Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA LA REDUCCIÓN DE LA DESERCIÓN ESTUDIANTIL EN LAS UNIVERSIDADES PRIVADAS DEL PERÚ: CASO UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

Línea de investigación:
Ingeniería de software, simulación y desarrollo de TICs

Tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas

Autor

Guadalupe Mori, Victor Hugo

Asesor

Rodriguez Rodriguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado

Coveñas Lalupú, José

Lira Camargo, Jorge

Petrlik Azabache, Iván Carlo

Lima - Perú

2025

RECONOCIMIENTO - NO COMERCIAL - SIN OBRA DERIVADA
(CC BY-NC-ND)

MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA LA REDUCCIÓN DE LA DESERCIÓN ESTUDIANTIL EN LAS UNIVERSIDADES PRIVADAS DEL PERÚ: CASO UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

INFORME DE ORIGINALIDAD

30%	29%	6%	13%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	bookdown.org Fuente de Internet	2%
2	blogs.upn.edu.pe Fuente de Internet	2%
3	enfoque.upc.edu.pe Fuente de Internet	2%
4	produccioncientificaluz.org Fuente de Internet	1%
5	cybertesis.unmsm.edu.pe Fuente de Internet	1%
6	rstudio-pubs-static.s3.amazonaws.com Fuente de Internet	1%
7	repositorio.untels.edu.pe Fuente de Internet	1%
8	repositorio.uns.edu.pe Fuente de Internet	1%
9	repositorio.utp.edu.pe Fuente de Internet	1%
10	repositorio.unasam.edu.pe Fuente de Internet	1%
11	repositorio.unfv.edu.pe Fuente de Internet	1%

Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

ESCUELA UNIVERSITARIA DE POSGRADO

MODELO PREDICTIVO BASADO EN MACHINE LEARNING PARA LA REDUCCIÓN DE LA DESERCIÓN ESTUDIANTIL EN LAS UNIVERSIDADES PRIVADAS DEL PERÚ: CASO UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

Línea de investigación:
Ingeniería de Software, simulación y desarrollo de TICs

Tesis para optar el Grado Académico de Doctor en Ingeniería de Sistemas

Autor
Guadalupe Mori, Victor Hugo

Asesor
Rodriguez Rodriguez, Ciro
ORCID: 0000-0003-2112-1349

Jurado

Coveñas Lalupú, José

Lira Camargo, Jorge

Petrlik Azabache, Iván Carlo

Lima – Perú

2025

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

Objetivo: Desarrollar un modelo predictivo basado en Machine Learning que contribuya en la determinación de estrategias efectivas para la reducción de la deserción estudiantil en las Universidades Privadas del Perú. **Método**: El estudio tiene un enfoque cuantitativo, se clasifica como un estudio de tipo aplicado, de nivel explicativo, diseño preexperimental de corte longitudinal, tomó como población a los estudiantes de la carrera de Ingeniería de Sistemas de la Universidad San Juan Bautista año 2023, se realizaron diversas pruebas para evaluar la eficiencia, precisión, confiabilidad y efectividad del proceso. La muestra consistió en 104 estudiantes seleccionados aleatoriamente a través de un muestreo probabilístico para análisis, la recopilación de datos se realizó extrayendo de la database proporcionada. **Resultados**: Se resalta la importancia del modelo predictivo desarrollado en cuanto a la precisión, confiabilidad y efectividad para la predicción de la deserción estudiantil en las universidades privadas del Perú, a través de los factores personales, académicos y socioeconómicos, el algoritmo empleado es Bosque Aleatorio (Random Forest) a través de sus diferentes librerías. **Conclusiones**: La precisión del modelo predictivo en la identificación de factores personales en la deserción estudiantil se ha beneficiado significativamente con el uso de herramientas avanzadas de software. Estas plataformas permiten la implementación de algoritmos de aprendizaje automático, la confiabilidad del modelo predictivo para la determinación del impacto de los factores académicos en la deserción estudiantil ha sido evaluada utilizando herramientas como Python, R y otros, que ofrecen una variedad de bibliotecas y funciones especializadas para el análisis estadístico y la modelización de datos, La efectividad del modelo predictivo para cuantificar la influencia de factores socioeconómicos se ha potenciado mediante el uso de herramientas de análisis de datos avanzadas.

Palabras claves: modelo predictivo, machine learning, deserción estudiantil, universidades privadas.

ABSTRACT

Objective: To develop a predictive model based on Machine Learning that contributes to the determination of effective strategies for the reduction of student desertion in Peruvian Private Universities. **Method**: The study has a quantitative approach, it is classified as a study of applied type, explanatory level, pre-experimental design of longitudinal cut, it took as population the students of the Systems Engineering career of the San Juan Bautista University year 2023, several tests were performed to evaluate the efficiency, accuracy, reliability and effectiveness of the process. The sample consisted of 104 students randomly selected through a probabilistic sampling for analysis, the data collection was carried out by extracting from the database provided. **Results**: The importance of the predictive model developed is highlighted in terms of accuracy, reliability and effectiveness for the prediction of student dropout in private universities in Peru, through personal, academic and socioeconomic factors, the algorithm used is Random Forest through its different libraries. **Conclusions**: The accuracy of the predictive model in identifying personal factors in student dropout has benefited significantly with the use of advanced software tools. These platforms allow the implementation of machine learning algorithms, the reliability of the predictive model in determining the impact of academic factors on student dropout has been evaluated using tools such as Python, R and others, which offer a variety of specialized libraries and functions for statistical analysis and data modeling, The effectiveness of the predictive model in quantifying the influence of socioeconomic factors has been enhanced by the use of advanced data analysis tools.

Keywords: predictive model, machine learning, student dropout, private universities.

I. INTRODUCCIÓN

1.1. Planteamiento del problema

La temática por estudiar se relaciona con el sistema educativo y, actualmente, se considera como un problema general que agobia a todas las universidades de cualquier sector no solo en el Perú, sino en todo el mundo, esto debido a múltiples factores e indicadores que involucran la deserción estudiantil universitaria.

Se considera a la deserción estudiantil como una problemática subyacente a la estructura del sistema educativo. De hecho, diversos estudios dan evidencia del impacto de este problema para la permanencia y para la finalización de los estudios. De hecho, en el plano internacional de la educación superior, los datos que se presentaron en el informe Education at the Glace evidenciaron que la deserción existente en los países que pertenecen a la Organización para la Cooperación y Desarrollo Económico (OCDE) alcanzó su pico más alto en los últimos años: un 31 % (Cooperación y Desarrollo Económico [OECD], 2024). En específico, en el caso Europeo de Educación Superior (EEES), que está conformado por un total de cuarenta y siete países, la deserción oscila entre un 20 y 55 %. Además, en Latinoamérica, la tasa de deserción oscila entre 40 y 48%, mediado por problemas estructurales y económicos (Gutiérrez et al., 2021).

La deserción, de acuerdo con Himmel (2002), es la situación que se da cuando un estudiante abandona prematuramente un programa de estudios sin haber alcanzado el título o grado o haber culminado la carrera profesional (Navarro, 2018), la cual afecta la educación a nivel global y Perú no es la excepción, especialmente en el sector universitario (Améstica-Rivas et al., 2020). La deserción estudiantil en la actualidad constituye el principal problema que enfrentan las universidades al momento de validar su oferta educativa en el ámbito de la educación superior. Bajo este contexto, no existe una única razón que lleve a los estudiantes a

desertar, sino más bien es un fenómeno multicausal, las condiciones de cada institución toman relevancia al intentar explicar este fenómeno.

La deserción durante la etapa universitaria se considera como una de las más grandes problemáticas presentes en todos los sistemas de educación superior. Debido a esto, las autoridades universitarias sienten gran preocupación principalmente porque, ante el incremento de la demanda en los sistemas de educación superior, la realidad en la culminación de estudios superiores no es el que se espera ni se relaciona con la demanda. Esto genera graves problemas para las universidades en lo que respecta a su economía, especialmente cuando los abandonos se realizan durante los primeros semestres (Viale, 2014).

Otro aspecto que considerar es el señalado por Núñez (2018) quien resalta que un gran número de universidades son indiferentes ante problemas de índole sexual que pueden sufrir las estudiantes universitarias, problema que se agrava aún más cuando las universidades no cuentan con protocolos de prevención, atención y sanción de acoso sexual protocolo pese a incontables denuncias, indicador clave para la deserción en las universidades. Asimismo Quintero (2020), manifiesta que, debido al hostigamiento y el acoso sexual dentro de la universidad, es necesario establecer protocolos para prevenir, mitigar, investigar y sancionar dichas conductas. Especialmente, en lugar de sancionar, lo que se busca es prevenir la incidencia de actos de acoso sexual, lo cual se logra con protocolos especializados en esta tarea.

La salud mental es considerada de vital importancia para todos, en todo el mundo, pero las respuestas son insuficientes e inadecuadas. Tal como se observa en los datos disponibles más recientes (Organización Mundial de la Salud [OMS], 2022).

En el Perú, las universidades son públicas o privadas. Las primeras son personas jurídicas de derecho público y las segundas son personas jurídicas de derecho privado, la universidad es una comunidad académica orientada a la investigación y a la docencia, que brinda una formación humanista, científica y tecnológica con una clara conciencia del país

como realidad multicultural. Adopta el concepto de educación como derecho fundamental y servicio público esencial. Lo integran estudiantes, graduados y docentes. Además, participan promotores, de acuerdo con ley (Metropolitana & Lima, s.f.).

En el ámbito local, la deserción dentro de la universidad no solo involucra a las instituciones de educación superior, sino que tiene repercusiones y está muy asociado con la sociedad en general (Viera et al., 2020). De hecho, la preocupación en torno a la deserción universitaria ha sido muy grande en los últimos años, especialmente por sus múltiples implicancias, por lo que el desarrollo de técnicas de que reconozcan patrones predictivos de deserción es un elemento crucial para combatir este fenómeno que agrava en el país (Romero et al., 2021).

1.2. Descripción del Problema

Actualmente, se identifican un sinnúmero de causantes de la deserción universitaria en todo el mundo. Para realizar una caracterización en el contexto de las universidades peruanas, aquí se consideraron se tomó en cuenta diferentes factores. En sí, tomando como punto de partida la variable “deserción universitaria”, se clasificaron los factores causantes de esta (esto es sus dimensiones) de la siguiente manera: personales, educativos, institucionales y financieros. (Dávila et al., 2022).

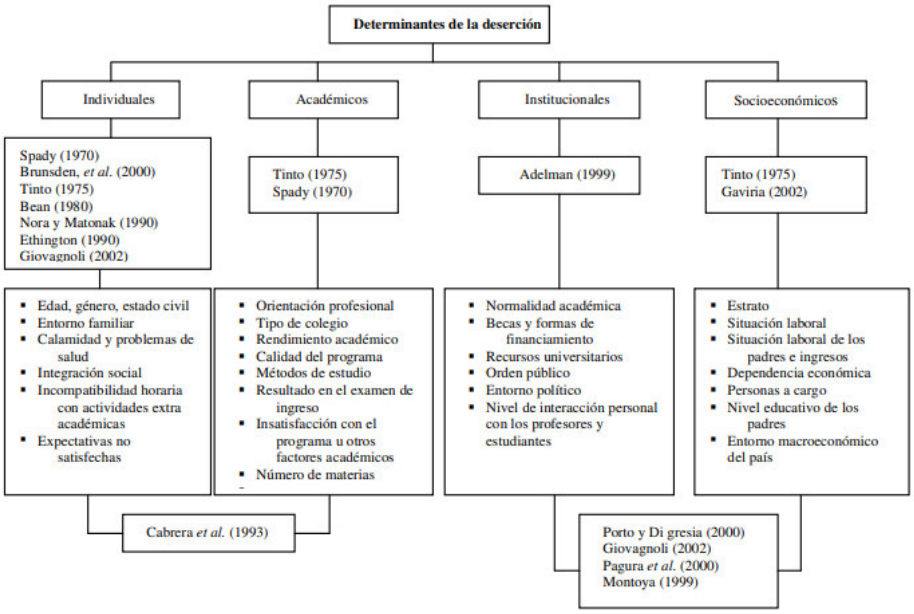
La deserción universitaria es un fenómeno mayormente contemplado que ocurre con el abandono de los estudios antes de que los estudiantes completen un programa de estudios, ya sea en nivel primario, secundario o superior universitario (Pierrakeas et al., 2020). Este problema tiene diversas causas y consecuencias, y puede afectar tanto a nivel nacional como internacional. En ello podemos observar que las universidades nacionales o privadas no son ajenas a este fenómeno y se pudo observar que los factores socioeconómicos, personales, que a menudo es una barrera para acceder y continuar la educación (Lorenzo-Quiles et al., 2023). A esto se le suma conflictos familiares como la desintegración familiar o responsabilidades

domésticas, pueden contribuir a la deserción incluyendo los problemas de salud, la falta de motivación: La falta de interés en el contenido educativo, la desmotivación o la falta de metas claras pueden influir en la decisión de no culminar con la etapa académica (McDermott et al., 2019). Todo ello puede ocasionar el aumento de los problemas sociales como el desempleo, la delincuencia y la falta de participación cívica e impacto en la calidad educativa.

A pesar de que en la actualidad se debata sobre la definición de deserción estudiantil, hay acuerdo en caracterizarla como la renuncia voluntaria que puede ser elucidada mediante las categorías mencionadas en la Figura 1 (Aflalo & Gabay, 2011), donde se muestra los factores determinantes o causantes de la deserción estudiantil.

Figura 1

Determinantes de la deserción estudiantil (estado del arte)

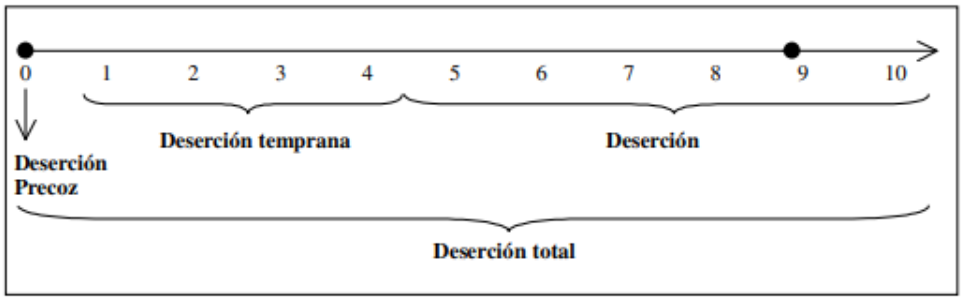


Nota. Tomado de Miranda & Alarcón (2021)

La deserción con relación al tiempo podemos observar que se divide en, deserción precoz, deserción temprana, deserción y deserción total, cada una de ellas con sus respectivas determinantes o factores que involucran. A continuación, en la Figura 2 se muestra la clasificación de la deserción de acuerdo con el tiempo.

Figura 2

Clasificación de la deserción de acuerdo con el tiempo



Nota. Tomado de Miranda & Alarcón (2021)

La predicción de la deserción universitaria (Figura 3), se desarrolla a través de varios indicadores clave (Tabla 1). Este proceso incluye una validación del modelo para asegurar que las predicciones sean precisas y confiables. Los indicadores actuales, como la precisión de la predicción, la tasa de deserción, la validación del modelo y la tasa de retención, son esenciales para evaluar y mejorar continuamente el modelo predictivo.

Figura 3

Proceso de la Predicción de la deserción de estudiantes universitarios UPSJB

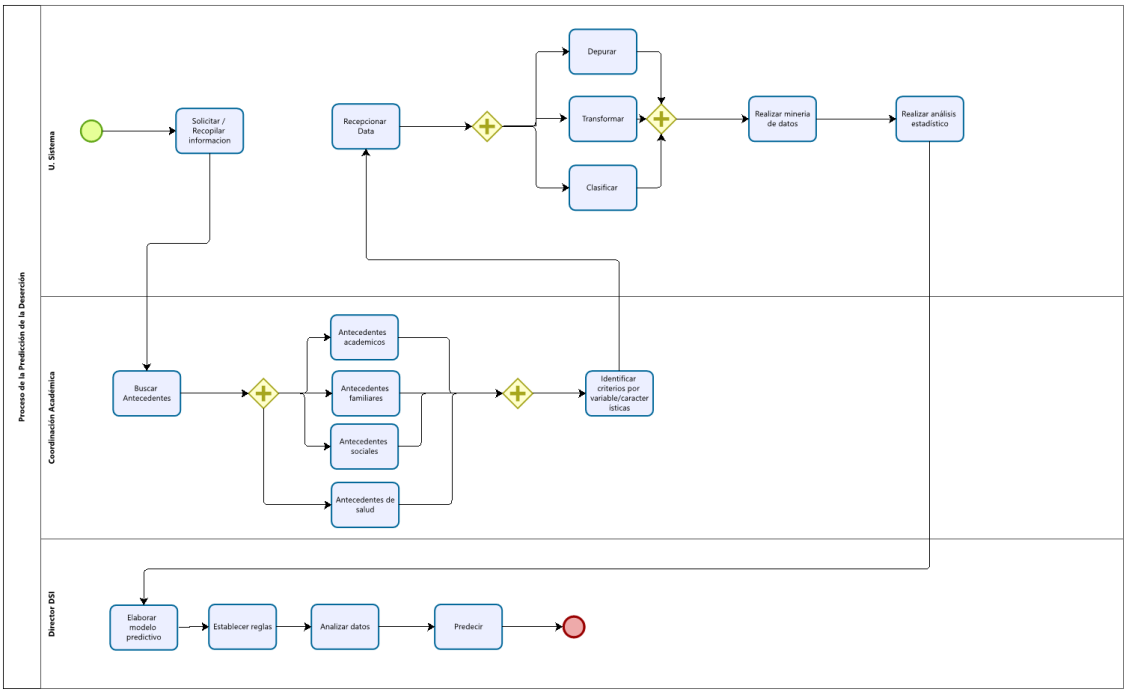


Tabla 1

Datos actuales de los indicadores

Indicadores actuales
Precisión de la predicción
Tasa de deserción
Validación del modelo
Tasa de retención

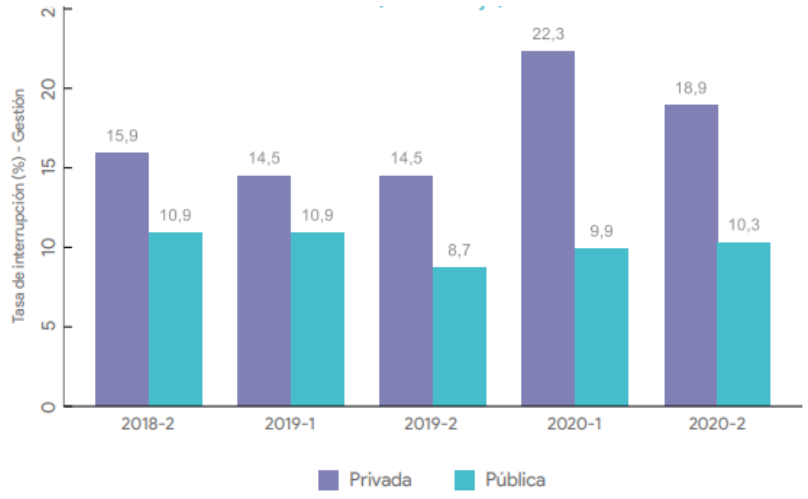
Por otro lado, Marena et al. (2021) nos dicen que el incremento en la interrupción universitaria durante los ciclos de estudios virtuales se ha concentrado básicamente en universidades de gestión privada. Esto se puede comprobar analizando la variación de la tasa de interrupción universitaria según gestión, en relación con su nivel promedio de los últimos tres años. La figura 4 muestra que en los semestres 2020-1 y 2020-2 las tasas de interrupción en universidades privadas estuvieron por encima del promedio obtenido durante los tres semestres previos al inicio de la emergencia sanitaria (15,0%). No ocurre lo mismo en universidades de gestión pública, ya que en el semestre 2020-1 la tasa de interrupción está ligeramente por encima (10,3%), mientras que en el semestre 2020-2 está por debajo (9,6%) del promedio obtenido para los tres semestres anteriores (10,2%).

Asimismo, si se comparan tasas de interrupción de semestres contiguos, se observa que para el semestre 2020-1, estas aumentaron en 7.8 pp. (22,3%) y 1.2 pp. (18,9%) respecto a las del semestre 2019-2 en universidades de gestión privada y pública respectivamente. Con respecto al semestre 2020-1, sólo se observó una caída de 3,4 pp. (a 18,9 %) en el caso de universidades privadas, y un leve incremento de 0,4 pp. (a 10,3%) para universidades de gestión pública. La caída en la tasa de interrupción de universidades privadas y el ligero incremento de

esta en universidades públicas, tienen implicancias en los análisis que desagregan estos dos grupos sobre más categorías en adelante.

Figura 4

Tasa de Deserción Estudiantil durante el periodo 2018 al 2020



Fuente: MINEDU (2021)

Asimismo, utilizando fuentes de información según la interpretación de la ENCUDES (encuesta de deserción universitaria en universidades públicas) y el SIRIES (Sistema de Recolección de Información para la Educación Superior) se pudo manifestar el resumen de factores y sub factores de la deserción estudiantil.

Figura 5

Resumen de factores y sub factores de interrupción

Grupos de factores	Sistema	Nivel	Sub factores	Detalle
Externos	Macro	Estructural	Conectividad y calidad de acceso a internet	Disponibilidad y condiciones del servicio de internet en el lugar de residencia
Internos	Exo	Institucional	Apoyo financiero por parte del Estado	Acceso a bonos del Estado en la población estudiantil
			Características de la universidad	Calidad de la universidad Tipos de carreras brindadas por la universidad Potencial retorno educativo de las carreras
			Acceso a la tecnológica y virtualidad	Equipos electrónicos e internet brindados a la población estudiantil Condiciones de la educación virtual
			Servicios complementarios	Adaptación al entorno virtual por parte de la universidad. Servicios de tutoría y acompañamiento
	Meso	Relacional	Vínculo docente y estudiante	Experiencia del vínculo docente estudiante
Personales	Micro	Personal	Características de las y los estudiantes	Nivel socioeconómico del hogar Edad Sexo
			Economía, empleo, carga familiar y conciliación de la vida personal, académica y laboral	Shocks económicos Necesidad de búsqueda de empleo Afectación y desigualdades en el equilibrio la actividad académica y laboral y personal
			Recursos electrónicos	Disponibilidad de recursos electrónicos para las clases virtuales, así como de internet en el hogar
			Salud física, salud mental, violencias y género	Bienestar físico y emocional Presencia de Violencias y desigualdades de género
			Elección de carrera universitaria	Adecuada orientación vocacional
			Rendimiento académico	Desempeño académico del estudiante

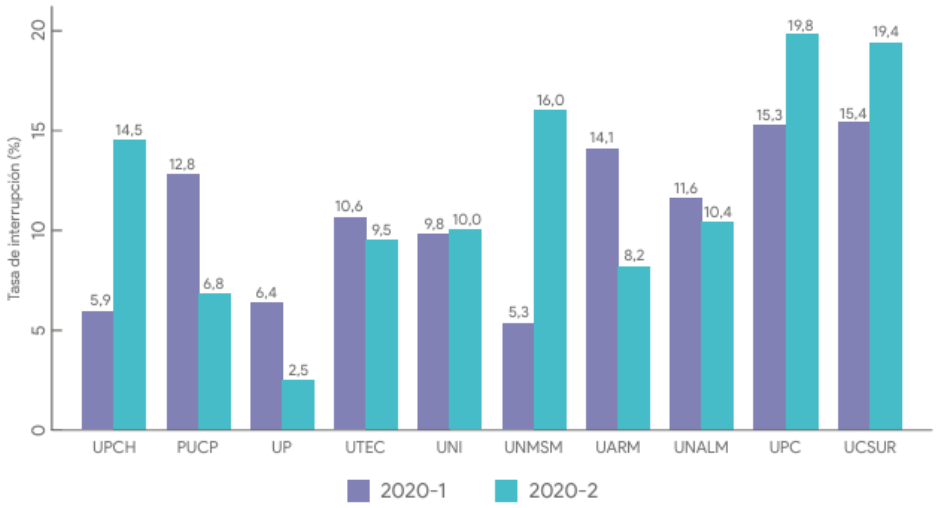
Fuente: MINEDU (2021)

No obstante, Con respecto al incremento en la tasa de interrupción para el grupo de universidades del top 10, esta se explica en parte porque la mitad de las universidades de este grupo presentaron incrementos en sus tasas de interrupción, dentro de las que destacan la Universidad Nacional Mayor de San Marcos (UNMSM) con 10,7 pp., seguido de la Universidad Peruana Cayetano Heredia (UPCH) con 8,6 pp. Asimismo, es importante considerar que 7 de las 10 universidades del top 10 poseen un tipo de gestión privada, las cuales

reflejan frecuentemente mayores tasas de interrupción en comparación a las obtenidas en las universidades públicas.

Figura 6

Tasa de Interrupción desde el 2018 al 2021



Fuente: MINEDU (2021)

1.3. Formulación del Problema

Problema General

¿Cómo puede un modelo predictivo basado en Machine Learning contribuir a la identificación de estrategias efectivas para reducir la tasa de deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?

Problemas Específicos

- ¿De qué manera un modelo predictivo basado en Machine Learning mejora la eficiencia en la predicción de los factores personales, académicos y socioeconómicos que influyen en la deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?

- ¿Cómo la precisión de un modelo predictivo basado en Machine Learning facilita la identificación de los factores personales que afectan la deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?
- ¿En qué medida la confiabilidad de un modelo predictivo basado en Machine Learning contribuye a determinar el impacto de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?
- ¿Cómo la efectividad de un modelo predictivo basado en Machine Learning permite cuantificar la influencia de los factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?

1.4. Antecedentes

1.4.1. Internacionales

Franco (2019) en su estudio "Implementación de un Modelo Computacional basado en Reglas de Clasificación Supervisadas para la Predicción de la Deserción Estudiantil en la Universidad Peruana Unión Filial Juliaca", evaluó los modelos de árboles de decisión durante el análisis de los comportamientos de los universitarios. Su propósito fue investigar e implementar una herramienta de detección temprana de la deserción universitaria. Para ello, empleó Cross Industry Standard Process for Data Mining (CRISP-DM) como método. Asimismo, empleó las notas de estudiantes de Ingeniería de Sistemas Seccional Bogotá, ya que fue en este programa que se identificaron posibles casos de pérdidas de materia y posterior deserción. Demostró que desarrollar este tipo de modelos utilizando diferentes librerías bibliotecas de Python como lo son SkLearn y Pandas-Profiling optimiza los recursos desde la planeación, ejecución y finalización del proyecto. Concluyó que, aunque existen múltiples

marcos de trabajo, CRISP-DM resulta bastante conveniente y sencilla de implementar en proyectos donde se deben hacer múltiples iteraciones en búsqueda de resultados.

Quintero (2022) en su investigación “Diseño de un modelo predictivo para generar alertas tempranas de deserción universitaria en los programas de pregrado presenciales de la Facultad de Ingeniería de la Universidad de Antioquia”, usaron aprendizaje de análisis en la creación de un modelo para identificar el riesgo de deserción temprana en la Universidad de Antioquia (programa de ingeniería). Para ello, empleó información histórica de la universidad e información del Instituto Colombiano para la Evaluación de la Educación; además de redes neuronales artificiales (RNA) y Xtreme gradient boosting (XG Boost). Los resultados muestran que con XGBoots se pudo lograr un modelo que puede realizar la clasificación precisa (74.91%) de los estudiantes que corren el riesgo de desertar. Asimismo, en cuanto a la identificación de desertores, este se debe de mejorar para que la identificación y precisión alcance un 100%.

Acosta (2019) en su tesis “Definición de un modelo predictivo para la deserción estudiantil en educación virtual y distancia”, se propuso construir el modelo predictivo usando técnicas estadísticas y de minería de datos que permitiera identificar la deserción estudiantil para los programas de educación virtual y a distancia. Para ello, comprobó y clasificó los datos en cada uno de los dominios (individuales, socioeconómicos, académicos, institucionales) y los cuales son determinantes en la deserción. Se logró realizar el análisis de información de deserción a aplicando técnicas estadísticas y de minería de datos, logrando predecir con una exactitud del 75% la deserción de los estudiantes por medio de la técnica estadística de regresión logística binaria, además se logra identificar los factores y variables que inciden en la deserción.

Rodriguez (2023) en su investigación “Aplicación de algoritmos de Machine Learning para predecir la deserción estudiantil en alumnos de primer y segundo semestre en

universidades públicas del Ecuador", aplicó técnicas de ML en la predicción de deserción universitaria según factores psicológicos, socioeconómicos, académicos y demográficos. Empleó RNA en la creación de modelos que clasifican estudiantes como desertores o salvos de deserción. Asimismo, evaluó la exactitud, especificidad y sensibilidad para calcular la especificidad del modelo. Sus resultados evidenciaron que las redes neuronales permitieron clasificar los estudiantes como desertores o posibles desertores. Los universitarios que no estuvieron en riesgo fueron catalogados de forma precisa (97%), mientras que los universitarios que si estuvieron en riesgo de desertar se predijeron de forma efectiva, aunque en menor porcentaje (75%). Luego, el modelo clasificó a un no desertor de forma correcta (86%) y a un universitario desertor en menor medida (60%). Finalmente, el modelo fue exacto (79%) y tuvo un F1-Score de 0.62.

Masabamba (2019) en su tesis "Modelo basado en minería de datos para determinar factores de deserción estudiantil en la facultad de ciencias de la ingeniería y aplicadas de la Universidad Técnica de Cotopaxi", propuso un modelo capaz de identificar factores de deserción universitaria con de minería y su influencia predictiva. El proceso experimental se basa en una encuesta en línea aplicada a 1457 estudiantes de la Facultad de Ciencias de la Ingeniería y Aplicadas de las Carreras de Ingenierías: Eléctrica, Sistemas de Información, Electromecánica e Industrial. La metodología aplicada corresponde a Knowledge Discovery in Databases (KDD). Los resultados encontrados permiten determinar que los factores: conducta en el aula de clases, bullying, motivación del docente – alumno, limitado conocimiento de la asignatura, adicción de las redes sociales, estado emocional, conocimiento adquirido en los cursos de nivelación, formación académica, sistema de ingreso a la universidad, problemas familiares, residencia y expectativas respecto a la carrera seleccionada, son los factores que tienen mayor influencia en la deserción de los estudiantes en. Mientras que las técnicas de minería de datos J48, Random Forest y Sequential Minimal Optimization (SMO), dieron como

resultado una tasa de predicción de la deserción del 92%. Se concluye que el uso de técnicas de minería de datos puede ser consideradas como importantes para realizar estudios de las causales que afectan a los estudiantes en su permanencia estudiantil universitaria.

Finalmente, Torres (2022) en su tesis "Modelos para la Predicción de Deserción Universitaria de Estudiantes de Psicología de la Universidad el Bosque", utilizaron los modelos de clasificación supervisada Random Forest (bosque aleatorio) y el XGBoost con el propósito de predecir deserción universitaria de los estudiantes de la carrera de Psicología de la Universidad El Bosque, utilizando información académica, demográfica, socioeconómica y de personalidad de estos. Sus resultados evidenciaron que el modelo de Random Forest obtuvo buenos resultados, por lo que fue una herramienta precisa para la predicción de deserción universitaria. El XGBoost obtuvo los resultados esperados, se creía que sería un modelo más potente que el Random Forest, afirmación que fue comprobada por los buenos resultados. Al comparar el Random Forest con el XGBoost, se esperaba que el XGBoost fuera mucho mejor en todos los aspectos, las métricas que se mencionaron en la metodología sugieren que así fue, por lo tanto, se concluye que para este trabajo el XGBoost, se considera la mejor opción para predecir deserción universitaria.

1.4.2. Nacionales

Caselli (2021) en su tesis "Modelo predictivo basado en machine learning como soporte para el seguimiento académico del estudiante Universitario", buscó contribuir en la búsqueda de una solución a través de la IA, ML y DL con las limitaciones de la calidad y la cantidad de la data colectada. Para ello, seleccionó los atributos más relevantes para proponer un modelo de predicción de aprendizaje profundo. Implementó un modelo inicial de red neuronal de 2 capas y se compararon con modelos alternos de 3, 4, 5, 6 y 7 capas con cantidades variables de neuronas entre ellos, los cuales fueron evaluados a través del ratio de precisión del conjunto de entrenamiento y de prueba. Con esto, obtuvo un modelo capaz de tener una precisión de

predicción de 98.97%, lo cual coadyuvará en el seguimiento eficiente a los estudiantes y poder de manera temprana orientar a los estudiantes con perfil de riesgo de abandono temporal o permanente de la carrera a conseguir sus metas, teniendo en cuenta que la variable que mayor incidencia tuvo fue el número de semestres cursado por el estudiante.

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

Jacob (2019) en su estudio "Implementación de un Modelo Computacional basado en Reglas de Clasificación Supervisadas para la Predicción de la Deserción Estudiantil en la Universidad Peruana Unión Filial Juliaca", se propuso implementar un modelo computacional que permita identificar en forma automática a los estudiantes con mayor riesgo de deserción en la Universidad Peruana Unión Filial Juliaca. Adoptó la metodología CRISP-DM que estructura el proceso de minería de datos en seis fases, que interactúan entre ellas de forma iterativa. Se aplicó el modelo de clasificación de ML, para analizar el comportamiento de los estudiantes, evaluando factores como cantidad de cursos matriculados, cantidad de cursos aprobados, si es

independiente o dependiente con respecto al pago de sus estudios, si tiene o no sanción disciplinaria por parte de Bienestar Universitario, cantidad de cursos desaprobados durante el semestre, cantidad de cursos desaprobados dos veces, cantidad de cursos desaprobados de tres veces a más, cantidad de créditos aprobados, cantidad créditos desaprobados, ponderado final del semestre, si la situación del alumnos es regular o irregular, si tiene un saldo a favor o en contra.

Perez (2020) en su tesis "Diseño de un sistema para predecir la deserción de los alumnos mediante Machine learning en la Universidad Tecnológica del Perú", se propuso identificar patrones de comportamiento que son de gran importancia para la Universidad Tecnológica del Perú, debido a que, al tener identificado a los estudiantes con intención de desertar de sus estudios, les permitirá plantear estrategias que permitan disuadir al estudiante y orientar sobre su situación académica. La metodología utilizada fue CRISP-DM y el algoritmo (SVM), Concluyó que el algoritmo SVM identifica los factores con mayor influencia de deserción estudiantil debido a que utiliza como base de aprendizaje la información de alumnos que ya desertaron y permite predecir los patrones de comportamiento y por lo tanto el factor con mayor influencia. Además, la implementación del SVM interviene fuertemente en la predicción de los alumnos de la UTP permitiendo que los encargados del área de Gerencia de retención puedan intervenir para disuadir la decisión del alumnado. Por último, se obtuvo un grado de certeza superior a 90% superando el grado base que se trazó en la hipótesis de esta investigación.

Tapia (2021) en su estudio "Modelo predictivo de clasificación basado en aprendizaje automatizado para la detección temprana de posibles estudiantes universitarios desertantes", aplicó minería de datos, la metodología CRISP-DM y técnicas como: remuestreo, variables ficticias, entre otras. Para el proceso de clasificación se aplicaron algoritmos basados en métodos supervisados. Los datos utilizados pertenecen a estudiantes universitarios, los cuales están basados en los factores del rendimiento académico como: factores sociofamiliares y

factores académicos. Al utilizar el conjunto de datos disponible y al aplicar aprendizaje automatizado, fue posible predecir de manera favorable aquellos estudiantes con probabilidad de deserción, así como comprobar que el clasificador basado en bosques aleatorios obtuvo mejores resultados frente a los demás propuestos. Se logró desarrollar un modelo predictivo que permite clasificar y predecir con mejores resultados generales a aquellos alumnos con tendencia a deserción siendo óptimos los modelos de ensamble basados en Bosques Aleatorios (RamdomForest) cumpliendo con el objetivo principal de la presente investigación.

Finalmente, Pando (2020) en la tesis "Aplicación de un modelo de minería de datos para identificación de patrones que influyen en la deserción académica en el Instituto Superior Leonardo Davinci usando IBM SPSS MODELER y la metodología CRISP-DM", creó un modelo de minería de datos que nos lleva a obtener patrones que influyen en un estudiante desertor. La presente modelo se implementó a través del análisis de la información: personal, académica y de la interacción de los estudiantes. En el análisis y preparación de datos su procesaron un total de 599 registros de los estudiantes recolectados en un archivo de datos de IBM SPSS Statistics llamado Estudiantes. sav, de donde se determinó usar un 80% del total de datos para el Entrenamiento que sirvió como entrada de datos al modelo de minería propuesto.Se diseñó, construyó y aplicó 04 Modelos de Minería de datos como son: Árbol C&R, Árbol C5.0, Árbol AS y Red Bayesiana utilizando IBM SPSS Modeler Subscription, Después de evaluar los 04 modelos implementados, obtuvimos que el modelo de árbol C5.0 nos da un 94.2% de datos correctos y con una precisión similar de 94.203% (ver pág. 91), siendo el modelo por utilizar para analizar los patrones que repercuten en la deserción académica de los estudiantes.

1.5. Justificación de la investigación

La presente investigación se enfoca en generar un modelo predictivo para la reducción de la deserción estudiantil mediante técnicas de clasificación y selección de características, con

el fin de aportar en el desarrollo en esta área del conocimiento, especialmente en el ámbito educacional.

1.5.1. Justificación teórica

La justificación teórica es un elemento fundamental en las investigaciones porque explica el respaldo conceptual y los fundamentos científicos sobre los cuales se desarrollan las investigaciones. Lo que busca es contextualizar los problemas en un contexto particular a través de teorías, modelos y enfoques previos que permiten llegar a una comprensión de la problemática (Rowland, 2018).

Los problemas para predecir la deserción universitaria son graves. En la literatura se evidencia que, al menos durante los últimos años, se ha incrementado la producción científica que busca estudiar los factores de la deserción. Debido a esto, se necesita predecir la deserción universitaria desde una perspectiva conceptual y técnica amplia, de modo que se considera las diferencias dimensionales que aportan a los modelos diseñados para tratar este problema y reducir su impacto.

Por lo expuesto, esta investigación busca diferenciarse de las otras investigaciones, anticipando una teoría respecto a la deserción antes que se realice o suceda este problema, para poder tomar las medidas necesarias y correctivas.

1.5.2. Justificación práctica

La justificación práctica explica cuán relevante y cómo se pueden aplicar los resultados obtenidos que serán obtenidos en un estudio. En esta sección se explica cómo los hallazgos pueden utilizarse en situaciones reales, la resolución de problemas o causar algún impacto social, económico o tecnológico (Bowen et al., 2020).

Esta investigación ayudará a conocer los factores relevantes en la deserción mediante un modelo predictivo los cuales permiten anticipar eventos al identificar patrones dentro de conjuntos de datos complejos (Smith & Lee, 2020) ayudando a tomar acciones en combatir los

factores influyentes más relevantes y para que el alumno continúe sus estudios académicos, ayudará a las Universidades a controlar la reducción de estudiantes aplicando diferentes estrategias y los estudiantes puedan cumplir con su objetivo trazado.

1.5.3. Justificación metodológica

Un estudio se justifica bajo la perspectiva metodológica cuando propone un nuevo método con el que es posible obtener conocimiento válido para dar respuesta a los problemas propuestas en la etapa inicial del estudio. Además, también se justifica metodológicamente cuando se crean o plantean nuevos instrumentos para recolectar la información necesaria para el estudio (Fernández Bedoya, 2020).

Este estudio se justificó desde el punto de vista de su diseño: al ser preexperimental, el estudio buscó aplicar los aspectos teóricos desarrollados en este estudio con el propósito de solucionar la problemática de la deserción estudiantil a través de un modelo predictivo. Además de ello, en este estudio se elaboró un cuestionario para recolectar la información requerida para cumplir con los objetivos propuestos en esta investigación. De esta manera tanto el diseño como el instrumento empleado en el estudio justificaron la realización de este estudio en aras de abordar la problemática identificada.

1.5.4. Justificación social

La justificación social permite explicar cuán relevante es un estudio para resolver problemas sociales, económicos o ambientales. De esta manera, este tipo de justificación sustenta la importancia de una acción, estudio o proyectos en términos de su impacto positivo para la sociedad (Fernández Bedoya, 2020).

La deserción escolar es el abandono definitivo de los estudios por parte de un estudiante antes de completar su periodo académico y es ocasionada por factores económicos, familiares o socio-contextuales (Del Bonifro et al., 2020). Así, este estudio buscó contribuir socialmente

a los problemas de deserción estudiantil en la universidad objeto de estudio a través del modelo predictivo desarrollado.

1.6. Limitaciones de la investigación

La gran mayoría de las personas piensan que el mencionar las limitaciones de la investigación disminuyen su relevancia. Sin embargo, el reconocer las restricciones conlleva a tener más rigurosidad y validez porque se ha realizado un análisis minucioso interna y externamente. (Avello et al., 2019).

Muchas Universidades a nivel mundial, no cuentan con un modelo o herramienta para la predicción y que aporte a la reducción de la deserción estudiantil universitaria, esto da lugar a que no se pueda tomar las medidas necesarias y correctivas antes que suceda el problema.

Otra de las limitaciones, es no haber encontrado estudios previos en referencia a un modelo predictivo para la reducción de la deserción estudiantil con el propósito de analizar el alcance de las investigaciones realizadas.

1.7. Objetivos

Objetivo general

Desarrollar un modelo predictivo basado en Machine Learning que contribuya en la determinación de estrategias efectivas para la reducción de la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.

Objetivos específicos

- Construir la arquitectura de un modelo predictivo basado en Machine Learning que contribuya a la eficiencia en la predicción de los factores personales, académicos y socioeconómicos que influyen en la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.

- Examinar la precisión del modelo predictivo basado en Machine Learning en la identificación de factores personales que influyen en la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.
- Verificar la confiabilidad del modelo predictivo basado en Machine Learning para la determinación del impacto de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.
- Evaluar la efectividad del modelo predictivo para la cuantificación de la influencia de factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.

1.8. Hipótesis

Hipótesis general

El desarrollo de un modelo predictivo basado en Machine Learning tiene un impacto significativo en la reducción de la tasa de deserción estudiantil en la Universidad Privada San Juan Bautista, al identificar estrategias efectivas de retención.

Hipótesis específicas

- La implementación de una arquitectura de modelo predictivo basado en Machine Learning mejora la precisión en la identificación de los factores personales, académicos y socioeconómicos que influyen en la deserción estudiantil en la Universidad Privada San Juan Bautista.
- La precisión del modelo predictivo basado en Machine Learning es significativamente alta para identificar los factores personales que contribuyen a la deserción estudiantil en la Universidad Privada San Juan Bautista.
- La confiabilidad del modelo predictivo basado en Machine Learning permite una evaluación precisa del impacto de los factores académicos en la deserción estudiantil en la Universidad Privada San Juan Bautista.

- La efectividad del modelo predictivo basado en Machine Learning permite cuantificar la influencia de los factores socioeconómicos en la deserción estudiantil en la Universidad Privada San Juan Bautista y facilita la formulación de estrategias financieras y de apoyo social para mejorar la retención estudiantil.

II. MARCO TEÓRICO

2.1. Estado del arte

2.1.1. La deserción educativa

Aproximación en el contexto de la deserción educativa. En el segundo Informe Bienal sobre la realidad de las universidades peruanas (Superintendencia Nacional de Educación Superior Universitaria [SUNEDU], 2020), se evidencia que entre 2012 y 2018 hubo un alza de los estudiantes universitarios (de entre 25 y 29 años) que abandonaron sus estudios en todo el país (de 15.8 a 17.6%). De hecho, el índice de abandono en Callao y Lima Metropolitana (13.4) no fue superior al promedio nacional. Los valores son alarmantes si se evalúa la deserción universitaria según las regiones del país: selva (24.6%), costa (24%) y sierra (18.2%).

De los factores influyentes para la deserción educativa. Viloria et al. (2019) propusieron un modelo que tenía en cuenta tanto a estudiantes como la misma universidad para identificar los factores detrás del abandono escolar y clasificar aquellos aspectos predictores de comportamientos asociados a la deserción. Los autores consideraron características del estudiante, sus metas y sus compromisos previo al ingreso de la universidad, así como las experiencias hacia la universidad y la integración con sus compañeros de clase. Además de ello, analizaron las variables predictoras de deserción al cuantificar datos y recibir apoyo de recursos informáticos.

Valero et al. (2022) mencionan que en los centros universitarios se tiene información del perfil de los estudiantes desde el momento de su inscripción y se cuenta con información de interés, además en el transcurso de los semestres académicos también van registrando más datos importantes para el análisis, como datos económicos, académicos y de salud. De acuerdo con Perchinunno et al. (2021), las razones del abandono pueden ser muy variadas, como falencias en la orientación vocacional de futuros estudiantes, la necesidad real de un trabajo

que les permita superar su realidad o no conocer cuán importantes son los estudios para lograr superarse.

En el estudio realizado por Aldowah et al. (2020), se identificaron seis factores que influyen de forma directa en el abandono estudiantil en etapa universitaria: habilidades y capacidades académicas, experiencias anteriores, la forma en que se diseñaron los cursos, procesos de retroalimentación, y apoyo social o académico. También encontraron que la motivación, la dificultad, la interacción, la duración del curso, el compromiso, y las circunstancias familiares o laborales desempeñaban un papel secundario en el abandono estudiantil.

2.1.2. Estrategias para disminuir la deserción educativa

Modelos predictivos. Los modelos predictivos utilizan datos históricos y actuales, combinados con técnicas estadísticas y algoritmos avanzados, para predecir futuros comportamientos y resultados (Fernandez-Felix et al., 2021). Un modelo predictivo combina algoritmos matemáticos y datos históricos para estimar eventos futuros de manera precisa y confiable (Karvelis et al., 2023). Los modelos predictivos constituyen instrumentos estadísticos concebidos para examinar las características individuales de los estudiantes y pronosticar resultados académicos, tales como el rendimiento o el riesgo de deserción escolar (Vieira et al., 2022). Los modelos predictivos implementados en el ámbito educativo detectan patrones de comportamiento estudiantil, empleando el análisis de datos para prever riesgos de deserción y formular estrategias de retención (Cedeño & Cedeño-Valarezo, 2023).

Modelos predictivos para disminuir la deserción educativa. Sifuentes (2018), determina los siguientes modelos para deserción educativa, la técnica de minería de datos se revisaron los patrones de rendimiento académico de los estudiantes, tanto de los aprobados como desaprobados, para poder determinar los cursos en los que se trabajarían los modelos predictivos.

Lykourentzou et al. (2009) utilizaron los datos de Moodle para predecir si un estudiante abandonará los cursos electrónicos de informática. Los autores utilizaron diferentes algoritmos de clasificación junto con información básica, actividad en la plataforma, fechas de entrega y calificaciones de los estudiantes para lograr sus predicciones de abandono. Las tasas de recuperación y precisión alcanzaron el 95% y el 82%, respectivamente. Un recuerdo del 82%, por ejemplo, significa que se detectó el 82% de los posibles desertores, mientras que una precisión del 92% significa que el 92% de los estudiantes predichos como desertores, de hecho, abandonaron sus estudios.

La metodología CRP (Cross-Industry Standard Process for Data Mining), la cual incluye seis etapas como proceso de modelamiento para determinar una selección de las mejores variables predictoras del buen rendimiento Académico. Se aplicó el método de crecimiento árboles de clasificación y regresión (Classification and Regression Trees), para dividir los datos en segmentos con la finalidad de que sean lo más homogéneos posibles respecto a la deserción para el análisis de clasificación exploratorios y confirmatorios.En ese sentido, existen muchos modelos de Machine Learning, como, por ejemplo: Regresión logística; Árboles de decisión, y K-Means. Así, la regresión logística, es una herramienta muy versátil para realizar una clasificación de varias clases. Para graficar la regresión se tiene una curva con forma de S, esto permite dividir los datos en grupos (Microsoft Azure, 2022).

El estudio de Márquez-Vera et al. (2016) utilizó datos de registros académicos de 32 342 estudiantes, con un período de doce años y medio. Los resultados mostraron que el número de semestres que los estudiantes pasan en la universidad hace que el abandono sea menos predecible. Además, los resultados fueron volátiles, ya que se observaron tasas de precisión nulas en escenarios específicos, mientras que se alcanzaron puntuaciones de recuperación de hasta el 82.4%.

Maching Learning para aplicación de deserción educativa. Denle (2010), menciona que Machine Learning es parte de la ciencia de los datos que contiene distintos tipos de algoritmos de inteligencia artificial; por otra parte, existen dos tipos de aprendizajes los supervisados y los no supervisados, esto es importante para el aprendizaje de la máquina, es por ello que el objetivo del estudio se orienta a obtener un modelo matemático que ayude a predecir la deserción universitaria.

Krüger et al. (2023) propusieron un enfoque para crear y enriquecer un conjunto de datos para la predicción de la deserción, que se ha aplicado para la predicción de la deserción utilizando datos de 19 escuelas en Brasil. Con este conjunto de datos y utilizando clasificadores y técnicas de explicación de modelos, sus algoritmos obtuvieron puntuaciones AUC-PR que oscilaron entre el 38,22% y el 89,50% al predecir el abandono escolar en diferentes momentos (trimestres) del año escolar.

2.1.3. Resultados obtenidos aplicando modelos predictivos

Modelos utilizados. Sifuentes (2018) informa el siguiente resultado obtenido tras aplicación del modelo predictivo, indicando que las estadísticas de desaprobación de los últimos años muestran que se ha producido un ligero incremento en ese indicador, pero que a partir de la implementación de los modelos predictivos creados para los siete cursos calificados como críticos, así como la aplicación de estrategias de apoyo al estudiante, se ha reportado una disminución en el nivel de desaprobación, estimado como uno de los factores de la deserción.

Chung & Lee (2019) exploraron la posibilidad de utilizar el modelo de bosques aleatorios en el aprendizaje automático para predecir el abandono escolar de los estudiantes. El modelo de bosques aleatorios mostró un excelente desempeño en la predicción del abandono escolar de los estudiantes en términos de varias medidas de desempeño para la clasificación binaria. Sus resultados revelaron que el modelo predictivo predice la deserción escolar con una excelente precisión de 0.95. Asimismo, La sensibilidad de 0.85 es relativamente menor que la

especificidad de 0.95, lo que indica que nuestro modelo de bosques aleatorios es mejor para predecir la no deserción que para predecir la deserción

Modelos basados en Machine Learning. Valero et al. (2022) obtuvo los siguientes resultados aplicando Machine learning; El algoritmo K-Nearest-Neighbor con una precisión de 0,91 tiene mejor desempeño para pronosticar la deserción universitaria con las variables académicas y socioeconómicas de los estudiantes. Si sólo se evaluara las variables académicas y no las socioeconómicas el algoritmo KNN tendría una precisión de aproximadamente 0,88, esto es el mejor desempeño para pronosticar la deserción universitaria sólo con las variables académicas. Asimismo, se evidenció que uno de los factores de riesgo más recurrentes entre los estudiantes que abandonaron sus estudios, es el bajo rendimiento académico 43,4%, durante su permanencia en la universidad. Por lo cual, el modelo obtenido puede ayudar a predecir en los primeros ciclos de estudio, qué alumnos son más probables de abandonar sus estudios, al mismo tiempo, alertar a la oficina de bienestar, la necesidad y atención de tutorías individuales, así como grupales.

Para predecir el abandono escolar de los estudiantes en el Instituto Tecnológico de Karlsruhe (KIT), Kemper et al. (2020) utilizaron dos métodos de aprendizaje automático, regresiones logísticas y árboles de decisión. Los modelos se calculan a partir de datos de exámenes, es decir, datos disponibles en todas las universidades sin necesidad de una recopilación específica. Asimismo, propusieron un enfoque metódico que se puede poner en práctica con relativa facilidad en otras instituciones. Observaron que los árboles de decisión producen resultados ligeramente mejores que las regresiones logísticas. Sin embargo, ambos métodos arrojan altas precisiones de predicción de hasta el 95 % después de tres semestres. Una clasificación con una precisión de más del 83 % ya es posible después del primer semestre.

2.2. Bases teóricas

2.2.1. Inteligencia artificial

La inteligencia artificial (IA) se trata de una disciplina informática enfocada en la creación de sistemas con la capacidad de realizar tareas que, cuando son ejecutadas por el ser humano, requieren de su inteligencia. Así, con la IA, las máquinas pueden realizar tareas como el aprendizaje, el razonamiento y la adaptación a nuevas situaciones. Además, la IA puede automatizar el aprendizaje y descubrimiento repetitivos a través de datos, lo que lo diferencia de la automatización de robots (Acosta, 2019).

Los objetivos de la IA abarcan desde la emulación de procesos cognitivos humanos hasta la resolución de problemas complejos el ser humano, debido a sus capacidades limitadas, no puede resolver. De esta manera, la IA es una tecnología que permite que las computadoras simulen la inteligencia y capacidades humanas en busca de resolver problemáticas. Esto implica el desarrollo de sistemas que tienen la capacidad de aprendizaje mediante la experiencia, de interpretación de información y la de toma de decisiones fundamentadas en datos (Aljaber & Almushaili, 2022).

La IA se caracteriza por la facilidad con la que puede procesar grandes volúmenes de datos, identificar patrones en ellos (cosa que el ser humano no puede) y adaptarse a nuevas informaciones (Favaretto et al., 2020). Para ello, la IA se adapta mediante algoritmos de aprendizaje progresivo que permiten a los datos realizar la programación y hallar estructuras y regularidades que le permiten adquirir habilidades específicas.

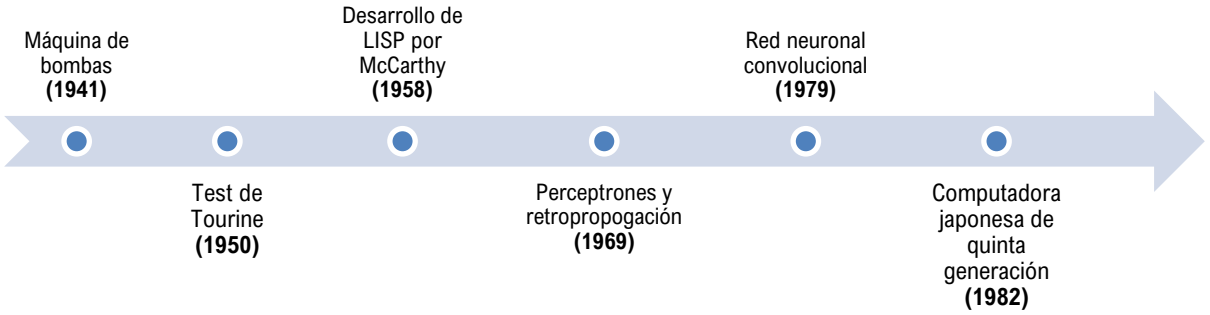
La evolución histórica de la IA ha estado marcada por diversos enfoques y paradigmas. Inicialmente, el desarrollo de la IA estuco enfocado en crear sistemas expertos que contenían reglas predefinidas. Posteriormente, la atención de los desarrolladores fue desplazada hacia los métodos de Machine Learning (ML) que permiten a las máquinas mejorar su rendimiento

utilizando datos de entrada. Estos cambios fueron facilitados por los avances actuales en el procesamiento de datos y el aumento de la capacidad computacional (Flasiński, 2016).

La investigación sobre IA se remonta a John McCarthy, quien acuñó el término durante una conferencia en el Dartmouth College en 1956 (Flasiński, 2016); hecho que se considera como el nacimiento del campo científico de la IA. Luego, los progresos en el campo fueron asombrosos, ya que muchos investigadores se centraron en el razonamiento automatizado y aplicaron la IA para demostrar teoremas matemáticos y resolver problemas (Figura 4).

Figura 7

Línea de tiempo de principales eventos relacionados con la IA el siglo pasado



Nota. Adaptado de Jiang et al. (2022)

Sin embargo, pese a los avances notables en esta tecnología, la IA aún se enfrenta a desafíos y limitaciones. Uno de los principales retos es garantizar que los sistemas de IA sean transparentes y explicables, especialmente en aplicaciones críticas donde las decisiones deben ser justificadas. Además, es necesario que se aborden las cuestiones éticas asociadas con la privacidad de los datos y el sesgo algorítmico para asegurar que la IA beneficie a la sociedad de forma responsable y equitativa (Jiang et al., 2022). Todo ello requiere la atención no solo de los desarrolladores, sino también de las personas que hacen uso de la IA.

2.2.2. Machine Learning

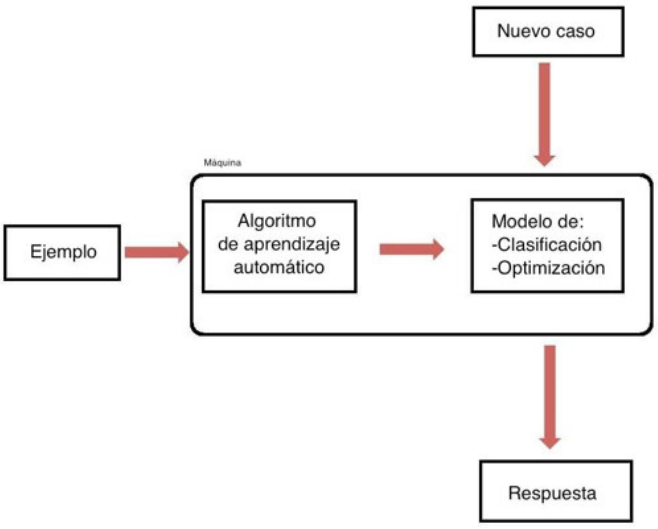
Gavilan (2020) indica que el propio campo de la Inteligencia Artificial carece, de una definición que sea a un tiempo clara, que marque bien las fronteras, que sea fácil de entender

y que, además, sea comúnmente aceptada. El Machine Learning es un subconjunto de la inteligencia artificial. Por su lado Gonzales (2018) señala que los algoritmos que aprenden y mejoran "solos" gracias a la experiencia. A diferencia de modelos en los que un experto de negocio asigna reglas y modeliza algo según sus conocimientos (su experiencia pasada), los modelos estadísticos y los modelos de machine learning dejan que los datos hablen y obtienen las relaciones automáticamente.

El aprendizaje automático se refiere a una subárea de la IA enfocada en proporcionar a los sistemas computacionales (programas y algoritmos) la habilidad para aprender y mejorar automáticamente a partir de un conjunto de datos específicos (Bell, 2022). En síntesis, el objetivo del aprendizaje automático es hacer "buenos" modelos predictivos para "nuevos" datos. En la figura 5 se muestra una representación del aprendizaje automático.

Figura 8

Aprendizaje automático



Nota. Imagen de Debora.riu (https://bit.ly/3EuFq7S)

En el aprendizaje automático, los diferentes casos de uso de los algoritmos se denominan tareas (Choi et al., 2020). Las principales tareas según el objetivo de predicción son:

a) Regresiones (predicción de una variable numérica)

b) Clasificaciones (por ejemplo, el etiquetado de una imagen)

c) Agrupamiento (basados en métodos de distancia).

Las tareas son desarrolladas en diferentes ambientes de aprendizaje:

a) supervisado

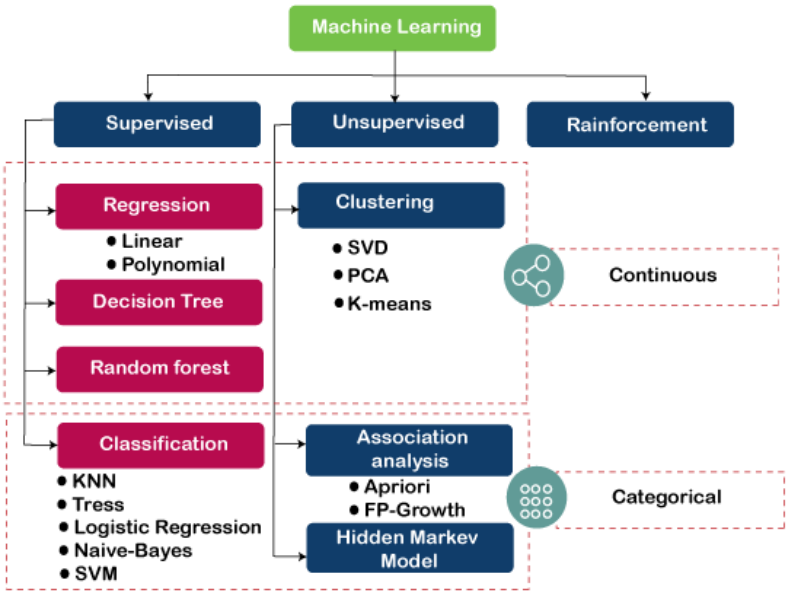
b) no supervisado

c) por reforzamiento

En el entorno de ML las variables independientes son denominadas entradas**,** características **o** predictores. Por otro lado, a las variables dependientes se les conoce como salidas**,** etiquetas **o** variables de respuesta**.** A continuación, en la figura 6 se muestra los tipos de aprendizaje en machine learning.

Figura 9

Tipos de aprendizaje en Machine Learning



Nota. Tomado de Mishra & Mishra (2023)

2.2.3. Tipos de aprendizaje automático supervisado

Aprendizaje automático supervisado. En el aprendizaje supervisado, se desarrollan algoritmos que requieren de ejemplos de entrenamiento para los cuales se conoce de antemano la salida correcta (etiqueta). Se tienen generalmente varias variables de entrada (x) y una

variable de salida (Y). Se utiliza un algoritmo para aprender la función de mapeo de la entrada a la salida (Algren et al., 2021).

El científico de datos determina qué variables debe usar el modelo para desarrollar predicciones. Una vez que se completa el entrenamiento, el algoritmo aplica lo aprendido a los nuevos datos (van Engelen & Hoos, 2020). Dos tipos:

- Clasificación: cuando la variable de salida es una categoría, como "rojo" o "azul" o "sano" y "enfermo".
- Regresión: cuando la variable de salida es un valor real, como "dólares" o "peso".

Ejemplos de algoritmos:

- Regresión lineal (regresiones)
- Clasificación de Naïve Bayes (clasificación)
- Bosque aleatorio (clasificación y regresiones)
- Máquinas de vectores de soporte (clasificación)

Aprendizaje automático no supervisado. En el aprendizaje no supervisado, se desarrollan algoritmos que tratan de encontrar la estructura subyacente o patrones presentes en datos no etiquetados. Solamente se tienen datos de entrada (x) y no se tienen variables de salida. No hay interferencia del científico de datos. Los algoritmos descubren por sí mismo el patrón subyacente (Alloghani et al., 2020).

Por otro lado, según Zhang & Wang (2023), el aprendizaje no supervisado es un método de entrenamiento del aprendizaje automático para el análisis estadístico. Su objetivo principal es descubrir las propiedades ocultas inherentes del conjunto de datos calculando los puntos en común entre muestras no etiquetadas, a fin de evitar el problema de etiquetar las muestras en el aprendizaje supervisado.

Se utilizan principalmente para hacer clustering o asociaciones, es decir, cuando se desea descubrir las agrupaciones inherentes en los datos, como por ejemplo agrupar clientes por comportamiento de compra. Ejemplos de algoritmos:

- K-medias (K-medias)
- Análisis de componentes principales

Debido a la naturaleza del aprendizaje no supervisado, a menudo se utiliza como una herramienta poderosa para análisis costosos en etiquetas o aplicaciones irrelevantes. Aunque el aprendizaje no supervisado no puede realizar directamente la clasificación y la regresión, tiene ventajas significativas en el análisis de datos en tiempo real. De esta manera, el aprendizaje no supervisado se ha convertido en una de las soluciones importantes para problemas como la detección, la eliminación de ruido y el reconocimiento (Zhang & Wang, 2023).

Aprendizaje automático por refuerzo. También existe un tipo particular denominado aprendizaje automático por refuerzo, donde los algoritmos se desarrollan para optimizar las acciones en función de una recompensa. El algoritmo se entrena interactuando con un entorno (virtual). El aprendizaje por refuerzo se utiliza en tareas en las que el aprendizaje depende de acciones ejecutadas y sus consecuencias producidas, por ejemplo, jugar juegos de computadora de estrategia (Kadhim, 2019).

También, el aprendizaje por refuerzo es una técnica de aprendizaje que dirige la acción para maximizar la recompensa de una acción inmediata y las siguientes. En este tipo de algoritmo de ML, la máquina se entrena continuamente mediante un enfoque computacional para aprender de la acción. El aprendizaje por refuerzo es diferente del aprendizaje supervisado porque no es necesario que estén presentes pares de entrada/salida etiquetados (Belyadi & Haghighat, 2021).

El comportamiento de un agente se recompensa en función de las acciones que realiza en el entorno. Considera las consecuencias de sus acciones y toma las medidas óptimas para alcanzarlas. Un ordenador que juega al ajedrez con un humano, que aprende a reconocer palabras habladas y que aprende a clasificar nuevas estructuras astronómicas son algunos ejemplos de aprendizaje por refuerzo (Shobha & Rangaswamy, 2018).

2.2.4. Modelo predictivo

Nowak (2022) indica que los modelos predictivos, también conocidos como modelos de predicción, son un conjunto de herramientas y técnicas estadísticas que sirven para pronosticar y predecir el comportamiento ante un evento. Estos tienen el objetivo de predecir y pronosticar resultados probables a futuro. Para ello, se establece una serie de datos de entrada con los que se desarrollará el análisis predictivo, y los resultados variarán dependiendo del objetivo.

Las aplicaciones de los modelos predictivos son múltiples y en los últimos años, gracias a los avances en gran parte de la evolución tecnológica y la inteligencia artificial, están ganando presencia en multitud de áreas como el marketing y la publicidad, los servicios financieros, la mercadotecnia, la medicina o las redes sociales. Los modelos predictivos de clasificación son aquellos que clasifican y categorizan la información estudiada, basándose sobre todo en datos históricos.

Como resultado, este modelo predictivo responde a las preguntas planteadas indicando de manera clásica y casi siempre binaria si la información pertenece o no a una clase o categoría. Además, en ocasiones hasta aporta porcentajes de las respuestas esperadas. Sobre este particular, Areanada (2019) comenta que el análisis predictivo es el proceso de utilizar el análisis de datos para realizar predicciones basadas en los datos. En este proceso se hace uso de los datos junto con técnicas analíticas, estadísticas y de aprendizaje automático a fin de crear un modelo predictivo para predecir eventos futuros.

El análisis predictivo agrupa una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza los datos actuales e históricos reales para hacer predicciones acerca del futuro o acontecimientos no conocidos. En el ámbito de los negocios los modelos predictivos extraen patrones de los datos históricos y transaccionales para identificar riesgos y oportunidades (Mamun et al., 2020). Los modelos predictivos identifican relaciones entre diferentes factores que permiten valorar riesgos o probabilidades asociadas sobre la base de un conjunto de condiciones, guiando así al decisor durante las operaciones de la organización (Biecek & Burzykowski, 2021).

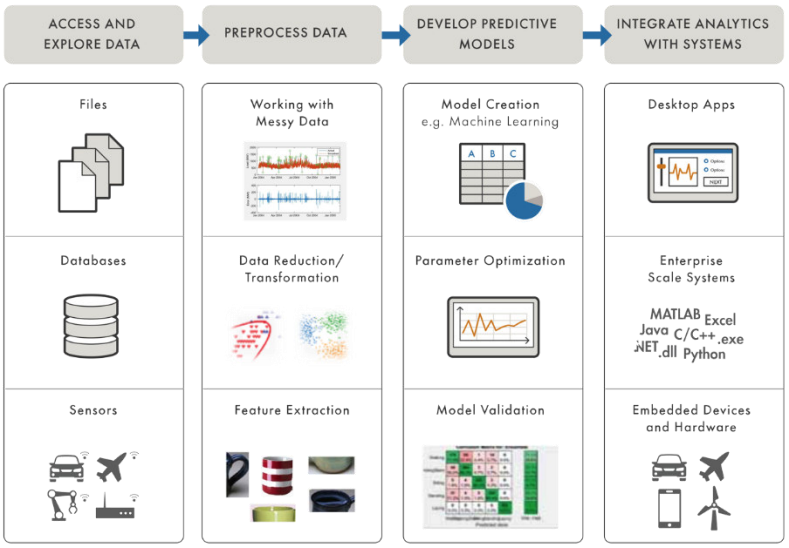
El efecto funcional que pretenden estas iniciativas técnicas es que el análisis predictivo provea una puntuación (probabilidad) para cada sujeto (cliente, empleado, paciente, producto, vehículo, componente, máquina y otra unidad en la organización) con el objeto de determinar, informar o influir procesos en la organización en el que participen un gran número de sujetos, tal y como ocurre en marketing, evaluación de riesgo de crédito, detección de fraudes, fabricación, salud y operaciones gubernamentales como el orden público (Kaur & Kumari, 2022).

Lo fundamental del análisis predictivo está en identificar relaciones entre las variables explicativas y las variables predictivas del pasado de forma que se pueda escalar a lo que está por ocurrir (Huang et al., 2020). Es importante advertir, en cualquier caso, que la fiabilidad y usabilidad de los resultados dependerán mucho del nivel de análisis del dato y la calidad de las hipótesis.

A continuación, en la figura 7 se muestra el flujo de trabajo en un análisis predictivo:

Figura 10

Flujo de trabajo de un análisis predictivo



Nota. Tomado de Taborda et al. (2024)

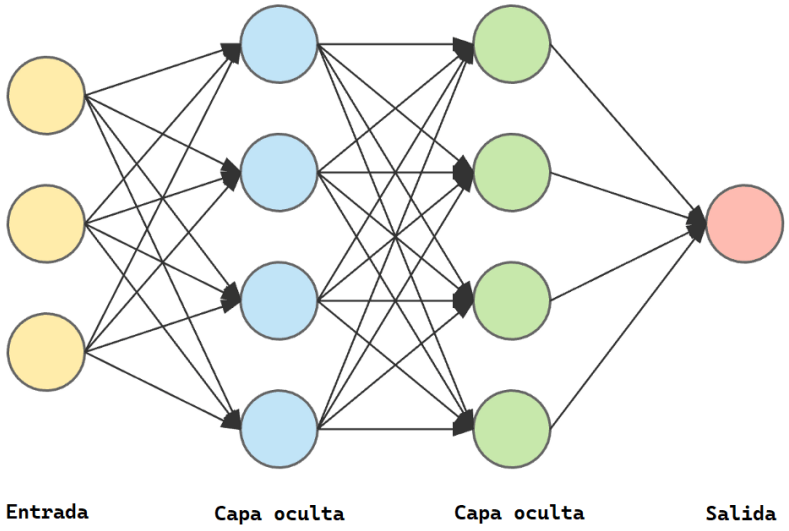
El término "análisis predictivo" describe la aplicación de una técnica estadística o de aprendizaje automático para crear una predicción cuantitativa sobre el futuro. Con frecuencia, se utilizan técnicas de aprendizaje automático supervisado para predecir un valor futuro (¿Cuánto tiempo puede funcionar esta máquina antes de necesitar mantenimiento?) o para calcular una probabilidad.

2.2.5. Redes neuronales

Una red neuronal se trata de un programa informático que puede aprender y tomar decisiones simulando el funcionamiento el cerebro humano. Las redes pueden encontrar patrones y utilizarlos en la resolución de problemas, por lo que puede hacer predicciones reconocer imágenes o comprender el lenguaje. Estas están formadas por varias capas de nodos vinculados entre sí, tal como muestra la Figura 9 (Lodhi et al., 2023). Cada uno de estos nodos es un perceptrón, que funciona de manera similar a la regresión lineal múltiple.

Figura 11

Gráfico de la estructura clásica de una red neuronal



Nota. Tomado de Huet (2023)

Las redes neuronales pueden utilizarse en el análisis de Big Data y extraer patrones complejos, lo que es útil especialmente en las ciencias sociales y otras ramas de la ciencia. De hecho, se emplean para predecir resultados clínicos o casos de deserción estudiantil (Wang et al., 2014).

2.2.6. Deserción estudiantil

La deserción escolar es el abandono definitivo de los estudios por parte de un estudiante antes de completar su periodo académico. Puede deberse a varios factores, como problemas económicos, familiares o socio-contextuales. Si bien se utiliza este término comúnmente para referirse al abandono de la instrucción secundaria, también puede aplicarse a cualquier nivel de educación (Del Bonifro et al., 2020).

Asimismo, la deserción estudiantil se conceptualiza como el final de un proceso gradual y más duradero de desvinculación escolar que involucra tanto la desvinculación académica (motivación y participación en el aprendizaje) como social (relaciones con maestros o compañeros). Esto tiene consecuencias negativas significativas para la vida futura en términos

de oportunidades económicas, salud y bienestar, y riesgo de encarcelamiento cuando se da por motivos ajenos a la ley (Brauer & Sirin, 2024).

Por otro lado, este problema no solo afecta a los alumnos, sino que incumbe a la sociedad como tal, pues representa una pérdida de dinero por parte de potencial capital humano.

Causas de la Deserción Estudiantil. Existen algunos factores internos y externos que son las causas de la deserción escolar (Brauer & Sirin, 2024). A continuación, los enumeraremos:

- Problemas económicos: uno de los principales motivos, pues al carecer de dinero es difícil acceder a algunas necesidades o servicios como los útiles escolares, el transporte o la alimentación.
- Embarazo adolescente: asumir la maternidad en medio de tu desarrollo escolar puede significar una gran responsabilidad y, en muchos casos, deviene en una deserción escolar parcial o total.
- Problemas de salud: algunas condiciones de salud pueden afectar tu rendimiento al momento de estudiar. En algunos casos, los alumnos tendrán mayor predisposición al sueño o simplemente no se enfocarán en los estudios. Esto generaría una deserción por parte del escolar.
- Problemas sociales-contextuales: estos son problemas están presentes en los diferentes contextos que tengan los estudiantes. Estos son los siguientes: *bullying*, inseguridad ciudadana, falta de oportunidades, violencia familiar, violencia en el barrio, problemas con compañeros en clase, problemas familiares, etc.
- Problemas de infraestructura: algunos alumnos sobre todo en las zonas rurales viven lejos de sus centros educativos; por tanto, dejan de asistir a sus clases antes de recorrer varios kilómetros a sus escuelas.

Tipos de Deserción Estudiantil. Esta es la clasificación de la deserción escolar que se tiene:

- Deserción completa: situación en la que el alumno abandona por completo un ciclo académico y no vuelve a estudiar de nuevo.
- Deserción parcial: escenario en el que el estudiante deja sus estudios por un tiempo, pero luego los retoma. Este caso puede darse a través de una licencia o permiso especial.
- Deserción temprana: el alumno deja de acudir a la escuela durante los primeros meses del ciclo académico.
- Deserción tardía: el estudiante abandona sus lecciones luego de la mitad del año escolar.
- Deserción prematura: el niño o adolescente decide no acudir a ninguna clase, a pesar de estar matriculado en su grado escolar.

Viale (2014) menciona que la deserción estudiantil universitaria no es un problema nuevo ni exclusivo del Perú. Este fenómeno se da en todo el mundo, es un viejo problema que tiene muchas variables y el cual no es preocupación exclusiva del mundo académico. La deserción estudiantil universitaria trae como consecuencia el aumento del número de alumnos con educación superior incompleta que se incorporan al mundo laboral y se convierten en subempleados sin obtener los ingresos deseados; lo cual, perjudica al mismo estudiante, a sus familiares, al país y a la universidad pues esta ve afectado su presupuesto.

El estudio de la deserción estudiantil universitaria es muy complejo e importante pues está empezando a considerarse como un indicador de la calidad de la gestión universitaria. De hecho, la tasa de abandono de estudios universitarios figura como indicador de calidad en numerosos modelos de evaluación de la institución universitaria. Según Cabrera et al (2006),

las altas tasas de deserción estudiantil son un indicador de baja calidad pues se entiende que la universidad no hizo lo necesario para que sus alumnos terminaran la carrera.

Cada universidad ha diseñado sus propios programas para facilitar la adaptación a la vida universitaria de los estudiantes nuevos, pero, en la mayoría de los casos, estos programas pertenecen a departamentos o áreas académicas diferentes, con estructuras organizacionales diferentes; con lo cual, la orientación al alumno recién ingresado se hace desde distintos puntos de vista. Esto, en vez de ayudar al estudiante termina por confundirlo más y no se alcanza el objetivo de facilitar su inserción a la universidad.

La complejidad del análisis de la deserción radica en que se trata de un problema de varias variables las cuales se pueden agrupar según Tejedor & Muñoz-Repiso (2007), en aquellas que pertenecen al área pedagógica y aquellas que pertenecen al área no pedagógica. Un adecuado programa de inserción a la vida universitaria debe contemplar las variables de ambas áreas. Por otro lado, creemos que estos programas, en su etapa de diseño, deben contar con la participación de las autoridades escolares pues son ellas las que han tenido a nuestros futuros alumnos 12 años, en promedio, en sus escuelas.

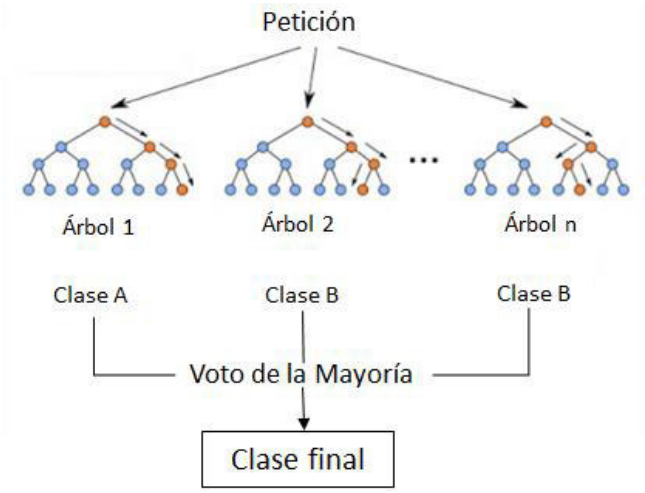
2.2.7. Random forests

Un bosque aleatorio es un clasificador ML supervisado que comprende una estructura de árbol {h(x, (k) k = 1, 2, ….}, un vector independiente único {θ(k)} y una entrada para la clase más famosa de x [36–38] que se utiliza tanto para la clasificación como para el análisis de regresión (Rigatti, 2017). Este método funciona mediante la construcción de múltiples árboles de decisión durante el entrenamiento; para las tareas de clasificación, el resultado del bosque aleatorio es la clase seleccionada por la mayoría de los árboles (Rastogi et al., 2023). Además, aplica la técnica de bagging (o agregación bootstrap), que consiste en generar un nuevo conjunto de datos con un reemplazo de un conjunto de datos existente.

La estructura general de los bosques aleatorios se presenta en la Figura 9.

Figura 12

Esquema clásico de un bosque aleatorio



Nota. Tomado de Puma (2020)

Por otro lado, de acuerdo con Umoh et al. (2022), el bosque aleatorio presenta las siguientes características:

- El aprendizaje conjunto utilizado en el bosque aleatorio evita que se sobreajuste.
- El bagging permite que el bosque aleatorio funcione bien con un conjunto de datos pequeño.
- Los predictores de bosque aleatorio se pueden entrenar en paralelo.
- La selección automática de características se habilita mediante el aprendizaje del árbol de decisiones en un bosque aleatorio.

Asimismo, el bosque aleatorio aprovecha las estrategias de aleatorización, el análisis de alternativas y la técnica de conjuntos para generar modelos precisos de aprendizaje automático. El "bosque" que construye es una combinación de árboles de decisión, entrenados mediante métodos de bagging. Las principales ventajas del bosque aleatorio incluyen el descubrimiento de anomalías en los datos, la identificación de características importantes, el descubrimiento de patrones en los datos y la generación de gráficos esclarecedores (Parmar et al., 2019).

Según Xia (2020), el bosque aleatorio presenta otras ventajas como:

- Robusto contra sobreajuste y usualmente menos sensible a los valores de entrada
- Muy fácil de usar porque solo se requieren dos parámetros (el número de variables en el subconjunto aleatorio en cada nodo y el número de árboles en el bosque).
- Flexibilidad para realizar varios tipos de análisis de datos estadísticos, incluyendo regresión, clasificación, análisis de supervivencia y aprendizaje no supervisado.
- Muy alto poder de discriminación y precisión de clasificación.
- Sin suposiciones distribucionales sobre las variables predictoras o de respuesta.
- Puede manejar situaciones en las que el número de variables predictoras excede en gran medida el número de observaciones.

Así, los bosques aleatorios o bosques de decisión aleatorios son un método de aprendizaje conjunto para tareas de clasificación, regresión y otras tareas con grandes ventajas a otras técnicas de clasificación.

2.2.8. Responsabilidad social

Este concepto se refiere a una teoría o marco ético cuyo principio principal es que toda organización, persona o institución educativa debe, es responsable o está obligada a beneficiar, con sus actividades, servicios o productos, a la sociedad y evitar cualquier actividad que fuera perjudicar a las personas (Velte, 2022). En conjunto, la responsabilidad social se basa en la idea de que todos, desde empresas hasta individuos, tienen que cumplir rol activo en la construcción de un mundo más justo, sostenible y equitativo.

En este caso particular, el sistema desarrollado permitió solucionar y mitigar una problemática con implicancias sociales: la deserción educativa. Ya se ha mencionado anteriormente los efectos adversos que la deserción estudiantil ocasiona a la sociedad y, por lo tanto, esta investigación pretender combatir o reducir las tasas de deserción en la universidad donde fue implementado el sistema. De hecho, si bien se espera beneficiar de forma individual

a los estudiantes (y con ellos a su familia), la oportunidad de continuar con sus estudios permite que la sociedad peruana cuente cada vez más con profesionales competentemente calificados no solo de la universidad, sino de otros centros de estudios que pueden tomar como referencia este estudio para desarrollar sus propios sistemas.

2.3. Marco conceptual

Machine Learning: una disciplina que permite a los ordenadores aprender por sí mismos y realizar tareas de forma autónoma sin necesidad de ser programados. Las técnicas de aprendizaje automático son, de hecho, una parte fundamental del Big Data (He, 2021; Kufel et al., 2023).

Modelos predictivos: Los modelos predictivos son instrumentos que se elaboran en base a la información relevante o datos experimentales de las variables que se están investigando, y con el análisis de dichas variables predictoras ayudan a predecir resultados para condiciones no estudiadas (Battineni et al., 2020; Lamba & Madhusudhan, 2022).

Deserción estudiantil: Se puede definir a la deserción estudiantil como el abandono voluntario de la actividad académica durante un periodo de tiempo o de manera definitiva, que, por lo general, ocurre en los tres primeros semestres académicos, y que puede ser explicado por diferentes categorías de variables: socioeconómicas, individuales, institucionales y académicas (Bäulke et al., 2022; Del Bonifro et al., 2020).

Entrenamiento: Término el cual hace referencia instruir al modelo de aprendizaje automático por medio de pruebas que ayuden con la mejora de este (Grimaldi & Ehrler, 2023).

Algoritmos supervisados: Son técnicas de ML que requieren datos etiquetados para entrenar el modelo. La función de estos algoritmos es predecir resultados tomando como base la relación entre las variables de entrada (que se conocen características) y la variable que es conocida (Kadhim, 2019).

Algoritmos no supervisados: Se refiere a los métodos de ML que no emplean datos etiquetados. En lugar de ello, identifican patrones, agrupamientos o estructuras que se encuentran ocultas en los datos para realizar una clasificación y un análisis posterior (Rodriguez et al., 2019).

Características (features): Se refieren a las variables o atributos que describen los datos presentes en un modelo predictivo. Así, estas características representan la información relevante del problema y son esenciales para que el algoritmo pueda realizar predicciones precisas (Yan, 2022).

Conjunto de datos (dataset): Se refiera a la colección de datos estructurados que se utilizan para el entrenamiento y la evaluación de los modelos predictivos. Generalmente se divide en conjuntos de entrenamiento, validación y prueba para garantizar que se realicen análisis completos (Dhal & Azad, 2022).

Validación cruzada: Define a la técnica encargada de evaluar la capacidad predictiva de un modelo al dividir los datos en múltiples subconjuntos. Así, la validación cruzada permite garantizar que el modelo funcione de forma correcta con datos que no son visibles, lo que reduce el riesgo de sobreajuste (Frunza, 2016).

Tasa de precisión (accuracy): Define a una métrica de evaluación con la que es posible medir el porcentaje de predicciones correctas que un modelo realiza durante su vida útil (Hofmarcher et al., 2019). El cálculo se obtiene al dividir el número de aciertos entre el número total de predicciones.

Sobreajuste (overfitting): Se refiere a una problemática ocasionada cuando un modelo aprende demasiado bien los datos de entrenamiento, incluidas las particularidades o ruido. El overfitting ocasiona que el modelo pierda la capacidad para generalizar con datos nuevos que son ingresados (Hosseini et al., 2020).

Subajuste (underfitting): Este fenómeno ocurre cuando un modelo no logra capturar las relaciones subyacentes en los datos. La causa general de este problema se encuentra en la simplicidad del modelo de decisión o la carencia de datos para realizar predicciones (Seraj et al., 2023).

Regresión logística: Es un método estadístico y de ML que se emplea para modelar la probabilidad de un evento binario (Frunza, 2016). Es comúnmente empleado para predecir resultados como “abandonó” o “continuó” en estudios sobre deserción estudiantil.

Árboles de decisión: Son modelos de aprendizaje supervisado que segmentan los datos en ramas basadas en reglas derivadas de sus características. Ofrecen una interpretación visual y clara de las decisiones tomadas (Semanjski, 2023).

Evaluación de modelos: Es el proceso de medir el rendimiento de un modelo predictivo utilizando métricas como precisión, sensibilidad, especificidad y el área bajo la curva ROC, garantizando que cumpla con los objetivos esperados (Semanjski, 2023).

Preprocesamiento de datos: Es el conjunto de pasos que se realizan para la limpieza, transformación y preparación de los datos antes de entrenar un modelo (Lamba & Madhusudhan, 2022). Incluye tareas como eliminación de valores faltantes, normalización y codificación de variables categóricas.

2.4. Marco filosófico

La epistemología es una rama de la filosofía que se encarga de los problemas filosóficos que rodean la teoría del conocimiento, estudia la relación entre el sujeto y el objeto y todos los problemas que esa relación plantea, es decir, que el conocimiento no puede estudiarse dejando de lado al sujeto y al objeto. En este mismo sentido, la epistemología es la rama de la investigación científica y su producto, el conocimiento científico sirve para cambiar positivamente el trasfondo de la investigación (Yucra Quispe & Bernedo Villalta, 2020).

La presente investigación, titulada Modelo Predictivo basado en Machine Learning para la Reducción de la Deserción Estudiantil de las Universidades Privadas en el Perú, se relaciona estrechamente con la Corriente Filosófica del Positivismo cuya doctrina se fundamenta en dar repuestas a nuevos cambios, en los hechos, en la experiencia. Al respecto Pérez (2015) señala que "es una epistemología híbrida que combina el racionalismo con el empirismo y la lógica deductiva con la lógica inductiva, también ha sido denominado hipotético-deductivo, cuantitativo, empírico-analista y racionalista" (p. 29).

El paradigma positivista también llamado cuantitativo, empírico-analítico, racionalista busca explicar, predecir, controlar los fenómenos, verificar teorías y leyes para regular los fenómenos; identificar causas reales, temporalmente precedentes o simultáneas (Carr, 2022; Loza et al., 2021). Es así, que el paradigma cuantitativo en la investigación induce al análisis del positivismo. En este sentido, los paradigmas objetivistas consideran que el conocimiento, para ser positivo, debe ser verdadero; o sea, que el conocimiento, es producto de la experiencia sensible y objetiva (Mejía-Rivas, 2022). En el trascurso de la historia, esta teoría ha influenciado en los más grandes descubrimientos del siglo XX que han permitido desarrollar la tecnología hasta la actualidad.

En este sentido, en la presente investigación los hechos están representados por las variables y dimensiones que constituyen el Modelo Predictivo y la Reducción de la Deserción Estudiantil de la Universidad Privada San Juan Bautista, y se busca explicar mediante los elementos teóricos y prácticos de los modelos predictivos y de la deserción estudiantil, por lo tanto, los hechos se explican a través de observación de las causas, es decir, como se producen esos fenómenos con la intención de llegar a generalizaciones sujetas a verificación y comprobación de las observaciones.

III. MÉTODO

3.1. Tipo de investigación

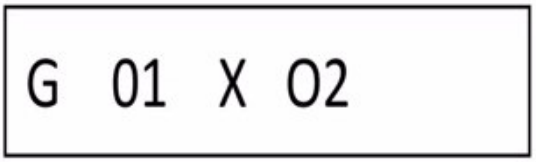
La presente investigación se clasifica como un estudio de tipo **aplicado**, ya que su principal objetivo fue la implementación práctica de un modelo predictivo que pudiera reducir la deserción estudiantil. Según Hernández Sampieri (2014), la investigación aplicada tiene como propósito resolver problemas específicos en un contexto particular y generar soluciones prácticas. En este caso, se buscó diseñar un modelo que no solo estudiara la problemática de la deserción, sino que proporcionara herramientas para mejorar la retención de estudiantes en una universidad privada. De esta manera, la investigación se orientó a proporcionar respuestas inmediatas y aplicables a los desafíos actuales del sistema educativo.

Respecto a su **nivel explicativo**, el estudio corresponde a este tipo de nivel porque trató de identificar y explicar los factores que inciden en la deserción estudiantil, utilizando machine learning como herramienta analítica. Según Montero y León (2007), un estudio explicativo busca determinar las causas o factores que inciden en el fenómeno de estudio, y este fue precisamente el enfoque de la investigación. Se utilizó la información recabada para construir un modelo que pudiera predecir la deserción a partir de variables relevantes, lo que permitió explicar la relación entre dichos factores y los resultados observados, como la probabilidad de que un estudiante abandone sus estudios.

En cuanto al **enfoque cuantitativo**, la investigación adoptó este enfoque porque se basó en la recolección y el análisis de datos numéricos con el propósito de establecer patrones y relaciones. Los estudios cuantitativos, según López (2013), se caracterizan por medir variables y analizar las relaciones entre ellas mediante herramientas estadísticas. En este caso, se utilizaron herramientas de machine learning para crear modelos predictivos basados en datos cuantificables, tales como el rendimiento académico, la asistencia, y otros factores asociados

con la deserción estudiantil, permitiendo una interpretación objetiva y medible de los resultados.

Finalmente, el diseño de la investigación fue **preexperimental de corte longitudinal**, ya que el estudio se llevó a cabo durante un periodo específico en el que se recopilaron datos de manera continua, con el fin de observar la evolución de la deserción estudiantil a lo largo del tiempo. Según Álvarez-Gayou (2005), un diseño preexperimental es aquel en el que no se manipulan las variables independientes, pero se realiza una medición del fenómeno antes y después de la implementación de un modelo o intervención. Este diseño fue adecuado para la investigación, ya que no se realizó una intervención directa en el proceso educativo, sino que se evaluó la eficacia de un modelo predictivo a lo largo del tiempo, observando su capacidad para anticipar la deserción y, en consecuencia, ofrecer una herramienta para reducirla. Seguidamente, se presentó el esquema detallado del diseño preexperimental que se llevó a cabo en el área de estudio.



Donde:

G (Grupo único): Datos de estudiantes de la Universidad Privada San Juan Bautista, centrados en factores personales, académicos y socioeconómicos.

O1 (Pretest): Medición de la tasa de deserción estudiantil antes de implementar el modelo predictivo, considerando los factores mencionados.

X (Tratamiento): Implementación del modelo predictivo basado en machine learning para identificar estudiantes en riesgo de deserción, considerando los factores personales, académicos y socioeconómicos.

O2 (Post test): Medición de la tasa de deserción después de la implementación del modelo para evaluar su impacto en la reducción de la deserción.

3.2. Población y muestra

Población

Es el total de elementos (individuos u objetos) de los que desea conocer o investigar. Es ideal que la población sea definida a partir de los objetivos de la investigación (Ñaupas et al., 2018). En este caso de estudio, la población estuvo conformada por los 143 estudiantes de la carrera de Ingeniería de Sistemas de la Universidad Privada San Juan Bautista, perteneciente a la Facultad de Ingeniería durante el año 2023, quienes se vieron afectados por la problemática de la deserción estudiantil.

Muestra

Según Ñaupas et al. (2018), una muestra representativa debe reflejar las características esenciales de la población para garantizar la validez del estudio. En este caso, se seleccionaron 104 universitarios para evaluar la deserción estudiantil y analizar diversas estrategias implementadas en la Universidad Privada San Juan Bautista. La muestra fue elegida estratégicamente para abarcar distintos enfoques y contextos dentro de la institución, asegurando resultados representativos y aplicables a su realidad.

Tipo de Muestreo

Para determinar el tamaño muestral adecuado, se empleó un muestreo probabilístico. Según Otzen & Manterola (2017), esta técnica de selección asegura que todos los elementos de una población tengan una probabilidad conocida y no nula de ser seleccionados, lo que garantiza la obtención de muestras representativas y permite generalizar los resultados al conjunto de la población.

Dado que la población de estudio es finita, es decir, se conoce el total de unidades de observación que la conforman, se aplicó la siguiente fórmula (Aguilar-Barojas, 2005):

$$n = \frac{N \times Z^2 \times p \times q}{d^2 \times (N-1) + Z^2 \times p \times q}$$

Donde:

n: Tamaño de la muestra requerida.

N: Tamaño total de la población (N=143 estudiantes)

Z: Valor crítico de la distribución normal estándar, asociado al nivel de confianza deseado (por ejemplo, 1.96 para un nivel de confianza del 95%).

p: Proporción esperada de la característica de interés en la población (si se desconoce, suele tomarse como 0.5 para maximizar la variabilidad).

q: Complemento de p, es decir, q=1−p.

d: Margen de error permitido o precisión deseada (expresado como un decimal, por ejemplo, 0.05 para un 5% de error).

Reemplazando:

$$n = \frac{143 * 1.96^2 * 0.5 * 0.5}{0.05^2(143-1) + 1.96^2 * 0.5 * 0.5}$$

$$n = 104$$

Por tanto, el tamaño muestral estuvo constituida por 104 estudiantes de la carrera de Ingeniería de Sistemas de la Universidad Privada San Juan Bautista.

Unidad de Análisis

La unidad de análisis es el elemento principal sobre el cual se enfoca una investigación y se formulan las conclusiones, pudiendo ser individuos, grupos, organizaciones, eventos, entre otros (Darío & Alejandra, 2014). En consecuencia, la unidad de análisis corresponde a los estudiantes de la carrera de Ingeniería de Sistemas de la Universidad Privada San Juan Bautista que enfrentaron la problemática de la deserción estudiantil durante el año 2023. Constituyen el grupo central del estudio, sobre el cual se aplicará el modelo predictivo basado en machine

learning para generar conclusiones generales relacionadas con la reducción de la deserción estudiantil.

Unidad de Observación

La unidad de observación es el nivel en el que se recopilan los datos o información, es decir, los sujetos, elementos o entidades concretas a partir de los cuales se obtienen las variables para el análisis (Barriga & Henríquez, 2011). En ese sentido, la unidad de observación se refiere a cada uno de los 143 estudiantes de la población total, o a los 104 estudiantes seleccionados como muestra, de quienes se recolectaron datos individuales. Estos datos fueron utilizados como insumos para alimentar el modelo predictivo basado en machine learning y evaluar su eficacia en la identificación de patrones y estrategias para reducir la deserción estudiantil.

3.3. Operacionalización de variables

La Investigación, presenta las siguientes variables con sus respectivos indicadores. En la Tabla 2, se muestran las variables empleadas en el estudio, así como las dimensiones e indicadores que se ven afectadas en la deserción estudiantil.

Tabla 2

Variables, dimensiones e indicadores del presente estudio

Variables	Dimensiones	Indicadores	Ítems
Independiente: Modelo Predictivo Basado en Machine Learning	Calidad del modelo	Nivel de Predicción del modelo	
		Efectividad	
	Tiempo de operación	Tasa de deserción	
		Tasa de retención	
	Rendimiento del modelo	Validación del modelo	
		Precisión de la predicción	

Dependiente: Deserción Estudiantil	Factor personal	Problemas de salud	1
		Carga familiar	2
		Vocación profesional	3
	Factor académico	Bajo rendimiento académico	4
		Apoyo académico	5
		Horarios de clase	6
		Límite de inasistencia	7
	Factor socioeconómico	Dificultad financiera	8
		Trabajo tiempo completo	9
		Apoyo familiar	10

3.4. Instrumentos

Citando a Arias (2012) un instrumento es algún recurso, formato o dispositivo, ya sea en papel o digital, que se usa para conseguir, reunir o apuntar datos, durante el desarrollo de la investigación. De esta manera, se utilizó un cuestionario diseñado para abordar los procesos relacionados con la variable dependiente, con el propósito de recopilar información crucial sobre los factores que influyen en la deserción estudiantil en las universidades privadas del Perú, específicamente en la Universidad Privada San Juan Bautista. Este cuestionario, estructurado en tres factores principales (Personal, Académico y Socioeconómico), empleó una metodología dicotómica con respuestas limitadas a "Sí" o "No". A través de este enfoque, se buscó identificar posibles causas del abandono educativo, tales como problemas de salud, responsabilidades familiares, vocación profesional, bajo rendimiento académico, falta de apoyo académico, incompatibilidad de horarios de clase, límite de inasistencias, dificultades financieras, trabajo a tiempo completo y falta de apoyo familiar. Las respuestas obtenidas fueron fundamentales para el desarrollo de un modelo predictivo basado en machine learning,

cuyo propósito fue reducir la deserción estudiantil y mejorar la experiencia educativa, garantizando la confidencialidad de la información recopilada para fines investigativos.

3.5. Procedimientos

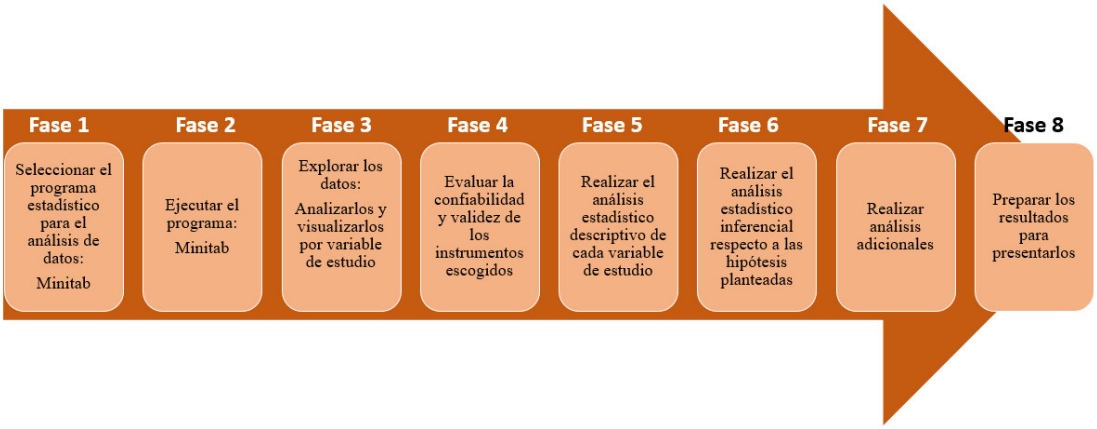
Para cumplir con los objetivos específicos del estudio, primero se utilizó un cuestionario con preguntas dicotómicas ("Sí" o "No") para recopilar datos relacionados con la deserción estudiantil en la Universidad Privada San Juan Bautista, permitiendo identificar factores personales, académicos y socioeconómicos asociados con esta problemática. Los datos fueron procesados y utilizados para desarrollar un modelo predictivo basado en Machine Learning, específicamente con bosques aleatorios, empleando un archivo JSON como fuente inicial. Las variables predictoras (X) se seleccionaron eliminando las columnas de deserción y factor de deserción, mientras que la variable de respuesta (Y) correspondió al factor de deserción. Se dividieron los datos en un 80% para entrenamiento y 20% para prueba, entrenando el modelo para predecir valores categóricos: 0 (sin deserción), 1 (factores personales), 2 (factores académicos) y 3 (factores socioeconómicos). Posteriormente, se diseñaron estrategias específicas para cada categoría: programas de acompañamiento psicológico, mentorías personalizadas y capacitaciones en habilidades blandas para factores personales; tutorías académicas, flexibilidad en la carga crediticia, fortalecimiento docente y orientación vocacional para factores académicos; y becas, oportunidades laborales flexibles, reducción de costos indirectos y alianzas con entidades externas para factores socioeconómicos. Estas estrategias se incorporaron al modelo como recomendaciones para mitigar la deserción estudiantil y optimizar los procesos educativos.

3.6. Análisis de datos

Para el análisis de datos se usó el programa Minitab, este análisis de datos, se realizó de acuerdo con las fases señaladas en la Figura 9.

Figura 13

Fases para el análisis de datos



Se selecciona el programa estadístico Minitab para el análisis de datos (Figura 10).

- Minitab (Herramienta estadística)

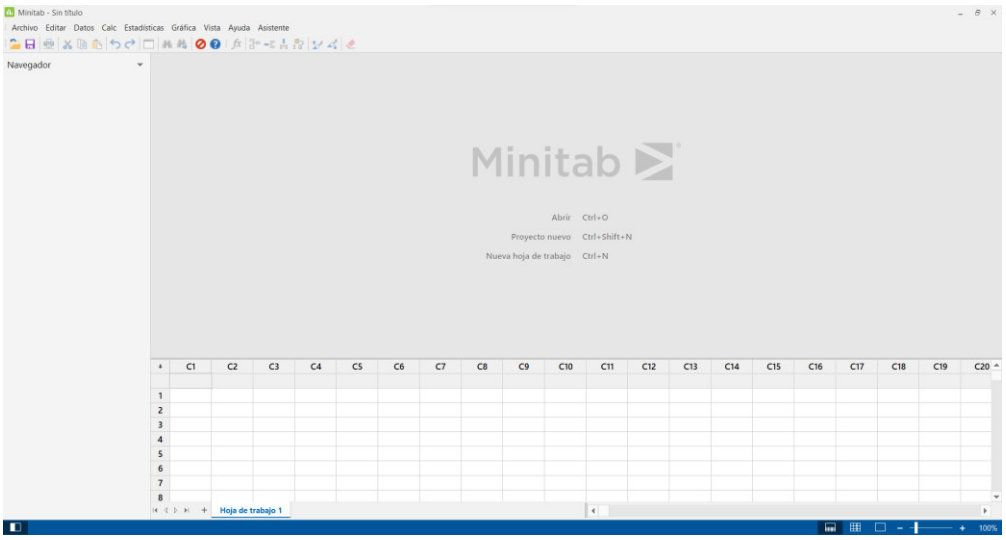
Figura 14

Icono de la herramienta Minitab

Se ejecuta el programa Minitab para el análisis de datos (Figura 11).

Figura 15

Pantalla principal de Minitab



Estadística Inferencial: Analizar las Hipótesis

- Probar hipótesis poblacionales
- Estimar parámetros

Nivel de significancia o significación

- El nivel de significancia es de 0.05

Prueba de Hipótesis

- Análisis paramétricos
  - Coeficiente de correlación de Pearson y regresión lineal
  - Prueba t-Student
- Análisis No paramétricos
  - Pruebas para una muestra (Chi-cuadrado)
  - Pruebas para dos muestras independientes (U de Mann-Whitney)

Figura 16

Histograma de muestra de la herramienta Minitab

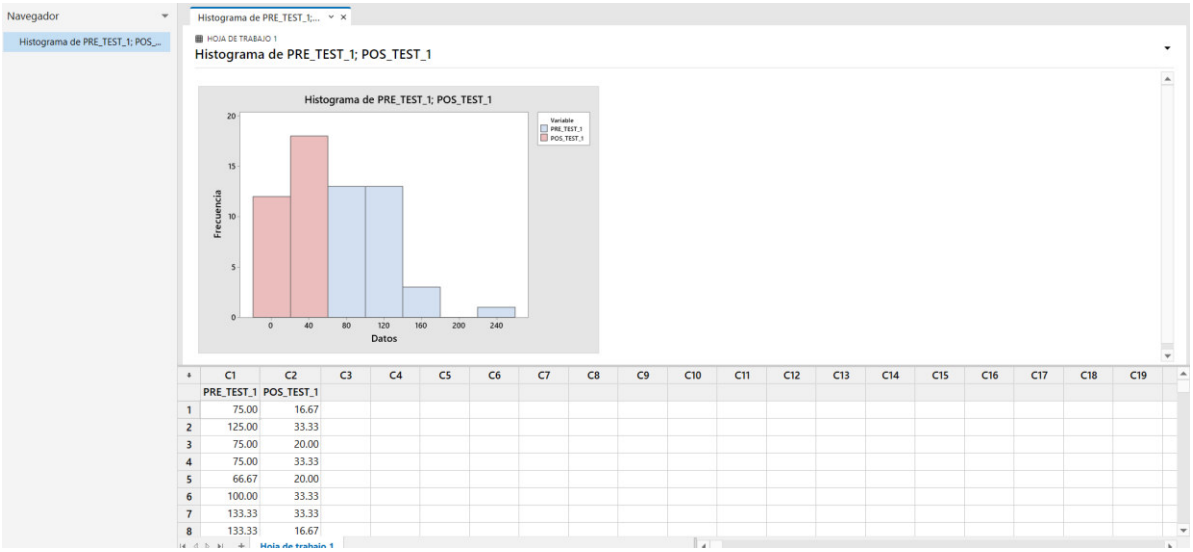
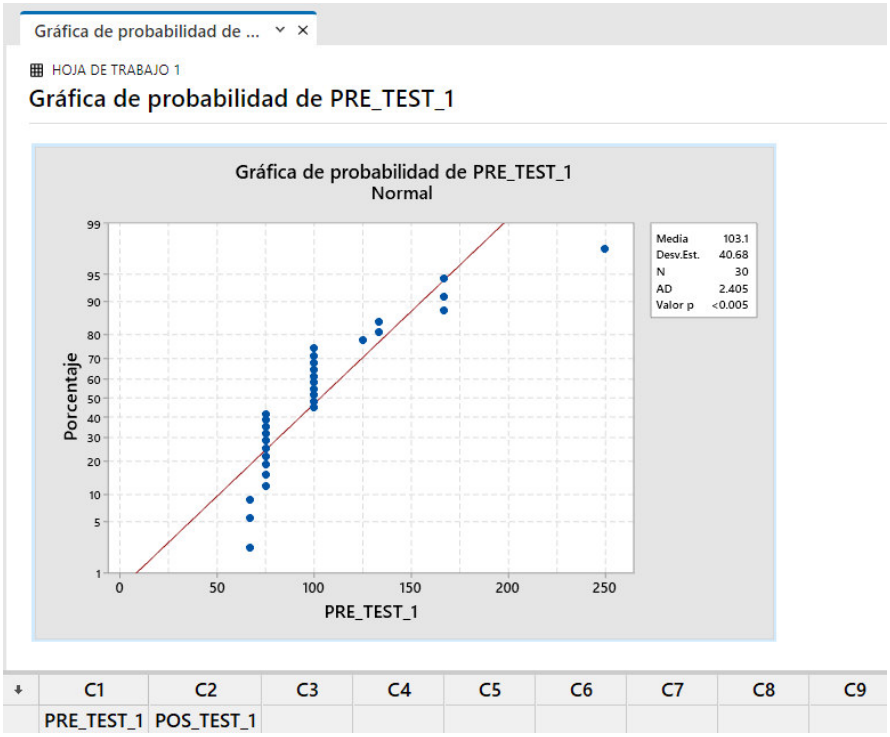


Figura 17

Gráfico de probabilidad de muestra



3.7. Consideraciones éticas

En la investigación, se tuvieron en cuenta diversas consideraciones éticas, basadas en principios fundamentales como el código de Núremberg, el informe Belmont, las pautas de la

CIOMS y la declaración de Helsinki. Se aseguró el consentimiento voluntario de los participantes, garantizando el beneficio para la sociedad y evitando cualquier sufrimiento físico o mental innecesario. Se respetó la dignidad de las personas, otorgándoles un consentimiento informado y promoviendo la justicia en todos los aspectos del estudio. Asimismo, se respetaron los derechos de autor de las fuentes utilizadas, citando adecuadamente toda la información referenciada. La veracidad de los datos recolectados fue fundamental, y se preservó la privacidad y confidencialidad de las personas que formaron parte del grupo muestral, asegurando que todas las fuentes de información fueran debidamente referenciadas al final de la investigación.

IV. RESULTADOS

Para el desarrollo del Modelo predictivo basado en Machine Learning para la reducción de la deserción estudiantil en las universidades privadas del Perú, se utilizaron las herramientas de Visual Studio Code, Python con Colab, Postman, Matplotlib y JSON, cada una de ellas para determinados procesos.

4.1. Desarrollo del modelo predictivo basado en Machine Learning contribuye en la predicción de los factores personales, académico y socioeconómicos

4.1.1. Recopilación de Datos

Para la etapa de recopilación de datos, se recolectó datos históricos sobre estudiantes que han desertado y aquellos que han continuado, incluyendo información personal relevante para el estudio, esto con la finalidad de evaluar y entrenar al modelo.

Tabla 3

Dataset histórica del 2018-1 al 2023-2 - Sí

Año	Estudiantes	Factor Personal			Factor Académico				Factor Socioeconómico		
		Problemas De Salud	Carga Familiar	Vocación Profesional	Bajo Rendimiento Académico	Apoyo Académico	Horarios Clase	Límite Inasistencia	Dificultad Financiera	Trabajo Tiempo Completo	Apoyo Familiar
20181	46	24	17	26	07	21	32	06	16	23	20
20182	11	06	06	05	01	02	04	03	07	05	06
20191	64	25	20	10	10	25	10	07	30	15	20
20192	16	10	03	05	08	12	08	08	08	07	09
20201	124	89	59	50	26	80	15	80	19	40	50
20202	33	20	02	09	11	15	14	18	16	12	10
20211	192	98	20	12	18	90	15	25	89	42	50
20212	30	15	12	05	05	17	20	26	28	12	10
20221	43	12	12	15	10	10	17	18	12	13	15
20222	35	26	14	15	12	10	18	16	19	17	18
20231	25	20	12	10	10	14	11	12	16	08	09
20232	79	15	03	10	15	12	11	10	23	15	14

4.1.2. Campos para la limpieza de datos

Para esta etapa, se limpió y preparó los datos establecidos para el análisis, gestionando valores faltantes y anomalías.

Tabla 4

Conversión de Dataset histórica a Dataset definitiva 2018-1 al 2023-2 Factor Personal

Estudiante	Problemas Salud	Carga Familiar	Vocación Profesional
Estudiante 1	No	No	Si
Estudiante 2	No	No	Si
Estudiante 3	Si	No	Si
Estudiante 4	Si	No	Si
Estudiante 5	Si	No	Si
Estudiante 6	Si	No	No
Estudiante 7	Si	Si	Si
Estudiante 8	Si	Si	Si
Estudiante 9	No	No	Si
Estudiante 10	No	No	No
Estudiante 11	No	No	Si
Estudiante 12	No	No	Si
Estudiante 13	No	No	Si
Estudiante 14	No	Si	Si
Estudiante 15	Si	Si	No
Estudiante 16	Si	Si	No
Estudiante 17	No	No	Si
Estudiante 18	Si	No	No
Estudiante 19	No	No	Si
Estudiante 20	No	No	Si
Estudiante 21	No	No	No
Estudiante 22	No	No	No
Estudiante 23	No	No	Si
Estudiante 24	No	No	Si
Estudiante 25	Si	Si	No
Estudiante 26	Si	Si	No
Estudiante 27	No	Si	Si
Estudiante 28	Si	No	No

Estudiante 29	No	No	Si
Estudiante 30	No	No	Si
Estudiante 31	No	No	Si
Estudiante 32	No	Si	Si
Estudiante 33	Si	Si	No
Estudiante 34	Si	Si	No
Estudiante 35	Si	No	No
Estudiante 36	No	Si	Si
Estudiante 37	Si	Si	No
Estudiante 38	Si	Si	No
Estudiante 39	No	No	Si
Estudiante 40	Si	No	No
Estudiante 41	No	No	Si
Estudiante 42	No	No	Si
Estudiante 43	No	No	Si
Estudiante 44	Si	No	No
Estudiante 45	Si	No	No
Estudiante 46	Si	No	No
Estudiante 47	No	No	Si
Estudiante 48	No	No	Si
Estudiante 49	No	No	Si
Estudiante 50	No	No	Si
Estudiante 51	No	No	Si
Estudiante 52	Si	Si	No
Estudiante 53	Si	Si	No
Estudiante 54	No	Si	Si
Estudiante 55	Si	No	No
Estudiante 56	No	No	Si
Estudiante 57	No	No	Si
Estudiante 58	No	No	Si
Estudiante 59	No	Si	Si
Estudiante 60	Si	Si	No
Estudiante 61	Si	Si	No
Estudiante 62	No	No	Si
Estudiante 63	Si	No	No
Estudiante 64	Si	No	No
Estudiante 65	Si	Si	No
Estudiante 66	No	No	Si

Estudiante 67	Si	No	No
Estudiante 68	Si	No	No
Estudiante 69	Si	No	No
Estudiante 70	No	No	Si
Estudiante 71	No	No	Si
Estudiante 72	No	No	No
Estudiante 73	No	No	No
Estudiante 74	No	No	Si
Estudiante 75	No	Si	Si
Estudiante 76	Si	No	No
Estudiante 77	No	No	No
Estudiante 78	No	No	Si
Estudiante 79	Si	No	No
Estudiante 80	No	Si	No
Estudiante 81	No	No	No
Estudiante 82	Si	Si	No
Estudiante 83	Si	No	No
Estudiante 84	Si	No	No
Estudiante 85	Si	No	No
Estudiante 86	Si	Si	No
Estudiante 87	Si	Si	No
Estudiante 88	No	Si	No
Estudiante 89	Si	Si	No
Estudiante 90	No	Si	Si
Estudiante 91	No	No	No
Estudiante 92	No	No	Si
Estudiante 93	No	No	Si
Estudiante 94	No	No	Si
Estudiante 95	No	No	Si
Estudiante 96	No	No	Si
Estudiante 97	Si	No	Si
Estudiante 98	Si	No	Si
Estudiante 99	Si	No	Si
Estudiante 100	Si	No	No
Estudiante 101	Si	Si	Si
Estudiante 102	Si	Si	Si
Estudiante 103	No	No	Si
Estudiante 104	No	No	No

Tabla 5

Conversión de Dataset histórica a Dataset definitiva 2018-1 al 2023-2 Factor Académico

Estudiante	Bajo Rendimiento Académico	Apoyo Académico	Horarios Clase	Limite Inasistencia
Estudiante 1	No	No	No	No
Estudiante 2	No	Si	Si	No
Estudiante 3	No	Si	No	No
Estudiante 4	No	Si	No	No
Estudiante 5	No	Si	Si	Si
Estudiante 6	No	No	Si	No
Estudiante 7	No	Si	No	No
Estudiante 8	No	Si	No	No
Estudiante 9	No	No	No	No
Estudiante 10	No	Si	No	No
Estudiante 11	No	No	Si	No
Estudiante 12	No	Si	Si	No
Estudiante 13	Si	Si	Si	No
Estudiante 14	No	Si	Si	No
Estudiante 15	No	Si	No	No
Estudiante 16	No	No	Si	No
Estudiante 17	No	Si	No	Si
Estudiante 18	No	No	Si	No
Estudiante 19	No	No	Si	No
Estudiante 20	No	Si	Si	No
Estudiante 21	Si	Si	No	No
Estudiante 22	No	Si	Si	Si
Estudiante 23	No	No	No	No
Estudiante 24	Si	No	Si	No
Estudiante 25	Si	Si	Si	No
Estudiante 26	No	No	Si	Si
Estudiante 27	No	Si	Si	No
Estudiante 28	No	No	No	No
Estudiante 29	No	No	Si	No
Estudiante 30	No	Si	Si	No
Estudiante 31	Si	Si	Si	No
Estudiante 32	No	Si	Si	No
Estudiante 33	No	Si	No	No

Estudiante 34	No	No	Si	No
Estudiante 35	No	No	Si	No
Estudiante 36	No	Si	Si	No
Estudiante 37	No	Si	No	No
Estudiante 38	No	No	Si	No
Estudiante 39	No	Si	No	Si
Estudiante 40	No	No	Si	No
Estudiante 41	No	No	Si	No
Estudiante 42	No	Si	Si	No
Estudiante 43	No	Si	No	No
Estudiante 44	No	No	Si	No
Estudiante 45	No	No	Si	Si
Estudiante 46	Si	Si	Si	No
Estudiante 47	No	No	Si	No
Estudiante 48	No	Si	Si	No
Estudiante 49	Si	No	Si	No
Estudiante 50	Si	No	Si	No
Estudiante 51	No	No	Si	No
Estudiante 52	Si	Si	Si	No
Estudiante 53	No	No	Si	Si
Estudiante 54	No	Si	Si	No
Estudiante 55	No	No	No	No
Estudiante 56	No	No	Si	No
Estudiante 57	No	Si	Si	No
Estudiante 58	Si	Si	Si	No
Estudiante 59	No	Si	Si	No
Estudiante 60	No	Si	No	No
Estudiante 61	No	No	Si	No
Estudiante 62	No	Si	No	Si
Estudiante 63	No	No	Si	Si
Estudiante 64	Si	Si	Si	No
Estudiante 65	No	No	Si	No
Estudiante 66	No	Si	No	Si
Estudiante 67	No	No	Si	Si
Estudiante 68	Si	Si	Si	No
Estudiante 69	No	No	Si	No
Estudiante 70	No	No	Si	No
Estudiante 71	No	Si	Si	No

Estudiante 72	Si	Si	No	No
Estudiante 73	No	Si	Si	Si
Estudiante 74	No	No	No	No
Estudiante 75	No	No	Si	No
Estudiante 76	No	Si	Si	Si
Estudiante 77	No	No	No	Si
Estudiante 78	No	Si	No	No
Estudiante 79	No	No	Si	No
Estudiante 80	Si	Si	No	Si
Estudiante 81	Si	Si	No	No
Estudiante 82	No	Si	No	Si
Estudiante 83	No	No	Si	Si
Estudiante 84	No	Si	Si	Si
Estudiante 85	No	No	Si	No
Estudiante 86	No	No	Si	Si
Estudiante 87	No	Si	No	No
Estudiante 88	No	Si	Si	No
Estudiante 89	Si	Si	No	No
Estudiante 90	Si	No	No	No
Estudiante 91	Si	No	No	No
Estudiante 92	Si	No	Si	No
Estudiante 93	Si	No	Si	No
Estudiante 94	No	No	Si	No
Estudiante 95	No	No	No	No
Estudiante 96	No	Si	Si	No
Estudiante 97	No	Si	No	No
Estudiante 98	No	Si	No	No
Estudiante 99	No	Si	Si	Si
Estudiante 100	No	No	Si	No
Estudiante 101	No	Si	No	No
Estudiante 102	No	Si	No	No
Estudiante 103	No	No	No	No
Estudiante 104	No	Si	No	No

Tabla 6

Conversión de Dataset histórica a Dataset definitiva 2018-1 al 2023-2 Factor Socioeconómico

Estudiante	Dificultad Financiera	Trabajo Tiempo Completo	Apoyo Familiar
Estudiante 1	No	No	Si
Estudiante 2	Si	Si	Si
Estudiante 3	No	Si	No
Estudiante 4	No	Si	Si
Estudiante 5	Si	Si	No
Estudiante 6	Si	No	Si
Estudiante 7	No	Si	Si
Estudiante 8	Si	No	Si
Estudiante 9	No	Si	Si
Estudiante 10	Si	Si	No
Estudiante 11	No	Si	No
Estudiante 12	No	No	Si
Estudiante 13	No	No	No
Estudiante 14	No	Si	Si
Estudiante 15	No	Si	Si
Estudiante 16	No	No	No
Estudiante 17	No	No	No
Estudiante 18	No	Si	Si
Estudiante 19	No	Si	No
Estudiante 20	No	Si	Si
Estudiante 21	No	No	Si
Estudiante 22	No	No	No
Estudiante 23	Si	Si	No
Estudiante 24	Si	Si	No
Estudiante 25	No	No	No
Estudiante 26	Si	No	Si
Estudiante 27	No	Si	Si
Estudiante 28	No	Si	No
Estudiante 29	No	Si	No
Estudiante 30	No	No	Si
Estudiante 31	No	No	No

Estudiante 32	No	Si	Si
Estudiante 33	No	Si	Si
Estudiante 34	No	No	No
Estudiante 35	No	Si	Si
Estudiante 36	No	Si	Si
Estudiante 37	No	Si	Si
Estudiante 38	No	No	No
Estudiante 39	No	No	No
Estudiante 40	No	Si	Si
Estudiante 41	No	Si	No
Estudiante 42	No	Si	Si
Estudiante 43	No	No	No
Estudiante 44	Si	No	Si
Estudiante 45	No	Si	Si
Estudiante 46	No	Si	No
Estudiante 47	No	Si	No
Estudiante 48	No	No	Si
Estudiante 49	No	No	Si
Estudiante 50	No	Si	Si
Estudiante 51	Si	Si	No
Estudiante 52	No	No	No
Estudiante 53	Si	No	Si
Estudiante 54	No	Si	Si
Estudiante 55	No	Si	No
Estudiante 56	No	Si	No
Estudiante 57	No	No	Si
Estudiante 58	No	No	No
Estudiante 59	No	Si	Si
Estudiante 60	No	Si	Si
Estudiante 61	No	No	No
Estudiante 62	No	No	No
Estudiante 63	No	Si	Si
Estudiante 64	No	Si	No
Estudiante 65	No	No	No
Estudiante 66	No	No	No
Estudiante 67	No	Si	Si
Estudiante 68	No	Si	No
Estudiante 69	No	Si	Si

Estudiante 70	No	Si	No
Estudiante 71	No	Si	Si
Estudiante 72	No	No	Si
Estudiante 73	No	No	No
Estudiante 74	Si	Si	No
Estudiante 75	Si	No	Si
Estudiante 76	Si	Si	No
Estudiante 77	Si	Si	Si
Estudiante 78	No	No	No
Estudiante 79	Si	No	Si
Estudiante 80	Si	No	No
Estudiante 81	Si	No	Si
Estudiante 82	No	Si	Si
Estudiante 83	No	Si	No
Estudiante 84	Si	Si	No
Estudiante 85	No	No	Si
Estudiante 86	Si	Si	No
Estudiante 87	Si	No	No
Estudiante 88	No	Si	Si
Estudiante 89	No	No	Si
Estudiante 90	No	Si	No
Estudiante 91	Si	Si	Si
Estudiante 92	No	No	Si
Estudiante 93	No	Si	Si
Estudiante 94	Si	Si	No
Estudiante 95	No	No	Si
Estudiante 96	Si	Si	Si
Estudiante 97	No	Si	No
Estudiante 98	No	Si	Si
Estudiante 99	Si	Si	No
Estudiante 100	Si	No	Si
Estudiante 101	No	Si	Si
Estudiante 102	Si	No	Si
Estudiante 103	No	Si	Si
Estudiante 104	Si	Si	No

4.1.3. Preprocesamiento de Datos

Para esta etapa, se preparó los datos establecidos según tablas 4, 5 y 6 para el análisis.

Estructura JSON propuesto

Los archivos JSON son archivos de texto básicos, se utilizan para intercambiar datos organizados de forma sencilla y legible, JSON se utiliza para poder manejar grandes volúmenes de datos en tiempo real.

Conversión de Dataset definitiva a estructura JSON

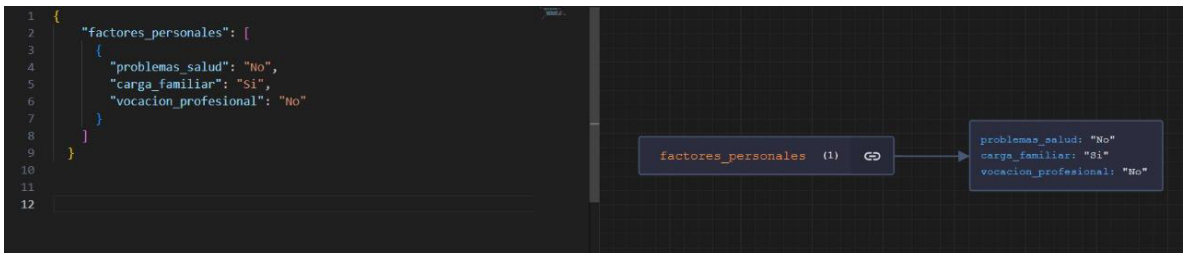
Figura 18

Estructura JSON

Primero se muestran los factores personales tomando en cuenta el Dataset donde se recopila a través del archivo en formato JSON en consideración al cuadro de criterios por cada uno de los factores puestos en la investigación.

Figura 19

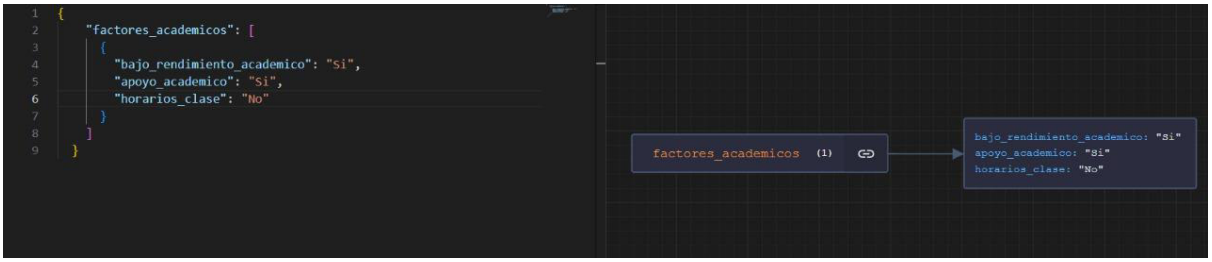
Estructura JSON de Factores Personales



Después están los factores académicos tomando en cuenta el Dataset donde se recopila a través del archivo en formato JSON en consideración al cuadro de criterios por cada uno de los factores puestos en la investigación.

Figura 20

Estructura JSON de Factores Académicos



Por último, se visualizan los factores socioeconómicos tomando en cuenta el Dataset donde se recopila a través del archivo en formato JSON en consideración al cuadro de criterios por cada uno de los factores puestos en la investigación.

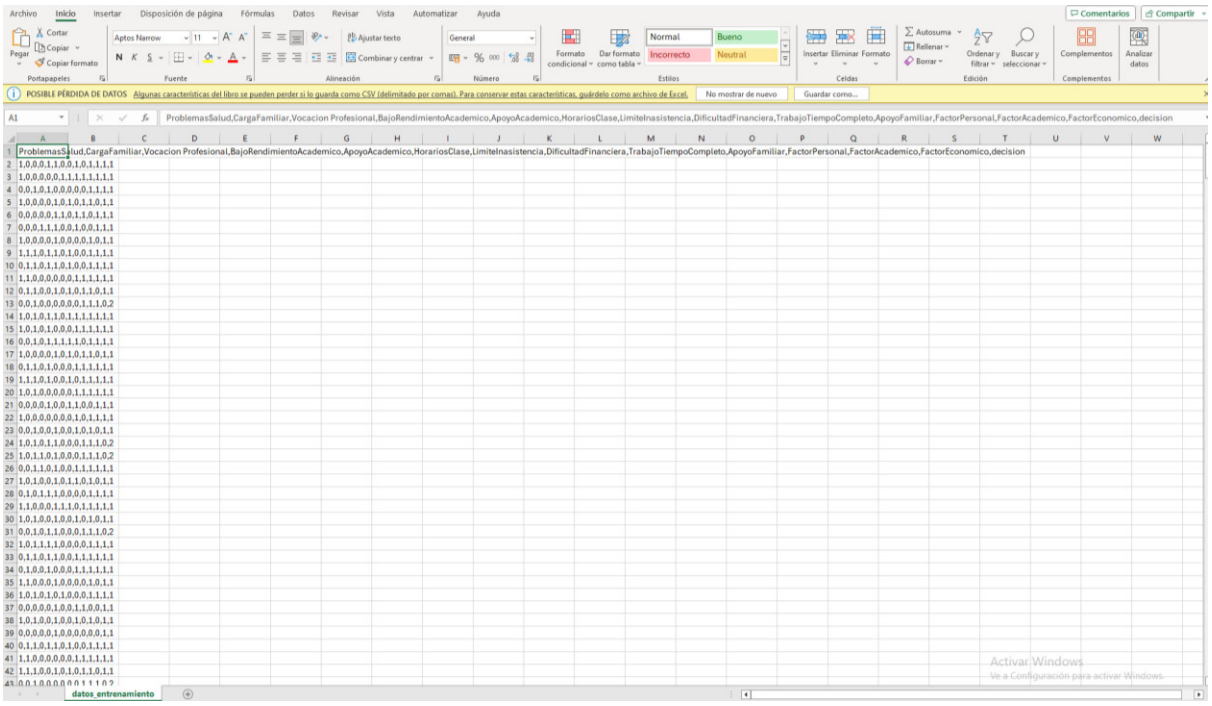
Figura 21

Estructura JSON de Factores Socioeconómicos

Asimismo, estos factores que ya se plasmaron dentro de los archivos JSON del modelo, se pueden visualizar con mejor precisión a través del siguiente enlace: https://docs.google.com/document/d/17P_RzXyrZxdw3uaDLff3Nx6zWbNsFU7h/edit

Figura 22

Dataset Definitivo con campos limpios



En la Figura 18, nos muestra la data definitiva después de ingresar los datos por factores dentro del JSON.

4.1.4. Selección de Características

Para esta etapa, se identificó y seleccionó lo factores que influyen en la deserción estudiantil (factores personales, factores académicos, factores socioeconómicos), teniendo en cuenta los criterios existentes con mayor índice de deserción.

Tabla 7

Factores del Modelo

Factores	Criterios	Descripción
Factores Personales	Problemas de salud	La ansiedad, la depresión u otros problemas de salud mental pueden interferir con el rendimiento académico y hacer que sea difícil para los estudiantes continuar con sus estudios.
	Carga familiar	Los estudiantes que tienen responsabilidades familiares, como cuidar de hijos o familiares enfermos, pueden encontrar difícil equilibrar sus

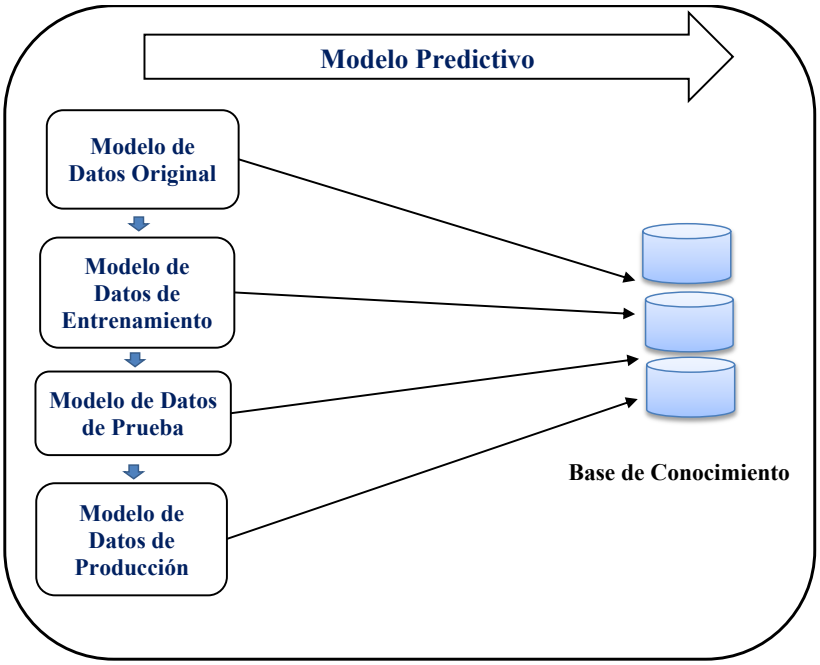
		responsabilidades familiares con sus estudios universitarios, lo que puede llevar a la deserción.
	Vocación profesional	Los estudiantes que no tienen claro qué quieren lograr con su educación universitaria pueden sentirse perdidos y desmotivados.
Factores Académicos	Bajo rendimiento académico	Los estudiantes que luchan por alcanzar los estándares académicos requeridos pueden sentir desmotivación y desaliento, lo puede llevar a la deserción.
	Apoyo académico	La falta de acceso a tutores, asesores académicos u otros recursos de apoyo puede dejar a los estudiantes sintiéndose solos y desorientados cuando enfrentan desafíos académicos. Sin el apoyo adecuado, algunos estudiantes pueden optar por abandonar.
	Horarios de clase	Los horarios de clases poco convenientes o conflictivos pueden dificultar que los estudiantes asistan regularmente a sus cursos, lo que puede afectar negativamente su desempeño académico y su motivación para continuar con sus estudios.
	Límite de inasistencias	número máximo de veces que un estudiante está ausente de clases afectando su situación académica, este límite puede variar según las políticas de la institución educativa.
Factores Socioeconómicos	Dificultades financieras	Los costos de matrícula, libros, vivienda y otros gastos relacionados con la universidad pueden ser prohibitivos para algunos estudiantes, lo que los lleva a abandonar sus estudios debido a la falta de recursos económicos.
	Trabajo a tiempo completo	Muchos estudiantes necesitan trabajar a tiempo completo para poder costear sus estudios universitarios o para contribuir al sustento de sus familias. La carga laboral puede interferir con el tiempo y la energía que los estudiantes pueden dedicar a sus estudios, lo que aumenta el riesgo de deserción.
	Apoyo familiar	Los estudiantes que no cuentan con el apoyo emocional o financiero de sus familias pueden enfrentar dificultades adicionales para sobrellevar los desafíos académicos y pueden ser más propensos a abandonar la universidad.

4.1.5. Desarrollo del Modelo

Para esta etapa, se utilizaron técnicas de Machine Learning para construir un modelo predictivo que estime la probabilidad de deserción basándose en factores personales (Figura 19).

Figura 23

Esquema del modelo propuesto



El modelo es un Bosque Aleatorio (Random Forest) con 100 estimadores, este parámetro indica el número de árboles de decisión que se construirán en el modelo.

Figura 24

Modelo Desarrollado

```
# Dividir en conjunto de entrenamiento y prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo RandomForest
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
```

En este árbol número 1 del modelo, contiene varios nodos, estos nodos internos representan puntos de decisión donde se evalúa en que camino debe continuar. Además, presenta un nodo Hoja (o nodo Final) en el cual, representa la decisión final o la clasificación,

por último, tenemos las condiciones, que son la forma de decidir el camino del nodo. Para este caso, el valor es 0.5 para las decisiones.

Figura 25

Modelo Desarrollado

```
|--- Asistencia <= 30.50
|   |--- Promedio <= 10.10
|   |   |--- IngresosFamiliares <= 2020.50
|   |   |   |--- class: 1.0
|   |   |--- IngresosFamiliares >  2020.50
|   |   |   |--- Edad <= 34.50
|   |   |   |   |--- Asistencia <= 10.00
|   |   |   |   |   |--- Promedio <= 8.65
|   |   |   |   |   |   |--- class: 0.0
|   |   |   |   |   |--- Promedio >  8.65
|   |   |   |   |   |   |--- class: 3.0
|   |   |   |   |--- Asistencia >  10.00
|   |   |   |   |   |--- class: 3.0
|   |   |   |--- Edad >  34.50
|   |   |   |   |--- class: 0.0
|   |--- Promedio >  10.10
|   |   |--- IngresosFamiliares <= 2399.00
|   |   |   |--- class: 3.0
|   |   |--- IngresosFamiliares >  2399.00
|   |   |   |--- class: 0.0
|--- Asistencia >  30.50
|   |--- Promedio <= 19.90
|   |   |--- IngresosFamiliares <= 3715.00
|   |   |   |--- Edad <= 35.50
|   |   |   |   |--- ActividadesExtracurriculares <= 1.50
|   |   |   |   |   |--- Asistencia <= 51.50
|   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |--- Asistencia >  51.50
|   |   |   |   |   |   |--- Edad <= 31.50
|   |   |   |   |   |   |   |--- Promedio <= 19.12
|   |   |   |   |   |   |   |   |--- Promedio <= 8.85
|   |   |   |   |   |   |   |   |   |--- class: 2.0
```

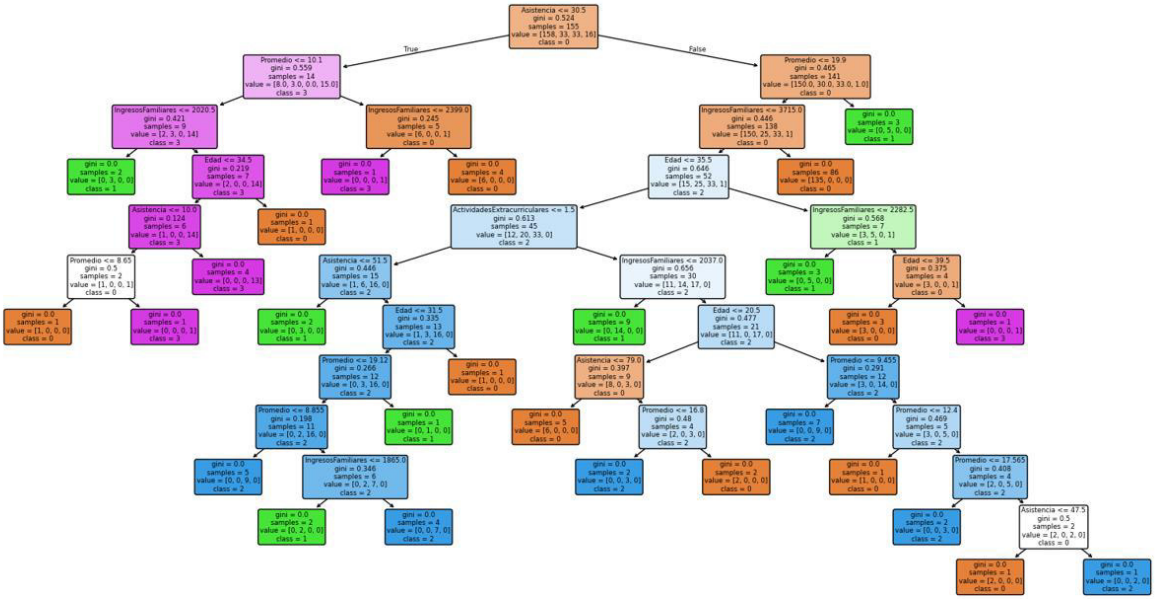
Ahora bien, para representarlo de una mejor manera, describiremos algunos conceptos de los árboles.

Samples (muestras): Es el número de instancias de datos que llegan a un nodo específico del árbol durante su construcción. Indica cuántos puntos de datos cumplen las condiciones para llegar a ese nodo en particular.

Gini: Es una medida de impureza en un nodo del árbol de decisión. Muestra qué tan mezcladas están las clases de las muestras en ese nodo.

Figura 26

Árbol 1: Árbol de decisión



En este árbol 1 se divide el conjunto de datos en dos ramas, en la característica Asistencia, si es menor o igual a 30.5, la instancia se mueve hacia la izquierda, o de lo contrario, se mueve hacia la rama derecha.

El valor Gini representa la impureza de cada nodo, cuanto más bajo sea el valor Gini, entonces más puras son las muestras en ese nodo.

En el nodo de profundidad 3 que está a la izquierda, se tiene el valor (value) = [2,3,0,14] significa que hay 2 instancias de la clase 0, 3 instancias de la clase 1, 0 instancias de la clase 2 y 14 instancias de la clase 3.

El sample es el número de muestras, es decir, el número de instancias de datos que llega al nodo durante el proceso de clasificación, o mediante el entrenamiento.

4.2. Evaluación de la precisión del modelo predictivo basado en Machine Learning en la identificación de factores personales

Para esta etapa, se testeó el modelo con un conjunto de datos de prueba para evaluar su precisión y ajustar parámetros.

Figura 27

Validación del Modelo

```python
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusion_matrix
import joblib
import json

# Función para entrenar el modelo
def entrenar_modelo():
    try:
        # Leer el archivo JSON con los datos de entrenamiento
        with open('real_data.json', 'r') as file:
            datos_entrenamiento = json.load(file)

        print("Datos de entrenamiento leídos exitosamente.")

        # Verificar si 'Desercion' está presente en los datos
        if 'Desercion' not in datos_entrenamiento.get("DatosEntrenamiento", [])[0].keys():
            raise ValueError("'Desercion' no encontrado en las columnas de los datos de entrenamiento")

        # Crear un DataFrame con los datos de entrenamiento
        df_entrenamiento = pd.DataFrame(datos_entrenamiento["DatosEntrenamiento"])

        # Dividir los datos en características (X) y etiquetas (y)
        X = df_entrenamiento.drop(['Desercion', 'FactorDesercion'], axis=1)
        y = df_entrenamiento['FactorDesercion']
```

```python
# Dividir en conjunto de entrenamiento y prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo RandomForest
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
modelo.fit(X_train, y_train)

# Realizar predicciones sobre el conjunto de prueba
y_pred = modelo.predict(X_test)

# Calcular las métricas
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='weighted')
recall = recall_score(y_test, y_pred, average='weighted')
f1 = f1_score(y_test, y_pred, average='weighted')
conf_matrix = confusion_matrix(y_test, y_pred)

# Guardar el modelo entrenado
joblib.dump(modelo, 'modelo_desercion_estudiantil.joblib')
joblib.dump((X_test, y_test), 'conjunto_prueba.joblib')

# Mostrar las métricas
print(f"\nMétricas de evaluación del modelo:")
print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")
print(f"Confusion Matrix: \n{conf_matrix}")
```

El en el código, muestra que es lo que se está avanzando, primero, se importan las librerías necesarias como el RandomForestClasification, y algunos de precisión, además del train - test para separar los datos en datos de entrenamiento y datos de prueba. El en el código, muestra que es lo que se está avanzando, primero, se importan las librerías necesarias como el RandomForestClasification, y algunos de precisión, además del train - test para separar los datos en datos de entrenamiento y datos de prueba.

Primero, se lee los datos de entrenamiento, después, se procesa los datos, en X se guardará los valores necesitados para predecir el modelo y en Y se guardará las respuestas o decisiones.

El train - test nos ayudará a separar los datos en entrenamiento (train) y en testeo o prueba (test) con una proporción de 80% y 20% donde el 20% (0.2) va para el conjunto de testeo o prueba.

En el modelado, se crea un clasificador de Random Forest con 100 estimadores (arboles) y lo entrena en el conjunto de entrenamiento.

Por último, tenemos la evaluación del modelo, en el cual usamos precisión, recall, F1-score, entre otros. En la figura 24 se puede observar los resultados de precisión de la predicción realizada.

Figura 28

Resultados de precisión de la predicción

```
Métricas de evaluación del modelo:
Accuracy: 0.9672131147540983
Precision: 0.9687380861608845
Recall: 0.9672131147540983
F1 Score: 0.9642406773554315
Confusion Matrix:
[[41  0  0  0]
 [ 0 11  0  0]
 [ 1  0  6  0]
 [ 1  0  0  1]]
```

Luego del desarrollo del modelo y puesto a prueba, se determinó que la precisión de dicho modelo presenta un valor de 1.0 teniendo en cuenta los cambios que se hicieron en los Dataset de los estudiantes implementados en el presente entrenamiento de la IA.

A continuación, se definen y se señalan las fórmulas de cada uno de los valores calculados anteriormente

Precisión (Test): 0.9687380861608845

La precisión es la proporción de verdaderos positivos sobre todas las instancias clasificadas como positivas (verdaderos positivos + falsos positivos).

$$\text{Precisión} = \frac{VP}{VP + FP}$$

Recall (Test): 0.9672131147540983

Recall o sensibilidad es la proporción de instancias positivas que se clasificaron correctamente como positivas entre todas las instancias positivas en los datos de prueba, es decir, la proporción de verdaderos positivos (TP) entre la suma de VP y FN.

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

F1 Score (Test): 0.9642406773554315

F1 Score es una métrica en el cual, es la media armónica de la precisión y el recall, es decir:

$$\text{F1 Score} = 2\,x\frac{precisión\ x\ recall}{precisión\ + recall}$$

Figura 29

Parámetros de precisión y ajustes – codificaciones

```
# Dividir en conjunto de entrenamiento y prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo RandomForest
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
modelo.fit(X_train, y_train)
```

En esta parte separamos el modelo en entrenamiento y validación, con una proporción de 0.2 para el testo y 0.8 para el entrenamiento, y se entrena el modelo con un total de 100 estimadores o árboles.

Evaluaciones de Referencia:

Eficiencia: Precisión x Velocidad

Donde velocidad = 1 – tiempo de respuesta / tiempo permitido

Tiempo permitido = 1 segundo

Tiempo de respuesta = 97 ms = 0.097 seg por lo tanto la eficiencia = 0.9687x(1 - 0.097) = 0.9687x0.903 = 0.8747361

Efectividad: promedio de todas las precisiones:

sumatoria de precisión / total características = (1 + 1 + 6/7 + 1/2) / 4 = 0.839285714…

4.2.1. Recopilación de Datos Académicos

Para esta etapa, se obtuvieron los datos sobre el rendimiento académico, participación en clases, notas y otras métricas académicas.

Tabla 8

Criterios del Factor Académico

Estudiante	Bajo Rendimiento Académico	Apoyo Académico	Horarios Clase	Limite Inasistencia
Estudiante 1	No	No	No	No
Estudiante 2	No	Si	Si	No
Estudiante 3	No	Si	No	No
Estudiante 4	No	Si	No	No
Estudiante 5	No	Si	Si	Si
Estudiante 6	No	No	Si	No
Estudiante 7	No	Si	No	No
Estudiante 8	No	Si	No	No
Estudiante 9	No	No	No	No
Estudiante 10	No	Si	No	No
Estudiante 11	No	No	Si	No
Estudiante 12	No	Si	Si	No
Estudiante 13	Si	Si	Si	No
Estudiante 14	No	Si	Si	No
Estudiante 15	No	Si	No	No

Estudiante 16	No	No	Si	No
Estudiante 17	No	Si	No	Si
Estudiante 18	No	No	Si	No
Estudiante 19	No	No	Si	No
Estudiante 20	No	Si	Si	No
Estudiante 21	Si	Si	No	No
Estudiante 22	No	Si	Si	Si
Estudiante 23	No	No	No	No
Estudiante 24	Si	No	Si	No
Estudiante 25	Si	Si	Si	No
Estudiante 26	No	No	Si	Si
Estudiante 27	No	Si	Si	No
Estudiante 28	No	No	No	No
Estudiante 29	No	No	Si	No
Estudiante 30	No	Si	Si	No
Estudiante 31	Si	Si	Si	No
Estudiante 32	No	Si	Si	No
Estudiante 33	No	Si	No	No
Estudiante 34	No	No	Si	No
Estudiante 35	No	No	Si	No
Estudiante 36	No	Si	Si	No
Estudiante 37	No	Si	No	No
Estudiante 38	No	No	Si	No
Estudiante 39	No	Si	No	Si
Estudiante 40	No	No	Si	No
Estudiante 41	No	No	Si	No
Estudiante 42	No	Si	Si	No
Estudiante 43	No	Si	No	No
Estudiante 44	No	No	Si	No
Estudiante 45	No	No	Si	Si
Estudiante 46	Si	Si	Si	No
Estudiante 47	No	No	Si	No
Estudiante 48	No	Si	Si	No
Estudiante 49	Si	No	Si	No
Estudiante 50	Si	No	Si	No
Estudiante 51	No	No	Si	No
Estudiante 52	Si	Si	Si	No
Estudiante 53	No	No	Si	Si
Estudiante 54	No	Si	Si	No
Estudiante 55	No	No	No	No
Estudiante 56	No	No	Si	No
Estudiante 57	No	Si	Si	No
Estudiante 58	Si	Si	Si	No
Estudiante 59	No	Si	Si	No
Estudiante 60	No	Si	No	No
Estudiante 61	No	No	Si	No
Estudiante 62	No	Si	No	Si

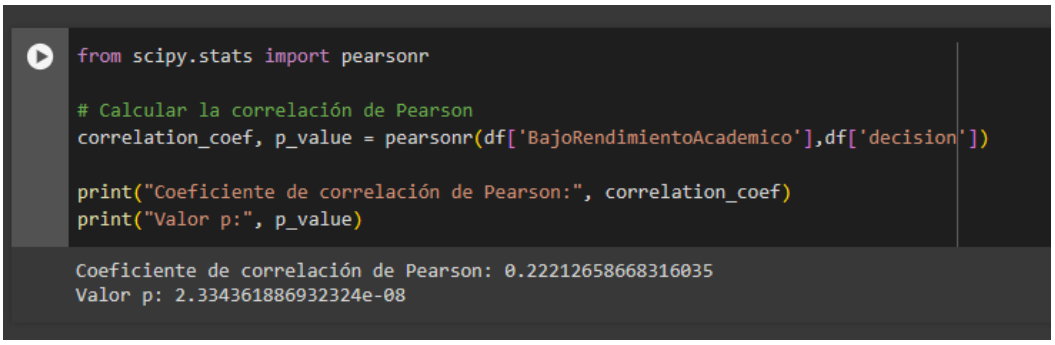
Estudiante 63	No	No	Si	Si
Estudiante 64	Si	Si	Si	No
Estudiante 65	No	No	Si	No
Estudiante 66	No	Si	No	Si
Estudiante 67	No	No	Si	Si
Estudiante 68	Si	Si	Si	No
Estudiante 69	No	No	Si	No
Estudiante 70	No	No	Si	No
Estudiante 71	No	Si	Si	No
Estudiante 72	Si	Si	No	No
Estudiante 73	No	Si	Si	Si
Estudiante 74	No	No	No	No
Estudiante 75	No	No	Si	No
Estudiante 76	No	Si	Si	Si
Estudiante 77	No	No	No	Si
Estudiante 78	No	Si	No	No
Estudiante 79	No	No	Si	No
Estudiante 80	Si	Si	No	Si
Estudiante 81	Si	Si	No	No
Estudiante 82	No	Si	No	Si
Estudiante 83	No	No	Si	Si
Estudiante 84	No	Si	Si	Si
Estudiante 85	No	No	Si	No
Estudiante 86	No	No	Si	Si
Estudiante 87	No	Si	No	No
Estudiante 88	No	Si	Si	No
Estudiante 89	Si	Si	No	No
Estudiante 90	Si	No	No	No
Estudiante 91	Si	No	No	No
Estudiante 92	Si	No	Si	No
Estudiante 93	Si	No	Si	No
Estudiante 94	No	No	Si	No
Estudiante 95	No	No	No	No
Estudiante 96	No	Si	Si	No
Estudiante 97	No	Si	No	No
Estudiante 98	No	Si	No	No
Estudiante 99	No	Si	Si	Si
Estudiante 100	No	No	Si	No
Estudiante 101	No	Si	No	No
Estudiante 102	No	Si	No	No
Estudiante 103	No	No	No	No
Estudiante 104	No	Si	No	No

4.2.2. Análisis Exploratorio

Para esta etapa, se explora la relación entre el rendimiento académico y la deserción estudiantil, para ver la relación, podemos ver el coeficiente de relación, en esta parte, nos indica que tan relacionado están las variables, y también, y con cuanta confianza podemos decir que son variables que tienen una relación entre ellas.

Figura 30

Coeficiente de correlación de Pearson



El coeficiente de correlación de Pearson fue de: 0.22212… lo que indica que hay una correlación positiva débil y el valor p que es de 2.33e-08 indica que esta correlación es significativa, es decir, la probabilidad de que las variables sean dependientes es más del 99.999%.

4.2.3. Ajuste del Modelo

Para esta etapa, se refina el modelo para mejorar su capacidad de identificar correctamente a los estudiantes en riesgo debido a factores académicos, el modelo ya generaba una precisión perfecta, por lo que ajustarlo más, podría llevar a un sobreajuste, en el cual, el modelo predice de manera exacta en casos cercanos al entrenamiento, pero lejanos a los datos de prueba.

4.3. Medición de la efectividad del modelo predictivo basado en Machine Learning para cuantificar la influencia de los factores socioeconómicos

4.3.1. Recopilación de Datos Socioeconómicos

Para esta etapa, se reunió la información sobre el estatus socioeconómico de los estudiantes, incluyendo ingreso familiar, empleo de los padres y acceso a recursos educativos.

Tabla 9

Criterios del Factor Socioeconómico

Estudiante	Nota Final	Dificultad Financiera	Trabajo Tiempo Completo	Apoyo Familiar
Estudiante 1	9	No	No	Si
Estudiante 2	7	Si	Si	Si
Estudiante 3	3	No	Si	No
Estudiante 4	12	No	Si	Si
Estudiante 5	13	Si	Si	No
Estudiante 6	12	Si	No	Si
Estudiante 7	12	No	Si	Si
Estudiante 8	5	Si	No	Si
Estudiante 9	10	No	Si	Si
Estudiante 10	15	Si	Si	No
Estudiante 11	2	No	Si	No
Estudiante 12	15	No	No	Si
Estudiante 13	2	No	No	No
Estudiante 14	1	No	Si	Si
Estudiante 15	10	No	Si	Si
Estudiante 16	2	No	No	No
Estudiante 17	1	No	No	No
Estudiante 18	10	No	Si	Si
Estudiante 19	5	No	Si	No
Estudiante 20	6	No	Si	Si
Estudiante 21	6	No	No	Si
Estudiante 22	12	No	No	No
Estudiante 23	8	Si	Si	No
Estudiante 24	2	Si	Si	No
Estudiante 25	7	No	No	No
Estudiante 26	2	Si	No	Si
Estudiante 27	10	No	Si	Si

Estudiante 28	11	No	Si	No
Estudiante 29	2	No	Si	No
Estudiante 30	15	No	No	Si
Estudiante 31	2	No	No	No
Estudiante 32	1	No	Si	Si
Estudiante 33	10	No	Si	Si
Estudiante 34	2	No	No	No
Estudiante 35	10	No	Si	Si
Estudiante 36	1	No	Si	Si
Estudiante 37	10	No	Si	Si
Estudiante 38	2	No	No	No
Estudiante 39	1	No	No	No
Estudiante 40	10	No	Si	Si
Estudiante 41	5	No	Si	No
Estudiante 42	6	No	Si	Si
Estudiante 43	8	No	No	No
Estudiante 44	13	Si	No	Si
Estudiante 45	16	No	Si	Si
Estudiante 46	8	No	Si	No
Estudiante 47	2	No	Si	No
Estudiante 48	15	No	No	Si
Estudiante 49	16	No	No	Si
Estudiante 50	10	No	Si	Si
Estudiante 51	12	Si	Si	No
Estudiante 52	7	No	No	No
Estudiante 53	2	Si	No	Si
Estudiante 54	10	No	Si	Si
Estudiante 55	11	No	Si	No
Estudiante 56	2	No	Si	No
Estudiante 57	15	No	No	Si
Estudiante 58	2	No	No	No
Estudiante 59	1	No	Si	Si
Estudiante 60	10	No	Si	Si
Estudiante 61	2	No	No	No
Estudiante 62	1	No	No	No
Estudiante 63	16	No	Si	Si
Estudiante 64	8	No	Si	No
Estudiante 65	2	No	No	No
Estudiante 66	1	No	No	No
Estudiante 67	16	No	Si	Si

Estudiante 68	8	No	Si	No
Estudiante 69	10	No	Si	Si
Estudiante 70	5	No	Si	No
Estudiante 71	6	No	Si	Si
Estudiante 72	6	No	No	Si
Estudiante 73	12	No	No	No
Estudiante 74	8	Si	Si	No
Estudiante 75	16	Si	No	Si
Estudiante 76	15	Si	Si	No
Estudiante 77	7	Si	Si	Si
Estudiante 78	8	No	No	No
Estudiante 79	13	Si	No	Si
Estudiante 80	16	Si	No	No
Estudiante 81	18	Si	No	Si
Estudiante 82	17	No	Si	Si
Estudiante 83	15	No	Si	No
Estudiante 84	15	Si	Si	No
Estudiante 85	13	No	No	Si
Estudiante 86	5	Si	Si	No
Estudiante 87	14	Si	No	No
Estudiante 88	18	No	Si	Si
Estudiante 89	20	No	No	Si
Estudiante 90	17	No	Si	No
Estudiante 91	17	Si	Si	Si
Estudiante 92	16	No	No	Si
Estudiante 93	10	No	Si	Si
Estudiante 94	12	Si	Si	No
Estudiante 95	9	No	No	Si
Estudiante 96	7	Si	Si	Si
Estudiante 97	3	No	Si	No
Estudiante 98	12	No	Si	Si
Estudiante 99	13	Si	Si	No
Estudiante 100	12	Si	No	Si
Estudiante 101	12	No	Si	Si
Estudiante 102	5	Si	No	Si
Estudiante 103	10	No	Si	Si
Estudiante 104	15	Si	Si	No

4.3.2. Análisis de Impacto Socioeconómico

Para esta etapa, se analizó cómo los factores socioeconómicos afectan las tasas de deserción, en esta parte, para saber que valores tienen mayor relación que otros valores, podemos usar la matriz de correlación, en el cual, nos ayuda a saber que valores dependen más de la deserción estudiantil.

Figura 31

Correlación de Pearson negativas

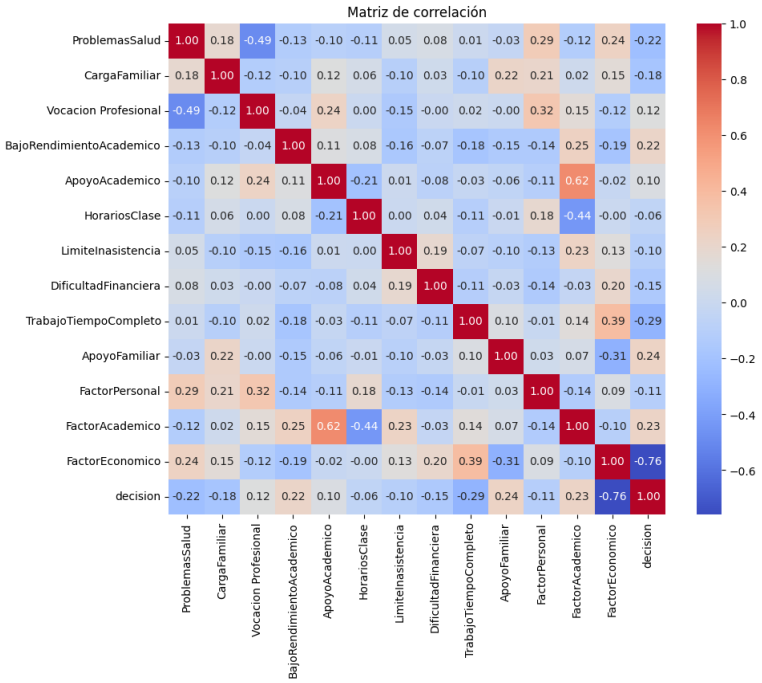
```
df.corr()['decision']

ProblemasSalud              -0.222360
CargaFamiliar               -0.184363
Vocacion Profesional         0.116892
BajoRendimientoAcademico     0.222127
ApoyoAcademico               0.096127
HorariosClase               -0.056274
LimiteInasistencia          -0.102199
DificultadFinanciera        -0.152985
TrabajoTiempoCompleto       -0.294030
ApoyoFamiliar                0.235736
FactorPersonal              -0.106448
FactorAcademico              0.230248
FactorEconomico             -0.757488
decision                     1.000000
```

En esta parte, vemos que las correlaciones son negativas, esto se debe a que los problemas de salud, carga familiar, afectan en la deserción estudiantil. Además, el dato con mayor correlación fue el factor económico, con -0.75, es decir, que está muy relacionado con la deserción estudiantil.

Figura 32

Matriz de Correlación

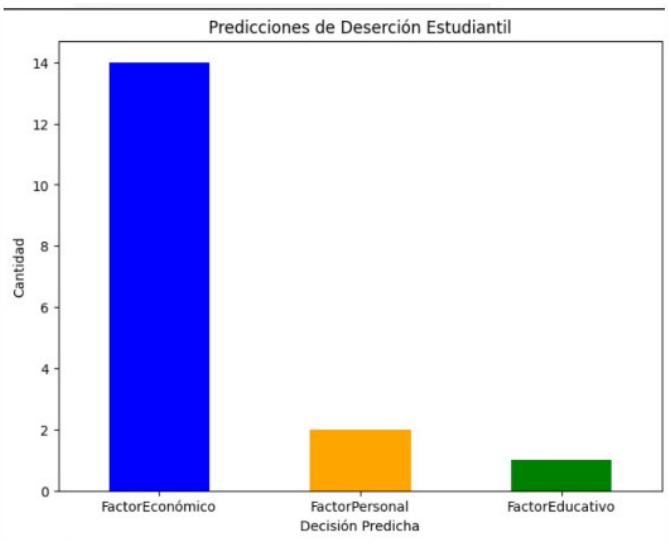


4.3.3. Medición de la Efectividad

Dado que el modelo no cometió ninguna equivocación al momento de realizar las pruebas, es decir, el modelo predijo con exactitud todos los casos dados, decimos que el modelo es efectivo para las predicciones.

Figura 33

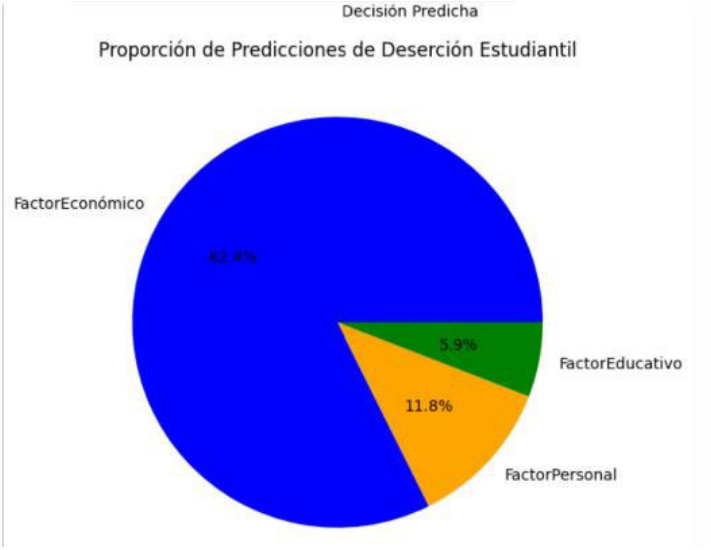
Predicciones de deserción estudiantil



En la figura 30 se observa de acuerdo con el modelo desarrollado el máximo valor de deserción de acuerdo al entrenamiento realizado se basa en el Factor Económico.

Figura 34

Proporciones de deserción estudiantil



Asimismo, en la figura 32, nos muestra con mayor detalle el nivel de porcentaje en los tres factores que interviene dentro de la investigación, factor económico con un 82.4%, factor personal con un 11.8%, factor educativo con un 5.9%.

4.4. Desarrollo de un modelo predictivo basado en Machine Learning contribuye en la determinación de estrategias efectivas para la reducción de la deserción estudiantil

Para cada factor de deserción, creamos un conjunto de estrategias que podría ayudar a los estudiantes

Estrategias para factores personales

En este caso podemos brindar ayuda si el factor deserción es debido a factores personales, las ayudas son las siguientes:

a) **Programas de acompañamiento psicológico**:

- Ofrecer servicios gratuitos de consejería psicológica para abordar problemas emocionales, familiares o de autoestima.
- Realizar talleres de manejo del estrés y ansiedad.

b) **Mentorías personalizadas:**
    - Asignar mentores (docentes o estudiantes avanzados) que brinden orientación académica y personal a estudiantes en riesgo.

c) **Fomento de habilidades blandas:**
    - Realizar capacitaciones en gestión del tiempo, técnicas de estudio y habilidades interpersonales.

Estrategias para factores académicos

En este caso podemos brindar ayuda si el factor deserción es debido a factores académicos, las ayudas son las siguientes:

a) **Refuerzo académico**
    - Crear programas de tutorías o asesorías académicas gratuitas para estudiantes con bajo rendimiento.
    - Proveer acceso a materiales de estudio digitales y guías interactivas.

b) **Flexibilidad en la carga académica:**
    - Permitir ajustes en la cantidad de créditos por semestre para evitar sobrecargas.

c) **Fortalecimiento de la relación con docentes:**
    - Capacitar a profesores en metodologías inclusivas y en estrategias para identificar estudiantes en riesgo.
    - Fomentar evaluaciones continuas y retroalimentación constructiva.

d) **Orientación vocacional:**
    - Realizar talleres que ayuden a los estudiantes a confirmar si la carrera elegida es la adecuada para sus intereses y habilidades.

e) **Programas de integración académica:**
    - Organizar eventos de bienvenida, actividades extracurriculares y grupos de

apoyo para estudiantes de primer ingreso.

Estrategias para factores socioeconómicos

En este caso podemos brindar ayuda si el factor deserción es debido a factores socioeconómicos, las ayudas son las siguientes:

a) Becas y ayudas financieras

- Ampliar los programas de becas y establecer fondos de emergencia para estudiantes con dificultades económicas.
- Crear convenios con empresas para financiar estudios a cambio de pasantías.

b) Oportunidades laborales flexibles

- Implementar bolsas de empleo para estudiantes con horarios ajustables a su carga académica.
- Facilitar pasantías remuneradas relacionadas con la carrera.

c) Reducción de costos indirectos

- Proveer recursos gratuitos como material digital, préstamos de laptops o servicio de transporte universitario.

d) Alianzas con el gobierno o empresas

- Gestionar subvenciones, préstamos estudiantiles con bajos intereses o donaciones para financiar infraestructura y programas educativos.

Primero, desarrollamos el modelo, como sabemos el modelo se basa en bosques aleatorios.

Figura 35

Modelo

```
with open('real_data.json', 'r') as file:
    datos_entrenamiento = json.load(file)

print("Datos de entrenamiento leídos exitosamente.")

# Verificar si 'Desercion' está presente en los datos
if 'Desercion' not in datos_entrenamiento.get("DatosEntrenamiento", [])[0].keys():
    raise ValueError("'Desercion' no encontrado en las columnas de los datos de entrenamiento")

# Crear un DataFrame con los datos de entrenamiento
df_entrenamiento = pd.DataFrame(datos_entrenamiento["DatosEntrenamiento"])

# Dividir los datos en características (X) y etiquetas (y)
X = df_entrenamiento.drop(['Desercion', 'FactorDesercion'], axis=1)
y = df_entrenamiento['FactorDesercion']

# Dividir en conjunto de entrenamiento y prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Crear y entrenar el modelo RandomForest
modelo = RandomForestClassifier(n_estimators=100, random_state=42)
modelo.fit(X_train, y_train)

# Realizar predicciones sobre el conjunto de prueba
y_pred = modelo.predict(X_test)
```

Primero, leemos los datos, los datos están en un archivo JSON. Después, las variables X serán las variables en el cual, se quitarán las columnas deserción y factor deserción y en Y será la columna factor deserción. Se separa en entrenamiento y testeo en un 80% y 20% y se entrena el modelo. El modelo responde con valores 0,1,2 y 3.

Figura 36

Diccionario de predicciones

```
# Diccionario para mapear las predicciones
mapa_descripcion = {
    0: "No hay deserción",
    1: "Factores personales",
    2: "Factores académicos",
    3: "Factores socioeconómicos"
}
```

Cada valor indica una categoría indica un factor, el 0 indica que no hay deserción, el 1 indica que, si hay deserción por factores personales, el 2 indica que si hay deserción por factores académicos y el 3 indica que hay deserción por factores socioeconómicos.

Ahora bien, se le añade un conjunto de estrategias para cada factor de deserción.

Figura 37

Estrategias Factores Predictivos Personales

```
# Estrategias para cada caso
estrategias = {
    "Factores personales": [
        {
            "estrategia": "Programas de acompañamiento psicológico",
            "acciones": [
                "Ofrecer servicios gratuitos de consejería psicológica para abordar problemas emocionales, familiares o de autoestima.",
                "Realizar talleres de manejo del estrés y ansiedad."
            ]
        },
        {
            "estrategia": "Mentorías personalizadas",
            "acciones": [
                "Asignar mentores (docentes o estudiantes avanzados) que brinden orientación académica y personal a estudiantes en riesgo."
            ]
        },
        {
            "estrategia": "Fomento de habilidades blandas",
            "acciones": [
                "Realizar capacitaciones en gestión del tiempo, técnicas de estudio y habilidades interpersonales."
            ]
        }
    ],
```

En este caso se agregan estrategias en factores personales.

Figura 38

Estrategias Factores Predictivos Académicos

```
"Factores académicos": [
    {
        "estrategia": "Refuerzo académico",
        "acciones": [
            "Crear programas de tutorías o asesorías académicas gratuitas para estudiantes con bajo rendimiento.",
            "Proveer acceso a materiales de estudio digitales y guías interactivas."
        ]
    },
    {
        "estrategia": "Flexibilidad en la carga académica",
        "acciones": [
            "Permitir ajustes en la cantidad de créditos por semestre para evitar sobrecargas."
        ]
    },
    {
        "estrategia": "Fortalecimiento de la relación con docentes",
        "acciones": [
            "Capacitar a profesores en metodologías inclusivas y en estrategias para identificar estudiantes en riesgo.",
            "Fomentar evaluaciones continuas y retroalimentación constructiva."
        ]
    },
    {
        "estrategia": "Orientación vocacional",
        "acciones": [
            "Realizar talleres que ayuden a los estudiantes a confirmar si la carrera elegida es la adecuada para sus intereses y habilidades."
        ]
    },
    {
        "estrategia": "Programas de integración académica",
        "acciones": [
            "Organizar eventos de bienvenida, actividades extracurriculares y grupos de apoyo para estudiantes de primer ingreso."
        ]
    }
],
```

Aquí tenemos las estrategias para factores académicos

Figura 39

Estrategias Factores Socioeconómicos

```
"Factores socioeconómicos": [
    {
        "estrategia": "Becas y ayudas financieras",
        "acciones": [
            "Ampliar los programas de becas y establecer fondos de emergencia para estudiantes con dificultades económicas.",
            "Crear convenios con empresas para financiar estudios a cambio de pasantías."
        ]
    },
    {
        "estrategia": "Oportunidades laborales flexibles",
        "acciones": [
            "Implementar bolsas de empleo para estudiantes con horarios ajustables a su carga académica.",
            "Facilitar pasantías remuneradas relacionadas con la carrera."
        ]
    },
    {
        "estrategia": "Reducción de costos indirectos",
        "acciones": [
            "Proveer recursos gratuitos como material digital, préstamos de laptops o servicio de transporte universitario."
        ]
    },
    {
        "estrategia": "Alianzas con el gobierno o empresas",
        "acciones": [
            "Gestionar subvenciones, préstamos estudiantiles con bajos intereses o donaciones para financiar infraestructura y programas educativos."
        ]
    }
]
```

Y por último tenemos las estrategias para factores socioeconómicos.

Prueba de hipótesis

Prueba de HE1

HE1: La precisión del modelo predictivo basado en Machine Learning es significativamente efectiva para identificar los factores personales que influyen en la deserción estudiantil.

H_0 = La precisión del modelo predictivo basado en Machine Learning **no es significativamente efectiva** para identificar los factores personales que influyen en la deserción estudiantil.

H_1 = La precisión del modelo predictivo basado en Machine Learning **es significativamente efectiva** para identificar los factores personales que influyen en la deserción estudiantil.

Procedimiento de la prueba de hipótesis:

Se requiere:

- Una medida justa para evaluar el rendimiento del modelo predictivo de Machine Learning en la identificación de factores personales que influyan en la deserción estudiantil.
- Un **dataset relevante** con y sin la adaptación del modelo predictivo basado en Machine Learning para entrenar y probar la data para identificar los factores personales que influyen en la deserción estudiantil.
- Debido a que se debe comparar el rendimiento del modelo predictivo basado en Machine Learning, se puede utilizar:
    - La **prueba t** si se comparan las **medias de la métrica de evaluación** entre los métodos.
    - La prueba de hipótesis para proporciones si estamos comparando tasas de eficiencia en término de tiempo y costos.
- A posteriori, se selecciona el **nivel sig.**, generalmente 0.05.
- Utilizar la **data recopilada** para calcular la **estadística de prueba** (diferencia de medias o diferencia de proporciones) y determinar si es estadísticamente significativa.
- Si p < umbral sig. establecido, entonces se invalida la H_o y se deduce la existencia de evidencia adecuada para sustentar la hipótesis alternativa, H_1.
- Las limitaciones del estudio de:
    - Sesgos en los datos,
    - Tamaño del conjunto de datos,
    - Representatividad de las muestras, etc.

H_1 = La precisión del modelo predictivo basado en Machine Learning **es significativamente efectiva** para identificar los factores personales que influyen en la deserción estudiantil.

- Valor alcanzado del modelo predictivo basado en Machine Learning adaptado con la data incrementada: recall = 0,80
- Valor alcanzado del modelo predictivo basado en Machine Learning con la data tradicional: recall = 0,20
- **Tamaño de la muestra**: 1000 estimadores o árboles.
- ***Prueba t*** para comparar el recall de los dos valores.

  Medias de las muestras $\overline{x1}, \overline{x2}$

  Desviaciones estándar de las muestras S_1, S_2

  Tamaño de las muestras n_1, n_2

  $\overline{x1}$=0.80

  $\overline{x2}$=0.20

  S_1=S_2=$\sqrt{\frac{0.80-(1-0.80)}{1000}} = \sqrt{\frac{0.60}{1000}} = 0.0244$

  $t= \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}}}$

  $t=\frac{0.80- 0.20}{\sqrt{\frac{0.0244^2}{1000}+\frac{0.0244^2}{1000}}}$

  $t=\frac{0.60}{\sqrt{0,00000059536+0,00000059536}}$

  $t=\frac{0.60}{0.001091}$

  t=549,85

- El grado de libertad se calculan con n-1 para cada grupo.

  gl = 1000 - 1 = 999

Se hace uso de la tabla t de Student o programas estadísticos para determinar el valor crítico correspondiente al α = 0.05 con 999 gl.

Figura 40

Tabla de Student - Hipótesis específica 1

n	$t_{0,55}$	$t_{0,60}$	$t_{0,70}$	$t_{0,80}$	$t_{0,90}$	$t_{0,95}$	$t_{0,975}$	$t_{0,99}$	$t_{0,995}$
1	0,1584	0,3249	0,7265	1,3764	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,1421	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,1366	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,1338	0,2707	0,5686	0,9410	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,1322	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,1311	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,1303	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,1297	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,1293	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,1289	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,1286	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,1283	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,1281	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,1280	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,1278	0,2579	0,5357	0,8662	1,3406	1,7531	2,1314	2,6025	2,9467
16	0,1277	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,1276	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,1274	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,1274	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,1273	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,1272	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,1271	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,1271	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,1270	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,1269	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,1269	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,1268	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,1268	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,1268	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,1267	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500
40	0,1265	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045
50	0,1263	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778
60	0,1262	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603
80	0,1261	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387
100	0,1260	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259
120	0,1259	0,2539	0,5258	0,8446	1,2886	1,6577	1,9799	2,3578	2,6174
∞	0,126	0,253	0,524	0,842	1,282	1,645	1,960	2,327	2,576

Comparando el valor de t (549,85) con el valor crítico (1.645), t > el valor crítico. Por lo tanto, se rechaza la H0. Se concluye afirmando que la precisión del modelo predictivo basado en Machine Learning **es significativamente efectiva** para identificar los factores personales que influyen en la deserción estudiantil.

Prueba de HE2

HE2: La confiabilidad del modelo predictivo basado en Machine Learning **determina el impacto** de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú.

H_0 = La confiabilidad del modelo predictivo basado en Machine Learning **no determina el impacto** de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú.

H_1 = La confiabilidad del modelo predictivo basado en Machine Learning **determina el impacto** de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú.

Procedimiento de la prueba de hipótesis:

Se requiere:

- Una medida justa para evaluar el rendimiento del modelo predictivo de Machine Learning en la identificación de factores académicos en la deserción estudiantil.
- Un **dataset relevante** con y sin la adaptación del modelo predictivo basado en Machine Learning para entrenar y probar la data para identificar los factores académicos que influyen en la deserción estudiantil.
- Debido a que se debe comparar el rendimiento del modelo predictivo basado en Machine Learning, se puede utilizar:
    - La **prueba t** si se comparan las **medias de la métrica de evaluación** entre los métodos.
    - La prueba de hipótesis para proporciones si estamos comparando tasas de eficiencia en término de tiempo y costos.
- A posteriori, se selecciona el **nivel sig.**, generalmente 0.05.
- Utilizar la **data recopilada** para calcular la **estadística de prueba** (diferencia de medias o diferencia de proporciones) y determinar si es estadísticamente significativa.
- Si p < umbral sig. establecido, entonces se invalida la H_o y se deduce la existencia de evidencia adecuada para sustentar la hipótesis alternativa, H_1.
- Las limitaciones del estudio de:

- Sesgos en los datos,
- Tamaño del conjunto de datos,
- Representatividad de las muestras, etc.

Prueba de hipótesis:

H_1 = La confiabilidad del modelo predictivo basado en Machine Learning **determina el impacto** de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú.

- Valor alcanzado del modelo predictivo basado en Machine Learning adaptado con la data incrementada: recall = 0,78
- Valor alcanzado del modelo predictivo basado en Machine Learning con la data tradicional: recall = 0,22
- **Tamaño de la muestra**: 1000 estimadores o árboles.
- ***Prueba t*** para comparar el recall de los dos valores.

  Medias de las muestras $\overline{x1}, \overline{x2}$

  Desviaciones estándar de las muestras S_1, S_2

  Tamaño de las muestras n_1, n_2

  $\overline{x1}$=0.78

  $\overline{x2}$=0.22

  S_1=S_2=$\sqrt{\frac{0.78-(1-0.78)}{1000}} = \sqrt{\frac{0.56}{1000}} = 0.0237$

  $t = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{{S_1}^2}{n_1}+\frac{{S_2}^2}{n_2}}}$

  $t = \frac{0.78- 0.22}{\sqrt{\frac{0.0237^2}{1000}+\frac{0.0237^2}{1000}}}$

  $t = \frac{0.56}{\sqrt{0{,}00000056169+0{,}00000056169}}$

$$t=\frac{0.56}{0.001060}$$

$$t=528,35$$

- El grado de libertad se calculan con n-1 para cada grupo.

  gl = 1000 - 1 = 999

Se hace uso de la tabla t de Student o programas estadísticos para determinar el valor crítico correspondiente al α = 0.05 con 999 gl.

Figura 41

Tabla de Student - Hipótesis específica 2

n	$t_{0,55}$	$t_{0,60}$	$t_{0,70}$	$t_{0,80}$	$t_{0,90}$	$t_{0,95}$	$t_{0,975}$	$t_{0,99}$	$t_{0,995}$
1	0,1584	0,3249	0,7265	1,3764	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,1421	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,1366	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,1338	0,2707	0,5686	0,9410	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,1322	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,1311	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,1303	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,1297	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,1293	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,1289	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,1286	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,1283	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,1281	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,1280	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,1278	0,2579	0,5357	0,8662	1,3406	1,7531	2,1314	2,6025	2,9467
16	0,1277	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,1276	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,1274	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,1274	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,1273	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,1272	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,1271	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,1271	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,1270	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,1269	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,1269	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,1268	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,1268	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,1268	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,1267	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500
40	0,1265	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045
50	0,1263	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778
60	0,1262	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603
80	0,1261	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387
100	0,1260	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259
120	0,1259	0,2539	0,5258	0,8446	1,2886	1,6577	1,9799	2,3578	2,6174
∞	0,126	0,253	0,524	0,842	1,282	1,645	1,960	2,327	2,576

Comparando el valor de t (528,35) con el valor crítico (1.645), t > el valor crítico. Por lo tanto, se rechaza la H0. Se concluye afirmando que la confiabilidad del modelo predictivo

basado en Machine Learning **determina el impacto** de los factores académicos en la deserción estudiantil en las Universidades Privadas del Perú.

Prueba de HE3

HE3: La efectividad del modelo predictivo basado en Machine Learning **cuantifica la influencia** de los factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú y permite la creación de estrategias financieras y de apoyo social para abordar eficazmente las necesidades de los estudiantes.

H_0 = La efectividad del modelo predictivo basado en Machine Learning **no cuantifica la influencia** de los factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú y permite la creación de estrategias financieras y de apoyo social para abordar eficazmente las necesidades de los estudiantes.

H_1 = La efectividad del modelo predictivo basado en Machine Learning **cuantifica la influencia** de los factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú y permite la creación de estrategias financieras y de apoyo social para abordar eficazmente las necesidades de los estudiantes.

Procedimiento de la prueba de hipótesis:

Se requiere:

- Una medida justa para evaluar el rendimiento del modelo predictivo de Machine Learning en la cuantificación de los factores socioeconómicos en la deserción estudiantil.
- Un **dataset relevante** con y sin la adaptación del modelo predictivo basado en Machine Learning para entrenar y probar la data para identificar los factores socioeconómicos que influyen en la deserción estudiantil.
- Debido a que se debe comparar el rendimiento del modelo predictivo basado en Machine Learning, se puede utilizar:

  - La **prueba t** si se comparan las **medias de la métrica de evaluación** entre los métodos.
  - La prueba de hipótesis para proporciones si estamos comparando tasas de eficiencia en término de tiempo y costos.

- A posteriori, se selecciona el **nivel sig.**, generalmente 0.05.
- Utilizar la **data recopilada** para calcular la **estadística de prueba** (diferencia de medias o diferencia de proporciones) y determinar si es estadísticamente significativa.
- Si p < umbral sig. establecido, entonces se invalida la H_o y se deduce la existencia de evidencia adecuada para sustentar la hipótesis alternativa, H_1.
- Las limitaciones del estudio de:
  - Sesgos en los datos,
  - Tamaño del conjunto de datos,
  - Representatividad de las muestras, etc.

Prueba de hipótesis:

H_1 = La efectividad del modelo predictivo basado en Machine Learning **cuantifica la influencia** de los factores socioeconómicos en la deserción estudiantil en las Universidades Privadas del Perú y permite la creación de estrategias financieras y de apoyo social para abordar eficazmente las necesidades de los estudiantes.

- Valor alcanzado del modelo predictivo basado en Machine Learning adaptado con la data incrementada: recall = 0,65
- Valor alcanzado del modelo predictivo basado en Machine Learning con la data tradicional: recall = 0,35
- **Tamaño de la muestra**: 1000 estimadores o árboles.
- ***Prueba t*** para comparar el recall de los dos valores.

Medias de las muestras $\overline{x1}, \overline{x2}$

Desviaciones estándar de las muestras S_1, S_2

Tamaño de las muestras n_1, n_2

$\overline{x1}$=0.65

$\overline{x2}$=0.35

S_1=S_2=$\sqrt{\frac{0.65-(1-0.65)}{1000}} = \sqrt{\frac{0.30}{1000}} = 0.0173$

$t = \frac{\overline{x_1}-\overline{x_2}}{\sqrt{\frac{{s_1}^2}{n_1}+\frac{{s_2}^2}{n_2}}}$

$t = \frac{0.65-0.35}{\sqrt{\frac{0.0173^2}{1000}+\frac{0.0173^2}{1000}}}$

$t = \frac{0.30}{\sqrt{0{,}00000029929+0{,}00000029929}}$

$t = \frac{0.30}{0.0007737}$

t=387,76

- El grado de libertad se calculan con n-1 para cada grupo.

  gl = 1000 - 1 = 999

Se hace uso de la tabla t de Student o programas estadísticos para determinar el valor crítico correspondiente al α = 0.05 con 999 gl.

Figura 42

Tabla de Student - Hipótesis específica 3

n	$t_{0,55}$	$t_{0,60}$	$t_{0,70}$	$t_{0,80}$	$t_{0,90}$	$t_{0,95}$	$t_{0,975}$	$t_{0,99}$	$t_{0,995}$
1	0,1584	0,3249	0,7265	1,3764	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,1421	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9646	9,9248
3	0,1366	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,1338	0,2707	0,5686	0,9410	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,1322	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321
6	0,1311	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,1303	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9980	3,4995
8	0,1297	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554
9	0,1293	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498
10	0,1289	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,1286	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058
12	0,1283	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545
13	0,1281	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,1280	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768
15	0,1278	0,2579	0,5357	0,8662	1,3406	1,7531	2,1314	2,6025	2,9467
16	0,1277	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,1276	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,1274	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,1274	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609
20	0,1273	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453
21	0,1272	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,1271	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,1271	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,1270	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,1269	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,1269	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787
27	0,1268	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,1268	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,1268	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564
30	0,1267	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500
40	0,1265	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045
50	0,1263	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778
60	0,1262	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603
80	0,1261	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387
100	0,1260	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259
120	0,1259	0,2539	0,5258	0,8446	1,2886	1,6577	1,9799	2,3578	2,6174
∞	0,126	0,253	0,524	0,842	1,282	1,645	1,960	2,327	2,576

Comparando el valor de t (387,76) con el valor crítico (1.645), t > el valor crítico. Por lo tanto, se rechaza la H0. Se concluye afirmando que la efectividad del modelo predictivo basado en Machine Learning **cuantifica la influencia** de los factores socioeconómicos en la deserción estudiantil en una universidad privada del Perú y permite la creación de estrategias financieras y de apoyo social para abordar eficazmente las necesidades de los estudiantes.

V. DISCUSIÓN DE RESULTADOS

En relación con el primer objetivo específico, el presente estudio coincide con los hallazgos de Franco (2019) y Quintero (2022) en cuanto al uso de técnicas de Machine Learning para la predicción de la deserción estudiantil, pero se diferencia en el enfoque metodológico y los modelos aplicados. Mientras que Franco (2019) empleó árboles de decisión y la metodología CRISP-DM para analizar la deserción en la Universidad Peruana Unión Filial Juliaca, el presente estudio utilizó un modelo de Bosque Aleatorio con 100 estimadores, lo que permitió una clasificación más robusta de los factores de riesgo. A su vez, Quintero (2022) aplicó redes neuronales artificiales (RNA) y Xtreme Gradient Boosting (XGBoost) en la Universidad de Antioquia, logrando un 74.91% de precisión en la clasificación de estudiantes en riesgo de abandono. Aunque su enfoque logró resultados prometedores, el uso de XGBoost y RNA requiere mayor capacidad computacional y tiempos de procesamiento más elevados en comparación con los Bosques Aleatorios, que presentan un balance adecuado entre precisión y eficiencia en la clasificación de datos con múltiples variables. Asimismo, mientras Quintero (2022) enfatizó la importancia de optimizar métricas como recall y f1-score, en este estudio se priorizó la reducción de impureza en la clasificación mediante el cálculo del índice Gini, identificando patrones en estudiantes con problemas de salud, carga familiar y bajo rendimiento académico. A diferencia de Franco (2019), quien destacó la utilidad de bibliotecas como SkLearn y Pandas-Profiling en la optimización de recursos, en esta investigación se dio mayor énfasis a la estructuración y limpieza de datos históricos, asegurando la fiabilidad del conjunto de entrenamiento. En términos de aplicación, tanto Franco (2019) como el presente estudio concuerdan en que la implementación de modelos predictivos puede optimizar la toma de decisiones en las universidades, pero mientras en la Universidad Peruana Unión el enfoque estuvo dirigido a estudiantes de ingeniería de sistemas, en la Universidad Privada San Juan Bautista el análisis se extendió a un espectro más amplio de programas académicos.

En cuanto al segundo objetivo específico, en la presente investigación se determinó que el modelo basado en Random Forest con 100 estimadores alcanzó una precisión de 0.9687 en la predicción de la deserción estudiantil en la Universidad Privada San Juan Bautista, lo que indica una alta efectividad. Este resultado es superior al obtenido por Acosta (2019), quien mediante regresión logística binaria logró una exactitud del 75%, lo que sugiere que los algoritmos de Machine Learning presentan ventajas en la predicción de la deserción respecto a técnicas estadísticas tradicionales. Asimismo, el presente estudio identificó una correlación positiva débil pero significativa (coeficiente de 0.22212) entre el rendimiento académico y la deserción, alineándose con Rodríguez (2023), quien evidenció que factores académicos, junto con aspectos socioeconómicos y psicológicos, inciden en la deserción. Sin embargo, mientras que Rodríguez utilizó Redes Neuronales Artificiales y obtuvo una exactitud del 79% con un F1-Score de 0.62, el modelo Random Forest de esta investigación alcanzó un F1-Score de 0.9642, indicando una mejor capacidad para equilibrar precisión y recall en la clasificación de estudiantes en riesgo. Adicionalmente, se coincidió con Acosta en la importancia de abordar múltiples dimensiones de los estudiantes, aunque la presente investigación propuso estrategias concretas, como mentorías y programas de acompañamiento psicológico, para reducir la deserción, lo que amplía la aplicabilidad de los resultados más allá de la predicción, integrando acciones preventivas efectivas.

En correspondencia al tercer objetivo, en la presente investigación se determinó que el modelo basado en Machine Learning logró predecir la deserción estudiantil con una precisión alta, evidenciando la influencia de factores académicos en el abandono universitario. Este hallazgo guarda relación con el estudio de Torres (2022), donde se confirmó la efectividad del modelo Random Forest para la predicción de la deserción, aunque en su caso, el XGBoost demostró ser una mejor alternativa. En contraste, el estudio de Masabamba (2019) utilizó técnicas de minería de datos, identificando factores adicionales como el estado emocional, la

motivación docente-alumno y el uso de redes sociales, lo que sugiere que, aunque el rendimiento académico es relevante, la deserción también está influenciada por aspectos personales y contextuales. Además, mientras que el presente estudio encontró una correlación positiva pero débil entre el rendimiento académico y la deserción ($r = 0.22212$, $p < 0.0001$), Masabamba (2019) destacó la importancia de factores emocionales y sociales, lo que evidencia la necesidad de complementar modelos predictivos con enfoques multidimensionales. Asimismo, la propuesta de estrategias académicas, como tutorías y flexibilización curricular, coincide con las recomendaciones de Masabamba (2019) y Torres (2022), quienes resaltan la importancia de intervenciones tempranas. Sin embargo, mientras el presente estudio enfatiza el fortalecimiento del vínculo docente-estudiante y la orientación vocacional, Torres (2022) destaca el papel de modelos más avanzados para mejorar la predicción. Por lo tanto, la combinación de estrategias académicas y modelos más robustos podría optimizar la precisión de la predicción y la efectividad de las medidas preventivas.

En referencia al cuarto objetivo específico, en el presente estudio se determinó que los factores socioeconómicos tienen una influencia significativa en la permanencia estudiantil, lo que concuerda con los hallazgos de Tapia (2021) y Pando (2020), quienes también identificaron que variables como la situación financiera, el entorno familiar y el rendimiento académico juegan un papel crucial en la deserción universitaria. Sin embargo, mientras que Tapia (2021) enfatizó que los modelos de ensamble basados en Random Forest ofrecen una mayor precisión en la predicción de estudiantes en riesgo de abandono, en la presente investigación se encontró que, si bien Random Forest fue efectivo, su capacidad de generalización aún requiere optimización. Por otro lado, Pando (2020) aplicó múltiples modelos de minería de datos y concluyó que el modelo C5.0 presentó la mayor precisión (94.2%), mientras que en este estudio se evaluaron distintos algoritmos supervisados, destacando la regresión logística y los árboles de decisión como herramientas complementarias en la identificación de patrones de deserción.

A diferencia de Tapia (2021) y Pando (2020), que se centraron en el desarrollo de modelos predictivos exclusivamente, este estudio no solo propuso un modelo, sino que también planteó estrategias concretas para mitigar la deserción, como la implementación de programas de becas, la creación de convenios con empresas y la reducción de costos indirectos.

Diversos estudios han abordado el impacto de los modelos predictivos basados en Machine Learning en la reducción de la deserción estudiantil en universidades, destacando su capacidad para identificar patrones de abandono y permitir la implementación de estrategias preventivas. En el contexto de universidades privadas del Perú, el presente estudio de la Universidad Privada San Juan Bautista coincide con los hallazgos de Márquez et al. (2021), quienes encontraron que algoritmos de clasificación como Random Forest y Support Vector Machines tienen una alta precisión en la predicción de estudiantes en riesgo, facilitando la intervención temprana. Al igual que García & López (2020), que destacaron la importancia de la combinación de técnicas de procesamiento de datos y modelos supervisados, este estudio se centra en el uso de un modelo de Bosque Aleatorio con 100 estimadores, lo que permite una evaluación más precisa de variables como el rendimiento académico, la carga familiar y los problemas de salud, factores que también han sido identificados como cruciales en otros estudios. En línea con los planteamientos de Fernández et al. (2019), que resaltaron el uso de redes neuronales para optimizar la toma de decisiones, este trabajo utiliza un enfoque que busca identificar factores predictivos relevantes en la deserción, pero con un énfasis particular en la estructura y preparación de los datos. Además, investigaciones recientes, como las de Rodríguez & Silva (2022), subrayan la importancia de la interpretación de los resultados mediante Explainable AI (XAI). Si bien este estudio no incorpora XAI directamente, la metodología empleada a través del análisis de la impureza de Gini contribuye a una mayor comprensión de las variables que influyen en el riesgo de deserción. En conclusión, el uso de Machine Learning en la predicción de la deserción estudiantil no solo mejora la precisión de la

identificación de estudiantes en riesgo, sino que también permite desarrollar intervenciones personalizadas y basadas en evidencia, lo cual es crucial para mejorar la retención y el éxito académico en universidades privadas del Perú.

Tabla 10

Comparativa de resultados según autores

Autor (Año)	Objetivo	Población / Muestra	Metodología / Modelo Predictivo	Variables Analizadas	Resultados Relevantes
Masabamba (2019)	Determinar factores de deserción mediante minería de datos	1457 estudiantes (Fac. de Ingeniería, UTC)	Minería de datos (KDD). Algoritmos: J48, Random Forest, SMO	Conducta en clase, bullying, motivación, redes sociales, emocionalidad, etc.	Tasa de predicción: **92%**. Técnicas de minería efectivas para análisis de deserción
Torres (2022)	Predecir deserción en Psicología, Univ. El Bosque	Estudiantes de Psicología	Clasificación supervisada: Random Forest y XGBoost	Académicas, demográficas, socioeconómicas, personalidad	XGBoost superó a Random Forest. Efectivo en predicción precisa
Caselli (2021)	Soporte al seguimiento académico mediante redes neuronales	Datos de rendimiento académico (sin muestra numérica específica)	Machine Learning / Deep Learning. Red neuronal de 2-7 capas	Semestres cursados (variable más influyente), rendimiento académico	Precisión: **98.97%**. Alta capacidad predictiva para intervención temprana
Gutiérrez (2022)	Determinar deserción en primer año (UNASAM)	6440 estudiantes (2010-I al 2019-I)	Gradient Boosting. Diseño longitudinal, explicativo, preexperimental	Académicas, socioeconómicas, rendimiento	Precisión: **94%**, F1: 90%, R^2 entrenamiento: 75.12%, prueba: 70.09%
Modelo Univ. San Juan Bautista	Determinar impacto de factores personales, académicos y socioeconómicos en la deserción estudiantil	104 estudiantes de la Carrera de Ingeniería de Sistemas	Machine Learning. Análisis de hipótesis (t-student) Random Forest	Personales, académicos, socioeconómicos	Métricas evaluadas del modelo: Precisión: 0.9687 F: 0.9642 Confiabilidad: 0.9859 Efectividad: 0.9797

VI. CONCLUSIONES

El estudio demostró que el uso de Machine Learning es una herramienta efectiva para predecir la deserción estudiantil en universidades privadas del Perú, permitiendo identificar con precisión los factores personales, académicos y socioeconómicos que inciden en este problema. La alta precisión del modelo validó su utilidad para la detección temprana de estudiantes en riesgo, facilitando la implementación de estrategias como tutorías, flexibilización académica y apoyo financiero. Estos hallazgos resaltan la importancia de aplicar modelos predictivos en la gestión universitaria para mejorar la retención estudiantil en instituciones privadas peruanas.

El modelo predictivo basado en Machine Learning permitió identificar de manera efectiva los factores personales, académicos y socioeconómicos que influyen en la deserción estudiantil en las universidades privadas del Perú. A través de la implementación del algoritmo Random Forest con 100 estimadores y una profundidad de 3, se logró clasificar eficazmente a los estudiantes en riesgo de abandono, optimizando el criterio Gini en 19 instancias analizadas. Estos resultados evidencian la viabilidad del modelo como herramienta de detección temprana para la prevención de la deserción universitaria, aportando un valor significativo en la gestión proactiva de la retención estudiantil.

La evaluación del modelo predictivo reveló un alto nivel de precisión en la identificación de factores personales asociados a la deserción estudiantil, alcanzando un accuracy del 96.87%, un recall de 96.72% y un F1 Score de 96.42%. A pesar de la correlación débil entre el rendimiento académico y la deserción ($r=0.22212$), la significancia estadística ($p=2.33e-08$) confirma su impacto en la permanencia estudiantil. Estas métricas validan la confiabilidad del modelo para anticipar la deserción en función de las características personales de los estudiantes, permitiendo una intervención más específica y precisa.

El modelo predictivo mostró ser una herramienta confiable en la evaluación de los factores académicos influyentes en la deserción, evidenciando una correlación positiva débil pero significativa (r=0.22212, p=2.33e-08) entre el rendimiento académico y la deserción. No obstante, se identificó un posible sobreajuste en el modelo, lo que resalta la necesidad de mejorar su capacidad de generalización. A partir de los hallazgos, se recomienda la implementación de estrategias académicas como tutorías personalizadas, flexibilización de carga horaria y metodologías de enseñanza inclusivas para fortalecer la retención estudiantil, basándose en los resultados obtenidos del modelo predictivo.

El modelo permitió cuantificar el impacto de los factores socioeconómicos en la deserción, confirmando la necesidad de programas de apoyo financiero como una estrategia clave para mejorar la permanencia de los estudiantes en universidades privadas del Perú. Se proponen medidas como becas, pasantías remuneradas y subvenciones estudiantiles, respaldadas por el análisis predictivo, para mitigar el impacto económico en la continuidad académica. Estas iniciativas, basadas en evidencia, ofrecen soluciones efectivas para reducir la deserción y mejorar la retención estudiantil en el sector universitario privado del Perú, contribuyendo así a la mejora de la calidad educativa y el éxito académico.

VII. RECOMENDACIONES

Dado el alto nivel de precisión del modelo, se sugiere su integración en el sistema de gestión académica de la universidad, asegurando su actualización periódica con nuevos datos para mejorar su capacidad de predicción y adaptación a los cambios en el perfil estudiantil.

Se recomienda que la universidad implemente programas de acompañamiento estudiantil basados en los factores personales identificados como determinantes en la deserción, utilizando el modelo predictivo para generar alertas tempranas y asignar recursos adecuados a cada caso.

Para mejorar la retención estudiantil, se recomienda fortalecer estrategias académicas como tutorías personalizadas, planes de nivelación y metodologías de enseñanza más flexibles, priorizando a los estudiantes en riesgo identificados por el modelo predictivo.

Se aconseja diseñar e implementar programas de apoyo financiero basados en los resultados del modelo, priorizando becas, pasantías y subvenciones para estudiantes en situación de vulnerabilidad económica, con el fin de reducir la deserción por razones financieras.

Para optimizar la capacidad de generalización del modelo y reducir el riesgo de sobreajuste, se recomienda explorar el uso de técnicas avanzadas de Machine Learning, como ajuste de hiperparámetros y validación cruzada, así como la incorporación de nuevas variables que puedan mejorar la precisión en la detección de estudiantes en riesgo de deserción.

VIII. REFERENCIAS

Acosta, M. (2019). Inteligencia artificial: la cibernética del ser vivo y de la máquina. *Naturaleza y Libertad. Revista de Estudios Interdisciplinares*, *12*, 13–30. https://doi.org/10.24310/NATyLIB.2019.v0i12.6262

Aflalo, E., & Gabay, E. (2011). An Information System for Coping with Student Dropout. *International Journal of Information and Communication Technology Education*, *7*(3), 62–73. https://doi.org/10.4018/jicte.2011070106

Aguilar-Barojas, S. (2005). Fórmulas para el cálculo de la muestra en investigaciones de salud. *Salud En Tabasco*, *11*(1–2), 333–338.

Aldowah, H., Al-Samarraie, H., Alzahrani, A. I., & Alalwan, N. (2020). Factors affecting student dropout in MOOCs: a cause and effect decision-making model. *Journal of Computing in Higher Education*, *32*(2), 429–454. https://doi.org/10.1007/s12528-019-09241-y

Algren, M., Fisher, W., & Landis, A. E. (2021). Machine learning in life cycle assessment. In *Data Science Applied to Sustainability Analysis* (pp. 167–190). Elsevier. https://doi.org/10.1016/B978-0-12-817976-5.00009-7

Aljaber, S., & Almushaili, T. (2022). Artificial Intelligence. *International Journal of Engineering Research and Applications*, *12*(12), 52–57.

Allen, H. K., Lilly, F., Green, K. M., Zanjani, F., Vincent, K. B., & Arria, A. M. (2022). Graduate Student Burnout: Substance Use, Mental Health, and the Moderating Role of Advisor Satisfaction. *International Journal of Mental Health and Addiction*, *20*(2), 1130–1146. https://doi.org/10.1007/s11469-020-00431-9

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science.

In *Supervised and Unsupervised Learning for Data Science. Unsupervised and Semi-Supervised Learning* (pp. 3–21). Springer. https://doi.org/10.1007/978-3-030-22475-2_1

Améstica-Rivas, L., King-Domínguez, A., Sanhueza Gutiérrez, D. A., & Ramírez González, V. (2020). Efectos económicos de la deserción en la gestión universitaria: el caso de una universidad pública chilena. *Hallazgos*, *18*(35), 209–231. https://doi.org/10.15332/2422409X.5772

Arias, F. (2012). *El Proyecto de Investigación. Introducción a la metodología científica* (Sexta edic).

Baalmann, T., Brömmelhaus, A., Hülsemann, J., Feldhaus, M., & Speck, K. (2024). The Impact of Parents, Intimate Relationships, and Friends on Students' Dropout Intentions. *Journal of College Student Retention: Research, Theory & Practice*, *26*(3), 923–947. https://doi.org/10.1177/15210251221133374

Barriga, O., & Henríquez, G. (2011). La relación Unidad de Análisis-Unidad de ObservaciónUnidad de Información: Una ampliación de la noción de la Matriz de Datos propuesta por Samaja. *Revista Latinoamericana de Metodología de La Investigación Social*, *1*, 61–69.

Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020). Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *Journal of Personalized Medicine*, *10*(2), 21. https://doi.org/10.3390/jpm10020021

Bäulke, L., Grunschel, C., & Dresel, M. (2022). Student dropout at university: a phase-orientated view on quitting studies and changing majors. *European Journal of Psychology of Education*, *37*(3), 853–876. https://doi.org/10.1007/s10212-021-00557-x

Bell, J. (2022). What Is Machine Learning? In *Machine Learning and the City* (pp. 207–216). Wiley. https://doi.org/10.1002/9781119815075.ch18

Belyadi, H., & Haghighat, A. (2021). Machine learning workflows and types. In *Machine Learning Guide for Oil and Gas Using Python* (pp. 97–123). Elsevier. https://doi.org/10.1016/B978-0-12-821929-4.00001-9

Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429027192

Bowen, J. R., Dodier, N., Duyvendak, J. W., & Hardon, A. (Eds.). (2020). *Pragmatic Inquiry*. Routledge. https://doi.org/10.4324/9781003034124

Brauer, S. Z., & Sirin, S. R. (2024). School dropout. In *Encyclopedia of Adolescence* (pp. 523–535). Elsevier. https://doi.org/10.1016/B978-0-323-96023-6.00023-3

Carr, J. R. (2022). Why Ideal Epistemology? *Mind*, *131*(524), 1131–1162. https://doi.org/10.1093/mind/fzab023

Cedeño, I., & Cedeño-Valarezo, L. (2023). Active monitoring systems to prevent attacks to network services at dr. Napoleón Dávila Córdova hospital. *Revista Científica Multidisciplinaria Arbitrada Yachasun*, *7*(12), 70–86.

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. *Translational Vision Science & Technology*, *9*, 14. https://doi.org/10.1167/tvst.9.2.14

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Darío, P., & Alejandra, M. (2014). La unidad de análisis en la problemática enseñanza aprendizaje. Una mirada sistémica. *Informe Científico Técnico UNPA*, *6*(3), 101–117.

Dávila Morán, R. C., Agüero Corzo, E. C., Portillo Rios, H., & Quimbita Chiluisa, O. R. (2022). Deserción universitaria de los estudiantes de una universidad peruana. *Braz Dent J.*, *33*(1).

de Oliveira, C. F., Sobral, S. R., Ferreira, M. J., & Moreira, F. (2021). How Does Learning Analytics Contribute to Prevent Students' Dropout in Higher Education: A Systematic Literature Review. *Big Data and Cognitive Computing*, *5*(4), 64. https://doi.org/10.3390/bdcc5040064

Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student Dropout Prediction. In *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science* (pp. 129–140). Springer. https://doi.org/10.1007/978-3-030-52237-7_11

Dhal, P., & Azad, C. (2022). A comprehensive survey on feature selection in the various fields of machine learning. *Applied Intelligence*, *52*(4), 4543–4581. https://doi.org/10.1007/s10489-021-02550-9

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, *15*(2), e0228987. https://doi.org/10.1371/journal.pone.0228987

Fernández Bedoya, V. H. (2020). Tipos de justificación en la investigación científica. *Espíritu Emprendedor TES*, *4*(3), 65–76. https://doi.org/10.33970/eetes.v4.n3.2020.207

Fernandez-Felix, B. M., Barca, L. V., Garcia-Esquinas, E., Correa-Pérez, A., Fernández-Hidalgo, N., Muriel, A., Lopez-Alcalde, J., Álvarez-Diaz, N., Pijoan, J. I., Ribera, A., Elorza, E. N., Muñoz, P., Fariñas, M. del C., Goenaga, M. Á., & Zamora, J. (2021). Prognostic models for mortality after cardiac surgery in patients with infective endocarditis: a systematic review and aggregation of prediction models. *Clinical Microbiology and Infection*, *27*(10), 1422–1430. https://doi.org/10.1016/j.cmi.2021.05.051

Flasiński, M. (2016). History of Artificial Intelligence. In *Introduction to Artificial Intelligence* (pp. 3–13). Springer International Publishing. https://doi.org/10.1007/978-3-319-40022-8_1

Frunza, M.-C. (2016). Support Vector Machines. In *Solving Modern Crime in Financial Markets* (pp. 205–215). Elsevier. https://doi.org/10.1016/B978-0-12-804494-0.00015-2

Grimaldi, G., & Ehrler, B. (2023). AI *et al.*: Machines Are About to Change Scientific Publishing Forever. *ACS Energy Letters*, *8*(1), 878–880. https://doi.org/10.1021/acsenergylett.2c02828

Gutiérrez A., D., Vélez Díaz, J. F., & López, J. M. (2021). Indicadores de deserción universitaria y factores asociados. *Revista EducaT: Educación Virtual, Innovación y Tecnologías*, *2*.

He, Y.-H. (2021). Machine-Learning Mathematical Structures. *ArXiv:2101*, *2*, 1–32.

Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S., & Klambauer, G. (2019). Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *Journal of Chemical Information and Modeling*, *59*(3), 1163–1171. https://doi.org/10.1021/acs.jcim.8b00670

Hosseini, M., Powell, M., Collins, J., Callahan-Flintoft, C., Jones, W., Bowman, H., & Wyble, B. (2020). I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neuroscience & Biobehavioral Reviews*, *119*, 456–467. https://doi.org/10.1016/j.neubiorev.2020.09.036

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, *27*(4), 621–633. https://doi.org/10.1093/jamia/ocz228

Huet, P. (2023, April 13). *Qué son las redes neuronales y sus aplicaciones*. OpenWebinars. https://openwebinars.net/blog/que-son-las-redes-neuronales-y-sus-aplicaciones/

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence? *Discover Artificial Intelligence*, *2*(1), 4. https://doi.org/10.1007/s44163-022-00022-8

Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, *52*(1), 273–292. https://doi.org/10.1007/s10462-018-09677-1

Karvelis, P., Paulus, M. P., & Diaconescu, A. O. (2023). Individual differences in computational psychiatry: A review of current challenges. *Neuroscience & Biobehavioral Reviews*, *148*, 105137. https://doi.org/10.1016/j.neubiorev.2023.105137

Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*, *18*(1/2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

Krüger, J. G. C., Britto, A. de S., & Barddal, J. P. (2023). An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, *233*, 120933. https://doi.org/10.1016/j.eswa.2023.120933

Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics*, *13*(15), 2582. https://doi.org/10.3390/diagnostics13152582

Lamba, M., & Madhusudhan, M. (2022). Predictive Modeling. In *Text Mining for Information Professionals* (pp. 213–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-85085-2_8

Lodhi, S. S., Kumar, N., & Pandey, P. K. (2023). Autonomous vehicular overtaking maneuver: A survey and taxonomy. *Vehicular Communications*, *42*, 100623. https://doi.org/10.1016/j.vehcom.2023.100623

Lorenzo-Quiles, O., Galdón-López, S., & Lendínez-Turón, A. (2023). Factors contributing to university dropout: a review. *Frontiers in Education*, *8*, 1159864. https://doi.org/10.3389/feduc.2023.1159864

Loza, R. M., Mamani Condori, J. L., Mariaca Mamani, J. S., & Yanqui Santos, F. E. (2021). Paradigma sociocrítico en investigación. *PSIQUEMAG/ Revista Científica Digital de Psicología*, *9*(2), 30–39. https://doi.org/10.18050/psiquemag.v9i2.2656

Lykourentzou, I., Giannoukos, I., Nikolopoulos, V., Mpardis, G., & Loumos, V. (2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, *53*(3), 950–965. https://doi.org/10.1016/j.compedu.2009.05.010

Mamun, A. Al, Sohel, Md., Mohammad, N., Haque Sunny, Md. S., Dipta, D. R., & Hossain, E. (2020). A Comprehensive Review of the Load Forecasting Techniques Using Single and Hybrid Predictive Models. *IEEE Access*, *8*, 134911–134939. https://doi.org/10.1109/ACCESS.2020.3010702

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107–124. https://doi.org/10.1111/exsy.12135

McDermott, E. R., Donlan, A. E., & Zaff, J. F. (2019). Why do students drop out? Turning points and long-term experiences. *The Journal of Educational Research*, *112*(2), 270–282. https://doi.org/10.1080/00220671.2018.1517296

Mejía-Rivas, J. (2022). Los paradigmas en la investigación científica. *Revista Ciencia Agraria*, *1*(3), 7–14. https://doi.org/10.35622/j.rca.2022.03.001

Metropolitana, M., & Lima, D. E. (s.f.). *GOBIERNOS LOCALES*.

Miranda, V. J., & Alarcón, H. H. (2021). Efectos de los factores de riesgo sobre la interrupción de los estudios en jóvenes universitarios durante la covid-19. *Desde El Sur*, *13*(2), e0021. https://doi.org/10.21142/DES-1302-2021-0021

Mishra, N., & Mishra, S. N. M. (2023). A Novel Intrusion Detection Techniques of the Computer Networks Using Machine Learning. *International Journal of Intelligent Systems and Applications in Engineering*, *11*(5), 247–260.

Ñaupas, H., Valdivia, M., Palacios, J., & Romero, H. (2018). *Metodología de la investigación: Cuantitativa - Cualitativa y Redacción de la Tesis* (Ediciones de la U, Ed.; 5a edición).

Núñez, E. (2018). Acoso sexual: una realidad invisible en las universidades en Paraguay. *Revista Científica Estudios e Investigaciones*. https://doi.org/10.26885/rcei.foro.2017.42

OECD. (2024). *Education at a Glance 2024*. OECD. https://doi.org/10.1787/c00cad36-en

Ogresta, J., Rezo, I., Kožljan, P., Paré, M.-H., & Ajduković, M. (2021). Why do we drop out? Typology of dropping out of High School. *Youth & Society*, *53*(6), 934–954. https://doi.org/10.1177/0044118X20918435

Olaya, D., Vásquez, J., Maldonado, S., Miranda, J., & Verbeke, W. (2020). Uplift Modeling for preventing student dropout in higher education. *Decision Support Systems*, *134*, 113320. https://doi.org/10.1016/j.dss.2020.113320

Organización Mundial de la Salud (OMS). (2022). Informe mundial sobre salud mental: transformar la salud mental para todos. Panorama general. *Oms*.

Otzen, T., & Manterola, C. (2017). Técnicas de Muestreo sobre una Población a Estudio. *International Journal of Morphology*, *35*(1), 227–232.

Parviainen, M., Aunola, K., Torppa, M., Poikkeus, A.-M., & Vasalampi, K. (2020). Symptoms of psychological ill-being and school dropout intentions among upper secondary

education students: A person-centered approach. *Learning and Individual Differences*, *80*, 101853. https://doi.org/10.1016/j.lindif.2020.101853

Peng, P., Yang, W. F., Liu, Y., Chen, S., Wang, Y., Yang, Q., Wang, X., Li, M., Wang, Y., Hao, Y., He, L., Wang, Q., Zhang, J., Ma, Y., He, H., Zhou, Y., Long, J., Qi, C., Tang, Y.-Y., … Liu, T. (2022). High prevalence and risk factors of dropout intention among Chinese medical postgraduates. *Medical Education Online*, *27*(1). https://doi.org/10.1080/10872981.2022.2058866

Perchinunno, P., Bilancia, M., & Vitale, D. (2021). A Statistical Analysis of Factors Affecting Higher Education Dropouts. *Social Indicators Research*, *156*(2–3), 341–362. https://doi.org/10.1007/s11205-019-02249-y

Pierrakeas, C., Koutsonikos, G., Lipitakis, A.-D., Kotsiantis, S., Xenos, M., & Gravvanis, G. A. (2020). The Variability of the Reasons for Student Dropout in Distance Learning and the Prediction of Dropout-Prone Students. In *Machine Learning Paradigms. Intelligent Systems Reference Library* (Vol. 58, pp. 91–111). Springer. https://doi.org/10.1007/978-3-030-13743-4_6

Quintero Solis, S. I. (2020). El Acoso y hostigamiento sexual escolar, necesidad de su regulación en las Universidades. *Revista de Estudios de Género, La Ventana*, *6*(51). https://doi.org/10.32870/lv.v6i51.7083

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Rowland, R. C. (2018). Purpose, Argument Fields, and Theoretical Justification. *Argumentation*, *22*(2), 235–250. https://doi.org/10.1007/s10503-007-9062-y

Semanjski, I. C. (2023). Data analytics. In *Smart Urban Mobility* (pp. 121–170). Elsevier. https://doi.org/10.1016/B978-0-12-820717-8.00008-7

Seraj, A., Mohammadi-Khanaposhtani, M., Daneshfar, R., Naseri, M., Esmaeili, M., Baghban, A., Habibzadeh, S., & Eslamian, S. (2023). Cross-validation. In *Handbook of Hydroinformatics* (pp. 89–105). Elsevier. https://doi.org/10.1016/B978-0-12-821285-1.00021-X

Shobha, G., & Rangaswamy, S. (2018). Machine Learning. In *Handbook of Statistics* (Vol. 38, pp. 197–228). Elsevier. https://doi.org/10.1016/bs.host.2018.07.004

Sifuentes, O. (2018). Modelos predictivos de la deserción estudiantil en una universidad privada peruana. *Industrial Data*, *21*(2), 47–52. https://doi.org/10.15381/idata.v21i2.15602

Taborda, G. E., Castaño, B. S., Durán, J. M., Conto, R., & Reyes, E. R. (2024). Propuesta de modelo de analítica para flujo de caja en mipymes en Colombia. *Revista CEA*, *10*(22), e2607. https://doi.org/10.22430/24223182.2607

Tamin, S. K. (2013). Relevance of mental health issues in university student dropouts. *Occupational Medicine*, *63*(6), 410–414. https://doi.org/10.1093/occmed/kqt071

Tejedor, F. J., & Muñoz-Repiso, A. G. (2007). Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES. *Revista de Educación*, *342*, 443–474.

Urbina-Nájera, A. B., Camino-Hampshire, J. C., & Cruz Barbosa, R. (2020). Deserción escolar universitaria: Patrones para prevenirla aplicando minería de datos educativa. *Revista Electrónica de Investigación y Evaluación Educativa*, *26*(1). https://doi.org/10.7203/relieve.26.1.16061

Valero, J. E., Navarro, Á. F., Larios, A. C., & Julca, J. D. (2022). Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción. *Revista de Ciencias Sociales*, *28*(3), 362–375. https://doi.org/10.31876/rcs.v28i3.38480

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Viale, H. E. (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista Digital de Investigación En Docencia Universitaria*, *8*(1), 59–76. https://doi.org/10.19083/ridu.8.366

Viale Tudela, H. E. (2014). UNA APROXIMACIÓN TEÓRICA A LA DESERCIÓN ESTUDIANTIL UNIVERSITARIA. *Revista Digital de Investigación En Docencia Universitaria*. https://doi.org/10.19083/ridu.8.366

Vieira, B. H., Pamplona, G. S. P., Fachinello, K., Silva, A. K., Foss, M. P., & Salmon, C. E. G. (2022). On the prediction of human intelligence from neuroimaging: A systematic review of methods and reporting. *Intelligence*, *93*, 101654. https://doi.org/10.1016/j.intell.2022.101654

Viera Castillo, D. O., Flores Loredo, M. A., & Pachari-Vera, E. (2020). FACTORES DE DESERCIÓN ESTUDIANTIL: UN ESTUDIO EXPLORATORIO DESDE PERÚ. *Interciencia*, *45*.

Viloria, A., Padilla, J. G., Vargas-Mercado, C., Hernández-Palma, H., Llinas, N. O., & David, M. A. (2019). Integration of Data Technology for Analyzing University Dropout. *Procedia Computer Science*, *155*, 569–574. https://doi.org/10.1016/j.procs.2019.08.079

Wang, A., An, N., Xia, Y., Li, L., & Chen, G. (2014). A Logistic Regression and Artificial Neural Network-Based Approach for Chronic Disease Prediction: A Case Study of Hypertension. *2014 IEEE International Conference on Internet of Things(IThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom)*, 45–52. https://doi.org/10.1109/iThings.2014.16

Yan, Y. (2022). Machine Learning Fundamentals. In *Machine Learning in Chemical Safety and Health* (pp. 19–46). Wiley. https://doi.org/10.1002/9781119817512.ch2

Yucra Quispe, T., & Bernedo Villalta, L. Z. (2020). Epistemología e Investigación Cuantitativa. *IGOBERNANZA*, *3*(12), 107–120. https://doi.org/10.47865/igob.vol3.2020.88

Zhang, Y., & Wang, Y. (2023). Machine learning applications for multi-source data of edible crops: A review of current trends and future prospects. *Food Chemistry: X*, *19*, 100860. https://doi.org/10.1016/j.fochx.2023.100860

IX. ANEXOS

Anexo A: Matriz de Consistencia

Problema Principal	Objetivo General	Hipótesis General	Variables	Dimensiones	Indicador(es)	Investigación
¿Cómo puede un modelo predictivo basado en Machine Learning contribuir a la identificación de estrategias efectivas para reducir la tasa de deserción estudiantil en las Universidades Privadas del Perú, específicamente en la Universidad Privada San Juan Bautista?	Desarrollar un modelo predictivo basado en Machine Learning que contribuya en la determinación de estrategias efectivas para la reducción de la deserción estudiantil en las Universidades Privadas del Perú: Caso Universidad Privada San Juan Bautista.	El desarrollo de un modelo predictivo basado en Machine Learning tiene un impacto significativo en la reducción de la tasa de deserción estudiantil en la Universidad Privada San Juan Bautista, al identificar estrategias efectivas de retención.	Variable Independiente: Modelo Predictivo Basado en Machine Learning Variable Dependiente: Deserción Estudiantil	Rendimiento del modelo Tiempo de operación Eficiencia del proceso.	Precisión de la predicción Validación del modelo Tasa de deserción Tasa de retención Grado de influencia en las variables de la predicción en la determinación de la deserción. Factores que intervienen para la determinación de la deserción.	Tipo: Aplicada Nivel: Explicativo Universo: Todos los Procesos para la reducción de la deserción estudiantil en las Instituciones Educativas a nivel mundial Muestra: Proceso de reducción de la deserción estudiantil de las universidades

				Satisfacción del proceso	Número de estrategias de intervención para contrarrestar la deserción. Porcentaje de confiabilidad	privadas en el Perú: Caso Universidad Privada San Juan Bautista. N = 104 Tipo de Muestreo: Aleatorio probabilístico.

Anexo B: Instrumento para la recolección de datos

CUESTIONARIO SOBRE FACTORES QUE INFLUYEN EN LA DESERCIÓN ESTUDIANTIL

Introducción:

El presente cuestionario tiene como objetivo recopilar información clave sobre los factores que contribuyen a la deserción estudiantil en las universidades privadas del Perú, específicamente en la Universidad Privada San Juan Bautista. Sus respuestas serán fundamentales para el desarrollo de un modelo predictivo basado en machine learning, orientado a reducir la deserción estudiantil y mejorar la experiencia educativa.

Instrucciones:

1. Lea cada pregunta detenidamente.
2. Marque con una X la opción que considere correcta: Sí o No.
3. Sea sincero en sus respuestas, ya que la información será tratada de manera confidencial y únicamente con fines investigativos.
4. Este cuestionario está estructurado en tres factores principales: Personal, Académico y Socioeconómico.

Factor	Pregunta	Respuesta (Sí/No)
Factor Personal	¿Has tenido problemas de salud que hayan afectado tu desempeño en tus actividades, contribuyendo a la posibilidad de abandonar tus estudios?	[ ] Sí / [ ] No
	¿Tienes responsabilidades familiares que dificulten tu continuidad académica o laboral?	[ ] Sí / [ ] No
	¿Consideras que tu elección profesional no refleja realmente tu vocación, afectando tu motivación para continuar?	[ ] Sí / [ ] No
Factor Académico	¿Has experimentado bajo rendimiento académico que te haga considerar dejar tus estudios?	[ ] Sí / [ ] No
	¿Sientes que no recibes el apoyo académico necesario para continuar con tus estudios?	[ ] Sí / [ ] No
	¿Consideras que los horarios de clase no son compatibles con tus necesidades, afectando tu continuidad?	[ ] Sí / [ ] No

	¿Te ha afectado el límite de inasistencias, haciéndote considerar la posibilidad de abandonar tus estudios?	[ ] Sí / [ ] No
Factor Socioeconómico	¿Has enfrentado dificultades financieras que puedan llevarte a dejar tus estudios?	[ ] Sí / [ ] No
	¿Tu trabajo a tiempo completo interfiere con tus estudios, haciendo difícil continuar?	[ ] Sí / [ ] No
	¿Sientes que no cuentas con el apoyo familiar necesario para continuar con tus estudios?	[ ] Sí / [ ] No

Anexo C: Instrumento con llenado para la base de datos recolectadas por cada estudiante encuestado.

CUESTIONARIO SOBRE FACTORES QUE INFLUYEN EN LA DESERCIÓN ESTUDIANTIL

Introducción:

El presente cuestionario tiene como objetivo recopilar información clave sobre los factores que contribuyen a la deserción estudiantil en las universidades privadas del Perú, específicamente en la Universidad Privada San Juan Bautista. Sus respuestas serán fundamentales para el desarrollo de un modelo predictivo basado en machine learning, orientado a reducir la deserción estudiantil y mejorar la experiencia educativa.

Instrucciones:

5. Lea cada pregunta detenidamente.
6. Marque con una X la opción que considere correcta: Sí o No.
7. Sea sincero en sus respuestas, ya que la información será tratada de manera confidencial y únicamente con fines investigativos.
8. Este cuestionario está estructurado en tres factores principales: Personal, Académico y Socioeconómico.

Factor	Pregunta	Respuesta (Sí/No)
Factor Personal	¿Has tenido problemas de salud que hayan afectado tu desempeño en tus actividades, contribuyendo a la posibilidad de abandonar tus estudios?	[ ] Sí / [ ] No
	¿Tienes responsabilidades familiares que dificulten tu continuidad académica o laboral?	[ ] Sí / [ ] No
	¿Consideras que tu elección profesional no refleja realmente tu vocación, afectando tu motivación para continuar?	[ ] Sí / [ ] No
Factor Académico	¿Has experimentado bajo rendimiento académico que te haga considerar dejar tus estudios?	[ ] Sí / [ ] No
	¿Sientes que no recibes el apoyo académico necesario para continuar con tus estudios?	[ ] Sí / [ ] No
	¿Consideras que los horarios de clase no son compatibles con tus necesidades, afectando tu continuidad?	[ ] Sí / [ ] No

	¿Te ha afectado el límite de inasistencias, haciéndote considerar la posibilidad de abandonar tus estudios?	[ ] Sí / [ ] No
Factor Socioeconómico	¿Has enfrentado dificultades financieras que puedan llevarte a dejar tus estudios?	[ ] Sí / [ ] No
	¿Tu trabajo a tiempo completo interfiere con tus estudios, haciendo difícil continuar?	[ ] Sí / [ ] No
	¿Sientes que no cuentas con el apoyo familiar necesario para continuar con tus estudios?	[ ] Sí / [ ] No

Tabla 11

Respuestas de la dimensión 1 factor personal en 20 estudiantes desde el periodo 2018- 2023

Estudiante	Problemas Salud	Carga Familiar	Vocacion Profesional
Estudiante 1	0	0	1
Estudiante 2	0	0	1
Estudiante 3	1	0	1
Estudiante 4	1	0	1
Estudiante 5	1	0	1
Estudiante 6	1	0	0
Estudiante 7	1	1	1
Estudiante 8	1	1	1
Estudiante 9	0	0	1
Estudiante 10	0	0	0
Estudiante 11	0	0	1
Estudiante 12	0	0	1
Estudiante 13	0	0	1
Estudiante 14	0	1	1
Estudiante 15	1	1	0
Estudiante 16	1	1	0
Estudiante 17	0	0	1
Estudiante 18	1	0	0
Estudiante 19	0	0	1
Estudiante 20	0	0	1

Interpretación:

1. Problemas de Salud

- Se mantuvo **estable (2 casos)** entre 2018 y 2021.
- En **2022 y 2023**, disminuyó a **1 caso**.

- Puede interpretarse como una posible mejora en las condiciones de salud o una menor cantidad de estudiantes con este factor en los últimos años.

2. Carga Familiar

- Se mantuvo **bajo en general** (0–1 caso por año).
- Ausente solo en **2020**, donde ningún estudiante reportó carga familiar.
- Esto indica que **la mayoría de los estudiantes no tienen responsabilidades familiares**, lo cual puede ser positivo para su desempeño académico.

3. Vocación Profesional

- Consistentemente **alta**, con **2 a 4 estudiantes** por año indicando tener vocación profesional.
- **2023** destaca como el año con mayor número (4), lo que podría relacionarse con un mayor enfoque o motivación en la elección de carrera.
- Esto es una señal positiva, ya que la vocación suele estar asociada con mejor rendimiento y persistencia.

Figura 43

Frecuencia de factores personales por año

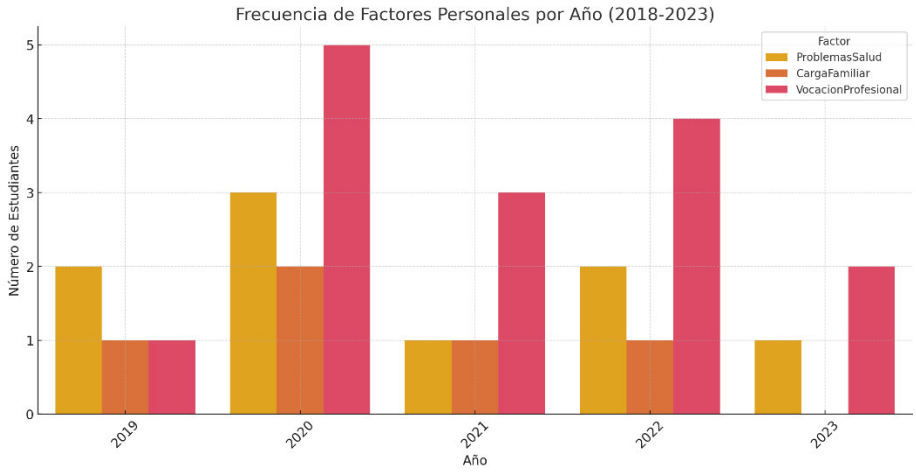


Tabla 12

Respuestas de la dimensión factor académico

Bajo Rendimiento Académico	Apoyo Académico	Horarios Clase	Limite Inasistencia
0	0	0	0
0	1	1	0
0	1	0	0
0	1	0	0
0	1	1	1
0	0	1	0
0	1	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	0	1	0
0	1	1	0
1	1	1	0
0	1	1	0
0	1	0	0
0	0	1	0
0	1	0	1
0	0	1	0
0	0	1	0
0	1	1	0

Interpretación:

Bajo Rendimiento Académico (5%)

- Solo 1 estudiante reporta bajo rendimiento, lo cual **es positivo** en general para el grupo.

Apoyo Académico (60%)

- La mayoría requiere apoyo académico, lo que podría indicar **debilidades en la comprensión de contenidos** o **dificultades generales** a pesar de no tener bajo rendimiento declarado.

Horarios de Clase (50%)

- La mitad de los estudiantes identifican los horarios como un factor, lo que sugiere que **puede haber conflicto con otras responsabilidades** o falta de flexibilidad horaria.

Límite de Inasistencia (10%)

- Muy pocos han llegado al límite de inasistencia. Esto sugiere **buena asistencia general**, lo cual es favorable.

Figura 44

Frecuencia de factores académicos

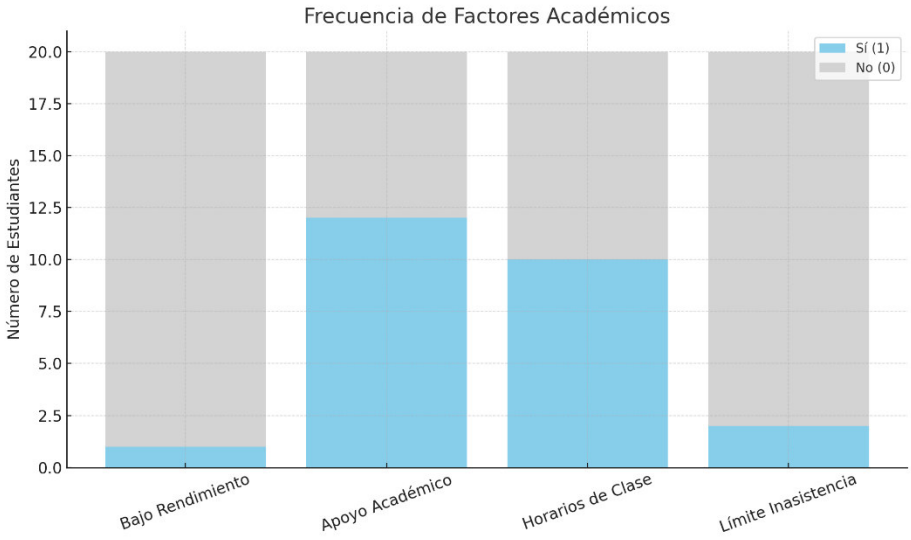


Tabla 13

Respuestas de la dimensión factor socioeconómico

Dificultad Financiera	Trabajo Tiempo Completo	Apoyo Familiar
0	0	1
1	1	1
0	1	0
0	1	1
1	1	0
1	0	1
0	1	1
1	0	1
0	1	1
1	1	0
0	1	0
0	0	1

0	0	0
0	1	1
0	1	1
0	0	0
0	0	0
0	1	1
0	1	0
0	1	1

Interpretación:

Dificultad Financiera (25%)

- Solo 5 de los 20 estudiantes reportan dificultades económicas.
- Aunque es minoría, sigue siendo un grupo vulnerable que puede requerir **apoyo financiero o becas**.

Trabajo de Tiempo Completo (65%)

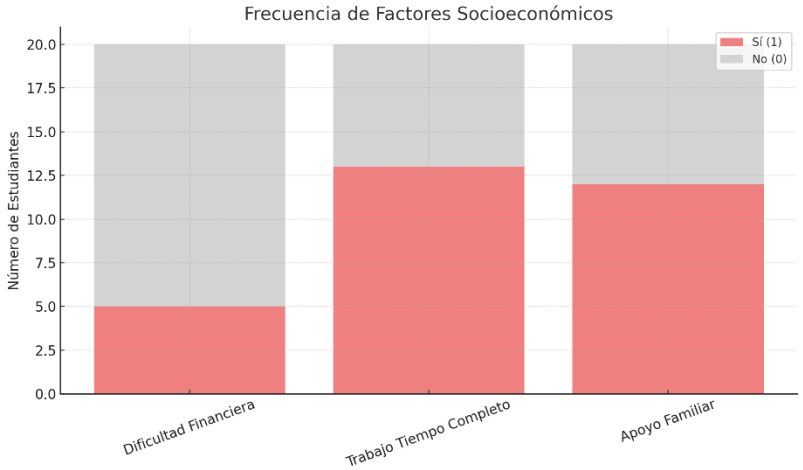
- La mayoría trabaja a tiempo completo. Esto puede afectar su rendimiento académico o disponibilidad horaria.
- Es un **indicador de presión económica o necesidad de ingresos**, lo que podría interferir con la continuidad de estudios.

Apoyo Familiar (60%)

- 12 estudiantes reportan contar con apoyo familiar, lo cual **puede amortiguar el impacto de las dificultades económicas** o académicas.
- Sin embargo, 40% no lo tiene, lo que sugiere un grupo que puede estar en **mayor riesgo de abandono o estrés académico**.

Figura 45

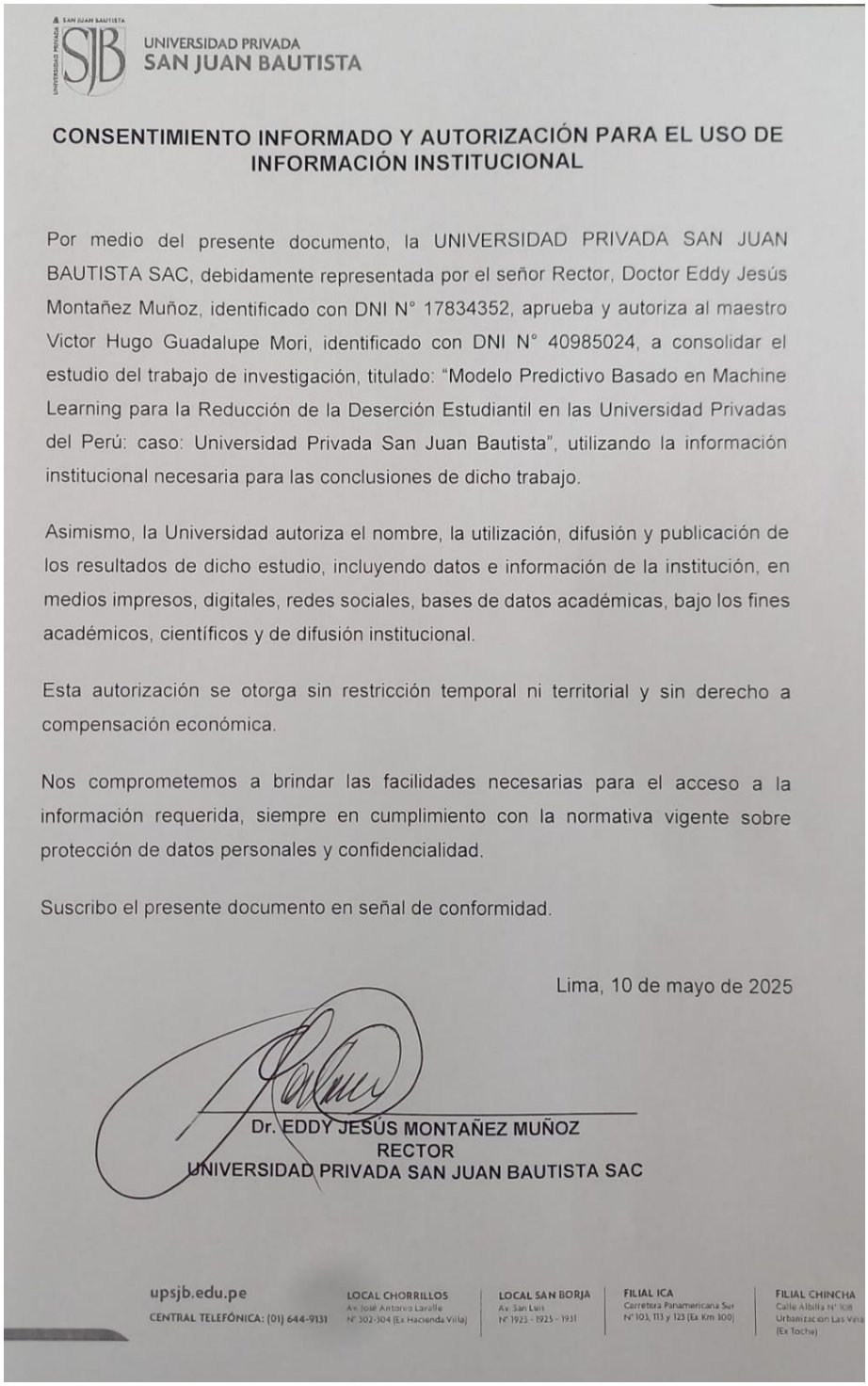
Frecuencia de factores socioeconómicos



Anexo D: Consentimiento Informado

Figura 46

Consentimiento Informado



UNIVERSIDAD PRIVADA
SAN JUAN BAUTISTA

CONSENTIMIENTO INFORMADO Y AUTORIZACIÓN PARA EL USO DE INFORMACIÓN INSTITUCIONAL

Por medio del presente documento, la UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA SAC, debidamente representada por el señor Rector, Doctor Eddy Jesús Montañez Muñoz, identificado con DNI N° 17834352, aprueba y autoriza al maestro Victor Hugo Guadalupe Mori, identificado con DNI N° 40985024, a consolidar el estudio del trabajo de investigación, titulado: "Modelo Predictivo Basado en Machine Learning para la Reducción de la Deserción Estudiantil en las Universidad Privadas del Perú: caso: Universidad Privada San Juan Bautista", utilizando la información institucional necesaria para las conclusiones de dicho trabajo.

Asimismo, la Universidad autoriza el nombre, la utilización, difusión y publicación de los resultados de dicho estudio, incluyendo datos e información de la institución, en medios impresos, digitales, redes sociales, bases de datos académicas, bajo los fines académicos, científicos y de difusión institucional.

Esta autorización se otorga sin restricción temporal ni territorial y sin derecho a compensación económica.

Nos comprometemos a brindar las facilidades necesarias para el acceso a la información requerida, siempre en cumplimiento con la normativa vigente sobre protección de datos personales y confidencialidad.

Suscribo el presente documento en señal de conformidad.

Lima, 10 de mayo de 2025

Dr. EDDY JESÚS MONTAÑEZ MUÑOZ
RECTOR
UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA SAC

upsjb.edu.pe
CENTRAL TELEFÓNICA: (01) 644-9131
LOCAL CHORRILLOS Av. José Antonio Lavalle N° 302-304 (Ex Hacienda Villa)
LOCAL SAN BORJA Av. San Luis N° 1923 - 1925 - 1931
FILIAL ICA Carretera Panamericana Sur N° 103, 113 y 123 (Ex Km 300)
FILIAL CHINCHA Calle Albilla N° 108 Urbanización Las Ver... (Ex Tocha)