PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EN LOS ESTUDIANTES DEL PROGRAMA DE PSICOLOGÍA DE LA FUNDACIÓN UNIVERSITARIA DE POPAYÁN UTILIZANDO DATOS DE LA IDENTIDAD CULTURAL Y EL ESTILO DE APRENDIZAJE

MARÍA ALEJANDRA MOLANO GUACÁN
CRISTIAN DAVID PIAMBA RODRIGUEZ

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN IMS
Popayán, JULIO de 2020

PREDICCIÓN DEL RENDIMIENTO ACADÉMICO EN LOS ESTUDIANTES DEL PROGRAMA DE PSICOLOGÍA DE LA FUNDACIÓN UNIVERSITARIA DE POPAYÁN UTILIZANDO DATOS DE LA IDENTIDAD CULTURAL Y EL ESTILO DE APRENDIZAJE

MARÍA ALEJANDRA MOLANO GUACÁN
CRISTIAN DAVID PIAMBA RODRIGUEZ

Informe final de Seminario de investigación como opción de grado

Director:
José Armando Ordoñez

FUNDACIÓN UNIVERSITARIA DE POPAYÁN
FACULTAD DE INGENIERÍA
PROGRAMA DE INGENIERÍA DE SISTEMAS
GRUPO DE INVESTIGACIÓN IMS
Popayán, JULIO de 2020

CONTENIDO

TABLAS

FIGURAS

CERTIFICACIÓN DE AUTORÍA

Certifico que conozco el concepto de plagiar según la Real Académica de la lengua (“Copiar en lo sustancial obras ajenas, dándolas como propias.”)

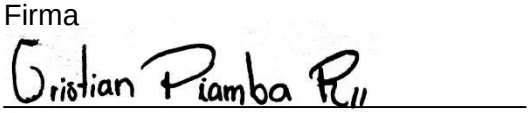
Y certifico que el contenido de este documento es de mi autoría, no hay contenido que haya sido copiado directamente y al pie de la letra de ninguna fuente. En el caso de ideas, teorías, conceptos, resultados y otros contenidos tomados de otros autores se menciona explícitamente la fuente original, y sólo en unos pocos casos se han mantenido el mismo texto, colocándolo entre comillas.

Reconozco las consecuencias académicas, jurídicas y económicas que conlleva el plagio.

Firma

Nombre del estudiante
CC. 1.061.773.743 de Popayán.

Firma



Nombre del estudiante
CC. 1.061.764.041 de Popayán.

RESUMEN

El bajo rendimiento académico ha sido el foco principal de atención en las instituciones de educación superior, provocando deserción por parte de los estudiantes, problemática que aqueja el sector laboral en el país, ya que las oportunidades laborales cada vez son pocas [1], de acuerdo a investigaciones se han tenido en cuenta que algunos de los factores influyentes son socioeconómicos, personales, laborales, entre otros.

Este trabajo plantea el desarrollo de un modelo de predicción del rendimiento académico en los estudiantes del programa de Psicología de la Fundación Universitaria de Popayán utilizando datos de la identidad cultural y el estilo de aprendizaje, en donde se realizaron encuestas para la captura de datos personales e información académica de los estudiantes del primer semestre de la carrera de Psicología, se procede a ejecutar la limpieza de datos para la optimización y creación de datos utilizando el modelo Linear Discriminant Analysis (LDA) en donde se obtuvo un 0,82 de precisión comparado con los demás modelos de predicción.

Palabras claves: Modelo de predicción, rendimiento académico, estilo de aprendizaje, identidad cultural.

CAPÍTULO I. ASPECTOS GENERALES DE LA INVESTIGACIÓN

1.1. DESCRIPCIÓN DEL PROBLEMA

El bajo rendimiento académico es uno de los principales problemas que se presentan hoy en día y que afecta directamente al desarrollo educativo de nuestro país [1]. Esto repercute en la disminución de las oportunidades de empleo de calidad [2], de igual manera a las altas tasas de desempleo, bajos ingresos, dificultades para el desarrollo personal y profesional [3].

Este problema ha obtenido un mayor enfoque en las instituciones de educación superior debido a las causas y a los factores económicos, motivacionales, sociológicos que inciden en el aumento del bajo desempeño académico que se está presentando actualmente a nivel nacional [4][5][6][7]. Cabe resaltar que esta problemática es una de las principales causantes de la deserción estudiantil afectando directamente al estado, la sociedad y las mismas instituciones de educación superior debido a los costos que se pueden generar [3].

Tomando como referencia otras investigaciones, algunas universidades a nivel nacional e internacional, se han dedicado a realizar procesos de investigación usando Minería de Datos para determinar los patrones e identificar las características de los estudiantes que presentan un bajo rendimiento ya sea en un programa académico o en una materia específica. En estos estudios se realiza la clasificación de los estudiantes con alto, medio y bajo rendimiento académico de acuerdo a las características comunes de datos socio-económicos, personales, académicos y demográficos de los estudiantes y su núcleo familiar [8]. Otros trabajos buscan Identificar los perfiles de los estudiantes que tienen un alto rendimiento académico y los que tenían bajo rendimiento, haciendo uso de técnicas de minería de datos (MD) para detectar las variables causantes del bajo rendimiento académico y poder tomar medidas al respecto.

A pesar de los avances conseguidos, los trabajos existentes no consideran la identidad cultural y el estilo de aprendizaje, siendo estos de suma importancia para el rendimiento académico debido a la existencia de la relación entre el grado de integración de la comunidad y el nivel de enseñanza de los estudiantes [9], así mismo, los estilos de instrucción influyen en las formas en que los estudiantes perciben, recuerdan y piensan, conservando y apoyando las iniciativas culturales de los alumnos ya que es vital para preservar su identidad y crecimiento personal.

Debido a lo anterior, y teniendo en cuenta que en la Fundación Universitaria de Popayán no se han evaluado este tipo de modelos, el presente proyecto de investigación se centra en analizar modelos de predicción sobre el rendimiento académico de los estudiantes del programa de psicología de dicha institución, a través de este análisis se busca definir un modelo que permita la predicción del rendimiento académico incluyendo el estudio de la identidad cultural y el estilo de aprendizaje, es decir, técnicas y métodos de aprendizaje propios de cada

estudiante, basados en un test de estilos de aprendizaje de los estudiantes de primer semestre de Psicología, realizado por el área de Bienestar Institucional.

1.2. FORMULACIÓN DEL PROBLEMA

¿Qué tanto influye la identidad cultural y el estilo de aprendizaje en el rendimiento académico y la deserción de los estudiantes de psicología de la Fundación Universitaria de Popayán?

1.3. OBJETIVOS

1.3.1. OBJETIVO GENERAL

Definir un modelo de predicción que permita identificar el rendimiento académico de los estudiantes del programa académico de psicología de la Fundación Universitaria de Popayán que incluya la identidad cultural y el estilo de aprendizaje.

1.3.2. OBJETIVOS ESPECÍFICOS

- Revisar el estado del arte sobre modelos de predicción del rendimiento académico.
- Definir un modelo de predicción para determinar el rendimiento académico que incluya la identidad cultural y el estilo de aprendizaje que se presente en el programa de psicología de la Fundación Universitaria de Popayán.
- Diseñar una estrategia de socialización de los resultados obtenidos por el modelo de predicción propuesto.

1.4. JUSTIFICACIÓN

Según Roberto Hernández Sampieri [10], en su obra destaca algunos criterios importantes para evaluar la calidad viable de una investigación, los cuales fueron adoptados para justificar el presente trabajo investigativo.

- **Valor metodológico de la investigación:** Existen trabajos similares a este realizados en otras instituciones de educación superior que no son adaptables a este trabajo de investigación, dado que un modelo de minería describe el comportamiento de un grupo particular de estudiantes [11].

  Por lo tanto, esta investigación busca examinar los datos socioeconómicos, socioculturales y académicos de los estudiantes para descubrir patrones y tendencias que permitan soportar nuevas suposiciones sobre el porqué se presenta dicho fenómeno y así definir un modelo evaluado mediante la identidad cultural y el estilo de aprendizaje que sea capaz de predecir el rendimiento académico.

- **Valor práctico de la investigación:** Ofrecer a Bienestar Institucional y a Vicerrectoría académica de la Universidad, un punto de vista más apropiado sobre el rendimiento académico de los estudiantes y así prevenir la deserción estudiantil.

- **Valor tecnológico:** El diseño de un sistema que contenga un modelo entrenado de minería de datos para predecir el rendimiento académico de los estudiantes de psicología de la Fundación Universitaria de Popayán a partir de las notas de otros períodos académicos cursados y datos socioeconómicos y socioculturales de los estudiantes.

  Por otra parte, se presenta este proyecto con el fin de generar un aporte a los sistemas de modelado de predicción de rendimiento académico aplicando metodologías de minería de datos y que será diseñado para uso institucional o en las comunidades donde surja la necesidad.

CAPÍTULO II. MARCO REFERENCIAL

2.1. MARCO CONCEPTUAL

En esta sección se presentará una explicación sencilla sobre los principales conceptos para el desarrollo de la presente investigación, esto con el fin de tener un mayor entendimiento de los impactos que pueden generarse y un mayor conocimiento de los conceptos a tratar.

2.1.1. Rendimiento académico:

Se señala el término de rendimiento académico como el nivel de beneficio que adquiere un individuo, teniendo en cuenta sus aptitudes y posibilidades de desempeño para cada materia. Su rendimiento académico puede ser bajo insuficiente, aún con notas suficientes, si su capacidad es alta o muy alta.

Por otra parte, Chadwick [5] define al rendimiento académico como la expresión de capacidades y de características psicológicas de los estudiantes desarrolladas y adquiridas a través del proceso de enseñanza-aprendizaje que posibilita obtener un nivel de funcionamiento y logros académicos a lo largo de un periodo o semestre, que se resume en un calificativo final.

2.1.2. Minería de datos:

Una de las definiciones para minería de datos o Data Mining es que se considera como un conjunto de técnicas que automatizan el descubrimiento de patrones relevantes, también se define como el proceso que permite transformar información en conocimiento de gran utilidad para el negocio, con la combinación de métodos que ayudan a la reducción de costos y riesgos.

La minería de datos o el Data Mining nació como una mezcla de diversas ciencias aplicadas tales como la estadística, el soporte a la toma de decisiones, el aprendizaje automático, la gestión y almacenamiento de bases datos y procesamiento en paralelo. Para la realización de estos procesos se aplican técnicas procedentes de muy diversas áreas, como pueden ser los algoritmos genéticos, las redes neuronales, los árboles de decisión, etc. [12]

2.1.3. Proceso CRISP-DM:

Como concepto básico se refiere al proceso de CRISP-DM (Cross-Industry Standard Process for Data Mining), como un método garantizado para orientar todo proyecto con base en minería de datos, este proceso como metodología, incluye descripciones de las fases y las tareas necesarias para desarrollarlo; como modelo de proceso la metodología CRISP-DM ofrece un resumen del ciclo vital de minería de datos [13].

2.1.4. Identidad cultural:

La identidad cultural se especifica como las características más relevantes y autóctonas de una región o comunidad, es todo aquello que hace que ese lugar sea único. Y todo esto gracias a patrimonios como monumentos, obras de arte, costumbres, el folclore, entre otros [14].

2.1.5. Estilo de aprendizaje:

Este concepto abarca muchas definiciones entre estas se puede definir al estilo de aprendizaje como los comportamientos, costumbres, la forma de escribir hasta la manera de expresarse de una persona [15].

2.2. ESTADO DEL ARTE

De acuerdo al progreso de la investigación, en esta sección se opta por emplear la técnica de mapeo sistemático, cuyo objetivo es obtener una amplia perspectiva de la propuesta de investigación [16], que en este determinado caso es sobre modelos de predicción que permitan identificar el rendimiento académico a través del análisis de datos de la identidad cultural y el estilo de aprendizaje, y así poder determinar el estado del arte del presente proyecto de investigación.

Para conocer las investigaciones divulgadas actualmente, su relación con el rendimiento académico de los estudiantes y si los factores como: la identidad cultural y estilo de aprendizaje afectan dicho rendimiento, es necesario realizar las etapas del procedimiento de revisión.

2.2.1. Pregunta de investigación

Se especifican las preguntas de investigación que se desean responder en el presente estudio.

- ¿A través de la identidad cultural y el estilo de aprendizaje se ha identificado el rendimiento académico?
- ¿Qué metodologías se han usado para la predicción del rendimiento académico?

2.2.2. Fuente de datos y estrategia de búsqueda

Una vez establecidas las preguntas de investigación se procede a la estrategia de búsqueda elaborada teniendo en cuenta el problema existente respecto al rendimiento académico. De modo que las cadenas de búsqueda están definidas de acuerdo a las preguntas de investigación a responder [17].

En este proyecto de investigación se seleccionaron motores de búsqueda como: IEEExplore, Ebsco y Scopus, que son bases de datos que pueden contener artículos relacionados al rendimiento académico.

2.2.3. Selección de estudios

Se definen los criterios para la clasificación de estudios.

- **Inclusión:** todos aquellos trabajos relacionados al tema de estudio que en este determinado caso trata de la predicción del rendimiento académico se tendrán en cuenta y que solo sean de revistas reconocidas. El periodo de las búsquedas comprendió desde el año 2015 hasta la presente fecha.
- **Exclusión:** todos aquellos trabajos que en su estudio no traten de predicción del rendimiento académico, información que no sea de utilidad para este estudio serán descartados en este proceso.

Cadena	IEEE	EBSCO	Scopus
(“Prediction” OR “predicting” OR “forecasting”) AND (“student academic learning” OR “student learning style” OR “student achievement”)	14	34	54

("prediction" OR "forecast" OR "predicting") AND ("academic performance" OR "school performance" OR "student performance") AND ("academic learning" OR "learning style" OR "cultural identity")	6	1	13

Tabla 1. Cadenas de búsqueda.

2.2.4. Clasificación de artículos

Una vez realizada la búsqueda de las cadenas se procede a realizar la clasificación de los artículos que, de acuerdo a los criterios anteriormente mencionados, se decide hacer la lectura completa de 10 artículos para su respectivo análisis.

Para analizar cada trabajo seleccionado, se establece una tabla donde se especifican los objetivos, técnicas o metodologías propuestas y conclusiones de cada uno de estos 10 artículos seleccionados.

2.2.5. Extracción de datos

En esta sección se presentará un breve resumen de trabajos relacionados elaborados en instituciones de educación superior de diversos países, los cuales desarrollaron tipos de modelos de predicción para determinar el bajo rendimiento de los estudiantes que cursaban sus programas académicos, teniendo en cuenta diversos factores sociales, económicos, familiares, entre otros.

	Objetivo	Características	Técnicas	Conclusiones
[8]	Brindar una breve descripción de aspectos relacionados con el almacén de datos construido y algunos procesos de minería de datos desarrollados sobre el mismo	-Información personal -Académica -Demográfica -Socio-económica - Núcleo familiar	- Clustering (o agrupamiento de datos). -Cluster demográfico -Almacén de datos - Minería de datos - Data Mart	El uso de la técnica Data Mart en un Warehouse donde se tenían datos académicos, socio económicos y demográficos correspondientes a alumnos de la asignatura de Sistemas Operativos, permitió clasificar como Libre los estudiantes que no han completado la práctica y trabajos de la materia, regular los que han ejecutado las actividades correspondientes a la materia pero con un promedio inferior a 7 en la escala 0-10 y promoción, los estudiantes que tienen un promedio igual o superior a 7, dejando como resultado que la mayoría de las personas que hacen parte de la categoría Libre, eran solteros y una cantidad pequeña de divorciados, ellos estaban más dedicados a desarrollar sus actividades de manera correcta y

				oportuna y le daban mayor atención al estudio que a la diversión y aplicaban lo de estudiar para aprobar, los regulares son solteros y algunos en una unión consensual, según los resultados indicaban que era más importante el estudio que la diversión y el trabajo, los promocionados tomaban la tecnología como moda, en su mayoría eran solteros y unos pocos casados. Los 3 grupos tenían en común el contacto directo con las tecnologías de la comunicación (TIC), permitiendo ampliar sus conocimientos respecto a un tema, dichos resultados son producto de la técnica agrupamiento de datos la cual permitió definir las categorías con las cuales se iban a trabajar e identificar las variables comunes para determinar el nivel de rendimiento académico y la importancia de tener datos de calidad para evitar tomar tiempos extensos al momento de realizar la selección y limpieza de los mismos.
[18]	Realizar un modelo para predecir los estilos de aprendizaje de cada uno de los estudiantes.	- Demográfico -Socio-económico -Uso de redes sociales	-Modelos de clasificación discriminativo -Modelo de clasificación generativos. -Algoritmo C4.5 - NB - NBTree - CART	Con el análisis realizado, se pudo detectar que el algoritmo C4.5 es más preciso con un 96%, el NB con un 79%, NBtree con 91% y CART con 81%. Permitiendo la detección de estilo de aprendizaje preferido por los estudiantes, los resultados obtenidos fueron: “el 30% de los estudiantes pertenecen a LSD1, el 22.6% pertenece a LSD2, el 28.7% pertenece a LSD3 y 18.7 pertenecen a LSD4. El estilo de aprendizaje dominado fue activo, reflexivo, sensor, visual y secuencial que era más del 61% entre el grupo de muestra”.
[19]	Construir un modelo que permita por medio de Informes de estado de rendimiento, visualizar el estado del estudiante en cuanto a su rendimiento académico y brindar técnicas para mejorar.	- Marca de participación - Promedio académico.	-Software BenchMark -Análisis Envolvente de Datos (DEA) -El proceso de jerarquía analítica (AHP)	- El BenchMark no permiten retroalimentar a los estudiantes para mejorar su rendimiento académico, el DEA no retroalimenta de manera conjunta sino individual y el AHP no soporta gran cantidad de datos lo que no permitió clasificar los promedios de los alumnos.
[20]	“Analizar si se	-Calificaciones	- Regresión (RG)	

	pueden predecir las puntuaciones de los cursos, qué elementos o variables afectan a las predicciones y de qué manera es posible anticipar las puntuaciones".	-Tiempo dedicado al Mooc (Curso masivo abierto en línea) o al foro. - Interacción en los foros	- Máquinas vectoriales de soporte (SVM) -Árboles de decisión (DT) - Bosque aleatorio (RF).	Se ejecutaron cada uno de los algoritmos y al final se tomó que los modelos más asertivos son los de regresión y bosque aleatorio, porque se adaptaron al modo acumulativo y no acumulativo de datos, actividades, prácticas y semanas. "Por lo tanto, se deben poner esfuerzos en el diseño de modelos con nuevos indicadores en lugar de crear algoritmos".
[21]	Determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años	- Académico -Socio-económico	-Algoritmo C4.5 -Algoritmo Mate-tree - A priori - FPGrowth - EquipAsso	Es importante tener un almacén de datos robusto con información de calidad para facilitar la clasificación y agrupación de los datos. Para clasificar los estudiantes con bajo, medio o alto rendimiento académico y que tienden a desertar se tomó en cuenta la variable que almacenaba el promedio de cada uno de los estudiantes dejando como resultado de la investigación y del proceso es que 90% de los estudiantes con bajo rendimiento tienen características como que son provenientes de la zona sur del departamento de Nariño, son de estratos bajos menores de 18 años entre otros, y tienden a desertar en los primeros semestres de las carreras, ellos hacen parte de la facultad de la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas y normalmente no se reintegran al programa académico, mientras que los estudiantes desertores de la facultad de ingeniera que son el 50%, reingresan al programa semestres después. Este análisis fue realizado con el fin de identificar las causas de la deserción estudiantil para que la universidad tome medidas al respecto, teniendo en cuenta que tienen las herramientas como TariyKDD que facilitó el desarrollo de la investigación.
[22]	Identificar los perfiles de los estudiantes que tienen un alto rendimiento académico y los que tienen bajo rendimiento, haciendo uso de técnicas de	-Escuela media de procedencia. -Nivel educativo de los padres -Socio-económico -Edad -Género -Herramientas de apoyo (campus virtual	-Data-driven -Clasificación con Árboles de decisión	Utilizar la técnica Data-driven permitió tomar los datos específicos para la investigación y el análisis, tomando el tiempo y las dimensiones es una manera precisa de identificar las variables comunes y poder clasificar los niveles de rendimiento académico (Libre, regular y promocionado).

	minería de datos (MD) para detectar las variables causantes del bajo rendimiento académico y poder tomar medidas al respecto.			
[23]	Mostrar un estudio por medio de técnicas de minería de datos que permitan determinar, a través de un clasificador, el rendimiento académico de los alumnos ingresantes de la carrera de Licenciatura en Sistemas de Información de la Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste (FACENA-UNNE).	Socioeconómicos -Personales -Nivel académico -Promedio académico	Algoritmo clasificador Logistic	Se puede identificar que el uso del algoritmo clasificador Logistic, permitió evaluar y luego usar modelos de regresión logística múltiple, dejando como resultados de mediada precisión un error de clasificación de 36,024% y de clasificación efectiva 63,97%. El uso de software libre como Weka permite ejecutar un análisis descriptivo ya que muestra los resultados con gráficos haciendo mucho más comprensible los resultados.
[24]	Construir modelos predictivos del rendimiento académico de los estudiantes de las diversas carreras de la FACENA de la UNNE.	-Socioeconómicas -Conocimientos matemáticos previos	-Árboles de decisión - Modelo de regresión logística -Redes neuronales.	La minería de datos cumple un factor muy importante al momento de realizar el análisis de grandes cantidades de datos, junto a los métodos simbólicos y estadísticos, para este caso se logró determinar el perfil y los patrones de los estudiantes con bajo rendimiento de acuerdo a las variables socioeconómicas y conocimientos matemáticos previos, gracias a este modelo la universidad ejecutará estrategias para brindar solución a este caso y hacer que el rendimiento académico de los alumnos mejore con la calidad de educación que brinda la misma.
[25]	Predecir el rendimiento educativo de un	-Socioeconómicos - Personales -Laborales	-Árbol de decisión C&R. -Redes neuronales	Utilizar el árbol de decisión C&R basado en el algoritmo CART de Leo Breiman, permitió crear particiones

	alumno en su año de ingreso usando únicamente datos más relevantes anteriores a su entrada en la universidad	-Familiares -Educativos previos	-Regresión multivariante -Método regresión del Clementine	binarias con el fin de hacer que cada una de las ramas sea diferente y así poder discriminar de manera correcta cada uno de los grupos de datos. Todas las técnicas de minería de datos utilizadas en este estudio, permitieron identificar las características de un estudiante con tendencia a tener un bajo rendimiento académico, dicha identificación hará que la universidad mejores la calidad de educación y busque estrategias para que los alumnos mejores su rendimiento escolar.
[26]	Predecir el bajo rendimiento académico en la asignatura "Métodos y Diseños de Investigación en Psicología I", del área de Metodología, de 175 alumnos de primer semestre del programa de Psicología.	- Calificaciones -La asistencia -La participación en clase	-Regresión lineal -Regresión logística	De acuerdo al estudio realizado se logró identificar que con el modelo de regresión logística es posible identificar el éxito o fracaso académico de un alumno, en este caso el resultado fue del 70% de efectividad, mientras que el modelo de regresión lineal no hizo posible tomar un diagnostico confiable ya que sus resultados fueron del 0,17, haciendo imposible la predicción del rendimiento académico.

Tabla 2. Extracción de datos seleccionados.

2.3. ANÁLISIS DE LOS TRABAJOS RELACIONADOS

En [8] en identificar el rendimiento académico de los estudiantes que pertenecen a la asignatura Sistemas Operativos de la Licenciatura en Sistemas de Información y alumnos que hacían parte de la antes llamada Licenciatura en Sistemas de la Faculta de Ciencias Exactas y Naturales y Agrimensura (FACENA), con "el objetivo brindar una breve descripción de aspectos relacionados con el almacén de datos construido y algunos procesos de minería de datos desarrollados sobre el mismo". Dicha investigación utilizó Data Warehouse (DW) para la recolección de datos sobre información personal, académica, demográfica y socioeconómica de los alumnos y de su núcleo familiar y Data Mining (DM) basada en Clustering [8].

La mayoría de las personas que se encontraban con un rendimiento académico alto, estaban más dedicadas a desarrollar sus actividades de manera correcta y oportuna, además estaban en contacto directo con las tecnologías de la comunicación (TIC), permitiendo ampliar sus conocimientos respecto a un tema, dichos resultados son producto de la técnica agrupamiento de datos la cual permitió definir las categorías con las cuales se iban a trabajar e identificar las variables comunes para determinar el nivel de rendimiento académico y la importancia de tener datos de calidad para evitar tomar tiempos extensos al momento de realizar

la selección y limpieza de los mismos. [8]

En [18] describe la importancia y cómo influye el análisis de aprendizaje teniendo en cuenta el entorno en el que un estudiante aprende de manera eficaz con diversos materiales y ambientes tecnológicos, el cómo toman y transmiten la información y sus preferencias académicas. Para iniciar con la creación del modelo, se realizó una encuesta para la captura de datos demográficos, ya que es de suma importancia saber los estilos de aprendizaje de los estudiantes y el uso de redes sociales. Después se categorizaron de acuerdo a índice de estilos de aprendizaje (ILS) denominados como "LSD1 (detección – intuitiva), LSD2 (visual – verbal), LSD3 (activa– reflexivo) y LSD4 (secuencial-global)".

Esta información fue ingresada en un sistema de analítico con técnicas de "agrupación, clasificación, asociación y análisis de texto" en donde se indicó que el algoritmo C4.5 tiene un mayor porcentaje de precisión el cual fue del 96%. A futuro se espera que el sistema de predicción permita asesorar a los docentes para mejorar las metodologías empleadas al momento de transmitir los conocimientos a los alumnos como la plataforma de Moodle. [18]

En [19] se identifica la inasistencia de estudiantes en las clases al inicio de los semestres, causando problemas en el transcurso de mismo ya que no cuentan con la información o no tienen conocimiento de cómo potenciar su rendimiento académico y aprobar sus cursos. El objetivo principal de este proyecto es realizar el estudio y pruebas de una aplicación de software llamada benchMark, un método de programación matemática Envolvente de Datos (DEA) y un método matemático proceso de jerarquía analítica (AHP) de acuerdo a una variable conocida como p-Mark que significa "Marca de participación", factores y pesos, para que los estudiantes puedan observar cómo va el rendimiento académico y que mejoras pueden tomar para aumentar el mismo por medio de informes de estado de rendimiento (performance status reports (PSRs)). Se ejecutaron los procesos para cada uno de los modelos seleccionados sin obtener resultados satisfactorios al no mostrar la retroalimentación sobre las mejoras para potenciar el rendimiento académico de los alumnos. Por ello se inicia el proceso con un modelo lineal formado por dos algoritmos, basado en las evaluaciones que el docente prepara normalmente para ejecutarlas en el semestre y las calificaciones que el alumno debe tener, el cual deja resultados éxitos, cumpliendo con la retroalimentación esperada. "Algunos aspectos que actualmente están recibiendo atención para mejorar el algoritmo propuesto el enfoque incluye el cálculo de la mejora máxima posible por estudiante individual en una clase y la generación automática de los planes de perspectiva futura; y desarrollo de una interfaz gráfica interactiva y fácil de usar que combina los diferentes algoritmos ocupaciones". [19]

En [20], muestra cómo Massive Open Online Course (Mooc) ha impactado en la educación, pero se ha identificado que al inicio de los cursos se presentan deserciones por parte de los estudiantes. Este proyecto tiene como objetivo de "predecir si los estudiantes aprueban o no los cursos, qué elementos o variables

afectan a las predicciones y de qué manera es posible anticipar las puntuaciones" además de permitir conocer que mejoras se requieren para el curso de Introducción a la Programación con Java teniendo en cuenta los datos sobre foro, vídeos y calificaciones pasadas; y optimizar el material educativo para brindar educación de mayor calidad a los alumnos. Para el modelo de predicción se trabajó sobre los algoritmos Regresión, Máquinas vectoriales de soporte, Árboles de decisión y Bosque aleatorio, los resultados más favorables fueron con el algoritmo de bosque aleatorio, el cual se comportó de manera correcta en modo acumulativo y no acumulativo de acuerdo a las semanas, al tipo de prácticas y ejercicios ejecutados dentro del curso. Dado a la no disponibilidad de algunos datos, los resultados fueron satisfactorios para el caso de participación de los estudiantes en el foro, pero además sirvió para identificar la no participación de algunos.

A futuro se espera mejorar el modelo para ampliar el análisis y la predicción a otras asignaturas, teniendo en cuenta variables como metodología, asistencia y participación en las actividades propuestas en los Mooc. [20]

En [21] se describe la investigación denominada "Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos" con el propósito de "determinar en la comunidad universitaria perfiles de bajo rendimiento académico y deserción estudiantil aplicando técnicas de descubrimiento de conocimiento, a partir de los datos almacenados en las bases de datos durante los últimos 15 años". Para ese proceso se hizo uso de la herramienta libre llamada TariyKDD creada por los laboratorios de Descubrimiento de Conocimiento en Bases de Datos (DCBD) en el departamento de ingeniería. TariyKDD permite la recuperación, selección, limpieza y transformación de datos, gracias a la ejecución de filtros, minería de datos para la clasificación utilizó el algoritmo C4.5 y Mate-tree y para las reglas de asociación, el algoritmo Apriori, FPGrowth y EquipAsso dichos algoritmos se encuentran en la herramienta nombrada anteriormente, cabe resaltar que cuenta con una interfaz intuitiva la cual facilita manejo.

Los resultados de la investigación y el proceso de minería de datos demostraron que los estudiantes matriculados en el primer semestre en la facultad de Ciencias Naturales y Matemáticas o en la facultad de Ciencias Humanas, la mayoría de estratos bajos y provenientes del sur de Nariño hacen parte del grupo de los estudiantes que presentan un bajo rendimiento académico, además tienen características de los estudiantes que se retiran de los programas académicos. Es importante tener en cuenta que los que hacen parte de las facultades mencionadas anteriormente no se reintegran, mientras que casi la mayoría de los alumnos de la facultad de ingeniería que se retiran, se reintegran semestres después [27].

El bajo rendimiento académico de los estudiantes está entre el 60% y 80% de las materias del primer semestre entre ellas Algoritmos y Estructura de datos del programa de Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional Facultad Regional Resistencia (UTNFRR), ubicada en Argentina en la ciudad de Resistencia, provincia del Chaco. De acuerdo a esta situación se inició

un proceso investigativo el cual fue llamado “Perfiles de Rendimiento Académico: Un Modelo basado en Minería de Datos”, con el fin de identificar los perfiles de los estudiantes que tienen un alto rendimiento académico y los que tienen bajo rendimiento, haciendo uso de técnicas de minería de datos (MD) para detectar las variables causantes del bajo rendimiento académico y poder tomar medidas al respecto.

Para iniciar el proceso se trabajó con la base de datos institucional y de la cátedra, donde se almacenan las calificaciones de los exámenes parciales de los estudiantes, su condición al finalizar y se realizó una encuesta online en la cual se tuvieron en cuenta las siguientes variables “escuela media de procedencia, nivel educativo de los padres, nivel socio-económico, edad, género, actitud general hacia el estudio, existencia del cursillo de ingreso, régimen de cursado (anual - cuatrimestral), uso de herramientas de apoyo (campus virtual)”, los datos de las calificaciones de las evaluaciones realizadas para clasificar los estudiantes con rendimiento académico bajo, medio y alto. Se realizó el pre procesamiento adecuado de la información recolectada haciendo una limpieza de datos inconsistentes y faltantes.

Para determinar los perfiles (bajo, medio y alto) de los estudiantes se creó el almacén de datos utilizando la técnica Data-driven la cual permite manejar datos de acuerdo a las necesidades del usuario, tomando un modelo de datos que consiste en hechos (tiempo) y dimensiones (estructura básica del diseño).

Figura 1. Modelo del Data Warehouse [22].

Para Data Mining se utilizó la técnica supervisada de Clasificación con Árboles de decisión con el fin de clasificar los datos comunes y poder detectar con facilidad los datos nuevos o desconocidos durante el proceso.

Los resultados obtenidos se basaron en varios factores como la escuela media de procedencia, nivel educativo de los padres, socioeconómico, edad, género, herramientas de apoyo (campus virtual), de acuerdo a la encuesta realizada inicialmente y a la técnica data-driven usada, se tuvo en cuenta el estado final de la materia de cada uno de los estudiantes, determinando cuantos parciales fueron ganados con una calificación superior, de acuerdo a este análisis se concluyó que 81.42% de alumnos tienen bajo rendimiento denominados en situación Libre, 10.62% rendimiento académico medio denominado Regular y sólo 7.96% rendimiento académico alto denominado Promocionado [22].

En Facultad de Ciencias Exactas de la Universidad Nacional del Nordeste en Argentina, se llevó a cabo el desarrollo del proyecto investigativo nombrado como “Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE” de la carrera de Licenciatura en Sistemas de Información, el cual consiste en determinar el bajo rendimiento de un estudiantes recién ingresado al programa académico, basado en datos almacenados en un almacén de datos sobre información personal y socioeconómica de los estudiantes, materias matrículas y el promedio de cada una de estas. Se empleará la minería de datos utilizando el algoritmo clasificador Logistic en la aplicación libre Weka (Waikato Environment for Knowledge Analysis). Obteniendo resultados con mediada precisión con un error de clasificación de 36,024% y de clasificación efectiva 63,97% [23].

Además, de este estudio se desarrolló otro proyecto investigativo llamado “Modelos predictivos y técnicas de minería de datos para la identificación de factores asociados al rendimiento académico de alumnos universitarios” el cual permitió conocer los estudiantes de distintas carreras que tienden a tener un bajo rendimiento académico de acuerdo a las variables socioeconómicas y/o de conocimientos matemáticos previos que se tienen en cuenta al momento de realizar el proceso de predicción, esto es con el objetivo de identificar de manera temprana los alumnos que están más propensos a disminuir su rendimiento escolar y así plantear y ejecutar estrategias para mitigar este caso. En el desarrollo de este modelo de predicción fue utilizado el método de regresión logística porque facilita clasificar los datos por sus características, los Árboles de Decisión para la identificación de espacios de los datos categorizados y numéricos, y las Redes Neuronales como herramienta que deja analizar y modelar de una manera más

sencilla la relación funcional que hay en las variables con las que se está trabajando [24].

En la Universidad Politécnica de Valencia (UPV) para los programas educativos o titulaciones de la facultad de Informática, se elaboró un proyecto de investigación nombrado como “Análisis del rendimiento académico en los estudios de informática de la Universidad Politécnica de Valencia aplicando técnicas de minería de datos”, tomando a los estudiantes recién ingresados a los programas como población a analizar, aplicando técnicas de minería de datos como árboles de decisión (árbol C&R), redes neuronales y la regresión multivariante (un método estadístico clásico), con el objetivo de predecir el rendimiento educativo de un alumno en su año de ingreso usando únicamente datos más relevantes anteriores a su entrada en la universidad como los son socioeconómicos, familiares, edad, estudios previos, entorno al inicio de sus estudios, actividad laboral durante los estudios, planes de estudios, métodos evaluativos [25].

En España la Universidad Complutense de Madrid, elaboró un proyecto académico “La predicción del rendimiento académico: regresión lineal versus regresión logística” con el objetivo de predecir el bajo rendimiento académico en la asignatura “Métodos y Diseños de Investigación en Psicología I”, del área de Metodología, de 175 alumnos de primer semestre del programa de Psicología. Usando y comparando el modelo de regresión lineal para predecir el rendimiento académico y regresión logística para identificar el éxito o fracaso escolar con variables como, calificaciones, la asistencia, la participación en clase y datos tomados de una encuesta realizada en donde se hicieron preguntas sobre cómo se desenvuelven en clase y cuáles son sus aptitudes y expectativas frente a la carrera. Los resultados de este proceso ayudaron a identificar que el modelo de regresión logística fue más acertado casi en un 70% al momento de predecir el fracaso o éxito académico de un alumno [26].

Con los trabajos relacionados que se presentaron en esta sección se puede concluir que las variables tomadas para ejecutar los modelos de predicción son similares como las variables socioeconómicas y personales, haciendo que cada proyecto investigativo muestre resultados relativamente similares, pero con diferentes técnicas de minería de datos, matemáticos y estadísticos; este proyecto tendrá en cuenta la cultura y el estilo de aprendizaje ya que es de suma importancia las metodologías que los docentes usan para transmitir conocimiento a sus estudiantes, de acuerdo a cada estilo de aprendizaje será un mecanismo para que la mente del alumno procese y asimile información además, tenga mayor percepción sobre las temáticas de los programas académicos de cada universidad [28].

CAPÍTULO III. CREACIÓN DEL MODELO

En este capítulo se presentará la metodología que se utilizó para continuar con el proceso de minería de datos de la presente propuesta de investigación, el cual se deriva en aplicar el modelo de referencia CRISP-DM para la construcción y evaluación de los modelos establecidos en el estado del arte [29][13], que consta de seis fases [30], especificadas a continuación:

3.1 Fase 1: Comprensión del negocio:

En esta fase se define las tareas del proceso de minería de datos, cuyo propósito es determinar los objetivos para crear un modelo utilizando técnicas de minería de datos, que permita predecir el rendimiento académico en los estudiantes, de acuerdo a la información que se obtuvo en la realización del estado del arte. De manera que el modelo de predicción contribuye en la prevención de la deserción universitaria.

3.2 Fase 2: Comprensión de los datos:

Una vez comprendido el negocio se procede a la recolección de los datos que se utilizaron para el desarrollo del proyecto de investigación, para esto el programa de psicología de la Fundación Universitaria de Popayán implementó una encuesta con el fin de obtener información sociocultural, socioeconómica como también información personal de los estudiantes, todas las respuestas fueron exportadas a un archivo de Excel con el objetivo de poder manipularlos con herramientas de minería de datos, en total se obtuvieron 170 registros de los estudiantes de primer semestre del presente año lectivo.

Dicha información se distribuye en 24 atributos de los cuales 18 son categóricos y 6 son numéricos, con el atributo de Promedio general se puede clasificar el nivel de rendimiento académico de los estudiantes, clasificados a continuación:

CLASIFICACIÓN PROMEDIO ACADÉMICO	
Bajo	Promedio académico inferior a 3.0.
Medio	Promedio académico entre 3.0 a 3.4.
Alto	Promedio académico superior a 3.5.

Tabla 3. Clasificación de promedio académico.

A continuación, se detallan los atributos de la encuesta el cual están en el archivo de Excel.

Ítem	Atributo	Clasificación
1	Semestre	Semestre académico actual del estudiante.
2	Código	Código estudiantil.
3	Edad	Valor numérico.
4	Género	Femenino, Masculino
5	Ocupación	Estudiante, Trabajador medio tiempo, Trabajador tiempo completo.
6	Etnia	Blanco, Mestizo, Indígena, Afro descendiente.
7	Tipo Vivienda	Arrendada, Familiar, Propia, Otro.
8	Estrato	Valor numérico 1 a 6.
9	Estado civil	Casado, Soltero, Unión libre, Viudo.
10	Nivel escolar	Secundaria, Técnico, Universitario.
11	Jornada académica	Diurno, Nocturno, Sabatino.
12	Acudiente	Si, No.
13	Departamento nacimiento	Departamentos Colombia.
14	Municipio nacimiento	Municipios Colombia.
15	Puntaje ICFES	Valor numérico.
16	Promedio general	Valor numérico.
17	Asignaturas preferidas	Si, No.
18	Lógica matemática	Si, No.
19	Neuroanatomía	Si, No.
20	Historia fundamentos	Si, No.
21	Epistemología	Si, No.
22	Actividad formativa	Si, No.

23	Proyecto vida	Si, No.
24	Electiva sociohumanística	Si, No.

Tabla 4. Descripción de atributos.

3.3 Fase 3: Preparación de los datos:

En esta fase se procedió a utilizar la herramienta de Python Notebooks para realizar las tareas de reprocesamiento con el fin de hacer uso de la librería de Scikit-learn para el análisis de minería de datos en la información que se encuentra agregada en el archivo CSV de Excel.

A continuación, se procede a cargar los datos obtenidos en la encuesta y que se encuentran consolidados en el archivo de Excel para posteriormente trabajar con ellos en la herramienta de Python.

Este formato se encuentra alojado en un repositorio de la plataforma de Github como se muestra en la figura 2.
https://github.com/Cristian-Piamba/Seminario.git

SEMESTRE	FECHA DE NACIMIENTO	LUGAR DE NACIMIENTO	ID	EDAD	GENERO	OCUPACION DEL ENCUESTADO	ETNIA	TIPO DE VIVIENDA	ESTRATO SOCIOECONOMIC
1	09/11/1996	ARMENIA-QUINDIO	731292251	21	MASCULINO	T. TIEMPO COMPLETO	MESTIZO(A)	ARRENDADA	3
1	10/01/1999	POPAYAN-CAUCA	73292263	23	FEMENINO	T. TIEMPO COMPLETO	MESTIZO(A)	PROPIA	2
1	30/11/1990	LA VEGA-CAUCA		35	FEMENINO	T. TIEMPO COMPLETO	INDIO(A)	FAMILIAR	
1		BALBOA-CAUCA		53	FEMENINO	T. TIEMPO COMPLETO	MESTIZO(A)	PROPIA	4
1		TAMBO-CAUCA	73292265	26	FEMENINO	ESTUDIANTE	MESTIZO(A)	ARRENDADA	2
1	02/05/1989	POPAYAN-CAUCA	73292227	29	FEMENINO	ESTUDIANTE	MESTIZO(A)	FAMILIAR	2
1	21/04/1997	POPAYAN-CAUCA	73292470	21	FEMENINO	T. TIEMPO COMPLETO		FAMILIAR	3
1	02/10/1997	POPAYAN-CAUCA	73292284	21	FEMENINO	T. TIEMPO COMPLETO	MESTIZO(A)	PROPIA	1
1	08/04/2000	POPAYAN-CAUCA		17	FEMENINO		MESTIZO(A)	PROPIA	2

Figura 2. Datos en repositorio Github.

Se realizaron actividades de transformación con el fin de obtener datos limpios, en otras palabras, se buscó evitar valores en blanco, erróneos o incomprensibles que imposibilitaron analizar de manera óptima la información, para esto se hizo uso de la herramienta Open Refine anteriormente llamado Google Refine, que permite la carga de archivos de Excel para posteriormente trabajar con ellos ordenando y limpiando los datos.

Figura 3. Plataforma Open Refine.

Se procede a cargar el archivo con toda la información de los estudiantes en la herramienta Open Refine para que a continuación se adecuen al trabajo, en ella se ejecutan las adecuaciones pertinentes para exportar el nuevo archivo y cargarlo en Python con Google Colaboratory.

En Google Colaboratory se importan las librerías necesarias para realizar las tareas requeridas e ir efectuando el modelamiento de los datos.

```
import pandas as pd
import numpy as np
import math
import matplotlib.pyplot as plt

from scipy.stats import pearsonr
from sklearn import linear_model
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier

from math import e
from sklearn.linear_model import LogisticRegression
%matplotlib inline

import seaborn as sns
```

Figura 4. Importación de librerías en Python.

En el repositorio donde se encuentran albergados los archivos Dataset se prosigue a llamar los datos guardados en la plataforma Github, por medio de la URL con el siguiente comando.

```
# Se procedio llamar los datos guardados en la plataforma GIT-HUB
dataset = pd.read_csv("https://raw.githubusercontent.com/Cristian-Piamba/Seminario/master/estudiantesDataset2.csv",
error_bad_lines=False)
dataset.head()
```

Figura 5. URL datos en Github.

Se carga el nuevo conjunto de datos con las modificaciones previamente realizadas, donde los datos categóricos se convirtieron a numéricos, así mismo se realizó el renombramiento de los atributos y corrección de los mismos para mejorar el acceso a estos.

Los atributos que se encuentran en el Dataset fueron codificados de la siguiente manera:

CODIFICACIÓN DE ATRIBUTOS	
GÉNERO	
Masculino	1
Femenino	0
OCUPACIÓN ENCUESTADO	
Estudiante	0
Trabajador medio tiempo	1
Trabajador tiempo completo	2
ETNIA	
Blanco	0
Mestizo	1
Indio	2

Afro	3
TIPO VIVIENDA	
Arrendada	0
Familiar	1
Propia	2
Otro	3
ESTADO CIVIL	
Casado(a)	0
Soltero(a)	1
Unión libre	2
Viudo(a)	3
JORNADA ACADÉMICA	
Diurno	0
Nocturno	1
Sabatino	2
ACUDIENTE	
Si	1
No	0
DEPARTAMENTO NACIMIENTO	

Cauca	0
Putumayo	1
Nariño	2
Valle del Cauca	3
Quindío	4
Atlántico	5
Cundinamarca	6
Huila	7
Risaralda	8
Santander	9

MUNICIPIO DE NACIMIENTO

Popayán	0	Cali	1	La Plata	2
La Vega	3	San Miguel	4	El Tambo	5
Timbío	6	Tumaco	7	Balboa	8
Barrancabermeja	9	Belalcázar	10	Bogotá	11
Bolívar	12	Cartagena	13	Hormiga	14
La Unión	15	Mocoa	16	Pereira	17
Quindío	18	Rosas	19	San Juan	20
San Sebastián	21	Sucre	22	Silvia	23

ASIGNATURAS PREFERIDAS	
Si	1
No	0
LÓGICA MATEMÁTICA	
Si	1
No	0
NEUROANATOMÍA	
Si	1
No	0
HISTORIA Y FUNDAMENTOS	
Si	1
No	0
EPISTEMOLOGÍA	
Si	1
No	0
ACTIVIDAD FORMATIVA	
Si	1
No	0
PROYECTO DE VIDA	
Si	1

No	0
ELECTIVA SOCIOHUMANÍSTICA	
Si	1
No	0
RENDIMIENTO ACADÉMICO	
Bajo	0
Medio	1
Alto	2

Tabla 5. Codificación de atributos.

Como resultado se obtiene el nuevo conjunto de datos, con el cual se trabajó para el modelamiento, este se encuentra en el repositorio de Github para posteriormente importarlo a la plataforma de trabajo.

	Semestre	Codigo	Edad	Genero	Ocupacion	Etnia	Tipo_vivienda	Estrato	Estado_civil	Nivel_escolar	Jornada_academica	Acudiente
0	1	73292251	21	1	2	1	0	3	0	1	0	1
1	1	73292263	23	0	2	1	2	2	0	1	0	1
2	1	73292258	35	0	2	2	1	3	0	0	0	0
3	1	73292301	53	0	2	1	2	4	1	2	0	1
4	1	73292265	26	0	0	1	0	2	0	1	0	0

Figura 6. Conjunto de datos procesado.

3.4. Fase 4. Modelado:

Conforme a las características del trabajo, en esta fase se realizaron diferentes actividades con el fin de elegir uno de los modelos de predicción que fueron encontrados en la revisión del estado del arte para que más adelante fueran evaluados. Cada actividad se puso en práctica con el propósito de determinar qué modelo presentaba una mayor precisión al momento de predecir el rendimiento académico de los estudiantes.

La primera actividad fue verificar que los atributos presentaran el tipo de dato entero para evitar que se generaran errores al momento de implementar las métricas, en caso contrario de que los atributos presentaron otro tipo de dato se efectuaría una conversión, en este caso los atributos presentaron el tipo de dato deseado.

```
dataset.dtypes
```

```
Semestre                    int64
Codigo                      int64
Edad                        int64
Genero                      int64
Ocupacion                   int64
Etnia                       int64
Tipo_vivienda               int64
Estrato                     int64
Estado_civil                int64
Nivel_escolar               int64
Jornada_academica           int64
Acudiente                   int64
Departamento_nacimiento     int64
Municipio_nacimiento        int64
Puntaje_icfes               int64
Promedio_general            int64
Asignaturas_preferidas      int64
Logica_matematica           int64
Neuroanatomia               int64
Historia_fundamentos        int64
Epistemologia               int64
Actividad_formativa         int64
Proyecto_vida               int64
Electiva_sociohumanistica   int64
dtype: object
```

Figura 7. Verificación del tipo de dato de los atributos.

Como segunda actividad se procede a comprobar que los datos que se encuentran en el atributo de promedio general se encuentren balanceados.

```
print(dataset.groupby('Promedio_general').size())
```

```
Promedio_general
0     1
1     9
2    46
dtype: int64
```

```
# Gráfico de barras de Promedio General
plot = dataset['Promedio_general'].value_counts().plot(kind='bar', title='Promedio General')
```

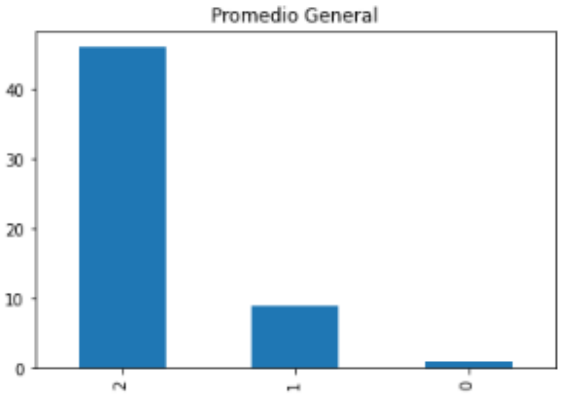


Figura 8. Balanceo de datos.

Continuando con las actividades se analizaron los atributos por medio de graficas con el objetivo de identificar cuál de estos presentaban baja influencia de respuesta, después fueron eliminadas con el propósito de que el modelo pudiera presentar una mayor efectividad.

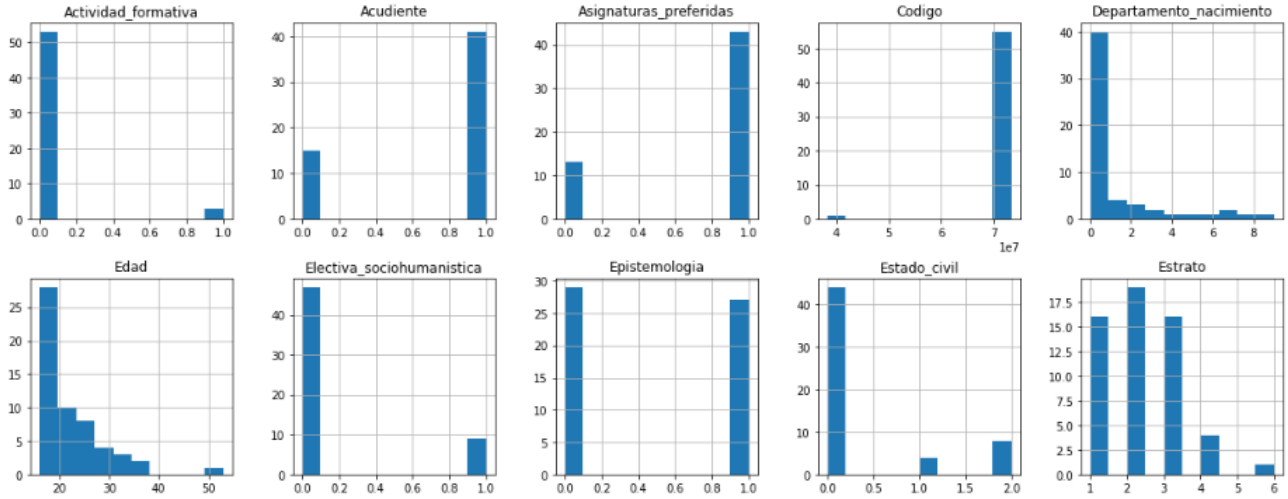


Figura 9. Histograma atributos 1.

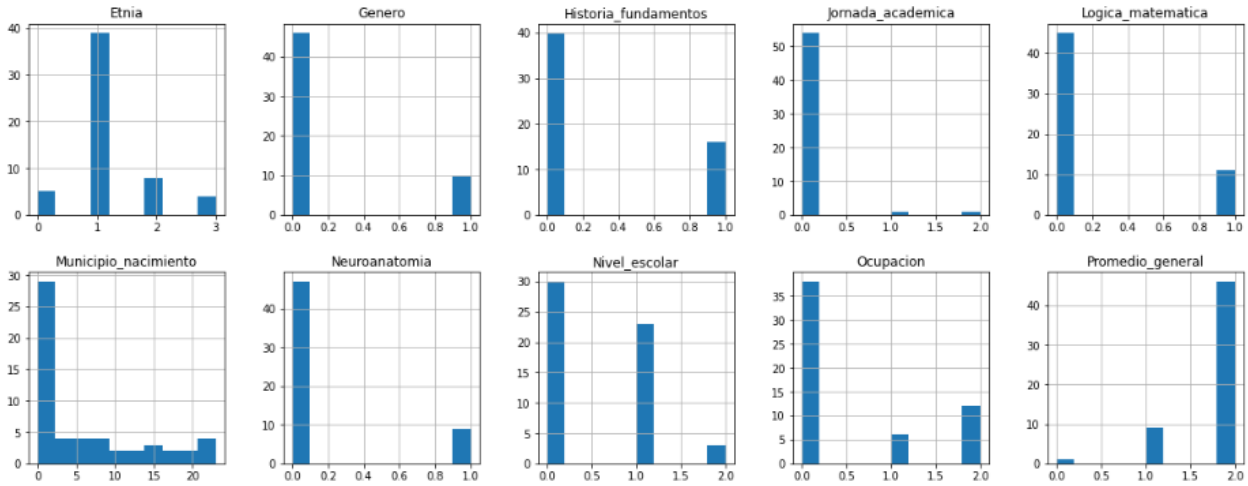


Figura 10. Histograma atributos 2.

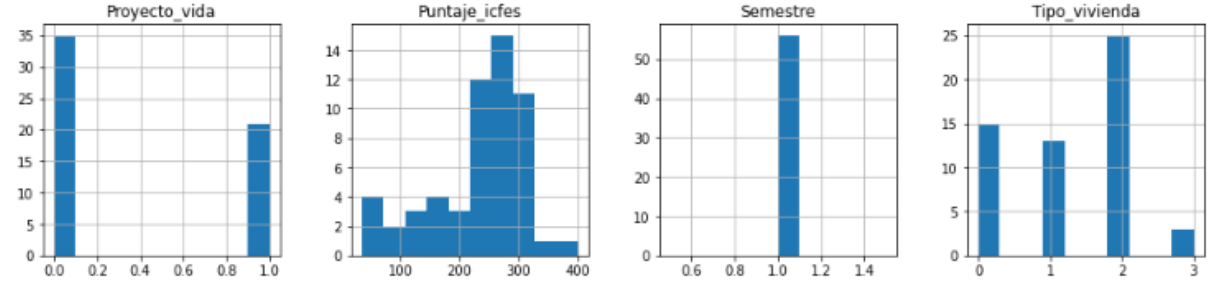


Figura 11. Histograma atributos 3.

De acuerdo a los anteriores resultados que arrojaron las gráficas se logró evidenciar que unos de los atributos no son de gran relevancia para el modelo, es decir, los atributos que no aporten mayor información al modelo serán descartados.

Enseguida se crea un gráfico de correlación con el objetivo de identificar las variables que presentan mayor similitud entre ellas y que ostenten alta intensidad a la variable dependiente promedio general para proceder a eliminarlas ya que pueden afectar los resultados de los modelos que serán evaluados más adelante.

```
def calcular_pvalue(dataset):
    dataset = dataset.dropna()._get_numeric_data()
    datasetcol = pd.DataFrame(columns=dataset.columns)
    pvalues = datasetcol.transpose().join(datasetcol, how='outer')
    for r in dataset.columns:
        for c in dataset.columns:
            pvalues[r][c] = round(pearsonr(dataset[r], dataset[c])[1], 4)
    return pvalues

pval = calcular_pvalue(dataset)
pval

# Vemos la correlacion y su significancia
rho = dataset.corr()
rho = rho.round(3)

# create three masks
r1 = rho.applymap(lambda x: '{}*'.format(x))
r2 = rho.applymap(lambda x: '{}**'.format(x))
r3 = rho.applymap(lambda x: '{}***'.format(x))

# apply them where appropriate
rho = rho.mask(pval<=0.1,r1)
rho = rho.mask(pval<=0.05,r2)
rho = rho.mask(pval<=0.01,r3)
rho
```

Figura 12. Codificación gráfica de correlación.

	Codigo	Edad	Genero	Ocupacion	Etnia	Tipo_vivienda	Estrato	Estado_civil	Nivel_escolar	Acudiente
Codigo	1.0***	-0.116	0.063	-0.24*	0.233*	-0.251*	-0.101	-0.12	0.117	-0.081
Edad	-0.116	1.0***	0.014	0.459***	0.018	-0.003	0.201	0.273**	0.447***	-0.23*
Genero	0.063	0.014	1.0***	0.15	0.272**	-0.246*	0.083	-0.232*	-0.014	0.071
Ocupacion	-0.24*	0.459***	0.15	1.0***	-0.028	0.104	0.033	0.099	0.416***	0.002
Etnia	0.233*	0.018	0.272**	-0.028	1.0***	-0.004	-0.255*	-0.213	0.056	-0.003
Tipo_vivienda	-0.251*	-0.003	-0.246*	0.104	-0.004	1.0***	-0.008	0.062	0.023	0.275**
Estrato	-0.101	0.201	0.083	0.033	-0.255*	-0.008	1.0***	0.112	-0.149	0.047
Estado_civil	-0.12	0.273**	-0.232*	0.099	-0.213	0.062	0.112	1.0***	-0.057	-0.036
Nivel_escolar	0.117	0.447***	-0.014	0.416***	0.056	0.023	-0.149	-0.057	1.0***	0.119
Acudiente	-0.081	-0.23*	0.071	0.002	-0.003	0.275**	0.047	-0.036	0.119	1.0***
Departamento_nacimiento	0.066	-0.04	0.412***	0.065	0.272**	-0.092	0.202	-0.134	0.038	-0.043
Municipio_nacimiento	0.106	-0.04	0.211	-0.021	0.253*	-0.257*	-0.032	-0.063	-0.022	0.03
Puntaje_icfes	-0.149	-0.228*	-0.24*	0.009	0.049	0.033	0.017	0.177	-0.181	0.021
Promedio_general	-0.06	0.095	0.208	-0.104	0.127	-0.082	0.324**	-0.061	-0.089	-0.087
Asignaturas_preferidas	0.245*	0.006	0.146	-0.002	0.095	-0.197	0.072	-0.139	0.335**	0.145
Logica_matematica	0.067	-0.016	0.004	0.224*	0.119	0.091	-0.058	-0.058	0.173	-0.107
Neuroanatomia	0.059	0.195	0.177	-0.049	0.227*	-0.03	0.143	-0.15	0.19	-0.065
Historia_fundamentos	0.085	-0.011	0.118	-0.172	0.049	-0.11	0.135	-0.094	-0.019	-0.153
Epistemologia	0.13	0.01	0.017	0.067	-0.067	0.011	0.178	-0.132	0.24*	0.261*
Actividad_formativa	0.032	-0.18	-0.111	-0.059	-0.068	-0.16	-0.2	-0.118	0.059	0.144
Proyecto_vida	0.104	-0.356***	0.024	-0.056	0.1	-0.04	-0.053	-0.077	0.069	0.302**
Electiva_sociohumanistica	0.059	-0.176	-0.077	-0.167	-0.194	-0.189	0.282**	0.189	-0.054	0.045

Figura 13. Captura gráfica de correlación.

```
dataset.drop(['Ocupacion','Puntaje_icfes','Municipio_nacimiento'], axis=1, inplace=True)

dataset.drop(['Acudiente','Tipo_vivienda'], axis=1, inplace=True)

dataset.drop(['Estado_civil'], axis=1, inplace=True )

dataset.drop(['Codigo','Actividad_formativa'], axis=1, inplace=True)


dataset.drop(['Jornada_academica','Semestre'], axis=1 ,inplace=True)
```

Figura 14. Eliminación de los atributos.

Con la gráfica de colores se indican las variables independientes que aún tienen mayor correlación con la variable dependiente, esto para conseguir que los modelos evaluados se puedan optimizar y que sean más precisas sus predicciones.

Se puede indicar que el coeficiente de correlación oscila entre los valores de -1.00 y 1.00, donde el color azul muestra la correlación positiva que se extiende del 0.00 al 1.00 y el color naranja muestra la correlación negativa que se extiende del 0.00

al -1.00, clasificando a 0.00 como un valor neutro. De esta manera se escogieron los atributos con mayor correlación.

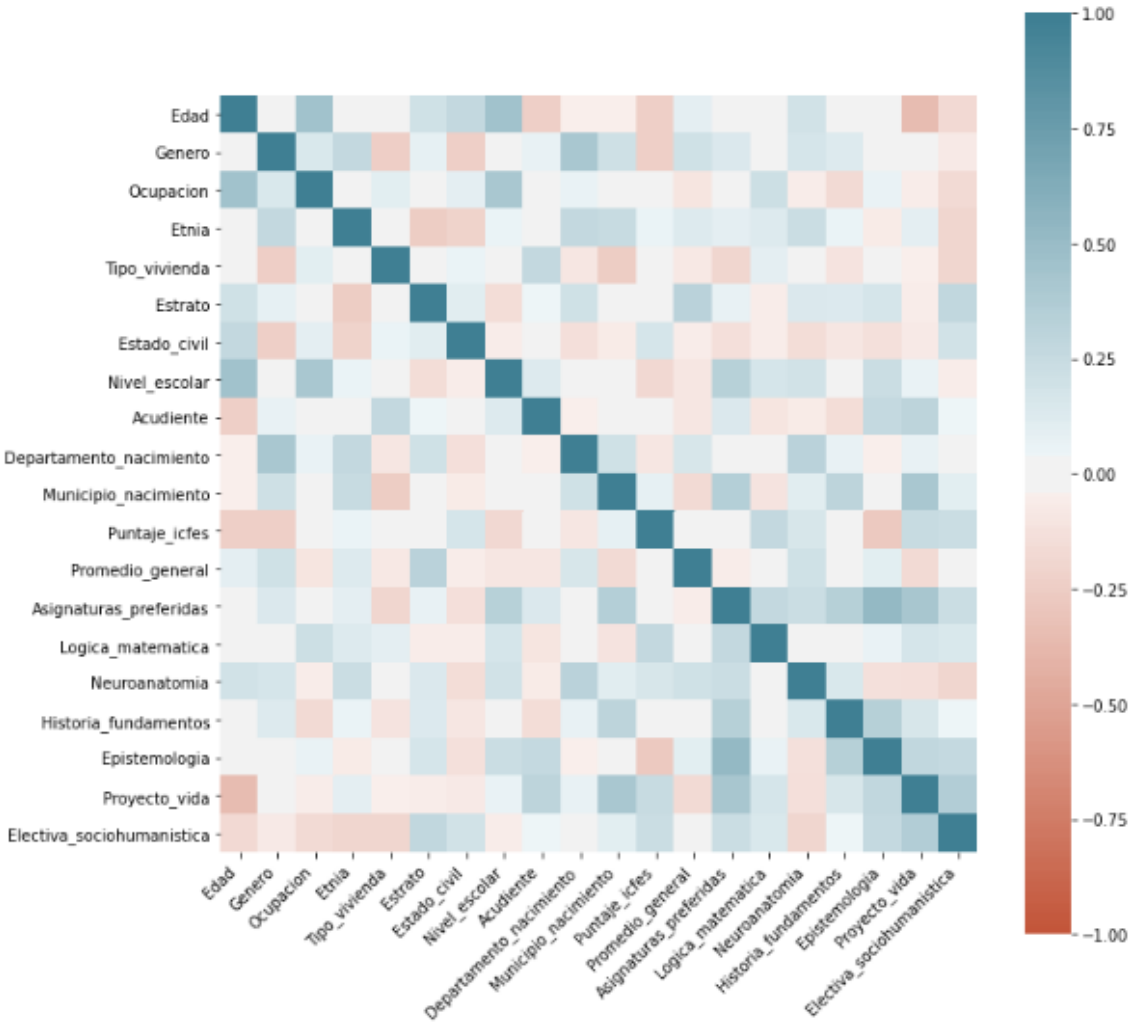


Figura 15. Mapa de calor.

Al analizar la gráfica de colores se observa la correlación que existe entre algunos de los atributos, es decir, el grado de asociación que presentan los atributos.

3.5. Fase 5. Evaluación:

Para esta fase de evaluación se generalizaron los atributos que presentaron mayor influencia con el caso de estudio de este proyecto de investigación, elegidos en la fase anterior para ser evaluados por los modelos que se escogieron en el estado del arte, teniendo en cuenta que el beneficio de este proyecto de investigación es determinar la clasificación de los estudiantes en los rangos de promedio académico.

Dando continuidad se realizó la selección de los atributos elegidos para evaluar el modelo.

```
X = dataset.iloc[:,dataset.columns != 'Promedio_general']  # columnas independientes
y = dataset.Promedio_general   # Columna objetivo

model = ExtraTreesClassifier()
model.fit(X,y)
#print(model.feature_importances_) #construye inbuilt class feature_importances de tree based classifiers
#grafica feature importances for better visualization
feat_importances = pd.Series(model.feature_importances_, index=X.columns)
feat_importances.nlargest(4).plot(kind='barh')
plt.show()
```

Figura 16. Selección de datos para el modelado.

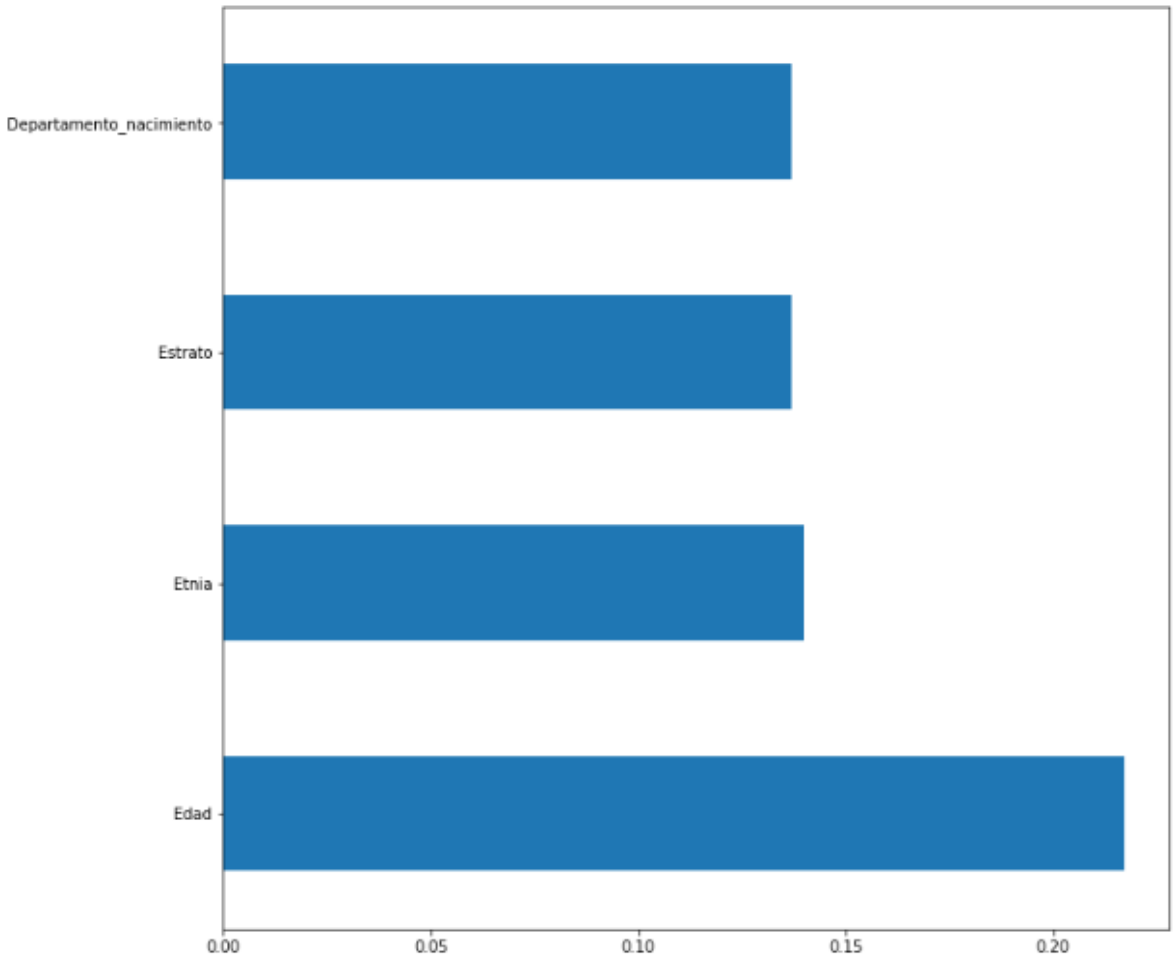


Figura 17. Diagrama de atributos importantes.

Con los atributos de la gráfica anterior se prueba su funcionamiento con los modelos de predicción.

```
feature_names = ['Departamento_nacimiento','Estrato','Etnia','Edad']

X = dataset[feature_names]
Y = dataset['Promedio_general']

dataframe = dataset

array = dataframe.values
# prepare configuration for cross validation test harness
seed = 7
# Preparar los modelos
models = []
models.append(('LR', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
models.append(('RF', RandomForestClassifier()))
#models.append(('SVM', SVC()))
# evaluate each model in turn
results = []
names = []
scoring = 'accuracy'
for name, model in models:
  kfold = model_selection.KFold(n_splits=10, random_state=seed)
  cv_results = model_selection.cross_val_score(model, X, Y, cv=kfold, scoring=scoring)
  results.append(cv_results)
  names.append(name)
  msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
  print(msg)
# Graficamos los resultados de evaluacion de los modelos
fig = plt.figure()
fig.suptitle('COMPARACION DE ALGORITMOS')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

Figura 18. Evaluación de los modelos de predicción.

La precisión que cada modelo efectuó fueron comparadas, donde se pueden observar los resultados de cada uno, de esta manera se compara y se elige el mejor. En cuanto a los resultados arrojados en este caso se escoge el algoritmo de LDA (Linear Discriminant Analysis), ya que obtuvo una presión de 0.82, valor superior a los resultados de los otros modelos que fueron evaluados y posteriormente comparados.

```
LR: 0.803333 (0.207338)
LDA: 0.823333 (0.215536)
KNN: 0.773333 (0.238421)
CART: 0.753333 (0.200666)
NB: 0.630000 (0.246509)
RF: 0.770000 (0.169607)
```

Figura 19. Precisión de modelos.

Damos inicio al entrenamiento de los datos con el algoritmo seleccionado para determinar la exactitud y la precisión que el modelo maneja.

```
model = LinearDiscriminantAnalysis()
model.fit(X,y)
print("precision de predicciones = ",model.score(X,Y))
validation_size = 0.40
from sklearn.model_selection import train_test_split

x_train,x_validation,y_train,y_validation = model_selection.train_test_split(X,Y, test_size = validation_size )
nombre= "Analisis Discriminante Lineal (LDA): "
muestra = model_selection.KFold(n_splits=10, random_state=seed)
cv_results = model_selection.cross_val_score(model,x_train, y_train, cv=muestra, scoring= 'accuracy')


predictions = model.predict(x_validation)
print("exactitud de las prediciones : ", accuracy_score(y_validation,predictions))
```

Figura 20. Entrenamiento del modelo.

```
precision de predicciones =  0.8214285714285714
exactitud de las prediciones :  0.9130434782608695
```

Figura 21. Precisión y exactitud del modelo entrenado.

Para finalizar esta etapa de modelamiento se mostrará como resultado el modelo de predicción con los datos reales, el cual presenta el rendimiento académico del estudiante y el desempeño obtenido del modelo a la hora de ser ejecutado.

De igual manera el modelo carga el conjunto de datos que se encuentra en el repositorio de Github y tiene establecidos los atributos que se tuvieron en cuenta para la predicción.

```
import pandas
from sklearn import model_selection
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn import metrics
import pandas as pd
from sklearn.metrics import classification_report
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis

url = "https://raw.githubusercontent.com/Cristian-Piamba/Seminario/master/estudiantesDataset2.csv"
atributos =['Departamento_nacimiento','Estrato','Etnia','Edad']
dataframe = pandas.read_csv(url)
array = dataframe.values
X = dataframe[atributos]
Y = array[:,7]
test_size = 0.9
seed = 7
validation_size = 0.60
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X,Y, test_size = validation_size )
# Fit the model on training set
model = LinearDiscriminantAnalysis()
model.fit(X, Y)

result = model.score(X_test, Y_test)

datos = pd.DataFrame({'Departamento_nacimiento':[1],'Estrato':[2],'Etnia':[2],'Edad':[27]})
print("Rendimiento academico: ", int(model.predict(datos)), "Puntualidad: ",round(result*100),"%")
```

```
Rendimiento academico:  2 Puntualidad:  71.0 %
```

Figura 22. Modelo predictivo propuesta de investigación.

3.6. Fase 6. Despliegue:

En esta etapa se da por terminado el desarrollo de la metodología, donde se incluye un modelo entrenado de minería de datos que tiene como objetivo predecir el rendimiento académico de los estudiantes de psicología de la Fundación Universitaria de Popayán, el cual puede definirse como desempeño alto, medio o bajo.

Para facilitar el registro de la propuesta de investigación y como estrategia de trabajo, los archivos se encuentran almacenados en el repositorio de Github y todas las pruebas fueron realizadas en Google Colaboratory, bajo el lenguaje de programación Python, y usando la librería Scikit-learn.

Repositorio Github:
https://github.com/Cristian-Piamba/Seminario.git
Modelo depositado en Google Colaboratory:
https://colab.research.google.com/drive/14j52EvEm1jyqghQ4R5_nLRvdqtxkcVa_#scrollTo=h4eBHJqanf4A

CAPÍTULO IV. COMUNICACIÓN DEL MODELO

En este capítulo se presenta la técnica de trabajo que se empleó para la comunicación del modelo de predicción desarrollado en el presente proyecto de investigación

Para la socialización del modelo de predicción al área de bienestar institucional de la Fundación Universitaria de Popayán, se organizó una reunión de manera virtual por medio de la aplicación Google Meet ya que en el desarrollo del modelo se presentó la emergencia sanitaria frente a la pandemia del COVID-19, esta reunión con el fin de presentar y comunicar los resultados obtenidos en las pruebas que se realizaron.

A continuación, se da a conocer la presentación en PowerPoint utilizada durante la socialización del modelo de predicción. Para la generación de la presentación se tomaron aspectos importantes para dar a conocer el modelo, ya que por cuestiones de tiempo no era posible alargar la socialización.

Figura 23. Inicio presentación.

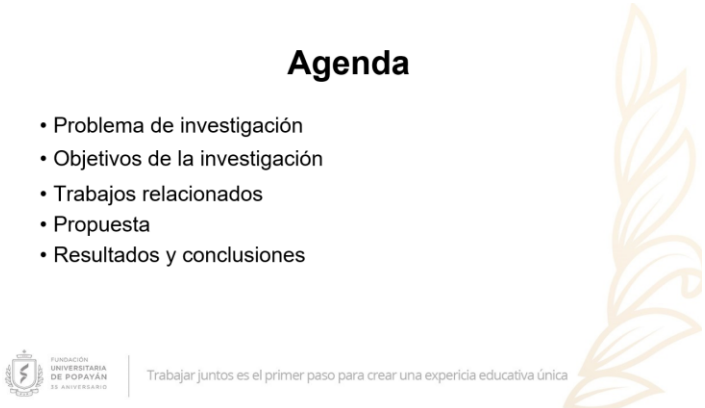


Figura 24. Agenda de la presentación.

Figura 25. Problema de investigación.

Figura 26. Objetivo general.

Objetivos de la investigación

Objetivos específicos

- Revisar el estado del arte sobre modelos de predicción del rendimiento académico.
- Definir un modelo de predicción para determinar el rendimiento académico que incluya la identidad cultural y el estilo de aprendizaje que se presente en el programa de psicología de la Fundación Universitaria de Popayán.
- Diseñar una estrategia de socialización de los resultados obtenidos por el modelo de predicción propuesto.

Figura 27. Objetivos específicos.

Trabajos relacionados

- Modelos de regresión
- Arboles de decisión
- Redes neuronales
- Modelos de clasificación

De acuerdo a las investigaciones algunos de los factores influyentes son socioeconómicos, personales, laborales, entre otros.

(Jena, 2018; Moreno-Muñoz, 2018; Porcel-Dapozo-López, 2009; Alcover et al., 2007)

Figura 28. Trabajos relacionados.

Propuesta

La metodología que se utilizo para continuar con el proceso de minería de datos , se deriva en aplicar el modelo de referencia CRISP-DM, para la construcción y evaluación del modelo.

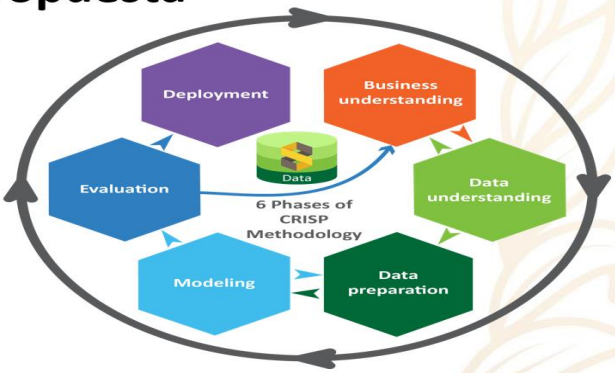


Figura 29. Propuesta.

Se presenta el proceso realizado para seleccionar el modelo con mayor precisión y los valores obtenidos.

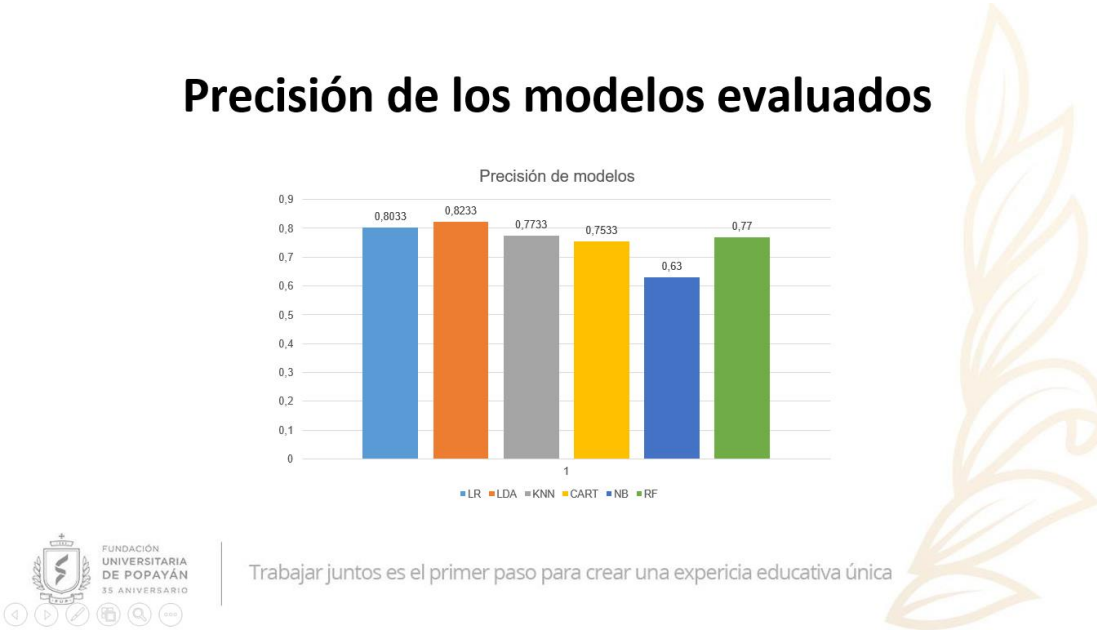


Figura 30. Precisión de los modelos evaluados.

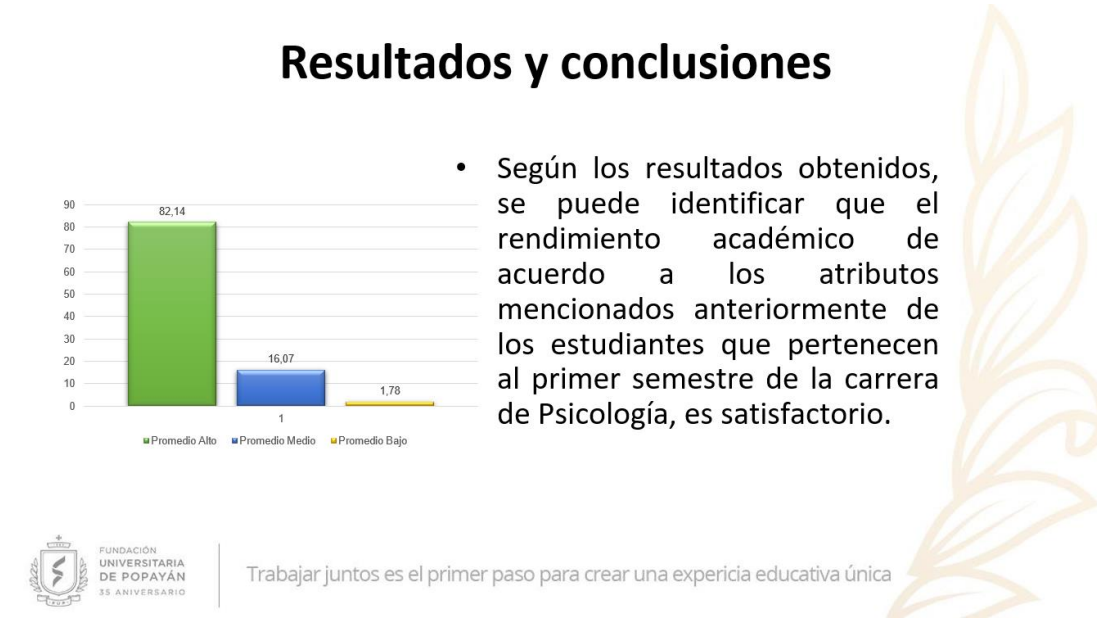


Figura 31. Resultados y conclusiones.

De igual manera se exponen las evidencias de la reunión que se tuvo con el área de bienestar institucional por medio de Google Meet.

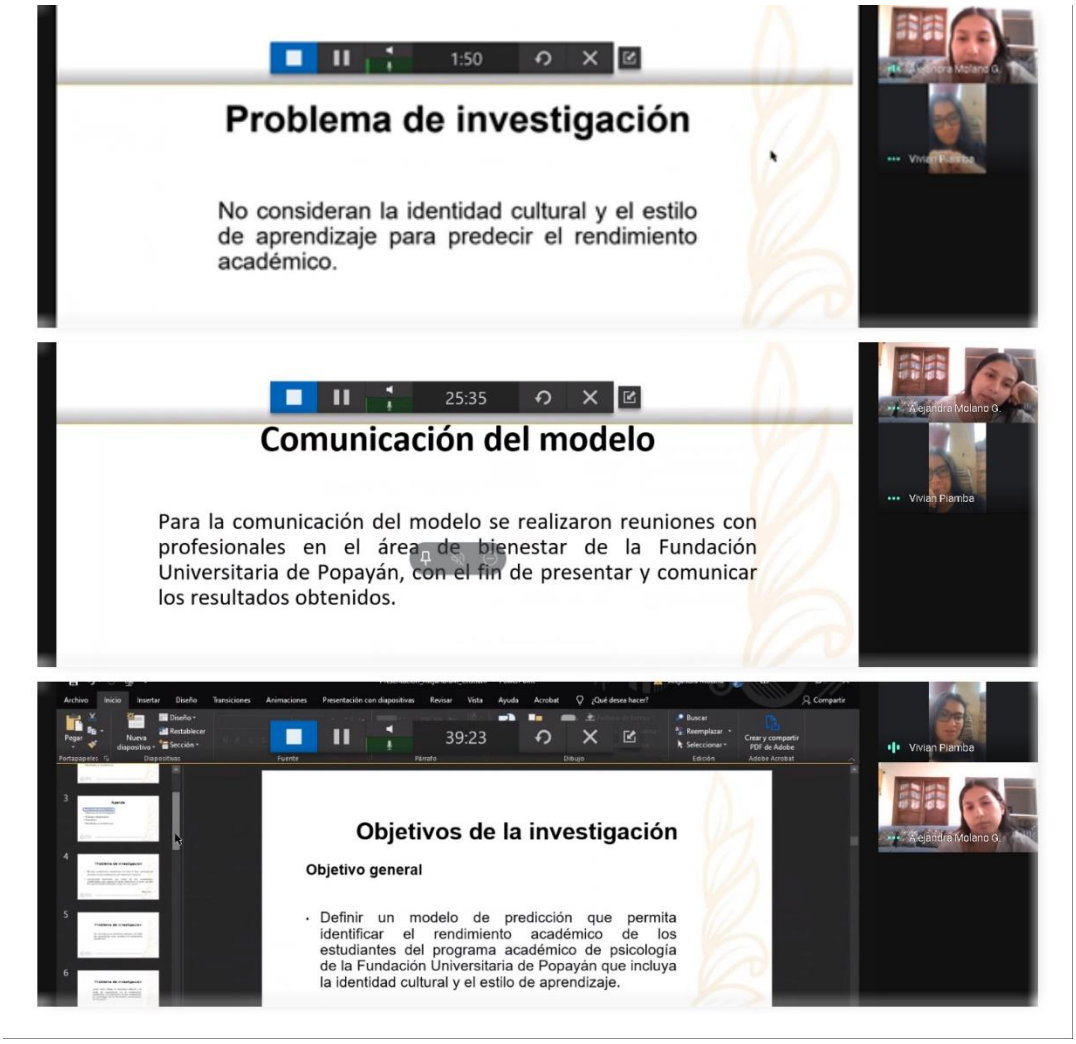


Figura 32. Evidencia de socialización.

Al finalizar la socialización, los resultados para la funcionaria del modelo son satisfactorio, ya que manifiesta la importancia del nivel de precisión de los datos presentados.

CAPÍTULO V. RESULTADOS

Este trabajo se desarrolló utilizando técnicas de minería de datos para la predicción del rendimiento académico haciendo uso de los datos proporcionados por los estudiantes del programa de psicología de la Fundación Universitaria de Popayán.

Inicialmente, se hizo uso del atributo **PROMEDIO GENERAL** que se encuentra consolidado en el Dataset, para ser usado como variable dependiente o variable objetivo y continuar con la realización de las pruebas de predicción del rendimiento académico.

Posteriormente de la limpieza y transformación de los datos, se determinaron los atributos con quien se realizarán las pruebas de predicción, dichos atributos se expresan a continuación:

Ítem	Atributo
1	Departamento de nacimiento
2	Estrato
3	Etnia
4	Edad

Tabla 6. Atributos escogidos para predicción.

La elaboración del modelo y de las técnicas usadas para la predicción se realizaron utilizando el lenguaje de programación Python, de igual manera las técnicas de clasificación usadas en la valoración comparativa para la predicción del rendimiento académico son:

Nombre	Modelo
“LR”	Regresión Logística
“LDA”	Análisis Discriminante Lineal
“KNN”	K Nearest Neighbors
“CART”	Arboles de Decisión y Clasificación
“NB”	Gaussian Naive Bayes
“RF”	Random Forest

Tabla 7. Modelos de predicción elegidos.

Para la predicción del rendimiento académico se opta por el modelo “LDA” Análisis Discriminante Lineal, ya que es una técnica de aprendizaje supervisado eficiente y que por medio de las pruebas de predicción resulto ser la más precisa según los resultados obtenidos con un valor equivalente a 0,82. Mencionada esta técnica de

minería de datos, podría usarse en la institución para clasificar y mejorar el proceso de enseñanza-aprendizaje de los estudiantes que el modelo determine con bajo rendimiento.

Los resultados del modelo “LDA” Análisis Discriminante Lineal, con los atributos, “Promedio general”, “Departamento de nacimiento”, “estrato”, “etnia” y “edad” son los siguientes:

PROMEDIO GENERAL	
Rendimiento académico	**Cantidad de estudiantes**
Bajo (0)	1
Medio (1)	9
Alto (2)	46

Tabla 8. Cantidad de estudiantes agrupados por el tipo de rendimiento académico.

Según los resultados obtenidos, se puede identificar que el rendimiento académico de acuerdo a los atributos mencionados anteriormente de los estudiantes que pertenecen al primer semestre de la carrera de Psicología, es satisfactorio porque del 100% de los alumnos, el 82,14% tiene un rendimiento académico alto; el 16,07% con un rendimiento de nivel medio; y tan solo el 1,78% presenta un bajo rendimiento.

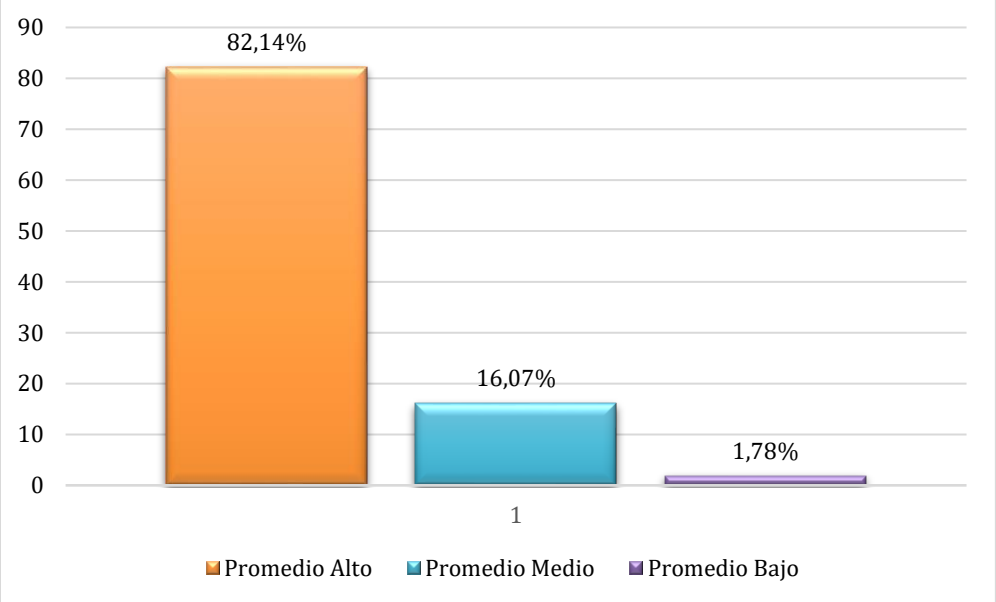


Figura 33. Grafica promedio general con porcentajes.

Por otra parte, en la socialización del modelo que se realizó al área de Bienestar Institucional de la Fundación Universitaria de Popayán, con el fin de dar a conocer el proyecto de investigación realizado y manifestar la importancia de tener

conocimiento acerca del rendimiento académico de los estudiantes, se obtuvieron los resultados esperados ya que se manifestó que el modelo de predicción cumple con los objetivos estipulados.

5.1. Conclusiones

De acuerdo al proyecto de investigación realizado se puede concluir que:

- Es factible el uso de minería de datos para el desarrollo de investigaciones que conlleven a la predicción.
- El objetivo de este trabajo de investigación fue definir un modelo de predicción del rendimiento académico de los estudiantes de primer semestre de la carrera de Psicología de la Fundación Universitaria de Popayán, donde inicialmente se realizó un estudio de los trabajos relacionados por medio de la técnica de mapeo sistemático, permitiendo tener bases para el desarrollo y evaluación de los modelos propuestos con datos obtenidos por medio de una encuesta previa a los estudiantes.
- Se ejecutó la evaluación de técnicas de minería de datos tales como: regresión logística, discriminación lineal, K vecinos más cercanos (KNN), arboles de decisión, Gaussian Naive Bayes y Random Forest donde se pudo determinar que la técnica de discriminación lineal fue la más precisa en las pruebas de predicción del rendimiento académico.
- El modelo de predicción para determinar el rendimiento académico de los estudiantes del programa de Psicología, se construyó con la ayuda del algoritmo de análisis de discriminación lineal y usando el lenguaje de programación de Python, para su posterior publicación en un repositorio público de Github, de este modo, se torne accesible a cualquier tipo de persona.
- Evidentemente, el modelo predictivo elaborado en esta investigación es eficiente para determinar la clasificación del rendimiento académico de los alumnos.

5.2. Recomendaciones

A continuación, se describen las recomendaciones que pueden ser aplicadas a futuro con la intención de proporcionarle mayor potencial y funcionamiento a la presente propuesta de investigación:

- Retomar este modelo predictivo aplicándolo en otros grupos poblacionales e implementarlo en otras universidades.

- Realizar una aplicación donde se incorpore el modelo propuesto que permita predecir y presentar el rendimiento académico de los estudiantes.

6. BIBLIOGRAFIA

[1] A. J. Salavarría Bravo, “Causas de la deserción escolar de los estudiantes de octavo año de básica del colegio fiscal Palestina,” Universidad de Guayaquil, 2011.

[2] B. P. C. Solis, “Deserción escolar en el sistema educativo del bachillerato,” Universidad de Guayaquil, 2019.

[3] G. Augusto and G. Méndez, “Factores asociados con la deserción de estudiantes de pregrado de la Universidad del Rosario,” Universidad Externado de Colombia, 2018.

[4] S. M. Ariza and D. A. Marín, “Factores intervenientes en la deserción escolar de la Facultad de Psicología, Fundación Universitaria Los Libertadores.,” *Tesis Psicológica*, vol. 4, no. 4, pp. 72–85, 2009.

[5] O. López Munguía, “La Inteligencia emocional y las estrategias de aprendizaje como predictores del rendimiento académico en estudiantes universitarios,” Universidad Nacional Mayor de San Marcos, 2008.

[6] L. F. Y. U. Digna y Vera, “Factores que influyen en la deserción de los estudiantes de la Facultad de Ciencias Económicas y Sociales de La Universidad del Zulia,” Universidad del Zulia, 2017.

[7] F. Apaza, Effer; Huamán, “Factores determinantes que inciden en la deserción de los estudiantes universitarios,” *Apunt. Univ. Univ. Peru. Unión.*, vol. 1, no. 1, pp. 77–86, 2012.

[8] D. L. La Red Martinez, J. C. Acosta, L. A. Cutro, V. E. Uribe, and A. R. Rambo, “Data Warehouse y data mining aplicados al estudio del rendimiento académico,” *CISCI 2010 - Novena Conf. Iberoam. en Sist. Cibern. e Informatica, 7to Simp. Iberoam. en Educ. Cibern. e Informatica, SIECI 2010 - Memorias*, vol. 1, no. 1, pp. 289–294, 2010.

[9] O. Doris, Millan; Claudia, Burbano; Armando, “Identidad cultural y el tipo de aprendizaje de los estudiantes de sicología de la Fundación Universitaria de Popayán. Hacia la predicción del rendimiento académico,” 2019.

[10] B. L. Roberto, Hernández Sampieri; Carlos, Fernandez Collado; Pilar, *Metodología de la investigación*, 6 Edición. México D.F: Marcela I. Rocha Martínez, 1996.

[11] R. B. Andres, “Uso de Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama,” Pontificia Universidad Javeriana Cali Cumplido, 2019.

[12] S. S. . Wu, A. . Elassal, R. Jordan, and F. . Schafer, “Minería de datos,” Benemérita Universidad Autónoma de Puebla, 1982.

[13] I. B. M. IBM, *Manual CRISP-DM de IBM SPSS Modeler*, 15th ed. Estados Unidos: IBM Corporation, 2012.

[14] J. Cepeda Ortega, “Una aproximación al concepto de identidad cultural a partir de experiencias: El patrimonio y la educación,” *Tabanque*, vol. 31, no. 31, pp. 244–262, 2018.

[15] G. Esguerra Pérez and P. Guerrero Ospina, “Estilos de aprendizaje y rendimiento académico en estudiantes de Psicología,” *Divers. Perspect. en Psicol.*, vol. 6, no. 1, p. 97, 2010.

[16] D. Carrizo and C. Ortiz, “Modelos del proceso de educción de requisitos: Un mapeo sistemático,” *Ing. y Desarollo*, vol. 34, no. 1, pp. 184–203, 2016.
[17] P. A. Vélez and A. Rey Piedrahita, “Control de calidad en sistemas crowdsourcing: un mapeo sistemático,” *Sci. Tech. Año XXII*, vol. 22, no. 1, p. 73, 2017.
[18] R. K. Jena, “Predicting students’ learning style using learning analytics: a case study of business management students from India,” *Behav. Inf. Technol.*, vol. 37, no. 10–11, pp. 978–992, 2018.
[19] A. Van der Merwe, H. Kruger, and J. Du Toit, “Mathematical modelling for academic performance status reports in learning analytics,” *ORiON*, vol. 34, no. 1, p. 31, 2018.
[20] P. M. Moreno-Marcos, P. J. Muñoz-Merino, C. Alario-Hoyos, I. Estévez-Ayres, and C. Delgado Kloos, “Analysing the predictive power for anticipating assignment grades in a massive open online course,” *Behav. Inf. Technol.*, vol. 37, no. 10–11, pp. 1021–1036, 2018.
[21] R. Timarán, “Detección de Patrones de Bajo Rendimiento Académico y Deserción Estudiantil con Técnicas de Minería de Datos,” *Iiisorg*, 2009.
[22] N. P. D. Martínez, M. Karanik, M. Giovannini, “Perfiles de Rendimiento Académico : Un Modelo basado en Minería de datos Academic Performance Profiles : A Model based on data Mining,” *Campus Virtuales*, vol. IV, no. 2015, pp. 12–30, 2015.
[23] G. N. Dapozo, E. Porcel, M. V. López, V. S. Bogado, and R. Bargiela, “Aplicación de minería de datos con una herramienta de software libre en la evaluación del rendimiento académico de los alumnos de la carrera de Sistemas de la FACENA-UNNE,” *VIII Work. Investig. en Ciencias la Comput.*, vol. 1, no. 1, pp. 1–6, 2006.
[24] E. Porcel, G. N. Dapozo, and M. V. López, “Modelos predictivos y técnicas de minería de datos,” *XI Work. Investig. en Ciencias la Comput.*, pp. 1–5, 2009.
[25] R. Alcover *et al.*, “Análisis del rendimiento académico en los estudios de informática de la universidad Politécnica de valencia aplicando técnicas de minería de datos,” in *XIII Jornadas de Enseñanza Universitaria de la Informática*, 1st ed., Thomson, Ed. Zaragoza: Thomson, Revista Novática, 2007, pp. 1–588.
[26] E. Yamao, “Predicción Del Rendimiento Académico Mediante Minería De Datos En Estudiantes Del Primer Ciclo De La Escuela Profesional De Ingeniería De Computación Y Sistemas, Universidad De San Martín De Porres, Lima-Perú,” Universidad De San Martín De Porres, 2018.
[27] S. Valero Orea, A. Salvador Vargas, and M. García Alonso, “Minería de datos: predicción de la deserción escolar mediante el algoritmo de árboles de decisión y el algoritmo de los k vecinos más cercanos,” *Recur. Digit. para la Educ. y la Cult.*, pp. 33–39, 2010.
[28] J. Serra-olivares, C. Leonel, M. Valverde, C. C. Armero, and P. G. Madrona, “Estilos de aprendizaje y rendimiento académico de universitarios de Educación Física chilenos,” *RETOS. Nuevas Tendencias en Educ. Física, Deport. y Recreación*, vol. 2041, no. 32, pp. 62–67, 2017.
[29] J. A. Gallardo Arancibia, “Metodología para el Desarrollo de Proyectos en

Minería de Datos CRISP-DM," vol. 84. pp. 487–492, 2013.
[30] M. GRÁNDEZ, "Aplicación De Minería De Datos Para Determinar Patrones De Consumo Futuro En Clientes De Una Distribuidora De Suplementos Nutricionales," Universidad San Ignacio de Loyola, 2017.