



EDITORIAL
MUNDO INTERDISCIPLINARIO

Modelo de Aprendizaje Automático Supervisado para Identificar Patrones de Bajo Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca



Autores:

Rudy Jhean Rojas Pari
Marcos Denys Choque Castro
Gedeoni Chura Mamani
Rocky Dennis Ari Goyzueta
Eder Gutiérrez Quispe

ISBN
978-628-97866-7-5

Autores

Rudy Jhean Rojas Pari Maestro en Investigación y Docencia en Educación Superior Universidad Nacional de Juliaca Docente Universitario de la Facultad de Ciencias de la Ingeniería rjrojas.doc@unaj.edu.pe  http://orcid.org/0009-0003-0351-8219	Marcos Denys Choque Castro Maestro en Ingeniería de Sistemas Universidad Nacional del Altiplano Puno Docente Universitario de la Escuela Profesional de Ingeniería de Sistemas denys.choque@unap.edu.pe  http://orcid.org/0000-0001-8972-7430
Gedeoni Chura Mamani Ingeniero de Sistemas Jefe de Oficina de Tecnologías de la Información Universidad Nacional de Carabaya erika.alvarez@unica.edu.pe  http://orcid.org/0000-0001-6135-1785	Rocky Dennis Ari Goyzueta Ingeniero de Sistemas Docente de la Universidad Nacional del Altiplano Puno rari@epg.unap.edu.pe  http://orcid.org/0000-0002-7509-5775
Eder Gutiérrez Quispe Maestro de Docencia Universitaria Docente Universitario de la Facultad de Ciencias de la Ingeniería Universidad Nacional de Juliaca egutierrezq.doc@unaj.edu.pe  http://orcid.org/0009-0003-0351-8219	



Editor: Nora González

Título

Modelo de aprendizaje automático supervisado para identificar patrones de bajo rendimiento académico en los ingresantes al Instituto de Educación Superior Pedagógico Público - Juliaca

Autores

Rudy Jhean Rojas Pari, Marco Denys Choque Castro, Gedeoni Chura Mamani, Rocky Dennis Ari Goyzueta, Eder Gutiérrez Quispe

ISBN Versión Digital: 978-628-97866-7-5

Sello Editorial:

Editorial Mundo Interdisciplinario

Editor de Publicaciones y Coordinador editorial: Fitzgerald Castro – Cartagena – Colombia

Diagramación:

LFSE Soluciones Empresariales

Portada:

Linda Luz Castro González

Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titular de los derechos patrimoniales

Esta obra está bajo una Licencia Creative Commons – Atribución – No comercial – Compartir igual 4.0 internacional / CC BY-NC-SA 4.0

<https://co.creativecommons.net/tipos-de-licencias/>



Cartagena –Colombia, Octubre 2026

***Modelo de aprendizaje automático supervisado para
identificar patrones de bajo rendimiento académico en los
ingresantes al Instituto de Educación Superior Pedagógico
Público - Juliaca***

Rudy Jhean Rojas Pari

Marcos Denys Choque Castro

Gedeoni Chura Mamani

Rocky Dennis Ari Goyzueta

Eder Gutiérrez Quispe

Colombia

Latinoamérica

2026

DECLARACIÓN DE ORIGINALIDAD Y PROCEDENCIA

La obra titulada "Modelo de aprendizaje automático supervisado para identificar patrones de bajo rendimiento académico en los ingresantes al Instituto de Educación Superior Pedagógico Público - Juliaca" constituye un trabajo original, resultado de un proceso de investigación rigurosamente ejecutado a partir del estudio dogmático que hizo el autor Rudy Jhean Rojas Pari para obtener el título profesional de Ingeniero de Sistemas en la Universidad Peruana Unión, el cual se encuentra publicado en el link: <https://repositorio.upeu.edu.pe/items/ee6a85a1-d380-4e92-b8fc-89f9294b9b2d>

Los demás autores ayudaron en dar forma a la tesis para que pueda ser considerado libro de investigación por la editorial. La presente declaración tiene como finalidad garantizar la transparencia académica, reconocer expresamente el proceso científico efectuado y dejar constancia de la autoría legítima y original de las contribuciones intelectuales que integran esta publicación.

Los autores

Contenido

RESUMEN	x
ABSTRACT	xi
CAPÍTULO I. El Problema	13
1.1. Identificación del Problema	13
1.2. Objetivos de la investigación	15
1.2.1. Objetivo general.	15
1.2.2. Objetivo específico.	15
1.3. Justificación	15
1.4. Presunción filosófica	16
CAPÍTULO II. Revisión de la Literatura	18
2.1. Antecedentes de la Investigación	18
2.1.1. Internacional	18
2.1.2. Nacional.....	19
2.1.3. Local	21
2.2. Marco Teórico	22
2.2.1. Teoría General de Sistemas	22
2.2.2. Machine Learning.....	22
2.2.3. ¿Quiénes utilizan Machine Learning?	23
2.2.4. Tipos de Aprendizaje de Machine Learning.....	27
2.2.5. Algoritmos de Machine Learning.....	30
2.2.6. Metodologías de Minería de Datos y Ciencia de Datos	38
2.2.7. Machine Learning con Python.....	45
2.2.8. Rendimiento Académico	47
CAPÍTULO III. Materiales y Métodos.....	50
3.1. Lugar de Ejecución.....	50
3.2. Población y tamaño de muestra	50
3.2.1. Población	50
3.2.1. Muestra	50
3.3. Materiales e Insumos	51
3.4. Metodología de la Investigación.....	51
3.4.1. Tipo de Investigación	51

3.4.2. Arquitectura de Solución	51
3.5. Aplicación de la metodología CRISP-DM	52
3.5.1. Fase 1: Comprensión del negocio.....	53
3.5.2. Fase 2: Comprensión de los datos	54
3.5.3 Fase 3: Preparación de los datos.....	62
3.5.4 Fase 4: Modelamiento	77
3.5.5. Fase 5: Evaluación.....	82
3.5.6 Fase 6: Implementación.....	83
CAPÍTULO IV. Resultados.....	85
4.1 Resultado 1	85
4.1 Resultado 2	85
4.1 Resultado 3	86
4.1 Resultados 4.....	88
4.1 Resultado 5	88
CAPÍTULO V. Conclusiones y Recomendaciones	90
5.1. Conclusiones.....	90
5.2. Recomendaciones	91
REFERENCIAS BIBLIOGRÁFICAS	92
ANEXOS	99

SÍMBOLOS USADOS

- I.E.S.P.P.J: Instituto de Educación Superior Pedagógico Público Juliaca
- SIGES: Sistema de Información Académica para Institutos de Educación Superior Pedagógica
- ML: Machine Learning (Aprendizaje Automático)
- CRISP-DM: Cross Industry Standard Process for Data Mining (Metodología más usada en la minería de datos)
- MINEDU: Ministerio de Educación
- IDE: Entorno de Desarrollo Integrado
- Taget: Columna a predecir.
- Accuracy: Puntuación de predicción.
- Features: Variables

RESUMEN

En presente estudio se llevó a cabo en el Instituto de Educación Superior Pedagógico Público Juliaca (IESPPJ), ubicado en el distrito de San Miguel de la Provincia de San Román, durante el año 2020, tuvo como objetivo general implementar un modelo de aprendizaje automático supervisado para identificar patrones de bajo rendimiento académico en los ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca, para su desarrollo de empleó la metodología de minería de datos denominado: CRISP-DM (Cross Industry Standard Process for Data Mining), el algoritmo Random Forest Classifier, dicho algoritmo fue entrenado con datos socioeconómicos, datos de admisión y datos académicos logrando un accuracy del 86%, permitiendo identificar las variables que más influyen en el rendimiento académico tales como: El promedio final del examen de admisión, edad, número de horas diarias que actualmente dedica al estudio, distrito en donde está ubicado el centro de estudios secundarios de procedencia, programa de estudios al que está postulando, número de dormitorios de su vivienda, cantidad de años en la que cursó la educación secundaria, idioma nativo que habla, ¿Cada cuánto tiempo recibes ayuda económica?, sexo, ¿La persona que mantiene su hogar es?, tipo de preparación que recibiste para postular al IESP, número de veces que postulaste a otros Institutos / Universidades, tipo de material de la vivienda, las variables identificadas influyen en el orden mencionado. Como trabajos futuros en el área de educación se propone profundizar el estudio utilizando nuevas fuentes de información, tales como información psicológica y/o historial médico de los estudiantes, para mejorar la toma de decisión.

Palabras Clave: Aprendizaje automático supervisado, Bajo rendimiento académico, Data Mining, CRISP-DM, Random Forest Classifier.

ABSTRACT

In this study, it was carried out at the Juliaca Public Pedagogical Higher Education Institute (IESPPJ), located in the San Miguel district of the San Román Province, during 2020, the general objective of which was to implement a supervised machine learning model To identify patterns of low academic performance in those entering the Public Pedagogical Higher Education Institute - Juliaca, for its development the data mining methodology called: CRISP-DM (Cross Industry Standard Process for Data Mining), the Random Forest algorithm Classifier, said algorithm was trained with socioeconomic data, admission data and academic data achieving an accuracy of 86%, allowing to identify the variables that most influence academic performance such as: The final average of the admission exam, age, number of hours daily that he currently dedicates to the study, district where the secondary school of proc is located age, study program to which you are applying, number of bedrooms in your home, number of years in which you completed secondary education, native language spoken, How often do you receive financial aid? Is it home? Type of preparation you received to apply to the IESP, number of times you applied to other Institutes / Universities, type of housing material, the identified variables influence the order mentioned. As future work in the area of education, it is proposed to deepen the study using new sources of information, such as psychological information and / or medical history of students, to improve decision-making.

Keywords: Supervised machine learning, Low academic performance, Data Mining, CRISP-DM, Random Forest Classifier.

Capítulo I

EL PROBLEMA

CAPÍTULO I. El Problema

1.1. Identificación del Problema

Según Pérez & Aldás (2019), España es uno de los países que se encuentran entre los países que menos aprovecha el esfuerzo en la educación superior, pues dicho país cuenta con una elevada tasa de abandono de estudios iniciados; De acuerdo a su informe indican que un 33% de los estudiantes españoles no terminan el nivel académico en el que se inscribieron, el 21% abandona la universidad sin obtener un título profesional y el 12% restante cambia de carrera profesional; y uno de los tantos factores que hacen que los estudiantes abandonen sus estudios es el bajo rendimiento académico. Estas elevadas tasas de abandono reflejan un importante desaprovechamiento dedicados a la formación universitaria y estos fracasos son pérdidas anuales que se acercan a los mil millones de euros.

Rico, Gaytán & Sánchez (2019), en su investigación identificaron como un problema el poco desarrollo de modelos predictivos con técnicas de minería de datos, para la prevención de reprobación y deserción escolar, dado que este tipo de aplicaciones brindan un potencial beneficio a una institución y/o organización en la mejora del rendimiento académico de los estudiantes. Por otro lado, Zainab, Noor, Wasan & Hazim (2020), identifican que la gestión de la de educación y el desempeño de los estudiantes tienen una estrecha relación. Esto se debe a que existen varios factores que afectan el rendimiento académico y luego la calidad de la educación.

Según Jara y otros (2008), en su investigación mencionan que el problema del bajo rendimiento académico en los universitarios está influenciado por múltiples factores que se manifiestan en los primeros años de estudio. Por otro lado Yamao, Celi, Campos & Huancas (2018), mencionan que los ingresantes a las universidades son los mas vulnerables a enfrentar problemas de rendimiento académico, resultando finalmente en una deserción académica.

Laura, Paredes & Baluarte (2017), el objetivo de la educación superior es garantizar e incrementar la calidad de la educación, por ende aumentar la tasa de egresados y disminuir la deserción, en la actualidad la deserción y el bajo rendimiento académico son los problemas que más abordan la mayoría de instituciones.

Por su parte, el Instituto de Educación Superior Pedagógico Público – Juliaca brinda servicios de educación superior con las carreras profesionales como: Educación Inicial, Educación Primaria y Educación Secundaria en las siguientes menciones: Matemática, Comunicación, Ciencia Tecnología y Ambiente y Ciencias Sociales, dicha casa de estudios cuenta con estudiantes de diferentes lugares de procedencia y en cada proceso de admisión se tienen nuevos ingresantes; Según la entrevista con el Director del Instituto de Educación Superior Pedagógico Público – Juliaca (Mamani Vargas, 2020), menciona que el problema que se tiene es que estos nuevos ingresantes, no cuentan un mismo nivel académico ya sea porque provienen de diferentes lugares de la región o son provenientes de una Institución Educativa Secundaria Pública o Privada con un bajo nivel académico.

Así mismo, el Instituto de Educación Superior Pedagógico Público – Juliaca cuenta con un sistema académico del Ministerio de Educación denominado; Sistema de Información Académica – SIGES (MINEDU, 2020), dicho sistema académico gestiona: datos personales del estudiante, proceso de matrícula, carga académica, gestión de notas, a su vez la oficina de admisión del Instituto de Educación Superior Pedagógico Público – Juliaca, cuenta con registros de los datos de los ingresantes en cada admisión, datos como: Lugar de procedencia, Institución de procedencia, nota obtenida en el examen de admisión, nota obtenida en la entrevista personal, modalidad de ingreso, periodo de ingreso y datos socioeconómicos.

Con lo expuesto, la presente investigación pretende analizar los datos de los estudiantes que cursaron entre el cuarto y sexto ciclo académico en el Instituto de Educación Superior

Pedagógico Público – Juliaca, para analizar los factores que influyen en el bajo rendimiento académico en los ingresantes, aplicando modelos de aprendizaje supervisado.

1.2. Objetivos de la investigación

1.2.1. Objetivo general.

Implementar un modelo de aprendizaje automático supervisado para identificar patrones de bajo rendimiento académico en los ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca.

1.2.2. Objetivo específico.

1. Obtener y preparar datos de los ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca.
2. Desarrollar el modelo predictivo basado en aprendizaje automático supervisado.
3. Identificar las variables que influyen en el bajo rendimiento académico de los estudiantes del nivel superior.
4. Evaluar los resultados del modelo predictivo para identificar al postulante con bajo rendimiento académico.
5. Desarrollo de una interfaz web para el uso del modelo predictivo para identificar al postulante con bajo rendimiento académico.

1.3. Justificación

El Modelo de Machine Learning para identificar al postulante con bajo rendimiento académico, aplicando la técnica de Aprendizaje Supervisado en el sector educación, será un gran aporte para poder realizar futuros proyectos investigación en esta rama.

Ayudará detectar a los ingresantes que podrían enfrentarse a problemas de bajo rendimiento académico y permitirá brindar un seguimiento personalizado a dichos estudiantes para que puedan mejorar su rendimiento académico.

Facilitará el incremento de la confianza de la sociedad en conjunto, al saber de que los jóvenes del distrito de San Miguel, que estudian en el Instituto de Educación Superior Pedagógico Público – Juliaca, reciben una educación superior de calidad.

Facilitará la toma de decisiones a los directores de las Unidades Académicas del Instituto de Educación Superior Pedagógico Público – Juliaca, para poder mejorar el rendimiento académico a través de un seguimiento personalizado de los estudiantes de las diferentes carreras profesionales que ofrece el Instituto, los resultados obtenidos no solo beneficiará al Instituto ya antes mencionado, sino también beneficiará a todas las Instituciones que brinda el servicio de Educación Superior.

1.4. Presunción filosófica

En (Éxodo 35:30-32) “Y dijo Moisés a los hijos de Israel: Mirad, Jehová ha nombrado a Bezaleel hijo de Uri, hijo de Hur, de la tribu de Juda; Y lo ha henchido de espíritu de Dios, en sabiduría, en inteligencia, en ciencia y en todo artificio, para proyectar inventos, para trabajar en oro, en plata y en metal”.

Capítulo II

REVISIÓN DE LA LITERATURA

CAPÍTULO II. Revisión de la Literatura

2.1. Antecedentes de la Investigación

Realizando la revisión de los antecedentes se pudo encontrar trabajos de investigación que hablan de Minería de Datos, Ciencia de Datos y Machine Learning; que son aplicadas en el área de educación, turismo y entre otros, dichos antecedentes ayudaran a complementar el presente trabajo de investigación:

2.1.1. Internacional

Rico, Gaytán & Sánchez (2019) en su investigación “Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes”, realizado en el Instituto Politécnico Nacional en la Ciudad de México, cuyo objetivo fue la construcción de un modelo predictivo para predecir el rendimiento académico de estudiantes universitarios, aplicando el algoritmo de minería de datos Naïve Bayes, las variables que fueron utilizadas son: Escolaridad del padre, escolaridad de la madre, ingresos familiares, promedio final obtenido en el bachillerato, cantidad de materias reprobadas actualmente, promedio actual, ¿cómo prefieres estudiar?, ¿cómo prefieres realizar actividades en clase?, ¿qué tan frecuentemente estudias?, para poder relizar su investigación mediante la técnica de minería de datos utilizaron la metodología KDD; la investigación les permitió identificar los factores que más influyen en el rendimiento académico, en donde el modelo predictivo identificó que el factor más influye en que un estudiante apruebe el curso, es si el estudiante prefiere estudiar solo.

La investigación realizada por Rico, Gaytán & Sánchez se encuentra enmarcada específicamente dentro de la minería de datos, tema de interés en la presente investigación. Así mismo muestra la importancia que tiene este tipo de estudios para los profesores, ya que es de gran ayuda al momento de diseñar estrategias de prevención e identificar a los estudiantes vulnerables.

García & Skarita (2018) en su investigación “Prediciendo el desempeño académico según el entorno familiar de los estudiantes: evidencia para Colombia usando árboles de clasificación”, fue realizado en el Instituto Colombiano de Evaluación Educativa (ICFES), en donde se tuvo como objetivo de determinar mediante los arboles de clasificación qué características familiares son las mejores para predecir el rendimiento académico de los estudiantes que presentaron el examen de estado de 2016 para acceder a la educación superior, las variables que usaron son: Lugar de residencia, número de habitaciones, material predominante del piso, estrato socioeconómico de la vivienda, bienes (teléfono fijo, lavadora, microondas, horno y auto), nivel de educación de la madre, ocupación de la madre, nivel de educación del padre, ocupación del padre, cantidad de libros, número de hermanos, servicio de internet, número de personas que residen en el hogar, etc. La investigación permitió determinar cuáles son las variables familiares que mejor predicen los resultados académicos, se presentaron en el siguiente orden: el nivel educativo de la madre, el estrato socioeconómico de la vivienda, el número de libros, el nivel educativo del padre y el poseer computador en la vivienda.

La investigación realizada por García Gonzales & Skarita, se encuentra dentro del tema de interés de la presente investigación. Así mismo, precisa la importancia de predecir el desempeño académico a partir de las observaciones y características familiares propias de los estudiantes.

2.1.2. Nacional

Bernuy (2018) en su tesis “Predicción del Rendimiento Académico Mediante Minería de Datos en Estudiantes del Primer Ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima - Perú” elaborado en la Universidad de San Martín de Porres, cuyo objetivo fue “Predecir el rendimiento académico mediante minería de datos en estudiantes del primer ciclo de la Escuela

Profesional de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres”; para lo cual utilizó datos de la oficina de admisión: “Semestre de ingreso, fecha de nacimiento, sexo, modalidad de ingreso, nombre del colegio de procedencia, departamento, provincia, distrito del colegio, tipo de colegio, puntaje obtenido en el examen de admisión y de la Facultad de Ingeniería y Arquitectura: Estado de matrícula, Escala de pensión, Dirección de domicilio, Cursos llevados, Sección y Nota”. Para poder realizar su modelo de predicción de regresión lineal, árbol de decisiones y support vector utilizó la metodología CRISP-DM. La investigación permitió determinar los factores más influyentes en el rendimiento académico, las cuales son: nota de examen de admisión, género, edad, modalidad de ingreso y distancia desde su casa hasta el centro de estudios.

El estudio realizado por Bernuy está enmarcado dentro de la minería de datos, tema de interés de esta investigación. Así mismo, abordó sobre los factores más influyentes en el rendimiento académico de los ingresantes.

La tesis de Holgado (2018) titulado “Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018”, cuyo objetivo fue “detectar los patrones de bajo rendimiento académico de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios, mediante el uso de minería de datos”, para el desarrollo de su investigación utilizó la metodología CRISP-DM, en donde llegó a la conclusión de que el algoritmo Random Forest permitió identificar que las variables: “cantidad de asignaturas cursadas, el servicio de comedor universitario, la carrera profesional, deuda con la universidad”, son las variables que más influyen en la predicción del rendimiento académico.

El estudio realizado por Holgado está enmarcado dentro de la minería de datos, tema de interés para la presente investigación. Así mismo, abordó sobre la comparación de

algoritmos de Random Forest, C5.0 y CART, en donde el algoritmo que mejor clasificación logró fue el algoritmo C5.0.

2.1.3. Local

Coyla (2016) en su tesis “Análisis de datos con BigData en proceso de admisión de la Universidad Nacional del Altiplano de Puno, 2016”, cuyo objetivo fue el de “identificar características y patrones de comportamiento con el desempeño académico de los ingresantes a la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano utilizando Bigdata”, para realizar su la investigación tomó a un total de 18 estudiantes ingresantes a la carrera profesional de Ingeniería de Sistemas de la UNAP del proceso de admisión CEPREUNA Enero Marzo del año 2015, la investigación permitió determinar el nivel de conocimiento de Matemática I, dicha investigación fue elaborada con la metodología SEMMA. Aplicando el lenguaje de programación R con el paquete Rattle, llegando a la conclusión de que los ingresantes razonan y demuestran proposiciones matemáticas, representan, analizan e interpretan datos matemáticos contextualizados y resuelven problemas matemáticos contextualizados.

El estudio realizado por Coyla, está enmarcado dentro del Big Data, tema de interés para la presente investigación. Así mismo, abordó temas de interés para mejorar el nivel de desempeño académico de los estudiantes de Ingeniería de Sistemas de la Universidad Nacional del Altiplano.

Escarcena & Velasquez (2017) en su tesis “Análisis de datos con R para determinar el nivel de cumplimiento del perfil del ingresante a la Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas de la UNA - Puno, 2017”, cuyo objetivo fue el de “realizar un análisis de datos con R para determinar el nivel del perfil del ingresante a la Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas de la Universidad Nacional del Altiplano”, para la investigación se tomó a los ingresantes a la Facultad de Ingeniería Mecánica Eléctrica,

Electrónica y Sistemas de la Universidad Nacional del Altiplano en el examen de modalidad general del 21 de mayo del año 2017, para el desarrollo de la investigación utilizó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), en donde los resultados demuestran que, en la escuela profesional de Ingeniería Mecánica Eléctrica, tiene un mejor perfil del ingresante.

El estudio realizado por Escarcena & Velasquez, está enmarcado dentro de la Minería de Datos, tema de interés para la presente investigación. Así mismo, abordó temas de interés para mejorar el nivel del perfil del ingresante a la Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas de la Universidad Nacional del Altiplano.

2.2. Marco Teórico

2.2.1. Teoría General de Sistemas

La teoría general de sistemas describe un nivel de formación de modelos teóricos que se encuentra entre las construcciones extremadamente generalizadas de las matemáticas puras y las teorías específicas de las disciplinas especializadas, y que en estas últimas fuentes fue influenciado por el sentimiento de una situación cada vez más fuerte de un espesor sistémico de construcciones teóricas que pueda discutir, examinar y dilucidar el compromiso generales del mundo empírico (Johansen, 2004).

El Biólogo Ludwig Von Bertalanffy (1901-1972) acuñó el nombre de "Teoría general de sistemas". La TGS debería ser un mecanismo de integración entre las ciencias naturales y sociales y al mismo tiempo una herramienta básica para la formación y preparación de científicos (Arnold Cathalifaud & Osorio, 1998).

2.2.2. Machine Learning

Según Joyanes (2019), el aprendizaje automático es una disciplina en informática y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que las computadoras

puedan usar para aprender. A su vez BSG INSTITUTE (2019), menciona que el Aprendizaje Automático es una disciplina de la Inteligencia Artificial (IA) que ofrece a los sistemas la oportunidad de aprender automáticamente a partir de la experiencia y mejorar. El Aprendizaje Automático utiliza algoritmos que son capaces de aprender cuando se exponen a nuevos datos, ofreciendo resultados más precisos para identificar comportamientos que permitan tomar las mejores decisiones.

Por otro lado, Hurwitz & Kirsch (2018), en su libro menciona que el Machine Learning se ha convertido en una de las tecnologías más importantes en las organizaciones empresariales que buscan formas innovadoras de utilizar los datos existentes para alcanzar un nuevo nivel de mercado. Las organizaciones empresariales están constantemente tratando de predecir cambios en la empresa para lograr sus objetivos comerciales.



Figura 1: Categorías generales que incluye Machine Learning.

Fuente: (Sanseviero, 2018)

2.2.3. ¿Quiénes utilizan Machine Learning?

En la actualidad la mayoría de las organizaciones que tienen a disposición una gran cantidad de datos almacenados, están trabajando con Machine Learning, dado que los datos

históricos que son almacenados son de gran valor que aporta a su organización y de esa manera lograr competitividad y tener ventaja sobre sus competidores.

2.2.3.1. Educación

Tan, Shi & Tang (2018), mencionan que en el sector educativo las técnicas de aprendizaje automático son utilizadas para comprender y explorar patrones que caractericen el comportamiento de los estudiantes, mediante métodos y algoritmos, basados en sus datos, evaluaciones y dominio de sus conocimientos. A su vez Martínez, Santos-Martínez, & Puche, (2018), mencionan que las técnicas de aprendizaje automático son utilizadas en la educación para poder procesar la gran cantidad de información que pueden ser registradas en las diferentes plataformas de e-learning (Moodle, Canvas LMS, Chamilo LMS, etc), gracias al análisis de esta información los docentes pueden descubrir los patrones de aprendizaje que llevan los estudiantes y de esa manera adaptar los cursos y lograr una educación personalizada.

Por otro lado, Biffi (2018) Afirma que la Inteligencia Artificial es una de las principales tecnologías que está transformando la educación, esta tecnología puede contribuir de muchas maneras, como diagnosticar problemas de aprendizaje, dislexia y el autismo.

2.2.3.2. Servicios financieros

Las entidades que prestan servicios financieros, en la actualidad utilizan Machine Learning para diferentes fines; que pueden ser el identificar patrones importantes en los datos que han almacenado a lo largo del tiempo. Los patrones que se pueden identificar para poder encontrar oportunidades de inversión o bien ayudar a las personas y/o empresas inversionistas a saber cuándo vender o comprar (SAS, 2019).

Según Joyanes (2019), en la industria bancaria aplicada la inteligencia de negocios, Minería de Datos y/o Machine Learning se pueden realizar las siguientes investigaciones:

- ✓ Detección de patrones de uso fraudulento de tarjetas y transacciones de banca en línea (online).
- ✓ Automatización de los procesos de concesión de préstamos para predecir, con la mayor precisión posible, los morosos más probables.
- ✓ Estudio de concesión de tarjetas de crédito.
- ✓ Determinación del gasto en tarjetas por segmentación de grupos.
- ✓ Identificación de reglas de comportamiento del mercado de valores a partir de los registros históricos de dichos mercados.
- ✓ Predicciones de hábitos y patrones de compra en grandes almacenes y en mercados en línea.
- ✓ Detección de segmentos de clientes predispuestos a la compra de determinados artículos, bien en el lanzamiento o cuando ya están en el mercado.
- ✓ Identificación de clientes fieles y también de fuga de clientes.

2.2.3.3. Medicina

El Machine Learning es una tecnología que está en rápido crecimiento, la industria de la medicina no es ajena a este crecimiento (SAS, 2019).

“La Minería de Datos en medicina es una de las aplicaciones más prácticas, debido a que complementa la investigación médica en análisis clínicos y en el trascendental campo de los diagnósticos, entre otras especialidades” (Joyanes, 2019, pág. 228).

Según Joyanes (2019), en la industria de la medicina se pueden realizar las siguientes investigaciones:

- ✓ Identificación de patrones novedosos para mejorar la supervivencia de pacientes con cáncer.
- ✓ Predicción de tasas de éxito en trasplantes de órganos a pacientes para desarrollar políticas de donantes/receptores en el tratamiento clínico.

- ✓ Descubrimiento de las relaciones entre síntomas y enfermedades, así como entre enfermedades y tratamientos con éxito.
- ✓ Estudio de factores de riesgo en diferentes patologías
- ✓ Segmentación de pacientes por grupos afines.
- ✓ Gestión hospitalaria y clínica para planificación temporal de habitaciones, quirófanos, salas de consulta, etc.

2.2.3.4. Marketing y ventas

Los sitios Web hoy en día que hacen constantemente recomendaciones de artículos que podrían gustarle, se basan en patrones de compras anteriores para poder realizar estas recomendaciones se utilizan Machine Learning, Con los rastros que uno va dejando cuando navega por internet, estos rastros son tomados y almacenado para luego ser analizados y de ese modo van promocionando artículos similares o que guarde alguna relación con lo que han estado buscando y que podrían interesarle al usuario (SAS, 2019).

2.2.3.5. Turismo

“En la industria del turismo existe una gran variedad de aplicaciones para hoteles, líneas aéreas, resorts, viajes, alquiler de automóviles, trenes, etc.” (Joyanes, 2019, pág. 230)

Según Joyanes (2019) en la industria del turismo se pueden realizar las siguientes investigaciones:

- ✓ Predicciones de ventas de diferentes servicios (reserva de asientos en diferentes clases, reserva de habitaciones en hoteles/resorts, reserva de autos en compañías de alquiler, etc.).
- ✓ Identificación de los clientes más rentables para proporcionarles mejores servicios (por ejemplo, las tarjetas de fidelización “millas” de los clientes “viajeros

frecuentes”, a los que se ofrecen beneficios como “prioridad en salas VIP”, upgrade (subida) de categoría, ofertas especiales en función de la tarjeta de fidelización, etc.).

- ✓ Predicción de ocupación en aviones, trenes, etcétera, dependiendo de rutas viajeras, épocas del año, entre otras.

2.2.4. Tipos de Aprendizaje de Machine Learning

En la actualidad existen varios tipos de aprendizaje de Machine Learning, pero los que más utilizan son el aprendizaje supervisado y el aprendizaje no supervisado.

2.2.4.1. Supervisado

El aprendizaje supervisado requiere una intervención humana, para poder indicar lo que está bien o lo que no está bien, muchas aplicaciones computacionales son intervenidas por los humanos y proporciona la semántica necesaria para que los algoritmos aprendan. (Joyanes, 2019)

Asimismo, Molina & García (2012), menciona que, en el aprendizaje inductivo supervisado, hay un atributo especial en todos los ejemplos, comúnmente conocido como una clase, que indica si el ejemplo pertenece o no a un concepto particular que será el objetivo del aprendizaje (pág. 98).

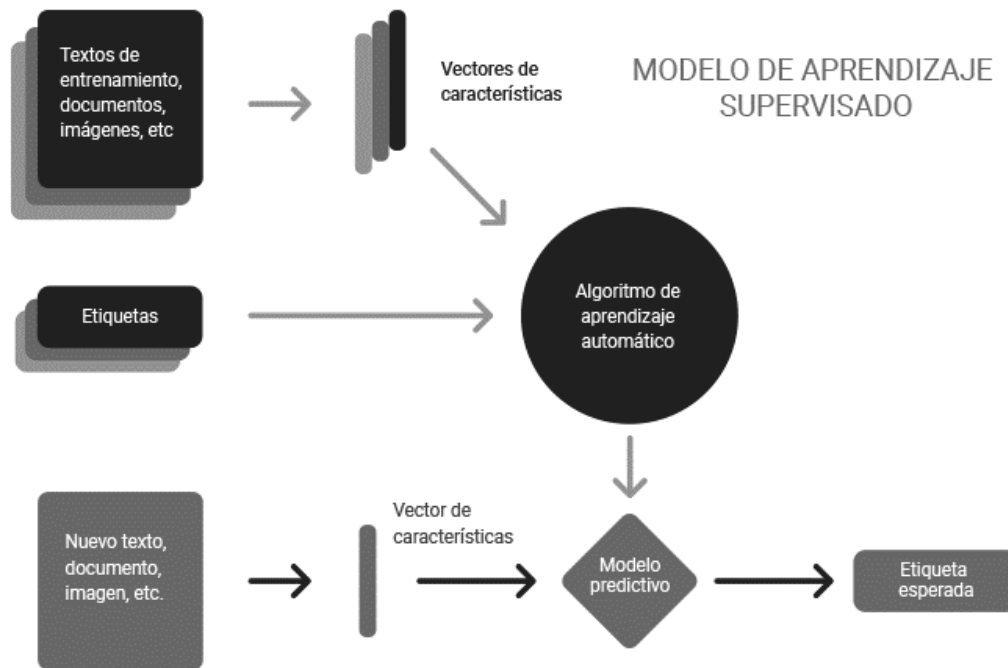


Figura 2: Diagrama de flujo del aprendizaje supervisado

Fuente: (Luna, 2018)

2.2.4.2. No Supervisado

El aprendizaje no supervisado consiste en entrenar una red que la expone a una variedad de ejemplos sin "decir" qué buscar. Más bien, la red aprende a reconocer características y agruparlas con ejemplos similares, lo que ayuda a identificar grupos, enlaces o patrones ocultos dentro de los datos (Joyanes, 2019).

A su vez Molina & García (2012), menciona que el aprendizaje no supervisado aprende sin la ayuda del maestro; Es decir, el aprendizaje no supervisado trata de ordenar los ejemplos en una jerarquía de acuerdo con las regularidades en la distribución de los pares de atributo-valor sin la guía de la clase de atributo específica. Este es el proceso de sistemas que realizan agrupaciones conceptuales y que también se supone que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es sintetizar conocimiento cualitativo o cuantitativo, el objetivo de los sistemas que realizan tareas de descubrimiento (pág. 97).

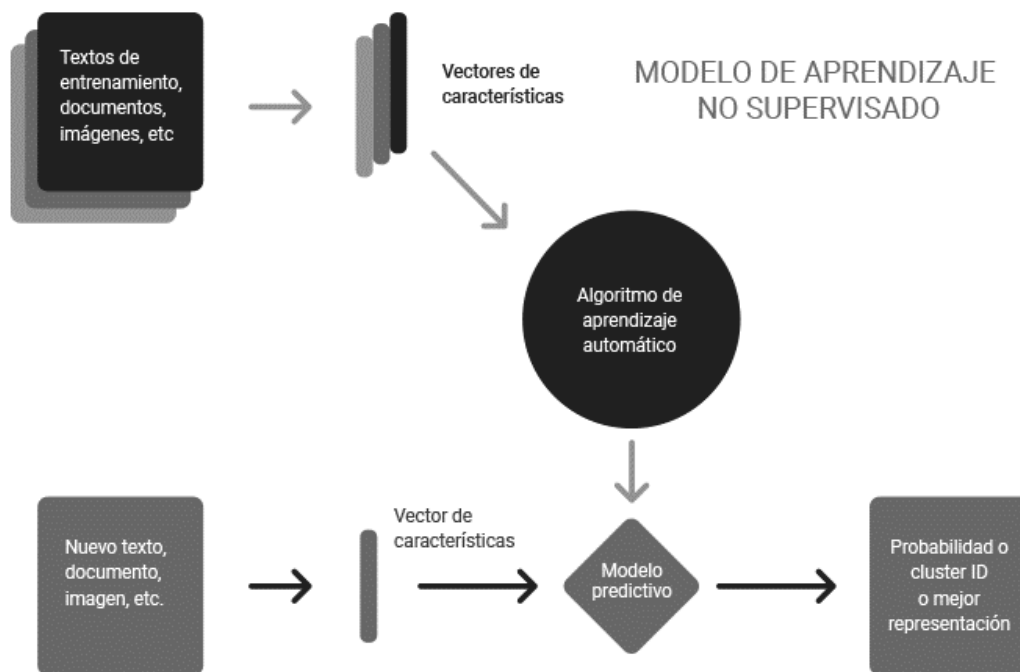


Figura 3: Diagrama de flujo del aprendizaje no supervisado

Fuente: (Luna, 2018)

2.2.4.3. Reforzado

El aprendizaje reforzado es una mezcla de aprendizaje supervisado y no supervisado. Se basa en la psicología del comportamiento e implica entrenar una red neuronal para interactuar con su entorno, ocasionalmente informando con una recompensa. Su entrenamiento es ajustar los pesos de la red para encontrar la estrategia que genere más recompensas de una manera más consistente (Joyanes, 2019, pág. 389).

MODELO DE APRENDIZAJE POR REFUERZO

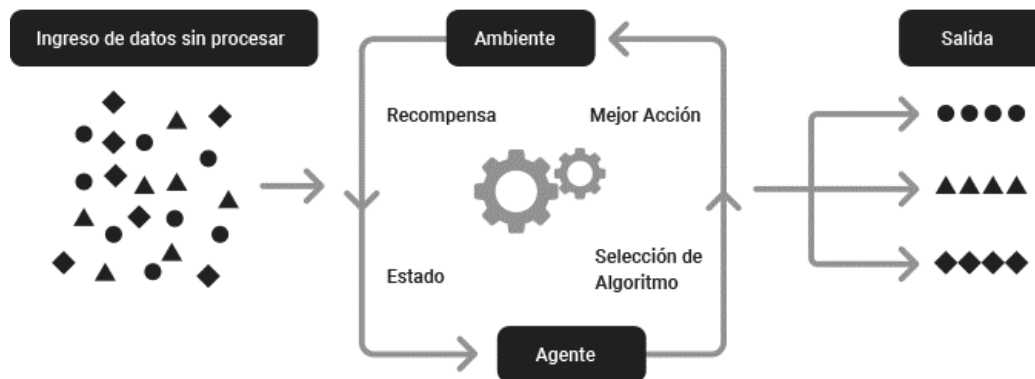


Figura 4: Diagrama de flujo aprendizaje por refuerzo

Fuente: (Luna, 2018)

2.2.5. Algoritmos de Machine Learning

2.2.5.1. Árboles de decisión

Los árboles de decisión es uno de los tantos algoritmos de aprendizaje supervisado; es uno de los modelos más populares y utilizados para realizar clasificaciones.

Barrientos y otros (2009), afirman que un árbol de decisión es un modelo predictivo, cuyo objetivo principal es el aprendizaje supervisado a partir de observaciones y construcciones lógicas, se utilizan para representar y clasificar una serie de condiciones que ocurren sucesivamente para resolver un problema.

También indican que este tipo de modelo se basa en la descripción narrativa de un problema porque proporciona una visión gráfica de la toma de decisiones y las variables a evaluar, las acciones a tomar y el orden en que se toma la decisión. Cada vez que se ejecuta este tipo de modelo, solo se sigue una ruta, según el valor actual de la variable evaluada. Los valores que las variables pueden tomar para este tipo de modelo pueden ser discretos o continuos (Barrientos, y otros, 2009)

Por otro lado, Mendoza (2018) los árboles de decisión son algoritmos utilizados como modelo predictivo en diversas disciplinas. Estos son similares a los diagramas de flujo, en el que se llega a puntos en los que se toman decisiones según una regla

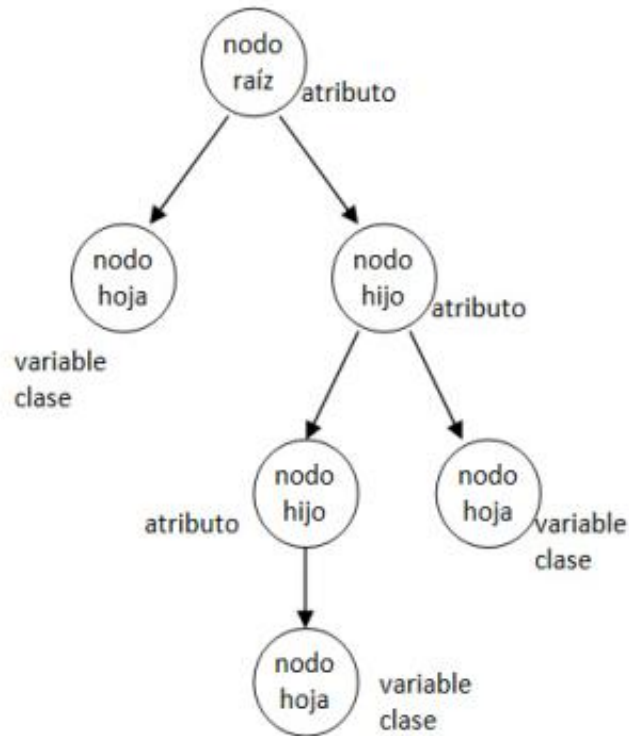


Figura 5: Estructura de un árbol de decisión

Fuente: (Barrientos, y otros, 2009)

2.2.5.1.1. Términos comunes utilizados con árboles de decisión.

Jain (2017), en su blog menciona que los términos más comunes que se utilizan para comprender los árboles de decisión son:

1. **Nodo raíz:** representa a toda la población o muestra y se divide en dos o más conjuntos homogéneos.
2. **División:** es un proceso de división de un nodo en dos o más sub nodos.
3. **Nodo de decisión:** cuando un sub nodo se divide en sub nodos adicionales, se llama nodo de decisión.

4. **Nodo hoja / terminal:** los nodos sin hijos (sin división adicional) se llaman nodo hoja o terminal.
5. **Poda:** cuando reducimos el tamaño de los árboles de decisión eliminando nodos (opuesto a la división), el proceso se llama poda.
6. **Rama / Subárbol:** Una subsección del árbol de decisión se denomina rama o subárbol.
7. **Nodo primario y secundario:** un nodo, que se divide en sub nodos, se denomina nodo primario de sub nodos, donde los sub nodos son secundarios del nodo primario.

2.2.5.1.2. ID3.

Jain (2017), el algoritmo ID3 es un algoritmo central que nos permite construir arboles de decisión, fue desarrollado por JR Quinlan, dicho algoritmo hace una búsqueda de arriba hacia abajo a través de sus posibles ramas sin retroceder. “El algoritmo ID3 usa entropía para calcular la homogeneidad de una muestra” (Sehra, 2018).

Jain (2017) Menciona en su blog para poder construir un árbol de decisión, se necesita calcular dos tipos de entropía usando tablas de frecuencia de la siguiente manera:

Entropía utilizando la tabla de frecuencias de un atributo:

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

- a. Entropía utilizando la tabla de frecuencias de dos atributos:

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

2.2.5.2. Random Forest

Según Villa, Carrión & Sozzi (2017), esta técnica pertenece a los algoritmos de aprendizaje supervisado, este algoritmo se basa en la construcción de árboles predictivos

utilizando bootstrap y Bagging, la cual garantiza la estabilidad del proceso. Cada árbol se crea utilizando muestras de bootstrap con reemplazo para corregir el error de predicción generado como resultado de la selección específica de una muestra. Para cada división de un nodo, la mejor variable de todas no se selecciona como en CART, sino que se selecciona al azar un conjunto de variables de un tamaño predeterminado y la selección de la variable de división se limita a este conjunto. De esta forma, se incluye una mayor variabilidad de los árboles y se reduce la dependencia del resultado en las subdivisiones anteriores.

El proceso OOB consiste en utilizar el conjunto de datos de entrenamiento T para crear k muestras bootstrap T_k , se construyen los árboles $h(x, T_k)$, y el promedio de estos será el predictor bagget. En adelante para cada (y, x) de T se construyen los árboles en cada T_k que no contienen a (y, x) , estas son las muestras que quedaron fuera de las muestras de bootstrap. El OOB permite estimar el error de clasificación, también son usadas para calcular la fuerza de predicción de cada una de las variables.

Villa, Carrión & Sozzi (2017), resumen el algoritmo de Random Forest en los siguientes pasos:

- Se crea B muestras bootstrap de tamaño N del conjunto train.
- Se crean $T_b, (b = 1, \dots, B)$ árboles con las muestras hasta que se obtiene el tamaño mínimo en el nodo terminal. Esto se alarga de forma recursiva mediante los siguientes pasos:
 1. Seleccionar aleatoriamente m_{try} variables del conjunto total de P variables.
 2. Seleccionar la óptima variable de división entre las p variables.
 3. Dividir el nodo en dos nodos hijos.
- El conjunto de salida es el ensamble (promedio) de los $\{T_n\}_1^B$ árboles, es decir:

$$f_{RF}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$$

- La estimación de la tasa de error o erros de clasificación de obtiene mediante el conjunto OOB.

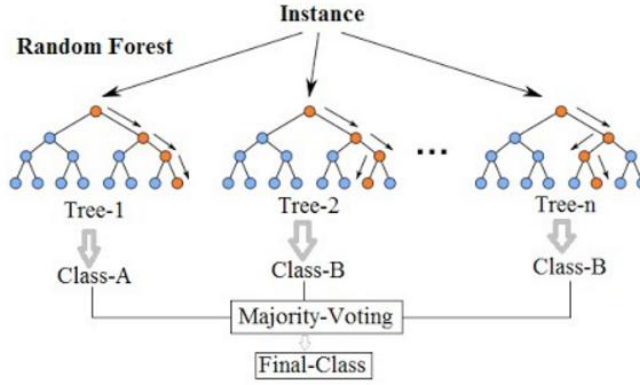


Figura 6: Estructura de Random Forest

Fuente: (Roman, 2019)

2.2.5.3. Extra Trees Classifier

Extra Trees Classifier es una variante del Random Forest, este algoritmo también es conocido como extremely randomized trees. Los Extra Trees difieren del Random Forest de la siguiente manera: (a) El procedimiento de ensacado no puede ser aplicado por extra trees para construir un conjunto de muestras de entrenamiento para cada árbol. El mismo conjunto de entrenamiento de entrada se usa para entrenar todos los árboles. (b) Los extra trees toman una división de nodo donde el índice de atributos y el valor de división del atributo se elige al azar. (c) Para un gran número de características ruidosas, los extra trees dan el peor rendimiento. (d) Para la selección óptima de funciones proporcionada, los extra trees se pueden calcular más rápido. Del análisis de sesgo / varianza, se puede concluir que, con un aumento de la aleatorización a un nivel óptimo, y hay una ligera disminución en varianza con un aumento significativo en el sesgo (Chandra, Bhateja, Mohanty, & Udgata, 2020).

2.2.5.4. Redes neuronales artificiales

Las Redes Neuronales Artificiales intentan imitar el comportamiento del cerebro humano, que se caracteriza por el aprendizaje a través de la experiencia y la extracción de conocimientos generales de un conjunto de datos. Estos sistemas imitan esquemáticamente la estructura neuronal del cerebro, ya sea mediante un programa informático o mediante el modelado a través de estructuras de procesamiento con cierta capacidad de computación paralela (Flóres & Fernández, 2008, pág. 11).

De acuerdo con Khepri (2018), las RNAs al margen de tener similitudes al cerebro humano, presentan una serie de características propias del cerebro. Por ejemplo, las RNAs aprenden de la experiencia, regionalizan de ejemplos previos a ejemplos nuevos y abstraen las características principales de una serie de datos.

Por otro lado, Villada, Muñoz & García (2012) las RNA son muy eficientes al momento de resolver problemas computacionales de clasificación y reconocimiento de patrones.

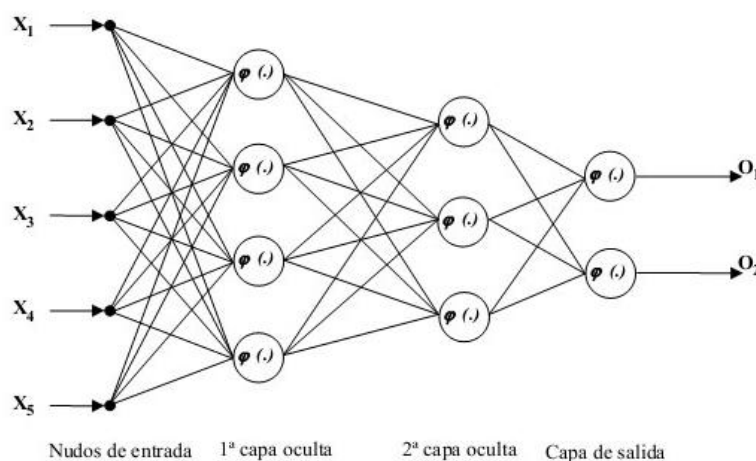


Figura 7: Red neuronal de propagación hacia adelante

Fuente: (Villada, Muñoz, & García, 2012)

2.2.5.5. KNN

El algoritmo KNN (K-Nearest Neighbours) es uno de los algoritmos de clasificación más básicos pero esenciales en el ámbito de Machine Learning. Dicho algoritmo pertenece al

aprendizaje supervisado y es aplicada para el reconocimiento de patrones (Sehra, K Nearest Neighbors Explained Easily, 2018), se podría decir que “el algoritmo hace predicciones calculando la similitud entre la muestra de entrada y cada instancia de entrenamiento” (Patil, 2018).

Por otro lado, Molina & García (2012) mencionan que el algoritmo KNN suele denominarse método porque es el marco básico de un algoritmo que admite el intercambio de la función de aproximación, que crea varias variantes. La función de aproximación puede decidir la clasificación de un nuevo ejemplo en función de la clasificación del ejemplo o la mayoría de los k ejemplos siguientes. También se admiten funciones de proximidad, que tienen en cuenta el peso o el costo de los atributos involucrados y, entre otras cosas, pueden eliminar atributos irrelevantes

Expresión matemática, basado en distancia bayesiana

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^n (x_{il} - x_{jl})^2}$$

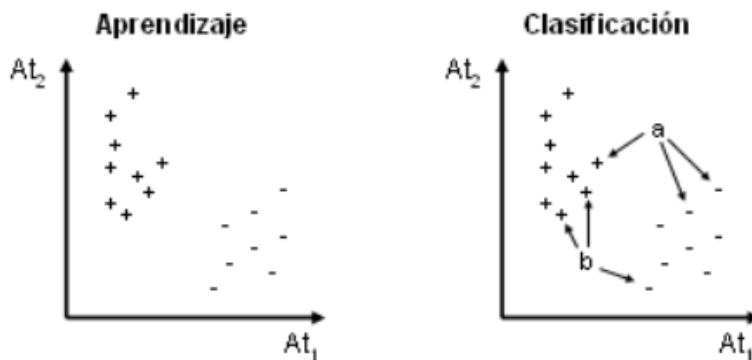


Figura 8: Ejemplo de Aprendizaje y Clasificación con KNN

Fuente: (Molina & García, 2012)

2.2.5.6. Regresión lineal

Es uno de los algoritmos de aprendizaje supervisado, “la regresión lineal permite identificar relaciones entre variables numéricas y crear modelos de regresión: 1 variable

salida y varias entradas numéricas. Se consideran relaciones de una variable de salida (dependiente) con varias variables de entrada (independientes)” (Molina & García, 2012).

Por otro lado, Astorga (2014) los modelos de regresión lineal se usan ampliamente en ingeniería porque se usan para analizar el comportamiento de las variables de entrada y de salida, para hacer predicciones y estimaciones.

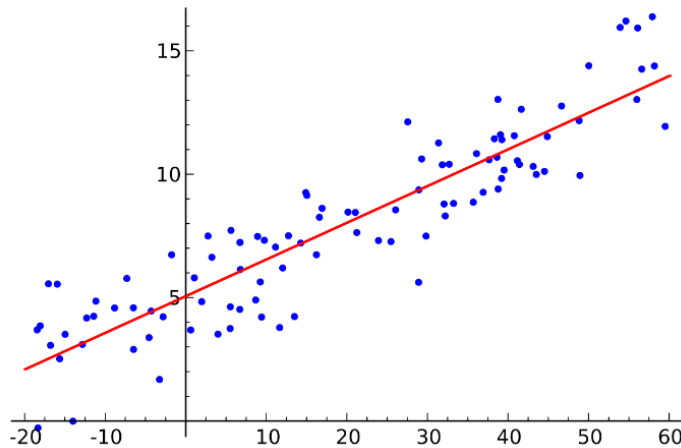


Figura 9: Ejemplo de línea con mejor ajuste de regresión lineal

Fuente: (Tanner, 2018)

2.2.5.7. K-Means

En el aprendizaje no supervisado uno de los algoritmos más utilizados para realizar clustering es el algoritmo k-means (k-medias), para poder utilizar este algoritmo en primera instancia se debe de especificar por adelantado la cantidad de clústers que se van a crear, este es el parámetro k, por tal motivo se seleccionan k elementos aleatorios, que representarán el centro o media de cada clúster (Molina & García, 2012).

Por otro lado, Singh (2018) menciona en su blog que “K-means intenta dividir x puntos de datos en el conjunto de k grupos donde cada punto de datos se asigna a su grupo más cercano. Este método está definido por la función objetivo que intenta minimizar la suma de todas las distancias al cuadrado dentro de un grupo, para todos los grupos”.

Por su parte Madushan (2017), afirma que “El algoritmo de agrupación de medios K se utiliza para buscar grupos que no se han etiquetado explícitamente en los datos y para encontrar patrones y tomar mejores decisiones”.

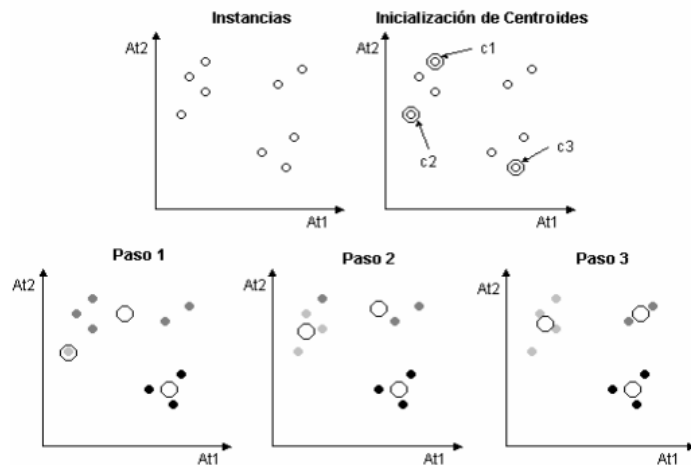


Figura 10: Ejemplo de clustering con k-medias.

Fuente: (Molina & García, 2012)

2.2.6. Metodologías de Minería de Datos y Ciencia de Datos

Las metodologías en la minería de datos y en la ciencia de datos son una serie de pasos para poder encontrar conocimientos. En la (figura 11) se observa una encuesta realizada en los años 2007 y 2014 en donde, se hace la comparación de los resultados obtenidos en los años ya antes mencionados. Donde en dichas encuestas podemos observar los resultados en donde la metodología que más se usa para aplicar minería de datos y ciencia de datos es la metodología CRISP-DM.

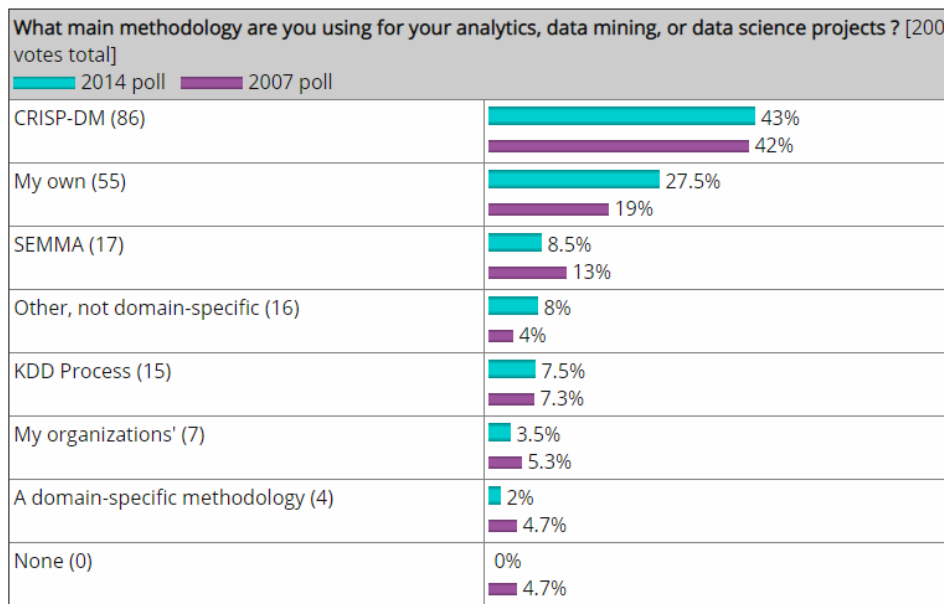


Figura 11: Resultado de encuestas de las metodologías más usadas

Fuente: (Piatetsky, 2014)

2.2.6.1. Metodología KDD

El descubrimiento de conocimiento en bases de datos (KDD) es el proceso de descubrir conocimiento útil de una colección de datos. Esta técnica de minería de datos ampliamente utilizada es un proceso que incluye preparación y selección de datos, limpieza de datos, incorporación de conocimientos previos sobre conjuntos de datos e interpretación de soluciones precisas a partir de los resultados observados. (Technopedia, 2019)

El proceso KDD ha alcanzado su punto máximo en los últimos 10 años. Ahora alberga muchos enfoques diferentes para el descubrimiento, que incluyen aprendizaje inductivo, estadísticas bayesianas, optimización de consultas semánticas, adquisición de conocimiento para sistemas expertos y teoría de la información. El objetivo final es extraer conocimiento de alto nivel de datos de bajo nivel. (Technopedia, 2019)

Según Madera (2014), el proceso KDD involucra las siguientes fases, que ayudan a descubrir conocimiento en bases de datos los cuales se definen a continuación:

- Limpieza e integración de los datos: En esta fase se procede a realizar la selección de fuentes de datos, como bases de datos y/o archivos de texto. A su vez se procede a eliminar los datos inconsistentes.
- Selección y transformación de los datos: En esta fase se procede a seleccionar los atributos (features), que será utilizados para el análisis y entrenamiento del modelo seleccionado.
- Minería de datos: La minería de datos es la parte medular del proceso KDD y su objetivo, como se mencionó anteriormente, es identificar y extraer patrones de comportamiento descriptivo y predictivos de grandes cantidades de datos.
- Evaluación de patrones y presentación del nuevo conocimiento: Es en esta fase del proceso donde se aplican distintas medidas, principalmente estadísticas para identificar los patrones más interesantes. También, se utilizan técnicas para visualizar los patrones descubiertos y así facilitar la interacción del usuario con el sistema.

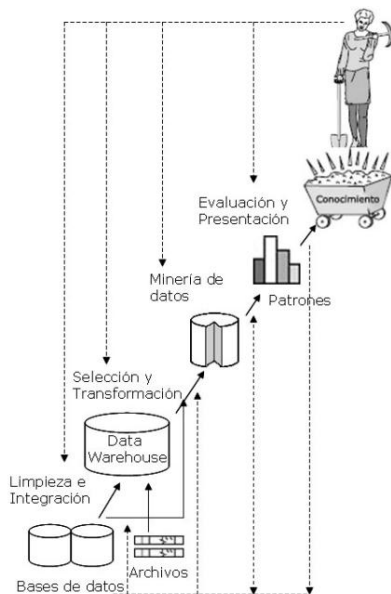


Figura 12: Fases de las etapas del proceso de descubrimiento del conocimiento en bases de datos. (KDD)

Fuente: (Han, Kamber, & Pei, 2011)

2.2.6.2. Metodología SEMMA

La metodología SEMMA es una de las tantas metodologías que existen. “La metodología SEMMA, abreviatura de Sample (muestreo), Explore (exploración), Modify (modificación), Model (modelado) y Asses (valoración) es también muy conocida y utilizada. Se puede definir como “el proceso de selección, exploración y modelado de grandes volúmenes de datos para descubrir patrones de negocio desconocidos”. Fue desarrollada por el SAS Institute (2005)”. (Joyanes, 2019, pág. 245)

En su libro Joyanes (2019), indica que el proceso de minería de datos aplicando la metodología SEMMA, se compone de las siguientes etapas:

- Muestreo. Genera una muestra representativa de datos. Se identifican los datos (Entrada de datos, ejemplos, partición de datos).
- Exploración. Visualización y descripción básica de los datos. Se exploran los conjuntos de datos para observar relaciones y patrones, y se generan análisis diversos, identificación de variables importantes y análisis de asociación (Exploración distribuida, múltiples particiones, intuición, asociación, selección de variables).
- Modificación. Selección de variables y transformación de la representación de variables. Se preparan los datos para el análisis (Transformación de variables, filtros a los datos fuera de rango, agrupamiento, ruido).
- Modelado. Utiliza técnicas diversas de estadística y modelos de aprendizaje automático (Regresión, arboles, redes neuronales, etc.).
- Evaluación (Valoración). Evalúa la precisión y utilidad de los modelos (Evaluación, medidas, reportes)



Figura 13: Etapas del proceso de la metodología SEMMA

Fuente: (Joyanes, 2019)

2.2.6.3. Metodología CRISP-DM

Este modelo es una secuencia idealizada de eventos. En la práctica, muchas de las tareas se pueden realizar en un orden diferente y, a menudo, será necesario retroceder a tareas anteriores y repetir ciertas acciones. El modelo no intenta capturar todas las rutas posibles a través del proceso de minería de datos (SMARTVISION, 2019); También podríamos decir que la metodología CRISP-DM, presta especial atención a la comprensión de las condiciones comerciales de los datos, puestos que esta “metodología CRISP-DM proporciona un enfoque estructurado para planificar un proyecto de minería de datos. Es una metodología robusta y bien probada” (SMARTVISION, 2019).

La metodología CrossIndustry Standard Processfor Data Mining (CRISP-DM) fue propuesto por un consorcio europeo en la segunda mitad de la década de 1990, sus fundadores fueron Daimler, Chrysler, SPSS y NCR, esta metodología se ha convertido en un método de minería de datos abierto y no patentado (Joyanes, 2019).

Joyanes (2019) Menciona que una de las ventajas de esta metodología es que fue diseñada y construida sobre la base de la experiencia real y no teóricamente por empresas de tecnología a su vez indica que la CRISP-DM funciona como metodología y como proceso. La metodología contiene descripciones de las fases normales de un proyecto y las tareas requeridas. Dicha metodología ofrece el siguiente ciclo vital de la minería de datos y son:

- Fase I. Comprensión del negocio: Esta primera fase se enfoca en comprender los objetivos del proyecto y definir las necesidades del cliente. Este conocimiento de

los datos se convierte en la definición de un problema de minería de datos y en un plan preliminar para lograr los objetivos. (Joyanes, 2019).

- Fase II. Comprensión de los datos: La fase de comprensión y estudio de los datos comienza con la recopilación y el aprendizaje de los datos, el reconocimiento preliminar de los datos y continúa con actividades que le permiten familiarizarse con los datos, identificar problemas de calidad y las primeras oportunidades para analizar y/o para descubrir subconjuntos, para formar hipótesis sobre información oculta (Joyanes, 2019).
- Fase III. Preparación de los datos: La fase de preparación de datos incluye todas las actividades que se requieren para crear el conjunto de datos final, a partir de los datos sin procesar. Las tareas incluyen seleccionar tablas, registros y atributos, y transformar y limpiar datos para herramientas de modelado. En resumen, el análisis de datos y la selección de las características se llevan a cabo en esta fase. (Joyanes, 2019).
- Fase IV. Modelado de datos: En esta fase se eligen las técnicas de modelado relevantes para el problema (cuanto más, mejor) se seleccionan y aplican, y sus parámetros se calibran para valores óptimos. Generalmente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requisitos especiales para la forma de los datos, por lo tanto, casi siempre regresa a la fase de preparación de datos en un proyecto. (Joyanes, 2019).
- Fase V. Evaluación: En esta fase del proyecto, se construyeron uno o más modelos que parecen ser de calidad suficiente desde la perspectiva del análisis de datos. Antes de proceder con el despliegue final del modelo, es importante evaluarlo a fondo, revisar los pasos para crearlo y comparar el modelo obtenido con los objetivos del negocio. Un objetivo importante es determinar si hay un tema

comercial importante que no se ha considerado lo suficiente. Al final de esta fase, se debe tomar una decisión sobre cómo aplicar los resultados del proceso de análisis de datos. El resultado final de esta fase es obtener resultados (Joyanes, 2019).

- Fase VI. Despliegue/Distribución o desarrollo (Implantación): En general, construir el modelo no es el final del proyecto. Incluso si el objetivo del modelo es mejorar el conocimiento de los datos, el conocimiento adquirido debe organizarse y presentarse de tal manera que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la creación de un informe o tan compleja como la implementación regular y posiblemente automatizada de un proceso de análisis de datos en la organización. El objetivo final de esta fase es la distribución o desarrollo (implementación) y la puesta en marcha (Joyanes, 2019).

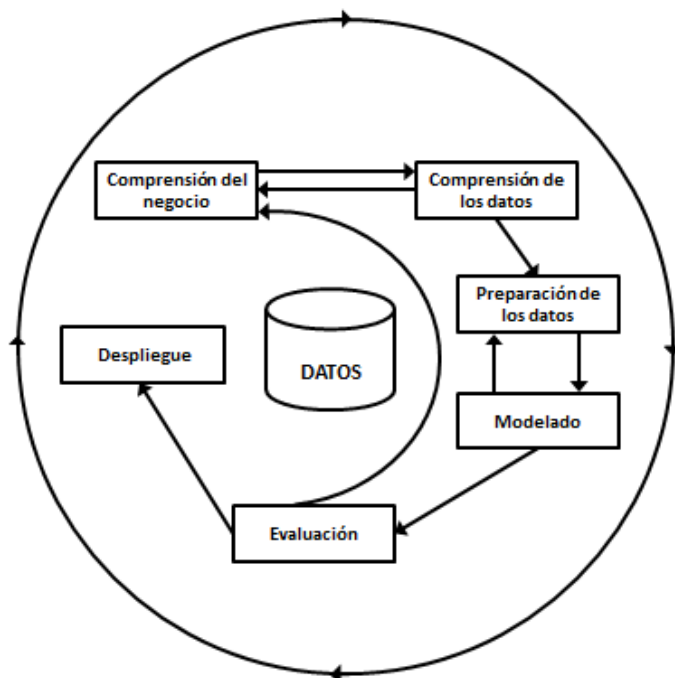


Figura 14: Fases del modelo de proceso de la metodología CRISP-DM

Fuente: (Peralta, 2014)

2.2.7. Machine Learning con Python

2.2.7.1. Python

Python es un lenguaje de programación interpretado que “fue creado por Guido van Rossum. Lo empezó a desarrollar a finales de 1989 y se considera hoy en día como el tercer lenguaje más ocupado por los desarrolladores. Además, es de código abierto lo que lo hace accesible a un mayor número de estudiantes, investigadores y desarrolladores” (Cervantes, Báez, Arízaga, & Castillo, 2017, pág. 30).

El lenguaje de programación Python aparte de ser un lenguaje muy expresivo, cuenta con muchas librerías que facilitan el proceso de análisis de información, tanto como para obtener datos, limpieza de datos, refinamiento, generación de modelos (Machine Learning) y visualización de datos, todo esto le hace una herramienta muy poderosa para poder realizar análisis de datos (Brey, 2018).

2.2.7.2. Scikit-Learn

Para poder relocalizar proyecto de aprendizaje automático hay una infinidad de bibliotecas, pero “podríamos decir que Scikit-Learn es la biblioteca de aprendizaje automático más popular para Python. Proporciona una gran cantidad de algoritmos de aprendizaje automático (clasificación, regresión y análisis de grupos) además de distintas técnicas de pre procesamiento y de evaluación de modelos” (Caballero, Martín, & Riesco, 2019, pág. 136).

2.2.7.3. Numpy

Numpy es una librería del lenguaje de programación Python que cuenta con un gran soporte para vectores y matrices.

Asimismo, Cuevas (2018) menciona en su libro que “NumPy, cuyo nombre proviene de “Numerical Python”, es una librería fundamental para el cálculo científico en Python, un

ámbito en el que los tipos estándar de Python serán insuficientes. Nos va a permitir trabajar con arrays multidimensionales de forma muy rápida y eficiente” (pág. 234).

Por otro lado, Caballero, Martín & Riesco (2019) mencionan que “NumPy permite realizar operaciones matemáticas del álgebra lineal o algoritmos más avanzados como la transformada de Fourier” (pág. 116).

A su vez Cuevas (2018) indica que NumPy en sus inicios fue parte del paquete científico SciPy, y en la actualidad es un paquete individual, que forma parte importante en el cálculo científico, dicho paquete está escrito en lenguaje de programación C.

2.2.7.4. Pandas

La biblioteca Pandas es muy utilizada en la actualidad junto al lenguaje de programación Python para poder realizar análisis de datos dicha biblioteca “Está construida sobre NumPy, y proporciona clases muy útiles para analizar datos como Series o DataFrame. Series permite representar una secuencia de valores utilizando un índice personalizado (enteros, cadenas de texto, etc.) para acceder a ellos. Por otro lado, DataFrame nos permite representar datos como si de una tabla o una hoja de cálculo se tratase” (Caballero, Martín, & Riesco, 2019).

2.2.7.5. Matplotlib

Con el lenguaje de programación Python tenemos distintas herramientas para poder crear gráficos pero “el estándar de facto para generar gráficos 2D (que también tiene una capacidad más que aceptable para representarlos en 3D) es Matplotlib, uno de los programas históricos dentro del ecosistema Python y usado ampliamente por su comunidad de usuarios, especialmente en el entorno científico debido a su capacidad para crear gráficos de gran calidad” (Cuevas, 2018, pág. 216).

A su vez Shetty (2018), en su blog menciona que la biblioteca Matplotlib emula a Matlab para poder realizar las visualizaciones de los gráficos; Matlab no es una herramienta

gratuita y por ende lo más recomendable es utilizar Matplotlib en Python, ya que es una biblioteca robusta, gratuita y fácil para la visualización de datos.

2.2.8. Rendimiento Académico

Según Reyes (2003), el rendimiento académico es un indicador del nivel de aprendizaje del estudiante, razón por la cual el sistema educativo otorga tanta importancia a este indicador. En este sentido, el rendimiento académico se convierte en una "tabla de medición imaginaria" para el aprendizaje en el aula, que es el objetivo central de la educación. Sin embargo, muchas otras variables fuera de la materia intervienen en el rendimiento académico, tales como: La calidad del maestro, el ambiente de enseñanza, la familia, el programa educativo, etc.

Por otro lado, Colonio (2017) menciona que “Es necesario tener en cuenta que el bajo rendimiento académico puede deberse a diferentes causas, como son la metodología de enseñanza empleada por el profesor, la falta de planificación y coordinación a la hora de encargar los trabajos de investigación, los problemas personales del estudiante y la situación del entorno familiar”.

2.2.8.1. Rendimiento académico de la Educación Básica Regular en el Perú.

El rendimiento constantemente se va midiendo mediante evaluaciones para poder ver cómo se van logrando el aprendizaje, y uno de las evaluaciones que se van realizando a nivel nacional es la Evaluación Censal Estudiantil (ECE), en donde se evalúa a todos los estudiantes del segundo grado de Educación Secundaria tanto de Instituciones Públicas y Privadas en las áreas de Matemática, Comunicación (Lectura) y Ciencia y Tecnología.

Según MINEDU - OFICINA DE MEDICIÓN DE LA CALIDAD DE LOS APRENDIZAJES (2020), en la evaluación ECE realizada en el año 2019, se obtuvo los siguientes resultados a nivel nacional, en matemáticas el 17,7% se encuentra en un nivel

satisfactorio, el 17,4% se encuentra en proceso, el 32,1% se encuentra en inicio y el 33% se encuentra en previo al inicio, en lectura el 14,5% se encuentra en nivel satisfactorio, el 25,8% se encuentra en proceso, el 42% se encuentra en inicio y el 17,7% se encuentra en previo al inicio, en Ciencia y tecnología el 9,7% se encuentra en un nivel satisfactorio, el 36,3 % se encuentra en proceso, el 43,8% se encuentra en inicio y el 10,1 se encuentra en previo al inicio.

Estos resultados que nos muestra la Oficina de Medición de la Calidad de los Aprendizajes, muestran un claro bajo rendimiento de los estudiantes del segundo grado de educación secundaria.

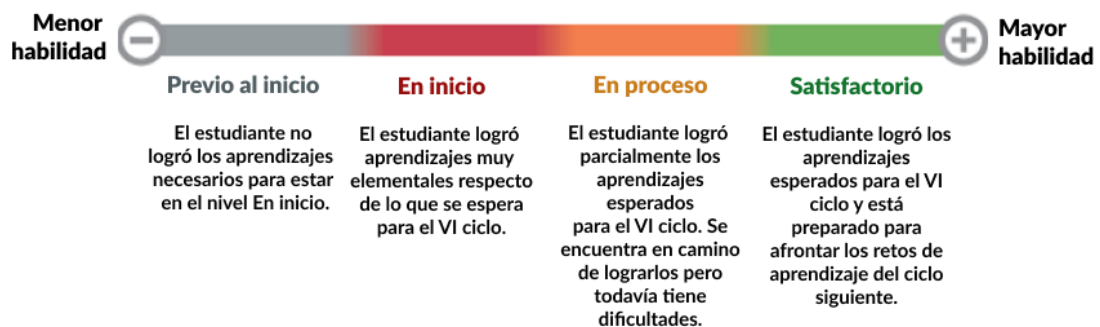


Figura 15: Interpretación de los resultados.

Fuente: (MINEDU - OFICINA DE MEDICIÓN DE LA CALIDAD DE LOS APRENDIZAJES, 2020)

Capítulo III

MATERIALES Y MÉTODOS

CAPÍTULO III. Materiales y Métodos

3.1. Lugar de Ejecución

El Presente trabajo de investigación se realizó en el Instituto Superior de Educación Pedagógico Público - Juliaca, ubicada en la Urbanización Pueblo Joven la Revolución en la Av. Infancia N° 303 del Distrito de San Miguel de la Provincia de San Román.

3.2. Población y tamaño de muestra

3.2.1. Población

La población para el presente estudio, estuvo constituida por los estudiantes matriculados desde el año 2013 - I hasta el año 2019 - II, del Instituto de Educación Superior Pedagógico Público Juliaca, que ascienden a 1242 registros.

Tabla 1.

Registros de proceso de admisión IESPPJ 2016 -2019

N°	Semestres	Total de ingresantes
1	2013-I	184
2	2014-I	234
3	2015-I	193
4	2016-I	100
5	2017-I	189
6	2018-I	193
7	2018-II	35
8	2019-I	124

Fuente: (Unidad Académica I.E.S.P.P.J., 2020)

3.2.1. Muestra

Dado que el presente estudio utilizará técnicas de Machine Learning para descubrir patrones en grandes cantidades de información, se optó por trabajar con toda la población.

3.3. Materiales e Insumos

Tabla 2.
Materiales e insumos

Herramientas	
Lenguaje de programación	Python
	Scikit-Learn
	Pandas
Librerías	NumPy
	Matplotlib
	Seaborn
Gestor de datos	Microsoft Excel
IDE	Anaconda
	Visual Studio
Metodología	CRISP-DM
Interfaz web	Framework Flask
	Framework Bootstrap

Fuente: Elaboración Propia

3.4. Metodología de la Investigación

3.4.1. Tipo de Investigación

El presente trabajo de investigación es de tipo tecnológico (Cegarra Sánchez, 2004) de propósito predictivo porque el modelo a desarrollar no solo explora, describe y explica, sino llega a predecir los comportamientos futuros de un objeto, fenómeno o hecho (Muñoz Rocha, 2015). En esta investigación se desarrollará un aplicativo web para identificar los patrones de bajo rendimiento académico en los ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca.

3.4.2. Arquitectura de Solución



Figura 16: Arquitectura de solución.

Fuente: Adaptado de (Joyanes, 2019)

3.4.2.1. Extracción

Es el proceso de identificación y recolección de datos de fuentes heterogeneas y homogéneas. Las herramientas de extracción que pueden ser utilizadas en este proceso soportan múltiples formatos de almacenamiento (Joyanes, 2019)

3.4.2.2. Almacenamiento

Es un repositorio de datos masivos que proporciona una visión global de un fácil acceso para almacenar datos. Data Lake es un tipo de almacenamiento en donde la información almacenada tiene una estructura variable, almacena los datos en su formato más básico, y están disponibles en todo momento, casi en tiempo real, este tipo de almacenamiento permite los análisis complejos y modelos predictivos (Joyanes, 2019).

3.4.2.3. Procesamiento

Procesamiento de los datos en bruto que fueron almacenados en el proceso de Data Lake mediante una herramienta básica de estadística o el uso de modelos predictivos que nos permiten identificar tendencias y comportamientos (Joyanes, 2019).

3.4.2.4. Visualización

La visualización es la presentación de los resultados de los análisis realizados, estos son representados mediante gráficos, diagramas, tableros de control, etc. Para facilitar la interpretación de los resultados, esto nos permite representar la información de la manera más intuitiva para poder consignar una comunicación simple, clara y efectiva (Joyanes, 2019).

3.5. Aplicación de la metodología CRISP-DM

En el presente trabajo de investigación, se utilizó la metodología Cross Industry Standard Process for Data Mining (CRISP-DM).

3.5.1. Fase 1: Comprensión del negocio

3.5.1.1. Contexto

El presente estudio se realiza en la oficina de Unidad Académica del Instituto de Educación Superior Pedagógico Público Juliaca, toda la información recaudada y los resultados obtenidos son de carácter académico.

3.5.1.2. Objetivos del negocio

(I.E.S.P.P.J., 2020) “Al 2023 ser una Escuela de Educación Superior Pedagógica líder, con docentes competitivos y exitosos, que ejercen la docencia con idoneidad, que fortalecen su profesionalidad para la transformación de la realidad educativa de la región.”

La Instituto de Educación Superior Pedagógico Publico Juliaca, busca transformar la realidad educativa de la región. Investigaciones que ayuden a lograr esta visión son de gran beneficio.

3.5.1.3. Producción del plan de proyecto

A continuación, se detalla las etapas del proyecto con el fin de una mejor organización y cumplimiento los objetivos del proyecto.

Primera etapa: Se realiza la solicitud de la base de datos histórica de los procesos académicos a la oficina de admisión y a la oficina de unidad Académica.

Segunda etapa: Análisis los datos recaudados.

Tercera etapa: Preparado de los datos: limpieza y transformación de los datos recaudados.

Cuarta etapa: Elección del modelo predictivo

Quinta etapa: Evaluación de los resultados del modelo predictivo seleccionado.

Sexta etapa: Implementación del modelo predictivo, mediante una interfaz web, utilizando el framework Flask.

3.5.1.4. Evaluación de herramientas

Tabla 3.
Herramientas empleadas para el Machine Learning

Herramientas	
Lenguaje de programación	
Librerías	
	
	
	
	
Gestor de datos	
IDE	
	
	
Interfaz web	
	
	

Fuente: Elaboración Propia

3.5.2. Fase 2: Comprensión de los datos

3.5.2.1 Recolección de data inicial

La recolección de datos se realizó en la oficina de Admisión y la oficina de Secretaría Académica del Instituto de Educación Superior Pedagógico Público – Juliaca.

De la oficina de admisión se extrajeron los datos que constan en su formulario de Proceso de Admisión (Anexo C); A su vez de la oficina de admisión se extrajeron los datos socio económico del estudiante (Anexo D).

La oficina de Secretaría Académica es encargada de administrar el Sistema de Información Académica – SIGES, del dicho sistema se extrajeron las notas de todos los estudiantes.

Figura 17: Reporte de notas del SIGES

Fuente: (Unidad Académica I.E.S.P.P.J., 2020)

3.5.2.2 Descripción de los datos

```
df_mate.dtypes
```

N°	int64
MATRICULA	int64
Sexo	object
Fecha de Nacimiento	datetime64[ns]
Edad	int64
...	
Créditos.10	float64
Opcional I / Seminario	float64
Créditos.11	float64
TOTAL DE CRÉDITOS	float64
PONDERADO	float64
Length: 82, dtype: object	

Figura 18: Descripción de la data inicial

Fuente: Elaboración Propia

3.5.2.3 Agrupamiento y selección de columnas

En esta etapa realizamos la selección de las columnas que servirán como variables predictivas del modelo, una vez realizado este proceso pasamos a realizar el consolidado de

todos los reportes (.xlsx), que la Unidad Académica del Instituto de Educación Superior Pedagógico Público – Juliaca nos proporción para realizar el modelo predictivo.

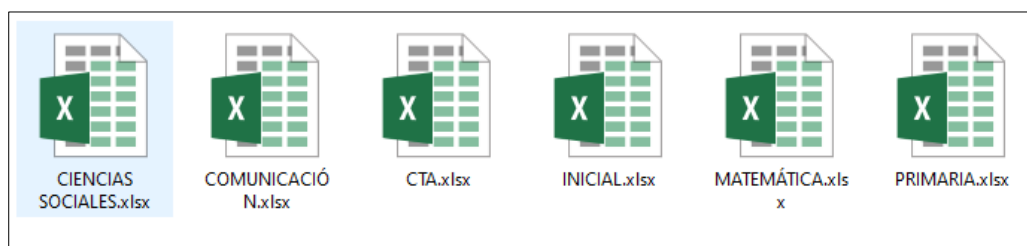


Figura 19: Consolidado de datos de estudiantes por carrera profesional

Fuente: (Unidad Académica I.E.S.P.P.J., 2020)

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Figura 20: Importación de librerías.

Fuente: Elaboración Propia

Para poder iniciar con el presente proceso, primeramente creamos un archivo (.ipynb). En la Figura 20 se observa la importación de librerías.

ESP. MATEMÁTICA

```
missing_values = ['NA', 'na', '--', '?', '-', 'None', 'none', 'non', 'nil']
df_mate = pd.read_excel('DATA/MATEMÁTICA.xlsx', sheet_name='MATEMÁTICA', na_values=missing_values)
print('df Shape:', df_mate.shape)
#df_mate.head()
df_mate.columns

df Shape: (114, 82)
Index(['Nº', 'MATRICULA', 'Sexo', 'Fecha de Nacimiento', 'Edad',
      'PROGRAMA DE ESTUDIOS / ESPECIALIDAD', 'PERIODO ACADÉMICO DE INGRESO',
      'Departamento', 'Provincia', 'Distrito',
      'Institución Educativa Secundaria', 'Tipo de Institución',
      'Departamento.1', 'Provincia.1', 'Distrito.1', 'Año de Inicio',
      'Año que culminó', 'Idioma Nativo', 'Modalidad de Ingreso',
      'Promedio final', 'PREG. 1', 'PREG. 2', 'PREG. 3.1', 'PREG. 3.2',
      'PREG. 4', 'PREG. 5', 'PREG. 6', 'PREG. 7', 'PREG. 8', 'PREG. 9',
      'PREG. 10', 'PREG. 11', 'PREG. 12', 'PREG. 13', 'PREG. 14', 'PREG. 15',
      'PREG. 16', 'PREG. 17', 'PREG. 18', 'PREG. 19', 'PREG. 20', 'PREG. 21',
      'PREG. 22', 'PREG. 23.1', 'PREG. 23.2', 'PREG. 23.3', 'PREG. 23.4',
      'PREG. 23.5', 'PREG. 23.6', 'PREG. 24.1', 'PREG. 24.2', 'PREG. 24.3',
      'PREG. 25', 'PERIODO ACADÉMICO', 'SEMESTRE ACADÉMICO', 'SECCIÓN',
      'Ciencias Sociales I', 'Créditos', 'Matemática I', 'Créditos.1',
      'Comunicación I', 'Créditos.2', 'Inglés I', 'Créditos.3',
      'Tecnologías de la Información y Comunicación I', 'Créditos.4',
      'Educación Física I', 'Créditos.5', 'Arte', 'Créditos.6',
      'Cultura Científico Ambiental I', 'Créditos.7',
      'Psicología I (General)', 'Créditos.8',
      'Desarrollo Vocacional y Tutoría I', 'Créditos.9', 'Práctica I',
      'Créditos.10', 'Opcional I / Seminario', 'Créditos.11',
      'TOTAL DE CRÉDITOS', 'PONDERADO'],
      dtype='object')
```

Figura 21: Leer un archivo (.xlsx),

Fuente: Elaboración Propia

En la Figura 21, se observa la codificación en Python para poder leer un documento (.xlsx), se inicia con el archivo de la especialidad de matemática.

```
#Se seleccionan las columnas que servirán como variables predictoras del modelo
df_mate_sel = df_mate[['MATRICULA','Sexo','Edad','Departamento.1','Provincia.1','Distrito.1',
'Año de Inicio','Año que culminó','Promedio final',
'PROGRAMA DE ESTUDIOS / ESPECIALIDAD','PERIODO ACADÉMICO DE INGRESO','Idioma Nativo',
'PREG. 1','PREG. 2','PREG. 3.1','PREG. 3.2','PREG. 4','PREG. 5','PREG. 6','PREG. 7','PREG. 8','PREG. 9','PREG. 10',
'PREG. 11','PREG. 12','PREG. 13','PREG. 14','PREG. 15','PREG. 16','PREG. 17','PREG. 18','PREG. 19','PREG. 20',
'PREG. 21','PREG. 22','PREG. 23.1','PREG. 23.2','PREG. 23.3','PREG. 23.4','PREG. 23.5','PREG. 23.6','PREG. 24.1',
'PONDERADO']]
print('df_mate_sel shape:',df_mate_sel.shape)
df_mate_sel.head()
```

df_mate_sel shape: (114, 46)

	MATRICULA	Sexo	Edad	Departamento.1	Provincia.1	Distrito.1	Año de Inicio	Año que culminó	Promedio final	PROGRAMA DE ESTUDIOS / ESPECIALIDAD	...	PREG. 23.2	PREG. 23.3	P
0	71552427	F	19	Puno	Azángaro	Achaya	2011.0	2015.0	11.12	EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA	...	4.0	5.0	
1	70810832	M	23	Puno	Azángaro	Arapa	2006.0	2010.0	17.12	EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA	...	NaN	NaN	
2	41819096	F	31	Puno	Azángaro	Azángaro	1998.0	2002.0	14.56	EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA	...	4.0	5.0	
3	80005432	M	38	Puno	Azángaro	Azángaro	1991.0	1995.0	12.48	EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA	...	NaN	NaN	
4	73736046	M	17	Puno	San Román	Juliaca	2011.0	2015.0	12.80	EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA	...	4.0	5.0	

5 rows x 46 columns

Figura 22: Seleccionar columnas y variables.

Fuente: Elaboración Propia

En la Figura 23 se observa la codificación en Python para poder seleccionar columnas, dichas columnas nos servirán como variables predictivas del modelo.

```
#Se procesa la columna 'PROGRAMA DE ESTUDIOS / ESPECIALIDAD';
#valores de 'EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA' para 'MATEMÁTICA' y
#valores de 'sin prog' para otro
df_mate_sel['PROGRAMA DE ESTUDIOS / ESPECIALIDAD'] = np.where(df_mate_sel['PROGRAMA DE ESTUDIOS / ESPECIALIDAD'] == 'EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA', 'MATEMÁTICA', 'sin prog')
```

Figura 23: Modificar valores de la columna programa de estudios.

Fuente: Elaboración Propia

La columna “PROGRAMA DE ESTUDIOS / ESPECIALIDAD” tiene siguiente valor: “EDUCACIÓN SECUNDARIA, ESPECIALIDAD: MATEMÁTICA”, dicho valor es un texto demasiado extenso, por tal motivo modificaremos el valor a “MATEMÁTICA” y los que no

cuentan con ningún valor se puso el valor de “sin prog”. En la Figura 23 se observa la codificación para poder realizar las modificaciones de la columna programa de estudios.

Los procesos vistos con anterioridad fueron aplicados a todos los reportes de la Figura 18, reportes que fueron proporcionados por la Unidad Académica del Instituto de Educación Superior Pedagógico Público – Juliaca.

Finalmente pasamos a consolidar toda la información en un solo archivo (.csv), con la función *concat()* de la librería Pandas.

```
#concat() nos permite agrupar toda la información.
df=pd.concat([df_mate_sel,df_comunic_sel, df_cta_sel,df_sociales_sel,df_primaria_sel,df_inicial_sel], axis=0)
print('df shape:',df.shape)
df.head()

df shape: (1222, 46)
```

Figura 24: Consolidar la información.

Fuente: Elaboración Propia

En la figura 24, se observa la codificación para poder consolidar nuestra data.

```
# Exportamos la información en un archivo .csv
df.to_csv(r'export_df.csv', index = False, header=True, encoding = "ISO-8859-1")
```

Figura 25: Exportar archivo .csv.

Fuente: Elaboración Propia

En la figura 25, se observa la codificación para exportar toda la data en un archivo (.csv).

3.5.2.4 Análisis exploración de data inicial

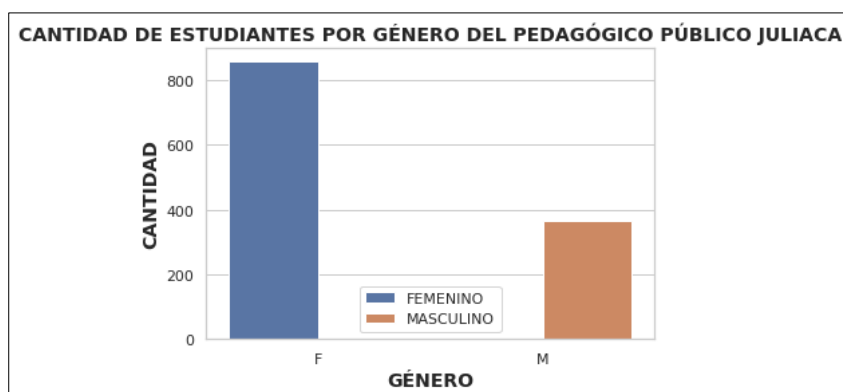


Figura 26: Cantidad de estudiantes según el género.

Fuente: Elaboración Propia

En la Figura 25, se observa la cantidad de estudiantes por género del Instituto de Educación Superior Pedagógico Público – Juliaca, en donde muestran un resultado de 858 estudiantes del género femenino y 364 estudiantes del género Masculino.

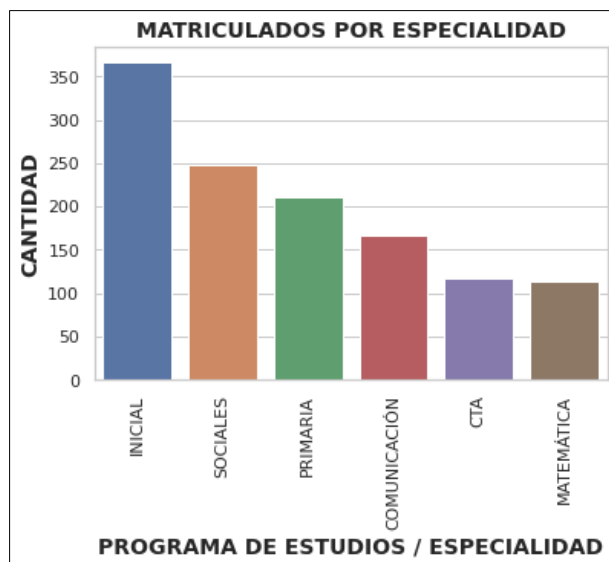


Figura 27: Cantidad de estudiantes matriculados por especialidad

Fuente: Elaboración Propia

En la Figura 27, se observa la cantidad de estudiantes matriculados en cada programa de estudios, en donde los resultados nos muestran que en la especialidad de comunicación hay un total de 166 estudiantes matriculados, en la especialidad de Ciencia, Tecnología y Ambiente hay un total de 118 estudiantes matriculados, en la especialidad de Educación Inicial hay un total de 366 estudiantes matriculados, en la especialidad de Matemática hay un total de 114 estudiantes matriculados, en la especialidad de Educación Primaria hay un total de 210 estudiantes matriculados y en la especialidad de Ciencias Sociales hay un total de 248 estudiantes matriculados. Según el grafico podemos observar que la especialidad con más demanda es la de Educación Inicial.

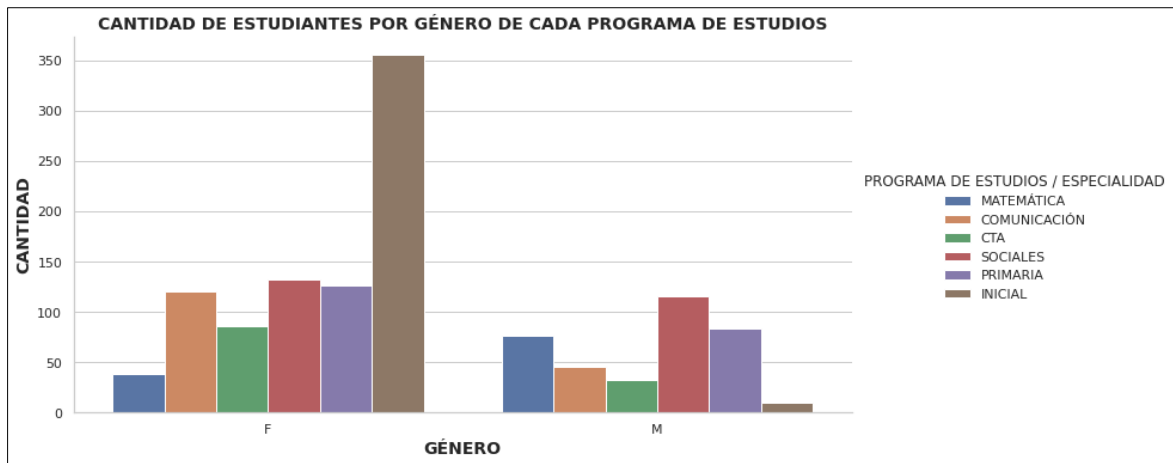


Figura 28: Cantidad de estudiantes matriculados por género en cada especialidad

Fuente: Elaboración Propia

En la Figura 28, se observa la cantidad de estudiantes matriculados en cada programa de estudios por género, en este grafico comparativo observamos que el Instituto de Educación Superior Pedagógico Público – Juliaca, alberga en mayoría a estudiantes del género femenino.

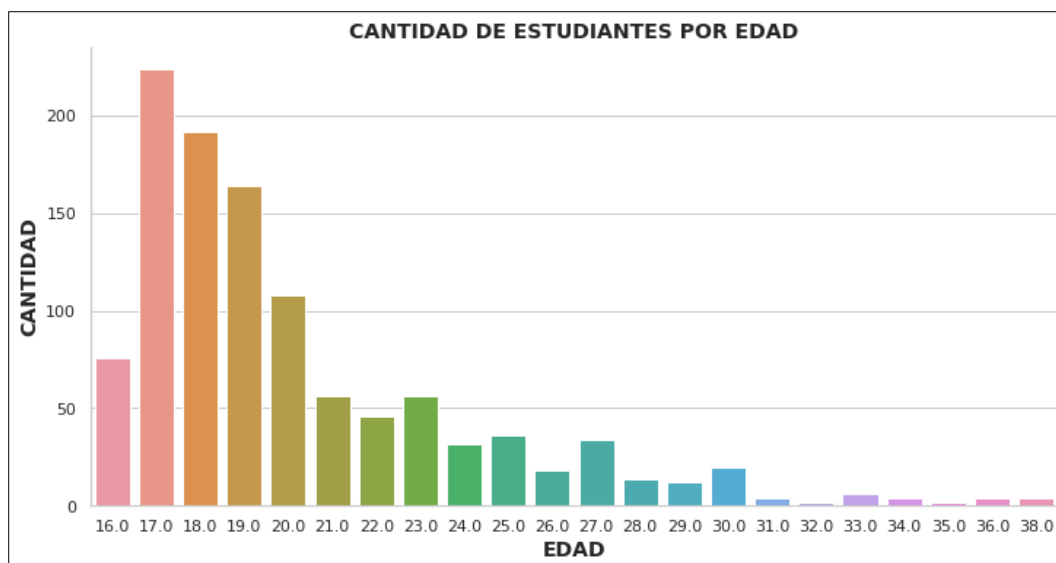


Figura 29: Ingresantes de estudiantes según edad.

Fuente: Elaboración Propia

En la Figura 29, se observa la cantidad de estudiantes ingresantes según edad, en donde se observa que en su gran mayoría los estudiantes ingresantes tienen 17 años de edad.

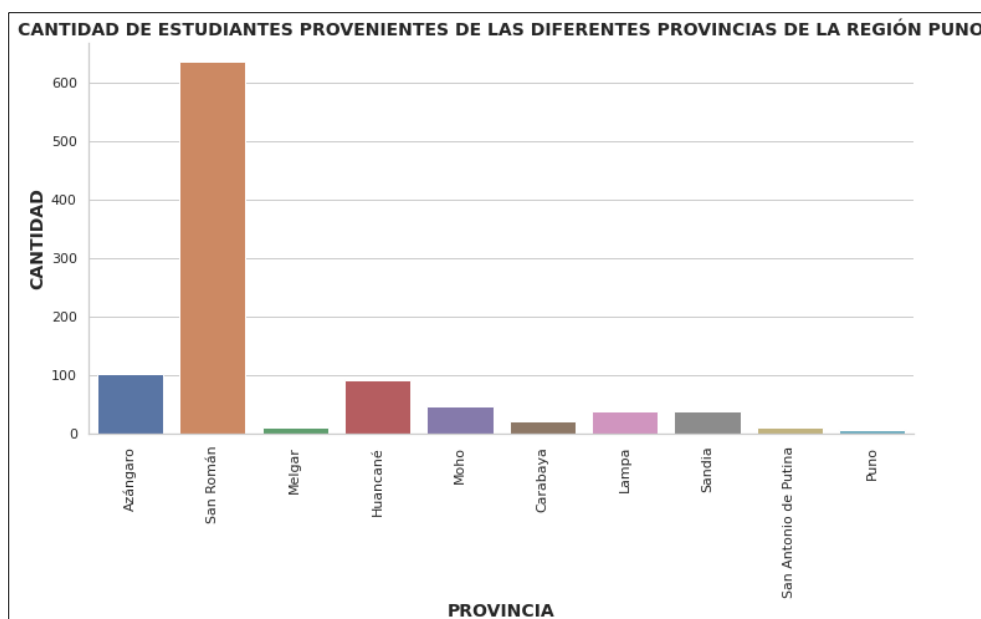


Figura 30: Estudiantes provenientes de las diferentes Provincias de la Región Puno.

Fuente: Elaboración Propia

En la Figura 30, se observa la cantidad de estudiantes ingresantes según las provincias de la Región Puno, en donde podemos observar que la gran mayoría de estudiantes son de la Provincia de San Román.

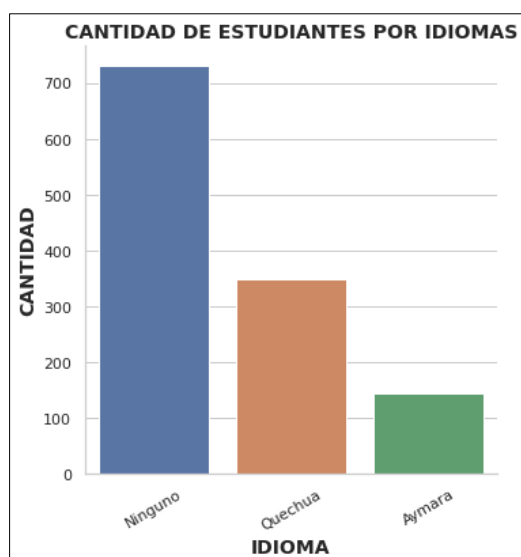


Figura 31: Cantidad de estudiantes según el idioma nativo que hablan.

Fuente: Elaboración Propia

En la Figura 31 se observa la cantidad de estudiantes ingresantes según el idioma nativo que hablan, en la imagen podemos observar que la gran mayoría de estudiantes no hablan ningún idioma nativo.

3.5.3 Fase 3: Preparación de los datos

3.5.3.1 Preparación de DataFrame

Para poder iniciar el entrenamiento de nuestro modelo debemos de hacer la limpieza de datos que no son de gran ayuda para el entrenamiento del modelo predictivo.

```
missing_values = ['NA', 'na', '--', '?', '-', 'None', 'none', 'non', 'nil']
df_mate = pd.read_csv('export_df.csv', na_values=missing_values, encoding = "ISO-8859-1")
print('df Shape:', df_mate.shape)
#df_mate.head()
df_mate1.columns
```

df Shape: (1222, 46)

Figura 32: Leer un archivo (.csv).

Fuente: Elaboración Propia

MATRICULA	int64
Sexo	object
Edad	float64
Departamento.1	object
Provincia.1	object
Distrito.1	object
Año de Inicio	float64
Año que culminó	float64
Promedio final	float64
PROGRAMA DE ESTUDIOS / ESPECIALIDAD	object
PERIODO ACADÉMICO DE INGRESO	object
Idioma Nativo	object
PREG. 1	float64
PREG. 2	float64
PREG. 3.1	float64
PREG. 3.2	float64
PREG. 4	float64
PREG. 5	float64
PREG. 6	float64
PREG. 7	float64
PREG. 8	float64
PREG. 9	float64
PREG. 10	float64
PREG. 11	float64
PREG. 12	float64
PREG. 13	float64
PREG. 14	float64
PREG. 15	float64
PREG. 16	float64
PREG. 17	float64
PREG. 18	float64
PREG. 19	float64
PREG. 20	float64
PREG. 21	float64
PREG. 22	float64
PREG. 23.1	float64
PREG. 23.2	float64
PREG. 23.3	float64
PREG. 23.4	float64
PREG. 23.5	float64
PREG. 23.6	float64
PREG. 24.1	float64
PREG. 24.2	float64
PREG. 24.3	float64
PREG. 25	float64
PONDERADO	float64
dtype: object	

Figura 33: Tipo de datos del DataFrame

Fuente: Elaboración Propia

En la Figura 33, se observa 7 campos de tipo object, estos campos tienen que ser modificados por un tipo int64 o float64.

```
df_mate2=df_mate

def substitute_idioma(a2):
    mapping={'Quechua':1, 'Aymara':2, 'Ninguno':3}
    a2['idioma_i']=a2['Idioma Nativo'].map(mapping)
    return a2
df_mate2.pipe(substitute_idioma)
print(pd.value_counts(df_mate2['idioma_i'], sort = True))

3      365
1      174
2        72
Name: idioma_i, dtype: int64
```

Figura 34: Sustituir el valor texto por un valor numérico en la columna idioma nativo

Fuente: Elaboración Propia

En la Figura 34, se observa la codificación para sustituir los textos por un número, en donde al idioma nativo Quechua le damos el valor 1, al idioma nativo Aymara le damos el valor 2 y al resto que no habla ningún idioma nativo le damos el valor 3.

Tabla 4.

Cantidad de estudiantes que hablan cada idioma nativo

Nº	Idioma Nativo	Cantidad
1	Quechua	348
2	Aymara	144
3	Ninguno	730

Fuente: Elaboración Propia

```
print(pd.value_counts(df_mate2['Departamento.1'], sort = True))

Puno      1004
Arequipa    16
Lima        14
Junín        2
Cusco        2
Name: Departamento.1, dtype: int64

#Se filtra solo puno
df_mate3 = df_mate2[df_mate2['Departamento.1']=='Puno']
print(pd.value_counts(df_mate3['Departamento.1'], sort = True))
print('df_mate3 Shape:',df_mate3.shape)

Puno      1004
Name: Departamento.1, dtype: int64
df_mate3 Shape: (1004, 47)
```

Figura 35: Seleccionar estudiantes del departamento de Puno

Fuente: Elaboración Propia

En la Figura 35, se observa la codificación para seleccionar solo los estudiantes que pertenecen en al departamento de Puno.

```
print(pd.value_counts(df_mate3['Provincia.1'], sort = True))
```

San Román	636
Azángaro	102
Huancané	92
Moho	48
Sandia	38
Lampa	38
Carabaya	22
Melgar	12
San Antonio de Putina	10
Puno	6

```
Name: Provincia.1, dtype: int64
```

```
#Se selecciona solo las provincias con más muestras
df_mate4 = df_mate3[np.logical_or(df_mate3['Provincia.1']=='San Román',
    np.logical_or(df_mate3['Provincia.1']=='Azángaro',
    np.logical_or(df_mate3['Provincia.1']=='Huancané',
    np.logical_or(df_mate3['Provincia.1']=='Moho',
    np.logical_or(df_mate3['Provincia.1']=='Lampa',
    np.logical_or(df_mate3['Provincia.1']=='Carabaya',
    df_mate3['Provincia.1']=='Sandia'))))))))
print(pd.value_counts(df_mate4['Provincia.1'], sort = True))
print('df_mate4 Shape:',df_mate4.shape)
```

San Román	636
Azángaro	102
Huancané	92
Moho	48
Sandia	38
Lampa	38
Carabaya	22

```
Name: Provincia.1, dtype: int64
df_mate4 Shape: (976, 47)
```

```
def substitute_prov(a2):
    mapping={'San Román':1, 'Azángaro':2, 'Huancané':3, 'Moho':4, 'Sandia':5, 'Lampa':6, 'Carabaya':7}
    a2['prov_i']=a2['Provincia.1'].map(mapping)
    return a2
df_mate4.pipe(substitute_prov)
print(pd.value_counts(df_mate4['prov_i'], sort = True))
```

1	636
2	102
3	92
4	48
6	38
5	38
7	22

```
Name: prov i, dtype: int64
```

Figura 36: Sustituir valor texto por un valor numérico en la columna provincia

Fuente: Elaboración Propia

En la Figura 36, se observa la codificación para seleccionar las provincias que pertenecen al departamento de Puno, luego seleccionamos a las provincias que tienen una mayor cantidad de muestra, posteriormente modificamos el valor textual por un valor numérico. Para la cada provincia se asignó un valor de acuerdo a la tabla N° 06.

Tabla 5.

Valor asignado a cada provincia

Provincia	Valor asignado
San Román	1
Azángaro	2
Huancané	3
Moho	4
Sandia	5
Lampa	6
Carabaya	7

Fuente: Elaboración Propia

```
def substitute_dep(a2):
    mapping={'Puno':1}
    a2['dep_i']=a2['Departamento.1'].map(mapping)
    return a2
df_mate4.pipe(substitute_dep)
print(pd.value_counts(df_mate4['dep_i'], sort = True))

1    976
Name: dep_i, dtype: int64
```

Figura 37: Sustituir valor texto por un valor numérico en la columna departamento

Fuente: Elaboración Propia

En la Figura 37, se observa la codificación para modificar el valor textual por un valor numérico de la columna departamento. Para el departamento de Puno se le asignó el valor 1.

```
print(pd.value_counts(df_mate4['Sexo'], sort = True))

F    666
M    310
Name: Sexo, dtype: int64

def substitute_sexo(a2):
    mapping={'F':1, 'M':2}
    a2['sexo_i']=a2['Sexo'].map(mapping)
    return a2
df_mate4.pipe(substitute_sexo)
print(pd.value_counts(df_mate4['sexo_i'], sort = True))

1    666
2    310
Name: sexo_i, dtype: int64
```

Figura 38: Sustituir valor texto por un valor numérico en la columna sexo

Fuente: Elaboración Propia

En la Figura 38, se observa la codificación para modificar el valor textual por un valor numérico de la columna sexo. Para el sexo Femenino (F) se asignó el valor 1 y para el sexo Masculino (M) se asignó el valor 2.

```

print(pd.value_counts(df_mate4['PROGRAMA DE ESTUDIOS / ESPECIALIDAD'], sort = True))
INICIAL      252
SOCIALES     202
PRIMARIA     174
COMUNICACIÓN 146
CTA          108
MATEMÁTICA   94
Name: PROGRAMA DE ESTUDIOS / ESPECIALIDAD, dtype: int64

def substitute_prog_estudios(a2):
    mapping={'INICIAL':1, 'SOCIALES':2, 'PRIMARIA':3, 'COMUNICACIÓN':4, 'MATEMÁTICA':5, 'CTA':6}
    a2['prog_estud_i']=a2['PROGRAMA DE ESTUDIOS / ESPECIALIDAD'].map(mapping)
    return a2
df_mate4.pipe(substitute_prog_estudios)
print(pd.value_counts(df_mate4['prog_estud_i'], sort = True))
1      252
2      202
3      174
4      146
6      108
5       94
Name: prog_estud_i, dtype: int64

```

Figura 39: Sustituir valor texto por un valor numérico en la columna programa de estudios

Fuente: Elaboración Propia

En la Figura 39, se observa la codificación para modificar el valor textual por un valor numérico de la columna Programa de Estudios. Para cada especialidad se asignó un valor de acuerdo a la tabla N° 07.

Tabla 6.
Valor asignado para cada programa de estudios

Programa de estudios	Valor asignado
Inicial	1
Sociales	2
Primaria	3
Comunicación	4
Matemática	5
CTA	6

Fuente: Elaboración Propia

```

print(pd.value_counts(df_mate4['distritox'], sort = True))
print('df_mate3 Shape:',df_mate4['distritox'].shape)

Juliaca      484
Otros        158
San Miguel   152
Moho         42
Taraco       32
Chupa        32
Sandia       26
Azángaro     26
Cojata       24
Name: distritox, dtype: int64
df_mate3 Shape: (976,)

def substitute_dist(a2):
    mapping={'Juliaca':1, 'San Miguel':2, 'Otros':3, 'Moho':4, 'Taraco':5,
             'Azángaro':6, 'Sandia':7, 'Cojata':8}
    a2['distrito_i']=a2['distritox'].map(mapping)
    return a2
df_mate4.pipe(substitute_dist)
print(pd.value_counts(df_mate4['distrito_i'], sort = True))

1.0      484
3.0      158
2.0      152
4.0       42
5.0       32
7.0       26
6.0       26
8.0       24
Name: distrito i, dtype: int64

```

Figura 40: Codificación para sustituir texto por un número de la columna distrito

Fuente: Elaboración Propia

En la Figura 40, se observa la codificación para modificar el valor textual por un valor numérico de la columna distrito. La asignación de valores se realizó de acuerdo a la tabla N° 09.

Tabla 7.
Valor asignado para cada distrito

Distrito	Valor asignado
Juliaca	1
San Miguel	2
Otros	3
Moho	4
Taraco	5
Azángaro	6
Sandia	7
Cojata	8

Fuente: Elaboración Propia

```
df_mate4["cant_anios_secund"] = df_mate4["Año que culminó"] - df_mate4["Año de Inicio"] + 1
```

Figura 41: Codificación para hallar la cantidad de años en la que curso la educación secundaria

Fuente: Elaboración Propia

En la Figura 40, se observa la codificación para hallar la cantidad de años en la que un estudiante ha cursado la Educación Secundaria.

MATRICULA	int64
sexo_i	int64
Edad	float64
dep_i	int64
prov_i	int64
distrito_i	int64
cant_anios_secund	float64
Promedio final	float64
prog_estud_i	int64
period_acad_i	int64
idioma_i	int64
PREG. 1	float64
PREG. 2	float64
PREG. 3.1	float64
PREG. 3.2	float64
PREG. 4	float64
PREG. 5	float64
PREG. 6	float64
PREG. 7	float64
PREG. 8	float64
PREG. 9	float64
PREG. 10	float64
PREG. 11	float64
PREG. 12	float64
PREG. 15	float64
PREG. 16	float64
PREG. 17	float64
PREG. 18	float64
PREG. 19	float64
PREG. 20	float64
PREG. 21	float64
PREG. 22	float64
PREG. 23.1	float64
PREG. 23.2	float64
PREG. 24.1	float64
PREG. 25	float64
PONDERADO	float64
dtype: object	

Figura 42: Tipo de datos del DataFrame

Fuente: Elaboración Propia

En la Figura 42, se observa que los 7 campos de tipo object que se visualizaron en la Figura 32, fueron modificados por el tipo int64.

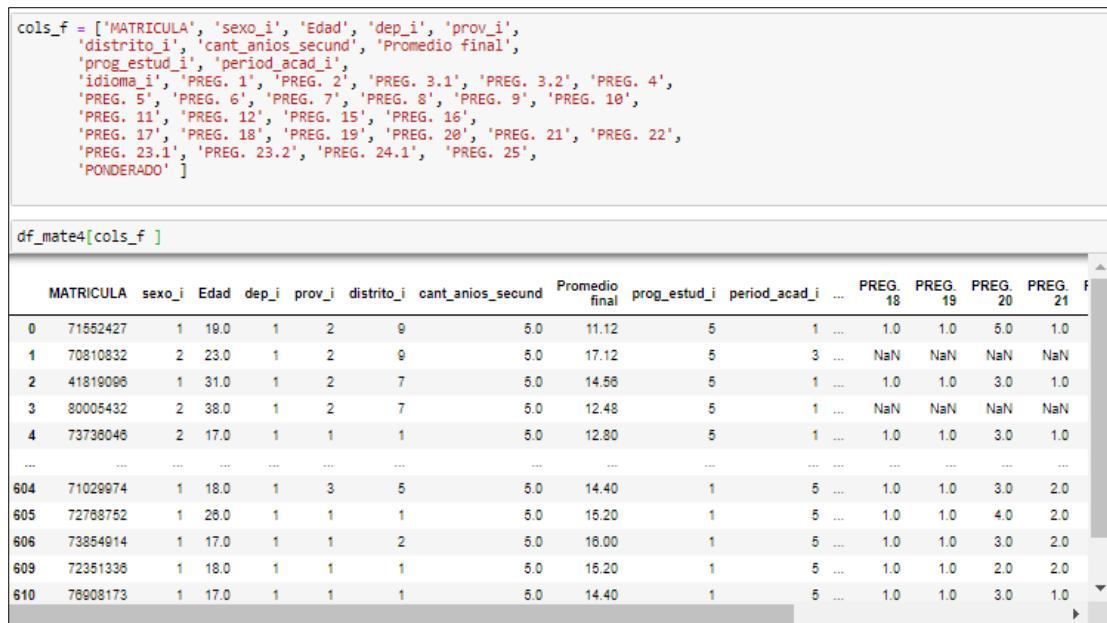


Figura 43: Datos del DataFrame de cada columna

Fuente: Elaboración Propia

En la Figura 43, se observa que toda nuestra data es de tipo int64 y float64.

3.5.3.2 Controlando valores nulos

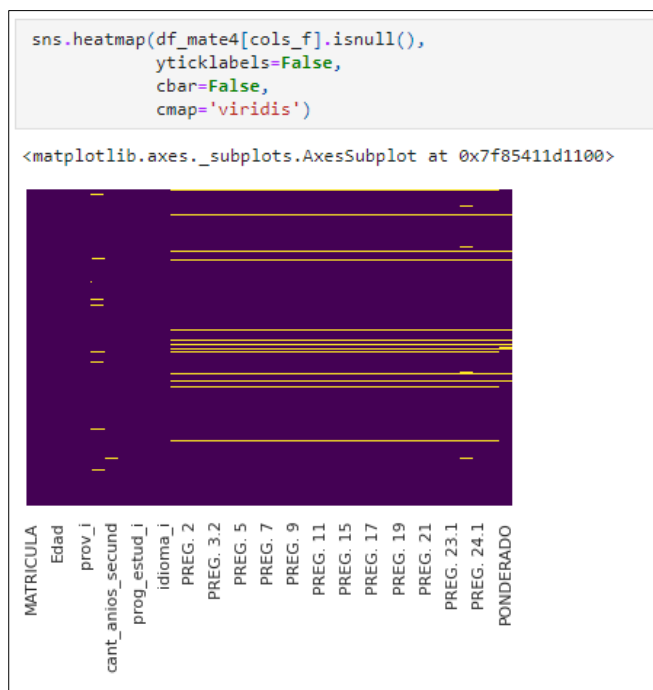


Figura 44: Valores nulos

Fuente: Elaboración Propia

En la Figura 44 se observa un gráfico en donde nos muestra todos los registros que cuentan con valores nulos o vacíos.

```
#porcentaje de valores null por columna
(df_mate4[cols_f].isnull().sum() / len(df_mate4[cols_f]))*100
```

MATRICULA	0.000000
sexo_i	0.000000
Edad	0.000000
dep_i	0.000000
prov_i	0.000000
distrito_i	3.278689
cant_anios_secund	0.409836
Promedio final	0.204918
prog_estud_i	0.000000
period_acad_i	0.000000
idioma_i	0.000000
PREG. 1	6.557377
PREG. 2	6.557377
PREG. 3.1	6.557377
PREG. 3.2	6.557377
PREG. 4	6.557377
PREG. 5	6.557377
PREG. 6	6.557377
PREG. 7	6.557377
PREG. 8	6.557377
PREG. 9	6.557377
PREG. 10	6.557377
PREG. 11	6.557377
PREG. 12	6.557377
PREG. 15	6.557377
PREG. 16	6.557377
PREG. 17	6.557377
PREG. 18	6.557377
PREG. 19	6.557377
PREG. 20	6.557377
PREG. 21	6.557377
PREG. 22	6.557377
PREG. 23.1	6.762295
PREG. 23.2	8.401639
PREG. 24.1	6.557377
PREG. 25	6.557377
PONDERADO	4.918033

dtype: float64

Figura 45: Porcentaje de valores nulos por columna

Fuente: Elaboración Propia

En la Figura 45, se observa los porcentajes de valores nulos por cada columna.



Figura 46: Eliminar valores nulos

Fuente: Elaboración Propia

En la Figura 46, se observa la codificación para eliminar valores nulos y el grafico nos muestra que en nuestra data actual no se encuentran valores nulos o vacíos.

```
#porcentaje de valores null por columna
(df_2[cols_f].isnull().sum() / len(df_2[cols_f]))*100
```

MATRICULA	0.0
sexo_i	0.0
Edad	0.0
dep_i	0.0
prov_i	0.0
distrito_i	0.0
cant_anios_secund	0.0
Promedio final	0.0
prog_estud_i	0.0
period_acad_i	0.0
idioma_i	0.0
PREG. 1	0.0
PREG. 2	0.0
PREG. 3.1	0.0
PREG. 3.2	0.0
PREG. 4	0.0
PREG. 5	0.0
PREG. 6	0.0
PREG. 7	0.0
PREG. 8	0.0
PREG. 9	0.0
PREG. 10	0.0
PREG. 11	0.0
PREG. 12	0.0
PREG. 15	0.0
PREG. 16	0.0
PREG. 17	0.0
PREG. 18	0.0
PREG. 19	0.0
PREG. 20	0.0
PREG. 21	0.0
PREG. 22	0.0
PREG. 23.1	0.0
PREG. 23.2	0.0
PREG. 24.1	0.0
PREG. 25	0.0
PONDERADO	0.0

dtype: float64

Figura 47: Porcentaje de valores nulos por columna

Fuente: Elaboración Propia

En la Figura 47, se observa los porcentajes de valores nulos por cada columna, a comparación de la Figura 44, nuestro porcentaje de datos nulos en la presente figura es del 0%.

3.5.3.3 Estructuración de los datos

```
#Se crea la columna 'TARGET' con las clases
df_2['TARGET'] = np.where(df_2['PONDERADO'] <= 10.49, 1, np.where((df_2['PONDERADO'] >= 10.50) & (df_2['PONDERADO']
np.where((df_2['PONDERADO'] >= 13.50) & (df_2['PONDERADO'] <=
print(pd.value_counts(df_2['TARGET'], sort = True))
```

3	494
2	242
1	124

Name: TARGET, dtype: int64

Figura 48: Creación de la columna TARGET

Fuente: Elaboración Propia

En la Figura 48, se observa la codificación para la creación de la columna TARGET que nos permitirá a realizar la predicción de nuestro modelo, dicha clasificación fue realizada según los indicadores de logros académicos del MINEDU.

Tabla 8.
Escala de evaluación de los aprendizajes

Clase	Valores	Indicador	Descripción
1	0-10	Previo al inicio	El estudiante no logró los aprendizajes necesarios para estar en el nivel En Inicio .
2	11-13	En inicio	El estudiante logró aprendizajes muy elementales respecto de lo que se espera para el VI ciclo.
3	14-17	En Proceso	El estudiante logró parcialmente los aprendizajes esperados. Se encuentra en camino de lograr, pero todavía tiene dificultades.
4	18-20	Satisfactorio	El estudiante logró los aprendizajes esperados y está preparado para afrontar los retos de aprendizaje del ciclo siguiente.

Fuente: Adaptado de (MINEDU - OFICINA DE MEDICIÓN DE LA CALIDAD DE LOS APRENDIZAJES, 2020)

df_2																
trito_i	cant_anios_secund	Promedio final	prog_estud_i	period_acad_i	...	PREG. 19	PREG. 20	PREG. 21	PREG. 22	PREG. 23.1	PREG. 23.2	PREG. 24.1	PREG. 25	PONDERADO	TARGET	
9	5.0	11	5	1	...	1.0	5.0	1.0	2.0	1.0	4.0	1.0	0.0	13	2	
7	5.0	15	5	1	...	1.0	3.0	1.0	1.0	1.0	4.0	1.0	0.0	15	3	
1	5.0	13	5	1	...	1.0	3.0	1.0	2.0	1.0	4.0	1.0	0.0	16	3	
1	5.0	12	5	1	...	1.0	5.0	1.0	2.0	1.0	4.0	1.0	0.0	14	3	
1	5.0	12	5	1	...	1.0	3.0	2.0	2.0	4.0	5.0	1.0	0.0	14	3	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1	5.0	15	1	5	...	1.0	4.0	2.0	1.0	1.0	4.0	1.0	0.0	16	3	
2	5.0	16	1	5	...	1.0	3.0	2.0	1.0	1.0	4.0	1.0	0.0	15	3	
13	6.0	14	1	5	...	1.0	3.0	1.0	1.0	1.0	4.0	1.0	0.0	13	2	
1	5.0	15	1	5	...	1.0	2.0	2.0	1.0	1.0	4.0	1.0	0.0	15	3	

Figura 49: Datos del DataFrame con la columna TARGET

Fuente: Elaboración Propia

En la Figura 49 se observa la nueva columna TARGET de nuestro DataFrame.

```
df_2[cols_f].corr()['TARGET'].sort_values(ascending=False)
```

TARGET	1.000000
PONDERADO	0.783792
MATRICULA	0.181873
PREG. 3.2	0.127162
period_acad_i	0.116141
Promedio final	0.091311
PREG. 5	0.084609
PREG. 4	0.082280
PREG. 25	0.063509
PREG. 18	0.056563
PREG. 24.1	0.040321
PREG. 19	0.038258
PREG. 23.2	0.037401
prov_i	0.033177
PREG. 21	0.033057
PREG. 11	0.028581
PREG. 23.1	0.021361
PREG. 20	0.010421
PREG. 12	0.009533
PREG. 22	-0.006520
cant_anios_secund	-0.016247
distrito_i	-0.026091
PREG. 15	-0.026303
PREG. 1	-0.031472
idioma_i	-0.042946
prog_estud_i	-0.043018
PREG. 17	-0.043368
PREG. 2	-0.046522
sexo_i	-0.055806
PREG. 7	-0.056496
PREG. 3.1	-0.074223
PREG. 16	-0.097519
PREG. 8	-0.101614
PREG. 6	-0.105136
Edad	-0.113379
PREG. 10	-0.146840

Name: TARGET, dtype: float64

Figura 50: Correlación de las columnas con la columna TARGET

Fuente: Elaboración Propia

En la Figura 50, se observa la correlación de las columnas del DataFrame con la columna TARGET.

3.5.3.4 Identificación de variables

VARIABLES
<pre>cols_f2 = ['sexo_i', 'Edad', 'distrito_i', 'Promedio final', 'prog_estud_i', 'cant_anios_secund', 'idioma_i', 'PREG. 2', 'PREG. 3.2', 'PREG. 4', 'PREG. 10', 'PREG. 16', 'PREG. 20', 'PREG. 22', 'TARGET']</pre>

Figura 51: Features seleccionados

Fuente: Elaboración Propia

En la Figura 50, se observa las variables (features) elegidas para el entrenamiento de nuestro modelo.

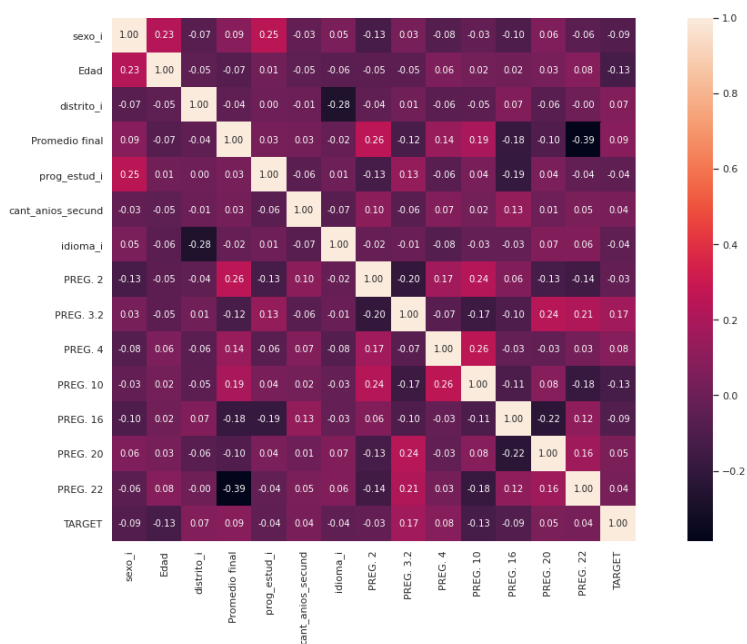


Figura 52: Correlación de Pearson

Fuente: Elaboración Propia

En la Figura 52, se observa la correlación de PEARSON con todos los features seleccionados.

3.5.3.5 Estructura del DataSet

Tabla 9.
Estructura del DataSet

Atributo	Descripción	Valores
sexo_i	Género del postulante	1: Femenino 2: Masculino
Edad	Edad del postulante	Valores numéricos
distrito_i	Ubicación del I.E.S. de procedencia.	1: Juliaca 2: San Miguel 3: Huancané 4: Taraco 5: Cojata 6: Moho 7: Azángaro 8: Chupa 9: Arapa 10: José Domingo Choquehuanca 11: Sandia

Promedio final	Promedio de examen de admisión.	12: Lampa 13: Macusani Valores numéricos
prog_estud_i	Programa de estudios del postulante.	1: Inicial 2: Ciencias Sociales 3: Primaria 4: Comunicación 5: Matemática 6: Ciencia Tecnología y Ambiente
cant_anios_secund	Cantidad de años en la que cursó la educación secundaria.	Valores numéricos
idioma_i	Idioma nativo que habla el postulante.	1: Quechua 2: Aymara 3: Ninguno
PREG. 2	¿La persona que mantiene su hogar es?	1: Tú mismo(a) 2: Mamá. 3: Papá. 4: Papá y Mamá. 5: Otros.
PREG. 3.2	Número de veces que postulaste a otros institutos/universidades:	Valores numéricos
PREG. 4	Tipo de preparación que recibiste para postular al IESP	1: Centro Pre 2: Centro Pre de otro IESP. 3: Academia. 4: Otros.
PREG. 10	Número de horas diarias que actualmente dedicadas al estudio	Valores numéricos
PREG. 16	¿Cada cuánto tiempo recibes ayuda económica?	1: Quincenal 2: Mensual 3: Una vez al año 4: Otros.
PREG. 20	Número de dormitorios de su vivienda	Valores numéricos
PREG. 22	Tipo de material de la vivienda	1: Cemento / Ladrillo 2: Adobe 3: Quincha 4: Esteras 5: Otro.
TARGET	Variable a predecir	1: 0-10 2: 11-13 3: 14-17 4: 18-20

Fuente: Elaboración Propia

3.5.4 Fase 4: Modelamiento

3.5.4.1 Comparación de modelos (algoritmos)

```
from sklearn.model_selection import train_test_split

df_f=df_2[cols_f2]

X = df_f.iloc[:, :-1].values
y = df_f['TARGET'].values

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

import scipy

scipy.stats.shapiro(y_train)

ShapiroResult(statistic=0.7140690888272095, pvalue=8.93127630297141e-31)

from numpy import array
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.svm import SVC
from sklearn.linear_model import SGDClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neural_network import MLPClassifier
from xgboost import XGBClassifier
from sklearn.model_selection import KFold, cross_val_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn import metrics

models = []
models.append(('LRN', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('SVM', SVC()))
models.append(('SGD', SGDClassifier()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('GNB', GaussianNB()))
models.append(('DTS', DecisionTreeClassifier()))
models.append(('RFS', RandomForestClassifier()))
models.append(('NNM', MLPClassifier()))
models.append(('XGB', XGBClassifier()))

results = []

names = []

for name, model in models:
    kfold = KFold(n_splits=10, random_state=7, shuffle=True)
    cv_results = cross_val_score(model, X_train, y_train, cv=kfold, scoring='accuracy')
    results.append(cv_results)
    names.append(name)
    msg = "%s: %f (%f)" % (name, cv_results.mean(), cv_results.std())
    print(msg)

LRN: 0.564973 (0.058088)
LDA: 0.579863 (0.049783)
SVM: 0.581475 (0.033608)
SGD: 0.421831 (0.117020)
KNN: 0.573197 (0.090042)
GNB: 0.310738 (0.076655)
DTS: 0.795738 (0.050818)
RFS: 0.812295 (0.049329)
NNM: 0.604754 (0.064883)
XGB: 0.810628 (0.040130)
```

Figura 53: Codificación para la comparación de modelos

Fuente: Elaboración Propia

En la figura 53, se puede observar los modelos entrenados para realizar una la comparación.

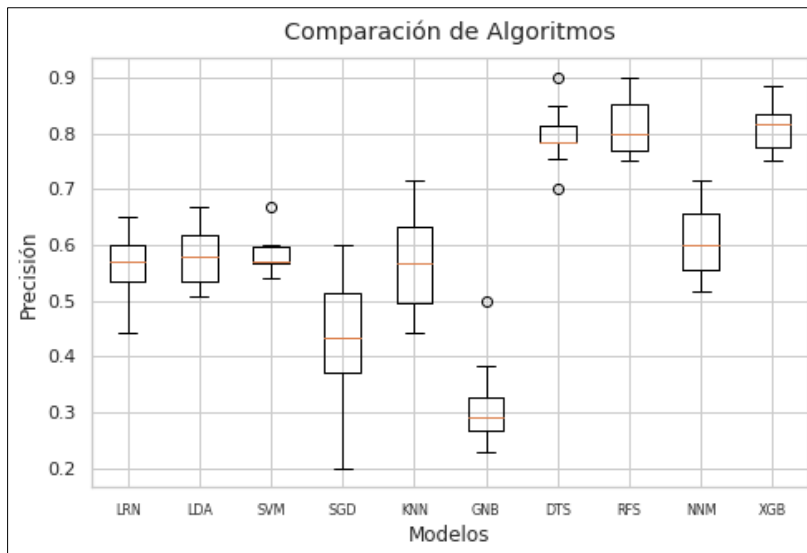


Figura 54: Diagrama de caja y bigotes con los resultados de evaluación

Fuente: Elaboración Propia

En la figura 54, se crea el diagrama de caja y bigotes con los resultados de evaluación de los modelos para comparar la distribución y precisión media para cada modelo. En donde el algoritmo Random Forest Classifier (RFS) tuvo una mejor distribución y precisión.

3.5.4.2 Construcción del modelo

```
print('model:', names[7])
results[7]

model: RFS
array([0.85245902, 0.7704918 , 0.86666667, 0.76666667, 0.9
       0.78333333, 0.76666667, 0.75      , 0.81666667, 0.85      ])
```

Figura 55: Selección del algoritmo más optimo

Fuente: Elaboración Propia

En la figura 55, se observa la codificación para poder elegir el modelo que tuvo una mejor distribución y precisión.

```
RFS = RandomForestClassifier()
RFS.fit(X_train, y_train)

RandomForestClassifier()
```

Figura 56: Entrenamiento del algoritmo Random Forest

Fuente: Elaboración Propia

```
#Se realiza las predicciones con los datos del array de prueba 'X_test'
predictions = RFS.predict(X_test)
```

Figura 57: Predicción con dataset de prueba

Fuente: Elaboración Propia

```
#Puntaje de clasificación de precisión
print("Accuracy: ",accuracy_score(y_test, predictions))

Accuracy: 0.8604651162790697
```

Figura 58: Accuracy obtenido por el algoritmo Random Forest

Fuente: Elaboración Propia

3.5.4.3 Variables más influyentes en el modelo de clasificación

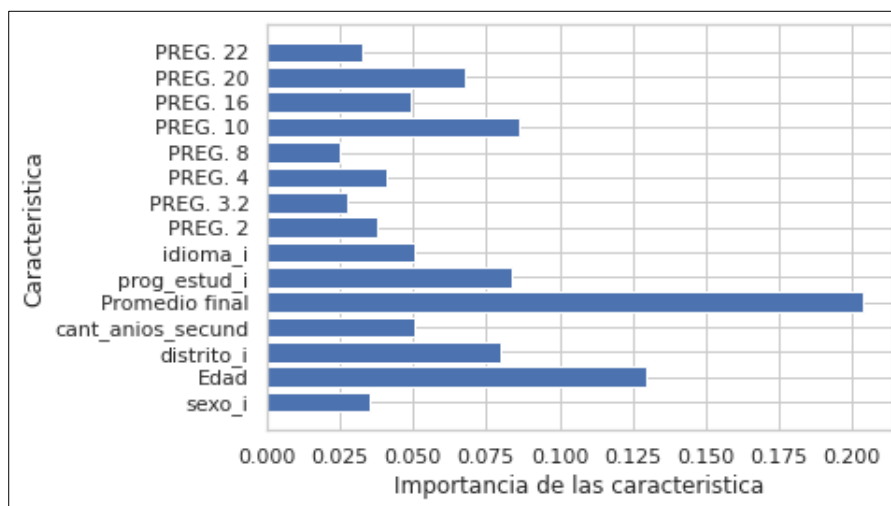


Figura 59: Variables más importantes para la predicción

Fuente: Elaboración Propia

En la figura 59, se observa el cuadro estadístico de cada factor, por lo tanto podemos decir que el factor que más influye en el rendimiento académico es la edad del estudiante, por otro lado la variable que menos influye es el tipo de material de la vivienda del estudiante.

```

feature_list=cols_f2

# Get numerical feature importances
importances = list(RFS.feature_importances_)
feature_importances = [(feature, round(importance, 2)) for feature, importance in zip(feature_list, importances)]
# Sort the feature importances by most important first
feature_importances = sorted(feature_importances, key = lambda x: x[1], reverse = True)
# Print out the feature and importances
[print('Variable: {:20} Importance: {}'.format(*pair)) for pair in feature_importances];

Variable: Promedio final      Importance: 0.2
Variable: Edad                Importance: 0.13
Variable: PREG. 10            Importance: 0.09
Variable: distrito_i          Importance: 0.08
Variable: prog_estud_i        Importance: 0.08
Variable: PREG. 20            Importance: 0.07
Variable: cant_anios_secund    Importance: 0.05
Variable: idioma_i            Importance: 0.05
Variable: PREG. 16            Importance: 0.05
Variable: sexo_i              Importance: 0.04
Variable: PREG. 2              Importance: 0.04
Variable: PREG. 4              Importance: 0.04
Variable: PREG. 3.2            Importance: 0.03
Variable: PREG. 22            Importance: 0.03
Variable: PREG. 8              Importance: 0.02

```

Figura 60: Nivel de importancia de cada variable

Fuente: Elaboración Propia

En la figura 60, se observa el nivel de importancia de cada factor, con respecto a la predicción del bajo rendimiento académico de los estudiantes del Instituto de Educación Superior Pedagógico Público Juliaca.

3.5.4.4 Calibración del modelo

```
## Random Forest Classifier
from sklearn.model_selection import StratifiedKFold
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report
from imblearn.combine import SMOTEENN
print(__doc__)

# Set the parameters by cross-validation
tuned_parameters = [{"max_depth": range(1,16),
                    "random_state": [7], 'min_samples_split' : range(10,500,20)}

scores = ['balanced_accuracy']
skf = StratifiedKFold(n_splits=2)

for score in scores:
    print("# Tuning hyper-parameters for %s" % score)
    print()

    clf = GridSearchCV(RandomForestClassifier(),tuned_parameters, cv=skf,
                      scoring= score)
    clf.fit(X_train, y_train)

    print("Best parameters set found on development set:")
    print()
    print(clf.best_params_)
    print()
    print("Grid scores on development set:")
    print()
    means = clf.cv_results_['mean_test_score']
    stds = clf.cv_results_['std_test_score']
    for mean, std, params in zip(means, stds, clf.cv_results_['params']):
        print("%0.3f (+/-%0.03f) for %r"
              % (mean, std * 2, params))
    print()

    print("Detailed classification report:")
    print()
    print("The model is trained on the full development set.")
    print("The scores are computed on the full evaluation set.")
    print()
    y_true, y_pred = y_test, clf.predict(X_test)
    print(classification_report(y_true, y_pred))
    print()

Detailed classification report:

The model is trained on the full development set.
The scores are computed on the full evaluation set.
```

	precision	recall	f1-score	support
1	0.80	0.24	0.37	50
2	0.71	0.47	0.57	64
3	0.70	0.97	0.81	144
accuracy			0.71	258
macro avg	0.74	0.56	0.58	258
weighted avg	0.72	0.71	0.66	258

Figura 61: Codificación de la calibración del modelo

Fuente: Elaboración Propia

En la Figura 61 se observa la codificación y los resultados de la calibración del modelo, en donde se obtuvo un accuracy de 0.71%, este valor es menor al que obtuvimos en el primer entrenamiento del modelo, por lo tanto, nos quedamos con el modelo del primer entrenamiento que obtuvo un 0.86% de accuracy.

3.5.4.5 Curva ROC

```
auc = roc_auc_score(testy, probs)
print('AUC: %.2f' % auc)

AUC: 0.95
```

Figura 62: Porcentaje del rendimiento de la clasificación AUC

Fuente: Elaboración Propia

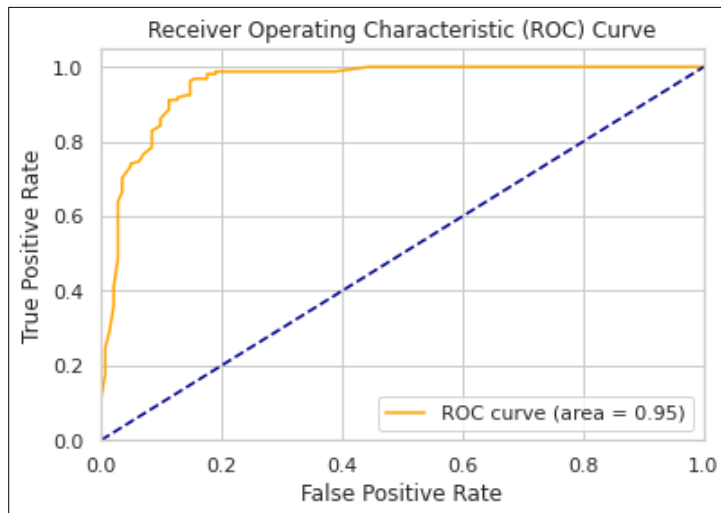


Figura 63: Curva ROC

Fuente: Elaboración Propia

3.5.5. Fase 5: Evaluación

3.5.5.1 Evaluación de los resultados

```
## xt=X_test[:1]
xt= np.array([[ 1., 18., 1., 13., 1., 6., 3., 3., 0., 3., 6., 2., 3.,
                1., 2. ]])

#Se realiza las predicciones con los datos del array de prueba 'X_test'

#cols_f2 = ['sexo_i', 'Edad', 'distrito_i', 'Promedio final', 'prog_estud_i',
#           'cant_anios_secund', 'idioma_i', 'PREG. 2', 'PREG. 3.2', 'PREG. 4',
#           'PREG. 10', 'PREG. 16', 'PREG. 20', 'PREG. 22', 'TARGET']
predictions2 = RFSloaded.predict(xt)
predictions2

array([3])
```

Figura 64: Test del modelo entrenado

Fuente: Elaboración Propia

En la figura 61, se observa datos ingresados de manera aleatoria para poder realizar el respectivo test, de esta manera se logró evaluar el modelo entrenado.

3.5.6 Fase 6: Implementación

En esta fase de la metodología CRISP-DM, para poder utilizar el modelo predictivo se realizará una interfaz web, que será entregada a la unidad académica del Instituto de Educación Superior Pedagógico Público Juliaca, para que puedan tomar acciones según los resultados de cada postulante.

```
from joblib import dump
dump(RFS, 'modelo_entrenado.pkl')

['modelo_entrenado.pkl']
```

Figura 65: Exportar el modelo entrenado

Fuente: Elaboración Propia

La figura 62, nos muestra la codificación para poder exportar el modelo entrenado, para luego poderlo utilizar con en la interfaz web mediante el framework flask.

APP ML

localhost:5000

INSTITUTO DE EDUCACIÓN SUPERIOR
PEDAGÓGICO PÚBLICO JULIACA

INSTITUTO DE EDUCACIÓN SUPERIOR PEDAGÓGICO PÚBLICO
JULIACA

Responde cada uno de las preguntas del formulario

SEXO	¿LA PERSONA QUE MANTIENE TU HOGAR ES?
EDAD	N° DE VECES QUE POSTULASTE A OTROS INSTITUTOS/UNIVERSIDADES
DISTRITO	TIPO DE PREPARACIÓN QUE RECIBISTE PARA POSTULAR AL IESP
CANTIDAD DE AÑOS EST. EN LA SECUNDARIA	NÚMERO DE HORAS DIARIAS QUE ACTUALMENTE DEDICAS AL ESTUDIO
PROMEDIO FINAL DE ADMISIÓN	¿CADA CUÁNTO TIEMPO RECIBES AYUDA ECONÓMICA?
PROGRAMA DE ESTUDIOS	NÚMERO DE DORMITORIOS EN TU VIVIENDA
IDIOMA NATIVO	TIPO DE MATERIAL DE LA VIVIENDA

PREDECIR

*El presente formulario nos permitirá ayudarte a mejorar tu rendimiento académico.

© 2021 - Instituto de Educación Superior Pedagógico Público Juliaca

Figura 66: Interfaz web

Fuente: Elaboración Propia

Capítulo IV

RESULTADOS

CAPÍTULO IV. Resultados

4.1 Resultado 1

La obtención y preparación de los datos de los ingresantes al Instituto de Educación Superior Pedagógico Público Juliaca, se realizaron aplicando la metodología CRIPS-DM, la recolección de datos se realizó en la oficina de Unidad Académica del IESPPJ, en la fase 3 se procedió con la preparación de los datos.

df_2																	
rito_i	cant_anios_secund	Promedio final	prog_estud_i	period_acad_i	...	PREG. 19	PREG. 20	PREG. 21	PREG. 22	PREG. 23.1	PREG. 23.2	PREG. 24.1	PREG. 25	PONDERADO	TARGET		
9	5.0	11	5	1	...	1.0	5.0	1.0	2.0	1.0	4.0	1.0	0.0	13	2		
7	5.0	15	5	1	...	1.0	3.0	1.0	1.0	1.0	4.0	1.0	0.0	15	3		
1	5.0	13	5	1	...	1.0	3.0	1.0	2.0	1.0	4.0	1.0	0.0	16	3		
1	5.0	12	5	1	...	1.0	5.0	1.0	2.0	1.0	4.0	1.0	0.0	14	3		
1	5.0	12	5	1	...	1.0	3.0	2.0	2.0	4.0	5.0	1.0	0.0	14	3		
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
1	5.0	15	1	5	...	1.0	4.0	2.0	1.0	1.0	4.0	1.0	0.0	16	3		
2	5.0	16	1	5	...	1.0	3.0	2.0	1.0	1.0	4.0	1.0	0.0	15	3		
13	6.0	14	1	5	...	1.0	3.0	1.0	1.0	1.0	4.0	1.0	0.0	13	2		
1	5.0	15	1	5	...	1.0	2.0	2.0	1.0	1.0	4.0	1.0	0.0	15	3		

Figura 67: Dataframe para el entrenamiento del modelo

Fuente: Elaboración Propia

En la Figura 67 se observa el Dataframe preparado para el entrenamiento del modelo predictivo.

4.1 Resultado 2

Para el entrenamiento y elección del modelo predictivo se utilizaron datos del semestre académico 2013-I al 2019-I. para el presente trabajo se utilizaron 10 modelos predictivos, para poder realizar una comparación de dichos modelos y seleccionar el modelo predictivo que mejor logre clasificar. (Figura 53). El modelo que mejor clasificación ha logrado fue el Random Forest Classifier (Figura 70).

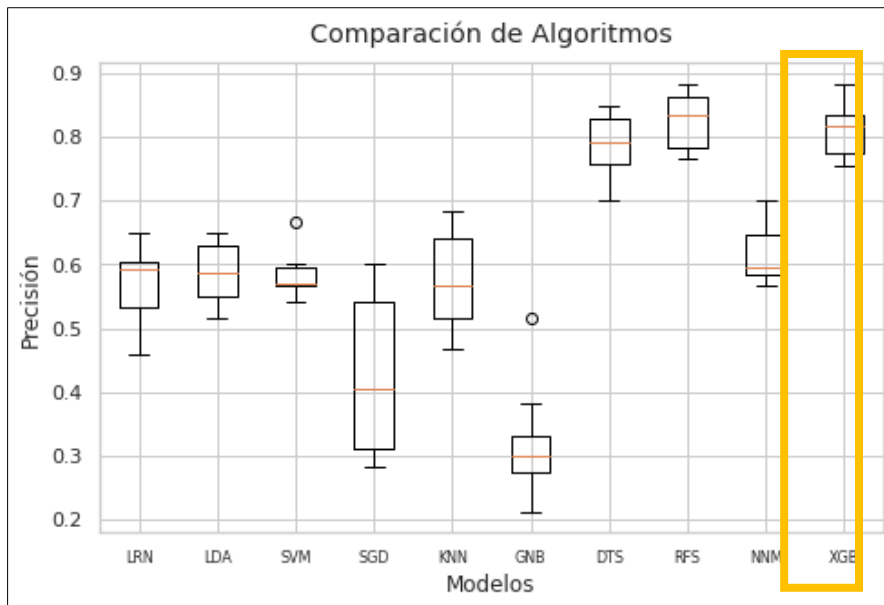


Figura 68: Diagrama de caja y bigotes con los resultados de evaluación

Fuente: Elaboración Propia

4.1 Resultado 3

A continuación, se detalla las Features (variables) que influyen en el bajo rendimiento académico de los estudiantes del nivel superior.

```
feature_list=cols_f2

# Get numerical feature importances
importances = list(RFS.feature_importances_)
feature_importances = [(feature, round(importance, 2)) for feature, importance in zip(feature_list, importances)]
# Sort the feature importances by most important first
feature_importances = sorted(feature_importances, key = lambda x: x[1], reverse = True)
# Print out the feature and importances
[print('Variable: {:20} Importance: {}'.format(*pair)) for pair in feature_importances];
```

Variable: Promedio final	Importance: 0.2
Variable: Edad	Importance: 0.13
Variable: PREG. 10	Importance: 0.09
Variable: distrito_i	Importance: 0.08
Variable: prog_estud_i	Importance: 0.08
Variable: PREG. 20	Importance: 0.07
Variable: cant_anios_secund	Importance: 0.05
Variable: idioma_i	Importance: 0.05
Variable: PREG. 16	Importance: 0.05
Variable: sexo_i	Importance: 0.04
Variable: PREG. 2	Importance: 0.04
Variable: PREG. 4	Importance: 0.04
Variable: PREG. 3.2	Importance: 0.03
Variable: PREG. 22	Importance: 0.03
Variable: PREG. 8	Importance: 0.02

Figura 69: Factores con importancia para la predicción del bajo rendimiento académico.

Fuente: Elaboración Propia

En la Figura 68 se puede observar el porcentaje de importancia que tiene cada factor con respecto a la predicción del bajo rendimiento académico. Siendo así que el promedio final del examen de admisión tiene un 20% de importancia con respecto a la predicción, luego está la

variable edad con un 13% de importancia, luego está la variable número de horas diarias que actualmente dedica al estudio con un 9% de importancia, luego está la variable distrito en donde está ubicado el centro de estudios secundarios de procedencia con un 8% de importancia, luego está la variable programa de estudios al que está postulando con un 8% de importancia, luego está la variable número de dormitorios de su vivienda con un 7% de importancia, luego está la variable cantidad de años en la que cursó la educación secundaria con un 5% de importancia, luego está la variable idioma nativo que habla con un 5% de importancia, luego está la variable ¿Cada cuánto tiempo recibes ayuda económica?, Con un 5% de importancia, luego está la variable sexo con un 4% de importancia, luego está la variable ¿La persona que mantiene su hogar es? Con un 4% de importancia, luego está la variable tipo de preparación que recibiste para postular al IESP con un 4% de importancia, luego está la variable número de veces que postulaste a otros Institutos / Universidades con un 3% de importancia, finalmente está la variable tipo de material de la vivienda con un 3% de importancia.

En la siguiente figura se puede observar gráficamente el nivel de importancia de cada factor.

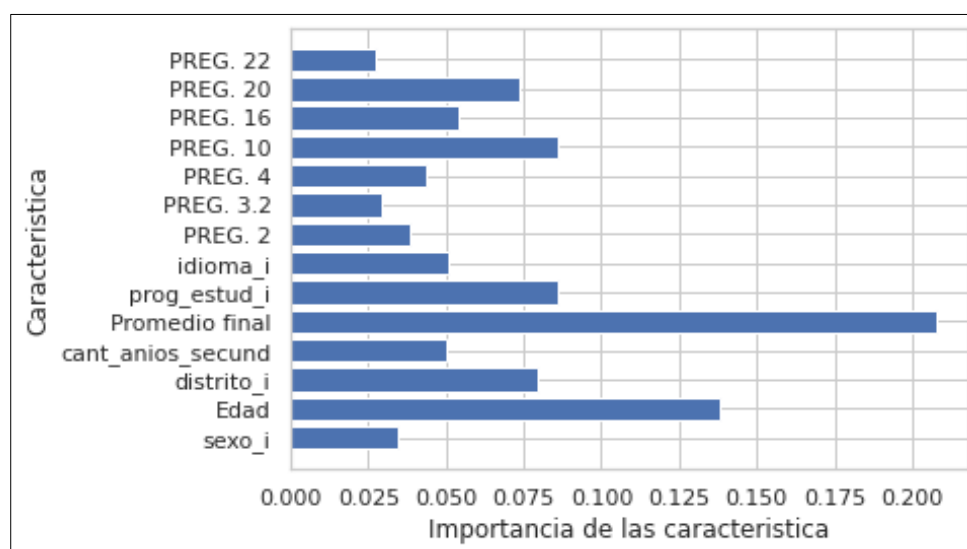


Figura 70: Variables importantes para la predicción

Fuente: Elaboración Propia

En la Figura 69 se observa el cuadro estadístico de cada factor, entonces se puede decir que la variable que más importancia tiene al momento de realizar la predicción con el modelo, es el promedio final del examen de admisión. Por otro lado, la variable que menos importancia tiene es: tipo de material de su vivienda.

4.1 Resultados 4

El modelo que se utilizó en el presente trabajo fue el de Random Forest Classifier, dicho modelo logró una mejor clasificación en comparación con otros modelos (figura 70). El modelo Random Forest Classifier logró un Accuracy del 86% (figura 58).

4.1 Resultado 5

La interfaz web para facilitar el uso de la del modelo predictivo fue desarrollado con el framework Flask (Figura 66).

Figura 71: Interfaz web

Fuente: Elaboración Propia

Capítulo V

CONCLUSIONES Y RECOMENDACIONES

CAPÍTULO V. Conclusiones y Recomendaciones

5.1. Conclusiones

1. Al aplicar la metodología CRISP-DM, en la Fase II en la etapa de comprensión de los datos se hizo la recolección, preparación y exploración de datos iniciales de los ingresantes al Instituto de Educación Superior Pedagógico Público Juliaca.
2. Para el desarrollo del modelo predictivo, primero se hizo la evaluación de los diferentes modelos de Machine Learning, en donde el algoritmo Random Forest Classifier (RFS) tuvo una mejor distribución y precisión de clasificación (figura 70), logrando un accuracy de 0.86% (figura 58).
3. Al aplicar Machine Learning mediante la Metodología CRISP-DM, con el uso del algoritmo Random Forest Clasifier, nos permitió identificar las variables más influyentes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes del Instituto de Educación Superior Pedagógico Público Juliaca: Primero se puede considerar al promedio final del examen de admisión que tiene un 20% de importancia con respecto a la predicción, luego está la variable edad con un 13% de importancia (figura 68), son las dos variable que más importancia tienen al momento de realizar la predicción del bajo rendimiento académico de los estudiantes.
4. La evaluación del modelo predictivo se realizó mediante la puntuación de la predicción (Accuracy) del modelo seleccionado (figura 58).
5. Se construyó la interfaz web (Figura 66), el cual facilita el uso del modelo predictivo para identificar al postulante con bajo rendimiento académico.

5.2. Recomendaciones

1. Se recomienda al Instituto de Educación Superior Pedagógico Público Juliaca, contar con sistemas de información propios y no depender del MINEDU, para tener una fuente de información y realizar nuevas investigaciones aplicando las técnicas de Machine Learning.
2. Se recomienda a futuros investigadores utilizar como conocimiento el presente trabajo y profundizar el estudio de las áreas del rendimiento académico de los estudiantes, utilizando nuevas fuentes de información, tales como información psicológica y/o historial médico de los estudiantes.
3. Se recomienda utilizar otras técnicas de Machine Learning, para identificar nuevas variables que influyen en el bajo rendimiento académico.

REFERENCIAS BIBLIOGRÁFICAS

- Arnold Cathalifaud, M., & Osorio, F. (1998). *Introducción a los Conceptos Básicos de la Teoría General de Sistemas*. Chile: Cinta de Moebio.
- Astorga, J. (2014). Aplicación de modelos de regresión lineal para determinar las armónicas de tensión y corriente. *Scielo*, 234-241.
- Barrientos, R., Cruz, N., Acosta, H., Rabatte, I., Gogeoascoechea, M. d., Pavón, P., & Blázquez, S. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica de la Universidad Veracruzana*, 19-24.
- Bernuy, A. E. (2018). *Predicción del Rendimiento Académico Mediante Minería de Datos en Estudiantes del Primer Ciclo de la Escuela Profesional de Ingeniería de Computación y Sistemas, Universidad de San Martín de Porres, Lima - Perú*. Lima - Perú: Universidad de San Martín de Porres.
- Benavidez-Acosta, M. Ángel, Vacca-González, H., & Giraldo-Ramos, F. N. (2021). La transformada Z en sistemas de control: revisión y un caso de aplicación. *Enfoque Disciplinario*, 6(1), 10-31. <https://doi.org/10.70165/enfdis.v6i1.277>
- Biffi, A. (03 de 09 de 2018). *Inteligencia artificial para la Escuela: ¿Cómo la Inteligencia Artificial está transformando las salas de clases?* Obtenido de <https://medium.com/@everislatam/inteligencia-artificial-para-la-escuela-c%C3%B3mo-la-inteligencia-artificial-est%C3%A1-transformando-las-a4158cb9ff26>
- Brey, G. (18 de 08 de 2018). *Introducción a ciencia de datos con Python*. Recuperado el 30 de 09 de 2018, de <https://medium.com/ingenia-architectural-journeys/introducci%C3%B3n-a-ciencia-de-datos-con-python-b7d027430e41>
- BSG INSTITUTE. (16 de 09 de 2019). *¿Qué es Machine Learning?* Recuperado el 16 de 09 de 2019, de <https://bsginstitute.com/area/Big-Data/Machine-Learning>

- Caballero, R., Martín, E., & Riesco, A. (2019). *Big data con python; recolección, almacenamiento y proceso*. Colombia, Bogota: Alfaomega Colombiana S.A.
- Cegarra Sánchez, J. (2004). *Metodología de la investigación científica y tecnológica*. España, Madrid: Ediciones Díaz de Santos.
- Cervantes, O. D., Báez, J. M., Arízaga, A., & Castillo, E. (2017). *Python con aplicaciones a las matemáticas, ingeniería y finanzas*. México: Alfaomega Grupo Editor, S.A. .
- Chandra, S., Bhateja, V., Mohanty, J., & Udgata, S. K. (2020). *Smart Intelligent Computing and Applications*. India: Springer.
- Colonio, L. A. (2017). *Estilos de aprendizaje y rendimiento académico de los estudiantes de los cursos comprendidos dentro de la línea de construcción-DAC-FIC-UNI*. Lima, Perú: Universidad Peruana Cayetano Heredia.
- Coyla, E. (2016). *Análisis de datos con BigData en proceso de admisión de la Universidad Nacional del Altiplano de Puno, 2016*. Puno, Perú.
- Cuevas, A. (2018). *Aplicaciones gráficas con Python 3*. España, Madrid: RA-MA Editorial.
- Escarcena, H. R., & Velasquez, T. X. (2017). *Análisis de datos con R para determinar el nivel de cumplimiento del perfil del ingresante a la Facultad de Ingeniería Mecánica Eléctrica, Electrónica y Sistemas de la UNA - Puno, 2017*. Puno, Perú.
- Flóres, R., & Fernández, J. M. (2008). *Las redes neuronales artificiales; Fundamentos teóricos y aplicaciones prácticas*. La Coruña, España: Netbiblo, S.L.
- García, J. (2019). *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*. Juliaca, Perú.
- García, J. D., & Skarita, A. (2018). Predicting Academic Performance Based on Students' Family Environment: Evidence for Colombia Using Classification Trees. *Psychology, Society & Education*, 299-311.

- Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: concepts and techniques* . United States of America: Morgan Kaufmann Publishers.
- Holgado, L. A. (2018). *Detección de patrones de bajo rendimiento académico mediante técnicas de minería de datos de los estudiantes de la Universidad Nacional Amazónica de Madre de Dios 2018*. Puno: Universidad Nacional del Altiplano.
- Hurwitz, J., & Kirsch, D. (2018). *Machine Learning For Dummies*. EEUU: John Wiley & Sons, Inc.
- I.E.S.P.P.J. (2020). *Visión*. Juliaca.
- Jain, R. (20 de 03 de 2017). *Decision Tree. It begins here*. Recuperado el 30 de 09 de 2019, de https://medium.com/@rishabhjain_22692/decision-trees-it-begins-here-93ff54ef134
- Jara, D., Velarde, H., Gordillo, G., Guerra, G., León, I., Arroyo, C., & Figueroa, M. (2008). Factores influyentes en el rendimiento académico de estudiantes del primer año de medicina. *Scielo*.
- Johansen, O. (2004). *Introducción a la teoría general de sistemas*. México: Linusa.
- Joyanes, L. (2019). *Inteligencia de negocios y analítica de datos: una visión global de business intelligence & analytics*. Bogotá, Colombia: Alfaomega Grupo Editor, México.
- Khepri, W. (02 de 11 de 2018). *Redes Neuronales, ¿qué son?*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@williamkhepri/redes-neuronales-que-son-a64d022298e0>
- Laura, L., Paredes, K., & Baluarte, C. (2017). Evaluación de Técnicas de Minería de Datos para la Predicción del Rendimiento Académico. *researchgate*, 19-21.
- Luna, J. (08 de 02 de 2018). *Tipos de aprendizaje automático*. Recuperado el 23 de 09 de 2019, de <https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2>

- Madera, M. (2014). *Modelo teórico-metodológico basado en el KDD para la integración y explotación de los bibliográficos de patentes*. Getafe, España: Universidad Carlos III de Madrid.
- Madushan, D. (01 de 12 de 2017). *Introduction to K-means Clustering*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@dilekamadushan/introduction-to-k-means-clustering-7c0ebc997e00>
- Mamani Vargas, M. P. (27 de 05 de 2020). *Entrevista*. Instituto de Educación Superior Pedagógico Público - Juliaca, Juliaca. Obtenido de https://drive.google.com/file/d/1hO2_vhQmMBXKKAeWSGbxtIcMCmcE0osh/view?usp=sharing
- Martínez, X., Santos-Martínez, C. J., & Puche, J. (2018). *Nueva enseñanza superior a partir de las TIC*. España: Editorial GEDISA.
- Menacho, C. H. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Dialnet*, 26-33.
- Mendoza, J. (23 de 04 de 2018). *Arboles de decisión con R - Clasificación*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@jboscomendoza/arboles-de-decisi%C3%B3n-con-r-clasificaci%C3%B3n-c6c583b16125>
- MINEDU - OFICINA DE MEDICIÓN DE LA CALIDAD DE LOS APRENDIZAJES. (2020). *Evaluaciones de logros de aprendizaje*. Lima, Perú: MINEDU.
- MINEDU. (27 de 05 de 2020). *Sistema de Información Académica*. Obtenido de Sistema de Información Académica: <https://sistema.siges-pedagogicos.pe/>
- Molina, J. M., & García, J. (2012). *Técnicas de análisis de datos: Aplicaciones prácticas utilizando microsoft excel y weka*. Madrid, España: Universidad Carlos III de Madrid.
- Morales, N. M. (2018). *Aplicación de la minería de datos a los registros académicos de los estudiantes de la Universidad Nacional Santiago Antúnez de Mayolo-Huaraz, periodo*

- 2000-2015. Tesis de Licenciatura, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz.
- Muñoz Rocha, C. (2015). *Metodología de la investigación*. México: Oxford University Press México, S.A.
- Patil, S. (31 de 07 de 2018). *K Nearest Neighbors*. Recuperado el 30 de 09 de 2019, de <https://medium.com/machinelearningalgorithms/k-nearest-neighbors-c9823dca611b>
- Peralta, F. C. (2014). Proceso de conceptualización del entendimiento del negocio para proyectos de explotación de información. *Revista Latinoamericana de Ingeniería de Software*.
- Pérez, F., & Aldás, J. (25 de 05 de 2019). *Indicadores Sintéticos de las Universidades Españolas*. Valencia: Fundación BBVA. Obtenido de <https://www.fbbva.es/noticias/un-33-de-los-alumnos-no-finaliza-el-grado-que-inicio-y-un-21-abandona-sin-terminar-estudios-universitarios/>
- Piatetsky, G. (01 de octubre de 2014). *Latest KDnuggets Poll asked*. Recuperado el 03 de julio de 2019, de <https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
- Reyes, Y. N. (2003). *Relación entre el rendimiento académico, la ansiedad ante los exámenes, los rasgos de personalidad, el primer autoconcepto y la asertividad en estudiantes del primer año de psicología de la UNMSM*. Lima, Perú: Universidad Nacional de San Marcos.
- Rico, A., Gaytán, N. D., & Sánchez, D. (2019). Construcción e implementación de un modelo para predecir el rendimiento académico de estudiantes universitarios mediante el algoritmo Naïve Bayes. *Scielo*, 1-18.
- Roman, V. (20 de 05 de 2019). *Proyecto de Clasificación de Machine Learning: Encontrar Donantes*. Obtenido de Proyecto de Clasificación de Machine Learning: Encontrar

- Donantes: <https://medium.com/datos-y-ciencia/proyecto-de-clasificaci%C3%B3n-de-machine-learning-encontrar-donantes-64a0c4dbcd34>
- Sanseviero, O. (30 de 01 de 2018). *AI en 3 minutos: Tipos de Machine Learning*. Obtenido de AI en 3 minutos: Tipos de Machine Learning: <https://medium.com/ai-learners/ai-en-3-minutos-tipos-de-machine-learning-945b708ac78>
- SAS. (16 de 09 de 2019). *Aprendizaje automático, qué es y por qué es importante*. Recuperado el 16 de 09 de 2019, de https://www.sas.com/es_pe/insights/analytics/machine-learning.html
- Sehra, C. (19 de 01 de 2018). *Decision Trees Explained Easily*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248>
- Sehra, C. (17 de 01 de 2018). *K Nearest Neighbors Explained Easily*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@chiragsehra42/k-nearest-neighbors-explained-easily-c26706aa5c7f>
- Shetty, B. (12 de 11 de 2018). *Visualización de datos usando Matplotlib*. Recuperado el 30 de 09 de 2019, de <https://towardsdatascience.com/data-visualization-using-matplotlib-16f1aae5ce70>
- Singh, S. (16 de 06 de 2018). *K-Means Clustering*. Recuperado el 30 de 09 de 2019, de <https://medium.com/datadriveninvestor/k-means-clustering-b89d349e98e6>
- SMARTVISION. (05 de julio de 2019). *What is the CRISP-DM methodology?* Recuperado el 05 de julio de 2019, de <https://www.sv-europe.com/crisp-dm-methodology/>
- Tan, Y., Shi, Y., & Tang, Q. (2018). *Data Mining and Big Data*. Shanghai, China: Springer.
- Tanner, G. (13 de 10 de 2018). *Linear regression explained*. Recuperado el 30 de 09 de 2019, de <https://medium.com/@gilberttanner/linear-regression-explained-8e45f234dc55>

- Technopedia. (02 de abril de 2019). *Knowledge Discovery in Databases (KDD)*. Recuperado el 02 de abril de 2019, de [https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd#targetText=Knowledge%20discovery%20in%20databases%20\(KDD,from%20a%20collection%20of%20data](https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd#targetText=Knowledge%20discovery%20in%20databases%20(KDD,from%20a%20collection%20of%20data).
- Unidad Académica I.E.S.P.P.J. (14 de 01 de 2020). *SIGES*. Obtenido de SIGES: <https://sistema.siges-pedagogicos.pe/>
- Villa, A., Carrión, A., & Sozzi, A. (2017). Optimización del diseño de parámetros: Método Forest-Genetic univariante. *Publicaciones en ciencias y tecnologías*, 12-24.
- Villada, F., Muñoz, N., & García, E. (2012). Redes neuronales artificiales aplicadas a la predicción del precio del oro. *Scielo*, 143-150.
- Yamao, E., Celi Saavedra, L., Campos, P. R., & Huancas Hurtado, V. D. (2018). Prediction of academic performance using data mining in first year students of peruvian university. *Revista Campus*, 151–160.
- Zainab Mohammed, A., Noor Hasan, H., Wasan Saad, A., & Hazim Noman, A. (2020). The Application of Data Mining for Predicting Academic Performance Using K-means Clustering and Naïve Bayes Classification. *International Journal of Psychosocial Rehabilitation*, 2143–2151.

ANEXOS

Anexo A. Cronograma del proyecto

Tabla 10.

Cronograma de actividades del proyecto de investigación

ACTIVIDADES	Enero			Febrero			Marzo			Abril			Mayo			Junio			Julio			Agosto			Setiembre			Octubre			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3
FASE 1: INICIO																															
1 Semblanza de caso de estudio	X	X																													
2 Elaboración de proyecto de investigación			X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X												
3 Dictaminación de proyecto de investigación																	X	X													
4 Corrección de proyecto de investigación																		X													
5 Aprobación de proyecto de investigación																			X												
FASE 1: DESARROLLO																															
6 Comprensión de la problemática del I.E.S.P.P.-J.																			X												
7 Recolección de los datos de los estudiantes del I.E.S.P.P.-J.																				X	X										
8 Preparación de los datos de los estudiantes del I.E.S.P.P.-J.																					X	X									
9 Desarrollo de los modelos predictivos																						X	X								
10 Evaluación de los modelos predictivos																							X	X							
11 Despliegue del modelo predictivo																								X							
12 Elaboración del informe final																									X	X					
13 Revisión del informe final por el asesor																										X					
14 Dictaminación del informe final																											X	X			
15 Corrección del informe final																													X		
FASE 3: FINAL																															
16 Presentación del informe final																															X
17 Aprobación del informe final																															X

*Instituto de Educación Superior Pedagógico Público – Juliaca (I.E.S.P.P.-J.)

Fuente: Elaboración Propia

Anexo B. Presupuesto

Tabla 11.

Presupuesto del proyecto de investigación

DESCRIPCIÓN	SUB TOTAL (S/)	TOTAL (S/)
BIENES		
Escritorio	150.00	
Laptop	2800.00	
Impresora	650.00	4 100.00
Útiles de escritorio Papel Bond A4	500.00	
SERVICIOS		
Servicio de internet	800.00	
Viáticos y movilidad	500.00	
Impresión	250.00	1 750.00
Copias	200.00	
SOFTWARE		
Python	0.00	
Scikit-Learn	0.00	
Anaconda	0.00	0.00
Google Drive	0.00	
Google Docs	0.00	
RR.HH.		
Analista de datos	2500.00	2 500.00
INSCRIPCIÓN DEL TEMA DE INVESTIGACIÓN		
Solicitud de asesor	300.00	
Inscripción de proyecto de investigación	600.00	1 700.00
Sustentación	800.00	
PRESUPUESTO TOTAL		10 050.00

Fuente: Elaboración Propia

Financiamiento del Proyecto: Recursos Propios del Investigador

"Año de la Universalización de la Salud"

SOLICITO: Autorización para realizar proyecto de tesis

Dr. Macías Platón Mamani Vargas
DIRECTORA GENERAL

Yo: **Rudy Jhean Rojas Pari** identificado con DNI **72672433**, egresado de la **Escuela Profesional de Ingeniería de Sistemas** de la **Universidad Peruana Unión Filial Juliaca**, con el debido respeto me presento y expongo.

Pido a Ud. autorización para realizar mi proyecto de tesis en el área de Unidad Académica, el cual lleva como título **"Modelo de Aprendizaje Supervisado para Identificar Patrones de Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público - Juliaca"**, para poder realizar el presente proyecto pido me brinde datos históricos de procesos académicos; dicho proyecto que se está planteando no conlleva ningún tipo de gasto para la institución.

Sin otro particular espero su apoyo para el presente.

San Miguel, 13 de marzo de 2020

Atentamente,

Rudy Jhean Rojas Pari
Bach. Ingeniería de Sistemas
DNI: 72672433

Anexo D. Autorización de ejecución del proyecto



MINISTERIO DE EDUCACIÓN
DIRECCIÓN GENERAL DE EDUCACIÓN SUPERIOR Y TÉCNICO PROFESIONAL
DIRECCIÓN REGIONAL DE EDUCACIÓN PUNO
INSTITUTO DE EDUCACIÓN SUPERIOR PEDAGÓGICO PÚBLICO - JULIACA
CREADO EL 02 - 02 - 64 POR LEY N° 14859 - JULIACA - PERÚ
Acreditado por Resolución de Presidencia del Consejo Directivo - Adm. N° 073-2015-COSUSINEACE/CDAH-P
"Rumbo al Licenciamiento"

"Año de la Universalización de la Salud"

AUTORIZACIÓN

El Director del Instituto de Educación Superior Pedagógico Público – Juliaca del distrito San Miguel de la Provincia San Román y Departamento Puno;

AUTORIZA:

Que el Estudiante **Rudy Jhean Rojas Pari**, Identificado con DNI N° **72672433**, egresado de la **Escuela Profesional de Ingeniería de Sistemas de la "Universidad Peruana Unión F-J"**, tiene la Autorización para aplicar su proyecto de investigación intitulado **"Modelo de Aprendizaje Supervisado para Identificar Patrones de Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca"**.

Se expide la presente autorización a petición del interesado para los fines que vea por conveniente.

San Miguel, 16 de marzo de 2020



Dr. Macías P. Mamani Vargas
DIRECTOR GENERAL
INSTITUTO DE EDUCACIÓN SUPERIOR PEDAGÓGICO PÚBLICO - JULIACA

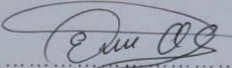
Anexo E. Ficha de Proceso de Admisión



PROCESO DE ADMISIÓN 2015
AL INSTITUTO DE EDUCACIÓN SUPERIOR PEDAGÓGICO PÚBLICO
JULIACA
Acreditada por Resolución de Presidencia del Consejo Directivo-Adhoc
N° 073-2015-COSUSINEACE/CDAH-P

APELLIDO PATERNO:.....
APELLIDO MATERNO:.....
NOMBRES:.....
CARRERA A LA QUE POSTULA: Ciencias Sociales
DOCUMENTO DE IDENTIDAD: DNI. 7-308111
SEXO: FEMENINO
DOMICILIO: Jr. CHILE N°232 TAPARACHI
TELÉFONO: 952101000
CORREO ELECTRÓNICO: analí.michel1@Hotmail.com
FECHA DE NACIMIENTO: 21-06-1996 EDAD: 19
DPTO.: PUNO Provincia: SAN ROMAN Distrito: Juliaca
ESTUDIOS REALIZADOS:
I.E.S.: C.E.B.A. PRIVADO SAN ROMAN
DPTO.: PUNO Provincia: SAN ROMAN Distrito: JULIACA
AÑO QUE INICIO: 2014 AÑO QUE CULMINO: 2014
IDIOMAS QUE HABLA:
a).- Quechua () b).- Aymara () c).- Castellano (X)
Juliaca: 09 de MARZO del 201


Firma del Postulante


Firma del Padre o Apoderado

Ficha Socio Económica del Estudiante

Características familiares:

- 1. ¿Con quién(es) vive(s)?**
 - a) Solo con Papá.
 - b) Solo con Mamá.
 - c) Ambos Padres.
 - d) Otros.
- 2. ¿La persona que mantiene su hogar es?**
 - a) Tú mismo(a)
 - b) Mamá.
 - c) Papá.
 - d) Papá y Mamá.
 - e) Otros.

Aspectos Educativos

- 3. Número de veces que postulaste al IESP:**

Número de veces que postulaste a otros institutos/universidades:

- 4. Tipo de preparación que recibiste para postular al IESP**
 - a) Centro Pre
 - b) Centro Pre de otro IESP.
 - c) Academia.
 - d) Otros.
- 5. Motivo por el cual elegiste en el IESP**
 - a) Por su prestigio
 - b) Tradición Familiar.
 - c) Porque allí estudian amigos.
 - d) Porque es Estatal.
 - e) Cercanía a mi domicilio.
 - f) Vocación.
 - g) Presión Familiar.
 - h) Otros.
- 6. Por qué motivo optaste seguir estudios de docencia (marca solo una alternativa)**
 - a) Por realización Personal
 - b) Por progresar económicamente
 - c) Por progresar socialmente.
 - d) Por ayudar a la comunidad.
 - e) Otros.
- 7. Motivo por el cual elegiste tu especialidad**
 - a) Vocación
 - b) Presión Familiar
 - c) Tradición Familiar
 - d) Test Vocacional.
 - e) Por el puntaje bajo
 - f) Es más rentable

- g) Es la carrera del momento.
- h) Otros.
- 8. ¿Tienes acceso a internet en casa?
 - a) SI
 - b) NO
- 9. ¿Actualmente usas correo electrónico?
 - a) SI
 - b) NO
- 10. Número de horas diarias que actualmente dedicadas al estudio

_____ Horas

- 11. Forma que prefiere estudiar
 - a) Grupo
 - b) Solo
 - c) Profesor particular

Aspectos Socio Económico

- 12. ¿Trabajas actualmente?
 - a) SI
 - b) NO

Si respondiste SÍ, responde a las preguntas 13 y 14, de lo contrario pasa a la pregunta 15

- 13. ¿Cuánto es tu ingreso mensual promedio?

S/. _____

- 14. Número de horas que laboras a la semana

- 15. ¿Recibes ayuda económica de tu padre o madre?
 - a) SI
 - b) NO
- 16. ¿Cada cuánto tiempo recibes ayuda económica?
 - a) Quincenal
 - b) Mensual
 - c) Una vez al año
 - d) Otros.
- 17. Tipo de apoyo que recibiste para su preparación pre-universitaria
 - a) Beca Parcial
 - b) Beca Integral
 - c) No recibo ningún beneficio

Aspectos de Vivienda

- 18. Tipo de vivienda
 - a) Casa Independiente
 - b) Departamento en edificio
 - c) Condominio
 - d) Casa de vecindad
 - e) Condominio
 - f) Casa de pensión
 - g) Condominio

- h) Cuarto Alquilado
 - i) Otro.
- 19. Situación de su vivienda**
- a) Propia
 - b) Alquilada
 - c) Alquiler Venta
 - d) Prestada
 - e) Otro.
- 20. Número de dormitorios de su vivienda**
- _____
- 21. Número de baños de la vivienda**
- _____
- 22. Tipo de material de la vivienda**
- a) Cemento / Ladrillo
 - b) Adobe
 - c) Quincha
 - d) Esteras
 - e) Otro.
- 23. Bienes y enseres en su vivienda**
- a) Cocina a Gas
 - b) Cocina eléctrica
 - c) Aspiradora
 - d) Televisor
 - e) DVD
 - f) Mini componente
 - g) Cámara de video
 - h) Computadora
 - i) Horno microondas
 - j) Lavadora
 - k) Secadora de ropa
 - l) Automóvil
 - m) Bicicleta
 - n) Motocicleta
 - o) Juego de video
 - p) Refrigeradora
 - q) Ninguna de las anteriores
- 24. Servicios con los que cuenta la vivienda**
- a) Empleada(o) doméstica(o)
 - b) Servicio de teléfono
 - c) Servicio de Cable
 - d) Servicio de Internet
 - e) Ninguna de las anteriores
- 25. ¿Cuántos hijos?**
- _____



EDITORIAL
MUNDO
INTERDISCIPLINARIO

Los derechos fundamentales de los profesores universitarios de las universidades privadas en los contratos laborales sujetos a movilidad en la interpretación del Tribunal Constitucional

Autores:

Marco Hernan Pantigozo Loaiza

Rosa Elvira Ruiz Reyes

Rafaela Maria Isabel Aquije Mendez

Fernando Eduardo Cano Legua



ISBN

978-628-97906-5-8