



INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

SISTEMA PREDICTIVO DE DESERCIÓN ESCOLAR DE LOS ALUMNOS DEL ITSM UTILIZANDO SQL SERVER DATA TOOLS DE MICROSOFT VISUAL STUDIO 2012

TESIS

PARA OBTENER EL GRADO DE MAESTRO EN
SISTEMAS COMPUTACIONALES

P R E S E N T A

**ARNULFO GAMALIEL HERNÁNDEZ
GONZÁLEZ**

DIRECTOR

MIA. Roberto Ángel Meléndez Armenta

CO-DIRECTOR

Dr. Abel García Barrientos

MISANTLA, VERACRUZ

JUNIO, 2016.



**INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA
DIVISIÓN DE ESTUDIOS PROFESIONALES
AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA**

FECHA: 16 de Junio de 2016.

ASUNTO: **AUTORIZACIÓN DE IMPRESIÓN
DE TESIS.**

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

ARNULFO GAMALIEL HERNÁNDEZ GONZÁLEZ

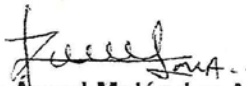
estudiante de la maestría en SISTEMAS COMPUTACIONALES con No. de Control 122T0553 ha cumplido satisfactoriamente con lo estipulado por el Lineamiento de Posgrado para la obtención del grado de Maestría mediante Tesis.


Por tal motivo se Autoriza la impresión del Tema titulado:


**SISTEMA PREDICTIVO DE DESERCIÓN ESCOLAR DE LOS ALUMNOS DEL
ITSM UTILIZANDO SQL SERVER DATA TOOLS DE MICROSOFT VISUAL
STUDIO 2012**

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

ATENTAMENTE


M.I.A. Roberto Angel Meléndez Armenta
Presidente


M.G.C. Eduardo Gutiérrez Almaraz
Secretario


Dr. Abel García Barrientos
Vocal



Archivo.

VER. 01/03/09

F-SA-39

Agradecimientos

Le agradezco a Dios por todos los dones que me regaló, por ser mi fortaleza y por hacerme día a día una mejor persona y por dame la excelente y hermosa familia que tengo.

Agradezco al Consejo Nacional de Ciencia y Tecnología (CONACYT) de México de quien recibí apoyo económico para cursar la maestría en Sistemas Computacionales en el Instituto Tecnológico Superior de Misantla.

Agradezco a la M.S.I. Ana Lilia Sosa y Durán de quien recibí mi primera clase de programación (Diseño de algoritmos), porque ha sido un pilar importante en mi crecimiento laboral, profesional y personal, gracias por brindarme todo su apoyo y la oportunidad de cursar la maestría, es usted un excelente ser humano.

Le agradezco al M.G.C. Eduardo Gutiérrez Almaraz por todos sus consejos, su amistad y por el conocimiento que me brindó ya que fue parte fundamental para la realización de esta tesis.

Agradezco el tiempo y los consejos, así como el apoyo de mis directores de tesis el M.I.A. Roberto Ángel Meléndez Armenta y al Dr. Abel García Barrientos. Sus enseñanzas me fortalecieron profesionalmente y personalmente.

Le agradezco al Dr. Luis Alberto Morales Rosales por sus enseñanzas dentro y fuera del aula de clase, por haber creído en mí ya que fue un pilar importante para que yo pudiera hacer esta maestría y por apoyarme en momentos difíciles de mi vida, mostrándome el gran ser humano que es.

Dedicatoria

A mi madre Felicitas Hernández González por haberme dado la vida y por hacer de mí el ser humano que soy a través de sus cuidados y sus enseñanzas demostrándome siempre su amor y lo orgullosa que está de mis logros. “Mamá siempre estaré agradecido de todo lo que tú tuviste que hacer para que mi hermana y yo fuéramos felices y pudiéramos ser personas de bien con una profesión, Gracias”

A mi esposa que es el amor de mi vida y que fue quien nunca dejó de insistir para que concluyera con la última etapa de la maestría dándome todo su apoyo. “Cuando me pregunten qué era lo que más me gustó de la vida, yo responderé que: Tu”.

Le dedico esta tesis a mi hermosa hija Daniella Ladiane Hernández Hernández quien ha sido mi mayor inspiración en la vida, ella iluminó mi vida desde que me enteré que había sido concebida desde entonces ha sido el motor que hace que mi vida valga la pena. “Nunca me cansaré de decirte que eres lo mejor que me ha pasado en la vida, gracias por existir”.

A mi hija la más pequeña Ella Danibeth Hernández Hernández quien ha venido a complementar mi vida haciéndome el hombre más feliz y afortunado del mundo, ahora puedo decir que no le hace falta nada a mi vida. “Eres el mejor regalo que Dios me ha dado”.

Resumen

En este trabajo se presenta el desarrollo de un sistema predictivo para detectar posibles alumnos desertores de los programas educativos del Instituto Tecnológico Superior de Misantla (ITSM). El sistema utiliza la información personal y académica de los alumnos existentes en la base de datos del sistema de control escolar (SIE) del instituto. El sistema es desarrollado con la herramienta SQL Server Data Tools de Microsoft Visual Studio 2012 de Microsoft SQL Server Analysis Services. El sistema utiliza la minería de datos y está basado en el algoritmo de Regresión Logística de Microsoft, con las cuales se pueden realizar las predicciones de deserción escolar y obtener el perfil de deserción escolar de los alumnos de la Ingeniería en Tecnologías de la Información y Comunicaciones (ITICs) del ITSM. La metodología empleada es la siguiente, primero se planteó el desarrollo de un sistema predictivo donde se adoptó el proceso de extracción, transformación y carga de los datos, extrayendo la información de alumnos del sistema "SIE" y haciendo uso de la herramienta DBFPlus para migrar las tablas a Excel. Posteriormente, se diseña una consulta de referencia cruzada para disponer de las calificaciones en columnas por alumnos utilizando Microsoft Access, migrando toda la información al servidor SQL Server 2012; el almacén de datos es utilizado para crear la estructura de minería de datos cuyas variables a utilizar son: número de control, nombre, edad y género del alumno, ciudad, municipio y escuela de procedencia y las materias cursadas por el alumno, en este caso se aplicó al programa educativo de la ingeniería en TICs. Finalmente, el 70% de los datos se utilizó para entrenamiento y el resto para pruebas, en el cual se agregaron los sistemas de minería de datos clúster y árbol de decisión, ambos de Microsoft. La aportación principal de esta tesis se centra en el desarrollo de un sistema predictivo para detectar alumnos con una probabilidad deserción alta, utilizando la información personal y académica de los alumnos del ITSM, empleando la herramienta SQL Server Data Tools de Microsoft Visual Studio 2012, con un 98.52% de probabilidad de predicción de alumnos desertores y proporcionar un perfil de deserción de alumnos del ITSM.

Contenido

Capítulo I. Generalidades	1
1.1. Introducción	1
1.2. Planteamiento del problema	2
1.3. Justificación	3
1.4. Hipótesis	5
1.5. Objetivo general.....	6
1.6. Objetivos específicos	6
1.7. Solución del problema	7
1.8. Alcances y limitaciones.....	8
1.8.1. Alcances	8
1.8.2. Limitaciones.....	8
Capítulo II. Análisis de los fundamentos	9
2.1. Marco teórico	9
2.1.1. Algoritmo de regresión logística	9
2.1.2. Algoritmo de clústeres	12
2.1.3. Algoritmo de árboles de decisión de Microsoft	14
2.1.4. Red Neuronal de Microsoft	17
2.1.5. Base de datos del sistema de información SIE	17
2.1.6. DBFPlus	18
2.1.7. Microsoft Access.....	19
2.1.8. Microsoft SQL Server Analysis Services	19
2.2. Estado del arte.....	20

Capítulo III. Desarrollo	24
3.1. Migración y transformación de datos	25
3.2. Creación de origen de datos.....	26
3.2.1. Consulta de tabla de referencias cruzadas.....	27
3.2.2. Entrenamiento	29
3.2.3. Prueba	32
3.3. Generación y selección del sistema de predicción	33
3.3.1. Estructura de minería de datos.....	34
3.3.2. Adición de sistemas a la estructura de minería de datos.....	37
Capítulo IV. Análisis de datos y resultados	43
4.1. Obtención de resultados	43
4.1.1. Matriz de Clasificación.....	43
4.1.2. Precisión y Sensibilidad (Recall)	46
4.1.3. Gráfico de elevación	47
4.1.4. Validación cruzada	48
4.1.5. Selección del sistema	53
4.1.6. Resultados.....	56
Capítulo V. Conclusiones y trabajo futuro	61
5.1. Conclusiones	61
5.2. Trabajo Futuro	62
Referencias.....	63
Anexos	67
7.1. Escuelas que contribuyen al perfil de deserción de los alumnos de la carrera de ITICs del ITSM.....	67

Índice de tablas

Tabla 2.1 Trabajos ya reportados sobre modelos para evaluar el desempeño del estudiante como la posibilidad de deserción.....	22
Tabla 3.1 Estructura de la tabla DKARDE	27
Tabla 3.2 Consulta de referencia cruzada	28
Tabla 3.3 Factores determinantes para predecir la deserción	30
Tabla 3.4 Tabla de entrenamiento	31
Tabla 3.5 Tabla de prueba	33
Tabla 3.6 Atributos finales para la estructura de minería de datos	35
Tabla 3.7 Estructura de minería de datos	35
Tabla 3.8 Parámetros del algoritmo de Regresión Logística.....	38
Tabla 3.9 Parámetros del algoritmo clúster.....	40
Tabla 3.10 Parámetros del algoritmo árbol de decisión	42
Tabla 4.1 Esquema de matriz de clasificación	44
Tabla 4.2 Matriz de clasificación para los sistemas de estudio.....	45
Tabla 4.3 F-Measure para los sistemas de estudio.....	46
Tabla 4.4 Leyendas para el gráfico de elevación para Desertor = 1 (desertor).....	47
Tabla 4.5 Estimaciones de validación cruzada de los sistemas de estudio	52
Tabla 4.6 Resultados de la matriz de confusión.....	54
Tabla 4.7 Resultados del gráfico de elevación.....	54
Tabla 4.8 Resultados de la matriz de confusión (Standard Deviation).....	56
Tabla 4.9 Resultados de predicciones con datos de prueba.....	57
Tabla 4.11 Resultados reales.....	58
Tabla 4.12 Perfil de deserción.....	59
Tabla 7.1 Escuelas de procedencia 1	67
Tabla 7.2 Escuelas de procedencia 2	71

Índice de figuras

Figura 2.1 Agrupación de clústeres.....	14
Figura 2.2 Diagrama de un árbol de decisión tradicional	16
Figura 2.3 Deserción y reprobación de la carrera de ITICs en el ITSM	18
Figura 3.1 Deserción y reprobación de la carrera de ITICs en el ITSM	25
Figura 3.2 Sistema de transformación de datos para el ITSM	26
Figura 4.1 Gráfico de elevación para los sistemas de estudio con Desertor = 1 (desertor).....	47

Índice de ecuaciones

Ecuación 2.1 Regresión Logística 1.....	10
Ecuación 2.2 Regresión Logística 2.....	10
Ecuación 2.3 Regresión Logística 3.....	10
Ecuación 2.4 Regresión Logística 4.....	11
Ecuación 2.5 Regresión Logística 5.....	11
Ecuación 2.6 Regresión Logística 6.....	11

Capítulo I. Generalidades

1.1. Introducción

En México y en otros países, el índice de deserción en el sector educativo es un fenómeno que afecta a la población estudiantil en general, los efectos que tienen estos fenómenos traen consigo una disminución en la eficiencia terminal y por lo tanto aumenta el rezago educativo nacional que finalmente se puede convertir en un problema social y económico. El Instituto Nacional para la Evaluación de la Educación (INEE) advirtió que la deserción escolar aumenta el desempleo y la incorporación de jóvenes al crimen organizado en México. La Organización para la Cooperación y el Desarrollo Económico (OCDE) dejó en claro que México ocupó el primer lugar en el número de desertores escolares de 15 a 18 años. La *Secretaría* de Educación Pública (SEP) señaló recientemente, que la deserción escolar en México provoca pérdidas de más de 34 millones de pesos, por el más de un millón de estudiantes que abandonaron sus estudios en los diferentes niveles de educación [24]. De manera general, se puede observar que la deserción es un tema común en la educación en todos los niveles, y en el ITSM no es ajeno a ello. El Instituto ha desarrollado campañas de apoyo a los estudiantes, asignando profesores tutores desde su ingreso al ITSM, además ha incorporado un programa de nivelación (curso propedéutico en matemáticas) para todos los alumnos de primer año, entre otras, esto con el fin de disminuir los índices de deserción. En el ITSM se han podido observar factores que inciden directamente en la deserción, siendo de los más conocidos: la reprobación, falta de vocación, problemas económicos, causas personales, incumplimiento de expectativas, problemas con docentes, entre otros factores; sin embargo a simple vista no se puede determinar las interrelaciones o los patrones marcados por los alumnos desertores, es ahí donde entra la minería de datos, la cual se concibe como el proceso de descubrir información útil a partir de grandes conjuntos de datos. Estos datos, mediante un análisis matemático, son utilizados para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden

detectar mediante una exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiados datos. Sin embargo, estos patrones y tendencias se pueden recopilar y definir como un sistema de minería de datos. En este caso, un sistema de minería de datos se crea mediante la aplicación de un algoritmo a los datos, pero es algo más que un algoritmo o un contenedor de metadatos: es un conjunto de datos, estadísticas y patrones que se pueden aplicar a los nuevos datos para generar predicciones y deducir relaciones. De ahí que, en este trabajo de tesis se optó por Microsoft SQL Server Analysis Services, ya que proporciona herramientas que puede utilizar para crear soluciones de minería de datos, utilizando directamente el Asistente para minería de datos de SQL Server Data Tools (SSDT) y el generador de consultas de predicción que proporciona SQL Server Management Studio y SSDT.

En el presente capítulo de este trabajo de investigación se enuncia el planteamiento del problema a resolver, la justificación del mismo, la hipótesis, el objetivo general, la metodología empleada, la solución propuesta y finalmente los alcances y limitaciones.

1.2. Planteamiento del problema

La deserción escolar es un factor que está directamente relacionado con la eficiencia terminal en cualquier institución educativa, y es un parámetro importante en la certificación de cualquier programa educativo. Para el caso del ITSM, la información proporcionada por el departamento de desarrollo institucional (al 2014) arroja que los principales factores que influyen en la deserción escolar de los alumnos son:

Estudio de carreras que no son primera opción para el alumno.

Alumnos sin el perfil requerido para la carrera.

Deficiencias en ciencias básicas.

Sin embargo, existen otros factores desconocidos que también influyen en la deserción escolar. En nuestro caso, el principal problema al que se enfrenta el ITSM al estudiar la deserción escolar es que no cuenta con un perfil de deserción de sus alumnos, debido a que no hay un mecanismo que permita captar información relevante de los mismos al ingresar a la institución ni durante su estancia en el Instituto, siendo el SIE el único almacén de datos que contiene información personal y académica de los alumnos. Para abordar la problemática anterior, se proponen las siguientes preguntas de investigación que son de interés para esta tesis:

¿Es posible construir un sistema predictivo utilizando la información personal y académica de los alumnos, almacenada en el SIE, del ITSM para detectar los posibles desertores en forma oportuna empleando el algoritmo de Regresión Logística de Microsoft?

¿Utilizando el sistema predictivo será posible determinar un perfil de deserción de los alumnos del ITSM?

1.3. Justificación

El ITSM nació para responder a las necesidades de educación superior de la región Misanteca y sus alrededores, ofreciendo educación de calidad. En el sistema educativo por competencias adoptado en el ITSM, se implementaron las tutorías como parte de un conjunto de estrategias para incrementar la eficiencia terminal y así disminuir la deserción escolar en el mismo instituto. En el ITSM se ha observado, desde su creación, que existen índices de reprobación altos y por ende sean blancos de deserción, aun cuando normalmente se les da un seguimiento adecuado, sin embargo, esta alerta se activa al final del primer parcial, o al final del semestre, por lo que los tutores no pueden tomar medidas a tiempo. Sin embargo, a través del tiempo se han podido observar factores que inciden directamente en la deserción siendo de los más conocidos: la reprobación,

falta de vocación, problemas económicos, causas personales, incumplimiento de expectativas, problemas con docentes, entre otros factores. De aquí la necesidad e importancia de predecir la probabilidad de deserción de un alumno desde que ingresa hasta que concluye sus estudios, esto es con la finalidad de aumentar la eficiencia terminal en el programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones (ITICs) del ITSM. Sin embargo, la metodología aplicada en este programa educativo puede ser aplicada para cualquier otro dentro del instituto y más aún, para cualquier institución educativa media y superior de nuestro país. De ahí el interés por desarrollar el presente trabajo, en predecir alumnos desertores del programa educativo de ITICs a partir de la información personal y académica almacenada en el SIE, del ITSM. Por lo tanto, es necesario desarrollar un sistema predictivo que ayude a detectar a tiempo los alumnos candidatos a desertar; y también que proporcione el perfil y probabilidad del alumno desertor. Al detectar alumnos con alta probabilidad de deserción, los tutores podrán actuar de forma oportuna dando al alumno diversas opciones tales como tutorías, asesorías, talleres, conferencias, diferentes tipos de becas que el mismo ITSM ofrece.

Por otra parte Microsoft SQL Server Analysis Services proporciona herramientas que facilitan el desarrollo del sistema predictivo, motivo por el cual se optó por su utilización para dar solución a la problemática citada en este documento. Microsoft SQL Server Analysis Services proporciona las siguientes herramientas que se pueden utilizar para crear soluciones de minería de datos: 1) El **asistente para minería de datos** de SQL Server Data Tools (SSDT) facilita la creación de estructuras y de sistemas de minería de datos, usando orígenes de datos relacionales o datos multidimensionales en cubos. En el asistente, se eligen los datos que desee utilizar y, a continuación se aplican técnicas de minería de datos específicas, como agrupación en clústeres, redes neurales o modelado de series temporales. 2) **SQL Server Management Studio** y **SQL Server Data Tools (SSDT)** disponen de visores de sistemas para explorar los sistemas de minería de datos una vez creados. Puede examinar los sistemas mediante visores adaptados a cada algoritmo o analizar con mayor profundidad utilizando el visor de contenido

del sistema. 3) El **Generador de consultas de predicción**, se proporciona en SQL Server Management Studio y SQL Server Data Tools (SSDT) para ayudarle a crear consultas de predicción. También se puede probar la exactitud de los sistemas respecto a un conjunto de datos de exclusión o datos externos, o utilizar validación cruzada para evaluar la calidad del conjunto de datos. 4) **SQL Server Management Studio** es la interfaz en la que administra las soluciones de minería de datos implementadas en una instancia de Analysis Services. Puede volver a procesar las estructuras y sistemas para actualizar los datos que contienen. 5) **SQL Server Integration Services** contiene herramientas que puede utilizar para limpiar datos, automatizar tareas como la creación de predicciones y actualización de sistemas y para crear soluciones de minería de datos de texto.

1.4. Hipótesis

Utilizando un sistema predictivo con la información personal y académica de cada uno de los alumnos, almacenada en el SIE, del ITSM se podrá detectar en forma oportuna los posibles alumnos desertores del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del mismo ITSM. Esta herramienta servirá como instrumento para identificar los posibles estudiantes con altas probabilidades de deserción, y ya con una detección oportuna nos permitirá implementar una estrategia para la retención del alumno desertor. Esto disminuirá los índices de deserción de los alumnos del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del ITSM. La deserción es un parámetro fundamental y preocupante para cualquier institución de educación superior que desee una certificación oficial, ya que se requiere un índice bajo.

1.5. Objetivo general

Desarrollar un sistema predictivo utilizando el algoritmo de Regresión Logística de Microsoft para detectar los posibles estudiantes desertores y obtener el perfil de deserción escolar de los alumnos del programa educativo de TICS utilizando la información personal y académica de los alumnos, almacenada en el SIE, del ITSM.

1.6. Objetivos específicos

- Seleccionar y migrar los datos académicos y personales de los alumnos del sistema SIE, del ITSM a Microsoft SQL Server 2012 con el fin de transformarlos en información útil para crear el sistema predictivo.
- Realizar el proceso de selección de variables de entrada para el sistema predictivo del programas educativo de TICS, con el fin de preparar datos de entrada con un formato adecuado para el proceso de minería de datos, utilizando SQL server 2012.
- Crear la estructura y los sistemas de minería de datos, para el sistema predictivo del programa educativo de TICS, utilizando la herramienta SQL Server Data Tools de Análisis Services de SQL server 2012.
- Validar el rendimiento de los sistemas de minería de datos usando datos reales con el fin evaluar la exactitud y determinar el grado de confiabilidad del sistema de minería de datos antes de implementarlo en el programa educativo de TICS del ITSM.
- Predecir los posibles alumnos desertores del programa educativo de TICS, del ITSM, utilizando el sistema de minería de datos basado en el algoritmo de Regresión Logística de Microsoft.
- Obtener el perfil de deserción escolar de los alumnos del programa educativo de TICS del ITSM, utilizando el modelo predictivo resultante de

esta investigación basado en el algoritmo de Regresión Logística de Microsoft.

1.7. Solución del problema

Para llevar a cabo el desarrollo de un sistema predictivo que permita detectar al alumno con alta probabilidad de deserción y más aún proporcionar el perfil de los alumnos desertores del programa educativo de Ingeniería en Tecnologías de Información y Comunicaciones, utilizando información personal y académica de alumnos almacenada en el SIE del ITSM. La solución propuesta consiste en cinco fases: la primera, consiste en migrar los datos del sistema SIE a Microsoft Excel, también se transforman los datos y se migran al servidor SQL Server 2012. En la segunda fase se seleccionan las variables de entrada para el sistema predictivo para preparar un origen de datos con un formato adecuado para el proceso de minería de datos. En la tercera, se crea la estructura del sistema de minería de datos y se agregan los sistemas de minería a utilizar, los cuales son validados mediante la matriz de clasificación, el gráfico de elevación y la validación cruzada de cada sistema. En la cuarta, se aplica el sistema de minería de datos basado en el algoritmo de Regresión Logística de Microsoft para predecir la deserción, y para obtener el perfil de deserción escolar. Por último, se adicionó una quinta fase que aunque ésta se encuentra fuera del área de investigación del presente trabajo se menciona ya que con la información proporcionada por el sistema la institución podrá emplear acciones preventivas o correctivas para evitar la deserción.

1.8. Alcances y limitaciones

1.8.1. Alcances

Este proyecto tiene como alcance la creación de un sistema con la finalidad de identificar a los alumnos con mayor probabilidad de deserción del programa educativo de Ingeniería en Tecnologías de Información del ITSM. Los aspectos puntuales que abarca la investigación del proyecto son los siguientes: proporcionar un reporte de los alumnos candidatos a desertar del primer semestre del programa educativo de TICS del ITSM, determinar la probabilidad de deserción de un alumno y proporcionar el perfil de deserción de los alumnos del primer semestre del programa educativo de TICS del ITSM. Se utilizará Microsoft SQL Server Analysis Services de SQL server 2012, el cual proporciona los algoritmos y herramientas de minería de datos con sistemas de regresión logística y clúster, algoritmos que se aplican a datos históricos comprendiendo desde el periodo agosto 2010 hasta agosto 2014.

1.8.2. Limitaciones

El proyecto de investigación está circunscrito en el ITSM, específicamente en el programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones, únicamente para primer semestre, donde por observación se ha detectado el mayor porcentaje de deserción de la carrera; sin embargo el sistema es completo y totalmente reproducible para cualquier carrera de cualquier tecnológico que utilice el sistema SIE.

Capítulo II. Análisis de los fundamentos

2.1. Marco teórico

2.1.1. Algoritmo de regresión logística

La regresión logística [9-11] es una técnica estadística conocida que se usa para modelar los resultados binarios. Existen varias implementaciones de regresión logística en la investigación estadística, que utilizan diferentes técnicas de aprendizaje. Por ejemplo, el algoritmo de Regresión logística de Microsoft es una variación del algoritmo de Red neuronal de Microsoft. Este algoritmo comparte muchas de las cualidades de las redes neurales, el cual tienen una característica importante que es el de fácil entrenamiento. Una de las ventajas de la regresión logística es que el algoritmo es muy flexible, pues puede tomar cualquier tipo de entrada y admite varias tareas analíticas diferentes. Algunas de las aplicaciones en la que se ha utilizado el algoritmo son las siguientes: 1) Usar datos demográficos para realizar predicciones sobre los resultados, como el riesgo de contraer una determinada enfermedad. 2) Explorar y ponderar los factores que contribuyen a un resultado. Por ejemplo, buscar los factores que influyen en los clientes para volver a visitar un establecimiento, 3) Clasificar los documentos, el correo electrónico u otros objetos que tengan muchos atributos.

Por otro lado, al preparar los datos para su uso en el entrenamiento de un sistema de regresión logística, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan. Los requisitos para un sistema de regresión logística son los siguientes: 1) Una columna de una sola clave: cada sistema debe contener una columna numérica o de texto que identifique cada registro de manera única y no están permitidas las claves compuestas. 2) Columnas de entrada: cada sistema debe tener al menos una columna de entrada que contenga los valores que se utilizan como factores en el análisis. Puede tener tantas columnas de entrada

como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el sistema. 3) El sistema debe contener al menos una columna de predicción de cualquier tipo de datos, incluidos datos numéricos continuos. Los valores de la columna de predicción también se pueden tratar como entradas del sistema, o se puede especificar que sólo se utilicen para las predicciones. No se admiten tablas anidadas en las columnas de predicción, pero se pueden usar como entradas.

La regresión logística analiza datos distribuidos binomialmente de la forma:

Ecuación 2.1 Regresión Logística 1

$$Y_i \sim B(p_i, n_i), \quad \text{para } i = 1, \dots, m,$$

Donde los números de ensayos Bernoulli n_i son conocidos y las probabilidades de éxito p_i son desconocidas. Un ejemplo de esta distribución es el porcentaje de semillas (p_i) que germinan después de que n_i son plantadas. El sistema es entonces obtenido a base de lo que cada ensayo (valor de i) y el conjunto de variables explicativas/independientes puedan informar acerca de la probabilidad final. Estas variables explicativas pueden pensarse como un vector X_i k -dimensional y el sistema toma entonces la forma

Ecuación 2.2 Regresión Logística 2

$$p_i = E\left(\frac{Y_i}{n_i} | X_i\right)$$

Los logits de las probabilidades binomiales desconocidas (i.e., los logaritmos de la razón de momios) son modeladas como una función lineal de los X_i .

Ecuación 2.3 Regresión Logística 3

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i}$$

Observe que un elemento particular de x_i puede ser ajustado a 1 para todo i obteniéndose una constante independiente en el sistema. Los parámetros desconocidos β_j son usualmente estimados a través de máxima verosimilitud.

La interpretación de los estimados del parámetro β_j es como los efectos aditivos en el logaritmo de la razón de momios (razón de oportunidades o razón de probabilidades) para una unidad de cambio en la j ésima variable explicativa. En el caso de una variable explicativa dicotómica, por ejemplo género, e^{β} es la estimación del odds ratio de tener el resultado para, por decir algo, hombres comparados con mujeres.

El sistema tiene una formulación equivalente dada por:

Ecuación 2.4 Regresión Logística 4

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \dots + \beta_k x_{k,i})}}$$

Esta forma funcional es comúnmente identificada como un "perceptrón" de una capa simple o red neuronal artificial de una sola capa. Una red neuronal de una sola capa calcula una salida continua en lugar de una función definida a trozos. La derivada de p_i con respecto a $X = x_1 \dots x_k$ es calculada de la forma general:

Ecuación 2.5 Regresión Logística 5

$$y = \frac{1}{1 + e^{-f(x)}}$$

Donde $f(X)$ es una función analítica en X . Con esta escogencia, la red de capa simple es idéntica al sistema de regresión logística. Esta función tiene una derivada continua, la cual permite ser usada en propagación hacia atrás. Esta función también es preferida pues su derivada es fácilmente calculable:

Ecuación 2.6 Regresión Logística 6

$$y' = y(1 - y) \frac{df}{dX}$$

De esta manera se plantea el algoritmo de regresión logística, el cual tiene infinidad de aplicaciones en tecnologías modernas, ya que nos ayudan a identificar predecir el resultado de una variable categórica.

2.1.2. Algoritmo de clústeres

El algoritmo de clústeres de Microsoft [15, 25-27] es un algoritmo de segmentación suministrado por Analysis Services. El algoritmo utiliza técnicas iterativas para agrupar los casos de un conjunto de datos dentro de clústeres que contienen características similares. Estas agrupaciones son útiles para la exploración de datos, la identificación de anomalías en los datos y la creación de predicciones. Los sistemas de agrupación en clústeres identifican las relaciones en un conjunto de datos que no se podrían derivar lógicamente a través de la observación casual. El algoritmo de clústeres se diferencia de otros algoritmos de minería de datos, como el algoritmo de árboles de decisión de Microsoft, en que no se tiene que designar una columna de predicción para generar un sistema de agrupación en clústeres. El algoritmo de clústeres entrena el sistema de forma estricta a partir de las relaciones que existen en los datos y de los clústeres que identifica el algoritmo.

El algoritmo de clústeres de Microsoft identifica, primero, las relaciones de un conjunto de datos y genera una serie de clústeres basándose en ellas. Un gráfico de dispersión es una forma útil de representar visualmente el modo en que el algoritmo agrupa los datos. Después de definir los clústeres, el algoritmo calcula el grado de perfección con que los clústeres representan las agrupaciones de puntos y, a continuación, intenta volver a definir las agrupaciones para crear clústeres que representen mejor los datos. El algoritmo establece una iteración en este proceso hasta que ya no es posible mejorar los resultados mediante la redefinición de los clústeres. Puede personalizar el funcionamiento del algoritmo seleccionando una técnica de agrupación en clústeres, limitando el número máximo de clústeres o cambiando la cantidad de soporte que se requiere para crear un clúster.

Al preparar los datos para su uso en el entrenamiento de un sistema de agrupación en clústeres, conviene comprender qué requisitos son imprescindibles para el algoritmo concreto, incluidos el volumen de datos necesario y la forma en que estos datos se utilizan. Los requisitos para un sistema de agrupación en clústeres son los siguientes: 1) Una columna key; cada sistema debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas. 2) Columnas de entrada; cada sistema debe tener al menos una columna de entrada que contenga los valores que se utilizan para generar los clústeres. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el sistema. 3) Una columna de predicción opcional; el algoritmo no necesita una columna de predicción para generar el sistema, pero puede agregar una columna de predicción de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas del sistema de agrupación en clústeres, o se puede especificar que sólo se utilicen para las predicciones.

El algoritmo de clústeres proporciona dos métodos para crear clústeres y asignar puntos de datos a dichos clústeres. El primero, el algoritmo mediana-k, es un método de agrupación en clústeres duro. Esto significa que un punto de datos puede pertenecer a un sólo clúster, y que únicamente se calcula una probabilidad de pertenencia de cada punto de datos de ese clúster. El segundo, el método Expectation Maximization (EM), es un método de agrupación en clústeres blando, el cual siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster. La implementación del algoritmo de clústeres utilizando el método EM proporciona dos opciones: EM escalable y no escalable. De forma predeterminada está en EM escalable y es el utilizado en el presente trabajo, tal como se muestra en la figura:

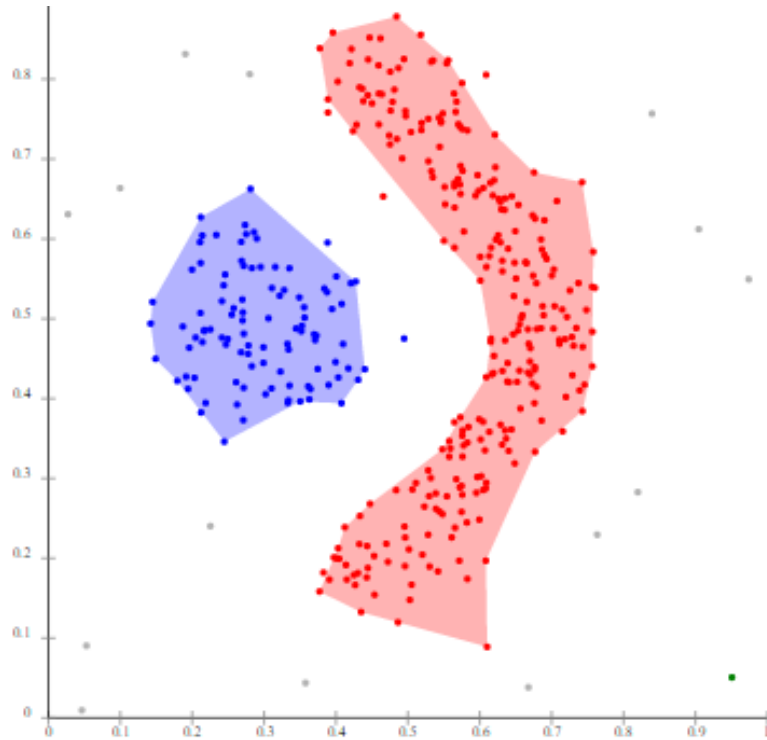


Figura 2.1 Agrupación de clústeres

2.1.3. Algoritmo de árboles de decisión de Microsoft

El algoritmo de árboles de decisión [12-17] de Microsoft es un algoritmo de clasificación y regresión proporcionado por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos, utiliza los valores, conocidos como estados, de estas columnas para predecir los estados de una columna que se designa como elemento de predicción. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto. Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión. Si se define más de una columna como elemento de predicción, o si los datos de entrada

contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción [19-23].

El algoritmo de árboles de decisión de Microsoft genera un sistema de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al sistema cada vez que una columna de entrada tiene una correlación significativa con la columna de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una columna discreta. El algoritmo de árboles de decisión de Microsoft utiliza la selección de características para guiar la selección de los atributos más útiles. Todos los algoritmos de minería de datos de Analysis Services utilizan la selección de características para mejorar el rendimiento y la calidad del análisis. La selección de características es importante para evitar que los atributos irrelevantes utilicen tiempo de procesador. Si utiliza demasiados atributos de predicción o de entrada al diseñar un sistema de minería de datos, el sistema puede tardar mucho tiempo en procesarse o incluso quedarse sin memoria. Un problema común de los sistemas de minería de datos es que el sistema se vuelve demasiado sensible a las diferencias pequeñas en los datos de entrenamiento, en cuyo caso se dice que está sobre ajustado o sobreentrenado. Un sistema sobre ajustado no se puede generalizar a otros conjuntos de datos. Para evitar sobreajuste de un conjunto de datos determinado, el algoritmo de árboles de decisión de Microsoft utiliza técnicas para controlar el crecimiento del árbol.

La forma en que el algoritmo de árboles de decisión de Microsoft genera un árbol, ver figura 2.2, para una columna de predicción discreta puede mostrarse mediante un histograma. Cuando el algoritmo de árboles de decisión de Microsoft genera un árbol basándose en una columna de predicción continua, cada nodo contiene una fórmula de regresión. Se produce una división en un punto de no linealidad de la fórmula de regresión.

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

- 1) Una columna key; cada sistema debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas,
- 2) Una columna de predicción; se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un sistema y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.
- 3) Columnas de entrada; se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

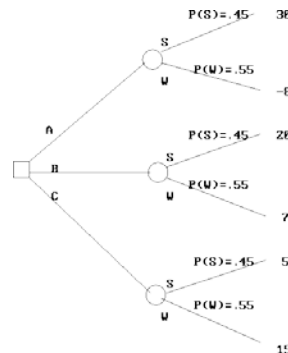


Figura 2.2 Diagrama de un árbol de decisión tradicional

De esta manera podemos establecer el algoritmo de árboles de decisión, este algoritmo es fácil de comprender y de implementar cuando se desea la búsqueda de alguna variable para su clasificación y regresión en específico.

2.1.4. Red Neuronal de Microsoft

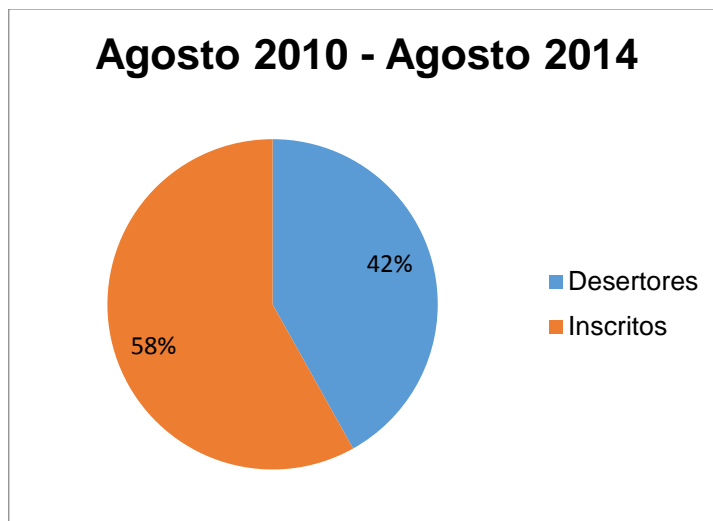
En SQL Server Analysis Services, el algoritmo de red neuronal [28-30] de Microsoft combina cada posible estado del atributo de entrada con cada posible estado del atributo de predicción, y usa los datos de entrenamiento para calcular las probabilidades. Posteriormente, puede usar estas probabilidades para la clasificación o la regresión, así como para predecir un resultado del atributo de predicción basándose en los atributos de entrada. Los modelos de minería de datos contruidos con el algoritmo de red neuronal de Microsoft pueden contener varias redes, en función del número de columnas que se utilizan para la entrada y la predicción, o sólo para la predicción. El número de redes que contiene un único modelo de minería de datos depende del número de estados que contienen las columnas de entrada y las columnas de predicción que utiliza el modelo. El algoritmo de red neuronal de Microsoft crea una red formada por hasta tres niveles de neuronas. Estas capas son una capa de entrada, una capa oculta opcional y una capa de salida. El modelo de red neuronal debe contener una columna de clave, una o más columnas de entrada y una o más columnas de predicción.

Los modelos de minería de datos que usan el algoritmo de red neuronal de Microsoft están muy influenciados por los valores que se especifican en los parámetros disponibles para el algoritmo. Los parámetros definen cómo se muestrean los datos, cómo se distribuyen o cómo se espera que estén distribuidos en cada columna, y cuándo se invoca la selección de características para limitar los valores usados en el modelo final.

2.1.5. Base de datos del sistema de información SIE

Para este caso de estudio, se obtuvo una muestra de alumnos del ITSM del periodo que comprende entre el mes de agosto del 2010 y el mes de agosto del 2015, donde se pudo contabilizar un total de 148 alumnos de la carrera de Ingeniería en Tecnologías de la Información y Comunicaciones (ITICs) de los

cuales 62 se reportaron como alumnos desertores, tal como se muestra en la figura 2.3.



Fuente: Elaborado a partir de datos proporcionados por Control Escolar (2014).

Figura 2.3 Deserción y reprobación de la carrera de ITICs en el ITSM

Como se puede observar en la figura 2.3, el porcentaje de desertores es muy elevado y por este motivo se seleccionó la carrera de ITICs para llevar a cabo la investigación y reportar resultados.

2.1.6. DBFPlus

DBFPlus es una herramienta compacta y de gran alcance para ver, editar e imprimir bases de datos en formato DBF. Es compatible con dBase, Clipper, FoxPro, Visual FoxPro y otros formatos de archivo DBF. DBFPlus es un programa para Windows, es fácil de usar e intuitivo, permite abrir y visualizar archivos DBF y convertir datos en formato DBF a archivos de Excel XLS; también puede convertir los datos a archivos tipo CSV [10]. Para la solución del problema se utilizó la herramienta DBFPlus para exportar los datos de las tablas en formato DBF a archivos de texto CSV.

2.1.7. Microsoft Access

Es un sistema de gestión de bases de datos incluido en el paquete ofimático denominado Microsoft Office. La razón por la cual se utilizó la herramienta Access es porque permite realizar consultas de referencia cruzada. Se define una consulta de referencias cruzadas cuando queremos representar una consulta resumen con dos columnas de agrupación como una tabla de doble entrada en la que cada una de las columnas de agrupación es una entrada de la tabla. En la base de datos donde se almacena el historial académico de los alumnos del ITSM, los datos de las calificaciones de las materias se encuentran organizados por filas y se requiere que las calificaciones se muestren en columnas por cada alumno y los nombres de éstas fueran las claves o los nombres de materias; esto se logró mediante el uso de consultas de referencia cruzada.

2.1.8. Microsoft SQL Server Analysis Services

Microsoft SQL Server Analysis Services [6] proporciona herramientas que facilitan el desarrollo del sistema predictivo y proporciona las siguientes herramientas que se pueden utilizar para crear soluciones de minería de datos:

- El **Asistente para minería de datos** de SQL Server Data Tools (SSDT) facilita la creación de estructuras y de sistemas de minería de datos, usando orígenes de datos relacionales o datos multidimensionales en cubos. En el asistente, elija los datos que desee utilizar y, a continuación, aplique técnicas de minería de datos específicas, como agrupación en clústeres, redes neurales o modelado de series temporales.
- **SQL Server Management Studio** y **SQL Server Data Tools (SSDT)** disponen de visores de sistemas para explorar los sistemas de minería de datos una vez creados. Puede examinar los sistemas mediante visores adaptados a cada algoritmo o analizar con mayor profundidad utilizando el visor de contenido del sistema.

- El **Generador de consultas de predicción**, se proporciona en SQL Server Management Studio y SQL Server Data Tools (SSDT) para ayudarle a crear consultas de predicción. También puede probar la exactitud de los sistemas respecto a un conjunto de datos de exclusión o datos externos, o utilizar validación cruzada para evaluar la calidad del conjunto de datos.
- **SQL Server Management Studio** es la interfaz en la que administra las soluciones de minería de datos implementadas en una instancia de Analysis Services. Puede volver a procesar las estructuras y sistemas para actualizar los datos que contienen.
- **SQL Server Integration Services** contiene herramientas que puede utilizar para limpiar datos, automatizar tareas como la creación de predicciones y actualización de sistemas y para crear soluciones de minería de datos de texto.

En este capítulo se analizaron los algoritmos que se utilizaron en este trabajo de tesis, los cuales ya son algoritmos estudiados y probados para algunas tareas ya realizadas. En este caso, éstos se utilizan para determinar un sistema, y más aún un perfil de los alumnos desertores.

2.2. Estado del arte

El desarrollo de un sistema automatizado para detectar a los alumnos con un alto porcentaje de abandonar sus estudios a nivel superior, no existe como tal [1-5]. Sin embargo, este es un parámetro importante para cualquier programa educativo a nivel superior que desee una certificación. Por tal razón, autoridades de las instituciones de educación superior han implementado estrategias que tienen como objetivo disminuir el índice de deserción en los alumnos de todos los niveles y ciclos escolares. Estas estrategias se realizan con base en la información que tiene el departamento de tutorías, ya que este último es el encargado de reducir los índices de deserción de cualquier institución de educación superior, además de poder guiar al alumno para superar cualquiera dificultad y así cumplir con su

objetivo de concluir sus estudios. Sin embargo, el departamento de tutorías, con en base en las herramientas que utiliza, obtiene la información del alumno de manera tardía, es decir que éste se entera de la deserción una vez que el alumno ya decidió abandonar sus estudios, por alguna razón que a veces se desconoce y ya sólo se tiene la oportunidad de aplicar una estrategia correctiva, como ofrecer una beca alimenticia, de transporte, etc. Sin embargo, con este método es imposible aplicar alguna estrategia preventiva. Y para mirar los datos crudos de esta problemática en nuestro país, basta con revisar las estadísticas que nos indican el índice de deserción para todos los niveles de educación en México, en el documento que publica la SEP México [29]. De ahí la necesidad de un sistema automatizado o semi-automatizado que nos proporcione la probabilidad de que un alumno tenga un alto porcentaje de deserción de manera temprana, y así poder aplicar una estrategia preventiva para evitar su deserción.

Algunos esfuerzos en la realización de sistemas han sido llevados a cabo para detectar de manera temprana las altas probabilidades que tiene un alumno de desertar en sus estudios, tal como se muestra en [6-7]. En la referencia [6] de los estudios más recientes sobre este tema, sin embargo, lo único que desean es este artículo es demostrar que puede obtener un sistema para los alumnos que tienden a desertar y en la referencia [7], se lleva a cabo un estudio, para una población muy pequeña. En ambas investigaciones utilizan técnicas de minería de datos para encontrar algunas predicciones, estos se basan principalmente en variables, que son a simple vista las que más influyen en el desempeño de los alumnos. La mayoría de los resultados de estas investigaciones han sido publicadas en la famosa conferencia de IEEE-Frontiers in Education Conference (FIE), sin embargo, los resultados mostrados [5-8] son principalmente en los alumnos de las áreas de ingeniería. Todos estos esfuerzos han sido exitosos, más aún en el área de matemáticas. Pero un sistema que se tomó como base la información de los alumnos que está concentrada en los departamentos de información de la institución y que por sí misma nos diga, qué estudiantes tienen altas probabilidades de desertar de sus estudios, no se ha realizado hasta ahora. En la tabla 1.1 se enlista los trabajos ya reportados, sobre este tema. Aunque los 3

primeros se basan principalmente a la experiencia que ya se tiene en casos anteriores, mientras los 3 últimos utilizan un modelo predictivo, tanto para evaluar el desempeño de los alumnos como su posible deserción. Sin embargo, ninguno de ellos muestra un perfil de deserción y más aún, este perfil nos puede aportar información relevante de un grupo de estudiantes.

Tabla 2.1 Trabajos ya reportados sobre modelos para evaluar el desempeño del estudiante como la posibilidad de deserción.

Referencia	Algoritmo	Aplicado a	Resultados
[2]	Acciones	Deserción y Evaluación	En este artículo principalmente muestra los resultados en la deserción escolar y en evaluación de los estudiantes. Los resultados son aceptables.
[3]	Modelo Cuantitativo	Deserción y Evaluación	Este artículo utiliza el modelo cuantitativo para analizar la deserción escolar. Los resultados son aceptables y comparables con otros autores
[4]	Difuso	Deserción	Este articulo crea un modelo utilizando lógica difusa para analizar la deserción escolar. Los resultados son aceptables y comparables con otros resultados.
[5]	Difuso	Evaluación	Este artículo crea un modelo con lógica difusa para analizar los resultados dela evaluación de los estudiantes. Los resultados son aceptables.

[6]	Modelo Predictivo	Deserción y Evaluación	En este artículo se utiliza un modelo predictivo para analizar la deserción y los resultados de la evaluación de los estudiantes. Los resultados son aceptables.
[7]	Modelo Predictivo	Deserción	En este artículo se utiliza un modelo predictivo para analizar la deserción de los estudiantes. Los resultados son aceptables.
[8]	Modelo Predictivo	Evaluación	En este artículo se utiliza un modelo predictivo para analizar la deserción y los resultados de la evaluación de los estudiantes. Los resultados son aceptables.

El sistema desarrollado en este trabajo de investigación utiliza la información personal y académica de los alumnos, el cual existente en la base de datos del sistema de control escolar (SIE) del mismo instituto. El sistema es desarrollado con la herramienta SQL Server Data Tools de Microsoft Visual Studio 2012 de Microsoft SQL Server Analysis Services. Además, este mismo sistema utiliza la minería de datos y está basado en el algoritmo de Regresión Logística de Microsoft, con las cuales se pueden realizar las predicciones de deserción escolar. Además el sistema hace uso del algoritmo Regresión Logística de Microsoft para obtener así el perfil de deserción escolar de los alumnos el programa educativo de TICS del ITSM.

Capítulo III. Desarrollo

En este capítulo se describe el desarrollo del proyecto, así como también los pasos que se llevaron a cabo para llegar al objetivo general del mismo, que es la de desarrollar e implementar un sistema predictivo utilizando el algoritmo de Regresión Logística de Microsoft que permita detectar a los posibles estudiantes desertores y obtener el perfil de deserción escolar de los alumnos del programa educativo de ITICs utilizando la información personal y académica de los alumnos, almacenada en el SIE, del ITSM. El desarrollo es fácil de implementar, en este trabajo de investigación sólo se ajustó para encontrar los posibles alumnos desertores del programa educativo de ingeniería en tecnologías de la información y comunicaciones de ITSM, sin embargo, este método puede ser aplicado para cualquier programa educativo del mismo instituto o más aun para cualquier plantel educativo que desee saber los posibles alumnos desertores en un cierto tiempo. Los resultados son interesantes porque, esto va a permitir a las autoridades de los planteles educativos a tomar medidas necesarias y así disminuir los índices de deserción.

El sistema propuesto para llevar a cabo esta investigación es el siguiente: como primer paso se tiene la migración y transformación de datos, de ahí viene una creación de origen de datos, con la información que más nos interesa. Después, viene la generación y selección del sistema de predicción, en el cual se utiliza después de ahí bien la obtención de resultados, y finalmente viene las acciones preventivas, por parte de nuestras autoridades para evitar la deserción de los estudiantes, tal como se muestra en la figura 3.1.

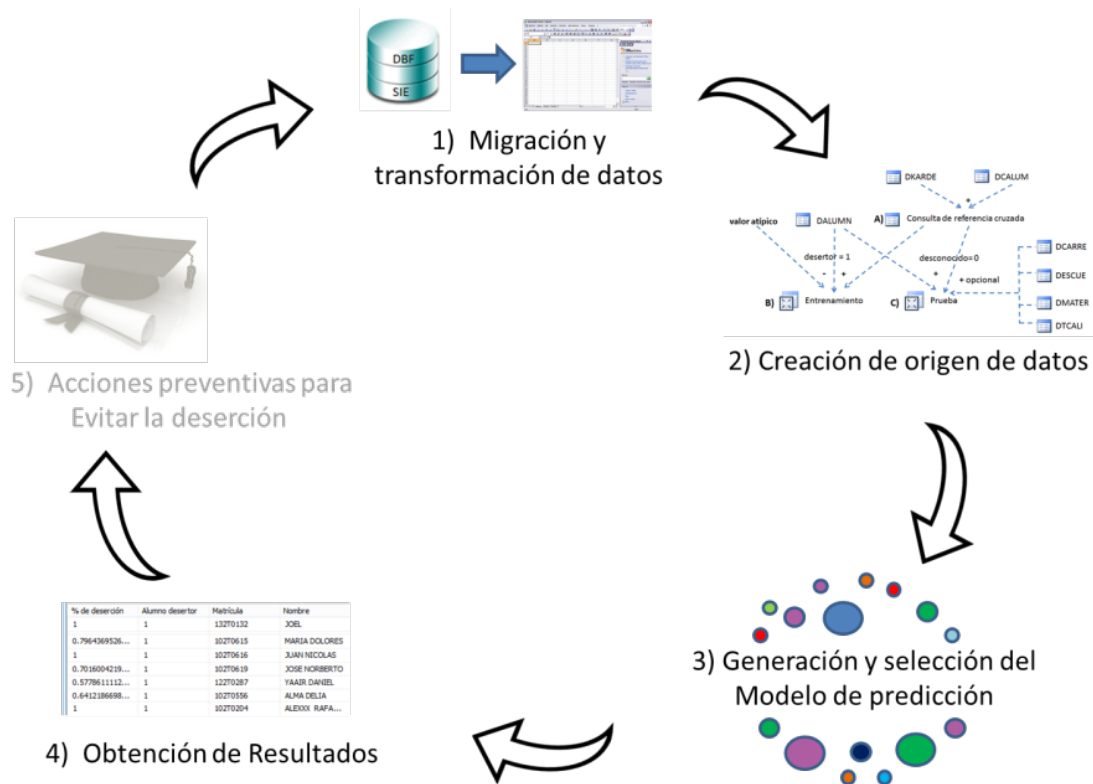


Figura 3.1 Deserción y reprobación de la carrera de ITICs en el ITSM

3.1. Migración y transformación de datos

En la primera etapa, migración y transformación de datos para el desarrollo del sistema, se debe de iniciar con obtener la información académica de los alumnos que se necesita, en este caso los alumnos de ITICs, estos datos los tiene el departamento de control escolar del ITSM, mediante el Sistema de Información Escolar denominado “SIE”. Este sistema consta de 103 tablas planas en formato DBF, de las cuales, para la solución del problema antes descrito, sólo se utilizaron 7 tablas: DKARDE, DCALUM, DALUMN, DCARRE, DESCUE, DMATER y DTCALI; éstas son las que deben ser emigradas a Microsoft Excel a través de la herramienta DBFPlus. DBFPlus, el cual permite exportarlas de formato DBF a Excel siempre y cuando la tabla no supere las 65536 tuplas, de ser así, se tendrá que exportar a archivos CSV y abrirlo en Excel; Tal es el caso de la tabla DKARDE que hasta este momento contiene 391759 registros. Una vez que éstas fueron

exportadas desde el Sistema Gestor de Base de Datos Relacional SQL Server 2012 mediante el asistente de SQL Server, se inicia el análisis y la creación de origen de los datos.

3.2. Creación de origen de datos

Los SGBD actuales brindan grandes bondades en cuanto a mejoría de capacidad de proceso que permiten realizar transformaciones complejas en SQL y también permiten tratar grandes volúmenes de información; por otro lado, la incorporación de capacidades de minería de datos, validación de datos o ejecución de algoritmo complejos in-database así como procesos de limpieza que hacen que los SGBD sean una excelente herramienta para realizar la transformación de datos, específicamente para este proyecto de investigación fue utilizado el SGBD SQLServer 2012, en el cual se procesaron en 3 etapas las siete tablas migradas anteriormente: A) Consulta de referencia cruzada, B) Entrenamiento y C) Prueba; estas etapas se pueden observar en la Figura 3.2 y serán descritas a continuación:

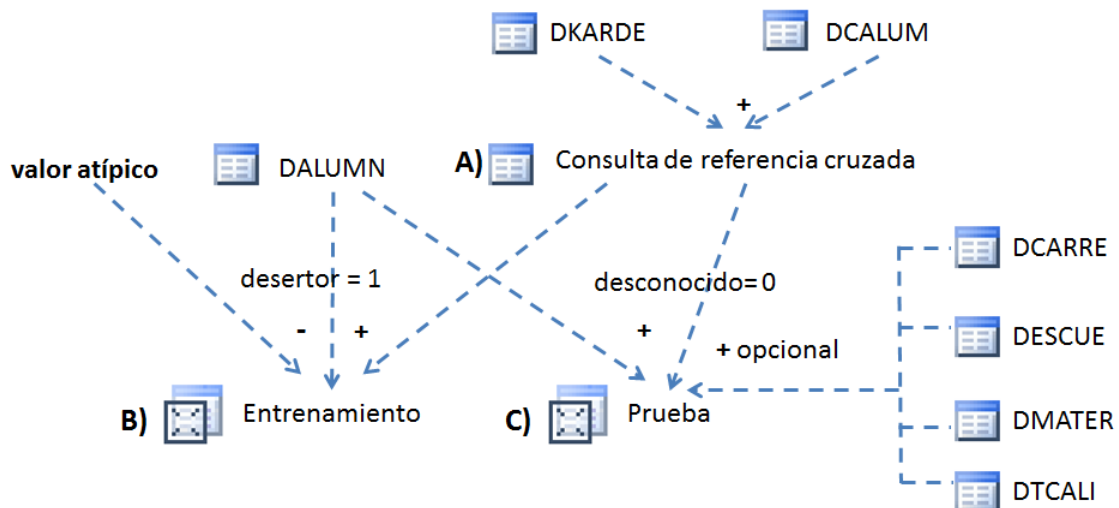


Figura 3.2 Sistema de transformación de datos para el ITSM

3.2.1. Consulta de tabla de referencias cruzadas

Los datos relativos a las calificaciones por materia se almacenan en la tabla DKARDE (del sistema SIE) en filas, como lo muestra en la Tabla 3.1 sin embargo no tiene la estructura requerida por los algoritmos a utilizar en esta investigación, por lo que se diseñó una consulta de referencia cruzada para crear una base de datos multidimensional, esta forma se garantiza una mejor visualización para su análisis y con ello se obtiene la estructura ideal para el estudio. Es importante no confundir la **consulta de tabla de referencia cruzada** con la **validación cruzada**. Ya que la primera se utiliza cuando queremos representar una consulta sumaria con dos o más columnas de agrupación como una tabla de doble entrada en la que cada una de las columnas de agrupación es una entrada y la segunda es una técnica para evaluar la precisión de un sistema de minería de datos, en este trabajo para validar el modelo predictivo resultado de la investigación, esta técnica se explica y analiza en el capítulo IV en el tema 4.1.4. La estructura y los datos quedaron como se muestran en la Tabla 3.2.

Tabla 3.1 Estructura de la tabla DKARDE

	ALUCTR	MATCVE	KARCAL	KARNPE1
1	102T0619	TIP1017	88	1
2	102T0623	ACA0907	88	1
3	102T0623	ACC0906	0	1
4	102T0623	ACF0901	0	1
5	102T0623	AEF1032	0	1
6	102T0623	TIF1019	92	1
7	102T0623	TIP1017	94	1
8	112T0348	ACA0907	88	1

Tabla 3.2 Consulta de referencia cruzada

	ALUCTR	ACA0907	ACC0906	ACF0901	AEF1032	TIF1019	TIP1017
1	102T0179	89	78	83	85	76	85
2	102T0180	98	90	100	95	90	93
3	102T0181	0	0	0	0	0	0
4	102T0182	93	79	78	80	73	84
5	102T0183	98	94	96	97	98	90
6	102T0184	97	90	95	97	81	95
7	102T0185	95	80	87	98	83	89
8	102T0186	97	85	96	99	87	88
9	102T0187	97	84	92	80	83	89

El significado de los nombres de los encabezados de la consulta de referencia cruzada es el número de control del alumno y las claves de las materias que cursó en primer semestre y se detallan en la Tabla 3.1 de la etapa de entrenamiento. La consulta de referencia cruzada se realizó utilizando datos de las tablas DKARDE y DCALUM, los campos de interés son: De la tabla DKARDE: ALUCTR (número de control), MATCVE (clave de la materia), KARCAL (calificación de la materia), KARNPE1 (semestre cursado) y de la tabla DCALUM: sólo se ocupó CARCVE para filtrar los alumnos por carrera. Para llegar de los datos originales a la tabla de “consulta de referencia cruzada” se hizo lo siguiente:

1. Crear una vista que contenga los datos de la tabla 3.1, filtrando la carrera objeto de estudio para este caso CARCVE=7 es la carrera de Ingeniería en Tecnologías de la Información y Comunicaciones.
2. Desde Microsoft Access fue importada la vista (“consulta de referencia cruzada”, creada en SQLServer) a mediante el uso de una fuente de datos ODBC previamente creada desde origen de datos de Windows.
3. En Access, una vez importada la vista, fue creada una consulta de referencia cruzada, usando el campo ALUCTR como encabezado de fila y el campo MATCVE como encabezado de columna y para cada intersección ALUCTR - MATCVE usar el valor del atributo KARCAL.
4. Una vez que fue creada con éxito la consulta de referencia cruzada, ésta fue exportada a un archivo de Excel.

5. Después de exportar la consulta a Excel, el archivo fue importado desde SQLServer para tener la tabla como parte del proyecto.
6. Se creó una base de datos, en este caso con el nombre de Tesis2015 para almacenar la información medular para esta investigación y en esta base de datos se crearon las tablas entrenamiento y prueba cuyos detalles de creación se explican en seguida:

3.2.2. Entrenamiento

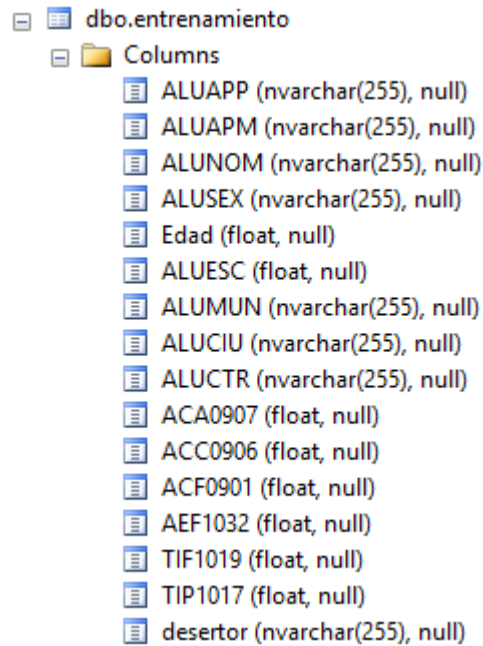
Como el objetivo principal de esta investigación es predecir alumnos desertores, para ello se debe crear un sistema de minería de datos adecuado para lograrlo. El Asistente de SQL Server Data Tools (SSDT) facilita su creación, usando orígenes de datos relacionales o datos multidimensionales en cubos. En el asistente, elija los datos que desee utilizar y, a continuación, aplique técnicas de minería de datos específicas, como agrupación en clústeres, redes neurales, árboles de decisión, etc. Esto para poder clasificar los datos, hacer recomendaciones o incluso predicciones. Por lo tanto, es necesario crear y alimentar un almacén de datos relacional que facilite la creación de la estructura y los sistemas de minería de datos. Para crear el conjunto de datos de entrenamiento primeramente se identificaron las variables más significativas a utilizar para este estudio, mismas que se detallan en la Tabla 3.3. Dicha tabla muestra los factores determinantes para la predicción de deserción, como son los de tipo personales, los académicos entre otros.

Tabla 3.3 Factores determinantes para predecir la deserción

Tipo	Atributo	Descripción	Posibles valores
Personales	ALUCTR	Número de control del alumno	Los números de control
	ALUAPM	Apellido materno del alumno	Todos los apellidos posibles
	ALUAPP	Apellido paterno del alumno	Todos los apellidos posibles
	ALUNOM	Nombre del alumno	Todos los nombres posibles
	ALUCIU	Ciudad de procedencia	Todas las claves de las ciudades
	ALUMUN	Municipio de procedencia	Todas las claves de los municipios
	ALUESC	Escuela de procedencia	Todas las claves de las escuelas
	ALUSEX	Género del alumno	Masculino (1), Femenino (2)
Académicos	ACA0907	Taller de ética	Rango de 0 a 100
	ACC0906	Fundamentos de investigación	Rango de 0 a 100
	ACF0901	Cálculo diferencial	Rango de 0 a 100
	AEF1032	Fundamentos de programación	Rango de 0 a 100
	TIF1019	Matemáticas discretas I	Rango de 0 a 100
	TIP1017	Introducción a las TICs	Rango de 0 a 100
Socioeconómicos	No se cuenta con ningún dato de este tipo en el almacén de datos		
Otros	EDAD	Edad del alumno	Rango 18 a 50
	DESERTOR	El estado a predecir	Desertó (1), Inscrito (0)

Con los atributos anteriores se creó la tabla denominada “entrenamiento” y cuya estructura se muestra en la tabla 3.4 para nuestro caso de estudio, se logró alimentar ésta (entrenamiento), utilizando la información de las generaciones 2010, 2011 y 2012 de la carrera de ITICs, el proceso para el llenado de la tabla se detalla a continuación:

Tabla 3.4 Tabla de entrenamiento



The image shows a screenshot of a SQL Server Enterprise Manager interface. It displays the 'dbo.entrenamiento' database and its 'Columns' folder. The columns listed are:

Column Name	Data Type
ALUAPP	(nvarchar(255), null)
ALUAPM	(nvarchar(255), null)
ALUNOM	(nvarchar(255), null)
ALUSEX	(nvarchar(255), null)
Edad	(float, null)
ALUESC	(float, null)
ALUMUN	(nvarchar(255), null)
ALUCIU	(nvarchar(255), null)
ALUCTR	(nvarchar(255), null)
ACA0907	(float, null)
ACC0906	(float, null)
ACF0901	(float, null)
AEF1032	(float, null)
TIF1019	(float, null)
TIP1017	(float, null)
desertor	(nvarchar(255), null)

- A). Primeramente, se reunieron los datos de la tabla DALUMN con los datos obtenidos en la CONSULTA DE TABLA DE REFERENCIA CRUZADA, tomando como base los datos de la tabla 3.3. La información obtenida en este proceso se almaceno en la tabla 3.4
- B). Se le asignó el valor 1 al atributo “desertor” de la tabla “entrenamiento”, a los alumnos registrados en el SIE y que han desertado de la carrera, esta decisión se apoyó de los datos proporcionados por el departamento de control escolar; cabe mencionar que el control de bajas es llevado en papel, de ahí lo complejo y tedioso de este proceso.
- C). Posteriormente, se asignó el valor 0 al mismo atributo, de la tabla entrenamiento, a los alumnos registrados en el sistema SIE como inscritos, control escolar también proporcionó un reporte escrito, esta vez tomado del SIE.
- D). Después se eliminaron las tuplas con casos atípicos (en inglés outlier), como es el caso de alumnos con excelentes calificaciones que se dan de baja por cambio de domicilio, escuela, trabajo, etc.

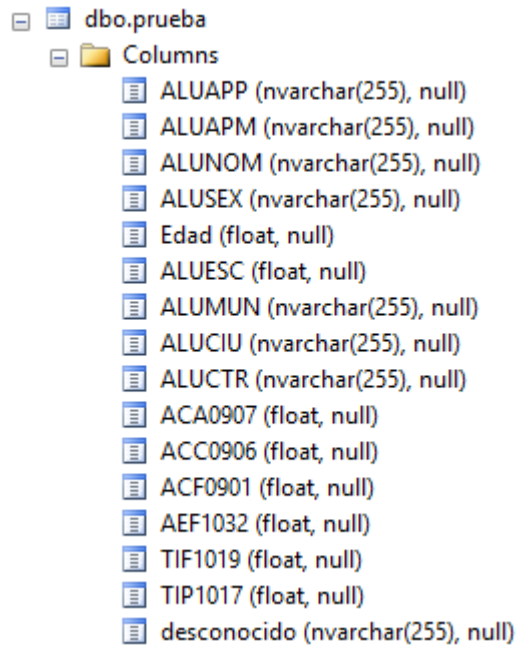
E). Finalmente, se procesaron los valores faltantes utilizando el enfoque del valor más pequeño, en este caso 0, puesto que alumnos que desertaron, ya no cursaron ciertas materias y el SIE les asigna un valor NULL el cual fue cambiado por un 0. Cabe mencionar que este simple cambio mejoró enormemente los resultados finales. Es importante mencionar que el sistema SIE asigna un cero como calificación final a los alumnos con calificación menor a 70.

De esta manera se llevó a cabo el proceso de llenado de la tabla 3.4, donde se descartan los valores extremos, por ejemplo, los alumnos excelentes y los alumnos que no tiene probabilidad de continuar con sus estudios, por problemas personales.

3.2.3. Prueba

Para los datos de prueba se diseñó una tabla denominada “prueba” y su estructura se puede observar en la tabla 3.5. que tiene la misma estructura que la de “entrenamiento”; las diferencias que existen entre ambas es que en la de “prueba” el valor asignado al atributo “desertor” es cero para todos los alumnos, esto bajo el supuesto de que no se sabe que alumnos están inscritos y que alumnos ya desertaron de la institución, dejando la tarea al sistema de minería de datos que haga la predicción de los posibles alumnos desertores, y que sea él quien coloque un 1 al alumno identificado como desertor. Otra diferencia es que los datos almacenados en esta tabla son de los alumnos de la generación 2013 de la carrera de ITICs, cabe mencionar que esta información no fue incluida en la de entrenamiento y por lo tanto no fue tomada en cuenta en la creación del modelo predictivo, con el fin de hacer pruebas reales mismas que son descritas y analizadas en el capítulo IV en el tema 4.1.6.1.2 predicciones reales.

Tabla 3.5 Tabla de prueba



Column Name	Data Type	Nullability
ALUAPP	nvarchar(255)	null
ALUAPM	nvarchar(255)	null
ALUNOM	nvarchar(255)	null
ALUSEX	nvarchar(255)	null
Edad	float	null
ALUESC	float	null
ALUMUN	nvarchar(255)	null
ALUCIU	nvarchar(255)	null
ALUCTR	nvarchar(255)	null
ACA0907	float	null
ACC0906	float	null
ACF0901	float	null
AEF1032	float	null
TIF1019	float	null
TIP1017	float	null
desconocido	nvarchar(255)	null

3.3. Generación y selección del sistema de predicción

La tercera fase de este desarrollo es sobre la generación y selección del sistema de predicción. En esta fase, se creó una estructura con los datos de entrenamiento que contiene un sistema de minería de datos individual basado en el algoritmo de regresión logística de Microsoft, se agregan los sistemas de minería de datos clúster y árbol de decisión, ambos de Microsoft. Posteriormente, se validarán sus características utilizando la matriz de clasificación, el gráfico de elevación y la validación cruzada de cada sistema. En esta etapa la principal tarea es la construcción de la estructura de un sistema de minería de datos, usando el algoritmo de regresión logística, además de utilizar el algoritmo de Clúster y Árboles de decisión, con el propósito de tener un parámetro de comparación respecto a qué sistema proporciona una mejor respuesta en cuanto a la estimación de la variable predictora de deserción escolar “desertor”. Para la construcción de los sistemas de minería de datos, se dispuso de los alumnos de la carrera de Ingeniería en Tecnologías de Información y Comunicaciones del Instituto Tecnológico Superior de Misantla, de las cohortes 2010 al 2012 y para

validar el sistema se dejó los alumnos de la cohorte 2013. Ésta etapa se dividió en cuatro secciones: 1.- Estructura de minería de datos, 2.- Adición de sistemas a la estructura de minería de datos, 3.- Evaluación de los sistemas y 4.- Selección del sistema, dichas etapas se detallan enseguida:

3.3.1. Estructura de minería de datos

La estructura de minería de datos define el dominio de los datos desde el que se generan, una sola estructura de minería de datos puede contener varios sistemas compartiendo los mismos datos. Para la creación del sistema predictivo se utilizó una estructura de minería de datos relacional ya que la ventaja de crearlas así son que puede reunir datos ad hoc y entrenar y actualizar un sistema sin la complejidad que supone crear un cubo; en este caso se utiliza la estructura y los datos de la tabla “entrenamiento” mostrada en la tabla 3.4 y se basa en el algoritmo de regresión logística. Cabe mencionar que para la construcción de una estructura de minería de datos se utiliza la unidad básica dispuesta en columnas, donde éstas almacenan información referente al tipo de datos, contenido y modo de uso. Para esta investigación se ha preparado una estructura, considerando los atributos descritos en la tabla 3.6, con el objeto de aplicar diferentes sistemas de minería de datos, y evaluar el comportamiento de cada sistema.

Tabla 3.6 Atributos finales para la estructura de minería de datos

COLUMNA	USO	TIPO CONTENIDO DE	TIPO DE DATOS
ACA0907	Input	Continuous	Double
ACC0906	Input	Continuous	Double
ACF0901	Input	Continuous	Double
AEF1032	Input	Continuous	Double
ALUAPM	Ignore	Discrete	Text
ALUAPP	Ignore	Discrete	Text
ALUCIU	Input	Discrete	Text
ALUCTR	Key	Key	Text
ALUESC	Input	Continuous	Double
ALUMUN	Input	Discrete	Text
ALUNOM	Ignore	Discrete	Text
ALUSEX	Input	Discrete	Text
Desertor	PredictOnly	Discrete	Text
Edad	Input	Continuous	Double
TIF1019	Input	Continuous	Double
TIP1017	Input	Continuous	Double

Siguiendo el proceso, es necesario presentar la estructura de minería de datos, y los parámetros asociados, para la base de datos de Tesis2015, utilizando la tabla “entrenamiento”, tal como se muestra en la tabla 3.7.

Tabla 3.7 Estructura de minería de datos

Property	Value
Advanced	
ErrorConfiguration	(default)
Language	
Basic	
Description	
ID	Entrenamiento
Name	Entrenamiento
Misc	
CacheMode	KeepTrainingCases
Collation	
HoldoutMaxCases	0
HoldoutMaxPercent	30
HoldoutSeed	0
Source	Tesis2015 (Data source view)

donde:

- `HoldoutMaxCases = 0`: Especifica el número máximo de casos en el origen de datos que se van a utilizar en la partición de exclusión que contiene el conjunto de pruebas para la estructura de minería de datos “Entrenamiento”. Los casos restantes en el conjunto de datos se usan para el entrenamiento. Un valor 0 indica que no hay ningún límite con respecto al número de casos que se pueden considerar como el conjunto de pruebas.
- `HoldoutMaxPercent = 30`: Especifica el porcentaje máximo de casos en el origen de datos que se van a usar en la partición de exclusión que contiene el conjunto de pruebas para la estructura de minería de datos “Entrenamiento”. Los casos restantes se usan para aprendizaje. Un valor 0 indica que no hay ningún límite con respecto al número de casos que se pueden considerar como el conjunto de pruebas. Para este proyecto de investigación se especificó el valor de la propiedad `HoldoutMaxPercent = 30`, esto indica el máximo de casos en el origen de datos que se van a usar para las pruebas, 30% para este caso; lo que quiere decir que el 70% de los datos se utilizará para entrenamiento.
- Si especifica los valores de `HoldoutMaxPercent` y `HoldoutMaxCases` el algoritmo limita el conjunto de pruebas al menor de los dos valores.
- Si `HoldoutMaxCases` está establecido en el valor predeterminado de 0 y no se ha establecido un valor para `HoldoutMaxPercent`, el algoritmo utiliza el conjunto de datos completo para entrenamiento.

Una vez definida la estructura de minería de datos “Entrenamiento”, se procede a definir los sistemas a desarrollar en esta investigación, estos son: Regresión Logística, Árbol de decisión, Cluster K-medianas, con los atributos indicados en la Tabla 3.6. En la siguiente sección se presentan los sistemas propuestos con sus respectivas características.

3.3.2. Adición de sistemas a la estructura de minería de datos

La estructura de minería de datos que creó, contiene un sistema de minería de datos individual que se basa en el algoritmo de Regresión Logística, es común que para asegurar que el análisis es detallado, crear sistemas relacionados usando algoritmos diferentes y comparar sus resultados, obteniéndose puntos de vistas diferentes.

3.3.2.1. Algoritmo de Regresión Logística

El algoritmo de regresión logística de Microsoft es una variación del algoritmo de red neuronal de Microsoft, en la que el parámetro `HIDDEN_NODE_RATIO` se establece en 0. Este valor creará un modelo de red neuronal que no contenga un nivel oculto y que, por consiguiente, sea equivalente a la regresión logística. En la Tabla 3.8 se muestran los parámetros que se pueden utilizar con el algoritmo de regresión logística de Microsoft y con los que se configuró el algoritmo, destacando:

`HOLDOUT_PERCENTAGE`: Especifica el porcentaje de casos en los datos de entrenamiento que se usan para calcular el error de exclusión. `HOLDOUT_PERCENTAGE` se utiliza como parte de los criterios de detención durante el entrenamiento del modelo de minería de datos. El valor predeterminado es 30.

`HOLDOUT_SEED`: Especifica un número que se utiliza para inicializar el generador pseudoaleatorio cuando se determinan aleatoriamente los datos de exclusión. Si `HOLDOUT_SEED` se establece en 0, el algoritmo genera la inicialización basada en el nombre del modelo de minería de datos, para garantizar que el contenido del modelo sigue siendo el mismo durante el nuevo procesamiento. El valor predeterminado es 0.

MAXIMUM_INPUT_ATTRIBUTES: Define el número de atributos de entrada que puede administrar el algoritmo antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características. El valor predeterminado es 255.

MAXIMUM_OUTPUT_ATTRIBUTES: Define el número de atributos de salida que puede administrar el algoritmo antes de invocar la selección de características. Establezca este valor en 0 para desactivar la selección de características. El valor predeterminado es 255.

MAXIMUM_STATES: Especifica el número máximo de estados de atributo que admite el algoritmo. Si el número de estados que tiene un atributo es mayor que el número máximo de estados, el algoritmo utiliza los estados más conocidos del atributo y pasa por alto los estados restantes. El valor predeterminado es 100.

SAMPLE_SIZE: Especifica el número de casos que se van a utilizar para entrenar el modelo. El proveedor de algoritmos utiliza el valor menor entre este número o el porcentaje del total de los casos que no están incluidos en el porcentaje de exclusión según se especifica en el parámetro HOLDOUT_PERCENTAGE. En otras palabras, si HOLDOUT_PERCENTAGE está establecido en 30, el algoritmo utilizará el valor de este parámetro o un valor que sea igual al 70 por ciento del número total de casos, según cuál sea menor. El valor predeterminado es 10000.

Tabla 3.8 Parámetros del algoritmo de Regresión Logística

Parameter	Value	Default	Range
HOLDOUT_PERCENTAGE		30	(0,100)
HOLDOUT_SEED		0	(...,...)
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_STATES		100	0,[2,65535]
SAMPLE_SIZE		10000	[0,...)

3.3.2.2. Algoritmo de clúster

El algoritmo de clúster de Microsoft utiliza el método Expectation Maximization (EM), es un método de agrupación en clústeres blando. Esto significa que un punto de datos siempre pertenece a varios clústeres, y que se calcula una probabilidad para cada combinación de punto de datos y clúster.

En la Tabla 3.9 se muestran los parámetros con los que se configuró el algoritmo, destacando que en el parámetro CLUSTER_COUNT se estableció en 0, con esto el algoritmo utiliza la heurística para determinar el mejor número de clústeres que debe generar, si se deja el valor predeterminado (el cual es 10) el algoritmo siempre genera 10 clústeres. Durante esta investigación se observó mejores resultados cuando se usó el valor 0 en comparación a cuando se usó el valor por defecto.

Por otra parte, el método predeterminado para la agrupación en clústeres es el método EM escalable. Para elegir el algoritmo a utilizar se dejó el valor por defecto en 1 para el parámetro CLUSTERING_METHOD.

Con este paso se determinaron las consideraciones que deben realizar para determinar los clusters, los cuales se enuncian en la tabla 3.5. Donde se puede observar, los valores de cada parámetro y el rango que puede tomar cada uno de éstos. En un principio pueden tomar los valores por default, sin embargo, estos pueden ser cambiados de acuerdo a la necesidad del objeto de estudio. En este caso se tomó una muestra de 50000, y una entrada de atributos de 255.

Para el resto de los parámetros se conservó los valores por preestablecidos.

Tabla 3.9 Parámetros del algoritmo clúster

Parameters:			
Parameter	Value	Default	Range
CLUSTER_COUNT	0	10	[0,...)
CLUSTER_SEED		0	[0,...)
CLUSTERING_METHOD		1	1,2,3,4
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_STATES		100	0,[2,65535]
MINIMUM_SUPPORT		1	(0,...)
MODELLING_CARDINALITY		10	[1,50]
SAMPLE_SIZE		50000	0,[100,...)
STOPPING_TOLERANCE		10	(0,...)

Donde CLUSTER_SEED especifica el número de inicialización usado para generar clústeres aleatoriamente para la fase inicial de generación del sistema. MINIMUM_SUPPORT especifica el número mínimo de casos requeridos para generar un clúster. Si el número de casos del clúster es inferior a este número, el clúster se trata como vacío y se descarta. MODELLING_CARDINALITY especifica el número de sistemas de ejemplo que se construyen durante el proceso de agrupación en clústeres. STOPPING_TOLERANCE especifica el valor que se usa para determinar cuándo se alcanza la convergencia y el algoritmo termina de generar el sistema. SAMPLe_SIZE especifica el número de casos que el algoritmo usa en cada paso si el parámetro CLUSTERING_METHOD está establecido en uno de los métodos de agrupación en clústeres escalables. MAXIMUM_INPUT_ATTRIBUTES especifica el número máximo de atributos de entrada que el algoritmo puede procesar antes de invocar la selección de características. MAXIMUM_STATES especifica el número máximo de estados de atributo admitido por el algoritmo. Si un atributo tiene más estados que el máximo permitido, el algoritmo usa los estados más conocidos y pasa por alto los estados restantes.

Además, un sistema de agrupación en clústeres debe contener una columna de clave y columnas de entrada. También se pueden definir columnas de entrada como columnas de predicción. Las columnas establecidas en Predict Only no se

usan para generar clústeres. La distribución de estos valores en los clústeres se calcula después de que se hayan generado los clústeres. De tal forma que el algoritmo de clústeres de Microsoft admite las columnas de entrada y de predicción específicas, en nuestro caso se pueden definir las más importantes, como son las de calificaciones.

3.3.2.3. Algoritmo de árbol de decisión

El algoritmo de árboles de decisión de Microsoft es un algoritmo de clasificación y regresión proporcionado por Microsoft SQL Server Analysis Services para el modelado de predicción de atributos discretos y continuos. Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos. Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción. Para el caso de estudio, es necesario predecir los alumnos que tienen mayor probabilidad de desertar de estudios a nivel ingeniería. El árbol de decisión realiza predicciones basándose en la tendencia hacia un resultado concreto. Para los atributos continuos, el algoritmo usa la regresión lineal para determinar dónde se divide un árbol de decisión. Si se define más de un atributo como elemento de predicción, o si los datos de entrada contienen una tabla anidada que se haya establecido como elemento de predicción, el algoritmo genera un árbol de decisión independiente para cada columna de predicción.

En este caso, el algoritmo de árboles de decisión de Microsoft genera un sistema de minería de datos mediante la creación de una serie de divisiones en el árbol. Estas divisiones se representan como nodos. El algoritmo agrega un nodo al sistema cada vez que una columna de entrada tiene una correlación significativa con la de predicción. La forma en que el algoritmo determina una división varía en función de si predice una columna continua o una discreta. Y los parámetros que utilizamos para el algoritmo de árbol de decisión se muestran en la tabla 3.10.

Tabla 3.10 **Parámetros del algoritmo árbol de decisión**

Parameters:			
Parameter	Value	Default	Range
COMPLEXITY_PENALTY			(0.0,1.0)
FORCE_REGRESSOR			
MAXIMUM_INPUT_ATTRIBUTES		255	[0,65535]
MAXIMUM_OUTPUT_ATTRIBUTES		255	[0,65535]
MINIMUM_SUPPORT		10.0	(0.0,...)
SCORE_METHOD		4	1,3,4
SPLIT_METHOD		3	[1,3]

Una vez procesado el sistema, los resultados se almacenan como un conjunto de patrones y estadísticas que se pueden usar para explorar las relaciones o para realizar predicciones.

El proceso de evaluar el rendimiento de los sistemas de minería de datos con datos reales se le conoce como validación. Antes de implementar un sistema de minería de datos es importante validar sus características y su calidad. Una vez que se generaron los sistemas de minería de datos se procedió a evaluarlos y determinar cuál de los sistemas era el mejor para realizar las predicciones, para evaluarlos se utilizó la matriz de clasificación de cada sistema, el gráfico de elevación y la validación cruzada; contando finalmente con suficientes argumentos para escoger uno de ellos.

Capítulo IV. Análisis de datos y resultados

4.1. Obtención de resultados

En este capítulo se muestran los resultados después de aplicar los algoritmos propuestos en el capítulo 2. Los resultados muestran que si es posible predecir a los alumnos que tienen altas posibilidades de desertar de sus estudios a nivel superior. En este caso, fue un análisis sobre los alumnos del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del Instituto Tecnológico Superior de Misantla, Veracruz, donde después de un análisis de 134 alumnos aproximadamente y utilizando la información de la base de datos del mismo instituto y aplicando los algoritmos se logró determinar los posibles alumnos desertores. Este método de identificación de alumnos desertores es una tarea difícil para realizarlo manualmente, debido a que es demasiada información y por tal razón un sistema automatizado es necesario. Los algoritmos implementados muestran resultados exitosos, ya que se logró identificar los posibles alumnos desertores y un perfil de los mismos. Para tomar una decisión de consideraron las siguientes métricas de clasificación.

4.1.1. Matriz de Clasificación

Una matriz de clasificación ordena todos los casos del sistema en categorías, determinando si el valor de predicción coincide con el valor real. La matriz de clasificación es una herramienta estándar de evaluación de sistemas estadísticos a la que a veces se denomina matriz de confusión.

El gráfico que se crea con la matriz de clasificación compara los valores reales con los de predicción. Las filas de la matriz representan los valores de predicción para el sistema, mientras que las columnas representan los valores reales. Las categorías usadas en el análisis son falso positivo, verdadero positivo, falso negativo y verdadero negativo.

Una matriz de clasificación es una herramienta importante para evaluar los resultados de la predicción, ya que hace que resulte fácil entender y explicar los efectos de las predicciones erróneas. Al ver la cantidad y los porcentajes en cada celda de la matriz, se podrá saber rápidamente en cuántas ocasiones ha sido exacta la predicción del sistema.

El esquema de la matriz de clasificación también conocida como matriz de confusión se muestra en la Tabla 4.1 y es para un caso de clasificación binaria.

Tabla 4.1 Esquema de matriz de clasificación

Categorías		Clase Actual	
		0	1
Clase Hipotética	0	TN	FN
	1	FP	TP
Columnas Totales		$N=FP+TN$	$P=TP+FN$

Los datos de la tabla 4.1, sirvieron para obtener métricas, mismas que se usaron para evaluar la matriz de confusión.

Como ya se observó los algoritmos son probabilísticos, y por lo tanto se buscan las entradas que más se repiten, esto ayuda a los mismos algoritmos a tomar decisiones, así como también a descartar las entradas que no tiene influencia en la toma de decisiones. En nuestro caso, los algoritmos implementados si ayudan a determinar un perfil de deserción de los alumnos de ITICs del ITSM. Esto se mostrará en el siguiente capítulo de esta tesis, el cual describe los resultados obtenidos.

La matriz de clasificación nos sirve para validar nuestros algoritmos, ya que estos entregan información importante sobre los resultados que arrojan a la aplicación de los algoritmos con la base de datos del sistema de información, en este caso del programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del ITSM. La tabla 4.2 muestra la matriz de clasificación para cada algoritmo aplicado, la de clúster EM escalable, la de árbol de decisión y regresión logística, con los valores reales y sus porcentajes correspondientes.

Tabla 4.2 Matriz de clasificación para los sistemas de estudio

Clúster EM escalable				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	2	100.00%	25.00%
1 (desertor)	0	6	0.00%	75.00%
correctas	32	6	100%	75%
incorrectas	0	2	0%	25%
Árbol de Decisión				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	8	100.00%	100.00%
1 (desertor)	0	0	0.00%	0.00%
correctas	32	0	100%	0%
incorrectas	0	8	0%	100%
Regresión Logística				
	0 (inscrito Real)	1 (desertor Real)	0 (inscrito Real)	1 (desertor Real)
0 (inscrito)	32	1	100.00%	12.50%
1 (desertor)	0	7	0.00%	87.50%
correctas	32	7	100%	88%
incorrectas	0	1	0%	13%

Para nuestro caso de estudio, de los 148 registros de alumnos de la carrera de ITICs, se seleccionaron 134 registros para el estudio, 94 registros al azar y se utilizaron para el entrenamiento (70%) y 40 registros (el resto) se utilizaron para las pruebas (30%). Por lo consiguiente, se excluyeron 14 registros del almacén de datos para aplicarle el sistema seleccionado como el mejor y obtener resultados reales. De esos 40 registros de prueba se puede apreciar en la Tabla 4.2 que los tres sistemas clasificaron correctamente (100%) a los alumnos inscritos, también se puede apreciar que el sistema de Clúster clasificó en un 75% a los alumnos desertores. Sin embargo, el sistema árbol de decisión falla al 100% en los alumnos desertores y en contraste el sistema Regresión Logística acierta al 88% al clasificar correctamente a los alumnos desertores.

4.1.2. Precisión y Sensibilidad (Recall)

La Tabla 4.3 representa las medidas de Precisión y Recall para los 3 sistemas de estudio, estas medidas están calculadas a partir de las matrices de clasificación, mostrada en la Tabla 4.1. La precisión se refiere a la fracción de ejemplares que se han clasificado como de la clase correspondiente y que, en realidad, son de esa clase. Por otro lado, el Recall (sensibilidad) se refiere a la fracción de ejemplos de la clase de todo el conjunto que se clasifican correctamente, es decir, mide la probabilidad de que si un alumno pertenece a una categoría el sistema lo asigne a esa categoría. Sin embargo existe otra métrica muy común para comparar sistemas: F-Measure que es una combinación de ambas, y se refiere a la media armónica de Precisión y Recall. A continuación se detallan los porcentajes obtenidos con estas métricas en los diferentes algoritmos.

Tabla 4.3 F-Measure para los sistemas de estudio

	Clúster	Árbol de Decisión	Regresión Logística
Precision	100%	0%	100%
Recall	75%	0%	88%
Accuracy	95%	80%	98%
F-Measure	86%	0%	93%
Lift	5	0	5

En la tabla anterior, claramente se puede observar que el sistema de Regresión Logística es el mejor sistema de acuerdo a las métricas, ya que clasificó a todos los estudiantes inscritos en la categoría que le corresponde, con una precisión del 100%, y una Sensibilidad del 88% debido a que de ocho estudiantes desertores siete los clasificó correctamente y un alumno desertor lo clasificó como inscrito, también se puede ver claramente que la Regresión Logística obtuvo los mejores porcentajes en Accuracy y F-Measure; sin embargo es necesario explorar otras métricas para asegurar que no haya sesgo en la investigación, esto se aclarará con los resultados de otras métricas (gráfico de elevación y validación cruzada) que a continuación se detallan

4.1.3. Gráfico de elevación

Un método para visualizar la mejora que se obtiene al utilizar un sistema de minería de datos, en comparación con una estimación aleatoria es el de gráfico de elevación. Con este gráfico se puede encontrar el número alumnos con posibilidades de desertar, tal como se muestra en la figura 4.1 cuyas leyendas del gráfico se encuentran en la tabla 4.4.

Figura 4.1 Gráfico de elevación para los sistemas de estudio con Desertor = 1 (desertor)

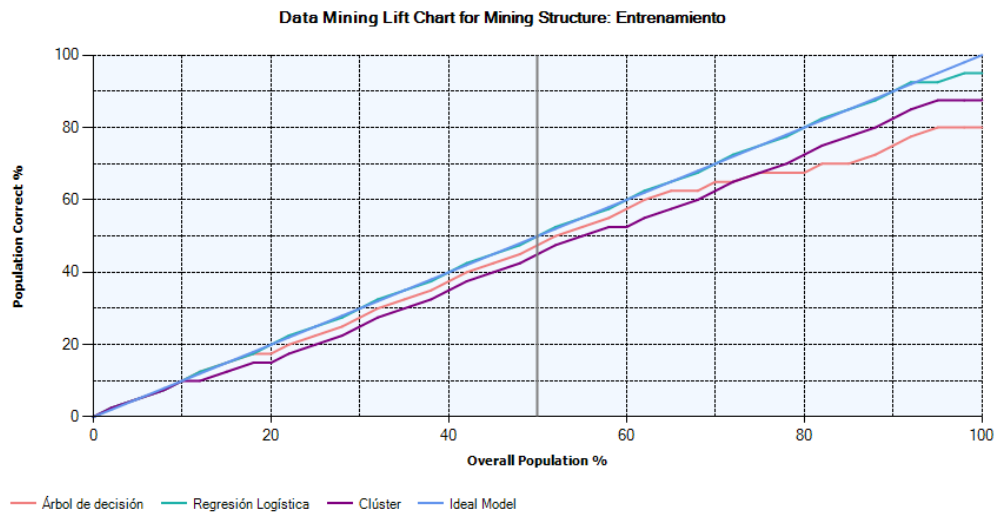


Tabla 4.4 Leyendas para el gráfico de elevación para Desertor = 1 (desertor)

Mining Legend			
Population percentage: 50.00%			
Series, Model	Score	Population correct	Predict probability
Árbol de decisión	0.90	47.50%	95.80%
Regresión Logística	1.00	50.00%	98.52%
Clúster	0.89	45.00%	100.00%
Ideal Model		50.00%	

En la Figura 4.1, la abscisa representa el porcentaje del conjunto de datos de prueba que se usa para comparar las predicciones. En la ordenada se representa el porcentaje de valores que se predicen con el atributo desertor = 1. En el gráfico,

la línea azul representa el sistema ideal. Por tal razón, el atributo de destino es “Desertor” y el valor de destino es 1 (significa desertor), lo que representa que el alumno es probable que deserte. El gráfico de elevación muestra así la mejora que el sistema proporciona al identificar a los estudiantes que es probable que deserten.

El valor de probabilidad de predicción representa el porcentaje para incluir un alumno en el grupo con probabilidad de desertar y el valor de puntuación ayuda a comparar los sistemas calculando la efectividad del sistema a través de una población normalizada. La mayor puntuación, representa al mejor sistema, estos datos se muestran en la tabla 4.4.

El algoritmo de Regresión logística es el que obtuvo la mejor puntuación (Score) con 100% con un porcentaje de población del 50%.

4.1.4. Validación cruzada

La validación cruzada es una técnica establecida en la comunidad de la minería de datos para evaluar la validez de un conjunto de datos y la precisión de un sistema de minería de datos en dicho conjunto. Esto consiste en dividir un conjunto de datos en subconjuntos y, después, va creando, entrenando y probando sistemas de forma iterativa en cada subconjunto. Por otro lado, la validación cruzada es una herramienta estándar de análisis que resulta muy útil a la hora de desarrollar y ajustar sistemas de minería de datos. Esta técnica también se usa después de crear un sistema de minería de datos para determinar la validez del sistema y comparar sus resultados con otros sistemas de minería de datos relacionados. En la tabla 4.5 se presenta el resumen de las medidas de precisión detalladas para cada partición para los sistemas Clúster, Árbol de decisión y Regresión Logística. En la estimación de validación cruzada, de la tabla 4.5, se presentan las 7 particiones y los resultados de las distintas métricas “True Positive (verdaderos positivos)”, “False Positive (falsos positivos)”, “True Negative (verdaderos negativos)”, “False Negative falsos negativos”, “Log Score (probabilidades

puntuación de registro)", "Lift (elevación)" y "Root Mean Square Error (error cuadrático medio)". Para cada una de ellas, se entrega en el resumen de cada partición el promedio y la desviación estándar.

4.1.4.1. Verdaderos positivos

Ésta métrica representa el recuento de alumnos clasificados correctamente como desertores, aquí el sistema clasifica alumnos desertores que realmente han desertado y de acuerdo a los resultados obtenidos, el algoritmo que tuvo menos dispersión de datos representado por la mejor desviación estándar es el árbol de decisión con un 0.4886 mientras que el Clúster y la Regresión Logística obtuvieron 0.499 con una diferencia casi insignificante de 0.010; tal como se puede ver en la tabla 4.5.

4.1.4.2. Falsos positivos

Ésta métrica indica el número de veces que el sistema predijo que alguien desertó cuando en realidad está inscrito con base en los resultados obtenidos. El sistema de Clúster obtuvo 0.2942 de desviación estándar obteniendo así mejores resultados, sin embargo, la regresión logística obtuvo 0.3083 en la misma métrica dejando una diferencia de 0.014, tal como se puede ver en la tabla 4.5; al analizar estos resultados es necesario decir que si el sistema clasifica algún alumno como posible desertor aún cuando el alumno no lo sea, lo que podría hacerse es que reciba algún tipo de asesoría o beca otorgada por la institución, siendo en el mejor de los casos no tan grave ya que es un alumno que no abandonará sus estudios.

4.1.4.3. Verdaderos negativos

Ésta métrica indica el número de casos que el sistema predijo correctamente como alumnos inscritos, obteniendo la mejor desviación estándar el sistema de Clúster con 0.4418, también se puede observar que la Regresión Logística obtuvo un 0.4473 dando una diferencia de 0.005 de dispersión del conjunto de datos. El árbol de decisión en esta métrica está muy próximo a 1 con un 0.9376 alejándose del objetivo enormemente.

4.1.4.4. Falsos negativos

Esta estadística indica que el sistema predijo que un alumno está inscrito cuando en realidad ya desertó, de acuerdo a los resultados los sistemas de Clúster y Regresión Logística fueron los que mejor clasificaron obteniendo los mismos porcentajes: para el promedio ambos obtuvieron 0.1064 y para la desviación estándar obtuvieron 0.3083, tal como se puede ver en la tabla 4.5.

4.1.4.5. Puntuación del registro (Log Score)

El significado de ésta métrica indica el grado en que la probabilidad de predicción de los sistemas se asemeja a la predicción aleatoria y en este sentido el sistema de regresión logística es el que más se acerca al sistema de predicción aleatoria una desviación estándar de 0.0698; por cierto muy cercano al cero.

4.1.4.6. Elevación (Lift)

Este indicador representa la proporción entre la probabilidad de predicción real y la probabilidad marginal en los casos de prueba y muestra hasta qué punto mejora la

probabilidad cuando se utiliza el sistema. En este sentido, se tiene que el sistema Regresión Logística presenta una mejor desviación estándar entre la probabilidad de predicción real y la probabilidad marginal en los casos de prueba con un 0.0761, respecto a los sistemas Clúster (0.1109), Árbol de decisión (0.1146), tal como se puede observar en la tabla 4.5.

4.1.4.7. Error cuadrático medio (*Root Mean Square Error*)

Esta métrica corresponde a la raíz cuadrada del error promedio para todos los casos de partición, dividido por el número de casos en la partición. Según los valores presentados en la tabla 4.5, se tiene que el sistema Regresión Logística (0.0467), tiene el mejor promedio con respecto al resto de los sistemas; sin embargo el sistema de Clúster tiene la menor desviación estándar de 0.0580 la diferencia entre el Clúster y la Regresión Logística es de un 0.0021.

Tabla 4.5 Estimaciones de validación cruzada de los sistemas de estudio

Partition Index	Partition Size	Test	Measure	Clúster EM	Árbol de decisión	Regresión Logística
1	8	Classification	True Positive	1	1	1
2	9	Classification	True Positive	1	1	1
3	10	Classification	True Positive	2	1	2
4	10	Classification	True Positive	2	0.00E+00	2
5	10	Classification	True Positive	1	0.00E+00	2
6	10	Classification	True Positive	2	0.00E+00	2
7	10	Classification	True Positive	2	0.00E+00	2
8	10	Classification	True Positive	2	1	1
9	9	Classification	True Positive	1	0.00E+00	1
10	8	Classification	True Positive	1	0.00E+00	1
			Average	1.5319	0.3936	1.5319
			Standard Deviation	0.499	0.4886	0.499
1	8	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
2	9	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
3	10	Classification	False Positive	0.00E+00	3	0.00E+00
4	10	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
5	10	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
6	10	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
7	10	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
8	10	Classification	False Positive	0.00E+00	1	0.00E+00
9	9	Classification	False Positive	1	0.00E+00	0.00E+00
10	8	Classification	False Positive	0.00E+00	0.00E+00	0.00E+00
			Average	0.0957	0.4255	0.1064
			Standard Deviation	0.2942	0.9395	0.3083
1	8	Classification	True Negative	7	7	7
2	9	Classification	True Negative	8	8	8
3	10	Classification	True Negative	8	5	8
4	10	Classification	True Negative	8	8	8
5	10	Classification	True Negative	8	8	8
6	10	Classification	True Negative	8	8	8
7	10	Classification	True Negative	8	8	7
8	10	Classification	True Negative	8	7	8
9	9	Classification	True Negative	7	8	8
10	8	Classification	True Negative	7	7	7
			Average	7.7340	7.4043	7.7234
			Standard Deviation	0.4418	0.9376	0.4473
1	8	Classification	False Negative	0.00E+00	0.00E+00	0.00E+00
2	9	Classification	False Negative	0.00E+00	0.00E+00	0.00E+00
3	10	Classification	False Negative	0.00E+00	1	0.00E+00
4	10	Classification	False Negative	2	0.00E+00	2
5	10	Classification	False Negative	1	2	0.00E+00
6	10	Classification	False Negative	0.00E+00	2	0.00E+00
7	10	Classification	False Negative	0.00E+00	2	0.00E+00
8	10	Classification	False Negative	0.00E+00	1	1
9	9	Classification	False Negative	0.00E+00	1	0.00E+00
10	8	Classification	False Negative	0.00E+00	1	0.00E+00
			Average	0.1064	1.2447	0.1064
			Standard Deviation	0.3083	0.7393	0.3083
1	8	Likelihood	Log Score	-1.41E-05	-0.099	-0.0159
2	9	Likelihood	Log Score	-0.0131	-0.0934	-0.016
3	10	Likelihood	Log Score	0.00E+00	-0.6357	-0.0163
4	10	Likelihood	Log Score	-0.0078	-0.2937	-0.0163
5	10	Likelihood	Log Score	-0.3261	-0.3458	-0.0163
6	10	Likelihood	Log Score	-0.0256	-0.3458	-0.0163
7	10	Likelihood	Log Score	-0.0102	-0.3458	-0.1757
8	10	Likelihood	Log Score	-0.0575	-0.4796	-0.1977
9	9	Likelihood	Log Score	-0.1195	-0.171	-0.0203
10	8	Likelihood	Log Score	-0.0095	-0.2567	-0.0157
			Average	-0.0589	-0.3158	-0.0528
			Standard Deviation	0.0983	0.1585	0.0698
1	8	Likelihood	Lift	0.3768	0.2778	0.3609
2	9	Likelihood	Lift	0.3357	0.2554	0.3328
3	10	Likelihood	Lift	0.5004	-0.1353	0.4841
4	10	Likelihood	Lift	0.4926	0.2067	0.4841
5	10	Likelihood	Lift	0.1743	0.1546	0.4841
6	10	Likelihood	Lift	0.4748	0.1546	0.4841
7	10	Likelihood	Lift	0.4902	0.1546	0.3247
8	10	Likelihood	Lift	0.4429	0.0208	0.3027
9	9	Likelihood	Lift	0.2294	0.1778	0.3286
10	8	Likelihood	Lift	0.3673	0.1201	0.361
			Average	0.3914	0.1345	0.3975
			Standard Deviation	0.1109	0.1146	0.0761
1	8	Likelihood	Root Mean Square Error	3.70E-05	0.142	0.0158
2	9	Likelihood	Root Mean Square Error	0.0194	0.1348	0.0159
3	10	Likelihood	Root Mean Square Error	0.00E+00	0.2835	0.0161
4	10	Likelihood	Root Mean Square Error	0.0122	0.2629	0.0161
5	10	Likelihood	Root Mean Square Error	0.031	0.3011	0.0161
6	10	Likelihood	Root Mean Square Error	0.0666	0.3011	0.0161
7	10	Likelihood	Root Mean Square Error	0.0171	0.3011	0.1217
8	10	Likelihood	Root Mean Square Error	0.1103	0.2198	0.1939
9	9	Likelihood	Root Mean Square Error	0.1937	0.1982	0.0214
10	8	Likelihood	Root Mean Square Error	0.014	0.263	0.0156
			Average	0.0468	0.2440	0.0467
			Standard Deviation	0.0580	0.0597	0.0601

4.1.5. Selección del sistema

Seleccionar el sistema de minería de datos a utilizar es una tarea no trivial que implica un análisis exhaustivo con la finalidad de encontrar el sistema que mejor clasifica, con la más alta probabilidad de predicción, con menores errores y que se acerque lo más posible a sistema ideal. El mejor sistema se determinó tomando los resultados arrojados por la matriz de confusión, el gráfico de elevación y la validación cruzada, la precisión y la sensibilidad; para mostrar lo tomado en cuenta para seleccionar el sistema se presentan los resultados en las siguientes tablas de valores. La tabla 4.6 muestra los resultados de la matriz de confusión, para los tres algoritmos que se analizaron, esto con la finalidad de hallar a los alumnos con más probabilidad de desertar de sus estudios a nivel superior. Para nuestro caso el de Ingeniería de Tecnologías de la Información y Comunicaciones del Instituto Tecnológico Superior de Misantla.

4.1.5.1. Resultados de la matriz de confusión

Se puede observar en la tabla 4.6, que el algoritmo de regresión logística obtuvo las mejores ponderaciones en las métricas: Precisión, Recall, Accuracy y F-Measure; Esto lleva a la primera conclusión de este trabajo de investigación, que utilizando los datos de la matriz de confusión, anteriormente descrita, el algoritmo que mejor clasificó a los alumnos en la categoría que deben estar es el sistema Regresión Logística, cabe mencionar que la matriz de confusión se alimentó del 70% de los datos de entrenamiento.

Tabla 4.6 Resultados de la matriz de confusión

	Precisión	Racall	Accuracy	F-Measure	Lift
Clúster	100%	75%	95%	86%	5
Árbol de Decisión	0%	0%	80%	0%	0
Regresión logística	100%	88%	98%	93%	5

4.1.5.2. Resultados del gráfico de elevación

En el gráfico de elevación Los sistemas de Regresión logística y Clúster son los que obtuvieron la mejor puntuación (Score) con el 100% y 89%, sin embargo el sistema Clúster fue el que obtuvo el más alto porcentaje de predicción; por otro lado, como se muestra en la tabla 4.7, (Gráfico de elevación), el sistema Regresión logística es el más ajustado al sistema ideal por lo tanto una vez más el mejor sistema es el de Regresión logística.

Tabla 4.7 Resultados del gráfico de elevación

	Score	Predict Probability
Clúster	89%	99.99%
Árbol de Decisión	90%	95.80%
Regresión logística	100%	98.52%

4.1.5.3. Resultados de la validación cruzada

Además de esto, se realizó la validación cruzada para corroborar que los resultados de los sistemas no se encontraran sesgados y de esta manera tener una métrica confiable y poder seleccionar el algoritmo a utilizar para este trabajo de investigación. De acuerdo a los resultados de las tablas 4.7 y 4.8, se eligió el algoritmo de Regresión Logística puesto que en la desviación estándar en las métricas Log Score y Lift fue el que mejor realizó la clasificación, dejando claro que su predicción es mejor con respecto a la predicción aleatoria (Log Score). Además, también mostró la mejor predicción real y probabilidad marginal en los casos de prueba (lift) y como en esta investigación lo que nos interesa es que clasifique correctamente a los alumnos desertores y el sistema Regresión Logística en la matriz de confusión y el gráfico elevado fue el que mejor clasificó, al analizar la validación cruzada el sistema de Regresión Logística y el del Clúster fueron los que mejor clasificaron alumnos desertores obteniendo la misma desviación estándar de 0.499 y ambos tuvieron la misma dispersión de datos al empatar en la clasificación de alumnos desertores como alumnos inscritos con una desviación estándar de 0.3083. Aunque los algoritmos de regresión logística y el de clúster, tienen mejor respuesta que los demás, y desde el punto de vista más práctico, el algoritmo de regresión legista es más efectivo y no solo por los resultados obtenidos en las diferentes métricas sino también por la naturaleza de los datos con los que se trabajaron en esta investigación el algoritmo de regresión legista se desempeña mejor que el resto de los algoritmos.

Por lo tanto, para llevar a cabo este trabajo de investigación se utilizó e implementó el algoritmo de Regresión Logística, con el objetivo de hallar a los alumnos con más probabilidades de desertar de sus estudios a nivel superior. Y se utilizará el mismo algoritmo para hallar un perfil de deserción.

Tabla 4.8 Resultados de la matriz de confusión (Standard Deviation)

	True Positive	False Positive	True Negative	False Negative	Log Score	Lift	Root Mean Square Error
Clúster	0.499	0.2942	0.4418	0.3083	0.0983	0.1109	0.058
Árbol de Decisión	0.4886	0.9395	0.9376	0.7393	0.1585	0.1146	0.0597
Regresión logística	0.4990	0.3083	0.4473	0.3083	0.0698	0.0761	0.0601

4.1.6. Resultados

4.1.6.1. Predicciones

En esta sección se utiliza el sistema de minería de datos basado en el algoritmo de Regresión Logística de Microsoft para predecir los posibles alumnos desertores y para obtener el perfil de deserción escolar. Estos algoritmos se aplicaron a los alumnos del primer semestre del programa educativo de ITICs del ITSM. Los algoritmos implementados no pueden operar para los 134 estudiantes en general, debido a que cada uno tiene información diferente. Por tal razón, se realizó con un conjunto de datos para alumnos del primer semestre. Esto es porque, según con datos reales que arroja el sistema de desarrollo institucional del mismo instituto, es en estos periodos donde más se nota la deserción, pues una cantidad de alumnos deja de asistir al instituto. Y, por otro lado, nos informan también que el más alto índice de deserción se da en el primer semestre.

4.1.6.1.1. Predicciones con datos de prueba

Ya teniendo resultados experimentales, esto es de acuerdo a la información real brindada por el mismo sistema de información del Instituto, se pueden comprar

con los datos obtenidos mediante la aplicación de los algoritmos antes descritos. Por lo tanto, para esta investigación se reportan los resultados del sistema de minería de datos, mismo que utilizó únicamente para entrenamiento los datos del primer semestre de la generación 2010, 2011, 2012 de la carrera de ITICs y posteriormente se implementó en el grupo de primer semestre de la misma carrera generación 2013. Antes de poner en producción el sistema fue probado con los mismos datos (generación 2010, 2011, 2012), esto sin especificar a los alumnos desertores, y el resultado fue que el sistema desarrollado en este trabajo de investigación predijo a 24 desertores, los mismos que ya habían desertado del ITSM, los datos de los alumnos se muestran en la Tabla 4.9. Para las otras generaciones de estudio sucedió lo mismo, en la mayoría el sistema desarrollado acertó entre un 90 y 100 % los alumnos desertores. Con esto, se puede afirmar que el sistema desarrollado tiene una confiabilidad aceptable y que es de gran ayuda para el instituto y cualquier otra institución de educación que desee conocer a sus estudiantes con altas probabilidades de deserción.

Tabla 4.9 Resultados de predicciones con datos de prueba

No	Porcentaje de deserción	Indicador de deserción	Matrícula del alumno	Nombre	Apellido paterno	Apellido materno	Edad
1	98.5507%	1	112T0360	ANGEL	BONILLA	HUERTA	25
2	98.5507%	1	112T0365	IVAN	FLORES	RUANO	22
3	98.5507%	1	112T0367	ARIANA	LOPEZ	ALVAREZ	22
4	98.5507%	1	102T0181	ABIGAILL	DOMINGUEZ	FRANZUA	25
5	98.5507%	1	102T0190	MARIA ANTONIA	HERNANDEZ	GOMEZ	23
6	98.5507%	1	102T0204	ALEX RAFAELL	MERINO	RAMOS	23
7	98.5507%	1	122T0302	ANDONI HERLICH	ZAVALETA	GARCIA	21
8	98.5507%	1	102T0616	JUAN NICOLAS	LANDA	POLO	28
9	98.5507%	1	102T0623	SARA GABRIELA	CASTRO	GONZALEZ	24
10	98.5507%	1	112T0347	JUAN ANTONIO	AGUILAR	VARELA	22
11	98.5507%	1	112T0348	JOSE ALBERTO	AGUILAR	ORTIZ	23
12	98.5507%	1	112T0359	RICARDO	RODRIGUEZ	RODRIGUEZ	31
13	98.5507%	1	112T0363	AARON GALDINO	PEREZ	AGUILAR	22
14	96.0928%	1	122T0289	MARTHA ELIANA	ESTEBAN	HUERTA	23
15	98.5507%	1	102T0316	EMMANUEL	MARTINEZ	MURRIETA	26
16	98.5507%	1	112T0368	NELLY JANET	HERNANDEZ	CASTILLO	24
17	98.5507%	1	102T0554	CARLOS ALFREDO	AGUILAR	ESCOBEDO	24
18	98.5507%	1	102T0555	LAURA ARLETTE	AGUIRRE	BAEZ	24
19	98.5507%	1	102T0562	ALICIA BERENICE	CASTELLANOS	ROJAS	23
20	98.5507%	1	102T0569	JOSE ANTONIO	FLORES	GOMEZ	23
21	98.5507%	1	102T0583	DAVID EDUARDO	HERNANDEZ	MENDEZ	29
22	98.5507%	1	102T0600	ALDO ANDRES	PEREZ	MORGADO	27
23	98.5507%	1	102T0606	JEIKOF	ROSADO	VAZQUEZ	24
24	98.5507%	1	102T0608	EFRE	SANCHEZ	BARRANCO	27

4.1.6.1.2. Predicciones reales

Otro ejemplo claro de este sistema, es que utilizó datos de la generación 2013, y donde muestra un resultado satisfactorio, ya que el sistema detectó a los 2 alumnos como posibles desertores de esa generación (ver tabla 4.11) y que, según información proporcionada por el departamento de control escolar a la fecha del reporte de esta investigación, los alumnos señalados como desertores por el sistema de minería de datos ya causaron baja de la institución. Por lo cual el sistema muestra un resultado satisfactorio. El sistema desarrollado en este trabajo arroja la matrícula, nombre y edad del posible alumno que va a desertar, por lo cual es fácilmente identificarlo.

Tabla 4.10 Resultados reales

No	Porcentaje de deserción	Indicador de deserción	Matrícula del alumno	Nombre	Apellido paterno	Apellido materno	Edad
1	100.0000%	1	132T0126	GUSTAVO	ORTIZ	MENDOZA	22
2	100.0000%	1	132T0132	JOEL	VASQUES	RUIZ	21

Finalmente, se puede decir que, aplicando el algoritmo de regresión logística, se puede predecir los alumnos que tienen índices altos deserción y más aún se puede tener un perfil de los alumnos que tienden a desertar de sus estudios universitarios. Este sistema desarrollado puede ser una gran herramienta para las autoridades de los diferentes planteles de educación superior de México, ya que, si se detecta con tiempo un alumno con alta probabilidades de deserción, se pueden implementar acciones que mantengan al estudiante y más aún si se tiene un perfil, se podrán implementar estrategias grupales que ayuden a disminuir los índices de deserción de los diferentes instituciones a nivel superior. Aunque, el sistema de educación actual ha implementado las tutorías, individuales y grupales, para el tutor resulta difícil identificar un alumno con altas posibilidades de deserción de ahí la necesidad de este sistema.

4.1.6.2. Perfil de deserción

Obtener el perfil de deserción es relativamente difícil, ya que el fenómeno del abandono escolar es multifactorial, los estudios indican como posibles causas las condiciones económicas, salud, orientación escolar, lugar de procedencia, asignaturas específicas, entre otros factores. Sin embargo, en el sistema desarrollado aquí, se logró obtener una tabla con datos específicos, como atributo que mayor influye en la deserción, el valor (lugar, escuela, etc.) y la probabilidad de deserción, ver tabla 4.12.

Tabla 4.11 Perfil de deserción

Atributo	Valor	Probabilidad
Ciudad de procedencia	LOC. LA REFORMA KM.9	100
Ciudad de procedencia	IGNACIO ALLENDE	80.94
Ciudad de procedencia	LOC. ARROYO HONDO	73.73
Ciudad de procedencia	RANCH. PIPIANALES	64.62
Municipio de procedencia	30009	55.17
Ciudad de procedencia	COLIPA	53.3
Municipio de procedencia	30197	46.96
Ciudad de procedencia	LOC. PLAN DE LA VEGA	43.37
Ciudad de procedencia	TENOCHTITLAN	36.9
Ciudad de procedencia	LAS LOMAS	35.28
Ciudad de procedencia	CONG. ARROYO NEGRO	34.86
Ciudad de procedencia	YECUATLA	34.62
Municipio de procedencia	30042	33.56
Taller de ética	0.308 - 62.322	33.48
Introducción a las tics	0.000 - 54.826	32.19
Fundamentos de programación	0.000 - 53.162	32.16
Fundamentos de investigación	0.000 - 55.367	32.11
Cálculo diferencial	0.000 - 47.185	30.51
Ciudad de procedencia	VILLA INDEPENDENCIA	27.76
Municipio de procedencia	30163	25.65
Matemáticas discretas I	0.000 - 60.341	25.61
Ciudad de procedencia	ALTO LUCERO	22.98
Ciudad de procedencia	FRANCISCO SARABIA	22.08
Ciudad de procedencia	LA GRAVERA	20.03
Ciudad de procedencia	MARTINEZ DE LA TORRE	17.71
Ciudad de procedencia	LOC. EL PORVENIR	16.36
Ciudad de procedencia	JUCHIQUE DE FERRER	12.65
Municipio de procedencia	30095	12.1
Escuela de procedencia	473.845 - 1,214.000	11.38
Ciudad de procedencia	LOC. BUENOS AIRES	10.01
Cálculo diferencial	47.185 - 69.064	7.34
Fundamentos de programación	53.162 - 73.638	7
Introducción a las tics	54.826 - 75.511	6.92
Fundamentos de investigación	55.367 - 75.000	6.66
Taller de ética	62.322 - 80.309	6.15
Ciudad de procedencia	LAS LAJAS	5.61
Ciudad de procedencia	CONGR. LAS PARCELAS	5.24
Ciudad de procedencia	LOC. INDEPENDENCIA	4.91
Matemáticas discretas I	60.341 - 78.362	4.79
Municipio de procedencia	30109	4.55
Edad del alumno	21.000 - 22.519	2.97
Municipio de procedencia	30102	2.56
Escuela de procedencia	222.106 - 473.845	2.3
Edad del alumno	22.519 - 24.426	1.06

Como se puede observar en la tabla 4.12, el perfil de deserción de 44 alumnos analizados se organizó de acuerdo a la probabilidad de deserción en forma descendente. En primer lugar, se muestra el alumno con la mayor probabilidad de deserción con un 100%, la cual muestra que la causa más influyente de su deserción es la ciudad de procedencia, el cual es LOC. LA REFORMA KM.9, así mismo destacan otras ciudades que tienen mayor influencia en la deserción escolar, siendo: loc. la reforma km.9, Ignacio Allende, loc. Arroyo Hondo, ranch. Pipianales, Colipa, loc. Plan de la vega, Tenochtitlan, las lomas, cong. Arroyo negro, Yecuatla, Villa independencia, Alto lucero, Francisco Sarabia, la gravera, Martínez de la Torre, loc. el porvenir, Juchique de Ferrer, loc. Buenos aires, las lajas, congr. las parcelas, loc. Independencia, así como sus respectivos municipios. Al mismo tiempo, también se analizó la influencia de Las escuelas de procedencia se discrizaron a través de la clave que control escolar les asignó; sin embargo, se detalla una lista completa de las escuelas en las tablas 7.1 y 7.2 que se encuentra en los anexos y que suman 150 escuelas, donde éstas tienen poca influencia en la deserción escolar. Las asignaturas que tienen mayor influencia en la deserción escolar son principalmente: Taller de ética, Introducción a las TICs y Fundamentos de programación, sin embargo, estas no son decisivas para la deserción.

Sin embargo, al registrar una combinación de una ciudad, un municipio, una escuela de procedencia, asignaturas que más influyen en la deserción, nuestro sistema arroja los alumnos con probabilidades más altas deserción, por lo que se puede identificar un perfil más claro. Si el alumno proviene de una ciudad, de cualquier otra escuela de procedencia, con bajos rendimientos en alguna asignatura, es más fácil identificar y aplicar estrategias de acción contra la deserción.

Capítulo V. Conclusiones y trabajo futuro

5.1. Conclusiones

Hasta antes de desarrollar este sistema, era difícil de detectar un alumno con altas probabilidades de deserción escolar. El mecanismo es manual, pues este mismo se realiza mediante un programa de tutorías, donde un grupo de alumnos son asignados a un profesor de tiempo completo, el cual tiene en un promedio de 25 estudiantes tutorados, y por más que el profesor realiza un esfuerzo para identificar a los alumnos con altas posibilidades, es realmente difícil lograrlo al 100%. Tal es el caso, que un profesor-tutor se da cuenta que el alumno ha desertado del programa educativo, cuando solicita su baja o definitivamente deja de asistir a la institución. El sistema desarrollado en este trabajo de investigación, pretende precisamente detectar de forma temprana y oportuna a los alumnos con altos índices de deserción. Para esto, se ha llevado a cabo un análisis y comparación de 3 algoritmos: el de árbol de decisión, regresión logística y el de clústers, en conjunto con técnicas de minería de datos y con el único objetivo, el de hallar a los alumnos con un mayor porcentaje de desertar de sus estudios. Sin embargo, los resultados muestran que el algoritmo con mejores resultados son el de regresión logística, con este algoritmo se pueden encontrar los posibles alumnos desertores; también se puede hallar un perfil de los estudiantes con las mismas probabilidades de deserción. Todos estos resultados fueron comparados con resultados experimentales que tiene el departamento de sistemas de información del Instituto Tecnológico Superior de Misantla. En este caso, sólo fue para el programa educativo de Ingeniería en Tecnologías de la Información y Comunicaciones del mismo instituto, sin embargo, sin ningún problema puede ser implementado para cualquier otro programa del Instituto o cualquier plantel de educación superior de México.

5.2. Trabajo Futuro

Como trabajo a futuro, se puede pensar en diseñar un test con variables significativas mediante una aplicación Web para recabar información demográfica, socio económica, de comportamiento y percepción de los alumnos a su ingreso al ITSM; el test servirá para desarrollar un sistema predictivo de deserción escolar en el ITSM más completo y así poder aplicar dicho sistema predictivo a los alumnos de nuevo ingreso al momento de notificarles su aceptación como alumnos del ITSM, es decir, antes de iniciar clases del primer semestre; de esta manera se podrá tener un repositorio completo que permita obtener una lista de los alumnos con alto porcentaje de deserción al momento de matricularse en esta institución y no esperar hasta que finalice el primer semestre, que es cómo funciona el sistema descrito en esta investigación. De esta manera se complementará y mejorará en gran medida el presente trabajo.

Referencias

- [1] Horacio Kuna, Ramón García Martínez And Francisco R. Villatoro, Pattern Discovery In University Students Desertion Based On Data Mining, Advances and Applications in Statistical Sciences, *Proceedings of The IV Meeting on Dynamics of Social and Economic Systems*, Vol. 2, Issue 2, Pages 275-285, 2010.
- [2] Guillermo López, María Posada, Claudia Cardozo, Diego José Cuartas, Specific actions for desertion reduction, competence identification and guidance for new students of an engineering program, a case study, *International Congress on Engineering Education (ICEED)*, 2010.
- [3] Jesús Alfonso Pérez Gama, Martha Isabel Rozo Arteaga; Roger Smith Londono Buritica ; Alejandro Marulanda Quinche, Quantitative models and software architecture, facing student Desertion and Permanence, *IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 2013.
- [4] Alfonso Pérez-Gama, Guillermo Hoyos, Leyini Parra-Espitia, Miguel Ortégón, Luis Giovanni Rozo-Pardo, Byron Perez-Gutierrez, Education software architecture: Facing student desertion in Colombia higher education with an intelligent knowledge based coaching system, *IEEE ANDESCON*, 2010.
- [5] Luis Felipe Zapata Rivera; Jorge Luis Restrepo Ochoa; Jaime L. Barbosa Perez, Improving student results in a statics course using a computer-based training and assessment system, IEEE, *Frontiers in Education Conference*, 2013.
- [6] Sandra Milena Merchan Rubiano; Jorge Alberto Duarte Garcia, Formulation of a predictive model for academic performance based on students academic and demographic data, *IEEE Frontiers in Education Conference (FIE)*, 2015.

- [7] John M. Mativo; Shaobo Huang, Prediction of students academic performance: Adapt a methodology of predictive modeling for a small sample size, *IEEE Frontiers in Education Conference (FIE)*, 2014.
- [8] Patrick D. Schalk; David P. Wick; Peter R. Turner; Michael W. Ramsdell, Predictive assessment of student performance for early strategic guidance, *IEEE Frontiers in Education Conference (FIE)*, 2011.
- [9] Setiono, R. ; Sch. of Comput., Nat. Univ. of Singapore, Singapore ; Azcarraga, A., An effective method for generating multiple linear regression rules from artificial neural networks, *IEEE 13th International Conference on Tools with Artificial Intelligence*, 2001.
- [10] D. D. Finlay, C. D. Nugent, P. J. McCullagh, N. D. Black, J. A. Lopez, Evaluation of a Statistical Prediction Model Used in the Design of Neural Network Based ECG Classifiers: A Multiple Linear Regression Approach, *Procc. of the 4th Annual IEEE Conf on Information Technology Applications in Biomedicine*, UK, 2003.
- [11] Zhi-cheng Zheng , Xin LIU, Analysis of regional logistics demand prediction based on support vector non-linear multiple regression, *International Conference on Management and Service Science*, 2009.
- [12] Amany Abdelhalim, Issa Traore, A New Method for Learning Decision Trees from Rules, *International Conference on Machine Learning and Applications*, 2009.
- [13] R. S. Michalski and I. F. Imam, "Learning problem-oriented decision structures from decision rules: the AQDT-2 system", *In Proceedings of 8th International Symposium Methodologies for Intelligent Systems. Lecture Notes in Artificial Intelligence*, 869, Springer Verlag, Heidelberg, 1994, pp. 416-426.
- [14] Y. Akiba, S. Kaneda, and H. Almuallim, "Turning majority voting classifiers into a single decision tree", *In Proceedings of the 10th IEEE International Conference on Tools with Artificial Intelligence*, 1998, pp. 224-230.

- [15] Masaki Kurematsu and Hamido Fujita, A Framework for Integrating a Decision Tree Learning Algorithm and Cluster Analysis, *12th IEEE International Conference on Intelligent Software Methodologies, Tools and Techniques*, Budapest, Hungar, September 22-24, 2013.
- [16] Saori Amanuma, Masaki Kurematsu and Hamido Fujita, "An Idea of Improvement Decision Tree Learning Using Cluster Analysis ", *The 11th International Conference on Software Methodologies, Tools and Techniques*, pp.351-360, 2012.
- [17] Quinlan, J. R. "Induction of Decision Trees", *Machine Learning*, Vol.1, No.1,pp.81-106(1986).
- [18] G. Bortolan, C. Brohet, S. Fusaro, "Possibilities of using neural networks for ECG classification," *Journal of Electrocardiology*, vol. 162, pp. 10-16, 1996.
- [19] Sam Chao, Fai Wong, An incremental decision tree learning methodology regarding attributes in medical data mining, *Procc. International Conference on Machine Learning and Cybernetics*, 2009.
- [20] P.E. Utgoff, N.C. Berkman, and J.A. Clouse, "Decision Tree Induction Based on Efficient Tree Restructuring", *Machine Learning*, Kluwer Academic Publishers, Vol. 29, pp. 5-44, 1997.
- [21] Y.L. Chen, C.L. Hsu, and S.C. Chou, "Constructing a Multi-Valued and Multi-Labeled Decision Tree", *Expert Systems with Applications*, Vol. 25, pp. 199-209, 2003.
- [22] I. Kononenko, "Inductive and Bayesian Learning in Medical Diagnosis", *Applied Artificial Intelligence*, Vol. 7, pp. 317-337, 1993.
- [23] Shen Bin, Liu Yuan, Wang Xiaoyi, Research on Data Mining Models for the Internet of Things, *International Conference on Image Analysis and Signal Processing (IASP)*, 2010.
- [24] http://www.sems.gob.mx/work/models/sems/Resource/11390/1/images/000_INTRODUCCION_Movimiento_contra_Abandono.pdf

- [25] H. Haggag, M. Hossny, S. Haggag, S. Nahavandi, D. Creighton, Efficacy comparison of clustering systems for limb detection, 9th International Conference on System of Systems Engineering (SOSE), 2014.
- [26] Lei Qiu; Yongqing Zheng; Yuliang Shi; Chengliang Sang, CET: Clustering Extension Table research in multi-tenant database for SaaS applications, International Conference on Information Science and Technology (ICIST), 2013.
- [27] D. Fetterly; M. Manasse; M. Najork, On the evolution of clusters of near-duplicate Web pages, Proceedings. First Latin American Web Congress, 2003.
- [28] Callejas Ivan; Pineros Juan; Rocha Juan; Hernandez Ferney; Delgado Fabio, Implementación de una red neuronal artificial tipo SOM en una FPGA para la resolución de trayectorias tipo laberinto, *II International Congress of Engineering Mechatronics and Automation (CIIMA)*, 2013.
- [29] J. A. Blakeley, C. Cunningham, N. Ellis, Balaji Rathakrishnan, M. -C. Wu Distributed/heterogeneous query processing in Microsoft SQL, Proceedings 21st International Conference on Data Engineering (ICDE) 2005.
- [30] Osamu Araki; Kazuyuki Aihara, Dual Information Representation with Stable Firing Rates and Chaotic Spatiotemporal Spike Patterns in a Neural Network Model, *Neural Computation*, Volume: 13, Issue: 12, 2001.

Anexos

7.1. Escuelas que contribuyen al perfil de deserción de los alumnos de la carrera de ITICs del ITSM

La lista de las escuelas que se muestran en las tablas 7.1 y 7.2 se filtraron de acuerdo a los municipios y ciudades de procedencia que marca el perfil de deserción en la tabla 4.12.

En la siguiente tabla se muestra una lista completa de las escuelas que contribuyen con un 11.38% al perfil de deserción, con un total de 157 escuelas.

Tabla 7.1 Escuelas de procedencia 1

Clave	Escuela de procedencia
1032	BACH. ÁLVARO GALVEZ Y FUENTES
1010	BACH. LIC. MARCO ANTONIO MUÑOZ
1045	BACHILLERATO "AGUSTÍN YAÑEZ"
1171	BACHILLERATO "GRAL. MANUEL RINCÓN"
1145	BACHILLERATO "LIC. MARCO ANTONIO MUÑOZ"
1115	BACHILLERATO "PAPANTECA"
1052	BACHILLERATO "POZA RICA"
1034	BACHILLERATO "VERACRUZ" SIST. ABIERTO (XALAPA)
1067	BACHILLERATO CATEMACO
1004	BACHILLERATO CHEDRAUI
1113	BACHILLERATO IVE
1092	BACHILLERATO LIC. BENITO JUÁREZ
1123	BACHILLERATO NAOLINCO DE VICTORIA
1089	BACHILLERATO PROFR. GABRIEL LUCIO
1083	BACHILLERATO XALAPA
1114	BACHILLERES ESTEBAN MORALES
1204	BACHILLERES "PASO DE OVEJAS"
1048	BACHILLERES DE MARTÍNEZ DE LA TORRE
1015	BACHILLERES DIURNA
1006	BACHILLERES NOCTURNA
1207	BACHILLERES OF. "PANUCO"
1002	BACHILLERES OFICIAL PAPANTECA
1017	BACHILLERES PATRIA
1158	BACHILLERES TAJÍN
1120	C.B.T.A. NO. 57

1088	CBTA NO. 99
1139	CBTIS 77
1194	CBTIS NO. 123
1100	CBTIS NO. 124
1001	CBTIS NO. 142
1008	CBTIS NO. 145
1212	CBTIS NO. 23
1149	CBTIS NO.134
1147	CECYT "COXQUIHUI" NO.7
1049	CECYTEV "AGUA DULCE"
1106	CECYTEV NO. 16
1119	CECYTEV EMILIANO ZAPATA
1156	CECYTEV NO. 15 EXT
1144	CECYTEV NO.15
1029	CENTRO DE ESTUDIOS "ALBERT EINSTEIN"
1163	CETIS 110
1154	CETIS NO. 002
1098	COBAEV – 21
1072	COBAEV NO. 22 CHICONTEPEC DE TEJADA
1125	COBAEV NO.13
1182	COBAEV NO.25
1134	COBAEV PLANTEL 27
1140	COBAEV PLANTEL 30 VILLA AZUETA
1198	COL. PREPARATORIO IVE
1080	COLEGIO DE ASIS JALAPA AC.
1129	COLEGIO DE BACHILLERES EMPRESAS TURÍSTICAS
1187	COLEGIO DE BACHILLERES (CLAVIJERO)
1091	COLEGIO DE BACHILLERES DEL ESTADO DE VERACRUZ
1066	COLEGIO DE BACHILLERES IZTACALCO
1019	COLEGIO PREPARATORIO PAPANTLA (IVEA)
1086	COLEGIO PREPARATORIO "IVEA" JOSÉ CARDEL
1084	COLEGIO PREPARATORIO "IVEA" SAN RAFAEL
1053	COLEGIO PREPARATORIO (IVEA) JUCHIQUE DE FERRER
1054	COLEGIO PREPARATORIO (IVEA) MARTÍNEZ DE LA TORRE
1195	COLEGIO PREPARATORIO DE ORIZABA
1087	COLEGIO PREPARATORIO IVEA MISANTLA
1035	COLEGIO PREPARATORIO IVEA (XALAPA)
1093	CONALEP (POZA RICA DE HGO.)
1094	CONALEP (POZA RICA DE HGO.)
1070	CONALEP ORIZABA 252
1133	CRYSTAL LAKE CENTRAL HIGH SCHOOL
1020	ESC. BACH. "JOSÉ MARTÍ"
1096	ESC. BACH. "LIC. BENITO JUÁREZ GARCÍA"
1056	ESC. BACH. HIDALGO, TIERRA BLANCA
1177	ESC. BACH. OFICIAL "DR. ALEJANDRO

	CERISOLA"
1058	ESC. BACHILLERES LIC. ANGEL CARVAJAL
1205	ESC. DE BACH. "CONSTITUCIÓN DE 1917 VESPERTINA"
1206	ESC. DE BACH. "LIC. ANGEL CARVAJAL"
1179	ESC. DE BACH. DIURNA "CONSTITUCIÓN DE 1917"
1060	ESC. DE BACH. JOSÉ PALACIOS ROJAS
1202	ESC. DE BACH. PART. "JOSÉ PALACIOS ROJAS"
1078	ESCUELA DE BACH. "BENITO JUÁREZ"
1063	ESCUELA DE BACHILLERES "JOSÉ VASCONCELOS"
1082	ESCUELA DE BACHILLERES "PROFR. GABRIEL LUCIO"
1185	ESCUELA DE BACHILLERES OFICIAL "PAPANTECA"
1197	INST, VER, DE EDUC, SUPERIOR APPP
1090	INST. TÉCNICO DEL GOLFO DE MÉXICO
1040	INST. TÉCNICO DEL GOLFO DE MÉXICO (BACHILLERATO)
1121	INST. VERACRUZANO DE EDUCACIÓN SUPERIOR A.P.P.P
1036	INSTITUTO "ANDERSEN" ,A,C.
1126	IVE APPP - XALAPA
1101	IVE MISANTLA
1122	IVE A.P.P.P.
1150	IVE APPP ACATLÁN
1180	IVE EMILIO CARRANZA
1132	IVE MARTÍNEZ DE LA TORRE
1186	IVEA (GUTIÉRREZ ZAMORA)
1105	IVEA TIERRA BLANCA
1151	IVEA XALAPA
1076	IVES APPP "YECUATLA"
1061	PREPARATORIA FED. DR. LUIS MORFIN ÁLVAREZ
1112	PREPARATORIA PAPANTECA
1050	TELEBACHILLERATO "EL JOBO"
1181	TELEBACHILLERATO EL FUERTE DE ANAYA
1007	TELEBACHILLERATO LEONA VICARIO
1097	TELEBACHILLERATO " PANCHO POZA"
1068	TELEBACHILLERATO "ANAYAL UNO"
1167	TELEBACHILLERATO "ATZALAN"
1153	TELEBACHILLERATO "CHICONQUIACO"
1028	TELEBACHILLERATO "CORRAL NUEVO"
1037	TELEBACHILLERATO "COYUTLA"
1193	TELEBACHILLERATO "EL AZOTAL"
1043	TELEBACHILLERATO "EL CIERVO"
1103	TELEBACHILLERATO "EL COLORADO"
1077	TELEBACHILLERATO "EL HUÉRFANO"

1142	TELEBACHILLERATO "EL REMOLINO"
1005	TELEBACHILLERATO "EL ZAPOTE"
1026	TELEBACHILLERATO "EMILIANO ZAPATA" ABIERTO
1075	TELEBACHILLERATO "ESPINAL"
1055	TELEBACHILLERATO "JILOTEPEC"
1152	TELEBACHILLERATO "LA GUADALUPE"
1022	TELEBACHILLERATO "LA PALMA"
1141	TELEBACHILLERATO "LAGUNA DE FARFÁN"
1175	TELEBACHILLERATO "LOMAS DE ARENA"
1176	TELEBACHILLERATO "LOMAS DE ARENA"
1039	TELEBACHILLERATO "LOS REYES"
1018	TELEBACHILLERATO "MACEDONIO ALONSO"
1021	TELEBACHILLERATO "MIAHUATLÁN"
1183	TELEBACHILLERATO "NAPOALA"
1148	TELEBACHILLERATO "ORILLA DEL MONTE"
1199	TELEBACHILLERATO "OTRA BANDA"
1191	TELEBACHILLERATO "PAHUA HUECA"
1102	TELEBACHILLERATO "PAPANTLA"
1104	TELEBACHILLERATO "PASO BLANCO"
1108	TELEBACHILLERATO "PASO DEL PROGRESO"
1025	TELEBACHILLERATO "PLAN DE LA VEGA"
1162	TELEBACHILLERATO "SALVADOR DÍAZ MIRÓN"
1116	TELEBACHILLERATO "SAN PEDRO ALTEPEPAN"
1130	TELEBACHILLERATO "SUCHILAPAN DEL RIO"
1201	TELEBACHILLERATO "TUZAMAPAN"
1143	TELEBACHILLERATO "URSULO GALVÁN"
1046	TELEBACHILLERATO "VALSEQUILLO"
1023	TELEBACHILLERATO "VEGA DE ALATORRE"
1081	TELEBACHILLERATO "VENUSTIANO CARRANZA"
1024	TELEBACHILLERATO "XIHUITLAN"
1038	TELEBACHILLERATO "ZANJAS DE ARENA"
1064	TELEBACHILLERATO "ZAPOTE REDONDO"
1031	TELEBACHILLERATO COACOATZINTLA
1071	TELEBACHILLERATO DOS ARROYOS
1192	TELEBACHILLERATO EL AZOTAL
1003	TELEBACHILLERATO EL PORVENIR 2
1030	TELEBACHILLERATO EMILIO CARRANZA
1085	TELEBACHILLERATO FRANCISCO I. MADERO
1016	TELEBACHILLERATO GUADALUPE VICTORIA
1168	TELEBACHILLERATO GUADALUPE VICTORIA
1208	TELEBACHILLERATO HUIPILTEPEC
1014	TELEBACHILLERATO LOS REYES
1117	TELEBACHILLERATO PALMA SOLA
1044	TELEBACHILLERATO PANCHO POZA
1000	TELEBACHILLERATO PLAN DELAS HAYAS
1073	TELEBACHILLERATO PUNTILLA ALDAMA
1213	UNIVERSIDAD DEL GOLFO DE MÉXICO

En la siguiente tabla se muestra una lista de las escuelas que contribuyen con un 2.3% al perfil de deserción, con un total de 2 escuelas.

Tabla 7.2 Escuelas de procedencia 2

Clave	Escuela de procedencia
246	COLEGIO VERACRUZ / MTZ. DE LA T
244	TELEBACHILLERATO LEONA VICARIO