Implementación de un clasificador binario para determinar si un estudiante puede incurrir en deserción en la sede Tuluá de la subregión centro de la Universidad del Valle

Carlos Daniel Marín Mosquera
Andrés Mosquera Alvarado

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

Implementación de un clasificador binario para determinar si un estudiante puede incurrir en deserción en la sede Tuluá de la subregión centro de la Universidad del Valle

Carlos Daniel Marin Mosquera
Código 201663787
carlos.daniel.marin@correounivalle.edu.co
Andrés Mosquera Alvarado
Código 201664103
andres.mosquera@correounivalle.edu.co

Director: Mauricio Lopez Benitez

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
Carlos Daniel Marin Mosquera,
Andres Mosquera Alvarado
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 implementando una técnica de minería de datos enfocado a la deserción en la sede Tuluá de la subregión centro de la Universidad del Valle, que sirva como herramienta para las directivas de la sede permitiendo generación de estrategias para trabajar en la disminución de este fenómeno. A lo largo de este documento se presenta la investigación que se llevó a cabo sobre la problemática, el proceso para obtener una base de conocimiento adecuada para el proyecto, las actividades de preprocesamiento de la información para la obtención del dataset, la selección de dos modelos de minería de datos y la posterior implementación de ellos para finalmente obtener los resultados que ayudaron a la culminación del prototipo y por ende del proyecto.

Palabras claves: Prototipo, dataset, deserción universitaria, minería de datos.

Abstract

This project was elaborated with the purpose of making a prototype implementing a data mining technique focused on desertion in the Tuluá branch of the central sub-region of the Universidad del Valle, which serves as a tool for the directives of the branch allowing the generation of strategies to work on the reduction of this phenomenon. This document presents the research that was carried out on the problem, the process to obtain an adequate knowledge base for the project, the activities of preprocessing of the information to obtain the dataset, the selection of two data mining models and the subsequent implementation of them to finally obtain the results that helped the completion of the prototype and therefore the project.

Keywords: Prototype, dataset, university dropout, data mining.

Capítulo 1

Introducción

Actualmente la Universidad del Valle cuenta con el sistema de registro académico y admisiones en el cual se registran los datos personales, socioeconómicos y académicos de sus estudiantes, toda esta información se mantiene incluso cuando el estudiante pasa a ser egresado o retirado de la universidad. Además de conservar la información, este tipo de sistemas permiten la generación de reportes o informes para los usuarios, sin embargo esta información es de tipo académico y general acerca del estudiante y no ayudan a comprender ciertas situaciones en las que algunos estudiantes pueden incurrir, como es el caso de la deserción universitaria.

Dado lo anterior, se buscó la obtención de esta información para aplicar dos técnicas de minería de datos sobre todos estos antecedentes que se tienen acerca de los estudiantes que han desertado, con el fin de obtener un clasificador binario que permita conocer qué estudiantes están en riesgo de abandonar sus estudios de acuerdo a los resultados que arroje este, teniendo en cuenta diferentes variables correspondientes a la información personal y académica de los estudiantes.

En el transcurso de este documento se encontrarán las investigaciones realizadas que cimientan el desarrollo del proyecto, los objetivos planteados, los antecedentes relacionados con esta problemática, definiciones de conceptos del trabajo, las actividades que permitieron avanzar en trabajo de grado, la seleccion de las tecnicas de mineria de datos y su configuracion, los resultados y finalmente las conclusiones que se obtuvieron una vez finalizado el desarrollo del proyecto.

1.1. Descripción del problema

En la Universidad del Valle, en los últimos diez años, se ha presentado una tasa de deserción promedio del 42 % en los estudiantes que han ingresado a los diferentes programas de formación en pregrado. Según el autor del artículo para el observador regional del Cidse[1], Jaime H. Escobar, son datos preocupantes, ya que la mayoría de estos estudiantes pertenecen a los primeros semestres, los cuales toman la decisión de desertar por diferentes motivos. Las razones por las cuales los estudiantes desertan son variadas, porque dependen de las características de cada programa de formación, dónde se requieren unas competencias y unas habilidades específicas.

Como se dijo anteriormente, la mayor parte de los desertores se encuentra en los primeros cuatro semestres y en algunos casos hasta quinto semestre, una de las razones identificadas se debe a que encuentran poco atractiva la carrera elegida, además de motivos personales, familiares, entre otros[2]. En la medida en que avanzan los semestres existe una mayor exigencia en los programa curriculares demandando del estudiante mayor dedicación.

En la mayoría de los casos, la deserción deja bastantes efectos negativos en la persona, la institución y la sociedad, como se menciona en el estudio sobre la deserción universitaria expuesto en el proyecto RAMON[3] del autor Ramón A. Benitez, debido a que pierde seguridad en sí mismo, hay una sensación de fracaso y en cierto sentido hay una interrupción en su proyecto de vida. también dependiendo en qué punto de la carrera desertó, pierde tiempo valioso para su formación como profesional, pierde oportunidades laborales, entre otras cosas que afectan a la persona.

1.2. Formulación del problema

De acuerdo al problema descrito, se puede decir con seguridad que la deserción es un problema que afecta a la Universidad del Valle en su totalidad, pero en este caso se enfocará a la sede Tuluá, por lo cual se plantea el siguiente interrogante.

¿Cómo predecir los casos de deserción en la sede Tuluá de la Universidad del Valle?

1.3. Objetivos

1.3.1. Objetivo general

Implementar un clasificador binario sobre la deserción estudiantil enfocado a la sede Tuluá de la subregión centro de la Universidad del Valle.

1.3.2. Objetivos específicos

1. Diseñar una base de conocimiento para el estudio de los casos de deserción en la sede Tuluá de la subregión centro de la Universidad del Valle.
2. Construir el dataset con todos los datos necesarios para realizar la respectiva minería de datos.
3. Implementar dos métodos de minería de datos a la base de conocimientos diseñada.
4. Comparar los resultados obtenidos de la aplicación de las técnicas de minería de datos a través de la matriz de confusión y la curva ROC para escoger la mejor técnica en función de precisión y confiabilidad.

1.4. Resultados esperados

Tabla 1.1: Relación entre los objetivos específicos y los resultados esperados

Objetivo específico	**Resultado**
Diseñar una base de conocimiento para el estudio de los casos de deserción en la sede Tuluá de la Universidad del Valle.	Una base de datos con la información correspondiente a los estudiantes que incurrieron en el problema de la deserción en una ventana de observación de diez años en la sede Tuluá de la Universidad del Valle.
Construir el dataset con todos los datos necesarios para realizar la respectiva minería de datos.	Un conjunto de datos procesados, limpios, con calidad, listos para la implementación de una herramienta de minería de datos.
Implementar dos métodos de minería para el procesamiento del dataset construido.	Obtención de resultados que incluyen, la codificación de las técnicas de minería de datos, patrones que reflejan las relaciones entre las dimensiones del dataset, los resultados de la aplicación de las técnicas de minería de datos transformados en información que permita comparar las herramientas de minería.
Comparar los resultados obtenidos de la aplicación de las técnicas de minería de datos a través de la matriz de confusión y la curva ROC para escoger la mejor técnica en función de precisión y confiabilidad.	La selección de la técnica a implementar a partir del análisis de la matriz de confusión y la curva ROC.

1.5. Justificación

De acuerdo a los distintos artículos y trabajos realizados relacionados con el problema de la deserción en las universidades de diferentes partes del país [4], se asumen varias teorías con respecto al estudiante desertor, donde dadas ciertas características (edad, sexo, condiciones socioeconómicas, condiciones demográficas y aspectos académicos), un estudiante es más propenso a desertar, por lo cual se buscará hallar patrones que ayuden a la prevención de este problema teniendo en cuenta todas estas características encontradas en las distintas investigaciones, tratando de generar una dimensionalidad que a partir de sus relaciones permita facilitar la identificación de patrones entre los estudiantes que ya incurrieron en la problemática.

Aunque el objetivo de este trabajo de grado no fue el de disminuir los índices de deserción en la sede Tuluá, se considera que puede ser una herramienta que sirva de apoyo a los directores de programas y las sedes, para tratar de ayudar de alguna manera al estudiante que sea indicado por el clasificador binario y no dejarlo a la deriva como ha pasado en casos, donde simplemente se dan cuenta de la deserción al momento en que ocurre y el estudiante abandona sus estudios o cambia de carrera. Con el clasificador binario propuesto en este trabajo, los directores de programa tendrán una herramienta de apoyo que les permita saber que hay estudiantes que presentan patrones ya observados, que señala que aquellos estudiantes pueden incurrir en deserción, por lo cual es necesario observar de manera detenida estos casos y ayudarlos para que no alimenten el porcentaje de deserción ya presentado.

Para llegar a la meta propuesta, que consiste en diseñar un clasificador binario que ayude a identificar posibles casos de deserción en la sede Tuluá de la Universidad del Valle, usamos el modelo CRISP, debido a que es una metodología estandarizada a nivel internacional y que ha mostrado resultados eficientes en proyectos de minería de datos, además que se considera que es la opción más óptima para el tratamiento de los datos y la posterior implementación de las técnicas de minería de datos, para aportar con este trabajo una herramienta que quizá sirva para disminuir los porcentajes de deserción en los distintos programas pertenecientes a la sede mencionada.

De acuerdo a la cantidad estimada de datos que fueron procesados, de los que se tiene una ventana de observación de ocho programas académicos que se ofertan en la sede Tuluá, además esta información proviene de los períodos académicos cursados en los últimos diez años y se analizaron mediante dos técnicas de minería de datos.

1.6. Alcance

El proyecto presenta la implementación de una técnica de minería de datos aplicada al problema de la deserción en la sede Tuluá de la Universidad del Valle, esta herramienta mediante ciertas caracteristicas permite determinar si un estudiante es desertor o no, esto con el fin de ir en pro de la prevención de la problemática en el dominio seleccionado. Se generaron patrones, que hacen referencia a la relación que tienen las características de los individuos en la base de conocimiento, donde se tendrán datos de los estudiantes de la sede Tuluá en una ventana de observación de diez años, incluyendo los estudiantes que se graduaron, los que están cursando alguna carrera y los que desertaron, todo esto en un plazo de 8 meses donde se recolectó dicha información y se hizo el respectivo estudio de los datos y posteriormente se llegó a los resultados por medio de las tareas propuestas. A partir de los datos obtenidos se generó un prototipo que puede servir como indicador a las directivas de la universidad del Valle, que les permitan

crear estrategias de apoyo estudiantil y prevención a posibles casos de deserción, cabe aclarar que para manejar los datos que fueron proporcionados por el profesor Jaime H. Escobar y su grupo de trabajo, se tuvo en cuenta la ley de protección de datos 1581 del 2012. [5].

1.7. Metodología

De acuerdo con las características del trabajo que se pretende realizar se llega a la conclusión de que la metodología que se va utilizar es la descriptiva, debido a que en el presente trabajo se busca que en sus tareas, la identificación de características en común, que presentan los afectados por el problema de la deserción nos permitan establecer patrones que nos ayuden a formular un clasificador binario sobre deserción en la sede Tuluá de la universidad del Valle basándonos en los hechos de nuestro problema, tratando como principal fuente de información los casos de deserción que se han presentado en las sedes en los últimos diez años. Estos hechos concretos permitirán generar información acerca de la problemática, que sirva como indicador a las directivas para generar estrategias de apoyo a los estudiantes involucrados en los resultados del modelo generado.

La información utilizada en el trabajo fue solicitada a la universidad, con el fin de obtener datos óptimos, claros y fidedignos acerca de los casos de deserción presentados en los últimos años, dicha información fue suministrada por el profesor Jaime H. Escobar y su grupo de trabajo. Esta base de datos permitió realizar las tareas que correspondan a la minería de datos aplicada en este problema para generar resultados cercanos a la realidad, que permita predecir futuros casos de deserción en la sede.

Para realizar y llevar a cabo todo lo anterior se siguió la metodología CRISP la cual es ideal debido a que es de ayuda para planear el proyecto, ejecutarlo, además reducir costos, este modelo cuenta con seis etapas las cuales consisten en:

1. **Analizar el problema:** Para lo cual se investigarán y analizarán diversos artículos relacionados con la problemática de la deserción, que permitan tener claridad de la problemática y como se puede abordar a partir de la minería de datos.

2. **Comprensión de los datos:** Etapa en la que se adquieren los datos y se determina su dimensionalidad.

3. **Transformación los datos:** Etapa de transformación de datos para filtrar datos que sean perjudiciales para efectos del desarrollo del trabajo.

4. **Selección de una técnica de modelación:** Etapa para la cual se tendrán en cuenta las características de las técnicas de minería y del dataset, para seleccionar las herramientas de mineria mas apropiadas para el trabajo.

5. **Evaluación de los resultados.**

6. **Despliegue de los resultados:** Etapa en la que se generan estrategias para el monitoreo, control y mantenimiento del modelo seleccionado, para el posterior reporte en el que se plasmarán las conclusiones del proyecto de minería aplicado al proyecto de la deserción.

Capítulo 2

Marco referencial

2.1. Marco teórico

Para el desarrollo de esta propuesta se tuvieron como referentes investigaciones respecto a la deserción estudiantil, cuya base conceptual provenía de la teoría del suicidio de Durkheim (1897), donde se permiten realizar una analogía entre el suicidio y al deserción para la sociedad, por otra parte estaba la consideración de que los centros de educación superior eran sistemas que tenían sus propios valores y estructuras sociales (Spady 1970), por lo cual dado este escenario, es natural que con bajos niveles de integración social, aumente la probabilidad de incurrir en deserción, pero con el pasar del tiempo dichas investigaciones quedaron en estudios que trataron el tema desde una perspectiva cualitativa e individual, donde se tomaban en cuenta la integración del individuo con la sociedad estudiantil y unos pocos aspectos externos que podrían afectar, principalmente el poder adquisitivo del individuo.[4]

A partir de ese momento, las investigaciones que trataban el problema de la deserción se dividieron en dos, aquellas que profundizaban en el tema desde el aspecto teórico y las investigaciones que buscaban hallar causas de este fenómeno a partir del conocimiento del problema. Actualmente la definición de deserción no es precisa, pero se puede decir que la deserción es el abandono de los estudios, bien sea aspectos socioeconómicos, por incapacidad académica y variables institucionales. Autores como Tinto (1989) indican que el estudio del fenómeno de la deserción es muy complejo, ya que implica una variedad de perspectivas , además la gama de tipos de abandonos que están definidos, por lo cual indica que ninguna de las definiciones formalizadas arropa todas las posibles perspectivas que requiere el realizar un estudio completo del fenómeno de la deserción, por lo cual la mayoría de investigaciones están enfocadas bajo una perspectiva, aquella que es escogida por el investigador.[4]

Por otra parte desde la perspectiva institucional, todos los estudiantes que incurren en este problema son catalogados como desertores y asocian dicha deserción al aspecto de bajo rendimiento académico o retiro forzoso. Cada estudiante que por la razón que sea decide abandonar sus estudios, genera una nueva vacante en la institución, aquella vacante que ocupó este estudiante bien pudo ser ocupada por otro individuo que podría persistir en los estudios, razón por la cual dada esta situación la institución pierde recursos financieros porque no va haber tasa de retorno en mano de obra calificada, este problema en gran medida puede desestabilizar una institución a nivel financiero.[4]

Teniendo en cuenta el problema, se desea implementar un modelo predictivo por medio de la minería de datos para tratar la problemática antes mencionada. Alrededor de la minería de datos se han manejado distintas concepciones en las cuales se encuentran tres corrientes diferentes en la literatura, de acuerdo

con Peacock (1998), el significado de la minería de datos se puede ver en tres aspectos, dependiendo de su amplitud, se puede definir como el descubrimiento automático de patrones o de modelos interesantes y no obvios ocultos en una base de datos, los cuales tienen un gran potencial para aportar en los aspectos principales de un estudio o en este caso en el problema de la deserción, la palabra interesante se refiere en cóomo este proceso puede ser aplicado en estrategias para tratar la investigación que se realiza. La minería de datos comprende, como sistema de extracción de relaciones, los métodos basados en la computadora, requiriendo poco involucramiento del analista para el descubrimiento de información relevante.[6]

Por otra parte, se habla de un concepto más amplio de la minería de datos, que engloba, aparte de lo comentado, un proceso más formalizado en el que se especifican fases para realizar el proceso de los datos, la identificación de patrones, la implementación de algoritmos y finalmente las conclusiones, englobando así un conjunto de actividades dentro de las cuales se encuentra el análisis de datos.[6]

Según Thuraisingham (1993), la minería de datos es una forma de hacer una variedad de preguntas y de utilizar datos útiles, patrones y tendencias previamente desconocidas desde grandes cantidades de información almacenada en bases de datos. Una de las definiciones es que la minería de datos se percibe como "un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos" (Fallad et al., 1996). Mientras tanto, M. Berry & Linoff (1997) adoptaron la definición de minería de datos como el anásis y la exploración, a través de medios automáticos y semiautomáticos, "de grandes cantidades de datos con el fin de descubrir patrones y reglas significativos".[6]

La minería de datos apareció por primera vez en la década de 1960 y utilizaba términos como: arqueología de datos, pesca de datos, "en donde se proponía el encuentro de correlaciones sin necesidad de plantear una hipótesis previa de trabajo en una investigación"(Olmos Pineda, 2010). Varias aplicaciones de minería de datos incluyen: evaluación de métodos de compra según características principales; conocer el grado de interés sobre tipos de productos, entre otros.[6]

Se considera que a partir de estas bases teóricas, conociendo en parte la problemática, la minería de datos puede ser una herramienta útil que proporcione conocimiento a las directivas de la universidad para generar estrategias que permitan la minimización de los impactos negativos de la deserción en la universidad, principalmente al dominio seleccionado.[6]

2.1.1. Técnicas de minería de datos

A continuación se presentará la caracterización de cada una de las posibles técnicas en minería de datos que se tuvieron en cuenta para la realización de este trabajo de grado. Sin entrar en detalles formales, se definirán las ideas principales detrás de cada metodología junto a algunas de sus características que permitan entender su funcionamiento.

Redes neuronales artificiales(ANN)

El propósito de una red neuronal artificial, es crear un modelo que sea capaz de reproducir el método de aprendizaje del cerebro humano. Las células involucradas en este aprendizaje son neuronas interconectadas entres sí a través de redes complejas. Las redes neuronales están formadas por neuronas las cuales son el núcleo del modelo[7]. El proceso de aprendizaje es el siguiente: cada neurona recibe un conjunto de entradas, las cuales llevarán un pero emitiendo una salida. Dicha salida viene dada por tres funciones:

- La función de propagación
- La función de activación
- La función de transferencia

Dados ciertos parámetros existe una forma de combinarlos para predecir un resultado en particular. Las redes neuronales son un modelo para encontrar dicha combinación de parámetros y aplicarla al mismo tiempo. El objetivo de esto es encontrar la combinación que mejor se ajusta al modelo para así entrenar a la red neuronal. Este entrenamiento y aprendizaje es la parte crucial de la ANN, ya que nos marcará la precisión del algoritmo[7].

Árboles de decisión

Es un modelo de predicción en varios campos pasando por la IA hasta la economía. Con la obtención de un grupo de datos se elaboran diagramas de construcciones lógicas, semejantes a los sistemas de predicción basado en reglas, que son útiles para representar y clasificar una sucesión de condiciones que se producen de forma seguida, para la resolución de un problema.[8].
Las características de los árboles de decisión son:

- Planteamiento de la problemática desde diferentes perspectivas de acción.
- Posibilita el análisis de manera completa contemplando todas las posibles soluciones.
- Suministra un esquema para calcular el costo de los resultados y probabilidad de uso.
- Ayuda en la escogencia de la mejor decisión basándose en la información disponible y en las mejores suposiciones.
- Su estructura permite el análisis de las alternativas, los eventos, las probabilidades y los resultados.

Algoritmo de agrupamiento o clustering

Es un método de agrupación de una secuencia de vectores acordes con algunos criterios. Esos criterios a menudo son la distancia o semejanza. La proximidad es definida de acuerdo a una función de distancia, siendo una de ellas la euclídea, aunque hay funciones más fuertes que permiten extenderla a variables discretas. La medida con más uso para la similitud entre los casos en la matriz de correlación entre los nxn casas. Ahora bien, existen diversos algoritmos basados en la maximización de la verosimilitud.[9].
Por lo regular, los vectores pertenecientes a la misma agrupación (o clusters) comparten cualidades comunes. El conocimiento de las agrupaciones puede posibilitar una descripción sintética de un grupo de datos multidimensional complejo. Por eso su implementación en minería de datos. La descripción sintética es resultado de sustituir la descripción de la totalidad de los elementos de un grupo por la de un representante característico.

Clasificador bayesiano

Según McCallum, Andrew (2019) "Es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de naive, es decir, ingenuo. En términos simples, un clasificador de Naive Bayes asume que la presencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra característica, dada la clase variable. Por ejemplo, una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro. Un clasificador de Naive Bayes considera que cada una de estas características contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características"[10].

Reglas de asociación

Las reglas de asociación son utilizadas para especificar acontecimientos que ocurren por lo general dentro de un determinado conjunto de datos. Se ha investigado acerca de diferentes métodos para aprendizaje de reglas de asociación que resultan ser interesantes para hallar correlaciones entre variables en bases de datos grandes[11].
Ejemplos:

- Alumnos que cursan IA propenden a cursar taller de Sistemas Multiagentes.
- Clientes que compran un producto lácteo propenden a adquirir productos panificados.

Máquinas de vectores de soporte

Dado un conjunto de puntos, que es subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a dos posibles grupos, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría es desconocida) pertenece a un grupo o al otro[12].
Las Máquinas de Vectores de Soporte (Support Vector Machines) permiten encontrar la forma óptima de clasificar entre varias clases. Esta clasificación se realiza maximizando el margen de separación entre las clases. Los vectores encargados de definir el borde de esta separación son los vectores de soporte. En el caso de que las clases no sean linealmente separables, se puede usar el truco del kernel, esto con el fin de añadir una nueva dimensión donde sí lo sean[13].

Las SVM se pueden utilizar para resolver varios problemas del mundo real:

- Las SVM son útiles en la categorización de texto e hipertexto , ya que su aplicación puede reducir significativamente la necesidad de instancias de entrenamiento etiquetadas tanto en la configuración estándar inductiva como transductiva . Algunos métodos para el análisis semántico superficial se basan en máquinas de vectores de soporte
- La clasificación de imágenes también se puede realizar utilizando SVM. Los resultados experimentales muestran que las SVM logran una precisión de búsqueda significativamente mayor que los esquemas de refinamiento de consultas tradicionales después de sólo tres o cuatro rondas de comentarios de relevancia.
- El algoritmo SVM se ha aplicado ampliamente en las ciencias biológicas y otras. Se han utilizado para clasificar proteínas con hasta un 90 % de los compuestos clasificados correctamente. Se han sugerido pruebas de permutación basadas en pesos de SVM como un mecanismo para la interpretación de modelos de SVM.

Máquina de aprendizaje basada en reglas

Es un término en ciencias de la computación destinado a abarcar los diferentes métodos de machine learning que identifiquen, aprendan o desarrollen "reglas" destinadas al almacenamiento, manipulación o aplicación. La característica principal de una máquina de aprendizaje automático basado en reglas es el identificar y utilizar un conjunto de reglas relacionales para representar el conocimiento obtenido por el sistema. Esto se diferencia de otras máquinas de aprendizaje que generalmente identifican patrones singulares que se aplican universalmente a cualquier caso para realizar una predicción.
Los modelos de aprendizaje basado en reglas comprenden los sistemas de clasificación de aprendizaje, aprendizaje de reglas de asociación, sistemas inmunitarios artificiales, y cualquier otro método que se base en un conjunto de reglas, cada una de las cuales cubre el conocimiento contextual.[14].

Ventajas:

- Representa de forma sencilla el conocimiento específico de los expertos, generalmente, los expertos humanos exponen el método para resolver problemas mediante expresiones condicionales tipo "si estamos en esta situación, entonces se hace esto...", ajustándose de manera fiel al método que se sigue en este modelo.
- Estructura homogénea, las reglas generadas poseen el mismo tipo de estructura "si... entonces...". Cada regla es una parte individual que no guarda relación de dependencia entre ellas.
- Separación entre la base de conocimiento y su procesamiento.
- Posibilidad de trabajo con información incompleta e incertidumbre.

Desventajas:

- **Opacidad de las reglas:** Independiente de la simplicidad de las reglas, la interacción producida entre la red de reglas puede ser bastante opaca, generando dificultades en el momento de analizar el papel que cumple cada regla en el razonamiento global de la resolución del problema.
- **Búsqueda ineficiente:** El problema principal con las máquinas de aprendizaje basado en reglas, es que al realizar búsquedas exhaustivas en todo el conjunto de reglas por ciclo de iteración, puede ser un proceso lento y en muchos casos inviable en problemas del mundo real.
- **Incapacidad de aprendizaje:** Los sistemas de reglas sin adiciones tienen poca capacidad de aprendizaje basándose en la experiencia, por lo que generar nuevos conocimientos no provee procedimientos para aprender más cosas de manera ágil.

Aprendizaje de reglas de asociación

Es un método de aprendizaje automático basado en reglas utilizado para encontrar relaciones interesantes entre variables en grandes bases de datos. Su objetivo es identificar reglas sólidas descubiertas en bases de datos con la ayuda de algunas medidas de interés.

Basado en el concepto de reglas estrictas, Rakesh Agrawal, Tomasz Imieliński y Arun Swami introdujeron reglas de asociación para descubrir regularidades entre productos en datos de transacciones a gran escala registrados por sistemas de punto de venta (POS) en los supermercados. Por ejemplo, la regla (cebollas, patatas) implica (hamburguesa) que se encuentra en los datos de ventas de un supermercado indicaría que si un cliente compra cebollas y patatas juntas, es probable que también compre carne de

hamburguesa. Dicha información se puede utilizar como base para decisiones sobre actividades de marketing como, por ejemplo, precios promocionales o ubicaciones de productos[15].

Además del ejemplo anterior del análisis de la canasta de mercado, las reglas de asociación se emplean hoy en muchas áreas de aplicación, incluida la minería del uso de la Web, la detección de intrusos, la producción continua y la bioinformática. A diferencia de la minería de secuencias, el aprendizaje de reglas de asociación generalmente no considera el orden de los elementos dentro de una transacción o entre transacciones[15].

2.2. Marco de antecedentes

En el artículo referenciado[16] aborda el tema de la deserción y retención de estudiantes en la educación superior en Chile desde una perspectiva conceptual, se analizan los diferentes enfoques para el análisis de la problemática los cuales son psicológicos, económicos, sociológicos, organizacionales o aspectos de las interacciones entre el estudiante y la institución, también se toman en cuenta variables pertenecientes a diferentes ámbitos las cuales pueden ser susceptibles a cambios externos, todos estos factores resultan predictores del abandono y la persistencia estudiantil.

La idea de la investigación es explorar estos lineamientos en el país (Chile), con el fin de encontrar cómo se combinan las distintas variables en los diferentes tipos y modalidades institucionales, de manera que éstas puedan actuar sobre los factores que son más controlables por las propias instituciones y así reducir el costo económico y social que conlleva el problema de la deserción.

En el documento referenciado[17] se aborda la problemática de la deserción en la universidad Simón Bolívar, se aplican técnicas de minería de datos que permiten identificar características para generar un modelo predictivo, que permiten tratar el problema, las herramientas escogidas fueron árboles de decisión C4.5 y el ID3, debido a que son herramientas eficientes para efectos de modelos predictivos, pero finalmente se decantaron por los árboles de decisión C4.5 o J4.8. Luego de realizar todas las tareas que requiere el proceso de minería de datos, se construyó un árbol con 201 instancias que hace referencia al conjunto de entrenamiento, arrojando probabilidades de deserción y clasificación de individuos en desertores o no, la precisión del modelo fue de 0.75 en una de sus ramas y 1 en la otra generando tablas con la clasificación de los estudiantes que podrían desertar o no.

En el artículo referenciado[18] se expone la construcción de un modelo predictivo de deserción estudiantil para determinar la probabilidad de abandono de sus estudios por parte de los estudiantes, con ayuda de técnicas de clasificación basadas en árboles de decisión. La metodología usada para el desarrollo de este fue KDD (Knowledge Discovery in Database), la cual cuenta con cinco etapas: selección, procesamiento, transformación, minería de datos y evaluación. Se aplicó el árbol de decisión CART (Classification and Regression Tree) de la herramienta R, se fabricó un árbol con una profundidad de cuatro, y finalmente con esto se llegó a concluir que las variables de nivel y notas del estudiante son las que tienen mayor influencia en la deserción.

En el documento referenciado[19] se analiza el fenómeno de la deserción para el segundo semestre del año 2007 en el programa académico de Ingeniería de Sistemas en las sedes Tuluá y Cali a partir de la herramienta SPADIES y según los informes del profesor Escobar (2006) “se categoriza dicha carrera como una de las unidades académicas con mayor proporción de deserción al interior de la Universidad (50 %).

El estudio presenta las características sociodemográficas de cada muestra de la población para ambas sedes y los perfiles del estudiante desertor, después se dan a conocer los resultados de la aplicación de la encuesta telefónica realizada a algunos de los estudiantes desertores contactados."

El artículo referenciado[20] cuenta la situación de un estudiante en particular que ingresó a la Universidad del Valle en Cali, nos describe la posición suya y de su familia para dar un perfil del estudiante promedio que podría incurrir en deserción, las frustraciones que puede sufrir el estudiante al verse inmerso en un nuevo "mundo", debido a que es completamente diferente al colegio, el ritmo de trabajo difiere y las exigencias son más superiores. Posteriormente nos indica, que parte de los estudiantes que alimentan los porcentajes de deserción en el país y más específicamente en la Universidad del Valle, tratan de entrar a carreras que no van con su vocación, pero que finalmente cursan porque al haber obtenido un buen puntaje en el ICFES le permite concursar por el cupo en dicho pregrado, pero en ocasiones ingresan y no tienen el conocimiento sobre lo que les espera, por otra parte la universidad no sabe si ciertos estudiantes cumplen con el perfil que apriori demandaría el realizar ciertos pregrados. Finalmente se concluye que para reducir los costos que implica sostener una tasa de deserción por encima del 40 %, se deben replantear los criterios de admisión, contribuyendo a sintonizarse con una universidad moderna y con proyección debido a tener cursando estudiantes en carreras que van con su orientación vocacional.

En el documento referenciado[21] se narra como con el pasar de los años se ha presentado el problema de la deserción en la Universidad del Valle, haciendo hincapié en que las facultades con mayor tasa de deserción son la de ingeniería y ciencias, además concluyendo que la mayoría de estudiantes que por alguna razón pausa sus estudios es más propenso a incurrir en deserción que aquel que no interrumpe sus estudios, razón por la cual es importante para la permanencia y posterior graduación en la universidad, no interrumpir la carrera en algún momento. Posterior a ello se habla del abordaje conceptual de la deserción desde un punto general, citando teorías de pioneros como Spady (1970), Tinto (1982) y Bean (1982), donde todos generaron modelos a partir de investigaciones para analizar el problema de la deserción.

Los autores de los artículos mencionados anteriormente hablan sobre la consecución de la información para realizar el estudio, la asociación de las variables pertenecientes a los estudiantes desertores con modelos matemáticos que permitieran sacar conclusiones acerca de la muestra de estudiantes. Posteriormente los datos de los estudiantes se ingresan a funciones matemáticas asociadas a la supervivencia de un estudiante, contemplando una variable por ejemplo: nivel de ingresos del hogar, ciudad de origen, género, estado civil, entre otras, cada una de estas variables es la entrada a una función de supervivencia dejando porcentajes de riesgo de acuerdo a los valores que toma la entrada generada por cada estudiante, para finalmente determinar en un modelo el riesgo total de un individuo teniendo en cuenta cada riesgo asociado a cada una de las variables.

2.3. Marco Conceptual

- **Minería de datos:** Es el análisis automático o semi-automático de grandes cantidades de datos con el fin de extraer patrones interesantes hasta ahora desconocidos, como lo son los grupos de registros de datos (análisis clúster), registros no convencionales (la detección de anomalías) y dependencias (minería por reglas de asociación).

- **Deserción universitaria:** Es aquella situación a la que se enfrenta un estudiante cuando aspira y no logra concluir su proyecto educativo, muchas veces el abandono de sus estudios es provocado por cuestiones socioeconómicas, el ambiente educativo, el ambiente familiar, lugar de residencia, entre otros[22]. Según Ramón A Benitez (2012) "muchos expertos en la materia no dudan en señalar a las graves falencias de la enseñanza media como la causante principal de esta debacle pues sus egresados, apenas armados con conocimientos básicos, al toparse con las nuevas exigencias de la universidad quedan desmoralizados. Otros, en cambio, agregan que la ansiedad propia de las nuevas generaciones las vuelve muy vulnerables a los tropiezos y las frustraciones, por lo que los traspiés normales de la inserción en el ámbito académico los lleva a tomar decisiones apresuradas y, sobretodo, irreversibles. También las presiones económicas cumplen un papel determinante al momento de tomar el difícil camino de la deserción. La realidad indica que muchos estudiantes trabajan en paralelo a los estudios y que esta doble vida a la larga puede resultar fatal para sus aspiraciones profesionales"[3].
  Por otro lado, muchos especialistas también asumen que parte de las causas del problema son propias del sistema universitario. Los cursos del primer año suelen ser multitudinarios, los recursos materiales siempre son escasos, la difícil situación que atraviesan los estudiantes en su ingreso a la enseñanza superior muchas veces no está contemplada en las modalidades pedagógicas, los diseños curriculares aparecen demasiado rígidos frente a las nuevas necesidades de los ingresantes, etc. En concreto, la poca flexibilidad de las estructuras universitarias también influye directamente en la construcción de la deserción de los estudios[23]. A partir de esta definición se observa que existen dos tipos de abandono: deserción con respecto al tiempo y al espacio.
  La deserción con respecto al tiempo se clasifica a su vez en:

  Deserción temprana: es cuando el estudiante abandona sus estudios en los primeros semestres del programa.

  Deserción tardía: es aquella donde el estudiante abandona sus estudios en los últimos semestres.

  La deserción con respecto al espacio, por su parte, se divide en:

  Deserción institucional: caso en el cual el estudiante abandona la institución.

  Deserción interna o del programa académico: se refiere al estudiante que decide cambiarse a otro programa que ofrece la misma institución de educación superior.

- **Modelo CRISP:** Es una guía de referencia utilizada para el desarrollo de proyectos de Data Mining, este modelo posee seis fases[24]:

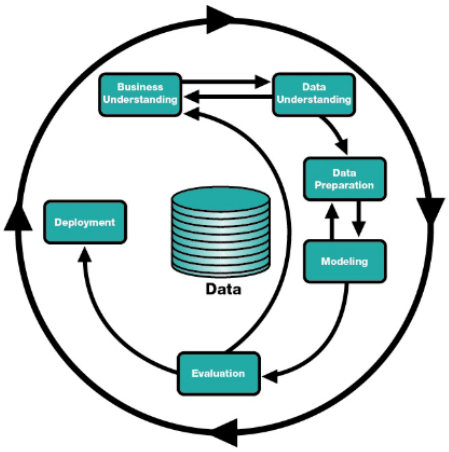


Figura 2.1: Modelo CRISP.
Fuente: Modelado y gestión de la información

El proceso que empieza con la fase de comprensión del negocio, considerada una de las más importantes, sino la más importantes, debido a que en esta fase se plantean los objetivos y requisitos que requiere el proyecto desde una perspectiva empresarial, dejando claro cuál va ser el problema que se desea resolver al final de todas las fases del modelo y a que dominio va arropar la solución que se plantea, y se determinan cuales van a ser los criterios de aceptación del proyecto.

A continuación está la fase de comprensión de datos, fase en la cual se recolectan los datos iniciales y se hace el primer contacto con el problema, identificando su calidad y estableciendo las relaciones más visibles entre los datos para generar las primeras hipótesis respecto a la problemática, esta y las siguientes dos fases son las que requieren más esfuerzos y cuidado en los procesos de data mining.

La fase de preparación de los datos, está relacionada con la fase de modelado, debido a que está en función de la técnica de modelado escogida, los datos requieren ser procesados de diferentes formas, seleccionando así un nuevo conjunto de datos, donde en sus características deben estar, ser limpios y con calidad, en este nuevo conjunto se pueden generar nuevos campos a partir de los datos anteriores.

La fase de modelado, busca que se escoja la herramienta de minería de datos más adecuada de acuerdo a las características del dataset que describe el problema, que la herramienta pueda cumplir los requisitos propuestos en la fase de comprensión del negocio, que permita realizar el modelo en el tiempo esperado y que haya un conocimiento acerca de ella, posterior a esto una vez implementada la técnica de minería de datos, se deben generar planes de pruebas para construir el modelo y finalmente evaluar la información que arroja este.

En la fase de evaluación se tienen en cuenta todos los objetivos fijados en la fase de comprensión del negocio, los requisitos y el criterio de aceptación o éxito, para valorar cuán exitoso fue el modelo a partir de los resultados obtenido después de la generación del modelo, para pensar en la siguiente fase.

La fase implementación busca la transformación del conocimiento obtenido en acciones dentro del proceso de negocio, además de la generación de estrategias que permitan el monitoreo y el mantenimiento del modelo, finalmente se sacan conclusiones sobre el proceso que llevó a cabo la implementación del modelo.

- **Proyecto RAMON:** El Proyecto de Refuerzo Académico Masivo Ordenado Normalizado (RAMON)[3] pretende unirse a la lucha contra todos los efectos que llevan a un estudiante a desertar bajo una premisa simple y efectiva: ayudar a los estudiantes recién ingresados a la universidad para que refuercen su desempeño académico y, al mismo tiempo, facilitar su adaptación a la vida universitaria. En este sentido, propone su realización basándose en un concepto distinto, innovador y completamente original: la participación activa de los propios estudiantes universitarios en la solución del problema. Concretamente, el Proyecto RAMON postula una respuesta creada por estudiantes, ejecutada por estudiantes para el beneficio de los estudiantes.

  El Proyecto RAMON sostiene que "la mejor manera de ayudar a los estudiantes en su adaptación a la vida universitaria se logra mediante su unificación en pos de solucionar las dificultades y circunstancias comunes que los afectan. Para ello ofrece una estructura organizativa sólida y equilibrada que, junto con las aplicaciones administrativas adecuadas, prometen simplificar la enorme tarea y amplificar los beneficios potenciales".(Benitez, 2012)

- **Precisión:** Es la razón donde está el número de verdaderos positivos y el número de falsos positivos. La precisión es intuitivamente la capacidad del clasificador de no etiquetar como positiva una muestra que es negativa. El mejor valor es 1 y el peor valor es 0.

- **Matriz de Confusión:** Es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada fila representa a las instancias en la clase real, mientras que cada columna de la matriz representa el número de predicciones de cada clase.

- **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:** En la clasificación de etiquetas múltiples, esta función calcula la precisión del subconjunto: el conjunto de etiquetas predichas para una muestra debe coincidir exactamente con el conjunto de etiquetas correspondiente en y_true.

- **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 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 según se varía el umbral de discriminación (valor a partir del cual decidimos que un caso es un positivo).

2.4. Marco legal

Este trabajo de grado está en el ámbito de las bases de datos, que contiene información perteneciente a personas que hicieron o hacen parte de la sede Tuluá de la Universidad del Valle, por lo cual se trataron bajo los conceptos de datos públicos y datos sensibles; razón por la que se hace mención a la ley 1581 del 2012 de protección de datos[5] que se tuvo en cuenta para el tratamiento de la información sin infringir las normas pertenecientes a la nación y sin vulnerar los derechos de las personas que tienen sus datos en las bases de la universidad, en este sentido, los datos proporcionados fueron entregados de forma anonimizada.

Capítulo 3

Desarrollo del proyecto

3.1. Caracterización y selección de los datos

De acuerdo al contexto del proyecto y la problemática que se abordó se consideraron varias características de los estudiantes para realizar un análisis de deserción, estas características pertenecen a varios de los siguientes aspectos:

1. **Aspecto institucional:** Hace referencia a hechos que son ajenos al estudiante pero que lo afecta en muchas situaciones, hasta el punto de hacerlo desertar esto se asocia a la oferta académica de la universidad y que tan afín es con su orientación vocacional, en los antecedentes esta información es muy relevante para los análisis de deserción. De este aspecto se extrajeron las siguientes características:

   - Apoyo económico, registra si un estudiante recibe apoyo por parte de la universidad.
   - Sede a la que pertenece
   - Condición de excepción
   - Programa académico
   - Jornada estudiantil

2. **Aspecto socioeconómico:** Este aspecto es importante en la vida de un estudiante que está realizando su proceso universitario, ayuda en gran medida el éxito estudiantil debido a la posibilidad de tener los recursos económicos y materiales necesarios para no pasar dificultades y dedicarse a los estudios permitiendo cumplir las metas y tareas requeridas. En los análisis varias características de este aspecto sumadas a otras hacen parte de de patrones que definen cuando un estudiante puede desertar. Se extrajeron las siguientes características:

   - Estrato socioeconómico
   - Composición familiar
   - Ciudad donde reside
   - Tipo de vivienda (propia o arriendo)
   - Tipo de colegio (público o privado)
   - Tipo de zona (rural o urbana)

3. **Aspecto académico:** Este aspecto también es importante porque va estrechamente relacionado con el desempeño académico que tenga el estudiante, es una de las medidas principales para determinar la posibilidad de deserción de un estudiante, porque tener un mal rendimiento puede generar deserción voluntaria por no cumplir metas o expectativas esperadas o directamente puede generar que la universidad lo saque, alimentando el porcentaje de deserción. Se extrajeron las siguientes características:

    - Cantidad de créditos aprobados por semestre
    - Cantidad de créditos reprobados por semestre
    - Nota de admisión
    - Ubicación semestral
    - Situación de bajo rendimiento académico
    - Estado del estudiante

Fuera de estos datos se definió que era necesario conocer la edad de los estudiantes al momento de su admisión en la universidad, el sexo y su estado civil.

3.1.1. Caracterización de los datos

La caracterización de los datos se consideró como una parte muy importante del proyecto, debido a que permite entender los registros de los atributos escogidos, identificar el tipo de dato al que pertenecen, los posibles valores que estos atributos pueden tomar y la información que estos representan en la base de conocimiento construida.

A continuación se presenta la tabla 3.1 con los atributos seleccionados y generados para el entrenamiento y evaluación de los algoritmos escogidos que permitieron cumplir con todas las tareas del proyecto.

Tabla 3.1: Atributos BD de procesamiento

Atributo	**Descripción**
Edad de ingreso	Este atributo representa la edad que tenía el estudiante cuando ingresó a la universidad, el valor que va tomar es numérico.
Sexo	Este atributo representa el sexo del estudiante, los valores que puede tomar son numericos (1, 2), donde el numero 1 representa el sexo masculino y el 2 el sexo femenino.
Situación de BRA	Este atributo representa la situación de bajo rendimiento académico en la que se encuentra el estudiante, tomando como representación valores numéricos del 0 al 3 donde 0 se refiere a que no tiene BRA y 3 que tiene 3 BRA.
Año de admisión	Este atributo en sus registros posee el año en el que el estudiante ingresó a la universidad

Atributo	Descripción
Periodo de admisión	Este atributo posee el periodo del año en el que el estudiante ingresó a la universidad, se representa de manera numérica (1, 2), donde el 1 se refiere al periodo Enero-Junio y el 2 se refiere al periodo Agosto-Diciembre.
Código de programa	Este atributo representa el código del programa que el estudiante matriculó y toma valores numéricos (2710,2711,2712,3249,3743,3753,3841,3845).
Jornada	Este atributo posee valor numérico (1, 2) representando con ese número la jornada en la que un estudiante se matriculó, donde 1 es la jornada diurna y el 2 la jornada nocturna.
Tipo de programa	Este atributo representa el tipo de programa que el estudiante matriculó, su valor es numérico (1, 2) donde el 1 es para los programas pregrado y 2 para los programas tecnológicos.
Número de periodos matriculados	Este atributo representa la cantidad de períodos académicos que el estudiante ha matriculado.
Condición de excepción	La universidad cuenta con varias condiciones de excepción, para el análisis de este atributo se tendrá en cuenta solo si el estudiante cuenta con algún tipo de condición o no, teniendo los valores 1 y 0 respectivamente.
Ciudad donde reside	En este atributo se tendrá un valor de 1 para aquellos estudiantes que sean residentes de Tuluá y en el caso de los estudiantes que sean residentes de otras ciudades ya sean cercanas o distantes de Tuluá se tendrá un valor de 2, en el caso de las que son bastante distantes se supone que el estudiante puede transportarse o se traslada a Tuluá.
Número de asignaturas matriculadas por periodo académico	Este atributo tiene en sus registros la cantidad promedio de matrículas que un estudiante matriculó por semestre académico cursado.
Promedio general de los estudiantes	Este atributo posee el promedio que el estudiante lleva en los periodos cursados.
Promedio de créditos matriculados por periodo	Este atributo representa la cantidad promedio de créditos que el estudiante matriculó en los períodos académicos cursados.
Cantidad de créditos aprobados	Es un atributo numérico que representa la cantidad de créditos que el estudiante ha aprobado en los periodos cursados.
Cantidad de créditos matriculados	Este atributo contiene la cantidad total de créditos que han sido matriculados por el estudiante en los periodos cursados.

Atributo	Descripción
Proporción entre créditos aprobados y matriculados	Este atributo contiene un valor numérico entre 0 y 1 representando la proporción de créditos aprobados respecto a los matriculados, donde 0 quiere decir que no aprobó ningún créditos de los matriculados y 1 quiere decir que los aprobó todos.
Tipo de zona	Este atributo representa de manera numérica (1, 2), si un estudiante vive en zona urbana (1) o si vive en zona rural (2), este atributo es construido a partir del barrio de su residencia.
Graduado	Atributo numérico (0, 1) que representa si un estudiante se graduó (1) o no se ha graduado (0).
Desertó	Atributo numérico (0, 1) que representa si un estudiante incurrió en deserción (1) o no (0).

3.1.2. Selección de la información

Una vez determinadas las características a tener en cuenta para realizar el análisis de deserción, se procedió a solicitar la respectiva base de datos a la universidad por medio profesor Jaime H. Escobar y su grupo de trabajo, para así hacer el respectivo tratamiento de la información y a la misma vez obtener una base de conocimiento confiable para seleccionar la información pertinente y de valor para el proyecto.

Después del proceso de solicitud y recepción de la información, se observó la base de conocimiento obtenida en la cual habian varios de los datos propuestos inicialmente pertenecientes a los aspectos mencionados, aunque sólo se encontraban algunos de los datos estimados anteriormente debido a que la base de datos cubre la información correspondiente a la admisión de los estudiantes y su progreso académico, por lo cual se obtuvieron en la base de conocimiento los siguientes atributos:

- ID, correspondiente a un identificador que ayuda a la organización de la información.
- SNP (Servicio Nacional de Pruebas), el cual es el código asignado por el icfes.
- Código de programa.
- Nombre de la sede de admisión.
- Código de la sede de admisión.
- Jornada de admisión.
- Periodo matriculado.
- Nombre del programa al que pertenece.
- Código del programa al que pertenece.
- Fecha de cancelación del periodo matriculado (si aplica).
- Promedio semestral.
- Número de BRA en el periodo matriculado.
- retiro por BRA.

- Código asignatura matriculada.
- Nombre asignatura matriculada.
- Créditos de la asignatura matriculada.
- Tipo de matrícula.
- Calificación asignatura.
- Habilitación.
- Estado asignatura.
- Año de admisión del estudiante.
- Semestre de admisión (Febrero/Junio - Agosto/Diciembre)
- Edad de admisión.
- Condición de excepción.
- Cupo condición excepción.
- Fecha de nacimiento.
- Sexo.
- Tipo programa
- Ciudad de residencia.
- Barrio
- Créditos del programa de admisión.
- Periodos del programa de admisión.
- Matrículas en el programa de ingreso.
- Matrículas en otros programas.
- Número de BRA en el programa.
- Número de BRA en otros programas.
- Total BRA.
- Total admisiones a primer semestre.
- Total admisiones a otros programas.
- Fecha de graduación.

Al analizar las características encontradas en la base de conocimiento y los registros en ella, se pudo observar mucha información redundante para cada uno de los estudiantes en la base de datos, en ese sentido se procedió a seleccionar los atributos que más información podrían proporcionar de manera directa o indirecta por medio de una transformación que permitiera la construcción del dataset final. Los atributos escogidos para la generación del dataset final fueron los siguientes:

- Código de programa.
- Periodo matriculado.
- Nombre del programa al que pertenece.
- Código del programa al que pertenece.
- Promedio semestral.
- Retiro por BRA.
- Nombre asignatura matriculada.
- Créditos de la asignatura matriculada.
- Calificación asignatura.
- Habilitación.
- Estado asignatura.
- Año de admisión del estudiante.
- Semestre de admisión (Febrero/Junio - Agosto/Diciembre)
- Edad de admisión.
- Condición de excepción.
- Sexo.
- Tipo programa.
- Ciudad de residencia.
- Barrio.
- Número de BRA en el programa.
- Fecha de graduación.

Después de seleccionar los datos apropiados para el desarrollo del proyecto, se pasó al siguiente nivel de la metodología usada el cual es el preprocesamiento de la información, de la cual se tienen más detalles en las siguientes secciones.

3.2. Preprocesamiento de la información

Una vez definida la caracterización de los datos, se inició el proceso de verificación de la calidad de los mismos, en ellos se encontraron varias novedades, como el ruido proveniente de los atributos de tipo fecha, que por defecto estaban en formato numérico y al ser transformados a formato fecha, este arrojaba fechas ilogicas, dejando la mayoría de fechas por debajo del año 1970, brindando información que no ayudaba con los objetivos del proyecto, también se encontró una cantidad moderada de datos faltantes en diversos registros, y asimismo se hallaron muchos datos duplicados, en este caso se debía a que la fuente de información manejan los datos de los estudiantes en el aspecto académico, generando tuplas por cada asignatura cursada, de esta manera se repitió el identificador del estudiante, el periodo cursado, el promedio semestral, la sede a la que pertenece, el número de BRA y muchos de los atributos que tenían un solo valor por estudiante o que cambiaban periódicamente se repitieron una y otra vez.

Después de verificar la calidad de la base de conocimiento, se procedió a realizar las tareas de preprocesamiento a los mismos, dando como primer lugar a la supresión de variables correspondientes a los atributos que se consideraron innecesarios para los objetivos del proyecto; siendo necesaria la imputación de sustitución a los registros con valores faltantes correspondientes en su mayoría a los atributos ciudad de residencia y barrio, esto teniendo en cuenta la media de valores que tomaron los registros en aquellos atributos.

Una vez finalizadas las etapas de limpieza y transformación, se observó la posibilidad de construcción de nuevos datos a partir de la base de conocimientos proporcionada, estos nuevos datos permitieron la obtención de resultados que dejaron un análisis y conclusiones importantes acordes a la importancia de la idea del proyecto de grado.

A continuación se presentan los atributos generados a partir de otros que se consideraron importantes para los objetivos del proyecto.

Tabla 3.2: Atributos generados

Atributo	**Obtención**
Número de asignaturas matriculadas por periodo académico	Este dato se obtiene mediante un análisis por semestre del estudiante que hay en la BD, en donde se describen las materias matriculadas y su respectiva nota.
Promedio general de los estudiantes	Este dato se genera a partir de la multiplicación de los atributos calificación de asignatura y los créditos por materia cursada y al final se divide por el número total de créditos cursados.
Promedio de créditos matriculados por periodo	Es un atributo que se puede generar a partir del conteo de los créditos de las asignaturas que fueron matriculadas en cada periodo.
Cantidad de créditos aprobados	Se genera a partir del campo créditos por asignatura y por estado de la materia donde se puede observar si fue aprobada o no.
Cantidad de créditos matriculados	Se genera a partir del conteo de los créditos correspondientes a las materias que han sido matriculadas en los diferentes periodos cursados por el estudiante. Son tomados los atributos id, materias cursadas y créditos asignatura.
Tipo de zona	Ya que en la BD se tienen la ciudad y los barrios de residencia del estudiante, mediante un análisis se podría obtener el tipo de zona en la que reside el estudiante.
Graduado	Con los datos que se tenían sobre un estudiante graduado se puede determinar si un estudiante se graduó o no y en la tabla colocar un valor de SI/NO o de 1/0 respectivamente.

Para cada uno de los atributos generados se tuvo en cuenta la información obtenida y su importancia en caso de ser transformado con otros, llegando a la conclusión que se podían generar datos con más valor teniendo en cuenta la importancia que pueden tener a la hora de generar patrones que permitieran identificar casos de deserción. Por lo cual se procede a explicar la razón de la transformación de cada uno de los datos.

- **Número de asignaturas matriculadas por periodo académico:** En los atributos seleccionados de la base de conocimiento inicial, estaban los atributos periodo académico y nombre de asignatura matriculada, que por sí solos en un modelo de machine learning pueden no decir mucho, razón por la cual se pensó que es importante saber la carga académica a la que un estudiante se está sometiendo obteniendo el número de asignaturas matriculadas por periodo académico cursado, este atributo se generó a partir del promedio entre el total de asignaturas matriculadas y el total de periodos matriculados.

- **Promedio general de los estudiantes:** Este atributo fue generado a través de los atributos calificación asignatura, créditos asignatura y habilitación. Y es que está claro que una de las características más recurrentes e importantes de un desertor es su promedio, debido a que muchas personas que están dentro de este grupo, tienen un mal desempeño académico, razón por la cual se

consideró que este atributo era un pilar fundamental para obtener un buen análisis.

- **Tipo de zona:** Es un atributo que se considera importante y para tener en cuenta en la generación de los patrones, ya que es interesante si el tipo de zona en la que vive el estudiante tiene alguna relación con el problema de la deserción académica. El atributo fue generado a partir de la ciudad de residencia del estudiante y el barrio en el que vive, porque en Tuluá hay barrios que no pertenecen a la zona urbana.

- **Graduado:** Este es uno de los atributos fundamentales para determinar si un estudiante es desertor o no, razón por la cual se decidió generar este atributo que además de lo que brinda por sí solo como componente de análisis, da un valor agregado el hecho que sirvió para generar el atributo desertor.

- **Deserto:** Este atributo fue generado para determinar si un estudiante ha desertado o no del programa académico, teniendo en cuenta los atributos graduado, periodo matriculado, número de BRA y retiro por BRA y a partir de la referencia [4] la cual indica que un estudiente se considera desertor cuando ha dejado de matricular dos periodos académicos seguidos.

Después de generar los atributos que ayudaron al desarrollo del proyecto, y de acuerdo a la naturaleza y el funcionamiento de los algoritmos seleccionados se realizó una transformación a los valores del dataset mediante el uso del OneHotEncoder y LabelEncoder que permitieron el rendimiento óptimo a las herramientas seleccionadas. Estos valores se representan por medio de números, los cuales reemplazan los datos iniciales y a su vez permiten identificar todos los posibles valores que cada atributo puede tomar, permitiendo mejorar el procesamiento y posterior uso de las técnicas de minería de datos seleccionadas para el desarrollo del proyecto.

La base de datos presenta desbalance de datos, pero para el caso de este proyecto se consideró no necesario aplicar herramientas de balanceo, esto debido a que la información en la base de datos son casos reales, razón por la cual el aplicar este tipo de técnicas podría ser contraproducente pudiendo generar sesgos en los resultados obtenidos en caso de aplicar lo mencionado anteriormente.

3.3. Selección de los dos algoritmos

Tabla 3.3: Características del proyecto

Características	**Modelo**					
	Redes Neuronales Artificiales (ANN)	**Árboles de decisión**	**Algoritmo de Agrupa-miento o Clustering**	**Clasificador Bayesiano**	**Máquinas de vectores de soporte (SVM)**	**Máquinas de aprendizaje basado en reglas**
Multi-dimensional	x	x	x	x	x	x
Variables categóricas	x	x		x	x	x
Predecir	x	x	x		x	x
Clasificar	x	x	x	x	x	x
Elementos clasificados etiquetados	x	x		x	x	x
Variables descriptivas	x	x	x	x	x	x
Aprendizaje supervisado		x		x	x	x
Análisis de varias características	x	x		x	x	x

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

Técnica	**Ventajas**	**Desventajas**
Redes Neuronales Artificiales (ANN)	-Las RNA tienen la habilidad de aprender mediante una etapa la cual es conocida como "etapa de aprendizaje". En esta etapa se le proporciona a la RNA los datos de entrada y a su vez se le indica cuál es el salida(respuesta) esperada. -Una RNA es auto-organizada, es decir, una RNA crea su propia representación de la información en su interior, descartando al usuario de esto. -Es bastante tolerante a fallos, debido a que esta almacena de manera redundante la información, con esto, puede seguir respondiendo de manera aceptable ante posibles daños. -Es bastante flexible. -Dependiendo de su implementación se pueden obtener respuestas en tiempo real.	-Complejidad de aprendizaje para grandes tareas, cuanto más se necesita que aprenda una red, más complicado será enseñarle. -Tiempo de aprendizaje elevado. -Elevada cantidad de datos para el entrenamiento, cuanto más flexible se requiere que sea una red neuronal, más información tendrá que enseñarle para que esta realice de forma adecuada la identificación. -La falta de reglas definitorias que ayuden a realizar una red para un problema dado.
Árboles de decisión	-Fácil de interpretar y entender. -Robusto al ruido y valores perdidos. -Preciso. -Excelente para aprender relaciones complejas, altamente no lineales. Por lo general, pueden lograr un rendimiento bastante alto. -Se puede trabajar tanto con variables cuantitativas como cualitativas.	-Es posible la duplicación dentro del mismo subárbol. -Muchas veces requiere una cantidad elevada de datos. -No se puede garantizar que el árbol generado sea el más óptimo.
Algoritmo de Agrupamiento o Clustering	-Implementación sencilla. -Bastante rápido.	-Sensible al ruido. -Algunas semillas pueden resultar en una tasa de convergencia menor. -Puede caer en mínimos locales. -El resultado puede variar dependiendo de las semillas escogidas al inicio.

Técnica	Ventajas	Desventajas
Clasificador Bayesiano	-Fácil y rápido de implementar. -No requiere demasiada memoria y se puede utilizar para el aprendizaje en línea. -Fácil de entender.	-Falla al estimar las características raras. -Sufre al tener características irrelevantes.
Máquinas de Vectores de Soporte	-La clasificación con SMV ofrece una buena precisión. -Realiza predicciones más rápidas comparado con otras técnicas (ej: clasificadores bayesianos). -No consume muchos recursos. -Funciona bien con un espacio dimensional elevado.	-No son adecuadas para grandes conjuntos de datos debido a su alto tiempo de formación. -Es sensible al tipo de núcleo utilizado. -Funciona mal con clases superpuestas.
Máquinas de aprendizaje basado en reglas	-Representan de forma natural el conocimiento explícito de los expertos. -Estructura uniforme. -Separación entre la base de conocimiento y su procesamiento. -Capacidad para trabajar con conocimiento incompleto e incertidumbre.	-Relaciones opacas entre reglas. -Estrategias de búsqueda muy ineficientes. -Incapaz de aprender.

De acuerdo a las características del problema, a los antecedentes revisados relacionados con la problemática de la deserción y las propias características de la base de datos, se consideró apropiado seleccionar las técnicas de minería de datos que se ajustaron por sus características y que permitieron la realización de todas las tareas que llevan al cumplimiento de los objetivos y realización del clasificador binario de deserción.

Árboles de decisión Se escoge la técnica árboles de decisión debido a que es una de las técnicas más usadas para este tipo de proyectos de clasificación, también porque de acuerdo a la documentación revisada, la técnica con un algoritmo adecuado permite la generación de reglas convenientes alcanzando una precisión superior al 75 % en el acierto de los casos de deserción analizados, como lo han mostrado varios resultados de trabajos e investigaciones hechas sobre el tema que trata este trabajo de grado y mencionados en la sección 2.2, tiene como ventaja la representación de las reglas que llevan a cabo la tarea de clasificar los datos de una manera clara y fácil de entender esto debido a que las reglas vienen en forma de condicional con conjunciones que hacen fácil el entendimiento de las mismas, los posibles escenarios con su resultado se representan en las ramas del árbol, además es una de las herramientas que observamos en los antecedentes los cuales tuvieron resultados óptimos en la predicción y clasificación, por eso se realizó la implementación de esta técnica teniendo en cuenta esto como dato no menor.

Máquinas de vectores de soporte (SVM) Las máquinas de vectores de soporte (SVM) fueron consideradas como una técnica útil para desarrollar este proyecto, debido a que el resultado final que se esperaba obtener era la clasificación de dos tipos de estudiantes, el estudiante desertor y el estudiante no desertor, por ende sería una clasificación binaria, en donde cada punto en el plano bidimensional sería las respectivas características que se tienen sobre un estudiante y con esto mediante un respectivo análisis decidir si es desertor o no. Se escogió esta técnica porque es una técnica bastante eficiente a la de hora

de clasificar un conjunto grande de datos, como se observó en la tesis referenciada[25], además de generar buenos resultados a la hora de realizar la respectiva clasificación. El análisis SVM intenta encontrar un hiperplano unidimensional (es decir, una línea) que separa los casos en función de sus categorías objetivo (desertor, no desertor en este caso), la idea es encontrar la línea óptima de separación entre las categorías.

3.4. Técnica de muestreo

Para efectos del proyecto se consideró necesario emplear una técnica de muestreo que permitió seleccionar los datos para el entrenamiento y sus respectivas pruebas, teniendo en cuenta lo anterior se eligió el muestreo aleatorio simple debido a que cada registro tiene la misma probabilidad de pertenecer a la muestra que se asignó a pruebas de los algoritmos (árboles de decisión y SMV), además también se empleó para cumplir el hecho de tener la misma distribución, un algoritmo pseudoaleatorio que cumple con las pruebas de distribución y uniformidad. Para esto a cada registro se le fue asignado un número que permitió emplear el algoritmo que retornó los números "aleatorios" que van a pertenecer al grupo de pruebas, cabe aclarar que se decidió dejar un 25 % del dataset para las pruebas de los algoritmos y el restante forman parte del conjunto de entrenamiento, esto debido a que es recomendable tener un subconjunto de pruebas lo suficientemente grande para generar resultados significativos estadísticamente y que representen al volumen total de los datos la proporción puede variar desde un 70 % - 30 % hasta un 80 % - 20 %.

3.5. Configuración de los modelos seleccionados

Para este tipo de técnicas primero se importó la base de datos en donde se encontraba la información transformada y los valores de cada campo de tipo numérico, esto con el fin de facilitar el procesamiento de los mismos, después de haber importado el dataset se asignaron a varias variables los valores de cada campo (tipo arreglo) por medio de DataFrame y el campo "Deserto" fué asignado a la variable "y" la cual sería la variable que contiene los valores a predecir y X tendría todas las variables anteriormente mencionadas con sus respectivos valores en una matriz, todo esto por medio de la librería de Python llamada Pandas. Con esto se tendría la construcción del dataset para después poder particionarlo en datos de entrenamiento y datos de testeo, para esto se utilizó "train_test_split" perteneciente a la librería "sklearn.model_selection" y se crearon las variables X_train, X_test, y_train, y_test encargadas de guardar el particionamiento del dataset con una proporción de 75 %-25 % para entrenamiento y testeo respectivamente.
Al tener los datos seleccionados se procedió a configurar cada modelo junto con la información recolectada de la siguiente manera:

Máquinas de Vectores de Soporte (SVM): Después de tener particionado el dataset en entrenamiento y testeo se utilizó SVC de sklearn.svm, se definió el tipo de kernel que se iba a utilizar para este, el cual fue "linear", los demás parámetros se dejaron como vienen predeterminados, ya que se observó de que en algunos casos que al cambiar estos parámetros el resultado al aplicar las métricas de evaluación no variaba mucho además algunos de los parámetros se usaban con un tipo diferente de kernel; después para realizar el respectivo entrenamiento de SVM con .fit se pasan como parámetros las variables X_train, y_train.

Árboles de decisión: Con los subconjuntos de datos ya definidos, se procede a utilizar el algoritmo de árboles de decisión de la librería sklearn donde se establece el criterio que se va tomar para la construcción del árbol, para este caso se utiliza el criterio "entropy" que permite medir la ganancia de información del nodo construido, también se asigna al parámetro max_deph el valor 10 debido a que con este nivel de profundidad se generan reglas concluyentes, poco repetitivas y ayuda a mejorar la calidad de la clasificación, así mismo al parámetro splitter se asigna el valor "best" porque construye el árbol a partir del mejor atributo y sus derivados son escogidos de la mejor relación entre estos, caso contrario al valor "random" donde se escogía un atributo de esta manera y el mejor valor posible de este atributo (divisor, min_samples, min_samples_leaf, min_eright_fraction_leaf, max_features, random state, min_impurity_decreace) Para los anteriores parámetros, se dejan sus valores por defecto debido a que es la manera en la que el árbol mejor se construye. En la documentación hacen la recomendación de customizar estos parámetros en el caso donde se esté consumiendo muchos recursos computacionales, esto por supuesto depende de la base de datos que se le entregue al algoritmo, para el caso del dataset construido y entregado al modelo, no hay problemas.

3.6. Plan de pruebas

Una vez preparada la base de datos y seleccionadas las dos técnicas de minería de datos, se procedió a implementar los algoritmos evaluando la viabilidad de las variables para seleccionar las opciones pertinentes que se presentan en las técnicas, en el caso de la máquina de vectores de soporte, el kernel adecuado que permitió obtener resultados positivos en la clasificación de la información y en los árboles de decisión el algoritmo óptimo que por las características de los datos posibilitó lograr resultados oportunos para el cumplimiento de los objetivos del proyecto.

Árboles de decisión: Para los árboles de decisión, una vez escogido el algoritmo apropiado, se procede a dividir el dataset en dos porcentajes, un porcentaje del 70 % para los datos de entrenamiento, el restante para los datos de prueba. Posterior a esto se implementó el algoritmo y se realizaron los ajustes necesarios para asegurar una eficiencia superior al 80 % de precisión al evaluar la implementación con los datos de prueba.

- Se dividió el dataset en dos partes, la primera corresponde al conjunto de datos de entrenamiento, tiene un porcentaje del 75 %, la segunda corresponde al conjunto de datos de prueba y tiene un porcentaje del 25 %.
- Se implementó el algoritmo de árboles de decisión a partir de x número de registros correspondientes a la información de los estudiantes en el dataset.
- Se obtuvo el árbol de decisión, que en sus ramas trae las reglas con sus respectivos ajustes a partir del entrenamiento.
- Se prueba el modelo resultante con el conjunto de datos de prueba para observar los resultados que este árbol arroja ante este segmento del dataset.
- Se realizan los análisis con las herramientas de evaluación para determinar si los resultados son adecuados o se debe modificar el árbol creado o los parámetros para la ejecución.

Máquinas de vectores de soporte (SVM): En las máquinas de vectores de soporte a la hora de realizar el respectivo entrenamiento este consta de dos fases:

- Se dividió el dataset transformado en dos partes, la primera corresponde al conjunto de datos de entrenamiento el cual tiene un porcentaje del 75 %, la segunda corresponde al conjunto de datos de prueba y tiene un porcentaje del 25 %.
- Se transformaron los predictores (datos de entrada) en un espacio de características altamente dimensional, para esta fase fue suficiente con especificar el kernel que se utilizará para llevar a cabo el proceso; los datos nunca se transforman explícitamente al espacio de características. Este proceso se conoce comúnmente como el truco kernel.

Después de este proceso de clasificación lo siguiente fue comparar los resultados obtenidos en la curva ROC que se presenta en la figura 4.1 y también mirando las demás métricas de evaluación que se obtuvieron en cada prueba hecha, para con esto llegar a un respectivo análisis y ver que tan eficiente es la técnica.

Capítulo 4

Análisis y discusión de resultados

Una vez ejecutadas las etapas de entrenamiento y test con las técnicas de minería de datos escogidas, se obtuvieron los resultados mostrados en la tabla 4.1, que fueron analizados con métricas de evaluación de modelos de aprendizaje supervisado: precisión, F1, exactitud, sensibilidad y curva ROC, esto con el fin de determinar cual modelo arroja mejores resultados respecto al otro y así escoger cual de los dos se va a implementar en el prototipo del clasificador binario.

Tabla 4.1: Resultados métricas de evaluación de los modelos

	Precisión		**Recall score**		**F1 score**		**Accuracy**		**Matriz de confusión**	
Prueba	SVM	Árboles de decisión	SVM	Árboles de decisión	SVM	Árboles de decisión	SVM	Árboles de decisión	SVM	Árboles de decisión
1	0.8310	0.8553	0.8213	0.8038	0.8261	0.8288	0.8571	0.8627	[[645 86] [92 423]]	[[661 70] [101 414]]
2	0.8310	0.8562	0.8213	0.8097	0.8261	0.8323	0.8571	0.8651	[[645 86] [92 423]]	[[661 70] [98 417]]
3	0.8310	0.8562	0.8213	0.8097	0.8261	0.8323	0.8571	0.8651	[[645 86] [92 423]]	[[663 68] [102 413]]
4	0.8310	0.8586	0.8213	0.8019	0.8261	0.8293	0.8571	0.8635	[[645 86] [92 423]]	[[661 70] [98 417]]
5	0.8310	0.8562	0.8213	0.8097	0.8261	0.8323	0.8571	0.8651	[[645 86] [92 423]]	[[650 81] [102 413]]

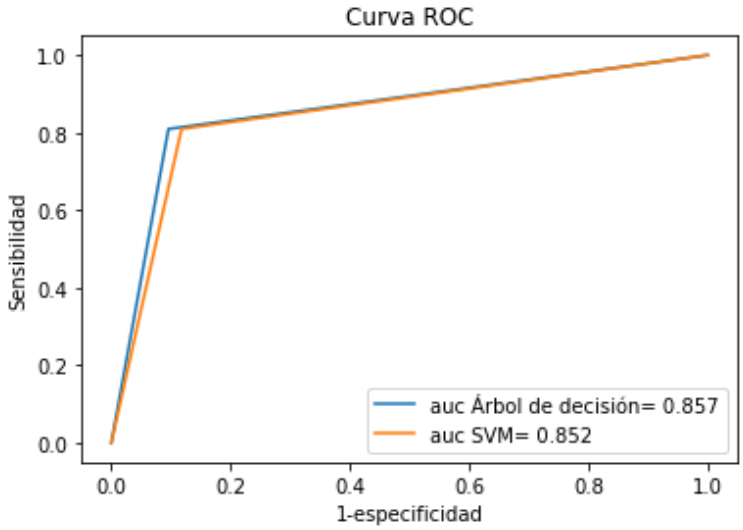


Figura 4.1: Curva ROC de SVM y Árbol de decisión.
Fuente: Elaboración propia

Con la curva ROC se pudo concluir que el modelo a implementar debía ser árboles de decisión debido a que su AUC o área bajo la curva como se puede ver en la figura 4.1 era superior al obtenido por las máquinas de vectores de soporte alcanzando un máximo de 0.85, con esto teniendo mayor probabilidad para distinguir las instancias positivas y las instancias negativas, haciendo de este algoritmo el más efectivo para el caso de estudio planteado.

Precisión: En ambos algoritmos se obtuvieron buenos valores para este dato como se muestra mejor en la imágen 4.2; la precisión es la que indica la razón que hay entre el número de verdaderos positivos y el número de falsos positivos, mientras el número sea más próximo a 1, mayor es el grado de precisión del modelo. Para obtener este valor se utilizó para ambas técnicas la librería 'sklearn.metrics' la cual brinda las diferentes métricas de evaluación para modelos de aprendizaje supervisado para probar la eficiencia del algoritmo; para este caso se utilizó el método 'precision_score' el cual recibe como parámetros y_test (el conjunto de testeo obtenido del dataset) y predict, la cual es una variable en donde se guardaron los resultados obtenidos de la predicción de los algoritmos, con esto retorna los valores postulados en la anterior tabla.

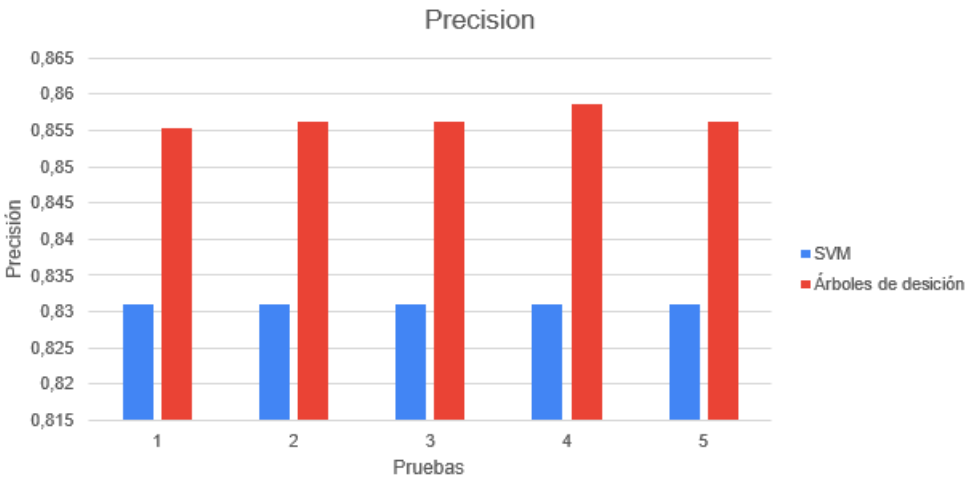


Figura 4.2: Gráfica comparativa precisión SVM vs Árboles de desición.
Fuente: Elaboración propia

Recall score: Para calcular esta métrica utilizamos la librería mencionada anteriormente, el valor obtenido en SVM para todas las pruebas fué de 0.8213 y el de árboles de decisión estuvo entre 0.8019 y 0.8097, para llevar a cabo este cálculo se hizo uso del método recall_score el cual recibe como parámetros y_test (el conjunto de testeo obtenido del dataset) y predict, la cual es una variable en donde se guardaron los resultados obtenidos de la predicción de los algoritmos, a continuación se muestra la gráfica comparativa de este resultado entre SVM y Árboles de desición.

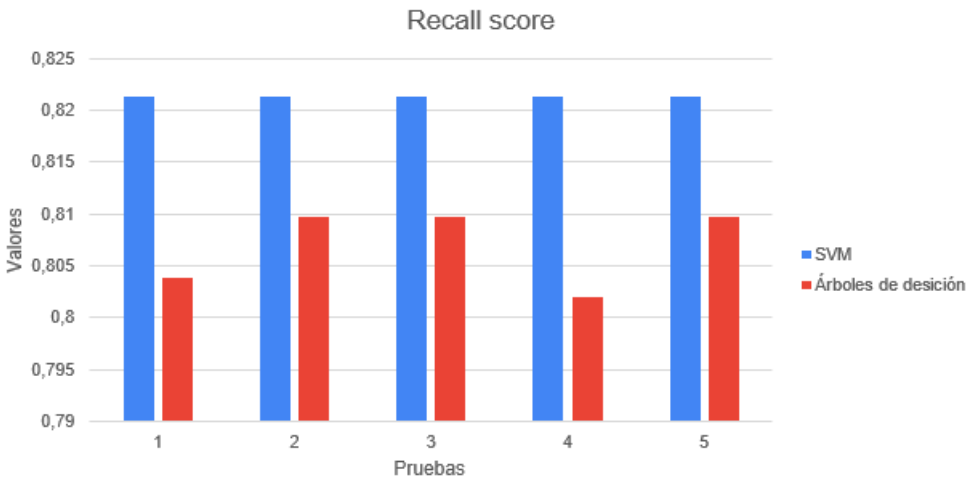


Figura 4.3: Gráfica comparativa Recall score SVM vs Árboles de desición.
Fuente: Elaboración propia

F1 score: En esta métrica se utilizó la librería mencionada en los puntos anteriores y de esta se obtuvo el método f1_score el cual recibe el conjunto de parámetros de testeo y_test y el conjunto de predicciones obtenidas por el algoritmo a evaluar, para así obtener cada uno de los valores que se encuentran en la tabla anterior y con esto analizar mientras sea más acercado a 1 mejor es la precisión y el recall score.

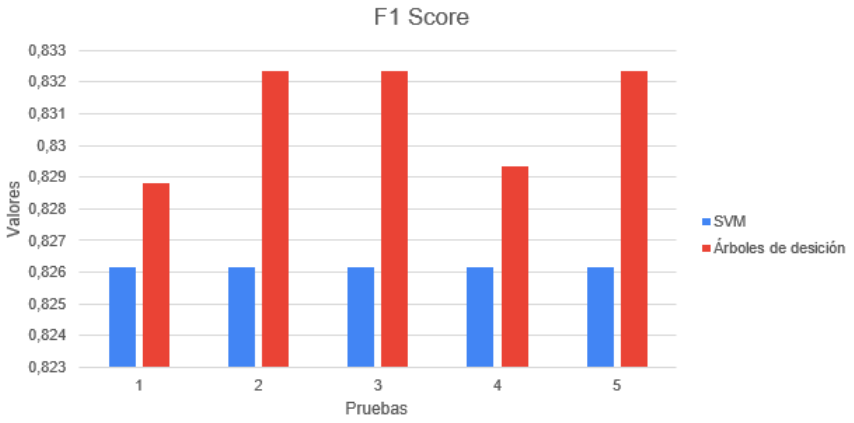


Figura 4.4: Gráfica comparativa F1 score SVM vs Árboles de desición.
Fuente: Elaboración propia

Accuracy: Para la métrica accuracy se utilizó la librería ya mencionada, se utilizó el metodo accuracy_score que es la razón entre la suma de los verdaderos positivos y los verdaderos negativos sobre la sumatoria de todos los campos de la matriz de confusión, se pasan como parámetros el subconjunto que contiene a la variable dependiente y las predicciones por parte de los modelos. Para los árboles de decisión se obtuvo 0.8651.

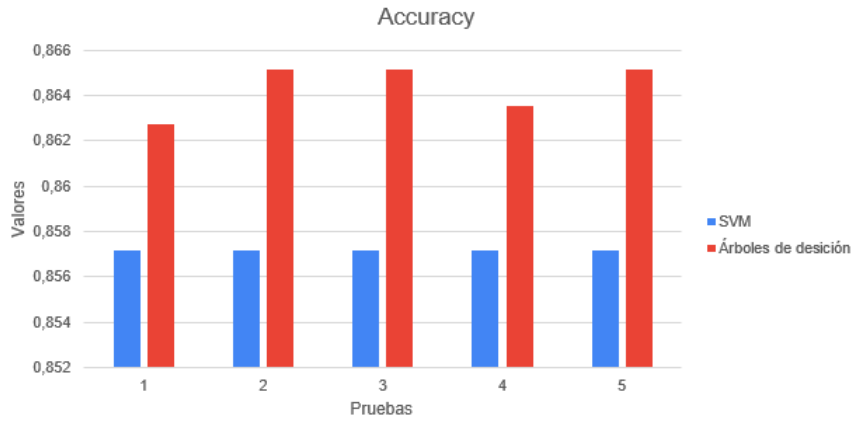


Figura 4.5: Gráfica comparativa Accuracy SVM vs Árboles de desición.
Fuente: Elaboración propia

Al obtener los resultados de los dos algoritmos seleccionados se pudo observar que ambos alcanzaron altos porcentajes de precisión, recall score, accuracy y demás estando entre 80 % y 86 %, teniendo como mejor modelo los árboles de decisión. Basado en lo anterior y siendo coherentes con los objetivos del proyecto, se procedió a elegir la técnica de los árboles de decisión debido a que en la mayoría de pruebas mantuvo resultados superiores al modelo de máquina de vectores de soporte, además dando el plus de la obtención de reglas que permiten de manera clara entender el razonamiento por el cual se da la clasificación en dicho modelo.

En el proceso de configuración del modelo de árboles de decisiones, se probaron varias alternativas en algunos de los parámetros para obtener los mejores resultados posibles, uno de los más importantes fue la profundidad del árbol, para el valor final en este parámetro se probó desde una altura de 8 hasta dejar total libertad en la construcción del árbol, para los casos en el que la altura era inferior a 9 se omitían algunos atributos que podrían ser determinantes generando pérdida de información, en el caso del parámetro default, se observó que las reglas a partir de la altura 10 u 11 se portaban repetitivas, razón por la cual la altura máxima fue de 10, esto debido que los resultados y reglas no variaban en caso de agregarse algunas unidades de profundidad y en algunas de las ramas estas se ajustaban de una manera más concluyente. Se presentan algunas de las reglas con más instancias:

1. Promedio general del estudiante $> 3.245 \rightarrow$ Graduado $> 0 \rightarrow$ No deserta.
   Esta es la regla con mayor número de instancias representadas en ella con un total de 863 instancias.

2. Promedio general del estudiante $\leq 3{,}245 \rightarrow$ Graduado $> 0 \rightarrow$ No deserta.
   Esta regla posee un total de 56 instancias.

3. Promedio general del estudiante $\leq 3.245 \rightarrow$ Graduado $< 1 \rightarrow$ BRA $< 1 \rightarrow 3753 = 0 \rightarrow$ Promedio general del estudiante $\leq 3.065 \rightarrow$ Periodo de admisión $< 2 \rightarrow$ Promedio de créditos matriculados por periodo $\leq 11.415 \rightarrow$ Promedio de créditos matriculados por periodo $\leq 12.83 \rightarrow$ Edad de ingreso $\leq 20.5 \rightarrow$ Número de asignaturas matriculadas por periodo $> 4.25 \rightarrow$ Deserta.
   Esta regla posee un total de 99 instancias.

4. Promedio general del estudiante $\leq 3.245 \rightarrow$ Graduado $< 1 \rightarrow$ BRA $> 1 \rightarrow$ Promedio de créditos matriculados por periodo $\leq 13.185 \rightarrow$ Jornada nocturna $= 1 \rightarrow$ Número de asignaturas matriculadas por periodo $\leq 4.52 \rightarrow$ BRA $> 1 \rightarrow$ Promedio general del estudiante $\leq 3.225 \rightarrow 2711 = 0 \rightarrow$ Proporción créditos matriculados por aprobados $\leq 0.665 \rightarrow$ Deserta.
   Esta regla posee un total de 108 instancias.

5. Promedio general del estudiante $\leq 3.245 \rightarrow$ Graduado $< 1 \rightarrow$ BRA $> 1 \rightarrow$ Promedio de créditos matriculados por periodo $> 13.185 \rightarrow$ Promedio general del estudiante $\leq 2.615 \rightarrow$ Programa tecnológico $= 0 \rightarrow$ Promedio general del estudiante $\leq 1.915 \rightarrow$ Proporción créditos matriculados por aprobados $\leq 0.455 \rightarrow$ Número de asignaturas matriculadas por periodo academico $> 5.8350 \rightarrow$ Otra ciudad de residencia $= 0 \rightarrow$ Deserta.
   Esta regla posee un total de 92 instancias.

6. Promedio general del estudiante $\leq 3.245 \rightarrow$ Graduado $< 1 \rightarrow$ BRA $> 1 \rightarrow$ Promedio de créditos matriculados por periodo $> 13.185 \rightarrow$ Promedio general del estudiante $\leq 2.615 \rightarrow$ Programa tecnológico $= 1 \rightarrow$ Número de asignaturas matriculados por periodo academico $> 4.365 \rightarrow$ Promedio de créditos matriculados por periodo $\leq 18.35 \rightarrow$ Promedio de créditos matriculados por periodo $\leq 17.365 \rightarrow$ Sexo femenino $= 0 \rightarrow$ Deserta.
   Esta regla posee un total de 164 instancias.

7. Promedio general del estudiante $>$ 3.245 $\rightarrow$ Graduado $<$ 1 $\rightarrow$ BRA $<$ 2 $\rightarrow$ Jornada diurna 0 $\rightarrow$ Promedio general del estudiante $\leq$ 3.835 $\rightarrow$ Periodo de admisión = 1 $\rightarrow$ Promedio de créditos matriculados por periodo $>$ 10.37 $\rightarrow$ Proporción créditos matriculados por aprobados $\leq$ 0.99 $\rightarrow$ Proporción créditos matriculados por aprobados $\leq$ 0.845 $\rightarrow$ Número de asignaturas matriculados por periodo académico $>$ 3.96 $\rightarrow$ No Deserta.
   Esta regla posee un total de 164 instancias.

8. Promedio general del estudiante $>$ 3.245 $\rightarrow$ Graduado $<$ 1 $\rightarrow$ BRA $<$ 2 $\rightarrow$ Jornada diurna = 0 $\rightarrow$ Promedio general del estudiante $\leq$ 3.835 $\rightarrow$ Periodo de admisión = 1 $\rightarrow$ Promedio de créditos matriculados por periodo $>$ 10.37 $\rightarrow$ Proporción créditos matriculados por aprobados $\leq$ 0.99 $\rightarrow$ Proporción créditos matriculados por aprobados $>$ 0.845 $\rightarrow$ Número de asignaturas matriculados por periodo académico $>$ 4.065 $\rightarrow$ No Deserta.
   Esta regla posee un total de 67 instancias.

9. Promedio general del estudiante $>$ 3.245 $\rightarrow$ Graduado $<$ 1 $\rightarrow$ BRA $<$ 2 $\rightarrow$ Jornada diurna = 1 $\rightarrow$ Promedio general del estudiante $\leq$ 3.835 $\rightarrow$ Promedio general del estudiante $>$ 3.365 $\rightarrow$ Promedio de créditos matriculados por periodo $>$ 17.685 $\rightarrow$ Número de asignaturas matriculados por periodo académico $>$ 6.39 $\rightarrow$ Edad de ingreso $>$ 14.5 $\rightarrow$ Promedio de créditos matriculados por periodo $>$ 20.415 $\rightarrow$ Proporción créditos matriculados por aprobados $>$ 0.515 $\rightarrow$ No Deserta.
   Esta regla posee un total de 88 instancias.

10. Promedio general del estudiante $>$ 3.245 $\rightarrow$ Graduado $<$ 1 $\rightarrow$ BRA $<$ 2 $\rightarrow$ Jornada diurna = 1 $\rightarrow$ Promedio general del estudiante $\leq$ 3.835 $\rightarrow$ Promedio general del estudiante $>$ 3.365 $\rightarrow$ Promedio de créditos matriculados por periodo $>$ 17.685 $\rightarrow$ Edad de ingreso $>$ 14.5 $\rightarrow$ Número de asignaturas matriculados por periodo académico $>$ 6.39 $\rightarrow$ Promedio de créditos matriculados por periodo $\leq$ 20.415 $\rightarrow$ Número de asignaturas matriculados por periodo académico $\leq$ 7.415 No Deserta.
    Esta regla posee un total de 108 instancias.

De forma general se podría decir que de acuerdo a los resultados, es mayor el número de estudiantes que tienden a desertar cuando su promedio es inferior a 3.245 y tiene bajos rendimientos académicos, este par de atributos son el pilar fundamental de las reglas generadas por el modelo de árboles de decisión, además a partir de esos nodos suele variar la generación de nuevas hojas cuando los siguientes atributos toman estos valores: Promedio de créditos matriculados por periodo inferior a 13.2, jornada diurna y promedio general del estudiante inferior a 3. Siendo estas las principales características después del promedio inferior a 3.245 y los BRA. A partir de ese punto los nodos toman los valores de los atributos: proporción de créditos matriculados por aprobados, número de asignaturas matriculadas por periodo, ciudad de residencia y código de programa. Por último se tiene que de acuerdo a las reglas generadas, los atributos: tipo de zona, tipo de programa y condición de excepción no son atributos fundamentales para determinar a un estudiante con tendencia a incurrir en deserción.

A continuación se muestran algunas gráficas que reflejan los resultados de la clasificación positiva de estudiantes que tienden a incurrir en deserción por parte del modelo de árboles de decisión.

Figura 4.6: Gráfica desertores por programa académico.
Fuente: Elaboración propia

De acuerdo a los resultados obtenidos en las métricas del modelo árboles de decisión en la gráfica 4.6 se puede observar la cantidad de desertores por programa académico, donde se observa que el programa de Tecnología en sistemas de información es el programa con mayor número de desertores.

Figura 4.7: Gráfica desertores por tipo de programa.
Fuente: Elaboración propia

Como se pudo observar en la gráfica anterior los tipos de programas con mayor índice de deserción eran Tecnologías, en esta gráfica 4.7 se muestra detalladamente este resultado en dónde la cantidad de desertores por programa tecnológico obtuvo un valor de 204 y para programas de pregrado un total de 292.

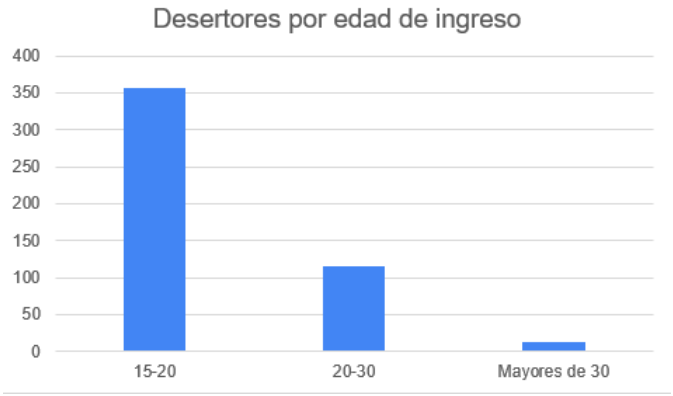


Figura 4.8: Gráfica desertores por rango de edad de ingreso.
Fuente: Elaboración propia

En la gráfica 4.8 se obtuvieron los rangos de edad de ingreso de los estudiantes que desertaron, como se puede observar el mayor índice de deserción está en los estudiantes menos de 20 años.

Figura 4.9: Gráfica desertores por cantidad de BRA.
Fuente: Elaboración propia

En la gráfica 4.9 se puede observar la cantidad de estudiantes que desertaron de acuerdo a su situación de bajo rendimiento académico, dejando como índice más alto el poseer un bajo rendimiento académico.

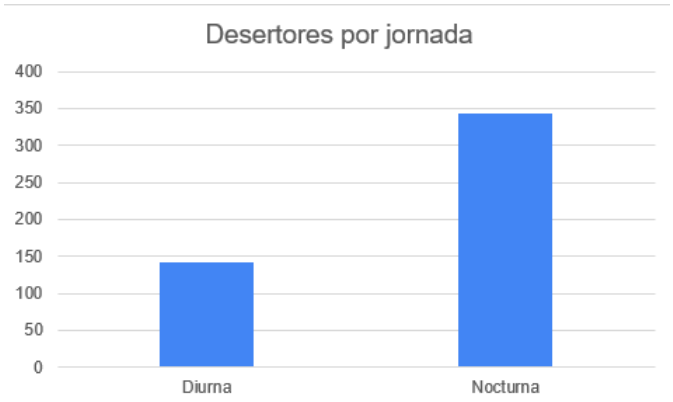


Figura 4.10: Gráfica desertores por jornada.
Fuente: Elaboración propia

En la gráfica 4.10 se puede observar la cantidad de estudiantes que desertaron de acuerdo a la jornada a la que pertenecen los estudiantes, dejando como índice más alto la jornada nocturna.

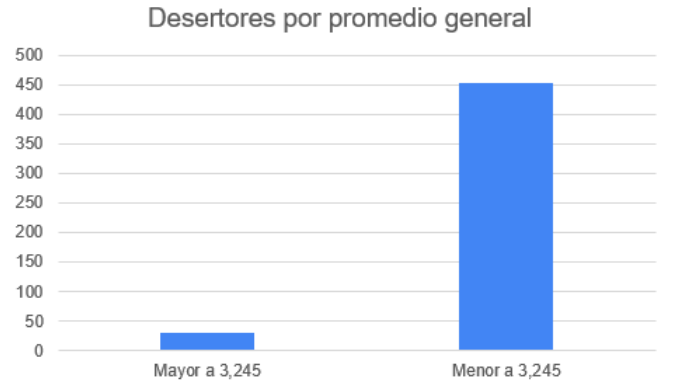


Figura 4.11: Gráfica desertores por promedio general.
Fuente: Elaboración propia

En la gráfica anterior se observa que la gran mayoría de estudiantes que tenían un promedio general inferior a 3,245 desertaron de la carrera en donde se encontraban matriculados esto también mencionado en las reglas generadas por el árbol de decisión.

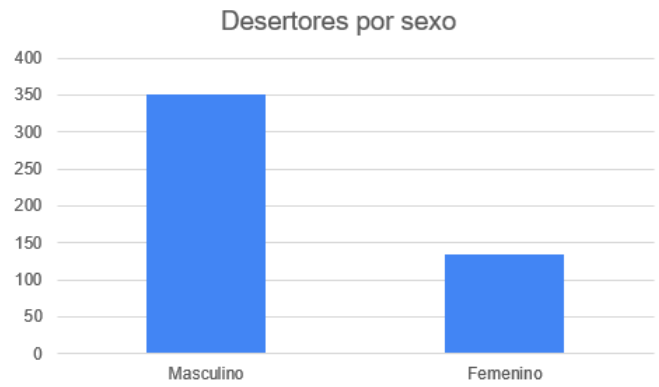


Figura 4.12: Gráfica desertores por sexo.
Fuente: Elaboración propia

Como se puede observar en la gráfica anterior la gran parte de los desertores en los diferentes programas académicos fueron hombres con un total de 369 estudiantes versus las mujeres que tienen un total de 127.

Figura 4.13: Gráfica desertores por tipo de zona.
Fuente: Elaboración propia

En la grafica 4.13 se evidencia que para el caso de la sede Tuluá de la Universidad del Valle gran parte de los estudiantes desertores se encuentran en una zona de tipo urbana.

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

1. Con el desarrollo del clasificador binario usando la técnica árboles de decisión, se lograron resultados superiores al 80 % de acuerdo a las métricas: precisión, F1, exactitud, sensibilidad y curva ROC, que son superiores a las obtenidas cuando se utiliza el modelo de máquinas de vectores de soporte con lo que se puede determinar con un mayor nivel de certeza cuando un estudiante puede incurrir en deserción.

2. El diseño de la base de conocimientos con las características correspondientes a los aspectos personales, académicos, socioeconómicos e institucionales permitió clasificar los casos de deserción en la sede Tuluá de la Universidad del Valle, logrando el desarrollo de las tareas y los objetivos planteados en el proyecto.

3. El análisis de los datos para conformar el dataset permitió identificar problemas de calidad como: el ruido, datos faltantes y outliers; la limpieza de dichos datos condujo a la obtención de características que dan valor agregado al análisis y a las tareas realizadas, dejando como resultado la implementación del clasificador binario.

4. La elección de las técnicas de árboles de decisión y las máquinas de vectores de soporte se basó en las características que proporcionaban estos modelos, en el aprovechamiento de ellos y la dimensión de los recursos obtenidos. Esto evidenció que algunos de los modelos revisados anteriormente tienden a ser más efectivos en la resolución de problemas más complejos y que otras herramientas tienden a ser más útiles con problemas simples; es por esto que frente al problema que trata el proyecto se comprueban como poco precisas.

5. Finalmente se concluye que el contar con una herramienta que alerte sobre un posible caso de deserción, permite a la universidad diseñar una estrategia de acompañamiento con el fin de ayudar al estudiante en su paso por ella y a la misma vez investigar más a fondo el fenómeno de la deserción que se vive en la Universidad del Valle generando impactos positivos para la sociedad dentro y fuera de ella.

5.2. Trabajos futuros

1. Ampliar el alcance del proyecto con la información que se obtenga de las demás sedes regionales, esto con el objetivo de determinar si las condiciones de deserción pueden variar de una sede a otra.

2. Se recomienda ampliar la base de conocimiento con la información socioeconómica de los estudiantes, en una primera etapa, a través de los datos registrados en el ICFES por medio de las pruebas SABER 11 y en una segunda etapa, con la implementación de un proceso de actualización esta información con el fin de incorporar nuevas variables en la identificación de posibles casos de deserción.

3. Se recomienda realizar una herramienta de predicción con ayuda de técnicas de minería de datos que tenga en cuenta la variación de características y competencias por facultad.

4. Se recomienda expandir la implementación de técnicas de minería de datos para evaluar los resultados que se puedan obtener.

5. Hacer uso de una estrategia con el fin de darle un manejo más adecuado a los datos categóricos.

Capítulo 6

Bibliografía

[1] B. Salazar, "¿qué debemos aprender de los desertores?: Deserción y decisión racional en la universidad del valle," *Cidse, El Observador Regional*, 2007.

[2] C. P., "Efectos de vecindario como determinantes de la deserción estudiantil y el logro académico en la universidad del valle," 2012.

[3] R. A. Benitez, "Proyecto r.a.m.o.n, refuerzo académico masivo ordenado normalizado," 2012.

[4] G. Carolina, D. Diana, F. Jorge, C. Elkin, G. Santiago, G. Karoll, and V. Johanna, "Deserción estudiantil en la educación estudiantil colombiana," *Ministerio de educación nacional*, 2009.

[5] "Ley 1581 de 2012, protección de datos personales," 2012.

[6] A. Echeverri, P. Retamoza, O. de la Rosa, V. Barros, O. Álvarez, and C. Guerrero, "Minería de datos como herramienta para el desarrollo de estrategias de mercadeo b2b en sectores productivos, afines a los colombianos: una revisión de casos," *Sotavento M.B.A. 22(2013)*, 2013.

[7] "La predicción del dato: Redes neuronales artificiales." url: https://www.merkleinc.com/es/es/blog/prediccion-dato-redes-neuronales-artificiales.

[8] "Árbol de decisión, una herramienta para decidir bien." url: https://www.altonivel.com.mx/liderazgo/management/36690-arbol-de-decision-una-herramienta-para-decidir-correctamente/.

[9] "Algoritmo de agrupamiento." url: https://es.wikipedia.org/wiki/Algoritmo_de_agrupamiento.

[10] "Clasificador bayesiano." url: https://es.wikipedia.org/wiki/Clasificador_bayesiano_ingenuo.

[11] "Reglas de asociación." url: https://es.wikipedia.org/wiki/Reglas_de_asociaci %C3 %B3n.

[12] "Support - vector machine." url: https://en.wikipedia.org/wiki/Support-vector_machineComputing_the_SVM_classifier.

[13] "Máquinas de vectores de soporte (svm)." url: https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/.

[14] "Aprendizaje automático basado en reglas." url: https://es.qaz.wiki/wiki/Rule-based_machine_learning.

[15] "Concepto de reglas estrictas." url: https://es.qaz.wiki/wiki/Association_rule_learning.

[16] E. Himmel, "Modelo de análisis de la deserción estudiantil en la educación superior," 2002.

[17] Y. K. Amaya, E. B. Avendaño, and D. J. Heredia, "Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos," Marzo 2014.

[18] B. CUJI, W. GAVILANES, and R. SANCHEZ, "Modelo predictivo de deserción estudiantil basado en arboles de decisión," 2017.

[19] O. Área de Análisis Institucional, "Estudio sobre deserción usando la herramienta spadies para el programa de ingeniería de sistemas en 2007_ii," Julio de 2009.

[20] J. H. Escobar, "La deserción universitaria: un problema que se debe abordar afrontar en varias dimensiones," 2007.

[21] J. H. Escobar, E. Largo, and C. A. Pérez, "Factores asociados a la deserción y permanencia estudiantil en la universidad del valle," Octubre de 2006.

[22] E. Jaime, P. Carlos, and L. Edwin, "Rendimiento académico en la universidad del valle: determinantes y su relación con la deserción estudiantil," 2008.

[23] M. Omar, "Factores que inciden en la deserción en el programa académico tecnología en gestión portuaria de la universidad del valle sede pacífico en el periodo 2009-2014," 2015.

[24] J. C. C. Bossa, "Modelado y gestion de la informacion." url: http://juliocarrreno.blogspot.com/2014/10/modelo-de-marcacion-telefonica-basado.html.

[25] J. C. Canales, "Clasificación de grandes conjuntos de datos vía máquinas de vectores soporte y aplicaciones en sistemas biológicos," 2009.

Anexo 1

Código fuente

El codigo fuente desarrollado en Python para llevar a cabo el desarrollo de este proyecto, se encuentra en un repositorio de Github, disponible en el siguiente enlace:

- https://github.com/carlosmarin2000/TrabajodeGrado

A continuación se muestra una imagen que representa la organización del repositorio en cuestión.

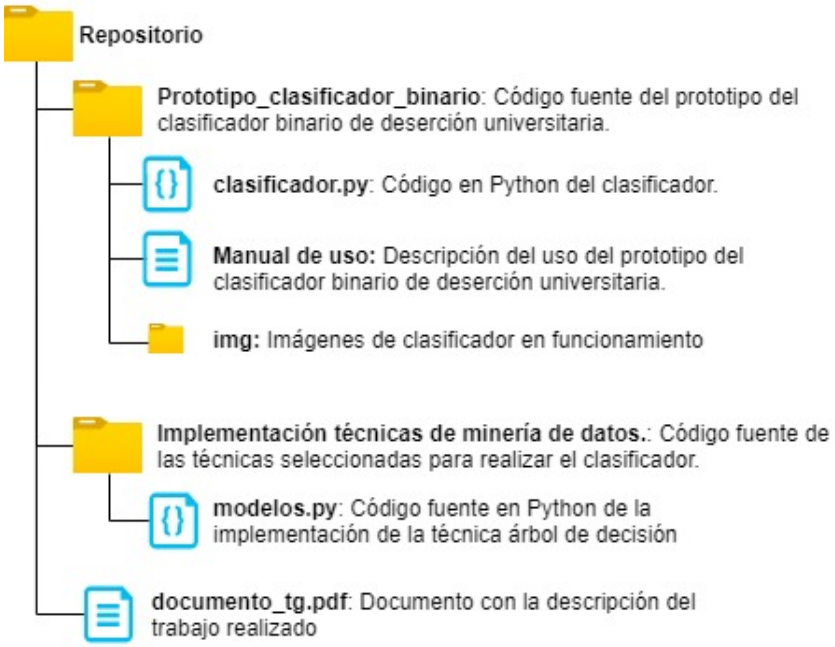
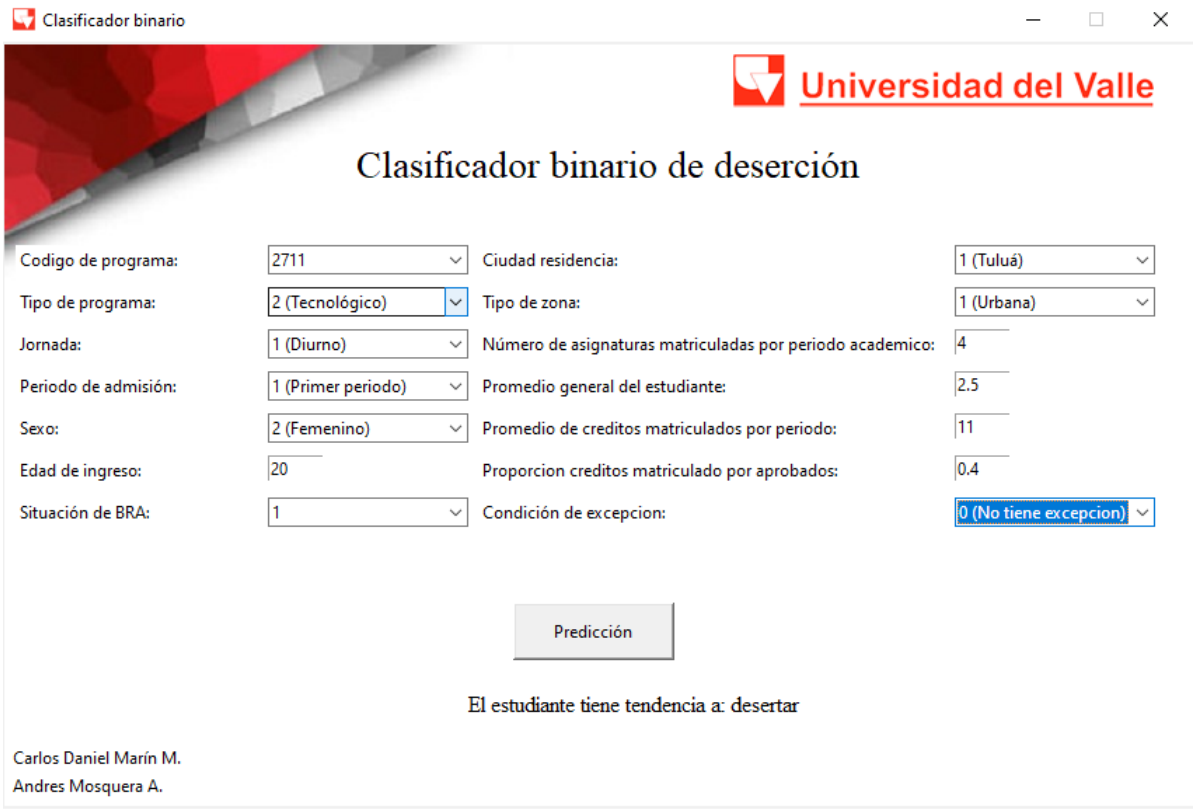


Diagrama repositorio Github
Fuente: Elaboración propia

Anexo 2

Uso del clasificador binario

El prototipo realizado en este trabajo de grado tiene como función determinar si un estudiante de la sede Tuluá de la subregión centro de la Universidad del Valle tiende a desertar dadas las características que se muestran en la interfaz. A continuación se encuentra una imagen representativa del prototipo en uso.



Aplicación del clasificador binario
Fuente: Elaboración propia

Anexo 2

En el siguiente enlace puede encontrar una demostración del prototipo desarrollado:

- https://youtu.be/VvYMb4TWNHo