UNIDAD ACADÉMICA DE INGENIERÍA ELÉCTRICA

INGENIERÍA DE SOFTWARE

TESIS DE LICENCIATURA

ANÁLISIS DE DESERCIÓN ESCOLAR EN SECUNDARIAS DE ZACATECAS UTILIZANDO TÉCNICAS DE MACHINE LEARNING

Tesis por José Guadalupe Bañuelos Vargas para obtener el grado de Ingeniero de Software en la Universidad Autónoma de Zacatecas

---

Asesorado por:
José Ramón Pasillas Díaz
Perla Velasco Elizondo

Abstract

Una bonita historia

Índice general

Índice de figuras

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

1.1. Contexto

Las ciencias computacionales así como las Tecnologías de la Información y Comunicación (TIC) tienen diferentes aplicaciones en diversas áreas de conocimiento; educación, entretenimiento o ciencia. Estas son sólo algunas de las muchas áreas de aplicación de las TIC y también son utilizadas para analizar grandes cantidades de información.

El almacenamiento y análisis de datos se convirtió en una tarea importante de los sistemas de información de todo tipo de organizaciones. Organizaciones de la "nueva economía", el e-commerce, telefonía, marketing, el sector educativo etc. Esta información almacenada se ha convertido en un tesoro para estas organizaciones, pues incluye todo tipo de transacciones, la contabilidad de sus procesos internos y la representación de lo valioso de cualquier organización. Pero almacenar estos datos no es suficiente ya que las acciones que se deben realizar sobre esos datos deben ser inteligentes para poder extraer y poder procesar dicha información, este, es el objetivo de la minería de datos. La minería de datos nos permite conocer información oculta en grandes cantidades de datos, información que es valiosa para quien la maneja.

Según Hand (1998), la minería de datos se define como el proceso de análisis secundario de grandes bases de datos dirigido a encontrar relaciones desapercibidas las cuales son de interés o valor para el o los dueños de las bases de datos.

Mena (1990) por su parte, la describe como el proceso iterativo de extracción de patrones predictivos ocultos de grandes bases de datos, usando tecnologías de inteligencia artificial así como técnicas de estadística.

La definición de Mena hace énfasis en cuáles son las raíces de la minería de

datos: la inteligencia articifial (Machine Learning) y la estadística. Machine Learning: Una rama de la inteligencia artificial que lidia con el diseño y aplicación de algoritmos de aprendizaje (Mena, 1990).

Estadística: Una rama de las matemáticas aplicadas y pueden ser categorizadas como matemáticas aplicadas a datos de observación. Las estadísticas pueden ser categorizadas como el estudio de población, estudio de variación o estudio de métodos de reducción de datos. (Fisher, 1925).

Hablar de minería de datos es hablar también de la estadística, ciencia que podemos dividir en dos tipos, una que se puede denominar Estadística Inferencial (modelado) y otra que podemos denominar Estadística Exploratoria (análisis de datos). Y es que a veces es difícil definir una línea entre los dominios de ambos tipos de estadística pero existe una diferencia conceptual que identifica cada tipo de estadística.

La estadística inferencial se refiere al problema del quehacer estadístico, por ejemplo, decidir entre distintas hipótesis a partir de los resultados observados.

Por otra parte, el análisis de datos trata la estadística aplicada de una forma muy general. En este enfoque la naturaleza aleatoria de los datos no se deja de lado, sino que primero se le da importancia a los datos que se están utilizando y es a partir de estos que se busca manifestar la información relevante para los problemas que se plantean. De esta manera se puede constatar que muchos de los problemas que el Análisis de datos aborda, son muy comunes con la inteligencia artificial.

En el ámbito académico sucede que se han desarrollado de manera contraria una a la otra, dando lugar a nomenclaturas totalmente diferentes para problemas iguales. Se puede concluir que la inteligencia artificial se preocupa más en ofrecer soluciones algorítmicas en relación a un costo de recursos computacionales aceptables, mientras que la estadística se ha preocupado más de poder generalizar los resultados obtenidos.

Por otro lado, el fenómeno del abandono y la deserción escolar es una problemática que cada vez se hace más aguda y compleja y ciertos contextos de la región de Zacatecas, creando barreras sociales, pedagógicas y culturales hacia el anhelo de crear una sociedad más justa, igualitaria y competente. Es evidente que uno de los principales desafíos para avanzar hacia esa dirección, es identificar y actuar sobre los factores que influyen directamente sobre este problema y así evitar que los estudiantes abandonen la sus estudios antes del término de su preparación básica y por lo menos identificar estos factores que hacen que los jóvenes deserten durante su educación secundaria.

Por ende, resulta indispensable no sólo conocer cuántos estudiantes abandonan su educación secundaria, sino que comprender los factores que los han llevado a elegir abandonar su proceso formativo a pesar de las consecuencias

que su decisión implica. Hablar de este problema orienta a las instituciones correspondientes y crea oportunidades para mejorar la capacidad de retención de los sistemas educativos.

1.2. Problemática

La deserción se ubica entre los problemas más frecuentes y de mayor complejidad a los que las instituciones de educación se enfrentan, destacando que en la trayectoria estudiantil, se sigue un proceso durante el cual, los estudiantes están de cierto modo sometidos a un conjunto de reglas comunes, las cuales les permiten progresar en su formación académica.

Entonces, los estudiantes tienden a adoptar comportamientos que afectan la continuidad de sus estudios, comportamientos que se caracterizan por el abandono de los estudiantes debido a deficiencias académicas, abandono debido a instalaciones deficientes de las instituciones, o el cambio de institución.

En México, la Constitución mexicana dicta en el artículo 3 que todo ciudadano tiene derecho a recibir educación básica, comprendida por tres años de preescolar, seis de primaria y tres de secundaria.

No hacer valer el derecho a la educación básica, tiene muchas consecuencias negativas para quienes sufren este problema, pues disminuyen sus oportunidades de obtener un trabajo formal, sufren desigualdad social, discriminación y mayor posibilidad de incidir en conductas delictivas (Carbonell, 2005).

Según el Instituto Nacional para la Evaluación de la Educación (INEE, 2005), se ha detectado mayor deserción en zonas rurales y en secundarias que se ubican en contextos socioeconómicos pobres (Escudero, 2005), la baja escolaridad de los padres (Saucedo, 2001), bajas expectativas dentro de la misma familia (Cueto, 2004) e incluso, hay casos en lo que los mismos padres hacen acciones directas para que sus hijos abandonen sus estudios para dedicarse a otras tareas (Blasco, 2003 y Parker, 2004).

1.3. Objetivos

1.3.1. Objetivos Generales

Desarrollar una herramienta que utilice técnicas y algoritmos de la minería de datos para poder descubrir factores de deserción en las escuelas secundarias de Zacatecas.

1.3.2. Objetivos Específicos

1.4. Estructura de la Tesis

Capítulo 2

Marco Teórico

2.1. Deserción Escolar

El término deserción frecuentemente se asocia con fracaso y abandono escolar. Aquí la deserción se entiende como sinónimo de abandono escolar y se diferencia de fracaso por ser parte de este. La deserción comprende a estudiantes que no concluyen sus estudios en el tiempo que fue previsto para ello y no fueron matriculados en otro plantel

Rumberger(2001) plantea que la deserción incluye a la personas que no tienen completo cierto nivel de estudios y no estan inscritos en escuelas que permitan terminar dicho nivel de estudios en un tiempo determinado.

El INEE(2005) lo entiende como el porcentaje de alumnos que abandonan sus actividades escolares antes de concluir el nivel educativo o un ciclo escolar determinado.

La Tasa de Deserción Escolar es el número estimado de alumnos que abandonan la escuela entre ciclos escolares consecutivos antes de concluir el nivel educativo de referencia, por cada 100 alumnos matriculados al inicio del ciclo escolar (INEE, 2008).

Por ejemplo, la tasa de deserción total en el estado de Aguascalientes en el periodo escolar 2008-2009 para las escuelas secundarias fue de 6.3 % siendo más alta en hombres con un 7.5 % y en mueres 5.2 %.

A nivel nacional, la Tasa de Deserción Escolar en escuelas secundarias en el periodo 2008-2009 fue de 6.4 %, en hombres fue de 7.6 % y en mujeres 5.2 %

Según la UNICEF, el Gobierno Federal de México y la INEE, las principales causas de a deserción escolar son la falta de recursos econónmicos y embarazo en adolescentes. También se menciona que en México de cada 100 niños que ingresan a preescolar, solo el 98 % termina su educación primaria

y de este 98 %, solo el 75 % termina su educación secundaria. Otras causas de la deserción escolar son el dificil acceso a la escuela y las pesimas condiciones escolares, la motivación y satisfacción personal, escuelas con recursos limitados y niños que trabajan por falta de recursos. La idea de que el trabajo es más importante que estudiar y la falta de transporte escolar también obstaculizan el progreso y la educación.

2.2. Minería de Datos

La minería de datos se ha definido como .[El] proceso de descubrir conocimiento interesante de grandes cantidades de datos almacenadas en bases de datos, data warehouses u otro repositorio de información"(Jiawei Han, Micheline Kamber 2001).

La minería de datos nació con la necesidad de automatizar aplicaciones y procesos en instituciones y organizaciones, también con la necesidad de desarrollar nuevas herramientas y técnicas para analizar el almacenaje masivo de información. Es parte del proceso de descubrir conocimiento a partir de grandes cantidades de información.

La minería de datos utiliza métodos badasos en tecnología de Bases de Datos, estadística, el aprendizaje automático (machine learning), reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, análisis de datos espaciales (Antonio Gonzáles, 2014).

¿Para qué se utiliza la minería de datos? Existen dos tareas principales en las cuales pueden actuar estos métodos de la minería de datos, la primera es predecir, en esta tarea se utilizan algunas variables en una base de datos para predecir los valores desconocidos de otras variables. La segunda tarea es describir, que consiste en encontrar patrones que puedan describir la información que estamos leyendo. Estas tareas se pueden realizar con el apoyo de técnicas que la minería de datos utiliza, algunas de estas son:

- Clasificación: Habilidad para adquirir una función que clasifique un elemento de dato a una de varias clases o atributos ya definidas.
- Regresión: Consiste en adquirir una función que clasifique un elemento a una variable de predicción de valor real .
- Agrupamiento o clustering: Es una tarea que busca identificar un conjunto de de categorías que permitan describir los datos.

- Sumarización (describir clases o conceptos): Consiste en encontrar un método que permita encontrar una descripción compacta de un conjunto o subconjunto de datos.

- Modelos de dependencias: Consiste en encontrar un model que permita describir las dependencias significantes dentre clases o variables.

- Detección de cambios o desvíos: Consiste en detectar los cambios más significantes en los datos en comparación a valores de datos anteriores, o valores que se consideren normales.

Estas técnicas nos permiten crear modelos predictivos o descriptivos. Los modelos predictivos responder preguntas sobre datos que no conocemos, por otro lado, los modelos descriptivos proporcionan información sobre las relaciones que tienen los datos y sus características.

2.3. Estudios Actuales (por cambiar)

Hay muchos estudios sobre la deserción escolar que abarcan aspectos psicológicos y socioeconómicos, pero no son suficientes ya que, son muy pocos los que involucran la minería de datos.

En la Universidad del Zulia, en Maracaibo, Venezuela, se llevó a cabo una investigación por parte de estudiantes de la licenciatura en computación, en la cual, el objetivo era obtener patrones de similitud entre estudiantes que no habían podido concluir sus estudios universitarios. Dicho estudio fue hecho bajo la metodología computacional Crisp-DM (Cross Industry Standard Process for Data Mining, por sus siglas en inglés), el cual es un modelo de proceso de minería de datos que describe los enfoques que abarca la minería de datos y con apoyo de Weka, un software libre escrito en el lenguaje de programación Java que se especializa en el aprendizaje automático y la minería de datos. Los datos fueron proporcionados por 3 fuentes: estudiantes de entre primer y tercer semestre, profesores de entre primer y quinto semestre y reportes de matrícula proporcionados por Control de Estudios que abarcaban los años 2008-2011. Con dichos datos se construyó un modelo computacional para predecir la deserción escolar empleando árboles de decisión C4.5 y el algoritmo de los k vecinos más cercanos (kNN). Entre los factores que tomaron para este análisis, se encuentran las siguientes preguntas:

- Si el alumno sabe el perfil que se necesita para ingresar a la licenciatura en computación

- Si cree que su educación media superior era suficiente para cursar dicha licenciatura
- La dedicación de horas al estudio
- Aspectos socioeconómicos
- Lugar de origen de los estudiantes
- Si trabajan
- Qué otras actividades realizan en su tiempo libre
- Si tienen buena o mala relación con sus profesores y compañeros de estudio.

El resultado de este ejercicio, demostró que las causas que hacen que los esudiantes abandonen sus estudios fueron:

- Los conocimientos adquiridos durante la educación media superior no fue sucifiente para cursar la licenciatura en computación por escasas competencias en matemáticas y lógica.
- La falta de recursos económicos para comprar equipo de cómputo y poder practicar fuera de clases
- La falta de concentración por parte de los estudiantes y las pocas horas dedicadas al estudio.

El modelo predictivo planteado permitió a los estudiantes de dicha licenciatura arrojar resultados en cuanto a cuáles son las causas de la deserción de sus compañeros. Después de reunir, integrar, seleccionar y limpiar dichos datos, compararon los resultados con los resultados que arrojaron los cuestionarios que aplicaron a las unidades seleccionadas. El estudio permitió afirmar que las técnicas de Minería de Datos proporcionan una herramienta muy útil que permite establecer una aproximación a las causas de la deserción escolar.

En la Universidad Tecnológica de Izúcar de Matamoros, se aplicó un estudio que utiliza técnicas de mineria de datos en base a datos proporcionados por el departamento de servicios escolares de dicha institución dentro de los cuales se incluyen bases de datos del EXANI-II, de alumnos inscritos, de alumnos que causaron baja y sus causas. En total, se recopilaron 11 bases de datos de todos los EXANI desde el año 2003 hasta el año 2008, 6 bases de datos de los alumnos inscritos y los memorandums de los alumnos que causaron baja y sus motivos. El primer paso fue en análisis de los tados que

se obtuvieron y para poder seleccionar los datos útiles para la investigación, se llevó a cabo el proceso de Extracción, Transformación y Carga de datos (ETL por sus siglas en inglés), proceso que permitió crear un modelo predictivo de calidad. Dentro de los datos que ofrecieron las bases de datos del examen CENEVAL, se seleccionaron los siguentes atributos:

- Año
- Si trabaja
- Las horas que trabaja
- Qué tipo de trabajo es
- El tipo de organización en la que trabaja
- El trabajo que desarrolla
- El ingreso personal

En el caso de la información de los estudiantes, se recurrió a homogeneizar dichos datos, pues estaban almacenados de distintas maneras y esto causa problemas en algúnas técnicas de la minería de datos. Al final, una vez limpios los datos y capturada la información de los alumnos que causaron baja (matricula, nombre y generación), se obtuvo un primer modelo con 16 atributos, los cuales fueron:

- Sexo
- Edad
- Bachillerato
- Promedio del Bachillerato
- Materias Reprobadas
- Intentos Previos
- Apoyo Económico
- Nivel del Inglés
- Habilidades en el Estudio
- Puntaje de Exani

- Escolaridad de la madre
- Escolaridad del padre
- Ingreso Familiar
- Tamaño de Familia
- Si trabaja
- Horas de trabajo
- Ha causado baja

Para este estudio, se optó por la técnica de clasificación utilizando árboles de desición mediante el algoritmo C4.5 y el algoritmo de los k vecinos más cercanos (kNN). Se crearon varios árboles de prueba para ver qué tan variados eran los resultados, así como un segundo modelo con kNN y se compararon ambos resultados, después se utilizó la herramienta Weka para generar los modelos predictivos. En total fueron 723 registros de alumnos, de los cuales Weka tomó 477 (66.6 %) para crear el modelo y 245 registros (33.4 % restante) para probarlo, con una precisión del 67.07 %. El segundo modelo se ejecutó con un k = 50 con lo que se obtuvo una precisión de 67.77 %, superior al primer modelo, aunque se optó por trabajar con la estructura de árboles de decisión por tener un nivel más alto de confiabilidad al trabajar con cantidades más grandes de datos. Los resultados del estudio arrojaron que la edad es un factor muy importante para la deserción, pues tiene que ver con la madurez y perspectiva de futuro de los estudiantes. Los ingresos familiares también representan un factor muy importante, pues muchos alumnos aún dependían de sus padres para costear su educación y por último el nivel de inglés para aquellos alumnos mayores de 18 años. Con este estudio, se pudo obtener una manera de determinar qué alumnos son candidatos a desertar.

Para poder comprobar estas afirmaciones, se compararon 2 escuelas con condiciones sociales diferentes. Uno de ellos, el Colegio Marista, A.C no registró deserción. Sus alumnos pertenecen a un estrato social medio, medio-alto y solo se presentaron cambios de institución por razones de trabajo de los padres, becas al extranjero o por mala conducta y reprobación, pero continuaron con sus estudios en otras instituciones.

La segunda institución fue el Telebachillerato Emiliano Zapata en el municiop de Pabellón de Arteaga, donde coinciden los datos estadísticos nacionales antes mencionados cn respecto a la deserción ya que de una matrícula escolar de inicio de 45 alumnos, solo 20 terminan sus estudios, siendo las principales causas de deserción, la escasez económica o falta de recursos para continuar con sus estudios.

2.4. Limitaciones de los estudios actuales (por cambiar)

Capítulo 3

Metodología del Proyecto

3.1. Selección de los datos

3.2. Pre-procesamiento de los datos

3.3. Análisis de los datos

3.4. Interpretación de los datos

Capítulo 4

Predicción de deserción usando árboles de decisión

4.1. Secciones no definidas aun

Capítulo 5

Discusión General

5.1. Secciones no definidas aun

Capítulo 6

Conclusiones y trabajos futuros

6.1. Secciones no definidas aun

Apéndice A

Más cosas

Aún faltan cosas por decir.

Apéndice B

Y más cosas aún

Y más cosas aún.

Bibliografía

[1] Aluja, T. *La Minería de Datos, entre la Estadística y la Inteligencia Artificial*. PhD thesis, 2001.

[2] Beltrán Martínez, M. B. Minería de datos. 67.

[3] BOSQUE, L. M. S., GUTIÉRREZ, L. C., PADILLA, E. A. L., and LÓPEZ, R. E. L. Deserción escolar en méxico.

[4] Marcano, Y. J., and Rodríguez, R. Mineria de datos aplicada a la desercion estudiantil. *Educare 18* (2014), 31–51.

[5] Rodallegas Ramos Erika, Torres González Areli ,. Gaona Couto Beatriz B, Gastelloú Hernández Erick, Lezama Morale Rafael As, V. O. S. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. *Recursos Digitales* (2010), 33–39.

[6] Román, M. Abandono y Deserción Escolar: duras evidencias de la incapacidad de retención de los sistemas y de su porfiada inequidad. *Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación (RIECE) 7*, 4 (2009), 3–9.

[7] Social, T., and Aplicades, S. Delgado, G. A. (2011) "Condiciones escolares asociadas a la deserción en educación secundaria. Análisis a partir de dos casos en México". 89–111.