Prototipo de un sistema de alerta temprana de bajo rendimiento académico en los programas de la Facultad de Ingeniería de la Universidad del Valle Sede Tuluá.

Daniel Meja Vélez

Universidad del Valle
Escuela de Ingeniería en Sistemas y Computación
Ingeniería en Sistemas y Computación
Tuluá – Valle del Cauca
2021

Prototipo de un Sistema de alerta temprana de bajo rendimiento académico en los programas de la Facultad de Ingeniería de la Universidad del Valle Sede Tuluá.

Daniel Mejía Vélez
mejia.daniel@correounivalle.edu.co

Director: Mauricio López Benítez

Universidad del Valle
Escuela de Ingeniería en Sistemas y Computación
Ingeniería en Sistemas y Computación
Tuluá – Valle del Cauca
2021

Trabajo de grado presentado por
Daniel Mejía Vélez
Como requisito parcial para la obtención del título de Ingeniero de Sistemas

_____

MAURICIO LOPEZ BENITEZ
DIRECTOR

_____ _____

Jurado **Jurado**

TABLA DE CONTENIDO

Lista de Figuras

Lista de Tablas

Resumen

Este proyecto fue elaborado con el fin de realizar, un prototipo de un sistema de alerta temprana de bajo rendimiento académico, en los programas de la Facultad de Ingeniería de la Universidad del Valle Sede Tuluá; que sirva como herramienta, para que las directivas de la universidad generen mecanismos, para trabajar la disminución de esta problemática. A lo largo de este documento, se irá presentando la investigación que se llevó a cabo sobre la problemática, el proceso para obtener la base de conocimiento para el proyecto, la fase del preprocesamiento de la información para la creación de dataSet, etapa de análisis y selección de dos modelos de minería de datos y posteriormente, implementación de los algoritmos, para poder obtener los resultados, que ayudaron a la culminación del prototipo y del proyecto.

Palabras claves: Prototipo, dataSet, bajo rendimiento academico, minería de datos.

Abstract

This project was developed with the purpose of creating a prototype of an early warning system of low academic performance in the programs of the Faculty of Engineering of the Universidad del Valle Sede Tuluá, which serves as a tool for the directors of the university to generate mechanisms to work on the reduction of this problem. Throughout this document, we will present the research that was carried out on the problem, the process to obtain the knowledge base for the project, the preprocessing phase of the information for the creation of dataSet, the analysis and selection stage of two data mining models and later, the implementation of the algorithms, to obtain the results, which helped the completion of the prototype and the project.

Keywords: Prototype, dataset, low academic performance, data mining

Capítulo 1 Introducción

En la actualidad, la Universidad del Valle sede Tuluá, presenta una gran problemática sobre la cantidad de estudiantes que incurren en bajo rendimiento, esta situación se presenta mayoritariamente en la Facultad de Ingeniería. En el historial académico de los estudiantes, además de las calificaciones, se registran las situaciones de bajo rendimiento cuando éstas ocurren, así como los, periodos cursados y otros datos relevantes, pero esta información no arroja un motivo o un por qué sobre esta tendencia.

Por lo anterior, se decidió diseñar el prototipo de una herramienta que alerte que estudiante puede incurrir en un bajo rendimiento académico. Todo el estudio fue realizado con técnicas de minería de datos y un dataSet, creado con base de toda la información, que la Universidad de Valle sede Tuluá nos suministró, sobre el rendimiento académico de los estudiantes de la Facultad de Ingeniería.

Con lo anterior, se aplicaron dos técnicas de minería de datos, las cuales fueron evaluadas y que al final se seleccionó una técnica para la implementación, considerando diferentes criterios como la precisión del modelo, la exactitud, la sensibilidad, entre otros criterios.

En este documento se describe más a fondo la problemática, el manejo que se le dio a los datos que se analizaron, el desarrollo con el uso de técnicas de minería y futuros trabajos que se puedan realizar basado en esta investigación.

1.1 Descripción del problema

Cuando un estudiante logra matricularse en la Universidad del Valle, esta matricula tiene asociado una cantidad de créditos permitidos (máximo 21 por semestre), estos créditos son proporcionados por la materia que se cursan, hay casos de materias de dos créditos y hasta los cuatro créditos, gracias a esto, se puede entender el cómo se incurre en un bajo rendimiento académico, desde un punto de vista normativo, en el cual el estudiante puede incurrir en esta falta de dos formas posibles, la primera es cuando el estudiante reprueba el 50%+1 de los créditos matriculados en el semestre, sabiendo que la nota mínima para aprobar una materia es 3.0, puesto que la universidad califica en una escala de 1 al 5; la segunda forma de incurrir en un bajo rendimiento es reprobar una materia en estado de repitente (repitente es aquel estudiante que curso la metería el semestre pasado y la está cursando en el semestre actual), sin embargo, si el estudiante no ha estado en situación de bajo rendimiento y su promedio acumulado es de tres punto cinco (3.5) o mayor, no habrá lugar a un bajo rendimiento. (ACUERDO No. 009 noviembre 13 de 1997, ARTICULO 59º), cuando se obtiene esta primer falta se le puede denominar como un llamado de atención al estudiante, pero en el caso de incurrir en un segundo bajo rendimiento académico la Universidad del Valle, comenzara un análisis para determinar si el estudiante posee el 60% o más de los créditos totales aprobados de la carrera, en caso de que no cumpla este requisito será retirado de esta por 2 años y dada la situación de incurrir en un tercer bajo rendimiento, se aplicara el mismo análisis anterior con la diferencia que evaluaran el 80% de los créditos, de no cumplir con este último requisito, el estudiante será retirado definitivamente del programa académico sin posibilidades de reingresar.

En los últimos 5 años la Universidad del Valle sede Tuluá, se presentó un elevado número de estudiantes que incurrieron en un bajo rendimiento, donde, se notó que la facultad de Ingeniería era una de las más afectadas. como gran ejemplo, tenemos que en el año 2017 se registraron 178 bajos rendimientos de los cuales 143 pertenecen a la facultad de ingeniería, toda esta información se conoce gracias a los registros suministrados por la secretaria académica de la Universidad del Valle sede Tuluá, esto demuestra una cifra alarmante para la facultad de ingeniería. Uno de los factores que están implicados, es la corta edad del estudiante al momento de ingresar a la educación superior, como se nombra en el documento [4] donde, se menciona que más del 60% de los estudiantes en primer semestre son menores de 18 años, lo cual implica, que al graduarse del colegio deben afrontar una decisión tan importante, como la de escoger a que carrera quieren ingresar, sabiendo que en el periodo escolar no se les da una orientación, sobre el tipo de profesión que desean ejercer en su futura vida de acuerdo a sus cualidades, gracias a esto, tenemos estudiantes que incurren un bajo rendimiento académico por el hecho de que no están en una carrera que deseen o donde no pueden exponer su máximo potencial.

Esta problemática lleva, a que muchos estudiantes sean retirados de la carrera o que decidan no continuar con sus estudios, lo cual le genera un daño económico a la universidad, puesto que se asignaron unos recursos para un determinado estudiante, el cual no aprovechara y se le quita la oportunidad a otros estudiantes que si rendirían; pero,

este problema económico también afecta al estudiante, ya que pierde el dinero que invirtió en su carrera, como el pago de semestre, materiales de estudio, entre otros gastos, a esto se le añade un daño emocional, puesto que el estudiante tiene como meta progresar en la vida y esto se ve frustrado por un mal rendimiento académico o una mala decisión en su vida. Como dato a tener en cuenta, la universidad ha tenido 1208 estudiantes que han incurrido en bajo rendimiento académico, de los cuales más del 50% ya no se encuentran matriculados.

Esta problemática va a seguir repitiendo, ya que hay mucho estudiante de semestres recientes que cometen los mismos errores que otros hace 5 años, esto demuestra que es necesaria una herramienta de predicción para alertar a la universidad, sobre estudiantes que puedan tender a un bajo rendimiento y de esta forma poder hacer un acompañamiento a estos, tratando de disminuir el número de estudiantes que son expulsados o donde ellos mismos desertan de la Facultad de Ingeniería.

1.2 Formulación del problema

De acuerdo a la descripción del problema, los bajos rendimientos en la Universidad del Valle sede Tuluá, es una problemática que afecta a la misma institución y a los estudiantes, por lo cual se plantea el siguiente interrogante.

¿Cómo alertar a la Universidad del Valle sede Tuluá sobre la posibilidad de que un estudiante de la Facultad de Ingeniería pueda incurrir en bajo rendimiento?

1.3 Objetivos

1.3.1 Objetivo general

Implementar el prototipo de un sistema de alerta temprana de bajo rendimiento académico en los programas de Ingeniería de la Universidad del Valle sede Tuluá.

1.3.2 Objetivos específicos

- Construir un dataSet que permita realizar un análisis sobre los bajos rendimientos en la facultad de ingeniería.
- Seleccionar un modelo de minería de datos para ser implementado en el prototipo de sistema de alerta temprana.
- Implementar el modelo de minería de datos para el procesamiento del dataset diseñado.
- Realizar pruebas funcionales para validar la efectividad del modelo.

1.4 Justificación

Analizando algunos trabajos investigativos [1][2][4][6] sobre “Bajos rendimientos académicos en Universidades” se logra detectar factores que están involucrados en esta problemática, como la edad de ingreso, de que escuela provienen, situación socio económica, dificultad de la materia, acceso a materiales de estudio, ETC. también se da entender que, los estudiantes más propensos a mostrar un bajo rendimiento son aquellos que cursan los primeros semestres. Tratando de generar un estudio sobre esta problemática, se busca identificar patrones con el historial de datos de los últimos 7 años [2013-2019] en la Universidad del Valle sede Tuluá, más específicamente en la Facultad de Ingeniería donde está la mayor parte de la problemática.

El prototipo de alerta temprana no busca reducir el número de bajos rendimientos, pero si, identificar estudiantes que sean susceptibles de incurrir en esta falta y así la universidad pueda generar mecanismos de ayuda para evitar que la problemática siga y aumente con el tiempo, la idea principal es disminuir estos números alarmantes y mejorar la calidad de aprendizaje del estudiante.

Para el desarrollo de este prototipo de alerta temprana para bajos rendimientos académicos, se empleó el modelo CRISP, el cual es muy utilizado en proyectos de minería de datos, mostrando una eficiencia en el tratamiento de los datos, con el fin, de crear una buena herramienta a partir de dataSet que se construya y poder obtener ese prototipo de alerta con una buena funcionalidad; también es una metodología estandarizada a nivel internacional con un gran reconocimiento.

Teniendo en cuanta la gran cantidad de registros que se procesaron durante 7 años, los cuales fueron entre el periodo 1 del 2013 y el periodo 2 del 2019, son 2245 estudiantes, los cuales provienen de la facultad de Ingeniería, que está compuesta por, Ingeniería de Sistemas, Ingeniería en Alimentos, Tecnología en Electrónica, Tecnología de Sistemas y Tecnología en alimentos, todo esto sirvió para ser analizado por las dos técnicas de minería de datos y proporcionarnos una buena ventana de observación.

1.5 Metodología

Analizando el tipo de trabajo que se realizó con sus respectivas características, se concluye que la metodología que mejor se adapta es la descriptiva, ya que en el desarrollo de este proyecto, se busca identificar patrones que presentan los estudiantes que han incurrido en bajos rendimientos, esto con el fin de poder generar el prototipo de alerta temprana de bajos rendimientos de la facultad de ingeniería, teniendo como fuente de información los registros de la sede Tuluá desde el 2013 hasta el 2019. Esto le podría permitir a las directivas de la Universidad del Valle sede Tuluá, crear campañas para mejorar la formación académica de los estudiantes.

Los datos sobre los bajos rendimientos de la sede Tuluá, que se emplearon en el trabajo, fueron solicitados a la Universidad con el fin de desarrollar la presente investigación,

utilizando una metodología de minería de datos conocida como CRISP-DM, la cual se divide en seis fases que son las siguientes:

1. **Analizar el problema:** fase donde se analiza varios documentos relacionados con la problemática y comprender cómo se puede abordar este conflicto con la minería de datos.
2. **Comprensión de los datos:** fase donde se adquieren los datos y se dimensiona.
3. **Preparación de Datos:** fase donde se transforman los datos para comenzar con el desarrollo del trabajo.
4. **Modelo:** fase donde se tendrá en cuenta el dataSet construido para determinar qué técnica de minería de datos es más apropiada
5. **Evaluación:** fase donde evaluar los resultados obtenidos.
6. **Despliegue:** fase donde se monitorea y se realiza un mantenimiento al modelo seleccionado, para realizar la debida documentación.

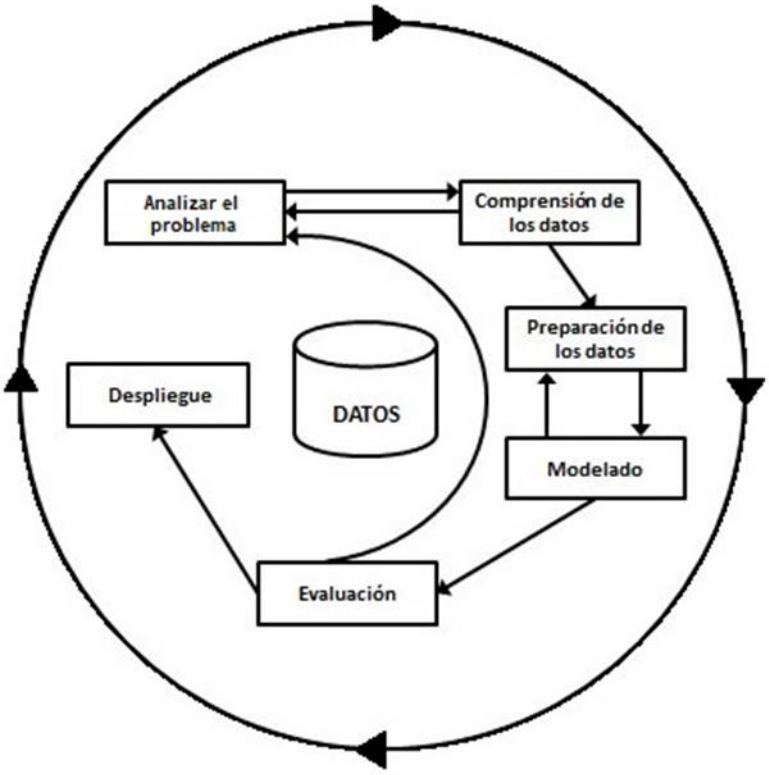

Figura 1.1: Grafica metodología CRISP-DM.

Capítulo 2 Marco referencial

2.1 Marco teórico

Para el desarrollo de este prototipo, se tuvo como referencia las investigaciones [2][4][6] en especial las referentes al rendimiento académico, donde abordaban el tema desde diferentes perspectivas, pero con el fin de orientar a las instituciones sobre qué medidas tomar para mejorar la calidad del proceso de educación que el estudiante está tomando. En específico, se define al rendimiento académico, como la productividad del sujeto, matizado por sus actividades, rasgos y la percepción casi correcta de los cometidos asignados (forteza, 1975). El rendimiento académico suele ser objeto de discusión, ya que existen muchos factores que pueden afectar el desempeño del alumno, determinar esto no es una tarea fácil. Se propone una alternativa muy viable, la cual establece que hay características que comparten un grupo de alumnos que ayudan a determinar perfiles de estudiantes con un bajo rendimiento.[2]

Analizando un patrón que tiene mucha importancia, el cual es la edad, puesto que, los estudiantes que son parte de estas investigaciones se encuentran, entre los 17 y los 22 años de edad, lo que corresponde a la adolescencia y a la etapa del adulto joven (Lefrançois, 2001 y Llínas, 2008). De acuerdo a una aproximación, se dice que más del 60% de los estudiantes son menores de 18 años (Boletín Estadístico, 2006). Esto da a entender y como lo define la psicología evolutiva que, en los primeros semestres de la Universidad, los estudiantes son adolescentes. A este factor se le puede añadir la difícil decisión que es escoger una carrera para ingresar, lo cual es complicado para muchos adultos, esta elección lo marcará toda su vida (Moreno, 2004).[4]
En la fase de minería de datos (Data Mining o DM) se genera un algoritmo, el cual clasifica por medio de patrones a partir de datos pre-procesados (fayyad et al., 2001), (hand et al., 2000), (frawley et al., 1992). La minería de datos funciona mucho con un historial de información almacenada, el cual le permite realizar una toma de decisiones (IbM Software group, 2003). Además, permite extraer patrones y tendencias, esto para predecir comportamientos futuros (Simon, 1997), (berson & Smith, 1997), (White, 2001).[2]

Cuando se construye la base de datos de conocimiento o dataset, en el cual se implementó las dos técnicas de minería de datos, toda la información pasa por un tratamiento de datos donde se escogen qué características nos sirve y cuáles deben ser omitidas, algo que se debe tener muy presente, es la ley de protección de datos, donde se eliminan algunos valores del alumno (DNI, nombre, apellidos, domicilio, etc.) también se debe tener en cuenta información concreta sobre los padres desde una posición social (estudio de padre, estudio de la madre, edad de los padres), todas estas variables se pueden tener en cuenta para la construcción de la base de datos.[6]

2.1.1. Técnicas de minería de datos

K-Nearest-Neighbor (KNN)

Es un algoritmo que se basa en ejemplos para un aprendizaje automático supervisado. Es utilizado para predecir o clasificar nuevas muestras (valores discretos) (regresión, valores continuos). Al ser un método simple es ideal para comenzar en el mundo del aprendizaje automático. Para realizar sus clasificación o predicciones, se puede utilizar de dos formas: [16]

- **Supervisado:** esto brevemente quiere decir que tenemos etiquetado nuestro conjunto de datos de entrenamiento, con la clase o resultado esperado dada "una fila" de datos.[16]
- **Basado en Instancia**: Esto quiere decir que nuestro algoritmo no aprende explícitamente un modelo (como por ejemplo en Regresión Logística o árboles de decisión). En cambio, memoriza las instancias de entrenamiento que son usadas como "base de conocimiento" para la fase de predicción.[16]

Naive Bayes

Según Víctor Román (2019) "Los modelos de Naive Bayes son una clase especial de algoritmos de clasificación de Aprendizaje Automático, o Machine Learning. Se basan en una técnica de clasificación estadística llamada "teorema de Bayes"".[17]

Estos modelos son llamados algoritmos "Naive", o "Inocentes" en español. En estos se asume que las variables predictoras son independientes entre sí, lo que significa que la presencia de una cierta característica en un conjunto de datos, no está en absoluto relacionada con la presencia de cualquier otra característica. [17]

Máquinas de vectores de soporte

Las Máquinas de Vectores de Soporte (Support Vector Machines) nos permiten encontrar la forma óptima de clasificar entre varias clases. Esta clasificación optima se realiza maximizando el margen de separación entre las clases, en palabras más claras, se busca la mejor posición para un vector (línea en el conjunto de datos) el cual separa la información en dos categorías, en el caso de que las clases no sean linealmente separables, podemos usar el truco del kernel para añadir una dimensión nueva donde sí lo sean.

Para muchas aplicaciones se prefería el uso de SVM en lugar de redes neuronales. La principal razón era que las matemáticas de los SVM se entienden muy bien y la propiedad para obtener un margen de separación máximo es muy llamativo.[19]

Algunos casos de éxito de las máquinas de vectores de soporte son:

- reconocimiento óptico de caracteres.
- detección de caras para que las cámaras digitales enfoquen correctamente.
- filtros de spam para correo electrónico.
- reconocimiento de imágenes a bordo de satélites (saber qué partes de una imagen tienen nubes, tierra, agua, hielo, etc.).

Arboles de Decisión

Un árbol de Decisiones es un esquema de los posibles resultados que se obtienen de una serie de decisiones relacionadas, permite que comparen posibles acciones según sus costos, probabilidades y beneficios, todo con el fin de que el algoritmo anticipe la mejor opción.[20]

Un árbol de Decisión por lo general, comienza con un único nodo (raíz) y luego se ramifica en resultados posibles. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades.[20]

Características de un árbol de decisión:[21]

- Plantea el problema desde distintas perspectivas de acción.
- Permite analizar de manera completa todas las posibles soluciones.
- Provee de un esquema para cuantificar el costo del resultado y su probabilidad de uso.
- Ayuda a realizar las mejores decisiones con base a la información existente y a las mejores suposiciones.
- Su estructura permite analizar las alternativas, los eventos, las probabilidades y los resultados.

Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (también conocidas como sistemas conexionistas) son un modelo computacional, que se fue formando a partir de aportaciones científicas registradas en la historia. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales.[22]

El objetivo principal de este modelo, es aprender modificándose automáticamente de forma que puede realizar tareas complejas, que no podrían ser efectuado por medio de la clásica programación basada en reglas. De esta forma se pueden automatizar funciones que en un principio solo podrían ser realizadas por personas.[23]

2.2 Marco de antecedentes

En el documento [1] se aborda el problema de los bajos rendimientos desde la Universidad Nacional Agraria La Molina, Lima – Perú, donde aplican cuatro métodos de minería de datos, que son, Regresión Logística, Arboles de Decisiones, Redes Neuronales y Red Naive de Bayes, también utilizan una base de datos de estudiantes matriculados en el curso de Estadística General en los semestres 2013 II y 2014 I, todo con la finalidad de predecir la clasificación final (Desaprobado o Aprobado).

Al final del desarrollo del trabajo, se dan a conocer la importancia que tiene algunas variables de la base de datos empleada como lo es el promedio ponderado, situación del curso, nota en diferencial, entre otras variables, puesto que estas influyen en los resultados obtenidos, donde la Red de Naive de Bayes logra la mayor tasa de buena clasificación con 71.0%, de ahí le siguen Regresión Logística con 68.4%, Arboles de Decisión con 68.3% y Redes Neuronales con 67.9%. El punto débil del trabajo es la falta de información socio económica del estudiante porque esta mejoraría la exactitud de los modelos seleccionados.

En el trabajo [2] se propone, la implementación de técnicas de minería de datos sobre la información del desempeño de los alumnos, con el propósito de caracterizar los perfiles de alumnos exitosos (buen rendimiento académico) y de aquellos, que no lo son (bajo rendimiento académico).

Una de las ventajas de esta investigación, es el manejo y la transformación de los datos para la construcción de una sólida base de datos depurada, habiendo empleado la metodología CRISP-DM menciona en la sección 1.5. Posteriormente se implementó la técnica Arboles de Decisión arrojando un 94.4% de exactitud, dando a conocer que hay factores que inciden en el rendimiento académico, como el aspecto socioeconómico, la formación educativa de los padres, entre otros datos.

En el documento [6] de la Universidad Politécnica de Valencia, busca analizar factores socioeconómicos, características personales, puntajes de ingreso de estudiantes de ingreso de la Faculta de Informática y la Escuela Técnica Superior de Informática Aplicada, con el fin de poder ver que influencia tienen estos factores en el rendimiento académico por medio de técnicas de minería de datos, como lo son, Arboles de Decisión y la regresión multivariante. Toda la información que suministro este trabajo, fue utilizada por la Universidad Politécnica de Valencia, para generar estrategias que mejoren la calidad de enseñanza.

En el artículo [7] se hace un estudio sobre las actividades cotidianas de los estudiantes, con el fin, de analizar si se beneficia o afecta el rendimiento académico, la información que se posee, fue adquirida por medio de encuestas a 208 estudiantes en su etapa escolar, obteniendo datos como las horas que pasa en el celular, deporte que practican, número de horas en el televisor, ETC. Al final la base de datos construida fue implementada con un

total de siete algoritmos de minería de datos, donde la mayoría tuvo una exactitud mayor a 80% como knn-1 la cual obtuvo una de las mejore precisiones con un 90.86%.

En el documento [4] se analiza el fracaso académico desde una perspectiva psicológica, mostrando factores como la temprana edad de ingreso a la carrera, factor económico y social; se trabajó con un número de 38 estudiantes de psicología, que están desde el primer semestre del 2007, donde se encontró que 78.9% de esos estudiantes ya no pertenecen al programa, lo cual es una cifra negativa para la Universidad, estos se hallaban entre el primer y cuarto semestre. Encontraron que 17 jóvenes universitarios recibieron una orientación vocacional en su etapa en el colegio, esto indica que el resto de los jóvenes no recibieron esa orientación adecuada, lo cual sugiere que no fundamentaron sus decisiones al escoger esta carrera.

2.3 Marco conceptual

- **Minería de datos:** Es un campo de la estadística y las ciencias de la computación, con el fin de analizar grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como lo grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación).
- **Bajo rendimiento académico:** es un problema que presentan los estudiantes, adonde su rendimiento no es el más optimo y su proceso de formación no se está dando de la mejor manera; Para la Universidad del Valle es una falta (o llamado de atención) que adquiere un estudiante por dos motivos en especial, uno es perder más del 50% de los créditos matriculados en el semestre, el segundo es perder una materia en estado de repitente, cuando un estudiante incurre en un segundo bajo rendimiento hay una gran posibilidad que lo veten de la carrera por dos años.

- **Modelo CRISP:** metodología para el desarrollo de proyectos de minería de datos, este se divide en seis fases. divide en seis fases.

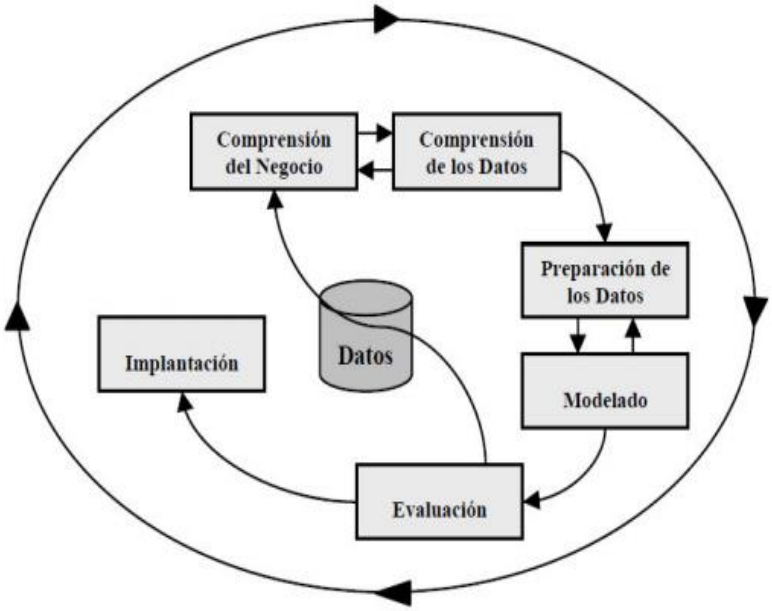

Figura 2.1: Grafica metodología CRISP-DM

1. La primera fase se denomina comprensión del negocio o problema, es una de las más importantes, pues estas buscan obtener una mejor comprensión de los objetivos y requisitos del proyecto desde la perspectiva empresarial o institucional. Para adquirir el mejor provecho del procesamiento de datos, es necesario entender de manera completa el problema que se desea resolver, lo cual nos ayudaría a una mejor recolecta de datos e interpretación de estos.
2. La segunda fase es la Comprensión de los datos, se inicia con la recolección de estos y así realiza un contacto inicial con el problema para familiarizarse con los datos, identificando su calidad y establecer relaciones evidentes.
3. En la tercera fase se realiza la preparación de los datos, después de recolectar estos, se iniciará un proceso de conversión y filtro de datos para adaptarlos a determinada técnica de minería de datos.
4. En la cuarta fase se realiza el modelado, donde se seleccionará una técnica de modelado más adecuada para el proyecto, las técnicas a utilizar en esta etapa se eligen en función de unos criterios, como ser apropiada al problema, disponer de datos adecuados, cumplir los requisitos de problema, tiempo adecuado para obtener el modelado y conocimiento de la técnica.
5. En la quinta fase se evalúa el criterio de éxito del modelado. Es preciso tener un registro de los resultados obtenidos, esto para repetir algún paso donde se pudo haber cometido un error. Se pueden emplear múltiples herramientas para medir el éxito como las “matrices de confusión”.

6. En la sexta fase, se realizará la implementación, donde se busca la conversión del conocimiento obtenido en acciones dentro del problema, también ayuda a generar estrategias de monitoreo y mantenimiento del modelo; esto para obtener finalmente unas conclusiones sobre el proceso que se llevó a cabo en la implementación del modelo.

- **Precisión:** En los campos de la ciencia, la Ingeniería y la estadística se refiere a la cercanía de los valores obtenidos, una forma de verlo es que el algoritmo no me etiquete una muestra como negativa, cuando en realidad es positiva.
- **Matriz de Confusión:** es una herramienta, que permite visualizar el desempeño de un algoritmo de aprendizaje supervisado, donde se almacenan verdaderos positivos, verdaderos negativos, falsos negativos y positivos.
- **Recall Score:** Es la proporción donde está el número de verdaderos positivos y el número de falsos negativos. Es intuitivamente la capacidad del clasificador de encontrar todas las muestras positivas. El mejor valor es 1 y el peor valor es 0
- **F1 Score:** La puntuación F1 se puede interpretar como un promedio ponderado de la precisión y la recuperación, donde una puntuación F1 alcanza su mejor valor en 1 y la peor puntuación en 0.
- **Accuracy:** La función de clasificación de etiquetas múltiples, calcula la precisión del subconjunto de etiquetas predichas para una muestra, esta debe coincidir exactamente con el conjunto de etiquetas correspondiente.
- **Curva ROC:** Es una representación gráfica, esta describe la sensibilidad frente a la especificidad para un sistema clasificador binario, según Curva ROC: Es una representación gráfica que describe la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Otra interpretación de este grafico es la representación de la razón o la proporción de los verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o proporción de los falsos positivos (FPR = Razón de Falsos Positivos) también se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo). Otra interpretación de este gráfico, es la representación de la razón o la proporción de los verdaderos positivos (VPR = Razón de Verdaderos Positivos) frente a la razón o proporción de los falsos positivos (FPR = Razón de Falsos Positivos) también se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo).

3.4 Marco legal

En el desarrollo del trabajo de grado que está orientado a la minería de datos, se trabajó con datos de estudiantes que hacen o hicieron parte de la Universidad del Valle sede Tuluá, por lo cual se manejan conceptos como, datos públicos y datos sensibles, razón por la cual se tiene en cuenta a la Ley 1581 del 2012 [3], teniendo en cuenta lo anterior, se trabajó con una base de datos anonimizada para garantizar la privacidad de la información, esto sin quebrantar las normas pertenecientes a la nación.

Capítulo 3 Desarrollo del Proyecto

3.1. Caracterización de los datos

Analizando la problemática y la forma en que se abordó el tema, es necesario caracterizar algunos aspectos de los estudiantes, como:

3.1.1. Aspecto institucional

En este aspecto se toma en cuenta, las condiciones en la que el estudiante se encuentra en su ambiente Universitario, los siguientes factores pueden generar que el estudiante incurra en un bajo rendimiento:

- Programa académico.
- Condiciones de excepción.
- Jornada.

3.1.2 Aspecto académico

Este aspecto es uno de las más importantes, puesto que, analizamos a que ambiente se enfrenta el estudiante durante el semestre, donde se tiene en cuenta las asignaturas, créditos, etc. mirando todos estos factores, se logró percibir algún patrón en estudiantes con bajos rendimientos, toda esta información servirá para la caracterización, los factores a tener en cuenta, son:

- Proporción de créditos aprobados.
- Proporción de créditos Reprobados.
- Puntaje de admisión.
- Ubicación semestral.
- Poseedor de algún bajo rendimiento.
- Estados del estudiante (graduado o no).
- Promedio por Facultades.
- Proporción de materias canceladas.
- Proporción de materias habilitadas.

En las proporciones de créditos aprobados y reprobados, se manipularon en porcentajes, puesto que, hay casos donde un estudiante incurre en bajo rendimiento en los dos primeros semestres y hay otros estudiantes, que lo adquieren en semestres más avanzados, la forma en cómo se calcularon los porcentajes fue, obtener el total de créditos vistos en la carrera, el cual será el 100% y después se analizó que porcentaje equivale los créditos aprobados y reprobados.

En el promedio por facultades se divide en tres secciones, las cuales son:

- Facultad de Ingeniería (los códigos de las meterías inician por el numero 7)
- Facultad de Ciencias (los códigos de las meterías inician por el numero 1)

- Otras Facultades (Economía, Humanidades, entre otras)

Con esta información se obtiene un promedio ponderado a cada de una de las tres secciones que se construyeron.

3.1.3 Otros Aspectos

Para este estudio, se tiene en cuenta otros datos muy importantes para el dateSet, como, la edad de inscripción, sexo, etc. estos datos pueden ser muy valiosos para predecir que estudiantes pueden incurrir en bajo rendimiento, ya que, en algunos estudios, muestran un mal rendimiento académico al ingresar con una temprana edad a la educación superior.

3.1.4 Estructura del DataSet

En la siguiente tabla se muestra la estructura final que tiene el dataSet, el cual será utilizado con las técnicas de minería de datos. Para la construcción se utilizaron los registros de los estudiantes desde el 2013 hasta el 2019 de la universidad del valle sede Tuluá, la siguiente tabla muestra como quedo la estructura de dataSet.

Tabla 3.1: Estructura del DataSet

Variable	Tipo de Datos	Descripsion
programaAdmicion	Numérico	Pregrado (1). Tecnología (0).
edadIngreso	Numérico	Edad del estudiante en el momento de la admisión.
graduado	Numérico	(0) No graduado. (1) graduado.
ciudad	Numérico	(0) Si vive en Tuluá. (1) si vive fuera de Tuluá.
sexo	Categórico	(F) Femenino. (M) Masculino.
condicionEx	Numérico	(0) No posee excepción de admisión. (1) Posee excepción de admisión.
jornada	Categórico	(DIU) Jornada diurna. (NOC) Jornada Nocturna.
CantidadSemestres	Numérico	Semestres cursados por el estudiante.
proSemestral	Numérico	Promedio total de Estudiante durante su vida universitaria.
ingeniería	Numérico	Promedio ponderado de las materias de la Facultad de Ingeniería.
ciencias	Numérico	Promedio ponderado de las materias de la Facultad de Ciencias.
otrasMaterias	Numérico	Promedio ponderado de las materias de otras Facultades.
creditosAprobados	Numérico	Porcentaje de créditos Aprobados.

Variable	Tipo de Datos	Descripsion
creditosReprobados	Numérico	Porcentaje de créditos reprobados.
materiasCanceladas	Numérico	Porcentaje de materias canceladas.
materiasHabilitadas	Numérico	Porcentaje de materias habilitadas.
BajoRendimiento	Numérico	(0) No Posee un bajo rendimiento. (1) Posee un bajo rendimiento.

3.1.2. Selección de la información

Cuando se obtuvo la base de datos suministrada por la Universidad del Valle sede Tuluá, contenía más de 42 mil datos registrados sobre el rendimiento de los estudiantes, desde el año 2013 hasta el 2019, esta información me fue suministrada por el profesor Jaime H. Escobar y su grupo de trabajo.

En el proceso de análisis, se observó toda la información con la que venía la base de datos, donde se eliminaron las columnas duplicadas, al final los atributos que se obtuvieron para la creación del dataSet fueron:

- ID, código para identificar cada estudiante.
- SNP (Servicio Nacional de Pruebas), el cual es el código asignado por el ICFES.
- Programa de admisión.
- Jornada admisión.
- Periodo de matrícula, (Febrero/Junio - Agosto/Diciembre).
- ID Programa matricula, código de la carrera.
- Jornada de matrícula.
- Fecha de cancelación.
- Promedio semestral (de cada periodo).
- Retiro por BRA (Bajo Rendimiento Académico).
- Código materia.
- Créditos de la asignatura.
- Tipo Matricula.
- Calificación.
- Habilitación.
- Estado de la asignatura, activa o cancelada.
- Fecha cancelación de asignatura.
- Estado de calificación, aprobada, reprobada o ninguna.
- Año de admisión.
- Semestre de admisión, primer o segundo periodo del año.
- Edad inscripción.
- Condición de excepción a la que aplico.
- Sexo.

- Ciudad de residencia.
- Barrio.
- Total de créditos del programa.
- Periodos programa.
- Matriculas en el programa de ingreso.
- Total de semestres que el estudiante asistió.
- Total de BRA (Bajo Rendimiento Académico).
- Fecha de grado.
- Programa de grado.

3.2. Preprocesamiento de la información

Con la caracterización de los datos, se continuó con la creación del dataSet, lo primero que se realizó, fue determinar la cantidad de estudiantes que se tenían registrado, puesto que, los 42 mil datos correspondían a 2.244 estudiantes, De esta forma comenzó a borrarse valores que eran únicos y se repetían por cada registro, como numero de BRA, jornada de admisión, programa de admisión, entre otros datos. En el caso de los atributos tipo fecha, había problemas ya que estos venían en un formato numérico general y al momento de transformarlos en datos tipo fecha, nos mostraban años por debajo de 1970, lo cual es ilógico ya que se tienen datos a partir del 2013, estos valores tipo fecha, solo nos sirvieron para verificar que el estudiante se graduó o no, pero no determina el año exacto de su graduación.

Después de eliminar los valores únicos que se repetían, se comenzó con el proceso de obtener datos calculados a partir de los registros de cada estudiante, los cuales fueron:

Tabla 3.2: Obtención de atributos

Atributo	**Obtención**
Graduado	Se genero a partir de los estudiantes que se les registrara una fecha de grado, donde “1” significara graduado y “0” no graduado.
Ciudad	Se determina si el estudiante vive en Tuluá o fuera de ella, donde “0” es estar donde de la ciudad de Tuluá y “1” es todo lo contrario.
Condición de excepción	Al ingresar a la Universidad muchos estudiantes son aceptados por excepciones dadas por leyes, lo que se realizó, fue determinar que estudiantes no posee ninguna excepción dándole un valor de “0” y en caso de poseer alguna, fue marcado con “1”.
Promedio de los semestres cursados	Se obtiene un promedio total de la duración del estudiante en su trayecto universitario.
Materias de la Facultad de Ingeniería.	Se buscan las materias donde el primer digito de su código es 7 y se comiza a calcular el promedio ponderado.
Materias de la Facultad de ciencias	Se buscan las materias donde el primer digito de su código es 1 y se comiza a calcular el promedio ponderado.

Atributo	Obtención
Materias de otras Facultades	Se Calcula el promedio ponderado al resto de materias que no pertenezcan a la facultad de Ciencias e Ingeniería.
Créditos aprobados	Se obtiene el total de créditos cursados por el estudiante y la cantidad de créditos aprobados, para después obtener un porcentaje (créditos aprobados/total créditos matriculados).
Créditos reprobados	Se obtiene el total de créditos cursados por el estudiante y la cantidad de créditos reprobados, para después obtener un porcentaje (créditos reprobados /total créditos matriculados).
Materias Canceladas	Obtener el porcentaje de materias canceladas al total de materias cursadas por el estudiante.
Materias habilitadas	Obtener el porcentaje de materias habilitadas al total de materias cursadas por el estudiante.
Posee Bajo Rendimiento	Es una etiqueta que se le asigna a los estudiantes con bajo rendimiento académico, donde posee un bajo rendimiento es “1” y en caso contrario es “0”

Como se puede analizar, se transformaron muchos datos a valores binario (0 y 1), puesto que la principal idea, es crear un dataSet bien estructurados para algoritmos de clasificación binaria, todos estos datos fueron manipulados en Python, con la Liberia Pandas la cual nos permite leer archivos tipo CSV (Comma Separated Values, que en español significa valores separados por comas) y crear un DataFrame con la base de datos original, ya con el archivo cargado se crearon y utilizaron funciones que nos permitían calcular todos estos datos que ya fueron explicados.

Ya con el dataSet creado, se obtiene un mejor modo de ver la información suministrada por la Universidad del Valle, sobre el rendimiento académico de los estudiantes.

3.3. Selección de los dos algoritmos

Tabla 3.3: Características del proyecto

Características	**Modelos**				
	K-Nearest-Neighbor (KNN)	**Naive Bayes**	**Máquinas de vectores de soporte (SMV)**	**Arboles de Decisión**	**Redes Neuronales Artificiales (RNA)**
Multi-dimensional		X	X	X	X
Variables categóricas	X	X	X	X	X
Predecir	X		X	X	X
Clasificar	X	X	X	X	X
Elementos clasificados etiquetados	X	X	X	X	X
Variables descriptivas		X	X	X	X
Aprendizaje supervisado	X	X	X	X	
Análisis de varias características		X	X	X	X

Tabla 3.4: Ventajas y desventajas de las técnicas de minería de datos

Técnicas	Ventajas	desventajas
K-Nearest-Neighbor (KNN)	- Alta precisión. - Manejo de Clasificaciones múltiples. - Sencillo y fácil de imple-mentar.	- Baja eficiencia depen-diendo si los datos de entrenamiento y prueba son muy grandes. - Depende demasido de los datos de entrenamiento, en caso de tener uno o dos datos erróneos, puede generar a futuro inexactitud en datos predichos. -KNN no es muy bueno procesando datos multidimensionales.
Naive Bayes	- Se pueden integrar múltiples variables para la clasificación los datos. - fácil y rápido de implementar. - No requiere demasiada memoria y se puede utilizar para el aprendizaje en línea.	- Los predictores se consideran independientes entre sí, lo cual puede que genere un modelo que no se ajuste correctamente a los datos. - Si hay un exceso de caracterización, habrá un aumento de esfuerzo computacional. - Sufre al tener características irrelevantes.
Máquinas de vectores de soporte (SVM).	- Buen manejo multidimensional. - Se pueden modelar relaciones complejas, no lineales. - En las funciones de decisión utiliza subconjuntos de puntos de entrenamiento, al final mostrando una mejor eficiencia en memoria.	- Las Máquinas de Vectores de Soporte no son adecuadas para grandes conjuntos de datos, ya que toma demasiado tiempo para entrenar. - Funciona mal con las clases superpuestas
Arboles de Decisión	- Simples de entender y de interpretar. - No requiere de datos excesivamente complejos. - Fáciles de combinar con otras herramientas en la toma de decisiones. -Utiliza un modelo de caja blanca: la respuesta del algoritmo es fácilmente justificable a partir de la lógica booleana implementada en él.	- Son inestables: cualquier pequeño cambio puede generar un árbol de daciones rotalmente diferente. - Un árbol de decisión puede llegar a ser demasiado complejo con facilidad, perdiendo su utilidad. - En ocasiones no es utilizado por ser un algoritmo tan sencillo y no tan poderoso para datos complejos.
Redes Neuronales Artificiales (RNA)	- Son modelos de vanguardia que capturan de una forma óptima y efectiva características complejas, obteniendo resultados con una alta precisión. - El procesado de la información es local, es decir que, al estar compuesto por unidades individuales de procesamiento,	- Complejidad de aprendizaje para grandes tareas, cuantas más cosas se necesiten que aprenda una red, más complicado será enseñarle. - Tiempo de aprendizaje elevado. Esto depende de dos factores: primero si se incrementa la cantidad de patrones a identificar o clasificar, y segundo, si se requiere

	dependiendo de sus entradas y pesos. - Alta tolerancia a fallos ya que esta almacena información de forma redundante, lo cual le permite seguir respondiendo de manera aceptable aun si s daña parcialmente.	mayor flexibilidad o capacidad de adaptación de la red neuronal para reconocer patrones que sean sumamente parecidos. - Elevada cantidad de datos para el entrenamiento, cuanto más flexible se requiera que sea la red neuronal, más información tendrá que enseñarle para que realice de forma adecuada la identificación.

De acuerdo al problema modelado, los antecedentes que se tienen sobre el rendimiento académico y las características del dataSet modelado, se toma la decisión de seleccionar dos algoritmos, que se ajusten a las necesidades que tenemos y de cumplimiento en los objetivos establecidos, con el fin de realizar el prototipo de alerta temprana.

Arboles de Decisión: Se escogió esta técnica dadas las ventajas identificadas en cuanto a la toma de decisiones, también se ha visto en trabajos donde se aplica esta para obtener resultados superiore al 75% de acierto, tomando en cuenta lo nombrado en la tabla 3.4; es muy fácil de implementar y entender los resultados obtenidos, posee la facilidad de combinar con otras herramientas de predicción, asimismo se tiene en cuenta que el objetivo principal, es determinar que estudiante puede incurrir en un bajo rendimiento académico o no incurre en este.

Máquinas de vectores de soporte (SVM): este algoritmo fue considerado como una técnica muy útil para el desarrollo del proyecto, gracias a los posibles resultados que se esperaban, puesto que al final es una decisión binaria, esto lo podría representar muy fácilmente en un plano bidimensional, también se tiene en cuenta que los atributos del dataSet, no tendrá un costo computación elevado y así se podrá realizar el trabajo de clasificar los estudiantes en las dos clases existentes separándolos mediante una línea en el plano donde se encuentra los datos.

3.4. Técnica de muestreo

Por el tipo de proyecto de considero indispensable el uso de una técnica de muestreo, que nos permitiera seleccionar de forma óptima los datos para el entrenamiento y la prueba de los algoritmos, Teniendo en cuenta la cantidad de datos que se posee y los algoritmos que se manejan (Arboles de decisiones y SVM), se elegio el muestreo aleatorio simple, dado que cada registro tiene la misma probabilidad de pertenecer a la muestra, cabe resaltar que no es recomendable implementar este técnica de muestreo en bases de datos pequeñas, ya que puede no representar de forma oportuna la población. Dicho lo anterior, se asigna un número que le permitirá generar “números aleatorios”, que van a corresponder al grupo de pruebas, Con respecto a la división de la base de datos, el 25% de los datos fueron asignados a pruebas y el resto serán utilizados para el entrenamiento.

3.5. Configuración de los modelos seleccionados

Para el uso de las técnicas escogidas, primero se importó el dataSet con la información filtrada, donde todos sus datos son numéricos, con el objetivo de facilitar el procesamiento de los datos, el siguiente paso es separar el campo “BajoRendimiento” del dataSet cargado, esta información se guardará en la variable “y”, dado que contiene los datos a predecir y la variable X tendrá el restante del dataSet, todo esto es realizado mediante la librería Pandas de Python, en relación con lo anterior, se comienza a particionar los datos en entrenamiento y prueba, para esto se utilizó “train_test_split” perteneciente a la librería “sklearn.model_selection”, donde se crean las variables X_train, X_test, y_train, y_test, con particiones de 75%-25%.
Al tener los datos seleccionados, se procede a configurar cada modelo, junto a la información recolectada de la siguiente forma:

Máquinas de Vectores de Soporte (SVM): posteriormente de tener los datos de entrenamiento y de testeo, se utiliza SVC de sklearn.svm, se especificó el tipo de kernel que se utilizó para este, el cual fue “linear”, las demás opciones se dejaron predeterminadas, ya que se observó que, al modificar estos parámetros, los resultados obtenidos no cambiaban; después se efectúa el entrenamiento con la función “.fit” donde se le pasan las variables X_train y y_train

Arboles de Decisión: Después de tener los datos particionados, se empieza a configurar el algoritmo de árboles de decisión de la librería sklearn, donde se fija el criterio para la construcción del árbol, en este caso se aplica “entrompy”, el cual trabaja con probabilidades y nos permite evaluar la ganancia del nodo construido, así mismo se le designa el valor de 12 al parámetro max_deph, dado que permite controlar la profundidad, lo cual genera normas concluyentes, poco repetitivas y ayuda a mejora la calidad de la clasificación, además se asignan el valor “best” al parámetro splitter, este permite construir el árbol a partir del mejor atributo, las demás opciones del algoritmo de dejaron predeterminadas, ya

que lo resultados no varían. En la documentación de recomiendan modificar estos parámetros, para que algoritmo no consuma demasiados recursos computacionales, pero depende del dataSet que se utilicé y en este caso no hay ese problema.

3.6. Plan de pruebas

Después de la preparación del dataSet y la selección de las dos técnicas de minería de datos, sigue la fase donde se implementaron los algoritmos, con sus respectivas configuraciones, en el caso de Máquinas de Vectores, el único ajuste fue especificar qué tipo de kernel, que nos permitió obtener resultados positivos y en el Árbol de Decisiones se modificaron tres características, las cuales dieron resultados oportunos para satisfacer los objetivos pactados.

Arboles de Decisiones: Para este algoritmo, se inicia con la fase del entrenamiento, la cual utiliza el 75% del dataSet y el restante fue utilizado en la fase de pruebas; luego se hacen determinados ajustes en el algoritmo, con el fin de obtener una eficacia mayor al 80%. Para dejara todo más claro, este es el paso a paso de las pruebas:

- Se entrena el algoritmo con el 75% del dataSet.
- Con el algoritmo ya entrenado, se procede a evaluar el modelo, con el restante de la información del dataSet.
- Con un conjunto de herramientas, se evalúa si los resultados obtenidos son adecuados o se debe modificar los parámetros o el árbol para la ejecución.

Máquinas de Vectores de Soporte (SVM): Para este modelo, la fase de entrenamiento es la misma que Arboles de Decisión, en la siguiente fase se transformaron los predictores (datos de entrada) en un espacio de características altamente dimensional, para esta fase solo basto con especificar el kernel que se utilizó para llevar a cabo el proceso; los datos nunca se transforman explícitamente al espacio de características. Esta fase se conoce comúnmente como el truco kernel.

Después de la fase de clasificación, el siguiente paso es comparar los resultados obtenidos, por medio de la curva ROC como lo muestra la figura 4.1 y también se detallan las métricas realizadas en cada prueba, todo con el finde de analizar y determinar que tan eficiente es la técnica.

Capítulo 4
Análisis y discusión de resultados

Una vez terminada la fase de entrenamiento y test, con las dos técnicas de minería de datos, se recogieron unos resultados obtenidos, por medio de métricas de evaluación de modelos de aprendizaje, como los son precisión, F1, exactitud, sensibilidad y curva ROC, toda la información se puede apreciar en la tabla 4.1, todo con el propósito de comparar los resultados de cada modelo y así escoger el algoritmo que fue implementado.

Tabla 4.1: Resultados métricas de evaluación de los modelos

	Precisión		Recall score		F1 Score		Accuracy		Matriz de Confusion	
Prueba	SVM	Arboles De Decision	SVM	Arboles De Decision	SVM	Arboles De Decision	SVM	Arboles De Decision	SVM	Arboles De Decision
1	0,875	0,8253	0,7915	0,8516	0,8312	0,8383	0,8378	0,8342	[[246 32] [59 224]]	[[227 51] [42 241]]
2	0,876	0,8345	0,7915	0,8551	0,8312	0,8447	0,8378	0,8414	[[246 32] [59 224]]	[[230 48] [41 242]]
3	0,877	0,8333	0,7915	0,8481	0,8312	0,8406	0,8378	0,8378	[[246 32] [59 224]]	[[230 48] [43 240]]
4	0,878	0,8316	0,7915	0,8551	0,8312	0,8432	0,8378	0,8396	[[246 32] [59 224]]	[[229 49] [41 242]]
5	0,879	0,8294	0,7915	0,8587	0,8312	0,8438	0,8378	0,8396	[[246 32] [59 224]]	[[228 50] [40 243]]

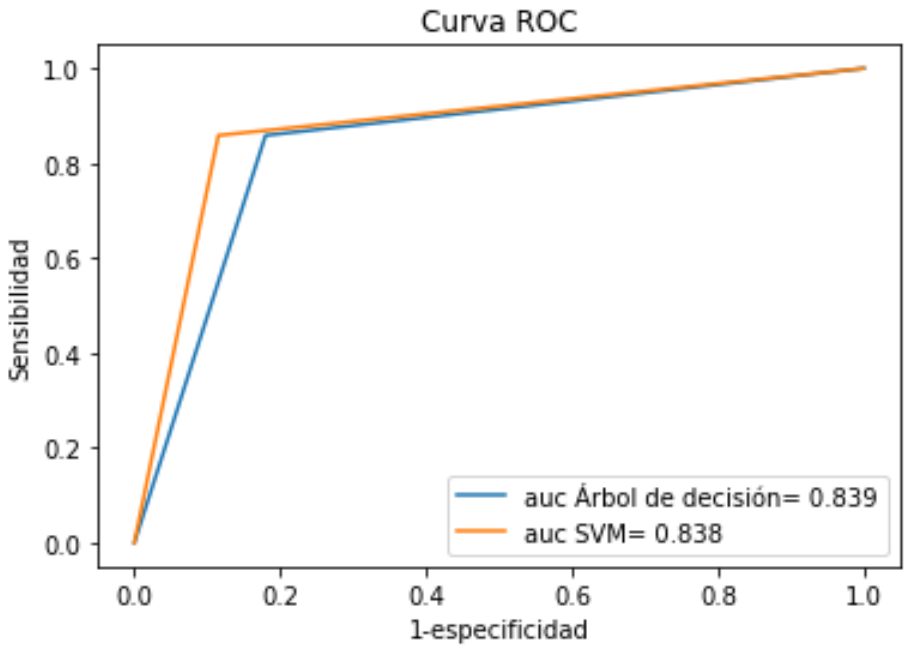

Figura 4.1: Curva ROC de SVM y Árbol de decisión.
Fuente: Elaboración propio

Al analizar los valores de la curva ROC, vemos ambos modelos son muy parejos en sus resultados, ya que poseen una diferencia de 0.001, donde Arboles de Decisión posee un 0.839 de AUC (Área Bajo la Curva), con esto hay una mayor probabilidad de distinguir instancias positivas y negativas, haciendo que el algoritmo sea más eficiente, para la problemática planteada.

Precisión: Los resultados que se obtuvieron de ambos modelos, son muy positivos, ya que son mayores a 0.8, como se puede apreciar en la figura 4.2, donde SVM obtuvo un 0.875 en los 5 intentos y en Arboles de Decisión oscilaba entre 0.82 y 0.84; es necesario recalcar que la precisión señala la razón que hay entre número de verdaderos positivos y falsos positivos. Para calcular estos datos, se utilizó la función “precisión_score” de la librería “sklearn.metrics”, este método recibe como parámetros el conjunto de testeo obtenido del dataSet (y_test) y el conjunto de predicciones de cada algoritmo, los valores de la precisión los puedes detallar en la Tabla 4.1.

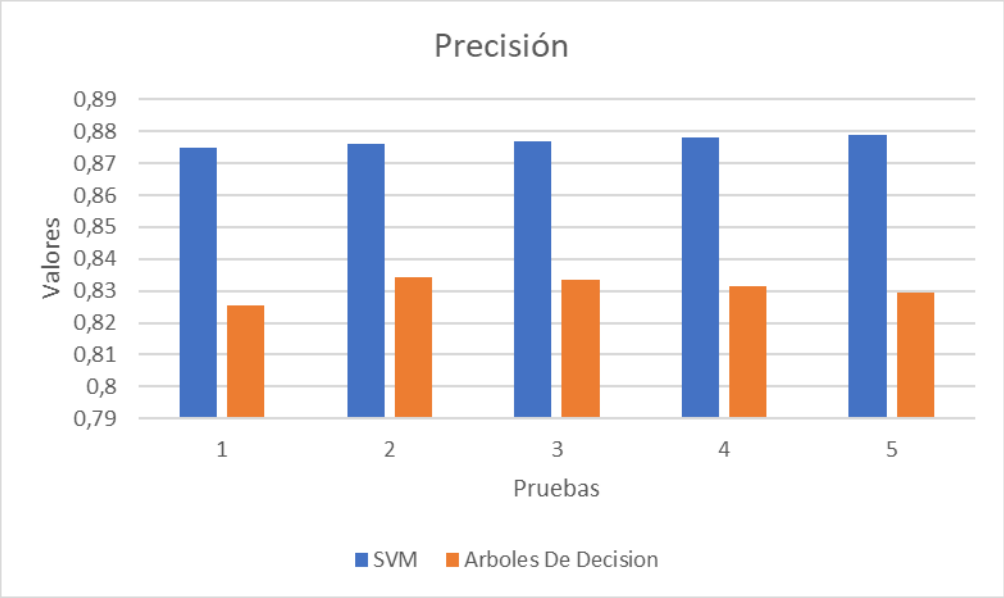

Figura 4.2: Grafica comparativa precisión SVM vs Arboles de decisión.
Fuente: Elaboración propia

Recall score: Para calcular esta métrica se utilizó la librería ya mencionada anteriormente, esta función nos determina, ¿qué porcentaje de predicciones se lograron identificar?, donde SMV obtuvo 0.7915 y en el caso de Arboles de Decisión, sus valores promediaban un 0.85, los parámetros que se utilizan para calcular Recall score son: el conjunto de testeo del dataSet (y_test) y el conjunto donde se guardaron las predicciones de los algoritmos, estos resultados los puede apreciar en la figura tabla 4.1 y en la figura 4.3.

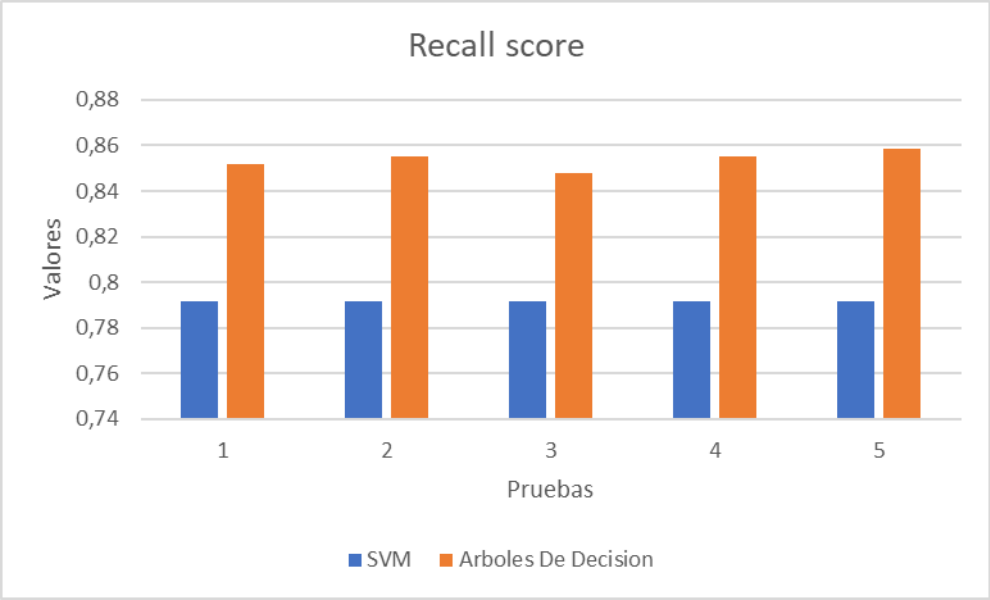

Figura 4.3: Grafica comparativa Recall score SVM vs Arboles de decisión.
Fuente: Elaboración propia

F1 score: la síguete métrica combina la Precisión y el Recall en una sola medida, es decir, entre más cercano sea el valor a 1, mejores son las medidas de Precisión y el Recall; esta función "f1_score" proviene de la librería anteriormente nombrada y recibe como parámetros el conjunto de testeo (y_test) y el conjunto de predicciones obtenidas por los algoritmos, para ver los valores específicamente, mirar la tabla 4.1, también se puede detallar las diferencias entre los resultados en la figura 4.4.

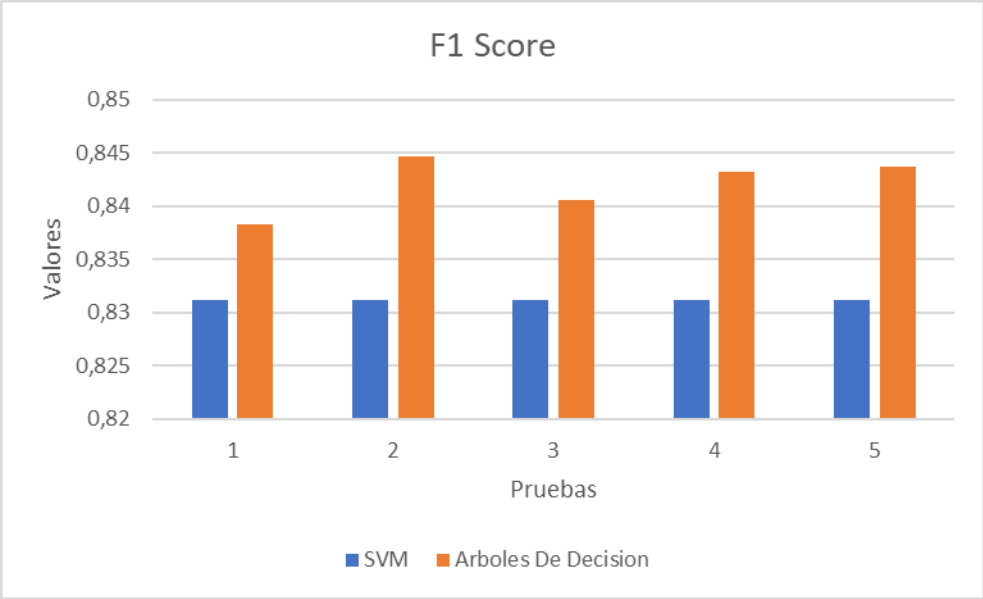

Figura 4.4: Grafica comparativa F1 score SVM vs Arboles de decisión.
Fuente: Elaboración propia

Accuracy: Para esta métrica se utilizó la función “accurecy_score”, la cual, suma los verdaderos positivos y los verdaderos negativos, sobre la sumatoria de los campos de la matriz de confusión, los parámetros que recibe son: el conjunto de testeo (y_test) y el conjunto de predicciones obtenidas por los algoritmos. El mejor resultado obtenido fue el de Arboles de Decisión en la segunda prueba con un 0.8414.

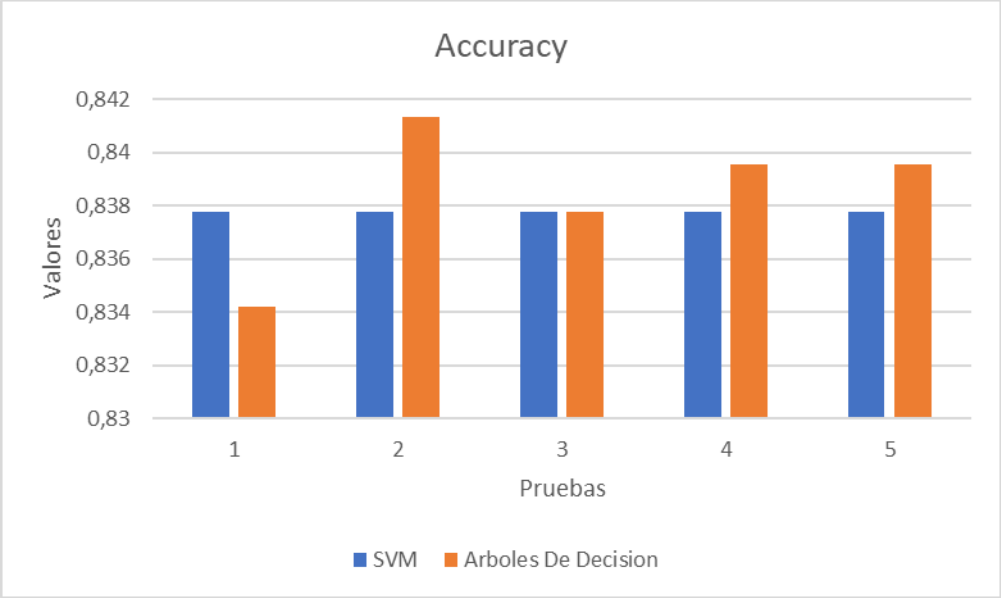

Figura 4.5: Grafica comparativa Accuracy SVM vs Arboles de decisión.
Fuente: Elaboración propia

Al obtener los datos proporcionados por las métricas, se puede observar que ambos algoritmos alcanzaron resultados entre el 79% y el 87%, además, las diferencia entre los valores, tampoco es muy significativa, pero cumpliendo con unos de los objetivos de este trabajo, se procede a escoger uno de los dos algoritmos implementados, en esta caso se optó por Árboles de Decisiones, ya que en muchos valores obtenidos por las métricas, nos mostraba una superioridad frente a Maquina de Vectores de Soporte, por otra parte, una ventaja de Arboles de Decisión, es poder visualizar las reglas, a la hora de realizar la clasificación.

En la fase de configuración del modelo de Árboles de Decisión, se modificaron algunos parámetros, pero el más fundamental fue la profundidad (max_depth), ya que, iniciando con 8 de altura, se omitían datos a la hora de clasificar, haciendo que las medidas de las métricas fueran disminuyendo, cuando se empieza a incrementar la altura, sus medidas mejoran; Después de varios intentos, se determina que en la altura 12, se obtienen buenos resultados y evitamos la perdida de información. En la figura 4.6 se puede apreciar el árbol construido, con una altura máxima de 3, cabe aclarar que la altura no modifica el orden de las reglas, en este sentido, si el árbol tiene altura 12 o 3, las primeras siempre serán las mismas condiciones.

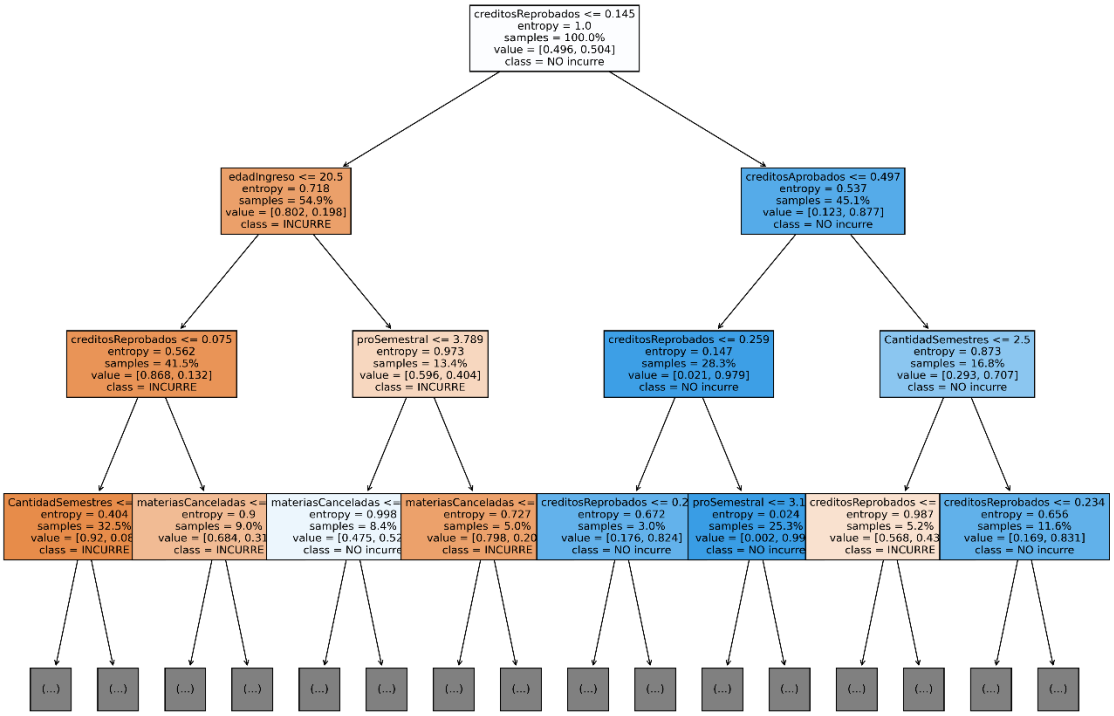

Figura 4.6: Grafica del Árbol de decisión.
Fuente: Elaboración propia

En la figura 4.7, se puede apreciar una pequeña parte de las reglas o condiciones, que el modelo de árboles de decisión diseño, a partir de la fase de entrenamiento. observando de forma mas completa las condiciones, se generan muchas condiciones para una misma variable, en gran ejemplo es la edad de ingreso, donde muchas veces se pregunta si es menor de 18 años, mayor a 20 años, entre otras condiciones, y así es para muchas variables, de acuerdo a la profundidad o el nodo por donde se está realizando la observación.

```
|--- creditosReprobados <= 0.15
|   |--- edadIngreso <= 20.50
|   |   |--- creditosReprobados <= 0.07
|   |   |   |--- CantidadSemestres <= 12.50
|   |   |   |   |--- ingenieria <= 4.01
|   |   |   |   |   |--- CantidadSemestres <= 7.50
|   |   |   |   |   |   |--- edadIngreso <= 18.50
|   |   |   |   |   |   |   |--- ciencias <= 3.22
|   |   |   |   |   |   |   |   |--- ingenieria <= 3.75
|   |   |   |   |   |   |   |   |   |--- edadIngreso <= 16.50
|   |   |   |   |   |   |   |   |   |   |--- programaAdmicion <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |--- edadIngreso >  18.50
|   |   |   |   |   |   |   |--- CantidadSemestres <= 1.50
|   |   |   |   |   |   |   |   |--- ciudad <= 0.50
|   |   |   |   |   |   |   |   |   |--- class: 0
|   |   |   |   |   |   |   |   |--- ciudad >  0.50
|   |   |   |   |   |   |   |   |   |--- sexo_M <= 0.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1
```

Figura 4.7: estructura de las reglas del Árbol de decisión.
Fuente: Elaboración propia

A continuación, se mostrarán algunas gráficas, que reflejan las clasificaciones acertadas por el modelo de Arboles de Decisión.

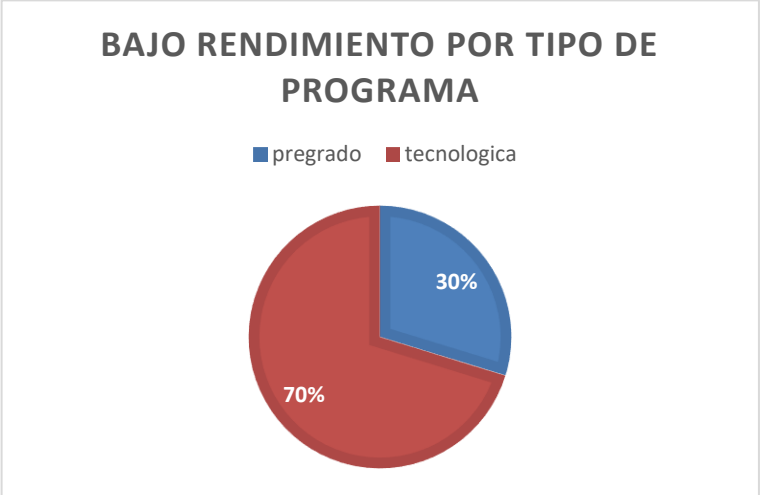

Figura 4.8: Bajo rendimiento por tipo de programa.
Fuente: Elaboración propia

Como se puede observar en la figura 4.8, nos divide la cantidad de estudiantes que incurrieron en un BRA. Se muestra que las carreras tecnológicas poseen 170 datos, que equivalen al 70% de un total de 242 datos, los registros restantes pertenecen a las carreras de pregrado.

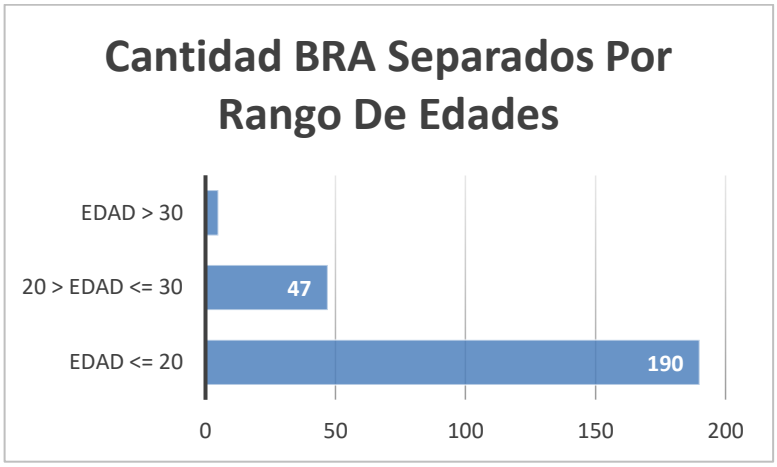

Figura 4.9: Bajo rendimiento por rango de edades.
Fuente: Elaboración propia

En la figura 4.9, se puede apreciar los rangos de edad de admisión, de los estudiantes que incurrieron en un BRA, donde los estudiantes con una edad inferior o igual a los 20 años, son los más afectados.

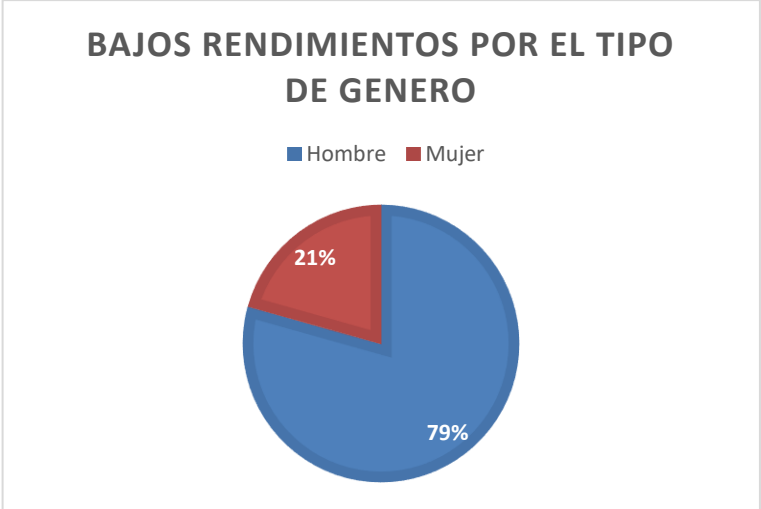

Figura 4.10: Bajo rendimiento por el tipo de género.
Fuente: Elaboración propia

En la figura 4.10, se divida los 242 registros en el tipo de genero del estudiante, donde la categoría hombre posee 192 datos, en comparación con los 50 datos de la categoría mujer, analizando estos valores, hay una gran diferencia, debido a que hay más hombre, que mujeres en la Facultad de Ingeniería.

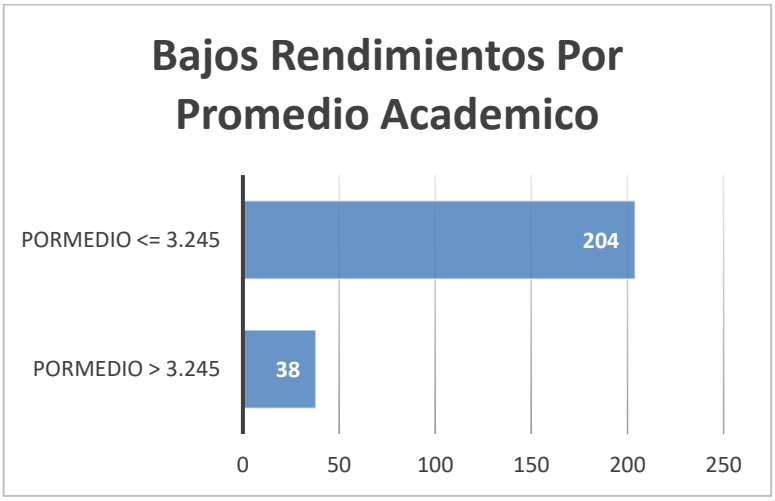

Figura 4.11: Bajo rendimiento por promedio académico.
Fuente: Elaboración propia

Observando la figura 4.11, se logra identificar, que la mayoría de estudiantes con un promedio menor a 3.245, tienen una tendencia muy alta a incurrir en un bajo rendimiento académico.

Capítulo 5
Conclusiones y trabajos futuros

5.1. Conclusiones

1. Con el análisis de las métricas: precisión, F1, exactitud, sensibilidad y curva ROC, se logran ver resultados superiores al 80%, en los dos médelos seleccionados de minería de datos, escogiendo Arboles de Decisión por motivos de superioridad, comparándolo con Maquinas de Vectores de Soporte, de tal forma, que puedo determinar con mayor certeza que estudiante puede incurrir en un bajo rendimiento académico.
2. La información de fue suministrada por la Universidad del Valle sede Tuluá, correspondiente a registros académico, institucionales, sociales y personales, nos permitió hallar patrones en estudiantes que incurren en BRA, facilitando el desarrollo de los objetivos y las tareas planteadas en el proyecto.
3. El análisis de datos y la construcción del dataSet, nos permitió encontrar problemas de calidad como: el ruido, datos faltantes, erróneos, y atípicos (outliers); el filtrar estos datos, mejora el dataSet y da un valor añadido, de tal forma que mejora los resultados, de la implementación los modelos de minería de datos.
4. La elección de las técnicas de Árboles de Decisión y Maquinas de Vectores de Soporte, se fundamento en el tipo de proyecto que se desarrolló y también en las características proporcionadas por las técnicas. A su vez, se revisaron varios algoritmos, donde se determinó que muchos de estos, son para problemas mas complejos y analizando la problemática del proyecto, se confirman como poco precisas.
5. En conclusión, poseer un prototipo de sistemas de alerta temprana de bajo rendimiento académico, en los programas de la Facultad de Ingeniería, le permitirá a la Universidad del Valle sede Tuluá, diseñar planes de acompañamiento a estudiantes que tengan este riego, con el fin de disminuir estas cifras tan alarmantes y al mismo tiempo investigar mas a fondo, la problemática del bajo rendimiento académico, que vive la Universidad del Valle, todo esto podría generar grandes impactos sociales dentro y fuera de ella.

5.2. Trabajo Futuros

1. Se recomienda ampliar la base de conocimiento, con información socioeconómica y con el registro de las pruebas SABER 11, que se encuentra por medio del ICFES, todo esto con el fin de mejorar la precisión de los modelos.
2. Otra recomendación, es poder tener un registro de los cortes evaluativos (notas de parciales), esto con el fin de poder realizar una predicción en el trascurso del semestre, puesto que se está perdiendo información valiosa, que es utilizada en este tipo estudios, como se evidencia en muchas investigaciones, donde se posee la misma problemática del bajo rendimiento académico.
3. Ampliar el alcance a otras facultades, teniendo en cuenta características y dificultades de estas, ya que este proyecto esta orientado, a la Facultad de Ingeniería de la Universidad del Valle sede Tuluá.
4. Se recomida expandir la implementación, de técnica de minería de datos, con el fin de evaluar los resultados que se pueden obtener.
5. Aumentar el alcance del proyecto a sede regionales, en relación con la información de estas misma, todo con el fin de mejorar el estudio y tener un panorama más amplio, sobre la problemática del bajo rendimiento académico.

Capítulo 6: Bibliografía

1. Cesar Higinio Menacho Chiok “Predicción del rendimiento académico aplicando técnicas de minería de datos”.
2. David Luis La Red Martínez, Marcelo Karanik, Mirtha giovannini, noelia Pinto “Perfiles de Rendimiento Académico: Un Modelo basado en Minería de datos”
3. “Ley 1581 de 2012 protección de datos personales,” 2012.
4. Katherine Contreras, Carmen Caballero, Jorge Palacio y Ana María Pérez “Factores asociados al fracaso académico en estudiantes universitarios de Barranquilla (Colombia)”
5. Ana María Luque-Clavijo, Félix Germán Fajardo-Prieto “APLICACIONES DE LA MINERÍA TECNOLÓGICA PARA LA GESTIÓN DE PROYECTOS DE INGENIERÍA”.
6. R. Alcover, J. Benlloch, P. Blesa, M. A. Calduch... “Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos”.
7. Huerta Luis, Ruiz Juan, Cabrera Nubia, Montiel Luis, Benítez Felipe, Ramírez Víctor “Minería de datos: Impacto de Actividades Cotidianas en el Rendimiento Estudiantil”.
8. Linda Marina Padua Rodríguez “Factores individuales y familiares asociados al bajo rendimiento académico en estudiantes universitarios”.
9. Karina Eckert, Roberto Suénaga “Aplicación de técnicas de Minería de Datos al análisis de situación y comportamiento académico de alumnos”.
10. Jessica Johanna Morales Carrillo, Viviana Katherine Trujillo Utreras, Sulay Katerine Cevallos Molina, Santana Cedeño Hiraida Monserrate. “MINERÍA DE DATOS EN EDUCACIÓN: UNA REVISIÓN LITERARIA”.
11. Norka Bedregal-Alpaca, Danitza Aruquipa-Velazco, Víctor Cornejo-Aparicio “Técnicas de Data Mining para extraer perfiles comportamiento académico y predecir la deserción universitaria”.
12. Minnieli Álvarez Suárez “Factores que inciden en el bajo rendimiento en matemáticas en las pruebas puertorriqueñas de aprovechamiento académico desde la perspectiva de los maestros de esta área de enseñanza”.
13. Dra. Fabiola Cruz Núñez, Dr. Abel Quinones Urquijo “Autoestima y rendimiento académico en estudiantes de enfermería de Poza Rica, Veracruz, México”.
14. Valle-Arias Antonio, Regueiro-Fernández Bibiana, Suárez-Fernández Natalia, etc “Rendimiento académico, enfoques de trabajo e implicación en los deberes escolares”.
15. María Antonieta Elvira-Valdés“Autorregulación y rendimiento académico en la transición secundaria-universidad”.
16. “Clasificar con K-Nearest-Neighbor” url: Algoritmo k-Nearest Neighbor | Aprende Machine Learning

17. “Algoritmos Naive Bayes: Fundamentos e Implementación” url: Algoritmos Naive Bayes: Fundamentos e Implementación | by Victor Roman | Ciencia y Datos | Medium
18. “Máquinas de vectores de soporte” url: Máquinas de vectores de soporte - Wikipedia, la enciclopedia libre
19. “Máquinas de Vectores de Soporte (SVM)” url: Máquinas de Vectores de Soporte (SVM) - IArtificial.net
20. “Qué es un diagrama de árbol de decisión” url: ¿Qué es un diagrama de árbol de decisión? | Lucidchart
21. “Árbol de decisión, una herramienta para decidir bien” url: Árbol de decisión, una herramienta para decidir bien - Alto Nivel
22. “Red neuronal artificial” url: Red neuronal artificial - Wikipedia, la enciclopedia libre
23. “Qué son las redes neuronales y sus funciones” url: Qué son las redes neuronales y sus funciones | ATRIA Innovation

Anexo 1

Código Fuente

El desarrollo de este prototipo, se realizó en Python y sen encuentra en un repositorio en Github, disponible en el siguiente enlace:

- https://github.com/daniell107/TrabajoDeGradoProtoripo.git

Anexo 2

Diseño del prototipo

El prototipo realizado en este proyecto, tiene la funcionalidad de predecir que estudiante puede incurrir en bajo rendimiento, esto es enfocado a estudiantes de la Facultad de Ingeniería, de la Universidad del Valle sede Tuluá. A continuación se encuentra una imagen del prototipo en su uso.