Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos

Analysis of Student desertion at Universidad Simón Bolívar, Faculty of Systems Engineering, with data mining techniques

Análise de abandono na Universidade Simon Bolivar, da Faculdade de Engenharia de Sistemas, com técnicas de mineração de dados.

Cómo referenciar este artículo:
Azoumana, K. (2013). Análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad Ingeniería de Sistemas, con técnicas de minería de datos. *Pensamiento Americano,* 41-51

Kamagate Azoumana*
kazoumana@unisimonbolivar.edu.co

Resumen
El presente artículo muestra los resultados del análisis de la deserción estudiantil en la Universidad Simón Bolívar, facultad de Ingeniería de Sistemas, con técnicas de minería de Datos. Se utilizó la herramienta Weka agrupando las causas de la deserción en 5 variables que son:
Pérdida de semestre, Dificultad financiera, Ingreso al mercado laboral
Otros intereses atraen al estudiante, Indeterminado. La muestra era de 707 sujetos entre los períodos 2007-2012. Se concluye que la causa mayor de la deserción es el factor indeterminado.

Palabras clave
Minería de datos, deserción estudiantil, Weka, ingeniería de sistemas.

Abstract
This article shows the results of the student desertion analysis at the Simon Bolivar University, Faculty of Systems Engineering with data mining techniques. Grouping the causes of desertion in five variables:
Loss of semester, Financial Difficulty, Access to job market, Others interests, Undetermined
A Sample of 707student was choose from 2007-2012. Using Weka to process these datas, the conclusion is that the major cause of student desertion is the factor undetermined.

Key words
Data Mining, student desertion, weka. Systemsengineering.

Resumo
Este artigo apresenta os resultados da análise de abandono estudante na Universidade Simon Bolívar, da Faculdade de Engenharia de Sistemas com técnicas de mineração de dados. Nós usamos a ferramenta Weka agrupar as causas de atrito em cinco variáveis:, A perda de metade, Dificuldade financeira, Junte-se à força de trabalho, Outros apelamaos interesses dos alunos
indeterminado, Exemplo 707 foi sujeito a períodos de 2007-2012. Nós concluímos que a principal causa da redução é o factor indeterminado.

Palavras-chave
Mineração de dados, abandono estudanteWeka, engenharia de sistemas.

* Docente Investigador en Ingeniería de Sistemas, grupo de Investigación Ingebiocaribe , Universidad Simón Bolívar Barranquilla.
Artículo recibido: Octubre 16/2012. Aceptado: Febrero 28/2013.

Introducción

La Universidad Pedagógica Nacional (UPN 2004) define la deserción estudiantil como el hecho de que un número de estudiantes matriculados no siga la trayectoria normal del programa académico, bien sea por retirarse de ella o por demorar más tiempo del previsto en finalizarla, por repetir cursos o por retiros temporales. El abandono o la interrupción pueden ser voluntarios o forzados. También puede presentarse cambió de carrera dentro de la misma institución o cambio de institución donde puede continuar con la misma carrera o con otra (U.P.N).

Desde el punto de vista institucional todos los estudiantes que abandonan una institución de educación superior pueden ser clasificados como desertores; en este sentido, muchos autores asocian la deserción con los fenómenos de bajo rendimiento académico y retiro forzoso. Así, cada estudiante que abandona la institución crea un lugar vacante en el conjunto estudiantil que pudo ser ocupado por otro alumno que persistiera en los estudios (Guzmán, et al, 2009).

Para predecir la probabilidad de que un estudiante abandone la institución educativa se han utilizado técnicas de minería de datos para lograr el objetivo; entre ellos tenemos a (Kuna, García y Villatoro, 2010) realizaron un trabajo basado en el uso del conocimiento, en reglas de descubrimiento y en el enfoque TDIDT (Top Down Induction of Decision Trees) sobre la base de datos de la gestión académica del consorcio SIU de Argentina (que reúne 33 universidades de Argentina), lo cual permite un interesante análisis para encontrar las reglas de conducta que contienen variables de ausencia. como los estudiantes financian estudios universitarios, el número de años desde el final de la escuela secundaria de acceso a la universidad.

Para poder obtener este conocimiento es necesario partir de la materia prima, que son los datos, los cuales se encuentran disponibles en gran cantidad gracias a las tecnologías de información y las comunicaciones. Estos datos por lo general se encuentran en forma no refinada y para poder analizarlos con fiabilidad es necesario que exista una cierta estructuración y coherencia entre los mismos.

Para realizar un análisis en profundidad de forma automática, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos, sin embargo, es la transformación de los datos en conocimiento y la aplicación de éste, lo que genera valor para una organización. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista.

Para las organizaciones que realizan algún tipo de mercado, el conocimiento es algo imprescindible para tener éxito, por tanto encontrar asociaciones o correlaciones interesantes en los registros de las transacciones de negocios puede ayudar a la toma de decisiones. En este entorno la minería de datos ofrece la posibilidad de llevar a cabo un proceso de descubrimiento de información automático.

Para brindar una solución acorde a las necesidades de las empresas es necesario entender los objetivos y requerimientos desde la perspectiva de lo que se busca, convirtiendo entonces este conocimiento en la definición de un problema de minería de datos, ya que dependiendo del problema de información que se desea solucionar, existe una serie de técnicas que son aplicadas en la solución de diversos problemas. De esta situación surge, entonces, la pregunta que guía este trabajo de investigación:

¿Las técnicas de Data Mining, permitirán determinar las causas de deserción estudiantil utilizando los datos históricos del programa de Ingeniería de Sistemas?

A continuación, se hablará del proceso de descubrimiento de conocimiento, el concepto de minería de datos, las operaciones de minería de datos, una descripción general de las técnicas de minería de datos más usadas, los criterios de selección del algoritmo escogido y finalmente, las conclusiones obtenidas.

El descubrimiento de conocimiento

La minería de datos es, en principio, una fase dentro de un proceso global denominado descubrimiento de conocimiento en base de datos (Knowledge Discovery in Database o KDD), aunque generalmente se asocia el concepto de minería de datos a todo el proceso, en lugar de la

fase de extracción de conocimiento. El proceso de KDD es útil en la deserción estudiantil universitaria para obtener conocimientos de las causales generales de ausencia, estudio de variables influyentes según el contexto evaluado, entre otros. KDD se constituye de varias etapas que se ejecutan interactivamente. El proceso es no trivial porque incluye acciones de cierta complejidad que involucran la búsqueda de estructuras, modelos y parámetros en la base de datos. Los patrones que se obtienen deben ser válidos con algún grado de certeza, preferiblemente novedosos para el usuario, al que deberán reportar algún tipo de utilidad.

El proceso KDD (ver figura 1) comienza con la definición y comprensión de un determinado problema y termina con el análisis de los resultados. Una de las propuestas más ampliamente extendida sobre las etapas o fases componentes del proceso KDD incluye la comprensión del problema, la selección de datos, su limpieza y pre procesamiento, la transformación y aplicación del método de descubrimiento (minería de datos) a utilizar y la interpretación de los patrones obtenidos o análisis de resultados.

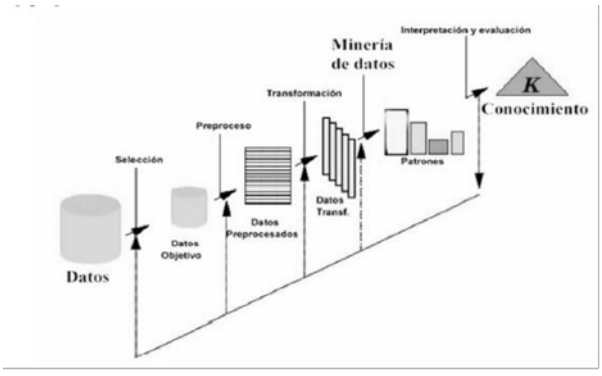

Figura 1. Proceso KDD

Minería de datos

La minería de datos es una fase dentro del KDD y se define como "el proceso de extracción de información previamente desconocida, válida y útil de grandes bases de datos y el uso de la información para tomar decisiones cruciales de negocios En pocas palabras, la minería de datos se refiere a la extracción o el conocimiento "minería" de grandes cantidades de datos. El término es en realidad un nombre inapropiado. Recuerde que la minería de oro de las rocas o la arena se conoce como la minería de oro en vez de roca o la extracción de arena. Por lo tanto, sería más apropiado el nombre de "minería de conocimiento a partir de los datos", que por desgracia es un poco largo. "La minería del Conocimiento", un término más corto, puede no reflejar el énfasis en la minería de grandes cantidades de datos. Sin embargo, la minería es un término vivo que caracteriza el proceso donde se encuentra un pequeño conjunto de pepitas de oro precioso de una gran cantidad de materia prima. Así, por ejemplo, dos términos equivocados como "datos" y "mineros" se convirtieron en una opción popular. Muchos otros términos tienen un significado similar o ligeramente diferentes a la minería de datos, como la extracción de conocimiento a partir de los datos, la extracción de conocimientos, datos o análisis de patrones, la arqueología de datos, y la filtración de información.

Muchas personas tratan la minería de datos como un sinónimo del término Descubrimiento de Conocimiento a partir de datos o KDD. Otros consideran que la minería de datos es simplemente un paso esencial en el proceso de descubrimiento de conocimiento que consta de una secuencia interactiva de pasos. La minería de datos emplea una serie de técnicas las cuales son aplicadas para la solución de diversos problemas, sus herramientas predicen futuras tendencias y comportamientos, permitiendo tomar decisiones conducidas por un conocimiento obtenido de los datos. Para conseguir esto hace uso de diferentes tecnologías que resuelven problemas típicos de agrupamiento automático, clasificación, asociación de atributos y detección de patrones secuenciales.

Técnicas de mineria de datos

La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos por un enfoque de análisis de datos.

La principal diferencia entre ambos se encuentra en que en el último se descubre 1a información sin necesidad de formular previamente una hipótesis. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos razón por la cual esta técnica es mucho más eficiente que el análisis

dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada. Dichas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones. Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Weiss y Indurkhya, 1998).

Los algoritmos predicen el valor de un atributo de un conjunto de datos, conocidos otros atributos. A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se desarrolla en dos fases: Entrenamiento (construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva. En ese caso hay que recurrir a los métodos que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. En la tabla siguiente se muestran alguna de las técnicas de minería de ambas categorías.

Clasificación de las técnicas de Minería de datos

Técnicas	Predictivas	*Regresión*	Descubrimiento
		Análisis de la varianza y Covarianza	
		Métodos bayesianos	
		Algoritmos genéticos	
		Arboles de decisión	
		Redes neuronales	
	Descriptivas	Asociación	
		Dependencia	
		Reducción de la dimensión	

Tabla 1
Clasificación de las técnicas de Minería de datos

Métodos descriptivos

Es un modelo que da una mejor comprensión de los datos, sin ninguna variable específica como objetivo único. Generalmente utiliza técnicas descriptivas que incluyen análisis de factores (para extraer dimensiones subyacentes de datos multivariados), análisis de clúster (de agrupación de una base de datos de clientes en segmentos) y el análisis de asociación (para descubrir relaciones entre los elementos tales como productos de venta al por menor) (Leventhal, 2010), que son orientados a la interpretación de datos y se enfocan en entender la forma en que los datos subyacentes se refieren a sus partes.
Las herramientas de Data Mining barren las bases de datos e identifican modelos previamente escondidos en un sólo paso. Otros problemas de descubrimiento de modelos, incluyen detectar transacciones fraudulentas de tarjetas de crédito e identificar datos.

Métodos predictivos

Son los que pretenden construir un modelo de comportamiento, el cual obtiene nuevas y ocultas muestras y es capaz de predecir valores de una o más variables relacionadas con la muestra.

Varias de las técnicas orientadas al descubrimiento están basadas en aprendizaje inductivo, en donde un modelo es construido, explícita o implícitamente, mediante la generalización de una cantidad suficiente de muestras de entrenamiento. El supuesto subyacente del enfoque inductivo es que el modelo entrenado es aplicable a futuras muestras ocultas. Los métodos de verificación incluyen las técnicas comunes de la estadística tradicional, como la bondad de ajuste, escala y técnicas de predicción capaces de manejar grandes residentes en una base de datos. Hay diversas técnicas básicas para la clasificación de datos, tales como la forma de construcción, clasificadores de árboles de decisión, clasificadores bayesianos, redes bayesianas, las creencias y los clasificadores basados en reglas. En el campo de la inteligencia artificial, se emplea frecuentemente la técnica de la Retropropagación (una técnica de red neuronal), además de un enfoque más reciente de clasificación conocido como máquinas de vectores, se basa en la prueba de hipótesis y el análisis de varianza. Estos métodos no se consideran pro-

pios de la Minería de datos, pues el objetivo de ésta es descubrir una hipótesis, en lugar de probar una duda.

Los métodos predictivos también se conocen como Aprendizaje Supervisado, los cuales intentan descubrir relaciones entre atributos de entrada (llamados Variables Independientes) y un atributo objetivo (Variable Dependiente). La estructura descubierta es representada como un modelo. Un modelo describe y explica un fenómeno, el cual está oculto entre un conjunto de datos y puede ser usado para predecir el valor del atributo objetivo conociendo los valores de los atributos de entrada. Los métodos supervisados pueden ser de dos tipos: Modelos de Clasificación y Modelos de Regresión.

Los métodos de Aprendizaje No Supervisado, por su parte, agrupan instancias sin un atributo dependiente pre-especificado; un ejemplo de este tipo de métodos es el clustering ("agrupamiento") (Maimon y Rokach, 2005).

Método de predicción, Clasificación:

Las Bases de datos son ricas en información oculta que puede ser utilizada para la toma de decisiones inteligente. La Clasificación y la Predicción son dos formas de análisis de datos que pueden ser utilizados para extraer los modelos que describen clases de datos importantes o para predecir las tendencias futuras de datos. Este análisis puede ayudar a proporcionarnos una mejor comprensión de los datos en general. Mientras que la clasificación predice categórico (discreto, sin ordenar) las etiquetas, los modelos de predicción funciones continuas valoradas. Por ejemplo, podemos construir un modelo de clasificación para categorizar las aplicaciones de préstamos bancarios, ya sea como seguro o peligroso, o un modelo de predicción para predecir los gastos en dólares de los clientes potenciales en los equipos informáticos debido a su ingreso y la ocupación. Muchos métodos de clasificación y predicción han sido propuestos por los investigadores de aprendizaje automático, reconocimiento de patrones, y estadística. La mayoría de los algoritmos son residentes en memoria, por lo general asumiendo un tamaño pequeño de datos. La investigación reciente de minería de datos se ha basado en dicho trabajo, el desarrollo de la clasificación escalable y técnicas de predicción capaz de manejar grandes residentes en disco de datos. Hay diversas técnicas básicas para la clasificación de datos, tales como la forma de construcción de clasificadores de árboles de decisión, clasificadores bayesianos, redes bayesianas, las creencias y los clasificadores basados en reglas. Retro propagación (una técnica de red neuronal) también se discute, además de un enfoque más reciente de clasificación conocido como máquinas de vectores soporte.

Otros enfoques de la clasificación, como el clasificador k-vecino más próximo, razonamiento basado en casos, algoritmos genéticos, juegos en bruto, y las técnicas de lógica difusa, se introducen. Los métodos para la predicción, incluyendo regresión lineal, regresión no lineal, y otros modelos basados en la regresión, se discuten brevemente. Clasificación y predicción tienen numerosas aplicaciones, incluyendo la detección del fraude, el marketing de destino, la predicción del rendimiento, la fabricación y el diagnóstico médico (Jiawei y Micheline, 2006).

Algoritmos de Clasificación (ClassificationAlgorithms)

En la Clasificación de Datos se desarrolla una descripción o modelo para cada una de las clases presentes en la base de datos. Existen muchos métodos de clasificación como aquellos basados en los árboles de decisión TDIDT como el ID3 y el C4.5, los métodos estadísticos, las redes neuronales, y los conjuntos Difusos, entre otros.

A continuación se describen brevemente aquellos métodos de Aprendizaje Automático que han sido aplicados a la Minería de Datos con cierto éxito: Algoritmos estadísticos: Muchos algoritmos estadísticos han sido utilizados por los analistas para detectar patrones inusuales en los datos y explicar dichos patrones mediante la utilización de modelos estadísticos, como, por ejemplo, los modelos lineales. Estos métodos se han ganado su lugar y seguirán siendo utilizados en los años venideros.

Redes Neuronales: las redes neuronales imitan la capacidad de la mente humana para encontrar patrones. Han sido aplicadas con éxito en aplicaciones que trabajan sobre la clasificación de los datos.

Algoritmos genéticos: técnicas de optimización que utilizan procesos como el entrecruzamiento genético, la mutación y la selección natural en un diseño basado en los conceptos de la evolución natural.

Método del vecino más cercano: es una técnica que clasifica cada registro de un conjunto de datos en base a la combinación de las clases de los k registros más similares. Generalmente se utiliza en bases de datos históricas.

Entre otras técnicas utilizadas tenemos:

Árboles de Decisión: En un árbol de decisión cada nodo representa una característica que puede tomar diversos valores, cada uno de los cuales genera una rama. Los nodos hojas representan las clasificaciones finales. Los árboles de decisión generalmente utilizan una técnica de aprendizaje supervisado que consiste en realizar particiones recursivas en un espacio de datos conformado por un conjunto de casos o ejemplos. La división de cada conjunto se realiza teniendo en cuenta el atributo cuya partición o agrupación por sus valores, se parezca más a la partición que produce el atributo clase. Con esto se persigue obtener particiones finales que contengan todos los ejemplares de una determinada clase distinguidos por valores de atributos comunes. En el árbol que se va construyendo con cada partición, sus nodos representan un atributo que sirvió para realizar una partición, los arcos que conectan los nodos entre sí representan los diferentes valores que pueden tomar los atributos y las hojas representan subconjuntos de casos pertenecientes a una clase. Todos los caminos del árbol desde la raíz hasta las hojas constituyen una regla de clasificación.

NaiveBayes: La técnica NaiveBayes o clasificador bayesiano simple, es una de las técnicas de clasificación más utilizadas, basada en la estadística. Se trata de un algoritmo de inducción probabilística que representa cada clase como un sumario de probabilidades. El fundamento principal del clasificador NaiveBayes es la suposición de que todos los atributos son independientes conocido el valor de la variable clase. La hipótesis de independencia asumida por esta técnica da lugar al modelo de una red bayesiana en la que existe un único nodo raíz (clase), y en la que todos los atributos son nodos hoja que tienen como único padre a la variable clase.

C-Means Clasico: C-means es un algoritmo iterativo que hace parte de las técnicas de agrupamiento no supervisado y tiene como objetivo encontrar patrones o grupos interesantes en un conjunto de datos dado, de tal manera que tales patrones, estructuras o grupos encontrados sirvan para clasificación, diseño de estrategias, soporte de decisiones, organización de la información, entre otras. C-means al igual que otras técnicas clásicas de agrupamiento realiza una partición dura del conjunto de datos, tal partición se caracteriza porque cada dato pertenece exclusivamente a un cluster (grupo o clase) de la partición, además, los clusters deben cubrir totalmente el conjunto de datos, es decir cada dato tiene que pertenecer a alguno de los clusters; la cantidad de clusters debe ser definida para inicializar el algoritmo.

Fuzzy C-Means: En muchas situaciones cotidianas ocurre el caso que un dato está lo suficientemente cerca de dos clusters de tal manera que es difícil etiquetarlo en uno o en otro, esto se debe a la relativa frecuencia con la cual un dato particular presenta características pertenecientes a clusters distintos y como consecuencia no es fácilmente clasificado; fuzzy c-means (FCM) es un algoritmo que se desarrolló con el objetivo de solucionar tales inconvenientes. El algoritmo FCM asigna a cada dato un valor de pertenencia dentro de cada cluster y por consiguiente un dato específico puede pertenecer parcialmente a más de un cluster. A diferencia del algoritmo c-means clásico que trabaja con una partición dura, FCM realiza una partición suave del conjunto de datos, en tal partición los datos pertenecen en algún grado a todos los clusters (Rojas, Chavarro y Moreno).

Metodología

El tipo de investigación fue descriptivo, ya que buscamos especificar las propiedades importantes de un grupo de estudiantes para su análisis, en este caso, su información académica y personal. La aplicación de los algoritmos de minería de datos requiere la realización de una serie de actividades previas encaminadas a preparar los datos de entrada debido a que, en muchas ocasiones dichos datos proceden de fuentes heterogéneas, no tienen el formato adecuado o contienen ruido.

Por otra parte, es necesario interpretar y evaluar los resultados obtenidos. El proceso completo consta de las siguientes etapas (Cabena, et al., 1998):

Selección: Identificación de las fuentes de información externas e internas y selección del subconjunto de datos necesario.

Reprocesamiento: estudio de la calidad de los datos y determinación de las operaciones de minería que se pueden realizar.

Conversión de datos en un modelo analítico.
Tratamiento automatizado de los datos seleccionados con una combinación apropiada de algoritmos.

Interpretación de los resultados obtenidos en la etapa anterior, generalmente con la ayuda de una técnica de visualización.

Aplicación del conocimiento descubierto. Aunque los pasos anteriores se realizan en el orden en que aparecen, el proceso es altamente iterativo, estableciéndose retroalimentación entre los mismos. Además no todos los pasos requieren el mismo esfuerzo. Generalmente la etapa de pre procesamiento es la más costosa ya que representa aproximadamente el 60 % del esfuerzo total, mientras que la etapa de minería sólo representa el 10%.

Criterio de selección de técnica de minería de datos.

Para la selección de una técnica de minería de datos se ha de tener en cuenta una serie de consideraciones previas que afectan al desempeño de la técnica. El entender estas características y su impacto, es útil para escoger la técnica que mejor se adecúe a una determinada aplicación. Teniendo en cuenta el entorno se seleccionaron los siguientes criterios:

Habilidad para manejar datos con ruidos: Las bases de datos a menudo contienen ruido en forma de imprecisiones o inconsistencias. Algunos procesos de validación de datos están mal diseñados y permiten introducir datos incorrectos a los usuarios.

Habilidad para manejar datos perdidos: es importante darle un manejo apropiado a los datos ya que se pueden producir pérdidas si los datos se obtienen de diferentes fuentes.

Procesamiento de grandes volúmenes de datos: es importante que la técnica posea la habilidad para manejar gran cantidad de información lo cual permite mayor precisión en el análisis.

Escalabilidad: esta es una propiedad muy deseable en una técnica de minería de datos para futuras actualizaciones.

Procesamiento de diferentes tipos de datos: es importante que la técnica seleccionada tenga lacapacidad para poder manejar diferentes tipos de datos numéricos, cadenas, etc.

Capacidad predictiva: Esta característica tiene gran influencia en la efectividad de la técnica de minería porque determina qué probabilidad existe para la solución de un problema.

Facilidad de Operación: La facilidad de integración y operación es otra característica importante para su utilización.

Capacidad explicativa:
Dependiendo de la técnica utilizada, el grado de procesamiento aplicable al dato varía, por tanto, una técnica que sea fácil de entender y que requiera de poco procesamiento previo es más interesante para un usuario final.

Complejidad de implementación:
Es importante que la técnica seleccionada no presente un alto grado de complejidad para su implementación, lo que resulta conveniente para el desarrollo de una herramienta de minería.

Inserciones a la base de datos: Es importante que los algoritmos minimicen el recorrido por la base de datos, pues el número de reglas crece exponencialmente con el número de ítems considerados, lo cual afecta el rendimiento del algoritmo cuando se accesa constantemente a la base de datos.

Costo computacional: es importante que el algoritmo no realice un gran número de operaciones que agoten los recursos de máquina.

Tiempo de Ejecución: se desea que el tiempo utilizado para la generación de reglas sea razonable.

Rendimiento: es importante que el algoritmo realice las operaciones y procesos de forma eficiente.

Éstas técnicas de Minería de datos aplicadas a problemas específicos permiten delimitar los requerimientos de cualquier proyecto y el uso de diferentes algoritmos rápidos y eficientes para la realización del proceso de minería de datos.

Aplicación de la Minería de datos para analizar la deserción estudiantil

En la investigación titulada Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos (Timarán, 2010), utilizaron TariyKDD, una herramienta de minería de datos de distribución libre, desarrollada en los laboratorios KDD del grupo de investigación Grias, del Departamento de Sistemas de la Facultad de Ingeniería de la Universidad de Nariño.

Las Variables que utilizaron fueron:
• Ingresos.
• Edad.
• Edad ingreso.
• Valor de la matrícula.

•Clase alterna: determina qué estudiantes han reingresado, se han retirado o no cumplen con ninguna de las condiciones anteriores.

• Rendimiento de la clase: determina la cantidad de materias perdidas por el estudiante.

• Promedio: promedio acumulado.

Otro estudio realizado en Tailandia por Nghe, Janecek y Haddwy (2007), fue Comparative Analysis of Techniques for Predicting Academic Performance, ellos utilizaron las Técnicas de Árboles de Decisión (J48) que como resultado fue más precisa, y las Redes Bayesianas la cual fue menos precisa.

En México, Valero, Salvador Vargas y García (2010), realizaron una Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos, usando las técnicas de Árboles de decisión C4.5 y Técnica de los k vecinos más cercanos.

Ellos utilizaron las siguientes variables:
• Sexo
• Edad
• Tipo de bachillerato
• Promedio Bachiller (ICFES)
• Materias reprobadas
• Apoyo Económico (Beca)
• Domino del idioma Inglés
• Hábitos de estudio
• Resultado de Exámenes
• Escolaridad del Padre
• Escolaridad de la Madre
• Ingreso mensual Familiar aproximado
• Tamaño de su Familia
• Trabaja actualmente
• Horas aproximadas del trabajo semanal

En Colombia, para el estudio de esta problemática de deserción estudiantil en la Universidad Simón Bolívar- sede Barranquilla, el grupo de investigación Ingebiocaribe del programa de Ingeniería de Sistemas, agrupó las causales definidas en 5 variables que son:

Pérdida de semestre.
Dificultad financiera.
Ingreso al mercado laboral.
Otros intereses atraen al estudiante.
Indeterminado.

Para el análisis, se seleccionó una población considerable (707 sujetos), abarcando desde el primer al décimo semestre del programa de Ingeniería de Sistemas, entre los períodos académicos 2007-2012.

Utilización de Weka para generar los resultados. Weka (Gallirallus australis) es un ave endémica de Nueva Zelanda.

Este ave da nombre a una extensa colección de algoritmos de Máquinas de conocimiento desarrollados por la universidad de Waikato (Nueva Zelanda) implementados en Java [1, 2]; útiles para ser aplicados sobre datos mediante los interfaces que ofrece o para embeberlos dentro de cualquier aplicación.

Además Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización. Weka está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla.

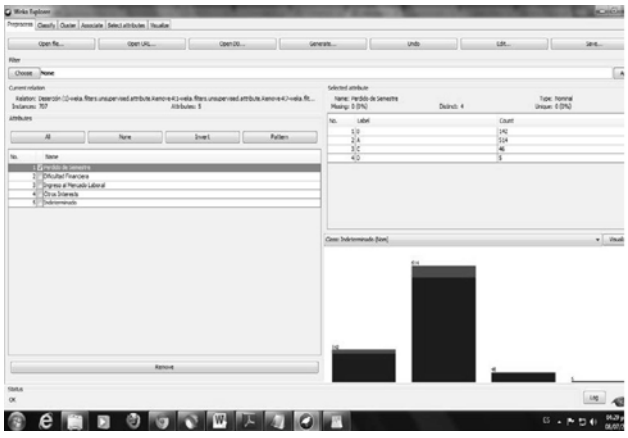

Figura 2.Ventana de selección de interfaz.

A continuación se realiza la clasificación según los atributos planteadoscon una muestra de 707 estudiantes usando el algoritmo de clasificación desarrollo de las metodologías de minería de datos.

PART -M 2 -C 0.25 -Q 1
Resultados
=== Run information ===
Scheme:weka.classifiers. rules.part
- M 2 -C 0.25 -Q 1
Relation: Desercion
Instances: 707
Attributes: 5
Test mode: 10-fold cross validation
Correctly Classified Instances637 94.5274%
Incorrectly Classified Instances 70 5.4726%

MATRIZ DE CONFUSION
=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
30	13	0	0	2	| a = Perdido de semestre
62	0	39	2	0	| b = Dificultad financiera
7	1	0	8	4	| c = Ingreso al mercado
51	1	1	0	10	| d = Otros intereses atraen
460	3	85	0		| e = INDETERMINADO

Vista general de las variables que se analizan

Interpretación de los resultados
Con un margen de confianza de 94%
A partir de estos datos, se puede destacar que la principal causa de deserción de los estudiantes, de acuerdo a los parámetros establecidos como causales definidas en 5 variables permite afirmar que la causa de la deserción es el factor Indeterminado.

Sin embargo, haciendo un análisis detallado de la matriz de confusión se encontró que:

30 desertaron por pérdida de semestre.
62 desertaron por Dificultad financiera.
7 por Ingreso al mercado laboral.
51 por Otros Interés.
460 por causa indeterminada.
13 desertaron por Pérdida de semestre y dificultad financiera.
2 fueron por pérdida de semestre y causa indeterminada.
39 desertaron por dificultad financiera y ingreso al mercado laboral.
2 fueron por dificultad financiera y Otros intereses.
1 estudiante por ingreso al mercado laboral y dificultad financiera.
8 por ingreso al mercado laboral y otros intereses.
4 por ingreso al mercado laboral y causa indeterminada.
10 desertaron por otros intereses y causa indeterminada.
3 por dificultada financiera y causa indeterminada.
8 por Ingreso al mercado laboral y causa indeterminada.
5 por otros intereses y causa indeterminada.
Este hallazgo debe generar un plan de acción,

para realizar mejores filtros en la búsqueda de estas causales indeterminadas, a través de diferentes estrategias de recopilación de la información necesaria. Estos registros son de suma importancia para las decisiones administrativas y financieras.
Seguidamente se encuentra la dificultad financiera, otros intereses atraen al estudiante, pérdida de semestre e ingreso al mercado laboral.
Una conclusión certera abre:

Conclusiones

• La investigación de las diferentes técnicas de minería de datos y su empleo en la solución de diversos tipos de problemas de análisis de información, ayudan a tener un conocimiento general del tema para desarrollar trabajos futuros en otras áreas de conocimiento.

• Para la escogencia de una técnica de minería de datos es necesario entender las necesidades propias del trabajo a desarrollar y tener en cuenta las consideraciones y criterios para seleccionar una técnica adecuada que satisfaga los requerimientos propios del mismo.

Lo importante del desarrollo de la investigación es lograr que la Universidad Simón Bolívar genere estrategias administrativas para mitigar de alguna manera las variables que afectan directamente en la deserción estudiantil, logrando sensibilizar a los directivos y otorgarles información base, para la toma de decisiones que competen a la población estudiantil en cuestión, para el diseño de planes de acción, tales como el otorgamiento de bonos de matrícula, monitoria social, vinculación laboral, reliquidación de matrícula, Estrategia comunicativa, Educación de las familias y acudiente, Intervención familiar, Estrategias de acompañamiento académico, Acompañamiento individual por psicología, entre otros.

Referencias

Amat Bedmar, A. (2005). Ingeniería De Conocimiento Minería De Datos Empresariales. M.S. / E.T.S. Ingeniería Informática de la Universidad de Granada España.

Bean, J. (1985). Interaction effect based on class level in an exploratory model of college student dropout syndrome. American EducationalResearchJournal, 35-64.

Christian Borgelt, C.Department of Knowtedge Processing and Language Engineering -School of Computer Science, Ottovon-Guerlcke - University of Magdeburg 2005. McGraw Hill. - 2000 .

Britos, P. (2005). Minería de Datos. Buenos Aires: Nueva Librería.

Cabena, H., y Stadler, V. Z. (1998). Discovering Data mining From Concept To Implementation. Irlanda.

Castañeda, J. A.. y Rodríguez, M. A. (2005). La minería de datos como herramienta de Marketing: Delimitación y Evaluación del resultado. España: Facultad de CC.EE. Departamento de Comercialización e Investigación de mercados. Universidad de Granada.

Consejo Nacional de Acreditación. (2011). cna.gov.co. Recuperado el 2011, de http://www.cna.gov.co/1741/article-187279.html

Darvger, R., y Berlanga, R. (2001). Informe técnico Búsqueda de Reglas de Asociación en bases de datos y colecciones de textos. Santiago de Cuba: Departamento de Computación, Universidad de Orlente.

El Diario, d. O. (19 de 04 de 2012). Analizan deserción universitaria. Recuperado el 31 de 05 de 2012, de http://www.eldiario.com.co/seccion/LOCAL/analizan-deserci-n-universitaria120418.html

Faculty of Computer and Slovenia Information Science, University of Liubliana. Orange, fruitful and fun. Recuperado de http://wvw.ailab.si/orange - 2007.

Guzmán, C., Durán, D., Franco, J., Castaño, E., Gallón, S., Gómez, K., y otros. (2009). Deserción estudiantil en la educación superior colombiana. Bogotá: ImprentaNacional de Colombia.

JlaweJ, H., y Kamber, M. (2002). Data Mining: Concepts and Techniques. Simón Fraser University. Morgan Kaufmann Publishers.

Kantardzic, M. “Data Mining: Concepts, Models, Methods, and Algorithms”, the textbook. Estados Unidos: IEEE Press & John Wiley, (First edition, November 2002; Second Edition, August 2011).

Kenneth, C., y Jane, P. Administración de la Información y toma de decisiones, Resúmenes de los principales capítulos del libro, Management Information Systems Organization and Technology. Documento. Chile: Universidad de Taparaca.

Kimball, R., y Ross, M. (2002). The Data Warehouse Toolkit The Complete Guide to Dimensional Modefing. McGraw Hül. Second Edition.

Núñez, F., y Lugones, F. (2001). Modelos de Negocios en Internet visión poscrisis. McGraw/Hill.

Rakotomalala. Tanagra project. Recuperado de http://chiroubte.univ- lyon2.fr/ricco/tanagra/en/tanagra.html 2007.

The CRISP-DM Consortium. (2002). CRISP-DM Step by step data mining guide. Documento. Recuperado de http://www.crisp-dm.org/CRISPWP-0800.pdf ,2007.

Waikato ML Group. The Waikato environment for knowledge analysis. Recuperado de http://www.cs.waikato.ac.nz/ml/weka - 2007

Weiss, S., y Indurkhya, N. (1998). Predictive Data Mining. A practical Guide. San Francisco: Morgan Kaufmann publishers.