Modelo de aprendizaje automático para la predicción del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí

Cuzme Romero, María Gabriela

Vicerrectorado de Investigación, Innovación y Transferencia de Tecnología

Centro de Posgrado

Maestría en Gestión de Sistemas de Información e Inteligencia de Negocios

Trabajo de titulación previo a la obtención del título de Magister en Gestión de Sistemas de Información e Inteligencia de Negocios

Dr. García Bermúdez, Rodolfo Valentín

26 de octubre de 2020

URKUND

Document Information

Analyzed document	INFORME DE TESIS GABRIELA CUZME.docx (D75738041)
Submitted	6/28/2020 6:02:00 PM
Submitted by	Rodolfo Garcia
Submitter email	rvgarcia@utm.edu.ec
Similarity	3%
Analysis address	rvgarcia.utm@analysis.urkund.com

Sources included in the report

W	URL: https://repositorio.espe.edu.ec/bitstream/21000/13530/1/T-ESPE-053889.pdf Fetched: 10/23/2019 10:52:36 PM	3
SA	ANDRADE SORIANO RONALD LEONEL.pdf Document ANDRADE SORIANO RONALD LEONEL.pdf (D63151072)	1
W	URL: https://docplayer.es/amp/155854357-Universidad-nacional-del-altiplano.html Fetched: 11/26/2019 5:06:46 PM	2
SA	TESIS_MAESTRIA_23987482_M.pdf Document TESIS_MAESTRIA_23987482_M.pdf (D52140349)	5
SA	M2.881_20192_PEC4 - Redacci&#65533;n de la memoria_12682769.txt Document M2.881_20192_PEC4 - Redacci&#65533;n de la memoria_12682769.txt (D75561886)	3
W	URL: https://docplayer.es/789942-Modelo-para-la-automatizacion-del-proceso-de-determina... Fetched: 1/31/2020 6:17:42 PM	1
W	URL: https://riunet.upv.es/bitstream/handle/10251/136752/Nocedal%20Gonz%C3%A1lez_memori... Fetched: 2/17/2020 5:41:51 PM	1
W	URL: https://doi.org/10.15446/esrj.v23n1.63860 Fetched: 6/28/2020 6:04:00 PM	1

Firmado electrónicamente por:
RODOLFO VALENTIN GARCIA BERMUDEZ

Dr. Rodolfo García Bermúdez
Director

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "**Modelo de aprendizaje automático para la predicción del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí**" fue realizado por la Señora **Cuzme Romero, María Gabriela** el mismo que ha sido revisado y analizado en su totalidad, por la herramienta de verificación de similitud de contenido; por lo tanto cumple con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, razón por la cual me permito acreditar y autorizar para que lo sustente públicamente.

Sangolquí, 26 de octubre de 2020

Firma:

Firmado electrónicamente por:
RODOLFO VALENTIN
GARCIA BERMUDEZ

.....

Dr. García Bermúdez, Rodolfo Valentín

Director

C.C.: 0959774795

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

RESPONSABILIDAD DE AUTORÍA

Yo **Cuzme Romero, María Gabriela**, con cédula de ciudadanía n° 1314306463, declaro que el contenido, ideas y criterios del trabajo de titulación**: Modelo de aprendizaje automático para la predicción del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí** es de mí autoría y responsabilidad, cumpliendo con los requisitos legales, teóricos, científicos, técnicos y metodológicos establecidos por la Universidad de las Fuerzas Armadas ESPE, respetando los derechos intelectuales de terceros y referenciando las citas bibliográficas.

Sangolquí, 26 de octubre de 2020

Firma

Cuzme Romero, María Gabriela

C.C.: 1314306463

VICERRECTORADO DE INVESTIGACIÓN, INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

AUTORIZACIÓN DE PUBLICACIÓN

Yo **Cuzme Romero, María Gabriela** autorizo a la Universidad de las Fuerzas Armadas ESPE publicar el trabajo de titulación: **Modelo de aprendizaje automático para la predicción del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí** en el Repositorio Institucional, cuyo contenido, ideas y criterios son de mi responsabilidad.

Sangolquí, 26 de octubre de 2020

Firma

Cuzme Romero, María Gabriela

C.C.: 1314306463

Dedicatoria

El camino ha sido largo, pero sentir ahora que estoy cumpliendo uno de mis objetivos me llena de satisfacción, por eso deseo dedicarle esta tesis con mucho cariño y amor a las siguientes personas:

A mi Padre Salomón Cuzme Briones, sé que no estás conmigo, pero desde el cielo me proteges y te sientes orgulloso de la hija que me he convertido a pesar de tu partida, que desde pequeña me enseñaste a ser quien soy, con valores y principios que me han permitido desarrollarme como persona.

A mi madre Ramona Romero López, quien ha sido uno de mis pilares fundamentales que siempre ha estado conmigo motivándome a seguir adelante, que con sus consejos, perseverancia, amor incondicional y apoyo me supo ayudar en cada momento, para que no me rindiera por alcanzar mis sueños.

A mi esposo Julio Cesar Delgado Mendoza, quien me ha brindado su apoyo en todo momento.

A mis Hermanas: Sergia, Mariana, Carmen y Anita, quienes me han acompañado en todo momento y con su alegría, entusiasmo y consejos siempre han estado presentes para poder realizarme como persona.

Agradecimiento

A nuestro creador, pilar fundamental sobre el que se edifican todas las cosas.

A mis padres, hermanas y familia quienes con su apoyo incondicional me guían para alcanzar mis metas.

A mis amigos que siempre estuvieron presentes con consejos y motivación para seguir adelante a pesar de las dificultades.

A mi tutor quien con sus conocimientos, experiencia y dedicación me orientó en el camino para alcanzar la meta.

A la Universidad de las Fuerzas Armadas – ESPE por los conocimientos adquiridos durante el proceso de formación académica.

Índice de Contenidos

Índice de Tablas

Índice de Figuras

Resumen

La deserción estudiantil es un problema que se presenta en cada institución de educación superior, por lo cual en este trabajo de investigación se propone la aplicación de métodos y algoritmos de aprendizaje automático que permitieron la elaboración de un modelo capaz de realizar la estimación del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí. Con este objetivo se realizó el proceso de minería de datos a la información demográfica y académica de los estudiantes, contenida en el Sistema de Gestión Académica de la institución. Con la metodología CRISP DM se detallaron fases y procesos, tomando como muestra una carrera de cada una de las facultades durante el periodo académico (mayo 2014/febrero 2019). Luego se definió el criterio de inclusión para verificar la regularidad de los estudiantes y comprobar a través del coeficiente de correlación de Pearson la relación entre los estudiantes desertores y los no regulares, como estimador del riesgo de deserción. A través del proceso realizado se obtuvieron tres escenarios y se ejecutaron los algoritmos de regresión logística, KNN, redes neuronales, máquinas de vector de soporte y el algoritmo random forest que generó los resultados esperados y una mejor predicción sobre la deserción estudiantil. Por lo cual el mejor escenario fue durante los niveles 2,3 y 4 con un margen de error de 0.05 valorado por las métricas de evaluación y existiendo una alta correlación en cada una de ellas, en el área bajo la curva de 0.95 y un Puntaje F1 de 0.95.

PALABRAS CLAVES:

- **DESERCIÓN ESTUDIANTIL**
- **APRENDIZAJE AUTOMÁTICO**
- **MINERÍA DE DATOS**
- **CRISP DM**

Abstract

Student desertion is a problem that occurs in each higher education institution. This piece of research proposes the application of methods and algorithms of automatic learning that allowed the development of a model capable of estimating the risk of desertion in students at Universidad Técnica de Manabí. For this purpose, the process of data mining was carried out on the demographic and academic information of the students, found at the Academic Management System in this higher institution with the CRISP DM methodology. Phases and processes were detailed, taking as a sample a career of each of the faculties during the academic period (May 2014 / February 2019). Then, the inclusion criterion was defined to verify the regularity of the students and to check the relationship between the dropout and non-regular students, as an estimator of the risk of desertion, through the Pearson correlation coefficient. Through the process carried out, three scenarios were obtained and the logistic regression, KNN, neural networks, support vector machines and the random forest algorithm were executed generating the expected results and a better prediction of students' desertion. Therefore, the best scenario was at levels 2, 3 and 4 with a margin of error of 0.05, measured by the evaluation metrics and existing a high correlation in each of them in the area under the curve of 0.95 and an F1 Score of 0.95.

KEYWORDS:

- **STUDENT DESERTION**
- **MACHINE LEARNING**
- **DATA MINING**
- **CRISP DM**

Capítulo I: Problema de Investigación

Introducción

Con el avance del tiempo las instituciones de educación superior públicas y privadas presentan incidencia negativa sobre los procesos sociales, económicos, políticos y culturales del desarrollo sostenible de la nación; teniendo mayor influencia en los estudiantes que tienen un bajo nivel académico dentro del país.

Por lo cual el autor (Rueda & Pinilla, 2014) menciona que la manera como se articula el sistema educativo con procesos transformadores de la realidad social, en procura de la supervivencia humana y dignidad en el marco de la ciencia y tecnología, organización y articulación de saberes con miras a garantizar una educación de calidad.

Debido a esto se implementan estrategias que promuevan la gestión de conocimiento y la calidad de la formación de los estudiantes; sin embargo, la deserción universitaria se presenta como un problema que limita la visión y misión de formar profesionales capacitados para mejorar el desarrollo del país, debido a los diferentes factores familiares, personales y pedagógicos.

Además, a través de distintas investigaciones, se reflejan resultados de un gran número de estudiantes que no logran culminar sus estudios universitarios, relacionado al costo que ocasiona al estado. Al respecto la educación se caracteriza por ser un mecanismo primordial para que los países alcancen niveles de desarrollo más elevados. Por lo cual a través de esta investigación se describen los principales inconvenientes que se presentan en los estudiantes de la Universidad Técnica de Manabí.

Antecedentes

A medida que pasa el tiempo las universidades públicas y privadas del Ecuador registran datos sobre deserción de los estudiantes de grado. Los factores personales, académicos e institucionales que afectan la etapa universitaria influyen para que los estudiantes abandonen sus estudios.

El Consejo de Aseguramiento de la Calidad de la Educación Superior (CACES), evalúa los niveles de titulación, retención y deserción estudiantil. Como resultado de la evaluación que se realizó en el 2013, para muchas instituciones superiores fue el problema de abandono académico, en los niveles próximos a obtener su titulación. Además, la deserción incide de forma preocupante, en los resultados de la inversión pública (Bazantes et al., 2016).

Además, es necesario obtener datos para reconocer las causas donde los estudiantes de los primeros niveles dejan su carrera y se retiran, permitiendo que las autoridades universitarias tomen decisiones para disminuir este índice, y así poder tomar acciones correctivas que estén enfocadas a formular políticas y estrategias relacionadas con los programas de retención estudiantil que actualmente se encuentran establecidos.

El objetivo de este estudio fue crear un modelo predictivo de deserción estudiantil en los primeros niveles de cada una de las carreras, que permitió determinar la probabilidad de que un estudiante abandone sus estudios, donde se tomó como referencia su rendimiento académico y variables de su entorno personal, que ayudaron a obtener resultados a través de los algoritmos supervisados y las métricas de evaluación aplicadas.

Planteamiento del Problema

La deserción universitaria que se viene presentando desde hace varios años en algunas universidades ecuatorianas, ha ido incrementándose de forma notoria con una proporción alta de estudiantes que han abandonado sus estudios, por lo general en los primeros niveles de la carrera.

Las razones para que un estudiante abandone la universidad son muy diversas, aunque existen causas de deserción que se vuelven muy comunes en la mayoría de los grupos, como factores económicos en relación al desempleo por la falta de preparación profesional, ambientes familiares, edad, modelos pedagógicos universitarios diferentes a los modelos de bachillerato o de una mala elección de la profesión a seguir. Esta situación no solo afecta al estudiante, sino también a la institución de educación superior el no cumplimiento de su visión por la deserción de alumnos, y a la sociedad ecuatoriana, ya que ser desertor de la educación superior retrasa los avances socioeconómicos y tecnológicos del país.

Todos estos elementos configuran un escenario que destaca la necesidad de establecer con la mayor certeza las posibles causas y modos de actuación para enfrentar el problema de deserción estudiantil en los primeros niveles de las carreras en la Universidad Técnica de Manabí, con el objetivo de elaborar un modelo predictivo para la estimación del riesgo de deserción en estudiantes, mediante la utilización de técnicas de aprendizaje automático y la selección de las características de mayor relevancia en la base de datos académica de la institución.

Justificación

La deserción en la Educación Superior es una problemática presente en el país, con especial relevancia en las instituciones universitarias que prestan el servicio de educación. Por lo cual existe la preocupación de que siga aumentando y sea un factor que afecte directamente a las universidades públicas y privadas. Debido a esto, se hace necesario analizar y buscar estrategias que ayuden a solucionar la problemática y que sensibilice a las directivas de las universidades ecuatorianas, para que los estudiantes sigan cursando su carrera y no detengan su proceso de formación profesional.

Por ello, a las universidades llegan una cantidad significativa de bachilleres, provenientes de diversas regiones de todo el país, gracias al compromiso y estrategias que diseñan las unidades educativas para que sus estudiantes se preparen en las pruebas que la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación realiza para el ingreso a las Instituciones de Educación Superior a escala nacional en las diversas áreas del conocimiento; sin embargo, en la realidad existen un alto porcentaje de estudiantes que no logran alcanzar el puntaje requerido y escoger la carrera de su elección afectando en su proceso académico. Donde este tipo de estudio, es muy importante, debido a que la tasa de deserción se está empezando a considerar como un indicador de la calidad en la gestión universitaria y en diversos modelos de evaluación de las universidades (Viale Tudela, 2014).

Desde esta perspectiva, se analizan diferentes enfoques, que permitan conocer, abordar e intervenir la situación que se presenta en el país debido a la deserción estudiantil en las Instituciones de Educación Superior, con el fin de plantear soluciones contundentes de carácter educativo y social.

En el caso de la Universidad Técnica de Manabí, se pretende a través de esta investigación elaborar un modelo predictivo para la estimación del riesgo de deserción en estudiantes de esta IES, ayudando en la toma de decisiones por parte de las autoridades, considerando los aspectos de ambientes educativos causales de estas deserciones desde el ámbito institucional como estudiantil.

Objetivos

Objetivo General

Elaborar un modelo predictivo para la estimación del riesgo de deserción en estudiantes de la Universidad Técnica de Manabí, mediante la utilización de técnicas de aprendizaje automático y la selección de las características de mayor relevancia en la base de datos académica de la institución.

Objetivos Específicos

OE1. Realizar una revisión del estado del arte que permita determinar las técnicas de creación de modelos de aprendizaje automático más adecuadas para la predicción del riesgo de deserción en estudiantes.

OE2. Determinar la situación actual de la deserción estudiantil en la Universidad Técnica de Manabí a partir del análisis de los datos del Sistema de Gestión Académica y la aplicación de entrevistas a estudiantes y docentes.

OE3. Implementar el modelo de aprendizaje automático que a partir de determinación de las características de mayor relevancia permita predecir el riesgo de deserción en estudiantes de la institución.

OE4. Realizar la validación del modelo utilizando métricas de evaluación del rendimiento de clasificadores en tareas de predicción, como son la sensibilidad, especificidad, puntaje F1 y la Curva AUC ROC[1].

Hipótesis

Un modelo de aprendizaje automático, que incluya la extracción de las características académicas más relevantes de los estudiantes de la Universidad Técnica de Manabí permitirá obtener una estimación del riesgo de deserción de estos.

Categorización de las Variables de Investigación

- **Variable dependiente:** Estimación del riesgo de deserción estudiantil.
- **Variable independiente**: Modelo de aprendizaje automático.

Para la demostración de la hipótesis planteada se considera el uso de varios métodos deductivos y técnicas de investigación. En primer lugar, la evaluación del modelo de aprendizaje automático por medio de métricas de uso común en estos casos, entre ellas AUC-ROC2, sensibilidad, especificidad, exactitud y el puntaje F1. Adicionalmente se prevé el uso de encuestas y entrevistas, así como la comprobación de los resultados obtenidos por medio de las técnicas estadísticas adecuadas.

---

[1] Curva ROC: (Receiver Operating Characteristic) presentan la sensibilidad de una prueba diagnóstica que produce resultados contínuos, en función de los falsos positivos (complementario de la especificidad), para distintos puntos de corte.
[2] AUC ROC: Significa "área bajo la curva ROC". Esto significa que el AUC mide toda el área por debajo de la curva ROC completa de (0,0) a (1,1).

Capítulo II: Estado del Arte

En el presente capítulo se presentan los resultados de la revisión bibliográfica sobre el tema planteado, donde intervienen los criterios de inclusión y exclusión, tomando en cuenta las siguientes fases:

- **Definición del objetivo:** En esta fase se relacionan las preguntas de investigación en concordancia con los objetivos específicos que direccionan al objetivo de la investigación planteada.
- **Revisión Inicial:** Se realizó una búsqueda de información previa, para constatar la existencia de estudios relacionados con las preguntas de investigación propuestas.
- **Validación cruzada de estudios**: La validación cruzada permite garantizar que los estudios cumplan con los criterios de inclusión y exclusión, como resultado de la búsqueda se ha podido constatar que todos los trabajos cumplen con los criterios de inclusión y exclusión.
- **Definición de los criterios de inclusión y exclusión:** Plantean las características del tema planteado, según los autores de los artículos encontrados, permitiendo verificar la viabilidad de la búsqueda de la información.

Preguntas de Investigación

En la investigación planteada, se procedió a realizar las siguientes preguntas por cada uno de los objetivos específicos que se menciona en la Tabla 1.

Tabla 1

Preguntas de Investigación

Objetivos Específicos	Preguntas de Investigación
OE1: Realizar una revisión del estado del arte que permita determinar las técnicas de creación de modelos de aprendizaje automático más adecuadas para la predicción del riesgo de deserción en estudiantes.	**OE1 – RQ1.1:** ¿Qué tipo de estudios aportan en la definición de las técnicas de aprendizaje automático? **OE1 – RQ1.2:** ¿Cuáles son las características de las técnicas de aprendizaje automático?
OE2: Determinar la situación actual de la deserción estudiantil en la Universidad Técnica de Manabí a partir del análisis de los datos del Sistema de Gestión Académica y la aplicación de entrevistas a estudiantes y docentes.	**OE2 – RQ2.1:** ¿Cuáles son los principales indicadores de deserción que se encuentran en el Sistema de Gestión Académica? **OE2 – RQ2.2:** ¿Cuál es el estado actual de la deserción escolar en la UTM? **OE2 – RQ2.3:** ¿Qué tipo de técnicas se van a utilizar para este tipo de estudio?
OE3: Implementar el modelo de aprendizaje automático que a partir de determinación de las características de mayor relevancia permita predecir el riesgo de deserción en estudiantes de la institución.	**OE3 – RQ3.1:** ¿Cuál es el modelo de aprendizaje automático a implementar? **OE3 – RQ3.2:** ¿De qué manera puede el modelo automatizado predecir el riesgo de deserción?
OE4: Realizar la validación del modelo utilizando métricas de evaluación del rendimiento de clasificadores en tareas de predicción, como son la sensibilidad, especificidad, puntaje F1 y la Curva AUC ROC.	**OE4 – RQ4.1:** ¿Cómo validar el modelo propuesto con técnicas de aprendizaje automático? **OE4 – RQ4.2:** ¿Qué método es el óptimo para garantizar las métricas de evaluación en el rendimiento de las tareas de predicción?

Criterios de inclusión y exclusión

Criterios de Inclusión

- Artículos publicados en los últimos 5 años
- Artículos que traten el tema de deserción escolar universitaria y aparecen en publicaciones catalogadas en bases de datos especializadas como del área de la informática.
- Artículos que tratan aspectos relacionados con los modelos predictivos de aprendizaje automático.
- Artículos que estén indexados en la Base de datos de Scopus.

- Artículos que incluían técnicas de aprendizaje automático como árboles de decisión, bayesianas, redes neuronales, entre otros; utilizadas en el proceso metodológico.
- Artículos que mencionaban definiciones y procesos de estudios que han utilizado la metodología CRISP – DM.

Criterios de exclusión

- Artículos que trataban el tema de deserción escolar universitaria y aparecen en publicaciones catalogadas en bases de datos especializadas como del área de la pedagogía o las ciencias sociales en general.
- Artículos que no incluían en el resumen términos de aprendizaje automático y los relacionados con la educación a distancia.
- Artículos que no tomaban en cuenta el proceso de deserción en los colegios, y la comparación de varias técnicas de aprendizaje automática.

Definición de la estrategia de búsqueda

Integración del Grupo de Control

La conformación del grupo de control está relacionada a los estudios que cumplen con las características de la investigación analizado por los investigadores, considerando el título del estudio, resumen y palabras claves. Los estudios del grupo de control utilizados se muestran en la Tabla 2.

Tabla 2

Artículos del grupo de control

GRUPO DE CONTROL	TÍTULO	PALABRAS CLAVES
EC1	Application of decision trees for detection of student dropout profiles	Extraction Patterns; Student Dropout; Decision trees.
EC2	Perspectives to Predict Dropout in University Students with Machine Learning	Dropout, university students, machine learning.
EC3	Applying Data Mining Techniques to Predict Student Dropout: A Case Study	Student drop out, student desertion prediction, educational data mining, prediction models
EC4	Factors to predict dropout at the universities: A case of study in Ecuador	University student desertion, factor, data mining
EC5	Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks	Learning Analytics; Bayesian networks; Computer Science studies dropout; student profile
EC6	A review of student's performance prediction using educational data mining techniques	Paper review, student performance, prediction, educational data mining, comparison, classification.
EC7	Applying CRISP-DM in a KDD process for the analysis of student attrition	Data mining, Student attrition, KDD, Patterns, CRISP-DM, Analysis
EC8	Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas	Abandono escolar, Bajo rendimiento académico, Árboles de clasificación, Sistema universitario español.
EC9	Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance	Data mining, Predictive modeling, Academic risk prevention, Academic performance, Educational data mining.

Construcción de la cadena de búsqueda

Para la construcción de la cadena de búsqueda se analiza los estudios del grupo de control, encontrando palabras comunes entre estudios y las palabras propias que están direccionadas al objetivo de la presente investigación, para ello se formaron los siguientes contextos: student dropout, machine learning, prediction models (Ver Tabla 3).

Tabla 3

Comparación de la Cadena de Búsqueda

Contexto	Palabras Clave	Ec1		Ec2		Ec3		Ec4		Ec5		Ec6		Ec7		Ec8		Ec9		Palabras Repetidas
		Cadenas de Búsqueda																		
		1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	
Student dropout	Student desertion	X			X		X		X	X	X	X		X		X	X		X	**11**
	Student performance		X	X		X		X					X		X			X		**7**
	Decision trees	X				X				X		X				X			X	**6**
Machine learning	Classification			X	X		X		X				X	X				X		**7**
Predictive	Bayesian networks		X					X			X				X		X			**5**
	Prediction		X				X	X							X			X		**5**
	Models	X									X	X	X	X						**5**
	Predictive modeling			X	X	X			X	X						X	X		X	**8**

La cadena de búsqueda se forma con la combinación de las palabras que más se repiten en cada contexto para unir en la cadena se utilizó el conector OR y para concatenar con el contexto se usa el conector AND, estableciendo las siguientes cadenas de búsqueda:

- (TITLE-ABS-KEY (dropout)) AND (student) AND (LIMIT-TO (PUBYEAR , 2019) OR LIMIT-TO (PUBYEAR , 2018) OR LIMIT-TO (PUBYEAR , 2017) OR LIMIT-TO (PUBYEAR , 2016) OR LIMIT-TO (PUBYEAR , 2015) OR LIMIT-TO (PUBYEAR , 2014) OR LIMIT-TO (PUBYEAR , 2013) OR LIMIT-TO (PUBYEAR , 2012) OR LIMIT-TO (PUBYEAR , 2011) OR LIMIT-TO (PUBYEAR , 2010)) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "MATH"))

- TITLE-ABS-KEY (desertion) AND (LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "COMP"))

Discusión de Resultados

A través de las cadenas de búsquedas de la base de datos de Scopus, se aplicaron los criterios de inclusión y exclusión con el objetivo de determinar los siete estudios primarios relacionados a la problemática planteada con el uso de técnicas de aprendizaje automático teniendo como resultado los siguientes artículos que presentan información relevante.

(Timaran Pereira & Caicedo Zambrano, 2018) **Application of decision trees for detection of student dropout profiles:**

Este artículo hace referencia a la identificación de los patrones de abandono escolar de los datos socioeconómicos, académicos, disciplinarios e institucionales de estudiantes de pregrado en la Universidad de Narinõ de la ciudad de Pasto (Colombia), utilizando técnicas de extracción de datos. Además, se utilizaron datos de los estudiantes que ingresaron en el período comprendido entre el primer semestre de 2004 y el segundo semestre de 2006, por lo cual se analizaron tres cohortes completas con un período de observación de seis años, es decir hasta el 2011. Debido a esto los perfiles de abandono socioeconómico y académico de los estudiantes fueron descubiertos utilizando una técnica de clasificación basada en árboles de decisión, permitiendo generar conocimiento que apoye a la toma de decisiones efectiva del personal de la Universidad enfocado en desarrollar políticas y estrategias.

(Solis et al., 2018) **Perspectives to Predict Dropout in University Students with Machine Learning**

En este estudio se analizó el rendimiento de cuatro algoritmos de aprendizaje automático con diferentes perspectivas para definir archivos de datos, en la predicción de la deserción de estudiantes universitarios. Los algoritmos utilizados fueron: Random Forest, Neural Networks, Support Vector Machines and Logistic Regression. Se encontró que el algoritmo de random forest con 10 variables muestreadas aleatoriamente como candidatos en cada división, fue el mejor para predecir el abandono escolar y que la perspectiva ideal para entrenar el algoritmo es utilizar información sobre todos los semestres que los estudiantes toman dentro de un período de tiempo determinado.

(Alban & Mauricio, 2018) **Factors to predict dropout at the universities: A case of study in Ecuador**

La deserción en las universidades se ha convertido en una preocupación en varios países del mundo, sus altas tasas generan consecuencias negativas para los estudiantes y las organizaciones. Sobre la base del análisis de lo educativo, las teorías organizacionales y el razonamiento lógico se establecieron 11 factores que influyeron en la deserción. Esta investigación tenía como objetivo diseñar un modelo para determinar nuevos factores para predecir la deserción en la cual la dimensión del análisis fueron los estudiantes, las instituciones, el contexto académico y el entorno social y económico. Además, probar el uso de Regresión logística, Árbol de decisiones y Máquina de vectores de soporte si los factores propuestos están relacionados o pueden contribuir a predecir el abandono escolar en las universidades de Ecuador.

(Lacave et al., 2018) **Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks**

Este artículo describe la metodología utilizada para abordar esta pregunta en el contexto del análisis de aprendizaje. Las redes bayesianas (BN) se han utilizado ya que

proporcionan métodos adecuados para la representación, interpretación y contextualización de los datos. El enfoque propuesto se ilustra a través de un estudio de caso sobre el abandono escolar en Informática (CS) en la Universidad de Castilla-La Mancha (España), que está cerca del 40%. Con ese fin, se obtuvieron varios BN de una base de datos que contenía 383 registros que representan datos académicos y sociales de los estudiantes matriculados en el grado CS durante cuatro cursos. Luego, estos modelos probabilísticos fueron interpretados y evaluados. Los resultados obtenidos revelaron que el algoritmo K2 proporciona el mejor modelo que se ajusta a los datos, aunque la gran heterogeneidad de los datos estudiados no permitió el ajuste del perfil de deserción del alumno con demasiada precisión. No obstante, la metodología descrita aquí puede tomarse como una referencia para trabajos futuros.

(Castro R. et al., 2018) **Applying CRISP-DM in a KDD process for the analysis of student attrition**

El desgaste estudiantil es un fenómeno común que preocupa a las universidades públicas y privadas, que se ven afectadas económica y socialmente. Varios estudios han abordado este tema, sin embargo, se han centrado principalmente en aspectos académicos, sociales, demográficos y económicos. En este documento, se propone un método para analizar la deserción académica al proporcionar una visión de esta problemática desde la perspectiva de KDD (descubrimiento de conocimiento en bases de datos) y utilizar técnicas para identificar patrones de comportamiento de los estudiantes. A diferencia de otras propuestas, también se consideró variables proporcionadas por la prueba BADyG. Esta propuesta es importante porque proporcionó un apoyo para la toma de decisiones y la creación de planes de acción por parte de las instituciones de educación superior para reducir la alta tasa de deserción estudiantil.

(Ortiz-Lozano et al., 2017) **Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas**

Este trabajo analiza si es posible obtener un perfil del estudiante que está en riesgo de tener un bajo rendimiento académico en su primer año en tres momentos diferentes: cuando se realiza la admisión, al inicio del curso académico, y después de los primeros exámenes. Este estudio ha utilizado la técnica de árbol de clasificación basada en los algoritmos CART y QUEST y ha utilizado datos de 844 estudiantes de primer año matriculados en la Licenciatura en Administración de Empresas de la Universidad Pontificia Comillas. Obteniendo un porcentaje del 56% de observaciones clasificadas correctas para aquellos estudiantes que terminan presentando un bajo rendimiento académico, con la información disponible al final del primer semestre.

(Merchan Rubiano & Duarte Garcia, 2016) **Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance**

Este documento presenta y analiza la experiencia de la aplicación de ciertos métodos y técnicas de extracción de datos en 932 datos de estudiantes de Ingeniería de Sistemas de la Universidad El Bosque en Bogotá, Colombia; esfuerzo realizado para construir un modelo predictivo para el rendimiento académico de los estudiantes. Se revisaron trabajos anteriores, relacionados con la construcción de modelos predictivos en entornos académicos utilizando árboles de decisión, redes neuronales artificiales y otras técnicas de clasificación. Como descubrimiento iterativo y proceso de aprendizaje, la experiencia se analiza de acuerdo con los resultados obtenidos en cada una de las iteraciones del proceso. Cada resultado obtenido se evalúa con respecto a los resultados esperados, la caracterización de la entrada y salida de datos, lo que dicta la teoría y la

pertinencia del modelo obtenido en términos de precisión de predicción. Dicha pertinencia se evalúa teniendo en cuenta detalles particulares sobre la población estudiada y las necesidades específicas manifestadas por la institución, como el acompañamiento de los estudiantes a lo largo de su proceso de aprendizaje y la toma de decisiones oportunas para evitar el riesgo académico y la deserción.

Por lo tanto, el estado del arte es una categoría central y deductiva que se aborda y se propone como estrategia metodológica para el análisis crítico de las dimensiones política, epistemológica y pedagógica de la producción investigativa en evaluación del aprendizaje. (Universidad Pedagógica Nacional & Guevara Patiño, 2016).

Conclusiones

De acuerdo al estudio del arte revisado, se puede visualizar que en América Latina existe un índice alto de deserción estudiantil en comparación a Europa, debido a los inconvenientes que en algunas ocasiones presentan los estudiantes para continuar con sus estudios universitarios.

Algunos autores mencionan que las principales características son referentes a los aspectos académicos, sociales, demográficos y económicos; asimismo, mostraron que los algoritmos más utilizados son los árboles de decisión, la regresión logística, Bayes y Random forest, permitiendo alcanzar niveles confiables de precisión para identificar los predictores de abandono en las universidades públicas y privadas.

En consecuencia, el presente estudio contribuirá con la generación de conocimiento sobre la problemática planteada; a través de una perspectiva más integral desde el proceso de obtener la información, sistematización y modelamientos de datos

utilizando técnicas de aprendizaje automático y encontrando los principales indicadores de la deserción estudiantil en esta Institución de Educación Superior.

Capítulo III: Marco Teórico

La presente investigación a través de la fundamentación teórica plantea conceptos básicos y conocimientos previos que son necesarios para la práctica del proceso de minería de datos, con base al planteamiento del problema realizado, en concordancia con la hipótesis y las variables dependientes e independientes; por lo cual es importante buscar las fuentes bibliográficas que permitan detectar, extraer y recopilar la información de interés para construir el marco teórico pertinente al tema de investigación planteado

Fundamentación de la Variable Independiente

Minería de Datos

Es el conjunto de técnicas y tecnologías que permiten explorar grandes cantidades de datos que se producen en la actualidad, sumadas a su heterogeneidad, en las cuales hacen que las herramientas tradicionales de análisis de datos no resulten adecuadas para su recopilación, almacenamiento, gestión y análisis. En este contexto se determina como el término Big Data, haciendo referencia a características como gran volumen, velocidad y variedad de producción de los datos, y a las herramientas que se utilizan para encontrar valor en las mismas. La posibilidad de hallar patrones y tendencias en estas grandes cantidades de datos impacta directamente en la toma de decisiones (De Battista et al., 2016).

Las técnicas de la minería de datos (datamining) permiten diseñar estrategias de manejo para explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto para intentar ayudar a comprender el contenido de un repositorio de datos (O & Molina, 2015).

Tipos de Aprendizaje Automático

El aprendizaje automático o Machine Learning por sus siglas en inglés se ocupa del desarrollo de sistemas computacionales diseñados con el propósito de aprender y adaptarse a partir de los datos, sin la necesidad de introducir explícitamente el nuevo conocimiento adquirido. El auge del aprendizaje automático se debe a su gran aplicabilidad, puesto que prácticamente todo conocimiento es susceptible de ser aprendido e interpretado. Resulta ineludible el hecho de que los métodos de aprendizaje automático constituyen una herramienta de gran utilidad en diversas ramas de la ciencia (Valero, 2018).

Las técnicas de aprendizaje automático son utilizadas en diferentes áreas y tienen la capacidad de transformar datos en conocimiento, se pueden clasificar de acuerdo a la naturaleza de la señal o entrada de aprendizaje como se muestra a continuación:

- **Aprendizaje Supervisado**

Se basan en la generación de conocimiento a través del análisis de datos etiquetados. En este proceso se incluyen en los datos de estudio un conjunto de ejemplos con resultados conocidos con anticipación y en el cual el modelo de aprendizaje comprende los parámetros de la muestra para progresivamente ir adaptando e incorporando los datos nuevos y clasificarlos de forma correcta. Además, permite realizar predicciones adecuadas del comportamiento de datos que aún no han ingresado al sistema o no han sido procesados.

(Pérez & Luis, 2014) menciona que existen dos modelos principales de análisis supervisado que son: los métodos de clasificación que tiene la finalidad de estimar las clases categóricas de un conjunto de datos basados en un patrón binario o multi-clases (valores discretos, no ordenados o pertenencia a grupos); y, los

métodos de regresión permiten predecir productos que son continuos, por lo cual se puede contar con diversos números de variables predictivas de orden explicativo y una variable de respuesta que puede ser el resultado, determinando si existen alguna relación entre dichas variables.

- **Aprendizaje No Supervisado**

Es otra modalidad de aprendizaje automático en la que se incluyen conjuntos de datos sin etiquetar para realizar análisis y clasificaciones a pesar de que no se conoce con anticipación la estructura que poseen los datos. Con este análisis tiene lugar en el cual solo se conocen los datos de entrada, pero no existen datos de salida que correspondan a una determinada entrada. Por ello tienen un carácter exploratorio (Pérez Verona & Arco García, 2016).

Dentro de este tipo de aprendizaje automático existen dos categorías específicas: clustering que consiste en una técnica exploratoria de análisis de datos en la que se clasifica la información por grupos.

- **Aprendizaje Semi-supervisado**

El aprendizaje semi-supervisado se encuentra en medio entre el aprendizaje supervisado y el no supervisado, es decir combina datos etiquetados y los que no están etiquetados para construir un modelo supervisado que provienen de la misma distribución. Por otro lado, puede existir un sesgo en la elección de datos no etiquetados.

Entre los métodos de aprendizaje semisupervisado se encuentran: Selft-training, Co-training, Assemble y Re-Weighting (Guanin-Fajardo et al., 2019).

- **Aprendizaje con refuerzo**

También llamado Aprendizaje reforzado consiste en aprender a decidir ante una situación determinada o decidir qué acción es la más adecuada para lograr un objetivo planteado. Además, recibe siempre algún tipo de valoración acerca de la idoneidad de la respuesta dada (Montero, 2014).

- **Aprendizaje Profundo o Deep Learning**

Este tipo de análisis forma parte de lo que se conoce como Aprendizaje Profundo o Deep Learning por sus siglas en inglés y tiene como objetivo principal la construcción de modelos que incrementen el rendimiento en base al análisis de resultados ya procesados.

Por ello el aprendizaje profundo se refiere a una clase completa de métodos y proyectos de aprendizaje automático, que reúnen la característica de usar muchas capas de datos procesados no lineales de carácter jerárquico. Debido al uso de estos métodos y proyectos, según (Chagas, 2019), abarca el aprendizaje en diversos niveles de representación e intangibilidad que ayudan en el proceso de comprensión de la información, imágenes, sonidos y textos.

Debido a esto el autor (López Ramos et al., 2019) menciona que los modelos de aprendizaje profundo tienen varias mejoras sobre las máquinas de aprendizaje tradicionales. Dos de sus más destacados aportes son que reducen la necesidad de construir los datos (mediante un pre-procesamiento inicial) y realizar ingeniería de características en los conjuntos de entrenamiento.

Durante este proceso también pueden aparecer características no detectadas por los humanos. Además, estos modelos consisten en técnicas de aprendizaje supervisado o no supervisado teniendo como estructura principal varias capas de Redes de Neuronas Artificiales (Neural Networks; NN) que son capaces de

aprender una representación jerárquica en arquitecturas profundas que están compuestas de varias capas de procesamiento, donde cada capa produce respuestas no-lineales basadas en la capa anterior y la entrada inicial.

Redes Neuronales Este sistema lo que busca es simular el cerebro humano, las redes neuronales cuentan con elementos que asemejan una neurona biológica, los cuales procesan la información y son capaces de aprender de la experiencia, generalizar de ejemplos previos a ejemplos nuevos y abstraer la información más importante de una base de datos; siendo así de gran utilidad en múltiples procesos en los cuales se tenga una serie de datos óptimos para la utilización en la red neuronal artificial (Jara Estupiñan et al., 2016).

- **Tensorflow**

Es una biblioteca de código abierto que se basa en un sistema de redes neuronales. Esto significa que puede relacionar varios datos en red simultáneamente, de la misma forma que lo hace el cerebro humano. Este tipo de algoritmos, se han utilizado por varios años en muchos productos y áreas en Google, como lo son búsqueda, traducción, publicidad, visión artificial y reconocimiento de voz e imágenes.

TensorFlow es un sistema de segunda generación, para implementar y desplegar redes neuronales. Fue lanzado al público como un marco de código abierto con Apache 2.0, en noviembre del 2015, donde en sus primeros inicios, se dedicó a proyectos internos de Google, pero, gracias a su estabilidad y flexibilidad, combinado con la experiencia de los Ingenieros de Google, que son el motor que da mantenimiento y que continúa desarrollando las librerías, han convertido a TensorFlow en un sistema líder para hacer aprendizaje profundo (Mendoza, 2019).

Algoritmos de Aprendizaje Automático Supervisado de Clasificación

Para realizar el proceso de minería de datos en la presente investigación es necesario la fundamentación teórica de cada uno de los algoritmos a ser utilizados que se detallan a continuación:

- **Regresión Logística**

Es una técnica de aprendizaje automático que proviene del campo de la estadística. A pesar de su nombre no es un algoritmo para aplicar en problemas de regresión, en los que se busca un valor continuo, sino que es un método para problemas de clasificación, en los que se obtienen un valor binario entre 0 y 1.

Además, se mide la relación entre la variable dependiente, la afirmación que se desea predecir, con una o más variables independientes, el conjunto de características disponibles para el modelo. Para ello utiliza una función logística que determina la probabilidad de la variable dependiente. Debido a esto lo que se busca en estos problemas es una clasificación, por lo que la probabilidad se ha de traducir en valores binarios. Para lo que se utiliza un valor umbral. Los valores de probabilidad por encima del valor umbral la afirmación es cierta y por debajo es falsa. Generalmente este valor es 0,5, aunque se puede aumentar o reducir para gestionar el número de falsos positivos o falsos negativos (Loor, 2018).

Por ello sus resultados son interpretables y es muy usada en diferentes áreas ya que funciona muy bien cuando hay muchísimos datos y las interrelaciones entre ellos no son muy complejas.

- **Máquinas de Vector Soporte (SVM)**

El autor (Jara Estupiñan et al., 2016) menciona que este algoritmo puede ser considerado una extensión del algoritmo “perceptron”. En SVM el objetivo es establecer una línea de decisión que separe las clases maximizando el margen entre esta línea y los puntos de muestra cercanos a este hiperplano, estos puntos se llaman vectores soporte.

Para establecer los márgenes máximos, se añaden dos rectas paralelas (márgenes) e intentando maximizar sus distancias a la línea de decisión original. Teniendo en cuenta los puntos sin clasificar (errores) y los que quedan entre los márgenes de la línea.

Normalmente, las líneas de decisión con márgenes grandes tienden a tener un error de generalización menor. Por otro lado, los modelos con márgenes pequeños tienen menor tendencia al “sobreajuste” (overfitting).

- **Bosques Aleatorios (Random Forests)**

Los Random Forests son bosques constituidos por árboles de clasificación o regresión, creados mediante un algoritmo que introduce dos fuentes de aleatoriedad en la generación de dichos árboles con el objetivo de reducir la correlación entre ellos y mejorar las predicciones. Una vez generado el bosque, la predicción se toma promediando las predicciones individuales de los árboles.

La primera fuente de aleatoridad es el Bootstrapping, técnica frecuentemente utilizada en algoritmos de aprendizaje automático que consiste en utilizar para la creación de cada árbol una muestra distinta, obtenida sobre el conjunto inicial mediante una elección aleatoria de los datos con reemplazamiento.

La otra fuente de aleatoriedad consiste en limitar el número de variables de entrada candidatas a provocar la partición en cada nodo a un número prefijado (mtry<n),

seleccionado al azar las mtry variables de entre las n posibles variables de entrada. De esta forma se crean árboles diferentes no correlacionados (González, 2017).

- **Los K-vecinos Más Cercanos (K Nearest Neighbors or KNN)**

Pertenecen a un tipo especial de modelos de aprendizaje automático que se llaman frecuentemente “algoritmos perezosos”, reciben este nombre porque no aprenden cómo discriminar el conjunto de datos con una función optimizada, en su lugar memorizan el conjunto de datos.

El nombre también se refiere a la clase de algoritmos llamados “no paramétricos”. Estos son algoritmos basados en instancia, que se caracterizan por memorizar el conjunto de datos de entrenamiento, y el aprendizaje perezoso es un caso particular de estos algoritmos, asociados con coste computacional cero durante el aprendizaje.

El algoritmo encuentra las k muestras que son más cercanas al punto que se quiere clasificar, basando sus predicciones en la distancia métrica, es decir calculando la distancia entre el ítem a clasificar y el resto de ítems del dataset de entrenamiento; además seleccionando los k elementos más cercanos (con menor distancia, según la función que se use) y por último se realiza una votación de mayoría entre los k puntos: los de una clase/etiqueta que dominen y después decidirán su clasificación final.

La principal ventaja es que se adapta a los nuevos datos de entrenamiento, al ser un algoritmo basado en la memoria. La desventaja es que el coste computacional se incrementa linealmente con el tamaño de los datos de entrenamiento (Roman, 2019).

Particionado del Conjunto de Datos

Validación Cruzada K-Fold (K-Fold Cross Validation)

El autor (Ziggah et al., 2019) menciona que la Validación Cruzada es un procedimiento de remuestreo utilizado para evaluar modelos de aprendizaje automático en una muestra de datos limitada. El procedimiento tiene un único parámetro llamado k que

se refiere al número de grupos en los que se dividirá una muestra de datos determinada. Como tal, el procedimiento a menudo se llama validación cruzada k-fold. Cuando se elige un valor específico para k, se puede usar en lugar de k en la referencia al modelo, como k = 10 convirtiéndose en una validación cruzada de 10 veces.

Se utiliza principalmente en el aprendizaje automático aplicado para estimar la habilidad de un modelo de aprendizaje automático en datos no vistos. Es decir, usar una muestra limitada para estimar cómo se espera que el modelo funcione en general cuando se usa para hacer predicciones sobre datos que no se usaron durante el entrenamiento del modelo (Brownlee, 2018).

Métricas de Evaluación

Para entrenar y probar un modelo se parten los datos en dos conjuntos: el conjunto de entrenamiento (training set) y el conjunto de prueba (test set). Esta separación es necesaria para garantizar que la validación de la precisión del modelo sea una medida independiente (Rodríguez, 2015).

Por lo cual es necesario la utilización de métricas de evaluación para verificar la confiabilidad de los resultados obtenidos en la aplicación de los algoritmos de clasificación de aprendizaje supervisado, por esto estas métricas se detallan a continuación:

- **Matriz de Confusión**

Es una herramienta fundamental a la hora de evaluar el desempeño de un algoritmo de clasificación, ya que dará una mejor idea de cómo se está clasificando dicho algoritmo, a partir de un conteo de los aciertos y errores de cada una de las clases en la clasificación. Así se puede comprobar si el algoritmo está clasificando mal las clases y en qué medida.

El desempeño de un sistema es usualmente evaluado usando métricas que se calculan en base a los datos de dicha matriz. La siguiente tabla muestra la matriz de confusión para un clasificador en dos clases:

Tabla 4

Explicación de la Matriz de Confusión

Valor Real	**Clasificador**	
	Negativos	Positivos
Negativos	A	B
Positivos	C	D

- a es el número de predicciones correctas de que un caso es negativo.
- b es el número de predicciones incorrectas de que un caso es positivo, o sea la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les denomina errores de tipo I.
- c es el número de predicciones incorrectas de que un caso es negativo, o sea la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les denomina errores de tipo II.
- d es el número de predicciones correctas de que un caso es positivo (Santana, 2018).

Métrica de Precisión : Con la métrica de precisión se puede medir la calidad del modelo de aprendizaje automático en tareas de clasificación (LUCA et al., 2018). Además, su cálculo se lo realiza del ratio[3] de los elementos clasificados correctamente de una clase

[3] Ratio: Es el cociente entre dos magnitudes que están relacionadas. El objetivo es poder establecer cálculos y realizar comparaciones a través de este instrumento.

concreta, respecto de todos los clasificados dentro de esa clase. La precisión es intuitivamente la capacidad del clasificador de no etiquetar como positivo una muestra que es negativa.

Métrica de Exhaustividad (Recall): Permite informar sobre la cantidad que el modelo de aprendizaje automático es capaz de identificar. Además, es el ratio de los elementos clasificados correctamente en su clase con respecto a todos los elementos que debería haber clasificado el modelo dentro de esa clase.

Puntaje F1: Se utiliza para combinar las medidas de precision y recall en un sólo valor y su cálculo se lo realiza en función del promedio ponderado de la precisión y la exhaustividad.

Métrica de Exactitud (Accuracy): Mide el porcentaje de casos que el modelo ha acertado. Esta es una de las métricas más utilizadas y, además el cálculo depende del número de predicciones correctas sobre el total de predicciones realizadas.

Área bajo la curva (AUC ROC): Representa de forma visual la efectividad del algoritmo que se está utilizando. Además, permite medir la precisión de los modelos de clasificación mediante tasa de los verdaderos positivos frente a los falsos positivos. Un área bajo la curva mayor, indica un modelo mejor (Fernández, 2019).

Técnicas de Desbalanceo de Datos

Las técnicas de minería de datos están encaminadas a desarrollar algoritmos que sean capaces de tratar y analizar datos de forma automática con el objetivo de extraer de cualquier tipo de información subyacente en dichos datos, por lo cual el desbalance en los datos se refiere a una situación en la que el número de observaciones no es el mismo para todas las clases en un dataset usado para clasificación. En algunas áreas los problemas con datos desbalanceados son muy comunes. Sin embargo, también se puede hallar situaciones de desbalanceo en clasificaciones con múltiples categorías.

- **Métodos de Remuestreo**

Los métodos de re-muestreo, también conocidos como métodos de pre-procesado de conjuntos de entrenamiento, pueden ser divididos en tres grupos: los que eliminan instancias de la clase mayoritaria (under-sampling), los que generan nuevas instancias de la clase minoritaria (over-sampling) o la hibridación de ambas. A continuación, se describen cada uno de los métodos:

Métodos de Over-Sampling: Entre las estrategias más conocidas para la generación de nuevas instancias con el fin de balancear conjuntos de entrenamiento se encuentra SMOTE (Synthetic Minority Over-Sampling TEchnique), este algoritmo para cada ejemplo de la clase minoritaria introduce ejemplos sintéticos en la línea que une al elemento con sus 5 vecinos más cercanos. Sin embargo, esta estrategia presenta el problema de que puede introducir ejemplos de la clase minoritaria en el área de la mayoritaria, es decir, crear malos ejemplos que posteriormente pudieran confundir a los clasificadores (Sánchez & Caridad, 2016).

Métodos de Under-Sampling: Esta técnica corresponde a un método no heurístico que tiene como objetivo equilibrar la distribución de las clases a través de la "eliminación aleatoria" de ejemplos de la clase mayoritaria. Dentro de los métodos más clásicos para realizar under-sampling se encuentran:

- El RU (random under-sampling), donde se selecciona de manera aleatoria instancias de la clase mayoritaria para ser eliminados sin reemplazamiento hasta que ambas clases queden balanceadas.
- El Tomek Links, donde se eliminan sólo instancias de la clase mayoritaria que sean redundantes o que se encuentren muy cerca de instancias de la clase minoritaria (Sánchez & Caridad, 2016).

Métodos híbridos: A pesar de que tanto el over-sampling como el under-sampling logran buenos resultados por separado, muchos investigadores del área han obtenido magníficos resultados hibridando ambos métodos, de estos se pueden citar:

- SMOTE-TomekHybrid: inicialmente se realiza el over-sampling con la clase minoritaria y luego se aplica el Tomek Link a ambas clases.
- CNN + Tomek links: es similar a la selección de un solo lado (one-sided selection), pero el método para encontrar el subconjunto consistente se aplica antes del Tomek Links (Sánchez & Caridad, 2016).

Métodos de codificación de variables categóricas:

Los datos categóricos o nominales, como su nombre lo indica, son usados para nombrar o categorizar información, por lo cual existen varios métodos para categorizar estas variables como se muestra a continuación.

- **Método de Variables Dummy**

Crear variables dummy o ficticias implica transformar datos de un formato “alto”, en el que cada columna contiene la información de una variable, a datos con un formato “ancho”, en los que múltiples columnas contienen la información de las dos variables, codificada de manera binaria, esto es, con 0 y 1.

- **Método de Variables One Hot Encoding**

El método One Hot Encoding tiene como estrategia la implementación de crear una columna para cada valor distinto que exista en la característica que se está codificando y, para cada registro, marcar con un 1 la columna a la que pertenezca dicho registro y dejar las demás con 0 (Echeverri, 2019).

Lenguaje de Programación Python

Python es un lenguaje de programación creado por Guido van Rossum a principios de los años 90, tiene una sintaxis muy limpia y que favorece un código legible. Entre sus principales características están: Ser un lenguaje interpretado o de script, tipado dinámico, multiplataforma (Windows, Linux, MAC) y orientado a objetos, es uno de los lenguajes más utilizados en el desarrollo de aplicaciones de ciencia de datos porque combina en gran forma su potencia, con una sintaxis muy clara lo que es atractivo para los programadores principiantes o personas que han dejado de programar por algún tiempo.

Sus librerías permiten una gran variedad de funcionalidades generales y específicas para la carga y visualización de datos, estadísticas, procesamiento de lenguaje natural y de imágenes, entre otras. Sus bibliotecas de manejo de tareas relacionadas con grandes volúmenes de datos son: Numpy y Pandas, éstas incluyen muchas de las capacidades del software R y MATLAB, pero son más intuitivas. Estas características hacen de Python un lenguaje superior a la par de los lenguajes disponibles, y es por esto, que actualmente se lo elija con más frecuencia para el desarrollo de un gran número de aplicaciones para análisis de datos (Robles & Roberto, 2019).

- **Framework Anaconda**

Anaconda es un sistema multiplataforma para la gestión de paquetes y entornos. Tiene licencia BSD, Open Source. En un principio fue creada para el manejo de programas en Python, pero actualmente permite compilar y gestionar diferentes tipos de software (Mejías, 2018). Para la programación de este proyecto de investigación se instaló la versión 2020 de Conda que viene junto la distribución Anaconda3 con Python 3.7 64-Bit.

- **Bibliotecas principales**

La biblioteca que se ha utilizado en gran parte para este estudio ha sido Scikit – Learn, centrada en el desarrollo con aprendizaje máquina en el lenguaje de programación

Python. Incluye soporte para la implementación con diversos algoritmos que proporcionan buenos resultados en gran variedad de conjuntos de datos.

Una de las ventajas más destacables de Scikit-Learn es el soporte, del que se beneficia por ser una biblioteca dentro de un lenguaje, es decir, puede utilizar todos los recursos disponibles del lenguaje para apoyar su labor y, dado que Python es un lenguaje que fue creado para facilitar el aprendizaje y que su programación fuese similar al lenguaje natural de las personas.

Por otro lado, en una primera aproximación del modelo elaborado con prueba de resultados, se ha hecho uso de la biblioteca TensorFlow, una biblioteca que permite una mayor parametrización en la utilización de modelos basados en algoritmos de redes neuronales, entre otros (Fernandez, 2019).

Metodología CRISP-DM

Para abarcar todas las necesidades del proceso de minería de datos es necesario seguir varios pasos, de esta manera permite adquirir el enfoque más idóneo para poder llegar a satisfacer los objetivos planteados. Es por ello que existen diferentes metodologías, pero en este trabajo de investigación se ha escogido la metodología conocida como CRISP-DM por sus siglas en inglés Cross-Industry Standard Process for Data Mining, que hace referencia a un modelo de trabajo que integra seis etapas de un proceso cíclico.

Etapas de la Metodología CRISP-DM

- **Comprensión del problema**

Incluye los objetivos, conocimiento previo de la situación de partida, los objetivos a alcanzar mediante la minería de datos y el desarrollo de un plan de proyecto.

- **Comprensión de los datos o información**

Esta etapa considera inicialmente una metodología para: conseguir datos, realizar una exploración de los datos para conocer la información que dan y la verificación de la calidad de éstos.

- **Preparación de los datos**

Una vez que se han recogido los datos suficientes, es necesario limpiarlos y adaptarlos a la forma deseada para poder trabajar con ellos posteriormente. La transformación de los datos puede facilitar la identificación de patrones y comportamientos de los datos.

- **Modelización**:

Esta etapa involucra la manipulación de softwares de minería de datos para el análisis o estudio de los mismos, para establecer agrupaciones de aquellos que muestran un comportamiento similar o para la creación de modelos de predicción, entre otros.

- **Evaluación:**

Esta etapa pretende descubrir el grado de fiabilidad del modelo o modelos obtenidos en la etapa anterior, para así saber cuál de ellos permite satisfacer de forma más completa los objetivos establecidos en las etapas previas.

- **Implementación:**

Es la etapa final de esta metodología, y llega cuando se pretende realizar la implementación del modelo para alcanzar los objetivos establecido (Mascort Colomer, 2019).

- **Coeficiente de correlación de Pearson**

Es una prueba que mide la relación estadística entre dos variables continuas. Si la asociación entre los elementos no es lineal, entonces el coeficiente no se encuentra representado adecuadamente.

El coeficiente de correlación puede tomar un rango de valores de +1 a -1. Un valor de 0 indica que no hay asociación entre las dos variables. Un valor mayor que 0 indica una asociación positiva. Es decir, a medida que aumenta el valor de una variable, también lo hace el valor de la otra. Un valor menor que 0 indica una asociación negativa; es decir, a medida que aumenta el valor de una variable, el valor de la otra disminuye (Hernández Lalinde et al., 2018).

Fundamentación de la Variable Dependiente

Universidad Técnica De Manabí

La Universidad Técnica de Manabí (UTM), es una universidad pública ubicada en la ciudad de Portoviejo, Manabí, fundada el 29 de octubre de 1952. Sus tres funciones sustantivas son: la investigación, la academia y la vinculación con la sociedad; interviene con calidad en todas las esferas y sectores tantos públicos como privados mediante el apoyo de estudiantes, docentes y autoridades (Universidad Técnica de Manabí (Ecuador) - EcuRed, s. f.).

La Universidad Técnica de Manabí posee plenas facultades para organizarse dentro de las disposiciones de la Constitución de la República del Ecuador, la Ley Orgánica de Educación Superior, su reglamento, otras leyes conexas, el Estatuto Orgánico de la Universidad Técnica de Manabí y los reglamentos expedidos para estructurar la organización de la institución. Actualmente la institución se encuentra acreditada dentro del Sistema de Educación Superior del Ecuador, ubicándose en la categoría B, de acuerdo a la resolución del Consejo de Evaluación, Acreditación y Aseguramiento de la Calidad de la

Educación Superior (CEAACES), emitida el 9 de mayo de 2016 sobre la base de la solicitud de recategorización y respectivo proceso de evaluación.

Deserción Estudiantil

La deserción estudiantil universitaria es un fenómeno causal que tiene diferentes factores que confluyen, como aspectos de orden académico, institucional, personal, sociodemográfico, no solo debe ser entendida como el abandono definitivo de las aulas de clase, sino también como el abandono de la formación académica que tiene serias repercusiones sociales (Rivera, 2015).

Tipos de deserción estudiantil

- **Deserción precoz:** Cuando un estudiante abandona un programa antes de comenzar habiendo sido aceptado.
- **Deserción temprana:** Cuando se abandona el programa durante los primeros cuatro semestres.
- **Deserción tardía:** Entendida como abandono desde el quinto semestre en adelante. Además, indica que de hecho hay una diferencia entre: Deserción total: cuando el alumno abandona por completo un plan educativo y decide no regresar.
- **Deserción parcial:** Cuando el alumno hace lo que generalmente se conoce como una baja temporal y cuando se siente seguro regresa al programa educativo para continuar con sus estudios (Rivera, 2015).

Causas de la deserción estudiantil

- **Factores personales:** Constituidos por motivos psicológicos, que comprenden aspectos motivacionales, emocionales, desadaptación e insatisfacción de expectativas; motivos sociológicos, debidos a influencias familiares y de otros

grupos como los amigos, condiscípulos, vecinos; y otros motivos no clasificados como la edad, salud, fallecimiento, entre otros.

- **Factores académicos:** Dados por problemas cognitivos como bajo rendimiento académico, repetición, ausencia de disciplina y métodos de estudio; deficiencias universitarias como dificultades en los programas académicos que tienen que ver con la enseñanza tradicional, insatisfacción académica generada por la falta de espacios pedagógicos adecuados para el estudio, falta de orientación profesional que se manifiesta en una elección inadecuada de carrera o institución y ausencia de aptitud académica.
- **Factores socio-económicos:** Generados por bajos ingresos familiares, desempleo, falta de apoyo familiar, incompatibilidad de horario entre trabajo y estudio.
- **Factores Institucionales:** Causados por el cambio de institución, deficiencia administrativa, influencia negativa de los docentes y otras personas de la institución, programas académicos obsoletos y rígidos, baja calidad educativa (Sánchez Amaya et al., 2009).

Estimación del riesgo de deserción estudiantil

Desde una perspectiva cualitativa, las principales causas y condicionantes de los procesos de deserción universitaria se presentan de este tipo: permanente y temporal, cuyos patrones explicativos son distintos: la deserción temporal se explica principalmente por razones vocacionales, socioculturales y motivacionales; la permanente, por razones socioeconómicas. Los factores explicativos de cada tipo de deserción dan cuenta de las diferentes oportunidades y limitaciones que los estudiantes enfrentan hoy en el sistema educativo. Por ello, es fundamental que sean consideradas al momento de delinear

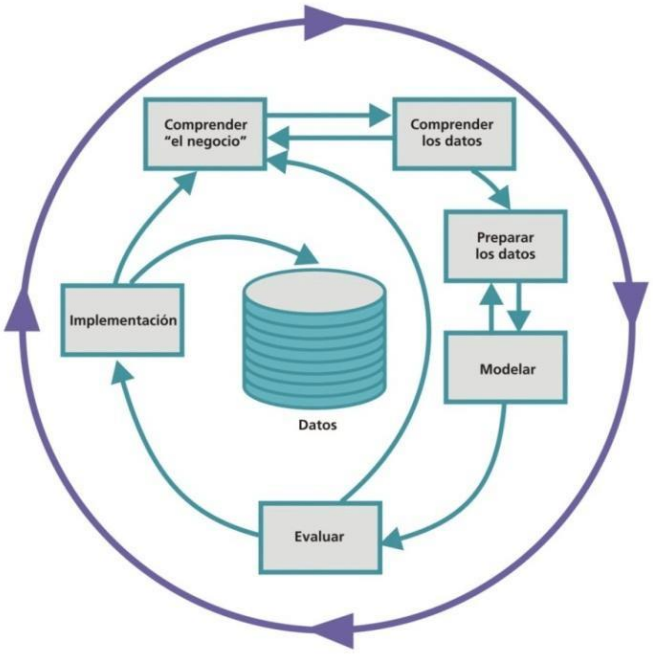
políticas educativas para reducir los efectos negativos asociados a estos procesos (Canales & Ríos, 2018).

Capítulo IV: Metodología

La minería de datos es una disciplina orientada al desarrollo de nuevos métodos y técnicas para explorar datos, por lo cual el presente trabajo es un tipo de investigación predictiva y está vinculado a la información de los estudiantes de la Universidad Técnica de Manabí como un estudio de caso. Debido a esto, el autor (Rodríguez León & García Lorenzo, 2016) menciona que para manejar gran cantidad de datos almacenados y para hacer menos complejos los procedimientos, se han diseñado metodologías que guíen estos procesos, por lo cual un estudio comparativo destaca la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) como la metodología más usada en este tipo de procedimientos, logrando llegar a un nivel de especificación más profundo en cada una de sus seis fases como se muestra en la figura 1:

Figura 1

Fases de la Metodología CRISP – DM



Nota: Tomado de (Cortina, 2015)

Fase 1: Comprensión del Negocio

Como antecedente a la presente investigación se analizó la problemática que se presentaba en la Universidad Técnica de Manabí acerca de la deserción temporal de estudiantes en las diferentes carreras, donde se habían realizados estudios anteriores por personal del Departamento de Bienestar Estudiantil a través de encuestas y procesos manuales en cada unidad académica. Por lo cual esta investigación detalla el proceso de la deserción temporal en estudiantes en relación a la irregularidad que presentan los mismos durante su carrera universitaria.

Objetivos del Negocio (Institución)

El objetivo de esta investigación es realizar predicciones de los estudiantes que se encuentran matriculados en la institución, en relación a las características obtenidas para verificar la regularidad de los estudiantes como un estimador del riesgo de deserción en los mismos.

Valoración de la Situación Actual

La Universidad Técnica de Manabí (UTM), es una universidad pública ubicada en la ciudad de Portoviejo, Manabí y actualmente cuenta con 10 Facultades que se detallan a continuación:

- Ciencias Humanísticas y Sociales
- Ciencias de la Salud
- Ciencias Informáticas
- Ciencias Veterinarias
- Ciencias Zootécnicas
- Ciencias Administrativas y Económicas
- Ciencias Matemáticas, Física y Química

- Ingeniería Agronómica
- Ingeniería Agrícola
- Filosofía, Letras y Ciencias de la Educación

Además, esta institución cuenta con 32 carreras presenciales, 8 carreras de modalidad virtual; y, cuatro Institutos: Instituto de Lenguas, Instituto de Ciencias Básicas, Instituto de Investigación y el Instituto de Posgrado, que ofrece varias especialidades y maestrías para la comunidad universitaria. También cuenta con tres sedes situadas: Lodana en el cantón de Santa Ana, Chone y Bahía de Caráquez.

Así mismo esta institución cuenta con una base de datos con la información académica y social de los estudiantes, por lo que se puede afirmar que se dispone de una gran cantidad de registros para realizar el proceso de minería de datos.

Objetivos de la Minería de Datos

La presente investigación tiene como objetivos del proceso de minería de datos los siguientes puntos:

- Identificar los estudiantes desertores por cada nivel
- Clasificar los estudiantes regulares y los que no son.
- Establecer la correlación de los estudiantes entre deserción e irregularidad académica.
- Descubrir patrones y características de los estudiantes que inciden en la irregularidad académica.
- Generar datos de entrenamiento y evaluación.
- Probar varios algoritmos de clasificación utilizados en el aprendizaje automático.

Plan de Proyecto

El presente trabajo se basa en las fases de la Metodología CRISP-DM, tal como se muestra a continuación para facilitar su organización y estimar el tiempo de realización del mismo:

- Fase 1: Comprensión del negocio

  Tiempo estimado: 1 mes (agosto)
- Fase 2: Comprensión de los datos.

  Tiempo estimado: 1 mes (septiembre)
- Fase 3: Preparación de los datos (selección, limpieza y transformación) para facilitar la minería de datos sobre los mismos.

  Tiempo estimado: 2 meses (octubre - noviembre)
- Fase 4: Modelado (Elección de las técnicas de modelado y ejecución de las mismas sobre los datos).

  Tiempo estimado: 2 meses (diciembre – enero)
- Fase 5: Evaluación (Análisis de los resultados obtenidos en la etapa anterior).

  Tiempo estimado: 1 mes (febrero).
- Fase 6: Implementación (Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos).

  Tiempo estimado: 1mes (marzo).

Evaluación inicial de las herramientas y técnicas

Para realizar el proceso de minería de datos se utilizó el lenguaje de programación Python con el framework Anaconda donde se instaló la versión 2020 de Conda que es parte de la distribución Anaconda3 con Python 3.7 64-Bit, que a través de sus librerías se pudo efectuar la selección y limpieza de los datos. Además, el balanceo

de los mismos a través de la librería SMOTE undersampling y oversampling, así como la utilización de varios algoritmos de clasificación: KNN, Random Forest, SVM con el kernel rfb y las redes neuronales a través de la librería scikit learn utilizando la plataforma Tensorflow que permitió desarrollar análisis más profundos de los datos.

Fase 2: Comprensión de los Datos

En esta segunda fase se realizó la recolección inicial de los datos para analizar y familiarizarse con la información del presente trabajo de investigación, verificando la calidad de los mismos.

Recolección de datos iniciales

Los datos utilizados en este trabajo de investigación son referentes a estudiantes que se encuentran matriculados entre mayo 2014 a febrero 2019, en cada una de las facultades de la UTM, por lo cual para obtener esta información fue necesario realizar una entrevista con el Director del Departamento de Tecnologías de la Información y Comunicación (TIC's), donde se solicitó a través de un oficio la correspondiente autorización para adquirir los datasets de los estudiantes.

Luego de la entrevista realizada y verificando que la población universitaria es elevada, se decidió utilizar como muestra una carrera de cada facultad como se detallan a continuación:

- Facultad de Ciencias de la Salud – Carrera de Medicina
- Facultad de Filosofía, Letras y Ciencias de la Educación – Carrera de Educación Física
- Facultad de Ingeniería Agronómica – Carrera de Ingeniería Agronómica
- Facultad de Ingeniería Agrícola – Carrera de Ingeniería Agrícola

- Facultad de Ciencias Matemáticas, Física y Química – Carrera de Ingeniería Civil
- Facultad de Ciencias Informáticas – Carrera de Ingeniería en Sistemas Informáticos
- Facultad de Ciencias Humanísticas y Sociales – Carrera de Trabajo Social
- Facultad de Ciencias Zootécnicas – Carrera de Ingeniería Zootecnia
- Facultad de Ciencias Veterinaria – Carrera de Acuicultura y Pesquería
- Facultad de Ciencias Administrativas y Económicas – Carrera de Administración

Así mismo se resolvió tomar como referencia para el proceso de minería de datos los periodos académicos detallados a continuación:

- Mayo del 2014 hasta septiembre del 2014
- Octubre del 2014 hasta febrero del 2015
- Mayo del 2015 hasta septiembre del 2015
- Octubre del 2015 hasta febrero del 2016
- Mayo del 2016 hasta septiembre del 2016
- Octubre del 2016 hasta febrero del 2017
- Abril del 2017 hasta: septiembre del 2017
- Octubre 2017 hasta febrero 2018
- Abril del 2018 hasta: agosto del 2018
- Septiembre 2018 hasta: febrero del 2019

Descripción de los datos

Los datos proporcionados por el Departamento de TIC’s fueron exportados en archivos .xls desde el Sistema Gestor de Base de Datos PostgreSQL donde almacenan

la información de los estudiantes de la UTM. Las muestras a ser utilizadas en el proceso de minería de datos por cada una de las carreras se presentan en la tabla 5.

Tabla 5

Muestra detallada de la población

CARRERA	NÚMERO DE ESTUDIANTES
Ingeniería Civil	1941
Medicina	1878
Administración de Empresas	1608
Ingeniería de Sistemas Informáticos	1314
Trabajo Social	1037
Ingeniería Agronómica	584
Ingeniería Agrícola	569
Ingeniería Zootécnica	445
Acuicultura y Pesquería	344
Educación Física, Deportes y Recreación	282
TOTAL DE ESTUDIANTES	**10002**

Además, se utilizaron datos demográficos y académicos de cada uno de los estudiantes que se encontraban matriculados en los diferentes periodos antes mencionados. En las tablas 6 y 7 se detallan las variables que formaron parte de la información proporcionada:

Tabla 6

Descripción de los Datos Demográficos de los estudiantes

N°	Campo	Tipo de Dato	Descripción
1	Cédula	Varchar	Identificador de cada uno de los estudiantes.
2	Fecha_Nacimiento	Date	Establece la fecha de nacimiento y a través de este campo se pudo calcular la edad.
4	Ocupacion_Padre	Varchar	Permite verificar a que se dedica el papá de cada estudiante, entre los datos su mayoría mencionaban agricultores y especificaban su profesión
5	Pais_Origen	Varchar	Registra el país de origen de los estudiantes
6	Provincia_Origen	Varchar	Registra la provincia de origen de los estudiantes
7	Canton_Residencia	Varchar	Registra el cantón de residencia de los estudiantes
8	Tipo_Parroquia	Varchar	Detalla el tipo de parroquia donde reside el estudiante, teniendo como referencia: cabecera cantonal, parroquia rural, parroquia urbana y no definido.
9	Tipo_Zona_Residencia	Varchar	Detalla el tipo de zona donde reside el estudiante, teniendo como referencia: cabecera cantonal, parroquia rural, parroquia urbana y no definido.
10	Genero	Varchar	Establece el género que pertenece el estudiante almacenando si es masculino o femenino
11	Estado_Civil	Varchar	Registra el estado civil del estudiante, entre los datos se establece: soltero, casado, unión libre, divorciado y viudo.
12	Etnia	Varchar	Detalla la etnia a la que pertenece el estudiante, entre los datos se establece: mestizo, afroecuatoriano, blanco, mulato y no definido.

N°	Campo	Tipo de Dato	Descripción
13	Nacionalidad	Varchar	Registra la nacionalidad a la que pertenece el estudiante.
14	Discapacidad	Varchar	Registra el tipo de discapacidad en el caso que tuviera el estudiante.
15	Salario_Mensual_Integrantes	Int	Detalla los integrantes que conforman la familia del estudiante
16	Salario_Mensual	Float	Registra el salario mensual que tiene el jefe de la familia.
17	Enfermedad	Varchar	Registra el tipo de enfermedad en el caso que tuviera el estudiante.
18	Edad	Int	Esta columna fue calculada en relación a la fecha de nacimiento y la fecha que inicia a estudiar en el periodo académico correspondiente, extrayendo el año de cada una de las variables y la diferencia corresponde a la edad de cada estudiante.
19	Tiempo_Cursado	Int	Esta columna fue calculada con la fecha en que se graduó del colegio con la fecha en que inicia a estudiar en la UTM, luego se extrajo el año en cada variable y la diferencia corresponde al Tiempo_Cursado.

Tabla 7

Descripción de los Datos Académicos de los estudiantes

N°	Campo	Tipo de Dato	Descripción
1	Tipo_Colegio	Varchar	Corresponde al tipo de colegio que curso el estudiante, teniendo como referencia los siguientes valores: fiscal, particular, fiscomisional y municipal.
2	Colegio_Fecha_Titulo	Date	Registra la fecha en la que cada estudiante se graduó del colegio y sirve como referencia para calcular la variable Tiempo_Cursado.

N°	Campo	Tipo de Dato	Descripción
3	Colegio_Calificacion	Float	Detalla la calificación con la que se graduó del colegio el estudiante.
4	Tipo_Parroquia_Colegio	Varchar	Registra la parroquia donde se encontraba localizado el colegio donde estudió el estudiante
5	IdNivel_Materia	Int	Identificador del nivel que corresponde cada materia.
6	IdMateria	Int	Identificador de la materia.
7	Materia	Varchar	Registra la materia correspondiente a cada nivel que tiene que cursar el estudiante.
8	Tipo_Materia	Varchar	Detalla el tipo de materia categorizándose de la siguiente manera: competencias generales, competencias básicas de carrera y competencias de especialidad de carrera.
9	Nota_Final	Float	Registra la nota final del estudiante por cada materia cursada.
10	Asistencia	Float	Detalla el porcentaje de asistencia que registra el estudiante por cada materia durante el periodo académico correspondiente.
11	Supletorio	Float	Establece la nota de supletorio por cada materia, en el caso que el estudiante no complete el rango correspondiente de la nota.
12	Aprobado	Varchar	Registra en abreviatura si el estudiante aprobó la materia.
13	Observacion	Varchar	Detalla los siguientes datos: acreditado y no acreditado correspondiente por cada materia.

N°	Campo	Tipo de Dato	Descripción
14	Periodo	Varchar	Registra el periodo académico correspondiente que se encuentra el estudiante.
15	Fecha_Inicio_Periodo	Date	Detalla la fecha de inicio del periodo académico.
16	Fecha_Final_Periodo	Date	Detalla la fecha final del periodo académico.
17	Revalidado	Varchar	Registra si el estudiante ha revalidado esa materia, en el caso que hubiera venido de otra Institución de Educación Superior.
18	Nivel_Estudiante	Int	Corresponde al nivel que cursa el estudiante.
19	Fecha_Egreso	Date	Registra la fecha que egreso el estudiante.
20	Veces_Asignatura	Int	Detalla el número de veces que el estudiante curso cada materia.
21	Malla	Varchar	Corresponde al nombre de la malla que cursa el estudiante.
22	IdMalla	Int	Identificador de la malla que cursa el estudiante.
23	Carrera	Varchar	Registra la carrera que cursa el estudiante.
24	Facultad	Varchar	Registra la facultad que cursa el estudiante.

Exploración de datos

Luego de la descripción de los datos, se procede a la exploración de los mismos, aplicando pruebas estadísticas a través de la herramienta Pandas Profiling que detalla las propiedades y permite crear gráficos de distribución de los datos de los estudiantes de las facultades. Además, este informe sirve principalmente para determinar la

consistencia y completitud de la información, por lo que se han obtenido algunos de los datos que se muestran a continuación:

- En la figura 2 se detalla el número de estudiantes en cada una de las facultades con un total de 10002 alumnos, donde se muestra que en la Facultad de Ciencias Matemáticas, Físicas y Químicas existe el mayor porcentaje de estudiantes (19,4%).

Figura 2

Detalle de los estudiantes por cada una de las Facultades

FACULTAD
Categorical

Distinct count	10
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	0

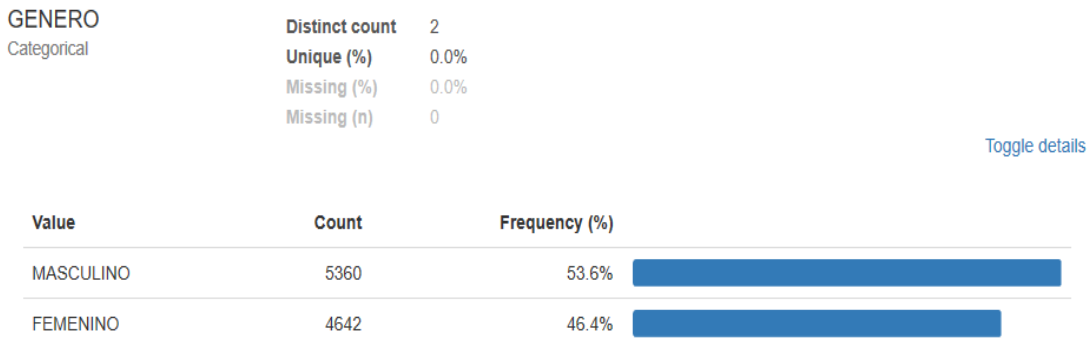
Toggle details

Value	Count	Frequency (%)
CIENCIAS MATEMÁTICAS FÍSICAS Y QUÍMICAS	1941	19.4%
CIENCIAS DE LA SALUD	1878	18.8%
CIENCIAS ADMINISTRATIVAS Y ECONÓMICAS	1608	16.1%
CIENCIAS INFORMÁTICAS	1314	13.1%
CIENCIAS HUMANISTICAS Y SOCIALES	1037	10.4%
INGENIERIA AGRONOMICA	584	5.8%
INGENIERIA AGRÍCOLA	569	5.7%
CIENCIAS ZOOTÉCNICAS	445	4.4%
CIENCIAS VETERINARIAS	344	3.4%
FILOSOFÍA LETRAS Y CIENCIAS DE LA EDUCACIÓN	282	2.8%

- En la Figura 3 se describe el total de estudiantes de acuerdo al género al que corresponden, observando un mayor porcentaje en hombres (53,6%) que en mujeres (46,4%)

Figura 3

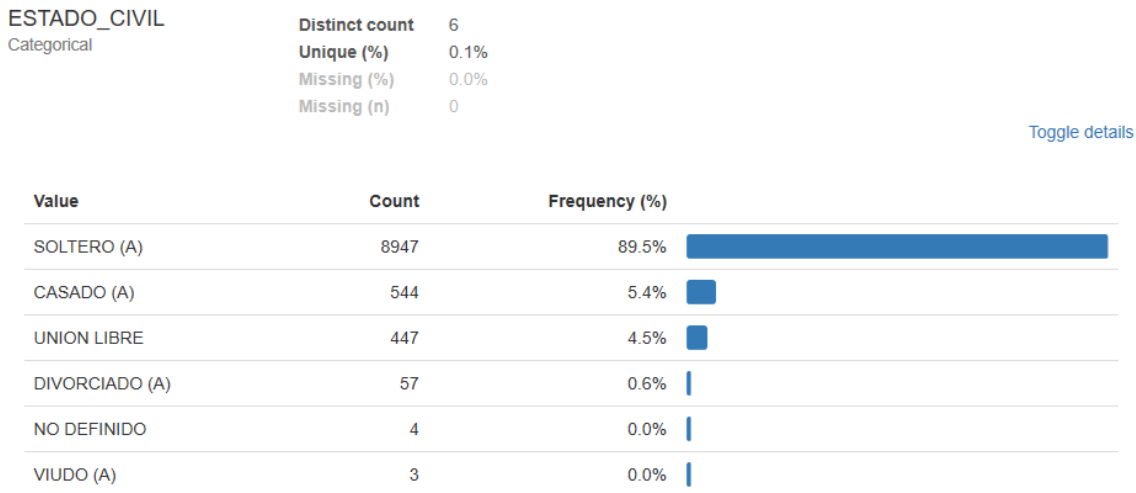
Descripción por género de los estudiantes



- En la figura 4 se describe el estado civil que registraron cada uno de los estudiantes, donde se muestra un mayor porcentaje en personas solteras (89,5%).

Figura 4

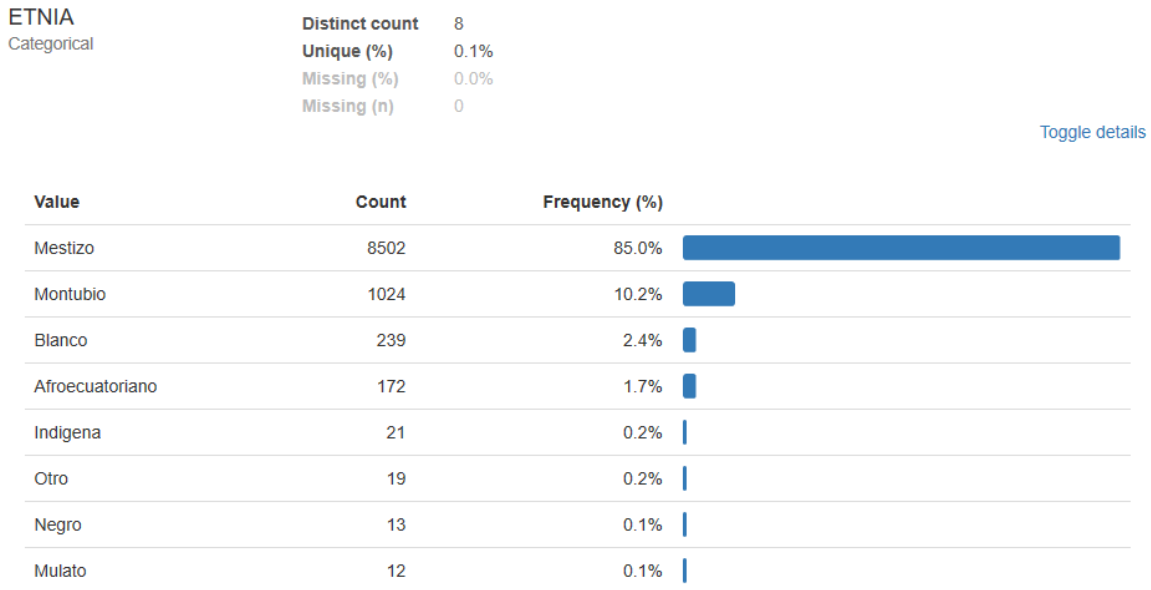
Distribución por estado civil de cada uno de los estudiantes



- En la figura 5 se describe la etnia que registraron cada uno de los estudiantes, observando un mayor porcentaje en las personas con etnia mestiza (85,00%)

Figura 5

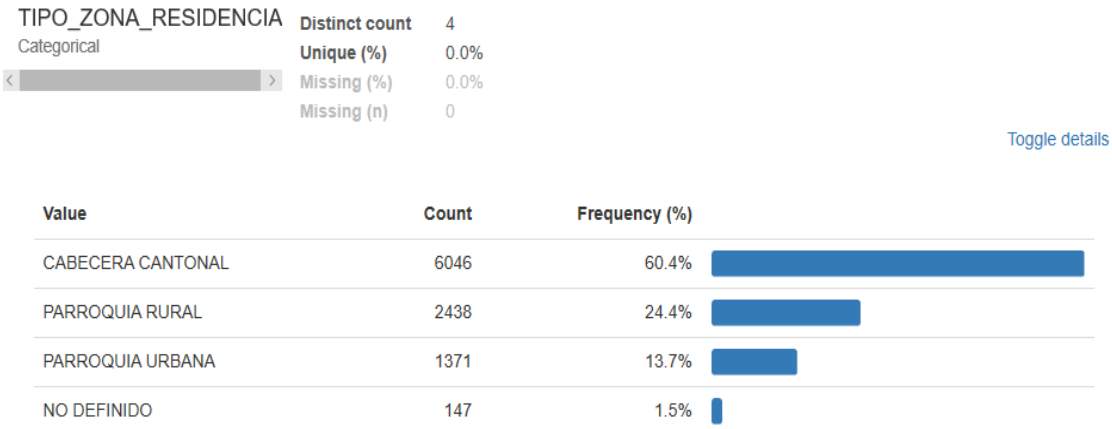
Distribución por etnia de cada uno de los estudiantes



- En la figura 6 se describe la zona de residencia que registraron cada uno de los estudiantes, siendo la cabecera cantonal el mayor índice de residencia con el 60,40%.

Figura 6

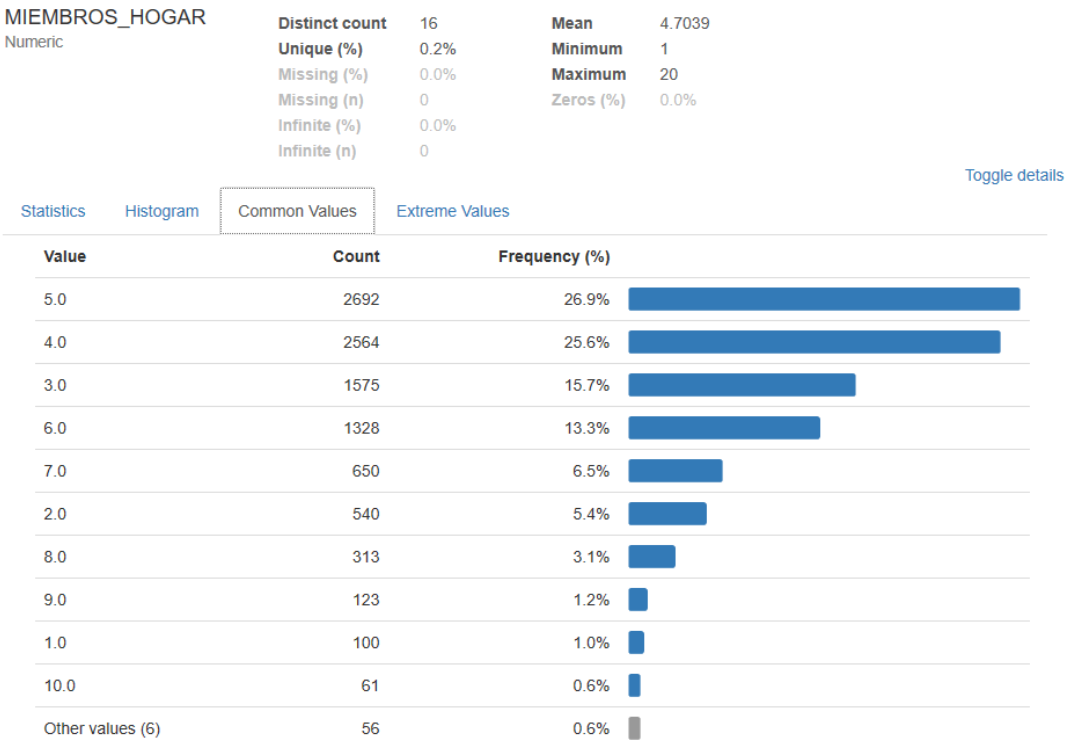
Distribución por etnia de cada uno de los estudiantes



- En la figura 7 se describe la cantidad de miembros que hay en el hogar de los estudiantes, constatando que un número de cinco personas equivale aun 26,9% siendo este el valor con mayor realce.

Figura 7

Distribución por miembros del hogar de cada uno de los estudiantes



MIEMBROS_HOGAR
Numeric

Distinct count	16	**Mean**	4.7039
Unique (%)	0.2%	**Minimum**	1
Missing (%)	0.0%	**Maximum**	20
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

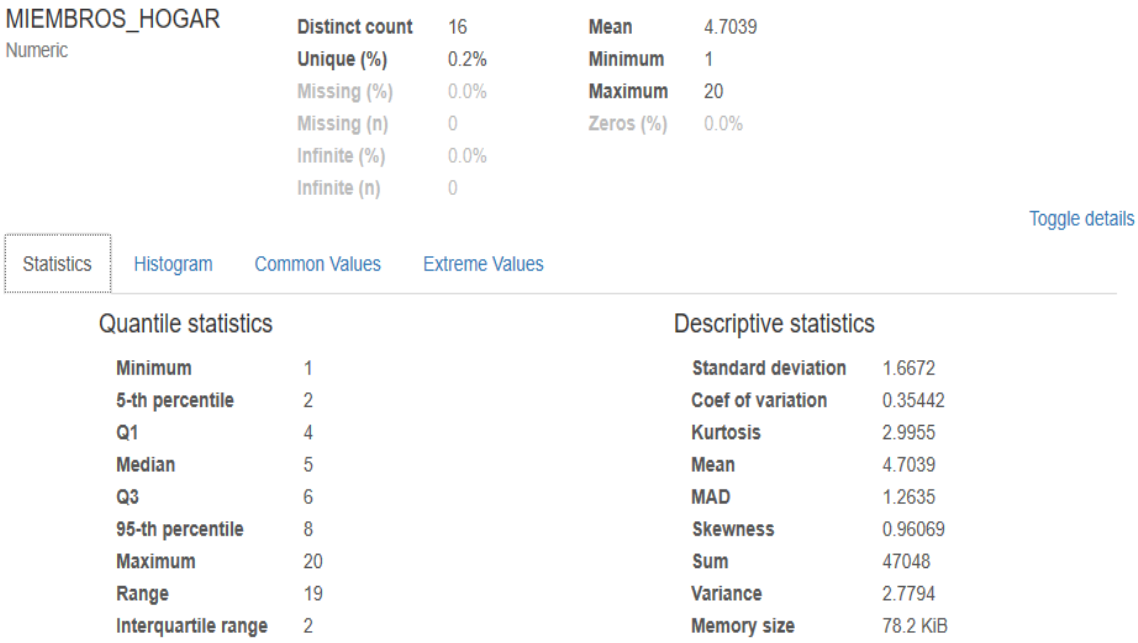
Statistics | Histogram | Common Values | Extreme Values

Value	Count	Frequency (%)
5.0	2692	26.9%
4.0	2564	25.6%
3.0	1575	15.7%
6.0	1328	13.3%
7.0	650	6.5%
2.0	540	5.4%
8.0	313	3.1%
9.0	123	1.2%
1.0	100	1.0%
10.0	61	0.6%
Other values (6)	56	0.6%

- En la figura 8 se describe de manera estadística los miembros que hay en el hogar de los estudiantes.

Figura 8

Distribuciones estadísticas de los miembros del hogar de cada uno de los estudiantes



Además, se hicieron varias consultas SQL al Sistema Gestor de Base de Datos de la universidad, que fueron necesarias para el proceso de minería de datos debido a la cantidad de registros y al cruce de tablas para obtener la información, conseguida en archivos .xls donde se registraron 286798 registros de estudiantes por cada asignatura.

Verificación de la calidad de los datos

Luego de realizar la exploración de los datos se verificaron que existen algunas observaciones en los registros que se detallan a continuación:

- Identificación correcta de cada uno de los registros de los estudiantes.
- Eliminación de registros nulos de las columnas: fecha_nacimiento, salarios_mensual_integrantes, fecha_egreso, enfermedad, ocupación_madre, ocupación_padre y discapacidad.

- Registro de la calificación_colegio, notal_final, asistencia y supletorio se encontraban fuera del rango establecido, es decir había valores que no se encontraban entre 1 y 10.
- Eliminación de los campos que tenían más del 60% de registros con valores nulos, en este caso las columnas que se eliminaron fueron: ocupación_padre, ocupación_madre, fecha_egreso, enfermedad y discapacidad.
- Eliminación de la columna aprobado, ya que el campo Observación contiene los mismos datos y expresado por categorías de una mejor manera.
- Se suprimieron los campos país_origen, provincia_origen, cantón_residencia debido de registros que contenían esta columna.
- La columna tipo_Parroquia fue eliminada porque en el campo tipo_zona_residencia se almacenaban los mismos datos.

Fase 3: Preparación de los Datos

Esta fase de la metodología permite preparar los datos, para luego seleccionarlos en las técnicas de minerías de datos. Esto implica escoger el subconjunto de datos que se va a utilizar, hacerle la limpieza correspondiente mejorando su calidad, ingresar nuevos datos si el caso lo amerita y darles el formato requerido por la herramienta de modelado.

Selección de los datos

En esta etapa de la metodología se hizo la selección de un subconjunto de datos, calculando como un estimador del riesgo de deserción en un periodo de tiempo, la condición de la regularidad académica de los estudiantes, debiendo cumplir al menos el 60% de asignaturas matriculadas del total del nivel.

Por lo antes expuesto esto se basa según lo establecido en el artículo 14 del reglamento de Régimen Académico del Consejo de Educación Superior y el artículo 10 del reglamento de Régimen Académico de la Universidad Técnica de Manabí vigente.

En base a este análisis se tomó como criterios de inclusión lo que se muestra a continuación:

- Se incluye si matriculó al menos el 60% de asignaturas del nivel actual.
- La información académica debe calcularse en relación al nivel anterior es decir solamente durante un periodo académico.
- Se debe incluir la información social de cada uno de los estudiantes.
- La etiqueta a tomar en cuenta es: ¿Es estudiante regular del siguiente nivel matriculando el 60% de asignaturas?

Limpieza de los datos

En la presente investigación se utilizaron datos en archivos .xls, los cuales fueron obtenidos a través de consultas SQL hacia la base de datos de la universidad, a través de las herramientas que presenta Python en la cual se hizo la limpieza de cada uno de los datos obtenidos de la siguiente manera:

- Se normalizó la columna de colegio_calificacion debido que como en los periodos académicos que se escogió para el proceso de minería de datos, existían calificaciones hasta 20 y en la actualidad el rango de las notas es de 0 a 10, por lo cual se sacó el promedio de dichas notas y quedaron solo calificaciones en ese rango.
- Se eliminaron los registros de asistencias y supletorio que no se encontraban en el rango de 0 a 10.

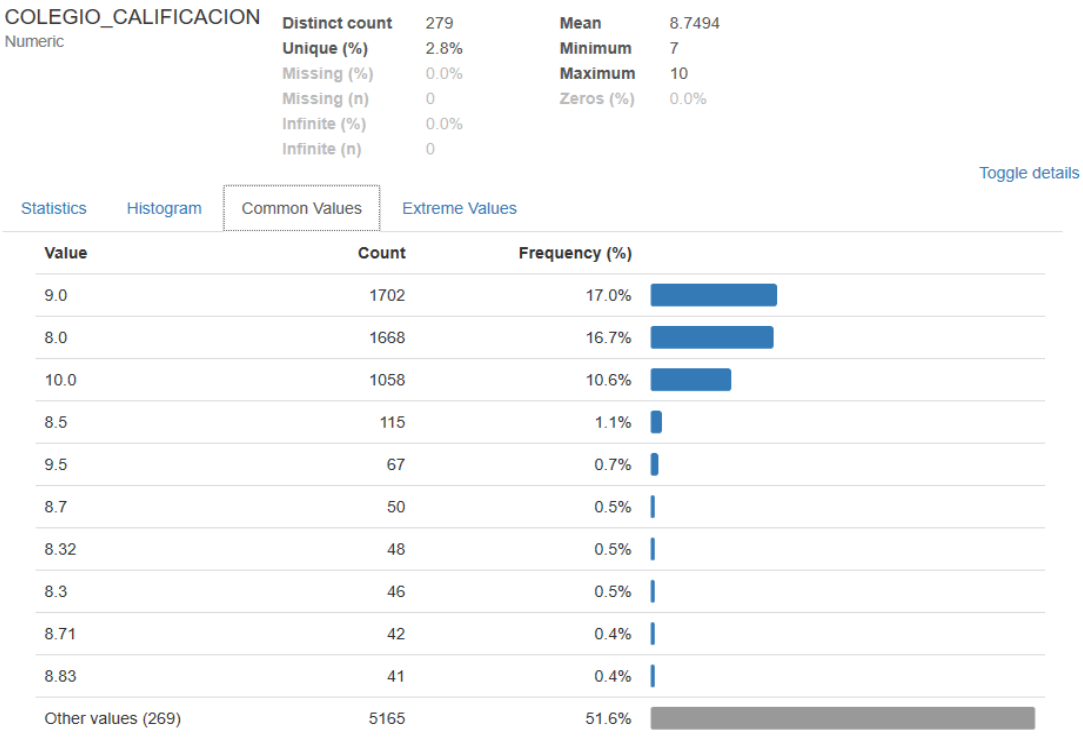
- Se eliminaron los registros de la columna estado_civil que no estaban definidos entre el rango especifico.
- Se reemplazaron los valores negativos de las columnas nota_final, asistencia y supletorio por cero, debido que por error involuntario fueron registrados de esa manera.
- La columna Observacion quedó con dos categorías acreditado y no acreditado de cada una de las materias, por lo cual los registros que tenían valores diferentes a los antes mencionados y según consultas con el administrador de la base de datos, decía que se generaban por error involuntario, quedando reemplazados como no acreditados.

Luego del proceso de limpieza de los datos el número de estudiantes total por las facultades que forman parte de la muestra es de 9994 y 286728 registros por cada asignatura, quedando de la siguiente manera los datos académicos más relevantes:

- En la figura 9 se describe la calificación general que registraron los estudiantes al ingresar a la universidad, teniendo como referencia el puntaje de nueve con el 17% de su totalidad.

Figura 9

Descripción de las calificaciones generales del colegio que registraban los estudiantes



COLEGIO_CALIFICACION
Numeric

Distinct count	279	**Mean**	8.7494
Unique (%)	2.8%	**Minimum**	7
Missing (%)	0.0%	**Maximum**	10
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Toggle details

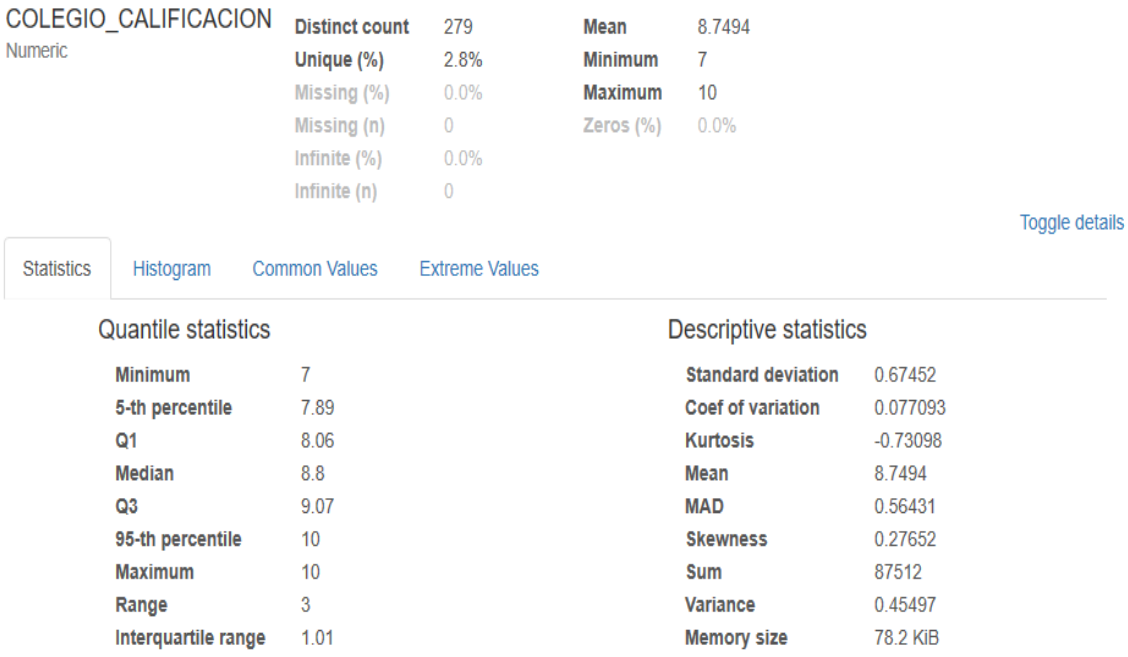
Statistics | Histogram | Common Values | Extreme Values

Value	Count	Frequency (%)
9.0	1702	17.0%
8.0	1668	16.7%
10.0	1058	10.6%
8.5	115	1.1%
9.5	67	0.7%
8.7	50	0.5%
8.32	48	0.5%
8.3	46	0.5%
8.71	42	0.4%
8.83	41	0.4%
Other values (269)	5165	51.6%

- En la figura 10 se describe estadísticamente la variable colegio_calificación donde registraron los estudiantes al ingresar a la universidad

Figura 10

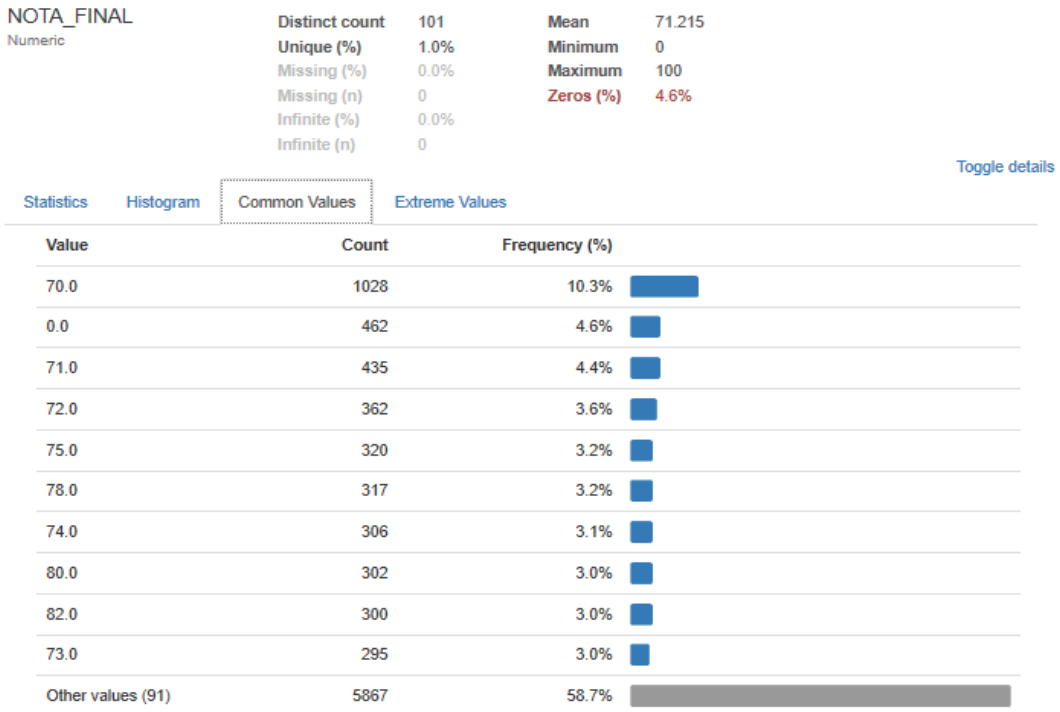
Análisis estadísticos de las calificaciones generales del colegio de los estudiantes



- En la figura 11 se describe la nota final de las asignaturas de cada uno de los estudiantes y en la figura 12 los valores estadísticos de la variable.

Figura 11

Análisis de las notas finales de las asignaturas de cada uno de los estudiantes



NOTA_FINAL
Numeric

Distinct count	101	Mean	71.215
Unique (%)	1.0%	Minimum	0
Missing (%)	0.0%	Maximum	100
Missing (n)	0	Zeros (%)	4.6%
Infinite (%)	0.0%		
Infinite (n)	0		

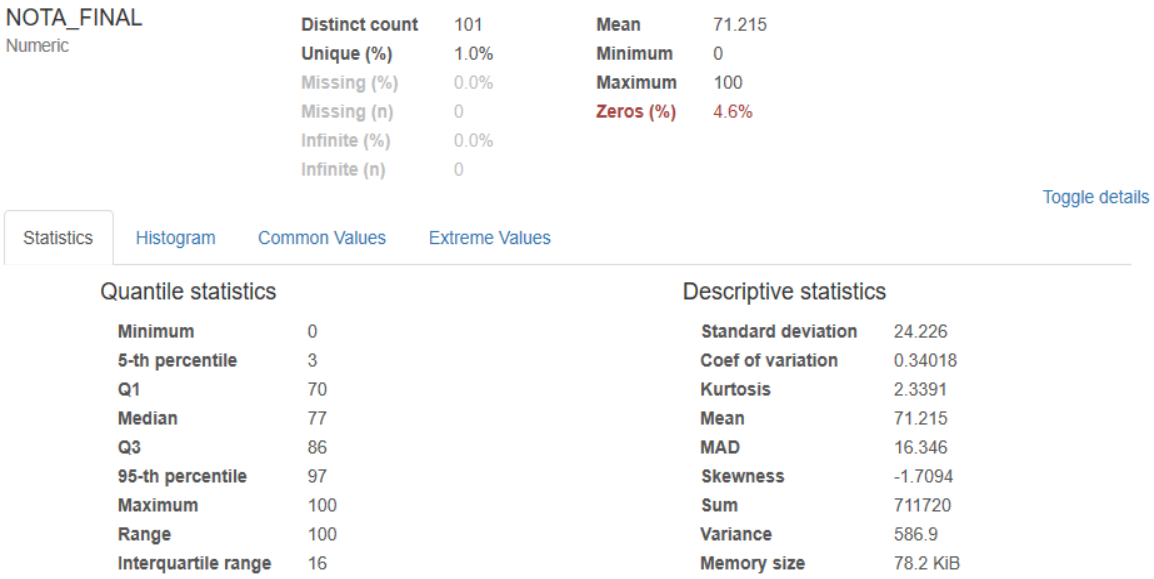
Toggle details

Statistics | Histogram | Common Values | Extreme Values

Value	Count	Frequency (%)
70.0	1028	10.3%
0.0	462	4.6%
71.0	435	4.4%
72.0	362	3.6%
75.0	320	3.2%
78.0	317	3.2%
74.0	306	3.1%
80.0	302	3.0%
82.0	300	3.0%
73.0	295	3.0%
Other values (91)	5867	58.7%

Figura 12

Análisis estadísticos de las notas finales de las asignaturas de los estudiantes



- En la figura 13 se describe la calificación de supletorio de los alumnos que en su momento tuvieron que registrar esta nota al no cumplir la nota final de las asignaturas y en la figura 14 su representación de manera estadística

Figura 13

Análisis de las calificaciones del supletorio de las asignaturas de los estudiantes

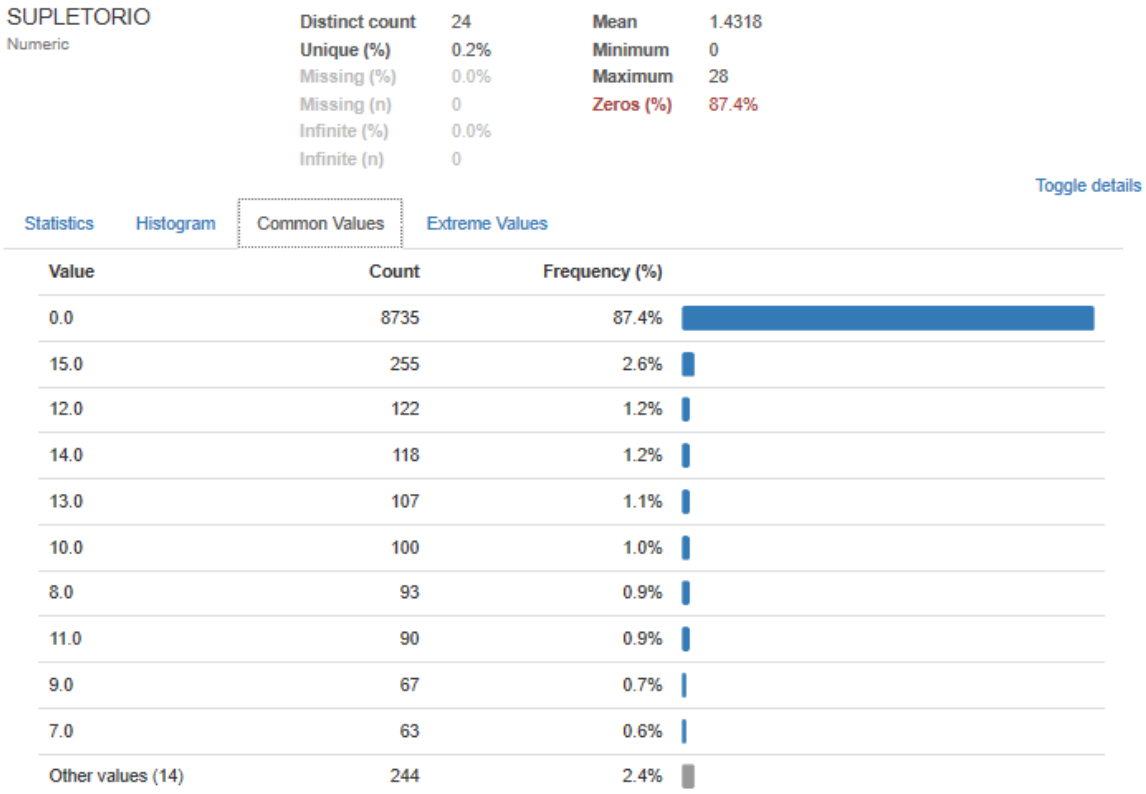
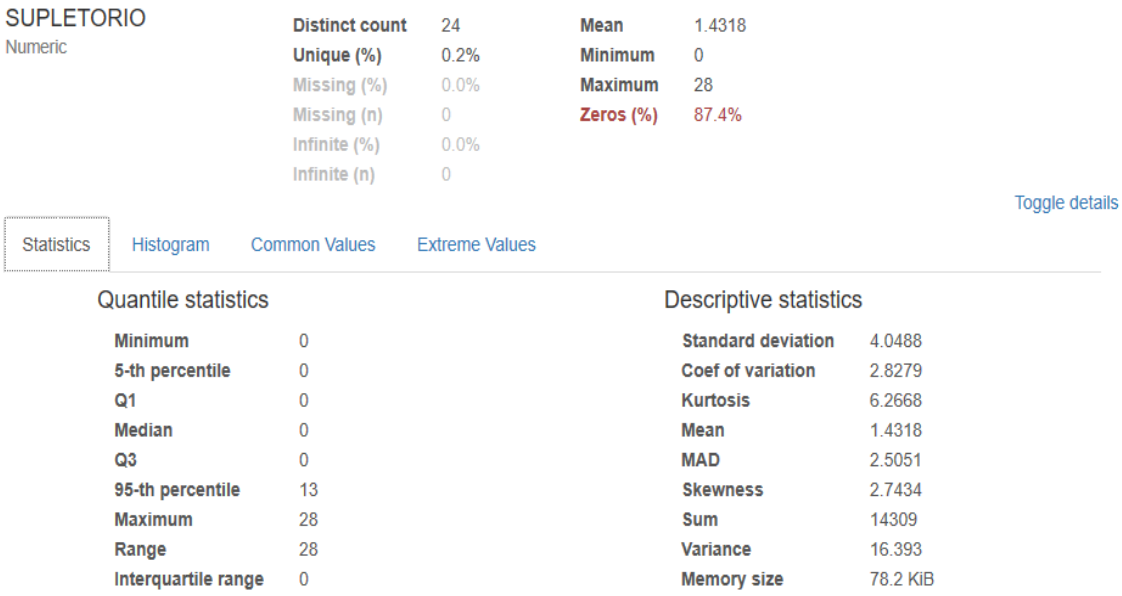


Figura 14

Análisis estadístico de las calificaciones del examen de supletorio de las asignaturas de los estudiantes



Estructura de los datos

A través de esta fase se han generado atributos a partir de otros campos para mejorar el proceso de minería da datos, detallándose a continuación:

- Se extrajeron solamente el año de las columnas colegio_fecha_titulo, fecha_nacimiento y fecha_inicio_periodo a través del importe de la herramienta datetime, utilizando la opción. year.

Integración de los datos

A partir de la extracción del año que se realizó en la fase de estructura de datos, se agregaron las siguientes columnas:

- El campo Edad que se calculó de la diferencia entre la fecha de nacimiento y la fecha que inicio el periodo académico al ingresar a la universidad.

- El campo tiempo_cursado es la diferencia entre la fecha que culminó los estudios del colegio y la fecha de inicio del periodo académico al ingresar por primera vez a la universidad.

Además, tomando como referencia los criterios de inclusión detallados en la selección de los datos, se agregaron los siguientes campos:

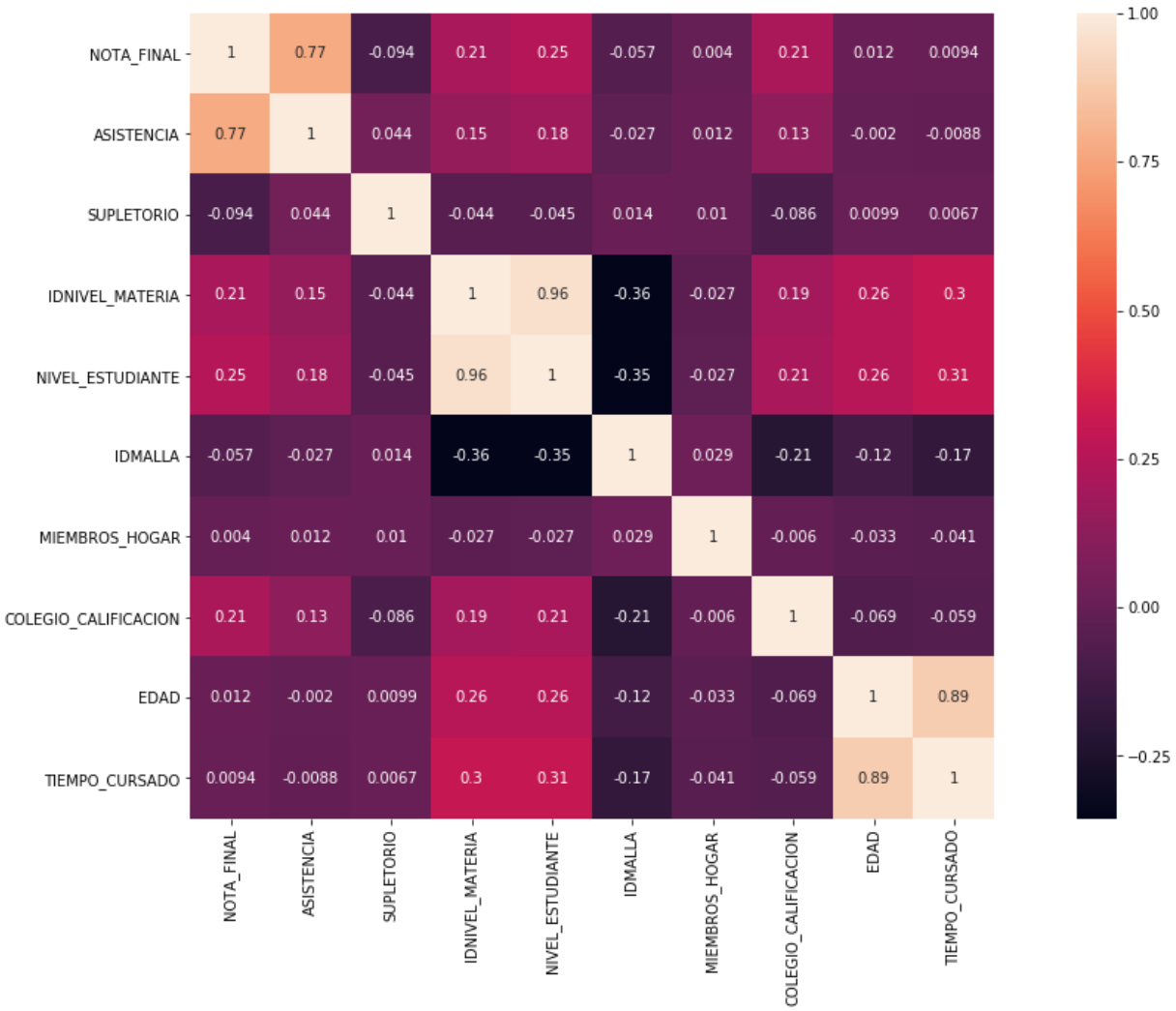
- Promedio del campo de nota_final y asistencia de todas las asignaturas del nivel anterior.
- Desviación estándar del campo de nota_final y asistencia de todas las asignaturas del nivel anterior.
- Total_supletorio, es decir un totalizado de las veces que estuvo en supletorio durante el nivel anterior en cada una de las asignaturas.
- NMateriasAprobadas, hace referencia al total de asignaturas aprobadas durante el nivel anterior.
- NMateriasReprobadas, hace referencia al total de asignaturas reprobadas durante el nivel anterior.
- Regular2, calcula si el estudiante es regular en el nivel actual, solo almacena en el caso que lo sea caso contrario se descarta.
- Regular3, calcula si el estudiante es regular o no en el siguiente nivel, siendo el target de la investigación, por lo cual es un campo de tipo binario, es decir, si cumple la condición es 1 caso contrario es 0.

Formateo de los datos

En esta fase de la metodología se toman en cuenta la correlación de las variables a través de herramientas de Python, por lo cual en primer instancias se hizo un análisis preliminar detallado en la figura 15.

Figura 15

Análisis estadístico de correlación de las variables



Según los datos de la figura 15, existe una correlación entre la nota final con la variable asistencia de 0,77 lo que define que en general la asistencia a las clases es un factor determinante en los resultados de los docentes y por tanto en la retención.

Además, se utilizó la herramienta de Python statsmodels.api que incluye un proceso matemático dentro de una función en la cual verifica la correlación de las variables categorizándolas y viendo su importancia.

Luego de la correlación de las variables se utilizaron los métodos de codificación numérica de variables categóricas como se detalla a continuación:

- La columna género utilizó la función Pandas pd.get_dummies, donde cambia la variable de tipo numérico representado entre 1 Masculino y 0 Femenino.
- La columna Estado_Civil, TipoZonaResidencia y Tipo_Colegio para la categorización de estas variables a través de la herramienta LabelEncoder(), que convierte estos campos en valores numéricos dependiendo del número de categorías.

Fase 4: Modelado

En esta fase de la metodología se escogieron las técnicas más apropiadas para el proceso de minería de datos, donde se detallan cada una de las variables para aplicar con las técnicas seleccionadas como lo muestra la tabla 8.

Tabla 8

Descripción de los variables generales para aplicar las técnicas de minería de datos

N°	Variable	Tipo de Dato	Descripción
1	Colegio_Calificacion	Float	Nota final que se graduaron cada uno de los estudiantes al culminar sus estudios del colegio
2	DesvAsistencia	Float	Desviación estándar del total de asistencia de las asignaturas vistas durante el periodo académico por cada uno de los estudiantes
3	DesvNotaFinal	Float	Desviación estándar del total la nota final de las asignaturas vistas durante el periodo académico por cada uno de los estudiantes
4	Edad	Int	Edad que tiene el estudiante cuando ingresa a la universidad

N°	Variable	Tipo de Dato	Descripción
5	Estado_Civil	Int	Estado civil que registra el estudiante entre ellos se encuentran los siguientes valores: soltero, casado, divorciado, unión libre y viudo
6	Miembros_Hogar	Int	Total de miembros que tiene el hogar de cada uno de los estudiantes
7	NMateriasAprobadas	Int	Total de asignaturas aprobadas durante el periodo académico de cada uno de los estudiantes
8	NMateriasReprobadas	Int	Total de asignaturas reprobadas durante el periodo académico de cada uno de los estudiantes
9	PromAsistencia	Float	Promedio del total de asistencia de las asignaturas vistas durante el periodo académico por cada uno de los estudiantes
10	PromNotaFinal	Float	Promedio total de la nota final de las asignaturas vistas durante el periodo académico por cada uno de los estudiantes
11	Tiempo_Cursado	Int	Tiempo transcurrido desde que culmino sus estudios de colegio hasta que ingresa a la universidad
12	Tipo_Colegio	Int	Tipo de colegio que estudió cada uno de los estudiantes entre ellos se incluyen los siguientes valores: fiscal, particular, fiscomisional y municipal
13	TipoZonaResidencia	Int	Lugar de residencia de los estudiantes entre ellos se presentan los siguientes valores: cabecera cantonal, parroquia rural, parroquia urbana y no definido
14	Total_Supletorio	Int	Total de veces que el estudiante se quedó en supletorio durante el periodo académico
15	Genero	Int	Género de cada uno de los estudiantes: Masculino y Femenino.

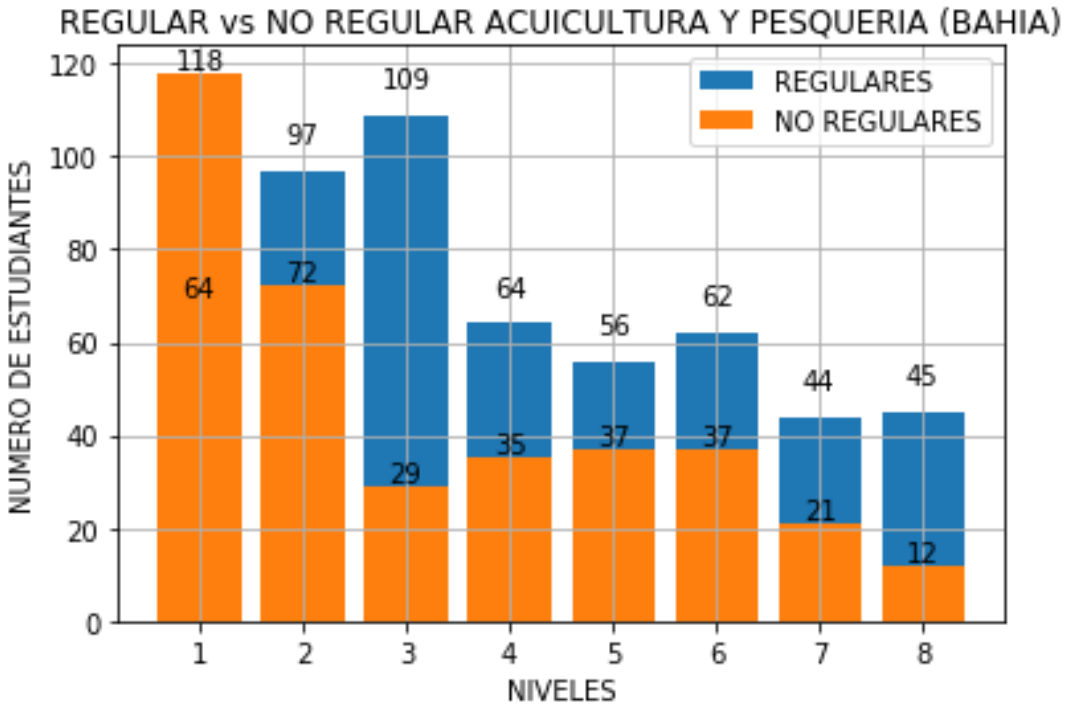
Selección de Técnicas

Luego de realizar la limpieza de los datos y teniendo las variables para aplicar las técnicas de minería de datos, fue necesario establecer la correlación entre los estudiantes no regulares, es decir que no matriculan al menos el 60% de asignaturas correspondientes al nivel que pertenecen; y los estudiantes desertores o que abandonan sus estudios en un periodo de tiempo determinado, como un estimador del riesgo de deserción en cada uno de los niveles.

Para el proceso realizado se tomó como muestra la facultad de Ciencias Veterinarias, con la carrera de Acuicultura y Pesquería, debido al correcto manejo de la información almacenada. La figura 16 muestra el gráfico representativo de la regularidad de los alumnos en cada uno de los niveles de la carrera correspondiente:

Figura 16

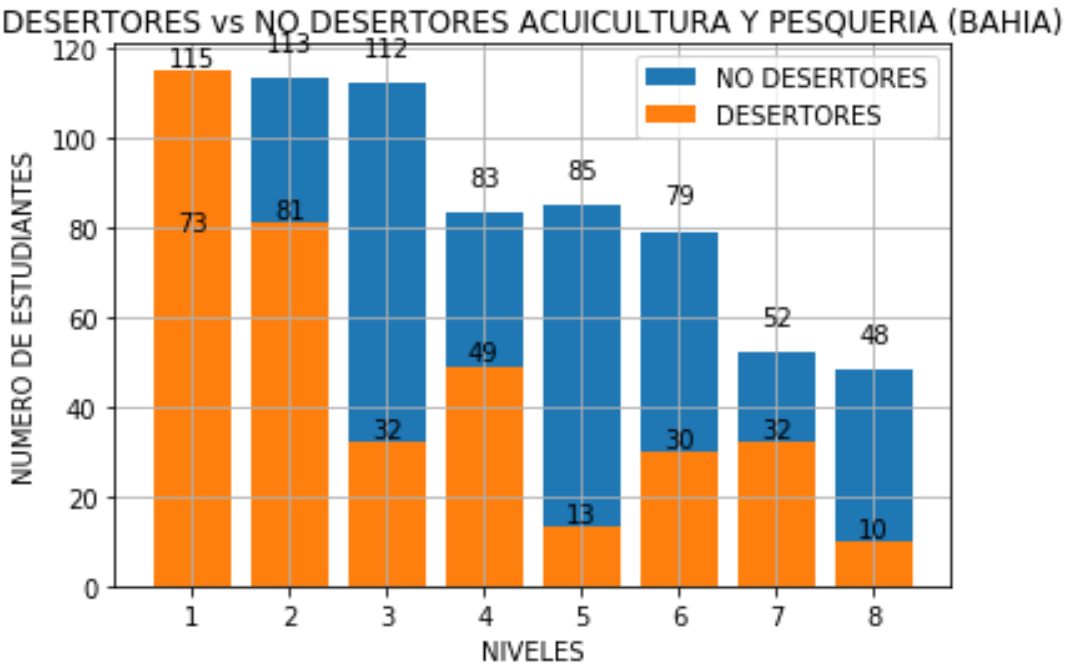
Análisis de la irregularidad de los estudiantes en la Carrera de Acuicultura y Pesquería



De manera similar la figura 17 muestra los valores de deserción por cada uno de los niveles de la carrera.

Figura 17

Análisis de la deserción de los estudiantes en la Carrera de Acuicultura y Pesquería



Además, se calcularon los valores de deserción correspondiente a cada nivel, mostrados en la tabla 9 en conjunto con la regularidad:

Tabla 9

Descripción de los datos de deserción e irregularidad de los estudiantes

Nivel	Regulares	No Regulares	Desertores	No Desertores
1	64	118	115	73
2	97	72	81	113
3	109	29	32	112
4	64	35	49	83
5	56	37	13	85
6	62	37	30	79
7	44	21	32	52
8	45	12	10	48

Debido a esto a través del coeficiente de correlación de Pearson con los datos obtenidos por cada nivel, se comprobó la correlación con un valor correspondiente a 0.94, entre los estudiantes no regulares y los estudiantes desertores como estimación del riesgo de deserción, como se muestra en la tabla 10:

Tabla 10

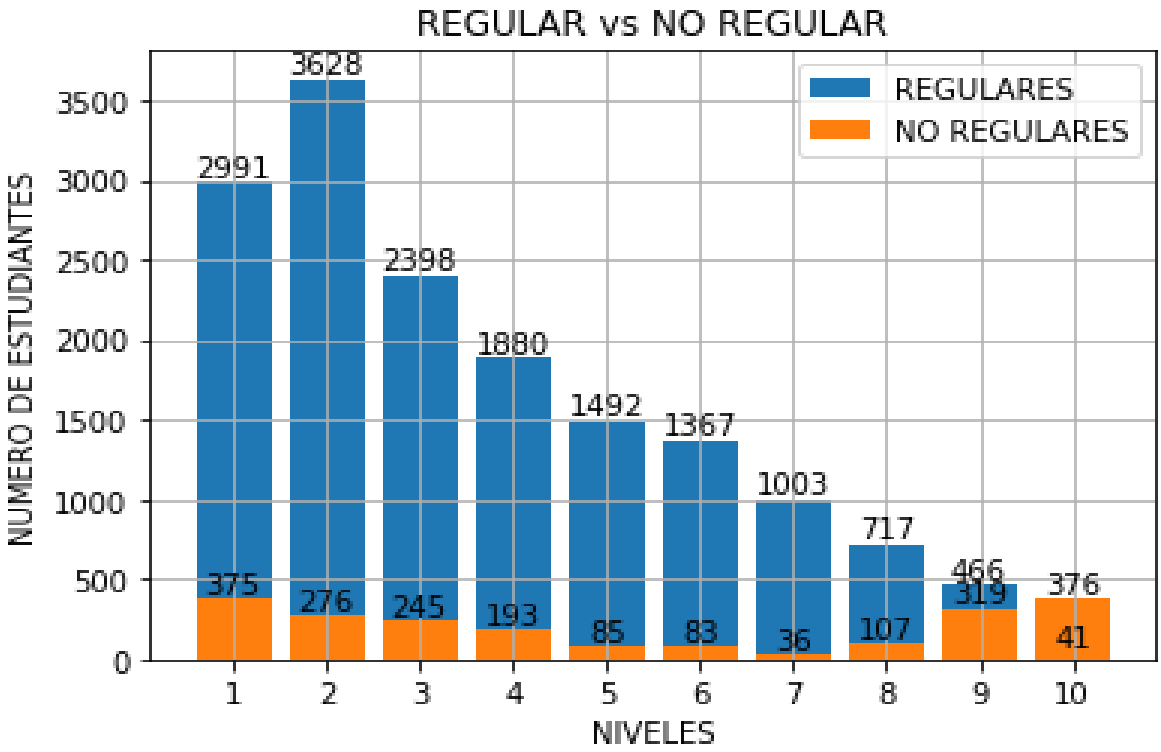
Matriz de Correlación de Pearson entre la deserción e irregularidad de los estudiantes

	Regulares	**No Regulares**	**Desertores**	**No Desertores**
Regulares	1	0.241326	0.307019	0.929028
No Regulares	0.241326	1	0.940415	0.267564
Desertores	0.307019	0.940415	1	0.271540
No Desertores	0.929028	0.267564	0.271540	1

Luego del análisis realizado y comprobando la correlación entre los desertores y los estudiantes no regulares, se hicieron la selección de las técnicas de minería de datos que más se ajuste a la presente investigación. Sin embargo, debido al desbalance de la información sobre la regularidad de los estudiantes en las facultades durante los periodos académicos seleccionados, fue necesario realizar otros procesos antes de aplicar estas técnicas y definir cada uno de los escenarios, el detalle se muestra en la figura 18.

Figura 18

Análisis de la regularidad de los estudiantes en cada uno de los niveles



A través de las figuras 19, 20 y 21 se detalla la regularidad de los estudiantes en los primeros niveles, que tiene mayor énfasis en el segundo nivel debido a que muchos alumnos cambian de universidades y por ello el número de estos resultados es mayor.

Además, es notorio que el porcentaje de irregularidad de los estudiantes a partir del octavo nivel es superior porque desde estos niveles los estudiantes en función de la malla curricular en que se encuentren y la carrera que sigan, empiezan el proceso de titulación correspondiente y como muchos de ellos adelantan asignaturas, existe un porcentaje mínimo de materias que toman en esos niveles y no cumplen los criterios de inclusión definido en las anteriores fases de la metodología.

Por lo cual luego de este análisis se constató el desbalance de las clases mayoritarias y minoritarias, creando tres escenarios como se muestra en las figuras 19, 20 y 21.

- **Primer escenario (Niveles 1,2,3)**
  - **Estudiantes Regulares:**

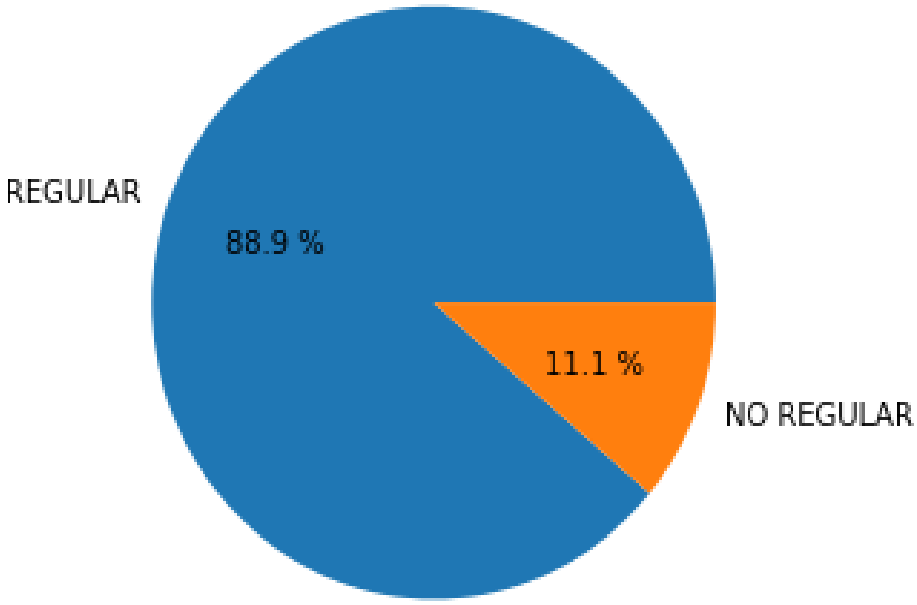
    2991 – 88.9%

  - **Estudiantes No Regulares:**

    375 – 11.1%

Figura 19

Descripción de los estudiantes regulares e irregulares (Nivel 1,2,3)



- **Segundo escenario (Niveles 2,3,4)**
  - **Estudiantes Regulares:**

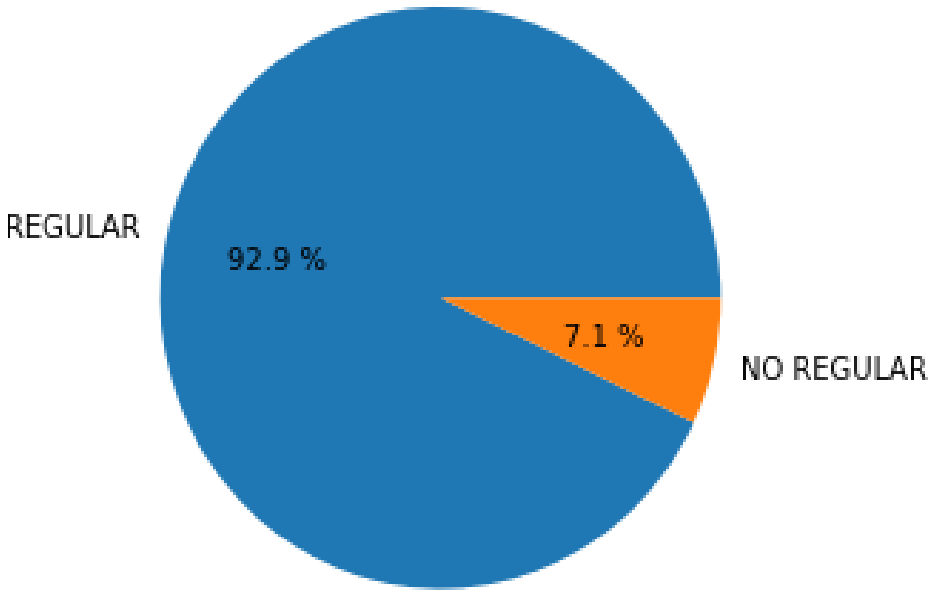
    3628 – 92.9%

  - **Estudiantes No Regulares:**

    276 – 7.1%

Figura 20

Descripción de los estudiantes regulares e irregulares (Nivel 2,3,4)



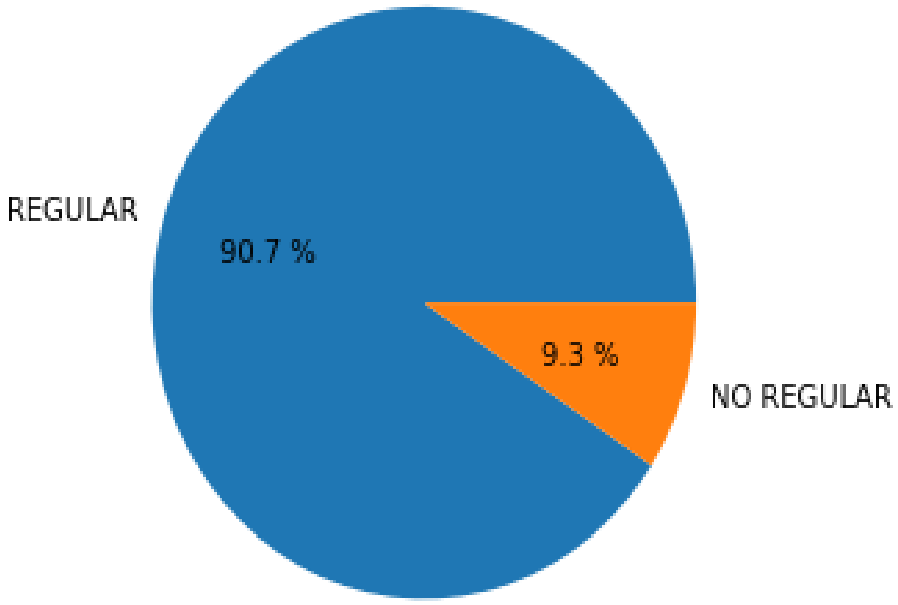
- **Tercer escenario (Niveles 3,4,5)**
  - **Estudiantes Regulares:**

    2398 – 90.7%
  - **Estudiantes No Regulares:**

    245 – 9.3%

Figura 21

Descripción de los estudiantes regulares e irregulares (Nivel 3,4, 5)



Para el proceso de desbalance de datos fue necesario verificar la correlación de las variables a través de la herramienta statsmodels que es un módulo de Python que proporciona clases y funciones para la estimación de muchos modelos estadísticos. Luego se utilizaron los métodos de remuestreo que permitieron eliminar instancias de la clase mayoritaria a través del under-sampling y los que generan nuevas instancias de la clase minoritaria utilizando el over-sampling con las siguientes herramientas que proporciona Python:

- from imblearn.over_sampling import SMOTE
- from imblearn.under_sampling import NearMiss,RandomUnderSampler
- from imblearn.combine import SMOTETomek
- from imblearn.ensemble import BalancedBaggingClassifier
- from imblearn.pipeline import Pipeline

Por lo cual luego de los resultados obtenidos según los criterios de inclusión se escogieron las siguientes técnicas aplicar:

- Regresión Logística
- Random Forest
- Los K-vecinos más cercanos (KNN)
- Red Neuronal
- Máquinas de Vector de Soporte (SVM – RBF)

Cada una de las técnicas antes mencionadas formaban parte de la herramienta de Python scikit learn que proporciona recursos para realizar estos tipos de algoritmos supervisados de clasificación.

Generación del Plan de Prueba

En esta fase de la metodología el procedimiento que se utilizará para probar la validez del modelo serán las métricas de evaluación del área bajo la curva (AUC ROC) y el puntaje F1 que se aplicaron en cada uno de los escenarios y en algoritmos de aprendizaje supervisado.

Para ello se utilizaron las siguientes herramientas que proporciona Python para este tipo de procesos:

- from sklearn.metrics import roc_curve
- from sklearn.metrics import roc_auc_score
- from sklearn.metrics import f1_score
- from sklearn.model_selection import cross_val_score

Para realizar las pruebas en cada uno de los escenarios se dividieron los datos en dos partes una que es el conjunto de datos de entrenamiento que se empleará para la creación de los distintos modelos de predicción en función del método utilizado. En cambio, para la verificación de la fiabilidad del modelo, el conjunto será el de prueba, que permita evaluar que tan bien es capaz de predecir dicho modelo a partir de nuevos datos, aplicando métodos de remuestreo utilizando el over-sampling y el under-sampling para el desbalanceo de los datos.

Construcción del Modelo

En esta fase de la metodología se realizó la creación de los modelos de minería de datos, los cuales permitirán llegar al cumplimiento de los objetivos planteados. Se han seleccionado cinco técnicas que serán aplicadas las cuales son:

- **Regresión Logística**

Este método estadístico es utilizado con frecuencia cuando la variable dependiente es binaria, como es el caso de la clasificación de los estudiantes regulares. Además, permite la generación de datos de entrenamiento y prueba.

Para el desarrollo del presente modelo se utilizaron los siguientes componentes:

- La herramienta de Python from sklearn.linear_model import LogisticRegression
- La etiqueta de predicción se la realizó con los siguientes parámetros: LogisticRegression(max_iter=2000).fit(X_train, y_train).predict(X_test).
- Se trabajó cada uno de los datos del modelo aplicando el resultado para el undersampling y oversamplig, que son métodos de desbalanceo de datos.
- Luego a través de la herramienta de Python sklearn.metrics se realizó el gráfico del área bajo la curva con los resultados de las métricas de evaluación aplicadas.

- **Random Forest**

El presente modelo se ha desarrollado utilizando esta técnica de aprendizaje automático, el cual permite cumplir los siguientes objetivos propuestos: la generación de datos de entrenamiento y prueba; y, la detección de patrones en la regularidad de los estudiantes en la UTM.

Para el desarrollo del presente modelo se utilizaron los siguientes componentes:

- La herramienta de Python from sklearn.ensemble import RandomForestClassifier
- La etiqueta de predicción se la realizó con los siguientes parámetros: RandomForestClassifier(n_estimators = 10, criterion = "entropy", random_state = 27).fit(X_train, y_train).predict(X_test)

- Se trabajó cada uno de los datos del modelo aplicando el resultado para el undersampling y oversamplig, que son métodos de desbalanceo de datos.
- Luego a través de la herramienta de Python sklearn.metrics se realizó el gráfico del área bajo la curva con los resultados de las métricas de evaluación aplicadas.

- **Los k-vecinos más cercanos (KNN)**

Este método tiene como objetivo clasificar correctamente todas las instancias nuevas, calculando la distancia entre el ítem a clasificar y el resto de ítems del dataset de los datos de entrenamiento, después seleccionar los k elementos más cercanos, es decir con menor distancia, según la función que se use, permitiendo a la etiqueta realizar la clasificación final.

Para el desarrollo del presente modelo se utilizaron los siguientes componentes:

- La herramienta de Python from sklearn.neighbors import KNeighborsClassifier
- La etiqueta de predicción se la realizó con los siguientes parámetros: KNeighborsClassifier(n_neighbors = 5, metric = "minkowski", p = 2).fit(X_train, y_train).predict(X_test)
- Se trabajó cada uno de los datos del modelo aplicando el resultado para el undersampling y oversamplig, que son métodos de desbalanceo de datos.
- Luego a través de la herramienta de Python sklearn.metrics se realizó el gráfico del área bajo la curva con los resultados de las métricas de evaluación aplicadas.

- **Redes Neuronales**

Para la utilización de este tipo de modelo fue necesario la instalación de la plataforma de aprendizaje profundo tensorflow, de esta manera poder trabajar el modelo de redes neuronales detallado a continuación:

Un pequeño modelo de red neuronal se construye con una sola capa oculta con 25 neuronas usando la función de activación de un rectificador (rectifier). La capa de salida tiene una sola neurona y utiliza la función de activación sigmoide para emitir valores similares a los de la probabilidad. El ritmo de aprendizaje del descenso de gradiente estocástico se ha fijado en un valor alto de 0,1. El modelo está entrenado para 60 épocas y el argumento de decaimiento se ha establecido en 0.00166, calculado como 0,1 / 60. Además, puede ser una buena idea utilizar el momentum cuando se utiliza un ritmo de aprendizaje adaptativo. En este caso se utiliza un valor de impulso (momentum) de 0,8. por lo que se hicieron dos ejemplos, utilizando una red oculta y usando dos capas ocultas.

Además, fue necesario de la utilización de las siguientes herramientas:

- from keras.models import Sequential
- from keras.layers import Dense
- from keras.optimizers import SGD

- **Máquinas de Vector Soporte (SVM – RBF)**

Este método tiene como objetivo clasificar o predecir nuevos datos, basado en un hiperplano que clasificará todas las observaciones de entrenamiento.

Para el desarrollo del presente modelo se utilizaron los siguientes componentes:

- La herramienta de Python from sklearn.svm import SVC

- La etiqueta de predicción se la realizó con los siguientes parámetros: SVC (kernel = "rbf", random_state = 27).fit(X_train, y_train).predict(X_test)
- Se trabajó cada uno de los datos del modelo aplicando el resultado para el undersampling y oversamplig.
- Luego a través de la herramienta de Python sklearn.metrics se realizó el gráfico del área bajo la curva con los resultados de las métricas de evaluación aplicadas.

Evaluación del Modelo

Esta etapa de la metodología está orientada a la evaluación de los resultados de los modelos, por lo cual se establecieron tres escenarios para aplicar cada uno de los algoritmos en relación a los diferentes niveles de los estudiantes y el porcentaje de regularidad de los mismos, según los puntos establecidos en los criterios de inclusión.

Además, se aplicaron las técnicas de aprendizaje detallada en el punto anterior, las cuales fueron evaluadas por las métricas de evaluación del área bajo la curva AUC ROC y el puntaje F1. Así mismo según los resultados obtenidos, uno de los mejores métodos para el desbalance de datos aplicado en este trabajo de investigación fue el undersampling y oversampling del Método de Remuestreo utilizado en los procesos de minería de datos.

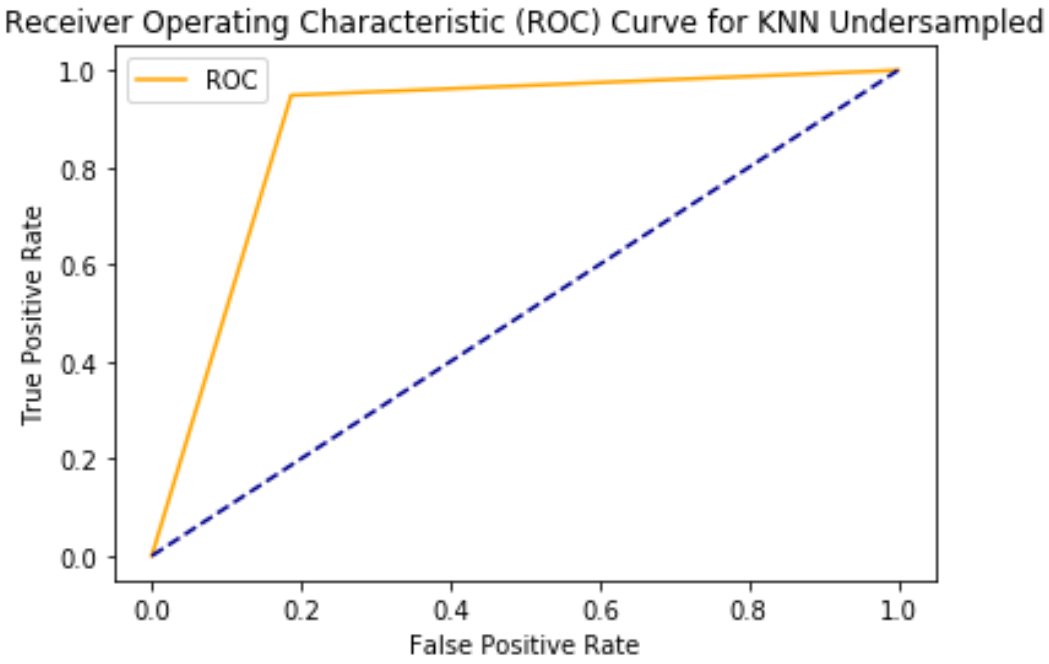
Por lo cual los mejores resultados en cada uno de los escenarios se detallan desde la figura 22 hasta la figura 36.

Primer escenario (Nivel 1,2,3)

- **KNN**

Figura 22

Curva AUC ROC del algoritmo de aprendizaje KNN

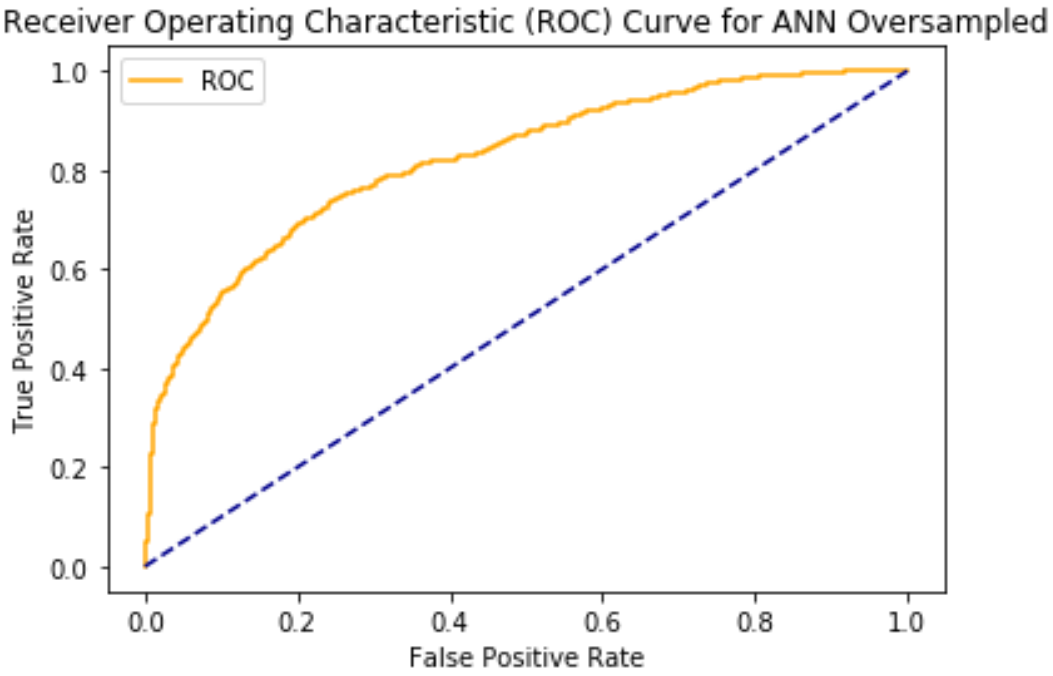


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.88, con un margen de error del 0.12, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante los primeros niveles. Así mismo el puntaje F1 es de 0.89 en relación al promedio ponderado de la precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Red Neuronal (2 capas ocultas)**

Figura 23

Curva AUC ROC del algoritmo de aprendizaje profundo de red neuronal

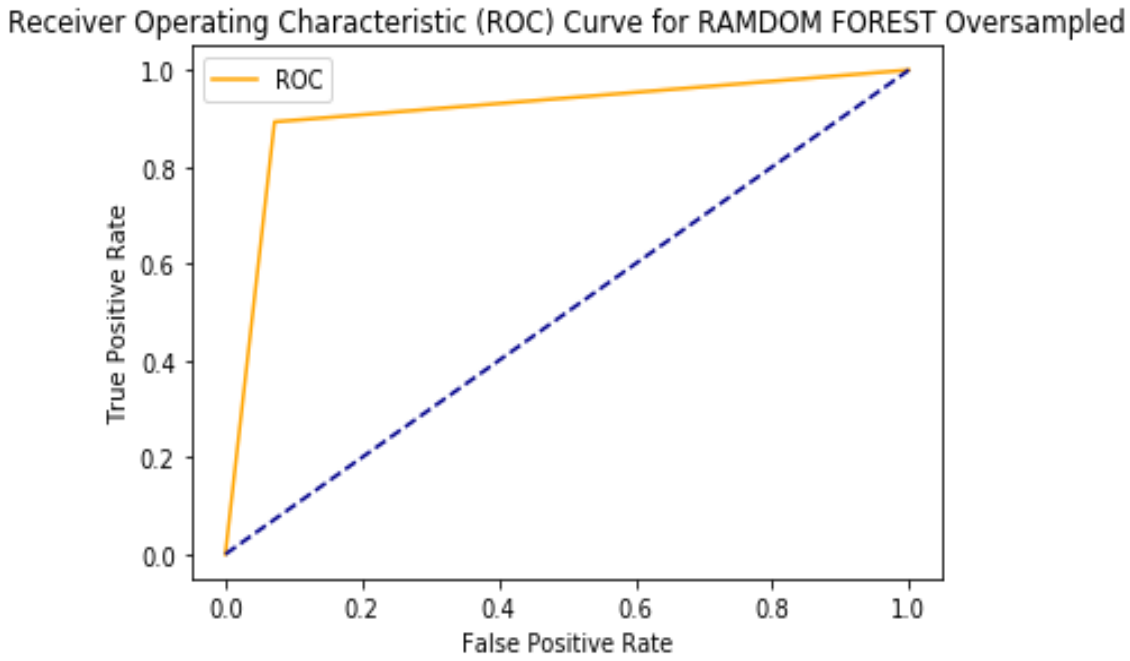


En el presente gráfico se utilizó el algoritmo de aprendizaje profundo a través de una red neuronal haciendo uso de la herramienta de tensorflow que contiene dos capas ocultas, aplicando el método de re muestreo oversampled que dio un buen resultado y además se detalla un área bajo la curva AUC ROC del 0.82, es decir permite interpretar la probabilidad de que un estudiante sea regular o no en los primeros niveles verificando que los clasifique correctamente debido al valor obtenido.

- **Random Forest**

Figura 24

Curva AUC ROC del algoritmo de aprendizaje de aprendizaje Random Forest

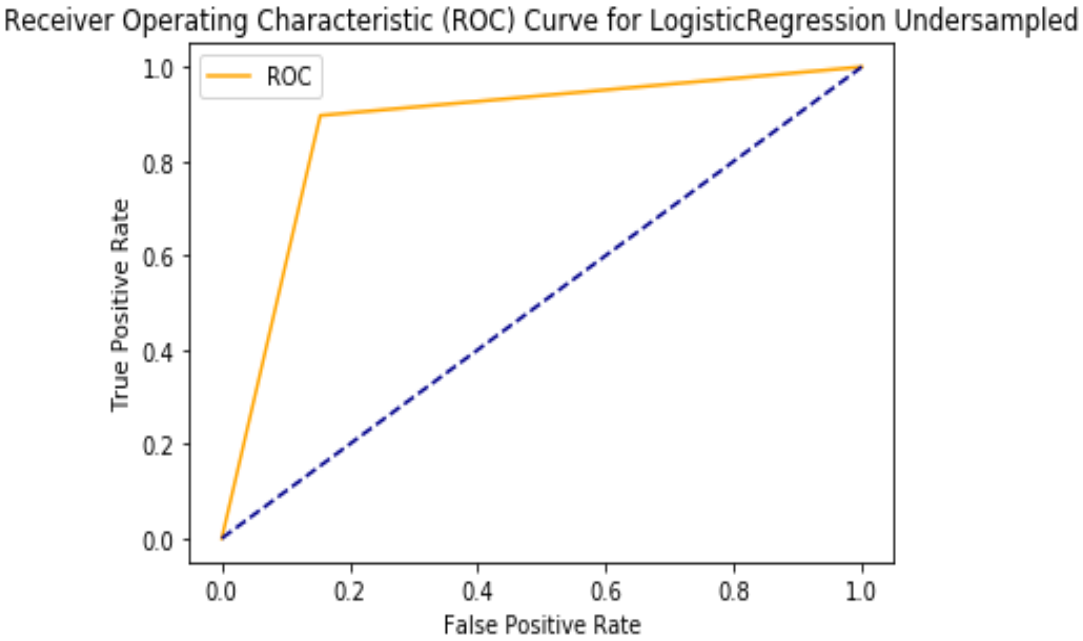


En el presente gráfico se utilizó el método de re muestreo oversampled, y se detalla un área bajo la curva AUC ROC del 0.91, con un margen de error del 0.09, es decir permite interpretar con un mínimo error la regularidad de los estudiantes en los primeros niveles. Así mismo el puntaje F1 es de 0.91 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual es uno de los mejores escenarios, debido a la concordancia de los valores que influyen en la irregularidad de los datos de los estudiantes.

- **Regresión Logística**

Figura 25

Curva AUC ROC del algoritmo de aprendizaje de aprendizaje Regresión Logística

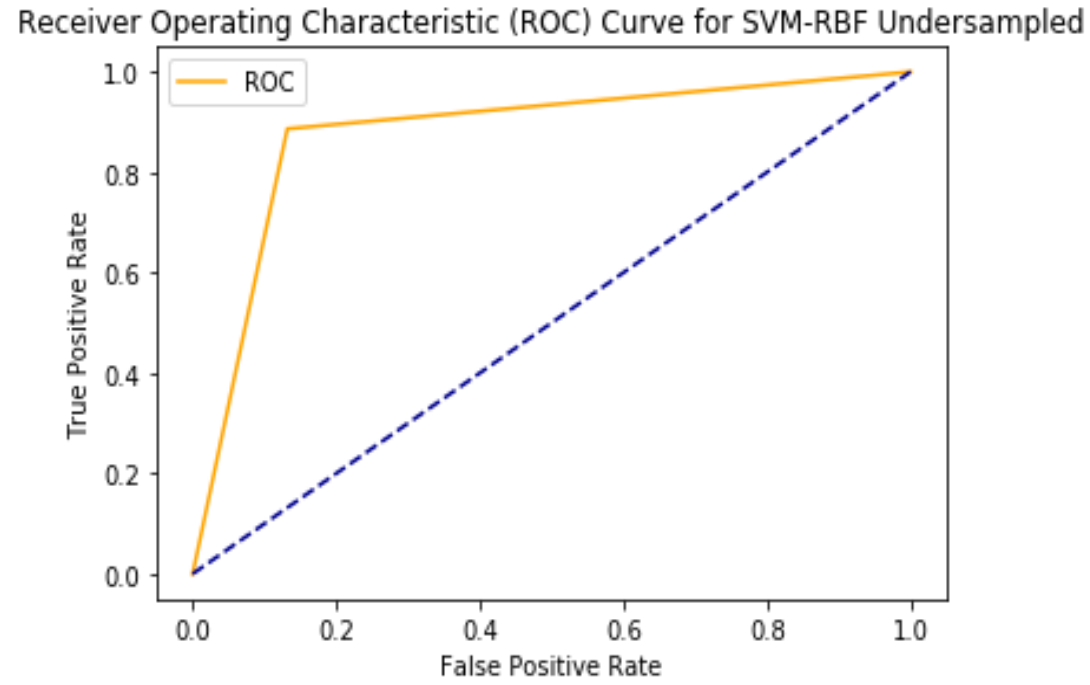


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.87, con un margen de error del 0.13, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante los primeros niveles. Así mismo el puntaje F1 es de 0.88 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Máquinas de Vector Soporte (SVM – RBF)**

Figura 26

Curva AUC ROC del algoritmo de aprendizaje de aprendizaje de Máquinas de Vector de Soporte (SVM-RBF)



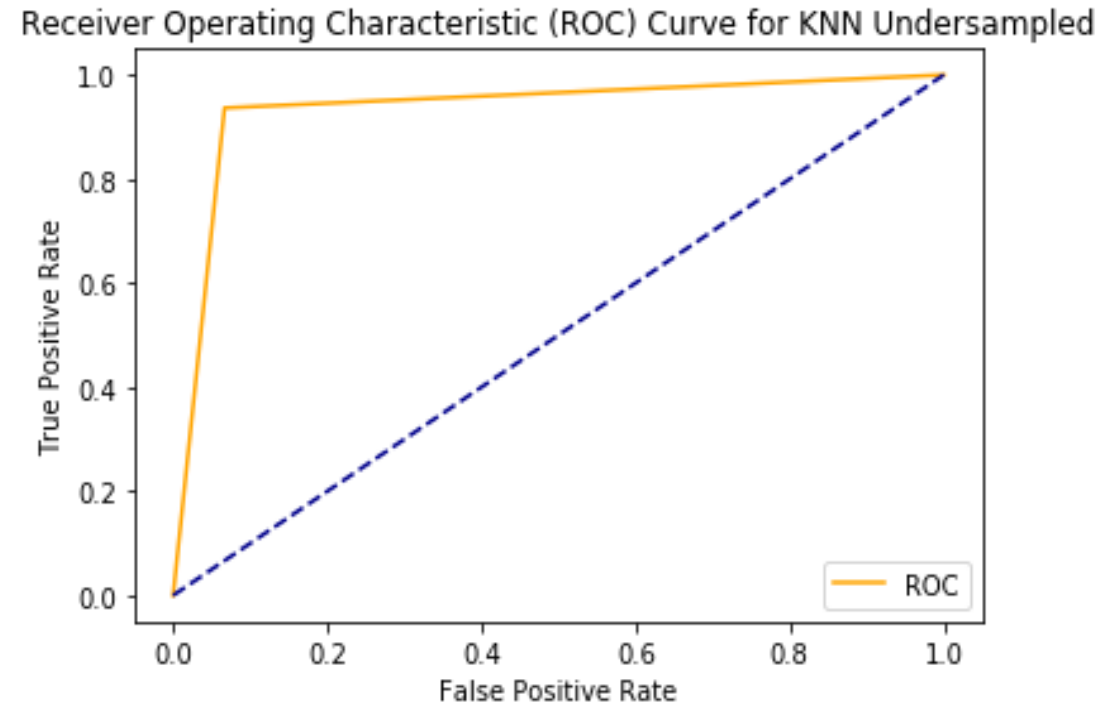
En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.88, con un margen de error del 0.12, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante los primeros niveles. Así mismo el puntaje F1 es de 0.88 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

Segundo escenario (Nivel 2,3,4)

- **KNN**

Figura 27

Curva AUC ROC del algoritmo de aprendizaje KNN

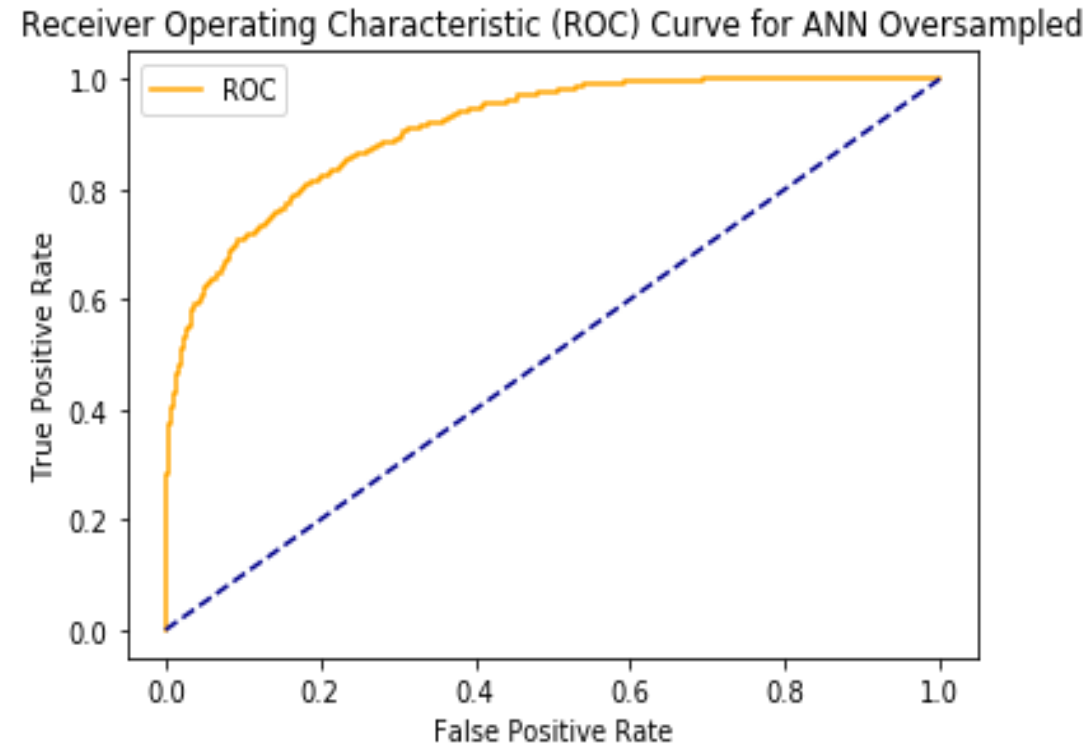


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.93, con un margen de error del 0.07, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 2,3 y 4 nivel. Así mismo el puntaje F1 es de 0.93 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores iguales entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Red Neuronal (2 capas ocultas)**

Figura 28

Curva AUC ROC del algoritmo de aprendizaje profundo Red Neuronal

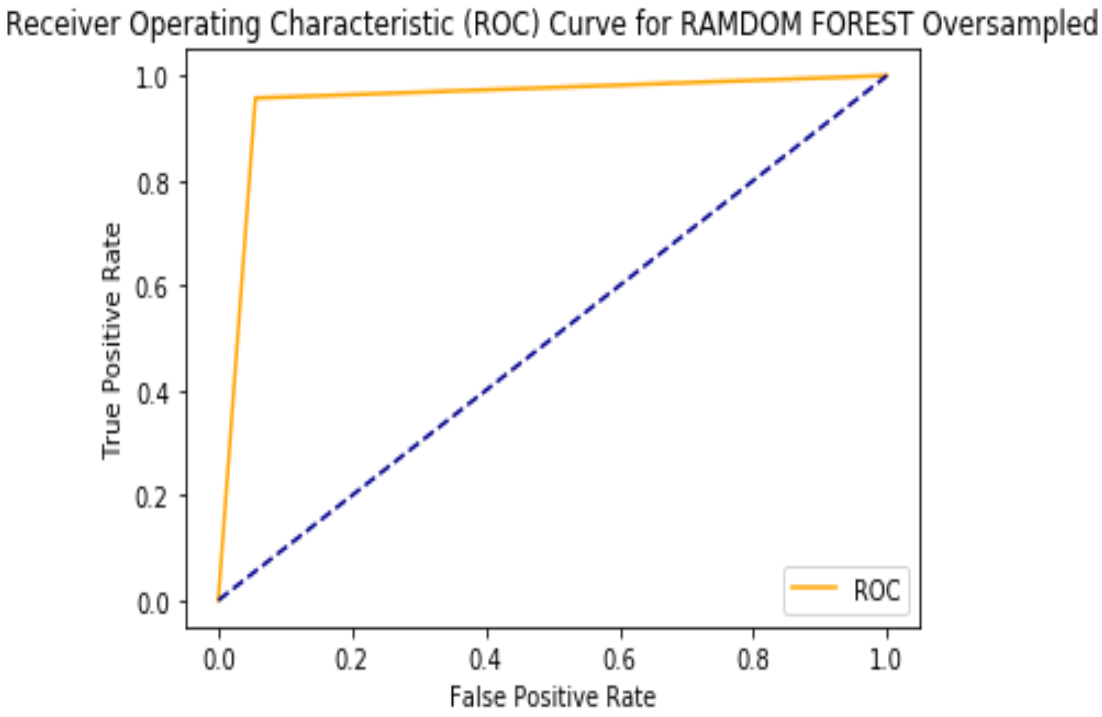


En el presente gráfico se utilizó el algoritmo de aprendizaje profundo a través de una red neuronal haciendo uso de la herramienta de tensorflow que contiene dos capas ocultas, aplicando el método de re muestreo oversampled que dio un buen resultado y además se detalla un área bajo la curva AUC ROC del 0.91, es decir permite interpretar la probabilidad de que un estudiante sea regular o no en los niveles 2,3 y 4, verificando que los clasifique correctamente debido al valor obtenido.

- **Random Forest**

Figura 29

Curva AUC ROC del algoritmo de aprendizaje Random Forest

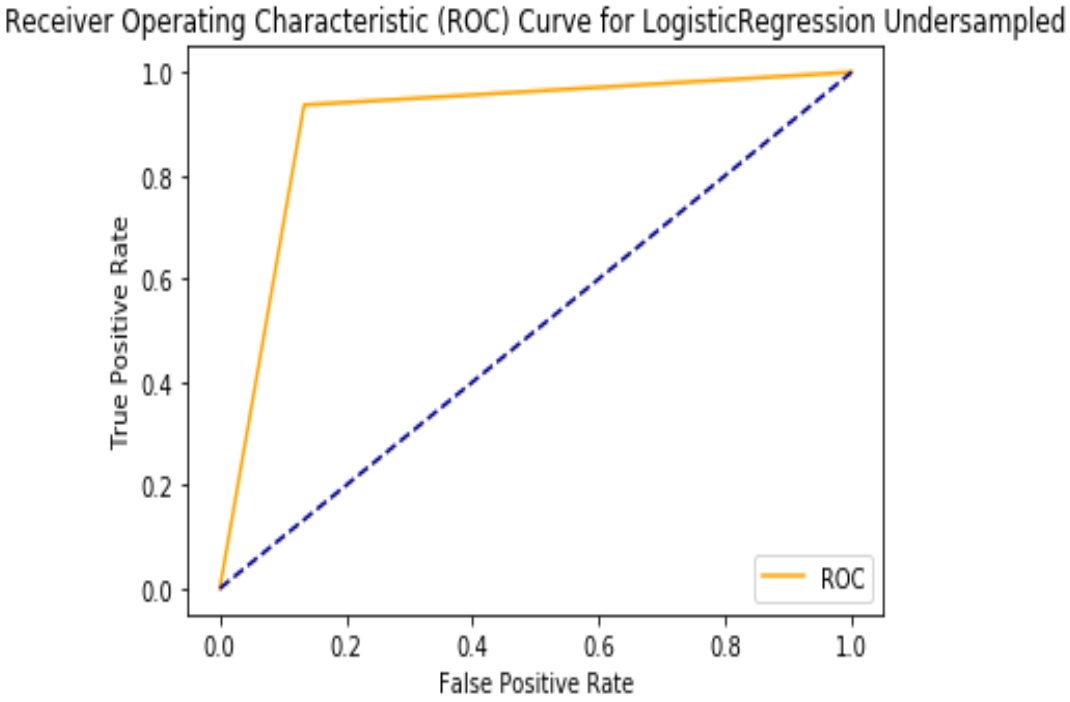


En el presente gráfico se utilizó el método de re muestreo oversampled, que detalla un área bajo la curva AUC ROC del 0.95, con un margen de error del 0.05, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante los niveles 2, 3 y 4. Así mismo el puntaje F1 es de 0.95 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual existe similitud entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Regresión Logística**

Figura 30

Curva AUC ROC del algoritmo de aprendizaje de Regresión Logística

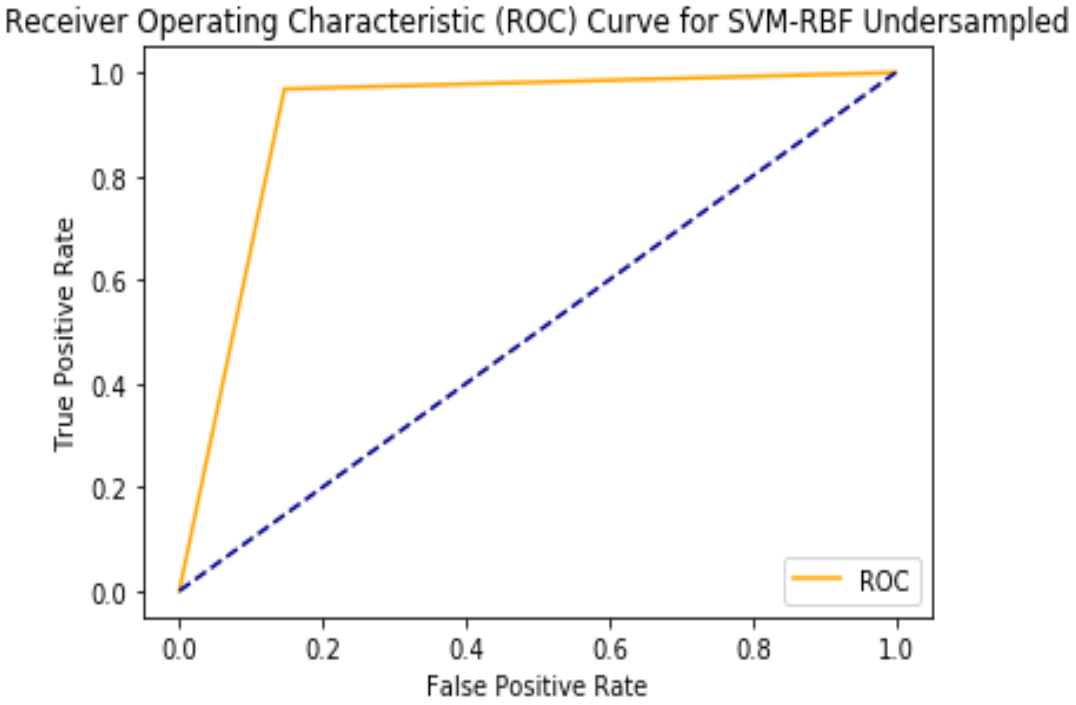


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.90, con un margen de error del 0.10, es decir permite interpretar un menos porcentaje de error en la regularidad de los estudiantes durante el 2, 3 y 4 nivel. Así mismo el puntaje F1 es de 0.89 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Máquinas de Vector Soporte (SVM – RBF)**

Figura 31

Curva AUC ROC del algoritmo de aprendizaje de Máquina de Vector de Soporte (SVM-RBF)



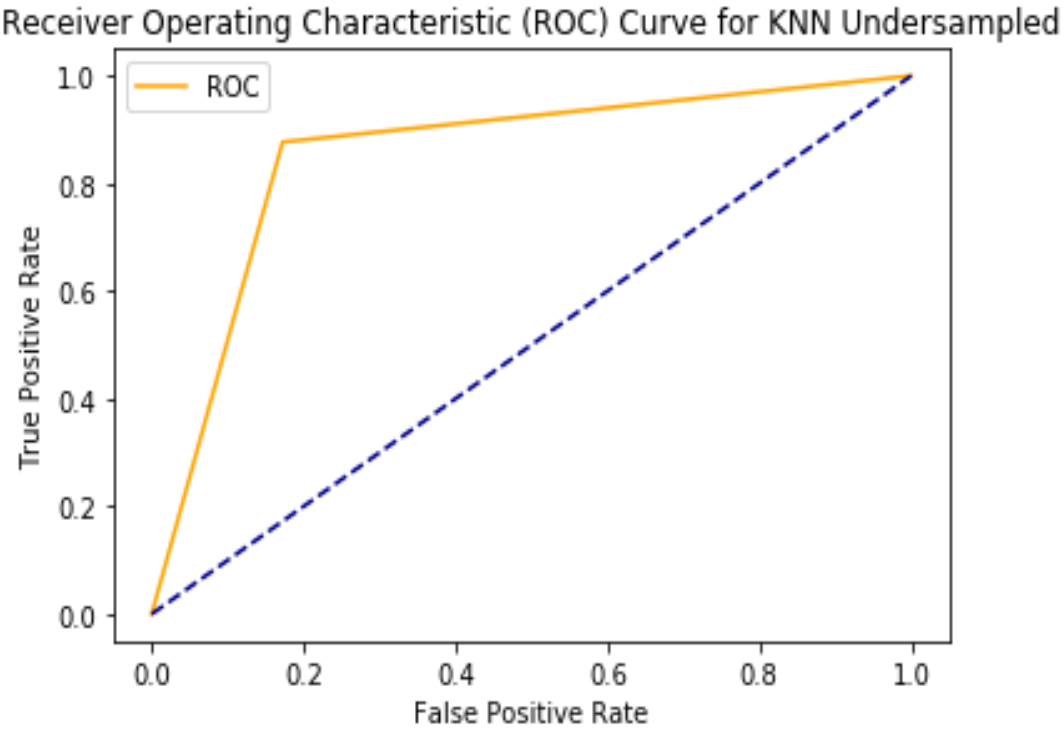
En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.91, con un margen de error del 0.09, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 2, 3, y 4 nivel. Así mismo el puntaje F1 es de 0.90 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

Tercer escenario (Nivel 3,4,5)

- **KNN**

Figura 32

Curva AUC ROC del algoritmo de aprendizaje KNN

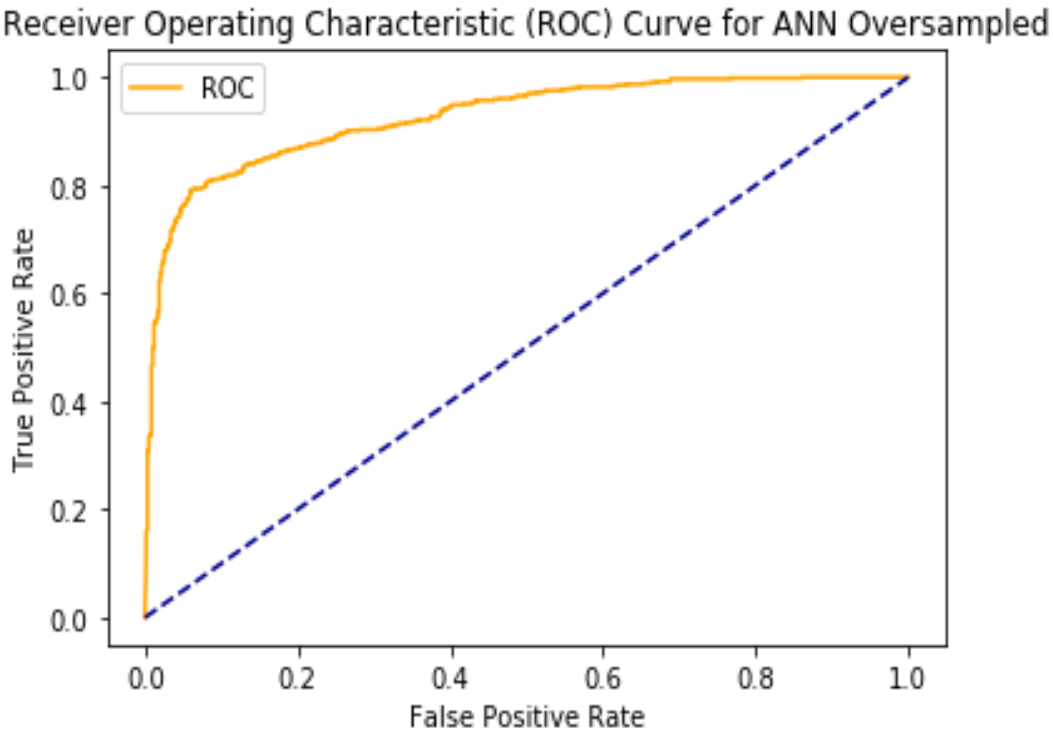


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.85, con un margen de error del 0.15, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 3,4 y 5 nivel. Así mismo el puntaje F1 es de 0.86 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Red Neuronal (2 capas ocultas)**

Figura 33

Curva AUC ROC del algoritmo de aprendizaje profundo Red Neuronal

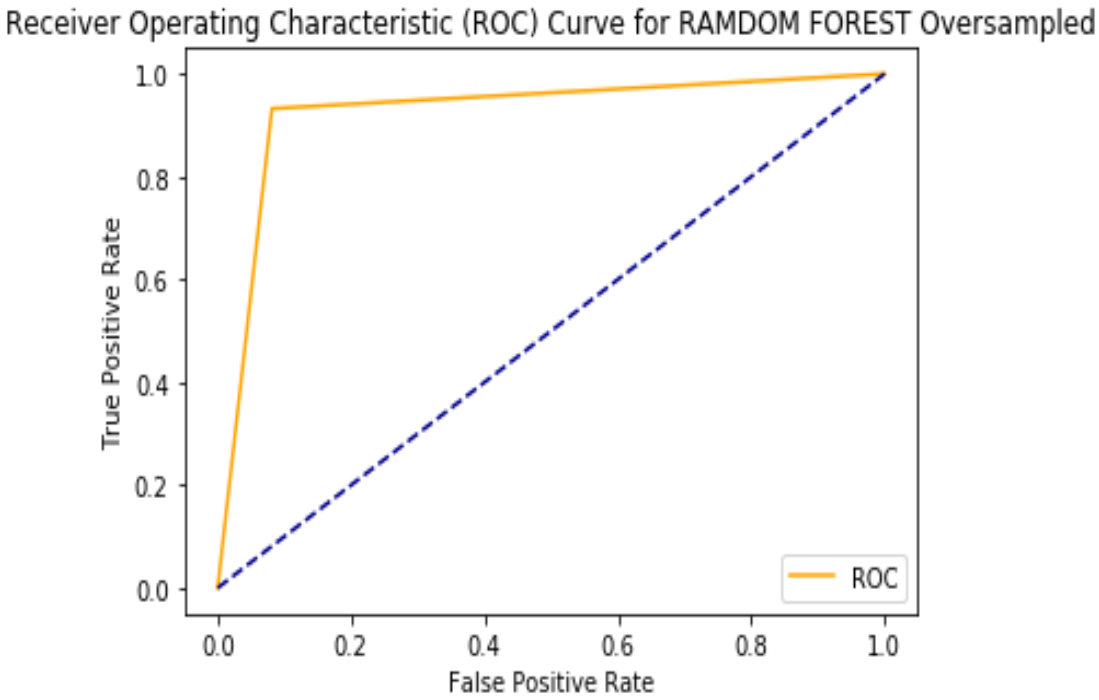


En el presente gráfico se utilizó el algoritmo de aprendizaje profundo a través de una red neuronal haciendo uso de la herramienta de tensorflow que contiene dos capas ocultas, aplicando el método de re muestreo oversampled que dio un buen resultado y además se detalla un área bajo la curva AUC ROC del 0.93, es decir permite interpretar la probabilidad de que un estudiante sea regular o no en los niveles correspondiente, verificando que los clasifique correctamente debido al valor obtenido.

- **Random Forest**

Figura 34

Curva AUC ROC del algoritmo de aprendizaje Random Forest

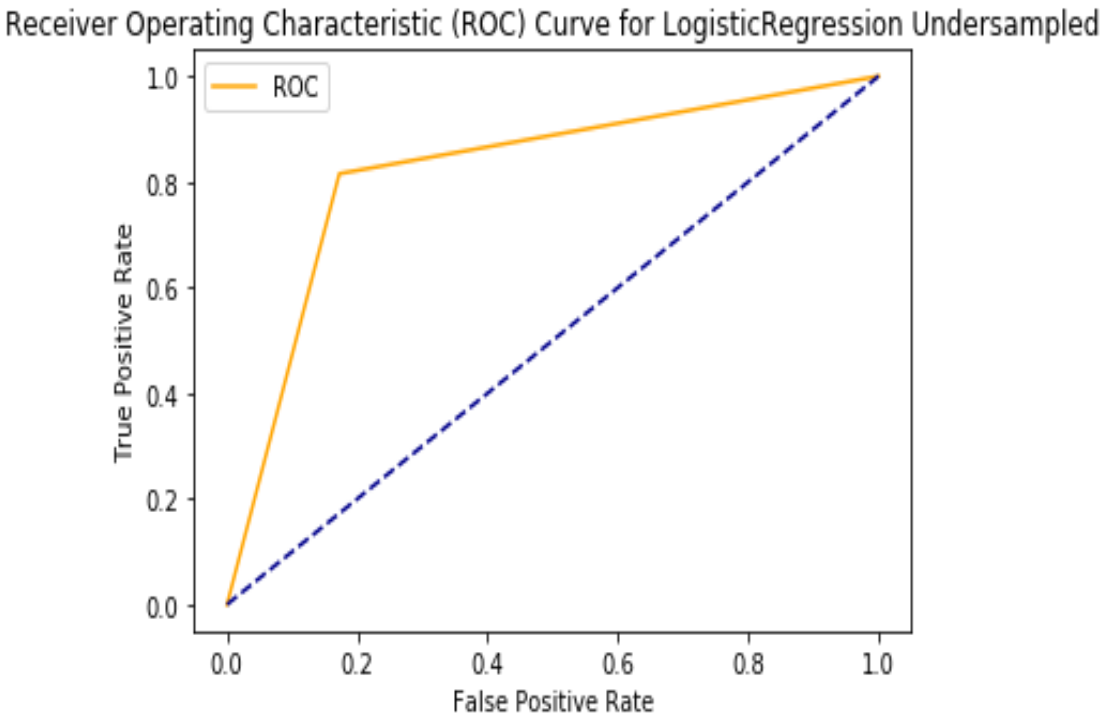


En el presente gráfico se utilizó el método de re muestreo oversampled, que detalla un área bajo la curva AUC ROC del 0.93, con un margen de error del 0.07, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 3, 4 y 5 nivel. Así mismo el puntaje F1 es de 0.93 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores iguales entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Regresión Logística**

Figura 35

Curva AUC ROC del algoritmo de aprendizaje Regresión Logística

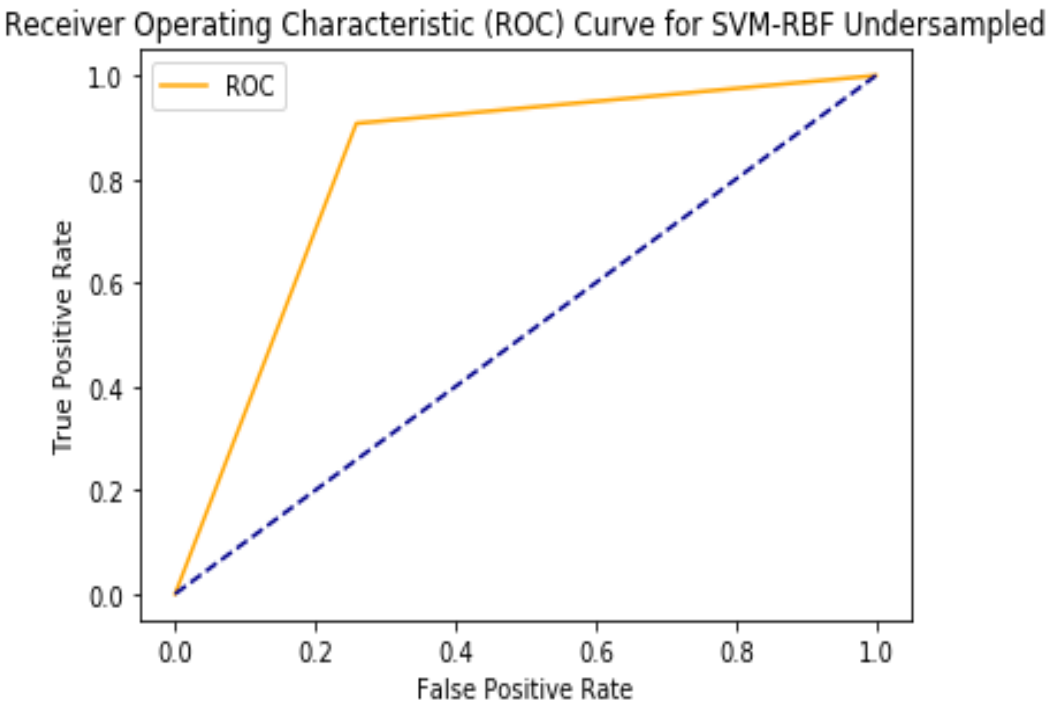


En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.82, con un margen de error del 0.18, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 3,4 y 5 nivel. Así mismo el puntaje F1 es de 0.83 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

- **Máquinas de Vector Soporte (SVM – RBF)**

Figura 36

Curva AUC ROC del algoritmo de aprendizaje e Máquinas de Vector de Soporte (SVM-RBF)



En el presente gráfico se utilizó el método de re muestreo undersampled, que detalla un área bajo la curva AUC ROC del 0.82, con un margen de error del 0.18, es decir permite interpretar un menor porcentaje de error en la regularidad de los estudiantes durante el 3, 4 y 5 nivel. Así mismo el puntaje F1 es de 0.85 en relación al promedio ponderado de precisión y sensibilidad del modelo utilizado; por lo cual son valores muy similares entre cada una de las métricas, obteniendo una mejor evaluación del modelo en relación al resultado obtenido.

Fase 5: Evaluación

Esta fase de la metodología evaluó cada uno de los escenarios generados desde el punto de vista de los objetivos del negocio, permitiendo verificar si estos han sido cumplidos.

En el desarrollo de la presente investigación se aplicaron cinco modelos en tres distintos escenarios, donde el porcentaje de acierto fue diferente en cada una de ellos, es así que a través de la tabla 11 se muestra cada uno de los resultados:

Tabla 11

Resultados Comparativos de los algoritmos de aprendizaje en cada uno de los escenarios

N°	ALGORITMO	ESCENARIO 1			ESCENARIO 2			ESCENARIO 3		
		NIVEL 1,2,3			NIVEL 2,3,4			NIVEL 3,4,5		
		PUNTAJE F1	ÁREA BAJO LA CURVA AUC ROC	TEST ERROR	PUNTAJE F1	ÁREA BAJO LA CURVA AUC ROC	TEST ERROR	PUNTAJE F1	ÁREA BAJO LA CURVA AUC ROC	TEST ERROR
1	KNN	0.89	0.88	0.12	0.93	0.93	0.07	0.86	0.85	0.15
2	RED NEURONAL		0.82			0.91			0.93	
3	SVM - RBF	0.88	0.88	0.12	0.90	0.91	0.09	0.85	0.82	0.18
4	RANDOM FOREST	0.91	0.91	0.09	0.95	0.95	0.05	0.92	0.93	0.07
5	LOGISTIC REGRESSION	0.88	0.87	0.13	0.89	0.90	0.10	0.83	0.82	0.18

En la tabla 11 se visualiza la iteración de los valores en cada uno de los algoritmos, en el cual el escenario con mejores resultados es el segundo que hace referencia a los niveles 2,3, y 4; donde el algoritmo de aprendizaje automático de clasificación Random Forest realiza una mejor predicción sobre la regularidad e irregularidad de los estudiantes en las diferentes carreras, valorado por las métricas de evaluación del área bajo la curva AUC ROC de 0.95, con un margen de error de 0.05, y el Puntaje F1 de 0.95; por lo que este escenario muestra valores similares y con un mínimo margen de error.

Fase 6: Implementación

En esta fase de la metodología en la presente investigación, se les informará a las autoridades de la institución sobre los resultados obtenidos, para la ayuda a la toma de decisiones y que sea una propuesta que permita incorporar un módulo dentro del Sistema de Gestión Académica que incluya un monitoreo de la regularidad e irregularidad de los estudiantes al finalizar el semestre en cada periodo académico ordinario, como un estimador del riesgo de deserción y de esta manera realizar un seguimiento más riguroso.

Capítulo V: Conclusiones y Recomendaciones

Conclusiones

- A través de la revisión del estado del arte se identificaron varias técnicas de minería de datos que se aplicaron en la presente investigación como son: Regresión Logística, Random Forest, Máquinas de Vector Soporte (SVM), Redes Neuronales y los K-vecinos más cercanos (KNN).
- Por medio de la información académica y demográfica de los estudiantes se realizó el proceso de minería de datos, comprobando la correlación entre los estudiantes desertores y los no regulares, por lo cual la irregularidad fue un estimador del riesgo de deserción que se presentó en la institución.
- Mediante el análisis de las características para definir los escenarios y realizar la ejecución del modelo de aprendizaje automático supervisado, se constató la irregularidad de los estudiantes en los primeros niveles académicos, por lo cual se verificaron que muchos estudiantes desertan en estos niveles durante su carrera universitaria.
- A partir de las métricas de evaluación los resultados esperados fueron calculados con el algoritmo de clasificación Random Forest que realizó una mejor predicción sobre la deserción estudiantil con un margen de error de 0.05, valorado por las métricas del área bajo la curva AUC ROC de 0.95 y el Puntaje F1 de 0.95, existiendo una alta correlación en los valores de cada una de las métricas de evaluación.

Recomendaciones

- Incluir en el ingreso de la información de los estudiantes en el sistema, campos de selección múltiple, de tal manera que evite ciertos errores que se presentan al momento de la manipulación de los datos.
- Actualizar periódicamente la información demográfica y personal de los estudiantes y que mencionados registros queden guardados de forma independiente.
- Hacer un seguimiento de medio ciclo a través del Sistema de Gestión Académica de los estudiantes durante el periodo académico ordinario para verificar el desenvolvimiento de cada uno de ellos y de esta manera evitar la deserción estudiantil.
- Realizar un sistema de minería de datos, que contenga visualizadores y generadores de reportes que facilite y provea datos consolidados y de calidad principalmente a las unidades académicas de cada una de las carreras, de tal forma que ayude en la toma de decisiones con las demás autoridades.

Bibliografía

Alban, M., & Mauricio, D. (2018). Factors to predict dropout at the universities: A case of study in Ecuador. IEEE Global Engineering Education Conference, EDUCON, 2018-April, 1238-1242. Scopus. https://doi.org/10.1109/EDUCON.2018.8363371

Bazantes, Z. P., Carpio, M. L. R., & Gutiérrez, M. L. Á. (2016). DESERCIÓN ESTUDIANTIL UNIVERSITARIA EN ECUADOR Y SU INFLUENCIA EN LA CALIDAD DEL EGRESADO. Revista Magazine de las Ciencias. ISSN 2528-8091, 1(4), 65-70.

Brownlee, J. (2018, mayo 22). A Gentle Introduction to k-fold Cross-Validation. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/

Canales, A., & Ríos, D. D. los. (2018). Factores explicativos de la deserción universitaria. Calidad en la Educación, 0(26), 173-201. https://doi.org/10.31619/caledu.n26.239

Castro R., L. F., Espitia P., E., & Montilla, A. F. (2018). Applying CRISP-DM in a KDD process for the analysis of student attrition (Vol. 885). Scopus. https://doi.org/10.1007/978-3-319-98998-3_30

Chagas, E. T. D. O. (2019). Deep Learning y sus aplicaciones hoy. Revista Científica Multidisciplinar Núcleo do Conhecimento, 04(05), 05-26.

Cortina, V. G. (2015). Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario. 120.

De Battista, A., Cristaldo, P., Ramos, L., Nuñez, J. P., Retamar, S., Bouzenard, D., & Herrera, N. E. (2016, mayo 19). Minería de datos aplicada a datos masivos. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://hdl.handle.net/10915/52901

Echeverri, A. V. (2019). Metodología para la generación de péptidos sintéticos antimicrobianos usando aprendizaje profundo y algoritmos de clasificación. 77.

Fernandez, J. (2019). USO DE MACHINE LEARNING CON PYTHON PARA LA DETECCI ÓN DE DAÑO COGNI TI VO EN PERSONAS CON SÍNDROME DE DOWN. http://castor.det.uvigo.es:8080/xmlui/bitstream/handle/123456789/341/TFG%20Jos%C3%A9%20Fern%C3%A1ndez%20G%C3%B3mez.pdf?sequence=1

Gonzalez, V. (2017). Modelado mediante RAMDOM FORESTS de las emisiones de autobuses urbanos en función de los ciclos cinemáticos. 2017. http://oa.upm.es/45914/1/TFG_VICTOR_PITA_GONZALEZ_CAMPOS.pdf

Guanin-Fajardo, J., Casillas, J., Chiriboga-Casanova, W., Guanin-Fajardo, J., Casillas, J., & Chiriboga-Casanova, W. (2019). Aprendizaje semi-supervisado para descubrir la escala de tiempo promedio de graduación de estudiantes universitarios. Conrado, 15(70), 291-299.

Hernández Lalinde, J. D., Espinosa Castro, J. F., Peñaloza Tarazona, M. E., Fernández González, J. E., Chacón Rangel, J. G., Toloza Sierra, C. A., Arenas Torrado, M. K., Carrillo Sierra, S. M., & Bermúdez Pirela, V. J. (2018). Sobre El Uso Adecuado Del Coeficiente De Correlación De Pearson: Definición, Propiedades Y Suposiciones. Archivos Venezolanos de Farmacología y Terapéutica. https://bonga.unisimon.edu.co/handle/20.500.12442/2469

Jara Estupiñan, J., Giral, D., & Martínez Santa, F. (2016). Implementation of algorithms based on support vector machine (SVM) for electric systems: Topic review.

Tecnura, 20(48), 149-170. https://doi.org/10.14483/udistrital.jour.tecnura.2016.2.a11

Lacave, C., Molina, A. I., & Cruz-Lemus, J. A. (2018). Learning Analytics to identify dropout factors of Computer Science studies through Bayesian networks. Behaviour and Information Technology, 37(10), 993–1007. https://doi.org/10.1080/0144929X.2018.1485053

Loor, F. (2018, julio 23). La regresión logística. Analytics Lane. https://www.analyticslane.com/2018/07/23/la-regresion-logistica/

López Ramos, D., Arco García, L., López Ramos, D., & Arco García, L. (2019). Aprendizaje profundo para la extracción de aspectos en opiniones textuales. Revista Cubana de Ciencias Informáticas, 13(2), 105-145.

LUCA, P. R. de los S. E. en generación de contenidos tecnológicos para los canales digitales de, vida", la unidad de datos de T. L. en C. F. y M. en T. E. A. por las "tecnologías para la, & Pedagogía, L. Q. N. H. L. V. M. F. P. L. (2018, enero 23). Machine Learning a tu alcance: La matriz de confusión - Think Big Empresas. Think Big. https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/

Mascort Colomer, A. (2019). Estudio del rendimiento de técnicas de minería de datos en la predicción de resultados académicos. https://upcommons.upc.edu/handle/2117/167732

Mejías, A. M. M. (2018). Programa basado en Python para integrar la gestión de documentos y procesos de trabajo en una empresa. 63.

Mendoza, L. E. E. (2019). 008. Tensorflow como alternativa a herramientas estadisticas, caso de aplicación: regresión lineal. 15.

Merchan Rubiano, S. M., & Duarte Garcia, J. A. (2016). Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance. IEEE Latin America Transactions, 14(6), 2783-2788. Scopus. https://doi.org/10.1109/TLA.2016.7555255

Montero, P. E. (2014). Aprendizaje por refuerzo en espacios continuos: Algoritmos y aplicación al tratamiento de la anemia renal [Http://purl.org/dc/dcmitype/Text, Universitat de València]. https://dialnet.unirioja.es/servlet/tesis?codigo=90545

O, J. A. G., & Molina, B. (2015). Identificación de técnicas de minería de datos para apoyar la toma de decisiones en la solución de problemas empresariales. Revista Ontare, 3(2), 33-51. https://doi.org/10.21158/23823399.v3.n2.2015.1440

Ortiz-Lozano, J. M., Rua Vieites, A., & Bilbao Calabuig, P. (2017). Aplicación de árboles de clasificación a la detección precoz de abandono en los estudios universitarios de administración y dirección de empresas. Recta, 18(2), 177–201. https://doi.org/10.24309/recta.2017.18.2.05

Pérez, R., & Luis, J. (2014). Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), 52-73.

Pérez Verona, I., & Arco García, L. (2016). Una revisión sobre aprendizaje no supervisado de métricas de distancias. A brief review on unsupervised metric learning. Revista cubana de ciencias informáticas, 10, 43-67.

Rivera, F. B. (2015). Investigación en deserción estudiantil universitaria: Educación, cultura y significados. Revista Educación y Desarrollo Social, 9(2), 86-101. https://doi.org/10.18359/reds.948

Robles, R., & Roberto, P. (2019). Desarrollo de una arquitectura conceptual para el análisis de contenidos en redes sociales sobre el tema del aborto usando Python. http://repositorio.utn.edu.ec/handle/123456789/9026

Rodríguez León, C., & García Lorenzo, M. M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. Revista universidad y sociedad, 8(4), 43-53.

Rodriguez, M. M. H. (2015). Grado de máster en ingeniería y tecnología de sistemas software. 75.

Roman, V. (2019, abril 1). Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos. Medium. https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407

Sánchez Amaya, G., Navarro Salcedo, W., & García Valencia, A. D. (2009). Factores de deserción estudiantil en la Universidad Surcolombiana. Paideia Surcolombiana, 1(14), 97. https://doi.org/10.25054/01240307.1083

Sánchez, C., & Caridad, A. (2016). Mejoras de la clasificación en interacciones de proteínas de la Arabidopsis Thaliana utilizando técnicas para conjuntos de datos desbalanceados [Thesis, Universidad Central “Marta Abreu” de Las Villas]. http://dspace.uclv.edu.cu:8089/xmlui/handle/123456789/10873

Santana, R. (2018). Matrices de confusión—EcuRed. https://www.ecured.cu/Matrices_de_confusi%C3%B3n

Solis, M., Moreira, T., Gonzalez, R., Fernandez, T., & Hernandez, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. 2018 IEEE International Work Conference on Bioinspired Intelligence, IWOBI 2018 - Proceedings. Scopus. https://doi.org/10.1109/IWOBI.2018.8464191

Timaran Pereira, R., & Caicedo Zambrano, J. (2018). Application of decision trees for detection of student dropout profiles. Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, 2018-January, 528–531. https://doi.org/10.1109/ICMLA.2017.0-107

Universidad Pedagógica Nacional, & Guevara Patiño, R. (2016). El estado del arte en la investigación: ¿análisis de los conocimientos acumulados o indagación por nuevos sentidos? Folios, 1(44), 165-179. https://doi.org/10.17227/01234870.44folios165.179

Universidad Técnica de Manabí (Ecuador)—EcuRed. (s. f.). Recuperado 15 de abril de 2020, de https://www.ecured.cu/Universidad_T%C3%A9cnica_de_Manab%C3%AD_(Ecuador)

Valero, C. S. (2018). Aplicación de métodos de aprendizaje automático en el análisis y la predicción de resultados deportivos. Retos: nuevas tendencias en educación física, deporte y recreación, 34, 377-382.

Viale Tudela, H. E. (2014). Una Aproximación Teórica a la Deserción Estudiantil Universitaria. Revista Digital de Investigación en Docencia Universitaria, 1, 59. https://doi.org/10.19083/ridu.8.366

Ziggah, Y. Y., Youjian, H., Tierra, A. R., Laari, P. B., Ziggah, Y. Y., Youjian, H., Tierra, A. R., & Laari, P. B. (2019). Coordinate Transformation between Global and Local Data Based on Artificial Neural Network with K-Fold Cross-Validation in Ghana. Earth Sciences Research Journal, 23(1), 67-77. https://doi.org/10.15446/esrj.v23n1.63860