UNIVERSIDAD NACIONAL

SANTIAGO ANTÚNEZ DE MAYOLO

FACULTAD DE CIENCIAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

MODELO PREDICTIVO PARA LA DESERCIÓN DE ESTUDIANTES EN EL PRIMER AÑO DE ESTUDIO EN LA UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO, HUARAZ – 2022

TESIS PARA OPTAR EL TÍTULO DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

PRESENTADO POR:

BACHILLER HENRY FRANCIS GUTIERREZ CHURATA

ASESOR: MS. FLORES CHACÓN ERICK GIOVANNY

HUARAZ – PERÚ

2022

N° Registro: T141

DEDICATORIA

A mi padre, Fermin Gutierrez, que siempre me apoya de manera incondicional con cada uno de mis objetivos que me eh propuesto.

A cada uno de mis docentes y amigos que me apoyaron para desarrollarme personal y profesionalmente.

AGRADECIMIENTO

Agradezco a Dios por protegerme durante todo mi camino y darme fuerzas para superar obstáculos y dificultades a lo largo de toda mi vida.

A mi familia por ser el apoyo contante que necesité en el trascurso de toda mi vida, en cada momento difícil y duro que pase ya sea en mi vida personal o mi vida profesional.

A mi asesor Ms. Erick Flores Chacón que siempre por sus conocimientos, orientaciones y motivaciones que me ayudaron y me guiaron para el cumplimiento de esta investigación.

A mis amigos en general, por haber logrado nuestro gran objetivo con mucha perseverancia y demostrarme que podemos ser grandes amigos durante los mejores y peores momentos.

HOJA DE VISTO BUENO

PRESIDENTE
Ing. Alberto Martin Medina Villacorta
CIP N° 143122

SECRETARIO
Ing. Wilfredo Manuel Trejo Flores
CIP N° 182621

PRESIDENTE
Ing. Erick Giovanny Flores Chacón
CIP N° 89540

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

La presente tesis denominada “Modelo predictivo para la deserción de estudiantes en el primer año de estudio en la universidad nacional Santiago Antúnez de Mayolo, Huaraz – 2022” tuvo como objetivo general determinar mediante un modelo predictivo la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo - 2022. La metodología fue de tipo longitudinal, con enfoque cuantitativo, nivel de investigación explicativo, con diseño pre experimental, la población fue conformada por los casos de estudio de cada estudiante que ingreso entre los semestres 2010-I al 2019-I que en total fueron 6440 casos de estudio y la muestra fue censal dado que necesario la mayor cantidad de datos para realizar los análisis de predicción. Dando como resultado que Gradient Boosting fue le mejor modelo y obtuvo 94% de precisión, 86% de sensibilidad, 90% en el score F1, 95% de accuracy, 75.12% en el score R-cuadrado de la data de entrenamiento y 70.09% en el score R2-cuadrado de la data de test. Llegando a concluir que con la aplicación de algoritmos de machine learning se puede tener la predicción de la deserción de estudiantes en su primer año de estudio.

Palabras clave: Deserción de estudiantes, algoritmo de machine learning, accuracy, score R-cuadrado.

ABSTRACT

The present tesis called "Predictive model for the desertion of students in the first year of study at the Santiago Antúnez de Mayolo National University, Huaraz - 2022" had the general objective of determining, through a predictive model, the desertion of students in the first year of study. at the Santiago Antúnez de Mayolo National University - 2022. The methodology was longitudinal, with a quantitative approach, explanatory research level, with a pre-experimental design, the population was made up of the case studies of each student who entered between the 2010 semesters. -I to 2019-I that in total there were 6440 study cases and the sample was census since the largest amount of data was necessary to carry out the prediction analysis. As a result, Gradient Boosting was the best model and obtained 94% precision, 86% sensitivity, 90% in the F1 score, 95% accuracy, 75.12% in the R-square score of the training data and 70.09%. in the R2-square score of the test data. Coming to the conclusion that with the application of machine learning algorithms it is possible to predict the dropout of students in their first year of study.

Keywords: Student dropout, machine learning algorithm, accuracy, R-square score.

I. INTRODUCCIÓN

Antecedentes de la investigación

Antecedentes internacionales

Zapata (2021) en su investigación denominada "Método para la Detección de Estudiantes en Riesgo de Deserción, Basado en un Diseño de Métricas y una Técnica de Minería de Datos" el cual tuvo como objetivo general: desarrollar un método para la detección de estudiantes en riesgo de deserción, basado en un diseño de métricas y una técnica de minería de datos educativos, en los niveles de educación básica secundaria y media aplicado a un caso de estudio en Colombia. De esta forma, se soporta una selección y representación óptima y suficiente de características para que luego puedan ser utilizadas como entrada a un clasificador experto para un determinado tipo de especifico de características. Posteriormente se usó la fusión de nivel de clasificador para obtener una respuesta más general porque diferentes clasificadores cometen errores en diferentes muestras. De esta forma, se puede mejorar el rendimiento del clasificador y se pueden interpretar fácilmente los resultados del algoritmo de aprendizaje automático. La validación en términos de precisión, sensibilidad e interpretabilidad del método propuesto en la presente tesis se realizó en comparación con una técnica de minería de datos y las características iniciales, lo que aprobó la capacidad de detectar la deserción de estudiantes mediante la transformación de características a partir de métricas. Se logro obtener un 82% de precisión y un 64% de recall, estos resultados representan una mejora significativa con respecto al 71% de precisión y el 57 % de recall obtenidos con características iniciales sin uso de métricas. Por lo mencionado anteriormente, se sugiere su potencial aplicación en el análisis de datos educativos para la predicción temprana del riesgo de deserción y el desarrollo de estrategias para persuadir a los estudiantes a permanecer en las instituciones educativas.

Behr, Giese, Teguim, & Theune (2020) en su investigación denominada "Early Prediction of University Dropouts – A Random Forest Approach" donde el autor predice la deserción universitaria utilizando Random Forest basados en conditional inference trees y en un amplio conjunto de datos alemanes que cubre una amplia gama de aspectos de la vida estudiantil y los cursos de estudio. Donde se modelo la decisión de abandono como una clasificación binaria (graduado o abandono) y nos enfocamos en la predicción muy temprana del abandono de los estudiantes mediante el modelado paso a paso de la transición de los estudiantes desde la escuela (antes de estudiar) durante la fase de decisión de estudio (fase de decisión) a los primeros semestres en la universidad. (fase de estudio inicial). Así mismo se

evalúo cómo cambia el rendimiento predictivo en los tres modelos y observamos un rendimiento sustancialmente mayor cuando se incluyen variables de las primeras experiencias de estudio, lo que da como resultado un AUC (área bajo la curva) de 0,86. Predictores importantes son la calificación final en la escuela secundaria, y también determinantes asociados con la satisfacción de los estudiantes y su autoconcepto académico subjetivo y autoevaluación. Un resultado directo de esta investigación es la provisión de información a las universidades que deseen implementar sistemas de alerta temprana y servicios de asesoramiento más personalizados para apoyar a los estudiantes en riesgo de abandonar los estudios durante una etapa temprana de los estudios.

Camargo (2020) en su investigación denominada “Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos” donde el objetivo principal del presente proyecto fue: crear un modelo para la predicción de la deserción de estudiantes de pregrado en la Universidad de la Costa - CUC, a partir del análisis de diferentes factores socioeconómicos y académicos. La investigación tuvo que pasar por varias fases: caracterización, experimentación, desarrollo y evaluación. En la fase de caracterización, se creó un conjunto de datos (dataset) a partir de la recopilación de datos sobre el perfil demográfico, cultural, social, familiar, educativo, socioeconómico y psicológico de cada estudiante de los semestres 2013-1 al 2018-2. Esta información se recopiló del formulario de registro que completaron los estudiantes cuando ingresaron a la universidad, donde se recopiló un total de 1,606 registros únicos de estudiantes. La fase experimental evalúa varios métodos de aprendizaje automático (machine learning) de las siguientes categorías: redes bayesianas, máquinas de vectores de soporte y árboles de decisión. El algoritmo que obtuvo el mejor número de aciertos fue el bosque aleatorio (de la categoría del árbol de decisión) con un 84,8 % de precisión. En la fase de desarrollo, el modelo se integra en una aplicación que permite predecir si un alumno o su grupo es propenso a desertar. Finalmente, en la fase de evaluación, la aplicación se sometió a varios tipos de pruebas para evaluar la funcionalidad de la interfaz gráfica con el usuario final y así mismo pruebas de acierto de la predicción de la deserción estudiantil, los resultados se compararon con los resultados de la fase experimental donde se puede denotar que ambos resultados llegan a coincidir.

Freitas, et al (2020) en su investigación titulada “Sistema IoT para Predicción de Abandono Escolar Utilizando Técnicas de Aprendizaje Automático Basado en Datos Socioeconómicos” el autor destaca que la presente investigación presenta un marco de Internet de las cosas (IoT) para predecir la deserción utilizando métodos de aprendizaje automático

como árbol de decisión, regresión logística, máquina de vectores de soporte, vecinos más cercanos K, perceptrón multicapa y aprendizaje profundo basado en datos socioeconómicos. Con el uso de datos socioeconómicos es posible identificar en el acto de preinscripción quiénes son los alumnos susceptibles de desertar, ya que esta información se llena en el formulario de preinscripción. Este artículo propone la automatización del proceso de predicción mediante un método capaz de obtener información que sería difícil y lenta de obtener para los humanos, contribuyendo a una predicción más precisa. Con la llegada de IoT, es posible crear una herramienta altamente eficiente y flexible para mejorar la gestión y los problemas relacionados con el servicio, que puede proporcionar una predicción de la deserción de los nuevos estudiantes que ingresan a cursos de nivel superior, lo que permite un seguimiento personalizado de los estudiantes para revertir una posible deserción. El enfoque se validó analizando la precisión, la puntuación F1, la recuperación y los parámetros de precisión. Los resultados mostraron que el sistema desarrollado obtuvo una precisión del 99,34 %, una puntuación F1 del 99,34 %, una recuperación del 100 % y una precisión del 98,69 % utilizando arboles de decisión. Por lo tanto, el sistema desarrollado se presenta como una opción viable para su uso en universidades para predecir la probabilidad de que los estudiantes abandonen la universidad.

Antecedentes nacionales

Alvarado (2022) en su trabajo de investigación titulado “Estudio comparativo del nivel de eficacia en los modelos algorítmicos al estimar la deserción de los estudiantes del nivel pregrado en la universidad de Huánuco”, el cual tuvo como objetivo principal comparar el nivel de eficacia en modelos algorítmicos al estimar la deserción de los estudiantes del nivel pregrado en la Universidad de Huánuco. Se define como investigación aplicada que utiliza el enfoque cuantitativo, con alcance descriptivo y tiene un diseño pre experimental. Se recolectó un total de 127,332 casos de estudio, cada uno de los cuales fue un conjunto de datos de cada estudiante que se matriculó entre los semestres 2010-0 y 2018-2, esta data consta de 17 atributos, donde uno de los cuales fue el indicador de deserción; Se seleccionaron 14,800 casos como muestra. Se aplicaron técnicas de ciencia de datos, minería de datos y aprendizaje automático; los modelos de comparados fueron: K vecinos más cercanos (KNN), máquinas de vectores de soporte, perceptrones multicapa y bosques aleatorios; con la ayuda del software desarrollado por los investigadores en el lenguaje Python. La fase de entrenamiento utilizó un conjunto de datos de la población general con un número igual al número de casos de la muestra para que el aprendizaje del modelo algorítmico sea consistente. En la fase de evaluación se procesan los casos correspondientes a las muestras, y la precisión del modelo estuvo cerca del 75%.

Mediante pruebas estadísticas se verificó que los niveles de eficiencia de los modelos algorítmicos difieren en la estimación de la tasa de deserción de los estudiantes de la Universidad de Huánuco. Considerando el nivel de eficiencia basado en la precisión, se concluye que bosque aleatorio es el mejor modelo y KNN es el peor modelo.

Shica (2022) en su investigación denominada “Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa 88331 en Chimbote - 2021” el cual tuvo como objetivo general: desarrollar un modelo de Data Science para la detección de la deserción escolar en la Institución Educativa 88331 en el centro poblado Rinconada, Chimbote. La presente investigación tuvo un enfoque cuantitativo, con un diseño pre experimental. Se analizó el historial de clases de secundaria matriculados desde el año 2011 al 2019 con un total de 804 alumnos en estos años. Las variables examinadas fueron género, fecha de nacimiento, sección, grado, calificaciones del curso, año de estudio, áreas deficientes, comportamiento y situación final. Para el desarrollo del modelo de ciencia de datos de regresión logística, usamos la plataforma Cloud Google Colab usando como lenguaje de programación Python, mediante la metodología CRISP-DM, se trabajó con el 70% de data de entrenamiento o conjunto de entrenamiento y 30% de prueba o testeo, y logramos la máxima precisión. Finalmente, la implementación de un modelo de ciencia de datos de regresión logística aumentó la tasa de retención de la escuela del 84,1 % al 95,5 % en un año académico.

Cevallos y Barahona (2021) en su trabajo de investigación denominado “Modelo para automatizar el proceso de predicción de la deserción en estudiantes universitarios en el primer año de estudio” el cual tuvo como objetivo principal: Implementar un modelo tecnológico para automatizar el proceso de predicción de la deserción en estudiantes universitarios en el primer año de estudio mediante análisis predictivo. Donde el estudio tuvo como finalidad proporcionar una solución que pueda ayudar a reducir la deserción universitaria mediante la aplicación de análisis predictivos y métodos de minería de datos para predeterminar la probabilidad de deserción de los estudiantes, brindando así una mejor visibilidad de las instituciones y oportunidades educativas realizar acciones para resolver este problema. Basado en la disciplina Educational Data Minig (EDM) soportada en la plataforma IBM SPSS Modeler, se desarrolló un modelo de análisis predictivo basado en 15 variables predictivas, 3 etapas de análisis y aplicación de los algoritmos. Para la validación se evaluó el uso de 4 algoritmos de predicción: Regresión Lineal, Redes Bayesianas, Árboles de Decisión y Redes Neuronales; el estudio se realizó en un recinto universitario de Lima. Los resultados muestran que la red bayesiana supera a otros algoritmos en términos de precisión, exactitud, especificidad y tasa de error. En

particular, la red bayesiana logró una precisión del 67,10 %, mientras que los árboles de decisión (el segundo mejor algoritmo) lograron un 61,92 % en iteraciones con una proporción de muestras de entrenamiento de 8:2. Además, las variables "persona deportista" (0,29%), "vivienda propia" (0,20%) y "calificaciones de preparatoria" (0,15%) fueron las variables que más contribuyeron al momento de realizar la predicción.

Novoa (2019) en su trabajo de investigación titulado “Reducción del Riesgo de Deserción Académica mediante seguimiento de alumnos en una universidad” el cual tuvo como objetivo principal: Implementar un sistema web para realizar el seguimiento de alumnos con riesgo de deserción académica. En la presente investigación se realizó una investigación aplicada tecnológica y como alcance o nivel de conocimiento se puede observar que es correlacional, así mismo es de enfoque cuantitativo, donde su población fue definida por 80 estudiantes de pregrado. El autor concluyó Esta tesis mejorará la calidad de los servicios en el proceso de seguimiento y evaluación académica a través de la gestión académica automatizada, que permite el seguimiento continuo y continuado del rendimiento académico de los estudiantes que se encuentran en riesgo de deserción a través de evaluaciones de asignaturas durante el semestre, así como la preparación para el curso académico. informes para gestionar información de rendimiento académico e interacciones con profesores y administradores en las unidades académicas encargadas de coordinar el seguimiento y evaluación de los estudiantes.

Antecedentes locales

Haciendo una investigación exhaustiva en los diversos repositorios a nivel local no se han podido determinar estudios que trabajen con mis variables.

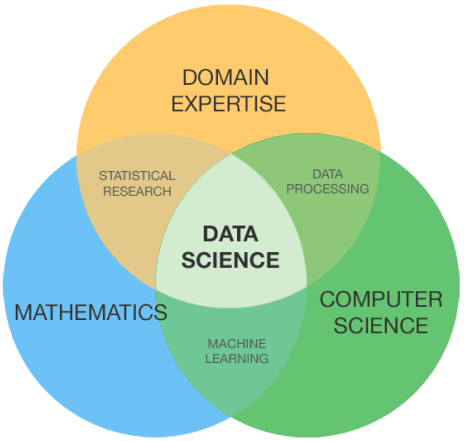
Bases teóricas

A. Ciencia de datos

Data Science hace referencia a un nuevo campo de trabajo relacionado con la recopilación, preparación, análisis, visualización, gestión y almacenamiento de grandes cantidades de información. Si bien el nombre Ciencia de datos parece conectarse más fuertemente con áreas como las bases de datos y la informática, donde se necesitan muchos tipos diferentes de habilidades, incluidas habilidades matemáticas. (Stanton, 2013)

Figura 1

Data Science



Nota. Palmer (2015)

Así mismo según Ozdemir & Kakade (2018), la ciencia de datos es el arte y la ciencia de adquirir conocimiento a través de datos, o dicho de otra manera se trata de cómo tomamos datos, los usamos para adquirir conocimiento y luego usamos ese conocimiento para; tomar decisiones, predecir el futuro, entender el pasado / presente o crear nuevas industrias /productos.

El data science es un campo de conocimiento relativamente nuevo y, aunque ha sido estudiado por la comunidad informática durante muchos años, sus límites aún son borrosos y dinámicos. Sus componentes incluyen álgebra lineal, modelado estadístico, visualización, lingüística computacional, análisis de gráficos, aprendizaje automático, inteligencia empresarial y almacenamiento y recuperación de datos (Boschetti & Massaron, 2018).

B. Análisis predictivo

Según Hashmi & Sheikh (2012) el análisis predictivo es una rama de la minería de datos, que combina el conocimiento las técnicas de análisis estadístico con los conocimientos del negocio, para sacar información predictiva oculta y para poder predecir tendencias y patrones futuros mediante el análisis de grandes cantidades de datos históricos con diversas variables. Variables que se pueden medir para predecir el comportamiento futuro de una persona o entidad.

Es importante comprender que la aplicación del análisis predictivo requiere un conjunto de datos grande y de suficiente calidad. Entonces, para mejorar la precisión de las predicciones, es importante definir claramente los conceptos que se van a predecir, porque solo así se pueden encontrar los modelos más valiosos para quienes tienen que predecir (Siegel, 2016).

C. Modelo predictivo

El modelo es una representación simplificada de la realidad creada para servir a algún propósito para el usuario y que se basará en ciertos datos. El propósito puede ser sobre varios temas, pero, generalmente, existe para preservar la información relevante o para simplificar aún más la información. Además de las dos cosas mencionadas anteriormente, un modelo también se puede usar para pronosticar o predecir lo que sucederá en el futuro, según los datos que la compañía tiene ahora, para que la empresa pueda tomar decisiones con anticipación y ayudarles a aumentar las ganancias, dar soporte a los clientes, ofrecer mejores productos o, al menos, reducir sus riesgos. (Jones, 2019)

Cuando estamos trabajando en la ciencia de datos, estos modelos están ahí para crear una buena imagen de los datos. Hace que los datos sean más fáciles de leer, por lo que es más fácil tomar buenas decisiones a partir de esos datos. (Jones, 2019)

Un modelo predictivo llega a ser una función matemática capaz de aprender la correlación entre un conjunto de variables de datos de entrada (normalmente contenidas en uno o más registros) y una variable de respuesta u objetivo. (Guazzelli, 2012)

Así mismo se puede comprender el modelo predictivo como un conjunto de técnicas de análisis matemático y estadístico encaminadas a encontrar relaciones lógicas y cuantitativas entre el objetivo, la respuesta o la variable dependiente y los predictores u otras variables independientes. La necesidad de definir un modelo de predicción es primero cuantificar el valor del predictor en una ubicación futura y luego conjugar todas las variables en una relación logístico-matemática para que el valor de la variable objetivo pueda cuantificarse en esa ubicación futura. (Dickey, State, & Raleigh, 2012)

Aplicar modelos predictivos a cualquier contexto o realidad implica definir exactamente cuál es la variable objetivo para la predicción y cuáles son los factores que afectan a los posibles valores que puede tomar la variable objetivo. Durante este análisis, es importante considerar que cada factor encontrado tiene una correlación e influencia diferente en el resultado objetivo. Esencialmente, estos factores pueden ser de bajo impacto (generalmente no relevante para el modelo), alto impacto (relevante en el modelo), donde aquellos con impacto medio se incluirán en el modelo si se requiere o se considera en el análisis. (Dickey, State, & Raleigh, 2012)

D. Machine Learning

Se dice que un programa de computadora aprende de la experiencia E con respecto a alguna tarea T y alguna medida de desempeño P, si su desempeño en T, medido por P, mejora con la experiencia E. (Mitchell, 1997)

La empresa multinacional de origen estadounidense, Microsoft (2022) define al machine learning como un subconjunto de inteligencia artificial que incluye técnicas (como el aprendizaje profundo) que permiten a las máquinas realizar tareas mejor con experiencia. El proceso de aprendizaje se basa en los siguientes pasos:

- Agregar datos al algoritmo. (En este paso, puede proporcionar información adicional al modelo, por ejemplo, extrayendo características).
- Utilizar estos datos para entrenar el modelo.
- Realizar pruebas e implementar el modelo.

- Use el modelo implementado para realizar una tarea de predicción automatizada.

Así mismo IBM (2020) lo define al aprendizaje automático como una rama de la inteligencia artificial (IA) y la informática que se centra en el uso de datos y algoritmos para imitar la forma en que los humanos aprenden y mejorar gradualmente su precisión.

El aprendizaje automático es una parte esencial del creciente campo de la ciencia de datos. Utilizando técnicas estadísticas, los algoritmos se entrenan para clasificar o predecir y descubrir información clave en un proyecto de minería de datos. Estos conocimientos luego impulsan la toma de decisiones comerciales y de aplicaciones, lo que idealmente impacta en las métricas de crecimiento clave. Donde se pedirá a los científicos de datos que ayuden a identificar las preguntas comerciales más relevantes y los datos para responder esas preguntas. (IBM, 2020)

Principales problemas del Machine Learning

En resumen, dado que su tarea principal es seleccionar un modelo y entrenarlo con algunos datos, las dos cosas que pueden salir mal son "modelo incorrecto" y "datos incorrectos". Por lo tanto, Géron (2022) plantea los siguientes problemas en el machine learning:

- Cantidad insuficiente de datos de entrenamiento. La mayoría de los algoritmos de aprendizaje automático requieren grandes cantidades de datos para funcionar correctamente; Incluso los problemas muy simples requieren miles de ejemplos, mientras que los problemas complejos, como el reconocimiento de imágenes, requieren millones de ejemplos.

- Datos de entrenamiento no representativos. Dada la descripción del párrafo anterior, dicha situación no es precisamente muy crítica, ya que muchas veces hay miles de registros de datos, pero no son representativos, en otras palabras, un conjunto de 100 registros de datos está muy estructurado y probablemente funcionará mejor que uno de miles sin correlación entre sus variables. Es muy importante hacer predicciones más precisas utilizando un conjunto de datos que represente el modelo.

- Datos de baja calidad. Si los datos están llenos de errores, los datos nulos lo que se conoce como "ruido de datos" y será más difícil para el modelo detectar patrones en los datos. Es muy importante y significativo tomarse el tiempo para limpiar los datos como se describe en las secciones anteriores.
- Características irrelevantes. Un modelo solo puede aprender si los datos contienen características relevantes y no demasiadas características irrelevantes. Lo importante es elegir las características más útiles y combinarlas para obtener la mejor generalización.
- Sobre entrenamiento de los datos. En machine learning esto se denomina Overfitting o sobreajuste, que significa que el modelo funciona muy bien en los datos de entrenamiento, pero no generaliza bien para los datos de validación.
- Sub entrenamiento de los datos. También llamado Underfitting, supone una adaptación insuficiente al modelo, en otras palabras, es cuando el modelo es muy simple para aprender del comportamiento de los datos.

E. Random forest

Random forest o bosque aleatorio es una técnica de enjambre. La base de estos métodos es realizar una combinación de diferentes clasificadores utilizando alguna técnica de agregación. Estas técnicas de conjunto contienen buenas propiedades contra el sobreajuste, como la reducción de la varianza del clasificador final. Esta característica aumenta la solidez del clasificador y, a menudo, da como resultado un rendimiento del clasificador muy bueno (Igual & Seguí, 2017).

El secreto de los bosques aleatorios es ensamblar árboles de decisión simples mientras son muy diferentes entre sí. Como son diferentes, tampoco están correlacionados, lo que es muy beneficioso a la larga, porque los diferentes resultados de cada árbol se agrupan, eliminando las diferencias (Boschetti & Massaron, 2018). Como se mencionó anteriormente, los bosques aleatorios tienen un rendimiento superior en comparación con los árboles de decisión, no solo en términos de resultados más confiables, sino también en términos de eficiencia de la CPU, por lo que los bosques aleatorios también son ideales cuando se trabaja con grandes cantidades de datos.

F. Validación del modelo

Esta parte del trabajo trata sobre el modelado de datos y requiere la validación de los resultados obtenidos por el modelo de predicción, es decir, el grado de ajuste del modelo a los datos, ya que existen dos indicadores para validar los resultados como se muestra a continuación.

Matriz de confusión

Una forma muy útil de evaluar un clasificador es observar y calcular la matriz de confusión. La principal idea de esta matriz es contabilizar la cantidad de veces que las instancias de una clase se clasifican como otra. Esto significa que esta matriz compara los valores reales con los valores pronosticados y contabiliza aquellas predicciones correctamente, así como aquellos que no se predijeron bien. Cada fila en una matriz de confusión representa una clase real, mientras que cada columna representa una clase predicha (Géron, 2022). El valor en la posición (0,0) de la matriz se denomina Verdaderos Negativos (VN), mientras el que se encuentra en la posición (0,1) se hace llamar Falso Positivo (FP). El valor en la posición (1,0) se denomina Falso Negativo (FN) y el que está en la posición (1,1) Verdaderos Positivos (VP). Un clasificador perfecto solo tendría verdaderos positivos y verdaderos negativos, por lo que esta matriz de confusión solo tendría valores distintos de cero en su diagonal (de arriba izquierda, a abajo derecha).

Para obtener aún más información sobre el rendimiento del modelo, debemos examinar otras métricas como la precisión, el recall, accuracy y el F1 score. (Kreiger, 2020)

- **La precisión** es el número de miembros correctamente identificados de una clase dividido por todas las veces que el modelo predijo esa clase.

$$Precisión = \frac{VP}{VP + FP}$$

- **El Recall** es el número de miembros de una clase que el clasificador identificó correctamente dividido por el número total de miembros de esa clase.

$$Recall = \frac{VP}{VP + FN}$$

- **Accuracy** Es el porcentaje de predicciones correctas

$$Accuracy = \frac{VN + VP}{VN + FP + FN + VP}$$

- **F1 score** es un poco menos intuitiva porque combina precisión y recuperación en una métrica. Si tanto la precisión como la recuperación son altas, F1 también lo será. Si ambos son bajos, F1 será bajo. Si uno es alto y el otro bajo, F1 será bajo. F1 es una forma rápida de saber si el clasificador es realmente bueno para identificar miembros de una clase o si está encontrando atajos (por ejemplo, simplemente identificando todo como miembro de una clase grande).

$$F1 = \frac{VP}{VP + \frac{FN + FP}{2}}$$

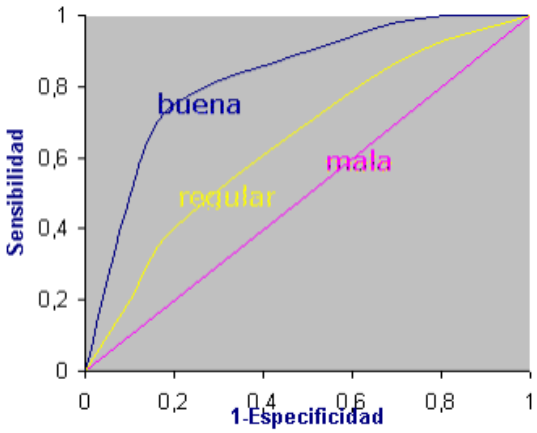
Curva ROC

La curva característica operativa (ROC por sus siglas en inglés Receiver Operating Characteristic) del receptor es otra herramienta común utilizada con clasificadores binarios. La curva ROC traza la Tasa de Verdaderos Positivos (otro nombre para el recall) frente a la Tasa de Falsos Positivos (TFP). Los TFP es la fracción de negativos que se clasifican erróneamente como positivos y es igual a "1 – la Tasa de Verdaderos Negativos (TFN)", el cual es la proporción de negativos que se clasifican erróneamente como negativos. El TFN también se llama especifidad. Por lo tanto, la curva ROC traza de recall versus 1 – especifidad. (Géron, 2022)

Una forma de comparar los clasificadores es medir el área bajo la curva (AUC). Un clasificador perfecto tendrá un AUC igual a 1, mientras que un clasificador puramente aleatorio tendrá un AUC igual 0.5.

Figura 2

Curva ROC



Nota. Géron (2022)

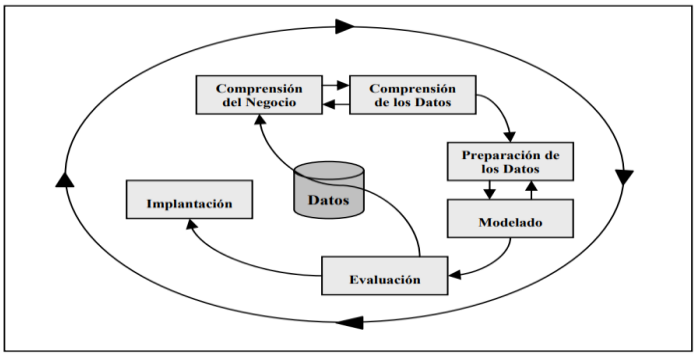
G. Metodología CRISP-DM

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining por sus siglas en inglés), fue desarrollada e introducida en el mercado en el año 1999 por las empresas europeas NCR (Dinamarca), AG(Alemania), SPSS (Inglaterra), OHRA (Holanda), Teradata, SPSS, y Daimer-Chrysler, la cual consistió en la desarrollar una guía de referencia distribuida gratuitamente para implementar la minería de datos en todo el mundo.

La metodología recomienda realizar un proyecto de minería de datos en las siguientes 6 etapas:

Figura 3

Modelo de proceso CRISP-DM



Nota. Gallardo (2009)

1. **Comprensión del Negocio**

La primera fase de la Guía de referencia de CRISP-DM, conocida como comprensión del negocio o problema y quizás la fase más importante, combina las tareas de comprender los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional para traducirlos en objetivos técnicos y en un plan del proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Sin comprender estos objetivos, no importa cuán sofisticado sea el algoritmo, no se pueden obtener resultados confiables. Sacar el máximo provecho de la minería de datos requiere una buena comprensión del problema que está tratando de resolver, lo que le permitirá recopilar los datos correctos e interpretar los resultados correctamente. En esta etapa, la capacidad de traducir el conocimiento comercial adquirido en problemas de minería de datos y planes preliminares destinados a lograr los objetivos comerciales es fundamental. (Gallardo, 2009)

2. **Comprensión de los Datos**

La segunda fase, fase de comprensión de los datos, consiste en la recogida inicial de datos, que pretende establecer un primer contacto con los problemas, familiarizarse con ellos, determinar su calidad e identificar las relaciones más evidentes para definir la primera hipótesis. Esta fase y las siguientes dos fases son las que requieren más esfuerzo y tiempo en un proyecto de minería de datos.

En general, si la organización tiene una base de datos corporativa, es mejor crear una nueva base de datos ad-hoc para el proyecto de minería de datos, ya que puede ocurrir un acceso frecuente y pesado a la base de datos durante el proceso de desarrollo del proyecto y se puede cambiar lo que puede causar muchos problemas. (Gallardo, 2009)

3. **Preparación de los Datos**

En esta etapa, cuando se completa la recopilación inicial de datos, están listos para adaptarlos a técnicas posteriores de minería de datos que se utilicen en su posterioridad, como técnicas de visualización de datos, búsqueda de

relaciones entre variables u otros fines de investigación. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. (Gallardo, 2009)

4. **Modelado**

En esta fase de CRISP-DM, se selecciona la técnica de modelado más adecuada para un proyecto de minería de datos específico. Elija la tecnología que utilizará en este paso en función de los siguientes criterios:

- Ser apropiada al problema
- Disponer de datos adecuados
- Cumplir los requisitos del problema
- Tiempo adecuado para obtener un modelo
- Conocimiento de la técnica

Antes de modelar los datos, debe decidir un método de evaluación del modelo que pueda determinar qué tan buenos o malos son los datos. Después de completar estas tareas generales, pasamos a la generación y evaluación del modelo. Los parámetros utilizados para generar el modelo dependen de las características de los datos y de las características de precisión que debe alcanzar el modelo. (Gallardo, 2009)

5. **Evaluación**

En este paso, el modelo se evalúa en términos de si cumple con los criterios de éxito del problema. También se debe tener en cuenta que la confiabilidad calculada para el modelo se aplica solo a los datos analizados. Con base en los resultados obtenidos, se debe revisar el proceso para repetir los pasos anteriores que pueden haber causado errores. Considere que se pueden usar múltiples herramientas para interpretar los resultados. (Gallardo, 2009)

6. Implementación

En esta fase, una vez que se ha creado y validado el modelo, el conocimiento resultante se traduce en acciones en el proceso de negocio, ya sea por los analistas que recomiendan acciones basadas en las observaciones del modelo y sus resultados, o aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como, por ejemplo, en análisis de riesgo crediticio, detección de fraude, etc.

A menudo, el proyecto de minería de datos no termina con la implementación del modelo, ya que los resultados deben registrarse y presentarse de forma comprensible para el usuario, con el objetivo de aumentar el conocimiento. Por otro lado, durante la etapa de desarrollo se debe asegurar el mantenimiento de la aplicación y la posible distribución de resultados. (Gallardo, 2009)

H. Deserción estudiantil

La deserción universitaria, entendida también como el abandono de los estudios, de manera voluntaria o involuntaria, el cual pueden ser por diferentes causas propias o fuera de las instituciones educativas, lo que significa una pérdida de capital para el estudiante, su familia, la sociedad y el país y conlleva un sentimiento de decepción o frustración. (Cevallos & Barahona, 2021)

Para Viale (2014), se comprende que “la deserción estudiantil está relacionado con intereses personales, motivacionales, los hábitos de estudio, el planes curriculares y reglamentos académicos propios de cada institución”. El mismo autor explica desde su estudio: “los casos de deserción más frecuentes se dan durante el primer ciclo de estudios de estudio y se pueden observar diversos tipos de deserciones, como aquellas que dejan de estudiar principalmente por motivos personales, como el embarazo o problemas de salud; la de aquellos que tienen un rendimiento académico deficiente; aquellos con bajo rendimiento académico y aquellos que sienten que la carrera elegida no es para ellos”. En otras palabras, la mayoría de los casos de deserción se producen durante los primeros ciclos de la educación superior, cuando los estudiantes están aprendiendo y conociendo el entorno universitario y tratando de adaptarse.

Rodríguez y Hernández (2008) expresa que la deserción estudiantil puede entenderse simplemente como una liberación de las obligaciones del estudiante o de la universidad bajo las reglas académicas de admisión por cualquier motivo. Esto tiene graves consecuencias económicas, académicas y sociales para ambas partes.

En ese mismo contexto, Himmel (2002) la deserción se refiere a abandonar prematuramente un programa de estudio antes de obtener un título o grado, y se considera que es por un período de tiempo suficiente para impedir que el estudiante regrese.

La deserción también puede entenderse como suspensión permanente o temporal, voluntaria o forzosa, que puede definirse de diferentes formas, por ejemplo: abandono de una ocupación, abandono de las instituciones y abandono del sistema de educación superior. (Mori, 2012)

Según Girón & González (2005), la deserción estudiantil universitaria tiene consecuencias emocionales, económicas y sociales no solo para el propio estudiante, sino también su entorno inmediato. Este problema de deserción estudiantil cuenta con muchas variables, por las que podemos dividir en aquellas que pertenecen al ámbito pedagógico y no pedagógico (Tejedor & García-Valcárcel, 2001), como por ejemplo variables psicológicas, sociofamiliares y de identidad, que influyen en el rendimiento académico de los estudiantes que recién ingresaron a la universidad. Considerando estos dos tipos de variables (las pedagógicas y las no pedagógicas), cada universidad desarrolla sus propios programas para facilitar la adaptación a los nuevos estudiantes dentro de la vida universitaria, sin embargo, en gran parte de los casos, estos programas pertenecen a departamentos o campos académicos distintos, con estructura organizacional diferente; con lo cual la orientación que reciben los alumnos que recién han ingresado se hace desde distintas concepciones. En lugar de ayudar al estudiante, termina confundiéndolo más y no se alcanza el objetivo de ayudarlo a ingresar satisfactoriamente a la universidad.

Factores de la deserción estudiantil

a) Factores personales

Según Zambrano et al. (2018) estos factores son las características personales del estudiante, como es la falta de actitud de logro en el crecimiento profesional, la falta de compatibilidad del tiempo empleado para trabajar y los estudios, el poco interés que tiene por los estudios, la carrera y la institución; también la poca motivación que brinda la universidad y las expectativas que tenía el estudiante sobre la importancia de la carrera que eligió estudiar.

b) Factores académicos

Castillo et al. (2019) indica que estos factores pueden ser la orientación profesional, tipo de colegio en el que se gradúo, rendimiento académico universitarios, hábitos de estudio, calificación en el examen de admisión, insatisfacción con el plan de estudio, número de créditos por ciclo, cantidad de veces que lleva un curso, entre otros, que llegan a optar por el abandono o deserción universitaria.

c) Factores socioeconómicos

De acuerdo a Albarrán (2019) un estudiante que cursa una carrera, al tener una falta de recursos económicos, tiene la opción de buscar un trabajo de medio tiempo para poder cubrir sus gastos, pero muchos estudiantes no logran aprobar el plan de estudios, ya que el trabajo obstaculiza el adecuado desarrollo de sus actividades académicas, teniendo como consecuencia que los estudiantes no culminen su formación académica en el tiempo previsto u opten por dejar sus estudios.

Para Zambrano et al. (2018) las condiciones económicas desfavorables de un estudiante, añadiendo la ausencia de financiamiento y desarticulación familiar causan la deserción o la desaprobación de las asignaturas del plan de estudio, dando lugar a problemas de salud física y mental, lo que en muchos casos llega a generar un problema de elevado costo para muchos países, por el abandono o deserción de los estudios.

Definición de términos

La terminología técnica empleada en esta presente investigación, se detalla a continuación:

- **Aprendizaje Supervisado.** Es un tipo de aprendizaje automático que consiste en algoritmos que aprenden de un conjunto de datos etiquetados que sirven como ejemplos de entrenamiento para generalizar a otro conjunto de datos de salida. (Igual & Seguí, 2017)

- **Datos:** Cualidad o propiedad derivada de la observación, medición o alguna forma de percepción. Es una forma de interpretar la realidad, transformándola en metas o categorías que conocemos y facilitará su procesamiento para diferentes fines. (Leek, 2015)

- **Dataset:** Es un conjunto de objetos o instancias que se utilizarán para entrenar el algoritmo y realizar su validación, que son los resultados de la información de la base de datos, mediciones, compilaciones, etc. Se requiere procesamiento porque los datos que se encuentran en el mundo real no están normalizados para su uso en los diferentes algoritmos. (Witten, Frank, & Hall, 2017)

- **Jupyter.** Originalmente llamado IPython, el entorno de desarrollo web integrado se limitaba al uso del lenguaje Python. El notebook de trabajo IPython ha sido desvinculados del software IPython y todos sus componentes viejos fueron heredados a Jupyter. (Igual & Seguí, 2017)

- **Matplotlib.** Es una biblioteca o librería construida bajo el lenguaje Python, que contiene todos los bloques de construcción necesarios para crear gráficos a partir de matrices. (Boschetti & Massaron, 2018)

- **NumPy.** Es una librería de análisis construida sobre el lenguaje de programación Python. Proporciona a los usuarios matrices multidimensionales con varias funciones matemáticas adjuntas para operarlas. (Boschetti & Massaron, 2018)

- **Pandas.** Es una librería construida bajo el lenguaje de programación Python. Con Pandas, puede cargar datos fácilmente desde cualquier fuente.

También puede dividir, cortar, manejar elementos faltantes, agregar y más. (Boschetti & Massaron, 2018)

- **Python.** Es un lenguaje de programación completo, pero tiene excelentes funciones para los nuevos programadores y es perfecto para aquellos que nunca han programado antes. Es uno de los lenguajes de programación más flexibles. (Igual & Seguí, 2017)

1.1. Justificación

1.1.1. Justificación social

La deserción de estudiantes puede llegar a ser un problema crítico dentro de la comunidad universitaria dado que esto se puede efectuar por diferentes factores ya sean personales de los estudiantes o factores donde intervenga la decisión de la universidad (Mori, 2012), por lo tanto, a partir de esta investigación se buscó disminuir esta brecha generada a lo largo de varios años dentro de la universidad, permitiendo localizar a los estudiantes con riesgo de deserción y se puedan tomar acción para su permanencia universitaria.

1.1.2. Justificación económica

Mediante el modelo predictivo de deserción de estudiantes en la universidad Nacional Santiago Antúnez de Mayolo la universidad se beneficiará a corto y mediano plazo, tales como la permanecía de los estudiantes estudiantil, gastos administrativos y educativos invertidos en cada estudiante, seguimiento a los estudiantes con riesgo de deserción (Mori, 2012).

1.1.3. Justificación tecnológica

Hoy en día las soluciones a partir de algoritmos de aprendizaje automático son más comunes para la solución de problemas cotidianos, dado que ofrecen una alternativa mucho más rápida para la toma de decisiones dentro una organización y adelantarse a los sucesos que se están por suscitar.

Así mismo, cada vez aparecen nuevas tecnologías que te ayudan a realizar este tipo de proyectos tales como Google Colab o Microsoft Azure que funcionan con los lenguajes de programación necesarios para realizar este tipo de tareas tales como Python y R (Boschetti & Massaron, 2018).

1.1.4. Justificación operativa

El modelo predictivo de deserción estudiantil, es una opción de reconocimiento de estudiantes con posibilidades de desertar a la universidad dando opción a la universidad de mejorar el modelo predictivo y usar las respuestas para la creación de estrategias para la retención de estudiantes.

1.1.5. Justificación legal

La presente tesis se justifica legalmente mediante la ley Nº 27658 Ley Marco de Modernización de la Gestión del Estado donde se menciona que la priorización de la labor de desarrollo social en beneficio de los sectores menos favorecidos, mejorando, entre otras acciones, la prestación de los servicios públicos por lo que es necesario buscar la permanencia de los estudiantes dentro de la institución evitando que estos deserten de sus respectivas carreras.

1.2. Planteamiento del problema

En las universidades en la actualidad sé puede notar que la mayoría de los estudiantes que acaban de ingresar a la universidad no se desempeñen satisfactoriamente. Existe una brecha académica entre lo que se enseña en las escuelas y lo que exigen las universidades. Este es un problema con muchas variables y está presente en mundo. Un alto porcentaje de alumnos que reprobaron una o más cursos en el primer ciclo los lleva a matricularse nuevamente en los mismos cursos. Sin embargo, muchos reprueban nuevamente y eventualmente abandonan la universidad, lo que provocó un aumento en el número de estudiantes que no se gradúan aumentan en los países, de esta manera también afecta a los presupuestos que se plantean para cada universidad. (Viale, 2014)

El fenómeno de la deserción de estudiantes se ha normalizado en la mayoría de los países latinoamericanos. Por ello, muchos de estos países la consideran una de las principales prioridades del sector educativo (Sánchez, Barboza, & Castilla, 2017). Por consiguiente, se han desarrollado varias propuestas de investigación que involucran la participación directa de las unidades de educación superior y las agencias gubernamentales, y se han logrado muchos avances importantes en el abordaje de este tema. Donde Colombia se destaca como el país que más información científica aporta, seguido de México, Chile y Argentina.

Culminar la educación superior es pieza principal para el desarrollo del Perú, dado que permite la generación de la fuerza capital humana, ente primordial de todo sistema. La educación superior es un tema primordial para lograr un desarrollo económico sostenible a largo plazo, y por ello, en los últimos años, las instituciones universitarias han puesto gran énfasis en el rendimiento académico de los estudiantes y los factores que los afectan, la investigación y análisis del tema brindan información como una herramienta para identificar indicadores que orienten a la toma de decisiones en la educación superior. (Castañeda & López, 2015)

En el peruano, la revisión de la literatura científica muestra que hay pocas investigaciones acerca de este tema. Las investigaciones generalmente se realizan para medir las tasas de deserción e informar los factores que influyeron para tomar dicha decisión. Al respecto, Mori (2012) indica que la tasa anual de deserción estudiantil universitaria alcanza el 17%. Además, el mayor porcentaje se encuentra en los alumnos de los primeros semestres de estudios.

En la Universidad Nacional Santiago Antúnez de Mayolo actualmente no se realiza correcto análisis de la deserción estudiantil y los efectos que este puede causar, mediante la presente investigación se busca aplicar las nuevas tendencias tecnológicas para poder abarcar el presente problema con una postura crítica y veraz ante las autoridades pertinentes.

1.2.1. Formulación del problema

Problema general

¿La aplicación de un modelo predictivo determina la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo -2022?

Problemas específicos

PE01: ¿Los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022?

PE02: ¿Los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022?

PE03: ¿Los factores socioeconómicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022?

1.3. Objetivos General

Determinar mediante un modelo predictivo la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo - 2022.

1.3.1. Objetivos específicos

OE01: Establecer si los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

OE02: Establecer si los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

OE03: Establecer si los factores socioeconómicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

1.4. Hipótesis Significativa

El modelo predictivo determina la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

1.5. Hipótesis Nula

El modelo predictivo no determina la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022

Hipótesis especificas

HE01: Los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HE02: Los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HE03: Los factores socioeconómicos influyen en para la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

II. MATERIALES Y MÉTODOS

2.1. Variables

2.1.1. Variable independiente

Modelo predictivo

2.1.2. Variable dependiente

Deserción de estudiantes

2.2. Operacionalización de variables

Tabla 1:

Matriz de operacionalización de variables

Variables	Definición conceptual	Definición operacional	Dimensiones	Indicadores	Escala de medición
Variable independiente: Modelo predictivo	Es la creación de una función matemática capaz de aprender mediante un conjunto de variables de entrada para obtener una respuesta o variables de destino (Guazzelli, 2012). Así mismo es un conjunto de técnicas matemáticas y de análisis estadístico que su finalidad es encontrar una relación lógica y cuantitativa entre un objetivo, respuesta o variable dependiente y factores de predicción u otras variables independientes (Dickey, State, & Raleigh, 2012).	La predicción se realizará mediante un modelo basado en diferentes enfoques de algoritmos de machine learning, a través de los datos proporcionados por los factores que intervienen en la deserción de estudiantes.	Funcional	Precisión	Escala de intervalo
				Recall	Escala de intervalo
				Accuracy	Escala de intervalo
				F1 score	Escala de intervalo
			Operativo	Tiempo de ejecución	Escala de intervalo
Variable dependiente: Deserción de estudiantes	Es una suspensión permanente o temporal voluntaria u obligatoria y se puede distinguir de diferentes formas, por ejemplo: retiro de la profesión, abandono de la institución y retiro del sistema de educación superior (Mori, 2012). También se aplica como deserción estudiantil a darse de baja de un programa de estudio antes de obtener un título o grado y considerando que el tiempo es lo suficientemente largo como para impedir que el estudiante se reincorpore. (Himmel, 2002).	La deserción de estudiantes será evaluada mediante la técnica de observación con los instrumentos de: - Ficha socioeconómica de los estudiantes (factores personales y económicos) - Ficha de récord académico (factores académicos).	Factores personales	Sexo	Nominal
				Edad	Escala de intervalo
				Lugar de procedencia	Nominal
				Estado civil	Nominal
				Modalidad de ingreso	Nominal
				Ciclo actual	Ordinal
			Factores académicos	Numero de cursos matriculados	Escala de intervalo
				Numero de cursos desaprobados	Escala de intervalo
				Créditos matriculados	Escala de intervalo
				Notas	Escala de intervalo
			Factores socioeconómicos	Situación socioeconómica	Nominal

Fuente: Elaboración propia

2.3. Definición conceptual

Modelo predictivo

Es la creación de una función matemática capaz de aprender mediante un conjunto de variables de entrada para obtener una respuesta o variables de destino (Guazzelli, 2012). Así mismo es un conjunto de técnicas matemáticas y de análisis estadístico que su finalidad es encontrar una relación lógica y cuantitativa entre un objetivo, respuesta o variable dependiente y factores de predicción u otras variables independientes (Dickey, State, & Raleigh, 2012).

Deserción de estudiantes

Es una suspensión permanente o temporal voluntaria u obligatoria y se puede distinguir de diferentes formas, por ejemplo: retiro de la profesión, abandono de la institución y retiro del sistema de educación superior (Mori, 2012). También se aplica como deserción estudiantil a darse de baja de un programa de estudio antes de obtener un título o grado y considerando que el tiempo es lo suficientemente largo como para impedir que el estudiante se reincorpore. (Himmel, 2002).

2.4. Definición operacional

Modelo predictivo

La predicción se realizará mediante un modelo basado en diferentes enfoques basados en algoritmos de machine learning, a través de los datos proporcionados por los factores que intervienen en la deserción de estudiantes.

Deserción de estudiantes

La deserción de estudiantes será evaluada mediante la técnica de observación con los instrumentos de:

- Ficha socioeconómica de los estudiantes (factores personales y económicos)
- Ficha de récord académico (factores académicos).

III. METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Tipo de estudio

- Según lo que mide: Longitudinal

  Los datos se recopilaron de la misma población en diferentes momentos durante un período de tiempo para examinar cómo cambiaron con el tiempo. (Bernal, 2010)

- **Enfoque**

  La presente investigación tuvo el enfoque cuantitativo dado que es secuencial y probatorio. Cada etapa precede a la siguiente y no podemos “brincar” o eludir pasos. El orden es riguroso, aunque desde luego, podemos redefinir alguna fase. (Hernández, Fernández, & Baptista, 2014)

- **Nivel de la investigación**

  Investigación explicativa

  La investigación explicativa se encarga de encontrar las causas de los hechos estableciendo relaciones causales. En este sentido, la investigación explicativa puede utilizar la prueba de hipótesis para determinar la causa (investigación post facto) y el efecto (investigación experimental). Sus hallazgos y conclusiones forman el nivel más profundo de conocimiento. (Arias, 2012)

3.2. El diseño de investigación

- Pre – Experimental:

  Esto implica dar un estímulo o tratamiento a un grupo y luego usar una medida de una o más variables para ver el nivel de esa variable en el grupo. Este diseño no cumple con los requisitos de un experimento "puro". No hay manipulación de la variable independiente (niveles) o grupos de contraste (ni siquiera el mínimo de presencia o ausencia). de variables independientes (niveles) o grupos de contraste (ni siquiera mínima presencia o ausencia). Tampoco hay referencia previa al nivel de grupo de la variable dependiente antes de la estimulación. Es imposible establecer la causalidad con certeza y controlar las fuentes de invalidación interna. (Hernández, Fernández, & Baptista, 2014)

$$G \rightarrow X \rightarrow O$$

G: Grupo de sujetos o casos

X: Tratamiento, estímulo o condición experimental (presencia de algún nivel o modalidad de la variable independiente).

O: Una medición de los sujetos de un grupo (prueba, cuestionario, observación, etc.). Si aparece antes del estímulo o tratamiento, se trata de una preprueba (previa al tratamiento). Si aparece después del estímulo se trata de una posprueba (posterior al tratamiento).

3.3. Población y muestra

- **Unidad de análisis**

La unidad de análisis considerada para esta investigación fueron los casos de estudio de cada estudiante que ingreso entre los semestres 2010-I al 2019-I, cada uno de estos casos de estudio representaron un conjunto de atributos de un determinado estudiante de la Universidad Nacional Santiago Antúnez de Mayolo.

- **Población**

Una población o más precisamente una población objetivo es un conjunto finito o infinito de elementos que comparten características comunes y las conclusiones del estudio fueron amplias. Está determinada por la pregunta y los objetivos de la investigación. (Arias, 2012)

La población para esta investigación fue conformada por los 6,440 casos de estudio de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo que ingresaron entre los semestres 2015-I al 2019-I.

- **Muestra**

La muestra es un subconjunto representativo y finito que se extrae de la población accesible. (Arias, 2012)

La muestra tomada para la presente tesis es de tipo no probabilístico, en específicamente es muestra intencionada. La muestra está conformada por los 5,657 casos de estudio de los estudiantes de la Universidad Nacional Santiago Antúnez

de Mayolo que ingresaron entre los semestres 2015-I al 2019-I dado que son los casos de estudio que cuentan con la información necesaria para realizar el estudio.

3.4. Técnicas de instrumentos de recolección de datos

Para esta investigación se utilizó la técnica de análisis documental, donde para la obtención de datos se tuvo acceso a los centros de recolección de datos de la universidad los cuales son la oficina de admisión y la oficina de general de estudios. El objetivo es poder considerar todos los aspectos o características que son relevantes para los estudiantes, de modo que podamos crear suficientes conjuntos de datos a partir de ahí para su posterior procesamiento y análisis. Mediante el data set final que se obtuvo a partir de validaciones estadísticas a los datos.

3.5. Técnicas de análisis y pruebas de hipótesis

Para la presente investigación se aplicó la metodología de CRISP-DM, dado que proporciona una descripción estandarizada del ciclo de vida del proyecto de análisis de datos estándar que corresponde al modelo de ciclo de desarrollo de software para completar la técnica de software.

Para analizar los resultados obtenidos por el modelo predictivo se utilizó la matriz de confusión, ya que es una herramienta que permite visualizar el rendimiento de los algoritmos de aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones para cada clase, mientras que cada fila representa las ocurrencias de la clase real.

IV. RESULTADOS DE LA INVESTIGACIÓN

4.1. Descripción de trabajo de campo

En este apartado se presentará la aplicación de modelos de aprendizaje automático basado en teorías de Data Science y CRISP-DM, mediante conceptos de minería de datos tales como técnicas de limpieza, exploración y modelamiento de datos en cada caso de estudio de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo.

4.1.1. Definición del problema

Para la evaluación desde una perspectiva metodología, es de relevancia establecer la problemática que existe en un contexto real, el cual fue evidenciado a través del análisis descriptivo de los casos de estudio que serán utilizados en la presente investigación los cuales están comprendidos entre los estudiantes de que ingresaron en los semestres 2015-I al 2019-I.

Durante los semestres que serán analizados se puede observar que hubo un total de 6440 ingresantes de los cuales al realizar un análisis descriptivo de las matrículas que tuvieron cada uno de estos estudiantes durante su primer año de estudios se puede evidenciar que 202 estuantes nunca realizaron ninguna matricula el cual representa al 3.14% de ingresantes y así mismo un grupo de estudiante en no tuvo permanecía académica durante el periodo de análisis.

Del total de alumnos que ingresaron en los semestres 2015-I y 2019-I y cuentan con al menos una matrícula en la Universidad Nacional Santiago Antúnez de Mayolo se puede verificar que el 15.96% de los ingresantes no tiene permanecía académica durante su primer año de estudios y se les considerar como desertores estudiantiles de acuerdo a los datos recopilados de la base de datos de la Oficina General de Estudios (OGE) y la Oficina General de Admisión (OGAD).

El porcentaje con mayor impacto fue en el semestre 2017-I siendo 18.12% de ingresantes que desertaron, considerando que ingresaron 1480, de forma que son cifras alarmantes que propician un análisis de los factores que han acondicionado el comportamiento ya mencionado. De los expuesto anteriormente, surge la pregunta a priori de: ¿Cómo se podría realizar la predicción de los estudiantes que presenten desertar en la Universidad Nacional Santiago Antúnez

de Mayolo?, de manera que, que nos podríamos adelantarnos a los posibles sucesos venideros, con la finalidad de ayudar en el futuro al establecimiento de decisiones acertadas por parte de los directivos para reducir la deserción. Por ende, en base la información obtenida y mencionada con anterioridad, se establecen factores que repercuten en la deserción que permitirán ser computadas a través de algoritmos de aprendizaje supervisado tales como como KNN, Random Forest y Gradient Boosting. De esta problemática encontrada y descrita en, se establecer como pregunta: ¿Podemos predecir la deserción de los estudiantes en la Universidad Nacional Santiago Antúnez de Mayolo?

Análisis descriptivo de los ingresantes durante los semestres 2015-I al 2019-I

Tabla 2:

Análisis descriptivo de los ingresantes durante los semestres 2015-I al 2019-I

Semestre	NM	Mat I	Mat II	Mat III	Alum. Mat.	No Des.	Des	Total	% Des
2015-1	87	1001	930	883	1005	865	140	1092	12.82%
2016-1	17	909	802	781	919	762	157	936	16.77%
2017-1	32	1468	1300	1236	1480	1206	274	1512	18.12%
2018-1	40	1319	1204	1142	1333	1116	217	1373	15.80%
2019-1	26	1489	1340	1283	1501	1261	240	1527	15.72%

Nota. Información proporcionada por la OGE y OGAD.

4.1.2. Preparación de los datos

Para obtener los datos fue necesario solicitarlos a las oficinas pertinentes las cuales son la Oficina General de Estudios (OGE) y la Oficina General de Admisión (OGAD), solicitando los datos de los ingresantes que se presentan a continuación:

Datos obtenidos

Tabla 3:

Datos de los ingresantes de los semestres 2015-I al 2019-I

Factores	Variable	Tipo	Categorías
Académicos	Facultad	Cualitativa	• Ciencias • Ciencias agrarias • Administración y turismo • […]
	Escuela profesional	Cualitativa	• Agronomía • Administración • Contabilidad • […]
	Modalidad de ingreso	Cualitativa	• Examen ordinario • 1roy 2do puesto • Graduado titulado • […]
	Puntaje de ingreso	Cuantitativa	• Puntaje obtenido para ingresar a la universidad.
	Ciclo actual	Cuantitativa	• Ciclo considerado en último semestre analizado por alumno.
	Promedio ponderado	Cuantitativa	• Promedio ponderado del último semestre analizado por alumno.
	Promedio del primer semestre	Cuantitativa	• Promedio del primer semestre analizado.
	Promedio del segundo semestre	Cuantitativa	• Promedio del segundo semestre analizado.
	Hábitos de estudio	Cualitativa	• Hábitos de estudio registrado por el estudiante.
	Número de créditos matriculados en el primer semestre	Cuantitativa	• Cantidad de créditos matriculado en el primer semestre analizado.
	Número de créditos matriculados en el segundo semestre	Cuantitativa	• Cantidad de créditos matriculado en el

			segundo semestre analizado.
	Número de créditos aprobados en el primer semestre	Cuantitativa	• Cantidad de créditos aprobados en el primer semestre analizado.
	Número de créditos aprobados en el segundo semestre	Cuantitativa	• Cantidad de créditos aprobados en el segundo semestre analizado.
	Deserción	Cualitativa	• SI • NO
Personales	Código de estudiante	Cualitativa	• Código de identificación del estudiante.
	Sexo	Cualitativa	• Masculino • Femenino
	Edad	Cuantitativa	• Edad de ingreso a la universidad.
	Estado civil	Cualitativa	• Soltero • Casado • […]
	Colegio de procedencia	Cualitativa	• Colegio de procedencia registrado.
	Año de egreso de la secundaria	Cuantitativa	• Año que culmino el colegio del alumno.
Socioeconómicos	Nivel de instrucción padre	Cualitativa	• Primaria • Secundaria • Superior • […]
	Nivel de instrucción madre	Cualitativa	• Primaria • Secundaria • Superior • […]
	Lugar de procedencia	Cualitativa	• Huaraz • Independencia • Yungay • […]
	Cantidad de hermanos	Cuantitativa	• Cantidad de hermanos declarado por el estudiante
	Situación económica	Cuantitativa	• Situación económica declarado por el estudiante.

Nota. Elaboración propia.

Para el procesamiento de los datos, modelamiento de datos, visualización de datos, etc. Es de importancia trabajar con diferentes herramientas que nos ayuden con este objetivo. Para esta investigación se escogió como principal entorno de trabajo el entorno de Anaconda que trabaja principalmente con el lenguaje de programación Python, donde se hizo uso de hizo uso de Jupyter Notebook. Así mismo se utilizó Transac-SQL para el análisis de los datos desde la base de datos SQL-Server y MySQL, por último, se utilizó Excel para el ordenamiento y concatenado del dataset.

Una vez definida las tecnologías que se utilizar primero se realiza la restauración de los scripts proporcionados por la oficina general de estudios del que se obtiene los siguientes esquemas:

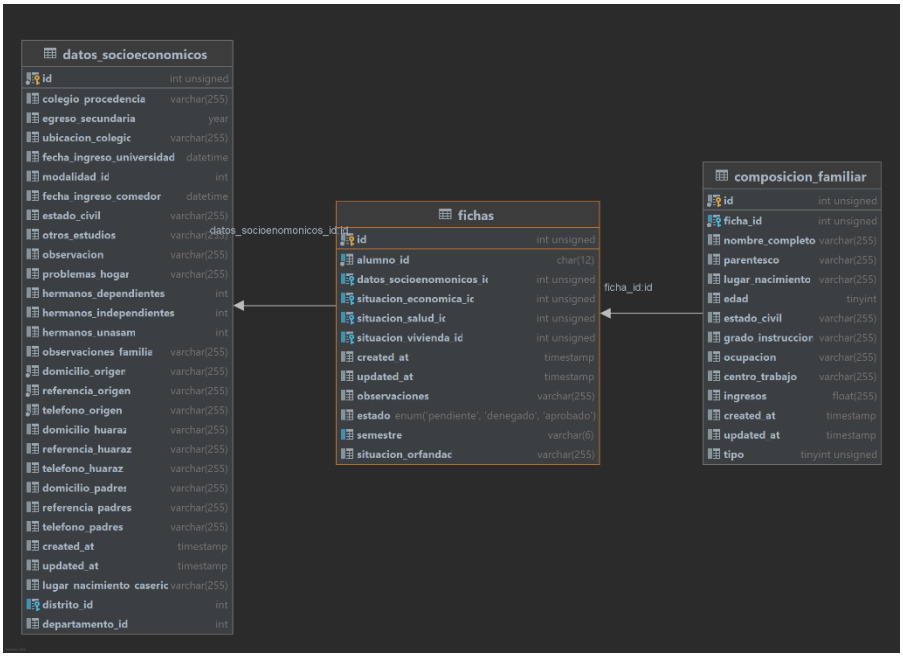
Figura 4

Esquema de la información académica de los estudiantes

Nota. Elaboración propia.

Figura 5

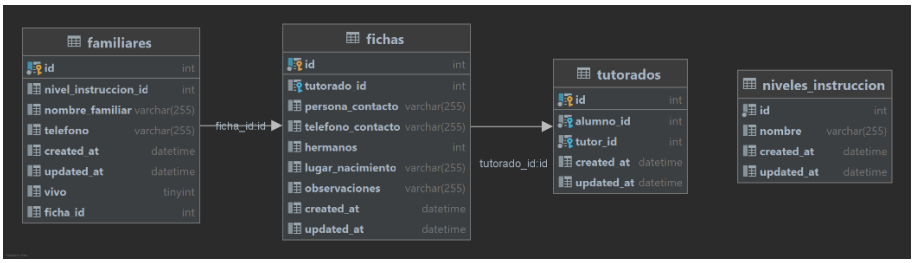
Esquema de información socioeconómica de los estudiantes



Nota. Elaboración propia.

Figura 6

Esquema de información de los familiares de los estudiantes



Nota. Elaboración propia.

A partir de los esquemas presentados se realizaron diversas consultar para extraer la información necesaria para el desarrollo de la investigación. De donde obtuvo los siguientes datasets.

- Datos socioeconómicos de alumnos que ingresaron el 2015-I
- Datos socioeconómicos de alumnos que ingresaron el 2016-I
- Datos socioeconómicos de alumnos que ingresaron el 2017-I
- Datos socioeconómicos de alumnos que ingresaron el 2018-I
- Datos socioeconómicos de alumnos que ingresaron el 2019-I
- Datos académicos de alumnos que ingresaron el 2015-I
- Datos académicos de alumnos que ingresaron el 2016-I
- Datos académicos de alumnos que ingresaron el 2017-I
- Datos académicos de alumnos que ingresaron el 2018-I
- Datos académicos de alumnos que ingresaron el 2019-I

Así mismo desde la oficina de admisión se obtuvo el listado de ingresantes desde el semestre 2015-I al semestre 2019-I. Por lo tanto, se necesitó el software Excel para unificar el dataset y contar con una única entrada de datos.

En la siguiente etapa es donde se aplica las técnicas de limpieza de datos, como por ejemplo tratar los registros vacíos y repetidos, buscar un orden en los registros, para poder explorar en ellos e identificar los posibles modelos que pueden aplicarse, para realizar la presente tarea se utilizó Jupyter Notebook y las librerías numpy y pandas.

Dataset de estudiantes del 2015-I al 2019-I

Figura 7

Dataset de estudiantes del 2015-I al 2019-I

	COD_ESTUDIANTE	FACULTAD	DESC_FACULTAD	ESCUELA	DESC_ESCUELA	MODALIDAD	DESC_MODALIDAD	PUNTAJE_INGRESO	SEXO	EDAD	...	INS
0	151.0103.172	2	CIENCIAS AGRARIAS	1	AGRONOMÍA	A	EXAMEN DE ADMISION ORDINARIO	0.0	0	25	...	
1	151.0103.173	2	CIENCIAS AGRARIAS	1	AGRONOMÍA	A	EXAMEN DE ADMISION ORDINARIO	0.0	1	16	...	
2	151.0103.174	2	CIENCIAS AGRARIAS	1	AGRONOMÍA	A	EXAMEN DE ADMISION ORDINARIO	0.0	1	20	...	
3	151.0103.175	2	CIENCIAS AGRARIAS	1	AGRONOMÍA	C	GRADUADO TITULADO	0.0	1	26	...	
4	151.0103.176	2	CIENCIAS AGRARIAS	1	AGRONOMÍA	A	EXAMEN DE ADMISION ORDINARIO	0.0	1	25	...	

Nota. Elaboración propia.

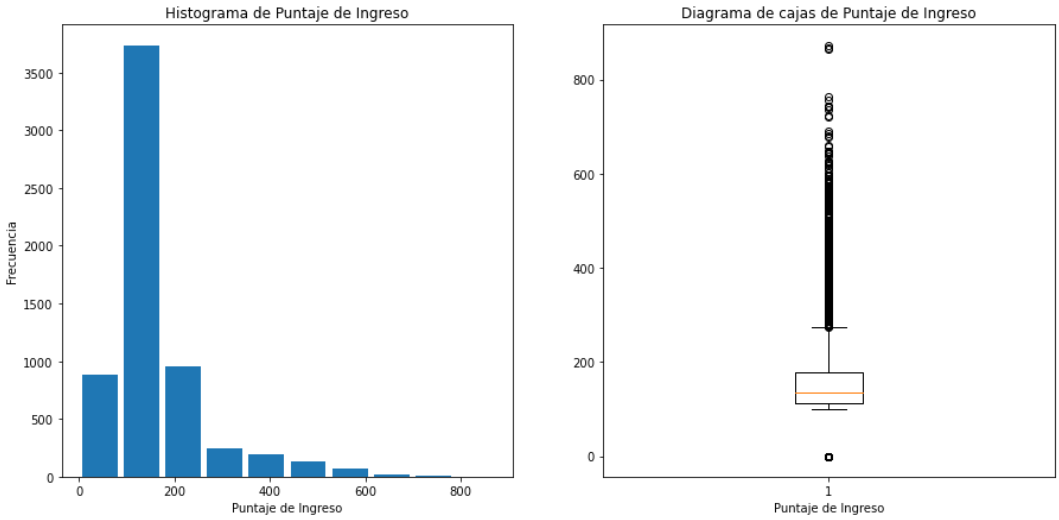
Para visualizar los datos en búsqueda de datos atípicos se realización de las variables numéricas y las variables categóricas.

Evaluación de la visualización de las variables numéricas:

- Puntaje de ingreso

Figura 8

Puntaje de ingreso

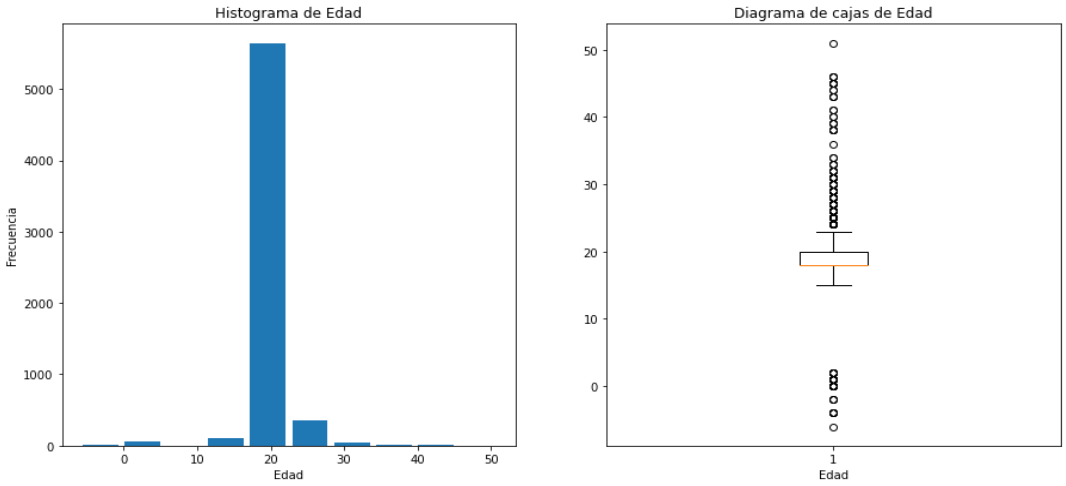


Nota. Elaboración propia.

- Edad

Figura 9

Puntaje de edad

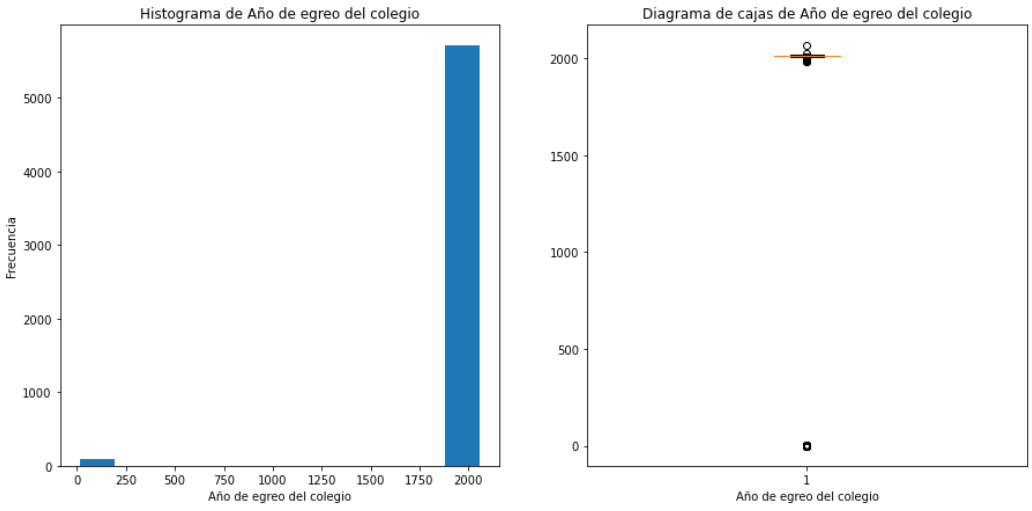


Nota. Elaboración propia.

- Año de egreso del colegio

Figura 10

Puntaje de año de egreso del colegio

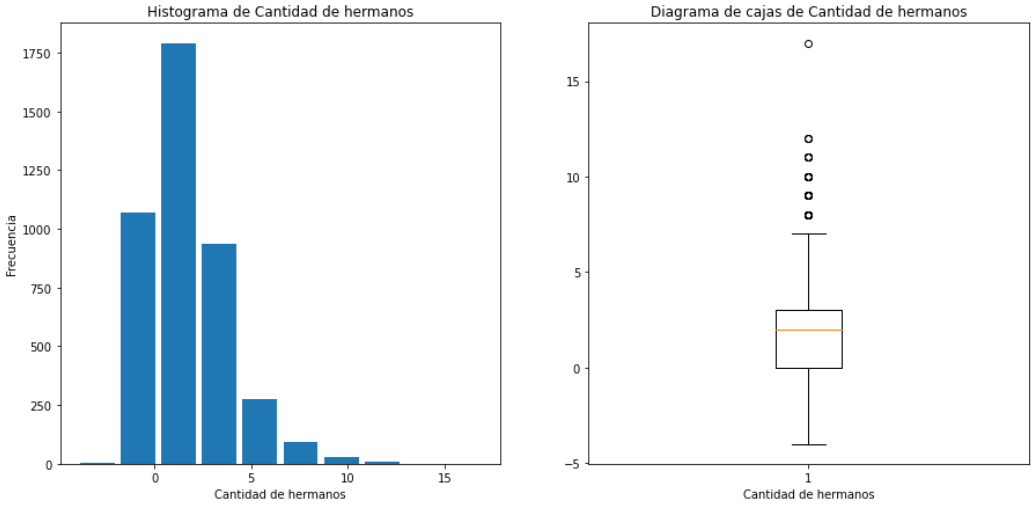


Nota. Elaboración propia.

- Cantidad de hermanos

Figura 11

Puntaje de cantidad de hermanos

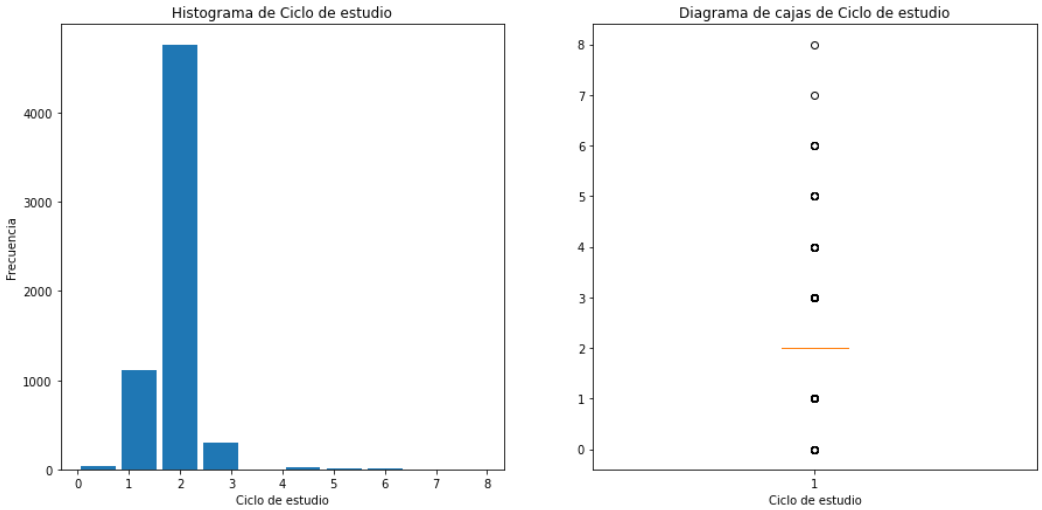


Nota. Elaboración propia.

- Ciclo de estudio

Figura 12

Puntaje de cliclo de estudio

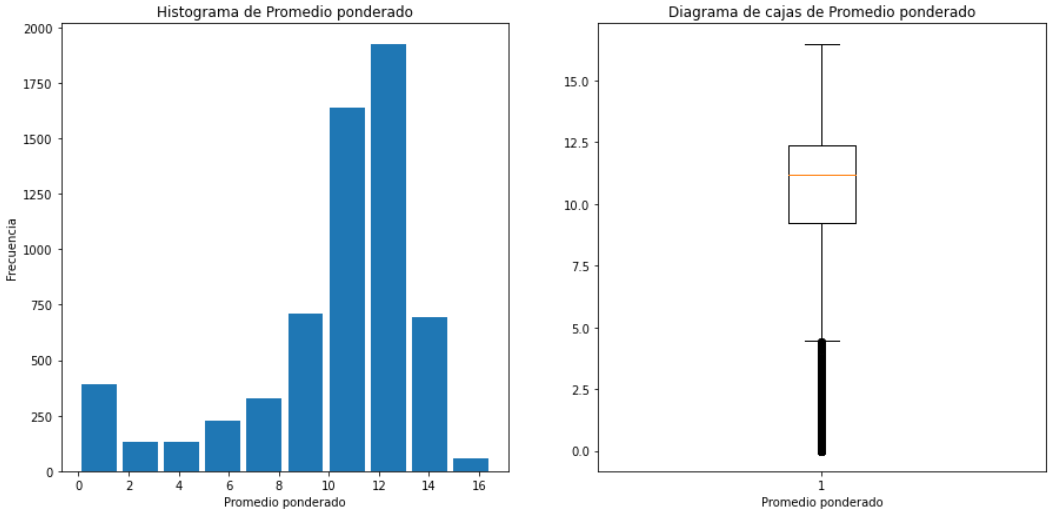


Nota. Elaboración propia.

- Promedio ponderado

Figura 13

Puntaje de promedio ponderado

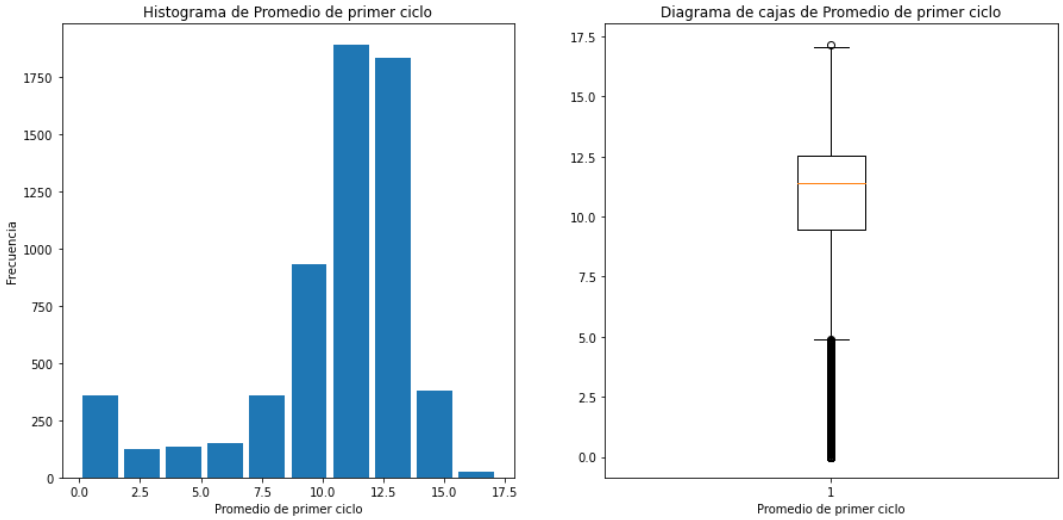


Nota. Elaboración propia.

- Promedio de primer semestre de estudio

Figura 14

Puntaje de primer semestre de estudio

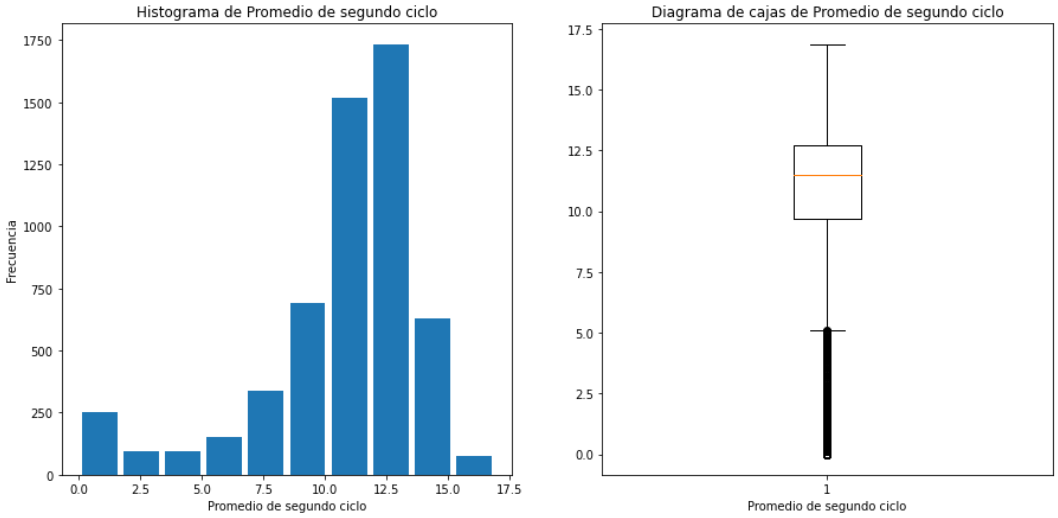


Nota. Elaboración propia.

- Promedio de segundo semestre de estudio

Figura 15

Puntaje de promedio de segundo semestre de estudio

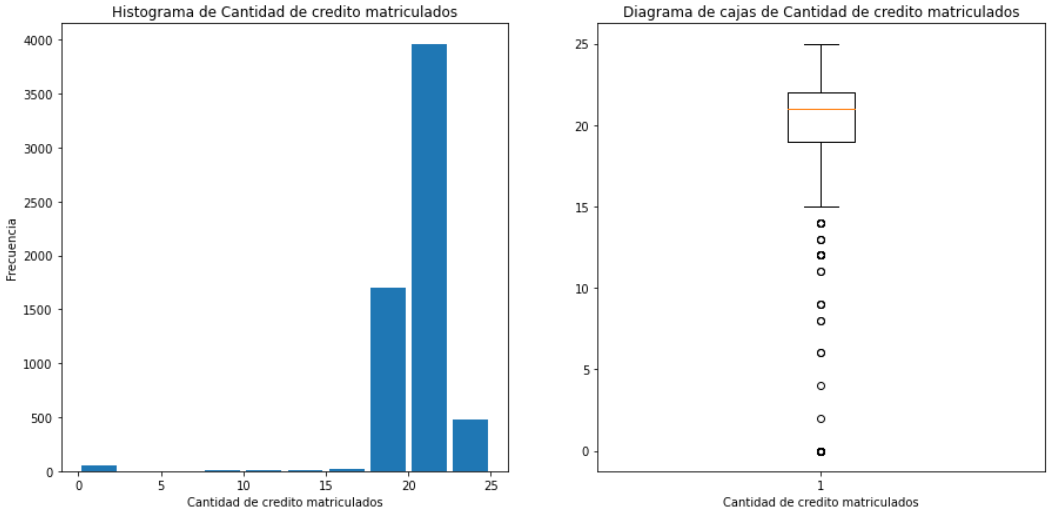


Nota. Elaboración

- Cantidad de créditos matriculados en el primer semestre

Figura 16

Cantidad de créditos matriculados en el primer semestre

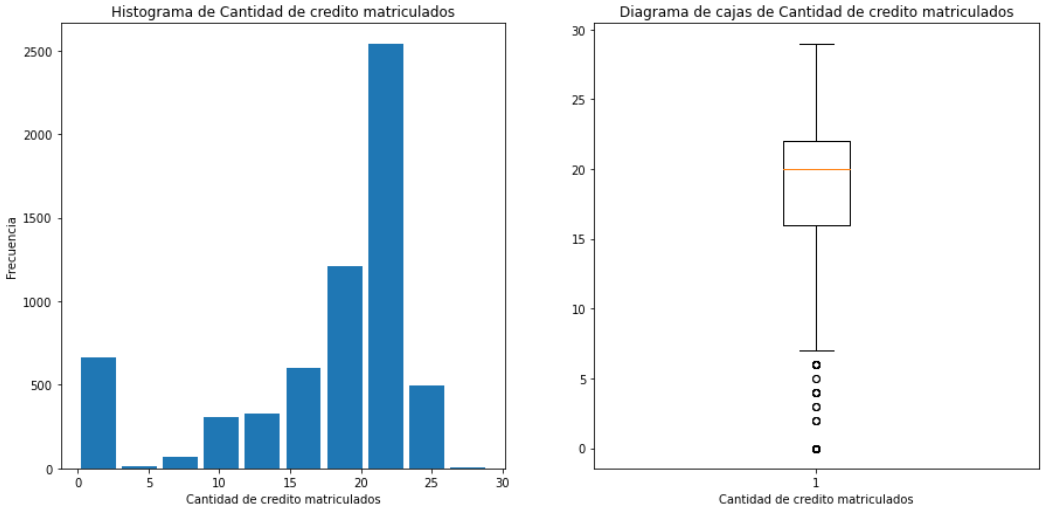


Nota. Elaboración propia.

- Cantidad de créditos matriculados en el segundo semestre

Figura 17

Cantidad de créditos matriculados en el segundo semestre

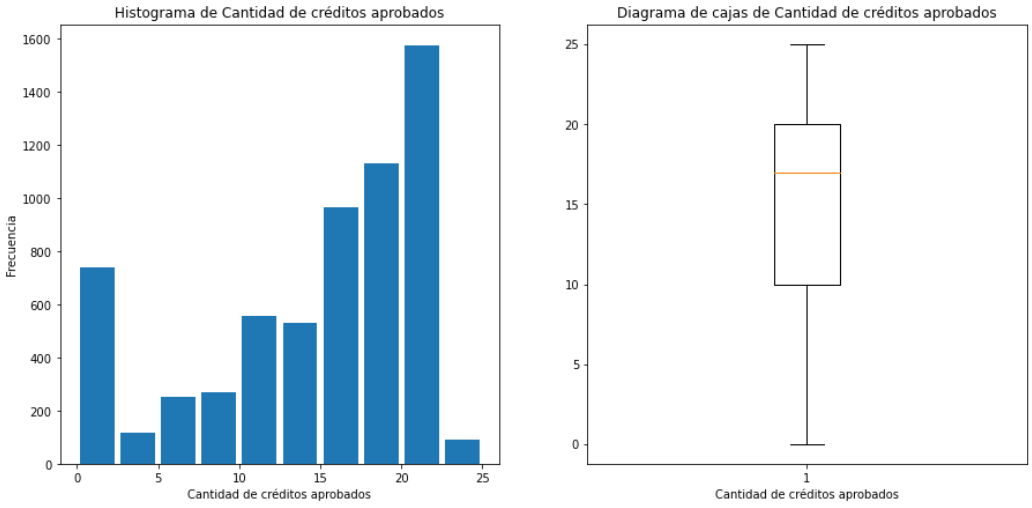


Nota. Elaboración propia.

- Cantidad de créditos aprobados en el primer semestre

Figura 18

Cantidad de créditos aprobados en el primer semestre

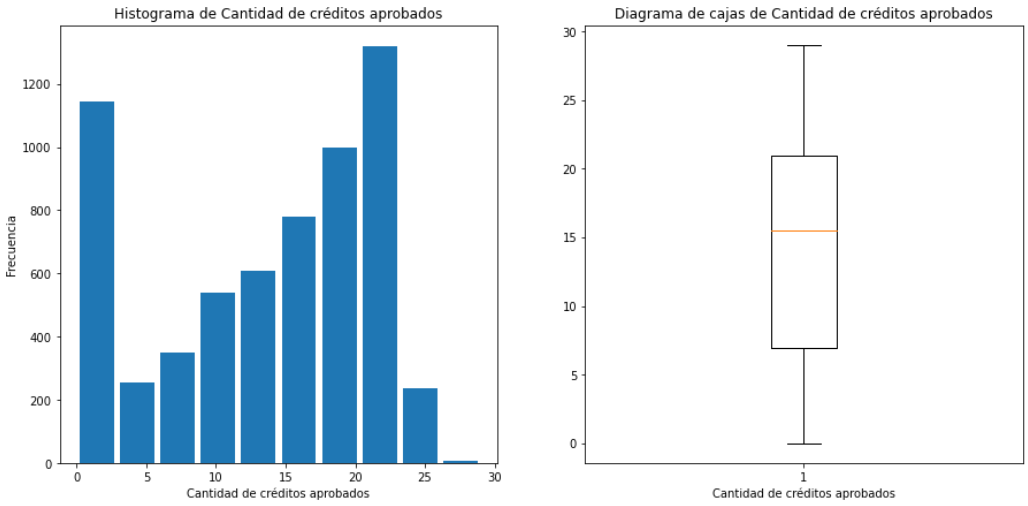


Nota. Elaboración propia.

- Cantidad de créditos aprobados en el segundo semestre

Figura 19

Cantidad de créditos aprobados en el segundo semestre



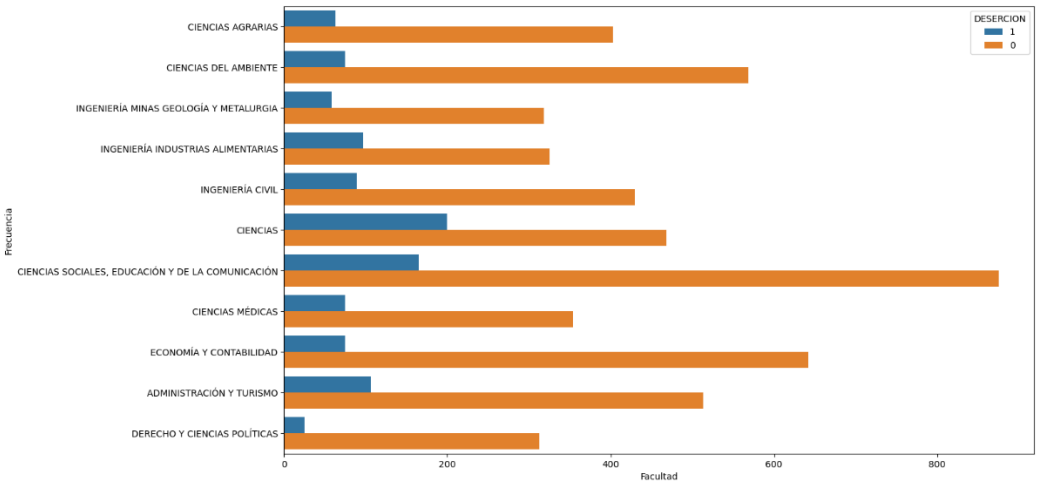
Nota. Elaboración propia.

Evaluación de la visualización de las variables categóricas:

- Facultad

Figura 20

Evaluación de la visualización de la variable categorica facultad

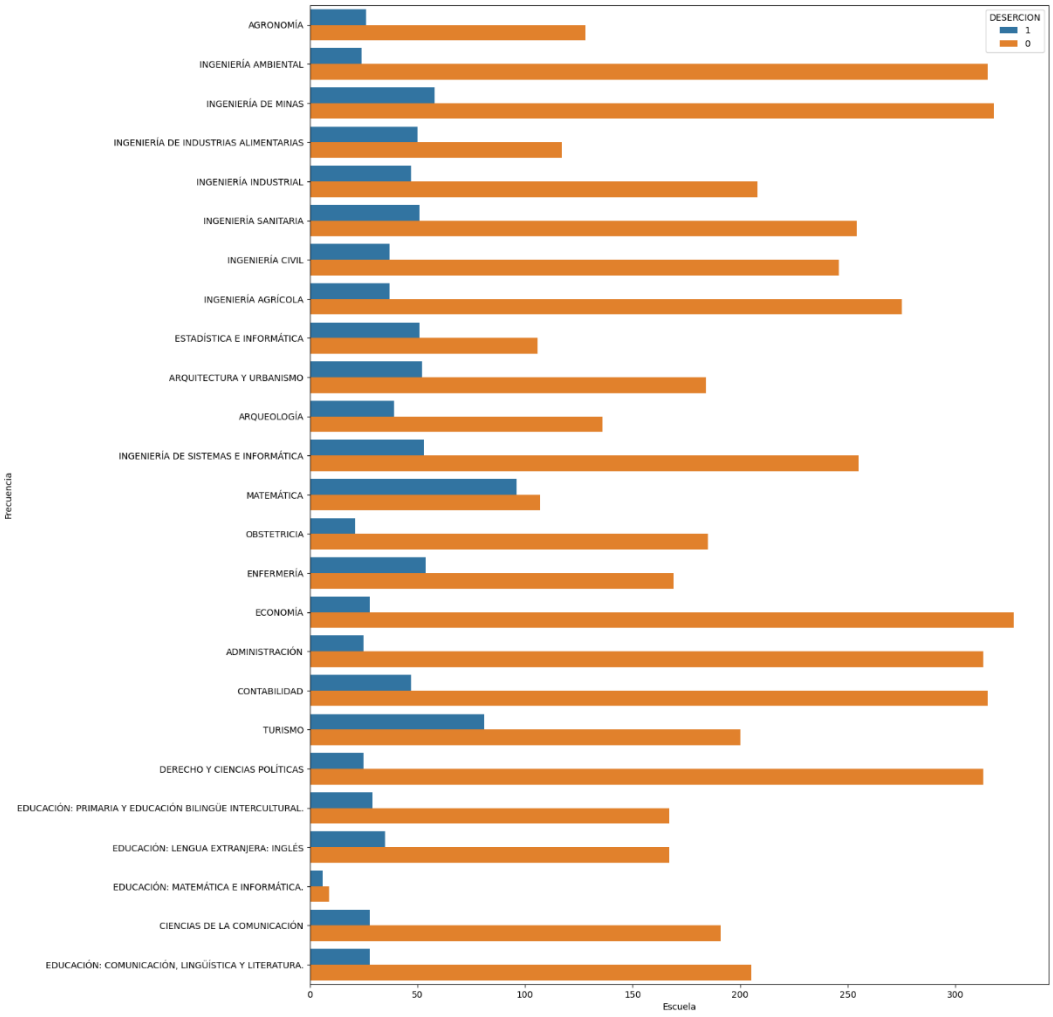


Nota. Elaboración propia.

- Escuela

Figura 21

Evaluación de la visualización de la variable categorica escuela

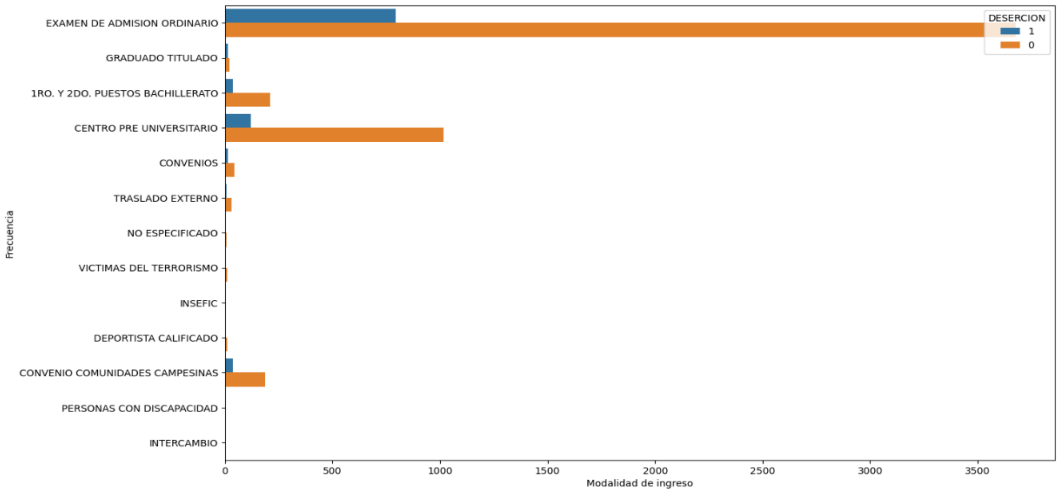


Nota. Elaboración propia.

- Modalidad de ingreso

Figura 22

Evaluación de la visualización de la variable categorica modalidad de ingreso

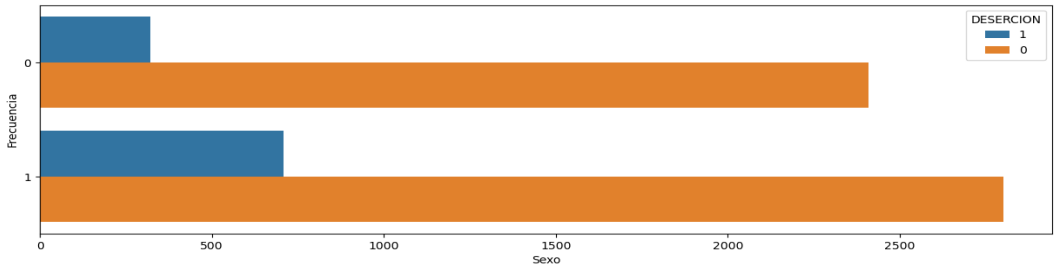


Nota. Elaboración propia.

- Sexo

Figura 23

Evaluación de la visualización de la variable categorica sexo

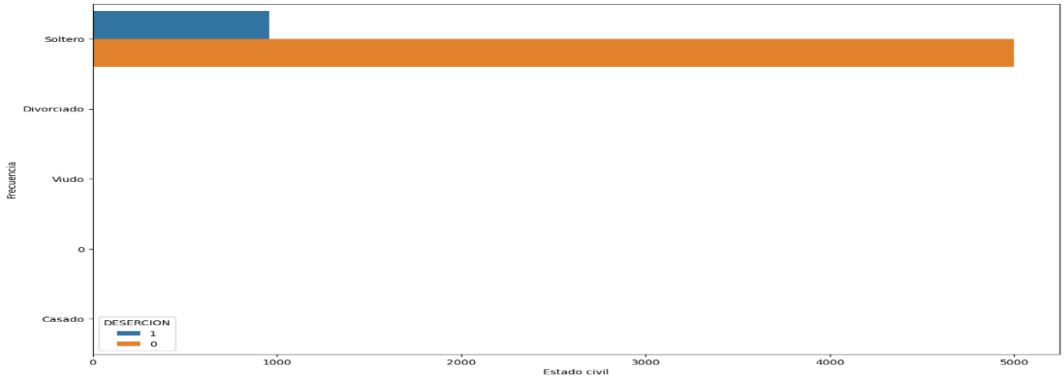


Nota. Elaboración propia.

- Estado civil

Figura 24

Evaluación de la visualización de la variable categorica estado civil

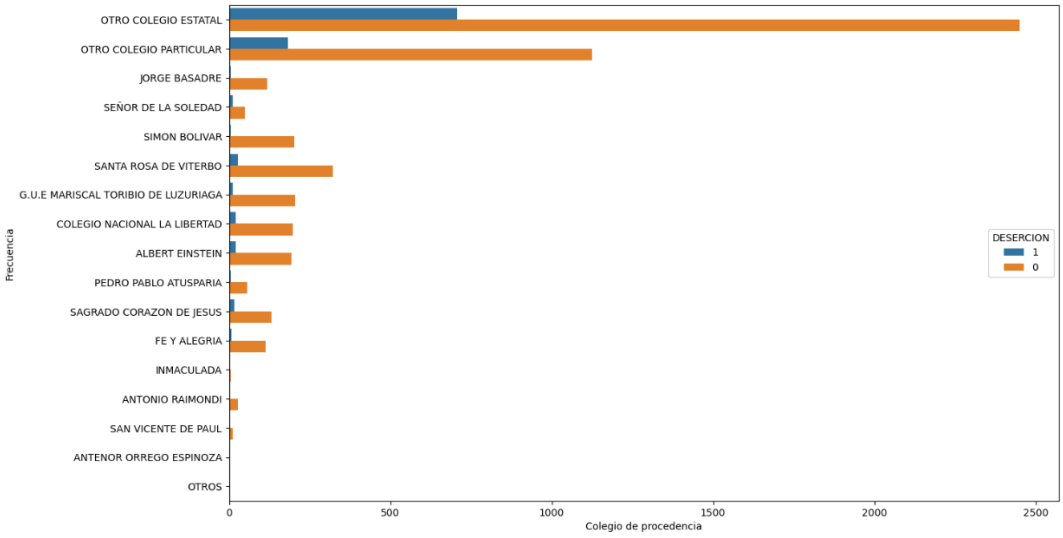


Nota. Elaboración propia.

- Colegio de procedencia

Figura 25

Evaluación de la visualización de la variable categorica colegio de procedencia

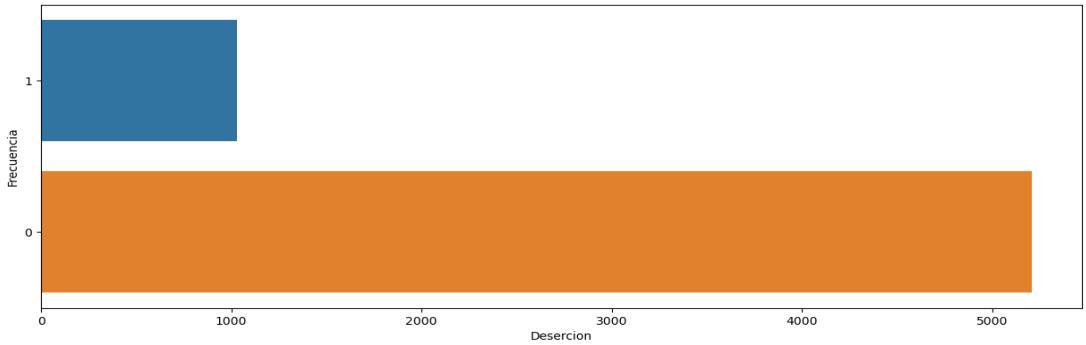


Nota. Elaboración propia.

- Deserción

Figura 26

Evaluación de la visualización de la variable categorica deserción.



Nota. Elaboración propia.

Así mismo se visualizaron la cantidad de nulos que hay en cada una de las variables:

Tabla 4:

Cantidad de nulos por variable

Variable	Datos nulos	No nulos
PUNTAJE_INGRESO	0	6238
EDAD	0	6238
ANIO_EGRE_COL	437	5801
CANT_HERMANOS	2034	4204
CICLO	0	6238
PROM_POND	0	6238
PROM_PRIM_CICLO	0	6238
PROM_SEG_CICLO	0	6238
CANT_CRED_MAT_PRIM_CICLO	0	6238
CANT_CRED_MAT_SEG_CICLO	0	6238
CANT_CRED_APOB_PRIM_CICLO	0	6238
CANT_CRED_APOB_SEG_CICLO	0	6238
COD_ESTUDIANTE	0	6238
FACULTAD	0	6238
DESC_FACULTAD	0	6238
ESCUELA	0	6238
DESC_ESCUELA	0	6238
MODALIDAD	0	6238
DESC_MODALIDAD	0	6238
SEXO	0	6238
ESTADO_CIVIL	262	5976
LUGAR_PROC	147	6091
COLEGIO_PROC	0	6238
DET_COLEGIO	4	6234

TIPO_COLEGIO	0	6238
INSTRUCCION_PADRE	2997	3241
INSTRUCCION_MADRE	2902	3336
DESERCION	0	6238

Nota. Elaboración propia.

Donde podemos apreciar claramente que claramente que la cantidad de hermanos, instrucción del padre e instrucción de la madre se cuenta con una gran concurrencia de datos nulos y al realizar el análisis correspondiente perdería 2369 representando un total del 37.97% de la data del dataset, por lo tanto, se optó por dejar estas variables fuera del estudio.

Al realizar el análisis respectivo de cada uno de los gráficos obtenidos de las variables numéricas, se observaron diferentes patrones que pueden ser outliers en las siguientes variables:

- Puntaje de ingreso

Se puede observar que hay puntajes de ingreso superiores a los 800 puntos y otros que son puntajes de ingreso igual a cero por lo tanto se pudo visualizar que los puntajes de ingreso que superan los 400 punto son aquellos estudiantes que cuentan con modalidad de ingreso Centro Pre Universitario donde se cuenta con un acumulado de varios exámenes y estos pueden puntajes llegar a ser muy superiores a las otras modalidades de ingreso.

En cuanto a los que cuentan con datos igual a cero se observa que hay algunas modalidades que no registran su puntaje de ingreso y otros datos no se llegaron a captar durante la recolección de información por lo tanto aquellos que datos que son cero y son de la modalidad de ingreso Examen Ordinario se marcaron como nulos y los demás si se dejaron como está.

- Edad

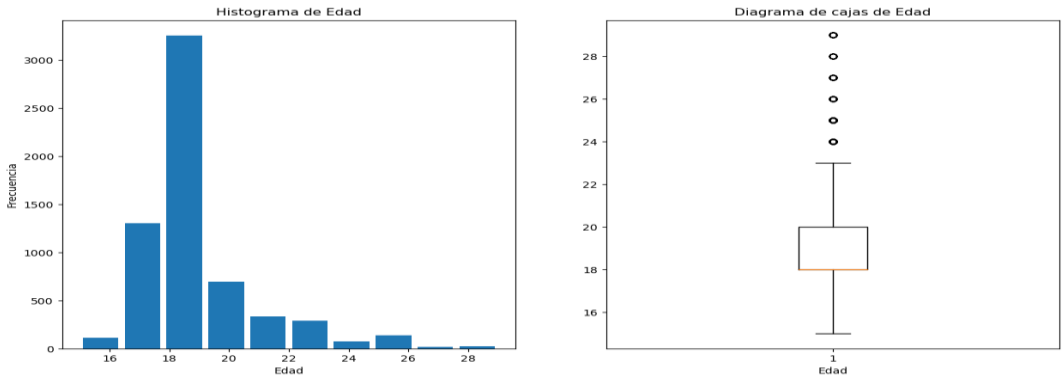
Se puede observar que la gran mayoría de estudiantes tiene 20 años sin embargo hay unos outliers que pueden estar afectando la calidad de la información por lo tanto se tomó como referencia los datos registrados por la INEI (2020) que mencionan que la población estudiantil universitaria está conformado por jóvenes de 15 a 29 años aproximadamente, por lo tanto todos los datos que sobrepasan esos

límites son remplazados por la media de los otros datos, quedando los gráficos de la siguiente manera:

Edad normalizada

Figura 27

Edad normalizada



Nota. Elaboración propia.

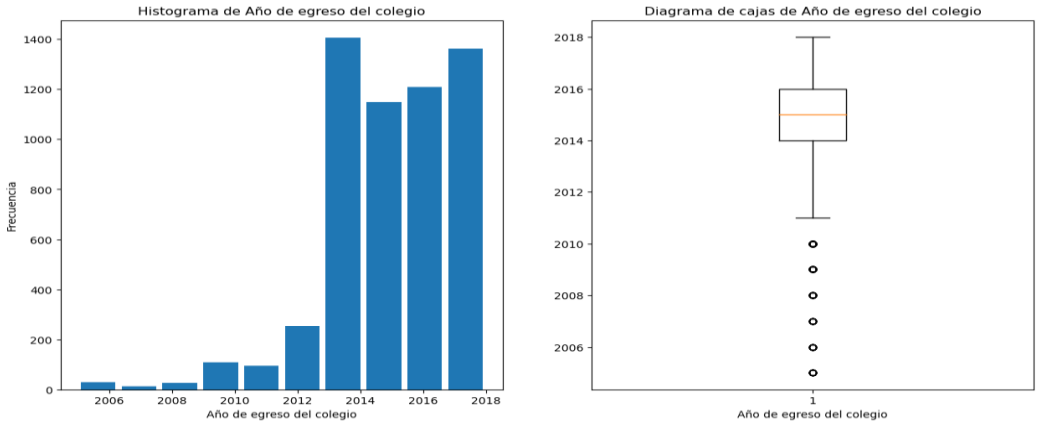
- Año de egreso del colegio

Se puede observar que en de los registros obtenidos hay un grupo de estudiantes que cuenta con año de ingreso igual a cero y otro grupo que sobrepasa el último año de estudio, por lo tanto, todos estos datos son eliminados por ser datos anómalos.

Año de egreso del colegio normalizado

Figura 28

Año de egreso del colegio normalizado



Nota. Elaboración propia.

- Ciclo de estudio

Se puede observar que la mayoría de estudiantes se encuentra en el segundo ciclo de estudio el cual se puede evidenciar dado que se está analizando a los estudiantes cuando terminan su primer año de estudio, sin embargo, se puede encontrar alumnos que cuentan con un ciclo superior al de su caso actual, estos se llevan a cabo porque su modalidad de ingreso es traslado externo o segunda carrera.

Al realizar el análisis respectivo de cada uno de los gráficos obtenidos de las variables categóricas, se observaron diferentes patrones que no serán de ayuda para la presente investigación, donde se puede evidenciar que:

- El estado civil

Se puede observar que el 95.49% de los datos son solteros, así mismo el 4.2% de los datos son nulos por lo tanto no existe variabilidad en la presente variable y se elimina para futuros análisis.

- Lugar de procedencia

En cuanto a este dato hay una variabilidad de 267 respuestas brindadas, por lo tanto, solo se analizará el top 5 de lugares más concurridos, donde se encuentra que solo el 57.55% de la data cumple con los requisitos, por lo tanto, esta variable queda fuera del estudio.

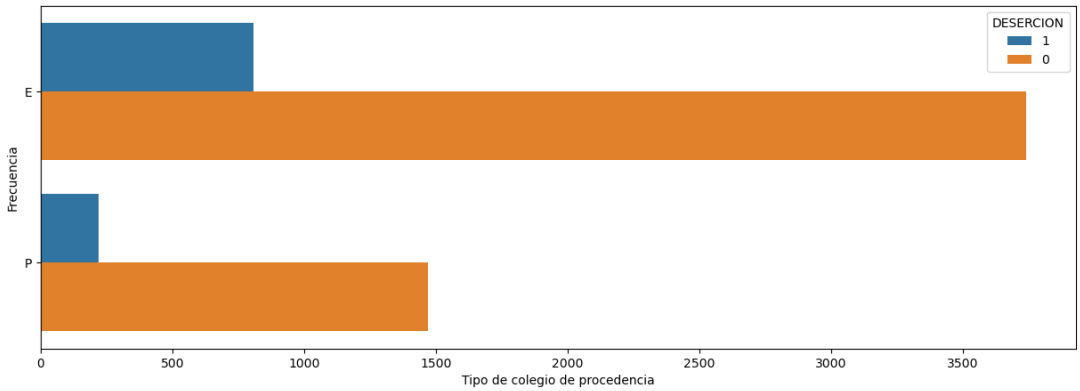
- Colegio de procedencia

Al observer la grafica se puede evidenciar que hay un total de 565 respuestas, por lo tanto se vuelve una variable que no tomara en cuenta para el presente estudio. Sin embargo, se tiene solo dos tipos de colegios los cuales son particulares o estatales y este dato si puede ingresar al estudio.

Tipo de colegio de procedencia

Figura 29

Tipo de colegio de procedencia



Nota. Elaboración propia.

Por lo tanto después de realizar todo el análisis respectivo y haber realizado la limpieza de los datos los datos que serán tomados en cuenta para la presente investigación van a ser: facultad, escuela profesional, modalidad de ingreso, puntaje de ingreso, sexo, edad, tipo de colegio de ingreso, año de egreso de colegio, ciclo, promedio ponderado, promedio de primer semestre de estudio, promedio de segundo semestre de estudio, cantidad de créditos matriculados en el primer semestre , cantidad de créditos matriculados en el segundo semestre , cantidad de créditos aprobados en el primer semestre, cantidad de créditos aprobados en el segundo semestre, deserción.

Para finalizar se eliminan todos los registros que cuentan con al menos un nulos dentro de sus datos, de esta manera nos quedamos con un data set conformado por 5657 filas y 17 columnas.

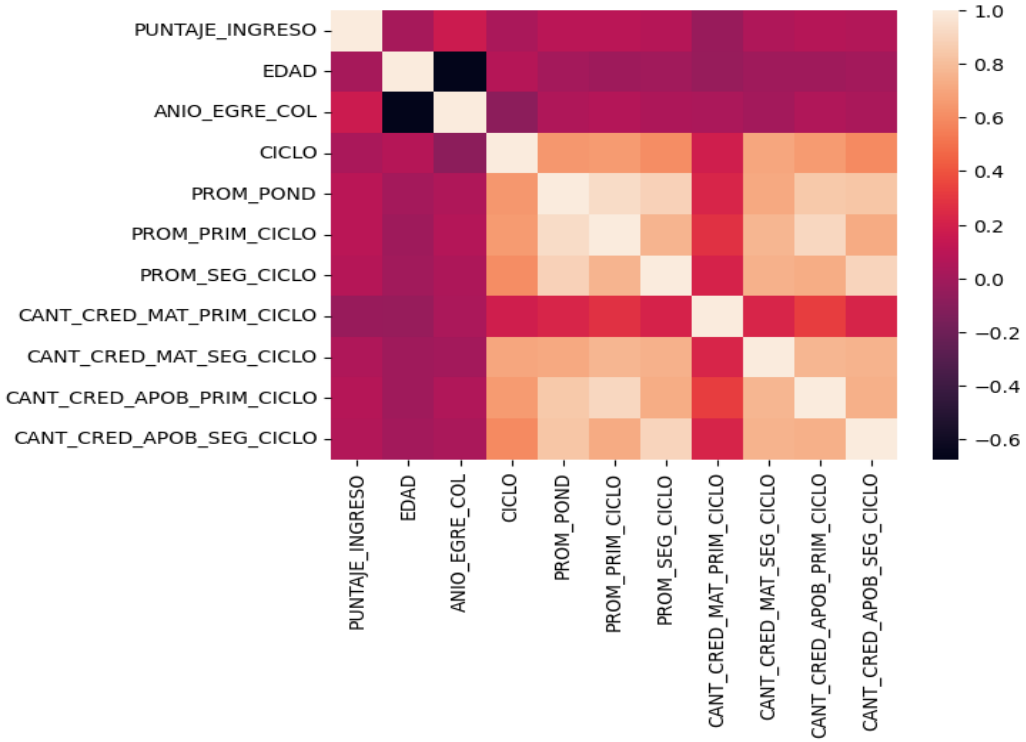
4.1.3. Representación de los datos

Con el dataset ya limpio se puede proceder a realizar el análisis exploratorio de los datos numéricos:

Correlación entre las variables numéricas

Figura 30

Correlación entre las variables numéricas



Nota. Elaboración propia.

En el grafico anterior podemos visualizar que no existe mucha correlación entre las variables numéricas, por lo tanto, cada una de las variables serán necesarios para realizar la predicción de la deserción de estudiantes en la Universidad Nacional Santiago Antúnez de Mayolo.

Creación de features

Para las variables categóricas se crearán features con los que se pueda trabajar posteriormente, para el cual se va generar variables ficticias que puedan contemplar las 6 variables categóricas con el que cuenta el dataset.

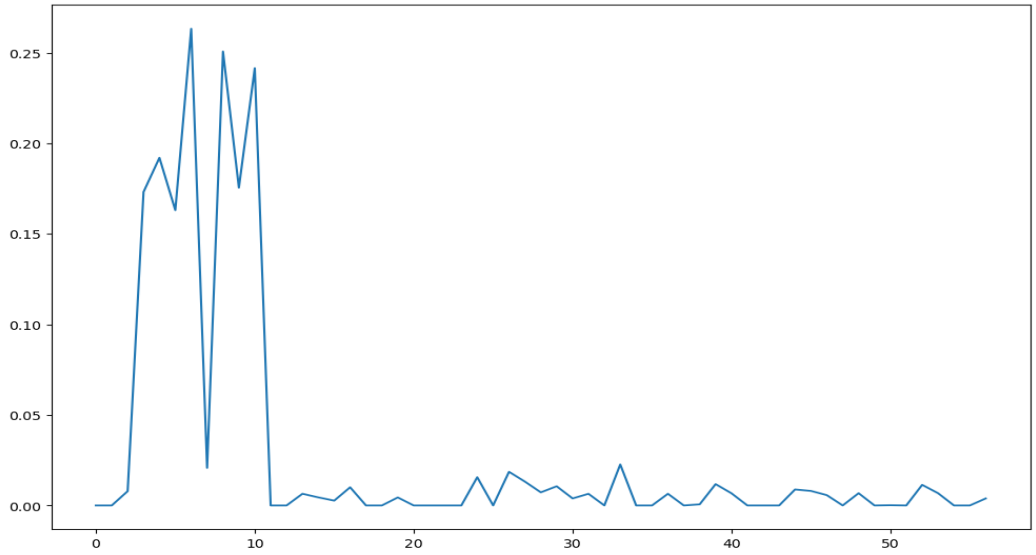
Después de la transformación de las variables categóricas se obtiene un nuevo dataset de 58 columnas listo para ser utilizado dentro de las predicciones.

Utilizando herramientas de scikit-learn podemos visualizar si alguna de las features es más influyente que las demás a la hora de realizar la predicción, en donde obtenemos el siguiente grafico de todas las features:

Importancia de features

Figura 31

Importancia de features



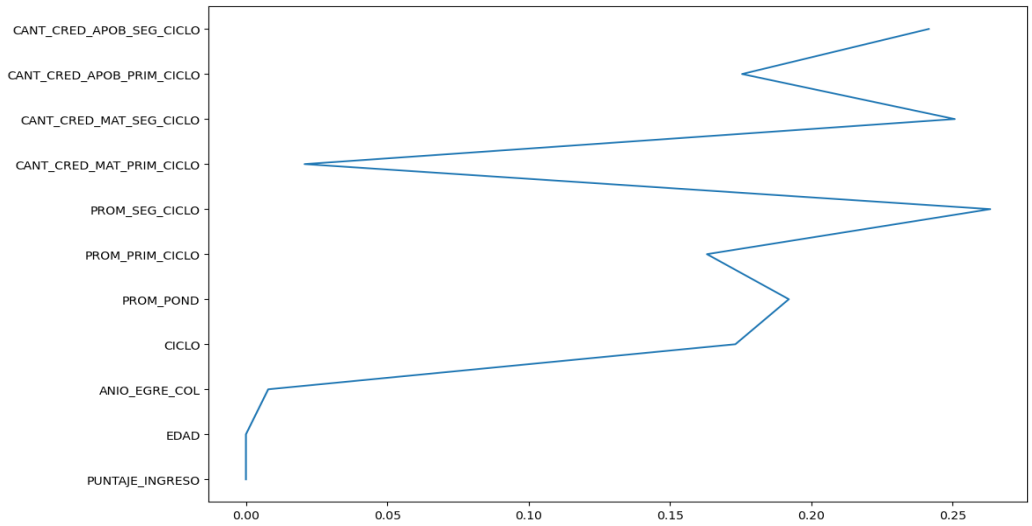
Nota. Elaboración propia.

En la presente imagen se puede visualizar que las features más importantes son las 11 primeras por tal motivo se muestra una ampliación de estas para revisar cuales son:

Features más importantes

Figura 32

Features más importantes



Nota. Elaboración propia.

Del grafico anterior podemos deducir que las features que van a influir más dentro de nuestra predicción serán: ciclo, promedio ponderado, promedio del primer ciclo, promedio del segundo ciclo, cantidad de créditos matriculados segundo ciclo, cantidad de créditos aprobados primer ciclo y cantidad de créditos aprobados segundo ciclo.

Relación entre las variables

Se realizo un análisis de relación entre las variables que quedaron con la variable de deserción, para valores de significancia menores a 0.05, donde se muestran las 16 variables que se quedaron después del análisis desarrollado anteriormente.

Tabla 5:

Análisis bivariado contra la variable deserción

Variable	Valor significancia	Relación con deserción	Valores de la frecuencia (%)
Facultad	0.00	Si	0.00
Escuela	0.00	Si	4.00
Modalidad	0.00	Si	27.30
Puntaje de ingreso	0.992	No	85.80
Sexo	0.00	Si	0.00
Edad	0.00	Si	14.30
Tipo de colegio	0.00	Si	0.00
Año de egreso del colegio	0.00	Si	14.30
Ciclo	0.00	Si	44.40
Promedio ponderado	0.00	Si	6.70
Promedio primer ciclo	0.00	Si	8.30
Promedio segundo ciclo	0.00	Si	6.70
Cantidad de créditos matriculados primer ciclo	0.00	Si	60.00
Cantidad de créditos matriculados segundo ciclo	0.00	Si	32.10
Cantidad de créditos aprobados de primer ciclo	0.00	Si	4.20
Cantidad de créditos aprobados de segundo ciclo	0.00	Si	12.50

Nota. Elaboración propia.

Cada una de las variables explicativas que inciden en la deserción de estudiantes son presentadas en presente análisis. Se muestra la relación de las variables con la deserción de estuantes y su valor de significancia mediante el estadístico chi cuadrado.

4.1.4. Modelamientos y aprendizaje

Para realizar el modelamiento y aprendizaje de los datos se procede a dividir el dataset en datos de entrenamiento y datos de prueba los cuales serás escogidos de manera aleatoria, en este caso se utilizará el estándar establecido donde el 75% de los datos serán de entrenamiento y el 25% de prueba.

En la presente investigación se seleccionó tres modelos para realizar las pruebas necesarias los cuales son:

- Algoritmo de vecinos más cercanos (KNN)
- Random Forest
- Gradient Boosting

Algoritmo de k vecinos más cercanos

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual. (IBM, 2022)

Para el entrenamiento del presente algoritmo se utilizó los valores por defecto de los hiper parámetros para que el mismo algoritmo decida cuantas comunidades es la mejor durante su entrenamiento.

Después del entrenamiento del modelo se pudo obtener el score de 0.9455, así mismo las siguientes métricas:

Métricas de medición de KNN

Figura 33

Metricas de medición KNN

	precision	recall	f1-score	support
0	0.95	0.99	0.97	1183
1	0.92	0.73	0.81	232
accuracy			0.95	1415
macro avg	0.94	0.86	0.89	1415
weighted avg	0.94	0.95	0.94	1415

Nota. Elaboración propia.

Métricas de medición de KNN

Tabla 6:

Métricas de medición KNN

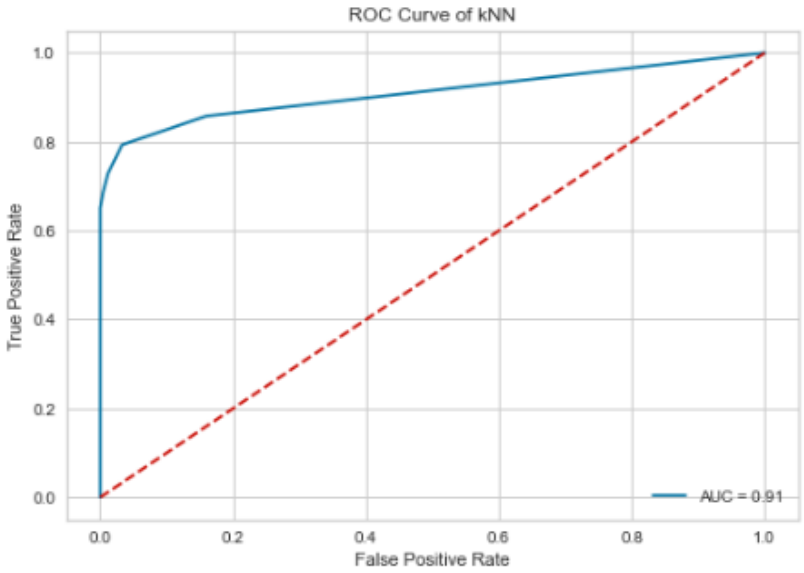
Precisión	Recall	F1 Score	Accuracy
0.94	0.86	0.89	0.95

Nota. Elaboración propia.

Podemos observar que la precisión es del 94% pero la sensibilidad es de 86% para este modelo, pero estos indicadores pertenecen a los datos de validación. Así mismo podemos verificar que el score F1 es del 89% y la accuracy total es del 95%. Teniendo unos muy buenos resultados en las métricas de medición del modelo.

Figura 34

Curva ROC para los datos de prueba del modelo KNN



Nota. Elaboración propia.

En la figura anterior se puede ver que el AUC para la deserción de estudiantes es del 91%, afirmando que el área debajo de la curva ROC es buena y el algoritmo de KNN tiene bueno resultados para predecir la deserción de estudiantes.

Random Forest

Random forest o bosque aleatorio es una técnica de enjambre. La base de estos métodos es realizar una combinación de diferentes clasificadores utilizando alguna técnica de agregación. (Igual & Seguí, 2017).

Para el entrenamiento del presente algoritmo se utilizó los valores por defecto de los hiper parámetros para que el mismo algoritmo decida cuantos selectores tendrá en su entrenamiento.

Después del entrenamiento del modelo se pudo obtener el score de 0.9540, así mismo las siguientes métricas:

Métricas de medición de Random Forest

Figura 35

Metricas de medición de Random Forest

	precision	recall	f1-score	support
0	0.96	0.99	0.97	1183
1	0.95	0.76	0.84	232
accuracy			0.95	1415
macro avg	0.95	0.88	0.91	1415
weighted avg	0.95	0.95	0.95	1415

Nota. Elaboración propia.

Métricas de medición de Random Forest

Tabla 7:

Métricas de medición de Random Forest

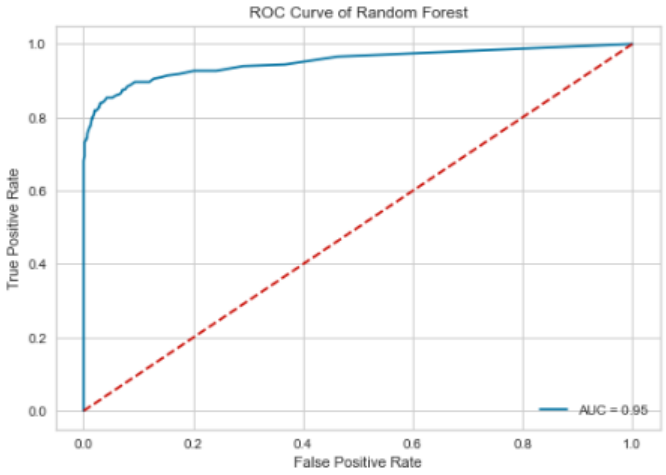
Precisión	Recall	F1 Score	Accuracy
0.95	0.88	0.91	0.95

Nota. Elaboración propia.

Podemos observar que la precisión es del 95% pero la sensibilidad es de 88% para este modelo, pero estos indicadores pertenecen a los datos de validación. Así mismo podemos verificar que el score F1 es del 91% y la accuracy total es del 95%. Teniendo unos muy buenos resultados en las métricas de medición del modelo.

Figura 36

Curva ROC para los datos de prueba de Random Forest



Nota. Elaboración propia.

En la figura anterior se puede ver que el AUC para la deserción de estudiantes es del 95%, afirmando que el área debajo de la curva ROC es buena y el algoritmo de Random Forest tiene bueno resultados para predecir la deserción de estudiantes.

Gradient Boosting

Gradient Boosting está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. (Amat, 2020)

Para el entrenamiento del presente algoritmo se utilizó los valores por defecto de los hiper parámetros para que el mismo algoritmo decida cuantos selectores tendrá en su entrenamiento.

Después del entrenamiento del modelo se pudo obtener el score de 0.9561, así mismo las siguientes métricas:

Métricas de medición de Gradient Boosting

Figura 37

Metricas de Gradiente Boosting

	precision	recall	f1-score	support
0	0.96	0.99	0.97	1183
1	0.96	0.77	0.85	232
accuracy			0.96	1415
macro avg	0.96	0.88	0.91	1415
weighted avg	0.96	0.96	0.95	1415

Nota. Elaboración propia.

Métricas de medición de Gradient Boosting

Tabla 8:

Métricas de medición de Gradient Boosting

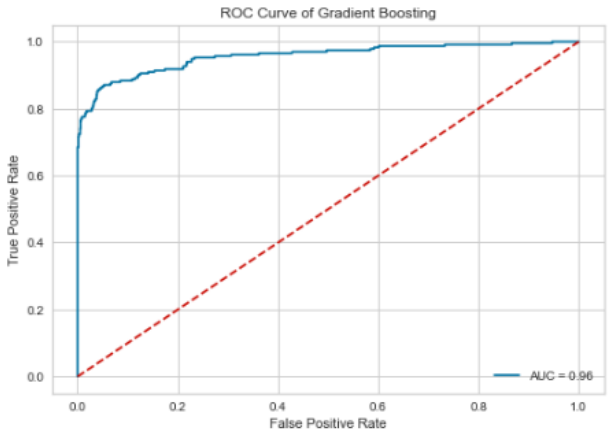
Precisión	Recall	F1 Score	Accuracy
0.96	0.88	0.91	0.96

Nota. Elaboración propia.

Podemos observar que la precisión es del 96% pero la sensibilidad es de 88% para este modelo, pero estos indicadores pertenecen a los datos de validación. Así mismo podemos verificar que el score F1 es del 91% y la accuracy total es del 96%. Teniendo unos muy buenos resultados en las métricas de medición del modelo.

Figura 38

Curva ROC para los datos de prueba de Gradient Boosting



Nota. Elaboración propia.

En la figura anterior se puede ver que el AUC para la deserción de estudiantes es del 96%, afirmando que el área debajo de la curva ROC es buena y el algoritmo de Gradient Boosting tiene bueno resultados para predecir la deserción de estudiantes.

4.1.5. Evaluación de modelos

Para la evaluación de los diversos modelos presentados se realiza la validación cruzada de todos datos recopilados, usando el coeficiente de determinación para el cálculo de las métricas, durante este proceso se evalúa cada uno de los modelos con 10 particiones de entrenamiento.

Algoritmo de k vecinos más cercanos

Figura 39

Validación del algoritmo KNN

```
ensemble = KNeighborsClassifier()

results_knn = cross_validate(ensemble, X, y, cv=10, scoring='r2', return_train_score=True)
results_knn
```

```
{'fit_time': array([0.00399971, 0.00400043, 0.00400233, 0.00500107, 0.00599933,
        0.00500321, 0.00499654, 0.00399899, 0.00399995, 0.00399852]),
 'score_time': array([0.2344892 , 0.13399792, 0.13899755, 0.13800097, 0.13719606,
        0.13399577, 0.13699913, 0.13499856, 0.13399935, 0.13399935]),
 'test_score': array([0.57539385, 0.66831951, 0.69383339, 0.59177786, 0.52799315,
        0.46420844, 0.52799315, 0.65247403, 0.61386003, 0.62673137]),
 'train_score': array([0.68865985, 0.67555647, 0.68409446, 0.68409446, 0.68409446,
        0.68551745, 0.70117043, 0.68582881, 0.68725039, 0.68725039])}
```

```
test_scores = results_knn['test_score']
train_scores = results_knn['train_score']
print(np.mean(train_scores))
print(np.mean(test_scores))
```

```
0.6863517162371963
0.5942584771688775
```

Nota. Elaboración propia.

Se observa que la puntuación del R-cuadrado con validación cruzada del algoritmo KNN es de 68.63%, pero así mismo se observa que en comparación del R-cuadrado de los datos de test es del 59.42% viéndose que no hay una gran dispersión dentro de los resultados, por lo cual es algoritmo va a funcionar como se espera.

Random Forest

Figura 40

Validación del Random Forest

```
ensemble_rnd = RandomForestClassifier()

results_rnd = cross_validate(ensemble_rnd, X, y, cv=10, scoring='r2', return_train_score=True)
results_rnd
```

```
{'fit_time': array([0.34627628, 0.36119342, 0.36135292, 0.36245394, 0.3689642 ,
        0.37251544, 0.35322547, 0.3565464 , 0.35905886, 0.3550005 ]),
 'score_time': array([0.01101351, 0.01104617, 0.01051211, 0.01051116, 0.01151347,
        0.01143789, 0.01051211, 0.0105114 , 0.0112052 , 0.01103687]),
 'test_score': array([0.60112755, 0.71934728, 0.68107645, 0.61729174, 0.65556257,
        0.4514515 , 0.56626397, 0.70395936, 0.66534536, 0.70395936]),
 'train_score': array([1.      , 1.      , 1.      , 0.998577, 1.      , 1.      ,
        1.      , 1.      , 1.      , 1.      ])}
```

```
test_scores = results_rnd['test_score']
train_scores = results_rnd['train_score']
print(np.mean(train_scores))
print(np.mean(test_scores))
```

```
0.9998577002049945
0.6365385142182671
```

Nota. Elaboración propia.

Se observa que la puntuación del R-cuadrado con validación cruzada Random Forest es de 99.98%, pero así mismo se observa que en comparación del R-cuadrado de los datos de test es del 63.65% viéndose que hay una gran dispersión dentro de los resultados, por lo cual es algoritmo no va a ser tan preciso como se espera.

Gradient Boosting

Figura 41

Validación del Gradient Boosting

```
ensemble_gb_2 = GradientBoostingClassifier(learning_rate=0.1,
                                           max_depth=1,
                                           n_estimators=500)

results_gb_2 = cross_validate(ensemble_gb_2, X, y, cv=10, scoring='r2', return_train_score=True)
results_gb_2
```

```
{'fit_time': array([1.29478645, 1.28883338, 1.27163672, 1.26226807, 1.2711699 ,
        1.27746534, 1.26944709, 1.26626301, 1.26919818, 1.28772998]),
 'score_time': array([0.00150943, 0.00100756, 0.0019989 , 0.00150466, 0.0019989 ,
        0.00200438, 0.00100017, 0.00150466, 0.00150847, 0.00100017]),
 'test_score': array([0.53679329, 0.69383339, 0.73210422, 0.56626397, 0.68107645,
        0.46420844, 0.55350703, 0.67821669, 0.65247403, 0.71683069]),
 'train_score': array([0.67586505, 0.65990349, 0.66986448, 0.68409446, 0.66274949,
        0.68124846, 0.67413347, 0.65597543, 0.67729927, 0.66450497])}
```

```
test_scores = results_gb_2['test_score']
train_scores = results_gb_2['train_score']
print(np.mean(train_scores))
print(np.mean(test_scores))
```

```
0.6705638559583628
0.6275308212318463
```

Nota. Elaboración propia.

Se observa que la puntuación del R-cuadrado con validación cruzada de Gradient Boosting es de 67.05%, pero así mismo se observa que en comparación del R-cuadrado de los datos de test es del 62.75% viéndose que hay una gran dispersión dentro de los resultados, por lo cual es algoritmo va a ser tan preciso como se espera.

Después de haber evaluado cada uno de los modelos presentados se puede deducir que el mejor algoritmo para este problema es Gradient Boosting, demostrando que es uno de los algoritmos más fuertes dentro del machine learning.

4.2. Presentación resultado y prueba de hipótesis

Presentación de resultados

Luego de la aplicación y evaluación de cada uno de los modelos los mejores resultados obtenidos fueron de Gradient Boosting que se obtuvo 96% de precisión, 88% de sensibilidad, 91% en el score F1, 96% de accuracy, 67.05% en el score R-cuadrado de la data de entrenamiento y 63.75% en el score R2-cuadrado de la data de test.

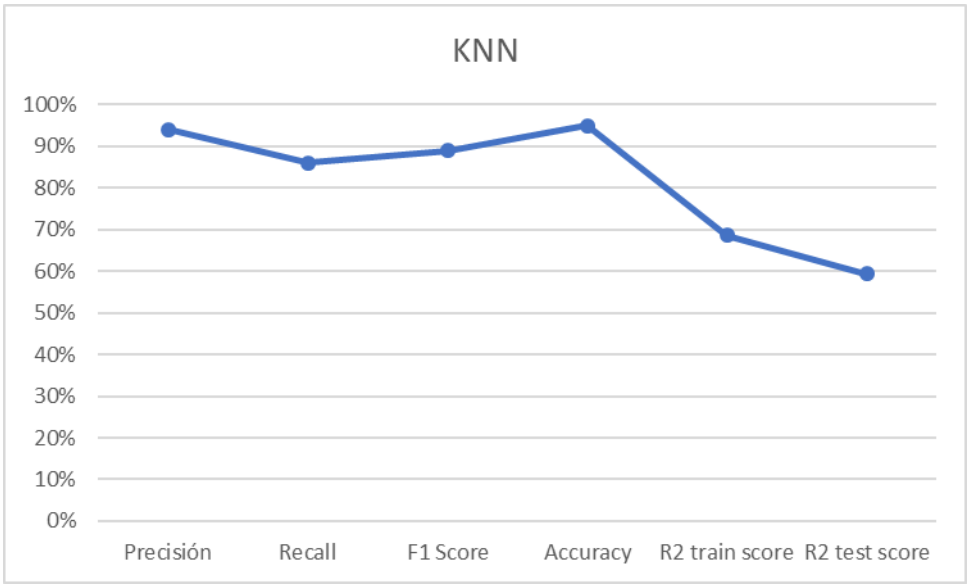
A se detalla los resultados obtenidos de cada uno de los modelos:

- **Algoritmo de k vecinos más cercanos:** Al aplicar se obtuvo una accuracy de 95% y R-cuadrado de entrenamiento del 68.63%

Resultados de KNN

Figura 42

Resultados de KNN



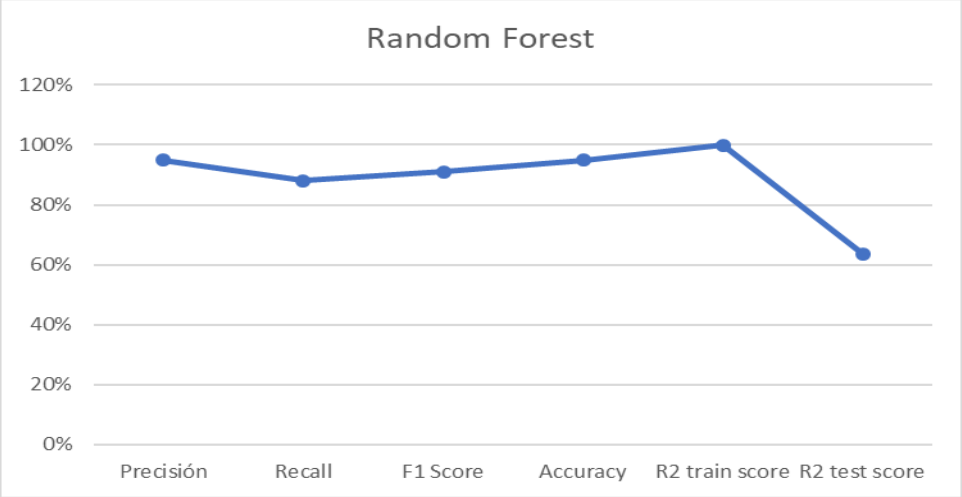
Nota. Elaboración propia.

- **Random Forest:** Al aplicar se obtuvo una accuracy de 95% y R-cuadrado de entrenamiento del 99.98%

Resultado de Random Forest

Figura 43

Resultados de Random Forest



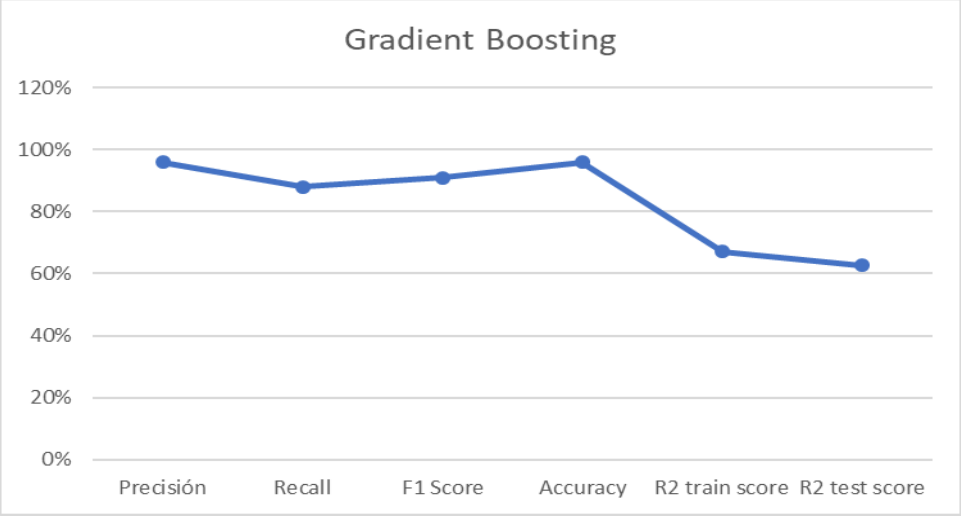
Nota. Elaboración propia.

- **Gradient Boosting:** Al aplicar se obtuvo una accuracy de 96% y R-cuadrado de entrenamiento del 67.05%

Resultados de Gradient Boosting

Figura 44

Resultados de Random Forest



Nota. Elaboración propia.

Dentro de estos tres modelos predictivos de machine learning se puede verificar una accuracy está entre el 95% y 96%, lo que es en gran parte consecuencia de las fases anteriormente realizada. Finalmente se presenta la comparación de los resultados obtenidos.

Cuadro comparativo de los resultados obtenidos:

Tabla 9:

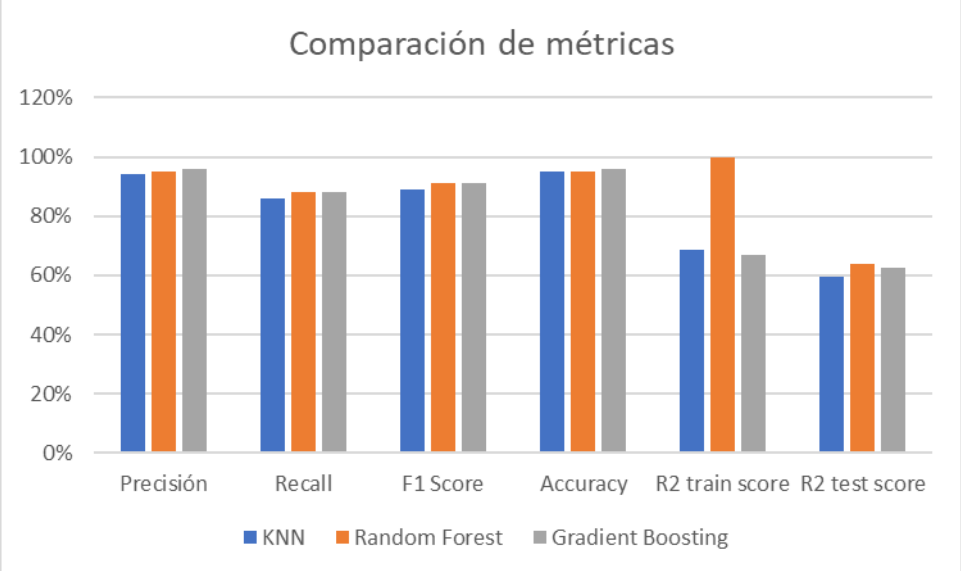
Cuadro comparativo de los resultados obtenidos

Métrica	KNN	Random Forest	Gradient Boosting
Precisión	94%	95%	96%
Recall	86%	88%	88%
F1 Score	89%	91%	91%
Accuracy	95%	95%	96%
R2 train score	68.63%	99.98%	67.05%
R2 test score	59.42%	63.65%	62.75%

Nota. Elaboración propia

Figura 45

Comparación de métricas



Nota. Elaboración propia.

Prueba de hipótesis

Hipótesis Estadística (HG)

HG_0: El modelo predictivo no determina de la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HG_1: El modelo predictivo determina de la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

Para la prueba de hipótesis se hace uso del estadístico R-cuadrado que mide la variación de una variable debido a la variación de otra variable, en tal sentido en la siguiente tabla se muestra los resultados obtenidos por cada modelo utilizado.

Tabla 10:

Prueba R-Cuadrado

Modelo Machine Learning	Datos de entrenamiento	Datos de validación
KNN	68.63%	59.42%
Random Forest	99.98%	63.65%
Gradient Boosting	67.05%	62.75%

Nota. Elaboración propia.

Después de una observación la tabla, los resultados señalan que el aprendizaje por cada modelo de machine learning se realizó de muy buena manera obteniéndose resultados por sobre el 63% de exactitud de los resultados previstos por los modelos y los resultados originales, por lo tanto, se rechaza la hipótesis nula y se acepta la hipótesis alterna.

Hipótesis especifica 1:

HE1_0: Los factores personales no influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HE1_1: Los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

Tabla 11:

Factores personales influyentes en la deserción estudiantil

Variable	Valor significancia	Relación con deserción	Valores de la frecuencia (%)
Sexo	0.00	Si	0.00
Edad	0.00	Si	14.30
Tipo de colegio	0.00	Si	0.00
Año de egreso del colegio	0.00	Si	14.30

Nota. Elaboración propia.

De los resultados obtenidos se puede visualizar que los factores personales influyentes dentro de la deserción estudiantil son Sexo, Edad, Tipo de colegio y Ano de egreso del colegio, donde mediante una prueba de chi-cuadrado se puede determinad que están correlacionadas con la deserción estudiantil de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo, por lo tanto, se rechaza la hipótesis nula y se acepta la hipótesis alterna.

Hipótesis especifica 2:

HE2_0: Los factores académicos no influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HE2_1: Los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

Tabla 12

Factores académicos influyentes en la deserción estudiantil

Variable	Valor significancia	Relación con deserción	Valores de la frecuencia (%)
Facultad	0.00	Si	0.00
Escuela	0.00	Si	4.00
Modalidad	0.00	Si	27.30
Puntaje de ingreso	0.992	No	85.80
Ciclo	0.00	Si	44.40
Promedio ponderado	0.00	Si	6.70
Promedio primer ciclo	0.00	Si	8.30
Promedio segundo ciclo	0.00	Si	6.70
Cantidad de créditos matriculados primer ciclo	0.00	Si	60.00
Cantidad de créditos matriculados segundo ciclo	0.00	Si	32.10

Cantidad de créditos aprobados de primer ciclo	0.00	Si	4.20
Cantidad de créditos aprobados de segundo ciclo	0.00	Si	12.50

Nota. Elaboración propia.

De los resultados obtenidos se puede visualizar que los factores académicos influyentes dentro de la deserción estudiantil son facultad, escuela, modalidad de ingreso, puntaje de ingreso, ciclo, promedio ponderado, promedio primer ciclo, promedio segundo ciclo, cantidad de créditos matriculados primer ciclo, cantidad de créditos matriculados segundo ciclo, cantidad de créditos aprobados primer ciclo, cantidad de créditos matriculados segundo ciclo, donde mediante una prueba de chi-cuadrado se puede determinad que todas las variables exceptuando el puntaje de ingreso están correlacionadas con la deserción estudiantil de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo, por lo tanto, se rechaza la hipótesis nula y se acepta la hipótesis alterna.

Hipótesis especifica 3:

HE3_0: Los factores socioeconómicos no influyen la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

HE3_1: Los factores socioeconómicos influyen la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

Posterior al análisis realizado se puede determinar que los factores socioeconómicos no son muy influyentes para la deserción de estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo dado que no se cuenta con la información necesaria para su análisis, por lo tanto, se rechaza la hipótesis alterna y se acepta la hipótesis nula.

4.3. Discusión de resultados

Se demostró que los modelos predictivos empleados en la presente investigación cuentan con buenas métricas de predicción entre los cuales el algoritmo que más destaco fue Gradient Boosting con una precisión del 96%, una sensibilidad del 88%, un score F1 del 91% y un accuracy del 96%, así mismo se pudo observar que a partir de la prueba estadística R-cuadrado de los datos de entrenamiento es del 67.05% y el R-cuadrado de los datos de validación es el 62.75%, lo que indica que este modelo tiene una gran asertividad en cuanto a la predicción de los estudiantes que son posibles a desertar dentro de su primer año de estudios en la Universidad Nacional Santiago Antúnez de Mayolo, así mismo se pudo visualizar que hubo algunas variables que se excluyeron del estudio por contar con demasiados nulos dentro de la recolección de datos. Los resultados hallados coinciden con los del antecedente Zapata (2021), en su investigación "Método para la Detección de Estudiantes en Riesgo de Deserción, Basado en un Diseño de Métricas y una Técnica de Minería de Datos" logro obtener un 82% de precisión y un 64% de recall, estos resultados representan una mejora significativa con respecto al 71% de precisión y el 57 % de recall obtenidos con características iniciales sin uso de métricas. Sin embargo, el modelo presentado en la tesis no detecta la clase desertora muy bien, pero cuando lo hace es considerablemente confiable. De igual manera, Behr, Giese, Teguim, & Theune (2020) en su investigación denominada "Early Prediction of University Dropouts – A Random Forest Approach" los autores realizan un estudio de la deserción de estudiantes mediante una clasificación binaria (graduado o abandono) donde el enfoque es una predicción muy temprana de abandono de estudiantes por lo cual se tomaron factores desde las calificaciones desde la escuela donde obtuvieron como resultado un AUC (área bajo la curva) de 0,86. Donde se determinó que los predictores importantes son la calificación final en la escuela secundaria, determinantes asociados con la satisfacción de los estudiantes y su autoconcepto académico subjetivo y autoevaluativo. Así mismo, Alvarado (2022) en su trabajo de investigación titulado "Estudio comparativo del nivel de eficacia en los modelos algorítmicos al estimar la deserción de los estudiantes del nivel pregrado en la universidad de Huánuco", donde el autor realizó una comparación de varios modelos de predicción y llego a concluir que el mejor de los modelos aplicados es Random forest y el peor modelo es KNN donde Random Forest llego a obtener una precisión del 77% y un recall del 76%.

Existen varios puntos de vista para poder predecir la deserción estudiantil pero sin embargo siempre será importante las técnicas de recolección de datos que se mantiene para cada una de las estrategias aplicadas, de la presente investigación se puede verificar que en la Universidad Nacional Santiago Antúnez de Mayolo se puede predecir la deserción estudiantil a partir de los fatores académicos y personales, no obstante se puede comenzar a masificar la recolección de los datos para mejorar las predicciones realizadas y este llegue a tener una mayor precisión en los alumnos que van a desertar dentro de la universidad, así mismo se puede ampliar a una investigación más profunda y realizarse en todos los semestres académicos.

V. CONCLUSIONES

1. Se concluye que con la aplicación de algoritmos de machine learning se puede tener la predicción de la deserción de estudiantes en su primer año de estudio dentro de la Universidad Nacional Santiago Antúnez de Mayolo, donde el mejor algoritmo obtenido en esta investigación fue Gradient Boosting con 96% de precisión, 88% de sensibilidad, 91% en el score F1, 96% de accuracy, así mismo se validó el modelo bajo el coeficiente R-cuadrado con valor igual al 67.05% en los datos de entrenamiento y 62.75% en los datos de validación.
2. Se concluyo que para la presente investigación los factores personales no son muy influyentes para la predicción de la deserción de estudiantes en su primer año de estudio dentro de la Universidad Nacional Santiago Antúnez de Mayolo, así mismo se puede evidenciar que existen escasos datos de los factores personales, y estos no pueden ser tomados dentro de la evaluación modelo de machine learning.
3. Se concluyo que para la presente investigación los factores académicos son los más influyentes para la predicción de la deserción de estudiantes en su primer año de estudio dentro de la Universidad Nacional Santiago Antúnez de Mayolo, donde se puede visualizar que el ciclo, promedio ponderado, promedio del primer ciclo, promedio del segundo ciclo, cantidad de créditos matriculados segundo ciclo, cantidad de créditos aprobados primer ciclo y cantidad de créditos aprobados segundo ciclo, son los datos más influyentes durante el aprendizaje del modelo de machine learning.
4. Se concluyo que para la presente investigación los factores socioeconómicos no son muy influyentes para la predicción de la deserción de estudiantes en su primer año de estudio dentro de la Universidad Nacional Santiago Antúnez de Mayolo, así mismo se puede evidenciar que existen escasos datos de los factores socioeconómicos, y estos no pueden ser tomados dentro de la evaluación modelo de machine learning.

VI. RECOMENDACIONES

1. Tomar como base la presente investigación para realizar nuevas estrategias de machine learning y tratar de buscar otras formas de predicción de la deserción de estudiantes en la Universidad Nacional Santiago Antúnez de Mayolo, de esta manera propiciar la investigación de modelos de machine learning dentro de la universidad.
2. Masificar la recolección de los datos socioeconómicos y personales de los estudiantes para realizar un mejor seguimiento de cada uno de ellos, de manera más personalizada.
3. Generar más filtros de calidad a cada uno de los datos que se registran dentro de la base de datos, dado que existen muchos campos nulos y otros muchos que son erróneos lo cual dificulta a la hora de realizar investigaciones acerca de los alumnos dentro de la universidad.
4. Consolidar información de las bases de datos de la universidad ya que dentro de las propias oficinas de la Universidad Nacional Santiago Antúnez de Mayolo se maneja diferente información de ciertos estudiantes y esto conlleva al análisis y descarte de datos obtenidos.
5. Implementar el modelo de predicción dentro de un API para su puesta en producción y apoyo del seguimiento estudiantil y continuar con la retención estudiantil de la universidad.
6. Ampliar la investigación a los estudiantes de todos los semestres de la Universidad Nacional Santiago Antúnez de Mayolo, para no tener la limitante que solo se pueda predecir la deserción de los estudiantes al finalizar su segundo semestre de estudio.

VII. REFERENCIAS BIBLIOGRÁFICAS

Albarrán, J. (2019). La deserción estudiantil en la Universidad de los Andes (Venezuela). *Revista Educación y Humanismo, 21*(36), 60-92. Obtenido de https://revistas.unisimon.edu.co/index.php/educacion/article/view/2806/4023

Alvarado, J. (2022). *Estudio comparativo del nivel de eficacia en los modelos algorítmicos al estimar la deserción de los estudiantes del nivel pregrado en la universidad de Huánuco.* Huánuco: Universidad de Huánuco.

Amat, J. (Octubre de 2020). *Gradient Boosting con Python*. Obtenido de www.cienciadedatos.net: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

Arias, F. (2012). *El proyecto de investigación.* Caracas: Episteme.

Behr, A., Giese, M., Teguim, H., & Theune, K. (2020). Early Prediction of University Dropouts – A Random Forest Approach. *Journal of Economics and Statistics*, 1-47.

Bernal, C. (2010). *Metodología de la investigación.* Colombia: Pearson.

Boschetti, A., & Massaron, L. (2018). *Python Data Science.* Birmingham: Packt.

Camargo, A. (2020). *Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos.* Barranquilla: Universidad de la Costa.

Castañeda, R., & López, E. (2015). *Factores determinantes del rendimiento académico en esetudiantes universitarios de la facultad de economía - UNCP en el periodo 2014-I.* Huancayo: Universidad Nacional del Centro del Perú.

Castillo, M., Gamboa, R., & Hidalgo, R. (2019). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas. *Revista UNICIENCIA, 34*(1), 219-245. Obtenido de https://www.scielo.sa.cr/pdf/uniciencia/v34n1/2215-3470-uniciencia-34-01-219.pdf

Cevallos, E., & Barahona, C. (2021). *Modelo para automatizar el proceso de predicción de la deserción en estudiantes universitarios en el primer año de estudio.* Lima: Universidad Peruana de Ciencias Aplicadas.

Dickey, D., State, C., & Raleigh. (2012). Introduction to Predictive Modeling with Examples. *SAS Global Forum 2012*, 1-14.

Freitas, F., Vasconcelos, F., Peixoto, S., Mehedi, M., Akber, A., Albuquerque, V., & Rebouças, P. (2020). Sistema IoT para Predicción de Abandono Escolar Utilizando Técnicas de Aprendizaje Automático Basado en Datos Socioeconómicos. *Electronics*, 1-14.

Gallardo, J. (2009). *Metodología para la Definición de Requisitos en Proyectos de Data Mining (ER-DM).* Madrid: Universidad Politécnica de Madrid.

Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.* United States of America: O'Reilly.

Girón, L., & González, D. (2005). Determinantes del rendimiento académico y ladeserción estudiantil, en el programa deEconomía de la Pontificia UniversidadJaveriana de Cali. *Economía Gestión y Desarrollo*, 173-201.

Guazzelli, A. (19 de junio de 2012). *IBM Developer*. Obtenido de IBM: https://developer.ibm.com/articles/ba-predictive-analytics2/?mhsrc=ibmsearch_a&mhq=model%20predictive

Hashmi, O., & Sheikh, S. (2012). Impact of social attributes on Predictive Analytics in telecommunication industry. *2012 15th International Multitopic Conference (INMIC)* (págs. 47–52). Islamabad: IEEE.

Hernández, R., Fernández, C., & Baptista, M. (2014). *Metodología de la investigación.* México D.F.: McGrawHillEducation.

Himmel, E. (2002). Modelos de análisis de la deserción estudiantil en la educación superior. *Calidad en la educación*, 91-108.

IBM. (15 de Julio de 2020). *Aprendizaje automático*. Obtenido de IBM: https://www.ibm.com/cloud/learn/machine-learning?mhsrc=ibmsearch_a&mhq=machine%20learning

IBM. (2022). *Algoritmo de k vecinos más cercanos*. Obtenido de IBM: https://www.ibm.com/co-es/topics/knn#:~:text=El%20algoritmo%20de%20k%20vecinos%20m%C3%A1s%20cercanos%2C%20tambi%C3%A9n%20conocido%20como,un%20punto%20de%20datos%20individual.

Igual, L., & Seguí, S. (2017). *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications.* Switzerland: Springer.

INEI. (2020). *Indicadores de Educación por Departamento, 2009-2019.* Lima: INEI.

Jones, H. (2019). *Ciencia de datos para empresas - Modelo predictivo, Minería de datos, Análisis de datos, Análsis de regresión, Consulta de bases de datos y Aprendizaje automático para principiantes.* Estados Unidos: Bravex Publications.

Kreiger, J. (Enero de 2020). *Evaluating a Random Forest model*. Obtenido de Medium: https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56

Leek, J. (2015). *The Elements of Data Analytic Style.* Leanpub.

Microsoft. (27 de Julio de 2022). *Aprendizaje profundo frente a aprendizaje automático en Azure Machine Learning*. Obtenido de Micorsoft: https://docs.microsoft.com/es-es/azure/machine-learning/concept-deep-learning-vs-machine-learning

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Mori, M. (Noviembre de 2012). Deserción universitaria en estudiantes de una universidad privada de Iquitos. *Revista Digital de Investigación en Docencia Universitaria*, págs. 60-83.

Novoa, E. (2019). *Reducción del Riesgo de Deserción Académica mediante seguimiento de alumnos en una universidad.* Lima: Universidad Ricardo Palma.

Ozdemir, S., & Kakade, S. (2018). *Principles of Data Science.* Birmingham: Packt.

Palmer, S. (2015). *Data Science Advisory*. Obtenido de Shelly Palmer: https://www.shellypalmer.com/data-science/

Rodríguez, J., & Hernández, J. (2008). La deserción escolar universitaria en Mexico. La experiencia de la Universidad Autónoma Metropolitana campus Iztapalapa. *Actualidades Investigativas en Educación*, 1-30.

Sánchez, G., Barboza, M., & Castilla, H. (2017). Análisis de la deserción y los factores asociados a la permanencia estudiantil en una universidad peruana. *Actualidades Pedagógicas*, 169-191.

Shica, Z. (2022). *Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa 88331 en Chimbote - 2021*. Trujillo: Universidad César Vallejo.

Siegel, E. (2016). *Predictive Analytics: The power to predict who will click, buy, lie, or die.* New Jersey: John Willey & Sons, Hoboken.

Stanton, J. (2013). *Introduction to Data Science.* New York: Syracuse University's School.

Tejedor, F., & García-Valcárcel, A. (2001). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES. *Revista de educación*, 443-473.

Viale, H. (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista Digital de Investigación en Docencia Universitaria*, 59-75.

Witten, I., Frank, E., & Hall, M. (2017). *Data Mining: Practical Machine Learning Tools and Techniques.* Estados Unidos: ELSEVIER.

Zambrano, G., Rodríguez, K., & Guevara, L. (2018). Análisis de la deserción estudiantil en las Universidades del Ecuador y América Latina. *Revista Pertinencia Académica*(8), 1-28. Obtenido de https://revistas.utb.edu.ec/index.php/rpa/article/view/2451/2059

Zapata, D. (2021). *Método para la Detección de Estudiantes en Riesgo de Deserción, Basado en un Diseño de Métricas y una Técnica de Minería de Datos.* Medellín: Universidad Nacional de Colombia.

ANEXOS

Anexo 1: Matriz de consistencia de la investigación

Tabla 13:

Matriz de consistencia de la investigación

Problema		Hipótesis		Objetivo	
General	**Específicos**	**General**	**Específicos**	**General**	**Específicos**
¿La aplicación de un modelo predictivo determina la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo -2022?	1. ¿Los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022? 2. ¿Los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022? 3. ¿Los factores socioeconómicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022?	El modelo predictivo determina de la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.	1. Los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022. 2. Los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022. 3. Los factores socioeconómicos influyen la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.	Determinar mediante un modelo predictivo la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo - 2022.	1. Establecer si los factores personales influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022. 2. Establecer si los factores académicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022. 3. Establecer si los factores socioeconómicos influyen en la deserción de estudiantes en el primer año de estudio en la Universidad Nacional Santiago Antúnez de Mayolo – 2022.

Fuente: Elaboración propia

Anexo 2: Instrumento de recolección de datos

ANÁLISIS DOCUMENTAL DE LOS DATOS DE LOS ESTUDIANTES DE LA UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO

Datos a recolectar de los estudiantes a la universidad nacional Santiago Antúnez de Mayolo:

Datos	Factor	Datos del estudiante
1. Código de estudiante	Académico	
2. Facultad	Académico	
3. Escuela profesional	Académico	
4. Modalidad de ingreso	Académico	
5. Puntaje de ingreso	Académico	
6. Sexo	Personal	
7. Edad	Personal	
8. Estado civil	Personal	
9. Lugar de procedencia	Personal	
10. Colegio de procedencia	Personal	
11. Año de egreso de la secundaria	Personal	
12. Nivel de instrucción padre	Socioeconómico	
13. Nivel de instrucción madre	Socioeconómico	
14. Cantidad de hermanos	Socioeconómico	
15. Situación económica	Socioeconómico	
16. Ciclo actual	Académico	
17. Promedio ponderado	Académico	
18. Promedio del primer semestre	Académico	
19. Promedio del segundo semestre	Académico	
20. Hábitos de estudio	Académico	
21. Número de créditos matriculados en el primer semestre	Académico	

22. Número de créditos matriculados en el segundo semestre	Académico	
23. Número de créditos aprobados en el primer semestre	Académico	
24. Número de créditos aprobados en el segundo semestre	Académico	
25. Deserción	Académico	

Anexo 3: Validación de experto

UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO
"Una nueva universidad para el desarrollo"
Facultad de Ciencias – Escuela Profesional de Ingeniería de Sistemas e Informática

MATRIZ DE EVALUACIÓN DEL INSTRUMENTO

Indicadores	Criterios	Totalmente en desacuerdo 1: 00 - 20				En desacuerdo 2: 21 - 40				Ni de acuerdo ni en desacuerdo 3: 41 - 60				Muy de acuerdo 4: 61 - 80				Totalmente de acuerdo 5: 81 - 100			
		05	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Esta formulado con lenguaje apropiado.																				X
2. OBJETIVIDAD	Esta expresado en conductas observables.																		X		
3. ACTUALIDAD	Adecuado al avance de la ciencia y tecnología.																			X	
4. ORGANIZACIÓN	Existe orden lógico de ideas.																			X	
5. SUFICIENCIA	Comprende las dimensiones de la investigación en cantidad y calidad.																		X		
6. INTENCIONALIDAD	Adecuado para valorar la variable seleccionada																			X	
7. CONSISTENCIA	Basado en el aspecto teórico científico y del tema de estudio.																			X	
8. COHERENCIA	Hay relación entre variables, dimensiones e indicadores.																			X	
9. METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto																				X
10. APLICABILIDAD	El instrumento es de fácil aplicación.																				X

POHL CACERES PAUL ELBIN
46408160
CIP: 178063

UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO

"Una nueva universidad para el desarrollo"

Facultad de Ciencias – Escuela Profesional de Ingeniería de Sistemas e Informática

MATRIZ DE EVALUACIÓN DEL INSTRUMENTO

Indicadores	Criterios	Totalmente en desacuerdo 1: 00 - 20				En desacuerdo 2: 21 - 40				Ni de acuerdo ni en desacuerdo 3: 41 - 60				Muy de acuerdo 4: 61 - 80				Totalmente de acuerdo 5: 81 - 100			
		05	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Esta formulado con lenguaje apropiado.																				X
2. OBJETIVIDAD	Esta expresado en conductas observables.																			X	
3. ACTUALIDAD	Adecuado al avance de la ciencia y tecnología.																				X
4. ORGANIZACIÓN	Existe orden lógico de ideas.																				X
5. SUFICIENCIA	Comprende las dimensiones de la investigación en cantidad y calidad.																		X		
6. INTENCIONALIDAD	Adecuado para valorar la variable seleccionada																				X
7. CONSISTENCIA	Basado en el aspecto teórico científico y del tema de estudio.																			X	
8. COHERENCIA	Hay relación entre variables, dimensiones e indicadores.																				X
9. METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto																				X
10. APLICABILIDAD	El instrumento es de fácil aplicación.																			X	

UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO

"Una nueva universidad para el desarrollo"

Facultad de Ciencias – Escuela Profesional de Ingeniería de Sistemas e Informática

MATRIZ DE EVALUACIÓN DEL INSTRUMENTO

Indicadores	Criterios	Totalmente en desacuerdo 1: 00 - 20				En desacuerdo 2: 21 - 40				Ni de acuerdo ni en desacuerdo 3: 41 - 60				Muy de acuerdo 4: 61 - 80				Totalmente de acuerdo 5: 81 - 100			
		05	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Esta formulado con lenguaje apropiado.																			X	
2. OBJETIVIDAD	Esta expresado en conductas observables.																		X		
3. ACTUALIDAD	Adecuado al avance de la ciencia y tecnología.																			X	
4. ORGANIZACIÓN	Existe orden lógico de ideas.																			X	
5. SUFICIENCIA	Comprende las dimensiones de la investigación en cantidad y calidad.																			X	
6. INTENCIONALIDAD	Adecuado para valorar la variable seleccionada																			X	
7. CONSISTENCIA	Basado en el aspecto teórico científico y del tema de estudio.																	X			
8. COHERENCIA	Hay relación entre variables, dimensiones e indicadores.																		X		
9. METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto																		X		
10. APLICABILIDAD	El instrumento es de fácil aplicación.																			X	

UNIVERSIDAD NACIONAL SANTIAGO ANTÚNEZ DE MAYOLO

"Una nueva universidad para el desarrollo"

Facultad de Ciencias – Escuela Profesional de Ingeniería de Sistemas e Informática

MATRIZ DE EVALUACIÓN DEL INSTRUMENTO

Indicadores	Criterios	Totalmente en desacuerdo 1: 00 - 20				En desacuerdo 2: 21 - 40				Ni de acuerdo ni en desacuerdo 3: 41 - 60				Muy de acuerdo 4: 61 - 80				Totalmente de acuerdo 5: 81 - 100			
		05	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
1. CLARIDAD	Esta formulado con lenguaje apropiado.																				X
2. OBJETIVIDAD	Esta expresado en conductas observables.																			X	
3. ACTUALIDAD	Adecuado al avance de la ciencia y tecnología.																			X	
4. ORGANIZACIÓN	Existe orden lógico de ideas.																			X	
5. SUFICIENCIA	Comprende las dimensiones de la investigación en cantidad y calidad.																			X	
6. INTENCIONALIDAD	Adecuado para valorar la variable seleccionada																		X		
7. CONSISTENCIA	Basado en el aspecto teórico científico y del tema de estudio.																			X	
8. COHERENCIA	Hay relación entre variables, dimensiones e indicadores.																			X	
9. METODOLOGÍA	El instrumento se relaciona con el método planteado en el proyecto																		X		
10. APLICABILIDAD	El instrumento es de fácil aplicación.																			X	

COLEGIO DE INGENIEROS DEL PERÚ
CONSEJO DEPARTAMENTAL ANCASH - HUARAZ
GONZALEZ BARRETO CRISTIAN YERSI
INGENIERO DE SISTEMAS E INFORMATICA
Cip. Nº 252995