UNIVERSIDAD NACIONAL DE COLOMBIA

Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos.

ALVARO AGUSTIN OÑATE BOWEN

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería de Sistemas e Industrial
Bogotá, Colombia
2016

Análisis de la Deserción y Permanencia Académica en la Educación Superior Aplicando Minería De Datos.

ALVARO AGUSTIN OÑATE BOWEN

Tesis o trabajo de investigación presentado como requisito parcial para optar al título de:
MSc. en Ingeniería de Sistemas y Computación

Director (a):

Ing. Fabio Augusto González Osorio, PhD.
Profesor Titular

Línea de Investigación:
Computación aplicada

Universidad Nacional de Colombia
Facultad de Ingeniería
Departamento de Ingeniería de Sistemas e Industrial
Valledupar, Colombia

2016

Este nuevo logro se lo dedico a mis padres, Alvaro y Miriam, que siempre han estado ahí apoyándome, a mi esposa Heidy y mi hijo Santiago José, por su gran corazón, entendimiento y tolerancia; sin ellos no hubiese sido posible.

Nunca consideres el estudio como una obligación, si no como una oportunidad para penetrar en el bello y maravillosos mundo del saber.

Albert Einstein

Agradecimientos

Este trabajo de grado fue realizado gracias al convenio realizado por las instituciones de educación superior la Universidad Popular del Cesar y la Universidad Nacional de Colombia, sede Bogotá.

Agradecer a la Oficina de Registro y Control Académico de la universidad popular del cesar, por el préstamo de la base de datos de los estudiantes matriculados durante los periodos académicos del 2010 al 2014.

Al Doctor Fabio A. González, mi director de tesis, por su orientación experta, rigurosa, atenta y precisa. Agradezco sus inestimables críticas y comentarios en cada lectura, sus sugerencias oportunas y su apoyo en cada etapa del proceso.

.

Resumen

Este trabajo presenta el estudio de minería de datos en la educación para modelar la perdida de la condición académica para estudiantes matriculados en los programas de Ingeniería Electrónica e Ingeniería de Sistemas de la Universidad Popular del Cesar. Se utilizaron dos tareas de minería de datos. En primer lugar, una tarea descriptiva basada en el algoritmo K-medias, que fue utilizado para seleccionar varios grupos de estudiantes. En segundo lugar, una tarea de clasificación soportada en dos técnicas conocidas como árbol de decisión y Naïve Bayes para predecir la pérdida de la condición académica debido a los malos resultados durante los cuatro primeros semestres de un estudiante. Para el entrenamiento y prueba de los modelos, se utilizaron los expedientes académicos y los datos recogidos durante el proceso de admisión de los estudiantes y se evaluaron utilizando la técnica de validación cruzada. Los resultados experimentales han demostrado que la predicción de la pérdida de la condición académica se mejora cuando se añaden los datos de la matrícula académica anterior.

Palabras clave: Educación, Minería de Datos, deserción.

Abstract

This paper presents the study of data mining in education to model the loss of academic status for students enrolled in both the Electronic Engineering and System Engineering programs at Universidad Popular del Cesar. Two tasks of data mining were used. Firstly, a descriptive task based on the *K-means* algorithm, which was utilized to select several student clusters. Secondly, a classification task supported on two classification techniques, known as Decision Tree and Naïve Bayes to predict the loss of academic status because of poor performance in a student's first four semesters. The academic records and data collected during the admission process of those students were used to

train and test the models, which were assessed using cross-validation technique. Experimental results have shown that the prediction of loss of academic status is improved when the data from the previous academic enrollment are added.

Keywords: Education, Educational Data Mining, dropout

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Introducción

La minería de datos representa un avance computacional significativo en la obtención de información a partir de relaciones ocultas entre variables; esta disciplina tiene como objetivo la extracción de conocimiento útil de un alto volumen de datos en el cual inicialmente este conocimiento es desconocido, pero que al aplicar las técnicas de minería estas relaciones son descubiertas. La aplicación de las técnicas y herramientas de la minería de datos en los diferentes contextos educativos, es lo que se conoce como Educational Data Mining (EDM) o minería de datos en educación; puntualmente, el estudio del rendimiento académico es la investigación pionera (de 1993 a 1999) de esta disciplina [47].

El volumen de información almacenada en las bases de datos de las instituciones educativas es de gran utilidad en el proceso de enseñanza y aprendizaje; es por ello que en los últimos años se ha mostrado un interés significativo sobre el análisis de los datos que allí reposan. Este proyecto busca aplicar las técnicas de minería de datos a los registros académicos de los estudiantes de los programas de ingeniería de sistemas y electrónica que ingresaron en los periodos académicos del 2010-02 y 2014-01 a través de la construcción de un modelo de minería de datos descriptivo, que permite crear los diferentes perfiles de los estudiantes admitidos con la información socioeconómica y los resultados de las pruebas saber 11 recogidos durante el proceso de admisión. En este orden de ideas, se requiere la construcción de un modelo de clasificación para identificar a los estudiantes que presentan bloqueo académico con la información inicial y los registros académicos obtenidos durante cuatro matriculas académicas.
Para el desarrollo del proyecto se utilizó la metodología CRISP-DM que estructura el ciclo de vida de un proyecto de minería de datos en seis fases, descrita en cuatros niveles, que interactúan entre ellas durante el desarrollo del proyecto [18].

La primera fase de la metodología es la comprensión del negocio, se establecen las bases teóricas para el desarrollo del proyecto, tomando como referencia el análisis de los datos

en el sector educativo; esta fase inicial se enfoca en los objetivos del proyecto, la evaluación de la situación actual y la elaboración de un plan de trabajo para alcanzar los objetivos de minería de datos. La segunda fase de la metodología es la comprensión de los datos disponibles en la minería, que implica estudiarlos más de cerca. La compresión de datos conlleva accederlos y explorarlos; inicia con la tarea de recolección inicial, la descripción y la exploración del número de registros, atributos y formatos de los mismos y finalmente se analiza en detalle los problemas de calidad de estos. La siguiente fase es la preparación de los datos considerado una de las tareas más importantes debido a que cubre todas las actividades necesarias para la descripción del conjunto de datos como; la selección de tablas, registros, y atributos para la aplicación del modelo descriptivo y predictivo. Finalmente la fase de construcción y evaluación del modelo.

Este documento contempla el desarrollo del proyecto y está organizado de la siguiente forma:

El Capítulo 1 contiene una revisión del estado del arte de la minería de datos y uso de sus técnicas en el sector educativo.

El Capítulo 2 presenta la problemática de la deserción y permanencia académica en la Universidad Popular del Cesar, la evaluación de la situación actual y la determinación de los objetivos de minería de datos.

En el Capítulo 3 se efectúa el entendimiento de los datos; basado en la comprensión de los mismos donde se aplican técnicas de visualización con las ayudas de tablas y gráficos, esto con el fin de realizar una exploración preliminar de los datos. La preparación de estos cubre todas las actividades necesarias para la construcción del conjunto de datos finales, la selección de tablas, registros, y atributos.

Los capítulos 4 y 5 se enfocan en el diseño y evaluación de un modelo descriptivo y de clasificación. Finalmente en el Capítulo 6. Se presentan las conclusiones y los trabajos futuros.

1. Análisis de datos en el contexto educativo

En este capítulo se establecen las bases teóricas de minería de datos y la aplicación en el contexto educativo, conocido como la minería de datos en la educación, (Sigla en Inglés EDM, Educational Data Mining). Esta disciplina se fundamenta en el desarrollo de técnicas, modelos o herramientas de la minería de datos para el análisis y la exploración de datos. Seguidamente se realiza una revisión del estado del arte y aplicación de las tareas y técnicas de minería de datos en el contexto educativo. Por último se presenta el del capítulo.

1.1 Minería de datos en la educación

La minería de datos es definida como el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos. [1]. Este tema de estudio en la educación es un área de investigación que ha evolucionado en los últimos años y se desarrolla a la par de otras tecnologías o aplicaciones tales como, la minería de datos en la estadística, aprendizaje de máquina, y algoritmos de minería de datos.

Las contribuciones de la minería de datos en la educación han sido utilizadas para obtener una mejor comprensión del proceso educativo con el principal objetivo de proporcionar a los docentes e investigadores recomendaciones para el mejoramiento del proceso de enseñanza - aprendizaje. El objetivo de la EDM, es aplicar la minería de datos a los sistemas tradicionales de enseñanzas, en particular a los sistemas de gestión de contenidos de aprendizaje y sistemas educativos inteligentes basados en la web. Cada uno de estos sistemas tiene diferentes fuentes de datos para el descubrimiento de conocimiento. Después del pre-procesamiento de los datos en cada uno de estos sistemas se aplican las diferentes técnicas de minería de datos: estadísticas y visualización, agrupamiento y clasificación reglas de asociación y la minería de datos. [1].

Existen varios métodos y algoritmos en el proceso EDM; la clasificación es una tarea del modelo predictivo, su objetivo es identificar y representar las similitudes existentes entre los datos. El agrupamiento es una tarea del modelo descriptivo que permite la identificación de los grupos donde los elementos guardan gran similitud entre si y muchas diferencias con otros grupos. [2].

1.2 Aplicación de la minería de datos en la educación

La aplicación de la minería de datos en la educación es un campo de investigación reciente, el objetivo es proporcionar información a los docentes y administrativos para mejorar el aprendizaje de los estudiantes y organizar los recursos educativos de una manera más eficiente. Una revisión del estado del arte de las diferentes técnicas y algoritmos de minería de datos en la educación se presentan en la Tabla 1-1.

Técnicas de minería de datos en la educación	PREDICTIVO	DESCRIPTIVO	
	Clasificación	Regresión	Agrupamiento
Árbol de Decisión	[3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [17]		
K-Means			[13], [14], [15], [16]
Bayesiano	[4], [9], [10], [11], [17]		
Red Neuronal	[4], [11]		
Regresión Logística		[3], [4]	
Máquinas de Vectores	[4], [9]		

Tabla 1-1 Tareas y técnicas de minería de datos en la educación

Los autores Araque, Roldan y Salguero [3], realizaron un estudio sobre los factores que inciden en la deserción universitaria, a partir de la creación de un modelo de predicción que permitiera medir el riesgo de abandono de un estudiante con la información socioeconómica y los registros académicos, a través de la técnica de árbol de decisión y regresión logística, para cuantificar los estudiantes con alto riesgo de abandonar los estudios.

La principal problemática en las universidades que ofrecen educación a distancia es el gran número de estudiantes que no terminan el curso, los autores Kotsiantis, Pierrakeas y Pintelas [4] presentan el estudio de un algoritmo de aprendizaje para la predicción de la deserción de los estudiantes que abandonan sus estudios. Un gran número de

experimentos se llevaron a cabo con los datos proporcionados por la universidad en estudio, se compararon los algoritmos árbol de decisión, red neuronal, *Naive Bayes*, regresión logística y máquinas de vectores. El análisis de los resultados mostró que el algoritmo *Naive Bayes* es el más apropiado para predecir el rendimiento de los estudiantes en un sistema de educación a distancia.

Los autores Kuna, García y Villatoro [5] en su trabajo "El descubrimiento de conocimiento obtenido mediante el proceso de Inducción de árboles de decisión (Sigla en Inglés TDIDT, Top Down Induction of Decision Trees)", utilizaron árboles de decisión para modelar las clasificaciones de los datos. Uno de los principales resultados obtenidos fue la caracterización de los estudiantes con alto riesgo de abandonar sus estudios universitarios.

Kovacic [6] estudió las variables socioeconómicas como la edad, el género, la etnia, la discapacidad, la situación laboral y el programa de estudio a distancia. El objetivo de la investigación fue identificar a los estudiantes con alto riesgo de abandonar sus estudios. Se utilizaron las técnicas de minería de datos árboles de decisión y la regresión logística, presentando ambos algoritmos correcta clasificación.

Los autores Yadav, Bharadwaj y Pal [7] presentaron un proyecto de minería de datos para generar modelos predictivos e identificar a los estudiantes con alto riesgo de deserción teniendo en cuenta los registros de los estudiantes en la primera inscripción. La calidad de los modelos de predicción se examinó con los algoritmos *ID3, C4.5 y ADT* de las técnicas de árbol decisión. Los algoritmos de aprendizaje automático *ADT* pueden aprender de modelos predictivos con los datos de los estudiantes de los años anteriores. En este mismo sentido, los autores Quadril y Kalyankar [8] presentaron el estudio de la minería de datos para construir y evaluar un modelo predictivo para determinar la probabilidad de la deserción de un estudiante en particular, utilizaron la técnica de árbol decisión para clasificar a los estudiantes con la aplicación del algoritmo *C4.5*.

Los autores Zhang y Oussena [9] propusieron la construcción de un sistema de gestión basado en minería de datos (Sigla en Inglés MCMS, Mining Course Management Systems), una vez procesados los datos en el sistema de gestión, los autores identificaron las características de los estudiantes que no tuvieron éxito en el semestre, utilizaron tres

algoritmos de clasificación máquina de vectores, *Naive Bayes* y árbol de decisión. La mayor precisión en la clasificación se presentó con el algoritmo *Naive Bayes*, mientras que el árbol de decisión obtuvo uno de los valores más bajos.

Para mejorar la calidad del sistema de educación superior mediante la evaluación de los principales atributos que pueden afectar en el rendimiento de los estudiantes los autores Radaideh, Al-Sahwakf y Al-Najjar [10] presentaron un modelo de clasificación donde se mostraron los resultados de la evaluación del modelo utilizando los algoritmos *ID3 y C4.5* de las técnicas de árbol decisión y el *Naive Bayes*. La clasificación para los tres algoritmos no es muy alta, para generar un modelo de clasificación de alta calidad es necesario agregar los atributos suficientes.

En este mismo estudio los autores Yudkselturk, Ozekes y Kilic [11] examinaron la predicción de la deserción en los programas académicos en línea, con el fin de clasificar a los estudiantes que abandonaron sus estudios, se aplicaron tres técnicas de minería: árbol de decisión, *Naive Bayes* y red Neuronal. Estos algoritmos fueron entrenados y evaluados utilizando la técnica de validación cruzada. Por otro lado, el autor Pal [12] presentó una aplicación de minería de datos para generar modelos predictivos teniendo en cuenta los registros de los estudiantes del primer periodo. El árbol de decisión es utilizado para la validación y el entrenamiento para encontrar el mejor clasificador para predecir a los estudiantes que abandonaron sus estudios.

Los autores Bhise, Thorat y Supekar [13] estudiaron los factores de evaluación de los estudiantes para mejorar el rendimiento, utilizando técnica de agrupamiento mediante el análisis del algoritmo *K-Means*, para caracterizar la población de los estudiantes. Así mismo, los autores Erdogan y Timor [14] presentaron la relación de los estudiantes universitarios entre los exámenes de entrada y los resultados de éxito. El estudio fue realizado mediante técnicas de algoritmos de análisis de grupo y *K-Means*, estudiada mediante técnicas de minería de datos de agrupamiento. Los autores Bhardwaj y Bhardwaj [15] presentaron la aplicación de la minería de datos en el entorno de la enseñanza de la ingeniería, la relación entre la universidad y los resultados obtenidos por los estudiantes, mediante el análisis de las técnicas del algoritmo de *K-Means*.

En el entorno colombiano, el autor Timarán [16] describió el proceso de descubrimiento de conocimiento que se llevó a cabo en la Universidad de Antonio Nariño para determinar

en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil, para lo cual se utilizó la base de datos histórica de los estudiantes de pregrado. El objetivo de la investigación fue la búsqueda y descubrimiento de patrones de bajo rendimiento académico y deserción estudiantil aplicando técnicas de minería de datos, como clasificación, agrupamiento y reglas de asociación.

López [17] en su tesis de maestría en la Universidad Nacional de Colombia, propuso un acercamiento a la minería de datos en la educación mediante la aplicación de técnicas de agrupamiento y clasificación. Se construyeron diferentes modelos para la predicción de la pérdida de calidad de estudiante en diferentes escenarios, en los primeros cuatro semestres; en un periodo académico específico usando inicialmente los datos de admisión y luego los registros académicos.

1.3 Metodología CRISP-DM

La metodología *CRISP-DM* (Sigla en Inglés CRISP-DM, Cross Industry Standard Process for Data Mining) es un método probado para orientar trabajos de minería de datos. Incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre tareas.

La metodología consta de cuatro niveles de abstracción, organizados de forma jerárquica en tareas que van desde el nivel más general hasta los casos más específicos. A nivel más general, el proceso está organizado en seis fases, cada una de ellas a su vez estructurada en varias tareas generales. En el segundo nivel las tareas generales se proyectan a tareas específicas, donde se describen las acciones que deben ser desarrolladas para situaciones específicas [18] [19].

La primera fase de la metodología, incluye la comprensión del negocio o análisis del problema, la primera tarea de esta fase tiene como finalidad obtener la máxima información posible de los objetivos y requerimientos del proyecto, con el fin de convertirlos en objetivos de minería datos para definir una solución técnica al problema y producir un plan para el proyecto.

La segunda fase comprensión o análisis de los datos implica estudiar más de cerca los datos disponibles de minería, acceder a estos y explorarlos con la ayuda de tablas y gráficos para determinar la calidad de los mismos y describir los resultados de estos pasos en la documentación del proyecto. Esta fase inicia con la tarea de recolección inicial de datos, identificando la calidad de los mismos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

Una vez realizado el análisis de datos, la metodología establece que se proceda con la fase de preparación de los datos, que es uno de las más importantes y con frecuencia la que más tiempo exige. Esta inicia con la tarea de selección de los datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

En la fase de modelado se seleccionan las técnicas más apropiadas para el proyecto de minería de datos elegidas en función de los siguientes criterios: Ser apropiada al problema de investigación, disponer de los datos adecuados, cumplir los requerimientos del problema, el tiempo necesario para obtener un modelo y conocimiento de la técnica de minería de datos. Finalmente en la fase de evaluación, se valora el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del proyecto.

1.4 Resumen del Capítulo

En este capítulo se presentaron las bases teóricas sobre los trabajos relacionados con minería de datos para la educación. Se inició con la descripción del concepto de minería de datos, tareas y técnicas más representativas para su aplicación en la educación, seguidamente se presentó una revisión del estado del arte sobres algunos trabajos realizados de minería de datos en el sector educativo y desempeño académico. Para la planificación y dirección del proyecto se utiliza la metodología *CRISP-DM*, de tal manera que se logre hacer un correcto y buen seguimiento del mismo.

2. Deserción y permanencia académica en la educación superior

En este capítulo, se presenta la problemática en estudio de la deserción y permanencia académica en la educación superior, en el contexto de la facultad de Ingenierías y Tecnológicas de la Universidad Popular del Cesar y la descripción de los recursos disponibles que deben ser considerados para alcanzar los objetivos del proyecto.

2.1 Deserción estudiantil en la educación superior

Uno de los principales problemas que enfrenta el sistema de educación superior colombiano concierne a los altos niveles de deserción académica en las instituciones universitarias. Pese a que los últimos años se han caracterizado por aumento de cobertura e ingreso de estudiantes nuevos, el número de alumnos que logra culminar sus estudios superiores no es alto, dejando entrever que una gran parte de éstos abandona sus estudios, principalmente en los primeros semestres. Según estadísticas del ministerio de educación nacional, indican que de cada cien estudiantes que ingresan a una institución de educación superior cerca de la mitad no logra culminar su ciclo académico y obtener la graduación [20].

Se entiende por deserción como una situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, considerándose como desertor el estudiante de una institución de educación superior que no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica, y la permanencia por su parte, se refiere a la cantidad de tiempo que tarda el estudiante en terminar su programa académico y obtener el título que lo acredita como profesional [15].

2.2 Universidad Popular del Cesar

La Universidad Popular del Cesar, sede Valledupar, es una universidad de orden nacional de carácter pública, con seis facultades conformado por 19 programas de pregrado. El proceso de admisión se realiza cada semestre alrededor de 2.200 aspirantes en cada periodo académico, sin embargo, sólo cerca de 200 estudiantes se admiten en los programas de ingenierías de sistemas y electrónica. Los estudiantes admitidos son seleccionados en base exclusivamente en un solo criterio, su desempeño en las pruebas de estado Saber 11.

Las pruebas de estado Saber 11 conocida también como el examen de estado de la educación media o examen de ingreso para la educación superior, se aplica dos veces al año, una durante el primer semestre para colegios de calendario B y otra en el segundo semestre para colegios de calendario A; este examen es obligatorio para poder cursar estudios en una Institución de educación superior. Con la autonomía universitaria establecida en la Ley 30 de 1992, son las Instituciones de Educación Superior las que fijan los criterios para su uso. La prueba es realizada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), entidad especializada en ofrecer servicios de evaluación de la educación en todos los niveles y en particular apoyar al Ministerio de Educación Nacional de Colombia 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.

2.3 Evaluación de la situación

Los tipos de datos disponibles para el análisis son suministrados por la oficina de planeación de la Universidad Popular del Cesar, que contiene la información socioeconómica y los resultados de las pruebas saber 11 de los estudiantes admitidos y el historial académico del semestre anterior entre los periodos académicos del 2010 al 2014.

La información se encuentra consistente y actualizada en el sistema de información Academusoft. El acceso a las fuentes de la información está a cargo de la oficina de registro y control académico de la universidad.

Teniendo en cuenta las directrices institucionales, la pérdida de la condición de estudiante por bajo rendimiento académico en la Universidad Popular del Cesar, está determinado por el promedio ponderado semestral que se calcula multiplicando la calificación definitiva en cada asignatura por el número de créditos correspondientes, los productos obtenidos se suman, y el resultado de esta suma se divide por el número total de créditos de las asignaturas. El promedio es una medida que indica el rendimiento académico del estudiante durante su permanencia en la universidad. Al finalizar un periodo académico el estudiante que obtenga un promedio ponderado acumulado inferior a tres punto cero (3.0) y pierda el 60% de los créditos académicos matriculados en este periodo quedará por fuera de la universidad por bajo rendimiento académico.

En el acuerdo 015 de 11 de diciembre del 2002 en el artículo primero el consejo académico reglamentó que: quien haya perdido la calidad de estudiante por bajo rendimiento académico, podrá inscribirse y participar en igualdad de condiciones en los procesos de selección, admisión y matrícula que se realicen para ingreso en programas académicos que ofrece la universidad después de transcurrido como mínimo un periodo académico.

2.4 Objetivos del Proyecto

Teniendo en cuenta el conjunto de datos obtenido con la información socioeconómica, los resultados de las pruebas Saber 11 y expedientes académicos de los estudiantes entre los periodos del 2010 y 2014 de los programas de ingeniería de sistemas y electrónica de la Universidad Popular del Cesar, se plantea el desarrollo de los siguientes objetivos específicos:

- Construir un conjunto de datos de estudiantes que contemplen información de tipo socioeconómico, historial académico y desempeño en prueba de estado saber 11.
- Construir y evaluar un modelo descriptivo que permita realizar la caracterización de los estudiantes admitidos y los matriculados por cuatro periodos académicos.
- Construir y evaluar un modelo predictivo para predecir la pérdida de la condición académica de un estudiante en los primeros cuatro semestres.

Con el desarrollo de estos objetivos se podrá obtener información para identificar las características y perfiles de los estudiantes admitidos y clasificación de los estudiantes en los cuatro semestres académicos que presentaron bloqueo por bajo rendimiento académico.

2.5 Evaluación inicial de herramientas y técnicas

Se presume que la información proporcionada es correcta y verificable, pero existe la posibilidad que existan datos faltantes y anómalos, ya que muchos de estos datos son ingresados por los estudiantes en el momento de la matrícula y en algunos casos no diligencian completamente los formularios. Los recursos del software para el manejo de la base de datos es necesario, algunas herramientas computacionales utilizadas para la manipulación de la información, fue el sistema de gestión de base de datos MySQL Server.

Como herramienta de trabajo en el proceso de extracción de conocimiento se utilizó el software RapidMiner, que es un programa de aprendizaje automático para el análisis y minería de datos; permite el desarrollo de procesos, análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Esta herramienta cuenta con algoritmos de minería de datos que se prevén para el desarrollo de este proyecto.

2.6 Resumen del Capítulo

En este capítulo se presentó la fase de la comprensión del negocio, la problemática de la deserción y permanencia académica en la educación superior, en el contexto de la Universidad Popular del Cesar, convirtiendo este conocimiento en la definición de un problema de minería de datos. En la evaluación de la situación se relacionan los recursos disponibles para el desarrollo del proyecto. Finalmente se identificaron los objetivos de minería de datos, para la construcción y evaluación de un modelo descriptivo y predictivo

3. Entendimiento de los datos

Este capítulo se enfoca en la comprensión de los datos donde se aplican las técnicas de visualización como los histogramas, con el fin de realizar una exploración preliminar de los registros y verificar la calidad de los datos. Una vez realizado el análisis, se procede con la fase de preparación datos, que incluye las tareas de selección de los datos a los que se le van aplicar la técnicas de modelado para su respectivo análisis.

3.1 Recolección de los datos iniciales

El objetivo de esta primera tarea es obtener las fuentes de los datos del sistema de información académica de la Universidad Popular del Cesar. El primer conjunto de datos agrupan la información socioeconómica y el resultado de las pruebas saber 11 como son: Edad de ingreso, género, lugar de procedencia, estrato socioeconómico, origen étnico, estado civil, sexo, SISBEN, situación económica del estudiante (Dependiente, Independiente, Empleado y Otro), quien costea sus estudios (Padres, Becas, Empresa, Crédito Educativo o Usted Mismo), carácter del colegio (Académico, Académico y Técnico, o Técnico), tipo de inscripción (Hijo de Empleado, Mejor Puntaje ICFES, Preuniversitario o Inscripción Regular) y puntaje obtenido por el estudiante en la prueba saber 11 en el componente o núcleo (Biología, Filosofía, Inglés, Geografía, Física, Matemáticas, Química y Lenguaje).

El segundo conjunto de datos lo conforman el historial académico y de calificaciones obtenido por los estudiantes: El año y periodo académico de ingreso del estudiante, programa en el que está matriculado, situación académica del estudiante (Bloqueo académico por Bajo Rendimiento Académico y No Bloqueo académico), número de créditos matriculadas, aprobadas, perdidos y cancelado. El promedio acumulado de cada periodo académico que indica el rendimiento académico del estudiante durante su permanencia en la universidad. Finalmente la información académica de las asignaturas

cursadas en cada periodo académico, algunos de los campos con respectos a las asignaturas son: Código de la asignatura, nombre de la asignatura, número de Créditos, calificación numérica (0 a 5), materias tomadas, ganadas, perdidas y canceladas.

3.2 Descripción de datos

Las consultas generadas se realizaron a través del sistema de gestión de base de datos MySQL. Se realizó un proceso de concatenación de los dos conjuntos de datos, obteniendo un archivo plano con 55 atributos y 1665 registros de los estudiantes admitidos y matriculados de los programas de ingeniería de sistemas y electrónica. En la Tabla 3-1 y la Tabla 3-2 se presenta la descripción del conjunto de datos: la información socioeconómica y los registros académicos del estudiante en cada periodo o semestre, además se incluye la identificación o el nombre del atributo o campo, la descripción y el tipo de dato.

Identificación	Tipo de Dato	Descripción
Estado Civil	varchar2(15)	Estado Civil del Estudiante
Sexo	varchar2(3)	Tipo de Sexo
Edad de Ingreso	número (12)	Edad de Ingreso
Pago de los estudios	varchar2 (30)	Quien Costea los estudios
Situación Económica	varchar2 (15)	Situación Económica
Estrato	varchar2(3)	Extracto del Estudiante
SISBEN	varchar2 (1)	Tiene SISBEN
Lugar de Procedencia	varchar2(30)	Lugar de Procedencia
Nombre Colegio	varchar2(30)	Nombre de la Institución
Carácter del Colegio	varchar2(30)	Público o Privado
Puntaje Pruebas	número (2)	Puntaje de las Pruebas
Puntaje Matemáticas	número (2)	Puntaje de la Prueba de Matemáticas
Puntaje Química	número (2)	Puntaje de las Prueba de Química
Puntaje Física	número (2)	Puntaje de las Prueba de Física
Puntaje Lenguaje	número (2)	Puntaje de las Prueba de Lenguaje
Puntaje Ingles	número (2)	Puntaje de las Prueba de Ingles
Puntaje Historia	número (2)	Puntaje de las Prueba de Historia
Puntaje Filosofía	número (2)	Puntaje de las Prueba de Filosofía

Tabla 3-1. Descripción de los atributos socioeconómico - Pruebas Saber 11.

Identificación·¤	Tipo·de·Dato·¤	Descripción¤
Nombre·Programa¤	varchar2(30)¤	Programa·que·fue·admitido¤
Periodo·de·Ingreso¤	número·(2)¤	Año·de·Ingreso¤
Semestre·Académico¤	número·(2)¤	Semestre·actual·del·Estudiante¤
Promedio·Acumulado¤	número·(2)¤	Promedio·Ponderado·Acumulado··¤
Créditos·Aprobados¤	número·(12)¤	Número·de·Créditos·Aprobados·¤
Créditos·Cancelado¤	número·(12)¤	Número·de·Créditos·Cancelados·¤
Créditos·Perdidos¤	número·(12)¤	Número·de·Créditos·Perdidos·¤
Materias·Tomadas¤	número·(12)¤	Número·de·Materias·Matriculadas·¤
Materias·Ganadas¤	número·(12)¤	Número·de·Materias·Ganadas·¤
Materias·Canceladas¤	número·(12)¤	Número·de·Materias·Canceladas¤

Tabla 3-2 Descripción de atributos de los registros Académicos y de calificaciones.

3.3 Descripción de atributos

El conjunto de datos contiene 1665 registros con 37 atributos, donde 11 atributos son categóricos y 26 numéricos. La Tabla 3-3 muestra los resultados estadísticos de cada uno de los atributos numéricos de los estudiantes admitidos.

Atributo	Máximo valor	Mínimo valor	Mediana	Media	Desviación Estándar	Varianza	Moda
Edad de Ingreso	37	16	19	19	2.060	4.244	18
Estrato	6	1	1	2	0.706	0.499	1
Semestre Académico	10	1	3	3,4	2,7	7,2	1
Promedio Acumulado	4,8	0,0	3,3	3,1	0,8	0,6	3,3
Créditos Matriculados	201	1	43	60,5	49	2396,2	19
Créditos Aprobados	185	0	29	43,5	42,6	1715,6	8
Créditos Perdidos	82	0	13	16,1	13	167,7	11
Créditos Cancelados	10	0	0	0,8	1,3	1,7	0
Materias Matriculadas	65	1	15	20,3	15,1	228,3	7
Materias Ganadas	62	0	11	15,3	13,7	186,9	3
Materias Perdidas	23	0	4	4,6	3,8	14,1	3
Materias Canceladas	15	0	1	1,6	2,2	4,9	0,0

Tabla 3-3 Resumen Estadístico de los atributos numéricos

La Tabla 3-4 presenta el resumen estadístico de los resultados de las pruebas Saber 11, en el área de las matemáticas, biología, física, química, lenguaje, inglés y biología.

Resumen Estadísticos	Media	Mediana	Moda	Desviación Típica	Varianza	Mínimo	Máximo
Biología	49	49	48	7,14	51,04	33	93
Física	49	49	49	6,82	46,58	32	68
Ingles	46	46	46	7,05	49,69	29	69
Lenguaje	49	49	49	7,08	50,13	29	69
Matemáticas	51	51	51	7,70	59,32	31	71
Química	49	49	49	6,73	45,34	31	67

Tabla 3-4 Resumen Estadístico de los resultados de las pruebas Saber 11

3.4 Exploración de datos

El análisis exploratorio es una tarea que permite analizar en detalle algunas variables e identificar características; para este se utilizaron algunas de las herramientas de visualización como las tablas y gráficos, con el propósito de describir los objetivos de minería de datos de la fase de comprensión. A continuación se consideran los siguientes interrogantes que permiten dar claridad a los objetivos planteados.

1. ¿El lugar de procedencia influye en la situación económica del estudiante?

En la Tabla 3-5 se observa que el mayor promedio de la población presenta situación económica dependiente con un 76%; donde aproximadamente el 47% son de Valledupar y el 30% provienen de otra ciudad. El 13% son independientes, donde el 8% provienen de Valledupar y el 5% de otra ciudad y el 11% restante son empleados o presentan otra situación económica.

Situación Económica	Lugar de Procedencia		Total General
	Otra Ciudad	Valledupar	
Dependiente	29.76%	46.65%	76.41%
Empleado	2.87%	4.97%	7.84%
Independiente	4.85%	7.84%	12.69%
Otro	1.14%	1.92%	3.05%
Total General	**38.62%**	**61.38%**	**100%**

Tabla 3-5 Situación económica y lugar de procedencia

1. ¿El estrato socioeconómico influye en el rendimiento de los resultados de las pruebas saber 11.?

En la Tabla 3-6 se puede observar que los mayores resultados se presentan en las pruebas de matemáticas en cada uno de los estratos, los puntajes más bajos se evidencian en las

pruebas de inglés en los estratos 1,2 y 3. Las pruebas de química, física, lenguaje y biología presentan resultados muy similares en cada estrato socioeconómico.

Estrato Socioeconómico	Puntaje Promedio					
	Matemáticas	Química	Inglés	Física	Lenguaje	Biología
1	50.71	48.13	45.32	48.80	48.90	48.73
2	51.55	49.05	45.97	50.07	48.60	49.35
3	51.29	48.07	47.23	48.55	48.91	51.32
4	54.85	51.31	47.15	47.77	50.77	53.31
5	51.00	45.00	46.00	49.00	55.00	48.00
6	52.00	59.00	50.00	48.00	51.00	56.00

Tabla 3-6 Estrato socioeconómico puntaje prueba saber 11

2. ¿La forma de pago de los estudios influye positivamente en los resultados de las prueba saber 11?

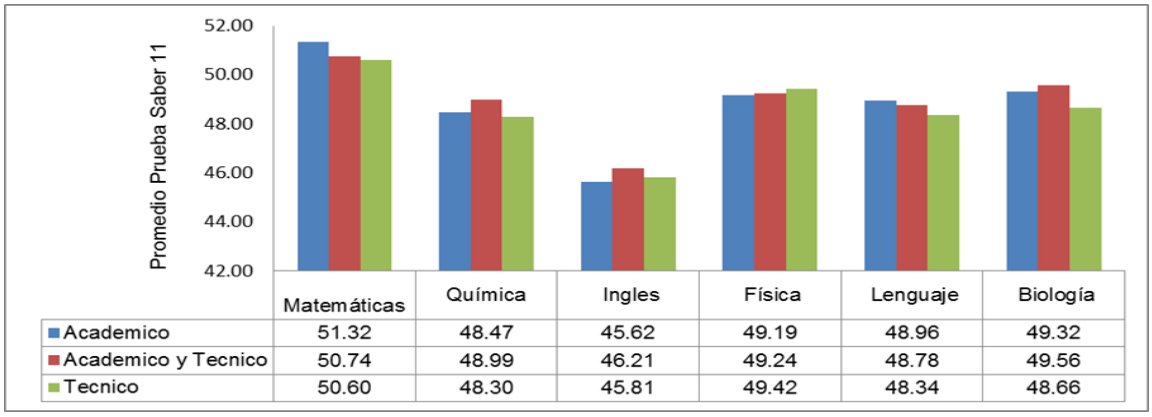
En la Tabla 3-7 se observa el puntaje promedio de las prueba Saber 11 vs quien costea los estudios. Los estudiantes que costean sus estudios con una beca registran mayor desempeño, en contraste con los estudiantes que pagan ellos mismos sus estudios, donde se evidencia que el nivel de desempeño es bajo en cada una de las áreas.

Quien Costea los estudios	Puntaje Promedio					
	Matemáticas	Química	Inglés	Física	Lenguaje	Biología
Beca	54.44	51.72	44.61	50.00	50.61	52.39
Crédito Educativo	49.58	46.08	43.63	46.71	48.75	50.50
Otro	50.56	48.96	46.93	47.20	50.15	48.40
Padres	51.53	48.56	45.85	49.58	48.58	49.23
Usted Mismo	45.42	46.76	42.42	49.00	49.24	49.91

Tabla 3-7 Quien Costea los estudios puntaje prueba saber 11

3. ¿El carácter el colegio es un factor determinante en los resultados de las prueba saber 11?

En la Figura 3-1 se puede observar que el carácter del colegio es un factor preponderante en los resultados de las pruebas en el área de inglés, donde se presentan los resultados más bajos. Sin embargo las matemáticas y física presentan los mayores resultados en colegios académicos, técnicos y la combinación de estos últimos.

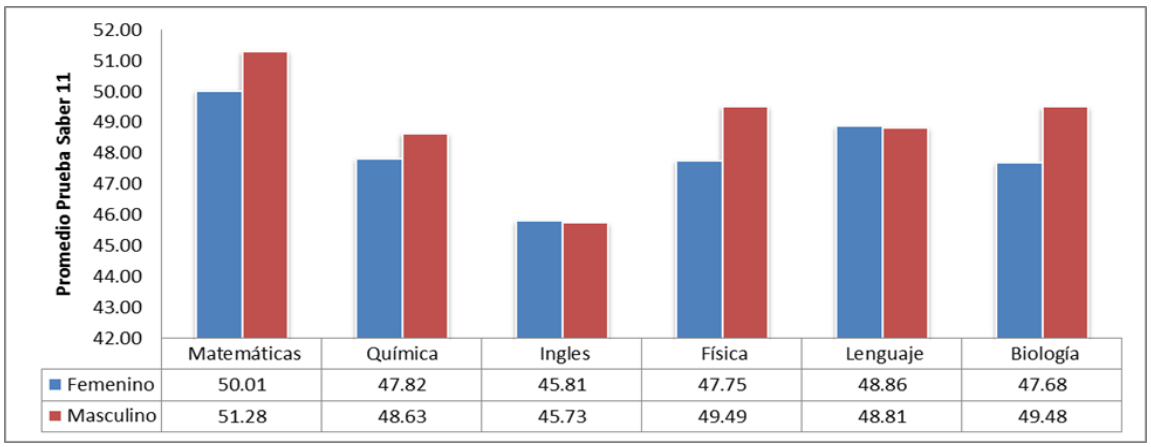


	Matemáticas	Química	Ingles	Física	Lenguaje	Biología
Academico	51.32	48.47	45.62	49.19	48.96	49.32
Academico y Tecnico	50.74	48.99	46.21	49.24	48.78	49.56
Tecnico	50.60	48.30	45.81	49.42	48.34	48.66

Figura 3-1 Carácter del colegio puntaje prueba saber 11

4. ¿El sexo es preponderante en la obtención de los resultados de las pruebas saber 11 en alguna área en particular?

En la Figura 3-2 se observa que los puntajes más altos en los resultados de cada una de las áreas los presenta el género masculino, mientras que en el área de inglés presentan el mismo promedio.



	Matemáticas	Química	Ingles	Física	Lenguaje	Biología
Femenino	50.01	47.82	45.81	47.75	48.86	47.68
Masculino	51.28	48.63	45.73	49.49	48.81	49.48

Figura 3-2 Sexo resultado de las prueba saber 11

3.5 Verificación de la calidad de los datos

La tarea de verificación de la calidad de los datos especifica una revisión de los mismos como los perdidos o los que tienen valores faltantes cometidos por errores de codificación. En esta sección se verifica la calidad de los datos correspondientes a la información

socioeconómica del estudiante admitido. En la Tabla 3-8 se muestran los atributos que presentan valores perdidos o faltantes de los estudiantes admitidos.

Atributo	Faltantes	Observación
ASPI_DIRECCIONRESIDENCIA	986	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_BARRIORESIDENCIA	785	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_TELEFONORESIDENCIA	1125	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_TELEFONOCELULAR	489	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_FECHANACIMIENTO	89	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_EMAIL	1080	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_ETNIA	1456	Vacío, en blanco, ?; No puede ser calculado por otro atributo
ASPI_TIPOSANGUINEO	1330	Vacío, en blanco, ?; No puede ser calculado por otro atributo
INST_NOMBREINSTITUCION	160	Vacío, en blanco, ?; No puede ser calculado por otro atributo

Tabla 3-8 Verificación de valores faltantes o perdidos de los estudiantes admitido.

Existen atributos como el SISBEN que contienen registros vacíos pero que se consideran no nulo, ya que estos campos contienen información de tipo Si y No, se infiere vacío cuando la respuesta es No; igual sucede con el atributo Etnia, que relaciona la etnia a la que pertenece o que no pertenece a ninguna etnia; existen otros atributos, adicionales que representan valores ID o códigos como; SNP, Codigoinstitucion, Pege_Id, Id_lugarnacimiento, Codigosnp, algunos son del tipo ordinal, estos atributos fueron eliminados por que pueden confundir al algoritmo de aprendizaje.

3.6 Selección de datos

En esta tarea se realiza el proceso de selección de los datos relevantes para el desarrollo de los objetivos de minería de datos. Un primer pre-procesamiento para la selección final de los datos, es la selección de atributos, en este se obtuvo que en el conjunto de datos existen 55 atributos o variables que contienen valores que pueden aportar o no al estudio, esto basado en la exploración inicial de los datos y en la descripción de los campos definidos en el diccionario de variables. En el conjunto de datos seleccionado para el modelamiento no se encontraron errores en los campos; a diferencias de los registros seleccionados, los errores que se presentaron en algunos casos fueron faltantes, debido a que el procesamiento no fue el adecuado al momento de la digitación como el correo electrónico, la dirección de residencia, número telefónico, fecha de nacimiento, tipo de

sangre, etnia y el número de la libreta militar que son atributos considerados no relevantes para el caso en estudio.

En La Tabla 3-9 se muestran los 24 atributos seleccionados; el primer conjunto de datos seleccionado lo conforman 14 atributos de los estudiantes admitidos en los periodos académicos del 2010-02 y 2014-01 con la información socioeconómica, los resultados de las pruebas saber 11. El segundo conjunto de datos lo conforman 10 atributos o campos del historial académico información obtenida en la matricula o periodo anterior.

Atributos Seleccionados	
Información Socioeconómica Resultado Pruebas saber 11	Lugar de Procedencia
	Carácter del Colegio
	Estado Civil
	Sexo
	Situación Económica
	SISBEN
	Edad de Ingreso
	Estrato
	Puntaje Pruebas de Biología
	Puntaje Pruebas de Física
	Puntaje Pruebas de Ingles
	Puntaje Pruebas de Lenguaje
	Puntaje Pruebas de Matemáticas
	Puntaje Pruebas de Química
Historial Académico	Situación
	Promedio Acumulado
	Créditos Matriculados, Ganado, Perdido
	Materias Matriculados, Ganado, Perdido

Tabla 3-9 Atributos Seleccionados para el procesamiento y modelado

3.7 Resumen del Capítulo

Con el desarrollo de la fase de comprensión y preparación de los datos se da por terminado el primer objetivo de la investigación, que inició con la preparación y recolección de los datos entre los periodos del 2010-1 al 2014-2 para ambos programas. En la tarea de descripción y exploración de los datos se analizaron en detalle algunas variables y se identificaron algunas características de los mismos. Se observó que el carácter del colegio y el estrato socioeconómico influyen significativamente en los resultados de las pruebas saber 11.

El proceso de verificación de los datos permitió conocer el conjunto de datos, identificando los atributos con valores perdidos o faltantes, algunos de ellos no pueden ser calculados, por lo que fueron eliminados se consideraron no relevantes para el caso de estudio. Finalmente se seleccionaron los datos para el procesamiento y la construcción del modelo que serán analizados en profundidad.

4. Modelo de minería de datos descriptivo

Este capítulo corresponde la fase de modelamiento descriptivo de minería de datos, la cual consiste en aplicar una técnica o algoritmo del modelo descriptivo a los datos seleccionados. En esta fase de la metodología fue necesario volver a la tarea de preparación de los datos para ajustarlos de acuerdo a los requerimientos del algoritmo seleccionado. Finalmente se presenta el proceso de diseño experimental y evaluación del modelo que permitirá alcanzar los objetivos propuestos.

4.1 Selección de la técnica del modelo descriptivo

La tarea del modelo descriptivo seleccionado es el de agrupamiento, dado que, en la revisión del estado del arte efectuado, se considera el más acorde con los objetivos de minería de datos propuestos en esta investigación. El algoritmo seleccionado es el K-Means no jerárquico, que es un algoritmo descriptivo de aprendizaje no supervisado. El objetivo es encontrar grupos con características similares, donde las distancias entre los grupos se maximizan y las distancias dentro los grupos se minimizan [21]. Para determinar la distancia entre grupos se utiliza la distancia euclidiana como medida de similitud de cada punto al centroide del grupo más próximo [22].

4.2 Diseño Experimental

Para la elaboración del diseño del modelo, se utilizó la aplicación RapidMiner para el aprendizaje automático para análisis y minería de datos, este programa permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Para la implementación del algoritmo se utilizó el operador K-Means de la biblioteca de *agrupación y segmentación* que utiliza la distancia euclidiana, para evaluar la calidad de los grupos encontrados. El algoritmo se encarga tanto de los valores numéricos y categóricos; sin embargo, fue necesario realizar un pre procesamiento

adicional para normalizar todos los atributos numéricos entre 0 y 1 con el operador *normalize*, todos los atributos deben tener la misma escala para una comparación justa entre ellos.

4.3 Construcción del modelo

Se aplicó un modelo de agrupamiento al conjunto de datos para la caracterización de los estudiantes admitidos, crear los diferentes perfiles de los estudiantes en los diferentes grupos encontrados y determinar que otros factores definen la separación de grupos producida por el algoritmo *K-Means*.

Para determinar cuál es el valor de K o el número de grupos que permita maximizar la agrupación del conjunto de datos se realizaron repetidas iteraciones donde el valor de k se varió de 2 a 14. Se evaluaron los resultados con base al error cuadrático de cada iteración; para la selección del número de grupo se utilizó el método del codo, que consiste en seleccionar el valor a partir del cual las variaciones no disminuyen de manera significativa.

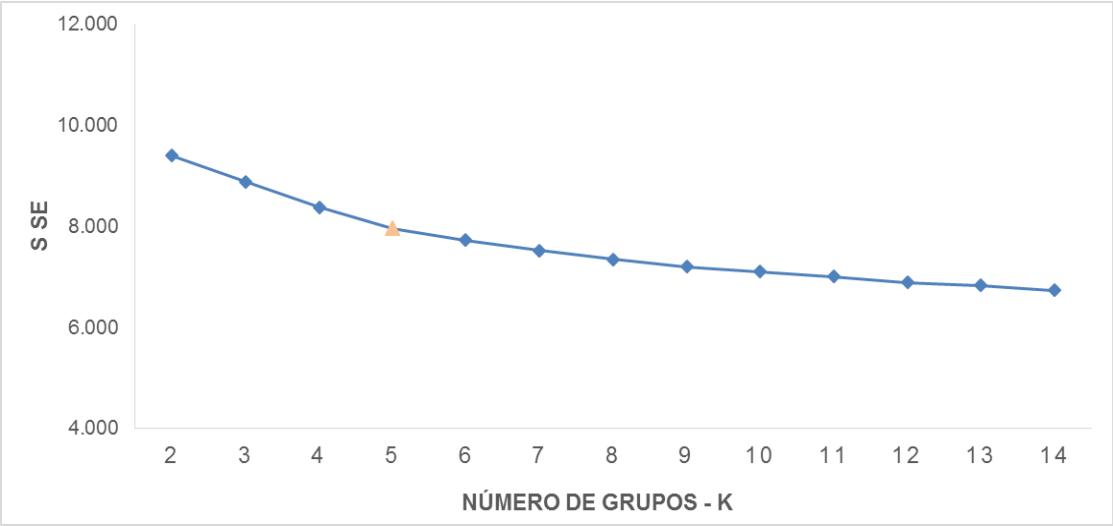


Figura 4-1 Selección del número de Grupo (K) para estudiantes Admitidos

En la Figura 4-1 se aprecian las iteraciones realizadas para hallar el valor de k en el primer conjunto de datos de los estudiantes admitidos, se seleccionó el k con valor de 5, donde el SSE es igual a 7.954. El algoritmo K-Means produjo un modelo con cinco grupos, a partir

de la descripción de estos grupos se espera poder caracterizar los perfiles de los estudiantes admitidos. En la Tabla 4-1 se muestra la distribución del número de registros y el porcentaje de cada uno de los grupos resultantes. El grupo 2 y 4 agrupan el mayor número de registros, por el contrario, el menor porcentaje de registros se encuentran en el grupo 1.

Grupos	Grupo_0	Grupo_1	Grupo_2	Grupo_3	Grupo_4
Número de Registros	317	130	418	389	409
Porcentaje	19%	8%	25%	23%	25%

Tabla 4-1 Distribución del número de registro en la aplicación del algoritmo K-Means

4.4 Análisis del modelo

Para la evaluación del modelo fue necesario "des-normalizarlo" con el operador *de-*normalize, para poner cada una de los valores de las variables en sus rangos originales. El análisis del modelo se realizó con la información socioeconómica y los resultados de las pruebas saber; luego, se realizó un análisis sobre la situación académica de los estudiantes que en cada grupo presentaban bloqueo académico con cuatro matriculas.

4.4.1 Análisis del modelo con respecto a la información socioeconómica y los resultados de las pruebas saber

En la Tabla 4-2 se muestra el puntaje promedio de las pruebas saber en cada grupo. El grupo 0 y 3 se caracteriza por presentar el mejor desempeño en las pruebas en todas las áreas, en contraste con lo observado en los grupos 1, 2 y 3. Por otro lado, se evidencia que las áreas de física, química y biología se caracterizan por presentar mayor promedio en cada uno de los grupos, sin embargo filosofía, geografía e inglés muestran los menores promedios.

Grupo	Puntaje Promedio							
	Biología	Filosofía	Física	Geografía	Ingles	Lenguaje	Química	Matemáticas
Grupo_0	54,5	49,8	53,5	55,7	51,5	54,3	57,2	54,1
Grupo_1	47,2	41,9	46,4	45,0	42,7	47,6	47,3	46,0
Grupo_2	46,0	41,0	46,4	41,2	41,7	45,6	46,5	44,5
Grupo_3	51,9	47,2	50,4	52,2	48,3	50,9	53,3	50,6
Grupo_4	46,6	41,3	48,8	43,9	43,9	46,4	50,2	47,1

Tabla 4-2 Puntaje promedio de las pruebas Saber 11

A continuación se presenta el análisis de la representación gráfica de la distribución de los valores de los diferentes atributos para cada uno de los grupos.

Grupo 0: Se caracteriza por agrupar el mayor promedio en las pruebas saber, el 65% de los estudiantes de este grupo provienen de Valledupar y el 35% de otra ciudad. El 65% pertenecen a colegios de carácter académico, 18% técnico y 17% académico y técnico. El 79% de la población presentan situación económica dependiente, el 54% de los estudiantes no registran SISBEN, la mayor concentración de estudiantes se encuentran en el estrato 1 con un 51%, en el estrato 2 un 37%, en el estrato 3 un 11% y en el estrato 4 sólo un 1%. (Ver Figura 4-2).

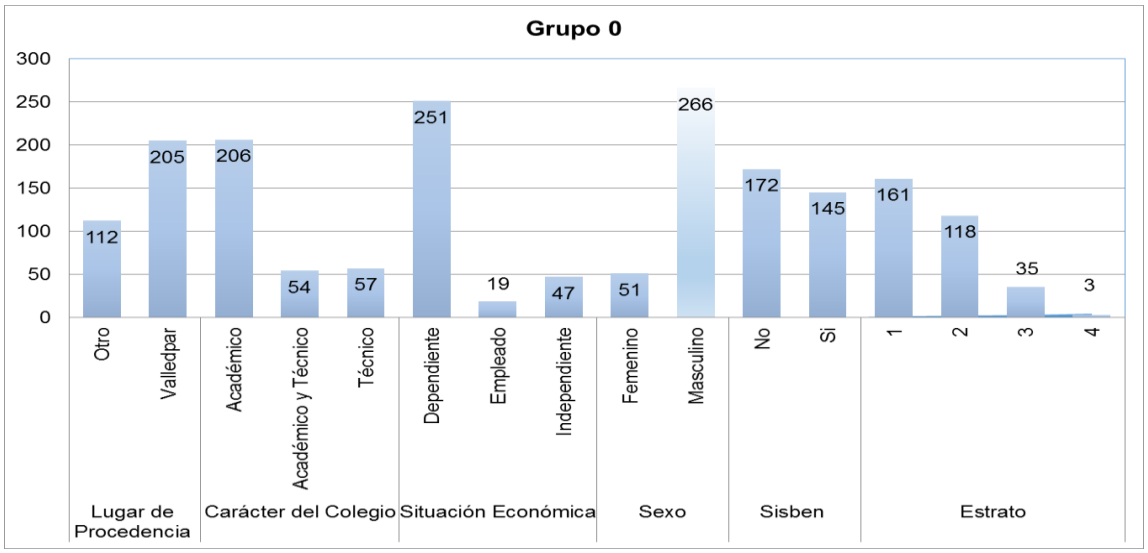


Figura 4-2 K-Means - Grupo 0 Distribución – Variables Socioeconómica.

Grupo 1: Este grupo presenta similitud con los grupos 2 y 4, se caracteriza por presentar el menor promedio en las pruebas saber 11. Este grupo presenta los menores resultados en el área de matemáticas. El 61% de los estudiantes provienen de Valledupar y el 39% de otra ciudad. También se observa que el mayor número de estudiantes son de colegios de carácter académico con un 73%. El 54% de la población presenta situación económica dependiente y el 62% no registran SISBEN. La mayor concentración de estudiantes se encuentra en el estrato 2 con un 46%, el estrato 1 con un 45%, el estrato 3 un 8% y en el estrato 4 un 2%. (Ver Figura 4-3).

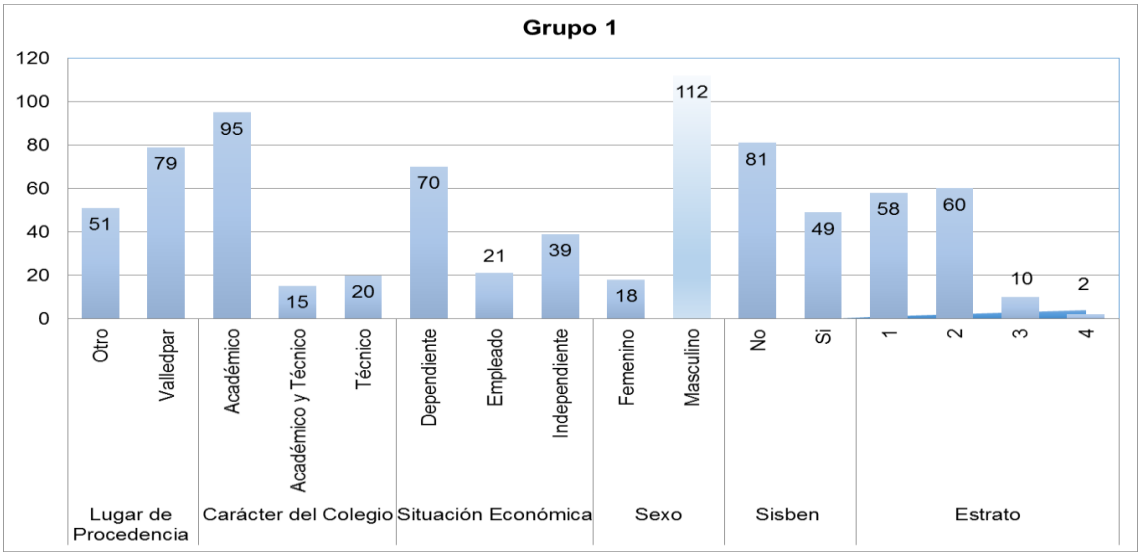


Figura 4-3 K-Means Grupo 1 Distribución – Variables Socioeconómica.

Grupo 2: Este grupo se caracteriza por agrupar el menor desempeño de las pruebas saber. El 69% de los estudiantes provienen de Valledupar y el 31% de otra ciudad. De forma similar a los grupos 1 y 4 el 64% de los estudiantes son de colegios de carácter académico, el 21% proceden de colegios técnicos y 15% de colegios académicos y técnicos. El 76% de los estudiantes presentan situación económica dependiente, el 8% son empleados y el 16% son independientes. La mayor concentración de estudiantes se presenta en el estrato 1 con un 53%, en el estrato 2 un 39% y en el estrato 3 un 8% (Ver Figura 4-4).

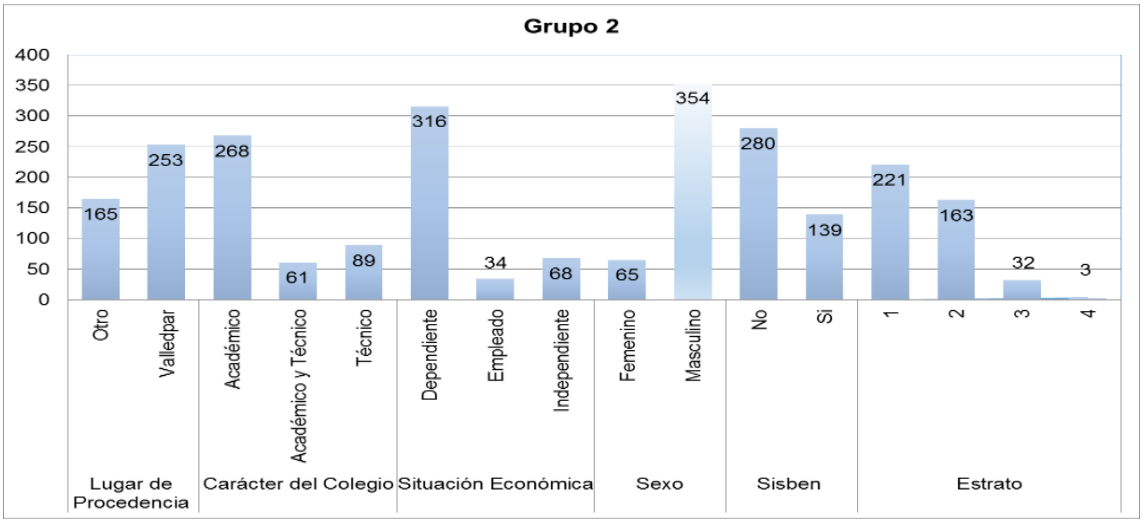


Figura 4-4 K-Means Grupo 2 – Distribución Estudiantes Admitidos

Grupo 3: Agrupa la población con el mejor desempeño en las pruebas saber en el área de las matemáticas con mayor promedio, de forma similar al grupo 0. El 60% de los estudiantes proceden de Valledupar y el 40% de otra ciudad, el 68% provienen de colegios de carácter académico. El 80% presentan situación económica dependiente, el 5% son empleados y el 15% son independientes. El 60% de los estudiantes no presentan SISBEN. La mayor concentración de estudiantes se presenta en el estrato 1 con un 48%, en el estrato 2 un 41%, en el estrato 3 un 10% y un 2% en el estrato 4. (Ver Figura 4-5).

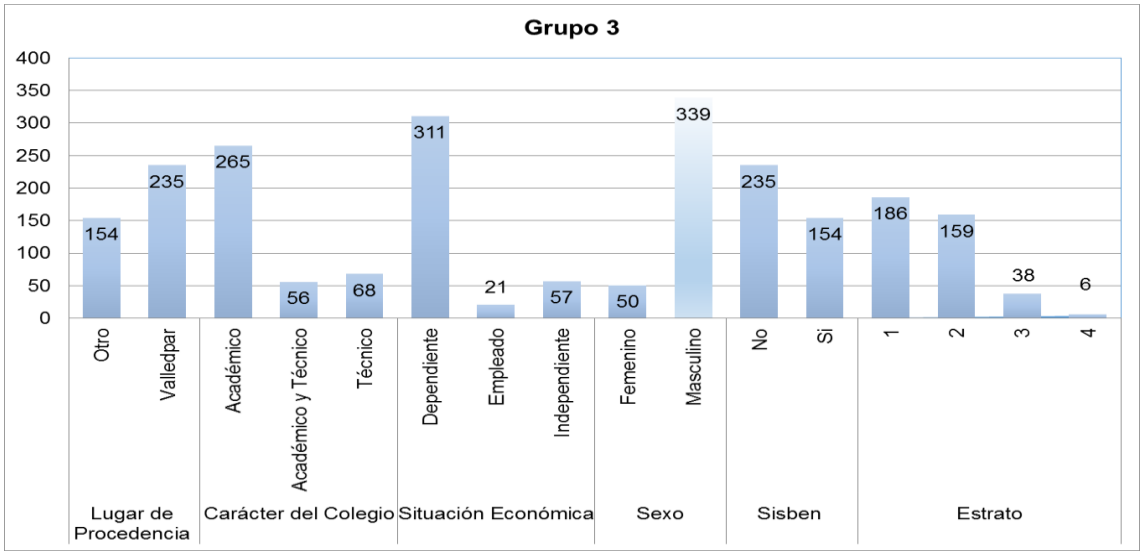


Figura 4-5 K-Means Grupo 3 – Distribución – Variables Socioeconómica.

Grupo 4: Este grupo presenta similitud con los grupos 1 y 4. Se caracteriza por agrupar el menor desempeño de las pruebas saber. El 61% de los estudiantes proceden de Valledupar y el 39% de otra ciudad; el 68% de los estudiantes provienen de colegios de carácter académico, el 22% de técnico y el 15% restante de académico y técnico. El 78% de los estudiantes presentan situación económica dependiente, el 9% están empleados y el 12% son independientes, el 64% de los estudiantes no registran tener SISBEN. Al igual que la mayoría de los grupos la mayor concentración de estudiantes son del estrato con un 57%, en el estrato 2 un 36% y en el estrato 3 un 7% (Ver Figura 4-6).

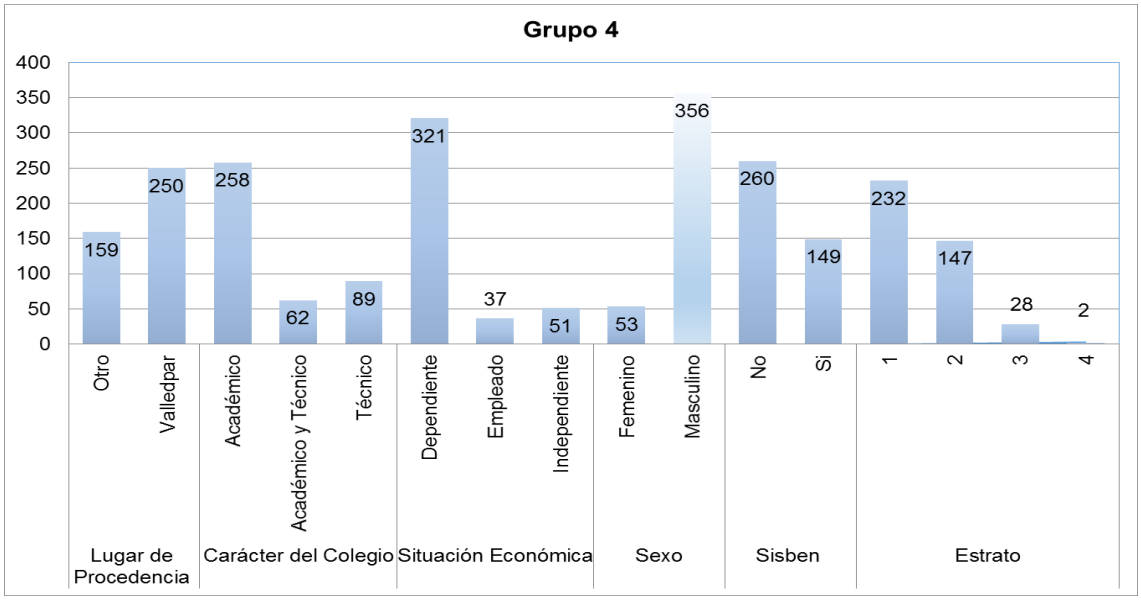


Figura 4-6 K-Means Grupo 4 Distribución – Variables Socioeconómica.

4.4.2 Análisis del modelo con respecto al desempeño académico con cuatro matriculas.

En la siguiente sección se analizó el modelo para examinar la pérdida de la condición académica de los estudiantes que presentaron bloqueo académico en los primeros semestres académicos, con el fin de comparar la población inicial de aquellos que continúan sus estudios después de tres períodos académicos.

Nro. De Grupos	Matricula 1	Matricula 2	Matricula 3	Matricula 4	Total
GRUPO_0	32	13	11	10	66
GRUPO_1	127	63	47	29	266
GRUPO_2	178	58	72	23	331
GRUPO_3	160	90	48	41	339
GRUPO_4	66	32	45	52	195
Número de Registros	563	256	223	155	1197
Porcentaje	47%	21%	19%	13%	100%

Tabla 4-3 Distribución del número de estudiantes con cuatro matriculas.

En la Tabla 4-3 se presenta la distribución del número de registros en cada uno de los grupos en los primeros cuatro semestres o matriculas académicas. Los grupos 2 y 3 se caracterizan por agrupar el mayor número de registros con un 28% en cada grupo. En el grupo 0 por el contrario se encuentra el menor número de registros con un 6%. El 47% de los registros son estudiantes del primer semestre, el 21% presentan segunda matricula, el 19% con tercera matricula y el 13% de los registros restantes presentan cuatro matriculas académica.

Nro. De Grupos	Matricula 1		Matricula 2		Matricula 3		Matricula 4	
	No BLQ.	BLQ.Acad	No BLQ.	BLQ.Acad	No BLQ.	BLQ.Acad	No BLQ.	BLQ.Acad
GRUPO_0	23%	26%	18%	2%	14%	3%	15%	0%
GRUPO_1	17%	30%	13%	11%	15%	2%	9%	2%
GRUPO_2	43%	11%	15%	3%	22%	0%	7%	0%
GRUPO_3	22%	26%	21%	6%	14%	0%	11%	1%
GRUPO_4	16%	17%	12%	5%	23%	0%	26%	1%

Tabla 4-4 La agrupación de los estudiantes con cuatro matriculas académicas

En la Tabla 4-4 se observa en cada grupo la condición académica de los estudiantes en los primeros cuatro matriculas. El grupo 1 se caracteriza por agrupar el mayor porcentaje de estudiantes con bloqueo académico, en contraste con el grupo 2 donde se puede ver el menor porcentaje de estudiantes con bloqueo académico.

El grupo 0 se caracteriza por presentar buen desempeño en las pruebas saber 11, agrupan el 26% de los estudiante con bloqueo en la primera matricula, el 2% en la segunda y el 3% en la tercera matricula. El grupo 1 agrupa los estudiantes con el menor desempeño de las pruebas saber 11 similar al grupo 2. El 30% de los estudiantes presentan bloqueo en la primera matricula, el 11% en la segunda y el 2% son de la tercera y cuarta matricula.

El grupo 2 se caracteriza por agrupar los estudiantes con el menor desempeño de las pruebas saber 11 y el menor número de estudiantes con bloqueo. El 11% de los estudiantes presentan bloqueo en la primera matrícula y el 3% en la segunda matricula.

El grupo 3 se caracteriza por agrupar los estudiantes con buen desempeño en las pruebas saber 11 similar al grupo 0. El 26% de los estudiantes presentan bloqueo en la primera matrícula, el 6% con dos y el 1% con cuatro matriculas. Finalmente el grupo 4 se caracteriza por agrupar el menor número de estudiantes con bloqueo. El 17% de los estudiantes con una matrícula presentan bloqueo, el 5% corresponde a los estudiantes con dos matrículas y el 1% con cuatro matriculas.

4.5 Resumen del Capítulo

En este capítulo se presentó la construcción de un modelo de agrupamiento con la aplicación del algoritmo K-Means para crear los diferentes perfiles de los estudiantes del programa de ingeniería de sistemas y electrónica de la Universidad Popular del Cesar.

Se realizó una evaluación sistemática del modelo propuesto, el primer análisis se realizó con la información socioeconómica y los resultados de las pruebas saber 11, en donde se pudo evidenciar que el entorno socioeconómico del estudiante incide en los resultados de su desempeño académico. Los estudiantes más representativos con altos niveles de desempeño en las pruebas saber 11 provienen de Valledupar, de colegios de carácter académico tienden a ser los que tienen una situación económica dependiente y pertenecen a los estratos socioeconómico 1 y 2.

De otra parte, también se analizaron los grupos resultantes con respecto al desempeño académico en las primeras cuatro matriculas. El grupo 2 se caracteriza por agrupar a los estudiantes con el menor desempeño de las pruebas saber 11 presentando el menor porcentaje de estudiantes con bloqueo. El 11% presentan bloqueo en la primera matrícula y el 3% de la segunda matricula.

5. Modelamiento Predictivo

Este capítulo presenta la metodología planteada por los autores López, Guzmán & González [34], en la cual proponen dos modelos de minería de datos para analizar los datos académicos y no académicos de los estudiantes de la Universidad Nacional de Colombia; los modelos utilizaron dos técnicas de clasificación, árbol de decisión y *Naïve Bayes*, con el fin de predecir la perdida de la condición académica por bajo rendimiento académico en las primeras cuatro matrícula del estudiante. Los expedientes académicos históricos y los datos recogidos durante el proceso de admisión se utilizaron para entrenar los modelos, los cuales fueron evaluados utilizando la validación cruzada.

El modelo de clasificación propuesto en esta investigación utiliza la información socioeconómica, los resultados de las pruebas saber 11 y los expedientes académicos de los estudiantes de la Universidad Popular del Cesar. La Tabla 5-1 presenta el número total de registros o estudiantes con primera matricula o inscripción y número de estudiantes con bloqueo académico por bajo rendimiento.

Condición Académica	Periodo Académico									
	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2
No Bloqueo	115	145	137	119	100	146	131	151	110	166
Bloqueo Acad.	70	32	49	30	31	22	25	35	27	23
Total Registros	185	177	186	149	131	168	156	186	137	189

Tabla 5-1 Número de matrícula y bloqueos académicas por período académico

La Tabla 5-2 muestra el número de estudiantes con bloqueo académico en cada periodo o matricula. El mayor número de estudiantes con bloqueo académico se presenta en la primera matricula. La segunda, tercera y cuarta matricula se evidencia una disminución del número de estudiantes con bloqueo. En el periodo de ingreso 2010-01 se presentó el mayor número de estudiantes con bloqueo académico en cada matricula académica.

Periodo Ingreso	Bloqueo Académico									
	2010-1	2010-2	2011-1	2011-2	2012-1	2012-2	2013-1	2013-2	2014-1	2014-2
2010-1	**40**	15	7	5	0	1	2	0	0	0
2010-2		**28**	0	0	3	0	0	0	0	0
2011-1			**39**	6	2	2	0	0	0	0
2011-2				**22**	8	0	0	0	0	0
2012-1					**20**	11	0	0	0	0
2012-2						**14**	8	0	0	0
2013-1							**15**	10	0	0
2013-2								**28**	7	0
2014-1									**26**	1
2014-2										**23**

Tabla 5-2 Bloqueo académico por período de Ingreso o primera inscripción.

5.1 Selección de la Técnica de Modelado

El modelo de clasificación utiliza dos técnicas ampliamente usadas, árboles de decisión y un clasificador bayesiano; la razón para seleccionar estos algoritmos es su gran sencillez e interpretabilidad.

El árbol de decisión es la primera técnica utilizada para la clasificación de los datos, este algoritmo genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información es decir, elige al atributo que mejor clasifica los datos [25, 26].

Es una técnica donde una instancia es clasificada siguiendo el camino de condiciones, desde la raíz hasta llegar a una hoja, la cual corresponderá a una clase etiquetada. Un árbol de decisión se puede convertir fácilmente en un conjunto de reglas de clasificación [27]. El algoritmo más representativo es el C4.5 que maneja atributos tanto categóricos como continuos, genera un árbol de decisión de forma recursiva al considerar el criterio de la mayor proporción de ganancia de información [28]. El nodo raíz será el atributo cuya ganancia es máxima. El algoritmo C4.5 utiliza poda pesimista para eliminar ramas innecesarias en el árbol de decisión y para mejorar la exactitud de la clasificación [24].

La segunda técnica a considerar para la construcción del modelo es un clasificador bayesiano, que es uno de los modelos de clasificación más efectivos [35], [36]. Los clasificadores bayesianos están basados en las redes bayesianas [37], éstos son modelos

gráficos probabilísticos que permiten modelar de una forma simple y precisa la distribución de probabilidad subyacente a un conjunto de datos. Las redes bayesianas son representaciones gráficas de las relaciones de dependencia e independencia entre las variables presentes en el conjunto de datos que facilitan la comprensión e interpretabilidad del modelo. Para estimar estas probabilidades se han propuesto numerosos algoritmos, entre los que cabe destacar el *Naive Bayes*, que es uno de los algoritmos de aprendizaje práctico más utilizados por su sencillez, su simplicidad, resistencia al ruido, poco tiempo para el procesamiento y alto poder predictivo. [29].

5.2 Modelo de predicción

Para predecir si un estudiante va ser bloqueado en una matrícula en particular fueron entrenados y probados diferentes modelos. El primer modelo analizó la pérdida de la condición académica con base a la información socioeconómica y los resultados de las pruebas saber 11 recogidos durante el proceso de admisión. El segundo modelo fue analizado con la información inicial del proceso de inscripción y los registros académicos de las primeras cuatro matriculas. La Tabla 5-3 describe el número de registros en las primeras cuatro matriculas con situación académica (No Bloqueo y Bloqueo Académico).

Situación Académica	Matricula 1	Matricula 2	Matricula 3	Matricula 4
No Bloqueo	309	190	214	145
Bloqueo Académico	255	66	9	10
Total de Registros	564	256	223	155

Tabla 5-3 Situación académica en las primeras cuatro matriculas.

5.2.1 Predicción de la pérdida de la condición académica utilizando los datos de entrada del proceso de admisión.

En este modelo se analizó la pérdida de la condición académica en un periodo determinado durante las primeras cuatro matriculas utilizando la información inicial o de entrada recogidos en el proceso de admisión. La Tabla 5-4 muestra las diferentes configuraciones, en cada matricula se representa la predición con flechas y los datos de entrada utilizados para entrenar el modelo.

Información de Ingreso	Matricula 1		Matricula 2		Matricula 3		Matricula 4	
Socioeconómica y Resultados Pruebas Saber 11.	Datos Ingreso	BLQ. Acad.	Datos Ingreso	BLQ. Acad.	Datos Ingreso	BLQ. Acad	Datos Ingreso	BLQ. Acad.
		⇧		⇧		⇧		⇧

Tabla 5-4 Modelo de predición de la pérdida de calidad de estudiante con la información de ingreso

En cada una de las técnicas seleccionadas se utilizó inicialmente la información de entrada del proceso de admisión. En la Tabla 5-5 se relacionan los atributos de entrada del modelo en cada una de las cuatro matriculas.

Información de Entrada del Modelo Proceso de Admisión	
Atributos de Entrada Socioeconómico y Resultados de las Pruebas Saber 11	Lugar de Procedencia, Carácter del Colegio, Estado Civil, SISBEN, Sexo, Programa, Situación Económica, Estrato, Edad de Ingreso, Resultado de las Pruebas Saber 11: Física, Ingles, Matemáticas, Biología, Filosofía, Geografía, Lenguaje, Ingles y Química.

Tabla 5-5 Información de Ingreso del proceso de admisión

5.2.2 Predición de la pérdida de la condición académica utilizando los datos académicos.

Para predecir la pérdida de la condición académica en una matrícula en particular el modelo utilizó como entrada la información inicial del proceso de admisión y el historial académico de la matricula anterior como; el promedio acumulado, total de créditos y materias (Matriculadas, Cancelados, Ganadas y Perdidas), la situación académica del estudiante (No Bloqueo y Bloque Académico por bajo rendimiento académico). Para hacer una predicción en un semestre, se utilizaron los datos de entrada del proceso de admisión y el historial académico del semestre anterior. La Tabla 5-6 muestra las diferentes configuraciones, en cada matricula se representa la predicción con flechas, los atributos de entrada y el historial académico del periodo anterior utilizados para entrenar el modelo.

Información de Ingreso del Modelo	Matricula 1		Matricula 2		Matricula 3		Matricula 4	
	Datos Ingreso	BLQ. Acad	Datos Ingreso	BLQ. Acad	Datos Ingreso	BLQ. Acad	Datos Ingreso	BLQ. Acad
Socioeconómica y Resultados Pruebas Saber 11 - Registros Académicos				⇧				
						⇧		
								⇧

Tabla 5-6 Modelo de predición de la pérdida de calidad de estudiante en un semestre determinado

5.3 Validación y diseño experimental

Después de la selección de las técnicas y la construcción del modelo se procedió a describir la forma cómo se divide el conjunto de datos disponibles. El primer conjunto corresponde a los datos de prueba y el segundo conjunto a los datos de entrenamiento y validación. Para la configuración de los experimentos se utilizó la validación cruzada con diez iteraciones para entrenar el modelo. La validación cruzada divide los datos etiquetados en conjuntos de entrenamiento y de prueba. Los modelos se aprenden sobre los datos de entrenamiento y se aplican sobre los datos de prueba. El conjunto de datos se dividió en diez grupos distribuidos por igual; nueve de ellos corresponde al conjunto de entrenamiento, el modelo se evaluó en el décimo grupo, que corresponde al conjunto de datos de validación.

5.3.1 Diseño experimental

Para la elaboración del diseño del modelo, se utilizó la aplicación *RapidMiner*; que es un programa para el aprendizaje automático y proceso de minería de datos, a través de un concepto modular, que permite el diseño de modelos de aprendizaje empleando operadores de cadena para diversos problemas. Para la validación del modelo de clasificación se utilizó la técnica de muestreo estratificado *Stratified Sampling.* El operador para realizar la partición del conjunto de datos de denomina *Split Data*; este operador crea partición al conjunto de datos en subconjuntos de acuerdo con el tamaño definido y la técnica seleccionada. Para la implementación del algoritmo árbol de decisión se utilizó el operador *Decisión Tree* y el algoritmo bayesiano *Naive Bayes.* La Tabla 5-7 presenta el número de registros en las primeras cuatro matricula, se tomó el 80% de los registros como conjunto de entrenamiento y validación cruzada de 10-fold y el 20% de la muestra fue empleada como conjunto de prueba.

Números de Matriculas	Total de Registros	Datos de Entrenamiento y Validación 80%		Datos de Prueba 20%	
		No Bloqueo	Bloqueo Acad.	No Bloqueo	Bloqueo Acad.
Matricula 1	564	247	204	62	51
Matricula 2	256	152	53	38	13
Matricula 3	223	171	7	43	2
Matricula 4	155	116	8	29	2

Tabla 5-7 Conjunto de datos de Prueba, de Entrenamiento y de Validación.

5.3.2 Medida de desempeño del modelo

Para estimar el rendimiento del modelo se utilizó el operador *X-Validation.* Este operador permite definir el proceso de validación cruzada con 10-fold sobre el conjunto de datos de entrada para evaluar el algoritmo de aprendizaje. El desempeño del modelo se midió con el operador *Performance Binomio Clasificación*, este operador presenta los resultados de desempeño del algoritmo en términos de exactitud, precisión, recall, error y curva ROC. Para analizar los errores generados a partir de un modelo clasificación se emplea la matriz de confusión [38]. Es una herramienta de visualización que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, como se muestra en la Tabla 5-8.

Predicción	***Positivo***	***Negativo***
Positivo	*Verdadero Positivo (VP)*	*Falso Negativo (FN)*
Negativo	*Falso Positivo (FP)*	*Verdadero Negativo (VN)*

Tabla 5-8 Matriz de confusión para la evaluación del modelo

De la tabla anterior se extraen las siguientes medidas:

a) **Exactitud (Accuracy):** Es la proporción del número total de predicciones que son correctas.

$$\text{Exactitud} = (\text{VP} + \text{VN})/\text{N} \;;\; \text{N} = \text{FN} + \text{FP} + +\text{VP} + \text{VN}$$

Donde N es el total de ejemplos del conjunto de validaciones

b) **Error de Clasificación:** Es la proporción del número total de predicciones que son incorrectas.

$$\text{Error} = (\#\ \text{errores})/(\#\ \text{ejemplo}) \;=\; (\text{FN} + \text{FN})/\text{N}$$

c) **Exhaustividad (Recall):** Es la Proporción de casos positivos que fueron clasificados correctamente

$$\text{Recall} = \text{VP}/(\text{VP} + \text{FN})$$

d) **Precisión:** Es la predición de casos positivos que fueron clasificados correctamente.

$\text{Precision} = \text{VP}/(\text{VP} + \text{FP})$

e) **Medida – F (f_measure):** Es una Combinación de la Precisión y el Recall.

$\text{f} = \frac{2\text{pr}}{\text{p} + \text{r}}$; Donde f, r y p son f_measure, recall y precisión, respectivamente.

f) **Especificidad (Specificity):** Es la Proporción de casos negativos que fueron clasificados correctamente.

$\text{Especificidad} = \text{VN}/(\text{FP} + \text{VN})$

g) **Sensibilidad (Sensitivity):** Es la Proporción de casos positivos que fueron clasificados correctamente.

$\text{Sensibilidad} = \text{VP}/(\text{VP} + \text{FN})$; Este parámetro es el mismo que el Recall.

h) **Tasa de Falsos Negativos (FN):** Es la Proporción de casos positivos que son clasificados incorrectamente como negativos.

$\text{FN} = \text{FN}/(\text{VP} + \text{FN})$

i) **Tasa de Falsos Positivos (FP):** Es la Proporción de casos negativos que son clasificados incorrectamente como positivos.

$\text{FP} = \text{FP}/(\text{VN} + \text{FP})$

j) **Área bajo la Curva (AUC):** Es una representación gráfica de la sensibilidad de los verdaderos positivos $\text{VP}/\text{VP} + \text{FN})$ Vs los falsos positivos $\text{FP}/\text{FP} + \text{VN})$. El punto (0,1) se llama una clasificación perfecta. La línea diagonal que divide el espacio de la ROC en áreas de la clasificación buena o mala. Los puntos por encima de la línea diagonal indican buenos resultados de clasificación, mientras que los puntos por debajo de la línea indican resultados equivocados. En la Figura 5-1 se presenta una representación gráfica de una curva ROC.

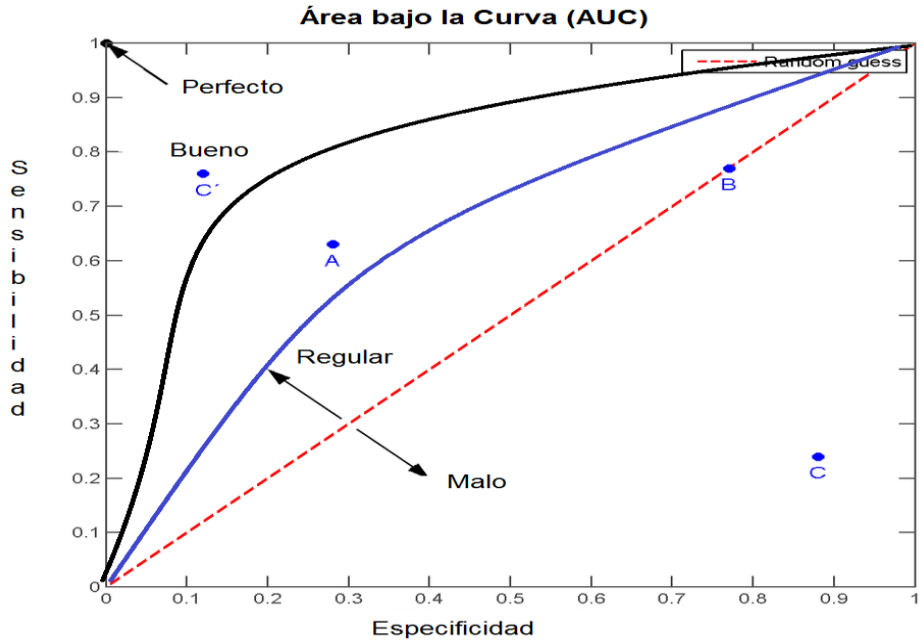


Figura 5-1 Representación gráfica Área bajo la Curva (AUC)

5.3.3 Resultado y evaluación de la predicción de pérdida de la condición académica utilizando los datos de entrada del proceso de admisión.

En esta etapa fueron entrenados y probados diferentes modelos, para clasificar estudiantes con bloqueo académico en las primeras cuatro matriculas académicas; mediante la utilización de la información socioeconómica y los resultados de las pruebas saber 11. Para la configuración de los experimentos, se utilizó la validación cruzada con 10-fold para entrenar los modelos y la evaluación del modelo se utilizó el conjunto de datos de prueba.

El desempeño del modelo fue evaluado con el 80% de los datos de entrenamiento y validación, el 20% de la muestra fue empleada como conjunto de prueba. En la técnica de árbol de decisión con los datos entrenamiento y validación se varió la profundidad del árbol de 1 hasta 20, el menor error de clasificación se encontró en la profundidad 3, donde el error comienza a mostrar cierta estabilidad en cada uno de los cuatro periodos académicos. Finalmente los modelos de entrenamiento y validación fueron evaluados con el conjunto de datos de prueba.

En la Tabla 5-9 se presentan los resultados del modelo de predición de la perdida de la condición académica con los datos de entrenamiento y validación, se comparan las diferentes técnicas de clasificación en cuanto a los diferentes parámetros de desempeño.

Predicción	Árbol de Decisión				Naive Bayes			
	Matricula 1	Matricula 2	Matricula 3	Matricula 4	Matricula 1	Matricula 2	Matricula 3	Matricula 4
Medida – F	0	0	30.77%	0	53.41%	38.51%	37.50%	22.22%
Precisión	0	0	33.33%	0.00%	56.27%	44.17%	33.33%	20.00%
Exhaustividad	0.00%	0.00%	28.57%	0.00%	51.45%	36.00%	42.86%	25.00%
Exactitud	54.76%	74.14%	94.93%	92.76%	59.43%	69.81%	94.41%	88.65%
Error	45.24%	25.86%	5.07%	7.24%	40.57%	30.19%	5.59%	11.35%
Curva (AUC)	0.500	0.500	0	0	0.608	0.630	0	0
Kappa	0.000	0.000	0.282	-0.015	0.177	0.190	0.346	0.119
Especificidad	100%	100%	97.61%	99.23%	66.02%	81.65%	96.50%	93.09%
Sensibilidad	0.00%	0.00%	28.57%	0.00%	51.45%	36.00%	42.86%	25.00%
Falsos Positivos	0%	0%	2%	1%	19%	14%	3%	6%
Falsos Negativos	45%	26%	3%	6%	22%	17%	2%	5%

Tabla 5-9 Modelo de predición de la pérdida de la condición académica con el conjunto de datos de Entrenamiento y Validación

Analizando los resultados del conjunto de datos de entrenamiento y validación con la información de ingreso del proceso de admisión, se observa como el clasificador bayesiano presenta la mejor precisión de registros con bloqueo académico que fueron clasificados correctamente. En la tercera matricula se incrementó un 7% con respecto al árbol de decisión. De igual forma, revisando el área bajo la curva (AUC) el árbol decisión en la primera y segunda matricula presenta un mal desempeño por debajo de 0.5. Todo lo contrario con el *Naive Bayes* con un regular desempeño por encima de 0.6. El algoritmo *Naive Bayes* presenta el mayor porcentaje de casos con No bloqueo académico que fueron clasificados incorrectamente con bloqueo académico. El árbol decisión presenta la mayor proporción de clase con bloqueo académico que fueron clasificados incorrectamente con No bloqueo académico.

En la Tabla 5-10 se presentan los resultados del modelo de predición de la pérdida de la condición académica con los datos de prueba y se comparan las diferentes técnicas de clasificación en cuanto a los parámetros de desempeño.

Predicción	Árbol de Decisión				Naive Bayes			
	Matricula 1	Matricula 2	Matricula 3	Matricula 4	Matricula 1	Matricula 2	Matricula 3	Matricula 4
Medida – F	0	0	0	0	43.30%	32.26%	0	0
Precisión	0	0	0	0	45.65%	27.78%	0	0
Exhaustividad	0.00%	0.00%	0.00%	0.00%	41.18%	38.46%	0.00%	0.00%
Exactitud	54.57%	74.51%	95.56%	90.32%	51.33%	58.82%	95.56%	90.32%
Error	45.13%	25.49%	4.44%	9.68%	48.67%	41.18%	4.44%	9.68%
Curva (AUC)	0.500	0.500	0.500	0,603	0,556	0.575	0.721	0.586
Kappa	0.00	0.000	0.00	-0.045	0,009	0.038	0,00	-0.045
Especificidad	100%	100%	100%	96.55%	59.68%	65.79%	100%	96.55%
Sensibilidad	0.00%	0.00%	0.00%	0.00%	41.18%	38.46%	0.00%	0.00%
Falsos Positivos	0%	0%	0%	3%	22%	25%	0%	3%
Falsos Negativos	45%	25%	4%	6%	27%	16%	4%	6%

Tabla 5-10 Modelo de predición de la pérdida de la condición académica utilizando los datos de Prueba.

Analizando los resultados anteriores, dado el conjunto de datos de prueba con la información de ingreso del proceso de admisión, se observa como el clasificador bayesiano en la primera y segunda matricula presenta la mejor precisión de registros con bloqueo académico que fueron clasificados correctamente. De igual forma, revisando el área bajo la curva (AUC), el algoritmo *Naive Bayes* presentan en la tercera matricula un buen desempeño. El árbol decisión en la cuarta matricula presenta un regular desempeño por encima de 0.5. El algoritmo *Naive Bayes* presenta el mayor porcentaje de casos con No bloqueo académico que fueron clasificados incorrectamente con bloqueo académico. El árbol decisión presenta la mayor proporción de clase con bloqueo académico que fueron clasificados incorrectamente con No bloqueo académico. Los anteriores resultados se pueden justificar dado el reducido número de ejemplos de clase con bloqueo académico utilizado en el conjunto de datos de prueba, lo cual ocasionó que los algoritmos se inclinaran por la clase con mayor presencia.

En la Tabla 5-11 y Tabla 5-12 se presentan los resultados de la técnica del árbol decisión y del modelo probabilístico de *Naive Bayes* en términos de la matriz de confusión para el conjunto de datos de pruebas.

Predicción	Matricula 1		Matricula 2		Matricula 3		Matricula 4	
	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.
No Bloqueo	62	51	38	13	43	2	28	2
Bloqueo Acad	0	0	0	0	0	0	1	0

Tabla 5-11 Matriz de Confusión Árbol Decisión - Conjunto de datos de prueba

Predicción	Matricula 1		Matricula 2		Matricula 3		Matricula 4	
	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.
No Bloqueo	37	30	25	8	43	2	28	2
Bloqueo Acad	25	21	13	5	0	0	1	0

Tabla 5-12 Matriz de Confusión *Naive Bayes* - Conjunto de datos de prueba

Analizando la información de la matriz de confusión de la Tabla 5-11 y la Tabla 5-12 en términos de los resultados del conjunto de datos de prueba se puede observar que el algoritmo árbol decisión en la primera y segunda matricula presenta la mayor proporción de clase con bloqueo académico que fueron clasificados incorrectamente con No bloqueo académico, sin embargo el algoritmo *Naive Bayes* presenta el mayor porcentaje de casos con No bloqueo académico que fueron clasificados incorrectamente con bloqueo académico.

5.3.4 Resultados y evaluación de la predicción de pérdida de la condición académica utilizando los datos académicos.

En esta sección se presentan los resultados y evaluación del modelo para la predición de la perdida de la condición académica en cada una de las técnicas seleccionadas; se utilizó la información inicial del proceso de admisión y el historial académico de la matricula anterior. Para la configuración de los experimentos, se utilizó la validación cruzada con 10-fold para entrenar los modelos y la evaluación del modelo se utilizó el conjunto de datos de prueba. La Tabla 5-13 se relacionan los atributos de entrada del modelo en cada una de las cuatro matriculas.

Información inicial del proceso de admisión y el historial académico	
Atributos de Entrada Socioeconómico y Resultados de las Pruebas Saber 11	Lugar de Procedencia, Carácter del Colegio, Estado Civil, SISBEN, Sexo, Programa, Situación Económica, Estrato, Edad de Ingreso, Resultado de las Pruebas Saber 11: Física, Ingles, Matemáticas, Biología, Filosofía, Geografía, Lenguaje, Ingles y Química.
Historial Académico	Número de Créditos: Cursados, Ganados, Perdidos. Número de Materias: Cursadas, Ganados, Perdidos.

Tabla 5-13 Información de Ingreso del proceso de admisión y el historial académico del semestre anterior

En la Tabla 5-14 se presentan los resultados del modelo de predición de la perdida de la condición académica con la información de ingreso del proceso de admisión y el historial

académico del semestre anterior con los datos de entrenamiento y validación, se comparan las diferentes técnicas de clasificación en cuanto a los diferentes parámetros de desempeño.

Predicción	Árbol de Decisión			Naive Bayes		
	Matricula 2	Matricula 3	Matricula 4	Matricula 2	Matricula 3	Matricula 4
Medida – F	74.42%	0	11.11%	71.00%	41.67%	40.00%
Precisión	66.40%	0.00%	10.00%	60.75%	29.41%	33.33%
Exhaustividad	86.67%	0.00%	12.50%	87.00%	71.43%	50.00%
Exactitud	84.45%	94.31%	87.05%	81.02%	92.06%	90.19%
Error	15.55%	5.69%	12.95%	18.98%	7.94%	9.81%
Curva (AUC)	0.851	0	0	0,912	0	0
Kappa	0.637	-0.224	0.042	0.578	0.295	0,350
Especificidad	83.58%	98.20%	92.12%	78.90%	92.93%	92.94%
Sensibilidad	86.67%	0.00%	12.50%	87.00%	71.43%	50.00%
Falsos Positivos	13%	2%	7%	16%	7%	6%
Falsos Negativos	1%	4%	6%	3%	1%	3%

Tabla 5-14 Modelo de predición de la pérdida de la condición académica utilizando los datos de Entrenamiento y Validación

Analizando los resultados del conjunto de datos de entrenamiento y validación se observa como el árbol decisión incremento su nivel de precisión en la segunda y cuarta matricula. El clasificador bayesiano incremento la precisión de registros con bloqueo académico que fueron clasificados correctamente. De igual forma, revisando el área bajo la curva (AUC), ambos algoritmos en la segunda matricula presentan un buen desempeño por encima de 0.7.

En la Tabla 5-15 se presentan los resultados del modelo de predición de la perdida de la condición académica con la información de ingreso del proceso de admisión y el historial académico del semestre anterior con los datos de prueba, se comparan las diferentes técnicas de clasificación en cuanto a los parámetros de desempeño.

Predicción	Árbol de Decisión			Naive Bayes		
	Matricula 2	Matricula 3	Matricula 4	Matricula 2	Matricula 3	Matricula 4
Medida – F	74.29%	0	0	70.27%	0	0
Precisión	59.09%	0	0	54.17%	0.00%	0.00%
Exhaustividad	100%	0.00%	0.00%	100%	0.00%	0.00%
Error	17.65%	4.44%	6.45%	21.57%	6.67%	9.68%
Exactitud	82.35%	95.56%	93.55%	78.43%	93.33%	90.32%
Curva ROC	0.882	0.500	0,534	0,913	0,907	0,828
Kappa	0.622	0.00	0.000	0,556	-0.031	-0.045
Especificidad	76.32%	100%	100%	71.05%	97.67%	96.55%
Sensibilidad	100%	0.00%	0.00%	100%	0.00%	0.00%
Falsos Positivos	18%	0%	0%	22%	2%	3%
Falsos Negativos	0%	4%	6%	0%	4%	6%

Tabla 5-15 Modelo de predición de la pérdida de la condición académica utilizando los datos de Prueba

Analizando los resultados del conjunto de datos de prueba, se observa como el árbol de decisión presentan el mayor número de predicciones con bloqueo académico que fueron clasificados correctamente en la segunda matricula. De igual forma, revisando el área bajo la curva (AUC), el algoritmo *Naive Bayes* presentan un buen desempeño con un área mayor al 0.9 en comparación con el algoritmo del árbol decisión.

En la Tabla 5-16 y Tabla 5-17 . Se presentan los resultados de la técnica del árbol decisión y *Naive Bayes* en términos de la matriz de confusión para el conjunto de datos de pruebas.

Predicción	Matricula 2		Matricula 3		Matricula 4	
	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.
No Bloqueo	29	0	43	2	29	2
Bloqueo Acad	9	13	0	0	0	0

Tabla 5-16 Matriz de Confusión Árbol Decisión - Conjunto de datos de prueba

Predicción	Matricula 2		Matricula 3		Matricula 4	
	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.	No Bloqueo Acad.	Bloqueo Acad.
No Bloqueo	27	0	42	2	28	2
Bloqueo Acad	11	13	1	0	1	0

Tabla 5-17 Matriz de Confusión Naive Bayes - Conjunto de datos de prueba

Analizando la información de la matriz de confusión de la Tabla 5-16 y la Tabla 5-17 en términos de los resultados del conjunto de datos de prueba se puede observar que ambos algoritmos en la tercera y cuarta matricula presenta igual proporción de casos con bloqueo académico que fueron clasificados incorrectamente con No bloqueo académico. A demás presentan igual proporción de clase No bloqueo académico que fueron clasificados incorrectamente con bloqueo académico.

5.4 Resumen del Capítulo

Los algoritmos de clasificación *Naive Bayes* y Árbol Decisión fueron utilizados en este capítulo para predecir la perdida de la condición académica por bajo rendimiento académico. Diferentes modelos fueron entrenados con diferentes configuraciones; inicialmente el modelo fue entrenado con el 80% de los registros con la información de ingreso del proceso de admisión y luego se le adicionaron los atributos académicos del

semestre anterior. Los modelos fueron probados con el 20% de los registros de cada matricula o periodo académico.

La evaluación del modelo con los datos de prueba el algoritmo *Naive Bayes* presentó el mejor desempeño en la primera y segunda matricula con los datos de ingreso del proceso de admisión. La predición de los clasificadores mejoró cuando se le adicionó a cada modelo el historial académico del semestre anterior. El árbol de decisión presentó el mayor número de registros clasificados correctamente.

6. Conclusiones y recomendaciones

6.1 Conclusiones

En los últimos años se ha mostrado un gran interés sobre el análisis de datos en las instituciones educativas, en las cuales se generan altos volúmenes de datos, dadas las nuevas técnicas y herramientas que permiten un entendimiento de los datos.

Para esta investigación se conformó un conjunto de datos a partir de la base de datos de la Universidad Popular del Cesar con la información socioeconómica, los resultados de las pruebas Saber 11 recogidos durante el proceso de admisión y el historial académico de la matricula anterior, para el entrenamiento y validación de los modelos descriptivos y predictivos.

La aplicación del algoritmo K-Means del modelo descriptivo, tuvo como objetivo analizar la población de estudiantes de la Universidad Popular del Cesar para identificar las características similares entre los grupos. Fue interesante establecer que algunas características iniciales socioeconómicas permitieron definir algunos perfiles o grupos. En la evaluación del modelo se observó que la información socioeconómica del estudiante incide en los resultados de su desempeño académico, mostrando que los grupos con mayor desempeño académico en los resultados de las pruebas saber se encontraron en los colegios de carácter académico con situación económica dependiente y de estrato dos. Los grupos con los menores niveles de desempeño en las pruebas saber 11, agruparon el mayor número de estudiantes con situación económica independiente y estrato 1 provenientes de la ciudad de Valledupar y de colegios de carácter académico.

El modelo de clasificación que se presentó en este trabajo analizó la información socioeconómica, los resultados de las pruebas saber 11 y el historial académico de la matricula anterior. El algoritmo de árbol de decisión con los datos de prueba presentó un mejor desempeño con la adición del historial académico del semestre anterior en comparación con el algoritmo *Naive Bayes*. Presento el mayor número de predicciones con bloqueo académico que fueron clasificados correctamente en la segunda matricula.

El análisis de los datos pudo evidenciar que efectivamente hay diferentes tipos de desempeño de acuerdo el perfil socioeconómico del estudiante y el historial académico, demostrando que es factible hacer predicciones y que esta investigación puede ser una herramienta muy útil para la toma de decisiones. .

4.1 Recomendaciones

Este trabajo puede ser utilizado para la toma de decisiones, por parte del programa de permanencia y graduación de la Universidad Popular del Cesar y puede ser utilizado como punto de partida para futuras investigaciones de minería de datos en la educación.

Para esta investigación se utilizaron solo dos programas de la facultad de ingenierías en el desarrollo de los modelos, sería importante la adición del resto de programas académicos con que cuenta la Universidad Popular del Cesar, con el fin de identificar las principales causas de pérdida de la condición académica y obtener datos más precisos con relación a grupos similares en sus características socioeconómicas que generen niveles de desempeño distintos.

Otra recomendación importante es que para mejorar el desempeño del modelo, se integren otras fuentes de datos, como la información del estudiante que se encuentra registrada en el ICFES, antes de ingresar a la universidad, además la información recopilada por el programa de permanencia y graduación de bienestar institucional.

A. Anexo: Consultas

A continuación se muestran las consultas desarrolladas en MySQL en las base de datos del sistema académico Academusoft de la Universidad Popular del Cesar, montada sobre el administrador MySQL - Entreprise.

A partir de estas consultas se obtuvo dos conjuntos de datos; la información socioeconómica y los resultados de las pruebas de los estudiantes admitidos y el otro conjunto de datos la información académica de los estudiantes matriculados La base de datos es administrada por la oficina de registro y control académico de la universidad.

- /*Consulta para extraer Datos de los Estudiantes Admitidos los periodos del 2010-02 entre 2014-01*/

```
SELECT asp.ASPI_PRIMERNOMBRE||' '||
  asp.ASPI_SEGUNDONOMBRE||' '||
  asp.ASPI_PRIMERAPELLIDO||' '||
  asp.ASPI_SEGUNDOAPELLIDO AS NOMBRE,
  GENERAL.TIPODOCUMENTOGENERAL.TIDG_ABREVIATURA,
  asp.ASPI_SEXO,
  asp.ASPI_NUMERODOCUMENTO,
  asp.ASPI_LUGAREXPEDICION,
  pr.PROG_NOMBRE,  estu.ESSE_SNP,
  ACADEMICO.INSTITUCION.INST_NOMBREINSTITUCION,
  asp.ASPI_DIRECCIONRESIDENCIA,
  asp.ASPI_BARRIORESIDENCIA,
  ACADEMICO.INFORMACIONSOCIOECONOMICA.INSO_ESTRATO AS ESTRATO,
  asp.ASPI_TELEFONORESIDENCIA,
  asp.ASPI_TELEFONOCELULAR,
  asp.ASPI_FECHANACIMIENTO,
  asp.ASPI_EMAIL,
```

```
 GENERAL.ESTADOCIVILGENERAL.ESCG_DESCRIPCION AS ESTADO_CIVIL,
 asp.ASPI_ETNIA,
 asp.ASPI_TIPOSANGUINEO,
 form.FOIN_ESTADOADMISION,
 co.COIN_NUMCONVOCATORIA, prog.PRXF_PUNTAJEOBTENIDO,
 ACADEMICO.INFORMACIONSOCIOECONOMICA.INSO_NUMEROFAMILIARES,
 ACADEMICO.INFORMACIONSOCIOECONOMICA.INSO_SITUACIONECONOMICA,
 ACADEMICO.INFORMACIONSOCIOECONOMICA.INSO_SISBEN,
 ACADEMICO.INFORMACIONSOCIOECONOMICA.INSO_NUMEROHERMANOS
FROM academico.ESTUDIOSSECUNDARIOS estu,
 academico.aspirante asp,
 academico.FORMULARIOINSCRIPCION form,
 academico.PROGRAMAXFORMULARIO prog,
 academico.unidadprograma unp,
 academico.programa pr,
 academico.convocatoriainscripcion co,
 GENERAL.TIPODOCUMENTOGENERAL,
 ACADEMICO.INFORMACIONSOCIOECONOMICA,
 GENERAL.ESTADOCIVILGENERAL,
 ACADEMICO.ESTUDIOSSECUNDARIOS,
 ACADEMICO.INSTITUCION
WHERE estu.ASPI_ID                    = asp.ASPI_ID
AND asp.ASPI_ID                       = form.ASPI_ID
AND form.FOIN_ID                      = prog.FOIN_ID
AND prog.COIN_ID                      = co.COIN_ID
AND prog.UNPR_ID                      = unp.UNPR_ID
AND unp.PROG_ID                       = pr.PROG_ID
AND GENERAL.TIPODOCUMENTOGENERAL.TIDG_ID    =
asp.ASPI_TIPODOCUMENTO
AND asp.ASPI_ID                       =
ACADEMICO.INFORMACIONSOCIOECONOMICA.ASPI_ID
AND ACADEMICO.ESTUDIOSSECUNDARIOS.ASPI_ID    = asp.ASPI_ID
AND ACADEMICO.ESTUDIOSSECUNDARIOS.INST_CODIGOSNP =
ACADEMICO.INSTITUCION.INST_CODIGOSNP
AND GENERAL.ESTADOCIVILGENERAL.ESCG_ID(+)    = asp.ESCG_ID
AND (
form.FOIN_ESTADOADMISION NOT          IN ('ANULADO')
AND co.COIN_NUMCONVOCATORIA               = '20071'
AND form.FOIN_ESTADOADMISION              IN ('ADMITIDO')
AND co.UNID_ID                        = 1
AND prog.UNPR_ID NOT  IN (267, 307, 310, 141, 142, 143, 144, 162, 224, 225, 226,
227, 228, 514))
```

```
 /*Consulta para extraer Datos de los Estudiantes Matriculados los periodos del 2010-02
entre 2014-01*/


SELECT peun.PEUN_ANO AS año,
  peun.PEUN_PERIODO  AS periodo,
  estp.ESTP_ID,
  pege.PEGE_DOCUMENTOIDENTIDAD,
  pege.PEGE_ID,
  peng.PENG_PRIMERNOMBRE,
  peng.PENG_SEGUNDONOMBRE,
  peng.PENG_PRIMERAPELLIDO,
  peng.PENG_SEGUNDOAPELLIDO,
  prog.PROG_NOMBRE,
  maac.SITE_ID,
  (SELECT COUNT(y.REAC_NOTAFINAL)
  FROM academico.registroacademico y
  WHERE y.ESTP_ID = estp.ESTP_ID
  AND y.PEUN_ID   = maac.PEUN_ID
  AND y.MATE_CODIGOMATERIA NOT LIKE '%ENFASIS%'
  AND y.MATE_CODIGOMATERIA NOT LIKE '%ELECTIVA%'  ) AS Materias_Tomadas,
  (SELECT COUNT(z.REAC_NOTAFINAL)   FROM academico.registroacademico z
  WHERE z.ESTP_ID       = estp.ESTP_ID   AND z.PEUN_ID         = maac.PEUN_ID
  AND z.REAC_NOTAFINAL >= 3.0   AND z.MATE_CODIGOMATERIA NOT LIKE
'%ENFASIS%'
  AND z.MATE_CODIGOMATERIA NOT LIKE '%ELECTIVA%'   ) AS Materias_Ganadas,
  maac.MAAC_PROMEDIOGENERAL,
  maac.MAAC_PROMEDIO,
  maac.CATE_ID,
  maac.MAAC_PERIODOACADEMICO,
  estp.ESTP_PERIODOCRONOLOGICO,
  estp.ESTP_CREDITOSAPROBADOS
FROM GENERAL.personanaturalgeneral peng,
  GENERAL.personageneral pege,
  academico.estudiantepensum estp,
  ACADEMICO.liquidacion LI,
  academico.unidadprograma unpr,
  academico.programa prog,
  academico.his_matriculaacademica maac,
  academico.periodouniversidad peun
WHERE pege.PEGE_ID     = peng.PEGE_ID
AND peng.PEGE_ID       = estp.PEGE_ID
AND estp.UNPR_ID       = unpr.UNPR_ID
```

```
AND unpr.PROG_ID      = prog.PROG_ID
AND estp.ESTP_ID      = maac.ESTP_ID
AND maac.PEUN_ID      = peun.PEUN_ID
AND estp.ESTP_ID      = LI.ESTP_ID
AND (prog.PROG_NOMBRE IN ('INGENIERIA DE SISTEMAS', 'INGENIERIA
ELECTRONICA')
AND maac.PEUN_ID      IN (205, 225, 245, 265, 285, 345, 366, 406, 426, 446))
```

Bibliografía

[1] Romero. C. and Ventura. S. 2007. Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications Volume, Issuse 1, July 2007, Pages 135-146.

[2] Han J., Kamber M., 2006, “Data mining: concepts and techniques”, The Morgan Kaufmann Publishers, USA, ISBN: 1558609016

[3]. F. Araque. C. Roldan. A. Salguero. “Factors influencing university drop out rates”. Computers & Education. Vol. 53 . 2009. pag. 563–574. España

[4]. S. Kotsiantis. C. Pierrakeas. and P. Pintelas. “Preventing student dropout in distance learning systems using machine learning techniques.” in Proc. Int. Conf. Knowl.-Based Intell. Inf.learning techniques.” in Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.. Oxford. U.K.. 2004. pp. 3–5.. Grecia

[5]. h. kuna, r. Garcia, f. Villatoro “Pattern discovery in university students desertion based on data mining,” advances and applications in statistical sciences proceedings of the iv meeting on dynamics of social and economic systems volume 2, issue 2, 2010, pages 275-285 © 2010 mili publications, Argentina

[6] Zlatko J. Kovacic “Predicting student success by mining enrolment data.” Research in Higher Education Journal.2012. New Zealand

[7] S. K. Yadav. B.Bharadwaj and S. Pal. “Mining Educational Data to Predict Student’s Retention :A Comparative Study”. International Journal of Computer Science and Information Security (IJCSIS). Vol. 10. No. 2. 2012. India

[8] Mr. M. N. Quadril. Dr. N.V. Kalyankar. “Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques.” Global Journal of Computer Science and Technology. P a g e | 2 Vol. 10 Issue 2 (Ver 1.0). April 2010. India

[9] Y. Zhang, S. Oussena. “Use Data Mining to Improve Student Retention in Higher Education – a Case Study”. ICEIS - 12th International Conference on Enterprise Information Systems. 2010. Inglaterra.

[10] Q. Al- Radaideh, E. Al-Sahwakf y M. Al-Najjar "Mining Student Data Using Decision Trees." The International Arab Conference on Information Technology (ACIT) , 2006, Jordan.

[11] E. Yudkselturk, S. Ozekes y Y. Kilic " Predicting Dropout Student : An Application of Data Mining Methods in a Online Education Program “Journal of European Journal of Open, Distance and E-Learning, Vol 17 / No.1 - 2014

[12] S. Pal, "Mining Educational Data to Reduce Dropout Rates of Engineering Students," International Journal of Information Engineering and Electronic Business, 2012,2, 1-7 Published Online April 2012

[13] R.B Bhise, S.S Thorat, A.K Supekar "Importance of Data Mining in Higher Education System," IOSR Journal Of Humanities And Social Science (IOSR-JHSS), ISSN: 2279-0837, ISBN: 2279-0845. Volume 6, Issue 6 (Jan. - Feb. 2013), PP 18-21

[14] S.Z Erdogan, M. Timor "A Data Mining application in a student database," Journal of Aeronautics and Space Technologies, vol. 2 number 2, pp. 53-57, 2005.

[15] A. Bhardwaj, A. Bhardwaj " Modified K- Means Clustering Algorithm for Data Mining in Education Domain " Internatiobnal Journal of Advanced Research in Computer Science and Software Engineering “ , vol. 3 ISSN 2277 128X, 2013.

[16] A. Timarán. “Una lectura sobre deserción universitaria en estudiantes de pregrado desde la minería de datos.” Revista Científica Guillermo de Ockham. vol. 8. núm. 1. enero-junio. 2010. pp. 121-130 Universidad de San Buenaventura. Cali. Colombia

[17] C. Lopez. “Data Mining Model to Predict Academic Performance at the Universidad Nacional de Colombia.” Thesis work to obtain the degree of: Master in Computer Science Universidad Nacional de Colombia, 2013

[18] Pete Chapman (NCR). J.C.S.. Randy Kerber (NCR). Thomas Khabaza (SPSS).Thomas Reinartz (DaimlerChrysler). Colin Shearer (SPSS) y Rüdiger Wirth(DaimlerChrysler) “CRISP-DM 1.0: Step-by-step data mining guide.” CRISP-DM 1.0:Stepby-step data mining guide.. The CRISP-DM consortium. 2000.

[19] IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785, Estados Unidos Manual CRISP-DM de IBM SPSS Modeler. IBM SPSS Modeler 15, @Copyright IBM Corporation 1994, 2012.

[20] K. G. J. Vásquez. S. Gallón y E. Castaño. "Deserción estudiantil en la educación Superior colombiana." Imprenta Nacional de Colombia. vol. Primera Edicion 2009. Bogotá

[21] Mohammed M. Abu Tair. A.M.E.-H.. Mining Educational Data to Improve Students' Performance: A Case Study. International Journal of Information and Communication Tecnology. 2012: p. 7.

[22] E. Ayers. R. Nugent. and N. Dean. “A comparison of student skill knowledge estimates.” in Proc. Int. Conf. Educ. Data Mining. Cordoba. Spain. 2009. pp. 1–10.

[23] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second. USA; Elsevier; Morgan Kaufmann Publishers, 2006, p. 772.

[24] P. Shekhawat and S. Dhande. “A Classification Technique using Associative Classification.” International Journal of Computer Applications. vol. 20. No.5. pp. 20-28. April 2011. (AC)

[28] Raunheitte. L. and R.d. Camargo. A Study on the application of Data Mining Methods in the analysis of Transcripts. iiis.org.

[29] Shah. N.. Predicting Factors That Affect Students'academic Performance By Using Data Mining Techniques. Pakistan Business Review. 2011.

[30] J. R. Quinlan. “Introduction of decision tree”. Journal of Machine learning”. pp. 81-106. 1986.

[31]. J. R. Quinlan. “C4.5: Programs for Machine Learning”. Morgan Kaufman Publishers 1993.

[33] Langley P. and Thompson K. Iba W. An analysis of Bayesian Classifiers.Proceedings of the 10th National Conferences on Artificial Intelligence. pages 223. 1992of the 10th National Conferences on Artificial Intelligence. pages 223. 1992

[34] Lopez Guarin, C. E., Guzman, E. L., & Gonzalez, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. In IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, Vol. 10, Issue 3, pp. 119-125. ISSN: 1932-8540. doi: 10.1109/RITA.2015.2452632

[35] Kononenko, I. (1990). Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. Current Trends in Knowledge Acquisition.

[36] Langley, P., Iba, W. y Thomas, K. (1992). An analysis of Bayesian classifiers. In Proceedings of the Tenth National Conference of Artificial Intelligence. AAAI Press. 223-228.

[37] Friedman, N., Geiger, D. y Goldszmidt, M. (1997). Bayesian network classifiers. Machine Learning, 29:2, 131–163.

[38] E. Alpaydin, Introduction to Machine Learning. The MIT press Cambridge, MA, 2004.