase de evaluación e interpretación se debe evaluar e interpretar estos patrones, con el fin de poder entender el resultado obtenido. (Galán, 2015:10)
- e) Difusión y uso: es la última fase del proceso de KDD tiene como objetivo utilizar el nuevo conocimiento adquirido y hacer que dicho conocimiento sea empleado por todos los usuarios posibles. (Galán, 2015:11)

Calderón menciona que las metas del KDD son “el procesar automáticamente grandes cantidades de datos crudos, identificar los patrones más significativos y relevantes, y presentarlos como conocimiento apropiado para satisfacer las metas del usuario” (2005: 18).

Por otro lado Galán (2015) señala que KDD nace como interfaz, se relaciona con otras disciplinas y se nutre de:

- Sistemas de información / bases de datos (DB): tecnologías de DB y bodegas de datos, maneras eficientes de almacenar, acceder y manipular datos.

- Estadística, aprendizaje automático / IA (redes neuronales, lógica difusa, algoritmos genéticos, razonamiento probabilístico): desarrollo de técnicas para extraer conocimiento a partir de datos.
- Reconocimiento de patrones: desarrollo de herramientas de clasificación.
- Visualización de datos: interfaz entre humanos y datos, y entre humanos y patrones.
- Computación paralela / distribuida: cómputo de alto desempeño, mejora de desempeño de algoritmos debido a su complejidad y a la cantidad de datos.
- Interfaces de lenguaje natural a DB.

2.4.2 Minería de Datos (Data Mining DM).

Una definición tradicional de Minería de Datos es un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos.

DM, término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos, capaz de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido de la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en que deben actuar y tomar decisiones. (Hernández et al., 2004)

Desde el punto de vista empresarial, se define 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 & García , 2012).

La DM surge como una tecnología que intenta ayudar a comprender el contenido de una base de datos, donde los datos representan la materia prima bruta, convirtiéndose en información cuando el usuario le atribuye algún significado en especial, es decir extraer conocimiento que coadyuve a la toma de decisiones inteligentes (Molina & García , 2012).

La DM trata de sacar toda la información posible de los almacenes de datos, no se conforma sólo con la visualización de estos datos como podría pasar con las consultas simples, sino que trata de obtener resultados en cuanto a la relación que existe entre los mismos y cómo podrían dar beneficios de algún modo al negocio.

La DM es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de DM muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta (Galán, 2015).

La DM es la etapa de descubrimiento en el proceso de KDD: siendo un paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos pre procesados (Fayyad et al., 1996); aunque se suelen usar indistintamente los términos KDD y Minería de Datos (Hernández et al., 2004).

Galán (2015) señala que una vez que se han recopilado los datos necesarios y éstos están bien organizados y limpios, es la hora de aplicar sobre ellos el proceso de minería de datos cuyo objetivo es descubrir patrones que deben ser válidos, novedosos y por supuesto, comprensibles.

Para ello existen diversas tareas de minería de datos y métodos o técnicas que permiten resolver dichas tareas.

A continuación se describen las tareas, metodologías y técnicas descritas por el autor:

- Las tareas de DM se define como un tipo de problema a ser resuelto por algoritmos de DM; implicando en cada tarea sus propios requisitos y que la información que se obtiene empleando una tarea en concreto puede ser muy distinta a la obtenida si se emplea otra tarea diferente. Se puede dividir en dos tipos: predictivas y descriptivas.
 - a) En las predictivas el objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes (variables predictivas). Este tipo de tareas incluye la clasificación o discriminación (en estadística), clasificación suave, estimación de probabilidad de clasificación, categorización y regresión.
 - b) El objetivo de las tareas descriptivas, es identificar patrones en los datos que los explican o resumen. Aquí se incluyen el tipo de tareas de: agrupamiento (clustering), correlaciones y factorizaciones y reglas de asociación.
- Los métodos permiten resolver cualquiera de las tareas anteriores, donde requieren aplicar diferentes técnicas, algoritmos y no sólo uno; puntualizando que un método puede servir para resolver más de una tarea.

Los tipos de técnicas y métodos que existen para poder resolver las tareas mencionadas según Galán (2015) son:

- a) Técnicas algebraicas y estadísticas: cuyo objetivo es expresar modelos y patrones mediante fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones o valores agregados estadísticos tales como medias, varianzas, correlaciones, etc.

Generalmente estas técnicas obtienen un patrón a partir de un modelo predeterminado del cual se estiman unos coeficientes o parámetros. Ejemplos de algoritmos dentro de estas técnicas son: regresión lineal (global o local), regresión logarítmica y regresión logística. (Galán, 2015:18)

- b) Técnicas bayesianas: El objetivo es estimar la probabilidad de pertenencia a una clase o grupo mediante la estimación de las probabilidades condicionales inversas o a priori, para lo cual utilizan el teorema de Bayes. Ejemplos de algoritmos dentro de este grupo son: el clasificador bayesiano naive, los métodos basados en máxima verisimilitud y el algoritmo esperanza-maximización (EM).
- c) Técnicas basadas en conteos de frecuencias y tablas de contingencia: se trata de contar la frecuencia con la que dos o más sucesos se dan conjuntamente. Cuando el conjunto de sucesos posibles es muy grande, sólo se cuentan aquellos casos en los que las frecuencias conjuntas superan un determinado umbral.
- d) Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas: estas técnicas además de sus representaciones en forma de reglas, también se basan en algoritmos del tipo “divide y vencerás” (por ejemplo el ID3/C4.5 o el CART) y “separa y vencerás” (por ejemplo el CN2).
- e) Técnicas relacionales, declarativas y estructurales: representan el modelo mediante lenguajes declarativos, como los lenguajes lógicos, funcionales o lógico-funcionales.
- f) Técnicas basadas en redes neuronales artificiales: Son técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido.

Existen multitud de variantes de organización con muchos algoritmos diferentes para cada organización, pero el más conocido probablemente sea el de retropropagación.

- g) Técnicas basadas en núcleo y máquinas de soporte vectorial: son técnicas que intentan maximizar el margen entre los grupos o las clases formadas. Para ello se basan en unas transformaciones que pueden aumentar la dimensionalidad, que se llaman núcleos o kernels.
- h) Técnicas estocásticas y difusas: en este grupo se incluyen la mayoría de las técnicas que junto a las redes neuronales forman lo que se denomina computación flexible (soft computing). Son técnicas en las que o bien los componentes aleatorios son fundamentales, como el simulated annealing, los métodos evolutivos y genéticos, o bien al utilizar funciones de pertenencia difusas (fuzzy). (Galán, 2015:19)
- i) Técnicas basadas en casos, en densidad o distancia: son métodos que se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos, o de una manera más sofisticada, mediante la estimación de funciones de densidad. Algunos algoritmos muy conocidos son los jerárquicos, como Two-step o COBWEB, y los no jerárquicos como K medias. (Galán, 2015:20)

En esta investigación, se analizarán cada una del tipo de tarea y metodología o técnica para aplicar la que cumpla con los objetivos planteados.

2.4.3 Minería de Datos Educativa (Educational Data Mining EDM).

Romero y Ventura (2013) definen la minería de datos aplicada a la educación o EDM como una disciplina emergente que desarrolla y aplica métodos de minería de datos o DM sobre datos que vienen de entornos educativos, y los usa para entender mejor a los estudiantes y los

entornos en los que aprenden. Además mencionan otros términos similares que son empleados por áreas muy afines o relacionadas con EDM como son:

- “Análisis del aprendizaje o *Learning Analytics* (LA) consiste en la medida, recolección, análisis e informe de datos sobre estudiantes/aprendices y su contexto, con el propósito de comprender y optimizar el aprendizaje y el entorno donde ocurre” (Márquez , 2015: 19).
- “Análisis académico o *Academic Analytics* (AA) consiste en la aplicación de técnicas estadísticas y de minería a datos institucionales para producir inteligencia de empresa y soluciones a universidades y administradores” (Márquez , 2015: 19).

EDM se ha posicionado como un campo de gran interés para investigadores de diferentes tipos de entornos educativos como un área emergente y multidisciplinaria en la educación tradicional, el aprendizaje electrónico (E-learning) y los Sistemas de gestión de aprendizaje (*Learning Management System*, LMS), Sistemas Tutoriales Inteligentes (*Intelligent Tutoring System*, ITS) y los Sistemas Hipermedia Adaptativos (*Adaptive Educational Hypermedia System*, AEHS).

Márquez menciona que se deben diferenciar aspectos de DM con EDM respecto a otros dominios:

- EDM tiene por objetivos fundamentales mejorar y guiar el proceso de aprendizaje de los estudiantes, además realizar investigación que permita profundizar en el entendimiento de los fenómenos educativos. Estas metas son en ocasiones difíciles de cuantificar y requieren de un propio conjunto de técnicas de medición.
- La Información en los ambientes educativos hay diferentes tipos de información disponible. La información es específica al área educativa y tiene intrínseca información

semántica, relaciones con otros tipos de información y múltiples niveles de significado jerárquico.

- Las Técnicas. los problemas educativos y su información tiene algunas características que requieren ser tratados de una manera particular. Aunque muchas de las técnicas tradicionales pueden ser aplicadas directamente, otras no es posible y deben ser adaptadas al problema específico a tratar. (Márquez , 2015: 21)

El autor resume en la **Figura 4**, un esquema del proceso general de EDM y sus actores principales, mismos que se analizan, durante el desarrollo de la investigación.

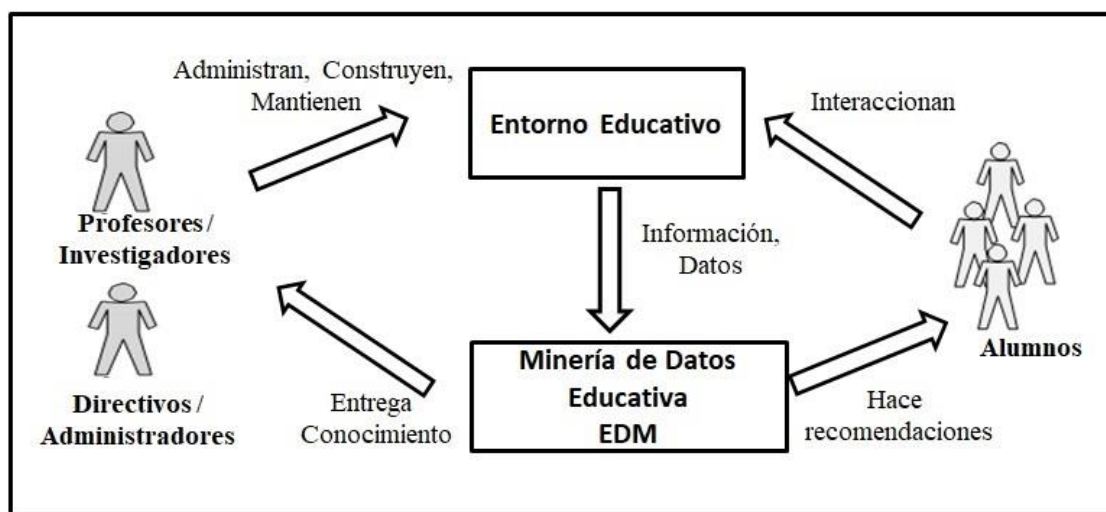


Figura 4. Proceso general de EDM.

Fuente: Márquez (2015).

2.4.4 Redes Neuronales Artificiales (RNA).

Las redes neuronales son una rama de la Inteligencia Artificial (IA), por ello se les denomina Redes Neuronales Artificiales (RNA) y por sus siglas en inglés *Artificial Neural Networks* (ANN's) que son sistemas de aprendizaje inspirados en el funcionamiento del cerebro humano. En este sentido, simulan e imitan sistemas permitiendo establecer relaciones no lineales

entre las variables de entrada y salida. La ventaja principal de las RNA es procesar información en paralelo en tiempo real que ha permitido su aplicación en la clasificación y reconocimiento de patrones en sistemas complejos. (Martínez , 2014, citado en Villada, Muñoz, & García-Quntero, 2016)

Las RNA no son más que un modelo artificial y simplificado del cerebro humano, es un nuevo sistema para el tratamiento de la información cuya unidad básica de procesamiento está inspirada en la célula fundamental del sistema nervioso humano: la neurona (Coello, Casas, & Pérez, 2015).

En cada neurona se recibe como entrada un conjunto de señales discretas o continuas, las pondera e integra y transmite el resultado a las neuronas conectadas a ella. Cada conexión entre dos neuronas tiene una determinada importancia asociada denominada peso sináptico o simplemente peso. En los pesos se suele guardar la mayor parte del conocimiento que la RNA tiene sobre la tarea en cuestión. El proceso sobre el cual se ajustan estos pesos para lograr un objetivo específico se denomina aprendizaje o entrenamiento y el procedimiento concreto para ello se denomina algoritmo de aprendizaje o algoritmo de entrenamiento (supervisado, no supervisado, reforzado e híbrido). El ajuste de pesos es la principal forma de aprendizaje en las redes neuronales aunque existen otras formas posibles.

Los elementos o componentes principales de una RNA que asemejan el comportamiento de cerebro humano son: -unidades de procesamiento (neurona artificial), -estado de activación de cada neurona, -conectividad entre neuronas y reglas de propagación, -función de transferencia y -reglas de aprendizaje. (Sanchez, 2004: 6)

Para Laurene (1994) una red neuronal se integra de tres componentes: un patrón de conexión entre las neuronas (arquitectura), métodos para determinar los pesos de las conexiones (algoritmos de aprendizaje o entrenamiento) y una función de activación. En la **Figura 5** se muestra la esquematización de una RNA a partir de una red neuronal biológica donde se identifican los elementos que la integran: patrones de entrada p_j , peso sináptico w_{ij} , polarización b_j , entrada neta r_i , la función de activación f resultante a partir de r_i y la salida de la neurona a_i (axón) después de aplicar la f .

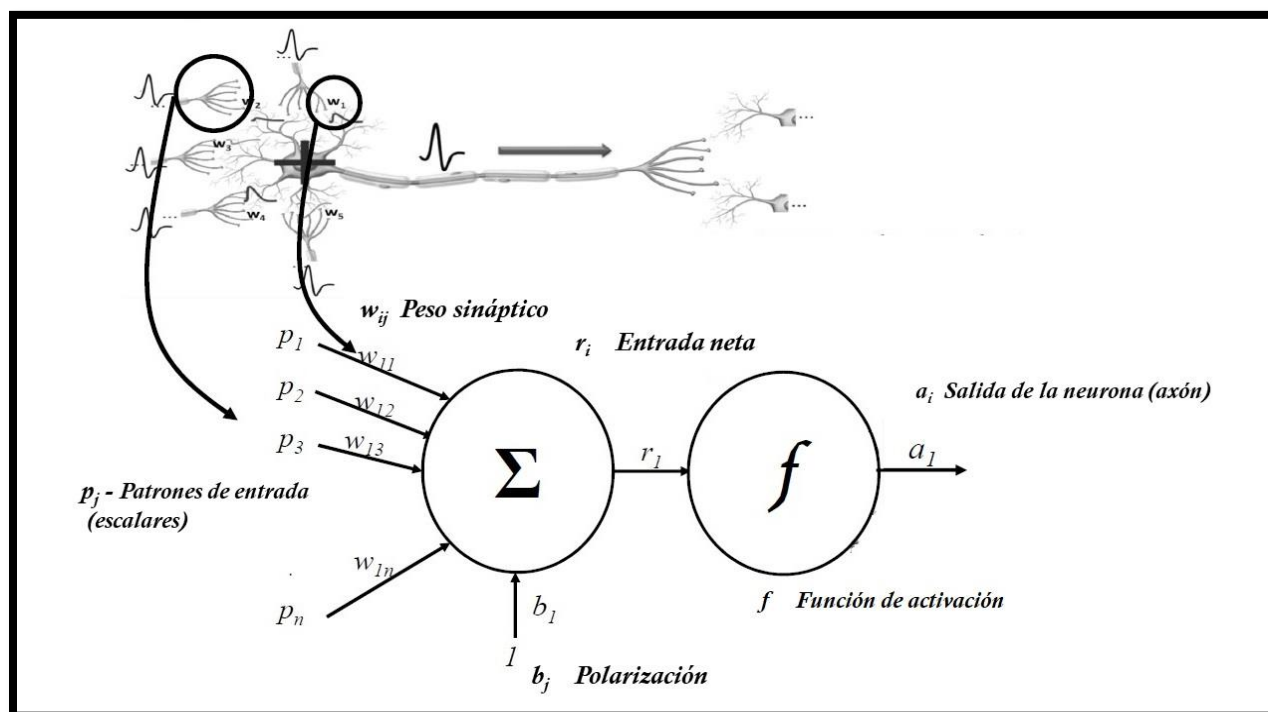


Figura 5. Esquematización de una RNA.

Fuente: Laurene (1994).

Plata (2008) define una RNA como un procesador distribuido en paralelo de forma masiva que tiene una tendencia natural para almacenar conocimiento de forma experimental y lo hace disponible para su uso, cuya semejanza respecto al cerebro humano está en dos aspectos:

- El conocimiento es adquirido por la red a través de un proceso de aprendizaje.
- Los pesos sinápticos o fuerza con que están interconectadas las neuronas se utilizan para almacenar la información.

Por su parte Rey (2014) en “Alternativas y determinantes en valoración de inmuebles urbanos” categoriza en seis criterios los modelos de RNA, que se indican a continuación:

- Número de Capas: Monocapa (Capa de entrada + Capa de salida) y Multicapa (Capas ocultas entre entrada y salida).
- Topología o arquitectura: No Recurrentes (alimentadas hacia delante o unidireccionales o feedforward) y Recurrentes (realimentadas o feedback).
- Tipos de respuesta: Autoasociativas (implican aprendizaje supervisado) y Heteroasociativas (implican aprendizaje no supervisado).
- Formas de datos de entrada: Entradas y Salidas analógicas o continuas, Entradas y Salidas discretas o binarias, y Entradas y Salidas combinadas.
- El grado de conexión: Totalmente conectadas y Parcialmente conectadas.
- Tipo de Aprendizaje: Supervisado, No supervisado, Reforzado, Híbrido.

El modelo neuronal resultante diferirá en función del número de capas presentes en la red, del tipo de conexiones (o topología de la red), del tipo de respuesta, de la forma de los datos de entrada y salida, así como del tipo de aprendizaje utilizado. De la multitud de modelos y variantes que existen, solo 50 son medianamente conocidos, aunque únicamente 15 son utilizados con mayor frecuencia en las aplicaciones prácticas.

Rey (2014) clasifica los tipos de aprendizaje o entrenamiento que se utilizan en el procesamiento o cómputo, aprendiendo a partir de un conjunto de patrones de la siguiente forma:

- **Supervisado:** se caracteriza por la presencia de un entrenador externo que comunica a la red del comportamiento que ésta, debe tener ante un determinado input. Se representa mediante ejemplos de entrada con su correspondiente salida deseada. Los pesos de la red se ajustan iterativamente, bajo la influencia combinada de sucesivos vectores de entrenamiento y las correspondientes señales de error (diferencias entre las respuestas proporcionadas por la red y las respuestas deseadas). (Rey, 2014:98)
- **No Supervisado (Autoorganizado):** no existe un entrenador externo que supervise el proceso, de manera que la red solamente cuenta con los inputs de entrada y, por sí misma, debe encontrar las propiedades o regularidades estadísticas de los datos de entrada y reflejarlas en la salida mediante la agrupación de patrones.
- **Reforzado:** se sitúa entre el aprendizaje supervisado y el autoorganizado. En este caso un entrenador externo proporciona a la red solo información cualitativa, es decir, informa sobre si la salida obtenida ha sido correcta o no. En definitiva, si la red ha actuado con éxito o si, por el contrario, se ha producido un fracaso. La idea es favorecer acciones que vayan seguidas de éxitos, así como atenuar aquéllas que van seguidas de fracasos. En ocasiones se denomina aprendizaje por premio-castigo. (Rey, 2014:99)
- **Híbrido:** este entrenamiento combina el aprendizaje supervisado y no supervisado en una misma red, siendo útil en determinadas situaciones, donde cada uno de los

aprendizajes se utilizan para la adaptación de los pesos de una de las capas (Rey, 2014).

En este sentido las RNA realizan este proceso mediante la toma de datos de un entorno externo y su almacenamiento en los pesos de las conexiones, cuyos cambios intrínsecos al procedimiento pretende obtener una mayor efectividad por parte de las RNA en el momento de realizar tareas similares. Señalando que el tipo de aprendizaje es determinado por la forma en que tiene lugar la adaptación de los pesos de la red. (Rey, 2014: 101)

La Tabla 5, presenta la clasificación de las RNA de acuerdo al tipo de Aprendizaje y de Topología, visualizando que existe un número mayor de los modelos de redes unidireccionales (no retroalimentadas) y de aprendizaje supervisado.

Tabla 5. Clasificación de las RNA por el tipo de Aprendizaje y de Topología

Tipo de Aprendizaje o Topología	RNA
Supervisado	• Realimentados: BSB, Fuzzy Cog. Map, BP Through Time. • Unidireccionales: Perceptrón, Adaline (Adaptative Linear Element), Madaline (Multiple Adaline), Perceptrón Multicapa (MLP), Back Propagation, Time-delay NN, CMAC, Red Neuronal de Regresión General (GRNN), Learning Vector Quantization (LVQ), Máquina de Boltzmann, Correlación en cascada.
No Supervisado	• Realimentados: Adaptive Resonance Theory (ART 1,2,3), Hopfield, Bidirectional Associative Memory (BAM). • Unidireccionales: LAM, OLAM, Mapas autoorganizados de Kohonen, Neocognitrón, Redes PCA.
Reforzado	• Premio-castigo asociativo • Crítico adaptativo
Híbrido	• Funciones de Base Radial (RBF) • Redes de Contrapropagación

Fuente: Elaboración propia a partir de Rey (2014).

Sánchez (2004) define una RNA como un conjunto de modelos matemáticos-computacionales reales e ideales y se emplea en estadística psicológica e inteligencia artificial.

Entre algunas aplicaciones importantes por señalar son: el reconocimiento de patrones (caracteres, voz, formas de onda EEG, imágenes), clasificación de células sanguíneas, e inspección de circuitos impresos (diagnósticos y tratamientos médicos), clasificación de patrones sin supervisión (clustering o agrupamiento), aproximación de funciones (procesamiento de señales electrónicas), predicción o forecasting en la toma de decisiones de negocios, ciencias e ingeniería (predicción de los cambios en la bolsa de valores, venta de productos, clima, sismos, etc.), optimización aplicada a problemas matemáticos, estadísticos, de ingeniería, ciencia, medicina y economía; así como el control adaptativo en procesos de manufactura entre otros.

Los beneficios que ofrece aplicar RNA son:

- No linealidad: Componentes no lineales, la no linealidad distribuida
- Mapeo de Entrada-Salida: Aprendizaje supervisado, inferencia estadística no paramétrica (estimación libre del modelo, no hay supuestos anteriores).
- Adaptabilidad: Ya sea mantener o adaptar. Puede hacer frente a entornos no estacionarios. Debe superar el dilema de estabilidad-plasticidad.
- Respuesta evidencial: La decisión más la confianza de la decisión puede ser proporcionada.
- La información contextual: Cada neurona en la red influye potencialmente cada una de las otras neuronas, por lo que la información contextual es tratada de forma natural.
- Tolerancia a fallos: El rendimiento se degrada con gracia.

- La uniformidad de análisis y diseño de componentes comunes (neuronas), se comparten teorías y algoritmos de aprendizaje, y una perfecta integración basado en la modularidad.
- Analogía neurobiológica: Redes Neuronales motivados por la neurobiología.

De todo, ello se remarca que existe un gran número de modelos de RNA que se han desarrollado en el campo de esta área de IA, por varios autores desde el primer modelo de RNA McCulloch y Pitts (1943), el Aprendizaje hebbiano de Hebb (1949), hasta los estudios reportados por Wan (1993) de Redes TDNN (time delay networks) para series temporales. En este trabajo solo se presentan los modelos más comunes, que se han utilizado en problemas reportados en este campo de investigación (Rey, 2014).

Plata (2008) menciona varios “modelos de redes y arquitecturas variadas: al modelo neuronal de McCulloch y Pitts, Adaline y Madaline, que constituyen un tipo de RNA desarrollada por BernieWidrow y MarcianHoff, en la Universidad de Stanford en 1959” (Citado en Coello, Casas, & Pérez, 2015: 16).

- Adaline (Adaptive Linear Element): fue desarrollada por Widrow&Hoff (1959), su estructura es muy parecida al perceptrón, con salida de un número real, donde el Elemento combinador adaptativo lineal, que recibe todas las entradas, las suma ponderadamente, y produce una salida, utiliza una nueva regla de aprendizaje llamada regla Delta o regla de mínimos cuadrados o por sus siglas en inglés Least Mean Squares (LMS), la diferencia principal está en cómo se crea la señal de error. La diferencia con respecto al perceptrón es la manera de utilizar la salida en la regla de aprendizaje. El perceptrón utiliza la salida de la función umbral (binaria) para el aprendizaje. Sólo se tiene en cuenta si se ha equivocado o no. En Adaline se utiliza directamente la salida de la red (real) teniendo en cuenta cuánto se

ha equivocado. En Adaline se basa en una generalización del aprendizaje de Widrow & Hoff para M neuronas monocapa, y ADALINE de aprendizaje de Widrow & Hoff Iterativo utilizando el método de descenso de gradiente. Es capaz de clasificar patrones linealmente separables, siendo pionera en la aplicación a problemas reales como el procesamiento adaptativo de señales y sistemas de control (Rey, 2014).

- Madaline: es una red multicapa de unidades ADALINE (Rey, 2014).

Sin embargo existen otros modelos dentro del campo de la investigación de RNA a considerar por otros autores:

- El Perceptron Multicapa o MLP (Multi-Layer Perceptron por sus siglas en inglés), es una red neuronal formada por múltiples capas que tienen como base un Perceptron, su propósito fundamental es resolver solamente problemas linealmente separable, representando esto, una limitante. (Plata, 2008, citado en Coello, Casas, & Pérez, 2015: 16)
- Redes Multicapa y Retropropagación cuyo objetivo es resolver problemas de clasificación con separabilidad NO Lineal (Sánchez, 2004).
- Redes Neuronales de Base Radial método propuesto por Broomhead y Lowe (1988) para hacer ajuste a funciones no lineales y clasificaciones de datos sin separabilidad lineal, se basa en la arquitectura de tres capas: la de entrada, la oculta y la de salida; así como de funciones de activación radial de carácter no lineal con sus centros gravitacionales propios en la capa oculta, y en la capa de salida mediante funciones lineales. Ofrecen mejor desempeño con un mayor volumen de datos de entrenamiento, con alta eficiencia

en la fase de entrenamiento. En este sentido se considera como un método alternativo para aproximar funciones y clasificar patrones (Sánchez, 2004).

- Mapas Autoorganizados o SOM (Self-Organizing Maps) propuesto por Teuvo Kohonen (1982) basado en ciertas evidencias descubiertas a nivel cerebral y con un gran potencial de aplicabilidad práctica.

Este tipo de red se caracteriza por poseer un aprendizaje no supervisado competitivo; cuya diferencia del aprendizaje supervisado, radica en que no existe ningún maestro externo que indique si la red neuronal está operando correcta o incorrectamente, donde no se dispone de ninguna salida objetivo hacia la cual la red neuronal deba tender. Por lo tanto el proceso de aprendizaje debe descubrir por sí misma rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones. En este sentido las neuronas deben autoorganizarse en función de los estímulos (datos) procedentes del exterior. (Palmer, Montaña & Jiménez, 2002: 2)

Finalmente, se señala que las RNA ofrecen ventajas que justifica su creciente interés de aplicación respecto a otras metodologías en el ámbito de la investigación de predicción tal es el caso de modelos estadísticos, entre otros.

2.5 Modelos de Predicción en indicadores educativos

2.5.1 Revisión de Modelos de Predicción.

En la Tabla 6, se resume el análisis realizado, sobre modelos de predicción que se ha venido desarrollando con Técnicas y Algoritmos de Minería de Datos para estudiar la problemática de estadísticos educativos como la deserción, reprobación y eficiencia terminal de

estudiantes de diferentes niveles educativos en estos últimos años, así como de los factores que afectan su rendimiento escolar.

Tabla 6. Revisión de modelos de predicción de estadísticos educativos

Título del Estudio / Autor / País	Técnica/ Algoritmo DM	Aportaciones / Contribuciones	Debilidades
Student Dropout Predictive Model Using Data Mining Techniques. (Amaya et al., 2015). Colombia	• Árboles de decisión ID3, C4.5 y Algoritmo j48 • Utiliza técnicas de transformación con el filtro de Discretize y Validación de pruebas a través de factor de confianza del software WEKA.	• Aplicación de técnicas de DM para generar un modelo predictivo y observar su comportamiento bajo ciertas circunstancias. • Modelo utilizado para predecir la probabilidad de que un estudiante abandone sus estudios, basados en sus características socioeconómicas y académicas. • El modelo predice un grado de precisión sobre la probabilidad de un alumno abandone sus estudios con base en sus valores de sus atributos y se pueda implementar líneas de acción para evitarlo.	• El modelo resultante es solamente una muestra del poder de los procesos de descubrimiento de conocimiento basado en herramientas de aprendizaje automático y minería de datos. • El modelo es específico a la población de objeto de estudio y a las características de los atributos, y no aplica con cambios dinámicos. Es decir no se puede replicarse en otras IES. • El modelo fue desarrollado con software WEKA donde el procesamiento DM fue tedioso para la puesta en marcha, por lo que se busca el desarrollo de un SW propio que permita un procesamiento DM más dinámico.
Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. Un caso de estudio. (Lanzarini et al., 2015). Argentina	• Algoritmos C4.5, PART y un multiperceptrón entrenado con el algoritmo de backpropagation • Técnicas de visualización	• Se mide la efectividad de los atributos seleccionados con los modelos que permiten clasificar a los alumnos en Regulares y No Regulares. (C4.5, PART y un multiperceptrón entrenado con el algoritmo de backpropagation). • Se muestra cómo utilizar la visualización para detectar características de un problema en un dominio específico en forma clara y precisa. • Construcción de un modelo de clasificación que permite describir y caracterizar a los alumnos según su condición de regularidad. • Se generan herramientas para obtener un modelo sencillo que posee una tasa de acierto equivalente al construido a partir de la información original.	• El modelo no se desarrolla con software propio, sino se apoya de herramientas de DM como RapidMiner. • Diseño de un Modelo específico para una IES.

Título del Estudio / Autor / País	Técnica/ Algoritmo DM	Aportaciones / Contribuciones	Debilidades
Predicting School Failure and Dropout by Using Data Mining Techniques. (Márquez et al., 2013). España-México	- Algoritmos de selección de atributos: CfsSubsetEval, ChiSquaredAttributeEval, ConsistencySubsetEval, FilteredAttributeEval, OneRAttributeEval, FilteredSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, ReliefFAttributeEval y SymmetricalUncertAttribute Eval. - Técnicas de rebalanceo de datos (SMOTE (Synthetic Minority Over-sampling Technique) y de clasificación sensible a costos DM - Árboles de Decisión (J48, SimpleCart, ADTree, RandomTree, REPTree) - Reglas de inducción: JRip, NNge, OneR, Prism, Ridor). - algoritmo evolutivo: Interpretable Classification Rule Mining (ICRM).	- Estas herramientas de visualización permiten a los docentes, independientemente del área en la que se desarrollen, analizar a sus alumnos, sin requerir conocimientos DM. - Desarrollo de dos metodologías: - Predecir a los estudiantes en riesgo de reprobar o abandonar sus estudios con la mayor exactitud posible. - Predecir lo más temprano posible a los estudiantes en riesgo de reprobar o abandonar sus estudios. - Los modelos son propuestos para la predicción del fracaso escolar o abandono y el algoritmo ICRM a datos procedentes de estudiantes de nivel medio superior. - El modelo es de tipo inductivo y con aprendizaje supervisado donde señalan obtener resultados de máxima exactitud de precisión.	- El Modelo fue basado en algoritmos de clasificación clásica de DM y programación Genética. Su utilización es compleja con usuarios no expertos en DM. - Se crea el modelo pero no se implementa en un SIAT real. - El modelo no se desarrolla con software propio, sino se hace primero un pre-procesado manual, se apoya de herramientas de DM como Weka y la ejecución desde línea de comandos del algoritmo ICRM. - No se integró toda la metodología propuesta dentro de un sistema real de alerta temprana de una institución educativa. Por tanto, para poder finalizar y evaluar todo este proceso de forma completa, es necesario disponer de información sobre los resultados obtenidos después de aplicar estas intervenciones a los estudiantes.
La deserción en cursos universitarios. Construcción de modelos sobre datos de la Universidad Nacional de Río Negro usando técnicas de Extracción de Conocimiento (Formia, 2013). Argentina	- Se basa en técnicas agrupamiento (k-means) para caracterización y elementos de análisis. - Modelado con Árbol de decisión, visor de para cada atributo. - Evaluación del desempeño del modelo con Clustering con el visor de clústeres de cada atributo, junto con la distribución del atributo en cada clúster - Utiliza técnicas de transformación de numerización (operador Nominal to Numerical y Map de RapidMiner). - Aplica normalización de rango sobre todos los atributos utilizando el método de transformación z (operador Normalize de RapidMiner). - Proceso de selección de atributos con selection forward que toma en cuenta la performance de un determinado modelo de aprendizaje para realizar la validación de los conjuntos de características, se implementan con el operador Optimize Selection de RapidMiner.	- Utiliza métodos no supervisados que sirvieron de base para la creación, comparación y evaluación de modelos analíticos para los datos, que aporte conocimiento útil a la problemática abordada. - Surge de la premisa de que no existe método DM aplicable a cualquier tipo de problema, es por eso que surge la necesidad de evaluar métodos conocidos en función de la problemática. - Se basa en técnicas descriptivas donde resume las características generales del conjunto de datos respecto a la información socioeconómica y/o al rendimiento académico de los estudiantes de las diferentes cohortes generacionales. Y métodos	- Se utiliza DM para apoyo en la interpretación del fenómeno abordado y las bases analíticas para futuros desarrollos de modelado de datos e iniciar la construcción de soluciones. - Solo se aplican técnicas de aprendizaje no supervisado en la investigación y dejan para trabajos futuros el aprendizaje supervisado que les permita obtener modelos que ayuden a predecir el fenómeno de deserción universitaria y definir estrategias de intervención solo se implementó para su evaluación con SOAP. - El modelo no se desarrolla con software propio, sino se hace primero un procesamiento manual de las primeras fases DM y se apoya de herramientas de DM como RapidMiner. - Diseño de un Modelo específico para una IES.

Título del Estudio / Autor / País	Técnica/ Algoritmo DM	Aportaciones / Contribuciones	Debilidades
A Framework to avoid Scholar Desertion using Artificial Intelligence Treviño et al. (2013)	- La evaluación de la selección con el método wrapper, CFS y una matriz de confusión. - La validación de atributos seleccionados se realiza con árboles de decisión con el algoritmo C4.5 implementado en operador W-J48 de la extensión Weka de RapidMiner.	de agrupamiento para organizar los datos en grupos similares. - Se evalúan los métodos de selección de atributos con SOAP basado en proyecciones que permite el tratamiento de atributos nominales y numéricos de manera supervisada. - Propuesta de una nueva metodología para evitar una posible deserción, bajo un framework basado en un algoritmo predictivo capaz de identificar tres aspectos importantes de trayectoria escolar del estudiante: - el nivel de logro - el nivel de satisfacción del investigador - las limitaciones de los estudiantes. - Marco completo con la idea principal de ofrecer a la IES con sistemas de utilidad y totalmente autónoma. Con esto, el departamento académico será capaz de actuar de forma preventiva hacia la deserción de otro estudiante. - Propuesta de un FRAMEWORK capaz de predecir cuando los estudiantes tienden a abandonar sus estudios en una forma autónoma. - El diseño del modelo se basa en lógica difusa para emular el análisis desarrollado por los seres humanos y capaz de examinar desde distintos puntos de la solución de las variables involucradas en particular y una situación especial (Deserción estudiantil) con un enfoque novedoso, útil e inteligente para reducir la tasa de deserción y mejorar el rendimiento académico.	- El modelo es específico y no puede replicarse en otras IES. - Se diseña el modelo pero no se implementa en un SIAT - No presentan resultados contrastados en una IES y técnicas de evaluación de modelo DM construido.
Modelo para la automatización del proceso de determinación de riesgo de	- Modelos de Redes Neuronales (RNA), Árboles de decisión y Cluster K-medianas. - Algoritmo de red neuronal de Microsoft	Metodología que permita identificar en forma automática a los estudiantes con mayor riesgo de deserción.	El modelo fue evaluado con datos de prueba, los cuales indican que ningún modelo predictivo arrojó resultados positivos, debido a esto se analizó el proceso y se llegó a la conclusión que es

Título del Estudio / Autor / País	Técnica/ Algoritmo DM	Aportaciones / Contribuciones	Debilidades
deserción en estudiantes universitarios (Fischer E. , 2012). República de Chile	• Algoritmo de árboles de decisión de Microsoft (entropía y las redes Bayesianas) • Algoritmo de clústeres de Microsoft (K-medianas, Expectation Maximization (EM). • Evaluación del desempeño del modelo DM con el método de validación cruzada. • Evaluación del desempeño del modelo RNA con el Visor de redes neuronales • Evaluación del desempeño del modelo Árbol de decisión con el Visor de árboles de decisión para cada atributo • Evaluación del desempeño del modelo con Clustering con el visor de clústeres de cada atributo, junto con la distribución del atributo en cada clúster.	• Metodología CRISP-DM que estructura el proceso EDM en seis fases, que interactúan entre ellas de forma iterativa. • El algoritmo de red neuronal de Microsoft crea modelos de DM datos de regresión y de clasificación mediante la generación de una red de perceptrón multicapa de neuronas. • El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles (entropía y las redes Bayesianas). • El algoritmo de clústeres de Microsoft, utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares, son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. • Evaluación de algoritmo de aplicados se basan en medidas DM en tres categorías de precisión, confiabilidad y utilidad. • Genera un gráfico de elevación como un método para visualizar la mejora de utilizar un modelo DM, comparado con una estimación aleatoria. • Utiliza una matriz de clasificación como método para ordenar las estimaciones buenas y malas en una tabla, para analizar con qué precisión predice el modelo el valor de destino ((RNA, Árboles de decisión y Cluster K-medianas). A través de utilizar F-Measure correspondiente a la media armónica de la precisión y el alcance.	muy probable que los datos de entrada no eran suficientemente confiables. • Una limitante de este trabajo era imposible conseguir datos fidedignos y completos, por lo que propone una metodología para enfrentar estudios de minería de datos educativa donde se eviten los problemas descritos. • Los valores obtenidos, no permiten validar positivamente dichos modelos, dado que su capacidad de predicción es menor que la estimación aleatoria. • Propuesta de implementar un proceso de captura de variables relevantes para la investigación, al momento del ingreso del estudiante a la institución superior, de tal manera de facilitar la generación de un almacén de datos con para ayudar a la toma de decisiones y tener la oportunidad de hacer un estudio serio de las variables que identifican a los alumnos que desertan de la carrera. • Selección de mejores atributos de forma manual, no hay rebalanceo del conjunto de datos ni apoya de un software. • El modelo no se desarrolla con software propio, sino se hace primero un pre-procesado manual, se apoya de herramientas de DM como Analysis Services de SQL server 2008 y algoritmo de red neuronal de Microsoft. • Modelo aplicado a una IES en específico de las carreras de Ingeniería de la Universidad de las Américas.

Título del Estudio / Autor / País	Técnica/ Algoritmo DM	Aportaciones / Contribuciones	Debilidades
Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. (Valero et al., 2010). México	- Algoritmo de árboles de clasificación C4.5 y el algoritmo de los k vecinos más cercanos (k nearest neighbors) - Se utiliza validación cruzada como un método de entrenamiento para la evaluación del modelo.	- Herramienta útil y práctica que coadyuva al cálculo de probabilidad de deserción de estudiantes, apoyando al proceso de tutorías. - Construcción de una interfaz WEB que permitiera interactuar con el modelo construido, que facilita la identificación de los alumnos vulnerables.	- Software específico y adaptativo solo al modelo desarrollado árboles de decisión. - Modelo aplicable solo a la IES en estudio (alumnos del Programa Educativo de TIC-SI, con datos históricos de éste y sólo puede utilizarse en la Universidad UTIM) - Solo emite el reporte del modelo probabilidad de deserción pero no genera una interpretación de resultados de la validación del Modelo con más técnicas que soporten comparativa de aprendizaje tanto no supervisado como supervisado

Fuente: Elaboración propia.

Del análisis de trabajos de investigación citados sobre modelos de predicción en la Tabla 6, se presenta una tabla comparativa con respecto al modelo predictivo PredATIS.

Tabla 7. Análisis Comparativo del Modelo PredATIS con los modelos de predicción estudiados

Modelo	Análisis de Fortalezas y Debilidades
Modelos Investigados	- Mayor porcentaje de Modelos construidos con técnicas, métodos y algoritmos no supervisados. - Trabajos desarrollados con técnicas, métodos y tareas adaptativas al software utilizado en el modelado. - Modelos desarrollados con aporte científico a la DM y EDM pero carecen de su implementación en un software propio. - Modelos adaptativos a IES exclusivamente y no implementados completamente en un SIAT.

Modelo	Análisis de Fortalezas y Debilidades
	• Se visualiza que los trabajos se han desarrollado, en su mayoría en otros países, dando un espacio de investigación para este fenómeno en estudio en México principalmente. • Son utilizados en el análisis y evaluación de factores implicados en el fenómeno multifactorial o “el problema de las mil causas” (Magaña, 2002), a través del procesamiento de grandes volúmenes de información de estudiantes de las IES para obtener reportes estadísticos educativos, donde se apoyan de las áreas matemáticas (métodos estadísticos para la toma de decisiones) y computación, específicamente en sus campos de estudio de Minería de Datos e Inteligencia Artificial. Sin embargo por el momento como lo señalan Treviño et al. (2013) ningún trabajo es capaz de contribuir con una predicción lo más cercano posible a la situación de deserción, reprobación y por ende eficiencia terminal, por lo que sigue representando un reto importante y un campo de investigación por parte de IES.
Modelo PredATIS	• Aportar a la investigación del fenómeno de indicadores educativos con técnicas, métodos DM a partir de algoritmos no supervisados y supervisados de aprendizaje automático para el diseño del Modelo. • Construir una CVA adaptable al proceso de KDD para la implementación del modelo predictivo de Alerta Temprana en indicadores Educativos, en un software de diseño propio, que aún los modelos analizados no han cubierto;

Modelo	Análisis de Fortalezas y Debilidades
	además de contribuir en el campo de investigación de EDM en México principalmente. • El modelo predictivo generado se pretende generar pruebas y entrenamiento con datos de una muestra de la CVA de un programa educativo (PE) de una IES del Estado de México y con proyección a futuro, que queda fuera del alcance de este trabajo su implementación y extensión a todos los PE que imparte la propia institución y de ser posible, a otras IES a nivel nacional.

Fuente: Elaboración propia.

2.6 Ingeniería y Arquitectura de Desarrollo de Software de la Comunidad Virtual

Someerville (2005), separa estas herramientas de apoyo para el desarrollo de software (SW) en Modelos y metodologías:

- Modelos de desarrollo de SW: es una representación simplificada del proceso para el desarrollo de software, presentada desde una perspectiva específica.
- Metodologías de desarrollo de SW: es un enfoque estructurado para el desarrollo de SW que incluye modelos de sistemas, notaciones, reglas, sugerencias de diseño y guías de procesos.

2.6.1 Modelos de Desarrollo de SW.

Un modelo para el desarrollo de SW es una representación abstracta de un proceso. Cada modelo representa un proceso desde una perspectiva particular y así proporcione información parcial sobre el proceso. Estos modelos generales no son descripciones definitivas de los procesos del SW más bien son abstracciones de los procesos que se pueden utilizar para el

desarrollo del SW. Puede pensarse en ellos como marcos de trabajo del proceso y que pueden ser adaptados para crear procesos más específicos (Cendejas, 2014). En la Tabla 8 se presenta una comparativa con respecto a características, ventajas y desventajas de algunos modelos utilizados de forma tradicional en el desarrollo de SW.

Tabla 8. Modelos tradicionales de desarrollo de SW

Cascada	Desarrollo evolutivo (espiral)	Desarrollo basado en componentes
Descripción		
“Considera las actividades fundamentales del proceso de especificación, desarrollo, validación y evolución. Los representa como fases separadas del proceso, tales como la especificación de requerimientos, el diseño del SW, la implementación, las pruebas, etcétera” (Cendejas, 2014: 87).	“Entrelaza las actividades de especificación, desarrollo y validación. Es decir surge de un sistema inicial que se desarrolla rápidamente a partir de especificaciones abstractas. Basándose en las peticiones del cliente para producir un sistema que satisfaga sus necesidades” (Cendejas, 2014: 87).	Se basa en un número significativo de componentes reutilizables. En el proceso de desarrollo se integran estos componentes en el sistema, más que en desarrollarlos desde cero (Someerville, 2005).
Ventajas		

Cascada	Desarrollo evolutivo (espiral)	Desarrollo basado en componentes
• “Se tiene todo bien organizado y no se mezclan las fases”. • “Es perfecto para proyectos que son rígidos, donde se especifiquen muy bien los requerimientos y se conozca muy bien la herramienta a utilizar” (EcuRed, 2010, párr. 4).	• Reduce los riesgos porque se revisa en cada iteración, antes de que se conviertan en problemas. • Es ideal para crear productos con diferentes versiones mejoradas y nuevos requerimientos como se hace con el SW actual, ya que este ciclo de vida no es rígido ni estático. • Incorpora objetivos de calidad (Someerville, 2005: 69).	• Reduce la cantidad de SW, ya que reutiliza código, reduciendo costos y riesgos. • Permite una entrega más rápida del SW. • Es uno de los modelos más utilizados en la práctica actual de la ingeniería de SW en el desarrollo de grandes sistemas de Información Pressman (2006).
Desventajas		
• Aplicable rara vez, los proyectos reales no sigue una secuencia lineal.	• Genera mucho tiempo en el desarrollo del sistema. • Requiere un alto costo de desarrollo.	• Requiere de alto compromiso a los requerimientos y esto da lugar a un sistema

Cascada	Desarrollo evolutivo (espiral)	Desarrollo basado en componentes
• Crea una mala implementación que lo lleva al fracaso. • El proceso de creación del software tarda mucho tiempo ya que debe pasar por el proceso de prueba y hasta que el software no esté completo no se opera (EcuRed, 2010, párr. 3).	• Requiere experiencia en la identificación de riesgos (Someerville, 2005: 69).	que puede no cumplir con las necesidades reales de los usuarios (Pressman, 2006).

Fuente: Elaboración propia a partir de Cendejas (2014); Pressman (2006) & Someerville (2005).

Señalando que los tres modelos descritos en la Tabla 8 son los que más se usan en la práctica actual de la ingeniería del SW y que no se excluyen mutuamente, donde a menudo se utilizan juntos especialmente para el desarrollo de grandes sistemas. Además de otras metodologías denominadas pesadas, entre las más populares son: Rational Unified Process (RUP) y Microsoft Solutions Framework (MSF). En estas metodologías pesadas se dice que:

Centran su atención en mantener una documentación exhaustiva del proyecto y cumplir con el plan previsto y definido con precisión en la fase inicial del desarrollo del proyecto, ofreciendo esta ventaja, con respecto a los modelos tradicionales que suelen enfatizar

la documentación, la planificación y seguimiento riguroso de múltiples actividades (mediante plantillas, técnicas de administración, revisiones formales, etc.) y están pensadas en avanzar en la construcción del SW de una manera iterativa e incremental, sin embargo carecen de este enfoque ya que representan altos costos al implementar cambios y buena solución en proyectos que operan en un entorno extremadamente incierto. (WikiUniversidad, 2015: párr.3)

Por otro lado Cendejas argumenta que:

El software de computadoras moderno se caracteriza por el cambio continuo, los tiempos de entrega son muy reducidos y una necesidad intensa de satisfacer al cliente/usuario. En muchos casos, el tiempo de llegada al mercado es el requisito de gestión más importante. Si se pierde una ventana del mercado, el mismo proyecto de software puede perder su significado. (2014: 93)

Sin embargo, con el progreso tecnológico global, se tiene que responder a nuevas oportunidades y mercados, condiciones económicas cambiantes y la aparición de productos y servicios de competidores, donde el SW juega un rol importante en casi todas las operaciones de negocio, se requiere un desarrollo rápido y eficiente ante estas necesidades y que aún las metodologías pesadas no cubren en la totalidad en este tipo de proyectos, en este sentido se utilizan las metodologías ágiles que se describen en la siguiente sección (Cendejas, 2014).

2.6.2 Metodologías para el Desarrollo Ágil de SW.

Las metodologías de desarrollo de SW han evolucionado de manera significativa en las últimas décadas, por lo que se requiere de una entrega rápida de los sistemas de información, donde los procesos de desarrollo de SW basados en una completa especificación de los

requerimientos, diseño, construcción y pruebas del sistema no se ajustan al desarrollo rápido de aplicaciones; es decir sí, se solicitan cambios o se encuentran problemas con ellos, el diseño o implementación del sistema se tiene que volver a realizar o probar, ampliando el tiempo de entrega al que se pactó con el cliente, donde muchas veces puede ser inútil. Derivado de estas desventajas en las metodologías tradicionales, da lugar a las metodologías de desarrollo Ágil de SW, que surge el término ágil aplicado al desarrollo de SW de la reunión de 2001 en Utah-EUA, donde participan un grupo de diecisiete expertos de la industria del SW y algunos creadores o impulsores de metodologías de SW; cuyo objetivo fue esbozar los valores y principios que deberían permitir a los equipos a desarrollar SW rápidamente y respondiendo a los cambios que surgen a lo largo de los proyectos en desarrollo(Cendejas, 2014: 96-98:).

Se pretendía ofrecer una alternativa a los procesos de desarrollo de software tradicionales, caracterizados por ser rígidos y dirigidos por la documentación que se genera en cada una de las actividades desarrolladas. Varias de las denominadas metodologías ágiles ya estaban siendo utilizadas con éxito en proyectos reales, pero les faltaba una mayor difusión y reconocimiento. Tras esta reunión se creó “The Agile Alliance” (2001), una organización, sin ánimo de lucro, dedicada a promover los conceptos relacionados con el desarrollo ágil de software y ayudar a las organizaciones para que adopten dichos conceptos. El punto de partida fue el manifiesto ágil, un documento que resume la filosofía "ágil". (Cendejas, 2014: 98)

Las metodologías de desarrollo ágil de SW buscan la satisfacción del cliente, la entrega temprana del SW incremental, equipos pequeños de alta motivación, métodos un tanto informales y una simplicidad general del desarrollo; cuyos procesos que se desarrollan

están diseñados para generar software útil de forma rápida e interactiva que entrelazan la especificación, el diseño, el desarrollo y las pruebas; permitiendo el éxito de muchos de los sistemas desarrollados para distintas áreas. (Cendejas, 2014: 97)

En la Tabla 9 se presenta una comparativa entre modelos tradicionales y metodologías pesadas con respecto a las metodologías ágiles, cuyas diferencias no solo se refieren al proceso en sí, sino también al contexto de equipo y organización que es más favorable a cada uno de estas filosofías de procesos de desarrollo de SW (Cendejas, 2014).

Tabla 9. Análisis Comparativo del Modelos tradicionales y metodologías ágiles en el desarrollo de SW

Metodología ágil	Modelos y Metodologías Tradicionales
• Pocos artefactos. El modelo es prescindible, modelos desechables. • Pocos roles, más genéricos y flexibles. • No existe un contrato tradicional, debe ser bastante flexible. • El cliente es parte del equipo de desarrollo. • Orientada a proyectos pequeños, y en el mismo lugar. • La arquitectura se va definiendo y mejorando a lo largo del proyecto.	• Más artefactos. El modelo es esencial, mantenimiento en los modelos. • Más roles y más específicos. • Existe un contrato prefijado. • El cliente interactúa con el equipo de desarrollo mediante reuniones. • Aplicables a proyectos de cualquier tamaño, pero suelen ser especialmente efectivas/usadas en proyectos grandes. • Se promueve que la arquitectura se defina tempranamente en el proyecto.

Metodología ágil	Modelos y Metodologías Tradicionales
• Énfasis en los aspectos humanos: el individuo y el trabajo en equipo. • Basadas en heurísticas provenientes de prácticas de producción de código. • Se esperan cambios durante el desarrollo del proyecto.	• Énfasis en la definición del proceso: roles, actividades y artefactos. • Basadas en normas provenientes de estándares seguidos por el entorno de desarrollo. • Se espera que no ocurran cambios de gran impacto durante el desarrollo.

Fuente: Cendejas (2014).

Los métodos de desarrollo ágil de SW más utilizados son: SCRUM, Extreme Programming (XP). Adicionalmente existen: Lean Software Development (LSD), Agile Modeling Adaptive Software Development (ASD), Agile Unified Process (AUP), Agile Data Method, Dynamic Systems Development Method (DSDM), Lean Software Development (LSD), y Microsoft Solutions Framework (MSF), etc. (EcuRed, 2010).

En este sentido y derivado del análisis comparativo de los tres modelos tradicionales de desarrollo de SW, las metodologías pesadas RUP y el desarrollo ágil de SW, no se puede decir que uno es mejor que el otro, por lo que representa una tarea difícil para el desarrollador al seleccionar ¿cuál será la metodología correcta?, ya que debe considerar las variables de: cambios tecnológicos, tipo y tamaño de proyecto que se va a realizar, recursos con los que se cuenta y el tiempo destinado para el desarrollo.

Como parte de la investigación, se presenta el diseño de la comunidad virtual académica CVAATI, que se realiza a través del Proceso Unificado Ágil por sus siglas en inglés Agile Unified Process (AUP). Derivado que esta metodología es una versión simplificada de RUP y la naturaleza y tamaño del proyecto que se desarrolla en este trabajo.

AUP describe de forma simple y fácil de entender la forma de desarrollar aplicaciones de SW usando técnicas ágiles y conceptos que aún se mantienen válidos en RUP, aplicando desarrollo dirigido de pruebas o TDD por sus siglas en inglés (test driven development), gestión de cambios ágiles y la refactorización de base de datos para mejorar la productividad del desarrollo y entrega (Flores, 2011).

Proceso Unificado Ágil (AUP). Es un marco de desarrollo de SW iterativo e incremental. Flores (2011) lo considera como un proceso altamente ceremonioso porque especifica muchas actividades y artefactos involucrados en el desarrollo de un proyecto software, que puede ser adaptado a RUP metodología denominada pesada y este a su vez ligero y ágil.

AUP se preocupa especialmente de la gestión de riesgos. Propone que aquellos elementos con alto riesgo obtengan prioridad en el proceso de desarrollo y sean abordados en etapas tempranas del mismo. Para ello, se crean y mantienen listas identificando los riesgos desde etapas iniciales del proyecto. Especialmente relevante en este sentido es el desarrollo de prototipos ejecutables durante la fase de elaboración del producto, donde se demuestre la validez de la arquitectura para los requisitos clave del producto y que determinan los riesgos técnicos. El proceso de AUP establece un modelo más simple que RUP por lo que reúne en una única disciplina las disciplinas de Modelado de Negocio, Requisitos y Análisis y Diseño. El resto de disciplinas (Implementación, Pruebas,

Despliegue, Gestión de Configuración, Gestión y Entorno) coinciden con las restantes de RUP. (Flores, 2011: 3)

En la **Figura 6** se indican cuatro flujos de trabajo de ingeniería y tres de apoyo del proceso de AUP y cuatro fases en el ciclo de desarrollo ágil de SW se describen a continuación (EcuRed, 2010):

- **Iniciación:** en esta fase se identifica el alcance y dimensión del proyecto, la propuesta de la arquitectura de desarrollo de SW y el presupuesto asignado.
- **Elaboración:** se identifica y valida la arquitectura de desarrollo de SW.
- **Construcción:** se desarrolla el SW en forma incremental con base en las prioridades funcionales de los participantes.
- **Transición:** se realiza la validación e implantación del sistema.

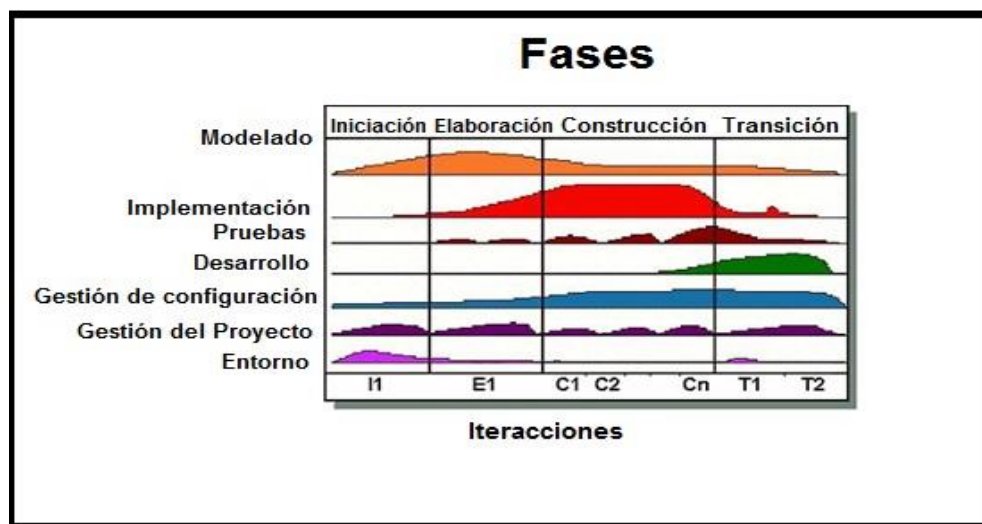


Figura 6. Fases de AUP.

Fuente: Ecured (2010).

Los flujos de trabajo de ingeniería de SW son el modelado, implementación, pruebas y despliegue.

- El *modelado* debe entender el negocio de la organización, el problema de dominio que se aborda en el proyecto y determinar una solución viable.
- La *implementación* transforma el modelo en código ejecutable y realiza un nivel básico de las pruebas específicamente la unidad de pruebas.
- En las *pruebas* se evalúa de forma objetiva la búsqueda de defectos, validar la funcionalidad de un sistema y verificar que se cumplan los requerimientos para garantizar la calidad del SW.
- El objetivo del *despliegue* es el plan de desarrollo del sistema y la ejecución de dicho plan, para que los usuarios finales pueda entrar al sistema.

Adicionalmente AUP describe los tres flujos de trabajo que sirven de apoyo (EcuRed, 2010).

- Gestión de configuración: consiste en gestionar el acceso a los artefactos del proyecto, incluyendo la traza de versiones, control de cambios y la gestión de los mismos.
- Gestión de proyectos: se dirigen las actividades entre las que se incluye la gestión de riesgos y recursos, así como la coordinación del proyecto.
- Entorno: debe apoyar el resto del esfuerzo asegurando que los procesos, métodos y herramientas están disponibles para el equipo de trabajo cuando los necesitan.

2.6.3 Modelado UML.

El Lenguaje de Modelado Unificado UML por sus siglas en inglés Unified Modeling Language, es una de las innovaciones conceptuales en el mundo tecnológico del desarrollo de SW que ha evolucionado en los últimos años, convirtiéndose en un estándar de facto en cualquier industria y no solo en el campo del SW, es decir aplicable en donde se requiere la construcción

de modelos, previo al diseño y construcción de prototipos y en general del desarrollo de SW.

(Rumbaugh, Jacobson, & Booch, 2000: 18)

Las propiedades de UML como lenguaje de modelado estándar son:

- Concurrencia, es un lenguaje distribuido y adecuado a las necesidades de conectividad actual y futuro.
- Ampliamente utilizado por la industria desde su adopción por OMG por sus siglas en inglés Object Management Group) que es, una asociación sin fines de lucro formada por grandes corporaciones, muchas de ellas de la industria del SW: IBM, Apple Computer, Sun Microsystems Inc y HewlettPackard; se encarga de la definición y mantenimiento de estándares para aplicaciones de la industria de la computación.
- Reemplaza a decenas de notaciones empleadas con otros lenguajes.
- Modela estructuras complejas.
- Las estructuras más importantes que soportan tienen su fundamento en las tecnologías orientadas a objetos, clase, componentes y nodos.
- Emplea operaciones abstractas como guías para variaciones futuras, añadiendo variables si es necesario.
- Modela el comportamiento del sistema a través de diagramas de casos de uso, diagramas de secuencia y de comunicación que sirven para evaluar el estado de las máquinas. (Rumbaugh, Jacobson, & Booch, 2000: 19)

UML es un lenguaje de modelado visual no es un lenguaje de programación; se utiliza para especificar, visualizar, construir y documentar artefactos de un sistema de SW, donde se captura decisiones y conocimiento sobre los sistemas que se deben construir. Permite entender,

diseñar, hojear, configurar, mantener, controlar la información sobre los sistemas orientados a objetos e incorporar mejores prácticas en el entorno de desarrollo de SW (Rumbaugh, Jacobson, & Booch, 2000).

La Tabla 10 resume las características principales de los trece diagramas que ofrece UML clasificado en dos categorías:

- a) Diagramas de estructura: “representan elementos de un sistema o función y reflejan las relaciones estáticas de una estructura como los diagramas de clases, paquetes, componentes, despliegue o arquitectura en tiempo de ejecución, tales como los diagramas de objetos o estructuras de composición” (Quirón, 2005:4).
- b) Diagramas de comportamiento: “representan las características de comportamiento de un sistema o proceso de negocio como representan los diagramas de actividades, de comunicaciones (antes de colaboración), de revisión de la interacción, secuencias, de máquinas de estado, de tiempos y casos de uso” (Quirón, 2005:4).

Tabla 10. Diagramas UML

Diagrama	Vista	Características principales
Diagramas de estructura		
• Clases	- Estática y gestión del modelo	- Muestra una colección de elementos de modelado declarativo (estáticos), tales como clases, asociación, generalización, dependencia, realización e interfaz.
• Paquetes	- Implementación	- Presenta cómo se organizan los elementos de modelado en paquetes y las dependencias entre ellos, incluyendo importaciones y extensiones de paquetes y componentes.

Diagrama	Vista	Características principales
Componentes	- Implementación	- Representa los componentes de una aplicación, sistema o empresa. Los componentes, sus relaciones, interacciones e interfaces públicas.
Despliegue	- De despliegue	- Muestra cómo y dónde se desplegará el sistema. Las máquinas físicas y los procesadores se representan como nodos y la construcción interna puede ser representada por nodos o artefactos embebidos. Como los artefactos se ubican en los nodos para modelar el despliegue del sistema, la ubicación es guiada por el uso de las especificaciones de despliegue.
Objetos	- Estática y gestión del modelo	- Presenta los objetos y sus relaciones respecto al tiempo. Se considera un caso especial de un diagrama de clases o un diagrama de comunicaciones.
Estructuras de composición	- Implementación	- Representa la estructura interna de un clasificador (tal como una clase, un componente o un caso de uso), incluyendo los puntos de interacción de clasificador con otras partes del sistema.
Diagramas de comportamiento		
Actividades	- Actividades de cada proceso	- Representa los procesos de negocios de alto nivel, incluidos el flujo de datos (estado, actividad, transición de terminación, división y unión). También puede utilizarse para modelar lógica

Diagrama	Vista	Características principales
		compleja y/o paralela dentro de un sistema.
• De estados	- De máquinas de estados	- Ilustra cómo un elemento, muchas veces una clase, se puede mover entre estados que clasifican su comportamiento, de acuerdo con disparadores de transiciones, restricciones y explican el movimiento (estado, evento, transición y acción).
• Secuencias	- De interacción	- Representa una interacción, poniendo el interés central en la secuencia de los mensajes que se intercambian, junto con sus correspondientes ocurrencias de eventos en las líneas de vida (objetos, mensaje y activación).
• Comunicaciones	- De interacción	- Muestra la interacción entre líneas de vida, centrada en la arquitectura de la estructura interna y cómo ella se corresponde con mensajes. La secuencia de los mensajes se da a través de un esquema de numerado de la secuencia.
• Revisión de la interacción	- De interacción	- Revisa el flujo de control, donde los nodos son Interacciones u ocurrencias de interacciones. Las líneas de vida los mensajes no aparecen en este nivel de revisión (interacción, colaboración, rol y mensaje).
• De tiempos	- De interacción	- Muestra el cambio de estado de un objeto a lo largo del tiempo, en respuesta a los

Diagrama	Vista	Características principales
		eventos o estímulos aceptados. Los eventos que se reciben se anotan, a medida que muestran cuándo se desea mostrar el evento que causa el cambio en la condición o en el estado.
• Casos de uso	- Casos de uso o escenarios del sistema	- Se muestra las relaciones entre los actores y el sujeto (sistema), asociación, extensión, inclusión, generalización del caso de uso.

Fuente: Elaboración propia a partir de Quirón (2005).

Génova, Fuentes, & Valiente (2006) señalan, que debido a la popularidad que ha ido adquiriendo el modelado de sistemas SW con diagramas UML se encuentra vinculada a la existencia de herramientas CASE para UML que faciliten su gestión. Las herramientas UML se pueden clasificar siguiendo distintos criterios y los clasifican según su capacidad a la hora de dibujar los diagramas en gráfica, sintáctica o semántica.

- Herramientas gráficas son aquellas que proporcionan algún tipo de ayuda para dibujar diagramas UML con plantillas predefinidas, de modo que resultan más convenientes que otras herramientas genéricas de dibujo (como podría ser Microsoft PowerPoint). El problema es que no imponen prácticamente ningún tipo de restricción en lo que el usuario puede dibujar, de modo que se facilita la construcción de diagramas UML, pero no se garantiza en absoluto su corrección sintáctica. (Génova, Fuentes, & Valiente, 2006:59)

- Las herramientas sintácticas en general, sólo permiten dibujar diagramas correctos según las reglas notacionales de UML. Por tanto, suponen una ayuda mayor para el usuario respecto a las herramientas gráficas, sin embargo no construyen internamente un modelo a la vez que los diagramas que lo expresan, de modo que los diagramas quedan desconectados unos de otros. Por lo tanto impide comprobar la coherencia entre los distintos diagramas que expresan un mismo modelo desde distintos puntos de vista y se no deben considerarse herramientas CASE satisfactorias para UML. (Génova, Fuentes, & Valiente, 2006:60)
- Las herramientas semántica son aquellas que tratan de garantizar la construcción de un modelo que esté correctamente expresado en diagramas que además sean coherentes entre sí. Aunque en algunos casos la puntuación pueda ser inferior a la obtenida por herramientas de los otros grupos, es dentro de esta categoría donde se encuentra las que con propiedad pueden llamarse herramientas CASE para UML. (Génova, Fuentes, & Valiente, 2006:60)

Existen muchas herramientas de SW licenciadas y libres de costo para el diseño del modelado UML entre las más usuales se describen en la Tabla 11.

Tabla 11. Herramientas de diseño UML

Categoría	Herramienta
• Gráfica	• UML Diagrammer
• Sintáctica	• Microsoft Visio
	• Visible Analyst
	• Visual Paradigm
	• Umbrello UML Modeller

Categoría	Herramienta
• Semántica	• Magic Draw • Objecteering • seCAKE • Object Domain • Visual UML • Rational Rose • ArgoUML

Fuente: Elaboración propia a partir de Génova, Fuentes & Valiente (2006).

En el modelado UML para la construcción de la CVAATI se utiliza la herramienta ArgoUML, debido a que es un SW libre de costo y nos ofrece los beneficios de un diseño semántico.

2.6.4 Modelo MVC arquitectura en el Desarrollo de SW.

La arquitectura MVC Modelo - Vista - Controlador permite definir un patrón de arquitectura de desarrollo de software (SW) que propone separar el código de los programas por sus diferentes responsabilidades, es utilizada para implementar sistemas que requiere el uso de interfaces de usuario, crear software más robusto con un ciclo de vida más eficaz y que potencialice la facilidad de mantenimiento, reutilización del código y la separación de conceptos como se muestran en la Figura 7 (ApacheCTLdesarrolloweb, 2014: párr. 1).

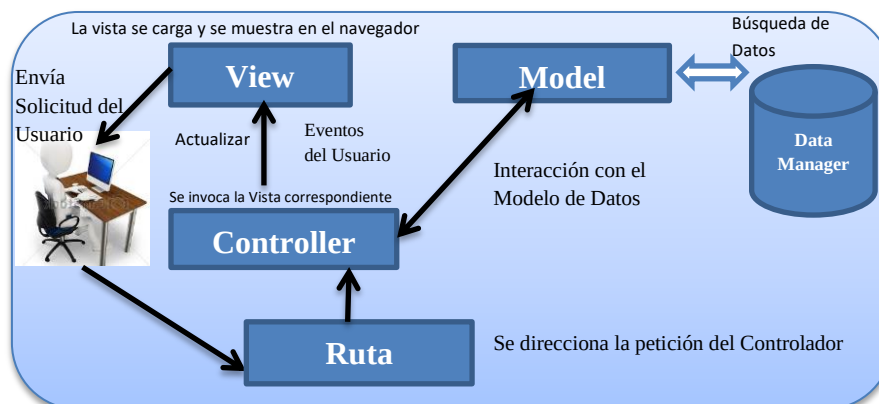


Figura 7. Modelo MVC.

Fuente: Elaboración propia.

Los tres componentes de esta arquitectura MVC se describen a continuación (ApacheCTLdesarrolloweb, 2014):

- El Modelo: es la capa donde se trabaja con los datos, por tanto contendrá mecanismos para acceder a la información y también para actualizar su estado. Los datos están almacenados en una DB, por lo que en los modelos deben tener todas las funciones que accederán a las tablas en cuanto a la ejecución de select, update, insert, etc. Además envía a las vistas la información que solicita el usuario y las peticiones de acceso o manipulación de la información llegan a través del controlador.
- El Controlador: esta capa sirve de enlace entre la vista y el modelo de acuerdo a las requerimiento de desarrollo de SW, ya que contiene el código necesario para responder a las acciones que solicita la aplicación como por ejemplo, visualizar un elemento, búsquedas de información, realizar una compra, entre otras.

- La Vista contiene el código de la aplicación o sistema de información que presentan la interfaz del usuario, por lo tanto requiere de la información que le envía el modelo para presentar como salida.

A continuación se resume el flujo de control de la interacción de componentes, considerando las diferentes implementaciones de MVC, que se publica en ApacheCTLdesarrolloweb (2014):

- “El usuario realiza una solicitud a un sitio web. Generalmente estará desencadenada por acceder a una página del sitio. Esa solicitud le llega al controlador” (Sexta Sección, párr. 3).
- a) “El controlador comunica al modelo y a las vistas. A los modelos les solicita datos o realiza las actualizaciones. A las vistas les solicita la salida correspondiente, una vez realizado las operaciones pertinentes según la lógica del negocio” (Sexta Sección, párr. 4).
- “Para producir la salida, las vistas pueden solicitar más información a los modelos. En ocasiones, el controlador será el responsable de solicitar todos los datos a los modelos y de enviarlos a las vistas, actuando de intermediario entre ellos” (Sexta Sección, párr. 5). Esto sucede dependiendo de la implementación de software que se realiza; como se visualizan en las flechas de la Figura 7.
- “Las vistas envían al usuario la salida. Aunque en ocasiones esa salida puede ir de vuelta al controlador y sería éste el que hace el envío al cliente” (Sexta Sección, párr. 6), como se indica en las flechas de la Figura 7.

Se presentan algunos ejemplos que visualizan las ventajas de MVC como se menciona en MasterWeb.La (2015):

- Aunque no tenga nada que ver, iniciar con algo tan sencillo como son el HTML y las CSS. Al principio, en el HTML se mezclaba tanto el contenido como la presentación. Es decir, en el propio HTML se tienen etiquetas como "font" que sirven para definir las características de una fuente, o atributos como "bgcolor" que definen el color de un fondo. El resultado es que tanto el contenido como la presentación estaban juntos y si algún día pretende cambiar la forma con la que se mostraba una página, se está obligado a cambiar cada uno de los archivos HTML que componen una web, tocando todas y cada una de las etiquetas que hay en el documento. Con el tiempo se observó que eso no era práctico y se creó el lenguaje CSS, en el que se separó la responsabilidad de aplicar el formato de una web. (Sección 2, párr. 3)
- Al escribir programas en lenguajes como PHP, cualquiera que comience mezclando tanto el código PHP como el código HTML (e incluso el Javascript) en el mismo archivo. Esto produce lo que se denomina el "Código Espagueti". Si algún día pretende cambiar el modo en cómo se debe mostrar el contenido, se está obligado a repasar todas y cada una de las páginas que tiene el proyecto. Sería mucho más útil que el HTML estuviera separado del PHP. (Sección 2, párr. 4)
- Si se quiere que en un equipo intervengan perfiles distintos de profesionales y trabajen de manera autónoma, como diseñadores o programadores, ambos tienen que tocar los mismos archivos y el diseñador se tiene necesariamente que relacionar con mucho código en un lenguaje de programación que puede no serle familiar, siendo que a éste quizás solo

le interesan los bloques donde hay HTML. De nuevo, sería mucho más fácil la separación del código. (Sección 2, párr. 5)

- Durante la manipulación de datos en una aplicación es posible que este accediendo a los mismos datos en lugares distintos. Por ejemplo, acceder a los datos de un artículo desde la página donde se muestra éste, la página donde se listan los artículos de un manual o la página de backend y se administran los artículos de un sitio web. Si un día cambian los datos de los artículos (se modifica la tabla para añadir nuevos campos o cambiar los existentes porque las necesidades de los artículos varían), se estaría obligado a cambiar, página a página, todos los lugares donde se consumían datos de los artículos. Además, si tiene el código de acceso a datos disperso por decenas de lugares, es posible que se repitan las mismas sentencias de acceso a esos datos y por tanto no se reutilice el código. (Sección 2, párr. 6)

Adicionalmente la estructura de programación MVC con Laravel incluye el concepto de Routing para direccionar o enrutar la petición del Controlador.

CAPÍTULO 3. METODOLOGÍA

3.1 Fundamentos de la Investigación

Al iniciar una investigación siempre se debe definir los objetivos claros, direccionados a la(s) pregunta(s) de investigación, hipótesis, alcances y delimitaciones, así como identificar que las fuentes de recopilación de datos estén disponibles o en caso contrario, considerar las posibles restricciones que se pueden presentar para no detener esta tarea de campo donde se desarrollará. Remarcando la diferenciación entre táctica y estrategia. La estrategia tiene que ver con el enfoque general que se adoptará, mientras que la táctica es sobre el cómo los datos serán recolectados (métodos de recolección de datos: cuestionarios, entrevistas, grupos de discusión...) y cómo serán analizados (Saunders, Lewis, & Thornhill, 2009).

La estrategia de esta investigación tomará como base el proceso de investigación denominado “cebolla”, planteado por Saunders 2009, en su libro: “Research Methods for Business Students”, mostrada en la **Figura 8**.

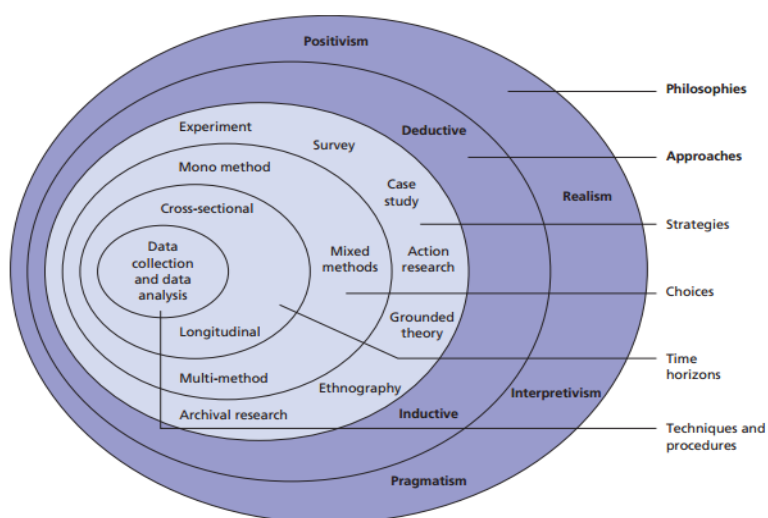


Figura 8. Estrategia de Investigación.

Fuente: Saunders (2009).

Como se observa el proceso “Cebolla” incluye en su filosofía de investigación (metodología científica) cuatro enfoques que dominan la literatura acerca del tema: positivismo, realismo, interpretativismo, y pragmatismo; dos métodos científicos deductivos e inductivos; y seis capas: filosofía, propósito, estrategias, elecciones, horizontes de tiempo, técnicas de recolección de datos y procedimientos de análisis. En esta estrategia el desarrollo del conocimiento en una investigación depende de la forma en que el investigador concibe el cómo responda a sus preguntas de investigación, donde rara vez cae perfectamente en un solo dominio y puede ir desprendiendo las capas que no apliquen (Saunders, et al., 2009).

En el desarrollo de una metodología de estudio, además de la estrategia “Cebolla”, también debe considerar los siete tipos de investigación que proponen Hernández, Fernández & Baptista (2010): Exploratorio, Cuasi experimental, Transaccional, Descriptiva, Correlacional, Documental y Análisis – síntesis; con el propósito de identificar cuál es la que se adapte al trabajo que se pretende investigar.

3.2 Diseño de la Investigación

La investigación presentada se divide en tres fases principales: la primera es, analizar los estudios desarrollados relacionados al fenómeno de la deserción y/o abandono escolar y con base en ello, continuar en una segunda etapa que es el diseño de la interfaz de la comunidad virtual académica para aplicar las fases precedentes a DM; mientras que, en la tercera etapa es, el diseño propio del modelo Predictivo con DM y RNA.

La investigación, se inicia utilizando el método científico deductivo, tomando como base la estrategia “cebolla” propuesta por Saunders et al. (2009) y un análisis de tipo exploratorio,

dado que ha sido poco abordada la aplicación de modelos de DM o EDM en este fenómeno del ámbito educativo en relación al desarrollo aplicado en áreas de medicina, negocios, mecatrónica, sistemas de decisiones entre otras (Calders & Pechenizkiy, 2011); incluyendo también un estudio correlacional y explicativo en las etapas subsecuentes, como lo describen Hernández et al. (2010), que identifique los factores y variables que inciden con mayor frecuencia en los indicadores de desempeño (deserción, reprobación y eficiencia terminal) que se estudiaron sobre los modelos predictivos relacionados a esta problemática. Además se basa en el enfoque positivista y realista de la estrategia de “cebolla” ya que se trabajará con fuentes de datos existentes de la IES prototipo (plataformas educativas, Sistemas de información de datos personales, académicos, entre otros), en tanto que el interpretativista sería obtener estadísticas de los resultados obtenidos de los dos enfoques y el tratamiento de la información con métodos, técnicas y algoritmos de DM para la extracción del conocimiento.

Señalando que el diseño de la CVA y del modelo Predictivo PredATIS, se realiza a través de la metodología propuesta en la **Figura 9**, denominada MSWD&DM por sus siglas en inglés (Methodology for Design Software and Data Mining) que combina la Ingeniería de Software (SW) y las fases del proceso de KDD que provee una representación completa del ciclo de vida de un proyecto **DM**, esto con la finalidad de cumplir el propósito de esta investigación.

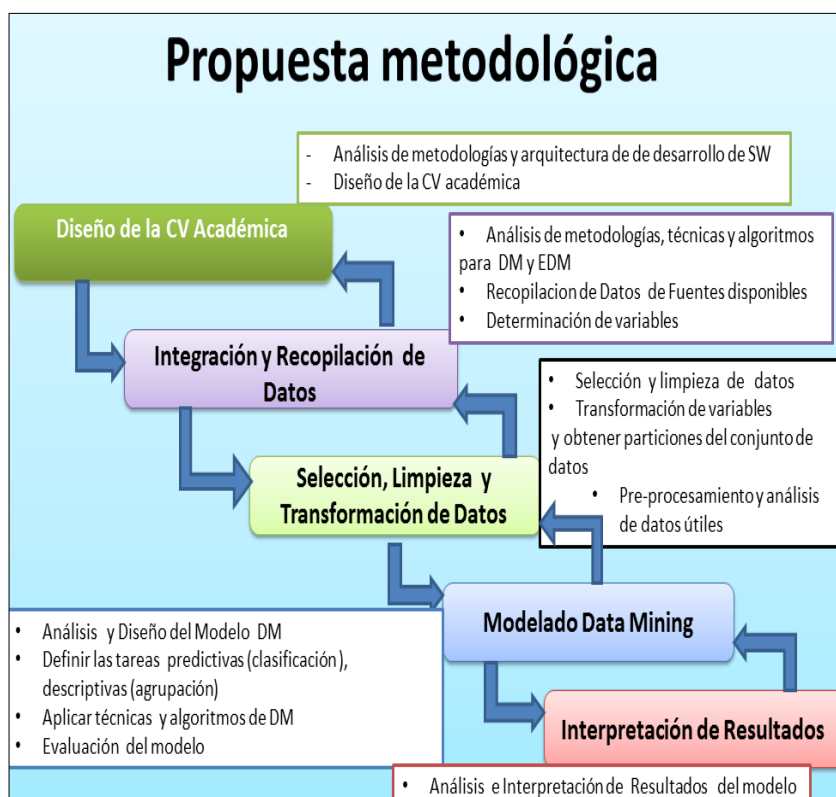


Figura 9. Metodología de Investigación MSWD&DM Propuesta.

Fuente: Elaboración Propia.

Se observa que la primera fase de la metodología descrita en la **Figura 9**, corresponde al análisis de metodologías y arquitectura MVC para el desarrollo del SW que se utilizan en el diseño de la comunidad virtual académica y a partir de la segunda fase representan el proceso de KDD que se aplicará en el diseño del modelo Predictivo propuesto.

3.3 Hipótesis

- El modelo PredATIS permite generar información útil de la comunidad virtual académica, que coadyuvará positivamente para alertas tempranas en los indicadores de educativos de la IES seleccionada.

- La comunidad virtual Académica CVAATI es una herramienta tecnológica que genera conocimiento a partir de los datos pre procesados, para el análisis de estudiantes en riesgo de deserción, reprobación de la Institución en estudio.

3.4 Selección de la muestra

Considerando la complejidad de trabajar una investigación en una población extensa como es el caso de una comunidad Educativa de Nivel Superior donde se ofrecen 6 diferentes perfiles académicos en sus áreas de Ingeniería entre las que se mencionan: Tecnologías de la Información y Comunicación en sus dos áreas de Técnico Superior Universitario (TSU) en Sistemas de Información y Redes y Telecomunicaciones, Mantenimiento Industrial, Tecnología Ambiental entre otras.

Tomando como base, estos espacios muestrales, se selecciona de forma no probabilística, como lo describe Hernández et al. (2010), donde los elementos seleccionados de la muestra no dependen de la probabilidad, sino de causas relacionadas con las características de la investigación o de quien la desarrolla. Señalando que la ventaja de utilizar muestras no probabilísticas es su utilidad para determinados tipos de diseños de investigación, los cuales no requieren tanto una representatividad de los elementos de su población, sino una cuidadosa y controlada elección de sujetos con ciertas características especificadas previamente en el planteamiento del problema. Para el objeto de estudio se tomará como muestra representativa a los estudiantes de una generación de Técnico Superior Universitario en TIC, donde se conoce el entorno y elementos esenciales que favorecen realizar esta investigación y sea posible aplicar el modelo PredATIS propuesto con datos de la CV académica.

3.5 Recolección de datos

Selección del instrumento de recopilación de información, medición y validación.

Según Hernández et al. (2010) un Instrumento de Medición es un recurso que utiliza el investigador para registrar información o datos sobre las variables que tiene en mente, que registra datos observables que representan verdaderamente los conceptos o variables a investigar. La medición es el proceso de asignar números o marcadores a objetos, personas, estados o hechos, según las reglas específicas para representar la cualidad (cualitativo) o cantidad (cuantitativo) de un atributo. En todo Instrumento de Medición debe reunir los tres requisitos esenciales: validez, confiabilidad y objetividad.

En la investigación que se presenta, se propone las siguientes fuentes para la obtención de la información:

- a) Primarias: Plataformas educativas, sistemas de control escolar, sistemas de información de datos personales y reportes estadísticos de estudiantes, reportes de control de notas de estudiantes por periodo escolar, información complementaria manejada por personal académico de la IES en estudio.
- b) Secundarias: sitios web oficiales, libros y revistas científicas tecnológicas, banco de datos de organismos oficiales.
- c) Terciarias: artículos científicos publicados en la web. Esto es solo para fines de análisis de técnicas y algoritmos de DM y RNA para el diseño del modelo PredATIS propuesto y las metodologías de desarrollo de SW en la construcción de la CVAATI.
- d) Recolectados: encuestas online para recopilar datos complementarios de otros factores que miden el desempeño del estudiante.

Se eligió utilizar encuestas ya que es una estrategia que Saunders et al. (2009) asocia con el método deductivo, utilizado en la investigación para identificar factores asociados directa o indirectamente al rendimiento y desempeño del estudiante, mismo que se aplica en alumnos de nuevo ingreso como al inicio de cada período escolar y la gestión que permite la recolección de una gran cantidad de datos a partir de la muestra descrita en la sección 3.3, la obtención de los datos se realizará de forma estandarizada, procurando la existencia de una comparación más estructurada de los datos que es de fácil aplicación y comprensión además de que puede dar un control sobre el proceso de la investigación. Considerando tener como base diferentes análisis comparativos que permitan recolectar información tanto cuantitativa como cualitativa. Además de las técnicas y algoritmos DM y RNA para la validación que apoye en la fase de Integración y Recopilación de Datos de la metodología descrita.

Aplicación del Instrumento de selección. El diseño y creación del instrumento de una encuesta individual, que se utiliza como base de datos, a través de un sitio web propio de la IES, donde se incluye preguntas cerradas, dicotómicas, de opción múltiple y mixtas (Saunders et al. (2009). Siguiendo las recomendaciones del manual de Oslo, que diseñó un cuestionario completo (OECD, 2005), con la finalidad de recopilar datos de la muestra sobre factores o variables que incluyen: datos generales, aspectos socioeconómicos, de salud, personales, vocacionales, académicos, laborales y adiciones de los estudiantes, generando de forma sistemática archivos Excel (.CSV), de texto (TXT) y SQL, así como la generación de un kardex e historial académico de cada período escolar; siendo la fuente principal para la recopilación de datos del modelo en la CVA y así continuar con las fases descritas en la metodología propuesta en la

sección 3.2 para identificar su participación en el rendimiento y desempeño del estudiante, mismo que se aplica en dos períodos escolares de la muestra seleccionada.

Codificación. Los instrumentos de medición se diseñan con el propósito de incluir solo los criterios de evaluación que sean útiles para la investigación y en cuanto a los datos que deben ser recopilados, pre-procesados y procesados conforme a los requerimientos de las fases de DM.

Preparación de datos. Una vez recopilados los datos en la comunidad virtual académica desde la fuentes disponible, se procede conforme a las fases de tratamiento de datos KDD, donde primero pasan por la fase de selección y limpieza de datos, transformarlos a variables, obtener particiones del conjunto de datos y realizar un pre-procesamiento y análisis de datos útiles para el procesamiento de Minería de Datos que deben cumplir con base a los requerimientos de la aplicación de técnicas y algoritmo de DM y RNA seleccionados, prácticamente aquí se trabaja con la creación y diseño del Modelo de DM donde se realiza el: análisis y diseño prototipo del Modelo, se definen las tareas descriptivas (agrupamiento, correlaciones y reglas de asociación) como predictivas (clasificación), así como de algoritmos de selección y búsqueda de mejores atributos con DM aplicando técnicas de árboles de decisión y sistemas de aprendizaje automático supervisado y no supervisado y RNA; donde se realizan en dos vertientes: una con datos de entrenamiento y otra con datos de prueba apoyándose con la suite de Minería de Datos WEKA y Matlab, para ir evaluando y controlando el balanceo de datos útiles para la construcción del modelo y ejecutarlo con el software desarrollado e incluido en la CVAATI propuesta, realizando comparativas de evaluación entre ellos, con el objeto de encontrar los factores o variables de mayor impacto en el fenómeno en estudio de esta

investigación. Por otro lado también se puntualiza que pueden ir retrocediendo a las fases anteriores la metodología hasta lograr todo un procesamiento de datos útiles y evaluando resultados hasta generar un modelo adaptativo a la CVA propuesta, para su análisis e interpretación de resultados obtenidos, que apoyen y faciliten información predictiva y emitir reportes de alerta temprana coadyuvando positivamente en los estadísticos educativos de la IES en estudio.

Análisis de datos. El análisis de los datos que se tengan en la CVA, una vez procesados con las de técnicas y algoritmo de DM y RNA seleccionados, se procede conforme a las fases de tratamiento del modelo implementado, con la evaluación, análisis e interpretación de resultados del modelo, introducirlo a la CVA hasta donde sea posible su difusión y uso a futuro, para la predicción de alertas tempranas de estudiantes con alto riesgo a deserción y reprobación. Ya que esto no depende de modelo directamente sino más bien de la gestión propia y comunicación estratégica de la CVA en donde se pretende implementar y que quedan fuera del alcance de esta investigación.

CAPÍTULO 4. DESARROLLO DEL MODELO PROPUESTO EN LA COMUNIDAD VIRTUAL ACADÉMICA

4.1 Diseño de la Comunidad Virtual Académica CVAATI

El diseño de la CVAATI se guía con la metodología MSWD&DM propuesta en la Figura 9 en su primera fase, realizando un análisis de metodologías y arquitectura MVC para el desarrollo del SW que mejor se adaptó al desarrollo tecnológico de esta investigación. Derivado del análisis realizado en la Tabla 8 donde se presenta una comparativa entre modelos tradicionales y metodologías pesadas con respecto a las metodologías ágiles, se observa que las metodologías ágiles ofrecen más ventajas, por lo que se selecciona la metodología del Proceso Unificado Ágil (AUP), que es una versión simplificada de RUP siendo la más utilizada en grandes proyectos de desarrollo de SW y con base a la naturaleza y tamaño del proyecto que se presenta en la investigación. Señalando que AUP tiene las mismas cuatro fases que RUP con la diferencia en que no está tan centrado en la documentación, reteniendo parte de la formalidad y combina técnicas ágiles con el propósito de mejorar la productividad. La gestión resulta sencilla, aun cuando algunos desarrolladores no lo utilizan, debido al gran número de artefactos que se requiere (Torrecilla, 2012). En este sentido se describen las cuatro fases de AUP que guiaron el diseño de la CVAATI en la Tabla 12.

Tabla 12. Fases del ciclo de desarrollo AUP utilizadas en la CVAATI

Fase de AUP	Descripción de objetivos y tareas principales e hitos
Iniciación	Se define el alcance del proyecto respecto al desarrollo tecnológico que incluye el diseño de una interfaz de la CVAATI y la programación únicamente de las técnicas, métodos y algoritmos de DM que se

Fase de AUP	Descripción de objetivos y tareas principales e hitos
	utilizarán para el diseño del Modelo PredATIS como se describe en la sección 1.7 Alcances y limitaciones. • Se diseña el plan de desarrollo del proyecto que se muestra en la Tabla 13, tomando como base la metodología MSWD&DM e identificando los recursos tecnológicos con respectos al Hardware (HW) y SW tanto para desarrollo como para la implementación, el tiempo asignado y la capacidad en el manejo de las tecnologías que se utilizan en el desarrollo de esta investigación, así como el presupuesto asignado y determinar la viabilidad del proyecto descrito en la sección 1.6. • Definición de Riesgos en el que se identifican tiempo asignado, el nivel de conocimiento en las herramientas tecnológicas viables, disponibilidad y acceso a la información escolar para la recopilación e integración en el desarrollo del proyecto. • Preparación del entorno del proyecto con base al análisis de modelos desarrollados en la predicción de la deserción escolar y rendimiento académico en diferentes niveles educativos presentados en la Tabla 6 metodologías de ingeniería de desarrollo SW descritas en la Tabla 9 y patrones de diseño más utilizados, flexibles y dinámicos.
<i>Hito: Objetivos del ciclo de vida (LCO) del Diseño de CVAATI.</i>	
Elaboración	• Se identifica y valida la arquitectura de desarrollo seleccionando la metodología ágil AUP y el patrón de diseño MVC para el desarrollo de

Fase de AUP	Descripción de objetivos y tareas principales e hitos
	la CVAATI con base al análisis presentado en la sección 2.6.4 Modelo MVC arquitectura en el Desarrollo de SW. • Para la ejecución del entorno del proyecto se guía con el plan de desarrollo que indica las principales actividades y el tiempo estimado. • Se determina que el equipo de desarrollo es de forma individual, dada la naturaleza del proyecto que se realiza en este trabajo, en tanto que la implementación del modelo se pretende la participación de personal académico de la IES en el que se aplica y la investigación que es el centro de interés de este trabajo, involucra al comité de dirección de tesis de UPAEP.
	<i>Hito: Arquitectura del ciclo de vida (LCA) del Diseño de CVAATI.</i> 
Construcción	• Modelado del desarrollo del SW: se plantea el diseño de la CVAATI inicialmente con el Lenguaje de Modelamiento Unificado (UML) utilizando el SW de argoUML, siendo la base para la construcción y que se describe a detalle en la sección 4.1.1 Modelado UML y arquitectura de Desarrollo de la CVAATI. • Construcción de CVAATI tomando como base el modelado UML se seleccionan herramientas open source para el desarrollo considerando el patrón de diseño MVC realizado con lenguaje de programación orientado a objetos (POO) en PHP puro, herramientas tecnológicas open source como Laravel, Composer, A(POO) en PHP puro Ajax, JQuery,

Fase de AUP	Descripción de objetivos y tareas principales e hitos
	Json, Bootstrap, HTML5, CSS3, Python y el gestor de base de datos de MySQL, como se presenta en la sección 4.1.2 Construcción de la CVAATI.desarrollo de CVAATI. ↓ <i>Hito: Capacidad de operación inicial (IOC) del Diseño de CVAATI</i>
Transición	• Preparación del diseño del Modelo PredATIS a través de la programación de las interfaces que integrarán las técnicas, métodos y algoritmos de DM en la CVA. • Las actividades adicionales que se enumeran a continuación, no se realizan propiamente en el diseño de la CVVATI, ya que se incluyen en el desarrollo planteado en la metodología MSWD&DM a partir de la segunda fase como se indica en la sección 3.2 Diseño de la Investigación que se presenta en este trabajo. • Programación e integración de técnicas, métodos y Algoritmos DM en la CVA. • Validación del diseño PredATIS en la CVA con la ejecución de pruebas de funcionalidad y usabilidad. • Despliegue de Resultados: generación del conocimiento para la interpretación con DM y la emisión de alertas de indicadores educativos. <i>Hito: Diseño del Modelo PredATIS</i> ↓ Entregable Final: <i>Implementación de PredATIS en la CVAATI</i>

Fuente: Elaboración Propia a partir de Torrecilla (2012).

Plan de desarrollo. Se diseña con base en la metodología MSWD&DM propuesta en esta investigación mostrada en la **Figura 9**, donde se combinan la metodologías de ingeniería de SW para el desarrollo de la CVAATI y se aplican las técnicas, métodos y algoritmos de DM y RNA para el diseño del modelo PredATIS. El plan que se describe procura atender principalmente los objetivos y actividades planteadas de cada fase de desarrollo, así mismo se asigna el tiempo que sea factible en su cumplimiento, señalando que pueden adaptarse a los requerimientos en cada etapa de desarrollo del proyecto de investigación.

Tabla 13. Plan de desarrollo

Actividades	Fecha de realización
Diseño de la investigación	Mayo – Julio 2016
Diseño de la CV Académica	
• Análisis de metodologías y arquitectura de desarrollo de SW	Agosto – Diciembre
• Diseño de la CV académica • Modelado UML de CVAATI • Diseño de la interfaz inicial de CVAATI • Diseño de interfaz inicial para la integración de datos y técnicas DM en la CVAATI.	2016 Enero-Agosto 2017

Integración y Recopilación de Datos en la CVAATI

- Análisis de metodologías, técnicas y algoritmos para DM y RNA. Julio 2017
- Recopilación de Datos de Fuentes disponibles de la IES
- Determinación de variables

Selección, Limpieza y Transformación de Datos

- Selección y limpieza de datos Julio 2017
- Transformación de variables y obtener particiones del conjunto de datos
- Pre procesamiento y análisis de datos útiles en la CVAATI. Agosto 2017
- Programación de interfaz inicial y de la interfaz para la integración de Datos a la CVAATI

Modelado de PredATIS

- Análisis y Diseño del Modelo DM y RNA
 - Definición de tareas predictivas (clasificación) y descriptivas (agrupación) Septiembre – Octubre
 - Programación y Aplicación de técnicas y algoritmos de DM y RNA del Modelado con SW de apoyo 2017
 - Programación de la interfaz de pre-procesamiento DM en la CVAATI Noviembre – 09 Diciembre
 - Programación de interfaz inicial del módulo para el modelado 2017
-

de PredATIS en la CVVATI.

- Evaluación del modelo PredATIS
 - Programación y evaluación del modelado PredATIS en la CVA
- 15 Enero –
08 Abril'2018

Interpretación de Resultados

- Análisis e Interpretación de Resultados del modelo
PredATIS
 - Conclusión de Resultados de investigación
- 09 Abril –
12 Mayo'2018
-

Fuente: Elaboración propia.

4.1.1 Modelado UML y arquitectura de Desarrollo de la CVAATI.

En la fase de elaboración y construcción de la CVAATI basado en metodología AUP seleccionada, se realiza primero el diseño, antes de iniciar propiamente el desarrollo de este SW, cuya intención es generar una herramienta tecnológica de calidad y funcionalidad a la que está destinada. El diseño se realizó a través del Lenguaje Unificado de Modelado por sus siglas en inglés (UML Unified Modeling Language). UML es un lenguaje de modelado orientado a objetos, es aceptado por la ingeniería del SW, que apoya a las organizaciones en grandes proyectos en el desarrollo de SW; además se ha adoptado como un estándar internacional, que permite visualizar, especificar, construir y documentar un sistema de SW, es decir describir un plano de la CVAATI, incluyendo aspectos conceptuales como procesos de negocios, funciones

del sistema y lenguaje de programación, esquemas de bases de datos y componentes de SW reutilizables (Fernández & García, 2008).

En el modelado de la CVAATI se utilizó ArgoUML que es un SW open source y que se adapta a la metodología y arquitectura de desarrollo seleccionada, donde se realizan cinco diagramas de los trece que ofrece UML (diagrama de clases, paquetes, componentes, estructura de composición, despliegue, objetos, secuencia, casos de uso, actividades, de estado, de comunicaciones, de interacción y de tiempo), derivado del objetivo de este trabajo del diseño de modelo PredATIS en la CVAATI no, de la construcción del SW; por lo tanto se presentan tres diagramas estructurales y dos diagramas de comportamiento. Los diagramas estructurales representan los elementos o componentes de la CVAATI que reflejan las relaciones estáticas de una estructura, como lo hacen los diagramas de clases, despliegue y de paquetes categorizados con prioridad alta, media y baja en el desarrollo de SW. Los diagramas de comportamiento representan las características de un sistema o procesos de la CVAATI como es el diagrama de secuencia y casos de uso categorizados con prioridad alta y media (Fernández & García, 2008). A continuación se presentan los cinco diagramas de despliegue, de paquetes, de clases, secuencia y casos de uso.

El Diagrama de Despliegue consiste en modelar los distintos componentes físicos utilizados en la implementación y la relación entre ellos de la CVAATI, como se visualiza en la **Figura 10**.

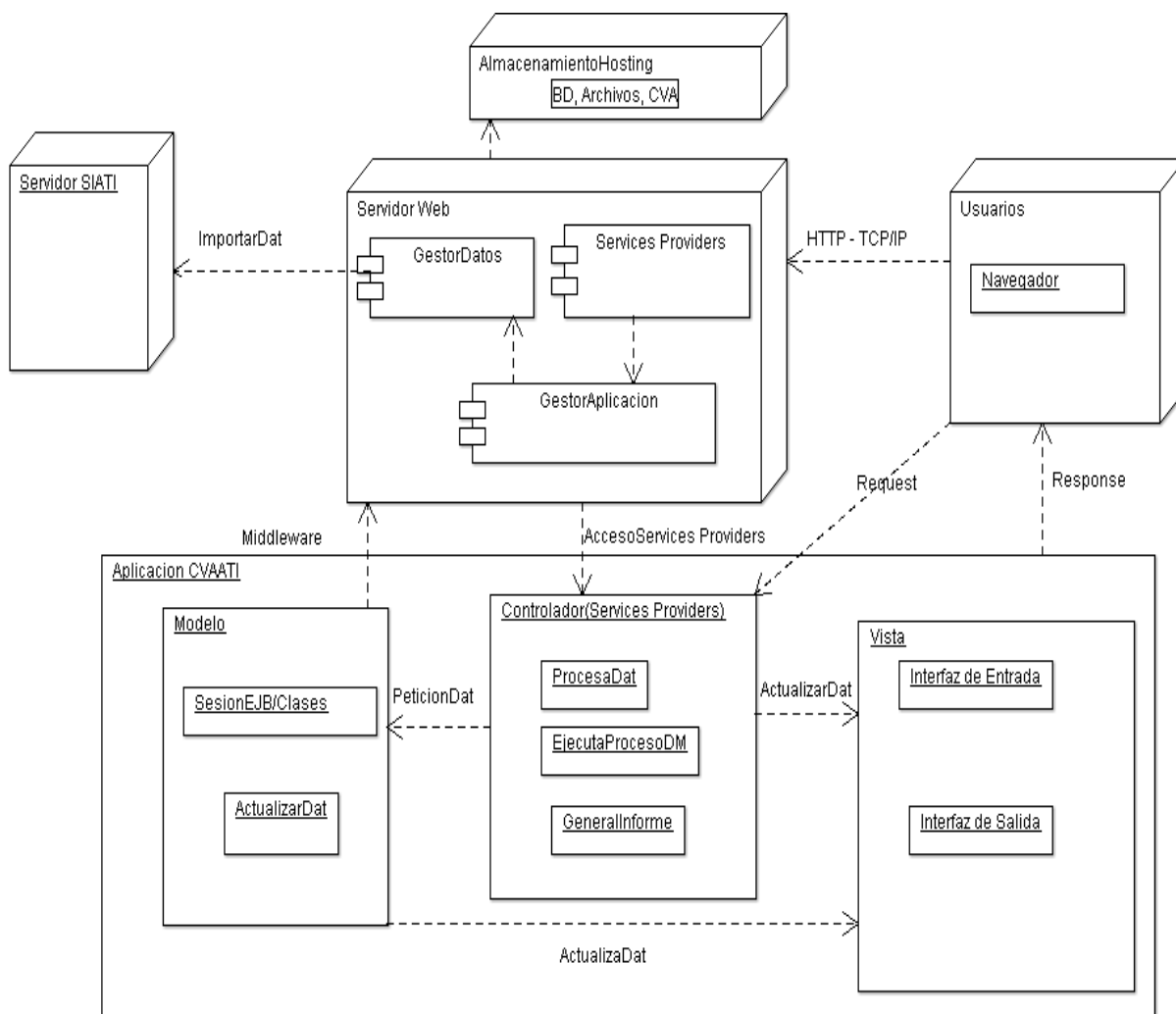


Figura 10. Diagrama de Despliegue.

Fuente: Elaboración Propia.

En el Diagrama de Paquetes se muestran los diferentes subsistemas lógicos en que se divide la CVVATI y las dependencias que existen entre ellos tomando como base la arquitectura MVC de desarrollo del SW, como se presenta en la **Figura 11**.

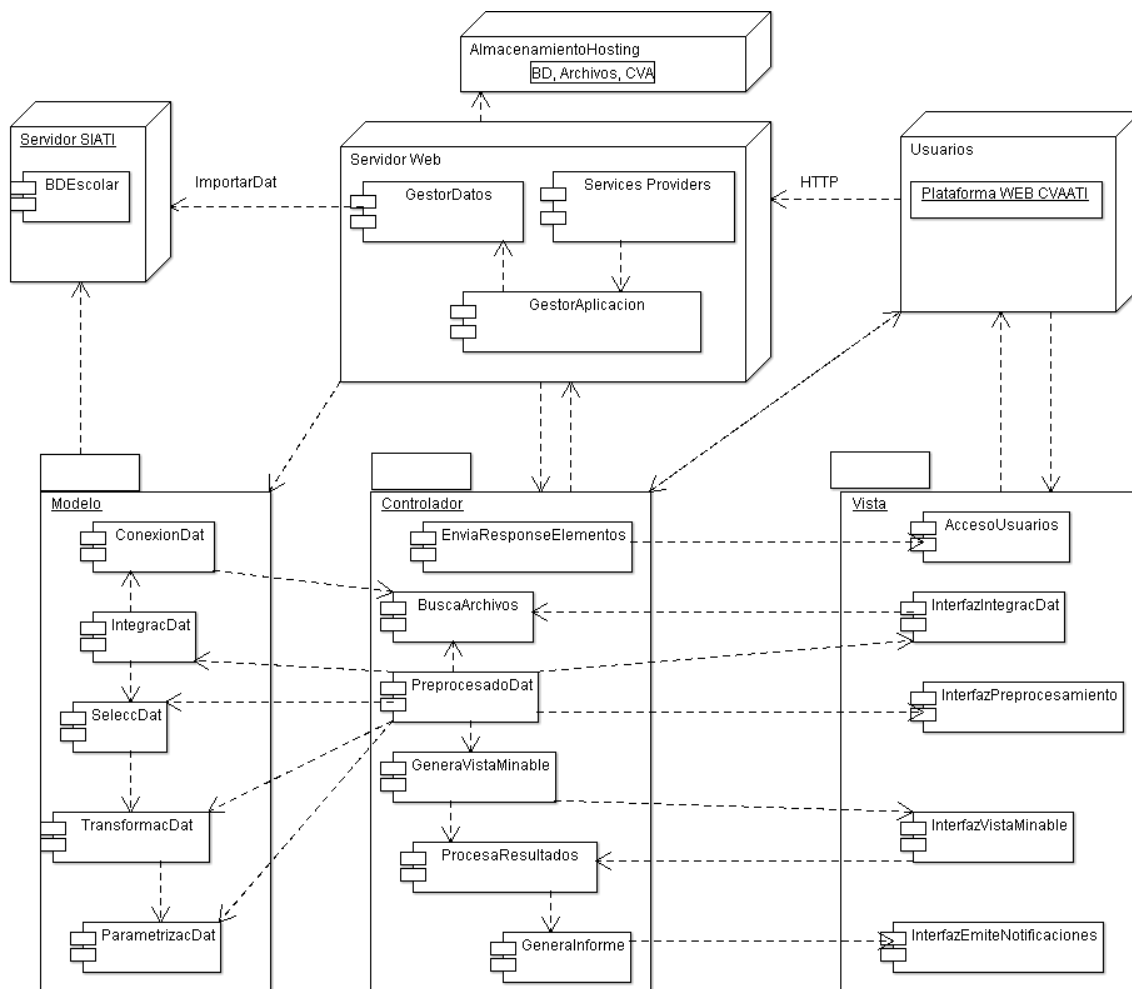


Figura 11. Diagrama de Paquetes.

Fuente: Elaboración Propia.

En el **Diagrama de Clases** se especifican los diferentes tipos de objetos relevantes al dominio de aplicación y la relación entre ellos dentro de la CVAATI, como se describe en la **Figura 12**.

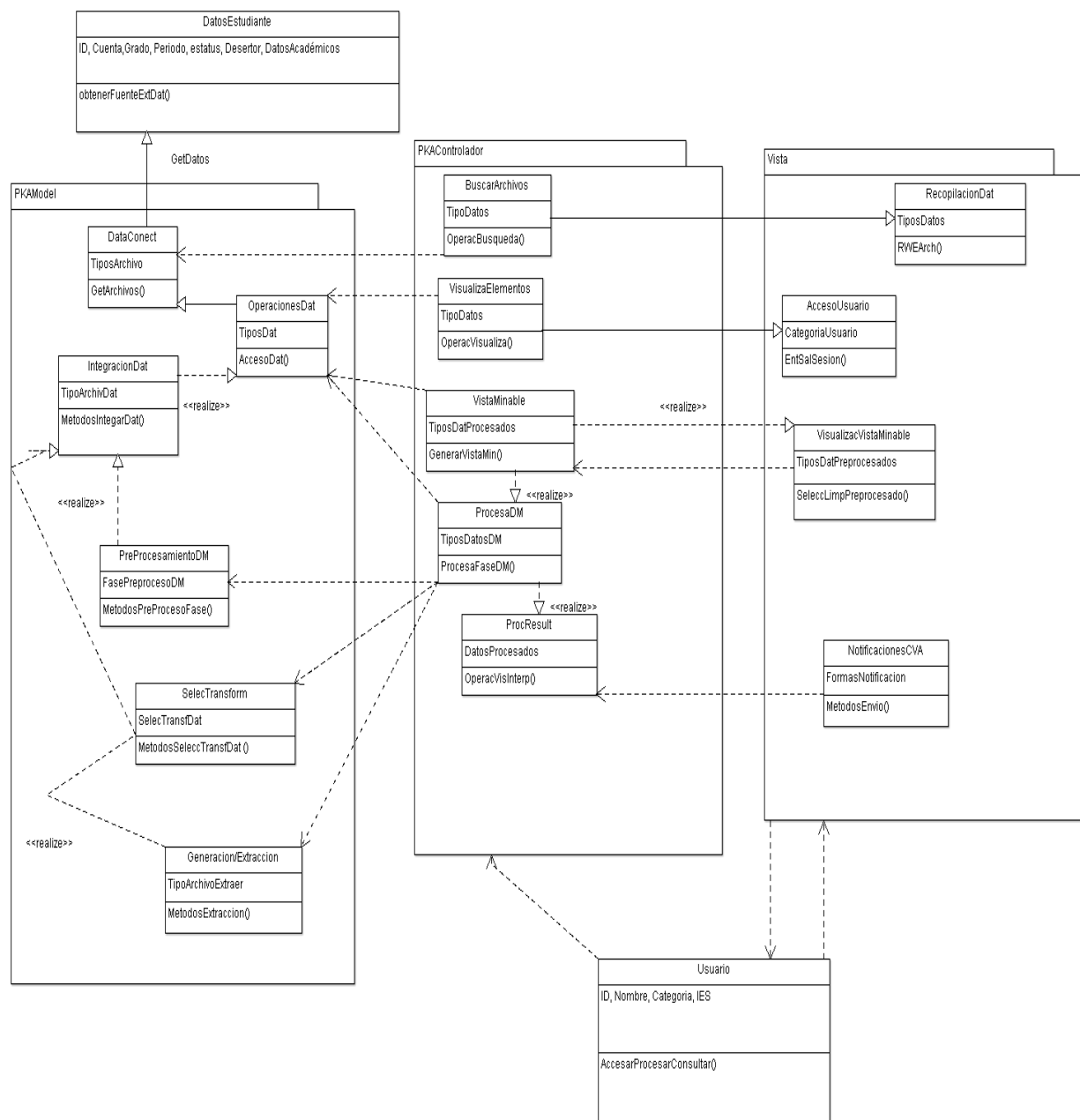


Figura 12. Diagrama de Clases.

Fuente: Elaboración Propia.

El **Diagrama de Secuencia** muestra la evolución en el tiempo de mensajes que intercambian los diferentes actores y objetos en un determinado escenario que se presentan en el diseño de la CVAATI, que se visualiza en la **Figura 13**.

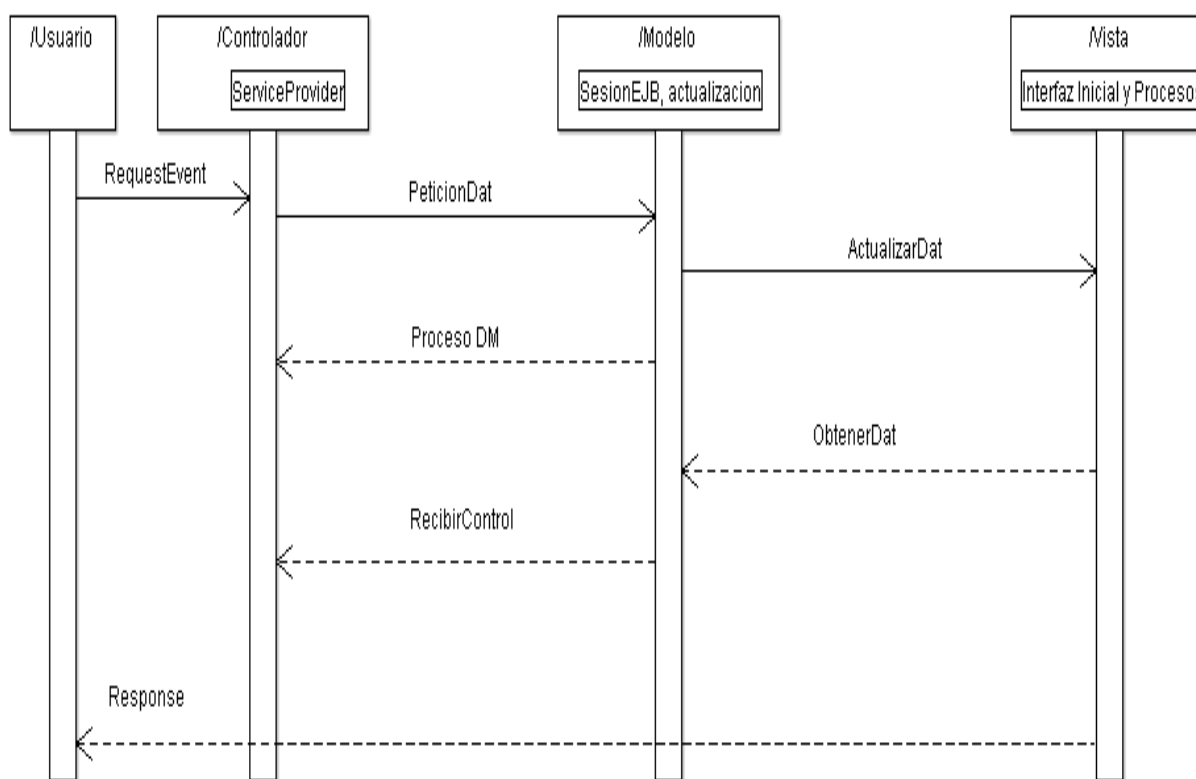


Figura 13. Diagrama de Secuencia.

Fuente: Elaboración Propia.

El **Diagrama de Casos de Uso** muestra las relaciones o interacción entre los actores y la CVAATI que participan en su implementación conjunta del Modelo PredATIS como se describe en la Figura 14.

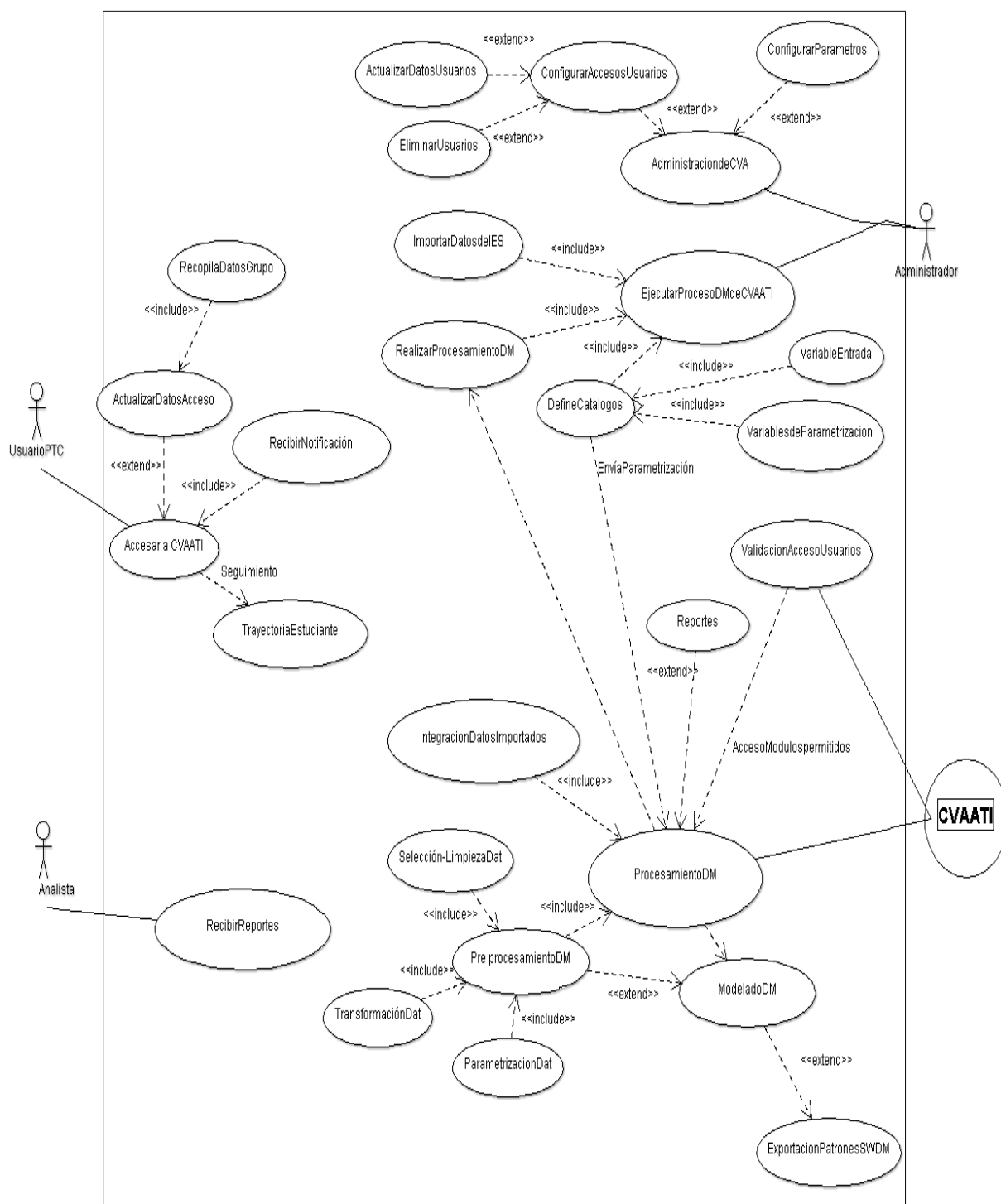


Figura 14. Diagrama de Casos de Uso.

Fuente: Elaboración Propia.

4.1.2 Construcción de la CVAATI.

En la construcción de la CVAATI se requiere un entorno flexible de desarrollo que integre tecnologías de código abierto, que en conjunto permita automatizar el proceso de recopilación de datos (importación o extracción), pre procesamiento, modelado y reportes de predicción de estudiantes probables a deserción. De esta forma se crea un sistema con arquitectura Modelo Vista Controlador (MVC), a través de una programación orientada a objetos (POO) en PHP puro, así como de otras herramientas tecnológicas open source como Laravel, Composer, Ajax, Jquery, Json, Bootstrap, HTML5, CSS3, Python y NuSoap. Adicionalmente se cuenta con una base de datos en MySQL, que auxilia el intercambio de información entre los diferentes procesos requeridos por la CVAATI, que permite almacenar la información clave y con acceso a un sitio web publicado en esta fase en <http://proyectototicsmn.com/cvaati01/public/index.php>.

De lo anterior con la finalidad de crear una CVAATI que permite la integración de datos de fuentes externas tal como los Sistemas de Información que maneja actualmente la IES en estudio, realizar el pre procesamiento de DM que incluye selección y limpieza, transformación y clasificación de datos, que se preparan antes de aplicar un modelo con algoritmos de aprendizaje supervisado, no supervisado o híbrido, cuyas actividades se incluyen: el diseño, programación y ejecución de pruebas.

Diseño de la CVAATI. Se realiza el diseño inicial de interfaces de la CVAATI en el que primero se presenta el inicio de sesión (Login) que se visualiza en la Figura 15, donde valida los niveles de acceso, definidos en cuatro roles: administrador, analista, docente, usuario normal e invitado. La autenticación de acceso del usuario es una estrategia de seguridad para el manejo de la información implementada en la IES de estudio.

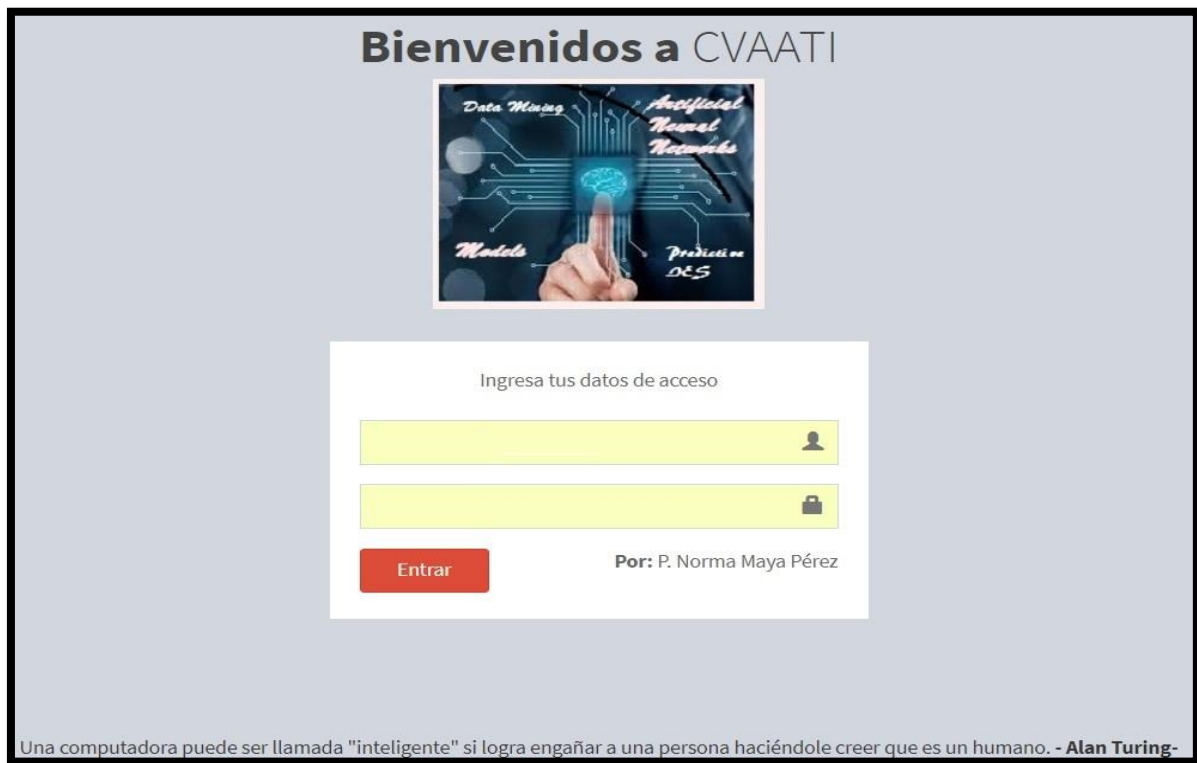


Figura 15. Pantalla de Login de CVAATI.

Fuente: Elaboración Propia.

La pantalla principal descrita en la **Figura 16**, incluye: la Integración y recopilación de datos, proceso de selección y limpieza, transformación y Pre procesamiento de datos que son la fuente de patrones de entrenamiento para posteriormente realizar el Modelado PredATIS apoyado en el SW de Matlab y WEKA en el modelado sobre el fenómeno que se aborda en este trabajo. Dejando como implementación futura la programación del modelo y emisión de reportes de alertas automatizados en este software de desarrollo propio y que no se reporta en este documento.

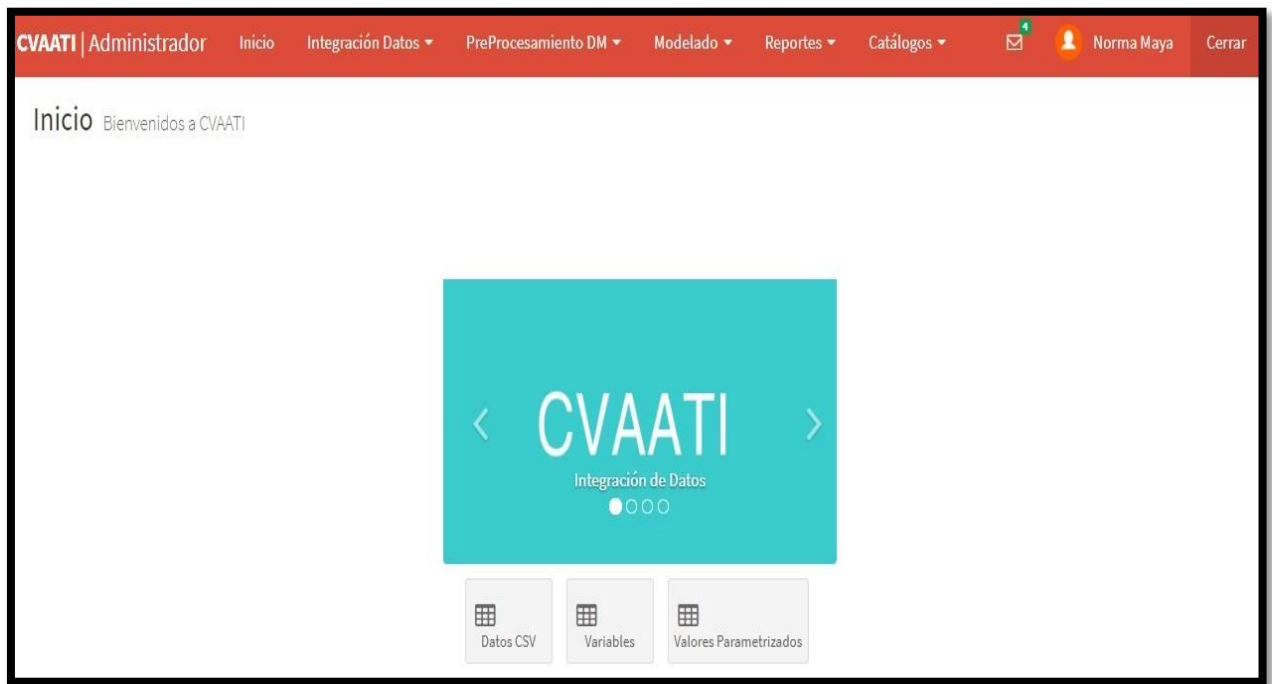


Figura 16. Pantalla Inicial de CVAATI.

Fuente: Elaboración Propia.

Programación de la CVAATI. Aquí se desarrollan y construyen las ventanas iniciales de la CVAATI, que incluye las siguientes actividades:

a) Pre-procesamiento

- Programación de la pantalla inicial de la CVAATI como se indica en la Figura 16.
- Programación del módulo de integración y recopilación de datos en la CVAATI.
- Programación del módulo de pre procesamiento de CVAATI que incluye:
 - La selección y limpieza de datos, en donde se realiza la depuración y actualización de los datos que serán pre procesados, antes de categorizar como patrones de entrenamiento del modelo.

- La transformación para identificar variables representativas en el diseño del Modelo PredATIS que incluye: asignar nueva variable, mostrar variables y eliminar variables.
- La parametrización consiste en categorizar a valores numéricos, que participarán como patrones entrada para el desarrollo del modelado con DM y RNA.

b) Modelado con DM y RNA

- Programación de la opción del Modelo de PredATIS en el que se incluyen un DM, RNA y gráficas, que se visualiza en la Figura 17. Señalando que la programación del modelado en CVAATI, no está definido en el alcance de este trabajo.

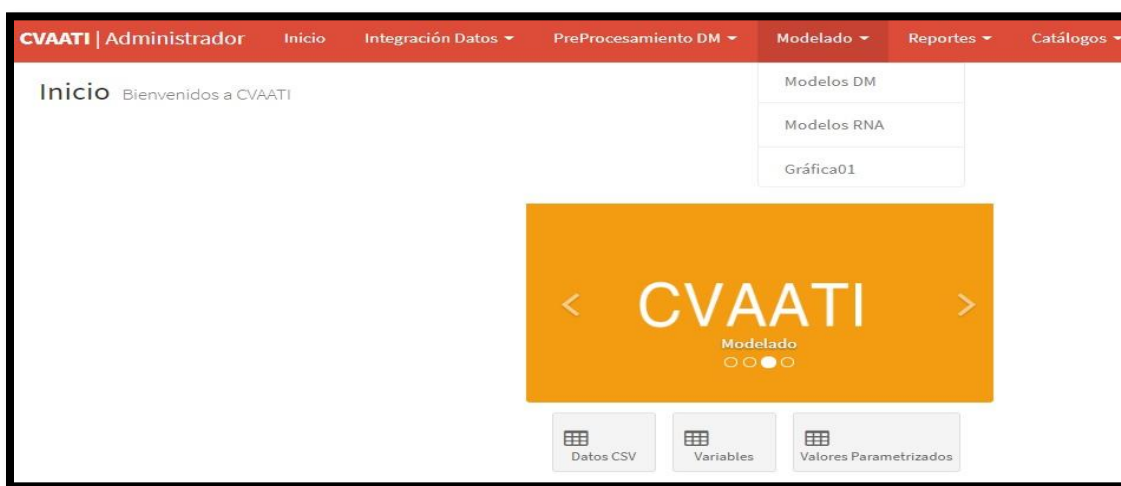


Figura 17. Pantalla de Modelado en CVAATI.

Fuente: Elaboración Propia.

- La programación del módulo de Interpretación de Resultados del Modelado que se visualiza en la Figura 18; **Error! No se encuentra el origen de la referencia.**, donde incluye: generación de estadísticas y reportes, la emisión de anuncios y alertas de la

CVAATI, así como el módulo de interpretación de resultados no se presentan en este documento, dejando como trabajo futuro.

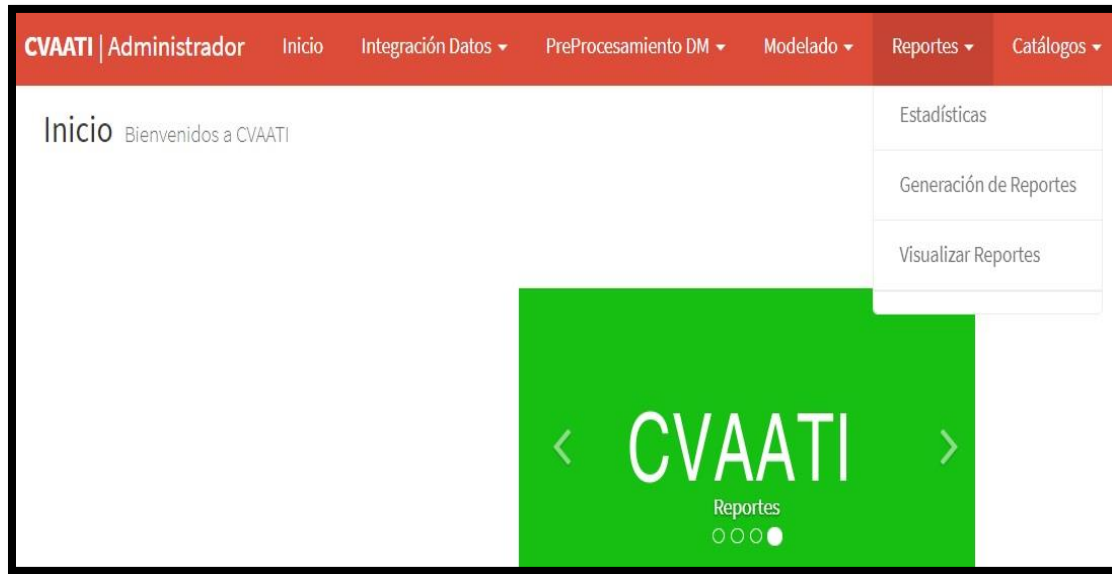


Figura 18. Pantalla de Interpretación de Resultados

Fuente: Elaboración Propia.

Pruebas de la CVAATI. En esta etapa de construcción de la CVAATI se realizan las pruebas de funcionalidad para el análisis comparativo que incluye:

- Ejecución de Pruebas de funcionalidad inicial con el software de Matlab y WEKA.

4.2 Diseño del Modelo PredATIS

El diseño del modelo PredATIS se basa en la metodología **MSWD&DM** propuesta en la **Figura 9** a partir de la segunda fase, que básicamente es el ciclo de vida de un proyecto de DM que incluye la selección del conjunto de datos, pre procesamiento de los datos y determinación del modelo, análisis de los resultados e integración. Derivado de ello, el trabajo que se presenta se desarrolla a través de las siguientes actividades:

- a) Análisis de algoritmos de aprendizaje con DM y RNA

- b) Integración y Recopilación de Datos
- c) Selección, limpieza y Transformación de Datos
- d) Diseño del Modelo de PredATIS empleando técnicas de clasificación DM y RNA
- e) Resultados presentados en el capítulo 5 para cada uno de los algoritmos de aprendizaje con DM entre los que se incluye RNA.

4.2.1 Análisis de Algoritmo de aprendizaje con DM y RNA.

En esta etapa se analizan las metodologías, técnicas y algoritmos de aprendizaje de DM y RNA a utilizar en el diseño de PredATIS, con el propósito de seleccionar aquellos, que se adapten al estudio que se presentan en este trabajo.

Análisis DM. De la revisión de modelos de predicción sobre el tema de estadísticos educativos (deserción o abandono escolar) descritos en la Tabla 6, y la literatura de la sección 2.4.2, se seleccionan las tareas, métodos, técnicas y algoritmos de DM que mejor se adaptan, al estudio que se presenta y posteriormente se describe en la sección de modelado, el procedimiento de entrenamiento o aprendizaje que se aplica en el diseño del Modelo predictivo propuesto.

- Las tareas definen un tipo de problema a ser resuelto por un algoritmo DM que pueden ser supervisado (predictivo) y no supervisado (descriptivo). Dada la naturaleza de la investigación que se presenta en este trabajo se decide utilizar algoritmos predictivos, ya que no se conocen todas las variables que impactan en la deserción del estudiante y por lo tanto se pretende predecir las variables o atributos del conjunto de datos desconocidos a priori, a partir de otros conocidos (Palomo, 2014) .

- Los métodos permiten resolver cualquiera de las tareas anteriores, donde se aplican diferentes técnicas, algoritmos y no sólo uno; puntualizando que un método puede servir para resolver más de una tarea Galán (2015).
- Las técnicas provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener resultados, donde las más representativas son: Redes Neuronales, Regresión Lineal, Árboles de Decisión, Modelos estadísticos, Clustering, Algoritmos genéticos (Palomo, 2014).

De forma que, para el diseño del modelo predictivo propuesto se seleccionan las técnicas de:

- Árboles de Decisión ya que permitirán investigar las correlaciones existentes entre las variables de mayor impacto en la deserción de los estudiantes.

Representan conjuntos de decisiones, y estas decisiones generan reglas para la clasificación de un conjunto de datos, son fáciles de usar, admiten atributos discretos y continuos, tratan bien los atributos no significativos y los valores faltantes. Su principal ventaja es la facilidad de interpretación. (Rodríguez & Díaz, 2009: 77)

- Redes Neuronales Artificiales en la que se utilizan diferentes modelos RNA que se describen en la Tabla 14. “Se caracterizan por ser capaces de trabajar con datos incompletos e incluso paradójicos, que dependiendo del problema puede resultar una ventaja o un inconveniente. Además esta técnica posee dos formas de aprendizaje: supervisado y no supervisado” (Rodríguez & Díaz, 2009: 76).

Análisis de Modelos de RNA. Se presentan un análisis comparativo en la **Tabla 14**, de los modelos de redes neuronales artificiales que se han aplicado en 12 estudios realizados; esto, con

la finalidad de seleccionar, cuales son los que mejor se adaptan al diseño del modelado predictivo propuesto PredATIS, siendo el propósito de este trabajo.

Tabla 14. Comparativa de modelos RNA

Tipo y Reglas de Aprendizaje / Arquitectura	Características	Ventajas	Desventajas	Estudios realizados-Modelo RNA/País/Autor-Año
Adaline (Adaptative Linear Element) o Adaline General e Iterativo				
Supervisado Monocapa: neuronas en paralelo	Unidireccional (feedforward) Se basa en un elemento combinador adaptativo lineal, que recibe todas las entradas, las suma ponderadamente y produce una salida que es un número real. Utiliza directamente la salida de la red (real) teniendo en cuenta cuánto se ha equivocado. Minimizar el error cuadrático de cada uno de los patrones de entrada individualmente Utiliza el método de descenso por gradiente para la función objetivo. (Bernard Widrow, 1960 citado en Rey, 2014)	Rápida Fácil de implementar Gráfica de errores hiperbólica único valor mínimo global a diferencia del perceptron de mínimos locales.	Sólo es posible clasificar patrones linealmente separables	Aplicación práctica en: - Procesamiento de señales: Filtros que eliminen el ruido en señales portadores de información - Filtros de ecualización adaptativos, en módems de alta velocidad - Cancelación del ruido materno de grabaciones EGC del latido del feto humano - Eliminación de ecos en circuitos telefónicos. - Reconocimiento de patrones - Traducción de textos para personas con problemas de lenguaje (Isasi & Galván, 2004).
Perceptron				
Supervisado Regla de aprendizaje basada en la corrección del error Monocapa: neuronas en paralelo	Capacidad de aprender a reconocer patrones, utiliza la salida de la función umbral (binaria) para el aprendizaje, utilizado en el Reconocimiento de caracteres impresos (Sánchez, 2004).	Pionera en la aplicación a problemas reales como el procesamiento adaptativo de señales y sistemas de control.	Es la RNA más antigua No puede reconocer caracteres complejos Sólo es posible para reconocer patrones que son linealmente separables (pueden ser separados por un	Estudio de aplicación de redes neuronales en la evaluación de riesgo crediticio- Perceptron, Chile (Méndez, 2009).

Tipo y Reglas de Aprendizaje / Arquitectura	Características	Ventajas	Desventajas	Estudios realizados-Modelo RNA/País/Autor-Año
			límite lineal)	
			Sólo se tiene en cuenta si se ha equivocado o no, pero no dice cuanto se ha equivocado.	
Perceptrón Multicapa o MLP				
Supervisado	Las neuronas se caracterizan en tres capas (entrada, oculta y salida), cada una posee su respectivo nivel de umbral, la función de transferencia puede ser la tangente hiperbólica.	Se remarca como una de las arquitecturas más útiles en la resolución de problemas de predicción a través de RNA por su alta precisión.	Principalmente se utiliza en la solución de problemas linealmente separables	La predicción de la quiebra y de la estimación de los rendimientos de cartera de valores en bancos españoles- MLP Back propagation, (Sanchez, 2003 citado en Del Carpio, 2005).
Algoritmo de Retropropagación	El proceso de aprendizaje se basa en encontrar un mínimo de la función error.	Reconocido por sus resultados en comparación con otros modelos como la mejor red neuronal para solucionar un problema de clasificación a partir de ejemplos.	La regla de aprendizaje empleada por el algoritmo de retropropagación puede decrecer el error cuadrático. Sin embargo puede conducir a un mínimo local más que a un mínimo global.	Sistema que puede pronosticar las señales de compra y venta, de acuerdo con la predicción de las futuras tendencias del mercado de valores y, de esta manera, proporcionar apoyo a la toma de decisiones por parte de los inversionistas- MLP Backpropagation, Perú (Zhang, Jiang , & Li, 2004 citado en Del Carpio, 2005)
Multicapa, tienen como base un Perceptron	El objetivo del aprendizaje es ajustar los parámetros de red, pesos y umbrales, donde las entradas produzcan las salidas deseadas (Plata, 2008, citado en Coello, Casas, & Pérez, 2015).	Útil en la solución de problemas con separabilidad no lineal.		- Aplicación de las Redes Neuronales al Pronóstico de Precios en el Mercado de Valores-MLP feedforward, Colombia (Villada, Muñoz, & García, 2012) - Redes neuronales artificiales en la producción de tecnología educativa para la enseñanza de la diagonalización- MLP, Cuba, (Coecillo , Casas, Pérez, & Caballero, 2015). - Predicción resultado académico de estudiantes de informática mediante el uso de redes neuronales-MLP, Chile, (Lau F, Álvarez, Pérez, & Leyva, 2016) . - Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro- Retropropagation feedforward, Colombia (Villada, Muñoz, & García-Quintero, 2016).
	Utiliza el algoritmo o regla de aprendizaje BackPropagation siendo una generalización de la regla LMS (Least Mean Square), basado en la corrección de error, con feedforward y backforward (Retropropagación) y las funciones de activación: Sigmoide logística, Sigmoide tangencial, y Lineal. Donde el número de neuronas puede variar en cada capa (Coecillo, Casas, Pérez, & Caballero, 2015).			

Tipo y Reglas de Aprendizaje / Arquitectura	Características	Ventajas	Desventajas	Estudios realizados-Modelo RNA/País/Autor-Año
RNA de Función de Base Radial (RBF)				
Híbridos Multicapa	Utiliza funciones de activación radial de carácter no lineal con sus centros gravitacionales propios en la capa oculta, y en la capa de salida, mediante funciones lineales. Utiliza método de entrenamiento NO supervisado (k-medias y mapas auto-organizados), y una adaptación supervisada de los pesos.	Es más eficiente en la fase de entrenamiento solo con un mayor volumen de datos. Ajustes a funciones no lineales y clasificación de datos sin separabilidad lineal. Método alternativo para aproximar funciones y clasificar patrones. Convergencia más simple. Entrenamiento más rápido. Arquitectura más simple. Combinación de distintos paradigmas de aprendizaje.	Requieren un mayor número de neuronas en los nodos en las capas ocultas que en MLP. Requiere un mayor volumen de datos en su fase de entrenamiento.	Redes Neuronales como aproximadoras de Funciones- FBR, Venezuela (Acquatella, 2009).
Mapas autoorganizados de Kohonen				
No Supervisado competitivo	Unidireccional (feedforward) Proceso de aprendizaje debe descubrir por sí misma rasgos comunes, regularidades, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones	Las neuronas se autoorganizan en función de los estímulos (datos) procedentes del Exterior. Tienen un gran potencial de Aplicación Práctica. Realiza mapas de características comunes de datos aprendidos. Útil para establecer relaciones desconocidas. Previamente se organiza la información de	Requiere mucho entrenamiento	- Aplicaciones de las redes neuronales. El caso de la Bibliometría, Cuba (Sotolongo & Guzmán, 2001). - Evaluación de Páginas web, Cuba (Kaski, 1996 citado en Sotolongo & Guzmán, 2001). - Clasificación de revistas en un determinado campo temático (Kohonen, 1997 citado en Sotolongo & Guzmán, 2001). - Minería de textos (text mining) sobre todo aplicado a la asociación de palabras o co-word, Cuba (Sotolongo & Guzmán, 2001)

Tipo y Reglas de Aprendizaje / Arquitectura	Características	Ventajas	Desventajas	Estudios realizados-Modelo RNA/País/Autor-Año
		entrada, de entre un gran volumen de datos, clasificándola automáticamente y permitiendo la visualización de relaciones que son importantes entre los datos. Interfaz amigable de los mapas que facilita su interpretación y utilidades prácticas. (Sotolongo & Guzmán, 2001)		

Fuente: Elaboración propia a partir de Sotolongo & Guzmán (2001); Del Carpio (2005); Acquatella(2009); Coello, Casas, & Pérez (2015); Lau F, Álvarez, Pérez, & Leyva(2016).

Con base en los modelos referenciados en la Tabla 14 y la revisión de la literatura en artículos científicos reportados en esta temática de investigación, se argumenta que no se pueden determinar que un modelo sea más eficiente que otro, dado que deben adaptarse a la problemática a resolver (Coecillo , Casas, Pérez, & Caballero, 2015); mediante un entrenamiento, a prueba y error.

El estudio se aborda en primera instancia aplicando métodos y técnicas de clasificación y selección de atributos de DM en el SW WEKA y continuando con Redes Neuronales Artificiales (RNA) mediante la programación de algoritmos de aprendizaje en Matlab, donde se observará las ventajas y desventajas de utilizar los diferentes algoritmos de aprendizaje en Tabla 14. Posteriormente se pretende como trabajo futuro y que esta fuera del alcance de este documento, realizar la implementación del Modelo PredATIS en la CVAATI que es un desarrollo de SW propio.

4.2.2 Integración y Recopilación de Datos en la CVAATI.

Después de realizar el análisis de modelos de DM y RNA se continua la metodología **MSWD&DM** descrita en la **Figura 9** a partir de la segunda fase, donde se realizan primero las actividades previas al proceso de DM como es la Recopilación de datos fuentes disponibles y la determinación de Variables que se utilizan en el estudio y diseño del Modelo propuesto, continuando con la selección, limpieza y transformación de las variables seleccionadas para ser pre procesadas antes de aplicar un modelo con DM.

Recopilación de Datos. Los datos son importados a través de archivos de tipo CSV, TXT y SQL que se generan del Sistema de Automatización de Información Integral de la IES del PE que se utilizan en el Modelado propuesto y que son cargados a la CVAATI a través de la primera pantalla de integración de datos, como una etapa inicial del modelado de PredATIS presentados en la Figura 19 y Figura 20.



Figura 19. Pantalla de Integración de Datos a CVAATI.

Fuente: Elaboración Propia.

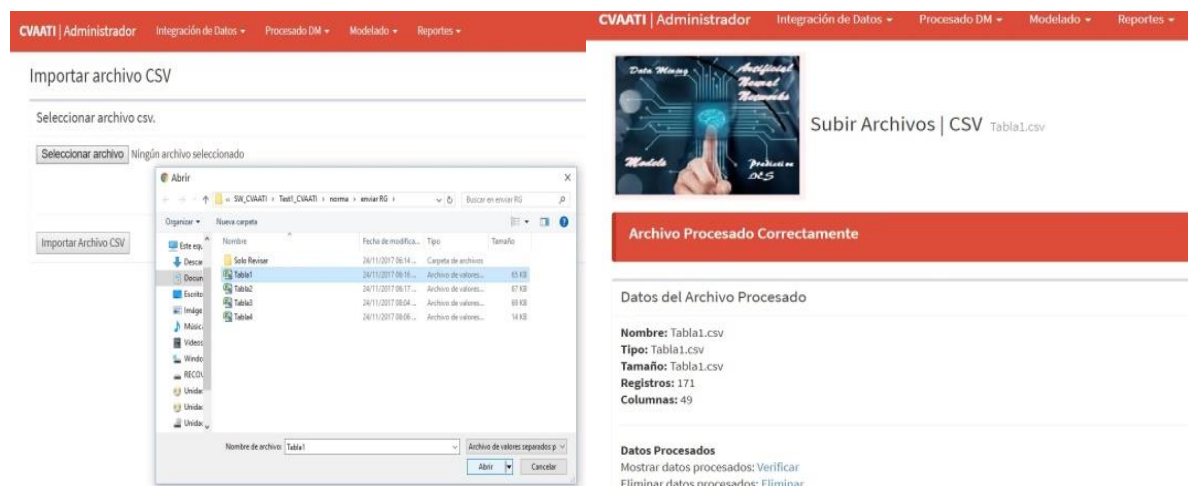


Figura 20. Pantalla de Importación de Archivos a CVAATI.

Fuente: Elaboración Propia.

Determinación de variables. De los datos recopilados de la muestra, se examinaron 106 variables que incluyen aspectos personales, vocacionales, socioeconómicos, de salud, académicos y otros aspectos de adiciones por parte de los estudiantes, de los cuales solo se consideran para el estudio, 39 variables de interés, mostradas en la Tabla 15, continuando en una etapa posterior de selección, limpieza, transformación y pre procesamiento de datos que serán utilizados en el entrenamiento para el diseño de Modelo PredATIS, que se describe en la sección 4.2.3 Selección, Limpieza y Transformación de Datos en CVAATI.

Tabla 15. Variables de estudio

Aspectos personales	Aspectos Vocacionales	Aspectos Académicos	Aspectos Socioeconómicos	Aspectos Salud	Otros Aspectos
• Edad • Género • Estado civil • Número de Hijos • Planes de matrimonio a corto plazo o en tu estancia en la IES	• Título Técnico o Bachillerato General • La IES es su primera opción. • El PE es su primera opción. • Tiene planeado presentar examen en otra IES	• Promedio de Nivel Medio Superior • Nuevo Ingreso al período escolar /Reingreso • Materias con dificultad en nivel • Número de asignaturas extraordinarias en	• Ingreso Mensual Familiar • Alumno Trabaja Si/NO, número de horas semanales. • ¿De quién depende económicamente el estudiante? • ¿Quién depende	• Estado de salud • Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura,	• Adicciones de fumar • Adicciones de ingerir bebidas alcohólicas. • Adicciones a drogas o no consume.

Aspectos personales	Aspectos Vocacionales	Aspectos Académicos	Aspectos Socioeconómicos	Aspectos Salud	Otros Aspectos
• Relación familiar • Planes de tu futuro profesional • Ocupación de su tiempo libre.	• Ha interrumpido estudios • Ha cursado otras carreras en otras IES diferentes al PE que estudia.	períodos escolares anteriores • Asignaturas extraordinarias en períodos escolares anteriores • Técnicas de Estudio • Tiene libros de tu perfil de carrera como apoyo para su estudio en casa • Tiene computadora en casa como apoyo a sus estudios • Tiene acceso a internet en tu domicilio.	económicamente de usted? • Gastos de transportación semanales a la IES. • Tiempo de transportación (hrs) Casa – IES • Tiene Beca (Si/No) • Tipo de Beca	visual, auditivo, discapacidad, Ninguna • Padece alguna enfermedad, tratamiento o alergia (Si/No, específica) • Frecuencia que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia) • Ha recibido atención psicológica.	

Fuente: Elaboración propia.

4.2.3 Selección, Limpieza y Transformación de Datos en CVAATI.

Aquí se preparan los datos importados, a través de una selección y limpieza de datos y posteriormente se transforman a variables y son pre-procesadas a datos útiles a partir de la Tabla 15 para generar la vista minable desde el software propiamente desarrollado de CVAATI, que representa una de las aportaciones de este trabajo y que son factores primordiales en la investigación del fenómeno de deserción o reprobación escolar de la IES en estudio. Las actividades que se incluyen en esta etapa son:

- a) **Selección y Limpieza:** aquí se revisa en el archivo de datos importado a la CVAATI como se visualiza en la Figura 21, de que no existan datos nulos o erróneos, se modifican y eliminan las variables descritas en la Tabla 15, que no participarán en el entrenamiento del

modelado de PredATIS aplicado a la predicción de los estudiantes en riesgo de baja de la IES y que se presenta en la sección 4.3, siendo el propósito de la investigación que se aborda en este trabajo.

Selección y Limpieza

Eliminar	Editar	1	Item	edad	sexo	edocivil	nhijos	planmatrimonio	relfamiliar	futuroprof	tiempolibre	especialidad	primeraspc_int	primeraspc_carr	otro_examen	estudio_inten	estudio_inten_razon	Estudio en otra IES	promedio_bach	maten
Eliminar	Editar	2	1	15	Masculino	Soltero(a)	0	Si	Regular	Enfocarme en lo que realmente quiero hacer desarrollando el conocimiento que tengo en el campo laboral	Aprendo cosas nuevas en mi computadora				0	en el 2012	Me dieron de baja de la primera preparatoria ya que se me dificultaba en ese tiempo aprender	Prefiero no responder		Nuevo Ingreso
Eliminar	Editar	3	2	19	Masculino	Soltero(a)	0	No	Buena	Trabajar en una buena empresa	Jugar futbol	Sin especialidad						Prefiero no responder	9.5	Nuevo Ingreso
Eliminar	Editar	4	3	21	Masculino	Soltero(a)	0	No	Buena	crear una empresa propia	Idioma proyectos				0	0	0	Prefiero no responder		Nuevo Ingreso
Eliminar	Editar	5	4	15	Masculino	Soltero(a)	0	No	Buena	Tener mi propia empresa	Practicar musica	Sin especialidad	Son sólo 3 años y 8 meses	Me gustan las tic	0	0	0	Prefiero no responder	7.7	Nuevo Ingreso
Eliminar	Editar	6	5	18	Femenino	Soltero(a)	0	No	Buena	tener un buen trabajo	hacer dibujos y manualidades	Informática	que la que me gusta	es lo que quiero estudiar				Prefiero no responder	9	Nuevo Ingreso
Eliminar	Editar	7	6	19	Femenino	Soltero(a)	0	No	Buena	TERMINAR MI INGENIERIA Y SEGURIDAD ACTUALIZANDOME	ver películas nadar divertirme estudiar				0	0	0	Prefiero no responder		Nuevo Ingreso
Eliminar	Editar	8	7	18	Masculino	Soltero(a)	0	No	Buena	conseguir un buen trabajo y que me guste	Jugar	Me catrónica		Porque me gusta				Prefiero no responder	8.9	Nuevo Ingreso
Eliminar	Editar	9	8	21	Masculino	Soltero(a)	0	No	Buena	Ser un buen profesionista y poder trabajar bien del resto	Hago deporte Juego				0	hace dos años	trabajo	Prefiero no responder		Nuevo Ingreso

Figura 21. Módulo de Selección y Limpieza de Datos en la CVAATI.

Fuente: Elaboración Propia.

b) **Transformación a variables**, donde se preparan los datos que representan las variables relevantes seleccionadas e identificadas con un nombre de variable como se muestra en la Tabla 16, para el entrenamiento con Algoritmos de Clasificación, Selección de DM en el SW WEKA.

Tabla 16. Variables para Modelado con Técnicas de Clasificación y Selección DM

Descripción de la Variable	Variable de Entrada Modelado DM
Edad	Edad
Género	Genero
Estado civil	EstadoCivil
Número de Hijos	NúmHijos
Planes de matrimonio a corto plazo o en su estancia en la IES	PlanesMatrimonioCtoPlazo
Relación familiar	RelacionFamiliar

Descripción de la Variable	Variable de Entrada Modelado DM
Planes de su futuro profesional	PlanesFuturoProf
Ocupación de su tiempo libre	OcupacTiempoLibre
Titulo Técnico o Bachillerato General	TituloTécnico/BachGral
La IES es su primera opción	La IES1a.Opc
El PE es su primera opción	El PEPrimOpc
Tiene planeado presentar examen en otra IES	TienePlaneadoExamOtIes
Has interrumpido estudios	InterrupcionEstudios
Ha cursado otras carreras en otras IES diferentes al PE que estudia	CursadoOtrasIES
Promedio de Nivel Medio Superior	PromedioNivelMedioSup
Nuevo Ingreso al período escolar /Reingreso	NuevoIngreso/Reingreso
Materias con dificultad en nivel medio superior	MateriasDificultadMSSup/IES
Número de Asignaturas extraordinarias en períodos escolares anteriores	NumAsigExtraPdosAnt
Asignaturas extraordinarias en períodos escolares anteriores	AsignExtraordinariasAnt
Técnicas de estudio	TecnicasEstudio
Tiene libros de tu perfil de carrera como apoyo para su estudio en casa	TieneLibrosPerfilCasa
Tiene computadora en casa como apoyo a sus estudios	TieneCcomputCasa
Tiene acceso a internet en tu domicilio	TieneInternetCasa
Ingreso mensual familiar	IngresoMensualFamiliar
Alumno trabaja Si/NO, Número de horas semanales	AlumnoTrabaja Si/NO, HrSem
¿De quién depende económicamente el estudiante?	DeQuienDependeEconomicamente Est
¿Quién depende económicamente de usted?	DependieteeEconomicos
Gastos de transportación semanales a la IES	GastosTransport-IES
Tiempo de transportación (hrs) Casa – IES	TiempoTranspCasaIES
Tiene Beca (Si/No)	TieneBeca
Tipo de Beca	TipBeca
Estado de salud	EdoSalud
Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura, visual, auditivo, discapacidad o ninguna	ProblemasSalud
Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica)	PadeceEnfermedad
Frecuencia que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia)	FrecuenciaPresEnfermMenores
Ha recibido atención psicológica(Si, No, cuanto tiempo)	AtencionPsicologica
Adicciones de fumar	AdiccionesFumar
Adicciones de ingerir bebidas alcohólicas	AdiccionesBebAlcoholicas
Adicciones a drogas	AdiccionesDrogas

Fuente: Elaboración propia.

La Tabla 17, describe las variables de entrada utilizadas en el modelado de PredATIS a través de RNA de FBR.

Tabla 17. Variables de Entrada para Modelado con RNA

Descripción de la Variable	Variable de Entrada Modelado
Edad	p1
Género	P2
Estado civil	P3
Número de Hijos	P4
Planes de matrimonio a corto plazo o en su estancia en la IES	p5
Relación familiar	p6
Planes de su futuro profesional	p7
Ocupación de su tiempo libre	p8
Titulo Técnico o Bachillerato General	p9
La IES es su primera opción	p10
El PE es su primera opción	p11
Tiene planeado presentar examen en otra IES	p12
Has interrumpido estudios	p13
Ha cursado otras carreras en otras IES diferentes al PE que estudia	p14
Promedio de Nivel Medio Superior	p15
Nuevo Ingreso al período escolar /Reingreso	p16
Materias con dificultad en nivel medio superior	p17
Número de Asignaturas extraordinarias en períodos escolares anteriores	p18
Asignaturas extraordinarias en períodos escolares anteriores	P19
Técnicas de estudio	p20
Tiene libros de tu perfil de carrera como apoyo para su estudio en casa	p21
Tiene computadora en casa como apoyo a sus estudios	p22
Tiene acceso a internet en tu domicilio	p23
Ingreso mensual familiar	p24
Alumno trabaja Si/NO, Número de horas semanales	p25
¿De quién depende económicamente el estudiante?	p26
¿Quién depende económicamente de usted?	P27
Gastos de transportación semanales a la IES	P28

Descripción de la Variable	Variable de Entrada Modelado
Tiempo de transportación (hrs) Casa – IES	P29
Tiene Beca (Si/No)	p30
Tipo de Beca	p31
Estado de salud	p32
Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura, visual, auditivo, discapacidad o ninguna	p33
Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica)	p34
Frecuencia que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia)	p35
Ha recibido atención psicológica(Si, No, cuanto tiempo)	p36
Adicciones de fumar	P37
Adicciones de ingerir bebidas alcohólicas	P38
Adicciones a drogas	P39

Fuente: Elaboración propia.

Una vez identificadas las variables, se asigna un nombre en la que se definen en el módulo de catálogos como se muestra en la Figura 22, antes de asignar y asociar como variables de entrada para el estudio.

The screenshot shows the 'Administrar Variables' window in the CVAATI system. The sidebar on the left has a 'Catálogos' dropdown and a 'Variables' tab, which is highlighted with an arrow. The main content area lists 14 variables being defined, each with a label and a text input field:

- 1) Variable #1: P1...
- 2) Variable #2: P2, Género
- 3) Variable #3: P3, Estado civil
- 4) Variable #4: P4
- 5) Variable #5: P5, Planes de matrimonio a corto plazo o en su estancia en la IES
- 11) Variable #11: P11, El PE es su primera opción
- 12) Variable #12: P12, Tiene planeado presentar examen en otra IES
- 13) Variable #13: P13, Has interrumpido estudios
- 14) Variable #14: P14, Ha cursado otras carreras en otras IES diferentes al PE que est...

Figura 22. Catálogo de definición de variables en CVAATI.

Fuente: Elaboración propia.

Una vez definido el nombre de las variables, se continúa con el pre procesamiento de DM, donde se transforman los datos en variables y se pueden asignar nueva variable, mostrar o eliminar variables que no se requiere para el estudio como se visualiza en la Figura 23.

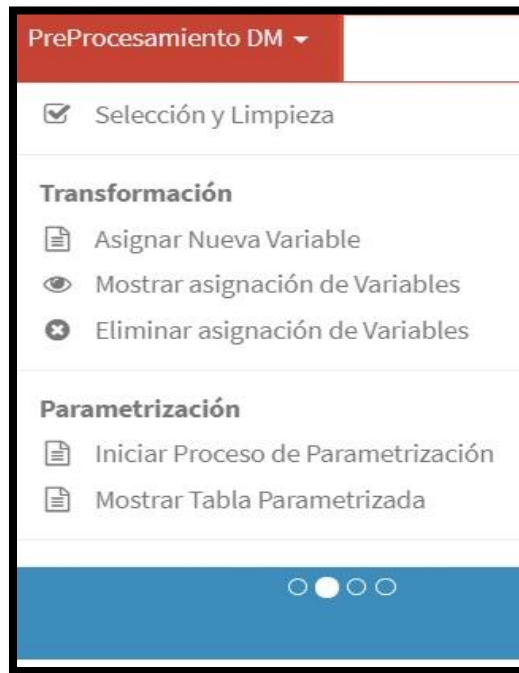


Figura 23. Opciones de Pre procesamiento DM.

Fuente: Elaboración propia.

- *Asignar nueva variable:* en la que se selecciona el id, descripción de la variable en estudio y uno o más aspectos que se desean incluir en la misma, presentados en la Figura 24, y representan los valores de entrada $p_1, p_2, \dots, p_{39}$ que generan la vista minable, para el entrenamiento del modelado con algoritmos de aprendizaje no supervisado y supervisado, que se aplican en el modelado de PredATIS que se describe en la sección 5.2.1 Modelado de PredATIS basado en RNA de FBR.

Figura 24. Asignación de nuevas variable en CVAATI.

Fuente: Elaboración propia.

Se observa en Figura 23 en la sección de transformación, adicionalmente dos opciones:

- *Mostrar variables asignadas*: donde se visualiza las variables que han sido seleccionadas para el modelado DM y en caso de no estar correctamente seleccionadas se pueden eliminar desde esta pantalla y volver a asignar de forma correcta desde la opción de asignar variable.
- *Eliminar variable*: se debe tener especial cuidado en realizar esta acción ya que elimina todas las variables que ya se habían seleccionado.

c) **Pre procesamiento de variables**, donde se preparan los datos que representan las variables seleccionadas, categorizando los valores alfanuméricos en numéricos como se presentan en la Tabla 18, con el objeto de ser clasificados como patrones de entrenamiento del aprendizaje con algoritmos de clasificación y selección DM que utiliza las variables categóricas nominales y RNA de FBR con variables numéricas en el diseño del modelo propuesto.

Tabla 18. Valores categorizados de Variables en estudio

Descripción de la Variable	Valores extraídos y categorizados numéricamente
Edad	Menor de 18 años=1, De 18 a 21 años=2, De 22 a 25 años=3, mayor de 25 años=4, No Respondió Pregunta (NRP)=5 y No realizó encuesta (NRE)=0
Género	Femenino= 1, Masculino=2, Otro=3, NRP=4 y NRE=0
Estado civil	Soltero=1, Casado=2, Divorciado=3, Viudo=4, Madre/Padre Soltero(a)=5, Unión libre=6, Prefiero no contestar=7 y NRE=0
Número de Hijos	Ninguno o 0=1, De 1 a 2 = 2, De 3 a 4=3, más de 4=4, NRP=5 y NRE=0
Planes de matrimonio a corto plazo o en su estancia en la IES	No=1, Si=2, Prefiero no contestar=3 y NRE=0
Relación familiar	Buena=1, Regular=2, Mala=3, Con Problemas Padres=4, Prefiero no responder=5 y NRE=0
Planes de su futuro profesional	Trabajar en perfil Empresa o Negocio=1, Graduarse TSU e Ingeniería=2, Especialidades y/o certificaciones del perfil=3, seguir estudios maestría y/o Doctorado=4, Otros= 5, Ninguno=6, NRP =7 y NRE=0
Ocupación de su tiempo libre	Estudiar o visitar bibliotecas=1, actividad cultural o Deportiva=2, Varios entretenimiento (visitar plazas comerciales, cines, parques, Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento,etc)=3, Chat o Redes Sociales/Navegar Internet=4, Platicar con amigos o Hablar por teléfono=5, otras actividades(domesticas/familiares/religiosas/trabajo)=6, No Respondió=7 y NRE=0
Título Técnico o Bachillerato General	Técnico a fin al PE=1, Técnico no afín a la Carrera=2, Bachillerato=3, NRP=4 y NRE=0
La IES es su primera opción	Si=1, No aprobé en otra IES=2, Indeciso=3, Obligado por padres/Cercanía=4, Prefiero no responder=5 y NRE=0
El PE es su primera opción	Si=1, No aprobé en otras IES o no me inscribe al Perfil deseado=2, Indeciso =3, Obligado por padres/Cercanía=4, mi última opción=5, No sin

Descripción de la Variable	Valores extraídos y categorizados numéricamente
	especificar motivo= 6, Prefiero no responder=7 y NRE=0
Tiene planeado presentar examen en otra IES	No=1, Indeciso=2, Si=3, Prefiero no responder=4 y NRE=0
Has interrumpido estudios	No=1, De 1 a 2 años por motivos personales=2, De 1 a 2 años por motivos económicos=3, De 1 a 2 años por motivos académicos o vocacionales=4, De 1 a 2 años por otros motivos =5, Más de 2 años por motivos personales/Económicos=6, Más de 2 años por motivos académicos o vocacionales=7, Más de 2 años por otros motivos =8, Prefiero no responder=9 y NRE=0
Ha cursado otras carreras en otras IES diferentes al PE que estudia	No=1, Si la del perfil=2, Si perfil diferente=3, Prefiero no responder=4 y NRE=0
Promedio de Nivel Medio Superior	De 9.6 a 10=1, De 9 a 9.5 =2, De 8 a 8.9=3, De 7 a 7.9=4, De 6 a 6.9 =5, Menor que 5.9 =6, NRP=7 y NRE=0
Nuevo Ingreso al período escolar /Reingreso	Nuevo Ingreso al ciclo escolar=1, Reingreso 1 vez=2, Reingreso 2 veces=3, Reingreso más de 2 veces=4, NRP=5 y NRE=0
Materias con dificultad en nivel medio superior	Ninguna=1, Matemáticas y lógica=2, Áreas Informáticas=3, Inglés=4, Matemáticas e Informática(BD/Programación/Redes/SO) = 5, Inglés y Asignaturas de Informática (Programación/Redes/SO/BD)=6, otras no afines al perfil=7, NRP=8 y NRE=0
Número de Asignaturas extraordinarias en períodos escolares anteriores	Ninguna=1, una asignatura= 2, dos asignaturas=3, Mayor o igual a 3 asignaturas=4, NRP=5 y NRE=0
Asignaturas extraordinarias en períodos escolares anteriores	Ninguna=1, Asignaturas no técnicas del perfil (formación sociocultural, expresión oral y escrita. etc.) =2, Desarrollo de Habilidades lógico-matemáticas-Física=3, Inglés=4, Asignaturas Técnicas del perfil (Programación, Base de Datos, Redes de cómputo, Soporte Técnico, Sistemas Operativos, etc.)=5, Inglés - Técnicas Perfil= 6, Ingles/Matemáticas= 7,

Descripción de la Variable	Valores extraídos y categorizados numéricamente
Técnicas de estudio	NRP=8 y NRE=0 Asesorías especiales=1, video tutoriales=2, páginas web/Internet=3, manuales/Libros/revistas académicas =4, Asesoría/Video Tutoriales/Web=5, Varias técnicas de estudio posterior= 6, Repaso Notas/Prácticas Clases=7, Ninguna solo las clases=8, NRP=9 y NRE=0
Tiene libros de tu perfil de carrera como apoyo para su estudio en casa	Si=1, Ninguno/No=2, NRP=3 y NRE=0
Tiene computadora en casa como apoyo a sus estudios	Si=1, No=2, NRP=3 y NRE=0
Tiene acceso a internet en tu domicilio	Si=1, No=2, NRP=3 y NRE=0
Ingreso mensual familiar	Más de \$6000=1, De 4001 a 6000=2, De 3001 a 4000=3, De \$2000-3000=4, Menor a \$2000=5, NRP=6 y NRE=0
Alumno trabaja Si/NO, Número de horas semanales	No trabaja=1, Si Menor a 20 horas (hrs)=2, Si de 20 a 30 hrs=3, Si de 30 a 40 hrs=4, Si más de 40 hrs=5, NRP=6 y NRE=0
¿De quién depende económicamente el estudiante?	Padres=1, Esposa (o) =2, Familiar=3, Hijos=4, Otro=5, Nadie, él mismo solventa gastos=6, NRP=7 y NRE=0
¿Quién depende económicamente de usted?	Nadie=1, Esposa (o) =2, Hijos=3, Padres=4, Familia=5, Otros=6, NRP=7 y NRE=0
Gastos de transportación semanales a la IES	Menor o igual a \$100=1, De \$101 a \$300=2, De \$301 a \$500=3, De 501 a 800 = 4, Mayor que \$800=5, No Respondió=6 y NRE=0
Tiempo de transportación (hrs) Casa – IES	Menos de una hr.=1, De 1 a 2 hr=2, Mayor que 2 y Menor = a 3 hrs=3, Más de 3 hrs=4, NRP=5 y NRE=0
Tiene Beca (Si/No)	Si=1 No=2 , Prefiero no contestar=3, NRP=4 y NRE=0
Tipo de Beca	Manutención/Bécalos=1, Programas Sociales(Prospera, otra)=2, Excelencia=3, Laptop= 4, otras(Jóvenes Valores de Paz)=5, No Tiene=6, NRP=7 y NRE=0
Estado de salud	Buena=1, regular=2, mala=3, pésima=4, NRP=5 y NRE=0
Problemas de salud (obesidad, delgadez extrema, manchas en	Ninguna=1, dentadura=2, Obesidad=3, Delgadez

Descripción de la Variable	Valores extraídos y categorizados numéricamente
la piel, falta de energía, dentadura, visual, auditivo, discapacidad o ninguna	extrema/falta de energía=4, visual=5, auditivo=6, discapacidad=7, Varios problemas (Manchas en la Piel/problemas de piel, visual, auditivo, obesidad, entre otros)=8, NRP=9 y NRE=0
Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica)	No=1, Si enfermedad crónica =2, Si enfermedad con tratamiento=3, Si otro tipo de enfermedad=4, Si es Alérgico a algún medicamento=5, NRP=6 y NRE=0
Frecuencia que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia)	Gripe, dolor de Cabeza, infecciones estomacales u otra enfermedad rara vez=1, Gripe, dolor de cabeza, infección estomacal u otra enfermedad alguna vez al mes =2, Gripe, dolor de cabeza, infección estomacal u otra enfermedad alguna vez a la semana =3, Gripe, dolor de cabeza, infección estomacal u otra enfermedad muy frecuentemente =4, Otras Enfermedades pocas veces= 5, NRP=6 y NRE=0
Ha recibido atención psicológica(Si, No, cuanto tiempo)	No=1, Si de 1 mes =2, Si de 2 meses a menor de 1 año=3, Si de 1 a 2 años = 4, Si más de 2 años=5, Prefiero no contestar =6 y NRE=0
Adicciones de fumar	No Fuma =1, rara vez fuma=2, Si una vez al mes=3, Si Algunas veces al mes=4, Si varias veces a la semana=5, Si una vez la semana=6, Si una vez al día=7, Si varias veces al día=8, NRP=9 y NRE=0
Adicciones de ingerir bebidas alcohólicas	No=1, rara vez ingiere=2, Si una vez al mes=3, Si Algunas veces al mes=4, Si varias veces a la semana=5, Si una vez la semana=6, Si una vez al día=7, Si varias veces al día=8, NRP=9 y NRE=0
Adicciones a drogas	No=1, rara vez consume=2, Si una vez al mes=3, Si Algunas veces al mes=4, Si varias veces a la semana=5, Si una vez la semana=6, Si una vez al día=7, Si varias veces al día=8, NRP=9 y NRE=0

Fuente: Elaboración propia.

Derivado del pre procesamiento, en la que fueron agrupadas en variables categóricas nominales, indicadas en la Tabla 18, siendo los datos de entrada para el diseño de PredATIS, que representan la base del modelado que se describe en la sección de 4.2.4 Diseño del Modelo PredATIS con DM. Estas variables también se pueden categorizar desde la opción de parametrización descrito en la Figura 23 y Figura 25 .

Variable ID	Valor = 0	Valor = 1	Valor = 2	Valor = 3	Valor = 4	Valor = 5	Valor = 6	Valor = 10
1	Hombre	Mujer	X	X	X			
2	Xfewew	Xewewe	Xvew	X	X			

1) Variable #1

- Hombre
- Mujer
- X
- X

Actualizar Información

Figura 25. Parametrización de variables en CVAATI.

Fuente: Elaboración Propia.

En tanto variables categóricas numéricas de la Tabla 18, son los datos de entrada mostrados en la Figura 26 para el diseño del modelado RNA que se presenta en la sección 4.2.5 Diseño del Modelo PredATIS con RNA.y 5.2 Modelado de PredATIS con algoritmos de RNA.

ITEM	Edad	Sexo	Nivel de estudios	Nivel de ingresos	Nivel de escolaridad	Nivel de salud	Nivel de empleo	Nivel de vivienda	Nivel de transporte	Nivel de alimentación	Nivel de recreación	Nivel de cultura	Nivel de tecnología	Nivel de comunicación	Nivel de energía	Nivel de medio ambiente	Nivel de seguridad	Nivel de justicia	Nivel de servicios	Nivel de bienestar	Nivel de calidad de vida	Nivel de satisfacción	Nivel de felicidad	Nivel de salud mental	Nivel de salud física	Nivel de salud emocional	Nivel de salud social	Nivel de salud espiritual	Nivel de salud integral
1	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
4	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
11	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
12	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
16	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
19	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
20	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
21	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
22	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
23	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
24	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
25	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
27	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
28	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
29	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
31	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
32	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
33	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
34	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
35	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
36	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
37	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
38	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
39	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
40	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figura 26 Pre procesado a datos discretos.

Fuente: Elaboración Propia.

4.2.4 Diseño del Modelo PredATIS con DM.

El diseño del modelo propuesto se guía en la metodología MSWD&DM que incluye las fases de un proyecto de DM que incluye: la integración y recopilación de datos, selección, limpieza y transformación de datos, descritas en la sección 4.2.2 Integración y Recopilación de Datos en la CVAATI. y 4.2.3 Selección, Limpieza y Transformación de Datos en CVAATI. en donde se realizó el pre procesamiento generando la vista minable, donde se obtienen los patrones de entrada y continuar, en el diseño del Modelo predictivo PredATIS con técnicas y métodos de DM supervisadas y no supervisadas, con el propósito de identificar las variables predictivas que influyen en la deserción del estudiante a través del SW WEKA y analizar dentro del contexto real que se reporta en la IES en estudio para medir la precisión obtenida en esta investigación.

- a) Las *No Supervisados* no toman como base una variable dependiente, endógena o variable a predecir, en donde se integran los Clustering o agrupación (EM, Simple k-Means, Cobweb, etc.) y Reglas de asociación (Apriori, FilteredAssociator y FPGrowth) (Herrero & Molina, 2012).
- b) Las *Supervisadas* tienen una variable dependiente, endógena o variable a predecir, que incluye:
- Técnicas de Clasificación: árboles de decisión, tabla de decisión, inducción de reglas, bayesianas, redes neuronales, lógica difusa, técnicas genéticas. En los árboles de decisión la variable a predecir o dependiente es categórica y por los resultados obtenidos pueden representar una *regresión logística* (Witten, Frank, & Hall, 2011, citado en Herrero & Molina, 2012).
 - Técnicas de Predicción: árbol de regresión (predicción), regresión, estimador de núcleos. Un árbol de regresión es análoga a una regresión lineal donde la variable a predecir o dependiente es numérica (discreta o continua) *logística* (Witten, Frank, & Hall, 2011, citado en Herrero & Molina, 2012).
- En un árbol de predicción numérica cada nodo hoja almacena un valor de clase consistente en la media de las instancias que se clasifican con esa hoja, representando un árbol de regresión, o bien un modelo lineal que predice el valor de la clase y se conoce como un árbol de modelos. Un algoritmo de esta técnica es el M5 o M5P en el SW WEKA (Witten, Frank, & Hall, 2011, citado en Herrero & Molina, 2012), donde se obtiene un árbol de modelos, que se puede utilizar para obtener un árbol de regresión (Herrero & Molina, 2012).

En la predicción numérica se emplea la *varianza del error* en cada hoja. Una vez construido el árbol que clasifica las instancias se realiza la poda del mismo, tras lo cual, se obtiene para cada nodo hoja una constante en el caso de los árboles de regresión o un plano de regresión en el caso de árboles de modelos. En éste último caso, los atributos que formarán parte de la regresión serán aquellos que participaban en el subárbol que ha sido podado. (Herrero & Molina, 2012: 111)

En el análisis de datos con DM para el diseño del modelo que se presenta en este trabajo, se utiliza técnicas de clasificación para medir la precisión sobre resultados predictivos.

La CLASIFICACIÓN consiste en dividir un conjunto de datos en grupos mutuamente excluyentes, donde cada miembro de un grupo esté lo más cerca posible de otros y los grupos diferentes estén lo más lejos posible de otros y la distancia se mide con respecto a las variables a predecir. (Herrero & Molina, 2012:120)

Señalando así, que el diseño del modelo PredATIS utiliza técnicas y algoritmos de: árbol de decisión, reglas de clasificación y selección de atributos.

a) El Árbol de Decisión

Es una clasificación supervisada y la representación del conocimiento es relativamente simple y su aprendizaje sencillo comparado con otros sistemas basados con lenguajes de representación más potentes, como redes semánticas, representaciones en lógica difusa, etc., no obstante, la potencia expresiva es menor. El aprendizaje suele ser más robusto frente al ruido y conceptualmente sencillo, aunque los sistemas que han resultado del perfeccionamiento y de la evolución de los más antiguos se complican con los procesos que incorporan para ganar

fiabilidad. La mayoría de los sistemas de aprendizaje de árboles suelen ser no incrementales, pero existe alguna excepción (Utgoff, 1988 citado en en Herrero & Molina, 2012).

Su representación se basa en una serie de reglas compactadas en forma de árbol, ejemplificando con vectores de pares ordenados atributo-valor, de acuerdo con el formato general en el aprendizaje inductivo a partir de instancias o datos, el concepto que estos adquieren durante el proceso de aprendizaje consiste en un árbol. Cada eje está etiquetado con un par atributo-valor y las hojas con una clase, de forma que la trayectoria que determinan desde la raíz los pares de un dato o instancia de entrenamiento alcanzan una hoja etiquetada -normalmente- con la clase. En el caso de la clasificación en el que se desconoce su clase se hace con la misma técnica, solamente que al atributo clase desconocido, se le asigna de acuerdo con la etiqueta de la hoja a la que se accede. (Herrero & Molina, 2012:124)

La inducción del árbol de decisión menciona Witten, Frank, & Hall (2011) como una de las formas más simples, pero más poderosas de aprendizaje automático y representa una función que toma como entrada un vector de valores de atributo y devuelve un valor de salida de "decisión" denominada clasificación.

Los problemas que se abordan con árboles de decisión, dependen del sistema de aprendizaje específico utilizado y deben cumplir con las siguientes características que mencionan Herrero & Molina, (2012):

- La representación sea mediante vectores de pares atributo-valor, especialmente cuando los valores son disjuntos y en un número pequeño. Los sistemas actuales están preparados

para tratar atributos con valores continuos, valores desconocidos e incluso valores con una distribución de probabilidad.

- El atributo que hace el papel de la clase sea de tipo discreto y con un número pequeño de valores, sin embargo existen sistemas que adquieren como concepto aprendido funciones con valores continuos.
- Las descripciones del concepto adquirido deban ser expresadas en forma normal disyuntiva.
- Que posiblemente existan errores de clasificación en el conjunto de pruebas de entrenamiento, así como valores desconocidos en algunos de los atributos, estos sistemas, por lo general, son robustos frente a los errores del tipo mencionado. (Herrero & Molina, 2012:125)

Actualmente existe una gran variedad de algoritmos de árbol de Decisión (Tree) como: Decisión Stump (árbol de un solo nivel), ID3 y C4.5 (J48), LMT, M5 (M5P), REPTree, etc., así como de reglas de clasificación entre las que se mencionan Tablas de Decisión, JRip, M5Rules, OneR, PART, ZeroR. Señalando que estos nombres pueden diferir dependiendo del SW que se utilice para el procesamiento de DM tal es el caso de WEKA, Rapid Miner, SPSS, R, Matlab, Oracle Data Mining entre otros.

El diseño del modelo PredATIS se desarrolló en el SW WEKA, a través de los algoritmos de árbol de decisión J48 para variable dependiente categórica nominal y REPTree (datos de entrada y variable dependiente numérica y categórica nominal), además se apoya en reglas de clasificación basadas en JRIP (datos categóricos nominales), OneR (datos categóricos

nominales) y ZeroR (datos categóricos nominales y numéricos), a continuación se explican los algoritmos utilizados en este trabajo:

- ID3: desarrollado por Quinlan (1986) que considera el más simple y a la vez potente al crear un árbol de decisión es capaz de tratar con atributos con valores discretos o continuos, el procedimiento se basa en seleccionar un atributo raíz y crear una rama con cada uno de los posibles valores de dicho atributo. Con cada rama resultante (nuevo nodo del árbol), se realiza el mismo proceso, esto es, se selecciona otro atributo y se genera una nueva rama para cada posible valor del atributo. Este procedimiento continúa hasta que se clasifiquen los datos a través de uno de los caminos del árbol. El nodo final de cada camino será un nodo hoja, al que se le asignará la clase correspondiente. Así, el objetivo de los árboles de decisión es obtener reglas o relaciones que permitan clasificar a partir de los atributos (Herrero & Molina, 2012).

En cada nodo del árbol de decisión se debe seleccionar un atributo para seguir dividiendo, y el criterio para elegir es: se selecciona el atributo que mejor separe (ordene) los datos de acuerdo a las clases, empleando la entropía, basada en la teoría de la información que calcula el número de bits (información, preguntas sobre atributos) que hace falta suministrar para conocer la clase a la que pertenece un dato o instancia. Cuanto menor sea el valor de la entropía, menor será la incertidumbre y más útil será el atributo para la clasificación. La definición de entropía que da Shannon en su *Teoría de la Información* (1948). (Herrero & Molina, 2012:125)

ID3 se fundamenta en la iteración, iniciando con un subconjunto total de datos de entrenamiento denominada window, seleccionada de manera aleatoria para formar un árbol

de decisión, que clasifica de manera correcta todos los objetos que pertenecen a window. El resto de los objetos, dentro del conjunto de datos de entrenamiento, es clasificado utilizando dicho árbol.

Si el árbol da una respuesta correcta para todos estos objetos, entonces también es correcto para el conjunto total de datos de entrenamiento, terminando así el proceso.

Si no, una selección de objetos clasificados incorrectamente es adicionada al subconjunto window, y el proceso continúa. (Torrado A., Barrientos A., & Vizcaino H., 2014: 5)

- Algoritmo J48: Es una implementación del algoritmo C4.5 que es una extensión del ID3 propuesto por Quinlan (1986). J48 es uno de los más utilizados en modelos de predicción de estadísticos educativos como se observa en la Tabla 6 y en múltiples aplicaciones.

Se argumenta que uno de los parámetros más importantes es el factor de confianza para la poda (confidence level) que influye en el tamaño y capacidad de predicción del árbol construido. Para cada operación de poda, define la probabilidad de error que se permite a la hipótesis de que el empeoramiento debido a esta operación es significativo. A probabilidad menor, se exigirá que la diferencia en los errores de predicción antes y después de podar sea más significativa para no podar. El valor por defecto es del 25%. Según baje este valor, se permiten más operaciones de poda y construyendo árboles cada vez más pequeños. (Jiménez & Álvarez S., 2010: 6)

- REPTree usa la ganancia de información y realiza una poda de error reducido. Se aplica en datos categóricos nominales y numéricos (Witten, Frank, & Hall, 2011 citado en Márquez, 2015)

El algoritmo de Clasificación con árbol de decisión referenciado por Witten, Frank, & Hall (2011) tiene como objetivo, encontrar un árbol pequeño consistente con los ejemplos de entrenamiento y con la Idea: (recursivamente) elegir el atributo "más significativo" como raíz del (sub) árbol.

```
function DECISION-TREE-LEARNING(examples, attributes, parent_examples)
  returns a tree

  if examples is empty then return plurality-value(parent_examples)

  else if all examples have the same classification then return the classification

  else if attributes is empty then return plurality-value(examples)

  else
     $A \leftarrow \operatorname{argmax}[a \in \text{attributes}] \text{IMPORTANCE}(a, \text{examples})$ 

    tree  $\leftarrow$  a new decision tree with root test A

    for each value vk of A do
      exs  $\leftarrow \{ e : e \in \text{examples} \text{ and } e.A = vk \}$ 

      subtree  $\leftarrow$  DECISION-TREE-LEARNING(exs, attributes - A, examples)

      add a branch to tree with label (A = vk) and subtree subtree

    return tree
```

b) Las Reglas de Clasificación o reglas de inducción

Generalmente emplean un enfoque que va de lo específico a lo general, en el cual, las reglas obtenidas son generalizadas hasta que se obtiene una descripción satisfactoria de cada clase. (Márquez, 2015: 47)

A continuación se describen los algoritmos que utilizan las reglas de inducción y se utilizan en este trabajo.

- JRIP propuesto por (Cohen, 1995), en este algoritmo se implementa un generador de reglas proposicionales, siendo uno de los más rápidos y efectivos. Se aplica en datos categóricos nominales. (Márquez, 2015)
- OneR referenciado por (Holte, 1993) el cual usa el atributo de mínimo error para la predicción de una clase, es uno de los primeros clasificador de reglas de inducción. Aplica a datos categóricos nominales. (Márquez, 2015)
- ZeroR, es uno de los primeros clasificadores de reglas de inducción, que predice la clase de la mayoría en el entrenamiento de datos para problemas con un valor de clase categórico o nominal y el valor de clase promedio para problemas con predicción numérica. Siendo útil para generar una línea de base de otras técnicas de clasificación (Witten, Frank, & Hall, 2011).

c) Selección de Atributos

Estos algoritmos intentan remover aquellos atributos que son irrelevantes o que influyen poco en la variable dependiente o endógena. Actualmente existe una gran variedad de ellos en la que pueden ser agrupados dos componentes y (Márquez, 2015:44)

- Atributo evaluador incluye: CfsSubsetEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, ReliefFAttributeEval y SymmetricalUncertAttributeEval, etc.
- Por método de búsqueda se citan: BestFirst, GreedyStepwise y Ranker.

La evaluación del clasificador de selección de atributos, se puede aplicar en dos modos:

- Cross-Validation: evaluación con validación cruzada. Se dividirán las instancias en tantas carpetas como indica el parámetro “Folds” y en cada evaluación se toman las instancias

de cada carpeta como datos de test y el resto como datos de entrenamiento para construir el modelo. Los errores calculados serán el promedio de todas las ejecuciones.

- Use training set: evaluación del clasificador sobre el mismo conjunto sobre el que se construye el modelo predictivo para determinar el error, que en este caso se denomina error de resustitución.

Los algoritmo de selección de atributos utilizados en el análisis que se presenta en este trabajo son: CfsSubsetEval y BestFirst.

d) Reglas de Asociación

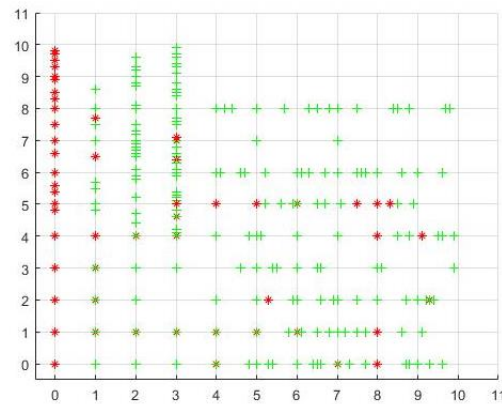
Permite realizar búsqueda automática de reglas de relacionan conjuntos de atributos entre sí, normalmente se aplican en análisis exploratorios; los algoritmos de asociación se les considera no supervisados, ya que no existen relaciones conocidas a priori con las que contrastar la validez de los resultados, sino que se evalúa, si esas reglas son estadísticamente significativas; las asociaciones identificadas pueden usarse para predecir comportamientos, y permiten descubrir correlaciones y co-ocurrencias de eventos. Por su forma se extracción del conocimiento se fundamenta en técnicas estadísticas, análisis de correlación y de variación. El algoritmo A Priori, es uno de los más utilizados (Jiménez & Álvarez S., 2010).

- A Priori algoritmo que sólo busca reglas entre atributos simbólicos, por lo cual todos los atributos numéricos deberían ser discretizados previamente (Jiménez & Álvarez S., 2010).

Resumiendo que en el análisis exploratorio que se presenta en este trabajo, se realizó con base en los algoritmos de clasificación J48, REPTree, JRIP, OneR y ZeroR, y en la selección de atributos con CfsSubsetEval y BestFirst, cuyo entrenamiento se realizó en el SW WEKA y que se detallan resultados en la 5.1.1 Modelado basado en algoritmos de Clasificación y Selección de DM.

4.2.5 Diseño del Modelo PredATIS con RNA.

En esta sección se presenta el diseño del modelo PredATIS cuyo objetivo es predecir a los estudiantes con mayor probabilidad de riesgo a baja de la IES en estudio, donde se propone algoritmos de DM y RNA cuyo proceso de extracción de conocimiento es a partir de un conjunto de datos pre procesados, como fue descrito en la sección 4.2.2 Integración y Recopilación de Datos en la CVAATI. y 4.2.3 Selección, Limpieza y Transformación de Datos en CVAATI., además del análisis de modelos predictivos relacionados a temas de estadísticos educativos en la Tabla 6. De esta forma se selecciona el algoritmo por aplicar, donde se inicia con pruebas en primera instancia con el entrenamiento supervisado de Adaline basado en el aprendizaje de Widrow & Hoff general e iterativo que solo es aplicable a problemas de separabilidad lineal, observando en la Gráfica 2 que es un problema de separabilidad NO lineal, por lo tanto no es posible abordar esta problemática, así entonces, se busca otras alternativas de solución y se decide realizar una segunda prueba con algoritmo de aprendizaje híbrido (combinación del No Supervisado con Supervisado) que son RNA de Función de Base Radial (FBR).



Gráfica 2. Comportamiento de los datos de los estudiantes del PE del primer período escolar
Septiembre – Diciembre 2016.

Fuente: Elaboración Propia.

Señalando que el modelo propuesto se realiza dentro del marco de trabajo de la DM, a través de métodos guiados por algoritmos supervisados (predictivos) y no supervisados (descriptivos), aplicando técnicas de IA y estadísticas. Derivado de este primer análisis se decide utilizar RNA de FBR y la clasificación basada en árboles de decisión y selección de mejores atributos, siendo los que más se adaptaron; aun cuando se reporta que los más utilizados son los de MLP en los estudios presentados en la Tabla 14.

Las RNA de FBR son aquellas cuyas funciones de activación en los nodos ocultos son radialmente simétricas, en donde su distancia depende entre un vector que almacena los datos de entrada y un vector de pesos sinápticos, denominado “centro o centroide”. El diseño básico consta de tres capas separadas como se visualiza en la Figura 27 (M. Awad, 2005).

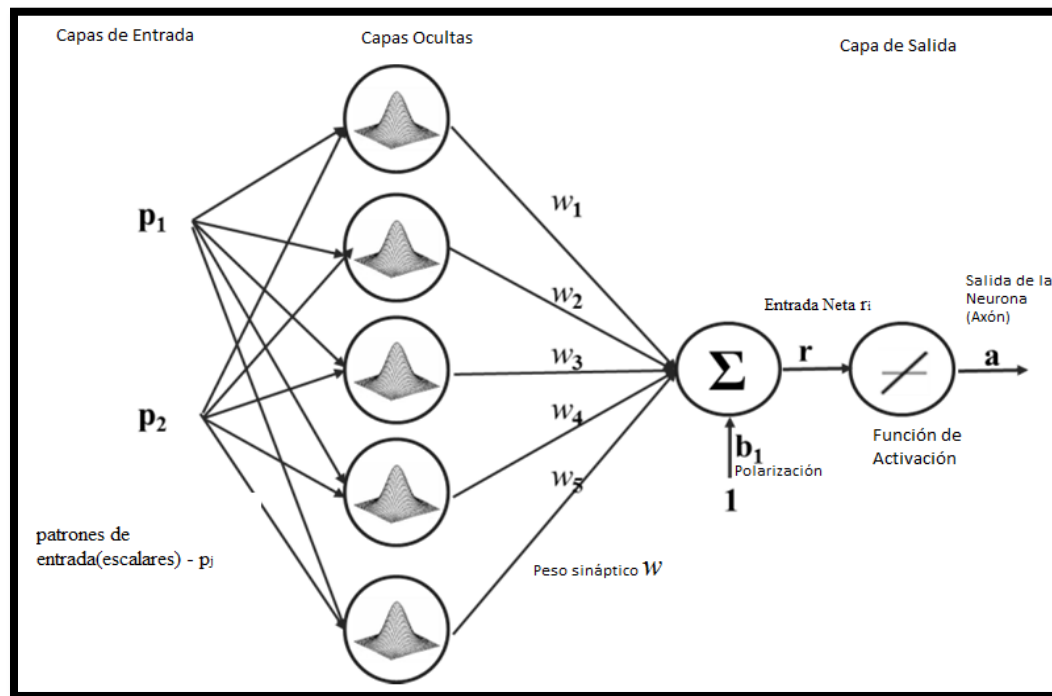


Figura 27. Arquitectura de una RNA de FBR.

Fuente: Elaboración propia a partir de M. Awad (2005).

- d) Capa de entrada: es el conjunto de patrones de entrenamiento o datos fuentes que representan las neuronas o los nodos fuente que reciben las señales del exterior, para ser transmitidas a la siguiente capa de la RNA sin realizar ningún procesamiento sobre dichas señales (M. Awad, 2005).
- e) Capa oculta: aquí las neuronas reciben las señales de la capa de entrada y realizan una transformación local y no lineal sobre estas señales y se activan en función de la distancia que separa cada patrón de entrada con respecto al centroide que cada neurona oculta almacena, aplicando una función radial denominada gaussiana como se visualiza en la Figura 28, finalmente este procesamiento se envía a la capa de salida (M. Awad, 2005).

- f) Capa de salida: realiza una combinación lineal de las activaciones de las neuronas ocultas a través del cálculo de la suma ponderada, siendo estas neuronas la que representan la respuesta de la RNA del modelo (Longoni, Porcel, López, & Dapozo, 2010).

Por su parte M. Awad (2005) menciona que las características principales de la RNA de FBR son: redes de capa de conexión hacia adelante, los nodos ocultos implementan un conjunto de funciones de base radial (funciones gaussianas), los nodos de salida implementan funciones de suma lineales, el entrenamiento se realiza en dos etapas primero se determinan los pesos de la entrada a la capa oculta y posteriormente se calculan los pesos de la capa oculta a la capa de salida, el aprendizaje es muy rápido, además de que son muy buenas en la interpolación y aproximación de funciones, como inicialmente se utilizaron.

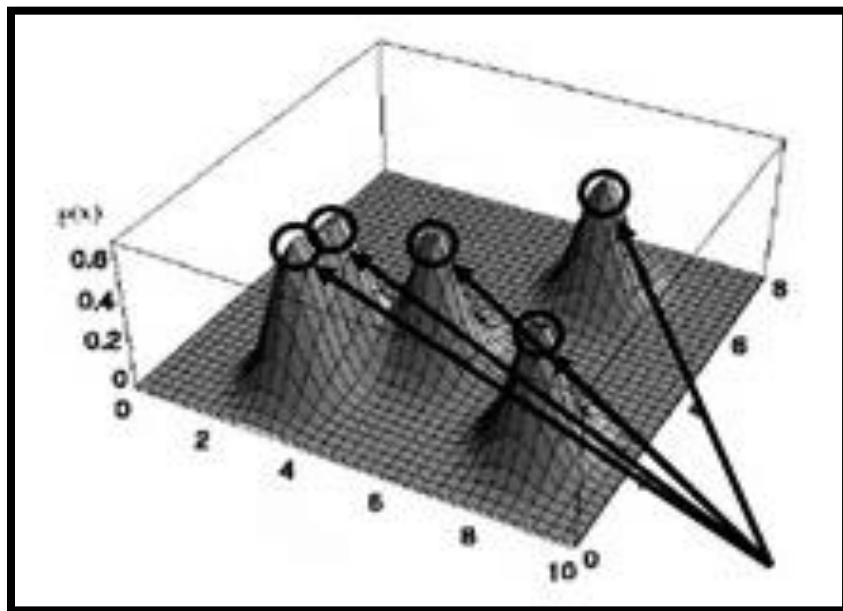


Figura 28. Función Gaussiana de RNA de FBR.

Fuente: Elaboración propia a partir de M. Awad (2005).

La función gaussiana mostrada en la Figura 28, se ajusta modificando los centroides μ y su desviación estándar σ , como se presentan en cada neurona de la capa oculta de la Figura 27 y calculando la aproximación a través de la fórmula:

$$p(\mathbf{x}) = \sum \mathbf{w}_j \phi(\|\mathbf{x} - \mu_j\|) \quad \text{Ecuación 1}$$

Las RNA de FBR se basan en el modelo matemático, dado en la ecuación:

$$a_k = \sum_{j=1}^M w_{kj} \phi(x_L) \quad \text{Ecuación 2}$$

De forma que ϕ_j se calcula con la siguiente ecuación:

$$\phi_j = e^{-\frac{\|x_L - \mu_j\|^2}{2\sigma_j^2}} \quad \text{Ecuación 3}$$

En tanto que el entrenamiento con redes de FBR se puede realizar con dos métodos como lo describen Sánchez (2004), Acquatella (2009) y M. Awad (2005):

- 1) El Método 1 aplica primero un proceso de aprendizaje no supervisado utilizando k-medias, mapas auto-organizados, entre otros y posteriormente la adaptación supervisada de los pesos sináptico W_{kj} (M. Awad, 2005).

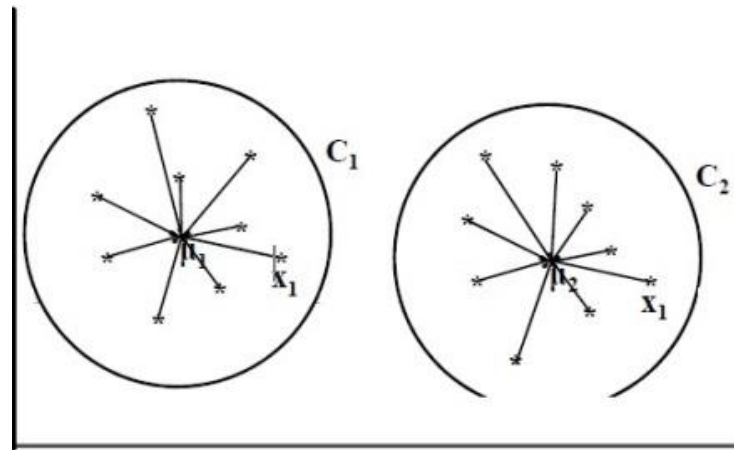
En el entrenamiento No Supervisado con k-Medias consiste en identificar:

- k conjuntos C de forma que: cada conjunto C_j sea lo más diferente posible a los demás y los patrones x_L dentro de cada C sean lo más parecidos posible entre ellos.
- El centroide μ_j de cada conjunto C_j .

El aprendizaje con k-medias (k-means) se basa en la siguiente ecuación de optimización.

$$\min \left\{ J = \sum_{j=1}^k \sum_{L \in C_j} \| \mathbf{x}_L - \boldsymbol{\mu}_j \|^2 \right\} \text{ Ecuación 4}$$

La ecuación de optimización identifica cada centroide como se visualiza en la **Gráfica 3**.



Gráfica 3. Centroides identificados con k-medias

Fuente: elaboración propia a partir de M. Awad (2005)

En tanto que la Figura 29 describe el procedimiento para el entrenamiento con k-medias y en la cual se realiza el modelado de PredATIS que se presentan en este trabajo.

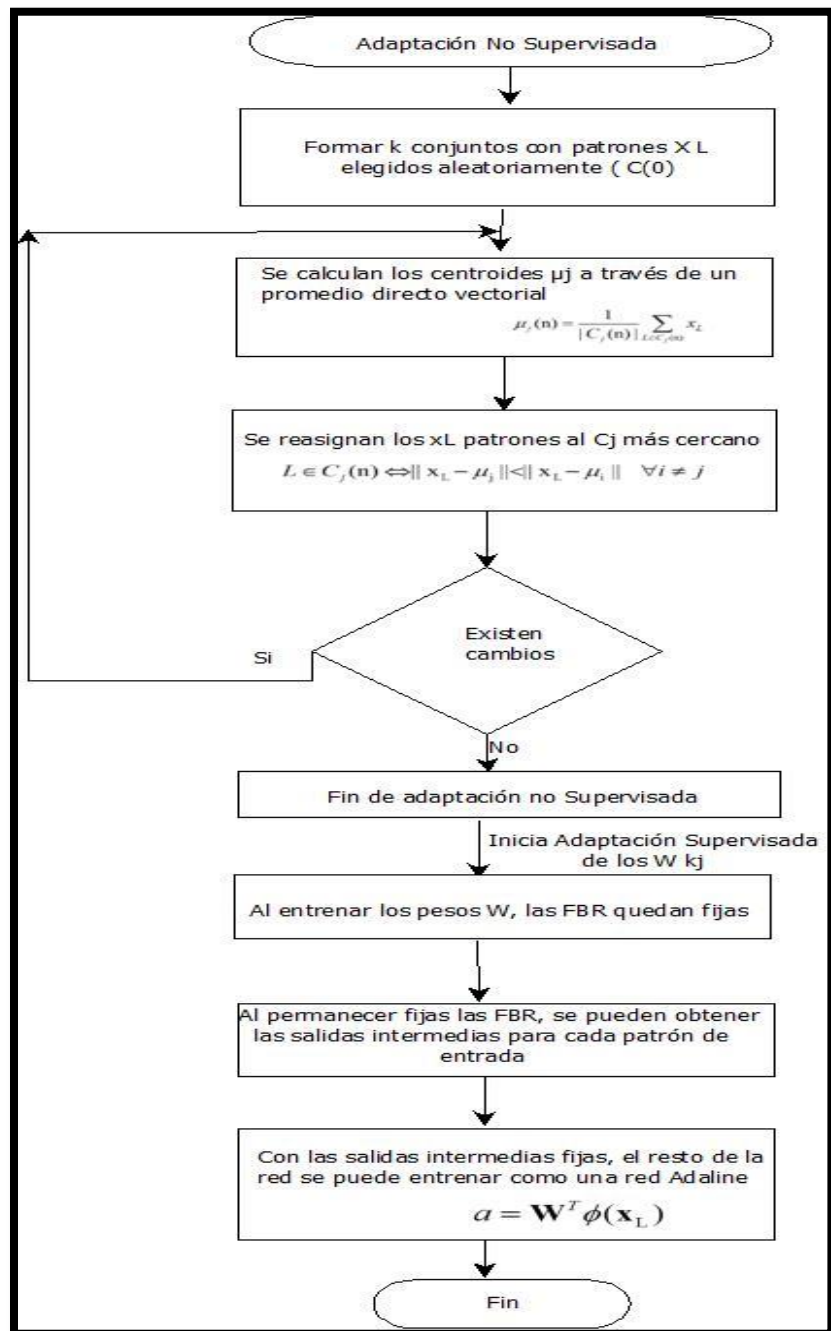


Figura 29. Proceso de Aprendizaje con k-medias

Fuente: Elaboración propia a partir de M. Awad (2005)

- 2) El Método 2 se inicializa primero por el método 1, luego se realiza la adaptación supervisada de las FBR través de la formula $\left(\frac{\delta e}{\delta \mu}, \frac{\delta e}{\delta \sigma}\right)$ y que no se incluyen en el estudio de este trabajo.

Se puede concluir que en general se adaptan las funciones de base radial y los pesos $\mathbf{w}$ por separado (Sánchez, 2004; Acquatella, 2009 & M. Awad, 2005).

Señalando que el modelado que se presentan en este trabajo se realiza a través del método 1 ya descrito, razón por la cual no se detalla el método 2.

CAPÍTULO 5. MODELADO DE PredATIS EN LA CVAATI

5.1 Modelado de PredATIS con algoritmos de DM

El modelo se diseña con técnicas de clasificación, que son: árbol de decisión, que permiten investigar las correlaciones existentes entre las variables de mayor impacto en la deserción de los estudiantes; así como de reglas de clasificación, asociación, selección de mejores atributos para identificar las variables predictivas o causas que inciden con mayor frecuencia en la deserción o reprobación del estudiante, siendo la base del análisis de dos experimentos desarrollados en esta investigación.

5.1.1 Modelado basado en algoritmos de Clasificación y Selección de DM.

El modelo se desarrolla considerando el conjunto de datos y variables de entrada para el entrenamiento, una muestra representativa de 170 estudiantes pertenecientes a un PE de nivel TSU del tercer período escolar mayo-agosto'2017 de la IES en estudio, con el propósito de validar los algoritmos aplicados, derivado de que ya se conocen las causas de deserción y reprobación de este ciclo escolar reportadas en las estadísticas de la propia Institución que son:

- Deserción: nunca se inscribió al cuatrimestre inmediato, nunca se presentó pero si se inscribió, sólo se presentó los primeros 15 días del inicio de cuatrimestre, por cambio a otra universidad, por cambio de carrera en la misma IES, por cambio de domicilio, por incumplimiento de expectativas del modelo educativo, por problemas de vocación (por parte del alumno), por la distancia de la universidad, por problemas de salud, por problemas económicos, por problemas familiares, por embarazo, por problemas laborales, matrimonio, motivos personales, faltas al reglamento (más del 20% de

inasistencias), falta de certificado de bachillerato, sin causas conocidas, cambio de especialidad o área de la misma carrera de TIC.

- Reprobación: faltas al reglamento (reprobó más de 2 materias), reprobación en el extraordinario y reprobación en prácticas profesionales.

Después de generar la vista minable se continúa el modelado, que en primer lugar se aplica un filtro no supervisados en el nivel de atributos Discretizados, para manipulación de los datos, en donde se transforma los atributos numéricos seleccionados en atributos simbólicos, con una serie de etiquetas que resultan de dividir la amplitud total del atributo en intervalos. Como por ejemplo, el PE es su primera Opción fue dividida en 6 intervalos: Si, No aprobé en otras IES o no me inscribe al Perfil deseado, Indeciso, Obligado por padres/Cercanía de la Universidad, Fue mi última opción, No sin especificar motivo, Prefiero no responder pregunta y No realizó Encuesta, como se observa en la Figura 30.

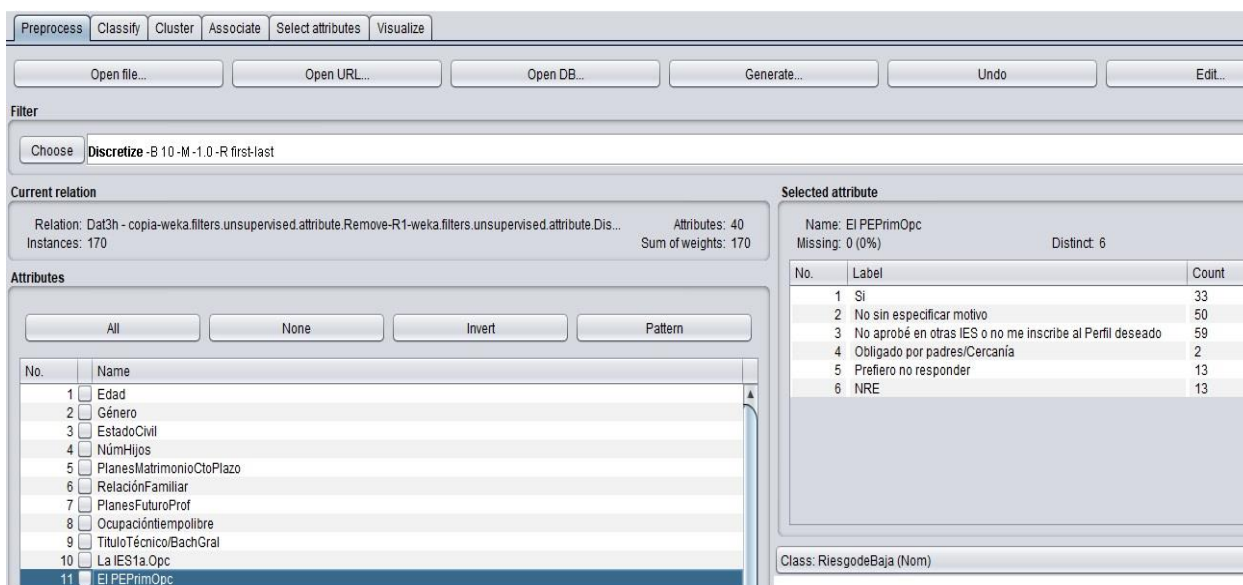


Figura 30. Filtro Aplicado al conjunto de datos de entrada del Modelo DM.

Fuente: Elaboración Propia.

Después de aplicar el Filtro Discretizado en los 39 atributos, se continúa con la ejecución de pruebas y entrenamiento desarrollados en WEKA, en donde se incluye tres tareas principales que son: Selección de Atributos, aplicación de reglas de clasificación, clasificación con árboles de decisión y visualización.

Selección de Mejores Atributos. En primera instancia, se aplica en la muestra representativa del conjunto de datos de las 39 variables descritas en la Tabla 16 y Tabla 18, los algoritmos de atributo evaluador CfsSubsetEval y método de búsqueda BestFirst, con validación cruzada (Cross-Validation- Folds 10 y seed 1); con el propósito de identificar las variables de interés de alto impacto en la deserción de los estudiantes, como se identifica en la Figura 31, donde se seleccionan 15 atributos más representativos del total de 39 variables de estudio, cuya variable a predecir es Riesgo de Baja (Si/No).

En tanto que la validación con el total del conjunto de datos (Use full training set:), atributo evaluador CfsSubsetEval y método de búsqueda BestFirst, no se utiliza en este trabajo, debido los resultados que resume en solo dos atributos: PlanesMatrimonioCtoPlazo y El PEPrimOpc, no son los más representativos en la IES. Es por ello que se toman como referencia para continuar con el modelado con la primera validación de Cross-Validation.

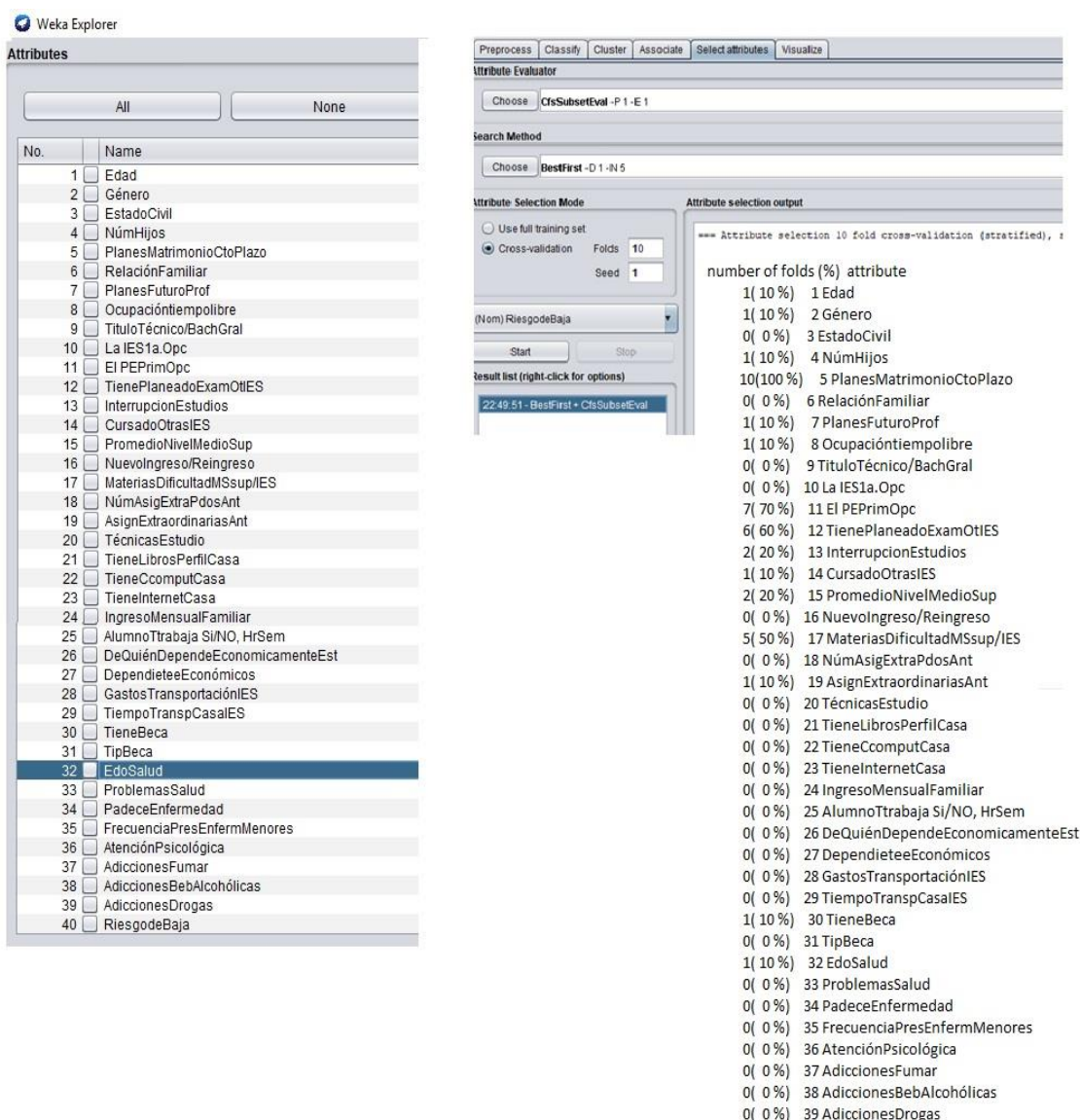


Figura 31. Aplicación de algoritmo de Selección de Atributos.

Fuente: Elaboración Propia.

De esta forma se realiza una comparativa de los atributos seleccionados por los algoritmos aplicados en el modelo, con los identificados en el reporte de la IES en estudio, donde se observa en la Tabla 19, que 10 de las 15 variables seleccionadas, si se relacionan como causas de deserción y reprobación de estudiantes.

Tabla 19, Factores identificados por Modelo DM

Variable de Entrada Modelado- Clasificación DM	Causa que se reporta en la IES
10(100 %) 5 PlanesMatrimonioCtoPlazo	Matrimonio Por embarazo Por cambio de carrera en la misma IES Cambio de especialidad de Sistema Informáticos a Redes y Telecomunicaciones Cambio de especialidad de Redes y Telecomunicaciones a Sistema Informáticos
7(70 %) 11 El PEPrimOpc	Telecomunicaciones a Sistema Informáticos
6(60 %) 12 TienePlaneadoExamOtIES	Por cambio a otra Universidad
5(50 %) 17 MateriasDificultadMSup/IES	Por problemas de vocación (por parte del alumno)
2(20 %) 15 PromedioNivelMedioSup	
2(20 %) 13 InterrupcionEstudios	
1(10 %) 32 EdoSalud	Por problemas de salud
1(10 %) 30 TieneBeca	Por problemas económicos
1(10 %) 19 AsignExtraordinariasAnt	Faltas al reglamento (Reprobó más de 2 materias)
1(10 %) 14 CursadoOtrasIES	Reprobación en el extraordinario Nunca se presentó pero si se inscribió Sólo se presentó los primero 15 días del inicio de cuatrimestre
1(10 %) 8 Ocupacióntiempolibre	
1(10 %) 7 PlanesFuturoProf	Por incumplimiento de expectativas del modelo educativo
1(10 %) 4 NumHijos	Matrimonio
1(10 %) 2 Genero	
1(10 %) 1 Edad	

Fuente: Elaboración Propia.

Reglas de clasificación. Una vez seleccionadas las variables representativas de los mejores atributos de causas de deserción o reprobación, se continúa con la aplicación de algoritmos de clasificación JRIP, OneR, ZeroR , realizados en sus tres tipos de validación: Porcentaje split 66, Cross-Validation Test y Training set, cuyos resultados se muestran en la Tabla 20; donde se observa que la precisión con el nivel de error absoluto menor es de 56.6650% con la validación del conjunto total de datos (Training set) y clasificador OneR – B.6, reflejando los resultados lógicos de la diagonal de matriz de confusión con 4 y 140 de clasificación correctamente positiva y negativa respectivamente.

Tabla 20. Resultados obtenidos con Reglas de Clasificación DM

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de Instancias	Error Absoluto Relativo	Matriz de Confusión
	Instancias	%	Instancias	%			
Aplicando el clasificador JRIP classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1							
Percentage split 66	44	75.8621%	14	24.1379%	58	107.3899%	a b classified as 2 8 a = Si 6 42 b = No
Cross- Validation Test	139	81.7647%	31	18.2353%	170	100.1004%	a b classified as 1 126 a = Si 5 136 b = No
Training set	147	86.4706%	29	13.5294%	170	86.3629%	a b classified as 4 23 a = Si 0 143 b = No
Aplicando el clasificador OneR – B.6							
Percentage split 66	44	75.8621%	14	24.1379%	58	87.5000%	a b classified as 1 9 a = Si 5 43 b = No
Cross- Validation Test	137	80.5882%	33	19.4118%	170	71.9996%	a b classified as 2 23 a = Si 8 135 b = No
Training set	144	84.7059%	26	15.2941%	170	56.6650%	a b classified as 4 23 a = Si 3 140 b = No
Aplicando el clasificador ZeroR							
Percentage split 66	48	82.7586%	10	17.2414%	58	100.0000%	a b classified as 0 10 a = Si 0 48 b = No
Cross- Validation Test	143	84.1176%	33	15.8824%	170	100.0000%	a b classified as 0 27 a = Si 0 143 b = No
Training set	143	84.1176%	33	15.8824%	170	100.0000%	a b classified as 0 27 a = Si 0 143 b = No

Fuente: Elaboración propia.

Árbol de decisión. Siendo las técnicas de clasificación más utilizadas, por ello, se aplican los algoritmos J48 y REPTree y las tres formas de validación; Percentage split 66, Cross-Validation Test y Training set; mostrando los resultados obtenidos en la Tabla 21. Donde se pretende dar un refinamiento al análisis, que se desarrolla en el modelo PredATIS, cuya variable a predecir o endógena es Riesgo de Baja categorizada en: Si o No es probable la deserción del estudiante.

Tabla 21. Resultados obtenidos con árbol de decisión DM

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de Instancias	Error Absoluto Relativo	Matriz de Confusión
	Instancias	%	Instancias	%			
Árbol decisión J48 con factor de confianza = 0.25							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5491%	a b classified as 0 10 a = Si 0 48 b = No
Cross- Validation Test	143	84.1176%	27	15.8824%	170	98.8897%	a b classified as 0 27 a = Si 0 143 b = No
Training set	143	84.1176%	27	15.8824%	170	98.9970%	a b classified as 0 27 a = Si 0 143 b = No
Árbol decisión J48 con factor de confianza = 0.50							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5491%	a b classified as 0 10 a = Si 0 48 b = No
Cross- Validation Test	141	82.9412%	29	17.0588%	170	86.5505%	a b classified as 2 25 a = Si 4 139 b = No
Training set	147	86.4706%	23	13.5294%	170	79.6133%	a b classified as 4 23 a = Si 0 143 b = No
Árbol decisión J48 con factor de confianza = 0.60							
Percentage split 66	40	68.9655%	18	31.0345%	58	106.4340%	a b classified as 1 9 a = Si 9 39 b = No
Cross- Validation Test	127	74.7059%	43	25.2941%	170	91.4633%	a b classified as 5 22 a = Si 21 122 b = No
Training set	155	91.1765%	15	8.8235%	170	47.5863%	a b classified as 13 14 a = Si 1 142 b = No
Árbol decisión J48 con factor de confianza = 0.75							
Percentage split 66	40	68.9655%	18	31.0345%	58	106.4340%	a b classified as 1 9 a = Si 9 39 b = No
Cross- Validation Test	127	74.7059%	43	25.2941%	170	91.4633%	a b classified as 5 22 a = Si 21 122 b = No
Training set	155	91.1765%	15	8.8235%	170	47.5863%	a b classified as 13 14 a = Si 1 142 b = No

Tipo Training/Test	Clasificación correcta		Clasificación Incorrecta		Número Total de Instancias	Error Absoluto Relativo	Matriz de Confusión
Árbol decisión J48 con factor de confianza = 1.0							
Percentage split 66	40	68.9655%	18	31.0345%	58	106.4340%	a b classified as 1 9 a = Si 9 39 b = No
Cross- Validation Test	127	74.7059%	43	25.2941%	170	91.4633%	a b classified as 5 22 a = Si 21 122 b = No
Training set	155	91.1765%	15	8.8235%	170	47.5863%	a b classified as 13 14 a = Si 1 142 b = No
Árbol decisión REPTree con trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1 -I 0.0							
Percentage split 66	48	82.7586%	10	17.2414%	58	98.5991%	a b classified as 0 10 a = Si 0 48 b = No
Cross- Validation Test	140	82.3529%	30	17.6471%	170	99.9465%	a b classified as 0 27 a = Si 3 140 b = No
Training set	143	84.1176%	27	15.8824%	170	98.9970%	a b classified as 0 27 a = Si 0 143 b = No

Fuente: Elaboración propia.

Observando en la Tabla 21, que el mejor resultado de clasificación se obtiene con el algoritmo J48 con un factor de confianza de 0.60, 0.75 y 1.0, en el conjunto total de datos de entrenamiento (Training set) categóricos y discretizados, comparado con JRIP, OneR y ZeroR.

5.2 Modelado de PredATIS con algoritmos de RNA.

Derivado del análisis y primer entrenamiento presentado en la sección 4.2, donde se observó un problema de separabilidad NO Lineal en la Gráfica 2, con datos de 156 estudiantes de la IES de este trabajo, por lo que no es posible aplicar algoritmos supervisados como Adaline basado en el aprendizaje de Widrow & Hoff general e iterativo cuyos resultados obtenidos de la programación y ejecución en Matlab se presentan en la Figura 32 y Figura 33, que no resulta útil

para resolver el problema que se aborda y solo se utilizó para fines de prueba. Seleccionando de esta forma el diseño de PredATIS con Redes Neuronales (RNA) de Función de Base Radial (FBR) basado en el aprendizaje no Supervisado con entrenamiento k-medias ya descrito.

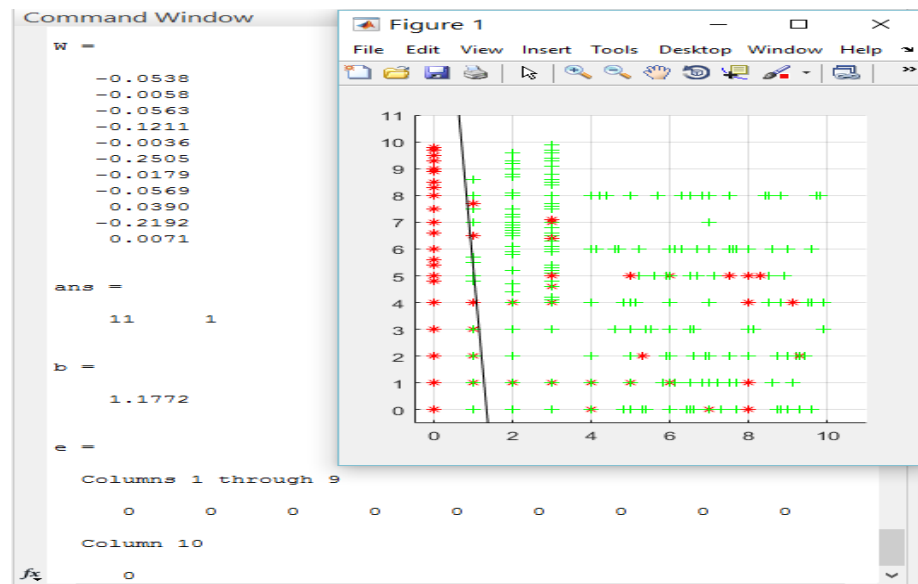


Figura 32. Resultados del comportamiento de patrones con Adaline -Widrow & Hoff.

Fuente: Elaboración Propia.

En la Figura 32 se observa que los * son estudiantes en riesgo de baja y los + son estudiantes sin riesgo de baja de la IES, por lo que no se puede visualizar una *frontera de decisión* que clasifique en estas dos categorías, dado que no existe una separabilidad lineal, siendo uno de los requisitos para iniciar un análisis con este algoritmo, aun cuando el resultado de precisión es de 0, una vez procesado el conjunto de patrones de entrenamiento que se muestra parte del conjunto en la Figura 33, en donde cada columna representa un valor de p_i , y cada fila los datos de cada estudiante; así, entonces se demuestra que no es útil para resolver este tipo de predicción del problema que se aborda en este trabajo.

P =

Columns 1 through 7

2.0000	2.0000	3.0000	5.0000	8.0000	1.0000	2.0000
1.0000	4.0000	3.0000	4.0000	6.5000	1.0000	5.0000
1.0000	1.0000	2.0000	6.0000	8.5000	1.0000	6.0000
1.0000	3.0000	1.0000	3.0000	7.0000	1.0000	5.0000
2.0000	0	0	0	0	0	8.0000
1.0000	1.0000	1.0000	6.0000	6.0000	1.0000	2.0000
2.0000	0	0	0	0	0	2.0000
2.0000	3.0000	3.0000	7.0000	6.0000	1.0000	4.0000
1.0000	3.0000	1.0000	6.0000	7.0000	1.0000	3.0000
1.0000	1.0000	2.0000	5.0000	7.5000	1.0000	6.0000
2.0000	0	0	0	0	0	1.0000
1.0000	1.0000	3.0000	4.0000	7.0000	1.0000	0
2.0000	3.0000	1.0000	6.0000	7.0000	1.0000	1.0000
2.0000	0	0	0	0	0	8.0000
1.0000	3.0000	3.0000	7.0000	4.0000	1.0000	2.0000
2.0000	1.0000	2.0000	6.0000	8.5000	1.0000	0
2.0000	0	0	0	0	0	8.0000
2.0000	1.0000	2.0000	3.0000	8.0000	1.0000	4.0000
1.0000	1.0000	2.0000	6.0000	8.0000	1.0000	1.0000
2.0000	3.0000	3.0000	7.0000	7.0000	2.0000	1.0000
1.0000	1.0000	2.0000	6.0000	7.5000	1.0000	0
1.0000	2.0000	3.0000	5.0000	7.0000	1.0000	2.0000
2.0000	1.0000	1.0000	0	4.0000	1.0000	4.0000
1.0000	3.0000	1.0000	6.0000	6.5000	1.0000	0
1.0000	1.0000	3.0000	7.0000	7.0000	1.0000	1.0000
1.0000	0	0	0	0	0	7.0000
2.0000	0	0	0	0	0	6.0000
2.0000	1.0000	2.0000	7.0000	8.0000	1.0000	4.0000
2.0000	3.0000	1.0000	3.0000	6.5000	1.0000	3.0000
1.0000	0	0	0	0	0	8.0000
1.0000	1.0000	3.0000	0	7.5000	1.0000	6.0000

Figura 33. Conjunto parcial de Patrones de entrenamiento del Algoritmo de Adaline - Widrow & Hoff General.

Fuente: Elaboración Propia.

5.2.1 Modelado de PredATIS basado en RNA de FBR.

El Modelo para la predicción de los estudiantes en riesgo de baja de la IES propuesta, se realiza primero con base en una serie de experimentos tomando dos muestras para el modelado, una se aplica al PE de TIC de nivel TSU área de sistemas del primer período escolar y la segunda se aplica al tercer periodo escolar del mismo PE que incluye área de Sistemas y Redes

y Telecomunicaciones; con la finalidad de validar los algoritmos aplicados, derivado de que ya se conocen las estadísticas de los indicadores de deserción de estos ciclos escolares.

El análisis se inicia, a partir de datos recopilados de la encuesta aplicada a estudiantes del PE en ambas muestras, además de datos extraídos del Sistema de Información Integral de la IES en estudio, donde se integran los datos, se realiza la selección, limpieza y transformación de Datos y variables descritas en la sección 4.2.2 y 4.2.3, así fueron definidas en la Tabla 15, Tabla 17 y Tabla 18, posteriormente son pre procesadas para generar la Vista Minable antes de diseñar el modelo basado en RNA de FBR, generando archivos de tipo .csv y .txt, siendo los datos de entrada o el conjunto de patrones de entrenamiento, que representan las variables del estudio para el modelado, donde se realizó una serie de experimentos que se describen a continuación.

Experimento 1. El estudio se realiza en la muestra 1 con datos de 156 estudiantes del primer periodo escolar Septiembre – Diciembre 2016 del PE de TIC de nivel TSU Sistemas, seleccionando en primera instancia las variables de interés de alto impacto en la deserción de los estudiantes que se han observado en la estadísticas reportadas de indicadores educativos de la IES en estudio, por lo que son consideradas como los valores de entrada identificadas con $p_1, p_2, \dots, p_{11}$, y que se describen en la Tabla 22, representando los patrones de entrenamiento de entrada en las neuronas de las capas ocultas, utilizadas en los algoritmos de RNA de FBR, en el análisis de predicción de estudiantes en riesgo de baja y posteriormente comprobar con las estadísticas reportadas en la IES, cuyo propósito es identificar la precisión y el nivel de confiabilidad del algoritmo respecto a los experimentos realizados.

Tabla 22. Variables de entrada del entrenamiento con RNA de FBR experimento 1

Variable de entrada	Dato Representativo del Estudiante	Aspecto que representa
p1	• Tiempo de Transportación	• Personal
p2	• El PE es tu primer opción	• Vocacional
p3	• Tienes planeado presentar examen en otra IES	• Vocacional
p4	• Materias con dificultad en nivel medio superior/IES	• Académico
p5	• Promedio Evaluación Diagnóstica de Ingreso	• Académico
p6	• Nuevo Ingreso o Reingreso	• Académico
p7	• No de Asignaturas no acreditadas de 1er. Parcial de período actual	• Académico
p8	• Ingreso Mensual Familiar	• Económico
p9	• Alumno Trabaja Si/NO	• Económico
p10	• Estado de salud	• De Salud
p11	• Problemas de salud	• De Salud

Fuente: Elaboración propia.

El primer diseño de la RNA de FBR se realiza en tres actividades, que fueron desarrolladas en el software de Matlab para fines de pruebas y posteriormente se comprobará la funcionalidad del algoritmo previamente entrenado en la CVA.

1. En el diseño de la arquitectura de RNA de FBR la capa de entrada se incluyen las variables de entrada (independiente) identificadas como p1, p2, p3, p4, p5, p6, p7, p8, p9, p10 y p11 representando los datos categorizados de 156 estudiantes de la muestra 1 descritos en la Tabla 22, que representan el conjunto de patrones de entrenamiento en el algoritmo de RNA de RBF que se visualiza una parte del conjunto P en la Tabla 23, que fueron extraídos de los archivos de tipo texto generados del Sistema de Información Integral de la IES y que fueron previamente pre procesados antes de diseñar y programar el modelo, adicionalmente se utiliza una variable de Salida (dependiente) $T1 = \{-1, 1\}$, donde **1** corresponde al alumno que está en riesgo de Baja y **-1** el que tiene probabilidad de continuar sus estudios.

Tabla 23. Conjunto parcial de Patrones de entrenamiento P experimento 1 muestra 1

Item	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11
1	2	2	2	4	3	1	2	2	1	1	1
2	1	4	2	3	5	1	5	0	1	2	1
3	1	1	1	7	3	1	6	1	1	1	5
4	1	3	3	2	4	1	5	1	1	1	4
5	0	0	0	0	0	0	0	0	0	0	0
6	1	1	3	7	5	1	2	1	1	1	5
7	5	3	2	8	6	1	6	4	4	2	5
8	2	3	2	1	5	1	4	3	1	1	5
9	1	3	3	7	4	1	3	1	1	1	1
10	1	1	1	4	4	1	6	0	4	1	1
11	0	0	0	0	0	0	0	0	0	0	0
12	1	1	2	3	4	1	0	4	1	3	3
13	2	3	3	7	4	1	1	0	1	2	1
14	2	1	3	6	4	1	7	6	3	2	8
15	1	3	2	1	6	1	2	3	1	1	1
16	2	1	1	7	3	1	0	1	1	1	1
17	0	0	0	0	0	0	0	0	0	0	0
18	2	1	1	2	3	1	4	1	1	1	1
19	1	1	1	7	3	1	1	1	1	1	4
20	2	3	2	1	4	2	1	3	3	1	4
21	1	1	1	7	4	1	0	2	1	1	5
22	1	2	2	4	4	1	2	1	1	1	1
23	4	4	2	8	7	2	6	5	6	2	1

Item	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11
24	1	3	3	7	5	1	0	1	1	1	1
25	1	1	2	1	4	1	1	4	1	1	1
26	5	3	3	5	7	1	8	5	5	2	1
27	0	1	2	8	6	2	7	6	3	2	1
28	2	1	1	1	3	1	4	2	1	1	1
29	2	3	3	2	5	1	3	1	1	1	4
30	4	4	2	5	4	1	7	4	4	3	1
31	0	1	3	6	7	2	6	3	4	2	1
.....											
154	1	2	3	7	6	1	3	2	1	1	1
155	2	1	2	6	4	2	8	6	3	4	4
156	0	0	0	0	0	0	0	0	0	0	0

Fuente: Elaboración propia.

2. El diseño y la programación del algoritmo de RNA de FBR se realiza en Matlab como una primera aproximación del modelo como se describe en la Tabla 24.

Tabla 24. Algoritmo de RNA de FBR

```

clc, clear, close all
% *****carga de archivos de Datos para el análisis
disp(' Asignando Patrones de Entrenamiento a partir del archivo de Datos...');
load ('Dat1i.txt')
P1 = Dat1i(:,:,);
P=[P1(:,3:14)] % extrae de datos de las 11variables evaluadas para deserción y la Salida T1
Sal1=[P1(:,14:14)]; % extrae los valores de clasificación para evaluar SI el estudiante está en Riesgo de baja o No
size(P);
% *****Grafica de distribución de Datos Clasificados para analizar el comportamiento
figure;
hold off;
figure (1)
hold on
axis([-0.5 9 -0.5 9])
grid on;
plot(P(1,1),P(2,1),'g+',P(1,2),P(2,2),'g+')
%*****plot(P(3,1),P(4,1),'g+',P(1,3),P(4,4),'g+')
title('Gráfica Clasificación de Estudiantes probables a Baja *Rojo');
for i=3:Tn-1
    for j=1:10
        if P(i,12)==1 % clasificación de estudiantes con riesgo de baja
            plot(P(i,j),P(i+1,j+2),'r*') % proyección de baja
        elseif P(i,12)==-1
            plot(P(i,j),P(i+1,j+2),'g+') % proyección de continúan
        end
    end
end

```

```

    end
end
end
%*****Asignando Patrones de Entrenamiento a partir del archivo de Datos*****
% =====
%   Entrenamiento para la red de función de Base Radial
% =====
ent_patrones=P;
% se toman número de k=4 para clasificar a los estudiantes en una sola
% salida para clasificar 1 es cuando es estudiante está en riesgo de baja y
% -1 cuando el estudiante puede continuar
k=6; % se clasifican en k grupos para el estudio % Número de centros a generar
num_salida=1; % número de neuronas de salida, solo clasifica SI es baja =1 del estudiante o NO es baja=-1
disp(' Entrenamiento para la RNA con función de Base Radial ...');
[n,m]=size(ent_patrones); % dimensión de la matriz para el conjunto de Datos
% Generando vector columna con valores entre 1 y k de tamaño n
index=round(1+(k+1).*rand(n,1));
size(index); % tamaño del vector índice
%
o_patrones=ent_patrones;
ent_patrones=[index ent_patrones];
pause;
cambio=true;
medias=zeros(k,m-num_salida)
% calculando centroides con k-medias
disp(' Calculando centroides con k-medias en el entrenamiento ...');
while(cambio)
    % Cambio la bandera a falso
    cambio=false;
    for i=1:(k*1) % para cálculo de las 12 variables
        conjuntoi=ent_patrones(ent_patrones(:,1)==i,2:end-num_salida); %%Seleccionando un Conjunto i de forma
        Aleatoria y almacenando matriz medias
        if isempty(conjuntoi)
            medias(i,:)=rand(1,m-num_salida)-0.999;
        else
            medias(i,:)=median(conjuntoi);
        end % fin del if
    end % fin del for i=k
    for i=1:(n*1)
        patron_e=ent_patrones(i,2:end-num_salida);
        % Calculo la distancia euclidiana del patrón a cada vector de medias
        for j=1:(k*1)
            distancia(j)=sum((patron_e - medias(j,:)).^2);
        end
        [~,grp]=min(distancia);
        if(grp~=ent_patrones(i,1)) % compara si cambia de grupo
            % Si cambia de grupo, se hace el cambio
            ent_patrones(i,1)=grp;
            % Si hay cambio, seguimos bucle para el entrenamiento
            cambio=true;
        end
    end
end
end
end

```

```

medias
conjuntoi;
disp(' Oprima tecla barra espaciadora o enter para continuar...')
pause
ent_patrones=o_patrones;
[K,col]=size(medias);
[n,~]=size(ent_patrones);
fgauss=zeros(n,K+num_salida);
ent_patrones(:,end-num_salida+1:end);
%***** calculando Gaussianas, distancia euclidiana
disp(' Desplegando Gaussianas en el entrenamiento ...');
fgauss(:,end-num_salida+1:end)=ent_patrones(:,end-num_salida+1:end)
for i=1:n
    Xu=ent_patrones(i,1:end-num_salida);
    for j=1:K*1
        dist=Xu-medias(j,:);
        fgauss(i,j)=exp(-1/1*dot((Xu-medias(j,:)),(Xu-medias(j,:))));
    end
    %i
end
disp(' Desplegando ID de Estudiantes probable Baja ...')
indx=find(ent_patrones(:,12)==1) % bajas
disp(' Desplegando ID de Estudiantes que continúan estudios ...')
indy=find(ent_patrones(:,12)==-1) % continúan
%***** graficando Centroides de cada gupo K
figure (2)
hold on
axis([0 3.5 0 4])
plot(medias(:,1),medias(:,2),'ok');
hold on;
title('Gráfica del Entrenamiento con Redes Neuronales de Base Radial');
pxy=-pi:0.001:pi;
for i=1:k
    if k>=10
        co=0.7
    elseif k>=6 & k<=9
        co=0.3
    elseif k<=5
        co=0
    end
    x=((k/10)-co)*cos(pxy)+medias(i,1);
    y=((k/10)-co)*sin(pxy)+medias(i,2);
    plot(x,y,'b');
    hold on
end
disp(' Aplicando Adaline para determinar el error cuadrático del entrenamiento ...');
fadalineFBRAnBaj(fgauss,1)

```

Fuente: Elaboración propia.

El código descrito en Tabla 24, representa la base del entrenamiento y la comparación de la ejecución en el SW desarrollado en la CVA. Se observa que el entrenamiento programado incluye el método 1, en el que realiza primero la adaptación no supervisado utilizando k-medias, calculando las funciones gaussianas y se continúa con el aprendizaje supervisado de los pesos sináptico W_{kj} con Adaline que a su vez llama la función de activación de hardlims que se describe en la Figura 34. Señalando que el código programado permite simular la RNA de FBR con tan solo cambiar el valor de k en el conjunto P que representa los patrones de entrenamiento de las diferentes muestras tomadas para este trabajo (archivos .txt), siendo posible obtener en este primer experimento la ejecución de varias pruebas que permiten la comparación de resultados del entrenamiento aplicado, como se muestra en la Tabla 25 y se describe en la sección de análisis y discusión.

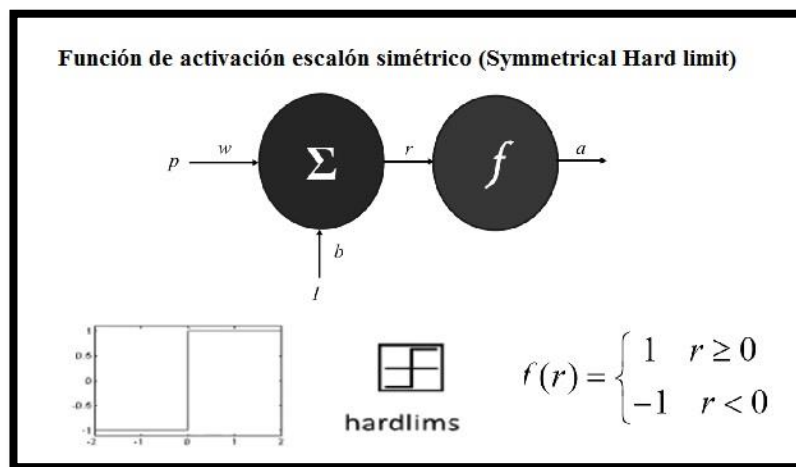


Figura 34. Función de Activación Hardlims.

Fuente: Mora (2011); Ruíz & Basualdo (s.f.).

3. Se analizan los resultados obtenidos del primer entrenamiento con RNA de FBR donde se realizaron 25 ejecuciones con base al conjunto de patrones P descrito en el paso 1 y definidos en la Tabla 22, que representan los aspectos personales, vocacionales, académicos, económicos y de salud, siendo factores detectados en reportes de indicadores educativos de la IES, que impactan, en conjunto con la asignación de forma aleatoria de valores -1 o 1 en la capa de salida T1 de la Red Neuronal, que identifica en la Gráfica 4, **SI** el estudiante es candidato a baja del PE representado con el símbolo de * color rojo y con + color verde cuando **NO** (-1) está en riesgo, es decir que continuará sus estudios en el siguiente período escolar.



Gráfica 4. Clasificación de estudiantes con predicción a Baja de la IES.

Fuente: Mora (2011).

Los resultados obtenidos en el experimento 1 se resumen en la Tabla 25, en donde se visualiza el error de predicción calculado en cada ejecución del algoritmo de aprendizaje de RNA de FBR en Matlab.

Tabla 25. Resultados de la ejecución de Pruebas experimento 1

Error calculado (e) en Matlab con base al número de centros k=4	Error calculado (e) en Matlab con base al número de centros k=5	Error calculado (e) en Matlab con base al número de centros k=6	Error calculado (e) en Matlab con base al número de centros k=9	Error calculado (e) en Matlab con base al número de centros k=11	Número de Estudiantes de Baja con predicción en Algoritmo FBR	Predicción Algoritmo del Número de Estudiantes de Baja con estadísticas reportadas reales	% de predicción obtenido en Algoritmo FBR con estadísticas reportadas	Diferencia de % obtenido en algoritmo con respecto al % real	% de predicción de bajas con Algoritmo FBR y que no fueron reportadas en las estadísticas reales
0.75	0.70	0.70	0.62	0.59	78	16	43.243%	56.757%	52.10%
0.23	0.23	0.23	0.22	0.22	26	16	42.243%	56.757%	8.40%
0.37	0.37	0.37	0.36	0.36	37	20	54.054%	45.946%	14.29%
0.33	0.33	0.32	0.31	0.31	36	24	64.865%	35.135%	10.08%
0.32	0.32	0.30	0.30	0.27	35	26	70.270%	29.730%	7.563%

Fuente: Elaboración propia.

En la Tabla 25, se describen los resultados de 25 ejecuciones a partir del conjunto P, que representan los patrones de entrenamiento tomados de los archivos de texto Dat1e.txt, Dat1f.txt, Dat1g.txt, Dat1h.txt y Dat1i.txt, que fueron extraídos del Sistema Integral de la IES, donde se asignan los valores de la variable de salida T1, para calcular el nivel de error y el porcentaje de predicción en Matlab, observando que el nivel de precisión del algoritmo para la muestra 1 que incluye las variables descritas en la Tabla 22, se obtiene un error de 0.75 a .027 y un porcentaje de predicción de error de 56.757% a 29.730% respecto a las estadísticas reportadas de 37 estudiantes en este período escolar, donde se evidencia el progreso de resultados va hacia la tendencia de llegar al error de 0. Además se señala que se realizaron 20 ejecuciones a prueba y error del experimento, con la finalidad de predecir un mejor el rendimiento del Algoritmo con FBR y que se comparará con la ejecución de entrenamiento en el SW propiamente desarrollado en la etapa de implementación futura en la CVAATI que no se contempla en este trabajo, y solo se tomaron 5 ejecuciones de mejor precisión.

Experimento 2. El segundo experimento se realiza con la metodología utilizada en el experimento 1, cuyo análisis se basa en la muestra 2 con datos de 170 estudiantes del tercer periodo escolar mayo – agosto 2017 del PE de TIC de nivel TSU área de Sistemas y de Redes y Telecomunicaciones, donde se incluyen treinta nueve (39) variables de estudio, que representa los patrones de entrada o escalares p_i ($p_1, p_2, p_3, \dots, p_{39}$) de las neuronas de capa oculta, presentadas en la Tabla 26, utilizadas en el entrenamiento a través del algoritmo de RNA de FBR, para la predicción de estudiantes con riesgo de baja, que son comprobadas con las estadísticas reportadas por la IES, con el propósito de verificar la precisión y el nivel de confiabilidad del algoritmo respecto a los experimentos realizados.

Tabla 26. Variables de entrada del entrenamiento con RNA de FBR experimento 2

Variable de Entrada Modelado	Descripción de la Variable
Aspectos Personales	
p1	• Edad
p2	• Género
p3	• Estado civil
p4	• Número de Hijos
p5	• Planes de matrimonio a corto plazo o en su estancia en la IES
p6	• Relación familiar
p7	• Planes de su futuro profesional
p8	• Ocupación de su tiempo libre
Aspectos Vocacionales	
p9	• Título Técnico o Bachillerato General
p10	• La IES es su primera opción
p11	• El PE es su primera opción
p12	• Tiene planeado presentar examen en otra IES
p13	• Has interrumpido estudios
p14	• Ha cursado otras carreras en otras IES diferentes al PE que estudia

Variable de Entrada Modelado	Descripción de la Variable
Aspectos Académicos	
p15	• Promedio de Nivel Medio Superior
p16	• Nuevo Ingreso al período escolar /Reingreso
p17	• Materias con dificultad en nivel medio superior o en la IES
p18	• Número de Asignaturas extraordinarias en períodos escolares anteriores
p19	• Asignaturas extraordinarias en períodos escolares anteriores
p20	• Técnicas de estudio
p21	• Tiene libros de tu perfil de carrera como apoyo para su estudio en casa
p22	• Tiene computadora en casa como apoyo a sus estudios
p23	• Tiene acceso a internet en tu domicilio
Aspectos Económicos	
p24	• Ingreso mensual familiar
p25	• Alumno trabaja Si/NO, Número de horas semanales
p26	• ¿De quién depende económicamente el estudiante?
p27	• ¿Quién depende económicamente de usted?
p28	• Gastos de transportación semanales a la IES
p29	• Tiempo de transportación (hrs) Casa – IES
p30	• Tiene Beca (Si/No)
p31	• Tipo de Beca
Aspectos de Salud	
p32	• Estado de salud
p33	• Problemas de salud (obesidad, delgadez extrema, manchas en la piel, falta de energía, dentadura, visual, auditivo, discapacidad o ninguna)
p34	• Padece alguna enfermedad, tratamiento o alergia (Si/No, especifica)
p35	• Frecuencia que presenta enfermedades menores como gripe, infecciones estomacales, dolores de cabeza. (especifica enfermedad y frecuencia)
p36	• Ha recibido atención psicológica(Si, No, cuanto tiempo)
p37	• Adicciones de fumar
p38	• Adicciones de ingerir bebidas alcohólicas
P39	• Adicciones a drogas

Fuente: Elaboración propia.

Señalando que el experimento 2 se desarrolla de manera similar al experimento 1, en Matlab para fines de pruebas en tres pasos: diseño de la RNA, programación y ejecución del algoritmo, así como la evaluación. Posteriormente se comprobará la funcionalidad del algoritmo previamente entrenado en la CVA en la sección de análisis y discusión.

- Diseño de la RNA de FBR. En la capa de entrada, se incluye las variables de entrada identificadas como p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, y p39 representando los datos categorizados de 170 estudiantes de la muestra 2 descritos en la Tabla 26, tomados como el conjunto de patrones de entrenamiento en el algoritmo de RNA de RBF que se visualiza en la Tabla 27, parte del conjunto P, que fueron extraídos de los archivos de tipo texto generados del Sistema de Información Integral de la IES especificados en la sección 4.2.2 Integración y Recopilación de Datos en la CVAATI., además de ser previamente pre procesados, así también se incluye la variable de Salida $T1 = \{-1, 1\}$, donde **1** corresponde al alumno que está en riesgo de Baja y **-1** el que tiene probabilidad de continuar en la Universidad.

Tabla 27. Conjunto parcial de Patrones de entrenamiento P experimento 2 muestra 2

Item	p1	p2	p3	p4	p5	p6	p7	p8	...	p36	p37	p38	p39
1	2	2	1	1	2	2	1	1	...	6	2	4	9
2	2	2	1	1	1	1	1	2	...	1	1	1	1
3	2	2	1	1	1	1	1	1	...	6	9	9	9
4	2	2	1	1	1	1	1	3	...	6	9	9	9
5	2	1	1	1	1	1	1	2	...	1	1	1	1
6	2	1	1	1	1	1	3	3	...	6	9	2	9
7	2	2	1	1	1	1	1	2	...	1	1	2	1
8	2	2	1	1	1	1	1	2	...	6	9	9	9
9	2	2	1	1	1	2	1	3	...	1	1	2	1
10	2	2	1	1	1	2	1	3	...	3	4	4	2
11	2	2	1	1	1	1	1	3	...	1	6	6	1
12	2	2	1	1	1	1	5	6	...	6	9	4	9
13	2	2	1	1	1	1	1	6	...	1	1	1	1
14	2	2	1	1	1	1	1	2	...	6	9	9	9
15	2	2	1	1	1	2	1	3	...	1	8	1	1
16	4	1	1	1	1	2	1	6	...	6	9	4	9
17	2	1	1	1	1	1	5	3	...	1	1	1	1

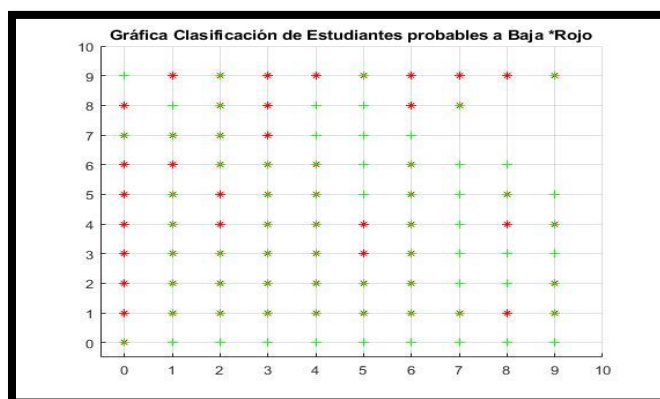
Item	p1	p2	p3	p4	p5	p6	p7	p8	...	p36	p37	p38	p39
18	2	1	1	1	1	1	2	1	...	6	9	9	9
19	2	2	1	1	1	1	5	2	...	4	1	4	1
20	2	2	1	1	3	5	7	7	...	6	9	9	9
21	2	1	1	1	1	1	1	3	...	1	1	1	1
22	2	2	1	1	1	1	3	2	...	1	1	1	1
23	2	2	1	1	1	1	1	1	...	1	1	2	1
24	2	2	1	1	2	1	5	1	...	1	2	2	1
25	2	2	1	1	1	1	1	2	...	1	1	1	1
26	2	2	1	1	1	1	2	2	...	1	1	1	1
27	2	2	1	1	1	2	1	1	...	6	9	2	9
28	2	1	1	1	1	1	1	2	...	1	1	1	1
29	2	1	1	1	1	1	2	6	...	4	1	1	1
30	2	2	1	1	1	1	1	2	...	1	1	1	1
.....													
162	2	1	1	1	1	1	1	1	...	1	1	1	1
163	2	2	1	1	1	1	1	2	...	1	1	1	1
164	3	2	2	2	1	1	1	6	...	3	6	2	1
165	2	2	1	1	1	1	1	3	...	1	4	1	1
166	2	2	1	1	1	1	3	2	...	1	1	1	1
167	2	2	1	1	1	1	5	2	...	1	1	1	1
168	2	2	1	1	1	1	1	6	...	1	1	1	1
169	2	2	1	1	1	1	3	1	...	1	1	1	1
170	2	2	1	1	1	1	1	3	...	1	6	1	1

Fuente: Elaboración propia.

- Programación y ejecución del algoritmo de RNA de FBR se realiza en Matlab de manera muy similar tan solo cambiando parámetros y archivo de estudio para el experimento 2, descrito en Tabla 24 , como una primera aproximación del modelo, a través del método 1 descrito en la sección de diseño del modelo, realizando primero la adaptación no supervisado utilizando k-medias, calculando las funciones gaussianas y se continúa con el aprendizaje supervisado de los pesos sináptico W_{kj} con Adaline para el cálculo del error y aproximación del algoritmo aplicado al conjunto P, que a su vez llama la función de activación de hardlims, quien determina el de valor Salida T1 como se describió en la Figura 34. Adicionalmente el código programado incluye simular la RNA de FBR con tan solo cambiar

el valor de k en el conjunto P que representa los patrones de entrenamiento de las diferentes muestras tomadas para este trabajo de los archivos de texto, siendo posible obtener en este segundo experimento la ejecución de varias pruebas que permiten la comparación de resultados del entrenamiento aplicado, siendo la base del entrenamiento y comparación de la ejecución en el SW propio desarrollado del Modelo PredATIS en la CVAATI, que se presentará en capítulo de discusión y resultados.

- Se analizan los resultados obtenidos del experimento 2, de la ejecución de varias pruebas para el entrenamiento con RNA de FBR, con base al conjunto de patrones P descrito en el paso 1 y definidos en la Tabla 26, que representan los aspectos personales, vocacionales, académicos, económicos y de salud, siendo factores detectados en las estadísticas reportadas de la IES, que impactan en la predicción que se realiza con la asignación de forma aleatoria de valores -1 o 1 en la capa de salida $T1$ de la Red Neuronal de que: **SI** el estudiante es candidato a baja del PE (1) que se visualiza en la Gráfica 5 con un símbolo de $*$ color rojo y con un $+$ color verde (-1) cuando **NO** está en riesgo, es decir que continuará sus estudios en el siguiente período escolar.



Gráfica 5. Clasificación de estudiantes con predicción a Baja de la IES muestra 2.

Fuente: Elaboración propia.

A continuación describen los resultados obtenidos del experimento 2 en la Tabla 28, donde se evalúa el error de predicción en cada ejecución del algoritmo de aprendizaje de RNA de FBR en Matlab.

Tabla 28. Resultados de ejecución de pruebas de experimento 2

Error calculado (e) en Matlab con base al número de centros k=4	Error calculado (e) en Matlab con base al número de centros k=6	Error calculado (e) en Matlab con base al número de centros k=9	Error calculado (e) en Matlab con base al número de centros k=13	Número de Estudiantes de Baja con predicción en Algoritmo FBR	Predicción Algoritmo del Número de Estudiantes de Baja con estadísticas reportadas reales	% de predicción obtenido en Algoritmo FBR con estadísticas reportadas	Diferencia de % error obtenido en algoritmo con respecto al % real	% error de predicción de bajas con Algoritmo FBR y que no fueron reportadas en las estadísticas reales
0.27	0.27	0.27	0.25	27	15	46.875%	53.125%	8.70%
0.27	0.27	0.27	0.25	32	18	56.250%	43.750%	10.14%
0.30	0.29	0.29	0.29	30	20	62.500%	37.500%	7.25%
0.22	0.22	0.22	0.21	29	22	68.750%	31.250%	5.07%
0.25	0.25	0.25	0.23	32	24	75.000%	25.000%	5.80%

Fuente: Elaboración propia.

En la Tabla 28, solo se describen los resultados de 5 del total de 20 ejecuciones realizadas, a partir del conjunto P, que representan los patrones de entrenamiento tomados de los archivos de texto Dat3h.txt, Dat3i.txt, Dat3j.txt, Dat3k.txt y Dat3l.txt que fueron extraídos del Sistema Integral de la IES, donde se asignan los valores de la variable de salida T1, para calcular el nivel de error y el porcentaje de predicción en Matlab, observando que el nivel de precisión del algoritmo para la muestra 2 que incluye las 39 variables descritas en la Tabla 26, se obtiene un error de 0.30 a .021 y un porcentaje de predicción de error de 53.125% a 25% respecto a las estadísticas reportadas de 32 estudiantes en este período escolar, donde se evidencia el progreso

de resultados va hacia la tendencia de llegar al error de 0. Adicionalmente se realizaron 25 ejecuciones a prueba y error del experimento, con la finalidad de predecir un mejor el rendimiento del Algoritmo con FBR y que se comparará con la ejecución de entrenamiento en el SW propiamente desarrollado en la CVAATI como trabajo futuro y que esta fuera del alcance de este documento.

CAPÍTULO 6. ANÁLISIS Y DISCUSIÓN

6.1 Análisis de Resultados del modelo PredATIS

En esta sección se presenta el análisis y discusión de resultados obtenidos de las propuestas de diseño del modelo de esta investigación. Primero se desarrolló con técnicas de clasificación y selección de mejores atributos DM y la segunda con algoritmos de RNA, descritos en la sección 5.1 Modelado de PredATIS con algoritmos de DM y 5.2.1 Modelado de PredATIS basado en RNA de FBR. Señalando la importancia de validar los algoritmos DM aplicados, que consiste en identificar el rendimiento en términos de calidad y características del modelo PredATIS desarrollado con datos reales de una IES seleccionada del estado de México. El desempeño de los algoritmos de clasificación y selección DM se mide con la exactitud (precisión) o tasa de error (clasificación correcta e incorrecta, error absoluto relativo) que determinan el grado de confiabilidad y comparar con otros algoritmos y conjunto de pruebas. En tanto que el desempeño de los algoritmos de RNA se valida con base al error cuadrático calculado en diferentes conjuntos de datos del entrenamiento con una tendencia de 0 para mayor precisión y confiabilidad del modelo.

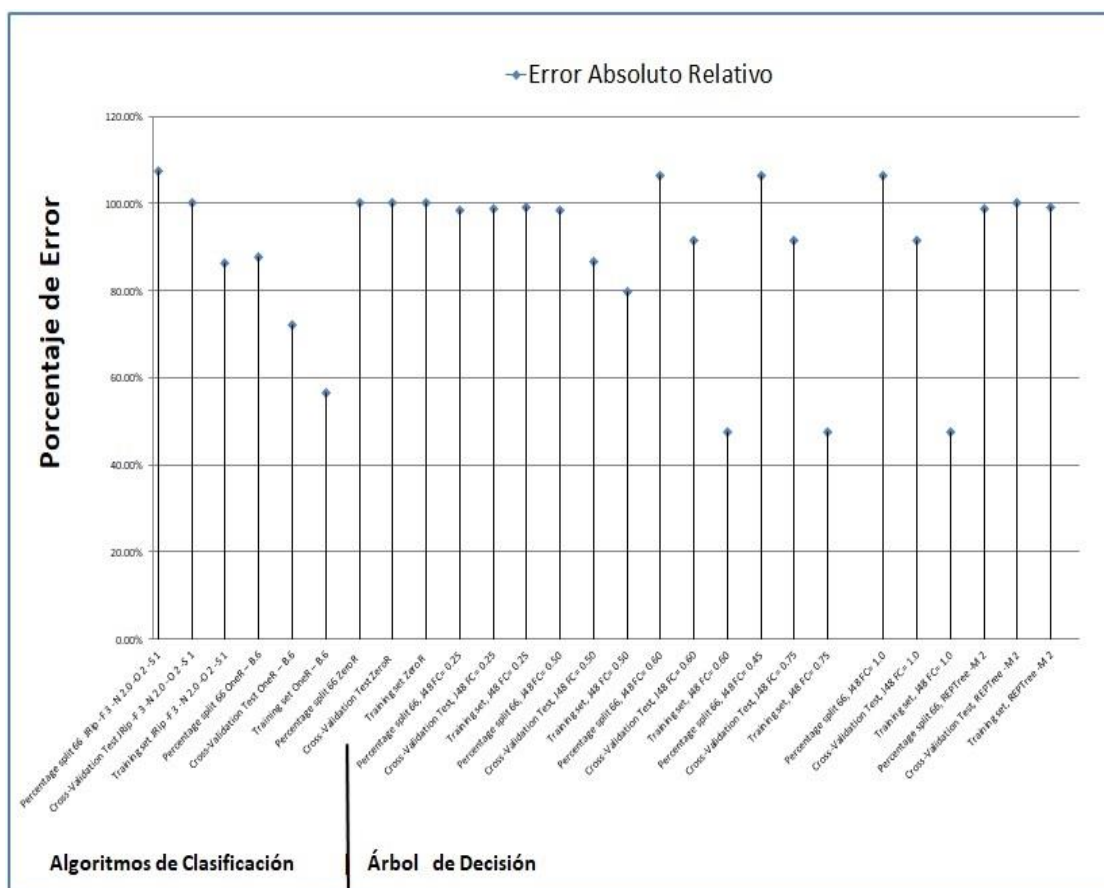
6.1.1 Validación del Modelo PredATIS basado en algoritmos de Clasificación DM.

El modelo se desarrolló con técnicas de selección de mejores atributos, reglas de clasificación DM y árboles de decisión, con el propósito de identificar las variables o factores de mayor impacto que causan la deserción o reprobación del estudiante. En la validación de los algoritmos de selección de atributos con CfsSubsetEval y BestFirst en el experimento 1, se aplican dos modos:

- Validación cruzada (Cross-Validation- Folds 10 y seed 1), en donde se predice que las 15 causas de mayor impacto en la deserción de los estudiantes descritas en la Tabla 19, y se verifica que 10 de ellas se relacionan en el reporte real emitido por la IES que son: Planes de Matrimonio a Corto Plazo, El PE es su Primera Opción, Tiene Planeado realizar Examen en Otra IES, Materias con Dificultad en Nivel Medio Superior o en períodos anteriores de la IES en la que cursa actualmente, su Estado de Salud, Tiene Beca, Asignatura en Extraordinarias en períodos escolares Anteriores, si ha Cursado en Otras IES, cuales son su Planes a Futuro en el ámbito Profesional y el Número de Hijos. Validando de esta manera la precisión del algoritmo en un 66.67 %. En tanto que las causas no reportadas son: el Promedio de Nivel Medio Superior, Interrupción de Estudios, Ocupación de tiempo libre, Género y Edad y representan solo un 33.33%. Evidenciando que se tiene buen resultado en la predicción con respecto al reporte real emitido por la IES en estudio.
- Validación con el conjunto total de datos (Use full training set) solo se obtienen dos causas: Planes de Matrimonio a Corto Plazo y El PE es su Primera Opción, esto representa tan solo un 13.33 % de precisión, lo cual la precisión del algoritmo no es factible y se toma la primera opción para esta investigación.

En lo que respecta a la evaluación de reglas de clasificación y árbol de decisión desarrolladas con los algoritmos JRIP, OneR, ZeroR, J48 y se aplicaron las pruebas (test) y el entrenamiento (training) en tres tipos: Percentage split 66, Cross-Validation Test y Training set. En este sentido se validan los resultados, observando de la Tabla 21, que la mejor precisión de clasificación se obtiene con el algoritmo J48 con un factor de confianza de 0.60, 0.75 y 1.0 con un porcentaje del error del 47.5863%, aplicado al conjunto total de datos de entrenamiento (Training set) categóricos

nominales y discretizados con filtro no Supervisado de Atributos, comparado con los test de Percentage split 66 y Cross-Validation Test, así como con los algoritmos JRIP, OneR y ZeroR de reglas de clasificación como se visualiza en la Gráfica 6.



Gráfica 6. Resultados de Algoritmos de Clasificación DM.

Fuente: Elaboración Propia

Adicionalmente se representa en la Tabla 29, el esquema del árbol de decisión del entrenamiento con J48 en el conjunto total de datos de Training set.

Tabla 29. J48 pruned tree

PlanesMatrimonioCtoPlazo = Si
MateriasDificultadMSsup/IES = Áreas Informáticas: Si (4.0)
MateriasDificultadMSsup/IES = Matemáticas y lógica: Si (0.0)
MateriasDificultadMSsup/IES = otras no afines al perfil: Si (0.0)
MateriasDificultadMSsup/IES = Ninguna: No (3.0)
MateriasDificultadMSsup/IES = Inglés y Asignaturas de Informática (Programación/Redes/SO/BD): Si (0.0)
MateriasDificultadMSsup/IES = NRP: Si (0.0)
MateriasDificultadMSsup/IES = Matemáticas e Informática(BD/Programación/Redes/SO): Si (0.0)
MateriasDificultadMSsup/IES = Inglés: Si (0.0)
MateriasDificultadMSsup/IES = NRE: Si (0.0)
PlanesMatrimonioCtoPlazo = No
El PEPrimOpc = Si: No (30.0)
El PEPrimOpc = No sin especificar motivo
MateriasDificultadMSsup/IES = Áreas Informáticas
Género = Masculino: No (4.0)
Género = Femenino
AssignExtraordinariasAnt = Asignaturas Técnicas del perfil (Programación, Base de Datos, Redes de cómputo, Soporte Técnico, Sistemas Operativos, etc.): No (3.0/1.0)
AssignExtraordinariasAnt = Ninguna: Si (1.0)
AssignExtraordinariasAnt = NRP: Si (0.0)
AssignExtraordinariasAnt = Inglés: Si (0.0)
AssignExtraordinariasAnt = Asignaturas no técnicas del perfil (formación sociocultural, expresión oral y escrita, etc.): Si (0.0)
AssignExtraordinariasAnt = Inglés - Técnicas Perfil: Si (2.0)
AssignExtraordinariasAnt = Ingles/Matemáticas: Si (0.0)
AssignExtraordinariasAnt = Desarrollo de Habilidades lógico-matemáticas-Física: Si (0.0)
AssignExtraordinariasAnt = NRE: Si (0.0)
Género = NRE: No (0.0)
MateriasDificultadMSsup/IES = Matemáticas y lógica: No (8.0/1.0)
MateriasDificultadMSsup/IES = otras no afines al perfil: No (0.0)
MateriasDificultadMSsup/IES = Ninguna
Género = Masculino
Ocupacióntiempolibre = Estudiar o visitar bibliotecas: No (6.0/1.0)
Ocupacióntiempolibre = actividad cultural o Deportiva: Si (4.0/1.0)
Ocupacióntiempolibre = Varios entretenimiento (visitar plazas comerciales, cines, parques, Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento,etc): Si (2.0)
Ocupacióntiempolibre = otras actividades(domesticas/familiares/religiosas/trabajo): No (3.0)
Ocupacióntiempolibre = NRP: No (0.0)
Ocupacióntiempolibre = Platicar con amigos o Hablar por teléfono: No (0.0)
Ocupacióntiempolibre = NRE: No (0.0)
Ocupacióntiempolibre = Chat o Redes Sociales/Navegar Internet: No (0.0)
Género = Femenino: No (5.0)
Género = NRE: No (0.0)
MateriasDificultadMSsup/IES = Inglés y Asignaturas de Informática (Programación/Redes/SO/BD): Si (1.0)
MateriasDificultadMSsup/IES = NRP: No (0.0)
MateriasDificultadMSsup/IES = Matemáticas e Informática(BD/Programación/Redes/SO): No (4.0)
MateriasDificultadMSsup/IES = Inglés: No (5.0)
MateriasDificultadMSsup/IES = NRE: No (0.0)
El PEPrimOpc = No aprobé en otras IES o no me inscribe al Perfil deseado: No (57.0/5.0)
El PEPrimOpc = Obligado por padres/Cercanía: No (2.0)
El PEPrimOpc = Prefiero no responder: No (10.0)
El PEPrimOpc = NRE: No (0.0)
PlanesMatrimonioCtoPlazo = Prefiero no contestar: No (3.0)
PlanesMatrimonioCtoPlazo = NRE: No (13.0/6.0)
Number of Leaves : 44
Size of the tree : 52

Fuente: Elaboración Propia

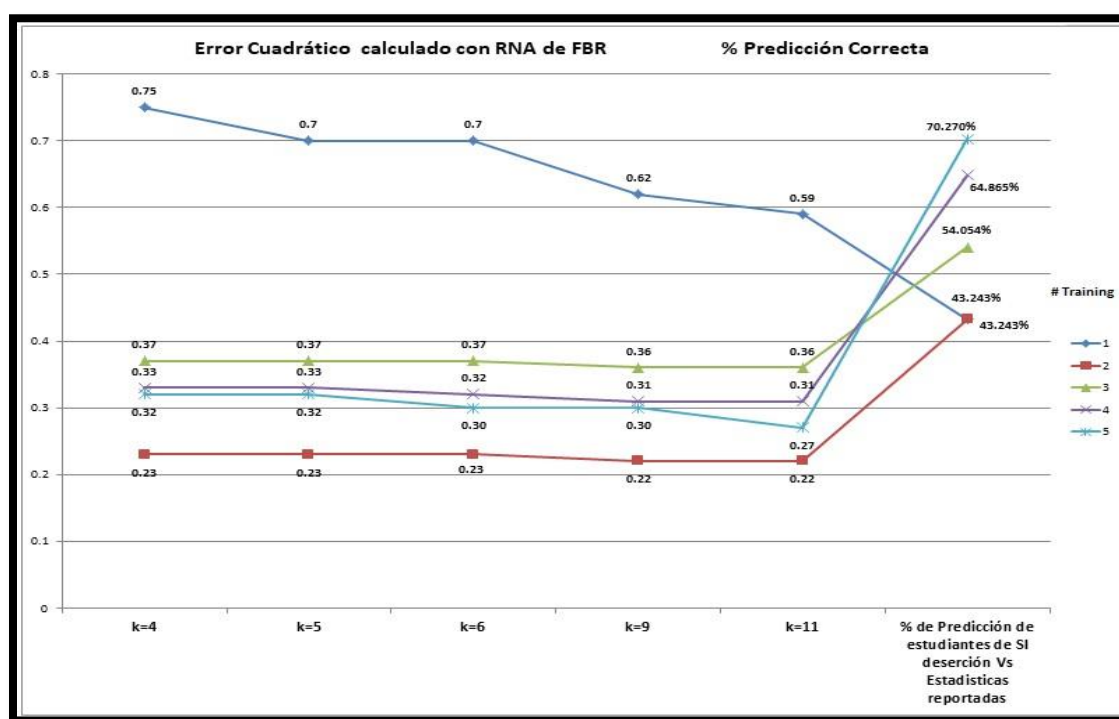
Como se observa en la estructura del árbol de decisión bajo el esquema SI – entonces sucede evento. Un estudiante está en riesgo de baja SI Tiene planes de Matrimonio, las asignaturas de Informática son las que más se le dificulta, ha presentado exámenes extraordinarios de las Asignaturas Técnicas del perfil como Programación, Base de Datos, Redes de cómputo, Soporte Técnico, Sistemas Operativos e Inglés, Su tiempo libre lo ocupa en Estudiar o visitar bibliotecas, actividad cultural o Deportiva, Varios entretenimiento (visitar plazas comerciales, cines, parques, Juegos, música, videojuegos, TV, teatro, leer revistas de entretenimiento, etc. y en otras actividades(domesticas/familiares/religiosas/trabajo, etc.) y El PE No es su primera opción, debido a que No aprobó en otras IES, NO se inscribe al Perfil deseado y es Obligado por padres o por Cercanía a la Universidad, en tanto que el Género del estudiante no es un factor representativo en la deserción o reprobación. Derivado de estos factores se puede evidenciar que las variables identificadas, SI se relacionan como causas que se reporta en la IES de análisis y por experiencia propia del autor de esta investigación.

6.1.2 Validación del Modelo PredATIS basado en algoritmos de RNA de FBR.

El segundo diseño del modelado se realiza con algoritmos de RNA de FBR donde se validan que los resultados de dos experimentos desarrollados en diferentes entrenamientos, se obtienen como error cuadrático aproximados a Cero, como se describe a continuación:

- Los resultado del Experimento 1 se obtiene un nivel de precisión en el entrenamiento con datos reales de 156 estudiantes de la muestra 1, como se observa en la Gráfica 7, que se inicia el entrenamiento con un error cuadrático calculado de 0.75 al 0.59 y el porcentaje de precisión de aciertos es del 43% es decir que identifican 78 estudiantes de los cuales solo 16 de ellos causaron baja de la IES; de esta forma al continuar con las pruebas de este algoritmo

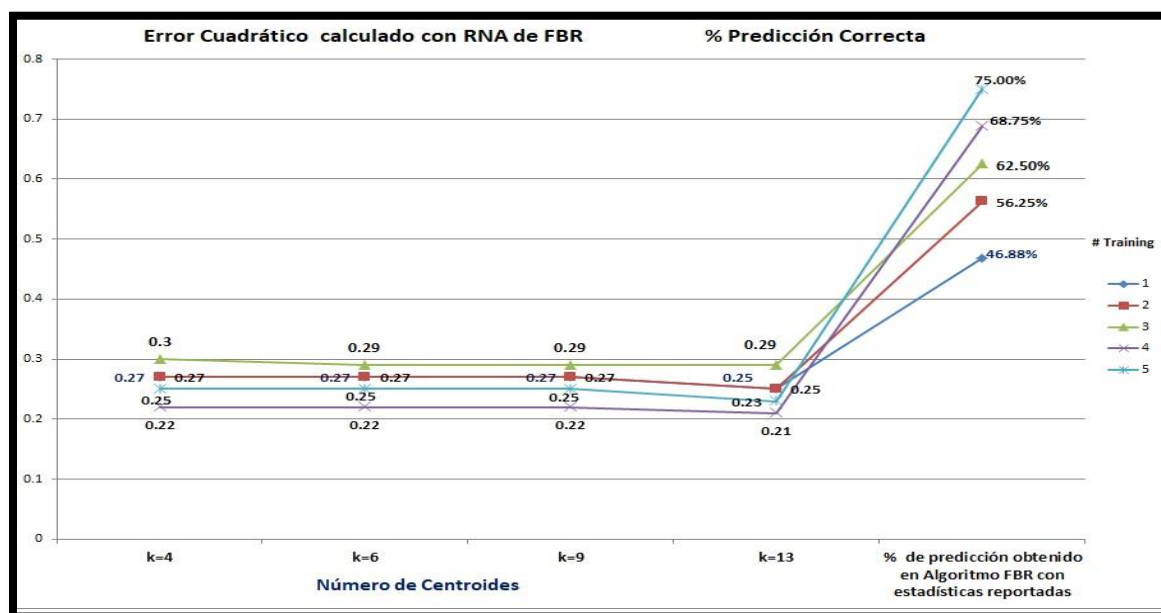
se va disminuyendo el error calculado que va del 0.23 al 0.26, hasta llegar a la última ejecución con 0.32 al 0.27 y se predice a 35 estudiantes en riesgo de baja, de la cual solo 26 de ellos se reportan en las estadísticas de deserción y reprobación de la IES, representando una predicción del 78%. Evidenciando con ello, el desempeño del algoritmo aplicado en esta muestra representativa, sin embargo se puede señalar que debe considerar otros parámetros debido, que cada generación de estudiantes es diferente.



Gráfica 7. Error Cuadrático Calculado con algoritmos de RNA de FBR Experimento1.

Fuente: Elaboración Propia

- Los resultados obtenidos del error cuadrático calculado en el Experimento 2, se presentan en la Gráfica 8.



Gráfica 8. Error Cuadrático Calculado con algoritmos de RNA de FBR Experimento 2.

Fuente: Elaboración Propia.

Observando que en el primer entrenamiento va de 0.27 con 4 centroides agrupados y conforme se van incrementado al valor $k=13$, el error va disminuyendo con valor a 0.25, y se identifica a 27 estudiantes, de los cuales solo 15 de ellos causaron baja de la IES, equivalente a un 46.8% de predicción del algoritmo. Continuando así, el crecimiento del porcentaje de predicción en 56 %, 62% , 68% y 75%, que este último entrenamiento identifica a 24 estudiantes bien clasificados como riesgo de baja y que fueron efectivamente baja de la IES, aun cuando se visualiza en la Gráfica 8 que el cuarto entrenamiento, tienen mejor nivel de precisión en cuanto a error cuadrático que va del 0.22 al 0.21 con una tendencia cercana al Cero, siendo la función objetivo de este algoritmo de RNA de FBR y validando el desempeño en este experimento, sin embargo es importante considerar otros parámetros directamente relacionados al fenómeno de deserción del estudiante debido a la variabilidad de aspectos que se presentan en generaciones escolares.

6.2 Discusión

De los resultados obtenidos en los experimentos desarrollados con técnicas de DM en el que se aplicaron algoritmos de clasificación y selección de mejores atributos y RNA de FBR en dos muestras de datos de la IES seleccionada, que permitieron medir el nivel de precisión en cuanto al valor de los errores de clasificación de instancias correctas e incorrectas, el porcentaje del error absoluto relativo, el error cuadrático con la tendencia aproximada 0, permitieron validar la confiabilidad y precisión del modelo PredATIS diseñado. Evidenciando esto, los resultados en las Gráfica 4, Gráfica 5, Gráfica 6, Gráfica 7 y Gráfica 8 reportada en la sección de análisis de resultados. En este sentido se argumenta, se logra un buen desempeño de los algoritmos aplicados en esta investigación, sin embargo no se puede garantizar el grado de confiabilidad debido a que las generaciones que ingresan a la IES son muy variables, por lo tanto debe incluir y analizar en conjunto, otros parámetros y aspectos directamente relacionados al fenómeno de deserción del estudiante y el nivel académico con el que ingresan.

Por otra parte no se puede generalizar que los algoritmos aplicados en el modelo PredATIS sean los de mejor precisión y confiabilidad, ya que es una tarea complicada en seleccionar y saber cuáles son los más efectivos, en la investigación del fenómeno de deserción o reprobación de estudiantes universitarios y queda como trabajo futuro evaluar y aplicar otras técnicas y métodos de DM.

CONCLUSIONES Y RECOMENDACIONES

El trabajo de tesis que se presenta en este documento centrado en el diseño del Modelo de un modelo predictivo denominado “PredATIS” con el objeto de identificar a estudiantes que están en Riesgo de Baja, así como patrones relacionados con los aspectos o factores que más influyen en la deserción o reprobación estudiantil, tomando como muestra una IES del estado de México. En este sentido se resume las actividades cumplidas en la investigación, desarrollada a través de la estrategia de cebolla y la metodología MSWD&DM propuesta, que combina la Ingeniería de Software (SW) y las fases del proceso de KDD que provee una representación completa del ciclo de vida de un proyecto DM.

- Análisis exploratorio, correlacional y explicativo enfocada a la revisión de trabajos reportados con este fenómeno educativo descritos en la Tabla 6 y Tabla 7, en la que se identifica que las técnicas de clasificación de DM y selección de atributos son las más usadas, en contraste con las RNA de FBR.
- Análisis comparativo de las técnicas y métodos de DM que mejor se adaptaron en esta temática, seleccionando con base a pruebas iniciales aplicadas con datos de una parte de la muestra del estudio que se presenta en este trabajo. Determinando que los algoritmos de clasificación DM y selección de atributos son los que se aplicarían al modelo en identificación de causas de deserción o reprobación. En tanto al algoritmo de RNA se revisa primero Adaline - Widrow & Hoff General que trabaja con separabilidad lineal y el fenómeno de deserción representa separabilidad no lineal, por lo que se resuelve con FBR para la predicción de estudiantes en riesgo de baja.

- Análisis metodologías y modelos de desarrollo de SW, seleccionando el Proceso Unificado Ágil (AUP), que es una versión simplificada de RUP siendo la más utilizada en proyectos similares a la naturaleza y tamaño de CVAATI.
- Diseño y Construcción de la CVAATI para el pre procesamiento de DM base del modelado propuesto.
- Modelado de PredATIS desarrollado con técnicas y métodos de Minería de Datos (DM) basadas en reglas de clasificación, árbol de decisión (JRIP, OneR, ZeroR, J48, REPTree) y selección de atributos (CfsSubsetEval y BestFirst), obteniendo el error de precisión más bajo el algoritmo J48 con factor de confianza del 0.60, .75 y 1.0 y un 66% de predicción de causas que impactan en la deserción y reprobación escolar que reporta la IES, evidenciando una aproximación satisfactoria para abordar este fenómeno.
- Modelado de PredATIS desarrollado con algoritmos de RNA de FBR donde se logra una precisión satisfactoria en los dos experimentos realizados para la predicción de estudiantes en riesgo de baja de la IES en estudio que se presenta en este trabajo.

Se concluye con base a los resultados mostrados, que se cumple con el propósito, hipótesis y objetivos planteados en la investigación de tesis que se presenta en este trabajo.

Finalmente como recomendaciones se sugiere aplicar el modelo propuesto con otras muestras de mayor volumen de datos e incluir otros parámetros de aspectos académicos como notas diagnósticas de ingreso a la IES, notas parciales, asistencia de asesorías y tutorías, para evaluar el algoritmo al menos en dos etapas y no solo aplicar el entrenamiento con notas finales, donde ya se detecta muy tarde el posible riesgo de Baja del estudiante y que muchas veces ya no se pueden

aplicar estrategias y remediar esta situación. Así, también de aspectos psicológicos de los estudiantes entre los que se mencionan conflictos personales y familiares; esto debido que también se ha detectado de manera empírica, que influye radicalmente en la deserción o reprobación escolar.

TRABAJO FUTURO Y DIVULGACIONES

Se puede señalar con los resultados obtenidos, el buen desempeño y precisión de los algoritmos de clasificación y selección de atributos, árbol de decisión y RNA de FBR en el escenario de la IES seleccionada de esta investigación, sin embargo se pretende continuar como trabajo futuro, el implementar el modelo en la CVAATI, con datos reales de todo el PE de la IES seleccionada y hasta donde sea posible en otras Instituciones de Nivel Superior interesadas en modelar este fenómeno de deserción y reprobación escolar con PredATIS, con el propósito de mejorar y realizar el refinamiento necesario en diferentes escenarios y difundir esta investigación.

Finalmente, se presenta la divulgación de los resultados obtenidos de la investigación de Tesis Doctoral.

1. En la primera etapa de investigación, se generó una ponencia y publicación de un artículo en revista indexada en el Journal ©ECORFAN internacional.
 - MAYA-PÉREZ, P. Norma. Propuesta del Diseño de un Modelo Predictivo de alerta temprana en indicadores educativos de Nivel Superior aplicando Minería de Datos. CONGRESO INTERDISCIPLINARIO DE CUERPOS ACADÉMICOS (CICA). Guanajuato, Gto. 17-18 Nov. 2016.

- MAYA-PÉREZ, P. Norma, AGUILAR-CISNEROS, Jorge R., ZAMORA-RÍOS, Rosa A. y BARRON-ADAME, J. Miguel. (2016). Propuesta del Diseño de un Modelo Predictivo de alerta temprana en indicadores educativos de Nivel Superior aplicando Minería de Datos. Revista de Aplicación Científica y Técnica ©ECORFAN, ISSN-2444-4928, 2-4: 29-40.
2. Se participó en la modalidad de poster con el tema “Diseño de un Modelo predictivo aplicando Minería de Datos para identificar causas de Deserción Estudiantil Universitaria” en el 11° Coloquio Interdisciplinario de Posgrado en UPAEP y se pretende publicar en la revista de investigación científica Strategy, Technology & Society y que actualmente se encuentra en proceso de revisión.
 3. Se envía el artículo titulado “WEB Platform Development for Data Mining preprocessing at Higher Education Institutions” donde se pretende publicar en 6th IEEE International Conference on Software Engineering Research and Innovation, CONISOFT 2018 y que actualmente se encuentra en proceso de revisión.

REFERENCIAS

- Secretaría de Educación Pública (SEP). (2014). *Sistema Nacional de Información Estadística Educativa (SNIEE)*. Recuperado el 1 de Julio de 2016, de <http://www.snie.sep.gob.mx/>
- Acquatella, P. (30 de Abril de 2009). *Redes Neuronales como aproximadoras de Funciones*. Valle de Sartenejas, Caracas, Venezuela: Universidad Simon Bolivar.
- Álvarez, L. (2009). Comportamiento de la Deserción y Reprobación en el Colegio de Bachilleres del Estado de Baja California: Caso Plantel Ensenada. Veracruz, Veracruz, México: X Congreso Nacional de Investigación Educativa.
- Amaya , Y., Barrientos, E., & Heredia, D. (2015). Student Dropout Predictive Model Using Data Mining Techniques. *En IEEE Latin America Transactions*, 13(9), 3127-3134. Recuperado de: http://www.ewh.ieee.org/reg/9/etrans/ieee/issues/vol13/vol13issue09Sept.2015/13TLA9_45Heredia.pdf.
- ApacheCTLdesarrolloweb. (2014). *DesarrolloWeb-MVC*. Obtenido de <http://www.desarrolloweb.com/articulos/que-es-mvc.html>
- Araque, F., Roldán, C., & Salguero, A. (2009). Factors Influencing University Drop Out Rates. *Computers & Education*, 53(3), 563 - 574. Recuperado de: <http://eric.ed.gov/?id=EJ848783>.
- Atlantia Search. (2014). *Reporte Cualitativo. Investigación sobre las causas de la deserción escolar en México*. Recuperado el 1 de Julio de 2016, de <http://atlantiasearch.com/wp-content/uploads/2014/12/CO-RS-2013-04-Reporte-Deserci%C3%B3n-escolar-en-M%C3%A9xico-1.pdf>
- Basnuevo, M., Avila, G., & Escalona, G. (2007). Gestión del conocimiento a través del Portal Holguín. *Redalyc*.
- Budiño, G. (2011). Gestión de comunidades virtuales: El rol del community manager. *Quantum*, 6(1), 4-11. Recuperado de: <http://www.quantum.edu.uy/2011-11-28%20-%20Vol.6Num.1/pdfs/QuantumCap1.pdf>.
- Calderón, D. (2005). Diseño e implementación de un modelo explicativo de las causas de las horas de sobretiempo que tiene una empresa que maneja personal de servicio. (*Tesis de grado en Estadística Informática, Instituto de Ciencias Matemáticas, Escuela Superior Politécnica del Litoral*). Ecuador.

- Calders, T., & Pechenizkiy, M. (2011). Introduction to the special section on educational data mining. *ACM SIGKDD Explorations Newsletter*.13 (2), Pages 3-6. doi:10.1145/2207243.2207245.
- Can Neural Networks Learn the Black- Scholes Model ?: A simplified approach. (2005). *The Center for Financial studies, Southern New Hampshire University, 2005(01)*.
- Cely, S. (2014). *Modelo para gestionar la creación de una comunidad virtual de formación con las familias de las Instituciones Educativas Oficiales del municipio de Itagüí. Magíster en Ingeniería con especialidad en Tecnologías de la Información para la Educación*. Obtenido de https://repository.eafit.edu.co/bitstream/handle/10784/2845/ShirleyCarolina_CelyNaranjo_2014.pdf?sequence=1&isAllowed=y
- Cendejas, J. (2014). Implementación del modelo integral colaborativo (MDSIC) como fuente de innovación para el desarrollo ágil de software en las empresas de la zona centro - occidente en México. (*Tesis Doctoral. Centro Interdisciplinario de Posgrados, Investigación y Consultoría. UPAEP*). Mexico.
- Coecillo, B., Casas, L., Pérez, O., & Caballero, Y. (2015). Redes neuronales artificiales en la producción de tecnología. *Revista Academia y Virtualidad*, 8(1), 12-20.
- Coello, L., Casas, L., & Pérez, O. (06 de 04 de 2015). Redes neuronales artificiales en la producción de tecnología educativa para la enseñanza de la diagonalización. (U. M. Granada, Ed.) *Academia y Virtualidad*, 8(1), 12-20. doi:10.18359/ravi.442.
- Del Carpio, G. (2005). Las redes neuronales artificiales en las finanzas. *Redalyc, Industrial Data revista de investigación*, 8(2), 1-6.
- Delen, D. (2010). A comparative analysis of machine learning techniques for student retention mangment. *Decision Support Systems*, 49(4), Pages 498-506. doi:10.1016/j.dss.2010.06.003.
- Desarrolloweb. (2015). *Desarrollo web*. Obtenido de APACHE CTL: <http://www.desarrolloweb.com/articulos/que-es-mvc.html>
- EcuRed. (2010). *Enciclopedia EcuRed Conocimiento con todos y para todos*. Obtenido de https://www.ecured.cu/Metodología_de_cascada
- eSandra. (2016). *Senior consultant*. Recuperado el 8 de Julio de 2016, de (<http://www.esandra.com/frameworks-y-grids-para-diseno-web-responsive/>)

- Fernández, J., & García, I. (Febrero de 2008). RICA O Recuperación e Integración de Contenidos Audiovisuales empleando Ontologías. (*Tesis Magister. Facultad de Informática. Universidad de Murcia*), 18-19. España.
- Fischer, E. (2012). Modelo para la Automatización del Proceso de Determinación de Riesgo de Deserción en Estudiantes Universitarios. (*Tesis de Magíster, Departamento de Ciencias de la Computación, Universidad Chile*). Santiago de Chile.
- Fischer, G., & Ostwald, J. (2005). Knowledge Communication in Design Communities. *Computer-Supported Collaborative Learning Series. En Barriers and Biases in Computer-Mediated Knowledge Communication*, 5, 213-242. doi:10.1007/0-387-24319-4_10.
- Flores, E. (2011). *Ingenieria de Software Mex, TL*. Obtenido de EcuRed: <http://ingenieriadesoftware.mex.tl>
- Fonseca, J. (2008). Tendencias sobre las comunidades virtuales da perspectiva dos prosumers. *RAE electrónica*, 7(2), 1-27. Recuperado de: http://rae.fgv.br/sites/rae.fgv.br/files/artigos/10.1590_S1676-56482008000200008.pdf.
- Formia, S. (2013). La deserción en cursos universitarios. Construcción de modelos sobre datos de la UNRN usando técnicas de Extracción de Conocimiento. (*Tesis Magsiter, Facultad de Informática, UNLP*). Argentina.
- Freud, D. (2008). *Conceptos de Administración Estratégica* (Décimo primera ed.). México: Pearson Educación.
- Galán, V. (2015). Aplicación de la Metodología CRISP-DM a un proyecto de Minería de Datos en el entorno Universitario. (*Tesis de grado.Escuela Politecnica Superior, Ingeniería en Informática, Universidad Carlos III de Madrid*), Recuperado de: http://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf?sequence=1. Madrid.
- Génova, G., Fuentes, J., & Valiente, C. (2006). Evaluación comparativa deherramientas CASE para UMLdesde el punto de vista notacional. *Novática. Sección técnica Tecnología de Objetos*, 59-64.
- Grasso, V. (2012). Early Warning Systems: State of Art Analysis and Future Directions. *United Nations Environment Programme (UNEP)*.
- Heppen, J., & Bowles, S. (2008). *Developing Early Warning Systems to identify potential high school dropouts*. Obtenido de <http://eric.ed.gov/?id=ED521558>

- Hernández, J., Ramírez, M., & Ferri, C. (2004). *Introducción a la Minería de Datos*. Pearson.
- Hernández, R., Fernández, C., & Baptista, M. (2010). *Metodología de la investigación* (5 ed.). México: Mc Graw-Hil.
- Herrero, J., & Molina, J. (2012). Técnicas de análisis de datos. *Aplicaciones Prácticas utilizando Microsoft Excel y WEKA*. Madrid, España: Universidad Carlos III.
- Iriberri, A., & Leroy, G. (2009). A life-cycle perspective on online community success. *ACM Computing Surveys (CSUR)*, 41(2), 11,1-29. doi: 10.1145/1459352.1459356.
- Isasi, P., & Galván, I. (2004). *Redes de Neuronas Artificiales. Un enfoque Práctico*. (P. E. C.V., Ed.) Madrid: Pearson Prentice Hall.
- Jiménez, M., & Álvarez S., A. (2010). Análisis de Datos en WEKA – Pruebas de Selectividad. *Departamento de Ingeniería en Telemática. Universidad Carlos III*. Madrid, España.
- Kaski, S. (1996). Creating an order in digital libraries with self-organizing map. *In:Proceeding of World Congress on Neural Networks, WCNN'96* (págs. 814-817). Mahwah, NJ: INNS Press.
- Kohonen, T. (1997). *Self-Organizing-Maps*. Berlín: Springer.
- Lanzarini, L., Charnelli, M. E., Baldino, G., & Díaz, F. J. (2015). Selección de atributos representativos del avance académico de los alumnos universitarios usando técnicas de visualización. *En TE & ET: Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*(15), 42-50. Recuperado de: <https://dialnet.unirioja.es/servlet/articulo?codigo=5386604>.
- Lau F, R., Álvarez, J., Pérez, S., & Leyva, E. (2016). Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales. *Ingeniare. Revista Chilena de Ingeniería*, 24(4), 715-727.
- Laurene, F. (1994). *Fundamentals of neural networks: architectures, algorithms, and applications*. (©. Inc. Upper Saddle River, Ed.) NJ,, USA : Prentice-Hall.
- Leimeister, E., Ebner, W., & Krcmar, H. (2005). *Diseño, Implementación y evaluación de componentes de Trust-Apoyar a comunidades virtuales para pacientes*. Diario de sistemas de información de gestión.
- Longoni, M., Porcel, E., López, M., & Dapozo, N. (2010). Modelos de Redes Neuronales Perceptrón Multicapa y de Base Radial para la predicción del rendimiento académico de

- alumnos universitarios. En *XVI Congreso Argentino de Ciencias de la Computación*. Argentina.
- López, J., Martínez, I., & Moreno, P. (2003). *Desarrollo de un sistema e-learning basado en estándares IMS*. Recuperado el 27 de Mayo de 2016, de <http://eprints.sim.ucm.es/9104/1/TC2004-6.pdf>
- M. Awad. (2005). Modelado de Sistemas Complejos mediante estructuras jerárquicas de Redes de Funciones de Base Radial. *Tesis Doctoral*. Universidad de Granada. España.
- Magaña, M. (2002). Causas del Fracaso Escolar. *XIII Congreso de la Sociedad Española de Medicina del Adolescente*, Recuperado de: <http://www.spapex.es/adolescencia02/ponencias.htm>.
- Márquez, C. (2015). Predicción del Fracaso y el Abandono escolar mediante Técnicas de Minería de Datos. (*Tesis Doctoral*, Universidad de Córdoba). España.
- Márquez, C., Romero, C., & Ventura, S. (2012). Predicción del fracaso escolar mediante técnicas de minería de datos. *IEEE-RITA: Revista Iberoamericana de Tecnologías del Aprendizaje*, 7(3), 109-117.
- Márquez, C., Romero, C., & Ventura, S. (2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *IEEE Journal of Latin-American Learning Technologies*, 8(1), 7-14.
- Martínez, M. (Ene-Abr de 2014). Metodología basada en redes neurales para interpretación de la resistividad del suelo en zonas urbanas. Ingeniería Energética. (Redalyc, Ed.) *Ingeniería Energética*, XXXV(1), 59-69.
- Más-Estellés, J., Alcover, R., Dapena, A., Valderruten, A., Satorre, F., Llopis, F., & otros. (2009). Rendimiento Académico de los Estudios de Informática en Algunos Centros Españoles. *XV Jornadas de Enseñanza Universitaria de la Informática*, 5-12 Obtenido de: <http://upcommons.upc.edu/bitstream/handle/2099/7904/p156.pdf>.
- MasterWeb.La. (2015). *MasterWeb.La-arquitectura MVC*. Obtenido de <http://www.masterweb.la/blog/desarrollo-web/arquitectura-mvc-que-es-como-funciona-ventajas>
- Matthews, J. (2008). The Globalization Reader. (J. Boli, & F. Lechner, Edits.) *Power shift*, 287-293. Blackwell Publishing.

- Meirinhos, M., & Osorio, A. (2009). Las comunidades virtuales de aprendizaje: el papel central de la colaboración. *Pixel-Bit. Revista de Medios y Educación*, 35, 45-60. Recuperado de: <http://www.sav.us.es/pixelbit/pixelbit/articulos/n35/4.pdf>.
- Méndez, V. (2009). Estudio de Aplicación de Redes Neuronales en la Evaluación de Riesgo Crediticio. *Informe Final Tesis de Grado. Facultad de Ingeniería. Pontificia Universidad Católica de Valparaíso*. Chile.
- Millington, R. (2008). *Buzzing Communities: How to Build Bigger, Better, and More Active*. s. L., FeeverBee.
- Molina, J., & García, J. (2012). Técnicas de Análisis de Datos, Aplicaciones prácticas utilizando Microsoft Excel y Weka. *Universidad Carlos III de Madrid*. Madrid.
- Mora, J. (Septiembre de 2011). Uso de redes neuronales artificiales para estimar la respuesta sísmica de sistemas estructurales. *Tesis Universidad Nacional Autónoma de México, Programa Maestría y Doctorado de Ingeniería*. México.
- OECD. (2005). *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*. France: OECD - European Communities. The measurement of scientific and technological activities.
- Palmer, A., Montaña, J., & Jiménez, R. (2002). Tutorial sobre Redes Neuronales Artificiales: Los Mapas Autoorganizados de Kohonen. *ELECTRÓNICA DE PSICOLOGÍA*, 6(1).
- Palomo, O. (15 de 01 de 2014). *Inteligencia en Redes de Comunicaciones. Minería de Datos*. (D. d. Telemática, Ed.) Recuperado el 01 de Julio de 2017, de Universidad Carlos III de Madrid: <http://www.it.uc3m.es/jvillena/irc/>
- Pavón, J. (2008). *Estructura de las Aplicaciones Orientadas a Objetos.El patrón Modelo-Vista-Controlador (MVC)*. Recuperado el 27 de Mayo de 2016, de <https://www.fdi.ucm.es/profesor/jpavon/poo/2.14.MVC.pdf>
- Plant, R. (2004). Online Communities. *Technology in Society*, 26, 51-65. doi:10.1016/j.techsoc.2003.10.005.
- Plata, R. W. (2008). MatLab & Redes Neuronales. (Dialnet, Ed.) *Información, Tecnología y Sociedad RITS online*(1), 96-102.
- Porter, M., & Kramer, R. (2011). *Shared value: How to reinvent capitalism and unleash a wave of innovation and growth*. (Harvard Business, Ed.) Obtenido de [http://www.bmw-stiftung.de/de/asset/index/mid/16/lang/x/file/o document de 387.pdf](http://www.bmw-stiftung.de/de/asset/index/mid/16/lang/x/file/o%20document%20de%20387.pdf)

- Porter, M., Anderson, C., & Torres, J. (2011). *Competitividad Desarrollo*. s.e.
- Pressman, R. (2005). *Ingeniería de software. Un enfoque práctico*. España: McGraw.Hill.
- Pressman, R. (2006). *Ingeniería de software. Un enfoque práctico*. España: McGraw.Hill.
- Quirón, A. (4 de Oct de 2005). *Epidata: Introducción a UML 2.0*. Obtenido de http://www.epidataconsulting.com/tikiwiki/tiki-read_article.php?articleId=15
- Rey , F. F. (2014). Alternativas y Determinantes en valoración de Inmuebles Urbanos. (*Tesis Doctoral. Universidad de Córdoba*). (S. d. Córdoba, Ed.) Córdoba, España.
- Rheingold, H. (1998). *The Virtual Community*. Obtenido de Homesteading on the Electronic Frontier: <http://www.rheingold.com/vc/book/intro.html>
- Rodríguez , Y., & Díaz, A. (2009). Herramientas de Minería de Datos. *Revista Cubana de Ciencias Informáticas*, 3(3-4), 73-80.
- Romero, C., & Ventura , S. (2010). Educational Data Mining: A Review of the State. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6), 601-618. doi:10.1109/TSMCC.2010.2053532.
- Romero, C., & Ventura, S. (2007). Educational Data Mining:A Survey from 1995 to 2005. *Expert System with Applications*, 33(1), 135-146. doi:10.1016/j.eswa.2006.04.005.
- Ruíz, C., & Basualdo, M. (s.f.). *Repositorio Institucional Abierto de Universidad Tecnologica Nacional*. Recuperado el 30 de Octubre de 2017, de Facultad Regional del Rosario: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf
- Rumbaugh, J., Jacobson, I., & Booch, G. (2000). *El Lenguaje Unificado de Modelado UML. Manual de Referencia* . Madrid: Pearson Educacion.
- Salinas, J. (2000). El aprendizaje colaborativo con los nuevos canales de comunicación. En J. Cabero, J. Salinas, A. Duarte, & J. Domingo, *Nuevas Tecnologías aplicadas a la Educación* (págs. 199-228). Madrid: Síntesis.
- Salinas, J. (2005). *La gestión de los entornos virtuales de formación.Seminario Internacional: La Calidad dela Formación en Red en el Espacio Europeo de Educación Superior*. Recuperado el 20 de Mayo de 2016, de https://www.researchgate.net/publication/232242328_La_gestion_de_los_Entornos_Virtuales_de_Formacion

- Sánchez, G. (Noviembre de 2004). Optimización de Estructuras de Redes Neuronales y Aplicaciones. (*Tesis Doctoral. Instituto Politécnico Nacional*). de México, México.
- Sanchez, J. (2003). Dos aplicaciones empíricas de las Redes Neuronales Artificiales a la clasificación y la predicción financiera en el Mercado español. *Revista Asturiana de Economía*(28), 61-87.
- Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students* (Fifth ed.). England: Pearson Education.
- SEMS. (2013). *Subsecretaría de Educación Media Superior: Sistema de Alerta Temprana - SIAT*. Recuperado el 03 de Junio de 2016, de SECRETARÍA DE EDUCACIÓN PÚBLICA: http://www.sems.gob.mx/en_mx/sems/sistema_alerta_temprana_siat
- SEP-SIGED. (2014). *Sistema de Información y Gestión Educativa*. Recuperado el 1 de Julio de 2016, de <http://www.siged.sep.gob.mx/>
- SEP-SNIEE. (2014). *Sistema Nacional de Información Estadística Educativa*. Recuperado el 1 de Julio de 2016, de Recuperado el 1 de julio de 2016 de (): <http://www.sniesep.gob.mx/>
- SEP-SNIEE. (2014). *Sistema Nacional de Información Estadística Educativa. Estadísticas Educativas*. Recuperado el 1 de Julio de 2016, de http://www.sniesep.gob.mx/estadisticas_educativas.html
- Society International Educational Data Mining. (29 de Jun de 2015). *Journal of Educational Data Mining*. Recuperado el 19 de Nov de 2015, de JEDM: <http://www.educationaldatamining.org/JEDM/index.php/JEDM>
- Someerville, I. (2005). *Ingeniería del software* (Séptima ed.). Madrid, España: Pearson.
- Sotolongo, G., & Guzmán, M. (2001). Aplicaciones de las redes neuronales. El caso de la Bibliometría. *Ciencias de la Información. Centro de Investigación-Producción de Vacunas y Sueros*, 32(1), 27-34.
- The Well Group, Inc. (1985). *Learn About The WELL*. Obtenido de <http://www.well.com/aboutwell.html>
- Torrado A., K., Barrientos A., E., & Vizcaino H., D. (2014). Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos. *Cooperación Latino Americana de Redes Avanzadas RedCLARA*, s.e.

- Torrecilla, P. (7 de Junio de 2012). *nosolopau.com*. Obtenido de <http://nosolopau.com/2012/06/07/mas-sobre-el-proceso-unificado-agil-fases-y-disciplinas/>
- Treviño, M., Ibarra, S., Castán, J., Laria, J., & Guzmán, J. (2013). A Framework to avoid Scholar Desertion using Artificial Intelligence. *Proceedings of the World Congress on Engineering*, 3, 1493-1497. Recuperado de: http://www.iaeng.org/publication/WCE2013/WCE2013_pp1493-1497.pdf.
- UIPPE. (2015). *Informe de indicadores UTVT*. Toluca.
- UNEP. (2012). *Early Warning Systems: A State of the Art Analysis and Future Directions*. (Division of Early Warning and Assessment, & United Nations Environment Programme, Edits.) Nairobi, Kenya.
- Utgoff, P. (1988). An Incremental ID3. *In Proc. Fifth Int. Conf. Machine Learning*, 107-120.
- Valdivia, J. (2009). La Comunidad de Práctica Online: Conocimiento y Aprendizaje. (Tesis Doctoral, Departamento de Didáctica, Organización Escolar y Didácticas Especiales). Universidad Nacional de Educación a Distancia. Madrid, España.
- Valero , S. (2009). *Aplicación de Técnicas de Minería de Datos para Predecir Deserción*. Obtenido de <http://www.utim.edu.mx/~svalero/docs/MineriaDesercion.pdf>
- Valero, S., Vargas, A., & García , M. (2010). Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. En M. Prieto, *Tecnologías del Aprendizaje* (pág. pag?). Merida: Kaambal.
- Vallejos, S., & Martínez, D. (2006). *Minería de Datos*. Obtenido de (Facultad de Ciencias Exactas, Naturales y Agrimensura, Universidad Nacional del Nordeste): http://exa.unne.edu.ar/informatica/SO/Mineria_Datos_Vallejos.pdf
- Villada, F., Muñoz, N., & García, E. (2012). Aplicación de las Redes Neuronales al Pronóstico de Precios en el Mercado de Valores. *Información Tecnológica*, 23(4), 11-20 doi:10.4067/S0718-07642012000400003.
- Villada, F., Muñoz, N., & García-Quintero, E. (2016). Redes Neuronales Artificiales aplicadas a la Predicción del Precio del Oro. (SciELO., Ed.) *Información tecnológica.*, 27(5), 143-150. doi:10.4067/S0718-07642016000500016.
- Villagómez, L. (2011). *Gestión de la innovación tecnológica y su vínculo con la oferta proveniente de las Universidades*. Obtenido de <http://www.gestiopolis.com/gestion-innovacion-tecnologica-vinculo-oferta-proveniente-universidades/>

- Waikato, W. U. (s.f.). *Machine Learning Group at the University of Waikato*. Recuperado el 17 de Jun de 2017, de <https://www.cs.waikato.ac.nz/ml/weka/book.html>
- WikiUniversidad. (Agosto de 2015). *Metodologías pesadas de desarrollo software*. Obtenido de https://es.wikiversity.org/wiki/Metodolog%C3%ADas_pesadas_de_desarrollo_software
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data mining, Practical machine learning tools and techniques*. USA: Morgan Kaufmann.
- Witten, I., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining. Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.
- Zea, C. M., & Atuesta, M. R. (2007). *Hacia una comunidad educativa interactiva*. Medellin: Fondo Editorial Universidad EAFIT.
- Zhang, D., Jiang , Q., & Li, X. (2004). Application of Neural Networks in Financial Data Mining. *International Journal of Computational Intelligence*, 1(2), 116-119.
- Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010). Use data mining to improve student retention in higher education - A case study. *Proceedings of the 12th International Conference on Enterprise Information Systems*, 1, 190-197. Recuperado de: <http://>.

ÍNDICE DE FIGURAS

Figura 1. Plan tecnológico.	39
Figura 2. Proceso KDD.....	42
Figura 3. Fases Proceso KDD.....	44
Figura 4. Proceso general de EDM.	52
Figura 5. Esquematización de una RNA.....	54
Figura 6. Fases de AUP.	77
Figura 7. Modelo MVC.....	86
Figura 8. Estrategia de Investigación.....	90
Figura 9. Metodología de Investigación MSWD&DM Propuesta.	93
Figura 10. Diagrama de Despliegue.....	107
Figura 11. Diagrama de Paquetes.	108
Figura 12. Diagrama de Clases.	109
Figura 13. Diagrama de Secuencia.	110
Figura 14. Diagrama de Casos de Uso.....	111
Figura 15. Pantalla de Login de CVAATI.	113
Figura 16. Pantalla Inicial de CVAATI.	114
Figura 17. Pantalla de Modelado en CVAATI.	115
Figura 18. Pantalla de Interpretación de Resultados.....	116
Figura 19. Pantalla de Integración de Datos a CVAATI.	123
Figura 20. Pantalla de Importación de Archivos a CVAATI.....	124
Figura 21. Módulo de Selección y Limpieza de Datos en la CVAATI.	126
Figura 22. Catálogo de definición de variables en CVAATI.....	129
Figura 23. Opciones de Pre procesamiento DM.	130

Figura 24. Asignación de nuevas variable en CVAATI.	131
Figura 25. Parametrización de variables en CVAATI.	136
Figura 26 Pre procesado a datos discretos.	137
Figura 27. Arquitectura de una RNA de FBR.....	149
Figura 28. Función Gaussiana de RNA de FBR.	150
Figura 29. Proceso de Aprendizaje con k-medias.....	153
Figura 30. Filtro Aplicado al conjunto de datos de entrada del Modelo DM.	156
Figura 31. Aplicación de algoritmo de Selección de Atributos.	158
Figura 32. Resultados del comportamiento de patrones con Adaline -Widrow & Hoff.....	163
Figura 33. Conjunto parcial de Patrones de entrenamiento del Algoritmo de Adaline - Widrow & Hoff General.	164
Figura 34. Función de Activación Hardlims.....	171

ÍNDICE DE GRÁFICAS

Gráfica 1. Tasa de Graduación del año 2005 y 2010 en Educación Superior de Licenciatura, grupo OCDE.	5
Gráfica 2. Comportamiento de los datos de los estudiantes del PE del primer período escolar Septiembre – Diciembre 2016.	148
Gráfica 3. Centroides identificados con k-medias	152
Gráfica 4. Clasificación de estudiantes con predicción a Baja de la IES.	172
Gráfica 5. Clasificación de estudiantes con predicción a Baja de la IES muestra 2.....	178
Gráfica 6. Resultados de Algoritmos de Clasificación DM.....	183
Gráfica 7. Error Cuadrático Calculado con algoritmos de RNA de FBR Experimento1.	186
Gráfica 8. Error Cuadrático Calculado con algoritmos de RNA de FBR Experimento 2.	187

ÍNDICE DE TABLAS

Tabla 1. Estadísticas modalidad escolarizada (cifras preliminares 2013-2014).....	3
Tabla 2. Estadísticas de indicadores de egreso por cohorte Generacional 2011-2015	6
Tabla 3. Estadísticas de indicadores educativos en México 2012-2014.....	19
Tabla 4. Variables o factores que influyen en los estudiantes de deserción o reprobación escolar	24
Tabla 5. Clasificación de las RNA por el tipo de Aprendizaje y de Topología.....	57
Tabla 6. Revisión de modelos de predicción de estadísticos educativos	62
Tabla 7. Análisis Comparativo del Modelo PredATIS con los modelos de predicción estudiados	66
Tabla 8. Modelos tradicionales de desarrollo de SW	69
Tabla 9. Análisis Comparativo del Modelos tradicionales y metodologías ágiles en el desarrollo de SW	74
Tabla 10. Diagramas UML	80
Tabla 11. Herramientas de diseño UML	84
Tabla 12. Fases del ciclo de desarrollo AUP utilizadas en la CVAATI.....	99
Tabla 13. Plan de desarrollo.....	103
Tabla 14. Comparativa de modelos RNA	119
Tabla 15. Variables de estudio.....	124
Tabla 16. Variables para Modelado con Técnicas de Clasificación y Selección DM	126
Tabla 17. Variables de Entrada para Modelado con RNA.....	128
Tabla 18. Valores categorizados de Variables en estudio.....	132
Tabla 19, Factores identificados por Modelo DM	159
Tabla 20. Resultados obtenidos con Reglas de Clasificación DM	160
Tabla 21. Resultados obtenidos con árbol de decisión DM.....	161

Tabla 22. Variables de entrada del entrenamiento con RNA de FBR experimento 1	166
Tabla 23. Conjunto parcial de Patrones de entrenamiento P experimento 1 muestra 1.....	167
Tabla 24. Algoritmo de RNA de FBR	168
Tabla 25. Resultados de la ejecución de Pruebas experimento 1	173
Tabla 26. Variables de entrada del entrenamiento con RNA de FBR experimento 2	174
Tabla 27. Conjunto parcial de Patrones de entrenamiento P experimento 2 muestra 2.....	176
Tabla 28. Resultados de ejecución de pruebas de experimento 2.....	179
Tabla 29. J48 pruned tree.....	184