UNIVERSIDAD CÉSAR VALLEJO

ESCUELA DE POSGRADO

PROGRAMA ACADÉMICO DE MAESTRÍA EN INGENIERÍA DE SISTEMAS CON MENCIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN

Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa 88331 en Chimbote - 2021

TESIS PARA OBTENER EL GRADO ACADÉMICO DE:

Maestra en Ingeniería de Sistemas con Mención en Tecnologías de Información

AUTORA:

Shica Julca, Zenaida Cristina (ORCID: 0000-0001-6142-1640)

ASESOR:

Dr. Pacheco Torres, Juan Francisco (ORCID: 0000-0002-8674-3782)

LÍNEA DE INVESTIGACIÓN:

Sistemas de Información y Comunicaciones

TRUJILLO – PERÚ
2022

Dedicatoria

A Dios por iluminar mi camino, ser mi soporte espiritual en mi desarrollo personal.

A mi Madrina Antonia y mi Padrino Eduardo, ejemplo de amor, paciencia y trabajo por su incondicional apoyo y consejos a lo largo de mi desarrollo profesional y personal.

A mis padres Zenobia y Félix; a Cristian, mi sobrina Emma y mis hermanos por ser siempre mi soporte para seguir adelante en cada paso de este proyecto y en mi vida.

Agradecimiento

A Dios por iluminar mi camino, ser mi soporte espiritual en mi desarrollo personal y profesional.

A mis docentes de la Escuela de Postgrado de la Universidad César Vallejo por sus conocimientos impartidos con dedicación que contribuyeron a mi formación académica y profesional.

Al Ing. Juan Francisco Pacheco Torres, por su apoyo como asesor a lo largo de toda la maestría y finalmente terminar el presente trabajo de investigación

Al Ex Director Luis Alberto Torres Álvarez, quien en vida me brindó su apoyo en la realización de esta investigación y gran contribución en mi desarrollo profesional y personal.

Índice de contenidos

Índice de tablas

Índice de figuras

RESUMEN

La presente investigación muestra un modelo de Data Science para la detección de la deserción escolar en la Institución Educativa 88331 en el centro poblado Rinconada, Chimbote. Bajo la Jornada escolar completa en el nivel Secundaria. Este proyecto nace a raíz de un problema latente en el sector educativo de la Educación Básica Regular: La deserción escolar; referido a los estudiantes de nivel básico regular que abandonan sus estudios antes de terminar su año escolar.

Se analizó un historial de notas del nivel secundaria matriculados desde el año 2011 al año 2019, con un total de 804 estudiantes a los largo de dichos años. Las variables estudiadas fueron género, fecha de nacimiento, grado, sección, año académico, notas por cursos, áreas desaprobadas, comportamiento y situación final.

Para el desarrollo del modelo de Data Science Regresión logística se utilizó la plataforma Cloud Google Colab, con el lenguaje de programación Python, bajo la metodología CRISP-DM, se trabajó con un primer grupo del 70% en Train o entrenamiento y un 30% en test o testeo, logrando obtener una óptima precisión.

Finalmente como resultado de implementar el modelo de Data Science de Regresión Logística se obtuvo una mejora de la Tasa de retención escolar de 84,1 % a un 95,5% en un año escolar.

Palabras Clave: Modelo de Ciencia de Datos, deserción escolar, abandono escolar, Aprendizaje Automático, Regresión Logística

ABSTRACT

The present investigation shows a Data Science model for the detection of school dropouts in the Educational Institution 88331 in the Rinconada town center, Chimbote. Under the full school day at the Secondary level. This project was born as a result of a latent problem in the educational sector of Regular Basic Education: School desertion; refers to regular basic level students who drop out before the end of their school year.

A history of high school grades enrolled from 2011 to 2019 was analyzed, with a total of 804 students throughout those years. The variables studied were gender, date of birth, grade, section, academic year, grades by courses, disapproved areas, behavior and final situation.

For the development of the Data Science Logistic Regression model, the Cloud Google Colab platform was used, with the Python programming language, under the CRISP-DM methodology, we worked with a first group of 70% in Train or training and 30% in test or testing, achieving optimum precision.

Finally, as a result of implementing the Logistic Regression Data Science model, an improvement in the school retention rate was obtained from 84.1% to 95.5% in one school year.

Keywords: Data Science Model, dropout, dropout, Machine Learning, Logistic Regression.

I. INTRODUCCIÓN

En el Perú, en diversas instituciones educativas existe gran número de información de sus alumnos de cada año de estudios, esta base de datos surge del registro de notas de estos y aspectos personales, institucionales o sociales. (Arce, 2015). La Data Science es aplicable para diversos aspectos como ayuda para resolver problemas, estos deben contar con un número de datos para el reconocimiento de algún patrón conductual de estos, lograr su representación de apoyo formativo y formular mejoras. (Arteaga, et al, 2018).

Educarse es fundamental para que países como Perú avancen, ya que concede generar el rendimiento capital humano, como un factor importante de todo sistema. (Asif, 2017). La educación es indispensable para el alcance de una economía sostenible a largo plazo, por ende, últimamente los responsables de establecimientos educativos ponen énfasis en los resultados académicos de sus alumnos y los aspectos influyentes para estos, la investigación y el analizar forman instrumentos sólidos para generar indicadores que dirijan a tomar decisiones educativas (Box, 2015).

La deserción académica forma parte de una gran problemática social, con repercusiones a nivel personal y familiar. (Cambruzzi, et al, 2015). La actual situación del ejercicio académico de los alumnos en el Perú, es alarmante, ya que a causa de este inadecuado desarrollo se conoce uno de varios problemas que resultan de esta circunstancia, lo difícil que es retomar los estudios. Asimismo, y para comprender mejor, se precisará el rendimiento estudiantil en diversas situaciones. (Orihuela, 2019).

Dentro de la Institución Educativa N° 88331, se halla que indudablemente uno el problema que más preocupa a los docentes es el nivel de pobreza de sus alumnos donde el 70% de ellos tiene condición de pobreza, además existen estudiantes que viven en los anexos del centro poblado Rinconada bajo la condición de extrema pobreza; otro factor es el abandono de los padres que es latente en las familias de los estudiantes quienes en algunos casos tienen

padres separados y viven con el padre o la madre bajo una condición de abandono en su alimentación y calidad de vida; también se encuentran algunos estudiantes que viven solos o con algún familiar como los abuelos.

Los estudiantes se ven influenciados por la cultura tradicional del cultivo del campo, sin mayor interés en la educación básica, que se repite a través de las generaciones dejando en segundo lugar la educación de los menores; otro factor latente es el consumo de drogas a temprana edad, se ha reportado casos de consumo de drogas en grupos desde primer grado de Secundaria con alta probabilidad de deserción del estudiante sino son controlados a tiempo; el factor embarazo precoz es producto de las familias disfuncionales donde las señoritas desde primer grado de secundaria han presentado casos de embarazo y un 90% de los casos dejan los estudios por el poco interés de los padres o su pareja; el factor de horas académicas mal organizadas.

El cambio de modalidad de estudio a Jornada Escolar Completa ha demandado 3 horas académicas más a la modalidad regular y esto a lo largo del año escolar ocasiona mucho cansancio y bajo interés por las clases al tener por ejemplo varios cursos de letras consecutivos; podemos agregar el cambio de aulas multifuncionales que trajo la nueva modalidad que involucra la creación de aulas especializadas para cada una de las 11 Áreas por ellos los estudiantes cada una, dos o tres horas académicas deben de cambiar de aula ocasionando ineficiente gestión de las horas efectivas de clases.

Podemos agregar a los factores el reducido tiempo programado para el almuerzo de los estudiantes luego de estudiar seis horas académicas, los estudiantes acceden a un tiempo de 30 minutos para el consumo de sus alimentos el cual lleva a un cansancio para terminar las últimas 3 horas académicas programadas, también es necesario mencionar que algunos estudiantes no reciben el almuerzo de sus padres, por residir en zonas alejadas al colegio y aún peor no tienen la condición económica para adquirir sus alimentos en el quiosco, sumando a ello el bullying por sus compañeros con mejores recursos económicos.

Finalmente se puede mencionar que algunas áreas no son adaptadas a la realidad de los estudiantes, los docentes presentan todos los días las sesiones de aprendizajes a los coordinadores académicos quienes lo revisan y devuelven para la ejecución de las sesión sin embargo muchas sesiones son descargadas de la plataforma del Ministerio de Educación y no son contextualizadas a la realidad de los estudiantes, generando sobrecarga de actividades, desmotivación y poca compresión de la sesión por parte de los estudiantes.

Todos estos factores han generado una preocupación en la deserción de los estudiantes en etapas de desarrollo del Año escolar, al año 2020 se encuentran matriculados 220 estudiantes, 27 docentes y 9 secciones; una situación preocupante porque dada la coyuntura el acceso a la educación remoto ha impactado sobre los estudiantes quienes en su gran mayoría no tienen acceso a una computadora y peor aún a internet.

El año 2020 se desarrolló bajo la modalidad virtual, en este contexto es aún más preocupante el logro de los aprendizajes, se contó con un 70% de estudiantes que cumplieron por lo menos con desarrollar sus actividades de manera regular y mantuvieron contacto con sus docentes. El 30% de matriculados es un aproximado de 60 estudiantes que pasaron a la etapa de recuperación que se desarrolló en enero y febrero, dentro de este grupo de estudiantes podemos encontrar los diversos factores mencionados líneas arriba, por lo cual la probabilidad de éxito de su recuperación académica tiende a ser baja.

Las instituciones educativas públicas y privadas realizan su gestión bajo las normativas establecidas por el Ministerio de Educación, el cual cada año emite las normas y orienta para desarrollar adecuadamente el año escolar. La gestión académica está alineada al cumplimiento de 7 compromisos de gestión, dentro de los cuales se pretende estudiar el compromiso número dos que es la Retención Anual de los alumnos en la institución educativa, este indicador mide la tasa de deserción escolar durante el año escolar, sin embargo, como se ha

descrito en párrafos anteriores la cantidad de estudiantes matriculados han ido reduciéndose año tras año ocasionando amonestaciones al equipo Directivo por parte de la Ugel Santa, incluso reducción de plazas docentes.

Además, se ha descrito posibles factores que pueden influenciar sobre la deserción del estudiante; es importante, una detección temprana de los posibles casos de deserción, para generar alternativas de solución que eviten la deserción de los estudiantes.

El objetivo del compromiso de gestión estudiantil Retener Anualmente a alumnos es: la casa de estudios permanece con la cantidad de alumnos matriculados al inicio de la época escolar, cuyos indicadores son: la matrícula oportuna a los alumnos y reportar en el SIAGIE, elaborar el análisis de alumnos desertores, aquellos en riesgo de desertar, identificarlo que origina la deserción de estudios escolares, control de asistencia de alumnos constantemente, a través del reporte mes a mes en el SIAGIE, establecer en el PAT(Plan anual de Trabajo) las acciones que prevengan y que corrijan para evitar la inasistencia y abandonen sus estudios.

El equipo directivo realiza la gestión académica en base a los indicadores del compromiso de gestión escolar N° 2 “Retención Anual de estudiantes en la Institución Educativa” y los indicadores de análisis del reporte de alumnos desertores o en riesgo a desertar, hallando los motivos de abandono de la IE, control de asistencia de alumnos constantemente, a través de un reporte por mes en el SIAGIE se advirtió una deficiente inferencia de información reportada por el SIAGIE, al no ser aprovechada de manera eficiente para detectar los posibles casos de deserción escolar de manera oportuna por lo tanto es imposible cumplir correctamente con el indicador de establecer en el PAT l (Plan anual de Trabajo) las medidas de prevención y corrección que eviten la inasistencia y abandono de estudiantes.

Ante lo expuesto, se planteó la siguiente pregunta de investigación, el cuál será: ¿De qué manera los modelos de Data Science influyen en la detección de la

Deserción Académica de los estudiantes de la Institución Educativa Nº 88331 de la ciudad de Chimbote en el año 2021?

El proyecto tiene una implicancia operativa, ya que a través de los modelos de Data Science permitirá a los responsables de la gestión de la I.E. la mejora en la detección de la deserción escolar, con fácil acceso a la información de manera confiable y con garantía de disponibilidad. A nivel tecnológico, hoy en día, la disponibilidad a costo cero, el software que se necesita para analizar estadísticamente la modelización y luego calibrar los modelos de Data Science. R y Python concedieron el desarrollo de los modelos con técnicas estadísticas más avanzadas.

El proyecto tiene una justificación económica ya que al contar con uno o más modelos de Data Science permitió optimizar la identificación de los posibles casos de estudiantes a desertar mejorando los procesos y estrategias, optimizando el tiempo. A nivel Social, el proyecto tiene justificación social porque permitió detectar la Deserción escolar, para su rápida intervención por los actores responsables, de manera que el estudiante culmine sus estudios sin interrupción. A nivel metodológico, la contingencia de áreas nuevas aplicadas a la investigación de datos y la sustentabilidad de información de investigaciones que emplean estos conocimientos nuevos muy cercanos a lo que se describe en este estudio, como otros ejemplos que clasifican, predijeron de forma acertada los posteriores sucesos a venir.

El estudio presentó el objetivo general: Determinar el efecto de los modelos de Data Science en la Deserción Académica de los estudiantes de la Institución Educativa Nº 88331 de la ciudad de Chimbote en el año 2021 a través de la aplicación de modelos Data Science para mejorar la detección de deserción de estudiantes de acuerdo a indicadores académicos, de conducta, sociales.

Apoyado de los objetivos específicos: (a) Incrementar la tasa de retención de estudiantes de la Institución educativa Nº 88331, (b) Reducir el tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución

educativa Nº 88331, (c) Incrementar el grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa Nº 88331, (d) Incrementar el grado de satisfacción del Equipo Directivo de la Institución educativa Nº 88331.

De acuerdo a la pregunta de investigación se propone la siguiente hipótesis: Los Modelos de Data Science mejorarán significativamente la detección de la Deserción Académica de los estudiantes de la Institución Educativa Nº 88331 de la ciudad de Chimbote en el año 2021.

Las hipótesis específicas son las siguientes: (1) El incremento de la tasa de retención de estudiantes de la Institución educativa Nº 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar. (2) La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa Nº 88331, se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos. (3)El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa Nº 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación. (4) El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa Nº 88331, se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science.

II. MARCO TEÓRICO

Para sustentar bajo aspectos teóricos y metodológicos, el estudio se dirige y considera como antecedentes referencias, captados de artículos de investigación y de tesis internacionales, nacionales y locales.

En la tesis de Vinueza (2021) denominado *"Diseñar un modelo matemático que estime el abandono de estudios a través de métodos analíticos multivariados en un instituto superior tecnológico"* publicado en el Repositorio de la Universidad Técnica de Ambato de Ecuador. El principal objetivo fue diseñar un modelo matemático que estime el abandono de estudios de los alumnos a través de técnicas de análisis multivariado y bajo un método bibliográfico y descriptivo se determinó en los resultados que Se rechazó la hipótesis nula, porque dos de las variables, carrera y repitencia son diferentes a 0. El coeficiente modelo de regresión logística IV clasificó con una precisión del 83% en la etapa de training y con un 79 % en la etapa de test. Además, se estimó un modelo de predicción con árboles de decisión, el cual fijó como variable explicativa el dato 'carrera'. Finalmente se concluye que la precisión del modelo de regresión logística IV es mejor que el modelo con árbol de decisión.

En la tesis de Cevallos y Barahona (2021) titulado *"Modelo que automatice el procedimiento de predicción del abandono de estudios universitarios en estudiantes del primer año de estudio"* publicado por el Repositorio de la Universidad Peruana de Ciencias Aplicadas. Para el estudio se mantuvo la propuesta de un patrón para automatizar el pronóstico del abandono de estudios de alumnos universitarios. En los resultados se halló que las redes bayesianas interactúan adecuadamente a comparación de otros algoritmos, si se comparan con las métricas precisas, exactas, específicas y margen de error.

Independientemente, la determinación de las redes bayesianas logra un 67.10% a comparación de los árboles de decisión que fue de un 61,92% en la muestra de entrenamiento para la iteración con razón de 8:2. También, las variables "persona deportista" (0,29%), "vivienda propia" (0,20%) y "calificaciones de preparatoria" (0,15%) son aquellas que contribuyen más al

patrón predictivo. Los autores concluyen que la herramienta más adecuada para el desarrollo de este proyecto es IBM SPSS MODELER ya que cuenta con el soporte y documentación necesaria para la implementación del modelo.

En el artículo de Ramírez y Grandón (2018) titulado *"Predecir la Deserción Académica en una Universidad Pública Chilena a través de la clasificación basada en Árboles de Decisión con Parámetros Optimizados*" publicado por la Revista Formación Universitaria de Chile. El principal objetivo fue dar a conocer una catalogación establecida en árboles de decisión (CBAD) parametrizada óptimamente en predestinar el abandono de estudios. En los resultados se determinó que esta aplicación alcanzó una precisión de 87.27%. Concluyendo que emplear este método de CBAD optimizando parámetros es más preciso si se compara a otros estudios con similar cantidad de datos.

En el artículo de Sifuentes (2018) titulado "Modelos predictivos de la deserción estudiantil en una universidad privada peruana" publicado por la Revista Industrial Data, en el que se planteó como finalidad establecer de qué manera emplear patrones predictivos en áreas de mayor criticidad ayudan a determinar qué alumnos se encuentran por desertar. Bajo un estudio descriptivo, se halló en los resultados que destacó que los patrones de predicción conllevan a disminuir en un 25 % y 40 % el grado de desaprobados y las variables de mejor predicción fueron la profesión que estudian (vocación), la cantidad de matrículas en el curso y la calificación obtenida en matemáticas o comunicación al cursar quinto de secundaria.

Se concluye que los siete modelos predictivos son efectivos, porque se alcanzó trabajando cada muestra precisamente por asignatura, por lo que especificaciones de las áreas y lo que necesitan los alumnos eran diversas, especificación es analizadas en el historial estudiantil de los alumnos brindado por los encargados de registrar matrículas en la universidad particular que formó parte en la investigación.

En el artículo de investigación de Martelo, Herrera y Villabona (2017) titulado *"Planes para aminorar la deserción universitaria a través de series temporales y multipol"* publicado por la Revista Espacios de Venezuela. Se tuvo como principal objetivo definir planes que disminuyan desertar en el programa Administración de Empresas de la Universidad de Cartagena y los resultados se tomaron mediante la entrevista a 272 estudiantes, 15 docentes de planta y 45 catedráticos, que forman parte del programa, conllevaron a patrones educativos accesibles, enseñanzas pertinentes y convenios con instituciones de educación media, como planificaciones aptas para aminorar el abandono estudiantil. Se concluye que lo idóneo para la prevención de la deserción es crearlo antes mencionado.

En la tesis de Cabanillas (2017) denominado *"Factores influyentes en la deserción escolar de los alumnos de secundaria de la red educativa Eduardo Villanueva del Distrito Eduardo Villanueva de la Provincia de San Marcos: 2011 – 2013"* publicado por el Repositorio de la Universidad Nacional de Cajamarca, se presentó el objetivo que fue establecer las causas influyentes en el abandono escolar de los alumnos de secundaria y bajo una metodología descriptiva simple y no experimental se halló como resultados que la economía familiar es un factor de incidencia elevada en el abandono escolar de 96 %, las bajas calificaciones influyen notoriamente a desertar y posee un 69 % en matemática, 65.5% en Ciencia Tecnología y Ambiente y 59.7 % en Comunicación y los problemas familiares alcanzan un grado notorio de 68 %.

Concluyendo así que las bajas calificaciones son de 69% en el área de Matemática como la más preocupante, igualmente se presentan niveles importantes en los otros aspectos como la enseñanza y entorno escolar con 10%, problemas familiares con niveles representativos de 68%, y problemas económicos con un 96% como el más relevante.

En la investigación de Delgado (2017) titulado *"Factores influyentes en la deserción estudiantil de estudiantes de secundaria de una institución educativa del distrito de Marmot, 2017"* publicado en el Repositorio de la Universidad

César Vallejo. El objetivo de la investigación fue encontrar los influyentes en el abandono escolar de estudiantes de secundaria. Bajo un estudio no experimental, descriptivo y metodología deductiva e inductiva se halló en los resultados que el grado influyente de Factores que dirijan a los alumnos a desertar clasificó con un nivel medio por el gran número de apoderados (42.7%), se estableció en el nivel medio los Factores influyentes en la Deserción Escolar de estudiantes de secundaria, lo que quiere decir que se nota un regular nivel en el abandono escolar. Concluyendo con la existencia de factores influyentes en el abandono escolar en estudiantes de secundaria de esta I.E.

En el artículo de Henríquez y Escobar (2016) titulado *"Elaboración de un modelo de alerta temprana para detectar estudiantes en peligro de desertar en la Universidad Metropolitana de Ciencias de la Comunicación"*, publicado por la Revista Mexicana de Investigación Educativa. El principal objetivo fue analizar las variables vinculadas con la maniobra de habilidades solicitadas en el ingreso a la Universidad. Para los resultados incluyen las siguientes variables, permitiendo predecir los déficits lingüísticos y matemáticos: resultados en la prueba de selección universitaria en las dos asignaturas, sexo y edad al ingreso a la UMCE. En base a esas variables se clasificó a los estudiantes (valores límite y probabilidades) con el fin de establecer diferentes estrategias de ayuda educativa. Los autores concluyen que estos patrones se recomiendan con fines de detección a tiempo para rastrear el desempeño académico de los ingresantes a la UMCE.

La Data Science según Navarro, et al (2017) es el talento cognitivo a través de datos. Se considera cómo se recopilan los datos, o en otras palabras, cómo se utilizan los datos para generar ideas, tomar decisiones, predecir lo que pasará y/o conocer el pasado/actualidad. La Data Science es un almacén de conocimiento nuevo, cuyos límites aún son difusos y dinámicos, aunque se ha estudiado durante muchos años. (Malik y Sudhakar, 2016). Sus componentes incluyen álgebra lineal, modelado estadístico, visualización, lenguaje

computacional, analizar gráficos, aprendizaje automático, cognición empresarial, almacenar y recuperar datos (Medina, et al, 2020).

Diferenciar entre diferentes tipos de datos es el paso indispensable en la ciencia de datos. Es llamativo comenzar a explorar, emplear métodos de estadística y utilizar herramientas de aprendizaje automático de inmediato. (2, 2017). Aun así, si se desconocen los datos utilizados, los pasos descritos pueden perder mucho tiempo aplicando ejemplares poco eficientes a datos precisos (Meedech, et al, 2016).

Los datos estructurados son los que se pueden asignar al concepto de observación o característica, se pueden distribuir a través de tablas (filas y columnas). Los datos no estructurados, son los que existen como una unidad libre y no corresponden a forma alguna de organización estándar. (Miranda, 2017). En general, los datos estructurados se consideran más maniobrables, más sencillos de procesar y estudiar. Esta propiedad de filas y columnas es más fácil de digerir para el ojo humano y mecánico, pero la relevancia de datos no estructurados es que representan entre el 80% y el 90% de los datos globales. (Ruíz, 2017).

Dado que mucha de la información está en formato libre, se deben usar ciertas técnicas para digerir estos datos, llamado preprocesamiento, que ocurre cuando se utilizan transformaciones para convertir datos no estructurados a contrapartes estructuradas (Sánchez, 2017).

Datos cuantitativos. Este tipo de datos se describen mediante números y se les aplica un tratamiento matemático básico. Datos cualitativos. Los números y las matemáticas básicas no pueden describir estos datos. (Siegel, 2013). La mayoría de estos datos se describen en términos categorizados y lenguaje natural. Los datos cuantitativos se dividen en dos categorías: datos discretos contables y con valores precisos, y datos continuos medibles en un rango infinito valorativo. (Viale, 2014).

Como cualquier otro esfuerzo físico, el proceso de lograr la ciencia de datos se estructura paso a paso preservando la seguridad de los resultados. Luego, para Ozdemir (2016), se describe en detalle cada una de estas etapas.

Elaborar una pregunta llamativa: Aquí se pretende cuestionar si se considera la existencia de información para resolver el problema planteado. (Valdez, 2018). En esta etapa, es indispensable tener conocimiento del contexto en el que se ubica la problemática, ya que los siguientes pasos requerirán los datos adecuados. (Yunia y Urbano, 2018).

Obtenga y prepare los datos: es hora de explorar el entorno buscando datos que se necesitan para dar respuesta a la pregunta establecida. Esta fase puede basarse en un estudio en espacios públicos y privados, complicando un poco la labor. En segundo lugar, los datos se cargan antes de prepararlos. (Subiria, 2019). Explorar datos: trata de la capacidad de distinguir entre tipos de información y posteriormente realizar el modelado. Aquí, el analista se la pasa manipulando e indagando la información y finalmente, tiene una idea bastante clara de lo que dicen los datos. En este paso, los datos se visualizan para localizar cualquier correlación o patrón en las mismas tendencias. (Hossen, 2016).

Modelado de datos: tiene como objetivo utilizar modelos de aprendizaje automático. El aprendizaje automático se refiere y habla de programar codificaciones que ajusten directamente el rendimiento en función de su exposición a la información de los datos. Esto se realiza gracias a un ejemplo con parámetros donde las variables son adaptadas directamente con los distintos aspectos. El aprendizaje automático se puede considerar como una rama de la inteligencia artificial (Mohammed, 2016).

Definición del abandono escolar como aspecto que obstaculiza el avance académico de un estudiante derivado de una ausencia escolar por motivos distintos a la afección. (García, 2014). Autores como Hannon (2017) se refieren a la deserción como el desinterés por continuar sus estudios del alumno por un

largo período de tiempo que hace que el estudiante desertado se niegue a seguirlo.

Para Bean y Metzner (1985), la deserción ocurre al momento en el que un estudiante que se ha matriculado en la escuela deja la escuela por una variedad de razones, incluidas económicas, sociales, culturales y familiares, lo que hace que fracase en su año académico regular. Estos autores intentan agrupar en una sola definición lo que tiene que ver con la deserción escolar porque creen que el alumno deja la escuela sin terminar sin ser trasladado a otro establecimiento.

Casarego (2014), manifiesta que la deserción escolar se debe a la ausencia del sistema educativo y que los apoderados piensan que la escuela no es la única forma de educarse, porque en zonas remotas se educan a través de las costumbres o se hacen la realidad aplicándola a su vida diaria.

Para Cruz (2017), estos factores son clasificados de la siguiente manera: Factores extraescolares: incluyen los que escapan a la influencia directa de la comunidad educativa. Por ejemplo, la falta de fondos en casa que cubran los gastos necesarios en el colegio. Además, los jóvenes desertan la escuela para laburar o buscar empleo, el desinterés de los jóvenes y sus allegados por educarlos, el bajo nivel educativo de sus apoderados, los problemas de la oferta educativa, en particular en las zonas rurales, bajas expectativas educativas, familias numerosas y embarazo y maternidad.

Asimismo, un estudio del INEI (2017) encontró que el embarazo fue motivo del abandono estudiantil de las adolescentes: Esta investigación mostró casi el 75% de las adolescentes, cuya gestación precede al abandono escolar regresa al sistema educativo; mientras que solo entre el 18% y el 30% regresan a sus estudios si su embarazo se ve interrumpido. La adolescente embarazada aminora las posibilidades de estudiar y trabajar ya que muchas se ven obligadas al abandono del sistema educativo, dejando a la estudiante con bajo nivel educativo que no le va a permitir encontrar un trabajo que le deje al menos

cumplir sus necesidades básicas, mientras que la reintegración a sus clases casi no ocurre inmediatamente después de que esta alumbre a su hijo.

En varios partes del Perú, el embarazo precoz forma parte del modelo cultural de la región y, por ende, se acepta en algunas comunidades o ciudades, por ejemplo: Loreto 30,4%; Amazonía 28,1%, Madre de Dios 24,4%, San Martín 24,1% y Ucayali 21,2%. En contraste, los porcentajes más bajos de embarazos adolescentes se encuentran en las regiones de Arequipa (8,4%) y Moquegua (7,0%). (Devasia, et al, 2016).

Factores académicos: vinculados con el servicio de la institución educativa, aspectos educativos y pedagógicos, mala calidad de la enseñanza, poco incentivo de este y desinterés, problemas para aprender, malos resultados, problemas conductuales y repetición. Para la autora Espinoza (2010), las causas que las provocan pueden ser internos a la escuela, notas bajas, problemas conductuales o enseñanza del maestro; asimismo el segundo aspecto sería la economía del estudiante, su familia, su baja probablemente por enfermedad, un embarazo en la adolescencia o las pocas expectativas de los padres en materia de educación. (Dong, 2016).

La Gestión de una Institución Educativa de nivel Básica Regular, bajo la ley N° 28044 en su artículo 79, Ley General de Educación, el Ministerio de Educación es el Órgano del Gobierno nacional cuya finalidad es definir. Dirigir y articular la política de educación, cultura, recreación y deporte. En dicha ley se establece que la Institución Educativa es la primera instancia de gestión del sistema educativo descentralizado, además que el Proyecto Educativo Institucional orienta su gestión y tiene un enfoque inclusivo. Además del PEI, el Plan Anual de Trabajo, Proyecto Curricular de la Institución Educativa y el Reglamento Interno son sus herramientas de gestión. (Fonseca, 2016).

Compromisos de gestión escolar (CGE), la Resolución Viceministerial N° 011-2019.MINEDU en el punto 6.3.1. precisa que los compromisos de gestión son aquellos que promueven y reflejan una gestión adecuada de la I.E., los CGE

son 5 divididos en dos grupos, los dos primeros como resultados y los tres últimos como condiciones para el funcionamiento de la I.E. A continuación, se detallan dichos compromisos: Progreso de los aprendizajes de las y los estudiantes de la institución educativa. Acceso y permanencia de las y los estudiantes en la institución educativa. Denominación de los CGE referidos a condiciones, calendarización y gestión de las condiciones operativas, acompañamiento y monitoreo para la mejora de las prácticas pedagógicas orientadas al logro de aprendizajes previstos en el CNEB, gestión de la convivencia escolar. (Elera, 2016).

La metodología CRISP-DM 1.0 propone un modelo de proceso de minería de datos donde los expertos del área abordan un problema, además brinda una descripción modelo sobre el ciclo de vida de un proyecto estándar de análisis de data, de igual forma que en la ingeniería del software se trabaja con modelos de ciclo de vida para el desarrollo de software, así mismo esta metodología es completa al mantener una aplicación en el entorno de negocio con los resultados. (Fontalvo, 2014).

Para esta investigación se emplea la metodología CRISP-DM 1.0 la cual se detalla bajo un modelo de jerarquía, que confiere una agrupación de actividades especificadas en cuatro etapas abstractas (que van desde la más general hasta la más específica). (Fontalvo, 2014).

En primera instancia se tiene al nivel superior, el proceso de data mining o minería de datos, está organizada en varios pasos; cada una de estas abarca numerosos de segundo nivel. En segunda instancia, está denominada genérica debido a que se dirige a ser lo general posible para abarcar toda la información en general, lo que se requiere es que las actividades genéricas sea la más completa y con estabilidad posible, para que las dos consideren toda la labor de data mining y los aplicativos que existan de esta, lo que quiere decir que el modelo debe ser el apto para ejecuciones imprevistas como métodos nuevos de modelado. (Elías y Molinas, 2016).

El tercer lugar de actividades especializadas, es el espacio para explicar la forma de realizar las actividades genéricas en sucesos específicos, un ejemplo de ello es el segundo nivel en la que puede existir una actividad denominada datos limpios. (Cati, et al, 2016). Este nivel explica la forma en la que esta actividad se distingue en diversos sucesos, por ejemplo, la limpieza de valores numerales ante la limpieza de valores de categoría o si la problemática es el conjunto del modelo predictivo. Describir las etapas y actividades como etapas discretas ejecutadas en orden establecido trata de una continuidad prevista de sucesos.

En la práctica, muchas actividades se pueden ejecutar en diferente orden y usualmente, será indispensable rehacer actividades anteriores y rehacer ciertas actividades. (Fernández, 2019). En la última y cuarta etapa, solicitar la actividad, es un registro de labores y resultados de la data mining. En primera instancia de actividad es organizar según labores establecidas en etapas superiores, pero confiere lo que pasó realmente en un suceso preciso, en vez de lo que pasa generalmente. (Fonseca, 2016).

III. METODOLOGÍA

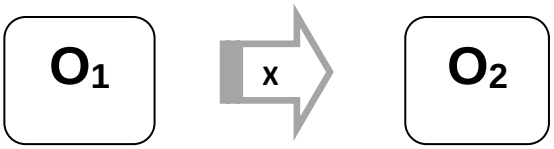
3.1. Tipo y diseño de investigación

El presente trabajo se basó en un enfoque cuantitativo, es decir los datos obtenidos fueron medibles, los datos recolectados a través de una base de datos, fueron procesados y se realizó la interpretación respectiva, la data fue cuantificable porque los datos obtenidos, permitieron determinar la correlatividad entre ambas variables de estudio (Hernández, et al, 2014).

Este trabajo es de tipo aplicada, porque se aplicaron los modelos de Data Science directamente a un problema de la comunidad (Hernández y Mendoza, 2018).

La investigación utilizó el diseño experimental de tipo Pre experimental, porque manipuló las variables en estudio (Baena, 2017).

Según lo descrito líneas arriba, la investigación se realizó bajo el esquema de investigación que se muestra en la figura 1.



Pre-Test Post-Test

Figura 1 Diseño de Investigación con un grupo experimental

***G1*:** Grupo experimental

***O_1* =** Detección de la deserción Académica antes de aplicar los modelos de Data Science

***X* =** Modelos de Data Science

***O_2* =** Detección de la deserción Académica después de aplicar los modelos de Data Science

3.2. Variables y operacionalización

Variable independiente:

Modelos de Data Science

Variable dependiente:

Detección de la Deserción Académica

La matriz de operacionalización de variables se visualiza en el Anexo 1 y la Tabla de indicadores de variable en el anexo 2

3.3. Población, muestra y muestreo

Población: Toda población o universo, se refiere a un grupo de personas o elementos con el que se busca obtener una conclusión acertada. Así mismo, la población estadística, está compuesta por diversos componentes con el que se busca investigar, y deben tener características en común. Dicha población permite aplicar la investigación estadística a fin de concluir con un estudio. (Galeno, 2004).

En la presente investigación se obtuvo la población 1 conformada por los alumnos (220) del nivel secundaria de la Institución educativa N.º 88331.
Como población 2 se contó con el Equipo directivo conformado por 3 docentes coordinadores, 6 personal administrativos y el Director de la Institución Educativa, en total 10 directivos.

Muestra: la muestra es un subgrupo que se desprende de la población en su totalidad, incluye componentes con características en común, a quienes se les aplica de manera directa de la base informativa para el estudio. (Carrasco, 2005).

Para la muestra de la población 1 se tomó a un total de 140 alumnos de la Institución educativa N.º 88331, luego de aplicar la fórmula para cálculo de muestra cómo se visualiza en al Anexo 3.

Para la muestra de la población 2 se estudió a todos los integrantes del equipo directivo, Institución educativa N.º 88331, por ser un grupo pequeño de 10 directivos.

Muestreo: El muestreo aplicado fue el no probabilístico por conveniencia, en el cual los elementos de la muestra en su totalidad conservaron esta opción de escogerse en la recopilación de información, mediante el desarrollo al azar (Hernández et. al, 2014).

El muestreo aplicado a la Población 1 se realizó tomando a los 140 estudiantes de los grados superiores.

Para la Población 2 no se consideró la técnica de muestreo por ser muy pequeña, es decir se estudió a toda la población 2.

Tabla 1: Población 1, muestra y muestreo.

Indicador	Tasa de retención de estudiantes, Tiempo promedio para analizar reportes, % de cumplimiento meta deserción escolar.	Nivel de detección de la deserción académica de los estudiantes
Población	Total de estudiantes de la IE N° 88331 (Información Documental)	220 estudiantes
Muestra	Grupo de estudiantes	140 estudiantes
Muestreo	Para mayor entendimiento de la investigación, se procedió a considerar parte de la población como estudio	Muestreo no probabilístico por conveniencia, tomando a los estudiantes de los grados superiores.
Unidad de análisis	Reportes Académicos de Estudiantes de la IE N° 88331	Estudiantes de la IE N° 88331

Criterios de evaluación	Nivel de detección de la deserción académica de los estudiantes	Estudiantes

Fuente: elaboración propia del autor.

Tabla 2: Población 2, muestra y muestreo.

Indicador	% de aprobación de encuesta de equipo	Nivel de detección de la deserción académica de los estudiantes
Población	Equipo Directivo de la IE N° 88331	10 Directivos
Muestra	Total de Directivos	10 Directivos
Muestreo	Para mayor entendimiento de la investigación, se procedió a considerar toda la población como estudio	
Unidad de análisis	Satisfacción de Equipo Directivo de la IE N° 88331	Equipo Directivo de la IE N° 88331
Criterios de evaluación	Equipo Directivo satisfechos de la IE N° 88331	Equipo Directivo de la IE N° 88331

Fuente: elaboración propia del autor.

3.4. Técnicas e instrumentos de recolección de datos

Técnica de recolección de datos: Se refiere a un conjunto de procesos agrupados con el fin de capturar los datos esperados de un lugar determinado. (Páramo y Arango, 2008).

Para la presente investigación las Técnicas e Instrumentos se emplearon de acuerdo a la Población en estudio.

Población 1: La técnica empleada fue el Análisis Documental, siendo un conjunto de operaciones que buscan representar el contenido y su forma facilitando su consulta. (Clauso, 1993)

La otra técnica empleada para el estudio de la población fue la encuesta, la cual es utilizada frecuentemente porque permite recolectar datos cuantitativos. (Hernández. Baptista y Fernández, 2014)

Instrumento de recolección de datos: Son las que permiten recolectar la información brindada por las técnicas como, formatos de recolección, registros de datos, validados o de elaboración propia (Páramo y Gómez, 2008). El instrumento empleado fue la Ficha de Registro de Datos ya que se recogió datos de los números de estudiantes, tiempo de análisis de reportes académicos y el porcentaje de cumplimiento de meta deserción además se contrastará entre un antes y un después de la implementación de un Modelo Data Science. Otro instrumento fue el cuestionario, ya que se aplicó un cuestionario para medir la satisfacción del Equipo Directivo, dicho instrumento está conformado por la formulación de 12 interrogantes, los cuales determinarán el grado de satisfacción del Equipo directivo con respecto al Modelo de Data Science.

Tabla 3: Técnicas e instrumentos de investigación.

Técnica	Instrumento	Fuente	Informante
Análisis documental	Ficha de Registro de datos	Estudiantes	Área administrativa
Encuesta	Cuestionario	Área administrativa	Equipo Directivo

Fuente: elaboración propia del autor.

Validez

En el presente trabajo se utilizó la validación en juicio de expertos, el cual estuvo compuesto por tres profesionales expertos en la materia con el grado mínimo de Maestría. Según Hernández. Baptista y Fernández (2014) indican que la validez es el valor con el que se cuantifica a la variable que intenta demostrar a través de un instrumento. La validez descrita en la presente investigación se logró aplicando el juicio de expertos donde participaron tres profesionales conocedores del tema; el juicio de expertos se basó por un grupo de personas, donde cada uno de ellos determinan una decisión sobre el instrumento, considerando si es esencial, útil pero prescindible o innecesario; basado en su experticia y su experiencia profesional.

Tabla 4: Expertos que validaron el instrumento e recolección de datos cuantitativos

DNI	Grado Académico, apellido y nombres.	Institución donde labora	Calificación
46663398	*Mg. Johan Max Alexander López Heredia*	Universidad Nacional del Santa	Aplicable
32888444	*Dr. Ricardo Ernesto Izaguirre Diego*	Siderperu - Gerdau	Aplicable
40197616	*Andrés David Epifanía Huerta*	Universidad Católica los Ángeles de Chimbote	Aplicable

Fuente: elaboración propia del autor.

Confiabilidad

Para garantizar la confiabilidad del cuestionario utilizado en la presente investigación se utilizó el Alfa de Cronbach, luego de aplicarse el instrumento a la población 2, conformado por los 10 integrantes del equipo directivo.

Se aplicó el método del Alfa de Cronbach y se trabajó con una muestra piloto de diez personas con atributos similares con la muestra, se obtuvo

un coeficiente de confiabilidad de $\propto = 0.973$, con el que se concluye que el instrumento a emplear es CONFIABLE y CONSISTENTE para medir la variable con 12 ítems. Ver anexo 7.

3.5. Procedimientos

Esta investigación inicia con el permiso de la IE N° 88331 por parte del director, el cual brindó su consentimiento para poder realizar el estudio en la institución educativa, después se procederá a recolectar datos de la tasa de retención de estudiantes actual, tiempo promedio para analizar reportes académicos, el porcentaje de cumplimiento de meta deserción escolar, aplicando los instrumentos de Fichas de Recolección de Datos, también se medirá el porcentaje de satisfacción de equipo directivo aplicando un cuestionario de 12 preguntas.

Posterior a ello, se aplicará los modelos de Data Science, con el objetivo de mejorar todos estos indicadores a fin de tener una mejor detección de la deserción académica de los estudiantes de la institución educativa N° 88331, se aplicará el Post Test con los instrumentos empleados en el Pre Test. Los cuatro instrumentos de Recolección de Datos fueron validados por el juicio de tres expertos como se muestra en el Anexo 3.

Finalmente, luego de la Implementación de los Modelos de Data Science se determinará la influencia que tuvo sobre la Deserción Escolar de la I.E. 88331 a través de la prueba de hipótesis.

3.6. Método de análisis de datos

Para el método de análisis de datos se hará de manera descriptiva e inferencial, en el cual se obtendrán los resultados descriptivos (%, frecuencia, tablas) y de manera inferencial (estadística, validación de hipótesis).

Análisis Descriptivo: Según Valderrama (2015), “el análisis descriptivo permite describir características básicas de los datos, como la media, mediana, asimetría, curtosis, desviar estándar”. Se refiere al análisis de la distribución de frecuencias, medidas de tendencia central y las medidas de la variabilidad relacionado al comportamiento de las variables para procesamiento de datos y plasmarlos en tablas, gráficos que permitan su interpretación.

En la presente investigación se implementará los Modelo de Data Science para lograr la Detección de la Deserción escolar, evaluando la tasa de retención de estudiantes actual, tiempo promedio para analizar reportes académicos, el porcentaje de cumplimiento de meta deserción escolar, y el porcentaje de satisfacción de equipo directivo, se utilizará el Pre Test, para conocer el estado actual de dichos indicadores.

En continuación a dicho análisis luego de dos meses de implementado el Modelo, se determinará la influencia de la propuesta sobre los indicadores planteados.

Análisis Inferencial: Según (Sampieri, 2014), indica, que busca demostrar la hipótesis a fin de generalizar los resultados encontrados de una muestra de la población o universo. La mayoría de veces, los datos son recolectados a partir de una muestra; los resultados de estadística se les menciona estadígrafos” (Sampieri, 2014). En el presente trabajo se utilizó el programa

SPSS que permitió procesar los datos y analizar los estadígrafos resultados con el objetivo de demostrar las hipótesis planteadas.

Para poder validar la hipótesis, se considera una confiabilidad del 95%, es decir el coeficiente Alfa será igual a 5% (= 0.005). H_1

Tabla 5: Hipótesis para la tasa de retención de estudiantes.

INDICADOR **Tasa de retención de estudiantes**	
TREa: Tasa de retención de estudiantes antes de aplicar los modelos de Data Science	TREd: Tasa de retención de estudiantes después de aplicar los modelos de Data Science
HIPÓTESIS	
Nula (H0)	Alterna (H1)
El incremento de la tasa de retención de estudiantes de la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar.	El incremento de la tasa de retención de estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar.
H0: TREd - Tera >= 0	H1: TREd - Tera < 0

Fuente: elaboración propia del autor.

Tabla 6: Hipótesis de tiempo promedio para analizar reportes.

INDICADOR **Tiempo promedio para analizar reportes**	
TPARa: Tiempo promedio para analizar reportes antes de aplicar los modelos de Data Science	TPARd: Tiempo promedio para analizar reportes después de aplicar los modelos de Data Science
HIPÓTESIS	
Nula (H0)	Alterna (H1)
La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.	La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.
H0: TPARd- TPARa >= 0	H1: TPARd- TPARa < 0

Fuente: elaboración propia del autor.

Tabla 7: Hipótesis de porcentaje de cumplimiento meta deserción escolar.

INDICADOR **Porcentaje de cumplimiento meta deserción escolar.**	
PCMDEa: Porcentaje de cumplimiento meta Deserción Escolar antes de aplicar los modelos de Data Science	PCMDEd: Porcentaje de cumplimiento meta Deserción Escolar después de aplicar los modelos de Data Science
HIPÓTESIS	
Nula (H0)	Alterna (H1)
El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.	El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.
H0: PCMDEd - PCMDEa >= 0	H1: PCMDEd - PCMDEa < 0

Fuente: elaboración propia del autor.

Tabla 8: Hipótesis de porcentaje de aprobación de encuesta de equipo.

INDICADOR **Porcentaje de aprobación de encuesta de equipo directivo**	
PAEEDa: Porcentaje de aprobación de encuesta de equipo directivo antes de aplicar los modelos de Data Science	PAEEDd: Porcentaje de aprobación de encuesta de equipo después de aplicar los modelos de Data Science
HIPÓTESIS	
Nula (H0)	Alterna (H1)
El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa N.º 88331, no se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science.	El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa N.º 88331, se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science.
H0: PAEEDd - PAEEDa >= 0	H1: PAEEDd - PAEEDa < 0

Fuente: elaboración propia del autor.

Esta investigación cuenta con una muestra de 10 personas por lo cual se utilizará Shapiro-Wilk, dado que es aplicable en muestras menores a 30. Además las pruebas de normalidad se utilizarán el software SPSS v.25, y dependiendo de los resultados obtenidos en la pruebas de normalidad se podrán emplear la prueba de hipótesis T- Student o Wilcoxon. Finalmente se determinará si la hipótesis planteada se acepta o se rechaza.

3.7. Aspectos éticos

La presente investigación se alinea a las presentes condiciones éticas, descritas en la normativa y los artículos de la Resolución del Consejo Universitario Nº 00126-2017-UCV. En relación al Art.14 de la publicación de las investigaciones, se redactó un permiso que permitió garantizar la originalidad de este proyecto de investigación, basándose en un compromiso ético y moral. Seguidamente el Art.15 de la Política anti plagio, el presente informe fue evaluado utilizando el software turnitin.

Además el Art.16 sobre los Derechos autor, se realizó una declaración de autenticidad y sobre la no realización de algún tipo de plagio y su alineación al artículo 15 de la Resolución del Consejo Universitario Nº 00126-2017-UCV.

En la aplicación del presente proyecto de investigación la Institución Educativa tuvo conocimiento sobre la investigación y los procedimientos realizados en su plantel. En el proceso de recolección de la información se solicitó el permiso de la Institución Educativa el cual figura en el anexo 8, que garantice la veracidad de la investigación.
Finalmente de acuerdo a los estatutos de la Universidad el presente trabajo se someterá al sistema Web Turnitin.

IV. RESULTADOS

4.1. Análisis Descriptivo

La presente investigación obtuvo la implementación de un Modelo de Data Science con la finalidad de mejorar la detección de la deserción escolar en la I.E. 88331, evaluando la Tasa de retención de estudiantes, el Tiempo promedio para analizar reportes, el Porcentaje de cumplimiento meta Deserción Escolar y el Porcentaje de aprobación de encuesta de equipo directivo, para ello, se realizó un pretest, que permitió conocer la necesidad inicial de cada indicador.

En seguida se implementó el Modelo de Data Science, y a través de un seguimiento durante un mes, se obtuvieron los nuevos resultados de cada uno de los indicadores antes mencionados.

Finalmente los indicadores se midieron a través de un post test y el resultado que se obtuvo se presenta en las tablas 9, 10, 11 y 12.

Indicador Tasa de retención de estudiantes.

Se aplicó el modelo de Data Science de regresión logística para modelar la deserción escolar de la I.E. 88331, previamente se aplicó la prueba pretest y finalmente el post test en donde se evaluó la tasa de retención de estudiantes.

Tabla 9: Análisis descriptivo del indicador: tasa de retención de estudiantes en el pretest y postest

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desviación estándar
TERa	10	,74	,90	,8410	,05322
TERd	10	,93	,98	,9650	,01354
N válido (por lista)	10				

Fuente: Base de datos

En la tabla 9, observamos N referido a los años, así mismo se obtuvo la cantidad del pretest que es un mínimo de 0,74 y el máximo de 0,90 de tasa de retención del año escolar, además se tiene la media 0,841 y la desviación estándar de 0,053. En el método del postest se obtuvo un valor mínimo de 0,93 y el máximo de 0,98 de tasa de retención del año escolar, además se tiene la media 0,965 y la desviación estándar de 0,0135.

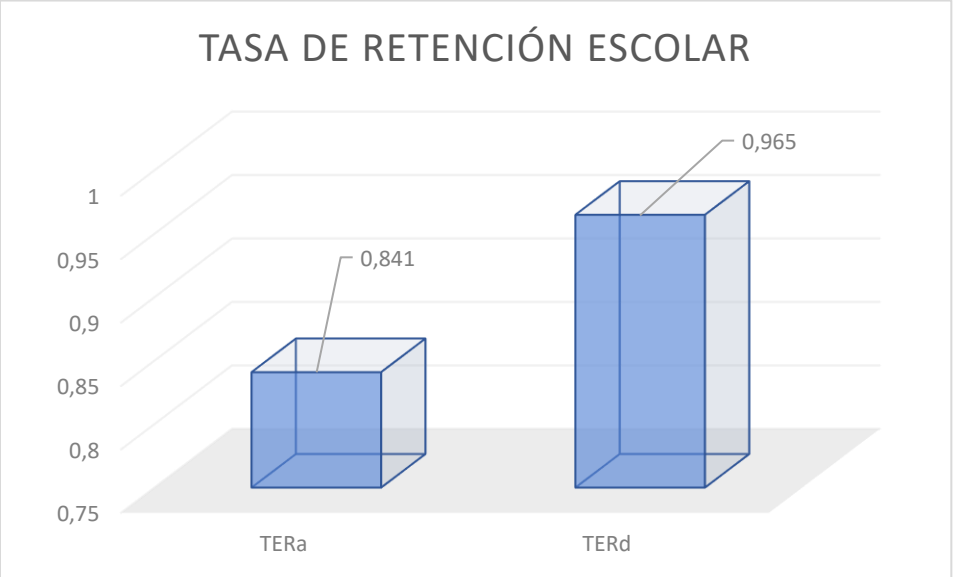


Figura 2: Comparativa del Indicador tasa de retención de estudiantes entre el pretest y postest

Fuente: Base de datos.

Como se observa en la figura 2, se tiene el pretest de 0,841 de tasa de retención de estudiantes antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 0,965 de tasa de retención de estudiantes. En efecto se evidencia que existe una diferencia antes y después de la aplicación del Modelo de Data Science Regresión logística

Indicador de Tiempo promedio para analizar reportes.

Se aplicó el modelo de Data Science de regresión logística para modelar la deserción escolar de la I.E. 88331, previamente se aplicó la prueba pretest y finalmente el post test en donde se evaluó el Tiempo promedio para analizar reportes

Tabla 10: Análisis descriptivo de tiempo promedio para analizar reportes en el pretest y postest

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desviación estándar
TPARa	10	67,00	83,00	73,5000	5,85472
TPARd	10	20,00	25,00	21,3000	2,00278
N válido (por lista)	10				

Fuente: Base de datos

En la tabla 10, observamos N referido a los bimestres, así mismo se obtuvo la cantidad del pretest que es un mínimo de 67 y el máximo de 83 de tiempo promedio para analizar reportes, además se tiene la media 73,5 y la desviación estándar de 5,8547. En el método del postest se obtuvo un valor mínimo de 20 y el máximo de 25 de tiempo promedio para analizar reportes, además se tiene la media 21,30 y la desviación estándar de 2,00278.

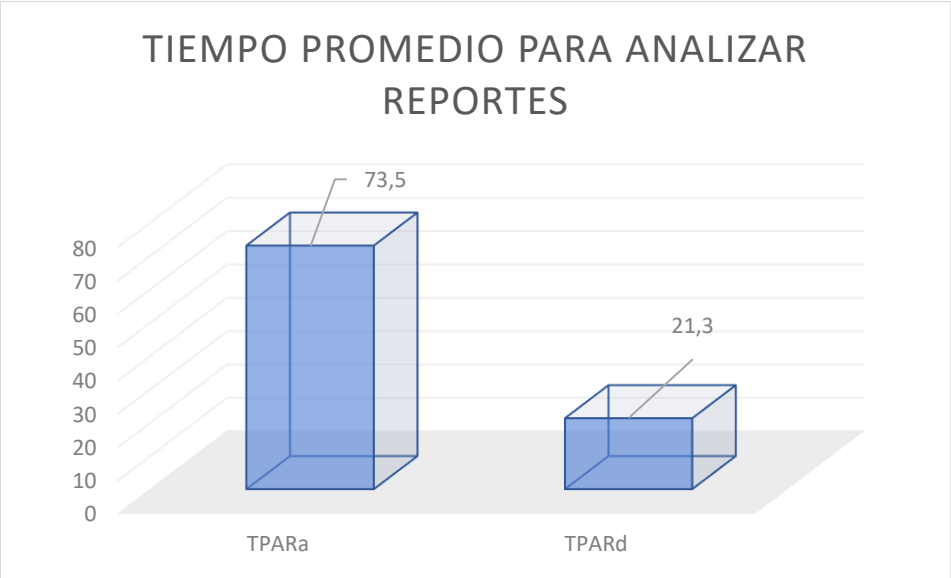


Figura 3: Comparativa del Indicador tiempo promedio para analizar reportes entre el pretest y postest

Fuente: Base de datos

Como se observa en la figura 3, se tiene el pretest de 73 minutos de tiempo promedio para analizar reportes antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 21,30 minutos de tiempo promedio para analizar reportes. En efecto se evidencia que existe una diferencia antes y después de la aplicación del Modelo de Data Science Regresión logística

Indicador Porcentaje de cumplimiento meta deserción escolar.

Se aplicó el modelo de Data Science de regresión logística para modelar la deserción escolar de la I.E. 88331, previamente se aplicó la prueba pretest y finalmente el post test en donde se evaluó el porcentaje de cumplimiento meta deserción escolar.

Tabla 11: Análisis descriptivo del porcentaje de cumplimiento meta deserción escolar en el pretest y postest

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desviación estándar
PCMDEa	4	81,34	93,30	87,9175	5,30474
PCMDEd	4	101,39	105,26	102,6325	1,82681
N válido (por lista)	4				

Fuente: Base de datos.

En la tabla 11, observamos N referido a los bimestres, así mismo se obtuvo la cantidad del pretest que es un mínimo de 81,34 y el máximo de 93,30 de porcentaje de cumplimiento meta deserción escolar, además se tiene la media 87,9175 y la desviación estándar de 5,30474. En el método del postest se obtuvo un valor mínimo de 101,39 y el máximo de 102,6325 de porcentaje de cumplimiento meta deserción escolar, además se tiene la media 102,6325 y la desviación estándar de 1,82681.

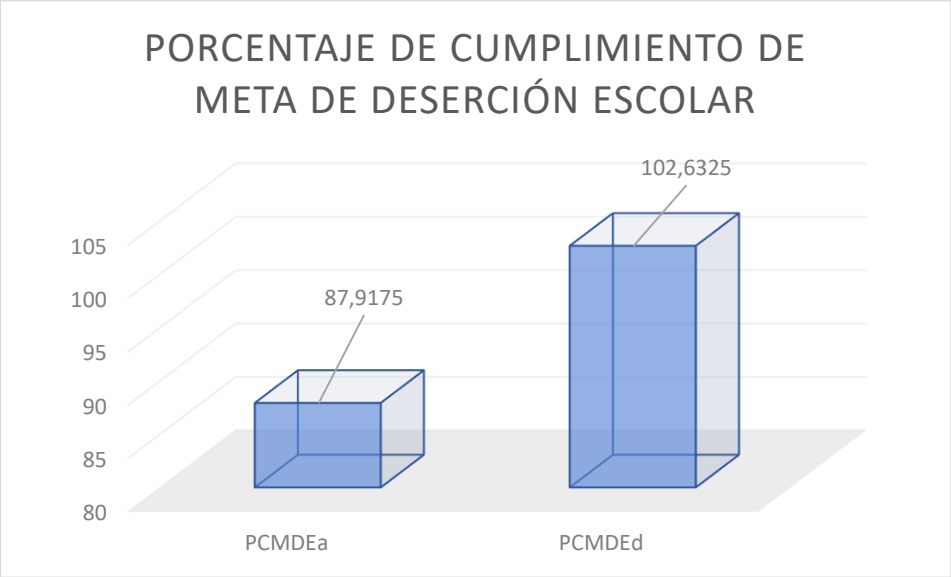


Figura 4: Comparativa del Indicador porcentaje de cumplimiento meta deserción escolar entre el pretest y postest

Fuente: Base de datos

Como se observa en la figura 4, se tiene el pretest de 87,9175 de porcentaje de cumplimiento meta deserción escolar antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 102,6325 de porcentaje de cumplimiento meta deserción escolar. En efecto se evidencia que existe una diferencia antes y después de la aplicación del Modelo de Data Science Regresión logística

Indicador de porcentaje de aprobación de encuesta de equipo directivo

Se aplicó el modelo de Data Science de regresión logística para modelar la deserción escolar de la I.E. 88331, previamente se aplicó la prueba pretest y finalmente el post test en donde se evaluó el porcentaje de aprobación de encuesta de equipo directivo.

Tabla 12: Análisis descriptivo del Porcentaje de aprobación de encuesta de equipo directivo en el pretest y el postest

Estadísticos descriptivos

Rango		Niveles	Pretest n	%	Postest n	%
16	60	Alto	0	0	10	100
13	15	Medio	6	60	0	0
0	12	Bajo	4	40	0	0
Total			**10**	**100**	**10**	**100**

Fuente: Base de datos.

En la tabla 12, observamos que el 60% de directivos están medio satisfechos con la gestión de la deserción escolar en el Pretest, siendo éste el grupo más representativo, además es seguido por el grupo de nivel bajo con un 40% de aprobación, finalmente ningún directivo se encuentra con un nivel alto de satisfacción. En el Postest el más representativo es el nivel de aprobación del equipo directivo logrando

un 100% de nivel alto, además que los niveles medio y bajo no representan ningún directivo. Podemos evidenciar que el Postest el nivel de satisfacción del equipo directivo alcanzó su total porcentaje de aceptación.

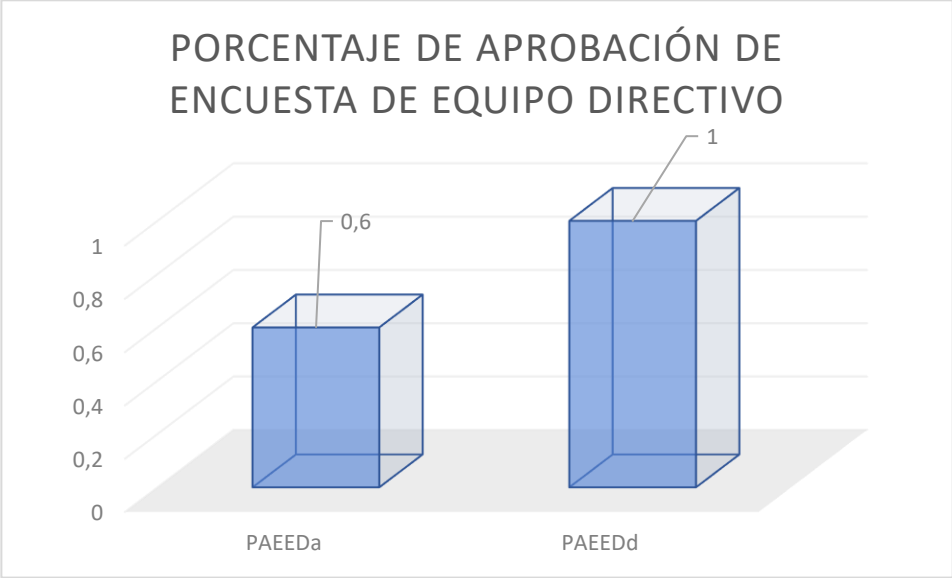


Figura 5: Comparativa del Indicador de aprobación de encuesta de equipo directivo entre el pretest y el postest.

Fuente: Base de datos

Como se observa en la figura 5, se tiene el pretest de 0,6 (60%) del Porcentaje de aprobación de encuesta de equipo directivo antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 1 (100%) del Porcentaje de aprobación de encuesta de equipo directivo. En efecto se evidencia que existe una diferencia antes y después de la aplicación del Modelo de Data Science Regresión logística.

4.2. Análisis Inferencial

4.2.1. Prueba de Hipótesis

Prueba de normalidad de la tasa de retención escolar

Se realizó la prueba de normalidad para la tasa de retención de estudiantes, donde los datos utilizados fueron 10, por ende se

trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%, tal como se observa en la tabla 13 del anexo 9.

Prueba de normalidad del Tiempo promedio para analizar reportes

Se realizó la prueba de normalidad para Tiempo promedio para analizar reportes., donde los datos utilizados fueron 10, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%, tal como se observa en la tabla 14 del anexo 9.

Prueba del porcentaje de cumplimiento meta deserción escolar.

Se realizó la prueba de normalidad para el porcentaje de cumplimiento meta deserción escolar, donde los datos utilizados fueron 4, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%, tal como se observa en la tabla 15 del anexo 9.

Prueba de Porcentaje de aprobación de encuesta de equipo directivo

Se realizó la prueba de normalidad para el porcentaje de aprobación de encuesta de equipo directivo, donde los datos utilizados fueron 10, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%, tal como se observa en la tabla 16 del anexo 9.

Prueba de hipótesis de la tasa de retención escolar

Formulación de la Hipótesis específica 1:

H_1 : El incremento de la tasa de retención de estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar.

Indicador: Tasa de retención escolar.

Hipótesis estadísticas:

Definición de variables:

- **TREa:** Tasa de retención de estudiantes antes de aplicar los modelos de Data Science
- **TREd:** Tasa de retención de estudiantes después de aplicar los modelos de Data Science

Hipótesis Nula H_0 : El incremento de la tasa de retención de estudiantes de la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar.

H_0: TREd - TREa >= 0

El indicador sin un Modelo de Data Science es más óptimo que el indicador afectado por el Modelo de Data Science

Hipótesis Alterna H_a : El incremento de la tasa de retención de estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna la deserción escolar.

H_a: TREd - TREa < 0

El indicador afectado por un Modelo de Data Science es más óptimo que el indicador sin un Modelo de Data Science

Tabla 13: Prueba de hipótesis Wilcoxon aplicado a las puntuaciones del pretest y postest del indicador de tasa de retención escolar.

Rangos

		N	Rango promedio	Suma de rangos
TREd - TREa	**Rangos negativos**	0[a]	,00	,00
	Rangos positivos	10[b]	5,50	55,00
	Empates	0[c]		
	Total	10		

a. TREd < TREa

b. TREd > TREa

c. TREd = TREa

Fuente: Base de datos.

Tabla 14: Estadístico de Prueba de Wilcoxon de tasa de retención escolar en el pretest y postest

Estadísticos de prueba[a]

	TREd - TREa
Z	-2,810[b]
Sig. asintótica (bilateral)	,005

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos negativos.

Fuente: Base de datos.

Observamos en la Tabla 18 que el valor de significancia 0,005 es menor al valor p=0,05, lo cual refleja una evidencia estadística que existe un incremento de la tasa de retención escolar, por lo tanto se rechaza la hipótesis nula y se acepta la hipótesis alterna con 95% de confiabilidad.

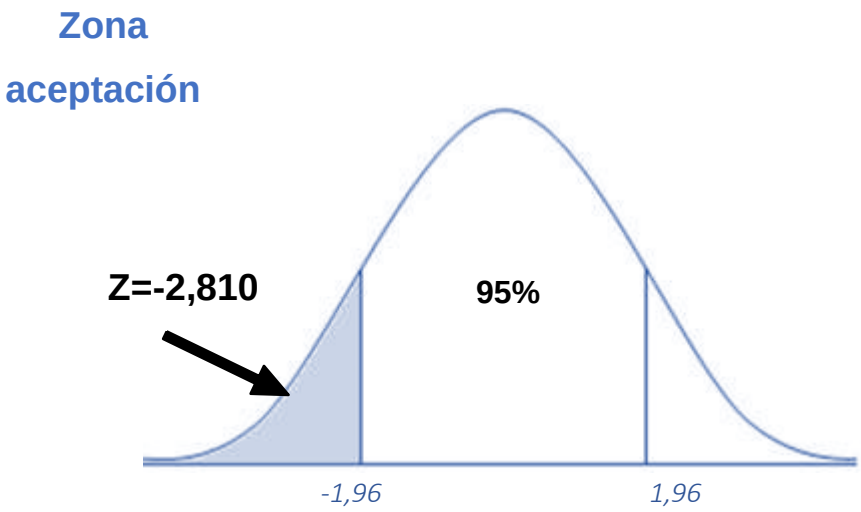


Figura 6: Contrastación de hipótesis del indicador de tasa de retención escolar

Fuente: Base de datos

Prueba de hipótesis del Tiempo promedio para analizar reportes.

Formulación de la Hipótesis específica 2:

H_2 : La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.

Indicador: Tasa de retención escolar.

Hipótesis estadísticas:

Definición de variables:

- **TPARa:** Tiempo promedio para analizar reportes antes de aplicar los modelos de Data Science.
- **TPARd:** Tiempo promedio para analizar reportes después de aplicar los modelos de Data Science.

Hipótesis Nula H_0 : La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.

H_0: TPARd - TPARa >= 0

El indicador sin un Modelo de Data Science es más óptimo que el indicador afectado por el Modelo de Data Science

Hipótesis Alterna H_a : La reducción del tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.

H_a: TPARd - TPARa < 0

El indicador afectado por un Modelo de Data Science es más óptimo que el indicador sin un Modelo de Data Science

Tabla 15: Prueba de hipótesis Wilcoxon aplicado a las puntuaciones del pretest y postest del indicador de Tiempo promedio para analizar reportes.

Rangos

		N	Rango promedio	Suma de rangos
TPARd - TPARa	Rangos negativos	10[a]	5,50	55,00
	Rangos positivos	0[b]	,00	,00
	Empates	0[c]		
	Total	10		

a. TPARd < TPARa

b. TPARd > TPARa

c. TPARd = TPARa

Fuente: Base de datos

Tabla 16: Estadístico de Prueba de Wilcoxon de Tiempo promedio para analizar reportes en el pretest y postest

Estadísticos de prueba[a]

	TPARd - TPARa
Z	-2,821[b]
Sig. asintótica (bilateral)	,005

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos positivos.

Fuente: Base de datos

Observamos en la Tabla 20 que el valor de significancia 0,0056 es menor al valor p=0,05, lo cual refleja una evidencia estadística que existe una disminución del Tiempo promedio para analizar reportes, por lo tanto se rechaza la hipótesis nula y se acepta la hipótesis alterna con 95% de confiabilidad.

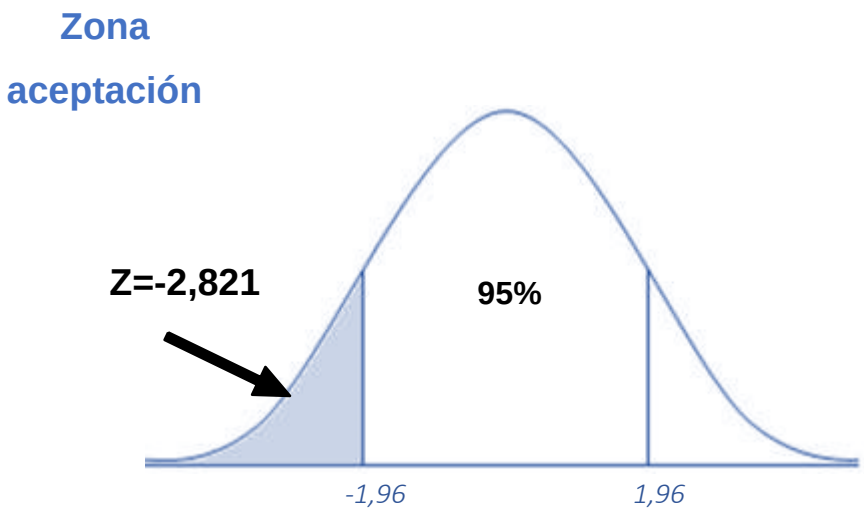


Figura 7: Contrastación de hipótesis del indicador Tiempo promedio para analizar reportes

Fuente: Base de datos

Prueba hipótesis del porcentaje de cumplimiento meta deserción escolar.

Formulación de la Hipótesis específica 3:

H_3 : El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.

Indicador: Porcentaje de cumplimiento meta deserción escolar.

Hipótesis estadísticas:

Definición de variables:

- **PCMDEa:** Porcentaje de cumplimiento meta Deserción Escolar antes de aplicar los modelos de Data Science.

- **PCMDEd:** Porcentaje de cumplimiento meta Deserción Escolar después de aplicar los modelos de Data Science

Hipótesis Nula H_0 : El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa N.º 88331, no se logra utilizando los modelos de Data Science, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.

$$H_0: PCMDEd - PCMDEa >= 0$$

El indicador sin un Modelo de Data Science es más óptimo que el indicador afectado por el Modelo de Data Science

Hipótesis Alterna H_a : El incremento del grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa N.º 88331, se logra utilizando los modelos de Data Science, que

detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.

H_a: PCMDEd - PCMDEa < 0

El indicador afectado por un Modelo de Data Science es más óptimo que el indicador sin un Modelo de Data Science

Tabla 17: Prueba de hipótesis T-Student aplicado a las puntuaciones del pretest y postest del indicador de Porcentaje de cumplimiento meta de deserción escolar

Estadísticas de muestras emparejadas

		Media	N	Desviación estándar	Media de error estándar
Par 1	PCMDEa	87,9175	4	5,30474	2,65237
	PCMDEd	102,6325	4	1,82681	,91340

Fuente: Elaborado con el Software IBM SPSS v22.

Tabla 18: Estadístico de Prueba T-Student de Porcentaje de cumplimiento meta deserción escolar en el pretest y postest

Prueba de muestras emparejadas

		Diferencias emparejadas							
					95% de intervalo de confianza de la diferencia				
		Media	Desviación estándar	Media de error estándar	Inferior	Superior	t	gl	Sig. (bilateral)
Par 1	PCMDEa - PCMDEd	-14,71500	3,92379	1,96190	-20,95863	-8,47137	-7,500	3	,005

Fuente: Elaborado con el Software IBM SPSS v22.

Observamos en la Tabla 22 que el valor de significancia 0,005 es menor al valor p=0,05, lo cual refleja una evidencia estadística que existe un incremento de porcentaje de cumplimiento meta deserción escolar. Por lo tanto se rechaza la hipótesis nula y se acepta la hipótesis alterna con 95% de confiabilidad.

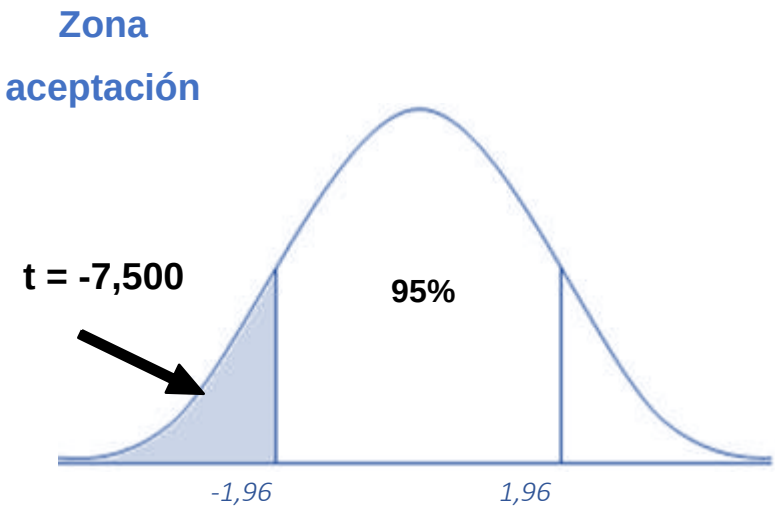


Figura 8: Contrastación de hipótesis del indicador Porcentaje de cumplimiento meta deserción escolar

Fuente: Elaboración propia

Prueba de hipótesis del porcentaje de aprobación de encuesta de equipo directivo

Formulación de la Hipótesis específica 4:

H_4 : El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa N.º 88331, se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science.

Indicador: Porcentaje de aprobación de encuesta de equipo directivo

Hipótesis estadísticas:

Definición de variables:

- **PAEEDa:** Porcentaje de aprobación de encuesta de equipo directivo antes de aplicar los modelos de Data Science.

- **PAEEDd:** Porcentaje de aprobación de encuesta de equipo después de aplicar los modelos de Data Science.

Hipótesis Nula H_0 : El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa N.º 88331, no se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science.

H_0: PAEEDd - PAEEDa >= 0

El indicador sin un Modelo de Data Science es más óptimo que el indicador afectado por el Modelo de Data Science

Hipótesis Alterna H_a : El incremento del grado de satisfacción del Equipo Directivo de la Institución educativa N.º 88331, se logra con la oportuna retención de los estudiantes utilizando los modelos de Data Science..

H_a: PAEEDd - PAEEDa < 0

El indicador afectado por un Modelo de Data Science es más óptimo que el indicador sin un Modelo de Data Science

Tabla 19: Prueba de hipótesis Wilcoxon aplicado a las puntuaciones del pretest y postest del indicador de porcentaje de aprobación de encuesta de equipo directivo.

Rangos

		N	Rango promedio	Suma de rangos
PAEEd - PAEEa	**Rangos negativos**	0[a]	,00	,00
	Rangos positivos	10[b]	5,50	55,00
	Empates	0[c]		
	Total	10		

a. PAEEd < PAEEa

b. PAEEd > PAEEa

c. PAEEd = PAEEa

Fuente: Elaborado con el Software IBM SPSS v22.

Tabla 20: Estadístico de Prueba de Wilcoxon de Porcentaje de aprobación de encuesta de equipo directivo en el pretest y postest

Estadísticos de prueba[a]

	PAEEd - PAEEa
Z	-2,805[b]
Sig. asintótica (bilateral)	,005

a. Prueba de rangos con signo de Wilcoxon

b. Se basa en rangos negativos.

Fuente: Elaborado con el Software IBM SPSS v22.

Observamos en la Tabla 24 que el valor de significancia 0,005 es menor al valor p=0,05, lo cual refleja una evidencia estadística que existe un incremento del Porcentaje de aprobación de encuesta de equipo directivo, por lo tanto se rechaza la hipótesis nula y se acepta la hipótesis alterna con 95% de confiabilidad.

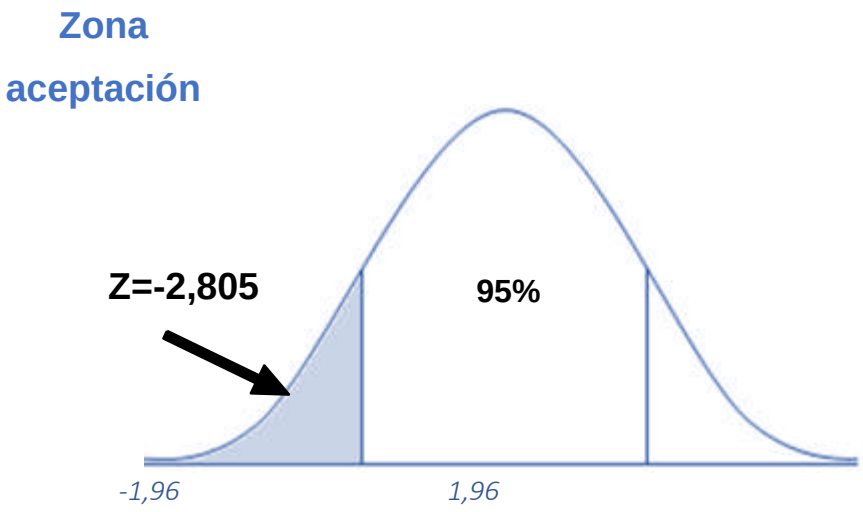


Figura 9: Contrastación de hipótesis del indicador Porcentaje de aprobación de encuesta de equipo directivo

Fuente: Elaboración propia

V. DISCUSIÓN

En relación a los resultados obtenidos en la presente investigación realizada muestran los cambios mostrados en los cuatro indicadores perteneciente a la variable dependiente Detección de deserción escolar, después de la implementación de la variable independiente Modelos Data Science en la Institución Educativa 88331, Chimbote 2021.

El indicador Tasa de retención escolar, en el análisis descriptivo se estudió 10 datos realizados. Además se observa el comportamiento en la Figura 2 del indicador Tasa de retención escolar antes y después de aplicar del Modelo de Data Science Regresión logística en relación a los datos obtenidos mediante las Fichas de registro, por ende, se logró incrementar la Tasa de retención escolar con una mejora de 84,1 % a un 95,5% anual. Adicionalmente, los datos descriptivos del indicador Tasa de retención escolar se observan en la Tabla 9.

Con respecto al análisis inferencial se realizó la prueba de normalidad obteniéndose un valor p mayor a 0,05, el cual fue de 0,216 en el pretest; mientras en el postest resultó un valor p menor a 0,05, un valor de 0,001, lo que significa que los datos no siguen una distribución normal, por lo tanto se trabajó el método no paramétrico de rango de Wilcoxon para la demostración de la Hipótesis, obteniendo un p-valor de 0,005, el cual es menor al valor alfa de 0,05. Por lo tanto se rechazó la Hipótesis nula H_0 y se aceptó la Hipótesis alterna H_1. Podemos concluir que el Modelo de Data Science Regresión logística mejora significativamente la Tasa de retención escolar en la detección de la deserción escolar en la Institución Educativa 88331, Chimbote 2021.

Estos resultados se relacionan con los antecedentes siguiente: Vinueza (2021) en la investigación denominada "*Diseño de modelo matemático para calcular la deserción estudiantil a través de técnicas de análisis multivariado en cualquier institución de educación superior tecnológica*", concluye que el modelo matemático estima la deserción estudiantil a través de métodos analíticos multivariados en un instituto superior tecnológico, en dicha investigación se analizó 849 estudiantes matriculados entre los años 2018 y 2020, se simuló el

modelo de regresión logística demostrándose la predicción de deserción en un 83% sobre los datos de entrenamiento y el 79% como resultado del testeo, con ello se puede predecir los posibles desertores, aumentando significativamente la tasa de retención estudiantil.

También encontramos el artículo de Sifuentes (2018) denominado "*Modelos predictivos de la deserción estudiantil en una universidad privada peruana*" que fue publicado por la Revista Industrial Data, donde se aplicó patrones predictivos en algunas áreas de alta criticidad que permiten determinar qué estudiantes son potenciales a desertar. Como resultado de este estudio descriptivo, se encontró en los resultados que los patrones de predicción permiten disminuir entre un 25 % y 40 % la tasa de desaprobados, teniendo como una variable de mayor predicción la calificación obtenida en matemáticas o comunicación al terminar el año académico. Se concluye que los modelos de predicción permiten conocer la aprobación de los curso y esto contribuye significativamente a la permanencia estudiantil por ende a mejorar la tasa de retención.

El segundo indicador tiempo promedio para analizar reportes, en el análisis descriptivo se estudió 10 datos realizados. Además se observa el comportamiento en la Figura 3 del indicador tiempo promedio para analizar reportes antes y después de aplicar del Modelo de Data Science Regresión logística en relación a los datos obtenidos mediante las Fichas de registro, por ende, se logró disminuir el tiempo promedio para analizar reportes escolar con una reducción de 73,5 % a un 21,3% por bimestre. Adicionalmente, los datos descriptivos del indicador tiempo promedio para analizar reportes se observan en la Tabla 10.

Con respecto al análisis inferencial se realizó la prueba de normalidad obteniéndose un valor p mayor a 0,05, el cual fue de 0,086 en el pretest; mientras en el postest resultó un valor p menor a 0,05, un p-valor de 0, lo que significa que los datos no siguen una distribución normal, por lo tanto se trabajó el método no paramétrico de rango de Wilcoxon para la demostración de la

Hipótesis, obteniendo un p-valor de 0,005, el cual es menor al valor alfa de 0,05. Por lo tanto se rechazó la Hipótesis nula H_0 y se aceptó la Hipótesis alterna H_a. Podemos concluir que el Modelo de Data Science Regresión logística reduce significativamente el tiempo promedio para analizar reportes en la detección de la deserción escolar en la Institución Educativa 88331, Chimbote 2021.

Podemos mencionar la tesis de Cevallos & Barahona (2021), denominada "*Modelo que automatice el procedimiento de predicción del abandono de estudios universitarios en estudiantes del primer año de estudio*" tuvo como contribución reducir la tasa de deserción universitaria, aplicando modelos predictivos, que permitan la detección temprana de estudiantes con posibilidades de interrupción o deserción escolar, brindando a los gestores académicos de las Instituciones educativas mayor claridad y oportunidades para tomar decisiones tempranas ante esta problemática, esta investigación tiene como resultado un modelo que facilita el acceso a la información, permitiendo a los responsables de la gestión académica un manejo en tiempo real de la información, lo cual reduce el tiempo de para analizar reportes de deserción estudiantil.

El tercer indicador Porcentaje de cumplimiento meta deserción escolar, en el análisis descriptivo se estudió 4 datos realizados. Además se observa el comportamiento en la Figura 4 del indicador Porcentaje de cumplimiento meta deserción escolar antes y después de aplicar el Modelo de Data Science Regresión logística en relación a los datos obtenidos mediante las Fichas de registro, por ende, se logró mejorar el Porcentaje de cumplimiento meta deserción escolar con una mejora de 87,9175 % a un 102,6325% por anual. Adicionalmente, los datos descriptivos del indicador Porcentaje de cumplimiento meta deserción escolar se observan en la Tabla 11.

Con respecto al análisis inferencial se realizó la prueba de normalidad obteniéndose un valor p mayor a 0,05, el cual fue de 0,798 en el pretest; además en el postest resultó también un valor p mayor a 0,05, un p-valor de 0,116, lo que significa que los datos siguen una distribución normal, por lo tanto

se trabajó la prueba paramétrica de T-Student para la demostración de la Hipótesis, obteniendo un p-valor de 0,005, el cual es menor al valor alfa de 0,05. Por lo tanto se rechazó la Hipótesis nula H_0 y se aceptó la Hipótesis alterna H_a. Podemos concluir que el Modelo de Data Science Regresión logística mejora significativamente el porcentaje de cumplimiento meta deserción escolar al lograr la detección temprana de la deserción escolar e implementar soluciones inmediatas para abordar el problema, en la Institución Educativa 88331, Chimbote 2021.

Podemos mencionar el artículo de Henríquez y Escobar (2016), denominada "*Elaboración de modelo de alerta temprana que permita detectar estudiantes en peligro de desertar en la Universidad Metropolitana de Ciencias de la Comunicación*" cuyos resultados mostraron dos modelos predictivos con un buen ajuste, además predicen la probabilidad de éxito del estudiante en cada prueba; esto permite conocer de manera prematuro las posibles debilidades en lenguaje o matemática. Este modelo permite detectar a los estudiantes con problemas de rendimiento académico de manera oportuna y garantizar un acompañamiento temprano en su proceso educacional.

Se midieron las capacidades predictivas por medio del análisis de la curva roc, obteniéndose valores en sus áreas de al menos 0.787 en efecto que son significativas.

Así mismo, se lograron resultados de clasificación correcta, especificidad y sensibilidad de al menos un 70%. Concluyendo que los modelos predictivos son recomendables para fines de alerta temprana que permita monitorear el rendimiento académico de los alumnos ingresantes.

Finalmente el cuarto indicador fue el porcentaje de aprobación de encuesta de equipo directivo, en el análisis descriptivo se estudió 10 datos recopilados con la técnica de entrevista. Además se observa el comportamiento en la Figura 5 del indicador Porcentaje de aprobación de encuesta de equipo directivo antes y después de aplicar el Modelo de Data Science Regresión logística en relación

a los datos obtenidos mediante un cuestionario, por ende, se logró mejorar el Porcentaje de aprobación de encuesta de equipo directivo con una mejora de 60 % a un 100% por anual. Adicionalmente, los datos descriptivos del indicador Porcentaje de aprobación de encuesta de equipo directivo se observan en la Tabla 12.

Con respecto al análisis inferencial se realizó la prueba de normalidad obteniéndose un valor p mayor a 0,05, el cual fue de 0.73 en el postest ; además en el pretest resultó un valor p menor a 0,05, un p-valor de 0,049 , lo que significa que los datos no siguen una distribución normal, por lo tanto se trabajó la prueba no paramétrica Wilcoxon para la demostración de la Hipótesis, obteniendo un p-valor de 0,005, el cual es menor al valor alfa de 0,05. Por lo tanto se rechazó la Hipótesis nula H_0 y se aceptó la Hipótesis alterna H_a.

Podemos concluir que el Modelo de Data Science Regresión logística mejora significativamente el porcentaje de aprobación de encuesta de equipo directivo al lograr la detección temprana de la deserción escolar e implementar soluciones inmediatas para abordar el problema, en la Institución Educativa 88331, Chimbote 2021.

Podemos mencionar la tesis de Vinueza y Loza (2021), denominada "*Diseño de modelo matemático para la estimación de la deserción estudiantil a través de técnicas de análisis multivariado en una Institución Educativa Superior Tecnológica*" el implementó un modelo de regresión Logística predecir la deserción de estudiantes en el Instituto Superior Tecnológico Luis Martínez Agronómico. Los resultados permitieron estimar la deserción estudiantil, además el modelo de regresión logística clasificó con una precisión correcta de 83% en la fase de entrenamiento y un 79% en la fase de Testeo.

En el mismo estudio se concluyó que el modelo de regresión logística tuvo mayor precisión con respecto al modelo de árbol de decisión, es decir que, el mejor rendimiento es del modelo de regresión logística fue de 88% comparado con el modelo árbol de decisión 74%.

En relación a la Metodología empleada, es el diseño pre experimental de tipo aplicada, donde se aplicó el Modelo de Data Science Regresión Logística, evidenciando con el cumplimiento de las cuatro hipótesis específicas en el grupo de muestra de la población 1 conformada por los estudiantes de la Institución Educativa 88331 y a toda la población 2 conformada por el Equipo Directivo de la Institución Educativa 88331.

La metodología empleada es el modelo CRISP-DM el cual consta de 6 etapas empezando por el entendimiento del negocio, seguido de la comprensión de los datos, preparación de los datos, el modelado, la evaluación y finalmente la distribución del modelo. (Fontalvo, 2014).

VI. CONCLUSIONES

A continuación se observan los resultados sobre la aplicación del Modelo de Data Science Regresión Logística en la Institución educativa 88331.

1. Se logró incrementar la tasa de retención de estudiantes de la Institución educativa Nº 88331, mediante la implementación del modelo de Data Science Regresión Logística, alcanzando incrementar la Tasa de retención escolar con una mejora de 84,1 % a un 95,5% en el año académico.

2. Se logró reducir el tiempo promedio para analizar reportes Académicos de los estudiantes de la Institución educativa Nº 88331, se obtuvo en el pretest 73 minutos de tiempo promedio para analizar reportes antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 21,30 minutos de tiempo promedio para analizar reportes, mediante la implementación del modelo de Data Science Regresión Logística, que permiten aprovechar los datos almacenados garantizando reportes académicos producto de la inferencia de los datos.

3. Se logró incrementar el grado de cumplimiento de la meta del indicador de deserción escolar en la Institución educativa Nº 88331, con un pretest de 87,9175 de porcentaje de cumplimiento meta deserción escolar antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 102,6325 de porcentaje de cumplimiento meta deserción escolar, que detectan de manera oportuna a los posibles desertores, para su inmediato control y reincorporación.

4. Se logró incrementar el grado de satisfacción del Equipo Directivo de la Institución educativa Nº 88331, con un pretest de 0,6 (60%) del Porcentaje de aprobación de encuesta de equipo directivo antes de la aplicación del Modelo de Data Science Regresión logística, y posterior a dicha implementación se logró un postest de 1(100%) del Porcentaje de aprobación de encuesta de equipo directivo con la oportuna retención de los estudiantes utilizando los modelos de Data Science.

VII. RECOMENDACIONES

1. Es recomendable cargar el modelo Regresión Logística a una plataforma Cloud, como lo es Google Colab para poder desplegarlo en la Institución educativa Nº 88331, dicha plataforma provee la capacidad física necesaria para ejecutar el modelo en óptimas condiciones de Hardware.

2. Es importante simular el modelo adicionando variables exógenas como si se infectó o no del COVID 19, o si algún familiar directo fue infectado y evaluar su impacto con la deserción de los infectados.

3. Es recomendable actualizar el modelo anualmente con un personal especializado, para garantizar su calidad.

REFERENCIAS

ARCE, M.E., CRESPO, B. y ÁLVAREZ Míguez, C. (2015). Higher Education Drop-Out in SpainParticular Case of Universities in Galicia. International Education Studies, 8(5), 247-264.

ARTEAGA González, Á., Torres Ávila, K., & López Cardona, L. (2 de 2018). Variables asociadas al riesgo de deserción y su relación con los beneficios y servicios ofrecidos por Bienestar al Aprendiz en el Centro de Servicios Financieros. Revista Finnova, 2(4)

ASIF, R., MERCERON, A., ALI, S., & HAIDER, N. (10 de 2017). Analyzing undergraduate students' performance using educational data mining. Computers & Education, 113.

BAENA, G. (2017). Metodología de la investigación. 3° ed. México: Mcgrawhill.

BEAN, J., & METZNER, B. (12 de 1985). A Conceptual Model of Nontraditional Undergraduate Student Attrition. Review of Educational Research, 55(4).

BOX, G. E., JENKINS, G. M., REINSEL, G. C., y LJUNG, G. M. (2015). Time series analysis: forecasting and control. New Jersey: John Wiley & Sons Inc.

CABANILLAS, B. (2017). Factores que influyen en la deserción escolar de los estudiantes del nivel secundario de la red educativa Eduardo Villanueva del Distrito Eduardo Villanueva De La Provincia De San Marcos: 2011-2013. Tesis para optar el grado académico de maestro en ciencias. Universidad Nacional de Cajamarca – Perú. Disponible en: https://repositorio.unc.edu.pe/bitstream/handle/UNC/2169/FACTORES%20QUE%20INFLUYEN%20EN%20LA%20DESERCI%C3%93N%20ESCOLAR%20DE%20LOS%20ESTUDIANTES%20DEL%20NIVEL%20SECUNDARIO%20DE%20LA%20RED%20E.pdf?sequence=1&isAllowed=y

CAMBRUZZI, W., RIGO, S. y BARBOSA, J. (2015). Dropout Prediction and Reduction in Distance Education Courses with the Learning Analytics Multitrail Approach, Journal of Universal Computer Science, 21(1), 23-47.

CATI, K. KETHUDA, O. y BELGIN, Y. (2016). Positioning Strategies of Universities: An Investigation on Universities in Istanbul. Article. Education and Sience. Recuperado de https://www.researchgate.net/publication/299404030_Positioning_Strategies_of_Universities_An_Investigation_on_Universities_in_Istanbul

CARRASCO, S. (2005). Metodología de la investigación científica. 1° ed. Perú: San Marcos

CASAREGO, G. (2014). Diseño de estrategias de retención para disminuir la deserción escolar de estudiantes de grado sexto del Instituto Politécnico de Bucaramanga. Ibagué: Universidad de Tolimam

CEVALLOS, E. y BARAHONA, C. (2021). Modelo para automatizar el proceso de predicción de la deserción en estudiantes universitarios en el primer año de estudio. Tesis para obtener título profesional de ingeniero de sistemas de información. Universidad Peruana de Ciencias Aplicadas – Perú. Disponible en: https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/656740/Cevallos_ME.pdf?sequence=3&isAllowed=y

CRUZ, E. (2017). Deserción escolar, factores determinantes de la institución educativa. San Nicolas de la Garza: Universidad de Nuevo León.

DELGADO, D. (2017). Factores que influyen en la deserción escolar de los alumnos del nivel secundario de una institución educativa del distrito de Marmot, 2017. Tesis para optar al grado académico de magíster en ciencias de comunicaciones. Universidad César Vallejo – Perú. Disponible en: https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/28323/delgado_os.pdf?sequence=1&isAllowed=y

DEVASIA, T., VINUSHREE T P, & HEGDE, V. (3 de 2016). Prediction of students performance using Educational Data Mining. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).

DONG M. (20216). Real-time residential-side joint energy storage management and load scheduling with renewable integration. IEEE Transactions on Smart Grid, 2016, vol. 9, no 1, p. 283-298.

ELERA, E (2016). Factores socioeconómicos que influyen en deserción estudiantil. Tesis maestría. Jaén. Universidad Nacional de Cajamarca.

ELÍAS, R y MOLINAS, J. (2016). La deserción escolar de adolescentes en Paraguay. Asunción: Universidad de Paraguay

FERNÁNDEZ, E., HOLANDA, M., VICTORINO, M., BORGES, V., CARVALHO, R., & ERVEN, G. (1 de 2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research, 94

FONSECA, S., & NAMEN, A. (3 de 2016). Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educação em Revista, 32(1)

FONTALVO Cerpa, W. (8 de 2014). Análisis comparativo entre las características más relevantes de deserción estudiantil en el programa de Ingeniería Industrial de la Universidad Autónoma del Caribe. Estudiantes activos en el periodo 2013-01 y desertores académicos de los periodos 2011-01. Escenarios, 12(1).

GALENO, B. (2004). Enfoque cuantitativo. Recuperado de https://www.slideshare.net/marypalma16/enfoques-de-investigacin-95626014

GARCÍA, A. (2014). Rendimiento académico y abandono universitario modelos, resultados y alcances de la producción académica en la Argentina.

HANNON, M. (2017). Examining Shifts In Institutional in the Evolving Iris Higher Education System. (Tesis doctoral). University of Bath School of Management. USA. Recuperado de https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=1349110

HERNÁNDEZ R, FERNÁNDEZ C y BAPTISTA M. (2014). Metodología de la Investigación. Quinta edición. México D.F.: Miembro de la Cámara Nacional de la Industria Editorial Mexicana, 2014. 613 pp.

HERNÁNDEZ, R. y MENDOZA, C (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta. México: Editorial Mc Graw Hill Education

HENRÍQUEZ, N y ESCOBAR, T. (2016). Construcción de un modelo de alerta temprana para la detección de estudiantes en riesgo de deserción de la Universidad Metropolitana de Ciencias de la Comunicación. Revista Mexicana de Investigación Educativa. Vol. 21(71) pp. 1221-1248. Disponible en: https://www.researchgate.net/publication/317449527_Construccion_de_un_modelo_de_alerta_temprana_para_la_deteccion_de_estudiantes_en_riesgo_de_desercion_de_la_Universidad_Metropolitana_de_Ciencias_de_la_Educacion ISSN: 0879-4323

HOSSEN A. (2016). An inventory model with price and time dependent demand with fuzzy valued inventory costs under inflation. Ann. Pure Appl. Math, 2016, vol. 11, no 2, p. 21-32.

INSTITUTO NACIONAL DE ESTADÍSTICA E INFORMÁTICA (2017). Recuperado de http://www.inei.gob.pe/biblionei/bancopub/Est/Lib0038/N10/cuadc022-027.HTML. 27 diciembre 2017

KAKOULLI E. (2017). A distributed file system with tiered storage management. En Proceedings of the 2017 ACM International Conference on Management of Data. 2017, vol. 25, no 3, p. 65-78.

MALIK, A. y SUDHAKAR, B. (2016). Brand positioning constructs and indicators for measurement of consumer´s positive psychology toward brands. Indian Journal of Positive Psychology, 7(1), 124-126. Recuperado de: http://search.proquest.com/docview/1788740643?accountid=37408

MARTELO, R; HERRERA, K y VILLABONA, N. (2017). Strategies to reduce university desertion through time series and multipol. Revista Espacios. Vol. 38 (45) pp. 1- 25. Disponible en: https://www.revistaespacios.com/a17v38n45/a17v38n45p25.pdf ISSN: 0798-1015

MEDINA, E., CHUNGA, C., ARMAS-Aguirre, J., & GRANDON, E. (6 de 2020). Predictive model to reduce the dropout rate of university students in Perú: Bayesian Networks vs. Decision Trees. 2020 15th Iberian Conference on Information Systems and Technologies (CISTI).

MEEDECH P., IAM-On N. y BOONGOEN T. (2016) Prediction of Student Dropout Using Personal Profile and Data Mining Approach. En Lavangnananda, K., Phon-Amnuaisuk, S., Engchuan, W. y Chan J. (eds.), Intelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization, vol 5. (pp. 143-155). Cham: Springer.

MIRANDA, M., & GUZMÁN, J. (2017). Análisis de la Deserción de Estudiantes Universitarios usando Técnicas de Minería de Datos. Formación universitaria, 10(3).

MOHAMMED O. (2016). Hybrid energy storage management in ship power systems with multiple pulsed loads. Electric Power Systems Research, 2016, vol. 141, p. 50-62.

NAVARRO, E. JIMÉNEZ, E. RAPPOPORT, S. y THOILLIEZ, M. (2017). Fundamentos de la investigación y la innovación educativa. España: Unir Editorial.

ORIHUELA, G. (2019). Aplicación de Data Science para la Predicción del Rendimiento Académico de los Estudiantes de la Facultad de Ingeniería de Sistemas de la Universidad Nacional del Centro del Perú. Tesis para optar al título profesional de ingeniero de sistemas. Universidad Nacional del Centro del Perú. Disponible en: http://repositorio.uncp.edu.pe/bitstream/handle/20.500.12894/5837/TESIS.pdf?sequence=1&isAllowed=y

PÁRAMO, P. y ARANGO, M. (2008). La investigación de las ciencias sociales. Bogotá: Universidad piloto de Colombia, Net educativa.

RAMÍREZ, P. y GRANDÓN, E. (2018). Prediction of Student Dropout in a Chilean Public University through Classification based on Decision Trees with Optimized Parameters. Revista SciELO, Vol. 11(3) pp. 1 – 12. Disponible en: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-50062018000300003&lng=pt&nrm=i.p&tlng=es ISSN: 0718-5006

RUIZ Larrocha, E. (2017). Nuevas tendencias en los sistemas de información. Madrid: Editorial Universitaria Ramón Areces.

SÁNCHEZ-Hernández, G., BARBOZA-Palomino, M., & CASTILLA-Cabello, H. (5 de 2017). Análisis de la deserción y los factores asociados a la permanencia estudiantil en una universidad peruana. Actualidades Pedagógicas (69).

SIEGEL, E. (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (Vol. 10). Hoboken: Wiley.

SIFUENTES, O. (2018). Predictive models of student desertion at a private Peruvian university. Revista Producción y Gestión. Vol. 21 (2) pp. 47 – 62. Disponible en: https://www.redalyc.org/jatsRepo/816/81658967008/html/index.html ISSN: 0983-4344

SUBIRIA, J y RAMÍREZ, A. (2019). Cómo Investigar en Educación. Cooperativa Editorial “Magisterio”. 1ra Edición.

VALDEZ, E. (10 de mayo de 2018). http://www.scielo.org.mx. Recuperado el 29 de julio de 2021, de http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1607-40412008000100007

VIALE Tudela, H. (12 de 2014). UNA APROXIMACIÓN TEÓRICA A LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA. Revista Digital de Investigación en Docencia Universitaria (1).

VINUEZA, C. (2021). Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior tecnológica. Tesis para optar el grado académico de magíster en matemática aplicada. Universidad Técnica de Ambato – Ecuador. Disponible en: https://repositorio.uta.edu.ec/bitstream/123456789/32219/1/t1765mma.pdf

YUNIA. J. y URBANO, C. (2018). Mapas y Herramientas para conocer la escuela: Investigación Etnográfica e Investigación – Acción. Madrid: “Brujas” – 3ra Edición.

ANEXOS

ANEXO 1: Matriz de operacionalización de las variables

Variables de estudio	Definición conceptual	Definición operacional	Indicadores	Escala de medición
Modelos de Data Science	La Data Science es el talento cognitivo a través de datos. Se considera cómo se recopilan los datos, o en otras palabras, cómo se utilizan los datos para generar ideas, tomar decisiones, predecir lo que pasará y/o conocer el pasado/actualidad. (Navarro, et al, 2017).	Los modelos de Data Science determinarán los factores que influyen en la deserción escolar. Logrando detectar de manera temprana los indicadores de mayor influencia.	Porcentaje de aprobación de encuesta de Satisfacción de equipo directivo (PAESED)	Razón
			Tiempo promedio para analizar reportes(TPAR)	Razón
Deserción Académica	Es un problema que afecta a la sociedad y se encuentra generalizado, donde se da por diversos factores culturales, sociales, familiares, económicos, etc. siendo un fenómeno no solo de los países en desarrollo sino además se da en los países industrializados, siendo en los primeros donde por las malas condiciones origina que los niños y adolescentes estén obligados a dejar la escuela para apoyar a su familia. (Delgado, 2017).	La Deserción Académica se determinará a través de la recolección de datos, empleando los instrumentos de Ficha de Registro de Datos y un Cuestionario.	Porcentaje de cumplimiento meta Deserción Escolar (PCMDE)	Razón
			Tasa de retención de estudiantes (TRE)	Razón

ANEXO 2: Tabla de Indicadores de la variable

Objetivos específicos	Indicadores	Descripción	Técnica/ Instrumento	Unidad de Medida	Fórmula
OE1: Incrementar la tasa de retención de estudiantes	Tasa de retención de estudiantes (TRE)	Determinar la Tasa de retención de estudiantes.	Análisis documental / Ficha de Registro	Anual	$\mathbf{TRE} = \frac{\sum_{i=1}^{n}(n)i}{EM}$*100 **TRE**= Tasa de retención de estudiantes **EM** = Número de estudiantes matriculados **n** = Número de estudiantes que terminan el año escolar
OE2: Reducir el tiempo promedio para analizar reportes Académicos de los estudiantes	Tiempo promedio para analizar reportes(TPAR)	Determinar el Tiempo promedio para analizar reportes.	Análisis documental / Ficha de Registro	Bimestral	$\mathbf{TPAR} = \frac{\sum_{i=1}^{n}(TAR)i}{n}$ **TPAR**= Tiempo promedio para analizar reportes académicos **TAR** = Tiempo de análisis de reportes académicos **n** = Número de reportes académicos analizados

OE3: Incrementar el grado de cumplimiento de la meta del indicador de deserción escolar.	Porcentaje de cumplimiento meta Deserción Escolar (PCMDE)	Determinar el porcentaje de cumplimiento meta Deserción Escolar.	Análisis documental / Ficha de Registro	Bimestral	$\mathbf{PCMDE} = \frac{\sum_{i=1}^{n}(NEMND)i}{NETAE}$*100 **PCMDE**= Porcentaje de cumplimiento meta Deserción Escolar Indicador meta Deserción **NEMND =** Número de Estudiantes META no deserción MINEDU **NETAE** = Número de estudiantes que terminan el año escolar
OE4: Incrementar el grado de satisfacción del Equipo Directivo	Nivel de Satisfacción del equipo directivo (NSED)	Determinar el Porcentaje de aprobación de encuesta de equipo directivo	Encuesta / Cuestionario	Bimestral	$\mathbf{PAESED} = \frac{\sum_{i=1}^{n}(\mathrm{TESRA})i}{n}$*100 **PAESED**= Porcentaje de aprobación de encuesta de equipo directivo **TESRA** = Total de encuestas de satisfacción con resultado aprobatorio **n** = Número de encuestas aplicadas

ANEXO 3: cálculo de tamaño de muestra

Para la muestra de la Población 1 se utilizó la siguiente fórmula:

$$\boldsymbol{n = \frac{Z^2 S^2 N}{E^2(N-1) + Z^2 S^2}}$$

Z = Nivel confianza 95%, 1,96

S = Desviación estándar

N = Tamaño de la población

E= % del estimador

$$\boldsymbol{n = \frac{(1,96)^2 * (0,5)^2 * 220}{(0,05)^2 * (220-1) + (1,96)^2 * (0,5)^2}}$$

n = 140 estudiantes de la Institución educativa N.º 88331.

ANEXO 4: Instrumentos de validación de datos.

Ficha de Registro Nº 1. Tasa de retención de estudiantes

Ficha de registro para determinar la tasa de retención de estudiantes				
Investigador				
Objeto de estudio				
Método de investigación				
Fecha de inicio			Fecha de fin	
Objetivo	Indicador	Medida	Fórmula	
Incrementar la tasa de retención de estudiantes	Tasa de retención de estudiantes (TRE)	Anual	$\mathbf{TRE} = \frac{\sum_{i=1}^{n}(n)i}{EM}$*100 **TRE**= Tasa de retención de estudiantes **EM** = Número de estudiantes matriculados **n** = Número de estudiantes que terminan el año escolar	
Año	Fecha	Número de estudiantes matriculados	Número de estudiantes que terminan el año escolar	**Tasa de retención de estudiantes**

Ficha de Registro Nº 2. Tiempo promedio para analizar reportes

Ficha de registro para determinar el tiempo promedio para analizar reportes				
Investigador				
Objeto de estudio				
Método de investigación				
Fecha de inicio			Fecha de fin	
Objetivo	Indicador	Medida	Fórmula	
Reducir el tiempo promedio para analizar reportes Académicos de los estudiantes	Tiempo promedio para analizar reportes(TPAR)	Bimestral	$\mathbf{TPAR} = \frac{\sum_{i=1}^{n}(TAR)i}{n}$ **TPAR**= Tiempo promedio para analizar reportes académicos **TAR** = Tiempo de análisis de reportes académicos **n** = Número de reportes académicos analizados	
Bimestre	Fecha	Tiempo de análisis de reportes académicos	Número de reportes académicos analizados	**Tiempo promedio para analizar reportes académicos**

Ficha de Registro Nº 3. Porcentaje de cumplimiento meta deserción escolar

Ficha de registro para determinar el porcentaje de cumplimiento meta deserción escolar				
Investigador				
Objeto de estudio				
Método de investigación				
Fecha de inicio			Fecha de fin	
Objetivo	Indicador	Medida	Fórmula	
Incrementar el grado de cumplimiento de la meta del indicador de deserción escolar.	Porcentaje de cumplimiento meta Deserción Escolar	Bimestral	$\mathbf{PCMDE} = \frac{\sum_{i=1}^{n}(NETAE)i}{NEMND}$*100 **PCMDE**= Porcentaje de cumplimiento meta Deserción Escolar Indicador meta Deserción **NEMND =** Número de Estudiantes META no deserción MINEDU **NETAE** = Número de estudiantes que terminan el año escolar	
AÑO	Fecha	Número de Estudiantes META no deserción MINEDU	Número de estudiantes que terminan el año escolar	**Porcentaje de cumplimiento meta deserción escolar indicador meta deserción**

Cuestionario de Satisfacción

Esta encuesta está destinada a evaluar algunos aspectos de calidad del Modelo de Data Science en la Detección de la Deserción Escolar, con el objetivo de Incrementar el grado de satisfacción del Equipo Directivo.

Esta encuesta debe ser respondida marcando con una X un casillero dentro de la escala, indicando el grado de acuerdo que tienes respecto al concepto que se expresa en cada ítem. La escala tiene cinco puntos, que van desde Nada de acuerdo hasta Totalmente de Acuerdo. Por favor responde a todos y cada uno de los ítems. Si piensas que en alguno de los ítems no puedes responder marca elpunto central de la escala.

Puntaje	Escala
1	Nada de acuerdo
2	Poco de acuerdo
3	Regularmente de acuerdo
4	Bastante de acuerdo
5	Totalmente de acuerdo

Nº	ITEM	PUNTUACIÓN				
1.	La información sobre la deserción escolar es confiable y sencilla de entender.	1	2	3	4	5
2.	Las consultas y reportes de la deserción escolar son exactas y no se presentan inconsistencias.	1	2	3	4	5
3.	La detección escolar se puede evaluar por bimestre	1	2	3	4	5
4.	Las consultas sobre deserción escolar son flexibles y variables en el tiempo.	1	2	3	4	5
5.	El tiempo ocupado en realizar reportes de deserción escolar es óptimo	1	2	3	4	5
6.	La disponibilidad de la información de deserción escolar es constante.	1	2	3	4	5
7.	La documentación para evaluar la deserción escolar es suficiente.	1	2	3	4	5
8.	Puedo conocer los factores que conllevan a la deserción escolar	1	2	3	4	5
9.	Realizó el seguimiento de los estudiantes por bimestre, para detectar la posible deserción escolar.	1	2	3	4	5
10.	Cuento con una interfaz amigable para conocer la deserción escolar	1	2	3	4	5
11.	La documentación para evaluar la deserción escolar, ha tenido una evolución continua y de mejora progresiva	1	2	3	4	5
12.	En general me encuentro satisfecho con la información obtenida sobre deserción escolar para toma de datos.	1	2	3	4	5

ANEXO 5: Validación de expertos

Evaluación por juicio de expertos

VALIDACIÓN DE EXPERTO 1

Respetado juez: Usted ha sido seleccionado para evaluar el instrumento que pretendo utilizar en la tesis para optar el grado de MAESTRO en Ingeniería de Sistemas, por la Escuela de Postgrado de la Universidad César Vallejo. La evaluación del instrumento es de gran relevancia para lograr que sea válido y que los resultados obtenidos a partir de éste sean utilizados eficientemente.

1. DATOS GENERALES DEL JUEZ

Nombre del juez:	JOHAN MAX ALEXANDER LOPEZ HEREDIA
Grado profesional:	Maestría (X) Doctor ( )
Área de Formación académica:	
Áreas de experiencia profesional:	5 Años
Institución donde labora:	Universidad Nacional del Santa
Tiempo de experiencia profesional en el área :	2 a 4 años ( ) Más de 5 años (x)

2. **PROPÓSITO DE LA EVALUACIÓN:**

a. Validar lingüísticamente el instrumento, por juicio de expertos.
b. Juzgar la pertinencia de los ítems de acuerdo a la dimensión del área según la autora.

Categoría	
ESENCIAL	
ÚTIL PERO PRESCINDIBLE	
INNECESARIO	

VALIDACIÓN DE EXPERTO 1

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO FICHA DE REGISTRO DE DATOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Ítems	Indicadores	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	Tasa de retención de estudiantes $\mathbf{X} = \frac{\sum_{i=1}^{n}(\text{Número de estudiantes retenidos})i}{número\ de\ estudiantes\ que\ terminan\ el\ año\ escolar}$*100	X			
2	Tiempo promedio para analizar reportes $\mathbf{X} = \frac{\sum_{i=1}^{n}(Tiempo\ de\ Análisis\ de\ Reportes)i}{número\ de\ reportes\ académicos}$	X			
3	Porcentaje de cumplimiento meta Deserción Escolar $\mathbf{X} = \frac{\sum_{i=1}^{n}(número\ de\ estudiantes\ que\ terminan\ año\ escolar)i}{Número\ de\ Estudiantes\ META\ no\ deserción\ MINEDU}$*100	X			

Sugerencias:

Opinión de aplicabilidad: **Aplicable (X) Aplicable después de corregir ( ) No aplicable ( )**

Nombre completo: JOHAN MAX ALEXANDER LOPEZ HEREDIA DNI: 46663398
Grado de Maestría en Ingeniería de Sistemas e Informática

Firma del Experto

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO CUESTIONARIO

MATRIZ DE EVALUACIÓN DE EXPERTOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Nº	Preguntas	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	La información que me brinda la herramienta es confiable.	X			
2	Las consultas y reportes que me brinda la herramienta son exactas y no se presentan inconsistencias.	X			
3	La navegación en la herramienta es fácil.	X			
4	La apariencia de la herramienta es estética y agradable, facilitando el trabajo cotidiano.	X			
5	Para operar la herramienta se requiere hacer una capacitación extensa y un continuo acompañamiento de los técnicos.	X			
6	La manera como se comunica la herramienta conmigo en la medida que trabajó con él (mensajes,	X			
7	advertencias, etc.) es entendible.	X			
8	La documentación de ayuda que tiene la herramienta es la apropiada.	X			
9	La herramienta presenta errores continuamente mientras se opera con ella.	X			
10	Cuando se solicita información en la herramienta, se despliega dicha información en el tiempo esperado.	X			
11	Considero que la herramienta es un activo para la empresa.	X			
12	Desde el inicio de mis labores con la herramienta, ha tenido una evolución continua y de mejora progresiva	X			

Sugerencias:

Opinión de aplicabilidad: Aplicable (X) Aplicable después de corregir ( ) No aplicable ( )

Nombre completo: JOHAN MAX ALEXANDER LÓPEZ HEREDIA

DNI: 46663398

Grado de Maestría en Ingeniería de Sistemas e Informática

Firma del Experto

Evaluación por juicio de expertos

VALIDACIÓN DE EXPERTO 2

Respetado juez: Usted ha sido seleccionado para evaluar el instrumento que pretendo utilizar en la tesis para optar el grado de MAESTRO en Ingeniería de Sistemas, por la Escuela de Postgrado de la Universidad César Vallejo. La evaluación del instrumento es de gran relevancia para lograr que sea válido y que los resultados obtenidos a partir de éste sean utilizados eficientemente.

3. DATOS GENERALES DEL JUEZ

Nombre del juez:	RICARDO ERNESTO IZAGUIRRE DIEGO
Grado profesional:	Maestría ( ) Doctor (X)
Área de Formación académica:	Administración de Empresas
Áreas de experiencia profesional:	15 Años
Institución donde labora:	Siderperu
Tiempo de experiencia profesional en el área :	2 a 4 años ( ) Más de 5 años (x)

4. PROPÓSITO DE LA EVALUACIÓN:

c. Validar lingüísticamente el instrumento, por juicio de expertos.
d. Juzgar la pertinencia de los ítems de acuerdo a la dimensión del área según la autora.

Categoría	
ESENCIAL	
ÚTIL PERO PRESCINDIBLE	
INNECESARIO	

VALIDACIÓN DE EXPERTO 2

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO FICHA DE REGISTRO DE DATOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Ítems	Indicadores	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	Tasa de retención de estudiantes $\mathbf{X} = \frac{\sum_{i=1}^{n}(número\ de\ estudiantes\ que\ terminan\ año\ escolar)i}{\text{Número de estudiantes retenidos}}$*100	X			
2	Tiempo promedio para analizar reportes $\mathbf{X} = \frac{\sum_{i=1}^{n}(Tiempo\ de\ Análisis\ de\ Reportes)i}{número\ de\ reportes\ académicos}$	X			
3	Porcentaje de cumplimiento meta Deserción Escolar $\mathbf{X} = \frac{\sum_{i=1}^{n}(número\ de\ estudiantes\ que\ terminan\ año\ escolar)i}{Número\ de\ Estudiantes\ META\ no\ deserción\ MINEDU}$*100	X			

Sugerencias:

Opinión de aplicabilidad: **Aplicable (X) Aplicable después de corregir ( ) No aplicable ( )**

Nombre completo: RICARDO ERNESTO IZAGUIRRE DIEGO **DNI: 32888444**

Grado de Maestría en Administración de Empresas

Firma del Experto

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO CUESTIONARIO

MATRIZ DE EVALUACIÓN DE EXPERTOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Nº	Preguntas	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	La información que me brinda la herramienta es confiable.	X			
2	Las consultas y reportes que me brinda la herramienta son exactas y no se presentan inconsistencias.	X			
3	La navegación en la herramienta es fácil.	X			
4	La apariencia de la herramienta es estética y agradable, facilitando el trabajo cotidiano.	X			
5	Para operar la herramienta se requiere hacer una capacitación extensa y un continuo acompañamiento de los técnicos.	X			
6	La manera como se comunica la herramienta conmigo en la medida que trabajó con él (mensajes,	X			
7	advertencias, etc.) es entendible.	X			
8	La documentación de ayuda que tiene la herramienta es la apropiada.	X			
9	La herramienta presenta errores continuamente mientras se opera con ella.	X			
10	Cuando se solicita información en la herramienta,se despliega dicha información en el tiempo esperado.	X			
11	Considero que la herramienta es un activo para la empresa.	X			
12	Desde el inicio de mis labores con la herramienta, ha tenido una evolución continua y de mejora progresiva	X			

Sugerencias:

Opinión de aplicabilidad: Aplicable (X) Aplicable después de corregir ( ) No aplicable ( )

Nombre completo: RICARDO ERNESTO IZAGUIRRE DIEGO

DNI: 32888444

Grado de Maestría en Administración de Empresas

Firma del Experto

Evaluación por juicio de expertos

VALIDACIÓN DE EXPERTO 3

Respetado juez: Usted ha sido seleccionado para evaluar el instrumento que pretendo utilizar en la tesis para optar el grado de MAESTRO en Ingeniería de Sistemas, por la Escuela de Postgrado de la Universidad César Vallejo. La evaluación del instrumento es de gran relevancia para lograr que sea válido y que los resultados obtenidos a partir de éste sean utilizados eficientemente.

5. DATOS GENERALES DEL JUEZ

Nombre del juez:	ANDRÉS DAVID EPIFANÍA HUERTA
Grado profesional:	Maestría (X) Doctor ( )
Área de Formación académica:	Ingeniería de Sistemas
Áreas de experiencia profesional:	20 Años
Institución donde labora:	Universidad Católica los Ángeles de Chimbote
Tiempo de experiencia profesional en el área :	2 a 4 años ( ) Más de 5 años (x)

6. PROPÓSITO DE LA EVALUACIÓN:

e. Validar lingüísticamente el instrumento, por juicio de expertos.
f. Juzgar la pertinencia de los ítems de acuerdo a la dimensión del área según la autora.

Categoría	
ESENCIAL	
ÚTIL PERO PRESCINDIBLE	
INNECESARIO	

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO FICHA DE REGISTRO DE DATOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Ítems	Indicadores	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	Tasa de retención de estudiantes $\mathbf{X} = \frac{\sum_{i=1}^{n}(\text{Número de estudiantes retenidos})i}{\textit{número de estudiantes que terminan el año escolar}}$*100	X			
2	Tiempo promedio para analizar reportes $\mathbf{X} = \frac{\sum_{i=1}^{n}(\textit{Tiempo de Análisis de Reportes})i}{\textit{número de reportes académicos}}$	X			
3	Porcentaje de cumplimiento meta Deserción Escolar $\mathbf{X} = \frac{\sum_{i=1}^{n}(\textit{número de estudiantes que terminan año escolar})i}{\textit{Número de Estudiantes META no deserción MINEDU}}$*100	X			

Sugerencias:

Opinión de aplicabilidad: **Aplicable (X) Aplicable después de corregir ( ) No aplicable ( )**

Nombre completo: ANDRÉS DAVID EPIFANÍA HUERTA
DNI: 40197616
Grado de Maestría en Ingeniería de Sistemas

Firma del Experto

CERTIFICADO DE VALIDEZ DE CONTENIDO DEL INSTRUMENTO CUESTIONARIO

MATRIZ DE EVALUACIÓN DE EXPERTOS

Título de la investigación:	**Modelos de Data Science para mejorar la detección de la Deserción Académica en la Institución Educativa Nº 88331 en Chimbote - 2021**
Línea de investigación:	SISTEMAS DE INFORMACIÓN Y COMUNICACIONES
El instrumento de medición pertenece a las variables:	VI: Modelos de Data Science VD: Detección de la Deserción Académica

Mediante la matriz de evaluación de expertos. Ud. Tiene la facultad de evaluar cada una de las preguntas marcando con una "x" en las columnas de SÍ o NO. Asimismo, le exhortamos en la corrección de los ítems indicando sus observaciones con la finalidad de mejorar la coherencia de las preguntas sobre la variable en estudio

Nº	Preguntas	Esencial	Útil pero prescindible	Innecesario	Observaciones
1	La información que me brinda la herramienta es confiable.	X			
2	Las consultas y reportes que me brinda la herramienta son exactas y no se presentan inconsistencias.	X			
3	La navegación en la herramienta es fácil.	X			
4	La apariencia de la herramienta es estética y agradable, facilitando el trabajo cotidiano.	X			
5	Para operar la herramienta se requiere hacer una capacitación extensa y un continuo acompañamiento de los técnicos.	X			
6	La manera como se comunica la herramienta conmigo en la medida que trabajó con él (mensajes,	X			
7	advertencias, etc.) es entendible.	X			
8	La documentación de ayuda que tiene la herramienta es la apropiada.	X			
9	La herramienta presenta errores continuamente mientras se opera con ella.	X			
10	Cuando se solicita información en la herramienta, se despliega dicha información en el tiempo esperado.	X			
11	Considero que la herramienta es un activo para la empresa.	X			
12	Desde el inicio de mis labores con la herramienta, ha tenido una evolución continua y de mejora progresiva	X			

Sugerencias:

Opinión de aplicabilidad: Aplicable (X) Aplicable después de corregir( ) No aplicable()
Nombre completo: ANDRÉS DAVID EPIFANÍA HUERTA
DNI: 40197616
Grado de Maestría en Ingeniería de Sistemas

Firma del Experto

ANEXO 6: Cuadro validación de Cuestionario con LAWSHE

Nº	enunciado / ítems	Lawshe	Tristan Lawshe	Decisión L	Decisión T-L
Ítem 1	La información sobre la deserción escolar es confiable y sencilla de entender.	**1,00**	**1,00**	excelente	excelente
Ítem 2	Las consultas y reportes de la deserción escolar son exactas y no se presentan inconsistencias.	**0,60**	**0,80**	excelente	excelente
Ítem 3	La detección escolar se puede evaluar por bimestre	**1,00**	**1,00**	excelente	excelente
Ítem 4	Las consultas sobre deserción escolar son flexibles y variables en el tiempo.	**1,00**	**1,00**	perfecto	excelente
Ítem 5	El tiempo ocupado en realizar reportes de deserción escolar es óptimo	**1,00**	**1,00**	excelente	excelente
Ítem 6	La disponibilidad de la información de deserción escolar es constante.	**1,00**	**1,00**	perfecto	excelente
Ítem 7	La documentación para evaluar la deserción escolar es suficiente.	**1,00**	**1,00**	excelente	excelente
Ítem 8	Puedo conocer los factores que conllevan a la deserción escolar	**1,00**	**1,00**	excelente	excelente
Ítem 9	Realizó el seguimiento de los estudiantes por bimestre, para detectar la posible deserción escolar.	**1,00**	**1,00**	excelente	excelente
Ítem 10	Cuento con una interfaz amigable para conocer la deserción escolar	**1,00**	**1,00**	excelente	excelente
Ítem 11	La documentación para evaluar la deserción escolar, ha tenido una evolución continua y de mejora progresiva	**1,00**	**1,00**	excelente	excelente
Ítem 12	En general me encuentro satisfecho con la información obtenida sobre deserción escolar para toma de datos.	**1,00**	**1,00**	perfecto	excelente
	LAWSHE INSTRUMENTO (CVI)	**0,97**	**0,98**		
	CVI ítems aceptables		**0,98**		

Se observa en la tabla mostrada anteriormente, los 12 Ítems del instrumento de recolección de datos encuesta, el cual fue validada por tres expertos, se puede observar que el valor de CVI de LAWSHE es de 0.98 (<75% - 100%]), con un nivel de confianza del 95%, esto significa que la validación es elevada.

ANEXO 7: Confiabilidad del instrumento de las variables.

	ÍTEMS												
ENCUESTADOS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**SUMA**
E01	5	5	5	5	5	4	5	5	5	5	5	5	59
E02	5	4	5	5	5	5	5	3	5	5	4	5	56
E03	5	5	5	5	1	5	5	5	5	5	5	4	55
E04	4	4	1	4	4	4	4	4	4	4	4	4	45
E05	4	5	4	4	4	4	4	4	4	3	4	4	48
E06	4	4	4	4	4	4	4	4	4	4	4	4	48
E07	5	5	5	5	5	5	5	5	5	5	5	5	60
E08	5	5	5	5	5	5	5	5	5	5	5	5	60
E09	4	4	4	4	4	4	4	4	4	4	4	4	48
E10	2	1	4	2	5	2	2	1	3	3	1	2	28

Estadísticas de fiabilidad

Alfa de Cronbach	N de elementos
,973	12

Estadísticas de total de elemento

	Media de escala si el elemento se ha suprimido	Varianza de escala si el elemento se ha suprimido	Correlación total de elementos corregida	Alfa de Cronbach si el elemento se ha suprimido
P1	46,3000	100,456	,983	,968
P2	46,4000	97,156	,882	,970
P3	46,4000	107,378	,438	,982
P4	46,3000	100,456	,983	,968
P5	46,5000	97,611	,888	,970
P6	46,4000	102,267	,911	,969
P7	46,3000	100,456	,983	,968
P8	46,6000	97,822	,838	,971
P9	46,2000	105,733	,961	,970
P10	46,3000	105,344	,830	,971
P11	46,5000	96,500	,940	,968
P12	46,4000	102,267	,911	,969

Análisis de la confiabilidad:

Utilizando el método del Alfa de Cronbach y aplicado a una muestra piloto de 10 personas con características similares a la muestra, obtuvo un coeficiente de confiabilidad de $\propto = 0.973$, lo que permite inferir que el instrumento a utilizar es CONFIABLE y CONSISTENTE para medir la variable con 12 ítems.

ANEXO 8: Aceptación

MINISTERIO DE EDUCACIÓN
UNIDAD DE GESTIÓN EDUCATIVA LOCAL SANTA INSTITUCIÓN EDUCATIVA Nº 88331 – RINCONADA

"Año del Bicentenario del Perú: 200 años de Independencia"

Chimbote 17 de noviembre del año 2021

OFICIO N° 101 – 2021 /ME/DRE-ANCASH/UGEL-SANTA/I.E. N° 88331 – DI

Sra: ZENAIDA CRISTINA SHICA JULCA

Presente:

ASUNTO: AUTORIZACIÓN PARA TRABAJO DE INVESTIGACIÓN
REF: CARTA DE PRESENTACIÓN Nº 148-2021 UCV.

Es grato dirigirme a usted en mi condición de Director de la Institución Educativa 88331, de la provincia del Santa, Departamento de Ancash, para hacer de su conocimiento mi ACEPTACIÓN, para que lleve adelante su Proyecto de Investigación el cual se denomina ***Modelos de Data Science para la detección de la Deserción Académica***, para ello les brindaré la información necesaria y relevante de acuerdo a sus requerimientos académicos.

Sin otro particular, auguro los mejores éxitos y parabienes en su investigación.

Atentamente

Mg. Alfredo G. Lozano Cabrera
DIRECTOR

ANEXO 9: Prueba de Normalidad

Pruebas de la normalidad de los indicadores

- ✓ **Prueba de normalidad de la tasa de retención escolar**

Se realizó la prueba de normalidad para la tasa de retención de estudiantes, donde los datos utilizados fueron 10, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%.

Tabla 21 Prueba de normalidad del indicador I

Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
TREa	,899	10	,216
TREd	,706	10	,001

a. Corrección de significación de Lilliefors

Fuente: Elaboración propia

Los resultados se visualizan en la tabla 13, una población de 10 (gl), a quien se le aplicó la prueba de Shapiro Wilk, entre lo obtenido en el Postest tenemos un valor (Sig) = 0.001 que es menor a 0.05, lo cual significa que los datos no siguen una distribución normal. En consecuencia, se trabajó la prueba no paramétrica de Wilcoxon para la validación de la hipótesis.

- ✓ **Prueba de normalidad del Tiempo promedio para analizar reportes**

Se realizó la prueba de normalidad para Tiempo promedio para analizar reportes., donde los datos utilizados fueron 10, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%.

Tabla 22: Prueba de normalidad del indicador 2

Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
TPARa	,864	10	,086
TPARd	,663	10	,000

a. Corrección de significación de Lilliefors

Fuente: Elaboración propia

Los resultados se visualizan en la tabla 14, una población de 10 (gl), a quien se le aplicó la prueba de Shapiro Wilk, entre lo obtenido en el Postest tenemos un valor (Sig) = 0.000 que es menor a 0.05, lo cual significa que los datos no siguen una distribución normal. En consecuencia, se trabajó la prueba no paramétrica de Wilcoxon para la validación de la hipótesis.

- ✓ **Prueba del porcentaje de cumplimiento meta deserción escolar.**

Se realizó la prueba de normalidad para el porcentaje de cumplimiento meta deserción escolar, donde los datos utilizados fueron 4, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%.

Tabla 23 Prueba de normalidad del indicador 3

Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
PCMDEa	,963	4	,798
PCMDEd	,807	4	,116

a. Corrección de significación de Lilliefors

Fuente: Elaboración propia

Los resultados se visualizan en la tabla 15, una población de 4 (gl), a quien se le aplicó la prueba de Shapiro Wilk, entre lo obtenido en el Pretest tenemos un valor (Sig) = 0.798 que es mayor a 0.05 y el Postest donde

se obtuvo un valor (Sig) = 0.116 es decir es mayor a 0.05, lo cual significa que los datos siguen una distribución normal. En consecuencia, se trabajó la prueba paramétrica de T - Student para la validación de la hipótesis.

- ✓ **Prueba de porcentaje de aprobación de encuesta de equipo directivo**

Se realizó la prueba de normalidad para el porcentaje de aprobación de encuesta de equipo directivo, donde los datos utilizados fueron 10, por ende se trabajó con Shapiro — Wilk. Así también, se realizó el procesamiento de los datos con ayuda del software SPSS IBM y se obtuvo un nivel de confianza del 95%.

Tabla 24: Prueba de normalidad del indicador 4

Pruebas de normalidad

	Shapiro-Wilk		
	Estadístico	gl	Sig.
PAEEa	,859	10	,074
PAEEd	,844	10	,049

a. Corrección de significación de Lilliefors

Fuente: Elaboración propia

Los resultados se visualizan en la tabla 16, una población de 10 (gl), a quien se le aplicó la prueba de Shapiro Wilk, entre lo obtenido en el Pretest tenemos un valor (Sig) = 0.049 que es menor a 0.05, lo cual significa que los datos no siguen una distribución normal. En consecuencia, se trabajó la prueba no paramétrica de Wilcoxon para la validación de la hipótesis.

ANEXO 10: Desarrollo del Modelo de Data Science

El sector de educación a lo largo de los últimos años ha buscado reducir la brecha de deserción escolar, a través de políticas de estado que promuevan y garanticen el término de la etapa escolar, sin embargo es necesario mencionar uno de los problemas más latentes y poco controlados en dicho sector es el tratamiento de los estudiantes con potencial a deserción tal es así que no se puede determinar si el estudiante terminará su año escolar o será desertor al inicio, durante o casi para concluir su año académico con las estrategias tradicionales.

Para Bean y Metzner (1985), la deserción ocurre al momento en el que un estudiante que se ha matriculado en la escuela y deja la escuela por una variedad de razones, incluidas económicas, sociales, culturales y familiares, lo que hace que fracase en su año académico regular.

El número de cursos desaprobados y las notas de comportamiento en muchas ocasiones puede influir sobre el resultado del término del año escolar.
El éxito del término del año académico depende de muchos factores los cuales se analizaron en el presente trabajo de investigación, dichos factores se llamarán en adelante features o drivers, quienes van a influir en el éxito o no del término del año escolar.

En esta investigación se aplicó la metodología de CRISP-DM, que según sus siglas son Cross-Industry Standard Process For Data Mining, bajo 6 pasos en su ciclo de vida, el cual ofrece flexibilidad entre las fases al poder regresar entre una y otra. Esta metodología permitió manejar un marco de trabajo para entender el negocio, comprender los datos, preparar los datos, modelar y finalmente evaluarlo, logrando finalmente implementar la solución.

Los features permiten un grado de influencia sobre el target u objetivo el cual es predecir la propensión del término del año escolar empleando algoritmos de Machine Learning de alta precisión.

METODOLOGÍA CRISP-DM

En la presente investigación se aplicó la metodología CRISP-DM, con el fin de predecir la propensión del término del año escolar. La metodología CRISP-DM está basada en 6 fases:

1. **Comprensión del negocio**
2. **Comprensión de los datos**
3. **Preparación de los datos**
4. **Modelado**
5. **Evaluación**
6. **Distribución**

El ciclo de vida CRISP-DM se visualiza en el siguiente gráfico:

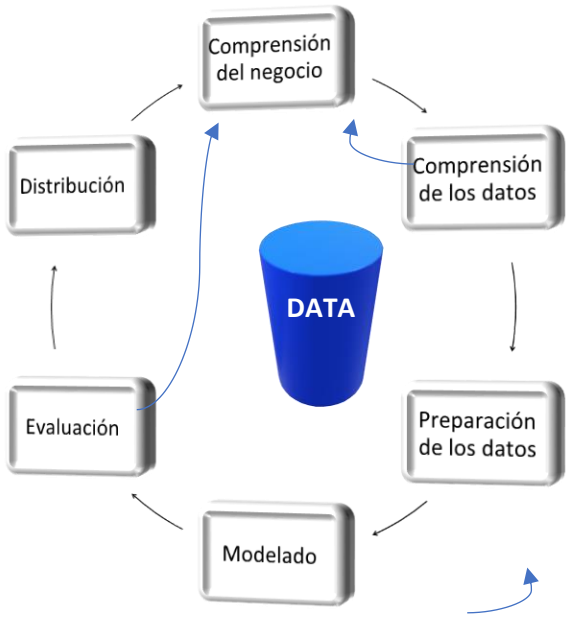


Figura 10: Pasos de la Metodología CRISP-DM

Fuente: Elaboración Propia del autor

1. Comprensión del negocio

1.1 Sector educación básica Jornada Escolar Completa: I.E. 88331

La Institución Educativa Nº 88331 del Centro Poblado de Rinconada en la ciudad de Chimbote pertenece al Modelo de Educación JEC, Jornada Escolar Completa en el nivel secundaria. Este modelo se implementó en algunos colegios focalizados en el año 2015 en la búsqueda de mejorar la calidad educativa agregando horas académicas, soporte emocional e implementando aulas equipadas con equipos de cómputo y conectividad a internet.

La presente investigación tuvo el objetivo de conocer la probabilidad de que los estudiantes terminen su año escolar o no, desde el inicio hasta el fin del año académico.

1.2 Definición del Problema

El equipo directivo tiene dificultad en estimar quiénes son los estudiantes potenciales desertores antes de finalizar el año escolar, si el estudiante termina el año escolar o si es un desertor del año académico.

Se logró identificar aquellos factores que están influenciando de manera negativa sobre el término del año escolar, y que generan que el problema se agudice sin mayor control, impactando negativamente en la vida de los estudiantes, así como la inestabilidad de la Institución Educativa. El objetivo del proyecto es predecir la propensión del término del año escolar usando algoritmos de Machine Learning de alta precisión.

2. Comprensión de los Datos

2.1 Recopilación inicial de Datos

Base de Datos Interno: La Institución Educativa cuenta con una base de datos sobre el estado académico de los estudiantes desde el año 2012 al 2019, al final del año escolar.

2.2 Descripción de los Datos

En el presente proyecto vamos a trabajar con el Dataset (Datos), Deserción.xls, con un total de 2212 Registros y está compuesta por 22 variables, clasificados como variables independientes y variable dependiente.

Tabla 25: Descripción de variables Features

Variable independiente	Descripción de la Variable
ID	Identificador único de cada estudiante
sexo	Categoría de género del estudiante (1- Femenino, 2-Masculino)
D_nac	Día de Nacimiento del estudiante
M_nac	Mes de Nacimiento del estudiante
A_nac	Año de Nacimiento del estudiante
Mat	Nota del área de Matemática por año académico
Com	Nota del área de Comunicación por año académico
Inglés	Nota del área de Inglés por año académico
Arte	Nota del área de Arte por año académico
Hist_Geo_E	Nota del área de Historia y Geografía por año académico
Form_ciu_Civ	Nota del área de Formación Ciudadana y Cívica por año académico
Per_F_R_H	Nota del área de Persona Familia y Relaciones Humanas por año académico
Edu_Fis	Nota del área de Educación Física por año académico
Edu_Rel	Nota del área de Educación Religiosa por año académico
C_T_A	Nota del área de Ciencia Tecnología y Ambiente por año académico
Edu_Trab	Notas del área de Educación para el Trabajo por año académico
Areas_Desap	Número de Áreas Desaprobadas por año académico
Comp	Categoría de Comportamiento del estudiante (1 = A, 2=B, 3 = C, 4=AD)
Grado	Categoría de Grado del estudiante (1 = Primero, 2 = Segundo, 3 = Tercero, 4 = Cuarto, 5 = Quinto)
Sección	Categoría de la Sección (1 = A, 2 = B, 3 = C)
Año	Año académico

Tabla 26: Clasificación de las variables del dataset

Variable	Tipo
Situacion_Final	Cualitativa nominal
ID	Cuantitativa discreta
sexo	Cualitativa nominal
D_nac	Cuantitativa discreta
M_nac	Cuantitativa discreta
A_nac	Cuantitativa discreta
Mat	Cuantitativa continua
Com	Cuantitativa continua
Inglés	Cuantitativa continua
Arte	Cuantitativa continua
Hist_Geo_E	Cuantitativa continua
Form_ciu_Civ	Cuantitativa continua
Per_F_R_H	Cuantitativa continua
Edu_Fis	Cuantitativa continua
Edu_Rel	Cuantitativa continua
C_T_A	Cuantitativa continua
Edu_Trab	Cuantitativa continua
Areas_Desap	Cuantitativa discreta
Comp	Cualitativa nominal
Grado	Cualitativa nominal
Sección	Cualitativa nominal
Año	Cuantitativa discreta

Tabla 27: Descripción de variable Target

Recordando que nuestro objetivo es “PREDECIR LA PROPENSIÓN DEL TÉRMINO DEL AÑO ESCOLAR”. Se tiene la siguiente variable.

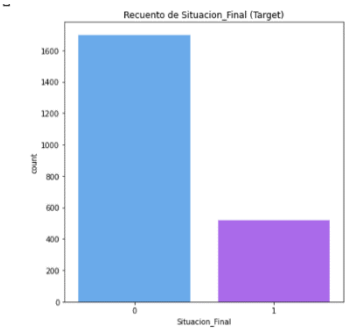
Variable Dependiente	Descripción de la Variable
Situacion_Final	Categoría de término del año escolar (1 = no Desertó, 2 = desertó)

2.3 Tipos de Variables:

- ✓ **Análisis Exploratorio de los Datos**

Figura 11: Distribución del Target Categoría de Daño del cultivo.

La variable objetivo está representada por la variable Situacion_Final (Categoría de término del año escolar) , distribuido de la siguiente manera:



1 -> No desertó, 1695 registros, representa un 76,6 % del total

2 -> Desertó, 517 registros, representa un 23,4 % del total.

- ✓ **Análisis Univariado**

Estudio de Variables Numéricas

La función descriptiva de Pandas muestra estadísticas descriptivas incluyendo: media, mediana, máx, mín, std y conteos para una columna en particular de los datos. La función describe solo regresa los valores de estas estadísticas para las columnas numéricas.

Tabla 28: Análisis de los indicadores estadísticos.

	count	mean	std	min	25%	50%	75%	max
D_nac	2212.0	15.902351	8.798280	1.0	8.0	16.0	23.25	31.0
M_nac	2212.0	6.287523	3.361334	1.0	4.0	6.0	9.00	12.0
A_nac	2212.0	2000.264467	3.164014	1978.0	1998.0	2000.0	2003.00	2009.0
Mat	2212.0	11.494575	3.939900	0.0	11.0	12.0	13.00	20.0
Com	2212.0	11.500000	3.852845	0.0	11.0	12.0	14.00	20.0
Ingles	2212.0	11.725588	3.847173	0.0	11.0	12.0	14.00	20.0
Arte	2212.0	12.090416	4.003273	0.0	11.0	13.0	14.00	20.0
Hist_Geo_E	2212.0	12.193942	3.879098	0.0	12.0	13.0	14.00	20.0
Form_ciu_Civ	2212.0	12.487342	3.989507	0.0	12.0	13.0	15.00	19.0
Per_F_R_H	2212.0	11.855787	3.928868	0.0	11.0	12.0	14.00	20.0
Edu_Fis	2212.0	13.014919	4.057619	0.0	13.0	14.0	15.00	20.0
Edu_Rel	2212.0	12.609403	4.621275	0.0	12.0	14.0	15.00	20.0
C_T_A	2212.0	12.150542	3.884054	0.0	11.0	13.0	14.00	19.0
Edu_Trab	2212.0	12.232821	3.995650	0.0	12.0	13.0	14.00	20.0
Areas_Desap	2212.0	0.839512	1.689892	0.0	0.0	0.0	1.00	11.0
Ano	2207.0	2014.856819	2.601411	2011.0	2013.0	2015.0	2017.00	2019.0

count: este campo muestra la cantidad de datos que contiene las columnas del dataset.

mean: este campo muestra el valor promedio de la columna del dataset.

std: este campo muestra la desviación estándar de la columna del dataset.

También se muestran los valores máximo y mínimo, así como el límite de cada uno de los cuartiles.

En la siguiente tabla podemos notar que existen valores nulos o Missing de las columnas Comportamiento de los estudiantes y año escolar.

Tabla 29: Valores missing de las columnas Comp y Ano

ID	0
sexo	0
D_nac	0
M_nac	0
A_nac	0
Mat	0
Com	0
Ingles	0
Arte	0
Hist_Geo_E	0
Form_ciu_Civ	0
Per_F_R_H	0
Edu_Fis	0
Edu_Rel	0
C_T_A	0
Edu_Trab	0
Areas_Desap	0
Comp	163
Grado	0
Seccion	0
Ano	5
Situacion_Final	0

Analizaremos las variables numéricas visualmente.

Edu_Rel: Observamos la distribución del Recuento de notas del área de Educación Religiosa por año escolar de los estudiantes, con el siguiente gráfico de histograma.

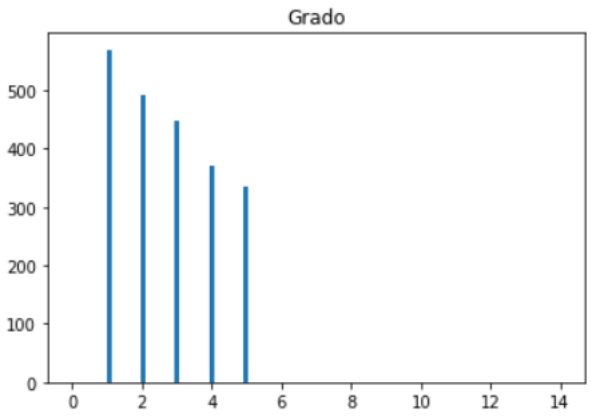


Figura 12: Histograma del Recuento de notas del área de Educación Religiosa por año académico.

Areas_Desap: Observamos la distribución del Número áreas desaprobadas, con el siguiente gráfico de histograma.

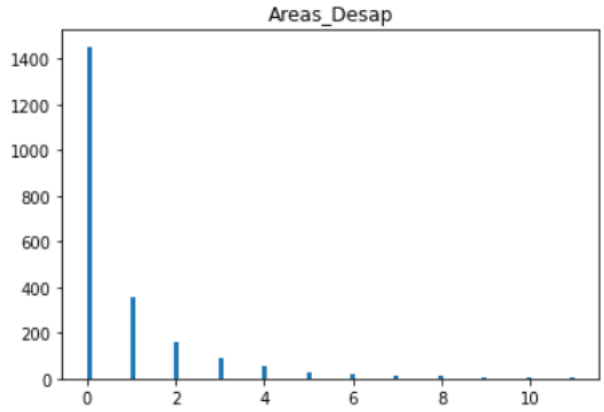


Figura 13: Histograma del Nº de áreas desaprobadas.

Mat: Observamos la distribución de Recuento de notas del área de Matemática por año escolar de los estudiantes, con el siguiente gráfico de histograma.

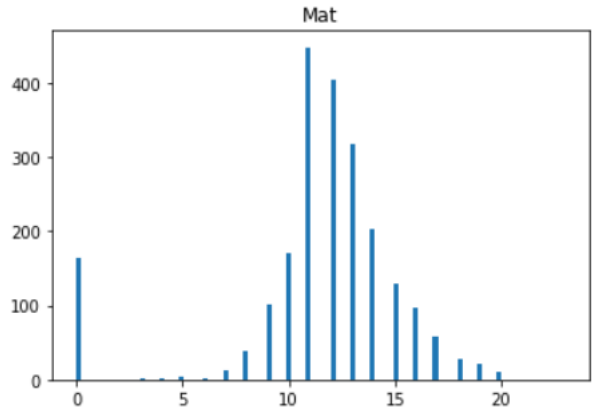


Figura 14: Histograma del Nº de notas del área de matemática por año académico

Edu_Fis: Observamos la distribución de Recuento de notas del área de Educación Física por año escolar de los estudiantes, con el siguiente gráfico de histograma.

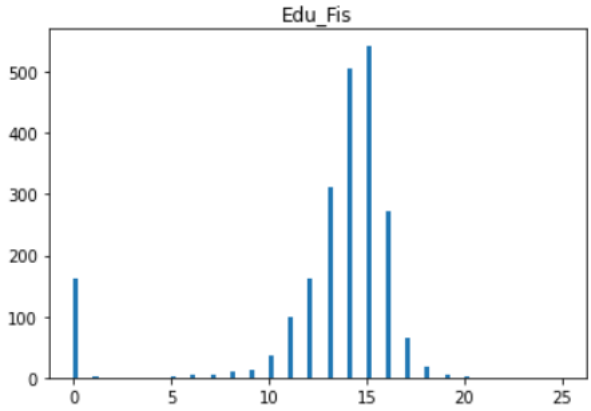


Figura 15: Histograma del Nº de notas del área de Educación Física por año académico

El gráfico de cajas es muy importante pues nos muestra, dispersión, forma y atípicos:

Edu_Rel: Observamos de manera gráfica el recuento estimado de notas del área de Educación Religiosa por año académico, con el siguiente gráfico de caja.

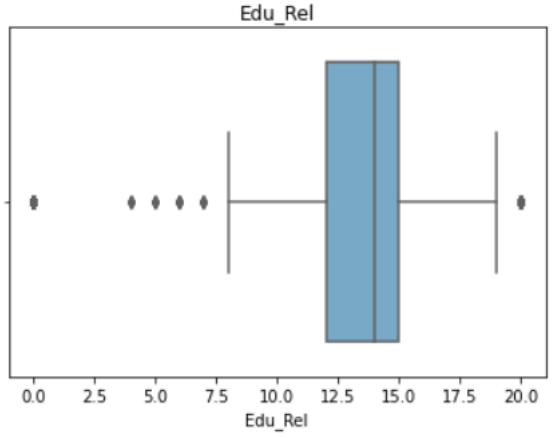


Figura 16: Gráfico de cajas Recuento de notas del área de Educación Religiosa por año académico.

Areas_Desap: Observamos de manera gráfica el recuento estimado de Número áreas desaprobadas, con el siguiente gráfico de caja.

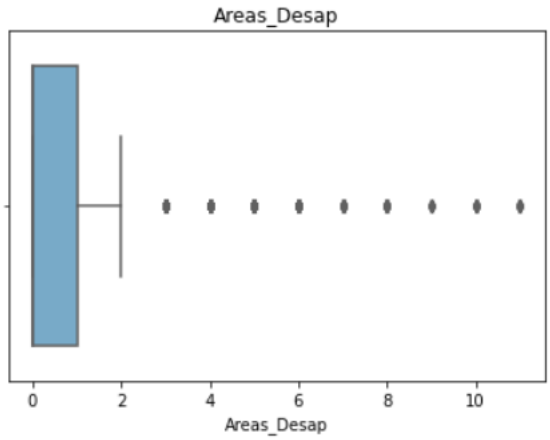


Figura 17: Gráfico de cajas del Nº de áreas desaprobadas

Mat: Observamos de manera gráfica el recuento estimado de notas del área de Matemática por año escolar de los estudiantes, con el siguiente gráfico de caja.

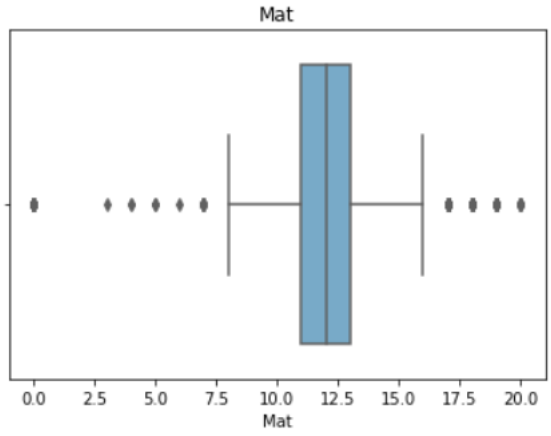


Figura 18: Gráfico de cajas del Nº de notas del área de matemática por año académico.

Edu_Fis: Observamos de manera gráfica el recuento estimado de notas del área de Educación Física por año escolar de los estudiantes, con el siguiente gráfico de caja.

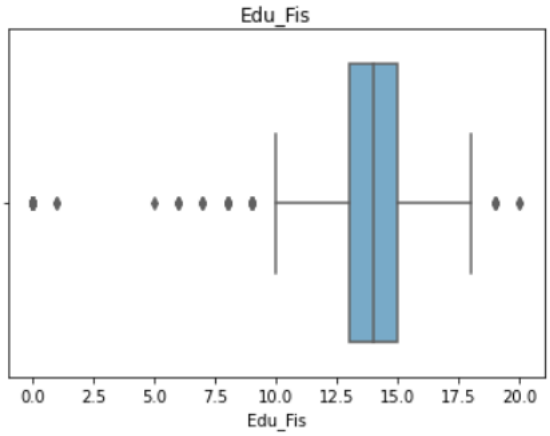


Figura 19: Gráfico de cajas del Nº de notas del área de Educación Física por año académico

Estudiamos Variables Categóricas

Analizaremos las variables Categóricas.

El Dataset en estudio cuenta con 5 variables Categóricas, incluyendo el target.

- ✓ Sexo
- ✓ Comp
- ✓ Sección
- ✓ Grado
- ✓ Situacion_Final

Mostramos la frecuencia de variables Categóricas para encontrar hallazgos

```
Situacion_Final
1    1695
2     517
```

```
sexo
1    1020
2    1192
dtype: int64
```

```
Comp
1.0    1180
2.0     604
3.0      59
4.0     206
dtype: int64
```

```
Seccion
1    1026
2     916
3     270
dtype: int64
```

```
Grado
1    570
2    491
3    447
4    370
5    334
dtype: int64
```

Observamos las variables categóricas visualmente.

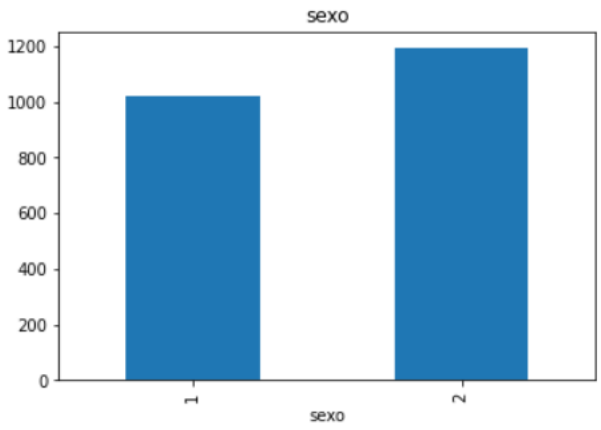


Figura 20: Gráfico de Barras de los Estudiantes por sexo.

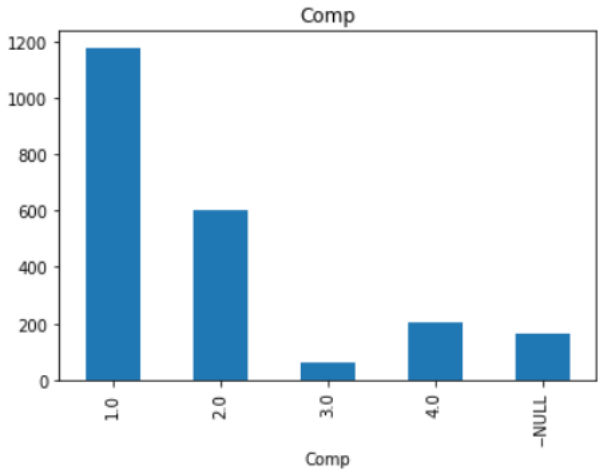


Figura 21: Gráfico de los Estudiantes por Comportamiento*.*

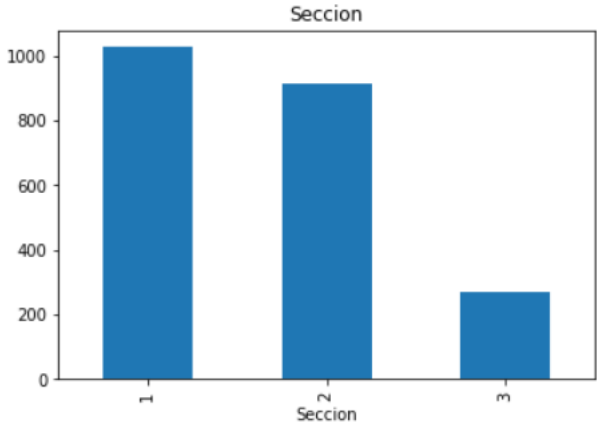


Figura 22: Gráfico de Barras de los Estudiantes por Sección.

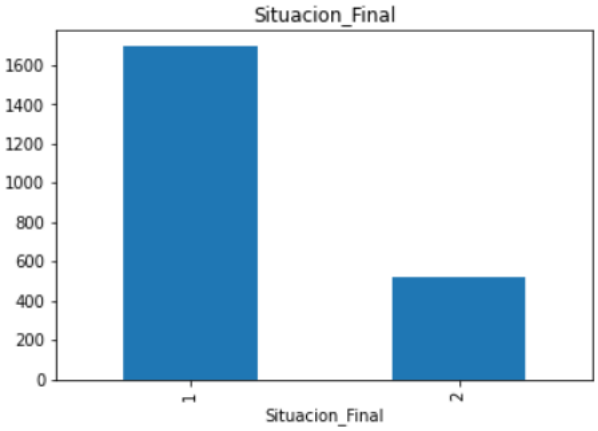


Figura 23: Gráfico de Barras de los Estudiantes por Situación Finalizado el año académico

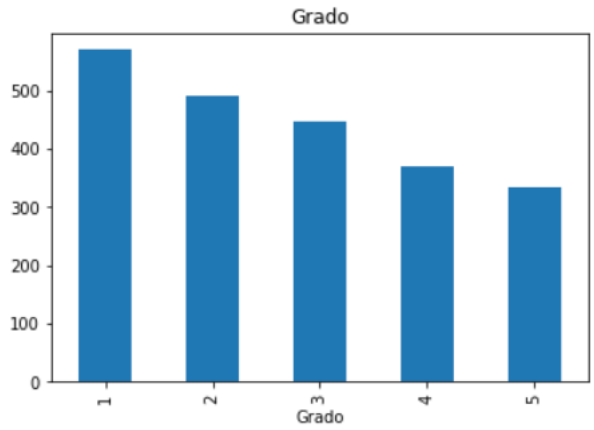


Figura 24: Gráfico de Barras de los Estudiantes por Grado*.*

Análisis Bivariado

Se analizó el target Categoría de Situación final del año académico (Situacion_Final) vs las variables independientes :

El gráfico de cajas es una excelente herramienta visual porque nos muestra , dispersión, forma y atípicos de las variables.

Situacion_Final vs Mat:

Comparamos la Categoría de Situación final del año académico con respecto al recuento estimado de las notas del área de Matemáticas por año académico, con el siguiente gráfico de cajas.

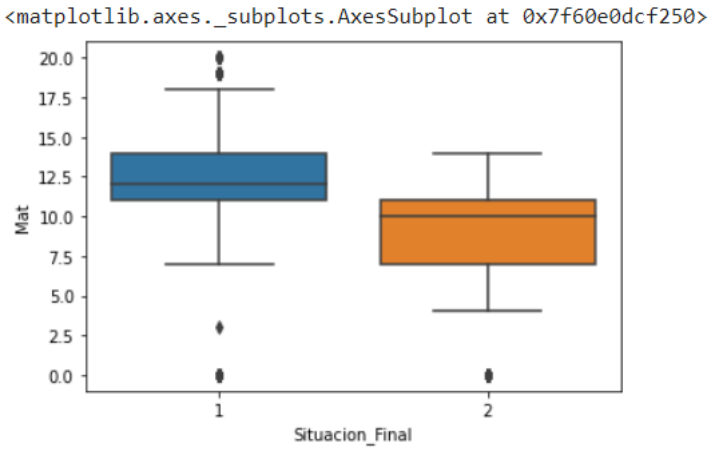


Figura 25: Gráfico de cajas Situación final del año académico con respecto al Recuento de notas del área de Matemáticas por año académico

Situacion_Final vs Areas_Desap:

Comparamos la Categoría de Situación final del año académico con respecto al recuento estimado de las áreas desaprobadas por año académico, con el siguiente gráfico de cajas.

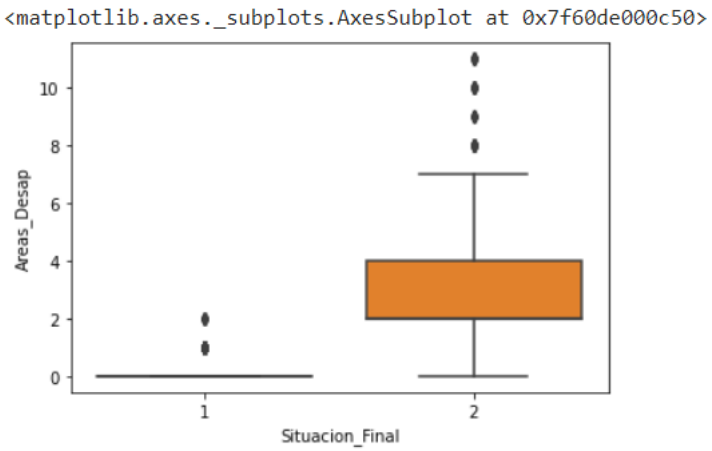


Figura 26: Gráfico de cajas Situación final del año académico con respecto al Recuento de áreas desaprobadas por año académico.

Situacion_Final vs Grado:

Comparamos la Categoría de Situación final del año académico con respecto al recuento estimado del grado por año académico, con el siguiente gráfico de cajas.

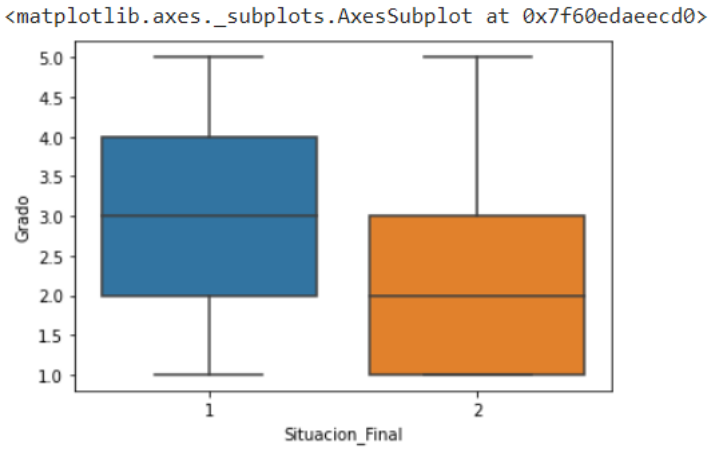


Figura 27: Gráfico de cajas Situación final del año académico con respecto al Recuento del grado por año académico.

Análisis Multivariado

El Análisis Multivariado es un método estadístico el cual se utiliza para estimar la contribución de varios factores en un simple evento o resultado.

Analizamos la correlación de las variable:

La correlación es una expresión numérica basada en la estadística cuya finalidad es evaluar la relación de dos o más variables, es decir, permite medir la dependencia de una variable con respecto de otra variable independiente.

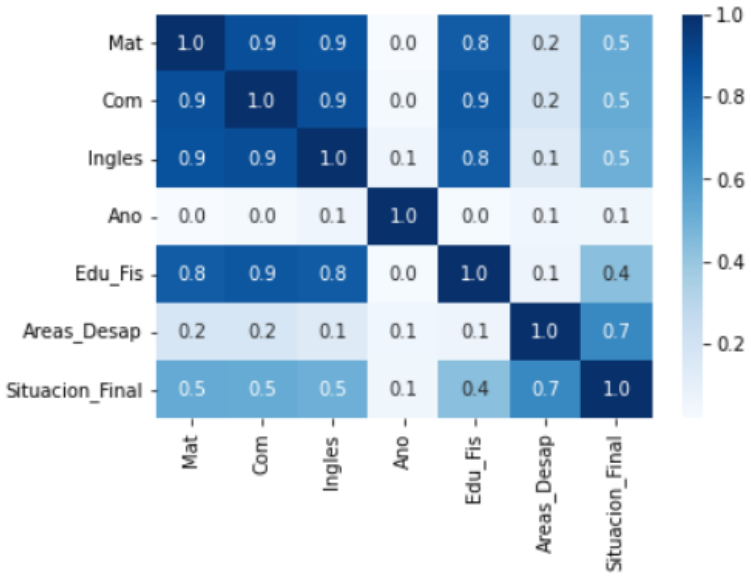


Figura 28: Gráfico de correlación de las variables numéricas.

En el gráfico anterior se visualiza que el daño a los cultivos tiene una correlación decente con las variables numéricas:

- ✓ Areas_Desap
- ✓ Inglés
- ✓ Com
- ✓ Mat

3. Preparación de los Datos

3.1 Completitud de los Datos

Observamos los primeros registros e identificamos que la Columna: Comp y Ano no presentan completitud de los datos.

Tabla 30: Valores missing en la columna Comportamiento de los estudiantes.

ID	0
sexo	0
D_nac	0
M_nac	0
A_nac	0
Mat	0
Com	0
Ingles	0
Arte	0
Hist_Geo_E	0
Form_ciu_Civ	0
Per_F_R_H	0
Edu_Fis	0
Edu_Rel	0
C_T_A	0
Edu_Trab	0
Areas_Desap	0
Comp	163
Grado	0
Seccion	0
Ano	5
Situacion_Final	0

Como se aprecia en la tabla anterior, dos de las variables tienen valores nulos o missings.

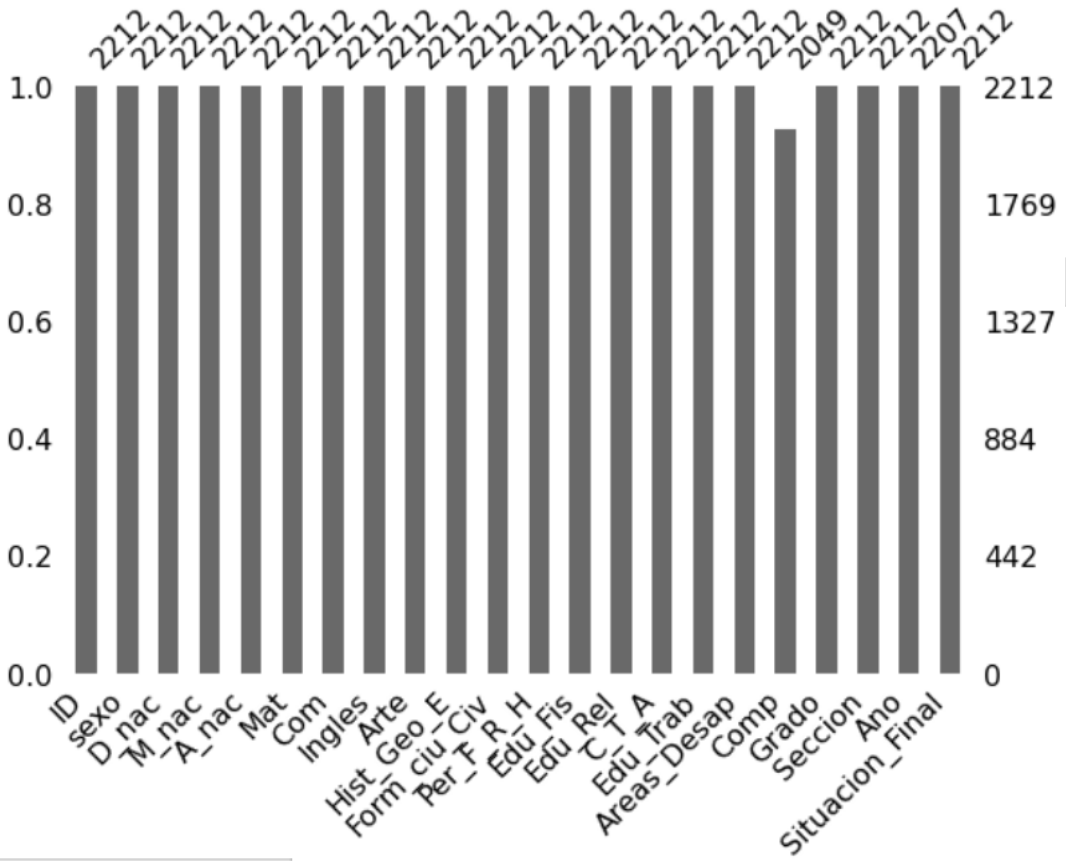


Figura 29: Diagrama de barras de los valores missing de las variables

La variable Comp tiene 163 registros nulos y la variable Año tiene 5 valores nulos.

Como observamos en el punto 3.1 las variables Comp y Ano no presentan completitud en los datos . Para dar solución a dicho problema se utilizó la función de imputación Univariada o Paramétrica de sklearn, SimpleImputer. Se validó la completitud de los datos y el resultado se muestra a continuación.

```
D_nac          0
M_nac          0
A_nac          0
Mat            0
Com            0
Ingles         0
Arte           0
Hist_Geo_E     0
Form_ciu_Civ   0
Per_F_R_H      0
Edu_Fis        0
Edu_Rel        0
C_T_A          0
Edu_Trab       0
Areas_Desap    0
Grado          0
Ano            0
sexo           0
Comp           0
Seccion        0
```

3.2 Selección de variables

Para el proceso de selección de variables, utilizaremos las técnicas Random Forest, WOE y Boruta.

- ✓ ***WOE:*** esta técnica utiliza el Módulo de Python WOE, nos brinda los grados de importancia de cada variable ordenando de mayor a menor grado, como se observa a continuación:

Tabla 31: Importancia de las variables con WOEs

	VAR_NAME	IV
5	Com	1.049382
2	Areas_Desap	0.990238
4	C_T_A	0.930226
17	Per_F_R_H	0.763845
16	Mat	0.760208
14	Ingles	0.742010
3	Arte	0.717455
13	Hist_Geo_E	0.589440
9	Edu_Rel	0.580001
11	Form_ciu_Civ	0.528776
10	Edu_Trab	0.388606
6	Comp	0.302467
8	Edu_Fis	0.290551
19	sexo	0.060043
12	Grado	0.047169
1	Ano	0.010576
18	Seccion	0.007149
15	M_nac	0.005706
0	A_nac	0.004616
7	D_nac	0.000021

- ✓ ***Random Forest:*** esta técnica utiliza del Módulo de Python sklearn, la Función Random Forest Classifier, nos brinda los grados de importancia de cada variable ordenando de mayor a menor grado, como se observa a continuación:

  Tabla 32: Importancia de las variables con Random Forest

	Driver	Importancia
0	Areas_Desap	0.362924
1	Mat	0.112581
2	Com	0.081480
3	Per_F_R_H	0.075257
4	Ingles	0.063377
5	Arte	0.046120
6	C_T_A	0.036402
7	Hist_Geo_E	0.028372
8	Edu_Trab	0.026051
9	D_nac	0.023174
10	Edu_Rel	0.022755
11	Edu_Fis	0.020905
12	M_nac	0.020057
13	Form_ciu_Civ	0.019758
14	Ano	0.018477
15	A_nac	0.016905
16	Grado	0.011966
17	Seccion	0.006501
18	sexo	0.004142
19	Comp	0.002794

- ✓ ***BORUTA:*** para esta técnica se instaló el paquete BORUTA, esta técnica también llamada permutación de árboles brinda el número óptimo de variables para el modelo:

  Luego de aplicar las técnicas de selección de las variables BORUTA trabajaremos el modelo con las siguientes 16 variables:

Variables seleccionadas

```
Tentative:      1
Rejected:       3
Iteration:      10 / 100
Confirmed:      16
Tentative:      1
Rejected:       3
Iteration:      11 / 100
Confirmed:      16
Tentative:      1
Rejected:       3
Iteration:      12 / 100
Confirmed:      16
Tentative:      1
Rejected:       3
Iteration:      13 / 100
Confirmed:      16
Tentative:      0
Rejected:       4


BorutaPy finished running.

Iteration:      14 / 100
Confirmed:      16
Tentative:      0
Rejected:       4
==============BORUTA==============
16
```

1. sexo
2. Mat
3. Com
4. Inglés
5. Arte
6. Hist_Geo_E
7. Form_ciu_Civ
8. Per_F_R_H
9. Edu_Fis
10. Edu_Rel
11. C_T_A
12. Edu_Trab
13. Areas_Desap
14. Comp
15. Grado
16. Año

4. Modelado

Para modelar el presente trabajo de investigación se utilizó 3 algoritmos muy potentes para problemas de Aprendizaje Supervisado de clasificación, es decir, lo que buscó el objetivo. El modelo de **Regresión Logística**, **Árbol CART**, **Random Forest.**

Los indicadores de los modelos aplicados fueron el Accuracy o Precisión Global del Modelo, Precision o Precisión, Recall o Sensibilidad.

El DataSet se dividió en dos para realizar el Entrenamiento o train (67% de la data) y la Validación o test (33% de la data), con la función train_test_split Tanto en el train como en el Test se aplicaron los indicadores antes descritos.

```
[ ]  ################################################
     # Creación de la data de train y la data de test
     ################################################

     # train_test_split (X,y,%y,Estratificar?,Semilla aleatoria)
     from sklearn.model_selection import train_test_split

     X_train, X_test, y_train, y_test = train_test_split(data_imp_f.drop('Situacion_Final',axis=1),
                                                         data_imp_f.Situacion_Final,
                                                         test_size=0.33,
                                                         stratify=data_imp_f.Situacion_Final,
                                                         random_state=100)
```

4.1 Modelo de Regresión Logística

El presente trabajo de investigación tiene por característica un problema de Clasificación, por ello se utilizó el modelo de clasificación de tipo binomial por la naturaleza de la target al cual tiene 2 categorías, por lo tanto se empleó la función ***LogisticRegression*** de sklean.

RESULTADO DE LA EJECUCIÓN: Como resultado de esta ejecución se modeló con la data balanceada del entrenamiento y se aplicó la técnica de Selección de variables.

```
Matriz confusion: Train
[[1103   33]
 [   1  345]]
Matriz confusion: Test
[[540  19]
 [  0 171]]
Accuracy: Train
0.9770580296896086
Accuracy: Test
0.9739726027397261
Precision: Train
0.9770580296896086
Precision: Test
0.9739726027397261
Recall: Train
0.9770580296896086
Recall: Test
0.9739726027397261
```

Figura 30: Matriz de Confusión e indicadores del Modelo Regresión Logística.

4.2 Modelo Árbol CART

El siguiente modelo de clasificación a emplear será el Árbol CART, para ello utilizaremos la función ***DecisionTreeClassifier*** de sklean.

RESULTADO DE LA EJECUCIÓN: Como resultado de esta ejecución se modeló con la data balanceada del entrenamiento y se aplicó Selección de variables.

```
Matriz confusion: Train
[[1088   48]
 [  82  264]]
Matriz confusion: Test
[[531  28]
 [ 51 120]]
Accuracy: Train
0.9122807017543859
Accuracy: Test
0.8917808219178082
Precision: Train
0.9122807017543859
Precision: Test
0.8917808219178082
Recall: Train
0.9122807017543859
Recall: Test
0.8917808219178082
```

Figura 31: Matriz de Confusión e indicadores del Modelo Árbol CART

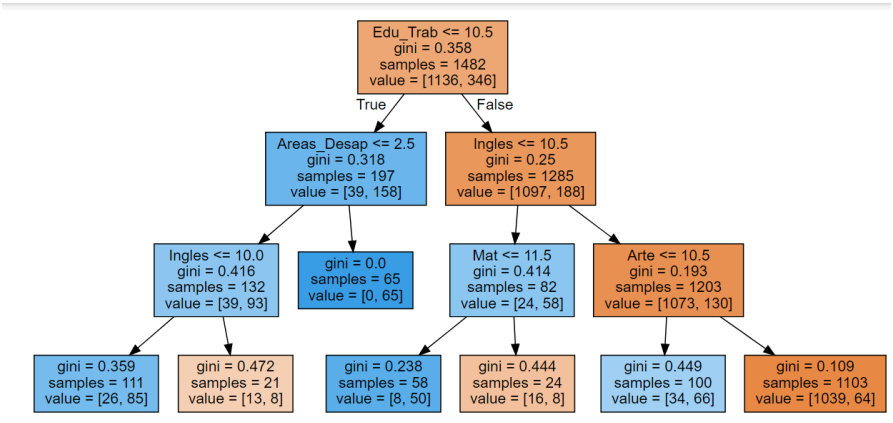


Figura 32: Gráfico del Modelo árbol de decisión

4.3 Modelo de Random Forest

El siguiente modelo de clasificación empleado es Random Forest, para ello utilizaremos la función ***Random Forest Classifier*** de sklean.

RESULTADO DE LA EJECUCIÓN: Como resultado de esta ejecución se modeló con la data balanceada del entrenamiento y se aplicó Selección de variables.

```
Matriz confusion: Train
[[1104   32]
 [   0  346]]
Matriz confusion: Test
[[540  19]
 [  0 171]]
Accuracy: Train
0.9784075573549258
Accuracy: Test
0.9739726027397261
Precision: Train
0.9784075573549258
Precision: Test
0.9739726027397261
Recall: Train
0.9784075573549258
Recall: Test
0.9739726027397261
```

Figura 33: Matriz de Confusión e indicadores del Modelo Random Forest

5. Evaluación

Luego de analizar y comparar los resultados, se observó que el modelo más preciso y más estable es del Modelo de Regresión Logística, para lograr el objetivo de predecir la propensión del término del año escolar de los estudiantes de la Institución Educativa 88331, del centro poblado Rinconada.

ANEXO 11: CONFORMIDAD

MINISTERIO DE EDUCACIÓN
UNIDAD DE GESTIÓN EDUCATIVA LOCAL SANTA INSTITUCIÓN EDUCATIVA Nº 88331—RINCONADA

"Año del Bicentenario del Perú: 200 años de Independencia"

Chimbote 31 de diciembre del año 2021

OFICIO Nº 120 – 2021 /ME/DRE-ANCASH/UGEL-SANTA/I.E. Nº 88331 – DI

Sra: ZENAIDA CRISTINA SHICA JULCA

Presente:

ASUNTO: CONFORMIDAD PARA TRABAJO DE INVESTIGACIÓN

Es grato dirigirme a usted en mi condición de Director de la Institución Educativa 88331, de la provincia del Santa, Departamento de Ancash, para hacer de su conocimiento mi CONFORMIDAD,por el cumplimiento de la entrega de los resultados y el Modelo de Data Science correspondientea su trabajo de investigación el cual se denomina *Modelos de Data Science para la detección de la Deserción Académica, el cual nos servirá para seguir mejorando como Institución Educativa.*

Sin otro particular, auguro los mejores éxitos y parabienes en su investigación.

Atentamente

Mg. Alfredo G. Lozano Cabanillas
DIRECTOR