Aplicación para la predicción de resultados en la prueba Saber 11º

Documento de Anteproyecto de Grado
Proyecto de Ingeniería

Estudiante
Jorge Iván Durán Páez
200859307
jdp0990@gmail.com

Director
Edgar Alberto Molina
Ingeniero de Sistemas
edgarmolina2@hotmail.com

Asesor
David Alejandro Przybilla
Ingeniero de Sistemas
paranoic.pum@gmail.com

Universidad del Valle - sede Tuluá
Escuela de Ingeniería de Sistemas y Computación
Junio 2012

Índice

1. Introducción

En el ámbito de la educación en Colombia, la prueba Saber 11°[1] es de gran importancia para medir la calidad de la enseñanza que se está impartiendo en los colegios del país. La prueba Saber 11° es mayoritariamente presentada por estudiantes de grado 11 de los colegios del país, pero no es exclusiva de estos, cualquier persona puede presentarla, siempre y cuando ya haya obtenido título de bachiller.

Los resultados que se obtienen en la prueba no han presentado mejorías importantes en los últimos 5 años [1, 2, 3, 4], causando preocupación por la calidad de la educación media en Colombia.

En [4] se presenta un análisis a los resultados obtenidos por los estudiantes del departamento de Valle del Cauca en las pruebas Saber 5°, 9°[2] y 11°. Para el caso de la prueba Saber 11°, se presentan los resultados obtenidos en los años de 2002 a 2009, una comparación de estos resultados con los del promedio nacional de cada una de las áreas evaluadas y la categorización[3] de los colegios en los años 2010 y 2011.

En las conclusiones y recomendaciones incluidas en [4], se establece la necesidad de brindar una educación que sea pertinente con las deficiencias académicas de los estudiantes, para así poder emprender planes de mejoramiento en las instituciones educativas y reducir los bajos rendimientos en la prueba Saber 11°.

En el presente documento se propone la construcción de un aplicación, para lograr determinar cuáles son los posibles resultados que obtendrán los estudiantes que están próximos a presentar la prueba Saber 11° y así conocer las áreas académicas en las cuales pueden poseer deficiencias.

[1] http://www.icfes.gov.co/examenes/saber-11o/informacion-general
[2] http://www.icfes.gov.co/examenes/pruebas-saber
[3] ftp://ftp.icfes.gov.co/SABER11/SB11-CLASIFICACION-PLANTELES/Clasificacion%20planteles%20SB11.pdf

2. Planteamiento del Problema

2.1. Descripción del Problema

Para la construcción de una aplicación, que permita conocer previamente como podría ser el puntaje de un estudiante en una o varias de las áreas académicas evaluadas en la prueba Saber 11°, se cuenta con las bases de datos del ICFES,[4][5] que almacenan información histórica sobre aspectos personales de los evaluados como por ejemplo: calendario académico del colegio al cual pertenece, carácter académico del colegio, ubicación del colegio (departamento y municipio), valor mensual de la pensión del colegio, si el hogar cuenta con servicio de alcantarillado y recolección de basuras, cantidad de automóviles que poseen en el hogar, si tiene o no computador, cantidad de televisores en el hogar, etnia a la cual pertenece, genero, fecha de nacimiento, si trabaja o no, nivel de educación de los padres, ocupación de los padres, etc. Y además información sobre los puntajes obtenidos por ellos en las áreas académicas evaluadas.[6]

El ICFES suministra estas bases de datos en archivos de Access,[7] un archivo por cada prueba realizada desde el año 2000 (se realizan 2 pruebas cada año), pero estas bases de datos no son homogéneas, ya que durante las 24 pruebas registradas hasta el momento la información almacenada de los evaluados no se ha mantenido constante. Durante 12 años la información ha sido recolectada a partir de encuestas en las cuales no siempre se han realizado las mismas preguntas. Por ejemplo, el dato sobre el nivel educativo de los padres se registró durante los años 2000 a 2004, pero no se registró en los años 2005 a 2007 y en 2008 vuelve a registrarse hasta ahora. Además los tipos de datos también se han modificado a lo largo de los años, en la época de 2000 a 2004 el valor "3" indicaba que el nivel de estudio de los padres llegaba a básica primaria, pero desde 2008 hasta ahora existen los valores "9" y "10" que significan "Primaria Completa" y "Primaria Incompleta" respectivamente. Causando esto que dentro de las bases de datos se encuentren muchos valores nulos y una falta de concordancia con los tipos de datos en muchos de los 84 atributos que se almacenan en las 24 bases de datos.

Por consiguiente, en el proceso de construcción de una aplicación que permita conocer cómo serán los resultados de un estudiante al momento de presentar la prueba Saber 11°, es necesario realizar la reestructuración de estas bases de datos, para que su información sea concordante y no almacene datos nulos, y así posteriormente utilizarlas como insumo en la construcción de la aplicación.

2.2. Formulación del Problema

¿Qué proceso se debe aplicar para la reestructuración de las bases de datos?

¿Cómo construir una aplicación que permita conocer previamente los resultados de un estudiante en la prueba Saber 11°?

[4] ftp://ftp.icfes.gov.co/SABER11/

[5] Estas bases de datos son de acceso gratuito pero previamente se deben registrar los datos personales en http://64.76.89.156/investigacion/index.php/bdicfes/solicitudregistro, para poder acceder al ftp.

[6] En ftp://ftp.icfes.gov.co/SABER11/SB11-Diccionario_de_Datos-v1-6.pdf se encuentra la información sobre todos los datos que se han recolectado desde el año 2000 hasta la fecha.

[7] http://office.microsoft.com/es-es/access

3. Justificación

Los puntajes obtenidos en la prueba Saber 11° son de gran importancia tanto para los estudiantes que la presentan, como para las instituciones educativas y para el gobierno nacional, ya que estos brindar una estimación de los indicadores de calidad de la educación media en el país.[8]

Un estudiante que desea continuar con su vida académica ingresando a la educación universitaria, se ve en la necesidad de obtener puntajes que le permitan no solo cumplir con los mínimos puntajes necesarios de inscripción en la carrera de su predilección, sino que también le permitan ingresar a ella en la universidad.

En las instituciones educativas, alcanzar un nivel medio o alto en la clasificación otorgada por el ICFES es de gran importancia para su prestigio dentro de la sociedad académica, los colegios utilizan además estos puntajes como un indicador de autoevaluación para medir la calidad de sus prácticas pedagógicas.

Como se puede observar en la educación media del país, prácticamente todos los involucrados obtienen beneficios si se mejoran los puntajes obtenidos en la prueba Saber 11°. Un primer paso para trabajar en estas mejoras es tener la capacidad de detectar las deficiencias de los estudiantes, con la aplicación que se propone construir en este documento, se podría dar ese primer paso.

4. Objetivos

4.1. Objetivo General

Desarrollar una aplicación que permita predecir los puntajes que obtendrá un estudiante en la prueba Saber 11º.

4.2. Objetivos Específicos

1. Aplicar el proceso de extracción, transformación y carga (*Extract, transform and load, ETL*) a las bases de datos suministradas por el ICFES.

2. Construir un clasificador, utilizando técnicas de minería de datos, a partir de la información procesada.

3. Implementar una interfaz en donde los usuarios puedan realizar consultas parametrizadas.

[8] http://www.icfes.gov.co/examenes/saber-11o/informacion-general/objetivos

4.3. Resultados Esperados

Objetivo Especifico	Resultados Esperados
1	▪ Bases de datos del ICFES, con estructuras homogéneas y sin datos nulos. ▪ Data warehouse construido con la información procesada en las bases de datos del ICFES.
2	▪ Informe sobre algoritmos de clasificación en el proceso de descubrimiento del conocimiento. ▪ Selección del algoritmo de clasificación que mejor se adapta a la cantidad y el tipo de datos con los que se cuenta en el data warehouse. ▪ Clasificador construido en base a la selección hecha.
3	▪ Documentación de los proceso de diseño, clasificación y pruebas de la interfaz de consultas.

Cuadro 1: Relación entre los objetivos específicos y los resultados que se desean obtener con cada uno de ellos.

5. Alcance

La presente propuesta establece la construcción de una aplicación, que usando solamente la información almacenada en las bases de datos históricas del ICFES, logre entregar a los usuarios información sobre cómo podrían ser los puntajes que obtendrá un estudiante en cada una de las áreas académicas evaluadas en la prueba Saber 11°.

Para que un usuario pueda conocer el posible resultado de un estudiante, deberá realizar una consulta en donde ingresará datos personales del estudiante, estos datos personales que se deben ingresar serán definidos en el proceso de construcción del clasificador, pero no serán distintos a los registrados en las bases de datos del ICFES.

Después de ser consultada la aplicación, la información que contendrá la respuesta estará constituida por los puntajes que podría obtener un estudiante en cada una de las áreas académicas evaluadas en la prueba Saber 11º: lenguaje, matemáticas, biología, química, física, filosofía, ciencias sociales e ingles.

El proyecto se desarrollará llevando a cabo la propuesta metodológica presentada por José Hernandez en [5].

6. Marco Referencial

6.1. Marco Conceptual

- **Descubrimiento de conocimiento en bases de datos (KDD):** Se define como el proceso de identificar patrones significativos en los datos que sean válidos, novedosos, potencialmente útiles y comprensibles para un usuario. El proceso global consiste en transformar información de bajo nivel en conocimiento de alto nivel. El proceso KDD es interactivo e iterativo conteniendo los siguientes pasos: comprender el dominio de aplicación, extraer la base de datos objetivo, preparar los datos minería de datos, Interpretación y utilizar el conocimiento descubierto [6].

- **Minería de datos:** Es un campo interdisciplinar con el objetivo general de predecir las salidas y revelar relaciones en los datos. Las tareas propias de la minería de datos pueden ser descriptivas, (i.e. descubrir patrones interesantes o relaciones describiendo los datos), o predictivas (i.e. clasificar nuevos datos basándose en los anteriormente disponibles). Para ello se utilizan herramientas automáticas que emplean algoritmos sofisticados para descubrir principalmente patrones ocultos, asociaciones, anomalías, y/o estructuras de la gran cantidad de datos almacenados en los data warehouses u otros repositorios de información, y filtran la información necesaria de las grandes bases de datos [6].

- **Clasificación:** Es la tarea de aproximar una *función objetivo* desconocida $\Phi : I \times C \to \{T, F\}$ (que describe cómo instancias del problema deben ser clasificadas de acuerdo a un experto en el dominio) por medio de una función $\Theta : I \times C \to \{T, F\}$ llamada el *clasificador*, donde $C = \{c_1, ..., c_{|c|}\}$ es un conjunto de categorías predefinido e I es un conjunto de instancias del problema. Comúnmente cada instancia $i_j \epsilon I$ es representada como una lista $A = \{a_1, a_2, ..., a_{|A|}\}$ de valores característicos, conocidos como *atributos*, i.e. $i_j = \{a_1, a_2, ..., a_{|A|j}\}$. Si $\Phi : i_j \times c_i \to T$, entonces es llamado un ejemplo positivo de c_i, mientras si $\Phi : i_j \times c_i \to F$ éste es llamado un ejemplo negativo de c_i.

  Para generar automáticamente el clasificador de c_i es necesario un proceso inductivo, llamado el *aprendizaje*, el cual por observar los atributos de un conjunto de instancias preclasificadas bajo c_i o $\bar{c_i}$, adquiere los atributos que una instancia no vista debe tener para pertenecer a la categoría. Por tal motivo, en la construcción del clasificador se requiere la disponibilidad inicial de una colección Ω de ejemplos tales que el valor de $\Phi(i_j, c_i)$ es conocido para cada $\langle i_j, c_i \rangle \epsilon \Omega \times C$. A la colección usualmente se le llama *conjunto de entrenamiento* (Tr)[7].

- **ICFES:** Instituto Colombiano para la Evaluación de la Educación, entidad especializada en ofrecer servicios de evaluación de la educación en todos sus niveles, y en particular apoyar al Ministerio de Educación Nacional en la realización de los exámenes de Estado y en adelantar investigaciones sobre los factores que inciden en la calidad educativa, para ofrecer información pertinente y oportuna para contribuir al mejoramiento de la calidad de la educación [8].
- **Prueba Saber 11º:** Antes conocida como Examen del ICFES, es un examen de estado que evalúa a los estudiantes que están terminando su ciclo de Educación Media. La prueba tiene como finalidad apoyar los procesos de selección y admisión que realizan las instituciones de Educación Superior. Además de este propósito, la prueba busca:

  - Brindar al estudiante información que contribuya a la selección de su opción profesional.
  - Proporcionar información a las instituciones de educación básica y media sobre el desempeño de los estudiantes.
  - Contribuir al desarrollo de estudios de tipo cultural, social y educativo.
  - Servir de criterio para otorgar beneficios educativos.

  Las áreas académicas evaluadas actualmente por la prueba Saber 11° son:

  - Lenguaje
  - Matemáticas
  - Biología
  - Química
  - Física
  - Filosofía
  - Ciencias sociales
  - Inglés
  - COMPONENTE FLEXIBLE (solo se presenta una de las siguientes opciones)
    - Profundización en lenguaje
    - Profundización en matemáticas
    - Profundización en biología
    - Profundización en ciencias sociales
    - Interdisciplinar violencia y sociedad
    - Interdisciplinar medio ambiente
- **Data Warehouse:** Es un repositorio de datos operacionales seleccionados y adaptados subjetivamente, que puede responder consultas de tipo ad hoc, estadísticas o analíticas. Está situado en el centro de los sistemas de apoyo a la toma de decisiones de una organización y contiene datos históricos, resumidos y detallados de esta. Es esencial para una inteligencia de negocios efectiva, para la formulación e implementación de estrategias donde la gran cantidad de datos requieren ser procesados de manera rápida para comprender su significado e impacto. Permite una fácil organización y mantenimiento de los datos para una rápida recuperación y análisis de la manera en que sean requeridos [9].

6.2. Antecedentes o Estado del Arte

6.2.1. Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos[3]

Su objetivo fue determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento del conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años. Este proceso se apoyó con TariyKDD, una herramienta de minería de datos de distribución libre.

Este trabajo desarrollo de manera organizada y bien estructurada, cada uno de los pasos planteados en [8]. El algoritmo de minería de datos utilizado fue arboles de clasicación con C4.5, los resultados entregados por esta investigación, fueron claros y mostraron cuales eran los perfiles de los estudiantes que podrían prever un bajo rendimiento.

La investigación se realizo en la Universidad de Nariño de Colombia. Las bases de datos recolectadas contenían información sobre el desempeño académico e información personal de 46173 estudiantes, acumuladas durante 15 años.

Inicialmente se contaba con 69 atributos en las bases de datos que describían las características de los estudiantes, pero después de pasar por el proceso de limpieza y transformación de los datos, se llego a una lista relevante de 26 atributos y una cantidad de registros de 20329.

Finalmente, se aplicaron las técnicas de minería de datos para clasificación, en este caso el algoritmo C4.5 y se obtuvieron las reglas que indicaban que tipo de situaciones personales de un estudiante podrían llevar a obtener un bajo rendimiento en la universidad. Unos ejemplos de las reglas obtenidas son:

- Si el estrato socioeconómico es 2, el ponderado de exámenes de estado ICFES está entre 50 y 70, es del Sur de Nariño, está en primer semestre y pertenece a la facultad de Ciencias Humanas, entonces su rendimiento es Bajo. El 68 % con estas características se clasifican de esta manera.

- Si la edad de ingreso es menor o igual a 18 años, el estrato socioeconómico es 2, género masculino, el ponderado ICFES está entre 50 y 70, vive con la familia, es del Sur de Nariño, está en primer semestre, está en la facultad de Ciencias Naturales y Matemáticas, entonces su rendimiento es Bajo. El 67 % con estas características se clasifican de esta manera.

6.2.2. Minería de Datos en la Educacion[4]

En este documento se describe el uso de la minería de datos aplicada a entornos educativos y su uso pedagógico. De manera muy clara y especifica brinda una vista de como se debe realizar un proceso de minería de datos en educacion y su importancia.

Además nos muestra un ejemplo: "Identicación de características de fracasos escolares en institutos". En este se usan arboles de decisión porque permiten encontrar cuales son las variables que tienen mayor relación con la variable que se desea predecir. El algoritmo de arboles de decisión utilizado fue CHAID (Chi-Squared Automatic Interaction Detection). CHAID realiza comparaciones en pares para encontrar la variable de predicción más altamente relacionada con la variable raíz. En sistemas de muchas variables, tener esta función implementada en un ordenador es esencial para picar amplios conjuntos de datos [4]. El resultado entregado fue un árbol, el cual se debe analizar para determinar su información.

6.2.3. Descubrimiento de Reglas de Predicción en Sistemas de e-learning utilizando Programación Genética[5]

Este artículo describe la utilización de técnicas de minería de datos en sistemas de e-learning para proporcionar retroalimentación a los desarrolladores de courseware. Aquí podemos observar un enfoque

distinto al de predecir rendimientos académicos, ya que aquí las reglas se desean utilizar para mejorar la forma en que se ofrecen los cursos basados en web. Un punto nuevo que se puede observar es el interés por generar resultados entendibles para cualquier persona que utiliza la herramienta que realizará la tarea de predicción de reglas, y es que estas reglas tendrán una gramática especíca, fácil de leer e interpretar. Para lograr esto se uso Programación Genética Basada en Gramática con técnicas de optimización multiobjetivo.

En este trabajo se implementaron varios algoritmos de mineria de datos con el fin de realizar una comparacion de estos y determinar cuales son los que mejor se adaptan a esta area de aplicacion. Se concluyo que la utilizacion de algoritmos geneticos entrega una mayor cantidad de reglas con informacion util para realizar modicacion a los cursos y esta informacion es novedosa, sin la aplicacion de la herramienta hubiese sido dicil determianr algunas de las relaciones que mostro la investigación.

6.2.4. Minería de Datos: Predicción de la Deserción Escolar Mediante el Algoritmo de Árboles de Decisión y el Algoritmo de los k Vecinos más Cercanos[6]

Se han aplicado técnicas de minería de datos para buscar predecir la deserción escolar en la Universidad Tecnológica de Izúcar de Matamoros, Mexico, tomando como base de análisis los datos del estudio socioeconómico del EXANI-II, elaborado por el CENEVAL, mismo que se aplica desde el año 2003 en la institución. Para esta investigación se utilizaron específicamente dos algoritmos: el algoritmo de árboles de clasicación C4.5 y el algoritmo de los k vecinos más cercanos[6].

En el proceso de transformación de los datos, el mas largo en muchos de los trabajos que se realizan, incluyendo este, se mofidicó e integr toda la información encontrada, esta se encontraba en distintos formatos de almacenamiento, además no mantenía un estructura clara, porque constaba de información de estudiantes de 14 cuatrimestres, por supuesto en cada registro no se realizaban las mismas preguntas, además de que muchas de estas se pueden responder como "No lo se", después de superar esta etapa y aplicar los algoritmos previstos, se encontró que el algoritmo de los k vecinos mas cercanos funciona bien con pocos datos (477 instancias), pero al momento de probarlo con 6525 instancias, el algoritmo C4.5, tuvo resultados muchísimo mas confiables (98,98 %).

Aquí también se realizo la creación de una herramienta que pueda ser accedida por cualquier usuario y le informa sobre la probabilidad de que un estudiante deserte. Los resultados mostraron que la edad, la situación económica y el nivel de ingles, tienen fuerte relación con que el estudiante deserte de la universidad.

6.2.5. Modelo Predictivo para la Determinación de Causas de Reprobación Mediante Minería de Datos[2]

En este trabajo, realizado en la Universidad Tecnológica de Puebla, Mexico, se llevó a cabo el análisis de los datos que nos permitirán generar un modelo que ayude a predecir, desde que los alumnos ingresan a la Universidad, las causas que los llevarán a reprobar, así como las materias con mayor riesgo de ser reprobadas[2].

El algoritmo de clasicación utilizado fue C4.5, y se recolectaron datos con información de todas las carrera ofrecidas por la universidad. Este proceso de nuevo fue demorado, porque sus fuentes de almacenamiento poseían distintas estructuras y no existía una homogeneidad en los datos.

Los resultados obtenidos mostraron que de las 157 materias distintas que ofrece la universidad en sus carreras, 64 materias tienen un porcentaje de reprobación menor a 40 % y por lo tanto no generan un árbol de predicción. En las materias con un porcentaje mayor al 50 % se determinó que el factor principal de reprobación es el profesor que imparte la materia y también la edad de los estudiantes.

6.2.6. La metodología del Data Mining. Una aplicación al consumo de alcohol en adolescentes[7]

Aunque en esta investigación su objetivo no estaba focalizado en descubrir conocimiento en el área de la educación. Si es un trabajo interesante y que puede servir de guía en este trabajo de grado, ya que en ella se realizo la construcción de 3 algoritmos distintos de predicción: redes neuronales, arboles de decisión y naive bayes.

El trabajo se centra en demostrar la importancia que tiene el descubrimiento del conocimiento en bases de datos al momento de predecir comportamientos en distintas áreas como educación, finanzas, comercio, telecomunicaciones, salud, entre otras.

Estos algoritmos fueron aplicados en un conjunto de datos que contenía 7030 registro que informaban sobre el consumo de alcohol en jóvenes de entre 14 y 18 años con una cantidad de 20 variables que incluían información de la personalidad como los constructos de autoestima, impulsividad, conducta antisocial y búsqueda de sensaciones.

Los mejores resultados se obtuvieron con el modelo construido con redes neuronales, 64,1 % de precisión al momento de predecir si un joven consumía o no alcohol. Los otros resultados fueron una precisión de 62,3 % con arboles de decisión usando el algoritmo CART y una precisión de 59,9 % usando Naive Bayes.

Como se pudo observar a lo largo de los trabajos revisados en el documento, siempre resaltan la importancia de las técnicas de minería de datos para lograr predecir eventos que pueden afectar el desempeño académico de los estudiantes y poder ejecutar acciones preventivas que minimicen el impacto negativos que estos eventos negativos pueden ocasionar sobre los estudiantes y las instituciones educativas.

En el área de la educación los trabajos se han orientado siempre a encontrar posibilidades de bajos rendimiento académicos y así lograr conocer perfiles o factores que hacen que los estudiantes no logren las metas propuestas en sus estudios. El trabajo se ha centrado mucho en educación superior y casi no se han realizado estudios que permitan conocer los riesgos de un estudiante que aun no es universitario, en este trabajo se propone trabajar con estos estudiantes que aun no han alcanzado la educación superior.

También se observó la variedad de algoritmos utilizados para realizar estos modelos de predicción, en algunos trabajos se utilizó mas de un algoritmo para lograr encontrar comparaciones de rendimiento y mejores niveles de precisión con cada algoritmo. En este trabajo se evaluaran varios algoritmos para la construcción de los modelos de predicción, con el fin de encontrar información sobre como se comportan los algoritmos con la cantidad de datos usados y con la estructura que posean estos datos al momento de ser presentados al modelo para realizar la predicción.

Uno de los aspectos que posee este trabajo y no fue encontrado en los trabajos revisados, es que se trabajará con bases de datos que contiene más de 5 millones de registros, recolectados a lo largo de 11 años y que presentan una gran variación en la integridad de los atributos de estos registros. Por eso este trabajo centrara mucho de su tiempo en la homogeneización de estas bases de datos.

En conclusión, todos los trabajos aquí presentados sirven de base para la realización de este trabajo, ya que sus fases de investigación fueron las mismas en la mayoría de los trabajos, iguales a las que se aplicaran en este trabajo, además los resultados obtenidos en ellos, muestran la utilidad de la construcción de clasificadores para predecir resultados de estudiantes en su vida académica.

6.3. Marco Teórico

Para la realización de este trabajo se utilizara la propuesta de fases claramente establecidas en [8] donde nos dice que: "Este proceso es iterativo e interactivo. Es iterativo ya que la salida de alguna de las fases puede hacer volver a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario (experto en el dominio del problema) interviene en la toma de muchas decisiones".

La teoría del proceso del desarrollo de un proyecto de descubrimiento del conocimiento en bases de datos, define los siguientes pasos:

6.3.1. Selección (*Extraction*)

En esta etapa del proceso, es donde se investigan las fuentes, en las cuales se podrán encontrar los datos que me servirán para general un almacén de datos con información útil. Se recolectan todos los datos obtenidos en estas fuentes y se organizan en bases de datos que se usaran para la etapa de transformación.

Comúnmente en los trabajos orientados a la academia, estos datos son la información de los estudiantes de la institución a lo largo del tiempo. Los datos pueden incluir información acerca de la situación socio-económica del estudiante, sus notas, su entorno académico, información familiar, entre muchas otras que defina la institución que es importante recolectar.

En [2], donde se buscaba encontrar las reglas que predigan si un estudiantes es propenso a reprobar una materia, se seleccionaron como datos útiles: calificaciones por área del examen de admisión EXANI II, datos relevantes del estudio socio-económico, calificación del test de intereses vocacionales (KUDER), calificación del test de coeciente intelectual (RAVEN), estilos de aprendizaje, evaluación a profesores, asignaturas cursadas y su promedio por cuatrimestre.

6.3.2. Transformación (*Transformation*).

En esta etapa el proceso se basa en realizar primero una limpieza de los datos, la limpieza permite obtener datos sin valores nulos o anómalos, además se estandarizan los datos para que sean del mismo tipo.

Para eliminar los datos nulos, se puede omitir este atributo o en algunos casos, cuando no son muchos los registros carentes de este valor, se pueden aplicar técnicas estadísticas como la media para insertar un valor valido y así no eliminar el registro o el atributo de la base de datos.

Estos procesos de limpieza, integración y agregación de los datos. Entregan como resultado una base de datos mofidicada con los atributos relevantes en los registros, para responder al objetivo de la investigación, muchas veces esto genera que la cantidad de registros que se tenían inicialmente, se disminuya signicativamente. Pero esto nos permitirá obtener resultados mucho más confiables, ya que, no habrán datos que generen conflictos a la hora de generar clasificaciones con respecto a algunos atributos.

6.3.3. Carga (*Load*)

Esta etapa, es la construcción del almacén de datos, con el diseño final entregado en el proceso de transformación. Las tablas con sus atributos, se generan y se realiza la carga de los registros obtenidos en el momento de la selección de los datos.

6.3.4. Minería de datos

El objetivo de esta etapa es la búsqueda y descubrimiento de patrones insospechados y de interés utilizando diferentes técnicas de descubrimiento tales como clasicación, clustering, patrones secuenciales, asociación, entre otras[1]. Las diferentes técnicas utilizadas pertenecen a campos como la inteligencia arfiticial y estadísticas. Algunos algoritmos usados comúnmente son: arboles de decisión C4.5, ID3 y CHAID (Detección Automática de Interacción basada en Chi-Cuadrado), algoritmo de los k vecinos más cercanos, Naive Bayes, redes neuronales, algoritmo de reglas de asociación Apriori, algoritmo de inducción de reglas como el Prism, diferentes versiones de algoritmos evolutivos, programación genética basada en gramática.

6.3.5. Interpretación y evaluación de los resultados

Usualmente cuando se tiene una cantidad de datos para trabajar en la reglas, se construyen los modelos con un 70 % de los datos y se deja el 30 % restante para realizar las evaluaciones del modelo, verificando así si las reglas generadas son validas o no. La manera de interpretar los datos es uno de los puntos críticos de la minería, ya que muchas veces estos resultados solo pueden ser interpretados por personas conocedoras del tema de la minería. Es por eso que muchas herramientas generadas en el ámbito académico, tiene una fase importante de generación de una interfaz de usuario y realizar la interpretación de los datos de manera interna y entregarle al usuario una información mejor tratada para sea entendida de manera mas fácil y determinar su utilidad y relevancia dentro del contexto que se este evaluando.

7. Metodología

7.1. Actividades a Realizar

Objetivo Especifico	Actividad	Resultado Esperado
1	▪ Recolección y selección de las bases de datos	▪ Bases de datos con información sobre estudiantes y sus resultados en la prueba Saber 11º
	▪ Limpieza de las bases de datos recolectadas	▪ Registros en las bases de datos sin valores nulos e integridad en los tipos de datos de cada uno de los atributos de estos
	▪ Integración de los datos	▪ Bases de datos con estructuras concordantes entre todas ellas y con los valores de los atributos discretizados

	▪ Diseño de los almacenes de datos	▪ Almacenes de datos con distintas estructuras sobre los cuales podamos aplicar los algoritmos de predicción
	▪ Carga de las bases de datos	▪ Bases de datos construidas con las estructuradas diseñadas y con todos los registros que cumplan con los atributos definidos en estas
2	▪ Construcción del modelo de predicción con el algoritmo C4.5	▪ Modelo basado en el algoritmo C4.5 en el cual se puedan ingresar los datos para conocer los resultados de la predicción
	▪ Evaluación del modelo con el algoritmo C4.5 en las bases de datos	▪ Informe sobre el rendimiento obtenido por este modelo al momento de realizar la predicción con las bases de datos
	▪ Construcción del modelo de predicción con el algoritmo SLIQ	▪ Modelo basado en el algoritmo SLIQ en el cual se puedan ingresar los datos para conocer los resultados de la predicción
	▪ Evaluación del modelo con el algoritmo SLIQ en las bases de datos	▪ Informe sobre el rendimiento obtenido por este modelo al momento de realizar la predicción con las bases de datos
	▪ Construcción del modelo de predicción con el algoritmo Naive Bayes	▪ Modelo basado en el algoritmo Naive Bayes en el cual se puedan ingresar los datos para conocer los resultados de la predicción

	▪ Evaluación del modelo con el algoritmo Naive Bayes en las bases de datos	▪ Informe sobre el rendimiento obtenido por este modelo al momento de realizar la predicción con las bases de datos
	▪ Construcción del modelo de predicción con técnicas de maquina de soporte vectorial	▪ Modelo basado en tecnicas de maquina de soporte vectorial en el cual se puedan ingresar los datos para conocer los resultados de la predicción
	▪ Evaluación del modelo con tecnicas de maquina de soporte vectorial en las bases de datos	▪ Informe sobre el rendimiento obtenido por este modelo al momento de realizar la predicción con las bases de datos
	▪ Construcción del modelo de predicción con tecnicas de redes neuronales	▪ Modelo basado en tecnicas de redes neuronales en el cual se puedan ingresar los datos para conocer los resultados de la predicción
	▪ Evaluación del modelo con tecnicas de redes neuronales en las bases de datos	▪ Informe sobre el rendimiento obtenido por este modelo al momento de realizar la predicción con las bases de datos
3	▪ Revisión de los informes de rendimiento de cada uno de los modelos	▪ Conocer cual será el algoritmo a usar para construir la aplicación
	▪ Definir el lenguaje en que se construirá la aplicación	▪ Información sobre la codificación de la aplicación

	▪ Codificar la aplicación	▪ Aplicación lista para la realización de pruebas
	▪ Realizar pruebas sobre la aplicación	▪ Informe sobre el desempeño de la aplicación

7.2. Cronograma de actividades

Actividad	2012															
	Ago		Sep				Oct				Nov				Dic	
	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2
▪ Recolección y selección de las bases de datos	X															
▪ Limpieza de las bases de datos recolectadas		X	X	X	X											
▪ Integración de los datos						X	X	X								
▪ Diseño de los almacenes de datos									X	X	X					
▪ Carga de las bases de datos												X	X			
▪ Construcción del modelo de predicción con el algoritmo C4.5														X	X	

▪ Evaluación del modelo con el algoritmo C4.5 en las bases de datos															X	X
	2013															
	Feb				Mar				Abr				May			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
▪ Construcción del modelo de predicción con el algoritmo SLIQ	X	X														
▪ Evaluación del modelo con el algoritmo SLIQ en las bases de datos		X	X													
▪ Construcción del modelo de predicción con el algoritmo Naive Bayes				X	X											
▪ Evaluación del modelo con el algoritmo Naive Bayes en las bases de datos					X	X										
▪ Construcción del modelo de predicción con tecnicas de maquina de soporte vectorial						X	X									
▪ Evaluación del modelo con tecnicas de maquina de soporte vectorial en las bases de datos							X	X								

▪ Construcción del modelo de predicción con tecnicas de redes neuronales								X	X							
▪ Evaluación del modelo con tecnicas de redes neuronales en las bases de datos									X	X						
▪ Revisión de los informes de rendimiento de cada uno de los modelos										X						
▪ Definir el lenguaje en que se construirá la aplicación											X					
▪ Codificar la aplicación												X	X	X	X	
▪ Realizar pruebas sobre la aplicación															X	X

8. Presupuesto

A continuación de presentan las tablas que indican el presupuesto necesario para la realizar del proyecto. Primero se presentara una tabla en donde se referencia la información general de los recursos que se utilizaran para la realización del proyecto, después se hace un detalle de cada uno de los gastos en que se incurrirá en cada recurso.

Las fuentes de financiamiento que tendrá el proyecto serán la Universidad del Valle y el estudiante desarrollador del trabajo de grado.

Rubro	*Valor (en pesos)*
Personal	14'560.000
Hardware	800.000
Software	260.000
Desplazamiento	840.000

Recursos Bibliográficos y Papelería	167.000
Total del Proyecto	16'627.000

Cuadro 5:

Tabla detallada de Personal

Concepto	Dedicación (horas/semana)	Costo por hora	Costo Total (en pesos)
Director	1	22.000	60.000
Codirector	1	22.000	60.000
Desarrollador	40	10.000	400.000
Total Gastos Personal (semana)			520.000

Tabla detallada de Hardware

Tipo	Descripción	Costo
Computador portátil	Equipo de computo para la realización de gran parte de las actividades	800.000
Total Gastos Hardware		800.000

Tabla detallada de Software

Tipo	Descripción	Costo
Microsoft Access 2010	Necesario para visualizar las bases de datos entregadas por el ICFES	260.000
Total Gastos Software		260.000

Tabla detallada de Desplazamiento

Tipo	Descripción	Costo
Viaje Buga - Tuluá	Diariamente se necesita el desplazamiento entre estas ciudades para reuniones con el director	6.000
Total Gastos Desplazamiento (semana)		30.000

Tabla detallada de Bibliografía y Papelería

Tipo	Descripción	Costo
Libro	Introducción a la minería de datos	117.000
Impresiones	Impresiones de artículos y avances de la documentación	50.000
Total Gastos Bibliografía y Papelería		167.000

Referencias

[1] Periódico El Colombiano, "Leve mejoría en pruebas Saber 11". [artículo en Internet]. http://www.elcolombiano.com/BancoConocimiento/L/leve_mejoria_en_pruebas_saber_11/leve_mejoria_en_pruebas_saber_11.asp [Consulta: 24 agosto de 2012].

[2] Periódico El Tiempo, "El 45% de los colegios presentó bajo rendimiento en pruebas Saber 11". [artículo en Internet]. http://www.eltiempo.com/vida-de-hoy/educacion/ARTICULO-WEB-NEW_NOTA_INTERIOR-8384822.html [Consulta: 24 agosto de 2012].

[3] Revista Dinero, "Las pruebas del Icfes no son el único indicador". [artículo en Internet] http://www.dinero.com/edicion-impresa/caratula/articulo/las-pruebas-del-icfes-no-unico-indicador [Consulta: 14 marzo de 2012].

[4] Departamento Del Valle Del Cauca, Secretaria De Educación. Informe Ejecutivo Análisis Pruebas Saber 5º ,9º Y 11º. Santiago de Cali, Febrero 15 de 2012.

[5] Hernández Orallo José, Ramírez Quintana Ma. José, Ferri, Ramírez César, Introducción a la minería de datos, Person Educación, S.A. Madrid 2004, ISBN: 978-84-205-4091-7.

[6] José C. Riquelme, Roberto Ruiz, Karina Gilbert, "Minería de Datos: Conceptos y Tendencias," *Revista Iberoamericana de Inteligencia Artificial*, No. 29 (2006), pp. 11-18.

[7] Téllez A., Extracción de Información con Algoritmos de Clasificación [Tesis de Maestría]. Tonantzintla, Puebla, México. Instituto Nacional de Astrofísica, Óptica y Electrónica. 2005.

[8] ICFES, Presentación. [artículo en Internet] http://www.icfes.gov.co/informacion-institucional/informacion-general [Consulta: 4 junio de 2012].

[9] C.S.R. Prabhu, Data Warehousing Concepts, Techniques Products and Applications, Prentice-Hall, India 2006, ISBN: 81-203-2068-9.

[10] Timarán Pereira, Ricardo. Una lectura sobre deserción universitaria en estudiantes de pregrado desde la perspectiva de la minería de datos.Revista Científica Guillermo de Ockham, vol. 8, núm. 1, enero-junio, 2010, pp. 121-130.

[11] Erika Rodallegas Ramos, Areli Torres González, Beatriz B. Gaona Couto, Erick Gastelloú Hernández, Rafael A. Lez ama Morales, Sergio Valero Orea. Modelo predictivo para la determinación de causas de reprobación mediante Minería de Datos. Universidad Tecnológica de Izúcar de Matamoros, Mexico.

[12] Ricardo Timarán Pereira. Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos. Departamento de Sistemas, Facultad de Ingeniería, Universidad de Nariño San Juan de Pasto, Nariño, Colombia.

[13] Álvaro Jiménez Galindo, Hugo Álvarez García. Minería de Datos en la Educación. Universidad Carlos III de Madrid.

[14] Cristóbal Romero, Sebastián Ventura, Cesar Hervás. Descubrimiento de Reglas de Predicción en Sistemas de e-learning utilizando Programación Genética. Universidad de Córdoba, Córdoba, España.

[15] Sergio Valero Orea, Alejandro Salvador Vargas, Marcela García Alonso. Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos. Universidad Tecnológica de Izúcar de Matamoros, Izúcar de Matamoros, Puebla, México.

[16] Elena Gervilla García, Rafael Jiménez López, Juan José Montaño Moreno, Albert Sesé Abad, Berta Cajal Blasco, Alfonso Palmer Pol. La metodología del Data Mining. Una aplicación al consumo de alcohol en adolescentes. Área de Metodología de las Ciencias del Comportamiento. Departamento de Psicología. Universitat de les Illes Balears.

[17] Oded Maimon, Lior Rokach, The Data Mining and Knowledge Discovery Handbook, Springer Science+Business Media, Inc 2005, ISBN-10: 0-387-24435-8.