UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

"MODELO DE APRENDIZAJE AUTOMATICO PARA LA PREDICCION DEL RENDIMIENTO ACADEMICO DE LOS ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERIA EN INDUSTRIAS ALIMENTARIAS DE LA UNAS"

Tesis

Para optar el título profesional de:

INGENIERO EN INFORMÁTICA Y SISTEMAS

PRESENTADO POR:

Bach. SANTOS VICTOR PONCE GUIZABALO

Asesor:

Dr. WILLIAM GEORGE PAUCAR PALOMINO

Tingo María – Perú

2023.

Universidad Nacional Agraria de la Selva Facultad de Ingeniería en Informática y Sistemas	**ACTA DE SUSTENTACIÓN DE TESIS N.º 02-2024**	**COMISION DE GRADOS Y TITULOS** **Fecha:** 01/abril/2024

PARTE 1. FASE INICIAL

Siendo las 17:00 **horas del día 01 de abril de 2024**; en la Sala de Conferencias de la FIIS, se instala el jurado calificador conformado por:

Jurado 1: Mg. Marco Arturo Canales Aguirre (presidente)

Jurado 2: Mg. Jorge Luis Pozo Malpartida

Jurado 3: Mg. Ronald Eduardo Ibarra Zapata

Oficializado mediante **RESOLUCIÓN N° 142-2023-D-FIIS-UNAS** del 18 de octubre de 2023, para el proceso de sustentación del informe final de Tesis del bachiller **Santos Víctor PONCE GUIZABALO,** titulado: **"MODELO DE APRENDIZAJE AUTOMATICO PARA PREDECIR EL RENDIMIENTO DE ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERÍA DE INDUSTRIAS ALIMENTARIAS DE LA UNAS".** ASESOR: **Dr. William George Paucar Palomino.**

Se manifiesta que el bachiller cumple con los requisitos exigidos de Ley y se le invita a disertar su Tesis por espacio de 30 minutos, asimismo se dispondrá de igual tiempo para la absolver preguntas y sugerencias.

PARTE 2. FASE DE PREGUNTAS Y RESULTADO

Culminada la exposición se inicia la fase de preguntas por parte del jurado calificador; también se invita a los asistentes a formular preguntas sobre el tema de Tesis.

Absueltas todas las peticiones, el jurado calificador procede a deliberar en privado la calificación y resultado.

Concluida la deliberación y en presencia del público, el jurado calificador anuncia que el resultado de la Sustentación de Tesis es: APROBADO POR MAYORIA

(NOTA: consignar una de la siguientes: DESAPROBADO, APROBADO POR MAYORIA o APROBADO POR UNANIMIDAD)

Con calificativo de: BUENO

(NOTA: consignar una de la siguientes: EXCELENTE, MUY BUENO, BUENO, DEFICIENTE, MUY DEFICIENTE)

Por lo que se comunicará a las instancias correspondientes para el trámite respectivo.

PARTE 3. CONFORMIDAD

De todo lo mencionado se firma al pie en señal de conformidad, siendo las 18:30 horas se da por finalizada la ceremonia de Sustentación de Tesis.

Firma:	Firma:	Firma:
Jurado 1: Marco Arturo Canales Aguirre	**Jurado 2:** Jorge Luis Pozo Malpartida	**Jurado 3:** Ronald Eduardo Ibarra Zapata
Firma:	Firma:	
Sustentante: Santos Víctor PONCE GUIZABALO	**Asesor:** William George Paucar Palomino	

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA
DIRECCIÓN DE GESTIÓN DE INVESTIGACIÓN - DGI
REPOSITORIO INSTITUCIONAL - UNAS
Correo: repositorio@unas.edu.pe

"Año del Bicentenario, de la consolidación de nuestra Independencia, y de la conmemoración de las heroicas batallas de Junín y Ayacucho"

CERTIFICADO DE SIMILITUD T.I. N° 165 - 2024 - CS-RIDUNAS

El Director de la Dirección de Gestión de Investigación de la Universidad Nacional Agraria de la Selva, quien suscribe,

CERTIFICA QUE:

El Trabajo de Investigación; aprobó el proceso de revisión a través del software TURNITIN, evidenciándose en el informe de originalidad un índice de similitud no mayor del 25% (Art. 3° - Resolución N° 466-2019-CU-R-UNAS).

Programa de Estudio:

Ingeniería en Informática y Sistemas

Tipo de documento:

Tesis	X	Trabajo de Suficiencia Profesional	

TÍTULO	AUTOR	PORCENTAJE DE SIMILITUD
MODELO DE APRENDIZAJE AUTOMATICO PARA LA PREDICCION DEL RENDIMIENTO ACADEMICO DE LOS ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERIA EN INDUSTRIAS ALIMENTARIAS DE LA UNAS	Santos Victor Ponce Guizabalo	23 % Veintitrés

Tingo María, 22 de mayo de 2024

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA
UNIDAD DE GESTION DE LA INVESTIGACION
Dr. Tomas Menacho Mallqui
JEFE

C.C. Archivo

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

MODELO DE APRENDIZAJE AUTOMATICO PARA LA PREDICCION DEL RENDIMIENTO ACADEMICO DE LOS ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERIA EN INDUSTRIAS ALIMENTARIAS DE LA UNAS

Autor : Santos Victor Ponce Guizabalo

Asesor : **Dr.** William George Paucar Palomino

Área de investigación : Sistemas de Información

Línea de investigación : Gestión de datos

Eje temático :

Lugar de ejecución : Facultad de Ingeniería en Industrias Alimentarias

Duración del trabajo : 6 meses

Financiamiento : Propio

Tingo María -Perú 2023.

VICERRECTORADO DE INVESTIGACION
OFICINA DE INVESTIGACION

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

REGISTRO DE TESIS PARA LA OBTENCION DEL TITULO UNIVERSITARIO, INVESTIGACIÓN DOCENTE Y TESISTA

(Resol. Nº 113-2019-CU-R-UNAS)

I. Datos Generales de Pregrado

Universidad	**:**	Universidad Nacional Agraria de la Selva.
Facultad	**:**	Facultad de Ingeniería en Informática y Sistemas.
Título de tesis	**:**	Modelo de aprendizaje automático para la predicción del rendimiento académico de los alumnos ingresantes en la Facultad de ingeniería en Industrias alimentarias de la unas.
Autor	**:**	Ponce Guizabalo, Santos Victor.
Asesor de tesis	**:**	William George Paucar Palomino.
Escuela Profesional	**:**	Escuela Profesional de Ingeniería en Informática y Sistemas.
Programa de investigación	**:**	Sistemas de Información.
Línea(s) de investigación	**:**	Gestión de datos.
Eje Temático	**:**	Modelo de aprendizaje automático para la predicción del rendimiento académico.
Lugar de ejecución	**:**	Facultad de Ingeniería en Industrias Alimentarias.
Duración	**:**	6 meses
Financiamiento	**:**	FEDU : S/0.00 Propio : S/5600.00 Otros : S/.0.00

Tingo María, Perú, mayo 2024.

Santos Victor Ponce Guizabalo
Tesista

William George Paucar Palomino
INGENIERO DE SISTEMAS
CIP. 97223
Asesor

DEDICATORIA

Dedicado a Dios por ser mi guía y protector, a mi hijo que en paz descanse en la gloria de Dios, a mis padres que tengan muchos años más de vida.

AGRADECIMIENTO

A todos los que brindaron aporte para el desarrollo de la presente investigación, a Dios por la vida y la salud, a mi asesor Dr. William George Paucar Palomino por bridarme el apoyo en el desarrollo de la tesis.

Índice General

Índice de Tablas

Tablas **Página**

Índice de Figuras

Figura **Página**

RESUMEN

La presente investigación tiene como objetivo determinar un modelo de aprendizaje automático para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS. Se planteó la hipótesis: La predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS con un modelo de aprendizaje automático es significativa. La técnica que se usó para recolectar los datos es una ficha de análisis documental, se obtuvo una población 204 datos. Con el software WEKA 3.9.5 y el análisis de cinco modelos Vote, k Vecinos más Cercanos (IBK), Random Forest, Naive Bayes, y Bagging. En conclusión, El rendimiento académico de los alumnos ingresantes es un tema muy complejo, usando la metodología CRISP-DM y técnicas de minería de datos se logró determinar un modelo de aprendizaje automático que permite predecir significativamente el rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias. Los resultados muestran que los indicadores claves para predecir el rendimiento académico son: la opción de ingreso, la nota de ingreso quien tiene mayor influencia, el sexo del ingresante y el número de hermanos dentro de la familia. Con las pruebas de Anova de Friedman se obtuvo igual exactitud para los modelos de Random Forest e IBK con 98,4%. El modelo con mejor precisión para los aprobados es Random Forest (98,34%).

Palabras claves: *minería de datos, Vote, k vecinos más cercanos (IBK), Random Forest, Naive Bayes, y bagging., metodología CRISP-DM.*

ABSTRACT

The objective of this research is to determine a machine learning model for the prediction of the academic performance of incoming students of the Faculty of Engineering in Food Industries at UNAS. The hypothesis was raised: The prediction of the academic performance of incoming students of the UNAS Faculty of Engineering in Food Industries with a machine learning model is significant. The technique that was used to collect the data is a documentary analysis sheet, a population of 204 data was obtained. With WEKA 3.9.5 software and the analysis of five Vote models, Nearest Neighbors (IBK), Random Forest, Naive Bayes, and Bagging. In conclusion, the academic performance of incoming students is a very complex issue. Using the CRISP-DM methodology and data mining techniques, it was possible to determine a machine learning model that significantly predicts the academic performance of incoming students of the Faculty of Engineering in Food Industries. The results show that the key indicators to predict academic performance are: the entry option, the entry grade who has the greatest influence, the sex of the applicant and the number of siblings within the family. With Friedman's Anova tests, equal accuracy was obtained for the Random Forest and IBK models with 98.4%. The model with the best accuracy for those approved is Random Forest (98.34%).

Keywords: data mining, Vote, nearest neighbors (IBK), Random Forest, Naive Bayes, and bagging., CRISP-DM methodology.

I. INTRODUCCION

1.1.Planteamiento del problema

A nivel mundial, en los últimos años, uno de los principales objetivos de las instituciones de enseñanza superior es la búsqueda permanente de la excelencia educativa. La educación superior surge ante los adolescentes como un medio fundamental para alcanzar sus metas de realización personal en una sociedad mundial donde se persigue la excelencia global y el mercado laboral y profesional es cada vez más selectivo y competitivo. Por ello, son muchos los jóvenes que cada año solicitan plazas en las instituciones públicas y privadas de nuestro país para recibir la formación profesional que necesitan para avanzar en un campo de estas características. En relación con lo anterior, el abandono escolar es especialmente interesante por su conexión con el rendimiento académico, con los procesos de selección y con el rendimiento académico del alumno (Ulloa Gallardo y otros, 2020). Incluso los estudiantes abandonan la universidad después de empezarlos durante unas semanas o meses. Las causas son muchas y variadas, entre ellas una mala orientación, la falta de una estrategia de estudio, un bajo nivel de capacidades, la falta de motivación y problemas económicos.

A nivel nacional, según los datos de Penta Analytics (2017), el 27% de los estudiantes que se matriculan en una universidad privada abandonarón sus estudios en el primer año; el porcentaje aumentaría si se contabilizara el número total de estudiantes que no terminan sus estudios. Hay cuatro razones principales por las que los estudiantes abandonan sus estudios antes de tiempo: bajo rendimiento académico, problemas económicos, incertidumbre profesional y problemas emocionales.

Para contribuir con la solución del problema sobre el rendimiento académico de los estudiantes ingresantes, se cuenta con una base de datos de los últimos cuatro años, así como las características que la identifican al interior de la universidad, con la finalidad de predecir el rendimiento académico que se ha visto un tanto deficiente en los últimos años en la Universidad

Nacional Agraria de la Selva (UNAS).

1.2.Formulación del problema

1.2.1. Problema general

La presente investigación plantea como interrogante general lo siguiente:

¿Qué modelo de aprendizaje automático permite la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?

1.2.2. Problemas específicos

Las interrogantes en los Problemas específicas planteadas son:

- ¿Qué modelo de aprendizaje automático tienen mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?
- ¿Qué modelo de aprendizaje automático tienen mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?

1.3.Justificación y alcance de la investigación

Justificación Práctica

La dirección general de estudios de la Facultad de Ingeniería en Industrias Alimentarias valorará muy positivamente este trabajo porque le permitirá implementar informes adecuados para la toma de decisiones apropiadas a las circunstancias académicas de los estudiantes universitarios que ingresen en el marco de la Ley Universitaria y de acuerdo con el Estatuto de la UNAS, además de ello es de suma importancia para los directivos de la escuela profesional de Ingeniería en Industrias Alimentarias ya que podrán conocer el rendimiento académico de los alumnos ingresantes y tomar medidas correctivas para lograr la mejora en el desempeño estudiantil.

El rendimiento académico de los alumnos es muy importante para la facultad y su correcta identificación del rendimiento de los alumnos permitirá tener mejores resultados y menos deserción por parte de los alumnos. En ese aspecto los modelos de machine learning son una herramienta muy valiosa para la predicción de que alumnos lograran aprobar y que alumnos tendrán dificultades, ya que pueden analizar grandes cantidades de datos tantos académicos como socioeconómicos para detectar patrones y relaciones entre las distintas variables que son muy difícil de detectar con métodos tradicionales.

Justificación Teórica

El alcance de la presente investigación solo abarcará predecir el rendimiento académico con la técnica de minería de datos y encontrar patrones que las determinan en alumnos ingresantes de la Facultad de Ingeniería en Industrias Alimentarias (CFFIA) de la UNAS.

La investigación propuesta, busca mediante la aplicación de la teoría de la minería de datos y los algoritmos de aprendizaje automático encontrar indicadores que influyen en la predicción del rendimiento académico. Además de obtener el algoritmo con mejor exactitud y precisión en la predicción del rendimiento académico, esto permite al investigador contrastar los diferentes conceptos de los modelos de aprendizaje automático y rendimiento académico en la realidad de la Facultad de Ingeniería en Industrias Alimentarias (CFFIA) de la UNAS.

1.4.Objetivos

1.4.1. Objetivos General

Determinar el modelo de aprendizaje automático con mejor predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS.

1.4.2. Objetivos Específicos

- Determinar el modelo de aprendizaje automático con mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS
- Determinar el modelo de aprendizaje automático con mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS.

1.5.Hipótesis

1.5.1. Hipótesis General

El algoritmo con mejor predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el modelo de aprendizaje automático árbol de decisión.

1.5.2. Hipótesis Específicos

- El modelo de aprendizaje automático con mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.
- El modelo de aprendizaje automático con mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.

II. REVISIÓN DE LITERATURA

2.1. Antecedentes

2.1.1. Antecedentes nacionales

Alvarez Gonzaga (2021) en su tesis tiene como objetivo general analizar el rendimiento de las distintas tecnicas de minería de datos en relación con las soluciones de inteligencia empresarial. El procedimiento no probabilístico del enfoque sugerido comenzó con la selección de dos enfoques de minería de datos de entre los que eran accesibles y habían sido reportados en varios estudios. Después, se creó un método de aplicación en cinco etapas, que incluía el procesamiento de datos, la implementación de la base de datos en SQL Server, el ETL, la implementación del algoritmo de minería de datos utilizando datos de entrada del proceso de inteligencia empresarial, y el análisis y la comprensión de las fuentes de datos. Los resultados mostraron que el modelo propuesto, que utilizaba datos de entrada procedentes de un procedimiento de inteligencia empresarial, funcionaba con una precisión superior al 90% de las veces. Ambos son métodos de minería de datos. Naive Bayes obtuvo un 93,67%, mientras que el árbol de decisión obtuvo un 93,69%. Naive Bayes también obtuvo los mejores resultados en el análisis de errores, con un error medio porcentual absoluto (MAPE) del 6,2%. El estudio llega a la conclusión de que se utilizan técnicas de minería de datos. Los mejores resultados proceden de Naive Bayes, que se deriva de un enfoque de inteligencia empresarial. Estas técnicas tienen una gran precisión en la predicción del rendimiento académico y pueden utilizarse para analizar otras características académicas como el abandono y la delincuencia.

Vega (2019) en su tesis de investigación el objetivo de este estudio fue utilizar algoritmos de aprendizaje automático para predecir cuántos estudiantes aprobarían y reprobarían cursos en el Programa de Estudios Básicos de la Universidad Ricardo Palma. La metodología de investigación es de tipo aplicado con un enfoque Cuantitativo; el método que se utiliza es de alcance descriptivo y Correlacional; el diseño de investigación es no-

experimental; la población se tiene un total de 9,118 alumnos y 578,283 calificaciones, posee una muestra de 1000 registros; la técnicas de recolección de datos está conformada por los archivos de datos, el instrumentó de recolección de datos esta desarrollado por el software Python que genera el archivo denominado “relacion.csv”. Se concluye en la investigación que el método de trabajo utilizado para el desarrollo de esta investigación fue la Metodología CRISP-DM. Su uso fue de gran ayuda para obtener ideas y poder estructurar un camino donde fuera fácil completar cada etapa, no porque la investigación fuera complicada, sino por el objetivo a lograr en cada una de ellas.

Candia (2019) en su tesis tuvo como objetivo de investigación utilizar métodos de aprendizaje automático para predecir el rendimiento académico de los estudiantes de primer semestre de la UNSAAC. En este estudio se utiliza una metodología de investigación cuantitativa, correlacional y no experimental; Los estudiantes inscritos en la UNSAAC entre los semestres 2014-I y 2018-I constituyen la población de estudio, que es el conjunto de variables a evaluar. Un total de 12,698 estudiantes inscritos en la UNSAAC a través de diversas modalidades conformarán la muestra. La técnica de investigación fue las encuestas a los postulantes. La técnica de recolección de datos se hizo la preparación y depuración de la data. Como resultados de la investigación tenemos utilizando el enfoque CRISP-DM (Cross Industry Standard Process for Data Mining), procesaremos, analizaremos e interpretaremos los datos proporcionados por el Centro de Cálculo. Como se muestra en la revisión bibliográfica, la conclusión del proyecto es que el rendimiento académico de los estudiantes es una cuestión compleja que depende de diversas variables, incluidos factores sociodemográficos y socioeconómicos, así como otros elementos como el estado emocional de los estudiantes y su familia. A pesar de ello, es posible predecir el rendimiento académico a partir de los datos de admisión o ingreso en la UNSAAC, mediante algoritmos de aprendizaje automático que alcanzan una eficacia del 69%.

Yamao (2018) en su tesis tuvo como objetivo principal: uso de minería de datos para predecir el rendimiento académico en estudiantes de primer ciclo de la Escuela Profesional de Ingeniería Informática y de Sistemas de la Universidad de San Martin de Porres, empleando una metodología de enfoque cuantitativo, nivel explicativo-correlacional y de diseño transaccional del tipo correlacional-causal, considerando una población a los estudiantes ingresantes a la carrera de Ingeniería de Computación y Sistemas de la Universidad de San Martín de Porres. Mediante un muestreo probabilístico, 1304 ingresantes. El instrumento fue la información histórica extraída de las bases de datos de la Oficina de Admisión y de la Facultad de Ingeniería y Arquitectura de la Universidad de San Martín de Porres. Como resultado, Se tomaron datos de 1304 participantes que se dividieron en tres grupos en función de criterios sociales, económicos y académicos. Se utilizaron tres métodos de predicción diferentes: regresión lineal, árboles de decisión y máquinas de vectores de apoyo. Los árboles de decisión obtuvieron los mejores resultados (82,87%). Las siguientes variables tuvieron los mayores efectos sobre el rendimiento académico: la nota del examen de acceso, el sexo, la edad, el método de ingreso y el tiempo de desplazamiento al centro de estudios desde el domicilio. Se pudo anticipar el rendimiento académico de los solicitantes utilizando la minería de datos. Esto permitió identificar a los solicitantes que podrían tener dificultades académicas.

Puga & Torres (2023) en su tesis tiene como objetivo estudiar el rendimiento académico de los estudiantes de Ingeniería de Sistemas de la Universidad Nacional de la Amazonía Peruana según las predicciones de las Redes Neuronales Artificiales. La Facultad de Ingeniería de Sistemas e Informática de la Universidad Nacional de la Amazonía Peruana realizó una investigación sobre la predicción del rendimiento académico en la asignatura de álgebra lineal. El objetivo del estudio era descubrir si las técnicas de aprendizaje automático podían aumentar la precisión a la hora de identificar a los estudiantes aprobados y reprobados. En la investigación de predicción se utilizaron todos los datos electrónicos accesibles y se emplearon redes

neuronales artificiales y herramientas analíticas, incluidas MS Excel y MATLAB Neural Network Toolbox. Con un error porcentual de 2,083 y un coeficiente de corrección de error de 0,196274, los resultados mostraron una precisión del 97,6 %, una integridad del 100 % y una precisión del 97,9 %, superando los de investigaciones anteriores. En resumen, el estudio superó investigaciones anteriores al demostrar la eficacia de los enfoques de aprendizaje automático para pronosticar el rendimiento académico en el campo del álgebra lineal.

2.1.2. Antecedentes internacionales.

Fabara Sarmiento & otros (2022) en su investigación que realizó tenía el objetivo de desarrollar una aplicación de ciencia de datos que prediga el éxito académico de los niños de primaria para evaluar la eficacia de las técnicas de enseñanza en línea. Este proyecto utilizó el enfoque del PMI, que define el ciclo de vida de un proyecto como la secuencia de etapas que atraviesa desde su concepción hasta su finalización. La división del trabajo en fases da al control una base formal. Cada fase comienza oficialmente con una descripción de lo que está permitido y lo que se espera de ella. 6 millones de niños y adolescentes de entre 0 y 17 años componen la población juvenil del país. De ellos, el 33% son menores de 5 a 11 años y necesitan ayuda para realizar trabajos virtuales. De estos jóvenes, el 50% vive cerca de la costa, el 33% en la sierra y el 7% en la Amazonia. Hacer una comparación de los empleados que se arriesgan y mejoran su técnica de estudio, viendo como resultado una buena base de conocimientos para su próximo curso académico, teniendo en cuenta los resultados del análisis de la ciencia de datos, puede ayudar a resolver este problema. El uso de grandes cantidades de información para análisis o predicciones, como se muestra en este trabajo, es uno de los beneficios del uso de la ciencia de datos, que se concluye que es una disciplina que incluye técnicas para el tratamiento y modelado de datos. Esto es muy útil para el desarrollo de modelos predictivos que puedan aplicarse con fines educativos.

Rico Páez (2022), en su revista publicada el estudio pretende crear modelos de

predicción progresiva del rendimiento académico de los estudiantes de una institución mexicana y evaluarlos mediante diversos métodos de aprendizaje automático. En este trabajo se presenta un enfoque que se basa en las calificaciones de los estudiantes en las tareas académicas realizadas mientras están inscritos en un curso de una universidad pública de México. La metodología consiste en hacer una predicción con un modelo predictivo creado a partir de la actividad 1, luego hacer una predicción con un modelo creado a partir de las actividades 1 a 2, y así sucesivamente hasta utilizar las actividades desde la 1 hasta la n. En este estudio participaron 260 estudiantes de una universidad pública de México y se emplearon 14 ejercicios en clase. A lo largo del curso, estas tareas se completan en tiempos comparables. Cualquier acción que reciba una calificación entre 6 y 10 se considera autorizada (A); en caso contrario, se considera suspensa (R), y también se considera aprobada (NP) si el alumno no presenta la actividad. Así, se crea una base de datos con 260 registros y 15 columnas (atributos). Esta tabla se utilizará para crear los modelos de predicción. El resultado final es un modelo de predicción creado mediante aprendizaje automático y un conjunto de datos conocidos como datos de entrenamiento. Naive Bayes, k vecinos más cercanos y el árbol de decisión C4.5 son los algoritmos de aprendizaje automático empleados en el conjunto de entrenamiento de este estudio, compuesto por 260 registros de una tabla (Hernández, Ramírez y Ferri, 2004). Todos los análisis de datos de este artículo se realizaron con la ayuda del programa gratuito Weka. El estudio produjo un modelo de predicción del rendimiento académico de los estudiantes basado en su actividad académica relacionada con el curso. Se demostró cómo crear modelos de predicción progresiva del rendimiento académico de los estudiantes de una universidad de México para realizar estas previsiones. Estos modelos se crearon utilizando 260 datos de estudiantes y métodos de aprendizaje automático, incluyendo naive Bayes, k vecinos más cercanos y árboles de decisión C4.5.

Molina & Fuentes (2021) en su artículo tiene como objetivo analizar estudios de los últimos diez años que abordan la predicción del rendimiento académico en dos situaciones: las modalidades de aprendizaje en línea y semipresencial, y el apoyo tecnológico a las modalidades presenciales. Como resultado, se mapean los cuatro campos computacionales de la analítica del aprendizaje, el aprendizaje automático, la minería de datos educativos, las redes neuronales artificiales y las teorías difusas. Se empleó una metodología de revisión que consta de tres fases: planificación (preguntas de investigación, criterios de selección y exclusión, una estrategia de búsqueda y un protocolo de revisión); realización (tres evaluadores evaluaron y extrajeron datos de los estudios); y comunicación de los resultados (análisis de la validez de los estudios primarios y esbozo de los datos). Los resultados obtenidos están relacionados con el uso de estrategias de minería de datos basadas en algoritmos genéticos difusos, redes neuronales y árboles de decisión; sin embargo, aún están latentes las pedagogías emergentes que orienten cómo interpretar los resultados de la predicción y la intervención pedagógica. Los resultados obtenidos, que concuerdan con estudios anteriores, están relacionados con estas estrategias de minería de datos. La conclusión del artículo es un análisis exhaustivo de cómo la tecnología educativa predice el rendimiento académico. Para esta predicción, la analítica del aprendizaje, el aprendizaje automático, la minería de datos educativos, las redes neuronales artificiales y las teorías difusas fueron los enfoques y algoritmos más empleados.

Contreras y Fuentes (2020) el objetivo de este estudio fue identificar el mejor escenario para predecir el desempeño de los estudiantes de Ingeniería Industrial de la Universidad Distrital Francisco José de Caldas, utilizando diversas técnicas para determinar cuáles de las 30 variables iniciales eran las más importantes para determinar su desempeño e implementar modelos utilizando algoritmos de clasificación (árbol de decisión, KNN, SVC y Perceptron). Se planteó una metodología para resolver el problema, que resultó ser tan complejo como determinar qué factores podían incidir en él. La metodología empleada en el estudio se expone

de manera resumida en los siguientes 7 pasos: (1) Participantes; (2) Variables; (3) Tratamiento de datos; (4) Estadísticas (5) Selección de Características; (6) Algoritmos de predicción; y (7) Métricas de evaluación. Los resultados son bastante exactos, con buena precisión y métrica de los modelos propuestos, aunque sin tener en cuenta todos los factores que, según la literatura, influyen en la ocurrencia. Se puede plantear la siguiente conclusión principal, es necesario utilizar las variables de otros factores que inciden en el rendimiento académico, como la gestión académica universitaria, los factores tecnológicos, bibliotecarios, institucionales, pedagógicos e intelectuales, para mejorar los resultados en las métricas de evaluación de los algoritmos y elevarlos por encima del 90%. Esto se hace sin comprometer los resultados anteriores, que examinaron las características académicas, demográficas y socioeconómicas preuniversitarias. Los resultados muestran que los mejores resultados los obtienen los algoritmos Stacking y Blending, con valores de precisión en cada semestre que varían entre 85% y 75% para entrenamiento y testing, respectivamente.

Quiñones y Carrasco (2020) El objetivo de este estudio fue predecir el éxito académico de los estudiantes de la Carrera de Ingeniería en Industrias Alimentarias de la Universidad Nacional de Jaén (UNJ). La metodología que se uso es CRISP-DM, y se utilizó el programa informático Weka para pronosticar los resultados de tres métodos de clasificación. Para obtener la matriz de datos se utilizaron el archivo y las oficinas de la institución. Una de las conclusiones fue que InfoGainAttributeEval de Weka permitía evaluar el valor de una variable calibrando la información obtenida con respecto a la variable de rendimiento académico. Concluye que Los tres algoritmos de minería de datos J48graft, J48 y PART, que se identificaron con ayuda del software Weka con un porcentaje de clasificación correcta superior al ochenta y tres (83%), nos permitieron identificar tres reglas para medir el éxito académico de los estudiantes.

Jiménez (2018) en su investigación tuvo como objetivo principal: Con los datos suministrados por el Ministerio de Educación Nacional en el campo de la ingeniería, identificar

las características sociales, económicas y demográficas más asociadas a los resultados promedio de los módulos genéricos de las pruebas Saber Pro a través de minería de datos. Empleando una metodología de enfoque cuantitativo. Sobre la base de un sistema de votación que utiliza métodos de análisis de correlación, ACP, árboles de decisión y reglas de asociación, se ha utilizado la minería de datos para encontrar patrones que puedan relacionarse con la puntuación de los módulos de lectura crítica, comunicación escrita, competencias ciudadanas, razonamiento cuantitativo e inglés en 26 variables económicas y sociodemográficas. Se descubre que el número de personas a cargo o ser cabeza de familia, el tipo de educación, si el hogar es regular o permanente, la reputación académica de la institución de educación superior, y si el hogar posee o no una motocicleta y un microhorno de gas son los factores que más influyen en los resultados de todos los módulos. Además, se observan variaciones notables entre las variables que influyen en los resultados de los distintos módulos. En particular, los estudiantes que pagan matrículas más caras obtienen mejores resultados en inglés, mientras que los que pagan matrículas más baratas obtienen mejores resultados en lectura crítica; el elemento de género influye en la comunicación escrita. A partir de las variables seleccionadas se construye un sistema de predicción del rendimiento con un porcentaje de casos correctamente identificados del 81,1159% y un ROC superior a 0,7. También se realiza un perfilado de los datos para identificar los tipos de estudiantes que realizaron las pruebas en los datos evaluados.

2.1.3. Estado del arte.

Describe Méndez & Otros (2020) en el campo de la inteligencia artificial conocido como aprendizaje automático, se utilizan algoritmos y herramientas para analizar y aprender de grandes volúmenes de datos históricos y sintéticos con el fin de producir predicciones y juicios que puedan, por ejemplo, ayudar en la toma de decisiones. En segunda instancia para Contreras y Fuentes (2020), La recopilación, el análisis y la difusión de datos sobre los agentes educativos con el fin de comprender y optimizar los componentes asociados del proceso de enseñanza-aprendizaje puede denominarse aprendizaje automático (machine learning, ML) aplicado a la educación.

El objetivo de Cruz y otros (2020) es comparar la eficacia de este método de aprendizaje automático con otro basado en expresiones regulares. las expresiones de negación y especulación de la colección de documentos.

En segundo lugar, Rico & Gaytán (2022) deducen que su estudio tiene como objetivo desarrollar una técnica para pronosticar el éxito académico utilizando características de los estudiantes de ingeniería de nuestra nación y comparar los modelos utilizando diversas medidas de evaluación. En este estudio participaron 228 estudiantes de una universidad pública de México. Los modelos de predicción se crearon utilizando tres técnicas de aprendizaje automático a partir de los datos que se recogieron al inicio del curso. Una vez examinadas las propiedades de cada modelo, se alcanzó una precisión de predicción de aproximadamente el 65%. También en su revista Páez & Ramírez (2022) Utilizando técnicas de aprendizaje automático y actividades académicas tempranas de estudiantes universitarios, el objetivo principal es construir y evaluar modelos predictivos del rendimiento académico.

En su metodología de Estrada & Fuentes (2022) Se aplicó el protocolo PRISMA, que consta de tres etapas: planificación (temas de investigación, criterios de selección y exclusión, estrategia de búsqueda y metodología de la revisión); realización (evaluación y extracción de

datos por tres revisores); y comunicación de resultados (evaluación de la validez de los estudios primarios y descripción de los datos). Otro punto de Rico & Gaytán (2022), afirma que el enfoque que se va a crear puede aplicarse a distintos cursos, y las características de los alumnos pueden recogerse antes del inicio del curso o antes, lo que abre la puerta a la aplicación de estrategias de intervención para los alumnos que corren peligro de reprobar.

Según Contreras (2021), Los resultados demuestran que la analítica del aprendizaje en la enseñanza superior se concentra en el resultado o rendimiento de los estudiantes que utilizan la modalidad virtual (E-Learning). Dado que lo que se necesita es la información que diversos dispositivos, plataformas y bases de datos de las instituciones han registrado o capturado en el trabajo académico, este tipo de técnica también puede concentrarse en la aplicación en otro tipo de modalidad. Por otro lado, Rico Páez (2022), todos los resultados de los análisis de datos mostrados en este artículo se realizaron con el apoyo del software libre Weka

Concluye Rojas (2021) que se logró determinar los factores más importantes en el modelo de clasificación para el bajo rendimiento académico de los estudiantes del Instituto de Educación Superior Pedagógico Publico Juliaca utilizando Machine Learning a través de la Metodología CRISP-DM y el algoritmo Random Forest Classifier: Los dos factores más determinantes en este momento para pronosticar el bajo rendimiento académico de los estudiantes son la variable edad y el promedio del examen final de admisión, que en conjunto tienen un valor combinado de 13% y 20%, respectivamente..

Definen Contreras y Fuentes (2020), de acuerdo al trabajo que presentan y a los resultados obtenidos, se plantea la siguiente conclusión principal, La edad, el sexo, la puntuación ICFES para la aptitud matemática, la puntuación ICFES global, el coste de la matrícula, la puntuación ICFES para la condición matemática y la cohorte son los siete factores de los examinados que más influyeron en el éxito académico de los estudiantes de ingeniería. También Molina & Fuentes (2021), En su estudio ofrecen un análisis exhaustivo de cómo la

tecnología educativa puede predecir el rendimiento académico. Para esta predicción, la analítica del aprendizaje, el aprendizaje automático, la minería de datos educativos, las redes neuronales artificiales y las teorías difusas fueron los enfoques y algoritmos empleados con más frecuencia.

En este artículo, Cruz & otros (2020) Describir un sistema de aprendizaje automático que pueda encontrar escepticismo y lenguaje especulativo en la literatura biomédica, particularmente en la colección de documentos BIOS cope.

Expresan Estrada & Fuentes (2022), El rendimiento académico es actualmente una preocupación, y refleja en parte lo bien que se lleva a cabo el proceso de formación y cómo ello afecta al aprendizaje. Se conceptualiza desde los puntos de vista social (definido por los rasgos de los entornos sociales y educativos), demográfico (diversidad y complejidad del alumnado) e individual (dominio de la información, habilidades, capacidades, etc.) de acuerdo con su naturaleza. Sirve como métrica crucial para determinar el grado de dominio del aprendizaje en general. El rendimiento académico está influido tanto por aspectos ambientales (escuela, familia, entorno social, habilidades del profesor y del tutor, entre otros) como internos (motivación, interés por la materia), aunque los alumnos son los responsables últimos de su aprendizaje. La aplicación de intervenciones didácticas preventivas será posible si se identifica previamente a los alumnos que pueden estar predispuestos a un bajo rendimiento académico.

También Páez & Ramírez (2022) mencionan que uno de los usos más intrigantes de las técnicas de aprendizaje automático es la predicción del rendimiento académico, ya que permite la detección temprana y la corrección de problemas académicos. Estos incluyen, entre otras cosas, el comportamiento del alumno y la presencia inspiradora del profesor.

De acuerdo con Rico & Gaytán (2022) la mayoría de los criterios empleados en el estudio, principalmente para identificar a los alumnos con riesgo de suspender, se ajustaron mejor al modelo Nave Bayes.

Ahora como aporte sobre la investigación de diferentes artículos y revistas científicas se predice

que el rendimiento académico es un componente crucial de la educación que permite a los profesores crear acciones didácticas preventivas. En este proceso predictivo intervienen diversas disciplinas, entre las que destacan las redes neuronales artificiales, la analítica del aprendizaje, el aprendizaje automático, la minería de datos educativos y las teorías difusas.

2.2. Bases teóricas.

2.2.1. Aprendizaje automático.

Las empresas ya utilizan la tecnología creada como resultado del avance de la inteligencia artificial, y esta tendencia continuará inevitablemente dado el estímulo que numerosas instituciones, tanto académicas como empresariales, están dando al estudio en esta área. El aprendizaje automático es una de las disciplinas de la inteligencia artificial que mantiene ocupados a los académicos. Este campo trata los datos de forma multidisciplinar utilizando diversos métodos, como algoritmos matemáticos, estadísticos y poco ortodoxos. (Ramírez Veliz, 2019).

2.2.2. ¿Qué es el aprendizaje automático?

En la actualidad, el aprendizaje automático se basa en representaciones del conocimiento obtenidas mediante procedimientos matemáticos y teoría estadística empleados junto con el procesamiento computacional de cantidades masivas de datos en bases de datos. En un principio, el aprendizaje automático se asoció a teorías cognitivistas y enfoques simbólicos. A pesar de que existen diferentes métodos de aprendizaje automático, el "aprendizaje profundo" es ahora el más popular y con frecuencia hace uso de "redes neuronales". El aprendizaje automático es, por tanto, uno de los principales candidatos para crear clases de referencia neutrales y objetivas debido a su conexión con las matemáticas y el procesamiento algorítmico de datos estadísticos. (Guersenzvaig Elisava & Casacuberta, 2022,p.45).

2.2.3. ¿Quiénes utilizan el aprendizaje automático?

Dado que los datos históricos que se almacenan son tan valiosos para su empresa y les ayudan a seguir siendo competitivos y a tener ventaja sobre sus rivales, la mayoría de las organizaciones con muchos datos registrados utilizan actualmente el aprendizaje automático. (Rojas Pari, 2021,p.30).

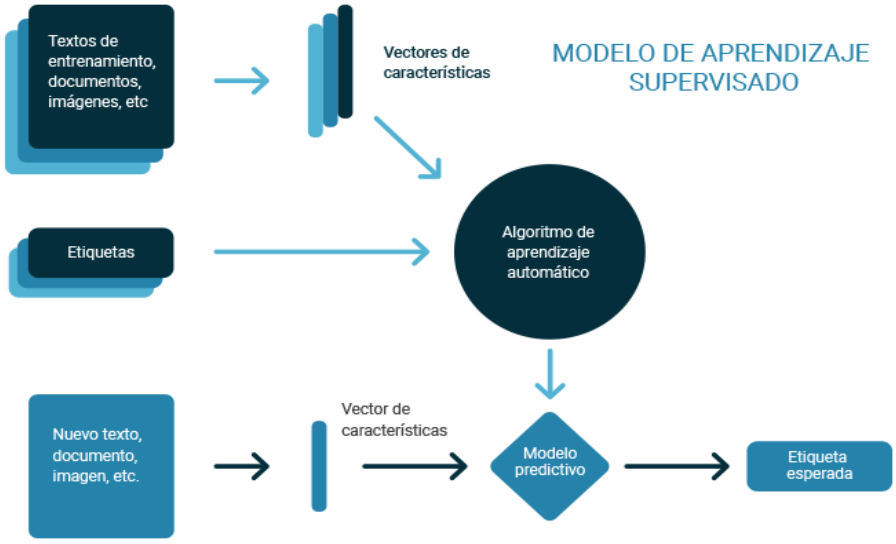
El rápido aumento de la popularidad del aprendizaje automático está directamente relacionado con la disponibilidad de conjuntos de datos más amplios y diversos, la reducción del coste y el aumento de la potencia de la informática, y la simplicidad del almacenamiento de datos que hace posible la computación en nube. Big Data es una tecnología que ha surgido en los últimos años como resultado de la producción y acumulación de cantidades masivas de datos procedentes de muchas fuentes por parte de diversas industrias. La posibilidad de extraer conocimientos significativos de esta información para ayudar a empresas y gobiernos a tomar mejores decisiones ha aumentado la demanda de herramientas que puedan procesarla. El aprendizaje automático ha demostrado ser una de las herramientas más eficaces. (Nieto Jeux, 2021,p.2).

El aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo son los tres tipos de aprendizaje reconocidos por el aprendizaje automático. En los párrafos siguientes se describen cada uno de ellos. (Ramírez Veliz, 2019).

En el **aprendizaje supervisado,** los desarrolladores de un sistema lo entrenan creando un conjunto de resultados de salida previstos para una serie de datos de entrada etiquetados. Una vez entrenado el modelo, el sistema puede asociar una etiqueta de salida a un nuevo valor. Los usuarios o programadores del sistema pueden utilizar sus propios datos para seguir entrenando el modelo e indicarle si la etiqueta asignada es correcta. (Guersenzvaig Elisava & Casacuberta, 2022,p.46).

Figura 1

El aprendizaje supervisado

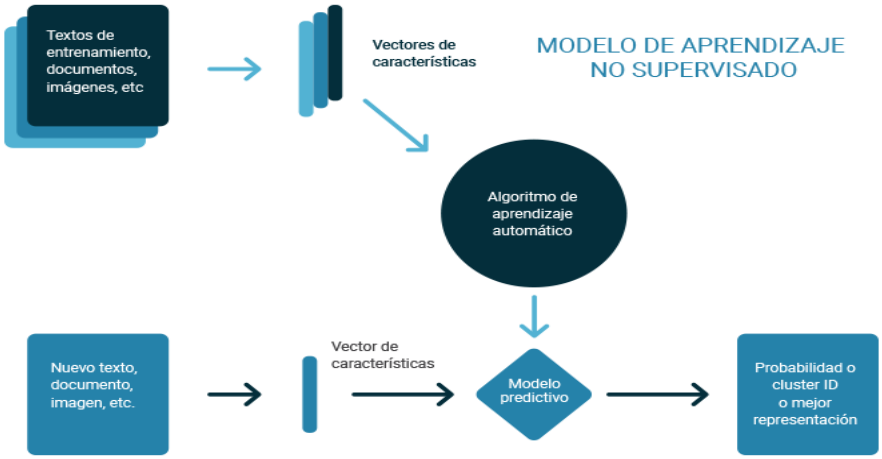


*Nota***:** Adaptado de diagrama de flujo del aprendizaje supervisado, de Javier Luna González, 2018.

El aprendizaje no supervisado, se basa en la premisa de que el propio sistema encuentra y reconoce de forma autónoma los patrones existentes en los datos (correlaciones estadísticas), en lugar de depender de que los programadores etiqueten los datos de entrenamiento o indiquen con precisión los resultados previstos. El clustering es un enfoque popular en el que se crea un grupo de clusters minimizando o maximizando un criterio de optimización. (Guersenzvaig Elisava & Casacuberta, 2022,p.46).

Figura 2

El aprendizaje no supervisado.

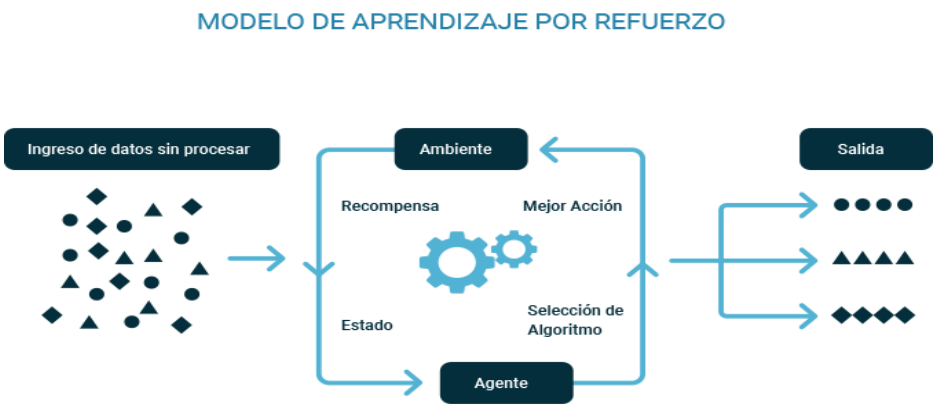


***Nota*:** Adaptado de diagrama de flujo del aprendizaje no supervisado, de Javier Luna González, 2018.

El Aprendizaje por refuerzo, El programa debe realizar una determinada tarea en relación con un entorno cambiante. Aprende por ensayo y error. Explora diversas opciones y es recompensado si sus esfuerzos hacen avanzar el objetivo y penalizado en caso contrario. Este método se emplea para enseñar a los ordenadores a jugar a videojuegos. (Nieto Jeux, 2021,p.6). Según Nieto (2021)**.**

Figura 3

Modelo de Aprendizaje por refuerzo.



***Nota*:** Adaptado de diagrama de flujo aprendizaje por refuerzo, de Javier Luna González, 2018.

Los principales algoritmos del aprendizaje automático son:

a) No probabilístico.

K- vecinos más cercanos (k-NN) es uno de los algoritmos de aprendizaje automático más utilizados. Su sencillez y excelente eficacia son sus principales ventajas. El concepto clave es que las etiquetas de los k datos más cercanos determinan la etiqueta del dato x objeto de estudio. Como resultado, la etiqueta dada a x es la que más aparece en los k datos más cercanos. El valor del parámetro k puede elegirse a voluntad. El algoritmo k- vecinos más cercanos se utiliza para resolver tareas de clasificación y regresión (Nieto Jeux, 2021,p.12). También Ramírez (2019) expresa que con K-Nearest Neighbors (KNN) se pueden resolver problemas de clasificación y regresión. Es un miembro de la familia del aprendizaje perezoso, que no requiere instrucción formal. Se guardan los datos iniciales y se clasifican los datos de entrada localizando las k muestras más cercanas. El grado de parentesco entre dos muestras puede determinarse utilizando diversas distancias o métricas de similitud. (p,10).

Los árboles de decisión son técnicas de aprendizaje automático muy utilizadas. Esto se debe principalmente a que son sencillas, intuitivas y visualmente atractivas. Además, los resultados son precisos, no exigen mucho trabajo de procesamiento y pueden gestionar grandes volúmenes de datos. Sin embargo, los principales inconvenientes de los árboles incluyen el sobreajuste (cuando el modelo no generaliza bien los datos porque se ha visto excesivamente influido por el ruido y los detalles insignificantes de los datos de entrenamiento), la alta varianza (cuando un pequeño cambio en los datos provoca un cambio significativo en el árbol) y el sesgo (cuando algunas clases tienen una influencia excesiva en el modelo y degradan los resultados). (Nieto Jeux, 2021,p.13). los árboles de decisión se componen de tres tipos de nodos:

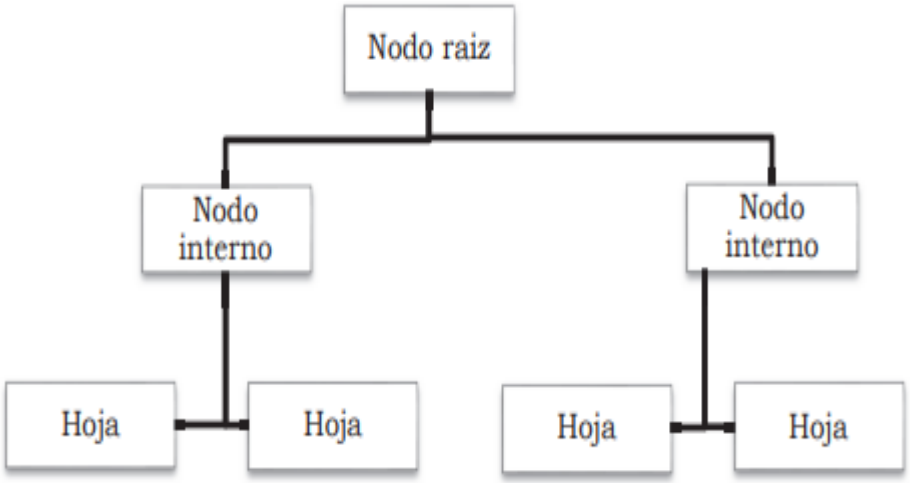
- **Nodo raíz**, es el nodo superior y el nodo inicial para hacer una división. Sólo tiene ramas de salida como resultado. Se dividen en dos o más sub-nodos, que representan los resultados potenciales de aplicar una función a un atributo [30]. Una categoría, un

valor o un rango de valores son los resultados. (Nieto Jeux, 2021).

- **Nodo interno**, son los nodos con ramas de entrada y salida. Las ramas de salida reflejan las respuestas a un atributo, al igual que el nodo raíz. (Nieto Jeux, 2021).
- **Hojas**, se sitúan estos nodos, que están en la parte inferior del esquema y sólo tienen ramas de entrada. Son todos los posibles resultados del árbol y, por tanto, del algoritmo que podrían haberse producido.. (Nieto Jeux, 2021).

Figura 4

Componentes de los árboles de decisiones.



***Nota*:** Elaboración propia.

b) Probabilístico.

- El **Análisis Discriminante,** es una técnica de categorización. Sus principales ventajas son la rapidez, la solidez y la sencillez. La ubicación de la frontera entre varias clases se determina mediante esta técnica utilizando los datos disponibles en ese momento. La densidad de cada clase se calcula utilizando muestras bajo el supuesto de gaussianidad. Por consiguiente, el vector de la media y la matriz de covarianza de cada clase deben utilizarse para estimar la distribución normal multivariante. Permite determinar el centro y el aspecto de la distribución. El conjunto de lugares en los que las probabilidades entre las clases son iguales puede

representarse tras aplicar este procedimiento a todas las clases. Como resultado, se descubre una ecuación que describe ese límite. La clase con mayor densidad está formada por las observaciones frescas que caen a ambos lados de esa ecuación. Por lo tanto, el procedimiento para utilizar este método consiste simplemente en determinar el límite entre cada clase y calcular la media y la covarianza de cada clase. (Nieto Jeux, 2021,p.23-24).

- La **regresión logística,** es una técnica de clasificación para el análisis predictivo. Debido a su éxito en la realización de clasificaciones utilizando datos continuos, es especialmente apreciada. El vínculo entre una variable de categoría que debe predecirse y uno o varios factores predictores se revela mediante la regresión logística. Se emplea con frecuencia para predecir el resultado de una variable binaria. Pronostica la probabilidad de caer en cada clase. (Nieto Jeux, 2021,p.24).

- Los **clasificadores basados en redes bayesianas,** son usados en los que existe una incertidumbre inherente. El objetivo es determinar en qué medida afectan estas evidencias a la distribución probabilística de las variables desconocidas (incertidumbre) en los casos en que se conocen determinadas variables (evidencias). La principal ventaja de estos clasificadores es que dan al conocimiento ambiguo una representación explícita, pictórica y comprensible. Estos clasificadores son cada vez más comunes, lo que resulta alentador. Se distinguen los clasificadores de redes bayesianas continuas y los clasificadores de redes bayesianas discretas. El clasificador de redes bayesianas discretas más sencillo es el Bayes ingenuo. Según la importancia de sus atributos, categoriza las ocurrencias. (Nieto Jeux, 2021,p.25).

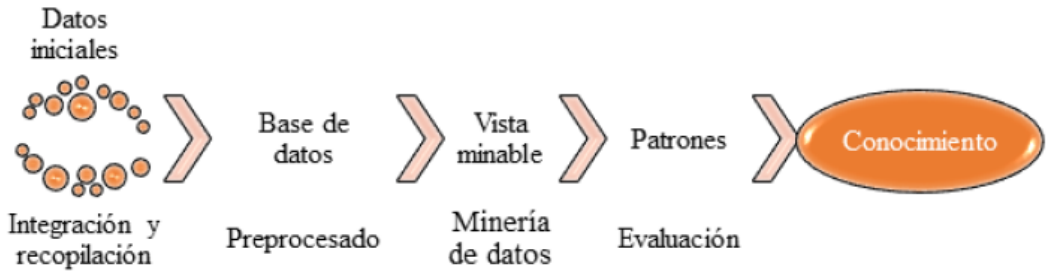
- **Métodos Bayesianos.** Para interpretar los datos observados y obtener una probabilidad a posteriori, la inferencia bayesiana combina los datos observados con los conocimientos previos en forma de probabilidad a priori. No nos asegurará que podamos llegar a la solución correcta, sino que nos dará la probabilidad de que cada una de las posibles soluciones sea correcta. Después, podemos utilizar este conocimiento para determinar qué respuesta es más

probable que sea la correcta. En otras palabras, nos permite hacer conjeturas basadas en los hechos disponibles. (Ramírez Veliz, 2019,p.9).

– **El conocimiento en bases de datos y algoritmo Naïve Bayes,** La integración y recopilación de datos, que determina las fuentes de información y la forma de reunirlas para crear la base de datos que se utilizará, constituye la etapa inicial del proceso de descubrimiento de conocimientos en bases de datos. El preprocesamiento, la etapa siguiente, consiste en seleccionar, limpiar y transformar los datos para crear el subconjunto que se extraerá o la vista extraíble. La siguiente etapa es la minería de datos, durante la cual se especifica el tipo de trabajo que se va a realizar y el algoritmo que se va a utilizar. La etapa de evaluación, que viene en último lugar, es en la que se establece la fiabilidad y validez de los conocimientos recopilados. (Rico Páez & Sánchez Guzmán, Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN, 2017).

Figura 5

Proceso de descubrimiento de conocimiento en bases de datos.



***Nota*:** Adaptado de Proceso de descubrimiento de conocimiento en bases de datos, de Espinosa et al., 2016.

– **Método bagging.** Un metaestimador conocido como método Bagging de construcción de comités utiliza clasificadores separados que se entrenan simultáneamente a partir del mismo conjunto de datos de entrenamiento, con reemplazos elegidos al azar. Para crear un comité de clasificadores (es decir, determinar cuántos clasificadores emplear), es necesario decidir cómo agregar los resultados de los clasificadores; el enfoque más directo es utilizar un voto

mayoritario para determinar en qué categoría entra un documento la que debe ser seleccionada por la mayoría de los clasificadores.una cantidad inusual (Mounier y otros, 2020)

- **Método Vote.** Son una colección de métodos que permiten fusionar las predicciones de muchos algoritmos de aprendizaje automático que, aplicados por separado, habrían obtenido buenos resultados en su labor de predicción. Utilizando un conjunto de datos de entrenamiento, construye inicialmente dos o más modelos independientes (aprendices de primer nivel). Un clasificador de votación, también conocido como metaaprendiz o aprendiz de segundo nivel, recibe los datos de salida de cada modelo y los utiliza para predecir la clase de la variable de salida. Para crear un único modelo (nivel 1) que se entrenará utilizando estos valores, el clasificador de votación utiliza esencialmente la salida de cada uno de los clasificadores débiles anteriores (modelos de nivel 0). El clasificador de votación, y no los clasificadores débiles (modelos de nivel 0) que se utilizaban anteriormente, realiza las predicciones para una nueva clase. Este grupo incluye las siguientes técnicas: votación estricta y un único tipo de algoritmo base. Múltiples tipos de algoritmos de base y procedimientos de votación estricta. La votación puede hacerse de forma suave o estricta, utilizando varios métodos básicos y ponderaciones. Combinación y ordenación (Contreras Bravo y otros, Revista redipe, 2021)

2.2.4. Minería de datos

Da a conocer Bernuy (2018), que la **minería de datos** es el proceso de seleccionar una gran cantidad de datos para encontrar patrones relacionados. Se considera el resultado inevitable del desarrollo de las tecnologías de la información y de la capacidad del sector de las bases de datos para crear capacidades vitales que incluyen la recopilación y generación de datos, su gestión y métodos analíticos sofisticados. Las bases de datos heterogéneas interconectadas globalmente y las potentes herramientas para el análisis de datos eran necesarias debido al volumen de datos producido por el desarrollo de la tecnología de Internet. Debido a la rápida evolución de la generación de datos, se ha creado una situación que se considera rica en datos,

pero pobre en información. Mediante la creación metódica de técnicas de minería de datos que puedan convertir estos datos en fuentes fiables de conocimiento, pretendemos cerrar esta brecha.

Por otro lado, Alania (2018), indica que la minería de datos es el proceso de extraer información relevante de los datos; se cree que sólo el conocimiento fresco es relevante (Luna Gonzales, 2018). Y también implementa que la minería de datos es el proceso de búsqueda de información importante en grandes cantidades de datos. En él colaboran personas y ordenadores.

Tendencias De La Minería De Datos.

A continuación, Asto (2020) identifica algunas de las siguientes tendencias en el uso de la minería de datos en la investigación:

- Aplicaciones particulares de la minería de datos en los campos de las telecomunicaciones, la banca, el comercio minorista y la biomedicina.
- Aplicaciones para la búsqueda de patrones web mediante minería de datos.
- Minería de datos educativos, que utiliza datos producidos en entornos educativos.
- Minería de datos en tiempo real para la detección de intrusiones en sistemas de seguridad de redes.
- Aplicaciones para la minería de datos procedentes de varias bases de datos dispersas entre sí. Estos conjuntos de datos pueden combinarse de diversas maneras para producir conjuntos de datos originales.
- Aplicaciones de la minería de datos para el conocimiento de bases de datos, la optimización semántica de consultas y la respuesta inteligente a consultas.
- Aplicaciones para la seguridad de la información y la protección de la privacidad.

La minería de datos para la educación.

Con el fin de investigar los datos creados en el entorno educativo, la minería de datos para la educación, una disciplina en desarrollo se centra en el desarrollo dimensional. Esta

información procede de diversas fuentes, como exámenes, herramientas educativas, cursos en línea y aulas tradicionales. Estas fuentes ofrecen continuamente nueva información que puede ser evaluada para responder a consultas que antes eran impracticables, como disparidades en la población estudiantil y cambios en el comportamiento de los alumnos, entre otras. (Yamao, 2018).

Tabla 1

Cuadro de FODA de la Minería de Datos para la Educación (EDM).

Fortalezas	**Debilidades**
• Gran volumen de datos disponibles. • Uso de algoritmos poderosos y validados ya existentes. • Modelos más precisos de los usuarios para la mejora y personalización de los sistemas. • Encontrar momentos críticos y patrones de aprendizaje. • Obtener una visión de las estrategias de aprendizaje y sus resultados.	• Errores en la interpretación de los resultados debido a los factores humanos. • Fuentes de datos heterogéneos. No existe todavía un estándar para los datos. • Los resultados en su mayoría son cuantitativos. Los métodos cualitativos no han brindado resultados significantes. • Sobrecarga de información. Sistemas complejos. • Incertidumbre debido a que solo los docentes o instructores con cierto nivel de habilidades pueden interpretar correctamente los resultados
Oportunidades	Debilidades
• Estandarización de la data y mejora de la complejidad entre las diferente aplicaciones y herramientas. • Aprendizaje multimodal y efectiva. • Capacidad de autoaprendizaje en sistemas inteligentes y autónomos. • Integración de los resultados obtenidos con otros sistemas de toma de decisiones. • Modelos de aceptación, describiendo usabilidad, expectativas, confiabilidad entre otros.	• Aspectos éticos como privacidad de los datos. • Sobre análisis. • Posibilidad de errores en la clasificación de patrones. • Confiabilidad: resultados contradictorios durante la implementación de modelos ya establecidos-

Nota: Elaboración propia.

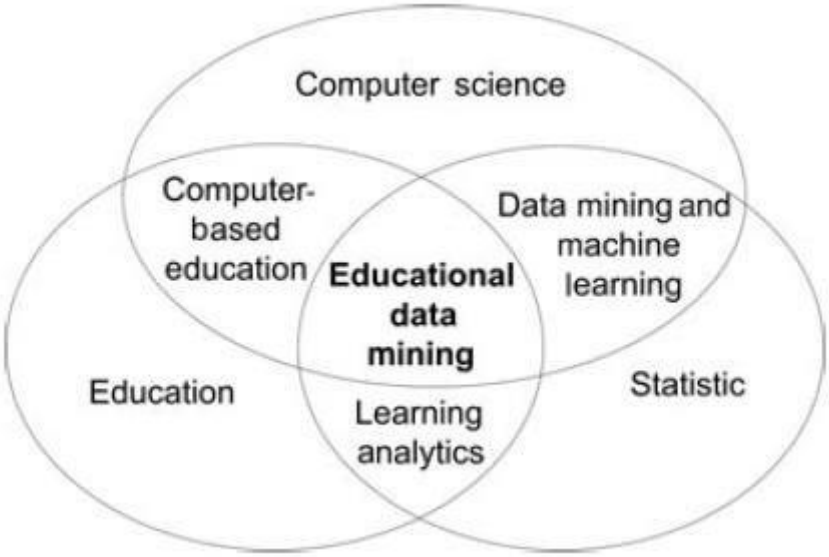
Para Bernuy (2018), la minería de datos para la educación (EDM) es un campo interdisciplinar que abarca, entre otras cosas, el análisis de datos, los sistemas de

recomendación, la psicopedagogía, la psicología del aprendizaje y el análisis de redes sociales. Combina tres disciplinas clave: estadística, educación e informática.

Sostienen Quiñones y Carrasco (2020) que la Minería de Datos Educacional (MDE) es uno de los usos de la minería de datos que permite prestar ayuda inmediata a los estudiantes mediante la predicción de patrones de abandono y aprendizaje. Sin embargo, no se ha aplicado en la enseñanza universitaria. La regresión lineal, la regresión logística, los árboles de decisión, las máquinas de vectores de soporte (SVM), las redes bayesianas, los bosques aleatorios, los algoritmos de reducción de la dimensionalidad y los algoritmos de refuerzo de gradiente son algunos de los algoritmos de minería de datos más populares.

Figura 6

Principales áreas relacionadas con minería de datos.



***Nota*:** Adaptado de Educational Data Mining, de Alaa Khalaf, 2018.

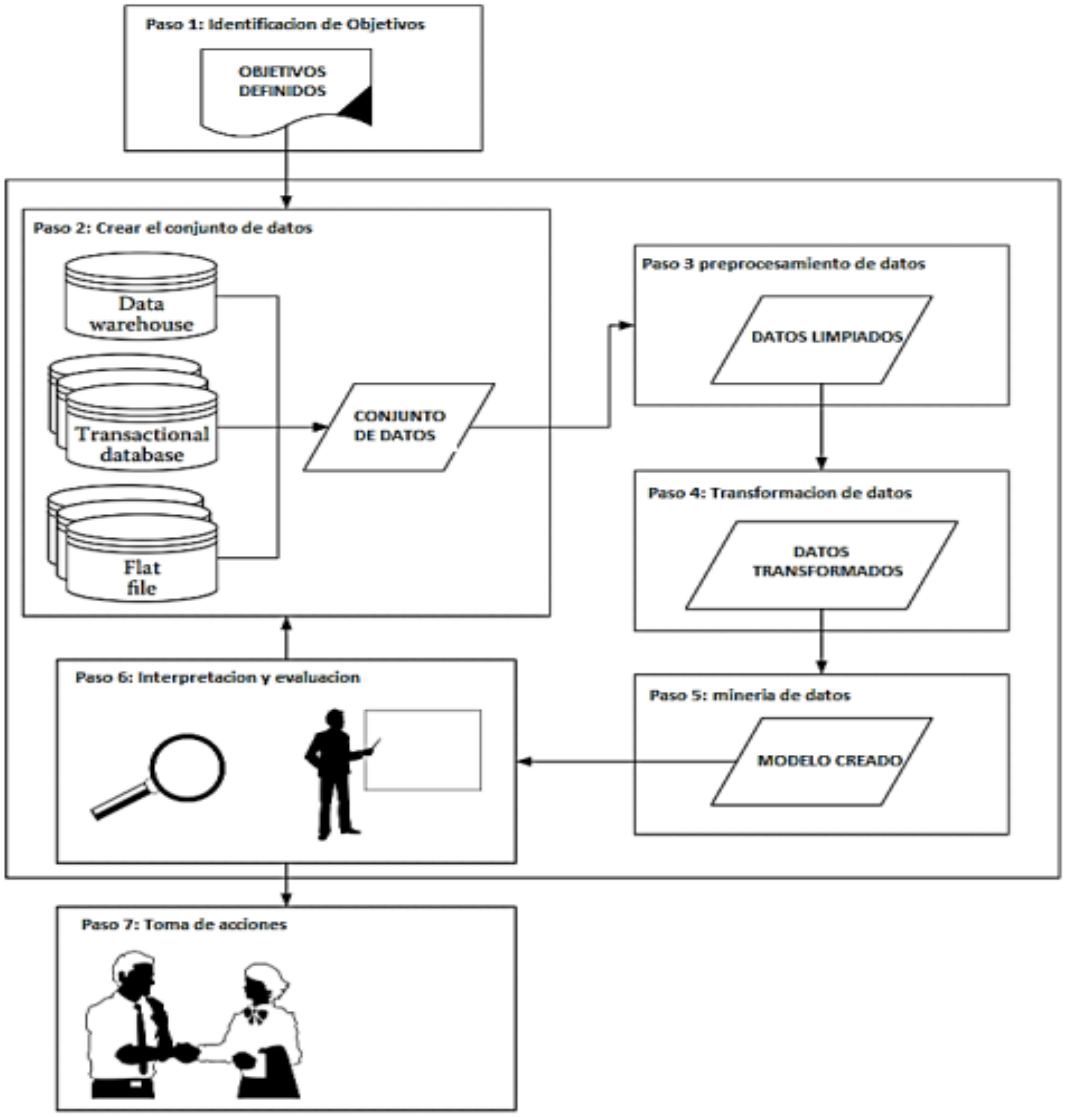
2.2.5. Técnicas De Minería De Datos.

A diferencia del enfoque tradicional, cuyo objetivo era la verificación, en la minería de datos se utilizan nuevas técnicas para tratar de descubrir información útil de forma automática. Estas nuevas técnicas evolucionan constantemente como resultado del avance tecnológico y de la colaboración entre numerosos campos de investigación relacionados con las bases de datos,

el reconocimiento de patrones, la inteligencia artificial, los sistemas expertos, la estadística, la visualización y la recuperación de información. (Asto Rodriguez, 2020).

Figura 7

Técnicas de la Minería de Datos.



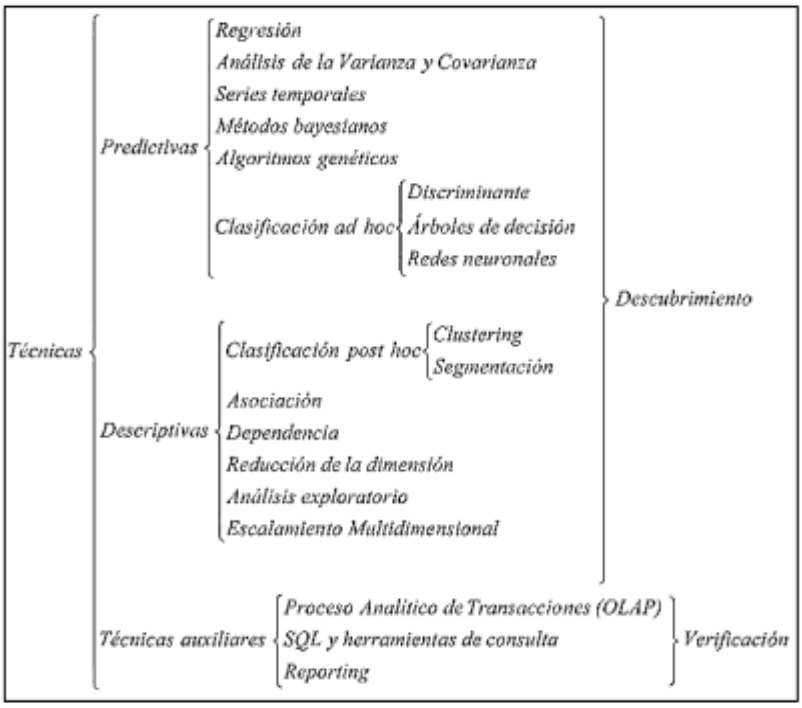
***Nota*:** Adaptado de A seven-step KDD process model, de Theodora Gilbert, 2023.

2.2.6. Clasificación De La Técnicas De La Minería De Datos.

Holgado (2018) los métodos de minería de datos pueden dividirse en tres categorías: auxiliares, descriptivos y predictivos. A continuación, repasaremos estas estrategias con más detalle.:

Figura 8

Clasificación de la técnica de la minería de datos.



Nota: Adaptado de Clasificación de las técnicas de Data Mining, Pérez y Santín, 2008.

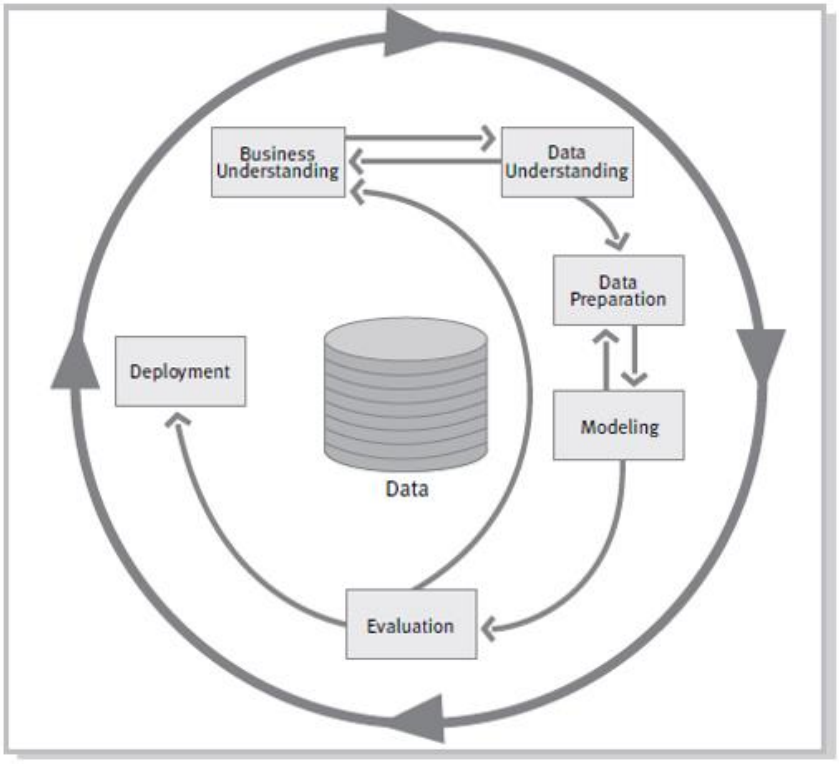
Está claro que los enfoques de categorización pueden incluirse tanto en categorías descriptivas como predictivas. Dado que clasifican a personas u observaciones dentro de agrupaciones previamente establecidas, los enfoques de clasificación predictiva suelen denominarse técnicas de clasificación ad hoc. Dado que ejecutan la clasificación sin especificar previamente los grupos, las técnicas descriptivas suelen conocerse como técnicas de clasificación post hoc. (Holgado Apaza, 2018).

2.2.7. **Metodología CRISP-DM**

Una metodología y un modelo de proceso conocidos como CRISP-DM (Cross-Industry Standard Process for Data Mining) definen un ciclo de seis acciones clave para crear un marco del ciclo de vida de la minería de datos. (Yamao, 2018).

Figura 9

Modelo CRISP-DM.



Nota: Adaptado de Fase de "Comprensión del negocio", de Mikel Niño, 2016.

Las relaciones más significativas y frecuentes entre las fases se muestran en el modelo mediante flechas. El modelo ilustra un proceso iterativo, por lo que el orden de las etapas no es rígido, y se puede ir y venir entre las fases tantas veces como sea necesario. El modelo pretende ser adaptable y puede utilizarse en cualquier circunstancia, independientemente del campo, la herramienta o la aplicación de minería de datos. (Yamao, 2018).

2.2.8. Proceso de extracción de Conocimiento.

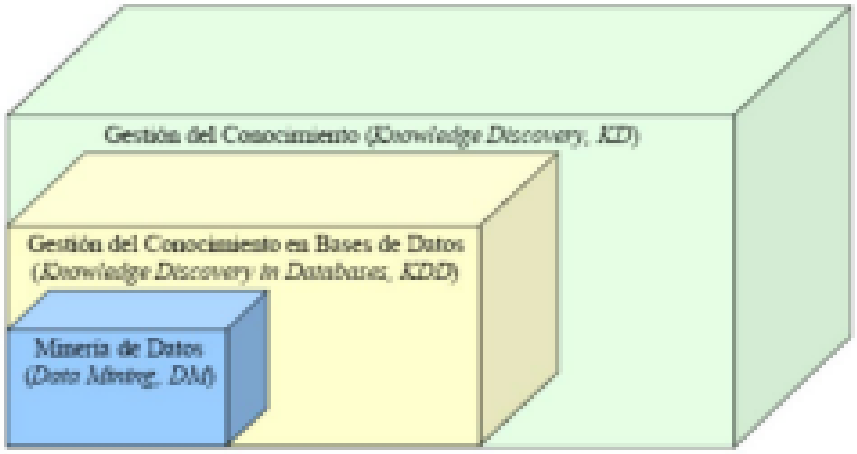
El Proceso de Extracción de Conocimiento (KDD) es un método desafiante para extraer patrones potencialmente beneficiosos, únicos y fiables de los datos. Mientras que la Minería de Datos es la parte de este proceso en la que se emplean métodos de inteligencia artificial para generar un modelo de los datos, lo contextualizan al proceso de encontrar y extraer

conocimiento de las bases de datos. (Alania Ricaldi, 2018).

Hoy en día, la extracción de conocimientos (KDD) y la minería de datos se utilizan a veces indistintamente. Como se muestra en la figura, la minería de datos es una etapa del proceso KDD.

Figura 10

Conceptos de la Minería de Datos.



***Nota*:** Adaptado de (Fleming, 2016).

2.2.9. Fases del Proceso de extracción de Conocimiento.

a) **Selección de datos.** El propósito de este paso es elegir las fuentes y categorías de datos que se utilizarán. En este paso se extraen los datos pertinentes de la(s) fuente(s) de datos (Alania Ricaldi, 2018).

b) **Pre procesamiento.** Como se necesitarán en las fases siguientes, en este paso se preparan y limpian los datos recuperados de las distintas fuentes de datos. Con el fin de establecer una estructura adecuada para modificar posteriormente los datos, se emplean diversos procedimientos para gestionar los datos que faltan, los datos incoherentes o los que están fuera de rango (Alania Ricaldi, 2018).

c) **Transformación.** En esta fase consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables partiendo de las existentes con una

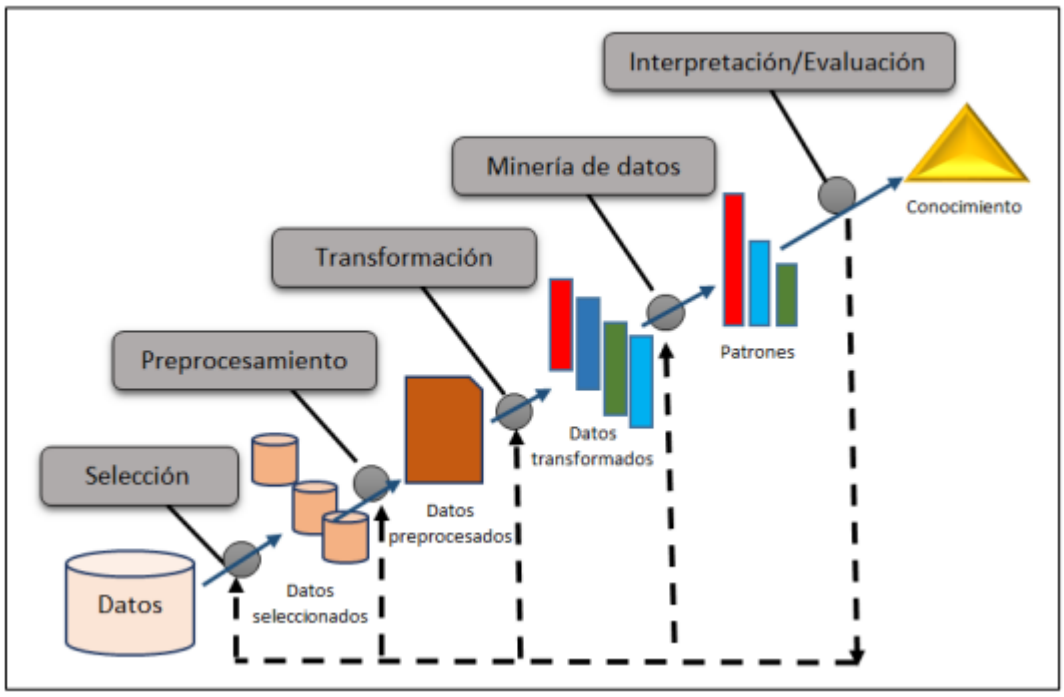
estructura de datos apropiada. En esta fase se realizan las operaciones de agregación o normalización, donde se consolidan los datos de una forma necesaria para la fase siguiente (Alania Ricaldi, 2018).

d) **Data Mining.** Es la fase de modelamiento, en donde métodos inteligentes son aplicados con la finalidad de extraer patrones previamente desconocidos, validos, nuevos y potencialmente útiles que están contenidos u ocultos (Alania Ricaldi, 2018).

e) **Interpretación y Evaluación.** Esta fase consiste en identificar los patrones encontrados, determinar qué patrones son realmente intrigantes, determinar qué patrones se basan en determinadas métricas y llevar a cabo una evaluación de los resultados (Alania Ricaldi, 2018).

Figura 11

Fases del proceso de KDD.



***Nota*:** (Flores Urgilés y otros, 2022).

2.2.10. Software para la Minería de Datos.

La minería de datos es posible gracias a diversas herramientas informáticas, algunas de

las cuales se enumeran a continuación:

- **RapidMiner.**

Es una conocida herramienta de minería de datos y análisis predictivo creada por la empresa del mismo nombre. Es gratuita, de pago y ofrece una interfaz de usuario intuitiva. RapidMiner contiene una colección de más de 1500 operadores que se utilizan para preprocesar, visualizar, convertir, construir, evaluar y optimizar modelos. Emplea un paradigma de flujo de trabajo para desarrollar modelos que resuelvan problemas complicados (Asto Rodriguez, 2020).

Figura 12

RapidMiner mostrando una comparación entre varios algoritmos.

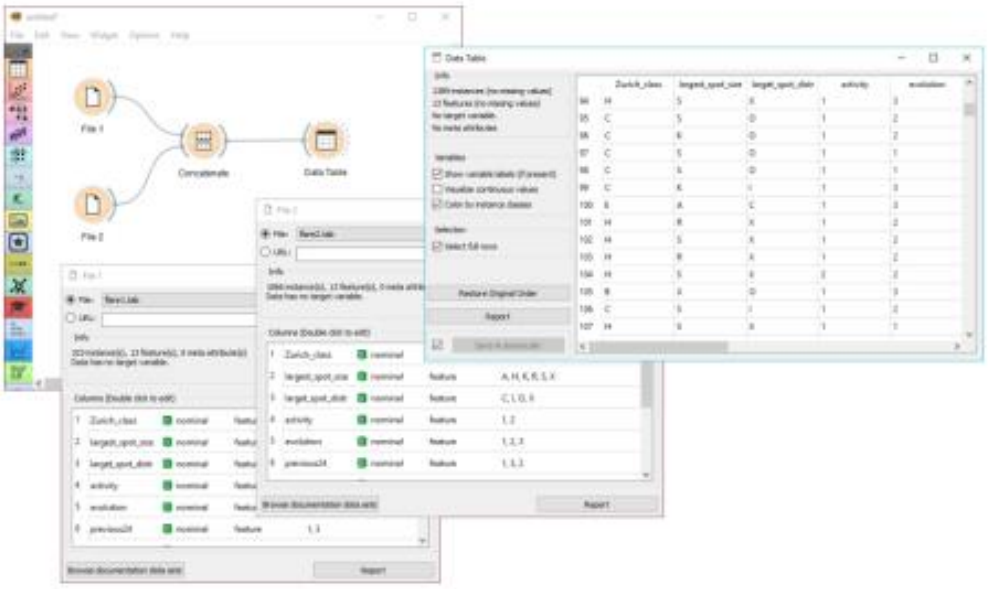
***Nota*:** (Valdivieso y otros, 2019).

- **Orange.**

Tanto novatos como expertos pueden utilizar este programa de aprendizaje automático y visualización de datos de código abierto. Ofrece una sólida caja de herramientas, complementos y un proceso gráfico interactivo para el análisis de datos cualitativos (Orange, 2018). Las diferentes ventanas del programa se representan en el siguiente diagrama (Asto Rodriguez, 2020).

Figura 13

Gráfica de software ORANGE.



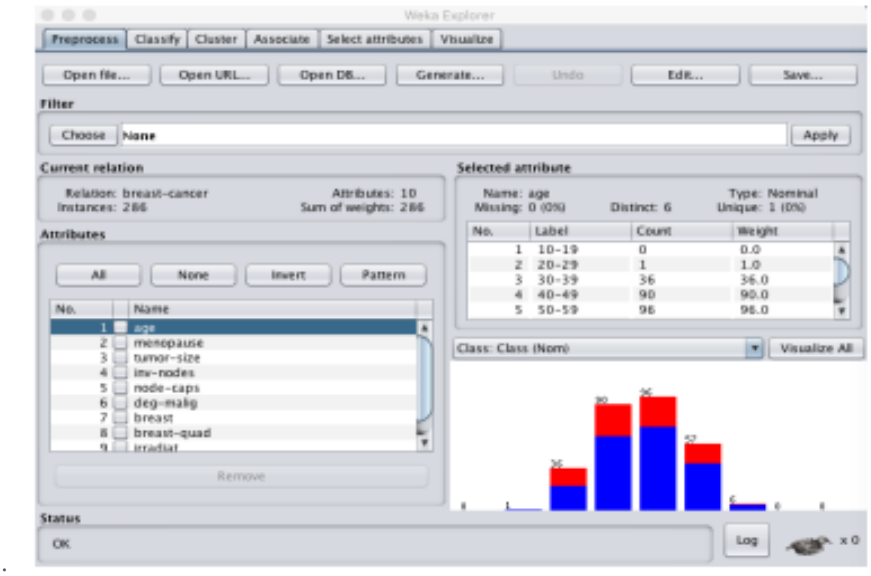
Nota: (Rodriguez & Toda, 2023).

- **Weka.**

Weka (Waikato Environment for Knowledge Analysis) es una herramienta de minería y análisis de datos de propósito general creada en la Universidad de Waikato (Nueva Zelanda). Está escrita en Java y se distribuye bajo la licencia GNU-GLP. Incluye una amplia gama de herramientas, como clasificadores, algoritmos de reglas de asociación, algoritmos de agrupamiento y funciones. Las funciones de preparación y visualización de datos también son buenas (Asto Rodriguez, 2020).

Figura 14

Gráfica de software WEKA



Nota: Pantalla tomada del software weka.

Cuadro comparativo de software de minería de datos

Además de weka existen una gran cantidad de herramientas de minería de datos, cada una con diferentes características, pero todas diseñadas con el mismo objetivo para predicción, en el siguiente cuadro se muestra la comparación de las diferentes herramientas:

Tabla 2

Cuadro comparativo de las herramientas de minería de datos

Características	SAS Enterprise miner	Tariykdd	Weka	Orange	Rapid miner	Clemetine
Licencia libre	No	Si	Si	Si	Si	No
Requiere conocimientos avanzados	No	No	No	No	No	No
Multiplataforma	Si	Si	Si	Si	Si	No
Acceso a SQL	No	Si	Si	SI	Si	Si
Requiere bases de datos especificas	Si	No	No	No	No	No
Soporte vectorial de métodos de maquinas	Si	No	Si	Si	Si	No
Combinan Modelos	Si	No	Si	Si	Si	No
Métodos bayesianos	----	No	Si	Si	Si	Si
Modelos de clasificación	Si	Si	Si	Si	Si	Si
Modelos de regresión	Si	No	Si	Si	Si	No
Interfaz amigable	Si	Si	Si	Si	Si	Si
Árbol de decisión	Si	Si	Si	Si	Si	Si
Permite visualización de datos	Si	Si	Si	Si	Si	Si
Plataforma	-----	-----	Todas	Todas	Windows, Linux	Windows, Linux

Nota: Elaboración propia

2.2.11. Algoritmos De Datos.

Asto (2020) Describe sobre las numerosas técnicas de minería de datos pueden dividirse en dos grupos: supervisadas y no supervisadas; las primeras se utilizan para aplicaciones con un objetivo de predicción y las segundas se aplican a aplicaciones con el objetivo de descubrir

nuevos datos. El objetivo de los algoritmos supervisados es predecir un valor desconocido (etiqueta) a partir de un conjunto de atributos conocidos. Para lograrlo, los algoritmos supervisados deben realizar primero un "entrenamiento" en registros en los que se conocen tanto los atributos como la etiqueta. Así se crean las reglas que se utilizarán para analizar los registros en los que se desconoce la etiqueta. Los algoritmos no supervisados, por su parte, no emplean registros con etiquetas porque el objetivo es dejar que el algoritmo encuentre cualquier patrón existente, incluida la creación del vínculo entrada-salida.

La categorización de los algoritmos de minería de datos se muestra en la siguiente tabla.

Tabla 3

Clasificación de los Algoritmos de Datos.

Supervisados	**No supervisados**
• Arboles de decisión	• Detección de desviaciones
• Inducción neuronal	• Segmentación
• Regresión	• Agrupación
• KNN	• Reglas de asociación
• Naybe Bayes	• Patrones secuenciales
• Algoritmos ensamblados	

Nota: En la tabla se muestra la clasificación de los Algoritmos de minería de Datos. **Fuente:** elaboración propia.

Tabla 4

Tareas y Aplicaciones De La Minería De Datos.

Tarea	Objetivo/Descripción	Aplicaciones claves
Predicción	Inferir la variable objetivo desde la combinación de otras variables. Se aplican técnicas de clasificación, regresión y estimación de densidad.	Predicción de rendimiento académico y detectar comportamiento de estudiantes.
Agrupamiento	Identificar grupos con características similares.	Crear grupos de estudiantes con características similares en su patrón de interacción y aprendizaje.
Minería de relaciones	Estudio de relaciones entre variables y descubrimiento de reglas.	Identificar relaciones en patrones de aprendizaje y diagnosticar dificultades en aprendizaje.
Destilación de datos	Representación de datos de manera más inteligible mediante resumen, visualización e interfaces interactivos.	Brindar herramientas de apoyo a los instructores para visualizar y analizar actividades relacionadas con sus estudiantes.
Descubrimiento con modelos	Emplear modelos validados previamente a nuevos datos.	Identificación de relación entre el comportamiento y las características de los estudiantes o de variables contextuales.
Detección de anomalías	Detección de individuos con diferencias significativas.	Detección de estudiantes con dificultades o con aprendizaje irregular
Minería de texto	Extracción de información de calidad desde los datos.	Análisis de contenido de foros, chats, páginas web y documentos.
Seguimiento de conocimiento	Estimar el dominio de habilidades de estudiantes, utilizando modelos cognitivos y los registros de la evaluación como evidencia.	Monitorear el nivel alcanzado por los estudiantes en el tiempo.
Factorización de matriz no negativa	Definición de matrices en base a resultados de evaluación que se descomponen en otras matrices que representan habilidades obtenidas.	Evaluación de habilidades.

Nota: Elaboracion propia.

2.2.12. Rendimiento académico

El rendimiento académico se considera un componente importante de las instituciones de enseñanza superior y se utiliza como uno de los criterios de evaluación de las universidades. Las investigaciones sugieren que las calificaciones obtenidas en diversas evaluaciones y cursos asociados pueden utilizarse para calibrar el éxito académico. Según algunas investigaciones, el rendimiento académico se mide por la capacidad de un estudiante para terminar con éxito sus estudios y obtener un título universitario (Yamao, 2018).

Para Morales (2018) el éxito académico en la universidad es consecuencia del aprendizaje, que se produce por la actividad educativa del profesor y se genera en el alumno, pero es obvio que no todo aprendizaje es resultado de la actividad instructiva del profesor. Una calificación, tanto cuantitativa como cualitativa, coherente y válida indicará un determinado aprendizaje o el cumplimiento de unos objetivos previamente establecidos. El éxito académico se comunica a través de las calificaciones.

Factores para Determinar El Rendimiento Académico.

Candia (2019) describe 3 factores:

1. Factores inherentes al alumno.

a) Falta de preparación para el acceso a la enseñanza superior o niveles de conocimientos insuficientes para cumplir los criterios de la Universidad.

b) Desarrollo inadecuado de las competencias en relación con el itinerario profesional elegido.

b) Factores relacionados con la actitud.

d) Falta de enfoques de estudio o de estrategias de productividad intelectual.

e) Hábitos de aprendizaje incompatibles con la profesión seleccionada.

2. Factores inherentes con el profesor.

a) Deficiencias pedagógicas

b) Atención personalizada insuficiente

c) Mayor compromiso insuficiente

3. Factores inherentes a la organización académica universitaria

a) Falta de objetivos claros

b) Falta de cooperación entre los cursos

b) Métodos de selección empleados

d) Normas objetivas de evaluación.

En conclusión, el rendimiento académico de los estudiantes universitarios depende exclusivamente de la institución en la que asisten a clase, de sus profesores y, lo que es más importante, de sus propias capacidades. Este último factor es especialmente complicado porque hay muchas variables que pueden hacer que un estudiante tenga éxito o no. En el presente análisis, nos centraremos principalmente en la información de los estudiantes, teniendo en cuenta únicamente la información de su admisión en la universidad, como se expondrá a continuación. (Candia Oviedo, 2019).

2.2.13. Rendimiento académico empleando minería de datos.

Las universidades están estudiando las causas profundas de este problema debido a su preocupación por el rendimiento de los estudiantes en sus carreras académicas y los malos indicadores de abandono y rendimiento académico. El sistema de enseñanza superior actual no satisface las demandas de crecimiento y desarrollo de la nación, lo que sirve de crudo recordatorio de los defectos de nuestro sistema educativo. En la enseñanza superior sigue imperando la memorización y se promueve la repetición de asignaturas; además, los profesores universitarios y de secundaria siguen empleando un enfoque dogmático y coercitivo de la enseñanza. El bajo rendimiento académico es el resultado de esta supresión, que restringe la expresión creativa y limita el desarrollo de talentos inventivos y creativos. Como resultado, los graduados de las escuelas no se integran eficazmente en la sociedad y no hacen avanzar el desarrollo y la tecnología en muchos campos (Quiñones Huatangari & Carrasco Vega,

"Rendimiento académico empleando minería de datos:Academic performance using data mining", 2020).

Deserción de los estudiantes universitarios.

Alania (2018), según esta definición, los estudiantes que faltan muchas semanas a clase en la universidad por motivos distintos de la enfermedad son los que abandonan los estudios. También se refiere el autor a la deserción, independientemente de que regresen u obtengan un título comparable, cuando un estudiante se retira de la universidad sin recibir un título. Por lo tanto, se dice que un estudiante que se matricula en un curso determinado y, por diversas razones, se retira sin terminar su preparación académica, ha desertado de su universidad.

El Learning Analytics.

Taya (2021) Para evaluar el rendimiento académico, prever el rendimiento e identificar posibles problemas, es necesario interpretar una cantidad significativa de datos creados y recibidos de numerosas fuentes. Learning Analytics es cada vez más importante. Según varias fuentes, uno de los avances más significativos en el aprendizaje y la enseñanza potenciados por la tecnología es la analítica del aprendizaje. Por lo tanto, no es de extrañar que se haya escrito un gran número de artículos académicos sobre Learning Analytics. Para mejorar las capacidades de enseñanza y aprendizaje de determinados estudiantes y profesores, el estudio y la mejora de Learning Analytics implica crear, utilizar e integrar nuevos procedimientos y herramientas. El proceso de aprendizaje es el foco principal de Learning Analytics. Learning Analytics es un área de estudio interdisciplinar que tiene vínculos con los campos de la investigación sobre la enseñanza y el aprendizaje, la informática y la estadística debido a sus conexiones con la enseñanza y el aprendizaje digitales. Los datos disponibles se recopilan, se examinan y las conclusiones extraídas se aplican para comprender el comportamiento del alumno y ofrecerle más ayuda.

2.2.14. Métricas de desempeño

Para facilitar la selección del algoritmo óptimo en función del objetivo del estudio, las métricas de rendimiento son cruciales en problemas de clasificación en los que el objetivo es discriminar entre varios algoritmos de aprendizaje automático y aprendizaje profundo (Danjuma, 2015). objetivo del estudio (Danjuma, 2015). En la biblioteca propuesta se utilizaron las siguientes métricas: recall, F1-score, curva ROC, AUC, índice kappa, tabla de confusión, precisión y exactitud (Borja Robalino y otros, 2020).

Los verdaderos positivos (VP) son conjunto de datos para los que tanto la clase esperada como la real son 1 (verdadero).

Los verdaderos negativos (VN) son conjunto de datos para los que tanto la clase prevista como la real son 0 (falso) o ambas 0 (falso).

Falsos positivos (FP): cuando la clase prevista de un conjunto de datos es 1 (verdadero), pero su clase real es 0 (falso).

Falsos negativos (FN): Cuando el valor previsto de un conjunto de datos es 0 (falso), pero su clase real es 1 (verdadero).

2.2.14.1 Exactitud.

La precisión es la proximidad en que un resultado se asemeja al valor real que se pretende obtener. Dicho de otro modo, representa el porcentaje en que ha alcanzado sus objetivos. Puede tratarse de un logro personal o de un objetivo estratégico. La precisión es alta cuando se alcanza con exactitud el valor objetivo. Si se desvía del objetivo, no es preciso. Después de un acontecimiento concreto, se puede determinar la precisión, pero si se quiere averiguar si se puede mantener como un éxito a largo plazo, hay que repetirlo (Raeburn, 2023)

$$Exactitud = \frac{\mathrm{VP} + \mathrm{VN}}{\mathrm{VP} + \mathrm{FP} + \mathrm{FN} + \mathrm{VN}}$$

2.2.14.2 Precisión

El grado de coincidencia de los resultados se mide por la precisión. La precisión puede controlarse a lo largo del tiempo, pero la exactitud es útil en determinadas situaciones. La razón es que la repetibilidad es necesaria para calibrar el grado de similitud entre cada conjunto de mediciones con el fin de cuantificar la precisión. Cuando los resultados son dispersos, la precisión es escasa, y cuando los resultados son comparables entre sí, la precisión es elevada. Dos situaciones en las que resulta muy útil medir la precisión son: cuando se intenta evitar un mismo error y cuando se quiere averiguar un procedimiento reproducible para obtener buenos resultados (Raeburn, 2023).

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

2.2.15. Definiciones conceptuales

- **Predicción**

Es la dispersion de los valores derivados de las clasificaciones. La precisión aumenta al disminuir la dispersión. Puede expresarse como una relación entre el número total de predicciones y el número de predicciones que fueron exactas. (Taya Acosta, 2021).

- **Minería de Datos**

A fin de recopilar datos relevantes para la toma de decisiones y extraer patrones y tendencias para prever comportamientos futuros, es necesario utilizar sistemáticamente determinados algoritmos que generen una lista de patrones a partir de una gran cantidad de datos.

- **Técnica de minería de datos**

Para descubrir patrones recurrentes que expliquen cómo se comportan estos datos, la minería de datos es una combinación de técnicas y tecnologías que permiten la exploración automática o semiautomática de bases de datos masivas.

- **Rendimiento académico**

El éxito académico incluye todos los aspectos de la educación, desde la obtención de un título hasta el crecimiento moral. El éxito académico se define como los logros académicos, la consecución de los objetivos de aprendizaje, la adquisición de habilidades y competencias deseables, la satisfacción, la perseverancia y el buen rendimiento después de la universidad. También se cree que el éxito académico está influido por una serie de elementos significativos, siendo el rendimiento académico el más utilizado por las universidades y cuantificado casi exclusivamente por las notas. Además, se afirma que las tareas y las evaluaciones pueden utilizarse para medir la consecución de los objetivos de aprendizaje y la adquisición de habilidades y competencias a nivel de curso, con un solapamiento significativo entre ambas mediciones.

- **Ingresantes**

Es innegable que las calificaciones que un estudiante obtiene durante sus experiencias curriculares se correlacionan directamente con su desempeño académico, la adquisición de habilidades y competencias y el logro de los objetivos educativos del programa; en otras palabras, constituyen un componente crucial de su perfil de egreso. Esta es la premisa de la presente tesis, que busca adquirir el perfil de egreso utilizando el registro académico de calificaciones.

- **Situación social**

La idea de la condición social de una persona está relacionada con su lugar en la sociedad. En otras palabras, el concepto se refiere a cómo actúa el sujeto en relación con su entorno o contexto.

- **Situación económica**

Alude al conjunto de elementos que constituyen el patrimonio de una persona (solvencia). Por tanto, el patrimonio está referido a la posición económica. Cuando una persona tiene un patrimonio importante, su situación económica es favorable.

III. MATERIALES Y MÉTODOS

3.1. Lugar de ejecución

La presente investigación se realizo en la Facultad de Ingeniería en Industrias Alimentarias de la Universidad Nacional Agraria de la Selva, universidad pública situada en la ciudad de Tingo Mara, distrito de Rupa Rupa perteneciente a la provincia de Leoncio Prado en el departamento de Huánuco.

3.2. Materiales y métodos

3.2.1. Materiales

Los materiales utilizados fueron:

Tabla 5

Los materiales utilizados.

Materiales		
DESCRIPCION	**Cantidad**	**Unidad de Medida**
Papeles	2	Millar
Lapiceros	3	Unidad
Fotocopias	100	Unidad
Impresiones	600	Unidad
Empastado	3	Unidad
TOTAL	708	

Nota: en la tabla se puede visualizar los materiales utilizados en el trabajo. **Fuente:** propia.

3.2.2. Equipos

Los equipos utilizados fueron:

Tabla 6

Los equipos de Hardware.

Equipos de Hardware		
Materiales	**Cantidad**	**Unidad**
Computadora	3	Equipo
Impresora	1	Equipo
CDs	3	Unidad
Total	7	

Nota: En la tabla se visualiza los equipos de Hardware que se utilizan en el proyectó. **Fuente:** propia.

Tabla 7

Equipos de Software.

Equipos de Software	
Descripción	**Cantidad**
Microsoft Office 2016	1
Windows 11	1
Software SPSS versión 25	1
Software Weka 3.9.5	1
TOTAL	4

Nota: En la tabla se visualiza los equipos de Software empleados en el trabajo de tesis. **Fuente:** propia.

Servicios

Los servicios utilizados fueron:

Tabla 8

Los servicios utilizados.

Servicios		
Descripción	**Cantidad**	**Unidad de medida**
luz	3	Meses
Movilidad	3	Meses
Internet	3	Meses
Telefonía móvil	3	Meses
Asesor de tesis	1	Unidad
TOTAL	13	

Nota: En la tabla se visualiza los servicios utilizados en el trabajo. **Fuente:** propia.

3.3. Metodología

3.3.1. Tipo de Estudio.

Por la naturaleza del estudio esta tesis es aplicada. Se hizo una descripción de los datos de accesos y rendimiento académico de los períodos académicos 2015 - 2018, para poderlas comparar y determinar diferencias (Taya Acosta, 2021).

3.3.2. Nivel de investigación

La investigación se encarga de determinar del grado de asociación existente entre dos o más variables en una misma muestra de sujetos, en nuestro caso de estudiantes (Taya Acosta, 2021).

En esencia, la investigación correlacional-causal busca identificar probables relaciones y explicaciones entre las variables estudiadas: modelo de aprendizaje automático y predicción del rendimiento académico.

3.3.3. Método de investigación.

El presente trabajo de investigación se desarrolló mediante método científico deductivo, que nos permitirá sistematizar los resultados obtenidos mediante los instrumentos de investigación utilizadas en el presente estudio (Vargas Saldivar, 2018).

3.4. Operacionalización De Variables.

3.4.1. Variable independiente: Modelo de aprendizaje automático.

Tabla 9

Operacionalización de las variables de estudio.

VARIABLES	Definiciones	DIMENSIONES	Elementos seleccionados	Indicadores	Definición dimensión
Variable Independiente: Modelo de aprendizaje automático	**Definición Conceptual:** (Hewlett, 2023) Es un archivo inteligente que ha sido condicionado con un algoritmo para identificar patrones particulares en conjuntos de datos y sacar conclusiones y predicciones a partir de ellos constituye un modelo de aprendizaje automático, para la predicción usaremos algoritmos supervisados. **Definición Operacional** Se entrena los algoritmos de aprendizaje automático generando un modelo, que luego es cargado para realizar las predicciones.	No tiene			Esta variable no se está midiendo, ya que solo tiene factores que son los modelos que vamos a utilizar
Variable dependiente: Predicción del Rendimiento Académico	**Definición Conceptual:** (Gonzales Tirados, 1989) El rendimiento académico de los estudiantes está determinado principalmente por su capacidad, sus profesores y la universidad a la que asisten. Esta idea se utiliza para describir cómo se evalúa el conocimiento de los estudiantes en entornos educativos de todo tipo, como lo demuestran los resultados de sus evaluaciones.	Exactitud	Académicos	Nota de ingreso Modalidad de ingreso Opción de ingreso Tipo de preparación	La exactitud es la proporción de predicciones correctas realizadas por el modelo de aprendizaje automático sobre el total de todas las predicciones.
			Económicos	Colegio de procedencia Tipo de colegio Dependencia economía	
		Precisión	Sociales	Lugar de procedencia Edad Sexo Como se enteró del examen Motivo de postulación Trabaja Numero de hermanos Lugar donde vive	La precisión es la proporción de identificaciones positivas que fueron de manera correcta en la evaluación de los modelos.

Nota: En la tabla 8 se muestra la variable Modelo de aprendizaje automático. **Fuente:** propia.

3.5. Población Y Muestra

3.5.1. Población.

La presente investigación tiene como población identificada los 204 alumnos de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS.

3.5.1. Muestra

La muestra de la presente investigación es de 204 alumnos de Facultad de Ingeniería en Industrias Alimentarias. Con un muestreo aleatorio.

3.6. Técnicas e Instrumentos de recolección de datos

3.6.1. Técnica

Para el desarrollo de la investigación se empleó las técnicas de análisis de data: Información Histórica, Análisis Documental de las variables de estudio que nos permitió sistematizar los datos recolectados para el desarrollo del presente trabajo de investigación.

3.6.2. Instrumento

Se aplicó el siguiente instrumento de investigación según las variables de estudio: Ficha de análisis documental (Anexo 4).

3.7. Métodos y tratamiento de datos

3.7.1 Métodos

La base de datos de los estudiantes que ingresaron a la facultad de Industrias Alimentarias desde el semestre 2015-I hasta el 2019-I, fue proporcionada por las oficinas de admisión de la Universidad Nacional agraria de la selva.

3.7.2 Metodología

3.7.2.1 Análisis descriptivo.

La información que se usó fueron los datos recolectados por la oficina de admisión, y por la Oficina de la Dirección de asuntos académicos (DICCA) donde se almacenan las calificaciones de los estudiantes universitarios, y a que a su vez se clasifican con aprobados y

desaprobados.

El análisis descriptivo nos permite obtener información relevante con respecto a los datos obtenidos por las oficinas de admisión y DICCA, como por ejemplo que modalidad son los que más ingresan.

Se realizó una prueba de chi cuadrado para identificar las variables que más influyen en el rendimiento académico de los estudiantes, dicha prueba se realzó en el software spss.

3.7.2.2 Métricas de desempeño

Exactitud

Para determinar la exactitud se utilizó un software libre weka 3.9.5 y se entrenó los siguientes modelos de aprendizaje automático:

- Modelos ensamblados Vote.
- Modelo de Random Forest.
- Modelos de K - Nearest Neighbors.
- Modelo de Nayve Bayes.
- Modeo ensamblado Bagging.

Dicho software nos muestra la matriz de confusión y la exactitud que se calcula con la siguiente formula:

$$\text{Exactitud} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{FN} + \text{VN}}$$

Precisión.

La precisión se calcula mediante la formula

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

Donde VP representa a la cantidad de categorías predichas correctamente para cada categoría de condición (aprobado o desaprobado) y VF representa la cantidad de predicciones erróneas de las categorías de la condición del alumno al final del ciclo. Posteriormente se comparó las precisiones de los modelos de machine learning mencionados anteriormente y se

propondrá una propuesta de caso para la aplicación del algoritmo más preciso. La propuesta de caso consistió en la identificación específica de una muestra para una determinada restricción de condiciones que responda a la pregunta ¿qué porcentaje de las categorías predichas son las correctas?, en este caso se necesitará al modelo más preciso.

Para escoger el modelo más adecuado para realizo las predicciones, se llevó todos los resultados de los entrenamientos de los modelos de machine learning al spss, donde se realizó una prueba no paramétrica, dicha prueba es el análisis de varianza de Friedman. Esta prueba se selecciona debido a que se comparan más de dos grupos relacionados sin asumir la normalidad de los datos con un nivel de significancia del $\alpha = 0.05$.

Luego se realizó un análisis de varianza de vías por rangos de Friedman para muestras relacionadas, esta prueba nos permite determinar si las distribuciones de los modelos son las mismas o son diferentes y con ello determinar el mejor modelo para la predicción.

3.7.3 Tratamiento de datos

Se siguió los pasos de la metodología CRISP-DM, según los autores (Candia Oviedo, 2019), (Vega García, 2019), (Alania Ricaldi, 2018), (Morales Agurto, 2018) y (Yamao, 2018) todos coinciden que es una metodología para las predicciones, usaremos el software weka 3.9.5 con cinco algoritmos de aprendizaje supervisado, modelo ensamblado Vote, modelo de Random Forest, modelos de K - Nearest Neighbors, modelo de Nayve Bayes, modelo ensamblado Bagging.

3.8. Aspectos éticos

La preocupación por la privacidad de los datos personales, tanto de alumnos como de profesores, es la principal dificultad ética de este tipo de estudios. Antes de utilizarse o hacerse pública, la utilización de información personal identificable requiere autorización y consentimiento.

Para garantizar la privacidad de los datos se utilizó un proceso de anonimización, dando

a cada registro de los datos un identificador distinto y eliminando los atributos que pudieran servir para identificar a las personas de los distintos registros, como nombres, apellidos y domicilios particulares, entre otros. De este modo, se garantiza la autenticidad de los resultados de las pruebas sin revelar ninguna información personal sobre los participantes.

IV. RESULTADOS Y DISCUSIONES

Vamos a desarrollar usando la metodología CRISP-DM de acuerdo con las fases y las tareas que propone.

4. 1 Comprensión Del Negocio

Aquí es la fase inicial nos enfocamos en entender los objetivos y requerimientos del proyecto para esta fase vamos a detallar cada uno de los pasos:

a) Determinación del objetivo de negocio.

La universidad Nacional Agraria de la Selva con su Facultad de Ingeniería Industrias Alimentarias es una comunidad integrada por alumnos, docentes, graduados y trabajadores administrativos. Se dedica a preparar profesionales con conocimiento en industrialización y mejora de alimentos. Sus programas de estudio con acorde a la región.

Entre sus objetivos está mejorar la calidad académica de los alumnos y egresados.

b) Evaluación de la situación.

Dentro de la Universidad la Facultad de Industrias tiene un problema que muy pocos egresan de la universidad debido a mucha deserción estudiantil, por diferentes motivos los cuales no son todos conocidos, sobre todo esto se da durante los primeros años de estudio en la universidad, lo que hace que la planificación de la facultad muchas veces se vea afectada.

Los datos de los alumnos son obtenidos desde el momento de la inscripción en la oficina de admisión como postulante o ingresante por el centro Preuniversitarios.

Los cuales contienen datos como apellidos y nombres, edad, sexo, tipo de colegio que egreso, lugar de procedencia, con quienes vive, dependencia económica. También aquí se obtuvieron el puntaje de ingreso y la modalidad por al que ingresaron.

Las notas de los ingresantes por el centro preuniversitario no se registraron en la oficina de admisión, por lo que estos datos se encontraron dentro de las oficinas del centro preuniversitario.

Todos los atributos han sido almacenados en Excel y se encuentran centralizados en la oficina de admisión.

Los recursos con los que se dispone para el desarrollo son los siguientes:

Los materiales disponibles son el Weka versión 3.9.5 se usó como herramienta para la minería de datos también los datos de Excel y el SPSS para los cruzar los datos de ser necesarios.

Los recursos humanos el autor de la investigación.

Se dispone de los datos de los alumnos ingresante desde el 2015 al 2019 que hacen un total de 204 registros con sus atributos.

c) Determinación de los objetivos de minería de datos

EL objetivo de la minería de datos es dar soporte mediante tecinas de minería de datos a los objetivos de la investigación.

Objetivos específicos de la investigación

Realizar estudios estadísticos de los datos.

Encontrar el rendimiento académico de los alumnos mediante los ámbitos académicos, económicos y sociales.

Conocer estos objetivos permitirá tener muy claro el inicio de una planificación estratégica en cada inicio de ciclo.

4. 2COMPRENSION DE DATOS

a) Recolección de datos iniciales

Los datos que se utilizó se han obtenido de la oficina de admisión de la UNAS, también se solicitó los datos de las oficinas de la Dirección de asuntos académicos (DICCA) donde se encuentra el promedio ponderado del semestre correspondiente al primer ciclo.

Las notas de los ingresantes por el centro preuniversitario se tuvieron que recolectar en

las mismas oficinas del centro preuniversitario.

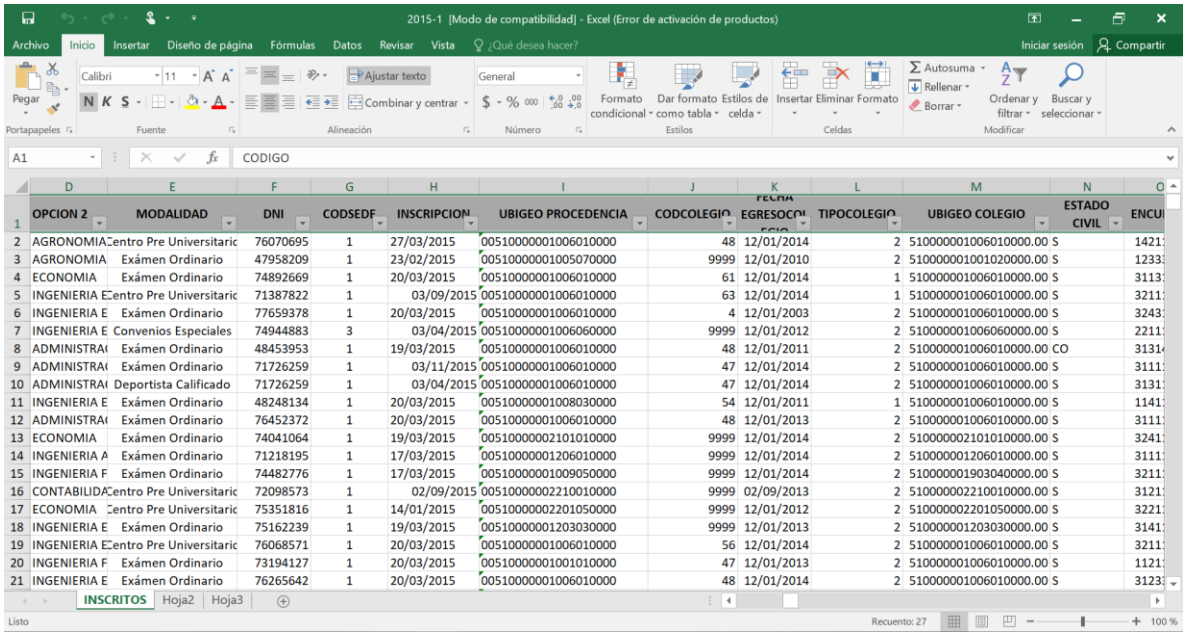
Los datos recolectados en las oficinas de admisión fueron:

Estos datos se recolectaron en hoja de cálculo Excel conteniendo los siguientes campos:

CODIGO
Apellidos y Nombres
OPCION 1
OPCION 2
MODALIDAD
DNI
CODSEDE
INSCRIPCION
UBIGEO PROCEDENCIA
CODCOLEGIO
FECHA EGRESOCOLEGIO
TIPOCOLEGIO
UBIGEO COLEGIO
ESTADO CIVIL
ENCUESTA
INGRESO
INGRESO A
SEXO
NOMBRE
COLEGIO
IDIOMA MAT
TELCELULAR
DIRECCION
UBIGEO
FECNAC
NOTAAC
NOTACO
RESPUESTA

Figura 15

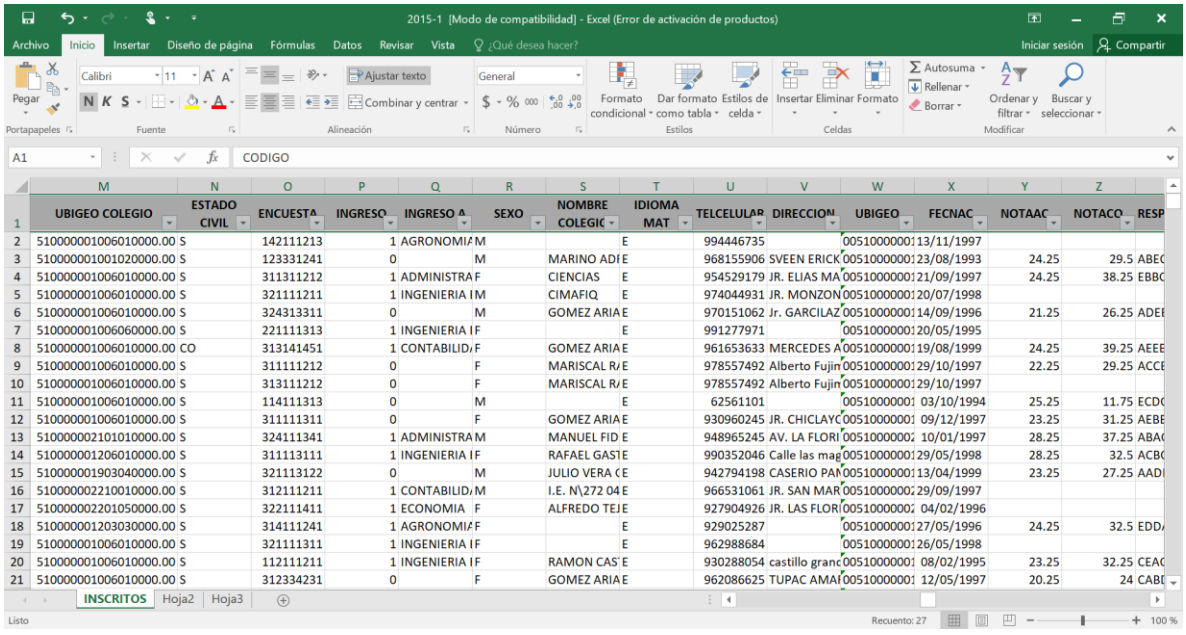
Base de datos Excel oficina de admisión



Nota: elaboración propia

Figura 16

Segunda parte de la base de datos de la oficina de admisión.



Nota: Elaboración propia.

Los datos recolectados en DICCA también se recolecto en Excel con los siguientes atributos:

TDOCUMENTO

CODALUMNO

PATERNO

MATERNO

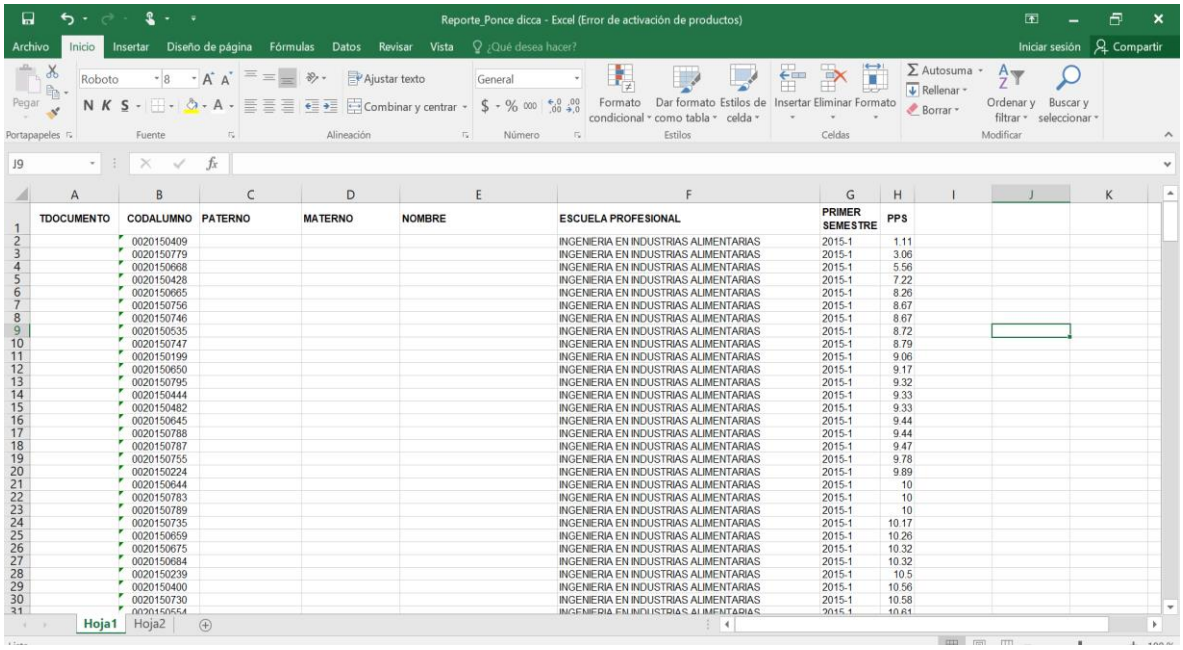
NOMBRE

ESCUELA PROFESIONAL PRIMER SEMESTRE

PPS

Figura 17

Datos recopilados de las oficinas de DICCA



TDOCUMENTO	CODALUMNO	PATERNO	MATERNO	NOMBRE	ESCUELA PROFESIONAL	PRIMER SEMESTRE	PPS
	0020150409				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	1.11
	0020150779				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	3.06
	0020150668				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	5.56
	0020150428				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	7.22
	0020150665				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	8.26
	0020150756				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	8.67
	0020150746				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	8.67
	0020150535				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	8.72
	0020150747				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	8.79
	0020150199				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.06
	0020150650				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.17
	0020150795				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.32
	0020150444				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.33
	0020150482				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.33
	0020150645				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.44
	0020150788				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.44
	0020150787				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.47
	0020150755				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.78
	0020150224				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	9.89
	0020150644				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10
	0020150783				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10
	0020150789				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10
	0020150735				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.17
	0020150659				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.26
	0020150675				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.32
	0020150684				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.32
	0020150239				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.5
	0020150400				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.56
	0020150730				INGENIERIA EN INDUSTRIAS ALIMENTARIAS	2015-1	10.58

Nota: Elaboración propia

Solo mostramos los datos que no incluyen los datos personales para conservar la privacidad de los datos.

b) Describir los datos. Procedemos a describir cada uno de los atributos de la tabla así como los tipos de datos.

- **Código.** Código único que se asigna a cada postulante para rendir el examen de admisión.

- **Apellidos y Nombres.** De cada alumno esta su apellido y nombre.
- **Opcion1.** Es la primera opción a la que postula el estudiante.
- **Opcion2.** A cada estudiante se le da la oportunidad de elegir dos opciones de carrera.
- **Modalidad.** Nos permitirá obtener la modalidad por la cual el estudiante logro ingresar a la universidad, puede ser por admisión, preuniversitario, primeros puestos, deportistas u otros.
- **DNI.** Documento Nacional de identidad de los estudiantes, será nuestro identificador para relacionar con los datos de la oficina de DICCA.
- **Codsede.** Es el código de la sede donde se tomó el examen de admisión o donde realizaron la preparación pre preuniversitaria.
- **Inscripción.** Es la fecha que se inscribieron para el examen de admisión que restado con la fecha de nacimiento nos permitirá obtener la edad a la que ingresaron a la universidad.
- **Ubigeo de procedencia.** Esto nos permitirá obtener el departamento de donde viene el estudiante.
- **Fecha de egreso del colegio.** Nos brinda la información del año en que el estudiante termino la secundaria.
- **Tipo de colegio.** Se expresa como colegio particular o estatal.
- **Ubigeo de colegio.** La ubicación del distrito, provincia y departamento donde está el colegio en el que culmino los estudios secundarios.
- **Estado civil.** La condición del estudiante si es soltero, casado u otro.
- **Encuesta.** Estas encuestas rellenan los estudiantes al momento de inscribirse y contra de 9 preguntas que serán de mucha importancia para nuestro estudio de predicción, a continuación, se detalla cada uno de las siguientes preguntas:

  C1. Tipo de preparación para su postulación

1) Autoestudio

2) Profesor particular.

3) Academia

4) Otros

C2. Como se enteró de las fechas de nuestro concurso de admisión.

1) Charlas dadas por el personal de la UNAS en el colegio

2) Familiares y amigos

3) Radio

4) Televisión

5) Internet

6) Otros

C3. ¿Cuál fue el principal motivo por el que se animó a postular a la UNAS?

1) Prestigio.

2) Nivel académico

3) La carrera que deseo no la ofrecen en otra universidad.

4) Recomendación de familiares o amigos

5) Económico

6) Servicios que ofrece comedor e internado

7) Otros

D1. ¿Trabaja?

1) No

2) Si, tiempo completo

3) Si, por horas

D2. Dependencia económica

1) Sus padres

2) Parientes

3) Si mismo

4) Otros

D3. ¿Viven tus padres?

1) Si los dos

2) Vive solamente el padre

3) Vive solamente la madre

4) Ninguno

D4. ¿Cuantos hermanos son?

1) Ninguno

2) De 1 a 3

3) De 4 a 5

4) De 6 a más hermanos

D5. ¿Con quién vive actualmente?

1) Con mis padres

2) Esposo (a)

3) Parientes

4) Solo

5) Otros

D6. Lugar donde vive

1) En la ciudad

2) Pueblo joven

3) Zona rural

- **Ingreso.** Es la condición del alumno si logro ingresar a la universidad, no ingreso (0), ingreso primera opción (1), ingreso segunda opción (2).

- **Ingreso A.** Indica la escuela profesional a la cual ingreso, por ejemplo, administración, contabilidad, economía entre otros.
- **Sexo.** Esta expresado como Masculino (M) o Femenino (F).
- **Nombre del colegio.** Nombre del colegio de egreso del estudiante.
- **Idioma Mat.** Es el idioma principal del estudiante.
- **TelCelular.** El número de celular del estudiante.
- **Dirección.** Dirección de vivienda del estudiante.
- **Ubigeo.** EL distrito, provincia y departamento de la vivienda del estudiante.
- **FecNac.** Este atributo nos permitirá obtener la edad del estudiante junto a la fecha de inscripción.
- **NotaAC.** Nota de aptitud academia correspondiente a razonamiento matemático y razonamiento verbal.
- **NotaCO.** Nota de conocimientos que incluye matemáticas (algebra, aritmética, geometría, trigonometría), física, química, biología y humanidades.
- **CODALUMNO.** Es el código asignado a cada estudiante al momento de ingresar.
- **Primer semestre.** Es el año en que estudio es primer semestre.
- **PPS.** Es el promedio ponderado del semestre en vigesimal.

La nota de ingreso de los alumnos por el centro preuniversitario no está en las oficinas de admisión, se recopilo en las oficinas de archivo central.

c) Explorar los datos.

Para la exploración de los datos vamos a usar el análisis descriptivo de las variables de la base de datos:

Tabla 10

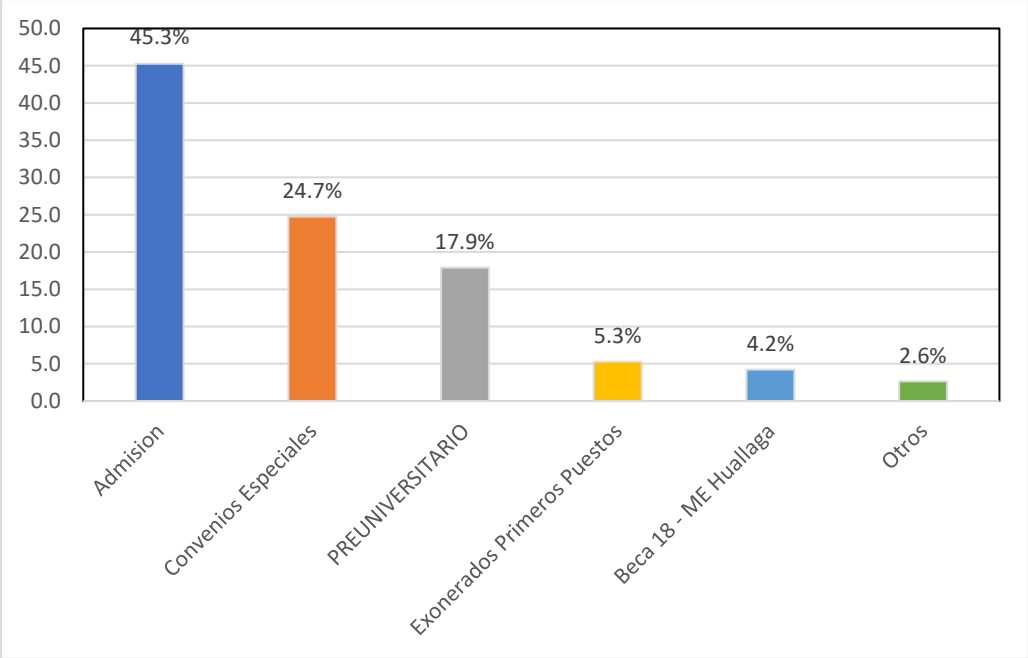
Número de estudiantes ingresantes por las distintas modalidades

MODALIDAD	Frecuencia	Porcentaje	Porcentaje acumulado
Admision	86	45.3	45.3
Convenios Especiales	47	24.7	70.0
PREUNIVERSITARIO	34	17.9	87.9
Exonerados Primeros Puestos	10	5.3	93.2
Beca 18 - ME Huallaga	8	4.2	97.4
Otros	5	2.6	100.0
Total	**190**	**100.0**	

Nota: La table nos muestra la cantidad de ingresantes por las distintas modalidades.

Figura 18

Distribución de estudiantes de acuerdo con la modalidad de ingreso



Nota: El grafico representa los porcentajes de alumnos ingresantes según las modalidad de ingreso a la universidad.

Como se observa en el gráfico el 45.3% de alumnos ingresaron por la modalidad de admisión mientras que también existe un gran porcentaje de ingreso por convenios especiales

que tiene la facultad.

Tabla 11

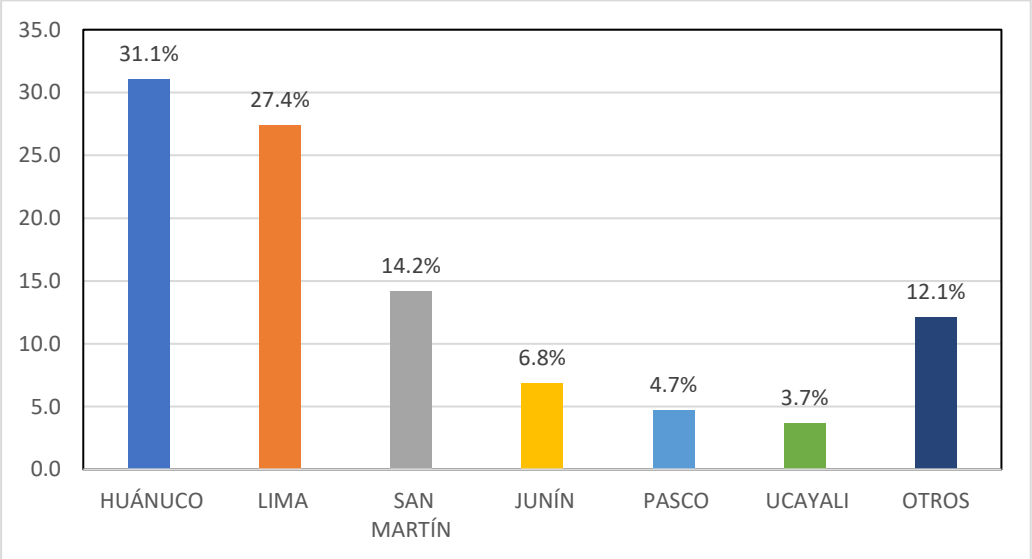
Número de estudiantes ingresantes por departamento de procedencia

DEPARTAMENTO	Frecuencia	Porcentaje	Porcentaje acumulado
HUÁNUCO	59	31.1	31.1
LIMA	52	27.4	58.4
SAN MARTÍN	27	14.2	72.6
JUNÍN	13	6.8	79.5
PASCO	9	4.7	84.2
UCAYALI	7	3.7	87.9
OTROS	23	12.1	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra que la cantidad de alumnos ingresantes procedentes de los distintos departamentos del Perú.

Figura 19

Distribución de estudiantes por departamentos de procedencia



Nota. El gráfico muestra el porcentaje de estudiantes ingresantes según su lugar de procedencia.

Como se observa en el grafico en su mayoría los estudiantes ingresantes son de los departamentos de Huánuco con 31.1%, también de la ciudad de Lima con 27,4% por el convenio que tiene la facultad con la municipalidad de Ate.

Tabla 12

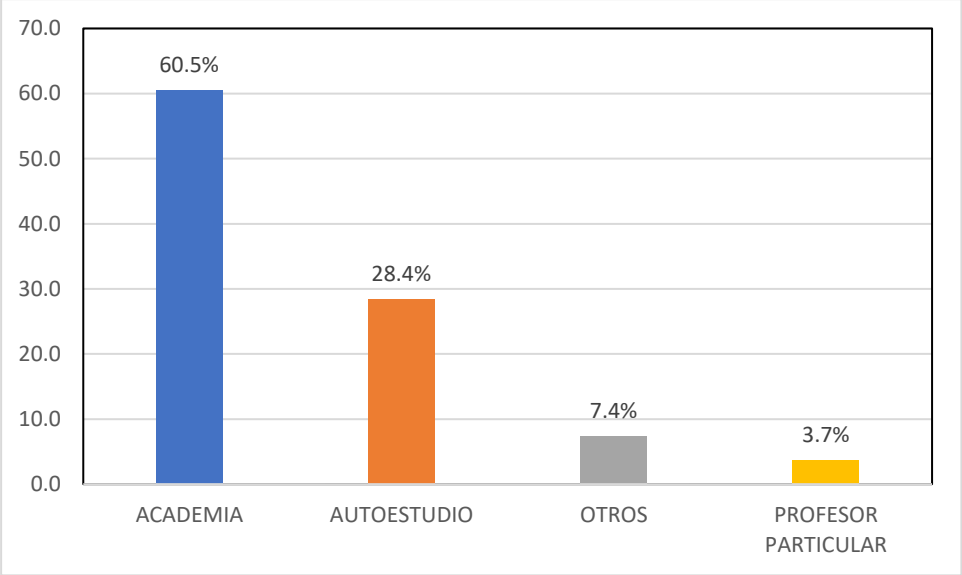
Número de alumnos ingresantes de acuerdo con el tipo de preparación.

Tipo de preparacion	Frecuencia	Porcentaje	Porcentaje acumulado
ACADEMIA	115	60.5	60.5
AUTOESTUDIO	54	28.4	88.9
OTROS	14	7.4	96.3
PROFESOR PARTICULAR	7	3.7	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra los alumnos ingresantes de acuerdo con el tipo de preparación

Figura 20

Estudiantes ingresantes de acuerdo con el tipo de preparación



Nota. La figura muestra el porcentaje de alumnos que ingresaron según el tipo de preparación.

Como se observa en el gráfico el 60.5% de los alumnos se prepararon en una academia preuniversitaria mientras que el 28,4% optaron por el autoestudio.

Tabla 13

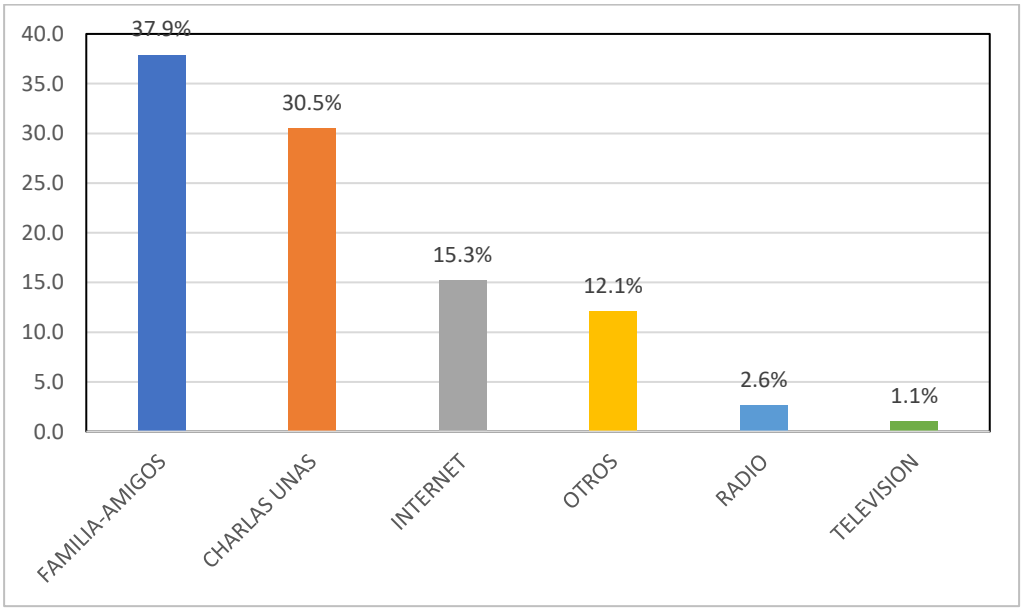
Número de estudiantes según como se informarón del examen de admisión

FORMA DE INFORMARSE	Frecuencia	Porcentaje	Porcentaje acumulado
FAMILIA-AMIGOS	72	37.9	37.9
CHARLAS UNAS	58	30.5	68.4
INTERNET	29	15.3	83.7
OTROS	23	12.1	95.8
RADIO	5	2.6	98.4
TELEVISION	2	1.1	99.5
7	1	0.5	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el porcentaje de alumnos que recibieron información por distingo medios o familia.

Figura 21

Distribución de estudiantes ingresantes según como se enteraron del examen de la UNAS



Nota. La figura muestra el porcentaje de alumnos de acuerdo con la forma como se enteraron del examen de admisión de la UNAS.

Como observamos en la figura el 37,9% de los ingresantes a la Facultad de Industrias

se enteraron del examen por familiares o amigos y un 30.5% mediante las charlas que brinda la universidad.

Tabla 14

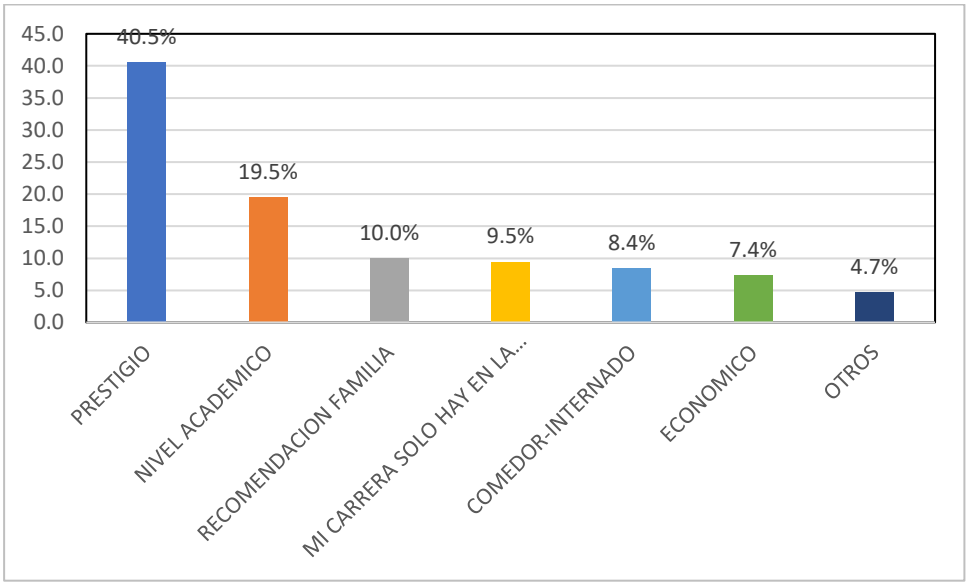
Número de alumnos de acuerdo al motivo de postulación a la Facultad de Ingeniería en Industrias Alimentarias.

Motivo postulacion	Frecuencia	Porcentaje	Porcentaje acumulado
PRESTIGIO	77	40.5	40.5
NIVEL ACADEMICO	37	19.5	60.0
RECOMENDACION FAMILIA	19	10.0	70.0
MI CARRERA SOLO HAY EN LA UNAS	18	9.5	79.5
COMEDOR-INTERNADO	16	8.4	87.9
ECONOMICO	14	7.4	95.3
OTROS	9	4.7	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos de acuerdo a los motivos por el cual postularon a la Facultad de Industrias.

Figura 22

Distribución de alumnos de acuerdo a los motivos por la cual postularon a la Facultad de Ingeniería en Industrias Alimentarias



Nota: La figura muestra la distribución de alumnos de acuerdo al motivo por el cual decidieron postular a la Facultad de Industrias.

Como se observa en la figura el 40.5% ingresaron a la facultad por el prestigio de esta, mientras que el 19.5% por el nivel académico que brinda.

Tabla 15

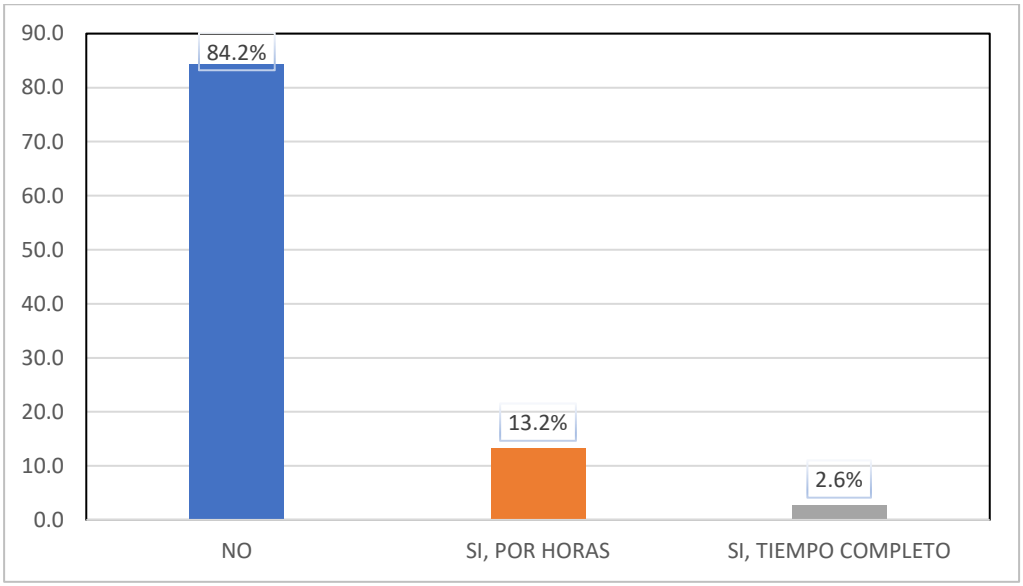
Número de alumnos de acuerdo con la pregunta ¿Trabaja?

TRABAJA	Frecuencia	Porcentaje	Porcentaje acumulado
NO	160	84.2	84.2
SI, POR HORAS	25	13.2	97.4
SI, TIEMPO COMPLETO	5	2.6	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos que trabajan y aquellos que lo hacen a tiempo parcial o a tiempo completo.

Figura 23

Distribución de alumnos de acuerdo con la pregunta ¿Trabajan?



Nota: La figura muestra el porcentaje de alumnos que no trabajan y los que si trabajan.

De acuerdo con la figura el 84,2 % no trabaja mientras que el 13,2 % trabaja solo por horas y un 2,6 % trabaja a tiempo completo.

Tabla 16

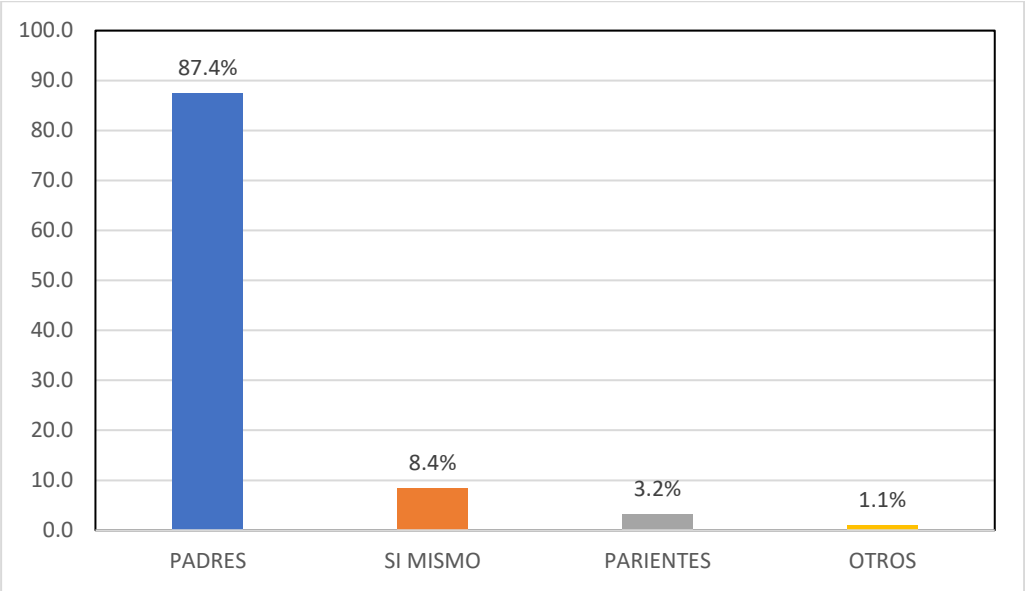
Número de alumnos de quienes dependen económicamente

DEPENDENCIA ECONOMICA	Frecuencia	Porcentaje	Porcentaje acumulado
PADRES	166	87.4	87.4
SI MISMO	16	8.4	95.8
PARIENTES	6	3.2	98.9
OTROS	2	1.1	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra de quienes dependen económicamente los alumnos ingresantes.

Figura 24

Distribución de alumnos de acuerdo con la dependencia económica



Nota. La figura muestra de quienes dependen económicamente los alumnos ingresantes a la Facultad de Industrias.

De acuerdo con el grafico el 87,4 % de los estudiantes dependen económicamente de sus padres mientras que un 8,4 % se solventan solos con sus gastos de la universidad.

Tabla 17

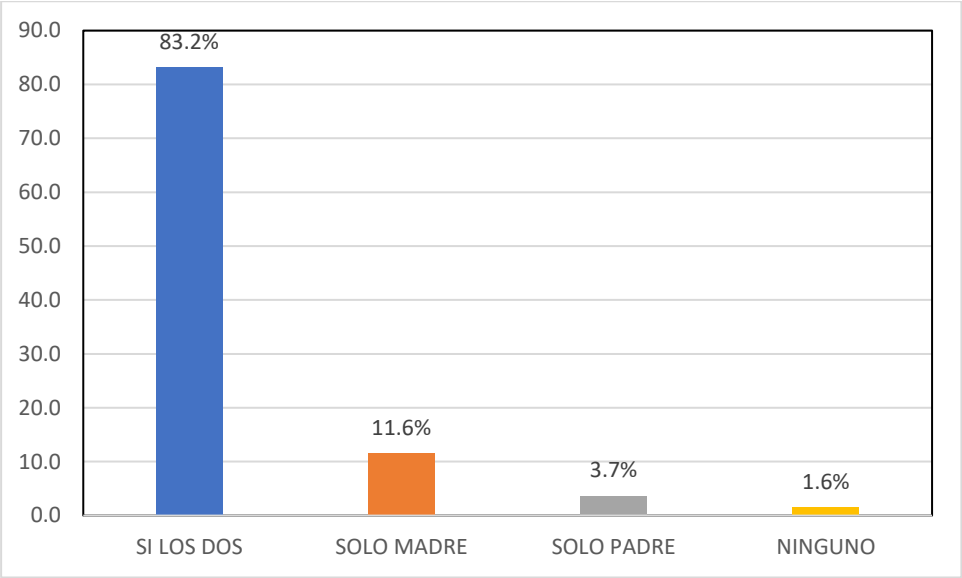
Número de alumnos de acuerdo con la pregunta ¿Viven tus padres?

Padres Vivos	Frecuencia	Porcentaje	Porcentaje acumulado
SI LOS DOS	158	83.2	83.2
SOLO MADRE	22	11.6	94.7
SOLO PADRE	7	3.7	98.4
NINGUNO	3	1.6	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos que tienen vivos a sus dos padres, solo a su madre, solo a su padre o ningunos.

Figura 25

Distribución de alumnos que de acuerdo con la pregunta ¿Viven tus padres?



Nota. La figura muestra el porcentaje de alumnos que tienen vivos a sus padres, solo padre, solo madre o ninguno.

Como se observa en la figura el 83,2 % tienen vivos a sus dos padres.

Tabla 18

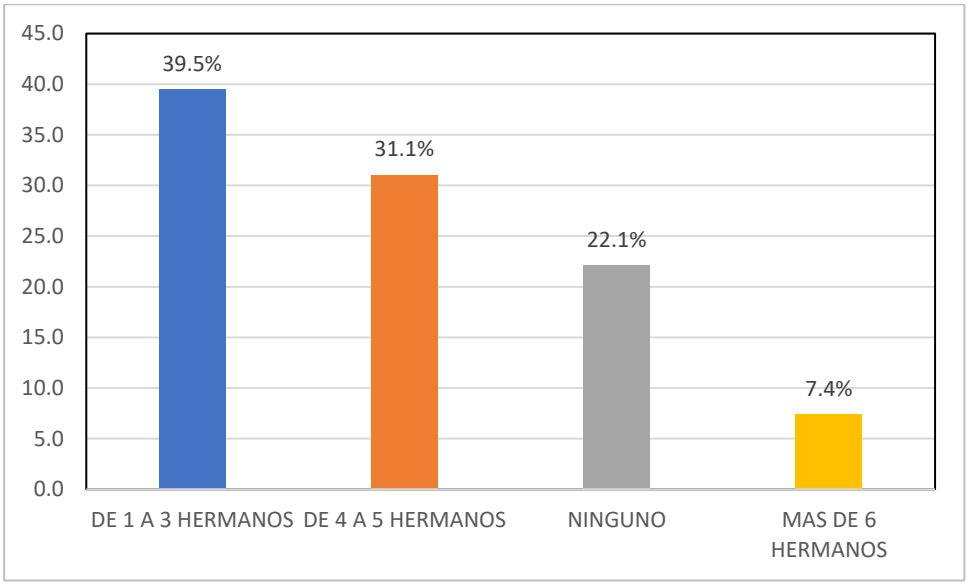
Número de alumnos de acuerdo con cuantos hermanos son en la familia

Número de hermanos	Frecuencia	Porcentaje	Porcentaje acumulado
DE 1 A 3 HERMANOS	75	39.5	39.5
DE 4 A 5 HERMANOS	59	31.1	70.5
NINGUNO	42	22.1	92.6
MAS DE 6 HERMANOS	14	7.4	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos de acuerdo con la cantidad de hermanos que tienen en su familia.

Figura 26

Distribución de alumnos de acuerdo con la cantidad de hermanos



Nota. La figura muestra el porcentaje de alumnos a acuerdo a la cantidad de hermanos que son en la familia.

Como se observa en la figura el 39,6 % tiene de 1 a 3 hermanos mientras que el 31,1% tienen de 4 a 5 hermanos y un 22,1 % son hijos únicos.

Tabla 19

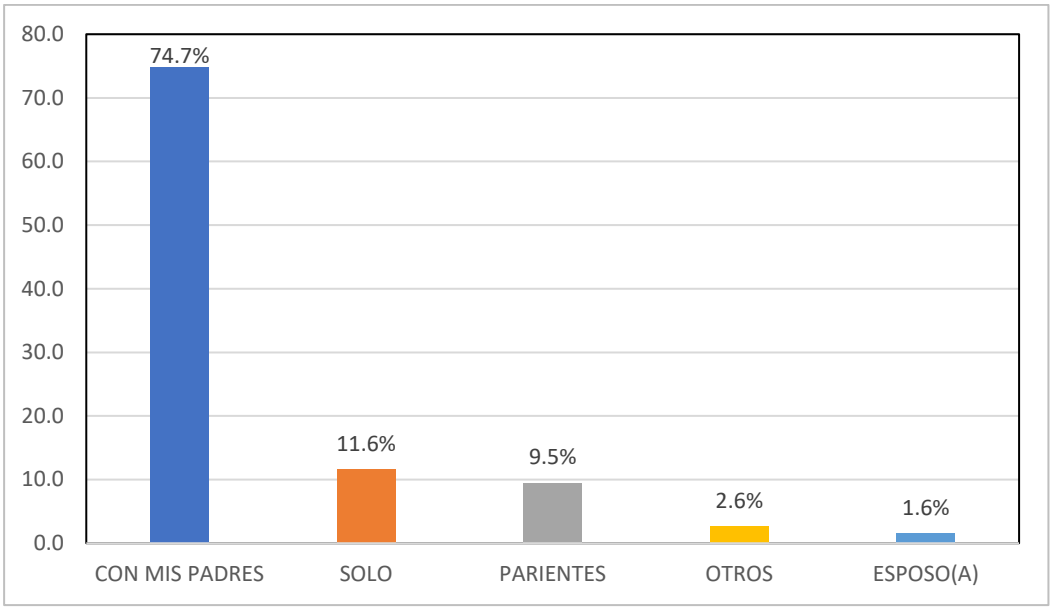
Número de alumnos de acuerdo con quien viven actualmente

Con quien vive	Frecuencia	Porcentaje	Porcentaje acumulado
CON MIS PADRES	142	74.7	74.7
SOLO	22	11.6	86.3
PARIENTES	18	9.5	95.8
OTROS	5	2.6	98.4
ESPOSO(A)	3	1.6	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos que viven o con sus padres, solos o con parientes.

Figura 27

Distribución de alumnos de acuerdo con quien viven actualmente



Nota. La figura muestra el porcentaje de alumnos que viven con sus padres, solos o parientes.

De acuerdo con la figura el 74,7 % viven con sus padres, mientras que un 11,6 % vive solo.

Tabla 20

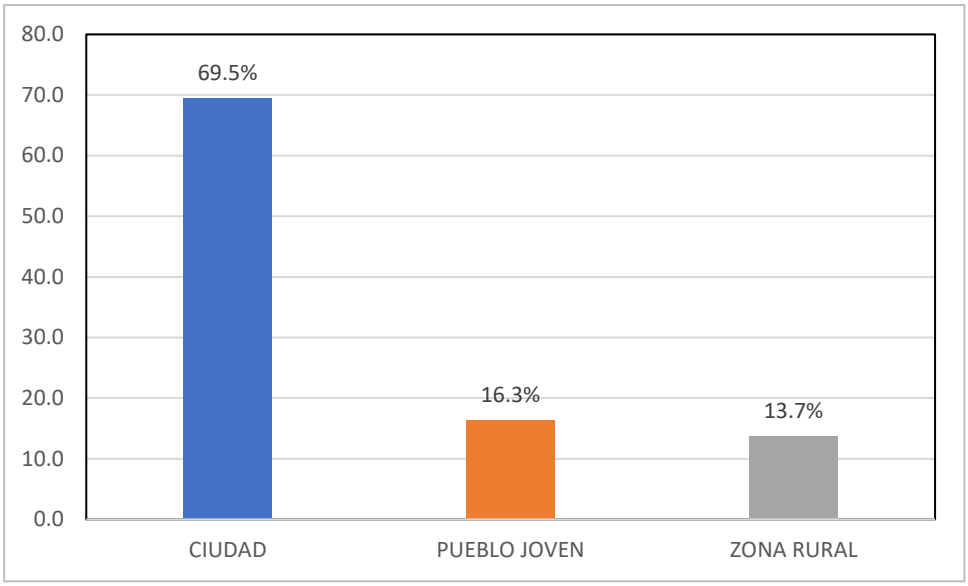
Número de alumnos donde están viviendo actualmente

Lugar donde vive	Frecuencia	Porcentaje	Porcentaje acumulado
CIUDAD	132	69.5	69.8
PUEBLO JOVEN	31	16.3	86.2
ZONA RURAL	26	13.7	100.0
Sistema	1	0.5	
Total	**190**	**100.0**	

Nota. La tabla muestra el número de alumnos según el lugar donde viven actualmente.

Figura 28

Distribución de alumnos de acuerdo con el lugar donde vive



Nota. La figura muestra el porcentaje de alumnos de acuerdo con el lugar donde viven actualmente.

En la figura ser puede observar que el 69.5 % vive dentro de la ciudad, el 16,3 % vive en un pueblo joven y el 13,7 % vive en una zona rural.

Tabla 21

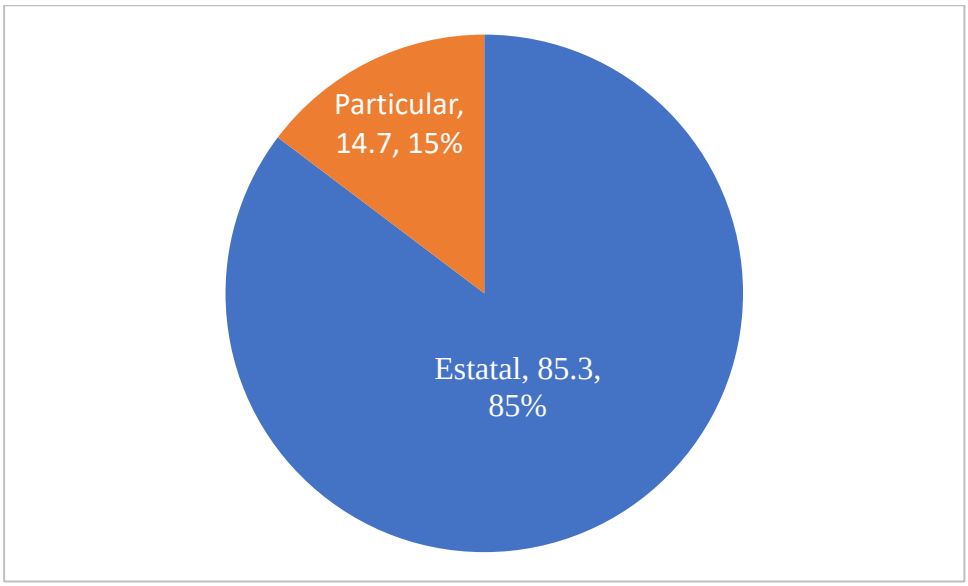
Ingresantes por tipo de colegio del 2015 - 2018

Colegio	Frecuencia	Porcentaje	Porcentaje acumulado
Estatal	162	85.3	85.3
Particular	28	14.7	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra la cantidad de ingresantes de los colegios estatales o particulares.

Figura 29

Distribución de ingresantes por tipo de colegio del 2015-2018



Nota. La figura muestra la cantidad y porcentaje de alumnos ingresantes de los colegios estatales y particulares.

Como se observa en la figura el 85 % de los ingresantes provienen de colegios estatales, esto significa que de cada 10 alumnos de la UNAS aproximadamente 8 alumnos son de colegios públicos.

Tabla 22

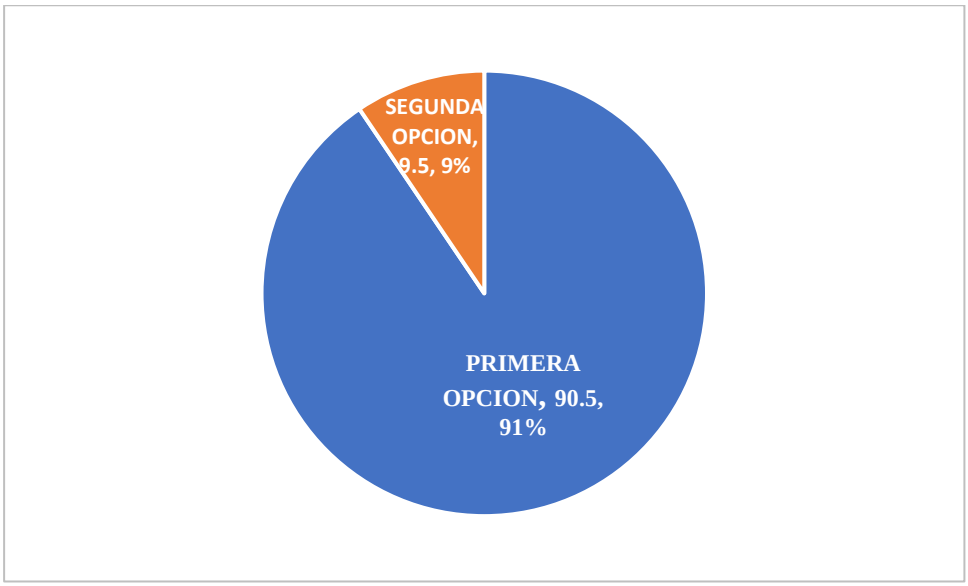
Ingresantes según la opción de ingreso

Opción de ingreso	Frecuencia	Porcentaje	Porcentaje acumulado
PRIMERA OPCION	172	90.5	90.5
SEGUNDA OPCION	18	9.5	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra la cantidad de alumnos que ingresaron por la primera y segunda opción.

Figura 30

Distribución de ingresantes según la opción que ingresaron



Nota. La figura muestra el porcentaje de alumnos que ingresaron por primera o segunda opción.

De acuerdo con el grafico se observa que el 91 % de los ingresantes a la Facultad de Industrias lo hicieron por la primera opción, esto significa que de cada 10 alumnos 9 de ellos ingresaron por la primera opción.

Tabla 23

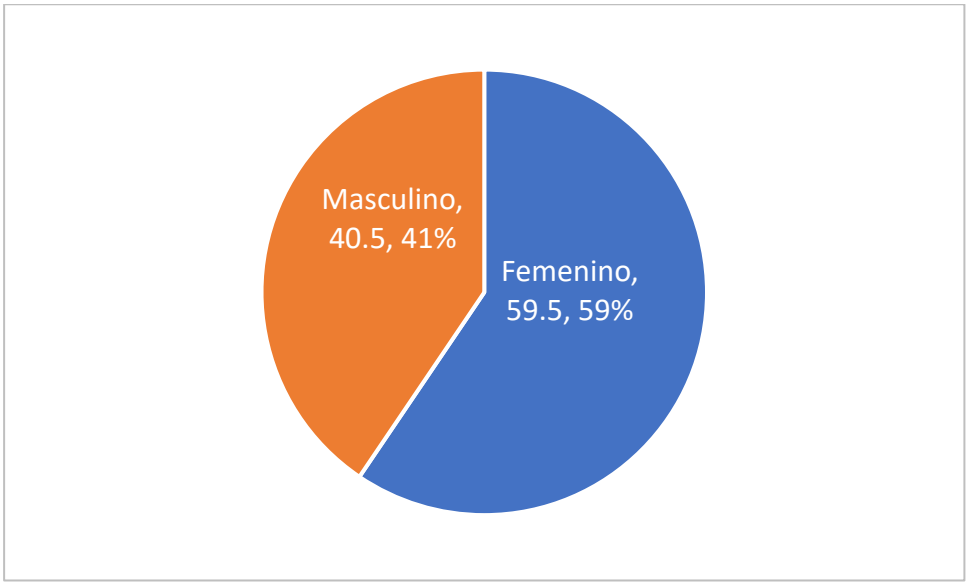
Ingresantes por genero del 2015-2018

Genero	Frecuencia	Porcentaje	Porcentaje acumulado
Femenino	113	59.5	59.5
Masculino	77	40.5	100.0
Total	**190**	**100.0**	

Nota. La tabla muestra los ingresantes a la Facultad de Industrias según el género.

Figura 31

Distribución de ingresantes por genero del 2015-2018



Nota. La figura muestra los porcentajes de ingresantes a la Facultad de Industrias por género.

Como se observa en la figura el 59 % de los ingresantes a la Facultad de Industrias son mujeres y el 41 % son varones, por cada 10 ingresantes, 6 de ellos son mujeres y 4 varones.

d) **Verificar la calidad de los datos.** En esta fase se hace un análisis de la calidad de los datos, es decir analizamos si los datos tienen inconsistencias teniendo en cuenta los siguientes detalles:

- Existen datos incompletos les falta la nota del promedio ponderado, muchos de los alumnos que ingresaron por modalidades no tienen nota de ingreso, por ejemplo, los alumnos que ingresaron por primeros puestos ingresaron sin dar examen en algunos años.
- Existen datos inconsistentes como por ejemplo la edad de ingreso, aparece que ingresaron con 115 años, debido al mal registro de la fecha de nacimiento.
- Algunos registros están vacíos o en algunos casos los datos ingresados no coinciden con la encuesta.

4. 3 PREPARACIÓN DE DATOS

En esta etapa se desarrolló las actividades para construir la base de datos final con el que vamos a construir los modelos, estos datos se obtienen de los datos brutos iniciales.

Estas tareas incluyen la selección de atributos y registros, también la transformación y la limpieza de los datos, para generar los modelos predictivos.

a) Selección de datos

Por ética hay variables que no se pueden usar, para el cual se procedió a eliminar las variables como apellidos y nombres y DNI del alumno, el código del alumno, también se realizó el cálculo de la edad de los alumnos de usando las columnas de fecha de inscripción y fecha de nacimiento, para que al final también eliminar las variables fecha de nacimiento y fecha de inscripción.

Luego la variable abigeo de procedencia se tuvo que extraer la ubicación de la región =EXTRAE([@Column5],11,2), para llenar la variable con la región de procedencia.

Como se desea conocer que métodos son más eficientes para predecir el rendimiento académico, es necesario primero conocer la condición del ingresante al finalizar el primer semestre si está con promedio ponderado mayor o igual a 11 (aprobado), caso contrario menor a 11 (desaprobado), agregamos una columna de datos condición del alumno de tipo categórico. A continuación, vamos a describir los atributos que hemos seleccionado para poder evaluar y aplicar los algoritmos de aprendizaje automático:

Tabla 24

Atributos seleccionados para la minería de datos

Atributo	Tipo	Descripción
Modalidad	Cualitativa – politómica	La modalidad por la que ingresaron a la universidad
Tipo_colegio	Cualitativa – dicotómica	Establece el tipo de colegio donde termino la secundaria
Ingreso	Cualitativa – dicotómica	Indica la opción con que ingreso a la universidad si es la primera o segunda opción
Procedencia	Cualitativa - politómica	Región de procedencia del alumno ingresante
Sexo	Cualitativa - dicotómica	Genero del estudiante
NotaIngreso	Cuantitativa – continua	Es la nota con la que ingreso el estudiante a la universidad
Edad	Cuantitativa – discreta	Edad con la ingreso el estudiante
TipoPrep	Cualitativa - politómica	Tipo de preparación que realizo el alumno para ingresar
Comosente	Cualitativa - politómica	La forma como se enteró de la UNAS y postular a la misma.
MOTPOSTULAR	Cualitativa - politómica	El motivo por el cual postula a la universidad
TRABAJA	Cualitativa - politómica	SI trabaja en el tiempo que estudia
DEP_ECONOMICA	Cualitativa - politómica	Establece de quien depende económicamente
VIVEPADRES	Cualitativa - politómica	Indica si sus padres están vivos
NUM_HERMANOS	Cualitativa - politómica	Establece la cantidad de hermanos que tiene
VIVE_CON	Cualitativa - politómica	Indica con quien vive actualmente mientras estudia
DONDEVIVE	Cualitativa - politómica	Indica el lugar donde vive actualmente
ESCUELA PROFESIONAL	Cualitativa - politómica	Estable la escuela profesional a la cual ingreso
PPS	Cuantitativa – continua	Promedio ponderado semestral del estudiante
CONDICION	Cualitativa – dicotómica	Si al final aprobó o desaprobó el semestre

Nota. La tabla muestra los atributos seleccionados para el modelo predictivo.

b) Limpieza de datos

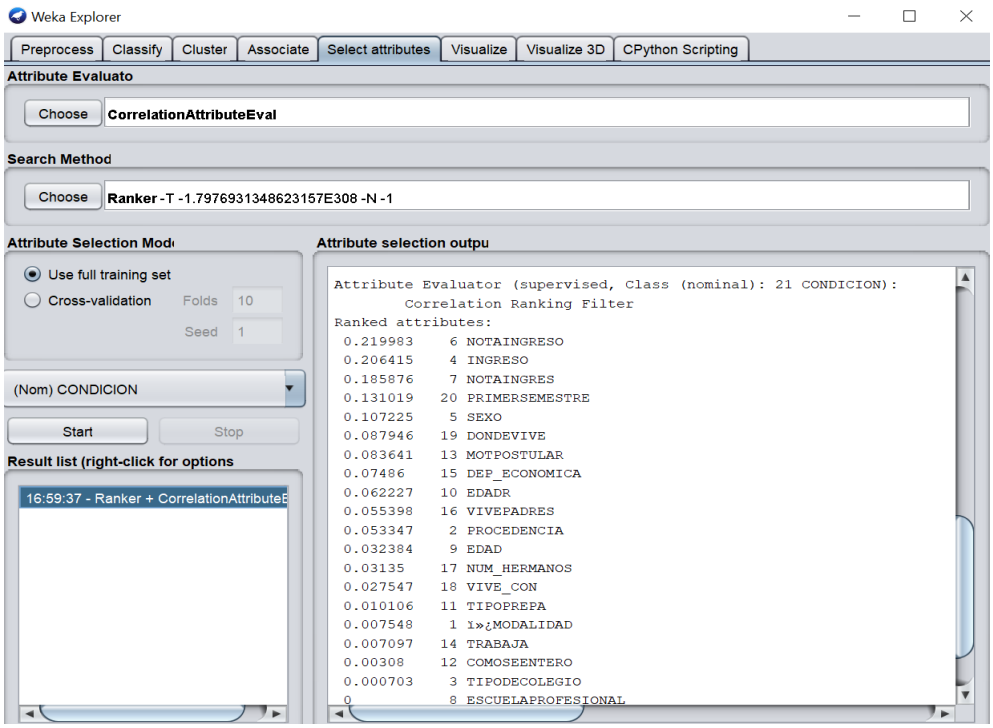
En esta fase describimos los factores más importantes que permitirán predecir el rendimiento académico, para ello primero de los 190 registros se procedió a eliminar los datos duplicados, se eliminó aquellos registros nulos en el promedio ponderado ya que ingresaron, pero no se matricularon en el ciclo correspondiente.

También se eliminaron algunos registros que no tenían las encuestas rellenadas, al igual que había datos inconsistentes como la edad de un estudiante con 118 años.

Para la limpieza se analizó cada uno de los atributos de la base de datos. En el software WEKA también se analizó que atributos permiten tener una mejor correlación, obteniéndose los siguientes resultados:

Figura 32

Correlación según los atributos en el software WEKA



Nota. La figura muestra la correlación de cada uno de los atributos con respecto al atributo condición.

c) Estructuración de datos

Para la construcción de los nuevos datos, algunos datos se normalizo, y los datos de promedio ponderado se clasifico como aprobado o desaprobado.

d) Integrar datos

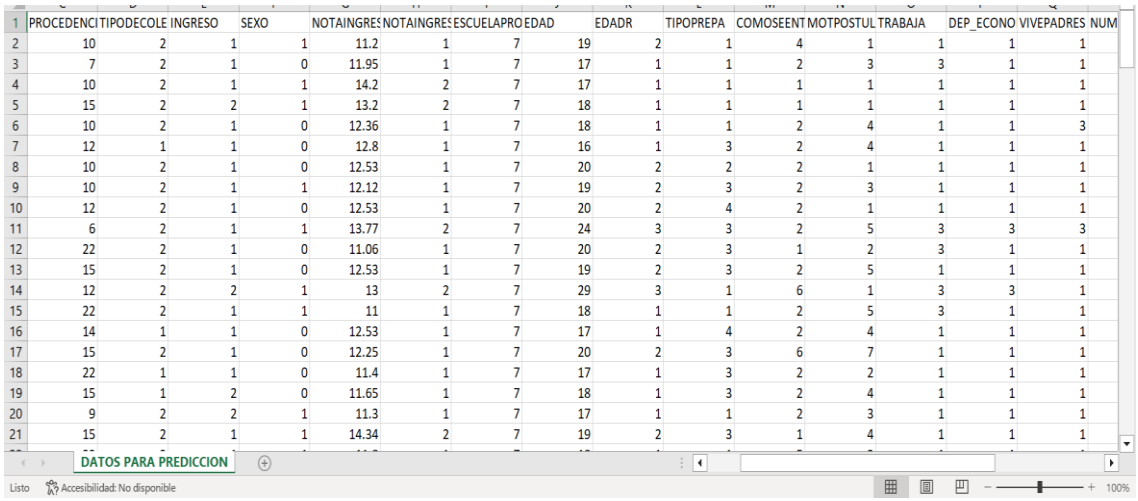
Los registros fueron integrados en una sola tabla llamada DATOS_PARA_PREDICCION que vamos a utilizar para aplicar la minería de datos las cuales fueron integrados de los registros obtenidos desde las oficinas de admisión, y de la base de datos de DICCA. Al final nos quedamos con un total de 155 registros.

e) Formatear datos

La tabla se convirtió en archivo tipo .csv para usar en el software WEKA, quedando de la siguiente manera:

Figura 33

Datos en el archivo .csv



	PROCEDENCI	TIPODECOLE	INGRESO	SEXO	NOTAINGRES	NOTAINGRES	ESCUELAPRO	EDAD	EDADR	TIPOPREPA	COMOSEENT	MOTPOSTUL	TRABAJA	DEP_ECONO	VIVEPADRES	NUM
2	10	2	1	1	11.2	1	7	19	2	1	4	1	1	1	1	
3	7	2	1	0	11.95	1	7	17	1	1	2	3	3	1	1	
4	10	2	1	1	14.2	2	7	17	1	1	1	1	1	1	1	
5	15	2	2	1	13.2	2	7	18	1	1	1	1	1	1	1	
6	10	2	1	0	12.36	1	7	18	1	1	2	4	1	1	3	
7	12	1	1	0	12.8	1	7	16	1	3	2	4	1	1	1	
8	10	2	1	0	12.53	1	7	20	2	2	2	1	1	1	1	
9	10	2	1	1	12.12	1	7	19	2	3	2	3	1	1	1	
10	12	2	1	0	12.53	1	7	20	2	4	2	1	1	1	1	
11	6	2	1	1	13.77	2	7	24	3	3	2	5	3	3	3	
12	22	2	1	0	11.06	1	7	20	2	3	1	2	3	1	1	
13	15	2	1	0	12.53	1	7	19	2	3	2	5	1	1	1	
14	12	2	2	1	13	2	7	29	3	1	6	1	3	3	1	
15	22	2	1	1	11	1	7	18	1	1	2	5	3	1	1	
16	14	1	1	0	12.53	1	7	17	1	4	2	4	1	1	1	
17	15	2	1	0	12.25	1	7	20	2	3	6	7	1	1	1	
18	22	1	1	0	11.4	1	7	17	1	3	2	2	1	1	1	
19	15	1	2	0	11.65	1	7	18	1	3	2	4	1	1	1	
20	9	2	2	1	11.3	1	7	17	1	1	2	3	1	1	1	
21	15	2	1	1	14.34	2	7	19	2	3	1	4	1	1	1	

DATOS PARA PREDICCION

Listo Accesibilidad: No disponible 100%

Nota. La figura muestra los datos que están en el archivo CSV.

4. 4 MODELAMIENTO

En esta fase, seleccionamos y aplicamos las técnicas de modelado en el software WEKA 3.9, se calibran los parámetros y hiperparametros para obtener los valores óptimos para la

predicción con los diferentes modelos que vamos a trabajar:

a) Selección de la técnica de modelado

Para el modelado utilizaremos el software WEKA 3.9, usaremos 5 algoritmos, divididos en 5 grupos de datos de los años 2015 al 2019.

EL primer paso es identificar las variables que son significativas para la predicción para ello vamos a usar el software WEKA.

Tabla 25

Indicadores que están relacionadas con el rendimiento académico

VARIABLE	Chi- cuadrado	gl	Sig.
MODALIDAD DE INGRESO	17.617	8	0.024
MODALIDAD DE INGRESO (6)	8.025	1	0.005
OPCION DE INGRESO (1)	6.604	1	0.010
NOTA DE INGRESO	7.501	1	0.006
TIPO DE PREPARACION	6.755	3	0.080
TIPO DE PREPARACION (3)	4.173	1	0.041
PRIMER SEMESTRE	8.396	1	0.004

Nota: La table muestra la significancia de la correlación de indicadores con el rendimiento académico.

Con los datos extraídos donde se observa indicadores sociales, económicos y académicos, se puede observar de los datos que existen atributos que son significativos cuyo nivel de confianza supera el 95% ya que se obtuvo un valor- P menor a 0.05. La significancia de las demás variables podemos observarlo en el ANEXO 3

Tabla 26

Indicadores significativos para la predicción del rendimiento académico

	Error estándar	gl	Sig.	Exp(B)
TIPO DE COLEGIO(1)	1.131	1	0.214	0.245
OPCION DE INGRESO(1)	1.220	1	0.001	0.015
SEXO(1)	0.728	1	0.044	4.336
NOTA DE INGRESO	0.269	1	0.000	2.582
TIPO DE PREPARACION(3)	1.333	1	0.127	7.652
TRABAJA		2	0.030	
TRABAJA(2)	2.185	1	0.010	0.004
NUMERO DE HERMANOS		3	0.045	
NUMERO DE HERMANOS(2)	1.306	1	0.028	17.758
NUMERO DE HERMANOS(3)	1.359	1	0.042	15.764
PRIMER SEMESTRE	0.000	1	0.000	1.000

Nota: La tabla muestra los indicadores que son significativas para la predicción del rendimiento académico.

De acuerdo con la tabla los atributos de las dimensiones sociales, económicos y académicos relacionados en con rendimiento académico de los estudiantes ingresantes se observa que los indicadores académicos tienen mayor significancia en el rendimiento académico, dentro de ellos tenemos la opción de ingreso, la nota de ingreso todos ellos con un p-valor < 0.05.

Los indicadores sociales como sexo, si trabajan y el número de hermanos son significativos para predecir el rendimiento académico ya que tienen un p-valor < 0.05.

La significancia de las demás variables podemos encontrarlo en el ANEXO 4.

Evaluación de a exactitud

Los modelos de aprendizaje automático evaluados con el software Weka nos mostró los diferentes porcentajes de exactitud. Los cuales fueron clasificados en la siguiente tabla.

Tabla 27

Porcentaje de exactitud de cada uno de los modelos

Vote	Random Forest	IBK	NaibeBayes	Bagging
95.83	100	100	79.17	81.25
53.33	93.33	93.33	76.67	83.33
84.62	100	100	71.79	97.43
71.05	100	100	78.95	84.21
70.32	98.71	98.71	63.22	82.58

*Nota***:** La tabla muestra la exactitud de acierto de cada uno de los modelos.

Observamos en la tabla que los modelos que mejor exactitud tienen son los modelos de Random Forest y KNN (IBK) mas conocido como k vecinos más cercanos con una igual exactitud en los diferentes grupos evaluados.

En comparación con la tesis de (Candia Oviedo, 2019) utilizo la metodología CRISP-DM y concluye que el rendimiento académico de los estudiantes depende de diversas variables tales como factores sociodemográficos y socioeconómicos, y a pesar de ello predice el rendimiento académico de los alumnos con una exactitud del 69%. En cambio (Yamao, 2018) obtuvo mejores resultados con el algoritmo arboles de decisión con una exactitud de 82.7% las variables que mas influyeron fueron la nota del examen de acceso, el sexo, la edad , el método de ingreso y el tiempo de desplazamiento del centro de estudios a su domicilio. Por otro lado (Contreras & Fuentes, 2020) utilizo el algoritmo de árbol de decisión , KNN, SVC y obtuvo resultados con una exactitud de 90 %. También (Quiñones & Carrasco, 2020) nos indica que usaron la metodologia CRISP-DM además del software weka con los algoritmos de J48graft, J48 y PART logrando obtener un exactitud de 83 %. De igual forma (Jiménez, 2018) nos indica que trabajo con 26 variables economicas y sociodemograficas logrando obtener con modelo arboles de decision, reglas de asociacion, analisis de correlacion logrando obtener una exactidu de 81,1159%.

Para identificar al modelo con mejores métricas de exactitud usamos el ANOVA de

Friedman aplicado de las diferentes exactitudes obtenidas con el software weka.

Tabla 28

Valores de los rangos de los modelos predictivos con el ANOVA de Friedman

Modelos	Rango promedio
Vote	1.80
RamdomForest	4.50
Ibk	4.50
NaiveBayes	1.40
Bagging	2.80

Nota: La tabla muestra los valores de los rangos de los diferentes modelos

De acuerdo con la tabla se observa que los modelos con mejor rango promedio son el Random Forest y k vecinos más cercanos (IBK).

Tabla 29

ANOVA de Friedman de la Exactitud de los Algoritmos de Machine Learning y prueba post hoc ambas a un nivel de significancia de α = 0.05

N	**Chi-cuadrado**	**gl**	**Sig. asintótica**
5	17.979	4	0.001

Muestra 1 - Muestra 2	**Estadistico de contraste**	**Error**	**Desv. Estadisctico de Contraste**	**Sig**	**Sig. Ajust.**
NaiveBayes:Vote	0.40	1.00	0.40	0.689	1.000
NaiveBayes-Bagging	-1.40	1.00	-1.40	0.162	1.000
NaiveBayes-RamdomForest	3.10	1.00	3.10	0.002	0.019
NaiveBayes-Ibk	3.10	1.00	3.10	0.002	0.019
Vote-Bagging	-1.00	1.00	-1.00	0.317	1.000
Vote-RamdomForest	-2.70	1.00	-2.70	0.007	0.069
Vote-Ibk	-2.70	1.00	-2.70	0.007	0.069
Bagging-RamdomForest	1.70	1.00	1.70	0.089	0.891
Bagging-Ibk	1.70	1.00	1.70	0.089	0.891
RamdomForest-Ibk	0.00	1.00	0.00	1.000	1.000

Nota: La tabla nos muestra los valores del ANOVA de Friedman para la exactitud y la prueba post hoc para la comparación de los modelos.

En la tabla cada fila prueba la hipótesis nula de las distribuciones de la muestra 1 y muestra 2 son las mismas. Se muestran las significancias asintóticas con el nivel de significancia de 0.05, los valores de significancia se han ajustado mediante la corrección de Bonferroni para varias pruebas.

De acuerdo a las pruebas de ANOVA los modelos con mejor exactitud tenemos a Random Forest e Ibk (k vecinos más cercanos) cada uno de ellos con una exactitud de 98.4%, también observamos que el que tiene menor porcentaje de exactitud es el modelo de naivebayes con 73,96 %, por otro lado, el modelo ensamblado vote tiene una exactitud de 75, 03% mientras que el otro modelo ensamblado tiene una exactitud de 85,76%.

Evaluación de la precisión

Los modelos de aprendizaje automático evaluados con el software Weka nos mostró los diferentes porcentajes de precisión tanto para los aprobados como para los desaprobados. La precisión de los aprobados se muestra en la siguiente tabla.

Tabla 30

Porcentaje de precisión de los aprobados de cada uno de los modelos

Vote	Random Forest	IBK	NaibeBayes	Bagging
100	100	100	69.2	87.5
53.3	92.9	87.5	88.9	100
83.3	100	100	82.4	95.7
73.1	100	100	75.9	87
67.1	98.8	98.6	66	91.2

Nota: La tabla muestra la precisión en cada uno de los modelos en el software weka.

En la tabla se observa que el modelo con mejor precisión es Random Forest con una precisión de 98.34% mientras que muy cerca está el modelo de k vecinos más cercanos (IBK) con 97.22%, también el que tiene menor porcentaje de precisión es el modelo de ensamblado vote con 75,36%.

Para analizar las diferencias de los modelos vamos a usar el ANOVA de Friedman para la precisión de los aprobados en los diferentes modelos.

Tabla 31

Valores de los rangos de los modelos predictivos de la precisión de los aprobados con el ANOVA de Friedman

Modelos	Rango promedio
Vote	2.00
RamdomForest	4.40
Ibk	3.80
NaiveBayes	1.60
Bagging	3.20

Nota: La tabla muestra los valores de los rangos de los diferentes modelos en la precisión de los aprobados.

Con el ANOVA de Friedman el que mejor rango de promedio tiene es el Random Forest pero muy cerca está el modelo de k vecinos más cercanos (IBK).

Tabla 32

ANOVA de Friedman de la Precisión de los aprobados de los Algoritmos de Machine Learning y prueba post hoc ambas a un nivel de significancia de α = 0.05

N	Chi-cuadrado	gl	Sig. asintótica		
5	11.915	4	0.018		
Muestra 1 - Muestra 2	**Estadistico de contraste**	**Error**	**Desv. Estadisctico de Contraste**	**Sig**	**Sig. Ajust.**
NaiveBayes:Vote	0.40	1.00	0.40	0.689	1.000
NaiveBayes-Bagging	-1.60	1.00	-1.60	0.110	1.000
NaiveBayes-RamdomForest	2.80	1.00	2.80	0.005	0.050
NaiveBayes-Ibk	2.20	1.00	2.20	0.028	0.278
Vote-Bagging	-1.20	1.00	-1.20	0.230	1.000
Vote-RamdomForest	-2.40	1.00	-2.40	0.016	0.164
Vote-Ibk	-1.80	1.00	-1.80	0.072	0.719
Bagging-RamdomForest	1.20	1.00	1.20	0.230	1.000
Bagging-Ibk	0.60	1.00	0.60	0.549	1.000
RamdomForest-Ibk	0.60	1.00	0.60	0.549	1.000

Nota: La tabla nos muestra los valores del ANOVA de Friedman para la precisión de los aprobados y la prueba post hoc para la comparación de los modelos.

En la tabla cada fila prueba la hipótesis nula de las distribuciones de la muestra 1 y muestra 2 son las mismas. Se muestran las significancias asintóticas con el nivel de

significancia de 0.05, los valores de significancia se han ajustado mediante la corrección de Bonferroni para varias pruebas.

Tabla 33

Porcentaje de precisión de los desaprobados de cada uno de los modelos

Vote	Random Forest	IBK	NaibeBayes	Bagging
94.3	100	100	82.9	80
53.3	93.8	100	71.14	76.2
86.7	100	100	63.6	100
66.7	100	100	88.9	80
74	98.6	98.8	61.8	77.6

Nota: La tabla muestra la precisión en cada uno de los modelos en el software weka.

En la tabla se observa que el modelo con mejor precisión para los desaprobados es el modelo de k vecinos más cercanos (IBK) con un 99,7 % de precisión, muy cerca esta el modelo de Random Forest con una precisión de 98,48 %, por el contrario, el algoritmo con menor porcentaje de precisión es el modelo de naive bayes con una precisión de 73,67%.

Para analizar las diferencias de los modelos vamos a usar el ANOVA de Friedman para la precisión de los desaprobados en los diferentes modelos.

Tabla 34

Valores de los rangos de los modelos predictivos de la precisión de los desaprobados con el ANOVA de Friedman

Modelos	Rango promedio
Vote	1.8
RamdomForest	4.2
Ibk	4.6
NaiveBayes	1.8
Bagging	2.6

Nota: La tabla muestra los valores de los rangos de los diferentes modelos en la precisión de los desaprobados.

Con el ANOVA de Friedman el que mejor rango de promedio tiene es el modelo de k vecinos más cercanos (IBK) pero muy cerca está el modelo de Random Forest.

Tabla 35

ANOVA de Friedman de la Precisión de los desaprobados de los Algoritmos de Machine Learning y prueba post hoc ambas a un nivel de significancia de α = 0.05

N	Chi-cuadrado	gl	Sig. asintótica
5	14.979	4	0.005

Muestra 1 - Muestra 2	Estadistico de contraste	Error	Desv. Estadisctico de Contraste	Sig	Sig. Ajust.
NaiveBayes:Vote	0.00	1.00	0.00	1.000	1.000
NaiveBayes-Bagging	-0.80	1.00	-0.80	0.424	1.000
NaiveBayes-RamdomForest	2.40	1.00	2.40	0.016	0.164
NaiveBayes-Ibk	2.80	1.00	2.80	0.005	0.050
Vote-Bagging	-0.80	1.00	-0.80	0.424	1.000
Vote-RamdomForest	-2.40	1.00	-2.40	0.016	0.164
Vote-Ibk	-2.80	1.00	-2.80	0.005	0.051
Bagging-RamdomForest	1.60	1.00	1.60	0.110	1.000
Bagging-Ibk	2.00	1.00	2.00	0.046	0.455
RamdomForest-Ibk	-0.40	1.00	-0.40	0.689	1.000

Nota: La tabla nos muestra los valores del ANOVA de Friedman para la precisión de los desaprobados y la prueba post hoc para la comparación de los modelos

En la tabla cada fila prueba la hipótesis nula de las distribuciones de la muestra 1 y muestra 2 son las mismas. Se muestran las significancias asintóticas con el nivel de significancia de 0.05, los valores de significancia se han ajustado mediante la corrección de Bonferroni para varias pruebas.

En su investigación (Alvarez Gonzaga, 2021) obtuvo una precisión de 93,67% para el modelo de naive bayes mientras que para el árbol de decisión su precisión fue de 93. 67%. Por otro lado (Puga & Torres, 2023) en su investigacion ubtuvo una presicion del 97, 6 % con resdes neuronales aritificiales y una integridad del 100% demostrando la eficacia de los modelos de aprendizaje automatico en la predicción del rendimiento automático. (Contreras & Fuentes, 2020) utilizo los modelo sde arbol de decisión, KNN, SVC y perceptron y tambien algoritmos emsamblados como sataking y blending que untuvieron una presicion que oscila entre 85% 7

75 % tanto para el entrenamiento como para el test.

b) Generación de la prueba de diseño

Las pruebas de diseños las hemos obtenido en la elección de las técnicas de modelado de acuerdo con la exactitud y la precisión de los modelos.

Primero agrupamos los datos tal como se obtuvieron desde la base de datos y los entrenamos a dos de los modelos que son Random Forest y k vecinos más cercanos (IBK), con dichos modelos se obtuvo los mejores porcentajes de exactitud y precisión.

c) Construcción del modelo

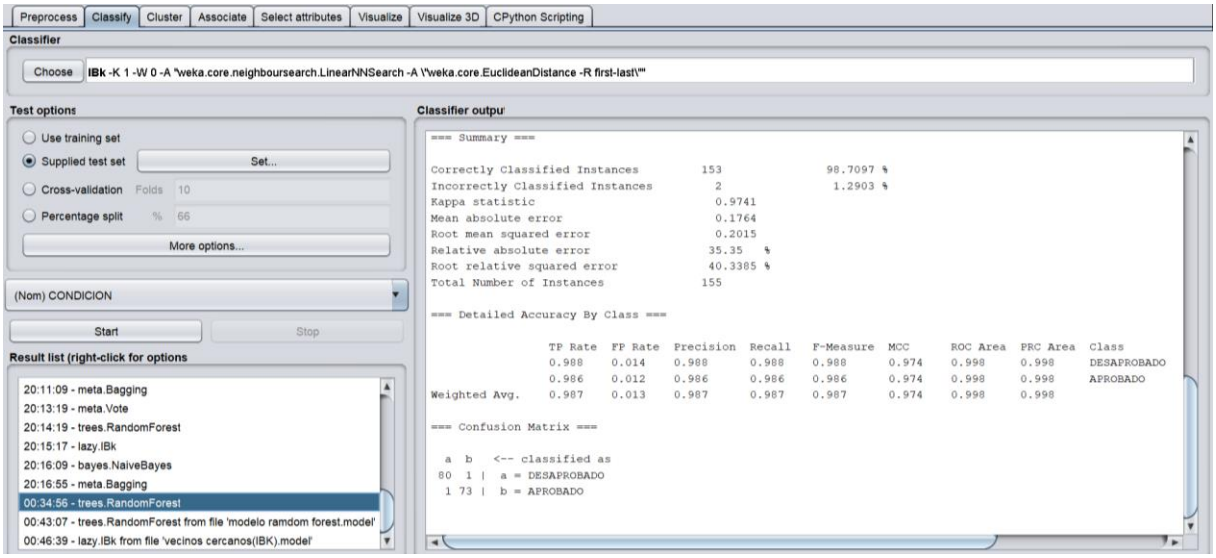
Para construir el modelo tenemos dos grupos de datos, primero para el entrenamiento tenemos 155 registros ya evaluados y procesados que son alumnos ingresantes desde los años 2015 al 2018.

Para la predicción o grupo (test) vamos a utilizar los datos de los alumnos que estudiaron el primer ciclo del año 2019-I que son 47 registros, los demás años 2020 al 2022 no se utilizó para la predicción por motivo que las clases fueron virtuales, el atributo de condición "Aprobado" o "Desaprobado", serán datos desconocidos para poder realizar la predicción, los mismos que tendrán signos de interrogación, el algoritmo ya entrenado será el encargado de clasificar si el rendimiento del estudiante tendrá la condición de aprobado o desaprobado al final de ciclo, el mismo que tendremos que corroborar con los valores reales que se tienen en la base de datos " DATOS_ALUMNOS_PREDICCION_COMPROBACION".

Se realiza el entrenamiento del modelo para poder guárdalo, en el siguiente grafico observamos el entrenamiento del modelo:

Figura 34

Entrenamiento del modelo Random Forest.



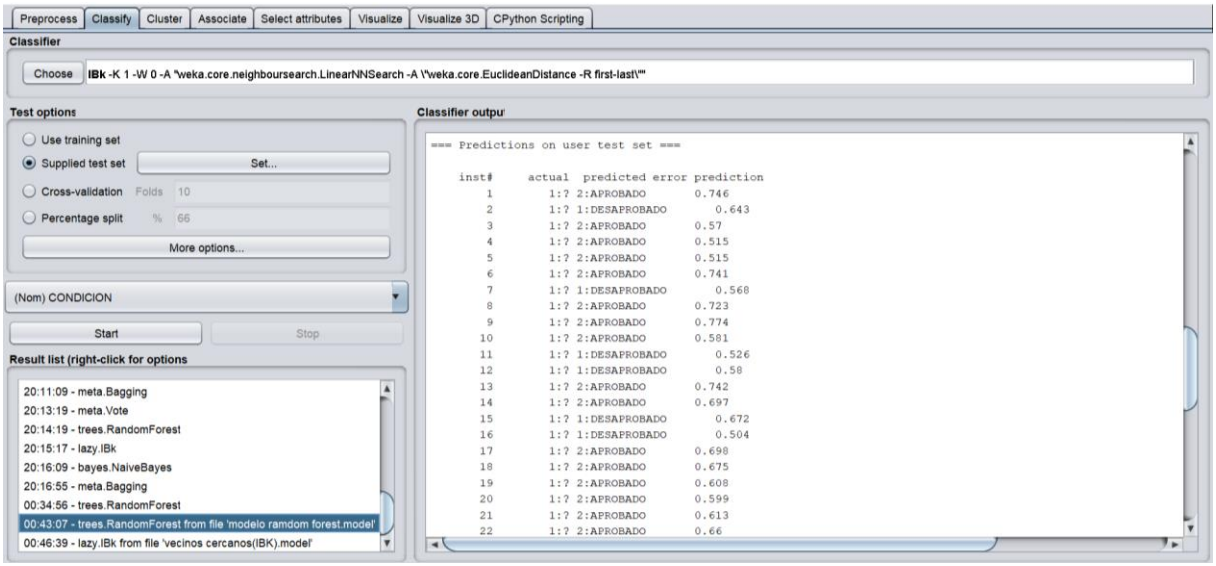
Nota. En la figura se muestra los resultados del entrenamiento de Random Forest.

d) Evaluación del modelo

Con el modelo ya construido y entrenado en el software WEKA 3.9 se realiza el test a la base de datos DATOS_ALUMNOS_PREDICCION.arff teniendo los siguientes resultados:

Figura 35

Predicción con el modelo Random Forest.



Nota. En la figura se observa el resultado de la predicción con el porcentaje de acierto.

4. 5 Evaluación

En esta fase vamos a evaluar el modelo construido en la fase anterior.

a) Evaluación de los resultados

El resultado obtenido en la predicción vamos a comparar con los resultados reales en una hoja Excel para determinar en las cuales las predicciones fueron CORRECTAS y en cuales INCORRECTAS.

Tabla 36

Instancias clasificadas de manera correcta e incorrecta

DATOS REALES	PREDICCION	CONCLUSION
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	DESAPROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO

DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	DESAPROBADO	INCORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	APROBADO	INCORRECTO
DESAPROBADO	APROBADO	INCORRECTO
APROBADO	APROBADO	CORRECTO
DESAPROBADO	DESAPROBADO	CORRECTO
APROBADO	DESAPROBADO	INCORRECTO

Nota. La tabla muestra las instancias con las que se realizó la predicción con el modelo entrenado de Random Forest.

b) Proceso de revisión

Se realizó la comparación de todas las instancias clasificadas con el modelo entrenado en el software WEKA 3.9

c) Determinación de futuras fases

Los pasos que sigue es realizar la predicción de los nuevos ingresantes, y cada que los datos se acumulan en cada año con los nuevos ingresantes, la predicción tendrá un mayor porcentaje de acierto.

4. 6 DESPLIEGUE DEL PROYECTO

Los resultados obtenidos serán entregados a las autoridades de la Facultad para que tomen las medidas correctivas para mejorar el rendimiento académico de los estudiantes ingresantes a la carrera de Ingeniería de Industrias Alimentarias.

4. 7 CONTRASTACIÓN DE HIPOTESIS

Los resultados obtenidos en el software WEKA 3.9 utilizando algoritmos de minería de datos en la información de los estudiantes que ingresaron en los años del 2015 - 2109 de acuerdo

con la hipótesis planteada en la investigación.

HG: La predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS se realiza de manera significativa con el modelo de aprendizaje automático árbol de decisión.

Las predicciones que se realizó con los distintos modelos de aprendizaje automático tales como modelo ensamblado Vote, modelo árbol de decisión Random Forest, modelo k vecinos más cercanos (IBK), modelo Naybe Bayes y modelo ensamblado Bagging se encuentro que los modelos de Random Forest y K vecinos más cercanos tienen porcentajes muy similares como se observa en las tablas 27, 30 y 33.

HE1: El modelo de aprendizaje automático con mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.

En la tabla 27 observamos que los modelos Random Forest y k vecinos más cercanos tienen los mismos porcentajes de exactitud por lo que podemos decir que ambos modelos son muy significativos para la predicción del rendimiento académico lo cual podemos constatar con la prueba de ANOVA de Friedman en la tabla 28 donde observamos que los rangos promedios de ambos modelos son 4.5 y son significativos lo cual se demuestra luego con la prueba post hoc en la tabla 29 con un nivel de significancia menor a 0.05 el cual nos muestra que los modelos Random Forest y k vecinos más cercanos (IBK) tienen un diferencia significativa con respecto a los demás modelos que as du ves podemos decir que son los modelos con mejore exactitud para la predicción del rendimiento académico ambos con una exactitud promedio de 98,4%.

HE2: El modelo de aprendizaje automático con mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.

Para la precisión se ha analizado tanto de precisión para los aprobados como para los desaprobados, para el caso de precisión de los aprobados (tabla 30) el que mejor precisión tiene es el modelo de Random Forest con 98,4% de precisión, mientras que vecinos cercnaos (IBK) tiene una precisión de 97.22%. en los rangos promedios de la prueba de ANOVA de Friedman (tabla 31) Random Forest tiene el mejor promedio (4.4) luego en la prueba de post host (tabla 32) con una significancia de 0.05 y comparación de los modelos de 1 en 1 se observa que hay una diferencia para Random Forest con una mejor diferencia, haciendo que sea el modelo con mejor precisión para los aprobados. Para el caso de la precisión de los desaprobados (tabla 33) el modelo que tiene mejor precisión es k vecinos más cercanos (IBK), con un porcentaje promedio de precisión de 99,7 % mientras que Random Forest tiene 98,48% de precisión, en la prueba de ANOVA de Friedman (tabla 34) el modelo con mejor rango promedio es k vecinos más cercanos (IBK), luego en la prueba post hoc se observa que la prueba es significativa lo cual nos indica que hay diferencias en los modelos, ambos modelos tanto Random Forest como k vecinos más cercanos son significativos en la precisión para la predicción del rendimiento académico de los alumnos ingresantes.

V. CONCLUSIONES

- El rendimiento académico de los alumnos ingresantes es un tema muy complejo y con los datos sociales, económicos o académicos, usando la metodología CRISP-DM y técnicas de minería de datos se determinó que los modelos de aprendizaje automático que predicen significativamente el rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería de Industrias Alimentarias son Random Forest y k vecinos más cercanos (KNN o IBK). Se identifico que los indicadores claves para predecir el rendimiento académico, en el caso de los indicadores académicos se tiene la opción de ingreso, la nota de ingreso que es la que más influencia tiene, en los indicadores sociales tenemos el sexo del ingresante y el número de hermanos dentro de la familia. En conclusión, la Facultad de Ingeniería en Industrias Alimentarias puede predecir el rendimiento académico de los alumnos ingresantes de manera significativa con ambos modelos, el uso de estos modelos de aprendizaje automático puede influir para mejorar el rendimiento académico de los alumnos ya que nos permite identificar los posibles alumnos aprobados y desaprobados, beneficiando a la planificación de estrategias dentro de la facultad.
- Se determino que los algoritmos con mejor exactitud son los modelos de Random Forest e Ibk (k vecinos más cercanos) cada uno con una exactitud de 98.4% en los entrenamientos. Luego con la prueba de ANOVA de Friedman de la Exactitud y las pruebas post hoc con una significancia de 0.05 se obtuvo que los modelos de Random Forest e Ibk obtuvieron un rango promedio de 4.5 y una significancia ajustada en la comparación de los modelos 0.019 logrando demostrar que son significativas dichos modelos.
- De determino que el modelo con mejor precisión para los aprobados es el modelo de Random Forest con una precisión de 98.34 %, también el modelo de k vecinos más cercanos (IBK) tiene una precisión para los aprobados del 97.22%, para el caso de la

precisión para los desaprobados el modelo de vecinos cercados obtuvo una precisión de 99.7% y Random Forest tiene una precisión de 98.48%, en la evaluación del ANOVA de Friedman para prueba no paramétrica se tiene la precisión para los aprobados el rango promedio para Random Forest es de 4.4 y el rango promedio para los desaprobados es k vecinos más cercanos es 4,6 logrando una significancia menor a 0.05 en ambas pruebas.

PROPUESTAS PARA FUTURAS INVESTIGACIONES

- Se recomienda a la Facultad de Ingeniería en Industrias Alimentarias realizar la predicción del rendimiento académico usando todos los indicadores ya que mejoran significativamente la predicción.
- Usar la metodología CRISP-DM, porque nos brinda una guía de 6 fases muy bien definidas en procesos de modelos predictivos, es una herramienta muy sencilla de entender independientemente de la herramienta de minería de datos a usar y algoritmo de aprendizaje automático a usar.
- A las autoridades de la Facultad de Ingeniería en Industrias Alimentarias, usar el modelo Random Forest y vecinos cercanos (IBK) así como el software WEKA para predecir el rendimiento académico de los estudiantes y de acuerdo con los resultados brindar especial atención a los alumnos que tienen mayor probabilidad de salir desaprobados y tomar medidas correctivas para brindarles asesorías para tener mejor calidad y éxito.
- Usar las técnicas de minería de datos y los algoritmos de aprendizaje automático de acuerdo con la data que se tiene y agregando otras variables de los datos históricos, sería una herramienta para apoyar a la toma de decisiones y a mejoras las políticas para tener un mejor rendimiento académico, además que esto permitirá tener un impacto en la sociedad y tener una mejor imagen de nuestros alumnos.
- Para futuras investigaciones se recomienda incluir variables como los cursos que se dictan, y los profesores encargados, así como su clasificación en SISTEMA DE FOCALIZACION DE HOGARES (sisfoh).

VI. REFERENCIAS BIBLIOGRAFICAS

Flores Urgilés , C., Flores Urgilés, C., & Quevedo Sacoto, A. (2022). Análisis de sentimiento político en redes sociales, como instrumento. *Pro Sciences: Revista De Producción, Ciencias E Investigación, 5*(45), 15. https://doi.org/: https://doi.org/10.29018/issn.2588-1000vol6iss45.

Russo, C. (2019). Mineria de datos aplicada a estrategias para minimizar el rezago academico y desercion universitaria en carreras de informatica de la UNNOBA. *Tesis para obtener grado de Doctor*. Universidad Nacional de la plata, Buenos Aires. https://doi.org/http://sedici.unlp.edu.ar/bitstream/handle/10915/79958/Documento_completo.pdf-PDFA1b.pdf?sequence=1&isAllowed=y

Alania Ricaldi, P. (2018). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR LA DESERCIÓN ESTUDIANTIL DE LA FACULTAD DE INGENIERÍA DE LA UNIVERSIDAD NACIONAL DANIEL ALCIDES CARRIÓN*. Pasco. http://repositorio.undac.edu.pe/bitstream/undac/829/1/T026_40573846_M.pdf

Alvarez Gonzaga, B. R. (2021). *ANÁLISIS COMPARATIVO DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A BUSINESS INTELLIGENCE*. repositorio.uss.

Analytics, P. (15 de setiembre de 2017). *Penta Analytics*. https://www.analytics.cl/educacion/peru-27-los-ingresantes-universidades-privadas-abandonan-carrera-primer-ano-estudios/#:~:text=Retail-,En%20Per%C3%BA%2C%20el%2027%25%20de%20los%20ingresantes%20a%20universidades%20privadas,el%20primer%20a%C3%B1o%20de%20e

Asto Rodriguez, E. M. (2020). *"Framework Basado En Minería De Datos Para La Obtención Del Perfil De Egreso De Los Estudiantes Del Programa De Ingenieria Mecatrónica De La Universidad Nacional De Trujillo Año 2019"*. Trujillo_ Perú. file:///D:/Victor%20Ponce_UNAS_Ing.%20de%20sistemas%20e%20Informatica/REP_MAEST.INGE_EMERSON.ASTO_FRAMEWORK.BASADO.MINER%C3%8DA.DATOS.OBTENCI%C3%93N.PERFIL.EGRESO.ESTUDIANTES.PROGRAMA.INGENIERIA.MECATR%C3%93NICA.UNIVERSIDAD.NACIONAL.TRUJILLO.2019.pdf

Borja Robalino, R., Monleón Getino, A., & Rodellar, J. (2020). *Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning.* Revista Ibérica de Sistemas e Tecnologias de Informação.

Candia Oviedo, D. I. (2019). *"Predicción Del Rendimiento Académico De Los Estudiantes De*

La Unsaac A Partir De Sus Datos De Ingreso Utilizando Algoritmos De Aprendizaje Automático". Cusco_Perú. https://repositorio.unsaac.edu.pe/bitstream/handle/20.500.12918/4120/253T20191024_TC.pdf?sequence=1&isAllowed=y

Cevallos Molina, S., & Trujillo Utreras, V. (2018). *Minería de datos aplicada a la clasificación del rendimiento académico.* Calceta. https://repositorio.espam.edu.ec/xmlui/bitstream/handle/42000/862/TTC10.pdf?sequence=1&isAllowed=y

Contreras Bravo, L. E., Fuentes López, H. J., & Rivas Trujilllo, E. (31 de Diciembre de 2021). *Revista redipe*. https://doi.org/https://doi.org/10.36260/rbr.v10i13.1737

Contreras Bravo, L. E., Tarazona Bermúdez, G. M., & Rodríguez Molano, J. I. (mayo-agosto de 2021). Tecnología y analítica del aprendizaje: una revisión a la literatura. *41*(2). https://revistas.udistrital.edu.co/index.php/revcie/article/view/17547

Contreras, L. E., & Fuentes, H. J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *13*(5). https://www.scielo.cl/scielo.php?pid=S0718-50062020000500233&script=sci_arttext

Cruz, N. P., Maña, M. J., & Mata, J. (2020). Aprendizaje Automático versus Expresiones Regulares en la Detección de la Negación y la Especulación en Biomedicina. *45*, 77-85. https://www.redalyc.org/pdf/5157/515751745008.pdf

Estrada Molina, O., & Fuentes Cancell, . R. (2022). ¿Se logra predecir el rendimiento académico? Un análisis desde la tecnología educativa. https://revistascientificas.us.es/index.php/fuentes/article/view/14278/16546

Fabara Sarmiento, Z. J., Diaz Vera, J. P., & Ruiz Ramirez, A. K. (2022). Data Science para la Predicción del Rendimiento Académico Data Science for Prediction of Academic Achievement Ciência de dados para a previsão do desempenho acadêmico. *8*(1), 99-112. https://dominiodelasciencias.com/ojs/index.php/es/article/view/2481

Gamboa Unsihuay, J. E., & Salinas Flores, J. W. (31 de enero de 2022). Predicción De La Situación Académica En Alumnos De Pregrado Usando Algoritmos De Machine Learning. *1*(27). http://ceaa.espoch.edu.ec:8080/revista.perfiles/faces/Articulos/Perfiles27Art1.pdf

Gonzales Tirados, R. M. (1989). *Analisis dc las causas del fracaso académico en la Universidad Politécnica dc Madrid.* CENTRO DE PUBLICACIONES -Secretaría Goneral Tknica.

Guersenzvaig Elisava, A., & Casacuberta, D. (20 de mayo de 2022). "La quimera de la objetividad algorítmica: dificultades del aprendizaje automático en el desarrollo de una noción no normativa de salud". *8*(1), 35-56. http://doi.org/10.12795/IESTSCIENTIA.2022.i01.03

Hewlett Packard. (23 de Agosto de 2023). *Modelo de aprendizaje automático*. https://www.hpe.com/lamerica/es/what-is/ml-models.html

Holgado Apaza, 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_ Perú. file:///D:/Victor%20Ponce_UNAS_Ing.%20de%20sistemas%20e%20Informatica/Luis_Alberto_Holgado_Apaza.pdf

Jiménez Giraldo, J. (2018). *Minería de datos educativos: análisis de los factores económicos, sociales y demográficos que influyen en el desempeño de las Pruebas Saber-Pro en estudiantes de ingeniería en Antioquia.* MEDELLÍN. https://repository.upb.edu.co/handle/20.500.11912/4317

Luna Gonzales, J. (8 de Febrero de 2018). *Medium*. https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2

Méndez, I. I., Ramírez Reyes, A., & Mora-Gutiérrez, R. A. (2020). Aprendizaje automático aplicado en física:Una revisión de la literatura científica. 803–816. https://rcs.cic.ipn.mx/2020_149_8/Aprendizaje%20automatico%20aplicado%20en%20fisica_%20Una%20revision%20de%20la%20literatura%20cientifica.pdf

Molina, O. E., & Fuentes Cancel, D. R. (15 de setiembre de 2021). ¿Se logra predecir el rendimiento académico? Un análisis desde latecnología educativa. 363-375. https://institucional.us.es/revistas/fuente/23_3/14278.pdf

Morales Agurto, 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".* Huaraz_ Perú. http://repositorio.unasam.edu.pe/bitstream/handle/UNASAM/2884/T033_46490471_T.pdf?sequence=1&isAllowed=y

Mounier, M., Acosta, K., Favret, F., & Zamudio, E. (2020). *Clasificación Automática de Estudios Epidemiológicos Referentes a Distintos Tipos de Cáncer Utilizando un Meta-estimador Bagging con Naïve Bayes.* Simposio Argentino de Inteligencia Artificial.

Nieto Jeux, A. (2021). Algoritmos de Aprendizaje Automático.Un Estudio de su Difusión y Utilización. https://oa.upm.es/68484/1/TFG_ALEJANDRO_NIETO_JEUX.pdf

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Mineria de datos educativos: Analisis del desempeño de estudiantes de ingeneria en las pruebas saber-pro. *Dialnet*, 1-13. https://doi.org/https://revistas.elpoli.edu.co/index.php/pol/article/view/1499/1219

Páez Rico, A., & Ramírez Gaytán, N. D. (31 de agosto de 2022). Predicción del rendimiento académico utilizando las primeras actividades académicas de estudiantes universitarios y técnicas de aprendizaje automático. (3). https://dilemascontemporaneoseducacionpoliticayvalores.com/index.php/dilemas/article/view/3177/3163

Puga de la Cruz, J., & Torres Monzon, R. (2023). *REDES NEURONALES ARTIFICIALES PARA PRONOSTICAR EL RENDIMIENTO ACADÉMICO DE ALUMNOS DE INGENIERÍA DE SISTEMAS E INFORMÁTICA DE LA UNIVERSIDAD NACIONAL DE LA AMAZONÍA PERUANA.* Repositorio Institucional Digital UNAP.

Quiñones Huatangari , L., & Carrasco Vega, Y. (2020). Rendimiento académico empleando minería de datos. *Espacios*, 1-9. https://doi.org/https://unj.edu.pe/wp-content/uploads/2021/09/a20v41n44p17.pdf

Quiñones Huatangari, L., & Carrasco Vega, Y. (2020). *"Rendimiento académico empleando minería de datos:Academic performance using data mining"*. Perú. https://unj.edu.pe/wp-content/uploads/2021/09/a20v41n44p17.pdf

Raeburn, A. (1 de Julio de 2023). *Asana*. https://asana.com/es/resources/accuracy-vs-precision

Ramírez Veliz, J. F. (2019). *"Estado Del Arte Del Aprendizaje Automático Relacionado Con La Lógica Difusa".* Callao. http://repositorio.unac.edu.pe/bitstream/handle/20.500.12952/5580/Informe%20Final-Ramirez%20Veliz-FIIS-2019.pdf?sequence=1&isAllowed=y

Rico Paez, A. (enero_ Junio de 2022). Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios. *12*(24). https://www.ride.org.mx/index.php/RIDE/article/view/1196/3544

Rico Páez, A. (24 de Enero-junio de 2022). Modelos preditivos progressivos de desempenho acadêmico de estudantes. *12*(24). https://www.ride.org.mx/index.php/RIDE/article/view/1196

Rico Páez, A., & Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *13*. https://www.rediech.org/ojs/2017/index.php/ie_rie_rediech/article/view/1426

Rico Páez, A., & Sánchez Guzmán, D. (15 de agosto de 2017). Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN. *8*(16).

https://www.redalyc.org/journal/4981/498159332010/html/

Rodriguez, L., & Toda, A. (2023). *Minería de datos en el aprendizaje gamificado.* SpringerEnlace. https://doi.org/https://doi.org/10.1007/978-3-031-31949-5_7

Rojas Pari, R. J. (2021). *Modelo de Aprendizaje Automático Supervisado para Identificar Patrones de Bajo Rendimiento Académico en los Ingresantes al Instituto de Educación Superior Pedagógico Público – Juliaca.* Juliaca. https://repositorio.upeu.edu.pe/bitstream/handle/20.500.12840/4505/Rudy_Tesis_Licenciatura_2021.pdf?sequence=5&isAllowed=y

Taya Acosta, E. A. (2021). *"Modelo de mineria de datos para evaluar el efecto del uso del aula virtual sobre el rendimiento academico de los estudiantes de la facultad de ingenieria de la universidad nacional Jorge Basadre Grohmann de Tacna, en Tiempos de pandemia,2020".* Tacna_ Perú. http://repositorio.unjbg.edu.pe/bitstream/handle/UNJBG/4358/97_2021_taya_acosta_ea_espg_doctorado_en_ciencias_de_la_educacion.pdf?sequence=1&isAllowed=y

Ulloa Gallardo, N., Miranda Castillo, R., & Holgado Apaza, L. A. (2020). *Técnicas de minería de datos para detectar patrones de bajo rendimiento académico.* Bogotá. https://www.ilae.edu.co/files/book-pdf/20200907144427211780664 8.pdf

Valdivieso, A., Díaz, C., & Sarmiento, J. (2019). *Mhealth con bigdata y machine learning como soporte tecnológico para la salud en Colombia.* INNOVACIÓN Y DESARROLLO TECNOLÓGICO.

Vargas Saldivar, H. T. (2018). *"Conocimientos Previos De Matemática Básica Y Su Relación Con El Rendimiento Académico De La Asignatura De Calculo I En Estudiantes Ingresantes A La Facultad De Ingeniería De Procesos De La Unsa, 2017".* Arequipa_ Perú. http://repositorio.unsa.edu.pe/bitstream/handle/UNSA/6863/EDDvasaht.pdf?sequence=3&isAllowed=y

Vega García, J. F. (2019). *"Modelo de pronóstico de rendimiento académico de alumnos en los cursos del programa de estudios básicos de la Universidad Ricardo Palma usando algoritmos de Machine Learning".* Lima_ Perú. https://repositorio.urp.edu.pe/bitstream/handle/20.500.14138/2914/DATO_T030_07616656_M%20%20%20VEGA%20GARCIA%20JAVIER%20FERNANDO.pdf?sequence=1&isAllowed=y

Yamao, 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. https://repositorio.usmp.edu.pe/bitstream/handle/20.500.12727/3555/yamao_e.pdf?sequence=3&isAllowed=y

Anexos

Anexo 1: Matriz de consistencia:

<table>
<tr><td colspan="6">Modelo de aprendizaje automático para la predicción del rendimiento académico de alumnos ingresantes en la Facultad de Ingeniería en Industrias Alimentarias de la UNAS</td></tr>
<tr><th>Problema</th><th>Objetivo</th><th>Hipótesis</th><th>Variables</th><th>Dimensiones</th><th>Metodología</th></tr>
<tr><td>PROBLEMA GENERAL

¿Qué modelo de aprendizaje automático permite la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?</td><td>OBJETIVO GENERAL
Determinar el modelo de aprendizaje automático con mejor predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS.</td><td>HIPOTESIS GENERAL
El algoritmo con mejor predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el modelo de aprendizaje automático árbol de decisión.</td><td>Variable independiente:

Modelo de aprendizaje automático</td><td>No tiene</td><td rowspan="2">TIPO: Aplicado

DISEÑO: Correlacional Causal

NIVEL: Predictivo

METODO: Deductivo

POBLACIÓN: N = 204

MUESTRA: n = 155 + 49= 204

MUESTREO: Aleatorio.

TÉCNICA:
Análisis de data
-Información Histórica.
-Análisis documental. (Del Excel se recogió en 3 lugares)
- Observación.

INSTRUMENTO:
-Ficha de Análisis documental.

PROCESAMIENTO DE DATOS:
Software SPSS
versión 26 en español+ wueka (Software de predicción)

4 años – 2015 al 2018 – Análisis 2019 – Para la predicción del modelo.</td></tr>
<tr><td>PROBLEMAS ESPECÍFICOS
PE1.- ¿Qué modelo de aprendizaje automático tienen mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?

PE2.- ¿Qué modelo de aprendizaje automático tienen mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS?</td><td>OBJETIVOS ESPECÍFICOS
OE1.- Determinar el modelo de aprendizaje automático con mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS

OE2.- Determinar el modelo de aprendizaje automático con mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS.</td><td>HIPOTESIS ESPECÍFICOS
HE1: El modelo de aprendizaje automático con mejor exactitud para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.

HE2: El modelo de aprendizaje automático con mejor precisión para la predicción del rendimiento académico de los estudiantes ingresantes de la Facultad de Ingeniería en Industrias Alimentarias de la UNAS es el árbol de decisión.</td><td>Variable dependiente: Predicción del rendimiento académico</td><td>Exactitud

Precisión</td></tr>
</table>

ANEXO 2

Tabla 36

Correlación de las variables con la variable predictora rendimiento académico

VARIABLE	Chi- cuadrado	gl	Sig.
MODALIDAD DE INGRESO	17.617	8	0.024
MODALIDAD DE INGRESO(1)	1.582	1	0.209
MODALIDAD DE INGRESO(2)	1.738	1	0.187
MODALIDAD DE INGRESO(3)	3.169	1	0.075
MODALIDAD DE INGRESO(4)	1.295	1	0.255
MODALIDAD DE INGRESO(5)	0.920	1	0.338
MODALIDAD DE INGRESO(6)	8.025	1	0.005
MODALIDAD DE INGRESO(7)	0.920	1	0.338
MODALIDAD DE INGRESO(8)	0.920	1	0.338
DEPARTAMENTO DE PROCEDENCIA	0.441	1	0.507
TIPO DE COLEGIO(1)	0.000	1	0.993
OPCION DE INGRESO(1)	6.604	1	0.010
SEXO(1)	1.782	1	0.182
NOTA DE INGRESO	7.501	1	0.006
EDAD DE INGRESO	0.163	1	0.687
TIPO DE PREPARACION	6.755	3	0.080
TIPO DE PREPARACION(1)	0.551	1	0.458
TIPO DE PREPARACION(2)	0.124	1	0.725
TIPO DE PREPARACION(3)	4.173	1	0.041
MODO EN LA QUE SE ENTERO DEL ADMISION	4.488	5	0.482
MODO EN LA QUE SE ENTERO DEL ADMISION(1)	0.506	1	0.477
MODO EN LA QUE SE ENTERO DEL ADMISION(2)	1.056	1	0.304
MODO EN LA QUE SE ENTERO DEL ADMISION(3)	1.223	1	0.269
MODO EN LA QUE SE ENTERO DEL ADMISION(4)	1.851	1	0.174
MODO EN LA QUE SE ENTERO DEL ADMISION(5)	0.467	1	0.495
MOTIVO DE POSTULACION A LA UNAS	4.285	6	0.638
MOTIVO DE POSTULACION A LA UNAS(1)	1.396	1	0.237
MOTIVO DE POSTULACION A LA UNAS(2)	0.327	1	0.568

MOTIVO DE POSTULACION A LA UNAS(3)	1.688	1	0.194
MOTIVO DE POSTULACION A LA UNAS(4)	1.558	1	0.212
MOTIVO DE POSTULACION A LA UNAS(5)	0.014	1	0.905
MOTIVO DE POSTULACION A LA UNAS(6)	0.014	1	0.905
TRABAJA	0.255	2	0.880
TRABAJA(1)	0.032	1	0.859
TRABAJA(2)	0.255	1	0.614
DEPENDENCIA ECONOMICA	2.432	3	0.488
DEPENDENCIA ECONOMICA(1)	0.485	1	0.486
DEPENDENCIA ECONOMICA(2)	0.851	1	0.356
DEPENDENCIA ECONOMICA(3)	1.688	1	0.194
VIVEN TUS PADRES	2.730	3	0.435
VIVEN TUS PADRES(1)	0.058	1	0.810
VIVEN TUS PADRES(2)	0.519	1	0.471
VIVEN TUS PADRES(3)	0.036	1	0.849
NUMERO DE HERMANOS	3.386	3	0.336
NUMERO DE HERMANOS(1)	0.819	1	0.366
NUMERO DE HERMANOS(2)	3.142	1	0.076
NUMERO DE HERMANOS(3)	0.414	1	0.520
CON QUIEN VIVE	4.126	4	0.389
CON QUIEN VIVE(1)	0.162	1	0.687
CON QUIEN VIVE(2)	0.004	1	0.949
CON QUIEN VIVE(3)	1.382	1	0.240
CON QUIEN VIVE(4)	1.459	1	0.227
LUGAR DONDE VIVE	1.836	2	0.399
LUGAR DONDE VIVE(1)	0.525	1	0.469
LUGAR DONDE VIVE(2)	0.197	1	0.657
PRIMER SEMESTRE	8.396	1	0.004

ANEXO 3

Tabla 37

Significancia de las variables para el rendimiento académico

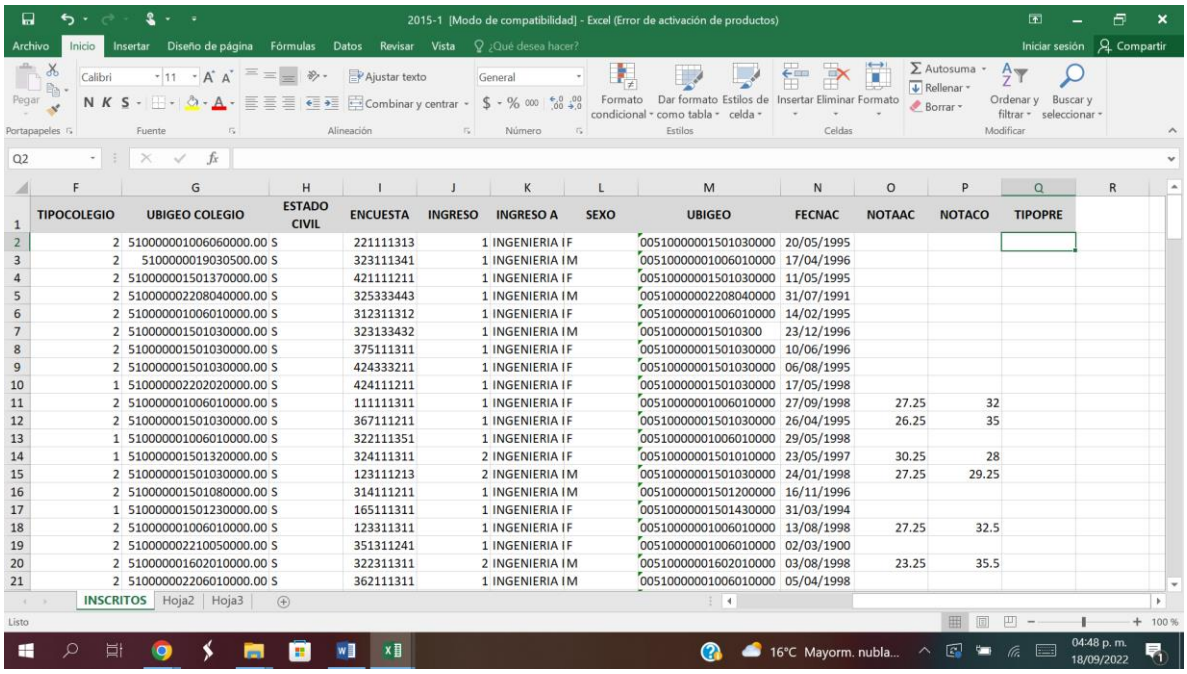
	Error estándar	gl	Sig.	Exp(B)
MODALIDAD DE INGRESO		8	0.309	
MODALIDAD DE INGRESO(1)	40193.034	1	1.000	286463783.679
MODALIDAD DE INGRESO(2)	40193.034	1	1.000	5178267944.540
MODALIDAD DE INGRESO(3)	40193.034	1	1.000	264525404.620
MODALIDAD DE INGRESO(4)	40193.034	1	1.000	532193895.742
MODALIDAD DE INGRESO(5)	56841.489	1	1.000	0.005
MODALIDAD DE INGRESO(6)	42808.181	1	0.999	1465721849915990000.000
MODALIDAD DE INGRESO(7)	56841.489	1	1.000	0.004
MODALIDAD DE INGRESO(8)	56841.489	1	1.000	35.150
DEPARTAMENTO DE PROCEDENCIA	0.058	1	0.832	0.988
TIPO DE COLEGIO(1)	1.131	1	0.214	0.245
OPCION DE INGRESO(1)	1.220	1	0.001	0.015
SEXO(1)	0.728	1	0.044	4.336
NOTA DE INGRESO	0.269	1	0.000	2.582
EDAD DE INGRESO	0.145	1	0.387	0.882
TIPO DE PREPARACION		3	0.368	
TIPO DE PREPARACION(1)	1.437	1	0.178	6.917
TIPO DE PREPARACION(2)	1.870	1	0.663	2.260
TIPO DE PREPARACION(3)	1.333	1	0.127	7.652
MODO EN LA QUE SE ENTERO DEL ADMISION		5	0.715	
MODO EN LA QUE SE ENTERO DEL ADMISION(1)	1.164	1	0.201	4.436
MODO EN LA QUE SE ENTERO DEL ADMISION(2)	1.153	1	0.167	4.918
MODO EN LA QUE SE ENTERO DEL ADMISION(3)	2.254	1	0.227	15.211
MODO EN LA QUE SE ENTERO DEL ADMISION(4)	26327.184	1	0.999	0.000
MODO EN LA QUE SE ENTERO DEL ADMISION(5)	1.219	1	0.107	7.127
MOTIVO DE POSTULACION A LA UNAS		6	0.752	
MOTIVO DE POSTULACION A LA UNAS(1)	1.297	1	0.428	2.799
MOTIVO DE POSTULACION A LA UNAS(2)	1.346	1	0.830	1.336
MOTIVO DE POSTULACION A LA UNAS(3)	1.602	1	0.865	1.313
MOTIVO DE POSTULACION A LA UNAS(4)	1.646	1	0.537	2.764
MOTIVO DE POSTULACION A LA UNAS(5)	1.691	1	0.643	0.457
MOTIVO DE POSTULACION A LA UNAS(6)	1.661	1	0.823	0.689
TRABAJA		2	0.030	

TRABAJA(1)	0.949	1	0.477	0.509
TRABAJA(2)	2.185	1	0.010	0.004
DEPENDENCIA ECONOMICA		3	0.329	
DEPENDENCIA ECONOMICA(1)	24918.862	1	0.999	1143364976.854
DEPENDENCIA ECONOMICA(2)	24918.862	1	0.999	58885281.077
DEPENDENCIA ECONOMICA(3)	24918.862	1	0.999	8212013067.029
VIVEN TUS PADRES		3	0.571	
VIVEN TUS PADRES(1)	37721.693	1	1.000	0.000
VIVEN TUS PADRES(2)	37721.693	1	1.000	0.000
VIVEN TUS PADRES(3)	37721.693	1	1.000	0.000
NUMERO DE HERMANOS		3	0.045	
NUMERO DE HERMANOS(1)	1.563	1	0.394	3.795
NUMERO DE HERMANOS(2)	1.306	1	0.028	17.758
NUMERO DE HERMANOS(3)	1.359	1	0.042	15.764
CON QUIEN VIVE		4	0.396	
CON QUIEN VIVE(1)	1.572	1	0.551	2.556
CON QUIEN VIVE(2)	2.748	1	0.341	13.702
CON QUIEN VIVE(3)	1.922	1	0.821	1.546
CON QUIEN VIVE(4)	2.122	1	0.138	23.297
LUGAR DONDE VIVE		2	0.600	
LUGAR DONDE VIVE(1)	0.949	1	0.968	1.039
LUGAR DONDE VIVE(2)	1.107	1	0.496	0.470
PRIMER SEMESTRE	0.000	1	0.000	1.000
Constante	49458.666	1	0.990	0.000

Nota: La table muestra la significancia de la variables y cuales son las que explican el rendimiento académico de los alumnos ingresantes

Figura 36

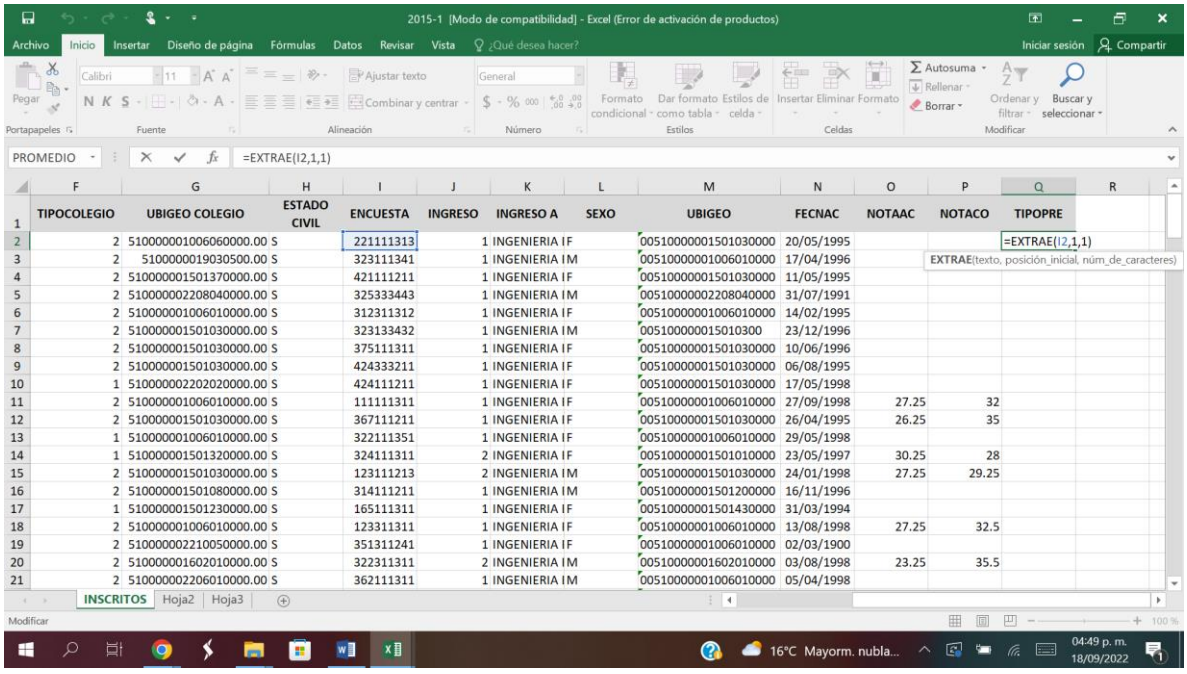
EXTRAER DATOS DE LA ENCUESTA DE EXCEL



Nota: Elaboración propia

Figura 37

Extracción de datos de encuesta Excel



Nota: Formula: =Extrae(texto,posición inicial, num_de_caracteres)

Este proceso de extracción de datos se hace para cada uno de las encuestas.

Anexo 4

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

PROYECTO DE TESIS

MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERÍA EN INDUSTRIAS ALIMENTARIAS DE LA UNAS

Programa de investigación : Sistemas de Información

Línea de investigación : Gestión de datos

TESISTA: Ponce Guizabalo Santos Víctor.

ASESOR(A): Dr. WILLIAM GEORGE PAUCAR PALOMINO

UNIVERSIDAD NACIONAL AGRARIA DE LA SELVA

FACULTAD DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA EN INFORMÁTICA Y SISTEMAS

FICHA DE ANÁLISIS DOCUMENTAL PARA LA TESIS:

"MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL RENDIMIENTO ACADÉMICO DE ALUMNOS INGRESANTES EN LA FACULTAD DE INGENIERÍA EN INDUSTRIAS ALIMENTARIAS DE LA UNAS

Datos de la aplicación:

Fecha de aplicación 1: _____ / _____ / _____

Documentos Revisados de la Universidad:

Documentos	Tiene		Se revisó	
	SI	NO*	SI	NO
Datos Sociales				
Datos Económicos				
Datos Académicos				

Los datos re recolectan a través de una encuesta a los postulantes y luego a los ingresantes, los datos académicos si aprobó o desaprobó el ciclo son registrados al finalizar el ciclo por la Dirección de Coordinación y Desarrollo Académico (DICDA). Todos los datos están en una base de datos.

1. Sobre Datos Sociales - Oficina admisión
   a) Se registra la edad del postulante Si ( ) No ( )
   b) Se registra el sexo del postulante Si ( ) No ( )
      Masculino ( ) Femenino ( )
   c) Se registra si Traja el postulante Si ( ) No ( )
      1) NO 2) Si, Tiempo Completo
      3) Si, por horas
   d) Se registra de quien depende económicamente Si ( ) No ( )
      1) sus padres 2) Parientes
      3) Si mismo 4) Otros
   e) Se registra el lugar donde vive Si ( ) No ( )
      1) Ciudad 2) Pueblo joven
      3) Zona Rural
   f) Se registra el lugar de procedencia (ubigeo) Si ( ) No ( )
      Departamento: _____
      Provincia: _____
      Distrito: _____
2. Sobre los Datos Económicos
   a) Se registra el Tipo de colegio Si ( ) No ( )
      1) Estatal 2) Particular
3. Sobre los Datos Académicos
   a) Se Registra la Facultad que ingresa Si ( ) No ( )
      - INGENIERIA AMBIENTAL ( )
      - ADMINISTRACION ( )
      - Agronomía ( )
      - CONTABILIDAD ( )
      - ECONOMIA ( )
      - INGENIERIA FORESTAL ( )
      - INGENIERIA EN INDUSTRIAS ALIMENTARIAS ( )

- INGENIERIA MECANICA ELECTRICA ( )
- INGENIERIA EN RECURSOS NATURALES RENOVABLES ( )
- INGENIERIA EN INFORMATICA Y SISTEMAS ( )
- INGENIERIA EN CONSERVACION DE SUELOS Y AGUA ( )
- ZOOTECNIA ( )

b) Se registra el tipo de preparación para su postulación Si ( ) No ( )

1) AUTOESTUDIO 2) PROFESOR PARTICULAR

3) ACADEMIA 4) OTROS

c) Se registra la modalidad de ingreso Si ( ) No ( )

EXAMEN ORDINARIO ( )

Primero y segundo puesto de educación secundaria ( )

GRADUADOS Y TITULADOS ( )

CENTRO PRE UNIVERSITARIO ( )

VICTIMAS DE TERRORISMO ( )

PERSONAS CON DISCAPACIDAD ( )

DEPORTISTAS ( )

ARTE Y CULTURA ( )

COMUNIDADES NATIVAS ( )

CONVENIOS ESPECIALES ( )

TALENTO BECA 18 ( )

d) Se registra la Puntaje de Ingreso Si ( ) No ( )

_____

e) Se registra de notal Final del I ciclo académico Si ( ) No ( )

_____

f) Se registra el colegio de Procedencia Si ( ) No ( )

_____