UNIVERSIDAD NACIONAL DE MOQUEGUA

FACULTAD DE INGENIERIA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

"ANÁLISIS DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA UTILIZANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL - CASO: EPISI-UNAM".

TESIS

PRESENTADO POR:

BACH. DINA ELISA LEON MARIN

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

MOQUEGUA - PERÚ

2022

UNAM
UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE
INGENIERÍA DE SISTEMAS E INFORMÁTICA

Folio Nº 007

ACTA DE SUSTENTACIÓN

I. Modalidad:

Tesis : [X]
Trabajo de Suficiencia Profesional : [ ]
Otro: ..... : [ ]

II. Documento de aprobación para la sustentación:
Resolución de Facultad: Nº 160-2022-FACIA-UNAM
Fecha: 12/setiembre/2022 como fecha de sustentación

III. El jurado ha dictaminado la (el) Tesis, TITULADO:
"Análisis del Rendimiento Académico de los estudiantes de Ingeniería utilizando técnicas de Inteligencia Artificial Caso: EPIST-UNAM"

Bachiller: Dina Elisa León Marín
Asesor Principal: Dr. Aníbal Fernando Flores García
Co-Asesor: .....

IV. Se dictamina:
Como: Aprobado por Unanimidad la sustentación de manera:
Presencial [X] Virtual [ ]
Con nota: Trece (13) por lo que; Sí [X] No [ ]
Opta el Título Profesional de INGENIERO DE SISTEMAS E INFORMÁTICA.

Siendo las 15:05 horas, del 12 de setiembre del año 2022 se procede con la firma en conformidad de lo dictaminado por el Jurado.

Ender Tito Chura
Presidente

Msc Victor Yanz Mamani
Primer Miembro

Dr. Salomón R. Ramos Rivera
Segundo Miembro

Bach Dina León Marín
Bachiller

DEDICATORIA

Esta tesis la dedico a mi familia porque siempre me han dado su apoyo incondicional y su paciencia, fueron siempre el soporte para tener éxito en mi vida profesional y a Dios sobre todo porque me ha bendecido grandemente.

AGRADECIMIENTO

A mi familia porque siempre han estado acompañándome en esta etapa de mi vida, me brindaron su apoyo y comprensión.

Un agradecimiento especial a mi hijo por ser el motivo de seguir preparándome profesionalmente y darle un ejemplo de que nunca es tarde para seguir estudiando que con perseverancia se consigue el éxito.

Gracias a mi alma master UNAM por darme la oportunidad de seguir con una carrera profesional y a todos los docentes que contribuyeron con sus conocimientos.

RESUMEN

El presente trabajo de investigación tiene como objetivo determinar causantes de bajo rendimiento académico de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM, se utiliza el método "Investigación - Acción" porque se pretende validar a través del caso de estudio. Los instrumentos utilizados para recolectar la información fue la encuesta virtual para el llenado de fichas socioeconómicas y la base de datos académicos de los alumnos donde vamos a seleccionar, preprocesar y conocer las características de los datos. El árbol de decisión es un modelo eficiente de Machine Learning que ayuda mucho a resolver múltiples tipos de problemas, es un modelo clásico de decisiones SI o NO, donde vamos a descubrir patrones y evaluar los resultados, los algoritmos utilizados son: Árbol de Decisión, Random Forest y Naive Bayes. El desarrollo del proyecto se basó en la metodología CRISP-DM.

Palabras clave: inteligencia artificial, algoritmos, clasificadores, rendimiento académico

ABSTRACT

The objective of this research work is to determine the causes of low academic performance of the students of the Professional School of Systems and Informatics Engineering of the UNAM, the "Research - Action" method is used because it is intended to validate through the case study. The instruments used to collect the information were the virtual survey for filling in the socioeconomic files and the academic database of the students where ve are going to select, preprocess and know the characteristics of the data. The decision tree is an efficient model of Machine Learning that helps a lot to solve multiple types of problems, it is a classic model of YES or NO decisions, where we are going to discover patterns and evaluate the results, the algorithms used are: Decisión Tree, Random Forest and Naive Bayes. The development of the project was based on the CRISP-DM methodology.

Keywords: artificial intelligence, algorithms, classifiers, academic performance

CONTENIDO

DEDICATORIA I
AGRADECIMIENTO II
RESUMEN III
ABSTRACT IV
CONTENIDO V
INTRODUCCION 1
CAPITULO I: PLANEAMIENTO DEL PROBLEMA 3
1.1 Descripción de la realidad problemática 3
1.2 Formulación del problema 5
1.2.1 Interrogante General 5
1.2.2 Interrogantes Específicos 5
1.3 Justificación e importancia de la investigación 5
1.4 Objetivos 6
1.4.1 Objetivo general 6
1.4.2 Objetivos específicos 6
1.5 Hipótesis 6
1.5.1 Hipótesis general 6
1.5.2 Hipótesis especificas 7
CAPITULO II: MARCO TEORICO 7
2.1 Antecedentes de estudio 7
2.2 BASES TEORICAS 9
2.2.1 Inteligencia artificial 9
2.2.2 Rendimiento Académico 20
2.2.2.1 Alto rendimiento académico 22
2.2.2.2 Rendimiento académico Medio 22
2.2.2.3 Bajo rendimiento académico 22
2.2.2.4 Factores que inciden en el rendimiento académico 24
2.2.3 Técnicas de inteligencia artificial 25
2.2.4 Árbol de decisión 26
2.2.5 Random Forest 32
2.2.6 Clasificador Naive Bayes 36
2.2.7 Metodología CRISP-DM 37
2.2.8 Software Weka (*Waikato Environment For Knowledge Analysis*) 41
2.2.9 PYTHON 42
2.3 DEFINICIÓN DE TÉRMINOS 43

CAPITULO III: MARCO METODOLÓGICO 45
3.1 Lugar de ejecución 45
3.2 Tipo y Diseño 45
3.2.1 Tipo de Investigación 45
3.2.2 Diseño de la Ejecución 46
3.3 Nivel de Investigación 46
3.4 Operacionalización de variables 47
3.4.1 Variables Independientes 47
3.4.2 Variables Dependientes 47
3.5 Población y muestra 48
3.5.1 Población: 48
3.5.2 Muestra: 48
3.6 Delimitación Espacial y Temporal 48
3.7 Método de Investigación 48
3.8 Técnicas, Instrumentos para recolección de datos 49
3.8.1 Encuesta 49
3.8.2 Procesamiento de datos 49
3.9 Validación y Confiabilidad de los Instrumentos 50
CAPITULO IV: PRESENTACION DE RESULTADOS 51
4.1 Comprensión del Negocio 51
4.1.3 Visión 52
4.1.4 Misión 53
4.1.5 Comprensión del problema 53
4.2 Comprensión de Data 54
4.2.1 Descripción de los Datos Generales (Fuente de Datos) 54
4.2.2 Datos seleccionados 57
4.3 Preparación de la DATA 58
4.4 Modelado 61
4.4.2 Algoritmo C4.5 63
4.4.3 Algoritmo RandomForest 67
4.4.4 Algoritmo Naive Bayes 70
4.4.6 Evaluación de precisión de las técnicas usadas en la Investigación 75
4.5 Despliegue (Implementación) 76
VI. CONCLUSIONES 76
VII. RECOMENDACIONES 78
VIII. REFERENCIAS 79
ANEXOS 83

ÍNDICE DE TABLAS

Tabla 01: Operacionalización de los factores ……... 47
Tabla 02: Total, de estudiantes del caso de estudio ..……. 48
Tabla 03: Cantidad de alumnos, según el ciclo académico ……. 50
Tabla 04: Diccionario de Datos.....……. 54
Tabla 05: Resultado de las técnicas usadas …….....……..... 76
Tabla 06: Resultados de las características más influyentes de los algoritmos..... 76
Tabla 07: Resultado de las validaciones de las técnicas usadas..... 77

ÍNDICE DE FIGURAS

Figura 01 Machine Learning Aprendizaje supervisado..... 12
Figura 02 Machine learning aprendizaje no supervisado..... 14
Figura 03 Machine learning aprendizaje por refuerzo 15
Figura 04 Idea general de deep learning …..... 17
Figura 05 Diagrama de redes neuronales 18
Figura 06 Símbolos utilizados en el diagrama de árboles de decisiones 27
Figura 07 Funcionamiento de random forest 33
Figura 08 Criterios de gini de random forest 35
Figura 09 Criterios de entropía de random forest 36
Figura 10 Secuencia del proceso CRISP-DM 38
Figura 11 Modelo de clasificación..... 41
Figura 12 Librerías de python 42
Figura 13 Datos no procesados académicos del caso de estudio año 2019, 2020 55
Figura 14 Datos no procesados ficha socioeconómicos año de estudio 2019, 2020 56
Figura 15 Descripción del dataset..... 58
Figura 16 Información de la data 59
Figura 17 Información del feature clases..... 59
Figura 18 Librería scikit-learn 59
Figura 19 Data procesada 60
Figura 20 Visualización de los valores de cada feature..... 61
Figura 21 Base de datos caso Episi – UNAM..... 61
Figura 22 División de los datos training y testing..... 62
Figura 23 Entrenamiento del árbol..... 63
Figura 24 Estructura del árbol creado..... 64
Figura 25 Árbol de decisión 64
Figura 26 Predicción de los datos de testing 64
Figura 27 Matriz de confusión del algoritmo árbol de decisión..... 65
Figura 28 Métricas del algoritmo árbol de decisión..... 66
Figura 29 Grupo de variables del algoritmo árbol de decisión..... 66
Figura 30 Creación del modelo bosques aleatorios..... 67

Figura 31 Configuración de la matriz de confusión de randomforest..... 68
Figura 32 Matriz de confusión del algoritmo randomforest..... 68
Figura 33 Validación del algoritmo de randomforest..... 69
Figura 34 Características más influyentes random forest..... 69
Figura 35 Importando librerías y módulo de naive bayes 70
Figura 36 Creación de naive bayes import gaussianNB 70
Figura 37 Categorías de codificación, utilizando labelEncoder..... 71
Figura 38 Métricas de desempeño, utilizando gaussianNB..... 71
Figura 39 Configuración de la matriz de configuración naive bayes 72
Figura 40 Métricas del algoritmo naive bayes..... 72
Figura 41 Precision del algoritmo naive bayes..... 73
Figura 42 Características más influyentes de naive bayes..... 73
Figura 43 Diferencia entre datos de entrenamiento y prueba 74
Figura 44 Características más importantes 74

INTRODUCCION

Con esta investigación se pretende analizar el rendimiento académico de los estudiantes de Ingeniería, con la finalidad de desarrollar métodos de inteligencia artificial para descubrir patrones de bajo rendimiento académicos de los estudiantes. Caso: Escuela profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua, a la fecha no contamos con ninguna herramienta o mecanismo que permita hacer seguimiento para la nivelación de las deficiencias de los estudiantes en los diversos cursos y mantener la permanencia de estudiantes en la Universidad.

Por consiguiente, proponemos la implementación de un modelo de análisis de datos basados en inteligencia artificial que permitirán determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial.

Algunas de las variables que se consideró para la implementación del modelo son factores socio-económicos y académicos, que inciden en el rendimiento académico las cuales corresponden a datos socioeconómicos del estudiante (lugar de procedencia, dependencia económica, nivel de ingreso, etc.), datos académicos (año de ingreso, número de cursos aprobados, número de cursos desaprobados, promedio final, ciclo repetido, etc.). Para tener éxito en este análisis tengo que considerar todos los factores ya mencionados y con el apoyo de la Universidad en brindar la base de datos de los estudiantes.

El objetivo principal es implementar un modelo de análisis académicos utilizando técnicas de inteligencia artificial (clasificadores) que incluye árbol de decisión. Con este estudio es pretender mejorar y sobretodo adecuar acciones y decisiones estratégicas que maximicen su rendimiento académico, si bien no es una tarea fácil para los educandos; pero, es un buen inicio si queremos lograr la acreditación de las carreras profesionales que consiste en cumplir 34 estándares enmarcados dentro de 4 dimensiones: Gestión Estratégica, Formación Integral, Soporte Institucional y Resultados.

A continuación, detallamos la estructura de la presente investigación:

En el capítulo I: Planteamiento del problema, se desarrolla la descripción y formulación del problema, antecedentes, objetivos, justificación e importancia y finalmente en este capítulo la hipótesis.

En el capítulo II: Marco Teórico, se incluye toda la recopilación de antecedentes, investigaciones previas debidamente sustentadas y el marco conceptual.

En el capítulo III: Marco metodológico, se describe el tipo de y nivel de investigación, el ámbito temporal y espacial, la población y muestra, los instrumentos utilizados y técnicas, procedimiento y el análisis de datos.

En el capítulo IV: Presentación de resultados de las técnicas usadas de Inteligencia Artificial y la comprobación de hipótesis y Discusión de resultados.

Finalmente se presentan las conclusiones, recomendaciones, referencias y anexos.

CAPITULO I: PLANEAMIENTO DEL PROBLEMA

1.1 Descripción de la realidad problemática

En los últimos años la problemática del rendimiento académico en los estudiantes se ha ido incrementando por diversos motivos, tanto en lo presencial y virtual. Por pandemia muchos estudiantes tuvieron que aprender nueva metodología de estudio.

Existen diversos factores que intervienen en el aspecto individual del estudiante, tanto positivo como negativamente. Hay varios estudios con la finalidad de reducir la deserción en las universidades, como en los alumnos de ingeniería.

Sin embargo, debemos entender porque hay estudiante con alto y bajo rendimiento académico, que característica influye mucho. En este análisis se utilizó la base de datos de la ficha socioeconómica y sus notas académicas de los alumnos de I y II ciclo de los estudiantes de la carrera de Ingeniería de Sistemas e Informática, matriculados en los años 2019 y 2020.

Debemos considerar que el rendimiento académico hace referencia a las evaluaciones del conocimiento adquiridos durante su ciclo académico, si el estudiante obtuvo calificaciones positivas, en todos sus cursos muy aparte de tener un alto rendimiento académico puedo lograr pertenecer al tercio superior o quinto superior de todo su ciclo.

En el caso de que el estudiante obtenga notas negativas, en algunos de sus cursos o todos, es considerado bajo rendimiento. Eso quiere decir que el rendimiento académico está vinculado a la aptitud.

En rendimiento académico también influye el factor socioeconómico de los estudiantes, el lugar de nacimiento, si estudia o trabaja, si posee una computadora, laptop, celular, internet, entre otros.

El bajo rendimiento académico en los estudiantes puede provocar abandono y repetición de cursos, si el alumno no tiene apoyo de los padres y tutores de la universidad se sienten abandonados, en la cual sus decisiones muchas veces tienden a dejar la carrera de lado. Otros factores pueden ser biofísicos, psicológicos, o puede ver estudiantes que tienen lentitud en el aprendizaje, y puede ver muchos factores que influyan y por eso hay estudiantes que, de una carrera de 5 años, lo hacen en 6 a 10 años. Si bien es difícil comprender por qué hay pocos estudiantes que entienden que una educación superior es para aprender y no para aprobar los cursos por aprobar.

Muchos estudiantes no ponen en contexto que los estudios superiores nos ayudan a adquirir conocimiento para servir en nuestra sociedad, ya sea en las distintas carreras u oficios. La evaluación del rendimiento académico a los estudiantes permite que la universidad identifique que áreas académicas necesita mayor apoyo.

Por lo expuesto, es importante llevar a cabo la INVESTIGACIÓN “ANÁLISIS DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA UTILIZANDO TÉCNICAS DE INTELIGENCIA ARTIFICIAL” – Caso: EPISI-UNAM

1.2 Formulación del problema

1.2.1 Interrogante General

¿Es posible determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM?

1.2.2 Interrogantes Específicos

IE1. ¿De qué tipo son los datos que tiene la UNAM sobre los estudiantes?

IE2 ¿Será posible descubrir patrones de bajo rendimiento con métodos de inteligencia artificial?

IE3 ¿Cuál será el nivel de precision de los métodos de inteligencia artificial?

1.3 Justificación e importancia de la investigación

Realizando la investigación sobre el tema, "Análisis del rendimiento académico de los estudiantes de ingeniería utilizando técnicas de inteligencia artificial", caso: Escuela Profesional de Ingeniería de Sistemas e Informática – UNAM, no se encontró estudios previos y fehacientes, la finalidad es desarrollar métodos de inteligencia artificial para descubrir patrones de bajo rendimiento académicos de los estudiantes, evaluar los resultados, dar a conocer cuál es el algoritmo más preciso.

Desarrollando este trabajo de Investigación damos a conocer las variables más influyentes que podrían presentarse durante el proceso formativo; por lo que, el estudiante debe perfeccionar y revalorar los conocimientos de acuerdo a la competencia que desea especializarse y enfrentarse al mundo competitivo. es necesario trabajar en las deficiencias de los estudiantes en los diversos cursos es que se hace necesaria la implementación de una solución.

La presente tesis se justifica socialmente, porque inicialmente beneficiará a los estudiantes universitarios de la Escuela Profesional de Ingeniería de Sistemas e Informática y en lo posterior a estudiantes de la universidad luego a estudiantes de otras Universidades del país y de la educación básica regular.

Técnicamente, se cuenta con los recursos humanos y conocimiento necesario para llevar a cabo cada una de las etapas del presente proyecto.

1.4 Objetivos

1.4.1 Objetivo general

Determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM

1.4.2 Objetivos específicos

OE1 Seleccionar, preprocesar y conocer las características de los datos socioeconómicos y académicos de los estudiantes de la Universidad Nacional de Moquegua.

OE2 Desarrollar métodos de inteligencia artificial para descubrir patrones de bajo rendimiento académicos de los estudiantes. Caso EPISI UNAM

OE3 Evaluar los resultados de los métodos de inteligencia artificial.

1.5 Hipótesis

1.5.1 Hipótesis general

Será factible determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM.

1.5.2 Hipótesis especificas

HE1. Se podrá seleccionar, preprocesar y conocer las características de los datos socioeconómicos y académicos de los estudiantes de la Universidad Nacional de Moquegua UNAM.

HE2. Se podrá descubrir patrones de bajo rendimiento con métodos de inteligencia artificial.

HE3. Se podrá evaluar los resultados de los métodos de inteligencia artificial

CAPITULO II: MARCO TEORICO

2.1 Antecedentes de estudio

En el estado del arte presento algunas investigaciones nacionales e internacionales que contribuyeron a mi investigación sobre el rendimiento académico de los estudiantes y el uso de distintas técnicas de Inteligencia artificial.

El estudio realizado por **J. Morales-Aquise et. al. (2019)** presentaron el paper "*Analysis of the Academic Performance of the Students of Mechanical and Electrical Engineering of the UNTELS*". Donde es motivo de estudio sobre el fracaso en el nivel superior y deserción estudiantil y que factores se involucran en el aprendizaje del estudiante en los diversos cursos, el autor de esta investigación es (Morales, 2019) .

Para el análisis del problema utilizaron el software IBM SPSS Statics (Producto de Estadísticas y Solución de Servicio) junto con el BMDP estos son utilizados para la investigación aplicada y comprensión de datos, analizar tendencias y planificar nuevos supuestos y conclusiones. Se muestran los resultados de la aplicación de técnicas informáticas que permitieron detectar los cursos que más desaprobaban los estudiantes de Ingeniería con un promedio del 45% y un máximo del 75% registrado en el período 2014-II y el mínimo del 16% en el período 2015-II, obteniendo una variación máxima del 59%. (pág. 3)

En la Investigación de **S. Merchan Rubiano et. al. (2015)** en el paper "Formulation of a predictive model for academic performance based on student's academic and demographic data" (S. Merchan Rubiano, 2015), presentan un modelo predictivo para rendimiento académico que consiste en el análisis, selección y definición de reglas de clasificación que predicen apropiadamente el rendimiento de estudiantes de Ingeniería de Sistemas en la Universidad El Bosque de Bogotá, Colombia.

En esta investigación ejecutaron procesos de minería de datos utilizando el J48 y PART, (pág. 2) son algoritmos que se utiliza con Weka el J48 pertenece a los grupos de árboles de decision y el PART obtiene reglas a partir del árbol de decisión. Este modelo tiene la capacidad de predecir rendimiento académico deficiente o pobre y en sus resultados resalta mucho la deserción estudiantil en la carrera de sistemas y sobre todo por los cursos de biología y física según el coordinador de la carrera van a empezar a pensar en estratégicas de prevención y en el plan de estudios van a restructurar por el puntaje del curso de biología es relevante en el proceso de admisión en especial a los alumnos que asisten en clases nocturnas. (pág. 6)

Considerando este estudio de S. Jafari Petrudi et. al (2015), en el paper "Application of fuzzy logic for performance evaluation of academic students" (Jafari,S. , Pirouz,M. , Pirouz, B., 2013), presentan una propuesta de un nuevo método para evaluar el rendimiento académico basado en sistemas de lógica difusa, el cual consiste en considerar tres características para un curso académico: examen 1, examen teórico 2, y examen práctico 3. (pág. 1) Una vez que los alumnos dan sus tres examenes el resultado se ingresa como variable de entrada de la lógica difusa, se utilizó el MATLAB en la evaluación del desempeño, este estudio demuestra un nuevo enfoque para evaluar estudiantes y con este método desean mejorar sus resultados en el desempeño de la universidad. (pág. 6)

El método que se utilizó en esta Investigación. Ramadhani et. al, en el paper "Finding similar clustering pattern between students academic performance and non curricular activities data" (Ramadhani,N. , Pujianto,U. , 2017) presentan un análisis de clústeres en variables relacionadas a actividades no curriculares y rendimiento académico del estudiante. Se utilizó el algoritmo k-means para segmentar el rendimiento académico en dos clústeres. Después de la experimentación se detectó que en más del 50% de estudiantes las actividades extra curriculares influían en el rendimiento académico.

2.2 BASES TEORICAS

2.2.1 Inteligencia artificial

El primer intento de definir la Inteligencia Artificial lo hizo el matemático Alan Turing, que es considerado como el padre de la computación. Este científico inglés es más conocido por su máquina Turing: una máquina conceptual que utilizó para formalizar los conceptos del modelo computacional que seguimos utilizando hoy en día. En concreto demostró que con las operaciones básicas que podía desarrollar su máquina, podía codificar cualquier algoritmo y que toda máquina capaz de computar tendría las mismas operaciones básicas que su máquina o un super conjunto de estas. (Garcia S., 2012)

El test de Turing, pese a los años que han pasado tiene una gran importancia, ya que exige una serie de capacidades a la máquina inteligente cuyo conjunto conforma a grandes rasgos, lo que es la inteligencia artificial hoy en día. En efecto, una máquina que sea capaz de pasar el test de Turing ha de tener las siguientes capacidades:

- Reconocimiento del lenguaje natural
- Razonamiento
- Aprendizaje

- Representación del conocimiento

La inteligencia artificial que (Thomas, 2001) se ha dado por objetivo el estudio y el análisis del comportamiento humano. De esta manera, las aplicaciones de la IA se sitúan principalmente en la simulación de actividades intelectuales del hombre.

La inteligencia artificial no se limita por lo tanto al sistema de datos de tipo numérico, como el de los computadores actuales. Su campo es mucho más extenso y se puede ilustrar a través de tres importantes puntos de vista:

- Aquellos que sostienen que es posible realizar "dispositivos realmente pensantes", punto de vista llamado la IA fuerte.
- Otros que piensan que es posible simular estados mentales - sin ser estados mentales - de nuestro cerebro por medio de computadores, punto de vista llamado IA débil.
- Finalmente, los "dualistas" quienes dan por separada, en resumidas cuentas - la dimensión del cuerpo y del espíritu; y que, de esa manera existirán "juicios de verdad" a los cuales los computadores no tendrán nunca acceso.

La inteligencia artificial se ha transformado en una ciencia importante, sino fascinante, puesto que los temas que abarca atraen a una amplia variedad de disciplinas. Muchos matemáticos, físicos, biólogos y filósofos esencialmente, convergen hacia el problema del cerebro y la conciencia. (pág. 4).

2.2.1.1 Machine Learning

Nació como una idea ambiciosa en la década de los 60, para ser más exactos es una subdisciplina de la Inteligencia Artificial, el principal objetivo del machine learning es abordar y resolver problemas prácticos en donde a través de algoritmos las máquinas usan estadísticas para

encontrar patrones en datos y poder predecir nuevos valores y comportamientos. Según (Rudolph, 2018) “es la práctica de programación de computadoras para aprender de los datos son conocidos como conjuntos de capacitación o ejemplos”. (pág. 9)

El término de Inteligencia Artificial nació en 1956, según (Camus, 2020), “En 1952, Arthur Samuel creó una de las primeras máquinas de aprendizaje temprano, con la capacidad de jugar a las damas. Se usó guías anotadas por jugadores expertos y jugó contra sí mismo para aprender a distinguir los movimientos correctos de los malos”. Luego de 4 años de investigación, nació el nombre AI.

- **¿Cuándo debemos usar Machine Learning?**

  Según (Rudolph, 2018) debemos usar machine learning:

  - Cuando tengas un problema que requiera de largas listas de reglas para encontrar la solución. En este caso, las técnicas de Machine Learning pueden simplificar tu código y mejorar el desempeño.
  - Problemas muy complejos para los que no haya solución con un enfoque tradicional.
  - Ambientes no estables: softwares con Machine Learning se pueden adaptar a nuevos datos. (pág. 12)

  Machine Learning se divide en 3 categorías de acuerdo al tipo y a la cantidad de supervisión humana durante su clasificación.

  - Machine Learning Supervisado
  - Machine Learning No Supervisado
  - Machine Learning de Refuerzo

Vamos a desglosar cada uno de las categorías de Machine Learning.

- **Machine Learning Supervisado**

  En Machine learning supervisado o aprendizaje supervisado, los datos son referidos como "labels" (etiquetas). El algoritmo entrena los datos dando las variables de entrada (input data), el algoritmo se entrena con un "histórico" de datos y así "aprende" a asignar la etiqueta de salida adecuada a un nuevo valor; es decir, predice el valor de salida.

  Según (Rudolph, 2018) estos dos tipos de aprendizaje supervisado, se distinguen por el tipo de variable objetivo y se suele usar en:

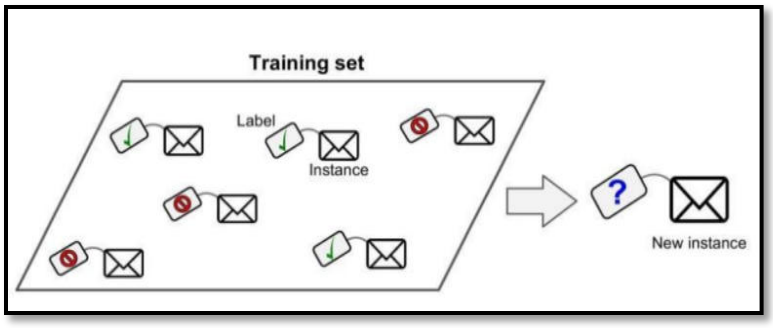
  - Problemas de Clasificación (tipo categórico)
  - Problemas de regresión (numérico)

  Los Algoritmos Supervisados Mas Importantes:

  - K-nearest neighbors (KNN, vecinos más cercanos K)
  - Máquinas de soporte de vectores
  - Regresión logística
  - Clasificación de Naïve Bayes.
  - Árboles de decisiones y bosques aleatorios

Figura 01

Machine learning aprendizaje supervisado



Nota. Se agrupa las tareas de clasificación, en este ejemplo ha sido capacitado con muchos correos electrónicos al mismo tiempo que sus clases. (Rudolph, 2018)

- **Machine Learning No Supervisado**

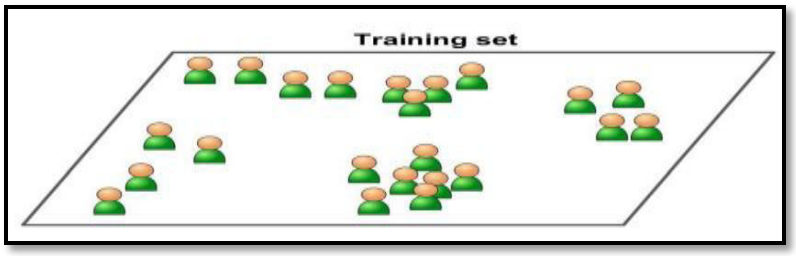
Machine Learning No Supervisado y/o aprendizaje no supervisado tiene lugar cuando no se dispone de datos "etiquetados" para el entrenamiento las tareas de clustering buscan agrupamientos basados en similitudes, pero nada garantiza que éstas tengan algún significado o utilidad.

Los más importantes algoritmos del Machine Learning No Supervisado (Rudolph, 2018)

- Agrupamiento (Clustering): Medios k, análisis de agrupamiento
- Jerárquico.
- Machine Learning de Asociación de Regla: Eclat y A priori
- Visualización y Reducción de Dimensionalidad: Núcleo PCA, distribuido de TPC

Figura 02

Machine learning aprendizaje no supervisado



Nota. Datos sin etiquetas, donde el algoritmo tendrá que agrupar y dividir en grupos o subgrupos, como una salida. (Rudolph, 2018)

- **Machine Learning de Refuerzo**

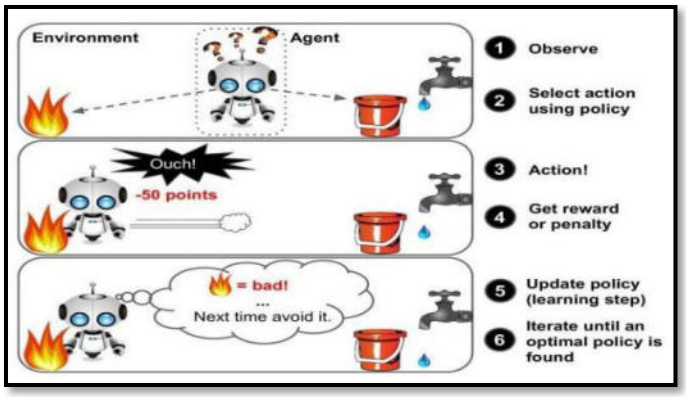
No todos los algoritmos ML se pueden clasificar como algoritmos de aprendizaje supervisado o no supervisado. Hay una «tierra de nadie» que es donde encajan las técnicas de aprendizaje por refuerzo. (Paloma, 2017).

En aprendizaje por refuerzo (Antonio, 2018) (Reinforcement Learning en inglés) no tenemos una "etiqueta de salida", por lo que no es de tipo supervisado y si bien estos algoritmos aprenden por sí mismos, tampoco son de tipo no supervisado, en donde se intenta clasificar grupos teniendo en cuenta alguna distancia entre muestras. (pág. 2)

Este tipo aprendizaje (Antonio, 2018), se basa en mejorar la respuesta del modelo usando un proceso de retroalimentación. El algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error.

Figura 03

Machine learning aprendizaje por refuerzo



Nota. Un agente "Sistema AI" observará el ambiente, ejecutando diferentes acciones, y luego recibe un ejemplo de aprendizaje por refuerzo. (Rudolph, 2018)

- **Pasos para construir un modelo de machine learning**

Se realiza todo un proceso que involucra varios pasos, para construir un modelo utilizando un algoritmo de aprendizaje o varios.

a) **Recolectar los datos.** Se puede utilizar diversas fuentes para extraer los datos.

b) **Preprocesar los datos.** Una vez obtenido los datos preprocesar sin antes asegurarnos que tengan el formato adecuado para nuestro algoritmo de aprendizaje.

c) **Explorar los datos.** Una vez que los datos tengan el formato correcto, realizar un pre análisis y corregir si es necesario datos faltantes.

d) **Entrenar el algoritmo.** Es una fase muy importante; ya que, empezamos a utilizar las técnicas de machine learning. Se tiene que nutrir el o los algoritmos que se van a utilizar.

e) **Evaluar el algoritmo**. Poner a prueba la información que el

algoritmo obtuvo del entrenamiento del paso anterior.

f) **Utilizar el modelo.** Una vez realizado los pasos anteriores nuestro modelo debe enfrentarse al problema real y medir su rendimiento.

- **Librerías de Python machine learning**

  El lenguaje de programación Python nos contribuye con una gran variedad de librerías de primer nivel que extienden las funcionalidades del lenguaje.

  Scikit-Learn: nos facilita las tareas de evaluación, diagnóstico y validaciones cruzadas.

  Statsmodels: nos facilita para hacer análisis estadístico, modelo predictivo y abarca tareas de validación.

  PyMC: es un módulo de Python, que nos ayudará a implementar los modelos bayesianos

  NTLK: es la librería líder para el procesamiento del lenguaje natural.

2.2.1.2 Deep Learning

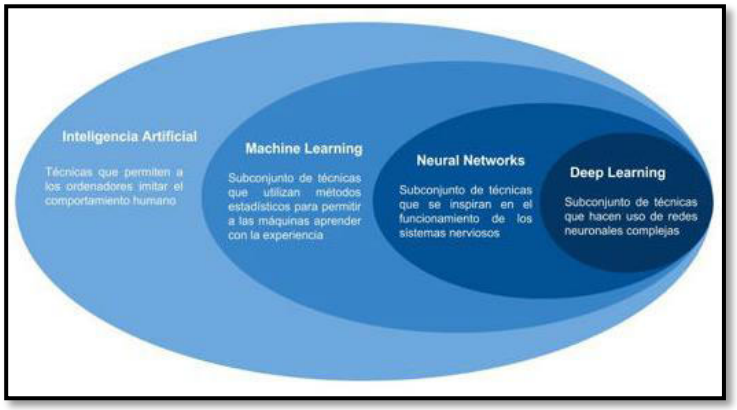
Es un subconjunto de técnicas que nacieron a inicios del siglo XXI, llamada aprendizaje profundo, comúnmente conocido como deep learning (DL). Según (Adrian, 2019), indica que "La popularidad del machine learning y el desarrollo de la capacidad de computación de los ordenadores permitieron llegar a esta nueva tecnología". Si bien machine learning y deep learning tiene conceptos similares, pero utilizan distintos algoritmos.

Deep learning usa redes neuronales que funcionan de forma muy parecida a las conexiones neuronales biológicas de nuestro cerebro,

forma parte del aprendizaje automático. Deep learning cuenta con redes capaces de aprender a realizar una tarea sin supervisión a partir de datos no estructurados o etiquetados, también recibe el nombre de aprendizaje neuronal profundo o red neuronal profunda.

Figura 04

Idea general de deep learning



Nota. Deep Learning está bastante relacionado con las redes neuronales (Adrian, 2019)

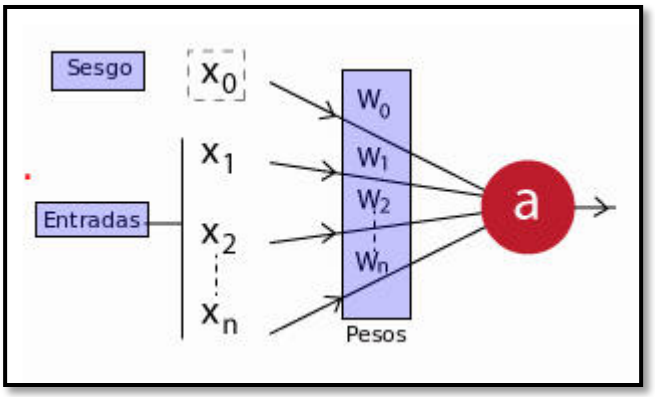
Los algoritmos de deep learning; según (IBM, s.f.), realizan una tarea repetitiva que ayuda a mejorar de manera gradual el resultado a través de "deep layers" lo que permite el aprendizaje progresivo. Este proceso forma parte de una familia más amplia de métodos de machine learning basados en redes neuronales.

- **Redes Neuronales**.

  Son un modelo computacional basado en un gran conjunto de unidades neuronales simples (neuronas artificiales), según (Raul, 2017) Cada una de estas neuronas simples, va a tener una forma similar al siguiente diagrama:

Figura 05

Diagrama de redes neuronales



Nota. Deep Learning: las redes neuronales donde se muestran los componentes (Raul, 2017)

En donde sus componentes son: (Raul, 2017)

- **x1,x2,…,xn**: Los datos de entrada en la neurona, puede ser que sean producto de la salida de otra neurona de la red.
- **x0**: La unidad de sesgo; un valor constante que se le suma a la entrada de la función de activación de la neurona.
- **w0,w1,w2,…,wn:** Los pesos relativos de cada entrada.
- **a:** La salida de la neurona. Que va a ser calculada de la siguiente forma:

$$a = f\left(\sum_{i=0}^{n} w_i \ . \ x_i\right)$$

Aquí f es la función de activación de la neurona. Esta función es la que les otorga tanta flexibilidad a las redes neuronales y le permite estimar complejas relaciones no lineales en los datos (pág. 2)

2.2.1.3 Data Science y Machine Learning

Data science siglas en inglés (*Ciencia de datos*), es dedicada al estudio y análisis de datos económicos, financieros y sociales estructurados y no estructurados. (Silveira, 2016) Machine Learning a menudo se incorpora al Data Science. Según, (Reguero, 2015). El Data Science lo definió William S. Cleveland en 2001 en su artículo: Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. En él presenta un plan de acción para ampliar las áreas de trabajo de la estadística definiendo un nuevo campo el del Data Science. según (Herbert, 2019), "se encarga de estudiar de dónde viene la información, qué representa y cómo se puede convertir en un recurso valioso en la creación de negocios y estrategias".

Data science, emplea las disciplinas de las matemáticas, estadística y la ciencia de la computación (UA, 2020). Según (Reguero, 2015), El Data Science lo definió William S. Cleveland en 2001 en su artículo: Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. En él presenta un plan de acción para ampliar las áreas de trabajo de la estadística.

- La metodología de trabajo es la siguiente:
  - ❖ Extrae los datos.
  - ❖ Genera una limpieza.
  - ❖ Pone en marcha un procesamiento.
  - ❖ Hace un rediseño de la información.
- Herramientas de Data Science

  algunas herramientas de Data Science que puedes ir conociendo y que serán de enorme utilidad en tus futuros proyectos dentro de una

empresa.

Scikit-learn: Hablamos de un módulo de Python, que basado en el Machine Learning, que implica un método de aprendizaje automático. Al mismo tiempo, se incluyen distintos paquetes de software, tales como Pandas, SymPy, ScyPy, entre otros.

Tensorflow: Genera patrones y razonamientos, tal y como lo haría un ser humano. emula cuáles son las estructuras lógicas subyacentes a la interpretación de la información.

R: Ideal para poder visualizar, analizar y comprender los datos, calculando y analizando estadísticamente.

2.2.2 Rendimiento Académico

Hace referencia a la evaluación del conocimiento adquirido en el ámbito escolar, terciario o universitario. Un estudiante con buen rendimiento académico es aquel que obtiene calificaciones positivas en los exámenes que debe rendir a lo largo de una cursada. En otras palabras, el rendimiento académico es una medida de las capacidades del alumno, que expresa lo que éste ha aprendido a lo largo del proceso formativo.

También supone la capacidad del alumno para responder a los estímulos educativos; en este sentido, el rendimiento académico está vinculado a la aptitud. (Ramadhani, N. , Pujianto, U. , 2017)

En otras palabras, se refiere al resultado cuantitativo que se obtiene en el proceso de aprendizaje de conocimientos, conforme a las evaluaciones que realiza el docente mediante pruebas objetivas y otras actividades complementarias.

Por ser cuantificable, el Rendimiento académico determina el nivel de conocimientos alcanzado, y es tomado como único criterio para medir el éxito o

fracaso universitario a través de sistemas de calificaciones de 0 a 20 en la mayoría de los centros educativos públicos y privados.

De acuerdo con lo anterior se derivan tres tipos de estudiantes, según (Covington, 1979):

- "Los orientados al dominio. Sujetos que tienen éxito escolar, se consideran capaces, presentan alta motivación de logro y muestran confianza en sí mismos.
- Los que aceptan el fracaso. Sujetos derrotistas que presentan una imagen propia deteriorada y manifiestan un sentimiento de desesperanza aprendido, es decir que han aprendido que el control sobre el ambiente es sumamente difícil o imposible y, por lo tanto, renuncian al esfuerzo.
- Los que evitan el fracaso. Aquellos estudiantes que carecen de un firme sentido de aptitud y autoestima y ponen poco esfuerzo en su desempeño; para "proteger su imagen ante un posible fracaso, recurren a estrategias como la participación mínima en el salón de clases, retraso en la realización de una tarea, trampas en los exámenes, etc." (pág. 77)

Covington nos da tres ideas según el orden en que se presentaron. El juego de la habilidad-esfuerzo. Si el alumno tiene éxito es porque invirtió poco o nada de esfuerzo, quiere decir que es brillante; ósea, es muy hábil, y se crea un sentimiento de orgullo y satisfacción. Sin embargo, cuando la situación de fracaso, las cosas cambian. El alumno invierte gran esfuerzo ya que posee poca habilidad, lo que genera sentimiento de humillación, porque el estudiante por más que se esfuerce para evitar que el profesor lo desapruebe sufre un sentimiento de humillación e inhabilidad. Dado la situación de fracaso, pone en duda su capacidad.

2.2.2.1 Buen rendimiento académico

Al buen rendimiento, hacemos referencia a la evaluación del conocimiento adquirido durante el ciclo en curso. Un estudiante con buen rendimiento académico es aquel, que aprueba todos los cursos con promedio ponderado mayor a 16 a lo largo de su educación, el estudiante que pertenece al tercio superior o quinto superior, se le hace el reconocimiento a través de un diploma de honor.

El estudiante para lograr tener alto rendimiento sigue ciertas estrategias:

- Participa en clases
- Realiza tareas en tiempo y forma
- Aprende a administrar su tiempo
- Toma apuntes de manea breve y puntualizados, entre otros.

La máxima nota es 20, se consideran aspectos cualitativos, por ejemplo, la conducta, los valores, la disciplina.

2.2.2.2 Rendimiento académico Medio

Un estudiante que tiene rendimiento académico medio está en el rango de nota de 12 a 16. Estos estudiantes muestran tendencia en subir o bajar su nota. Por ello, se considera imprescindible retroalimentar el proceso de enseñanza y logran cubrir todos sus créditos e incluso pueden pertenecer en el tercio superior o quinto, según el ciclo matriculado.

2.2.2.3 Bajo rendimiento académico

Según en la revista electrónica de psicología nos indica que (Garcia-Cruz, 2007) “el bajo rendimiento académico significa que el estudiante no

ha adquirido de forma adecuada y completa los conocimientos, además de que no posee las herramientas y habilidades necesarias para la solución de problemas referente al material de estudio" (pág. 25). Si bien también debemos ver cuáles son las causas y en la revista de estudios pedagógicos nos indica (Planck, 2014) "las causas del bajo rendimiento académico deben buscarse más allá del estudiante mismo". En otro estudio sobre el rendimiento académico dice que (Martinez, 2016)"el bajo rendimiento académico es una representación de escasos niveles de participación en el proceso de aprendizaje, el cual se ve influenciado por distintos factores, reflejándose en una nota bien sea cualitativa o cuantitativa; así pues "es una característica negativa del rendimiento académico" (pág. 34)

Para conocer más afondo las características de un estudiante debemos tomar en cuenta el total de estudiantes de la carrea profesional. Se va a desagregar en tres: identificación, académicas y abandono:

a. **Identificación:** Nos va a permitir delinear el perfil del estudiante, tanto en lo social, económico.
b. **Resultados académicos:** Las características académicas más relevantes de los estudiantes. Se va a trabajar con los datos de estudiantes que pertenecen a diferentes áreas académicas, también se debe tener en cuenta la data actitudinal.
c. **Abandono:** El abandono a una carrera profesional se puede dar al inicio, a mitad y muy rara vez finalizando, muchas veces suele pasar porque desaprueban cursos, entre otras características.

Los factores que influyen son diversos, como los amigos en el desempeño académico. Si los amigos son empeñosos en los estudios tendrá

buen rendimiento, pero si los amigos no son bien adaptados en el estudio, se perjudicará.

La influencia de los padres, según, (Peña, 1992) nos indica que, "La comprensión y el tratamiento de los aspectos psicopatológicos y psicosociales que rodean al estudiante hace imprescindible que se recoja información tanto de las pautas de interacción intrafamiliar como de las características psicosociales del medio escolar en que se desenvuelve".

Muchas veces el nivel educativo de los padres es muy importante en el desempeño académico de los hijos, según (Halpern, 1986). "Está relacionado con la pobreza, los hábitos de vida, los modelos de interacción familiar, entre otros". (pág. 2) El rendimiento académico de los estudiantes incide negativamente; ya que, al momento de ayudar en el estudio sus estrategias son poco efectivas por lo que interactúan escasamente en su aprendizaje.

2.2.2.4 Factores que inciden en el rendimiento académico

En el rendimiento Académico tenemos varios factores que intervienen en los estudiantes universitarios, ya que es, multicausal ya que pueden ser internas o externas al individuo. Si bien los factores presentan subcategorías o dimensiones. (Garbanzo Vargas, 2010)

a) Factores Socio - económico

- Entorno Familiar
- Contexto económico
- Nivel Educativo

b) Factores Académicos

- Elección de la Carrera

- Servicios Institucionales
- Ambiente Estudiantil

El bajo rendimiento académico es un problema que enfrentan estudiantes y educadores en todos los niveles educacionales. (Yaritza, 2014) Su trascendencia para el individuo y la sociedad es palpable a partir de 2 elementos fundamentales: primero, cuando el bajo rendimiento académico afecta la autorrealización profesional de los educandos, y segundo, cuando el nivel de conocimientos y habilidades que pueden adquirir, resulta limitado a las exigencias de su práctica profesional. (pág. 1)

En muchos casos el bajo rendimiento académico se da por muchos factores, según la revista Cubana de Investigaciones Biomédicas nos dice que (Giovanny, 2014) pueden ser: el fumar a temprana edad afecta la salud del adolescente ya que no solo afecta la salud sino también afecta negativamente al rendimiento académico, otro factor que también afectan es el consumo de bebidas alcohólicas esto a muy corto tiempo hará que los adolescentes pierdan oportunidades de formación y llegar a ser profesionales en el futuro. (pág. 36)

2.2.3 Técnicas de inteligencia artificial

Según la investigación de (Paola, 2010) nos dice que las técnicas de la Inteligencia Artificial son soportes valiosos y poderosos para abordar estos nuevos paradigmas que dejan clara la necesidad de contar con soportes de ayuda poderosos para tomar decisiones en la actividad empresarial. Estos soportes deben ser capaces de considerar la información cualitativa y a partir de ella diseñar e implementar modelos estadísticos y computacionales que asistan a los decisores en la resolución de los diversos problemas empresariales.

Las más destacadas aplicadas al campo de la gestión financiera empresarial son:

- Sistemas Expertos
- Redes neuronales
- Algoritmos Genéticos
- Lógica Difusa.

Estas técnicas pueden combinarse para obtener una solución más adecuada del problema en estudio. (pág. 2)

2.2.4 Árbol de decisión

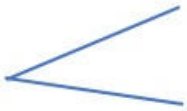
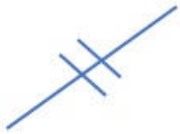
El árbol de decision o árboles de decisiones es una técnica más de inteligencia Artificial que nos ayuda a analizar decisiones secuenciales es un tipo de sistemas de experto. (Gamboa, 2014) Los árboles de decisión se clasifican en instancias, ordenadas a partir de la raíz, recorriendo diversos nodos de decisión, hasta que llegue a un nodo tipo hoja, el cual proporciona la clasificación de la instancia considerada. El aprendizaje por medio de árboles de decisión es generalmente más apropiado para problemas que poseen las siguientes características:

- Las instancias son representadas por pares atributo – valor.
- La función objetivo posee valores de salida discretos.
- Las descripciones son disyuntivas.
- Los datos de entrenamiento, pueden contener errores.
- Los datos de entrenamiento pueden tener atributos con valores incompletos o faltantes (pág. 132)

Con una base de datos se construyen diagramas de construcciones lógicas, similares a los sistemas de predicción basados en reglas, nos ayudan a tomar la decisión “más acertada” que nos sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema.

Figura 06

Símbolos utilizados en el diagrama de árbol de decisiones

FIGURA	NOMBRE	SIGNIFICADO
	Nudo de decisión	Indica una decisión que se tomará
	Nudo de probabilidad	Muestra múltiples resultados inciertos
	Nudo terminal	Indica un resultado definido
	Ramificaciones alternativas	Cada ramificación indica un posible resultado o acción
	Alternativa rechazada	Muestra una alternativa que no estaba seleccionada

Nota. Un árbol de decisión está conformado por una serie de nodos de decisiones con ramas que llegan y salen, como podemos observar se utiliza una serie de elementos. (Mendez, 2019)

I. Elementos de un árbol de decisión

En la técnica de árbol de decisión están formados por nodos, vectores de números, flechas y etiquetas

a) Los Nudos

Como parte de los elementos de un árbol de decisión tenemos a (Rafael, 2018) los nudos atienden a hitos de gran trascendencia en el caso, sugiriendo al intérprete que en dicho punto se habrán de tomar decisiones relevantes, o que ocurrirán hechos de cualquier naturaleza que influyan directamente en nuestros planteamientos, sin que nosotros podamos

interferir. (pág. 29)

Las cuales se dividen en:

- **Los nudos de decisión**

  Representadas por un cuadrado las decisiones son aquellas actuaciones donde de forma discriminatoria, se seleccionará una de entre las alternativas existentes. (Rafael, 2018) Este hecho repercute exclusivamente sobre el sujeto decisor.

  Las decisiones que se tomen proceden de la percepción del sujeto, de la intuición o sexto sentido, no estarán influenciadas por ninguna probabilidad, ya que dichas alternativas son excluyentes. (pág. 29)

- **Los nudos aleatorios**

  Para facilitar la explicación los (Rafael, 2018) nudos aleatorios bajo el modelo del árbol de decisión, y son representados con una circunferencia. Estos sucesos vienen condicionados por el entorno, por cualquier elemento que afecte al caso de estudio. (pág. 30)

- **Los nudos finales**

  No obstante, se dicen que los nudos finales nos ayudaran a tomar la decisión final. (Rafael, 2018) Con los nudos de fin concluyen las extensiones del diagrama de árbol, exponiendo cada uno de ellos los resultados, en términos de costes o en búsqueda de la mayor rentabilidad, en cada alternativa posible. (pág. 34)

  (Rafael, 2018) En ellos se resumen cuantitativamente las diferentes alternativas, dichos resultados aparecen representados por un triángulo y enumerados cuantifican cada alternativa. (pág. 35)

b) Ramificaciones alternativas:

Cada ramificación representa un resultado probable.

c) Alternativa Rechazada:

Una vez desarrollado el árbol las alternativas que no se seleccionan se marcan con dos líneas.

II. Pasos para crear un árbol de decision

Para crear un Árbol de Decision realizaremos los siguientes pasos: (Abel, 2019)

1. Iniciar el árbol dibujando un nodo cuadrado que representa la decisión principal del problema.
2. Agregar las ramas y nodos que se presentarán en los posibles caminos que podemos elegir a partir del nodo de decisión inicial.
3. En los nodos de probabilidad, colocar los valores de probabilidad en cada rama.
4. Determinar los valores esperados de cada nodo terminal, teniendo en cuenta los gastos e ingresos esperados de cada camino.
5. Calcular los valores esperados de los nodos intermedios. Para el caso de los nodos de probabilidad se calcularán con un promedio ponderado y para los nodos de elección se elegirá la opción más conveniente de las disponibles.
6. Para el primer nodo (el de decisión), se elegirá el valor esperado más conveniente dependiendo de la decisión que se desea tomar (pág. 2).

III. Clasificación de árboles de decisión

Los árboles de decisión utilizados en la minería de datos son de dos tipos principales: (Gamboa, 2014)

- Es cuando el resultado predicho es la clase a la que pertenecen los datos.
- Árboles de regresión es cuando el resultado predicho se puede considerar un número real (por ejemplo, el precio de una casa, las notas de un estudiante, etc.)

Hay muchos algoritmos específicos de árbol de decisiones. Entre los más destacados están:

3.1 ID3 (Iterative Dichotomiser 3)

El algoritmo ID3 selecciona la mejor característica en cada paso mientras crea un árbol de decisión, es un algoritmo desarrollado por J. Ross Quinlan en 1993, según estudio de (Lopez Takeyas, 2006) "objetivo de este algoritmo ID3 es construir un árbol de decisión que explique cada instancia de la secuencia de entrada de la manera más compacta posible a partir de una tabla d inducción." (pág. 1)

Un árbol de decision es una estructura con nodos, bordes y se construye a partir que hay una base de datos y se representa a través de características y atributos. En este algoritmo se selecciona la mejor función, es que ID3 usa Gian para encontrar la mejor característica.

Se utiliza la entropía que es la medida de la incertidumbre que hay, es decir la probabilidad de que ocurra cada uno de los posibles resultados.

Su expresión es con logaritmos base 2, más usada es:

$$entropía\ (S) = \sum_{i=1}^{n} -pi\ \log_2 pi$$

Donde:

S: es una colección de objetos

Pi: es la probabilidad de los posibles valores

i: las posibles respuestas de los objetos

Ross Quinlan define el esquema del algoritmo ID3, el cual adopta los siguientes:

- Calcular la entropía para todas las clases.
- Seleccionar el mejor atributo basado en la reducción de la entropía.
- Iterar hasta que todos los objetos sean clasificados.

3.2 C4.5(Sucesor de ID3)

El C4.5 es una extensión del ID3 que acaba con muchas de sus limitaciones, según investigaciones dice que el C4.5 se mejora mucho. (Servente, 2003) En el nodo del árbol de selección (atributo) ID3 es seleccionar el atributo con el mayor valor de ganancia de información como nodo. Mientras C4.5 introduce un nuevo concepto de "tasa de ganancia de información", C4.5 selecciona el atributo con la tasa de ganancia de información más grande como el nodo del árbol. (pág. 1)

El ID3 y el C4.5 utilizan la estrategia de “divide y reinarás” para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento. Árboles de clasificación, algoritmo J48 es una implementación libre que fue adaptado por Weka en código de programación java, se utiliza el concepto de entropía en cada una de las clases. La idea original de esta estrategia de divide y vencerás se basa en un trabajo de Hoveland y Hunt de los años 50, culminado en el libro

Experiments in Induction [Hunt y otros, 1966].

2.2.5 Random Forest

Random Forest es un algoritmo que utiliza una técnica de aprendizaje automático supervisada basada en árboles de decisión. Es un modelo que está formado por un conjunto de árboles de decisión individuales, cada uno entrenado. Para entender cómo funcionan los modelos Random Forest es necesario conocer conceptos de ensemble y bagging, según (Joaquin, 2017).

Random Forest es una combinación de árboles predictivos (clasificadores débiles); es decir, una modificación del Bagging, el cual trabaja con una colección de árboles incorrelacionados y los promedia, según (Hastie, 2001), en el cual se tiene que cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque.

En la etapa de entrenamiento, el algoritmo intenta optimizar los parámetros de las funciones de split a partir de las muestras de entrenamiento. (Breiman L. F., 1984)

$$\theta_k^* = argmax_{\theta\ j\epsilon\tau_j} I_J$$

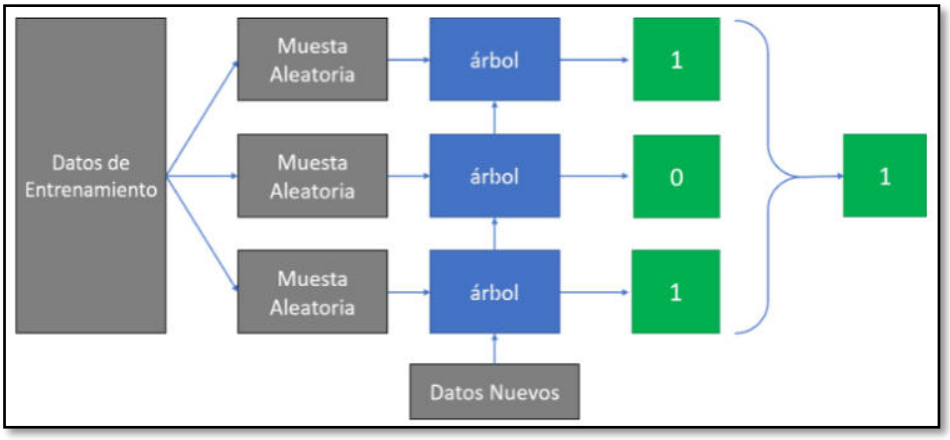
Para ello se utiliza la siguiente función de ganancia de información:

$$I_J = H_{(J)} - \sum_{i\epsilon 1,2} \frac{S_j^i}{S_j} H(S_j^i)$$

Donde S representa el conjunto de muestras que hay en el nodo por dividir, y S^i son los dos conjuntos que se crean de la escisión. La función mide la entropía del conjunto, y depende del tipo de problema que abordamos (Breiman L. , 2001).

Figura 07

Funcionamiento de random forest



Nota. Random Forest es un conjunto de árboles la cual los datos se entrenan en forma aleatoria, la cual se combinan sus resultados. (Jose, 2020)

2.2.5.1 Por qué utilizar el algoritmo de Random Forest

Las principales ventajas de estos modelos frente a otras familias son: (Daniel, 2019).

- **Reduce la varianza.** Debido a la utilización de múltiples árboles de decisión la varianza de la predicción disminuye respecto a un solo árbol de decisión. Lo que reduce la posibilidad de obtener un modelo sobre ajustado. Aun así, el sesgo del modelo es el mismo que se obtiene con un solo árbol de decisión.
- **No requiere la asunción de suposiciones.** Cuando solamente se quiere un modelo para realizar predicciones Random Forest es una solución excelente. Esto es así porque no es necesario realizar suposiciones sobre el modelo o los conjuntos de datos. (pág. 2)

2.2.5.2 ¿Cómo se combinan las predicciones?

Según, (Daniel, 2019). En la página web artificial nos indica que, para los problemas de clasificación, se suelen combinar los resultados de los

árboles de decisión usando soft-voting (voto suave). En el voto suave, se le da más importancia a los resultados en los que los árboles estén muy seguros. Para problemas de regresión, la forma más habitual de combinar los resultados de los árboles de decisión, es tomando su media aritmética (pág. 1)

2.2.5.3 Construcción de los Random Forest

A cada árbol se le asigna

- Una parte aleatoria de los datos (tree bagging)
- Una parte aleatoria de los features (feature sampling)

RF = Tree bagging + feature sampling

Tree Bagging

La construcción de B árboles se hace con sigue:

- Se tiran al azar, y con reposición, B muestras del problema (X, Y), que notamos (Xb, Yb) ($b \in \{1, B\}$)
- Se entrena un árbol sobre cada par (Xb, Yb) Eso baja el overfitting

Feature Sampling

- Sobre los n features, cada árbol tiene acceso a solamente una parte (típicamente $\sqrt{n}$).
- Esto baja la correlación entre los árboles (notado por el coeficiente ρ).

2.2.5.4 Criterios de split de Random Forest

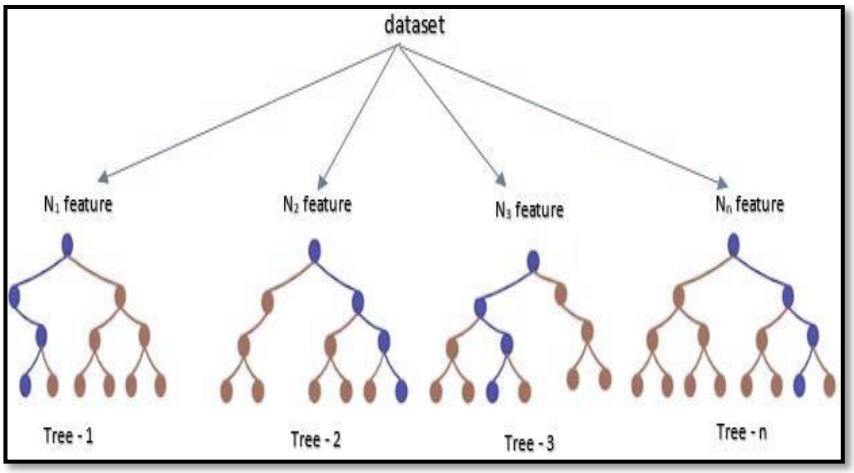
Existen dos criterios para hacer una división de ramas (split):

- **El criterio de Gini**

  Principio: Tomar la clase l más representada, y ve por qué *feature* se distingue.

Figura 08

Criterios de gini de random forest



Nota. Ejemplo de aplicación del criterio de Gini (Romain, 2016)

- **El criterio de Entropía**

  Sea S un conjunto de datos labelizados {+, -}

  Sea p+ la proporción de población positiva en S, y p−

  negativa. Entonces:

  Entropía(S) = −p+ log p+ − p− log p−

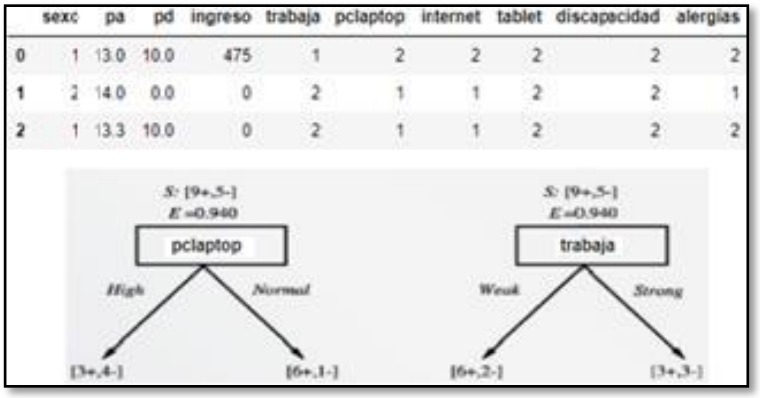
  Ganancia de información

  Sea R un nuevo nodo N dando H nuevas hojas

$$Ganancia\ (S,R) = Entropía\ (S) - \sum_{h\epsilon\ H} \frac{|S_h|}{|S|}\ Entropía\ (S_h)$$

Figura 09

Criterio de entropía de random forest



Nota. Ejemplo de aplicación del criterio de entropía (Romain, 2016)

2.2.6 Clasificador Naive Bayes

Es uno de los algoritmos de clasificación que busca clasificar o etiquetar correctamente un conjunto de datos, la clasificación está basado en el Teorema de Bayes con una suposición de independencia entre los predictores. Según, (Naive, 2021) La fórmula del teorema de Bayes es la siguiente: (Ratanamahatana, 2002)

$$P\left(\frac{h}{D}\right) = \frac{P\left(\frac{D}{h}\right)P(h)}{P(D)}$$

Donde:

- **P(h):** es la probabilidad de que la hipótesis h sea cierta (*independientemente de los datos*). Esto se conoce como la probabilidad previa de h.
- **P(D):** probabilidad de los datos (*independientemente de la hipótesis*). Esto se conoce como probabilidad previa.
- **P(h|D):** es la probabilidad de la hipótesis h dada los datos D. Esto se conoce como la probabilidad posterior.
- **P(D|h):** es la probabilidad de los datos d dado que la hipótesis h era cierta. Esto se conoce como probabilidad posterior.

En caso de que se tenga una sola característica, el clasificador Naive Bayes calcula la probabilidad de un evento en los siguientes pasos: (Rish, 2001)

- Paso 1: calcular la probabilidad previa para las etiquetas de clase dadas.
- Paso 2: determinar la probabilidad de probabilidad con cada atributo para cada clase.
- Paso 3: poner estos valores en el teorema de Bayes y calcular la probabilidad posterior.
- Paso 4: ver qué clase tiene una probabilidad más alta, dado que la variable de entrada pertenece a la clase de probabilidad más alta

Según, (Jaimes, 2015) Diferenciando los clasificadores basados en redes bayesianas (Naive Bayes, (Kononenko, 1991) Red Bayesiana aumentada a árbol TAN, Semi Naive Bayes y otros) el clasificador Naive Bayes ha demostrado comportarse bastante bien a pesar de que asume que las variables predictoras son condicionalmente independientes dada la clase, lo que generalmente no se cumple. (Rish, 2001) Este clasificador Naive Bayes es uno de los modelos más efectivos, debido a su simplicidad, su resistencia al ruido, poco tiempo de procesamiento y alto poder predictivo (pág. 27)

2.2.7 Metodología CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining), está compuesta por seis fases las cuales dependen entre sí tanto en forma secuencial como cíclica pudiendo encontrarse interacciones que permitan mejorar la aproximación obtenida en otras fases anteriores.

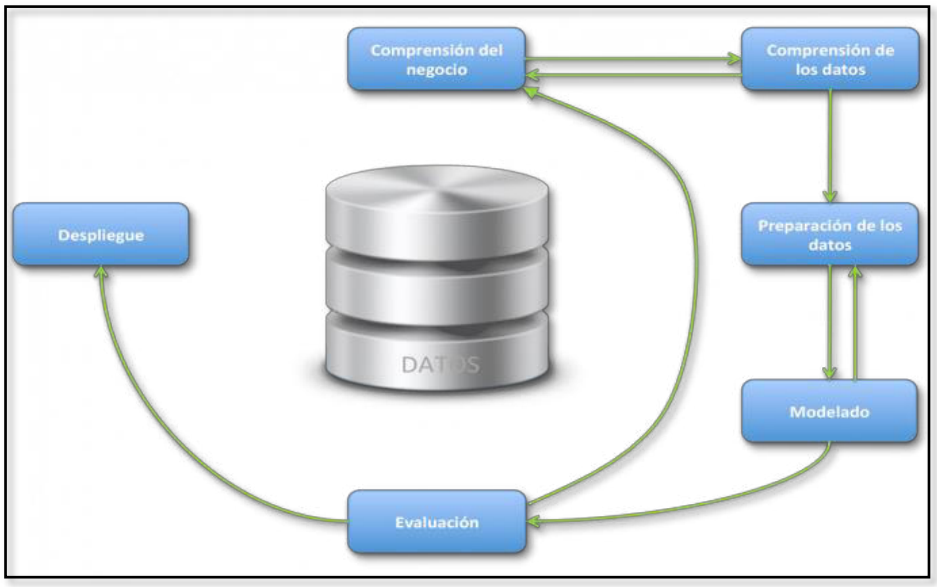
Las fases son las siguientes:

- Comprensión del negocio.

- Comprensión de los datos.
- Preparación de los datos.
- Modelado.
- Evaluación.
- Despliegue.

Figura 10

Secuencia del proceso CRISP-DM



Nota. Como se observa son 6 procesos de la metodología CRISP-DM (Rodriguez, 2010)

2.2.7.1 CRISP-DM: Comprensión del negocio

En la primera etapa debemos comprender en profundidad en profundidad el problema que se quiere resolver y estableciendo los requisitos y objetivos del proyecto desde una perspectiva empresarial para luego trasladarlos a objetivos técnicos y a un plan de proyecto. (Rodriguez, 2010)

2.2.7.2 CRISP-DM: Comprensión de los datos

Es esta fase se lleva a cabo la recolección y exploración inicial de los datos, con el objetivo de establecer un primer contacto con el problema. Esta fase suele ser crítica en el proyecto dado que un mal entendimiento de los datos tiene como consecuencia un aumento en el tiempo global del proyecto y también reduce las garantías de éxito. (Rodriguez, 2010)

2.2.7.3 CRISP-DM: Preparación de los datos

La fase de preparación trata de seleccionar, limpiar y generar conjuntos de datos correctos, organizados y preparados para la fase de modelado. Esta es una fase sumamente crítica en un proyecto de minería de datos. Los errores en los datos que se pasan por alto y que no son resueltos en esta fase se trasladan hasta la fase de modelado, lo que genera una reducción en la exactitud de los modelos o incluso, es posible entregar al cliente resultados basados en datos que aún contienen errores no detectados. (Rodriguez, 2010)

Por esta razón, esta fase es crucial y generalmente demanda siempre el mayor esfuerzo y tiempo del proyecto, aproximadamente un 75% del tiempo total.

2.2.7.4 CRISP-DM: Modelado

En esta fase se lleva a cabo la creación de modelos de conocimiento a partir de los datos suministrados desde la fase anterior. Estos modelos de conocimiento pueden ser de distintos tipos, por ejemplo, se pueden crear modelos de clasificación o regresión con el objetivo de estimar o inferir el valor de una determinada variable. (Rodriguez, 2010)

Para afrontar esta fase hay que seguir una serie de pautas que nos ayudarán a obtener mejores resultados se debe seguir:

a) **Escoger la técnica de modelado:** Antes de escoger la técnica, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. (Cortina, 2015)

b) Generar el plan de prueba

Se debe generar un procedimiento destinado a probar la calidad y validez del modelo elegido una vez que éste esté construido. Por ejemplo, en una tarea supervisada de minería de datos como la clasificación (Cortina, 2015)

c) Construir el modelo.

Se ejecuta la técnica seleccionada sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. (Cortina, 2015)

2.2.7.5 CRISP-DM: Evaluación

Si los modelos obtenidos cumplen con las expectativas de negocio, se procede a la explotación del modelo. Si no, se evalúa en esta fase si se procede a iterar nuevamente sobre los pasos anteriores con el objetivo de encontrar nuevos resultados. (Rodriguez, 2010)

2.2.7.6 CRISP-DM: Despliegue

Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados. (Cortina, 2015)

2.2.8 Software Weka (*Waikato Environment For Knowledge Analysis*)

Es un software de código libre que tiene una colección extensa de algoritmos según el manual de Weka (Garcia, 2005) La licencia de Weka es GPL, lo que significa que este programa es de libre distribución y desarrollado por la universidad de Waikato (Nueva Zelanda) implementados en Java [1, 2]; útiles para ser aplicados sobre datos mediante los interfaces que ofrece o para embeberlos dentro de cualquier aplicación. Además, Weka contiene las herramientas necesarias para realizar transformaciones de los datos, tareas de clasificación, regresión, clustering, asociación y visualización. Weka está diseñado como una herramienta orientada a la extensibilidad se añade nuevas funcionalidades, una tarea sencilla. (pág. 8)

Este software ayuda mucho cuando uno desea implementar diversos algoritmos, ya que nos una mejor visión para la toma de decisiones según el nivel de precisión que genera de acuerdo a las instancias y clases de nuestra base de datos. Tiene diversas herramientas para una mejor visibilidad de los datos, también nos ayuda a saber si hemos tomado bien nuestros atributos para el estudio respectivo.

Figura 11.

Modelo de clasificación

EjemploClasificador.java

```
/*
* Ejemplo de clasificador
* -----------------------
* Calcula del conjunto de entrenamiento la media
* de los valores de todos sus datos, y compara con la media
* de la instancia a clasificar. Si es mayor clasifica con la
* clase cuyo indice valor sea 0.0 si no clasifica con la clase
* cuyo indice valor es 1.0.
*/
import weka.classifiers.*
import weka.core.*
import weka.util.*

public class EjemploClasificador extends Classifier {

   private double media_entrenamiento = 0; //Media del conjunto de entrenamiento

   public void buildClassifier(Instances instancias) throw Exception {

     // Recorremos todas las instancias
     for (int i=0; i<instancias.numClasses();m_numclase.length; i++) {
       // Recorremos todos sus atributos
       for (int j=0; j < instancias[i].numAttributes(); j++) {
         media_entrenamiento += instancias[i].value[j]
           / instancias[i].numAttributes();
       }
     }
   } // buildClassifier

   public double classifyInstance(Instance instance) {
     float media_instancia = 0;

   // Recorremos todos sus atributos
   for (int j=0; j < instancias[i].numAttributes(); j++) {
     media_instancia += instancias[i].value[j] / instancias[i].numAttributes();
   }

   if (media_instancia > media_entrenamiento) {
     return 0.0;
   } else if {
     return 1.0;
   }
   } // classifyInstance
} // Fin clase
```

Nota. Este código de programación fue desarrollado en java, para Weka es muy sencillo y para el usuario es algo enrevesada, en parte por su mal diseño, pero muy útil para utilizar sus algoritmos.

2.2.9 PYTHON

Python fue desarrollado por Python Software Foundation, maneja varias extensiones como .py, .pyc, pyd, .pyo, .pyw, .pyz, pyi. (Alberca, 2020). Se trata de un lenguaje de programación multiparadigma ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma. Este lenguaje de programación apareció por los años 1991 fue diseñado por Guido Van Rossum.

Principales ventajas de Python: (Alberca, 2020)

- Es de código abierto (certificado por la OSI).
- Es interpretable y compilable.
- Es fácil de aprender gracias a que su sintaxis es bastante legible para los humanos.
- Es un lenguaje maduro (29 años).
- Es fácilmente extensible e integrable en otros lenguajes (C, java).

Está mantenido por una gran comunidad de desarrolladores y hay multitud de recursos para su aprendizaje. (pág. 8)

Figura 12.

Librerías de python

```
import pandas as pd
import numpy as np
import seaborn as sns
# Diferentes tipos de clasificadores Naive Bayes
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
import matplotlib.pyplot as plt
```

Nota. Una de las librerías muy utilizadas en Algoritmos de Clasificación para importar datos

2.3 DEFINICIÓN DE TÉRMINOS

2.3.1 Definición de Académico.

El término académico proviene del griego akademia (*el lugar ubicado en las afueras de Atenas donde Platón se reunía a estudiar*) y es aquel que es utilizado para denominar no sólo a individuos sino también a entidades, objetos o proyectos que se relacionan con niveles superiores de educación. (Bembibre, 2009)

2.3.2 Definición de modelo predictivo

Es un modelo de datos basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a una promoción de marketing directo. El modelo predictivo utiliza estadísticas para predecir los resultados. (arimetrics, 2008)

2.3.3 Definición de Inteligencia artificial

Lo cierto que no hay un consenso entre los científicos e ingenieros sobre lo que es inteligencia artificial, y muchos menos se ha llegado a un definición exacta y concisa que nos permita dirimir que programas son o no inteligentes. El problema es que ni siquiera tenemos la certeza de que seamos capaces de definir que es inteligencia (*no artificial*). (Garcia S., 2012)

2.3.4 Definición de Árbol de Decisión

Utilizan una técnica de aprendizaje supervisado que consiste en realizar particiones recursivas en un espacio de datos conformado por un conjunto de casos o ejemplos. El árbol que se va construyendo con cada partición, sus nodos representan un atributo que sirvió para realizar una partición, los arcos que conectan los nodos entre si representan los diferentes valores que pueden tomar los atributos y las hojas representan subconjuntos de casos pertenecientes a una clase. (Lopera, 2018)

2.3.5 Definición de Redes neuronales artificiales

Son redes interconectadas masivamente en paralelo de elementos simples (*usualmente adaptativos*) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico. (Ruiz, C., Basualdo M. S., 2001)

2.3.6 Definición de Redes Bayesianas

Son una representación gráfica de dependencias para razonamiento probabilístico, en la cual los nodos representan variables aleatorias y los arcos representan relaciones de dependencia directa entre las variables (Sucar, 2018).

2.3.7 Definición de Rendimiento Académico

Es un logro alcanzado por el educando en todo el proceso de enseñanza – aprendizaje. En tanto, (Nováez, 1986) sostiene que el rendimiento académico es el Quantum obtenido por el individuo en determinada actividad académica.

Otra forma de definir el rendimiento académico según, (Chadwick, 1979) es como la expresión de capacidades y de características psicológicas del estudiante desarrolladas y actualizadas a través del proceso de enseñanza-aprendizaje que le posibilita obtener un buen funcionamiento y logros académicos. (pág. 34)

CAPITULO III: MARCO METODOLÓGICO

3.1 Lugar de ejecución

- **Lugar:** Universidad Nacional de Moquegua (UNAM) de la Escuela Profesional de Ingeniería de Sistemas e Informática (EPISI)
- **País:** Perú
- **Región:** Moquegua
- **Provincia:** Ilo
- **Distrito:** Pacocha - Urb. Ciudad Jardín S/N

3.2 Tipo y Diseño

3.2.1 Tipo de Investigación

Por su nivel de profundidad

3.2.1.1 **Investigación exploratoria** (Babbie, 1996)

Según, (Babbie, 1996) "Se explora un tema y se estructura". Nos ayuda realizar el análisis de las causas, donde encontraremos los procedimientos adecuados para una investigación posterior es flexible; es decir, estudiar fenómenos nuevos respaldándonos con todo el material existente.

3.2.1.2 **Investigación descriptiva** (Babbie, 1996)

Según (Babbie, 1996) "Describe situaciones y acontecimientos". Se describe los datos y este debe tener un impacto en la investigación realizada a los alumnos de la Escuela profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua.

3.2.1.3 Investigación campo.

En este tipo de investigación los datos e información socioeconómica, personal e institucional nos brindó la Dirección de Bienestar universitario; si bien estaba en físico se tuvo que extraer. Con respecto a los datos académicos se solicitó a la Dirección de Actividades y Servicios Académicos. Se realizó una encuesta para recolectar información extra para la investigación.

3.2.2 Diseño de la Ejecución

Diseño Experimental

En el diseño experimental se manipula una o más variables de estudio. Dicho de otra forma, un experimento consiste variables dependientes y variables independientes. (Ronal, 1962)

3.3 Nivel de Investigación

Se ha utilizado el nivel experimental; ya que, conforme vamos a determinar las características de los estudiantes con bajo rendimiento académico, los estudiantes que pueden abandonar la carrera, los cursos que presentan mayor número de desaprobados, entre otros, se experimentará a través de la clasificación si tengo dos o más categorías y la regresión si deseo predecir algo.

3.4 Operacionalización de variables

Operacionalización de la variable, factores que inciden en el rendimiento académico: Factores sociales, factores económicos, factores académicos y factores actitudinales.

Tabla 01

Operacionalización de variables

Variable	Tipo	Indicadores	Valores
Variables independientes	Factores socio-económico	Género	M, F
		ingreso	numérico
		trabaja	Si, No
		Pc-laptop	Si, No
		Internet	Si, No
		Tablet	Si, No
		Discapacidad	Si, No
		Alergias	Si, No
	Factor académico	Promedio aprobado	Numérico
		Promedio desaprobado	
		....	

Nota. Operacionalización de la variables donde visualizamos los factores que inciden en el rendimiento académico fue adaptado de (Garbanzo Vargas, 2010) (pág. 8)

3.4.1 Variables Independientes

Factores:

- Factores Socio - económico
- Factores académicos

3.4.2 Variables Dependientes

Datos académicos (Clases: buen y bajo rendimiento).

3.5 Población y muestra

3.5.1 Población:

Estudiantes universitarios de la Escuela Profesional de Ingeniería de Sistemas e Informática - Universidad Nacional de Moquegua.

Tabla 02

Total, de estudiantes del caso de estudio

Total, de Estudiantes	Varones	Mujeres
167	128	39

Nota. Se contabiliza el total de estudiantes, según el caso de Estudio

3.5.2 Muestra:

Se tomó como muestra a los estudiantes matriculados en los años 2019-I, 2019-II, 2020-I, 2020-II.

3.6 Delimitación Espacial y Temporal

El presente trabajo de investigación se realizará en la región Moquegua datos del año 2019 -2020.

3.7 Método de Investigación

El método de la investigación a utilizar se conoce, como "Investigación - Acción" porque se pretende implementar una solución y validarla a través de un caso de estudio.

Representa un esfuerzo conjunto entre los profesionales y los agentes locales durante todo el proceso de la investigación: desde la definición del problema por investigar hasta el análisis de los resultados. Implica el uso de múltiples métodos en el recojo de la

información y en el análisis de los resultados, tiene como objeto resolver un problema en un contexto determinado aplicando método científico. (EcuRed, 2016)

3.8 Técnicas, Instrumentos para recolección de datos

3.8.1 Encuesta

Esta técnica es aplicada a los estudiantes de la carrera de Ingeniería de Sistemas e Informática, el instrumento es el cuestionario online, en el aula virtual, por lo que esta información es muy útil para el investigador.

3.8.2 Procesamiento de datos

Sistema de Base de Datos

- Data académica extraída del total de matriculados 2019-I
- Data académica extraída del total de matriculados 2019-II
- Data académica extraída del total de matriculados 2020-I
- Data académica extraída del total de matriculados 2020-II

Forma de Tratamiento de los Datos

- Tabulación de datos.
- Filtración de datos

Tabla 03

Cantidad de alumnos, según el código académico

Código Académico	Cantidad
2017	5
2018	20
2019	73
2020	69

Nota. Este cuadro nos muestra la cantidad de alumnos según su código académico, año de matrícula 2019 – 2020 de los ciclos I – II, figuran en el sistema de la UNAM. Datos proporcionados por OTIN – autorizado por VIPAC

3.9 Validación y Confiabilidad de los Instrumentos

Son datos obtenido de la Universidad Nacional de Moquegua – Escuela de Ingeniería de Sistemas e Informática

CAPITULO IV: PRESENTACION DE RESULTADOS

El objetivo de este capítulo es dar a conocer el desarrollo del problema planteado, utilizando la data obtenida del caso de estudio de EPISI – UNAM.

4.1 Comprensión del Negocio

La Universidad Nacional de Moquegua, (UNAM, 2019) es una entidad pública de educación superior universitaria, ubicada en la ciudad de Moquegua e Ilo, cuenta con 8 carreras. Mediante Resolución de Consejo Directivo N° 103-2017-CD la SUNEDU autoriza a la UNAM a prestar el Servicio Educativo Superior Universitario por seis años.

La Universidad está presidida por una comisión organizadora que consta de presidente, vicepresidente académico y vicepresidente de Investigación. (pág. 1)

4.1.1 Funciones

La Universidad Nacional de Moquegua tiene como funciones básicas.

En lo Académico

a. Desarrollar Investigación Científica, Tecnológica y Humanística como soporte del desarrollo académico y la integración con el entorno.
b. Participar con entidades de desarrollo del estado, entidades privadas y otras instituciones, en la ejecución de estudios y/o proyectos de investigación que contribuyan al desarrollo regional y nacional.
c. Desarrollar la enseñanza para la formación profesional.
d. Promover la integración con la comunidad a través de programas de Proyección y Extensión Universitaria.
e. Implementar programas de capacitación docente e intercambio cultural.
f. Desarrollar programas o procesos de Producción de Bienes y Prestación de Servicios, para complementar la enseñanza teórica.

En lo Administrativo

a. Realizar una gestión eficiente, transparente, ética y honesta; con apego a las normas legales vigentes y lo establecido en los Sistemas Administrativos.
b. Contribuir con mejorar el desempeño laboral y educativo, implementando servicios de bienestar y promoción social, salud, deporte y recreación, para todos los miembros de la comunidad universitaria.
c. Mantener las relaciones y comunicaciones con otras instituciones educativas, culturales, estatales, sociales y económicas con fines de coordinación, cooperación y conocimiento recíproco.

4.1.2 **Objetivos Estratégicos** (UNAM, 2019)

OEI.01. Mejorar la formación académica de los estudiantes universitarios.

OEI.02. Promover la investigación científica y tecnológica en la comunidad universitaria.

OEI.03. Fortalecer las actividades de extensión cultural y de proyección social para la comunidad universitaria.

OEI.04. Fortalecer la Gestión Institucional.

OEI.05. Implementar la gestión del riesgo de desastres.

4.1.3 **Visión**

La visión de la Universidad (UNAM, 2019) al 2050, somos un país democrático, respetuoso del Estado de derecho y de la institucionalidad, integrado al mundo y proyectado hacia un futuro que garantiza la defensa de la persona humana y de su dignidad en todo el territorio nacional.

Estamos orgullosos de nuestra identidad, propia de la diversidad étnica, cultural y lingüística del país. Respetamos nuestra historia y patrimonio milenario, y protegemos nuestra biodiversidad.

El Estado constitucional es unitario y descentralizado. Su accionar es ético, transparente, eficaz, eficiente, moderno y con enfoque intercultural.

Juntos, hemos logrado un desarrollo inclusivo, en igualdad de oportunidades, competitivo y sostenible en todo el territorio nacional, que ha permitido erradicar la pobreza extrema y asegurar el fortalecimiento de la familia.

4.1.4 Misión

La misión de la universidad (UNAM, 2019). Somos una comunidad académica que forma profesionales e investigadores que sean innovadores, competitivos, éticos y multiculturales para contribuir con el bienestar de la sociedad.

4.1.5 Comprensión del problema

Toda carrera profesional debe de acreditarse y en este trabajo de Investigación busca hacer seguimiento para la nivelación de las deficiencias de sus estudiantes y que podrían presentarse durante el proceso formativo; por ello, se hizo la gestión para obtener la base de datos de: ficha socioeconómica, base de datos académica de los años 2019, 2020 lo que se busca es, saber porque el estudiante tiene bajo rendimiento académico cuales son los factores que influyen.

4.2 Comprensión de Data

4.2.1 Descripción de los Datos Generales (Fuente de Datos)

En la metodología CRISP-DM una de las etapas es el levantamiento de información; por ello, se muestra la data extraída de la ficha socio económica y la base de datos académica del caso de Estudio: Escuela Profesional de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua.

De la base de datos obtenida se realizó un diccionario de datos para entender mejor las variables independientes que ayudará mucho en el análisis y poder interpretar el modelo que deseo diseñar. Esta base de datos se obtuvo de la misma universidad con previa gestión.

Tabla 04

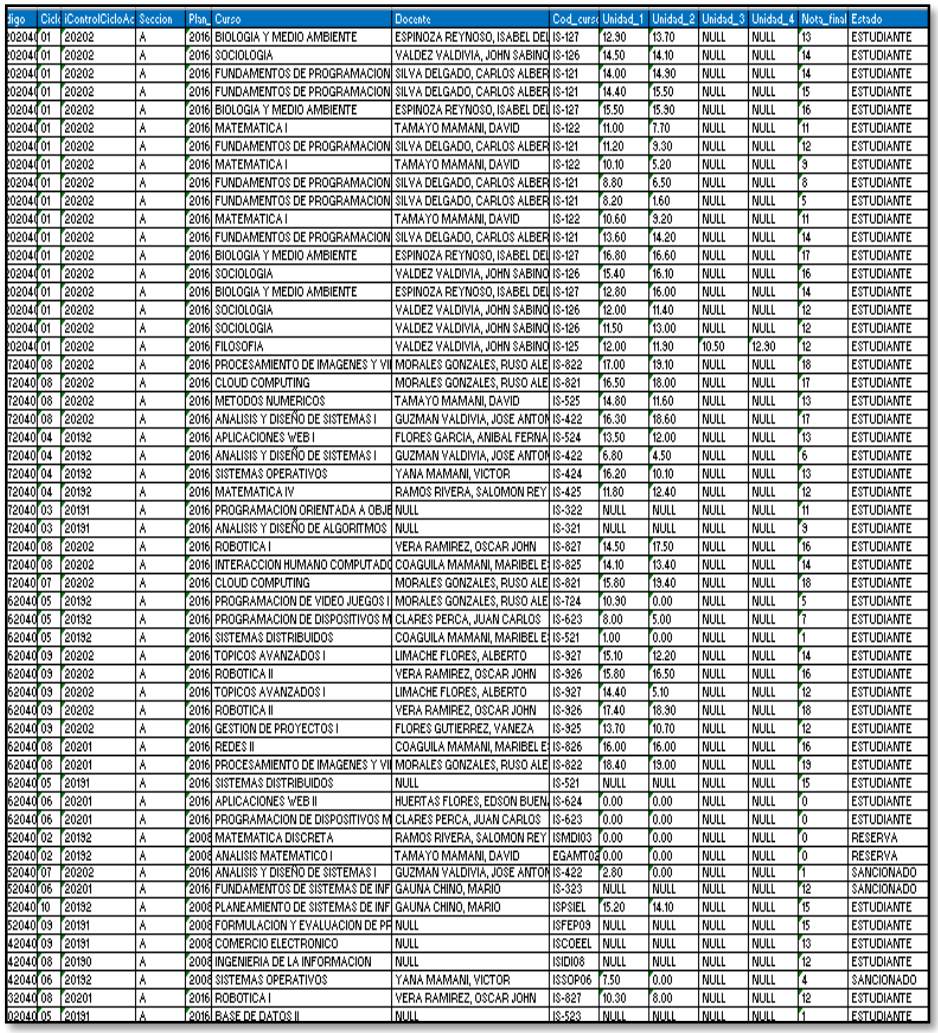
Diccionario de datos

Indicadores	Descripción	Valor
Género	Género	M, F
Ingreso	Cantidad de Ingreso	Numérico
trabaja	trabajas	Si, No
Pc-laptop	Cuenta con Pc-laptop	Si, No
Internet	Cuenta con Internet	Si, No
Tablet	Cuenta con Tablet	Si, No
Discapacidad	Sufres de Discapacidad	Si, No
Alergias	Sufres de alergia	Si, No
Pa	Promedio Aprobados	Numérico
Pd	Promedio Desaprobados	Numérico

Nota. En la tabla colocamos todas las variables seleccionadas con su previa descripción y a que factor corresponde

Figura 13

Datos no procesados académico del caso de estudio año 2019, 2020

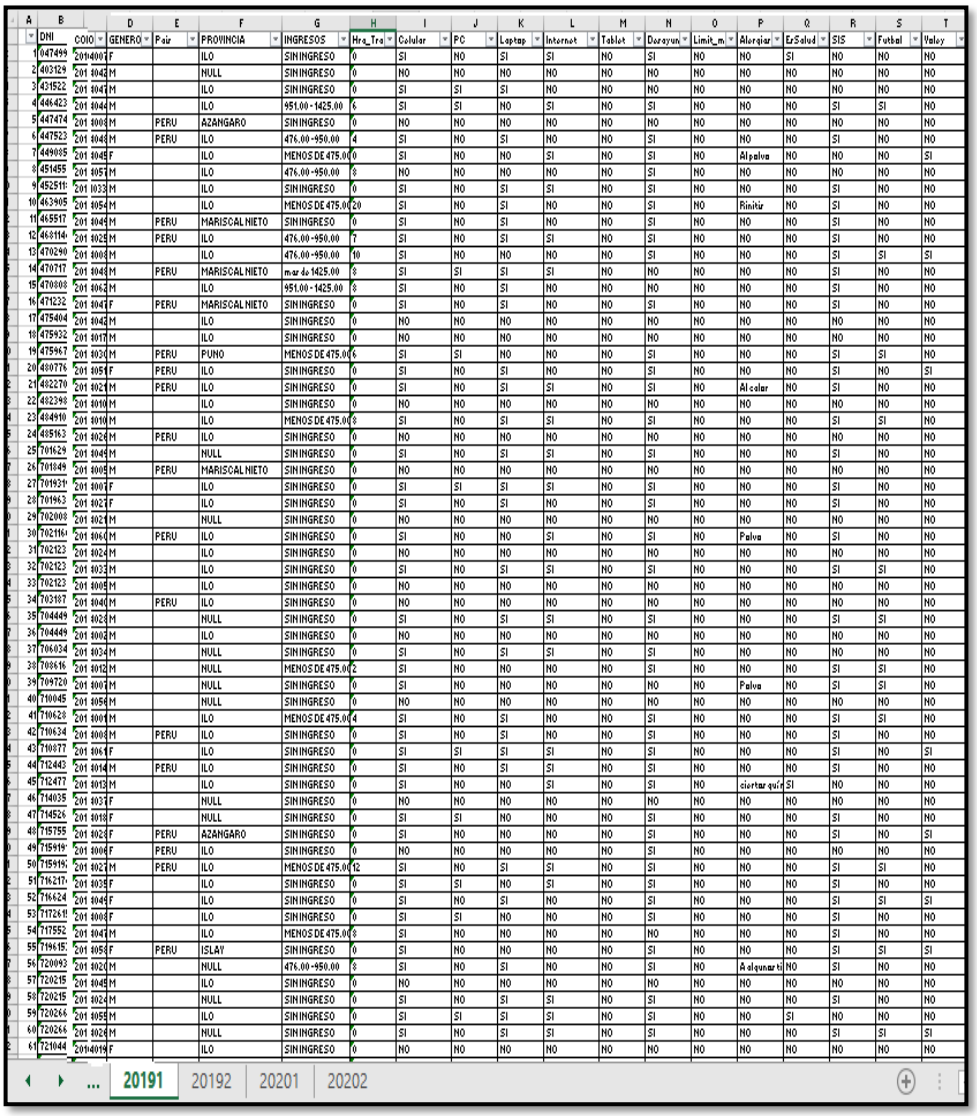


digo	Ciclo	iControlCicloAc	Seccion	Plan_	Curso	Docente	Cod_curs	Unidad_1	Unidad_2	Unidad_3	Unidad_4	Nota_final	Estado
02040	01	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	12.90	13.70	NULL	NULL	13	ESTUDIANTE
02040	01	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	14.50	14.10	NULL	NULL	14	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	14.00	14.30	NULL	NULL	14	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	14.40	15.50	NULL	NULL	15	ESTUDIANTE
02040	01	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	15.50	15.30	NULL	NULL	16	ESTUDIANTE
02040	01	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	11.00	7.70	NULL	NULL	11	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	11.20	3.30	NULL	NULL	12	ESTUDIANTE
02040	01	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	10.10	5.20	NULL	NULL	9	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	8.80	6.50	NULL	NULL	8	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	8.20	1.60	NULL	NULL	5	ESTUDIANTE
02040	01	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	10.60	9.20	NULL	NULL	11	ESTUDIANTE
02040	01	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBER	IS-121	13.60	14.20	NULL	NULL	14	ESTUDIANTE
02040	01	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	16.80	16.60	NULL	NULL	17	ESTUDIANTE
02040	01	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	15.40	16.10	NULL	NULL	16	ESTUDIANTE
02040	01	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	12.80	16.00	NULL	NULL	14	ESTUDIANTE
02040	01	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	12.00	11.40	NULL	NULL	12	ESTUDIANTE
02040	01	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	11.50	13.00	NULL	NULL	12	ESTUDIANTE
02040	01	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	12.00	11.90	10.50	12.90	12	ESTUDIANTE
72040	08	20202	A	2016	PROCESAMIENTO DE IMAGENES Y VI	MORALES GONZALES, RUSO ALE	IS-822	17.00	19.10	NULL	NULL	18	ESTUDIANTE
72040	08	20202	A	2016	CLOUD COMPUTING	MORALES GONZALES, RUSO ALE	IS-821	16.50	18.00	NULL	NULL	17	ESTUDIANTE
72040	08	20202	A	2016	METODOS NUMERICOS	TAMAYO MAMANI, DAVID	IS-525	14.80	11.60	NULL	NULL	13	ESTUDIANTE
72040	08	20202	A	2016	ANALISIS Y DISEÑO DE SISTEMAS I	GUZMAN VALDIVIA, JOSE ANTON	IS-422	16.30	18.60	NULL	NULL	17	ESTUDIANTE
72040	04	20132	A	2016	APLICACIONES WEB I	FLORES GARCIA, ANIBAL FERNA	IS-524	13.50	12.00	NULL	NULL	13	ESTUDIANTE
72040	04	20132	A	2016	ANALISIS Y DISEÑO DE SISTEMAS I	GUZMAN VALDIVIA, JOSE ANTON	IS-422	6.80	4.50	NULL	NULL	6	ESTUDIANTE
72040	04	20132	A	2016	SISTEMAS OPERATIVOS	YANA MAMANI, VICTOR	IS-424	16.20	10.10	NULL	NULL	13	ESTUDIANTE
72040	04	20132	A	2016	MATEMATICA IV	RAMOS RIVERA, SALOMON REY	IS-425	11.80	12.40	NULL	NULL	12	ESTUDIANTE
72040	03	20131	A	2016	PROGRAMACION ORIENTADA A OBJE	NULL	IS-322	NULL	NULL	NULL	NULL	11	ESTUDIANTE
72040	03	20131	A	2016	ANALISIS Y DISEÑO DE ALGORITMOS	NULL	IS-321	NULL	NULL	NULL	NULL	9	ESTUDIANTE
72040	08	20202	A	2016	ROBOTICA I	VERA RAMIREZ, OSCAR JOHN	IS-827	14.50	17.50	NULL	NULL	16	ESTUDIANTE
72040	08	20202	A	2016	INTERACCION HUMANO COMPUTADO	COAGUILA MAMANI, MARIBEL E	IS-825	14.10	13.40	NULL	NULL	14	ESTUDIANTE
72040	07	20202	A	2016	CLOUD COMPUTING	MORALES GONZALES, RUSO ALE	IS-821	15.80	19.40	NULL	NULL	18	ESTUDIANTE
62040	05	20132	A	2016	PROGRAMACION DE VIDEO JUEGOS I	MORALES GONZALES, RUSO ALE	IS-724	10.90	0.00	NULL	NULL	5	ESTUDIANTE
62040	05	20132	A	2016	PROGRAMACION DE DISPOSITIVOS M	CLARES PERCA, JUAN CARLOS	IS-623	8.00	5.00	NULL	NULL	7	ESTUDIANTE
62040	05	20132	A	2016	SISTEMAS DISTRIBUIDOS	COAGUILA MAMANI, MARIBEL E	IS-521	1.00	0.00	NULL	NULL	1	ESTUDIANTE
62040	03	20202	A	2016	TOPICOS AVANZADOS I	LIMACHE FLORES, ALBERTO	IS-927	15.10	12.20	NULL	NULL	14	ESTUDIANTE
62040	03	20202	A	2016	ROBOTICA II	VERA RAMIREZ, OSCAR JOHN	IS-926	15.80	16.50	NULL	NULL	16	ESTUDIANTE
62040	03	20202	A	2016	TOPICOS AVANZADOS I	LIMACHE FLORES, ALBERTO	IS-927	14.40	5.10	NULL	NULL	12	ESTUDIANTE
62040	03	20202	A	2016	ROBOTICA II	VERA RAMIREZ, OSCAR JOHN	IS-926	17.40	18.30	NULL	NULL	18	ESTUDIANTE
62040	03	20202	A	2016	GESTION DE PROYECTOS I	FLORES GUTIERREZ, VANEZA	IS-925	13.70	10.70	NULL	NULL	12	ESTUDIANTE
62040	08	20201	A	2016	REDES II	COAGUILA MAMANI, MARIBEL E	IS-826	16.00	16.00	NULL	NULL	16	ESTUDIANTE
62040	08	20201	A	2016	PROCESAMIENTO DE IMAGENES Y VI	MORALES GONZALES, RUSO ALE	IS-822	18.40	19.00	NULL	NULL	19	ESTUDIANTE
62040	05	20131	A	2016	SISTEMAS DISTRIBUIDOS	NULL	IS-521	NULL	NULL	NULL	NULL	15	ESTUDIANTE
62040	06	20201	A	2016	APLICACIONES WEB II	HUERTAS FLORES, EDSON BUEN	IS-624	0.00	0.00	NULL	NULL	0	ESTUDIANTE
62040	06	20201	A	2016	PROGRAMACION DE DISPOSITIVOS M	CLARES PERCA, JUAN CARLOS	IS-623	0.00	0.00	NULL	NULL	0	ESTUDIANTE
52040	02	20132	A	2008	MATEMATICA DISCRETA	RAMOS RIVERA, SALOMON REY	ISMDI03	0.00	0.00	NULL	NULL	0	RESERVA
52040	02	20132	A	2008	ANALISIS MATEMATICO I	TAMAYO MAMANI, DAVID	EGAMT02	0.00	0.00	NULL	NULL	0	RESERVA
52040	07	20202	A	2016	ANALISIS Y DISEÑO DE SISTEMAS I	GUZMAN VALDIVIA, JOSE ANTON	IS-422	2.80	0.00	NULL	NULL	1	SANCIONADO
52040	06	20201	A	2016	FUNDAMENTOS DE SISTEMAS DE INF	GAUNA CHINO, MARIO	IS-323	NULL	NULL	NULL	NULL	12	SANCIONADO
52040	10	20132	A	2008	PLANEAMIENTO DE SISTEMAS DE INF	GAUNA CHINO, MARIO	ISPSIEL	15.20	14.10	NULL	NULL	15	ESTUDIANTE
52040	09	20131	A	2008	FORMULACION Y EVALUACION DE PF	NULL	ISFEP09	NULL	NULL	NULL	NULL	15	ESTUDIANTE
42040	09	20131	A	2008	COMERCIO ELECTRONICO	NULL	ISCOEEL	NULL	NULL	NULL	NULL	13	ESTUDIANTE
42040	08	20130	A	2008	INGENIERIA DE LA INFORMACION	NULL	ISIDI08	NULL	NULL	NULL	NULL	12	ESTUDIANTE
42040	06	20132	A	2008	SISTEMAS OPERATIVOS	YANA MAMANI, VICTOR	ISSOP06	7.50	0.00	NULL	NULL	4	SANCIONADO
32040	08	20201	A	2016	ROBOTICA I	VERA RAMIREZ, OSCAR JOHN	IS-827	10.30	8.00	NULL	NULL	12	ESTUDIANTE
02040	05	20191	A	2016	BASE DE DATOS II	NULL	IS-523	NULL	NULL	NULL	NULL	1	ESTUDIANTE

Nota. Estos datos no procesados son de los alumnos del ciclo académico 2019-I, 2019-II, 2020-I y 2020-II corresponden a la parte académica. Estos datos van a ser presentados a través de los cuadros estadísticos para luego analizarlos y utilizando las técnicas de Inteligencia artificial.

Figura 14

Datos no procesados de la ficha socio-económico, caso de estudio año 2019, 2020



DNI	CODI	GENERO	Pair	PROVINCIA	INGRESOS	Hrs_Trab	Celular	PC	Laptop	Internet	Tablet	Duerme	Limit_mi	Alergias	EsSalud	SIS	Futbol	Voley
1 047499	20194007	F		ILO	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	SI	NO	NO	NO
2 403129	201 4042	M		NULL	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
3 421522	201 4047	M		ILO	SININGRESO	0	SI	SI	SI	NO	NO	NO	NO	NO	NO	NO	NO	NO
4 446423	201 4044	M		ILO	951.00 - 1425.00	6	SI	SI	NO	SI	NO	SI	NO	NO	NO	SI	SI	NO
5 447474	201 4006	M	PERU	AZANGARO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
6 447523	201 4048	M	PERU	ILO	476.00 - 950.00	4	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	NO
7 449095	201 4045	F		ILO	MENOS DE 475.00	0	SI	NO	NO	SI	NO	SI	NO	Alpalva	NO	NO	NO	SI
8 451455	201 4057	M		ILO	476.00 - 950.00	8	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
9 452511	201 1033	M		ILO	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
10 463905	201 4054	M		ILO	MENOS DE 475.00	20	SI	NO	SI	NO	NO	SI	NO	Rinitis	NO	SI	NO	NO
11 465517	201 4049	M	PERU	MARISCAL NIETO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	NO
12 468114	201 4025	M	PERU	ILO	476.00 - 950.00	7	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
13 470290	201 4008	M		ILO	476.00 - 950.00	10	SI	NO	NO	NO	NO	SI	NO	NO	NO	SI	SI	SI
14 470717	201 4048	M	PERU	MARISCAL NIETO	mas de 1425.00	8	SI	SI	SI	SI	NO	NO	NO	NO	NO	SI	NO	NO
15 470808	201 4063	M		ILO	951.00 - 1425.00	8	SI	NO	SI	NO	NO	NO	NO	NO	NO	SI	NO	NO
16 471232	201 4041	F	PERU	MARISCAL NIETO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	NO
17 475404	201 4042	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
18 475922	201 4017	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
19 475967	201 4024	M	PERU	PUNO	MENOS DE 475.00	6	SI	SI	NO	NO	NO	SI	NO	NO	NO	SI	SI	NO
20 480776	201 4054	F	PERU	ILO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	SI
21 482270	201 4021	M	PERU	ILO	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	Al calor	NO	SI	NO	NO
22 482398	201 4010	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
23 484910	201 4010	M		ILO	MENOS DE 475.00	8	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	SI	NO
24 485163	201 4026	M	PERU	ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
25 701629	201 4044	M		NULL	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
26 701849	201 4005	M	PERU	MARISCAL NIETO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
27 701921	201 4001	F		ILO	SININGRESO	0	SI	SI	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
28 701963	201 4021	F		ILO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	NO
29 702008	201 4024	M		NULL	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
30 702116	201 4064	M	PERU	ILO	SININGRESO	0	SI	NO	NO	SI	NO	SI	NO	Polvo	NO	SI	NO	NO
31 702123	201 4024	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
32 702123	201 4033	M		ILO	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	SI	NO
33 702123	201 4005	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
34 703107	201 4040	M	PERU	ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
35 704449	201 4028	M		NULL	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	SI	NO
36 704449	201 4002	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
37 706034	201 4024	M		NULL	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	NO	NO	NO
38 708616	201 4012	M		NULL	MENOS DE 475.00	2	SI	NO	NO	NO	NO	SI	NO	NO	NO	SI	SI	NO
39 709720	201 4007	M		NULL	SININGRESO	0	SI	NO	NO	NO	NO	NO	NO	Polvo	NO	SI	SI	NO
40 710045	201 4056	M		NULL	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
41 710628	201 4001	M		ILO	MENOS DE 475.00	4	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	SI	NO
42 710634	201 4006	M	PERU	ILO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	NO	NO
43 710877	201 4061	F		ILO	SININGRESO	0	SI	SI	SI	SI	NO	SI	NO	NO	NO	SI	NO	SI
44 712443	201 4014	M	PERU	ILO	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
45 712477	201 4013	M		ILO	SININGRESO	0	SI	NO	NO	SI	NO	SI	NO	ciertas qui	SI	NO	NO	NO
46 714025	201 4031	F		NULL	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
47 714526	201 4018	F		NULL	SININGRESO	0	SI	SI	NO	NO	NO	SI	NO	NO	NO	SI	NO	NO
48 715755	201 4023	F	PERU	AZANGARO	SININGRESO	0	SI	NO	NO	NO	NO	SI	NO	NO	NO	SI	NO	SI
49 715919	201 4006	F	PERU	ILO	SININGRESO	0	NO	NO	NO	NO	NO	SI	NO	NO	NO	NO	NO	NO
50 715919	201 4023	M	PERU	ILO	MENOS DE 475.00	12	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	SI	NO
51 716217	201 4035	F		ILO	SININGRESO	0	SI	SI	NO	SI	NO	SI	NO	NO	NO	SI	NO	NO
52 716624	201 4049	F		ILO	SININGRESO	0	SI	NO	SI	NO	NO	SI	NO	NO	NO	SI	SI	SI
53 717241	201 4006	F		ILO	SININGRESO	0	SI	SI	NO	NO	NO	SI	NO	NO	NO	SI	NO	NO
54 717552	201 4047	M		ILO	MENOS DE 475.00	8	SI	NO	NO	NO	NO	NO	NO	NO	NO	SI	NO	NO
55 719615	201 4054	F	PERU	ISLAY	SININGRESO	0	SI	NO	NO	NO	NO	SI	NO	NO	NO	SI	SI	SI
56 720093	201 4020	M		NULL	476.00 - 950.00	8	SI	NO	SI	NO	NO	SI	NO	Alajunas	NO	SI	NO	NO
57 720215	201 4045	M		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO
58 720215	201 4024	M		NULL	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	NO	NO
59 720266	201 4059	M		ILO	SININGRESO	0	SI	SI	SI	SI	NO	SI	NO	NO	SI	NO	NO	NO
60 720266	201 4024	M		NULL	SININGRESO	0	SI	NO	SI	SI	NO	SI	NO	NO	NO	SI	SI	SI
61 721044	20194019	F		ILO	SININGRESO	0	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO

20191 | 20192 | 20201 | 20202

Nota. Estos datos no procesados son de los alumnos que llenaron la ficha socio-económica en los años académicos 2019-I, 2019-II, 2020-I y 2020-II Estos datos van a ser presentados a través de los cuadros estadísticos para luego analizarlos y utilizando las técnicas de Inteligencia artificial.

4.2.2 Datos seleccionados

Se realiza una descripción de los datos que proporcionó la Universidad Nacional de Moquegua, previa gestión con la oficina de Dirección de Actividades y Servicios Académicos, Dirección de Bienestar Universitario. Los datos corresponden a los alumnos de Ingeniería de Sistemas e Informática año de matrícula 2019 – 2020.

La información fue proporcionada en Libros de Excel con diferentes hojas de cálculo, información que sólo pudo ser compartida:

La información general del alumno como su nombre completo, código de estudiante, ciclo que cursa, control de ciclo académico en la cual se ha matriculado, sexo, sección.

El registro del Año y semestre académico que corresponde la malla curricular, los nombres de todos los docentes según el ciclo que cursa, el registro de todos sus cursos, código curso, las notas de las unidades según el curso.

El registro de los cursos con su nota final donde podemos ver lo que obtuvo el alumno en sus asignaturas, además del número de matrícula que corresponde, el año y semestre en que lo llevó ahí podemos verificar si el alumno es regular o irregular, si aprobó o desaprobó los cursos en los ciclos matriculados, aunque también se visualiza, si solicitó reserva o está sancionado, según el número de matrículas que tuvo con un curso. En la columna estado observamos la condición del estudiante al finalizar el curso.

4.3 Preparación de la DATA

4.3.1 Estructura de data

En las variables independientes tenemos muchos datos categóricos, es por eso que necesariamente se debe codificar las etiquetas para convertir a valores numéricas (binarios 0 y 1). Con la función *sklearn. preprocessing. LabelEncoder*. el método *fit* lo entrena (creando el mapeado entre las etiquetas y los números) y el método *transform* transforma las etiquetas que se incluyan como argumento en los números correspondientes. El método *fit_transform* realiza ambas acciones simultáneamente.

Figura 15

Descripción del dataset

Out[4]:

	sexo	pa	pd	ingreso	trabaja	pclaptop	internet	tablet	discapacidad	alergias	clases
0	1	13.0	10.0	475	1	2	2	2	2	2	BUENREND
1	2	14.0	0.0	0	2	1	1	2	2	1	BUENREND
2	1	13.3	10.0	0	2	1	1	2	2	2	BUENREND
3	1	12.9	0.0	0	2	2	2	2	2	2	BUENREND
4	1	14.9	0.0	0	2	1	1	2	2	2	BUENREND

Nota. Se presenta los datos del dataset.

Figura 16

Información de la data

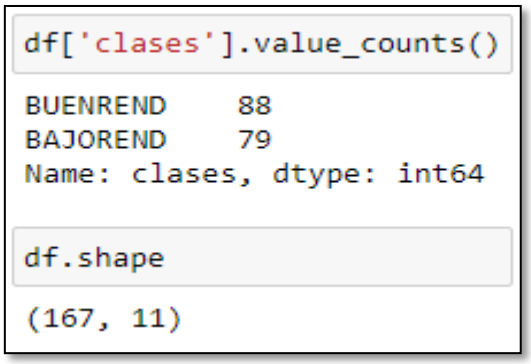
```
dataset.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   sexo          167 non-null    int64
 1   pa            167 non-null    float64
 2   pd            167 non-null    float64
 3   ingreso       167 non-null    int64
 4   trabaja       167 non-null    int64
 5   pclaptop      167 non-null    int64
 6   internet      167 non-null    int64
 7   tablet        167 non-null    int64
 8   discapacidad  167 non-null    int64
 9   alergias      167 non-null    int64
 10  clases        167 non-null    object
dtypes: float64(2), int64(8), object(1)
memory usage: 14.5+ KB
```

Nota. Información de la DataFrame

Figura 17

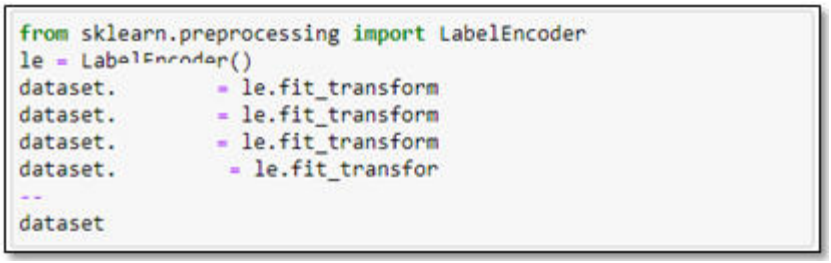
Información del feature clases



Nota. Información del feature, la cantidad total de cada clase. Shape (), Devuelve el número de elementos del DataFrame.

Figura 18

Librería scikit-learn



Nota. Se utiliza el LabelEncoder y el método fit_transform en todos los datos categóricos.

Figura 19

Data procesada

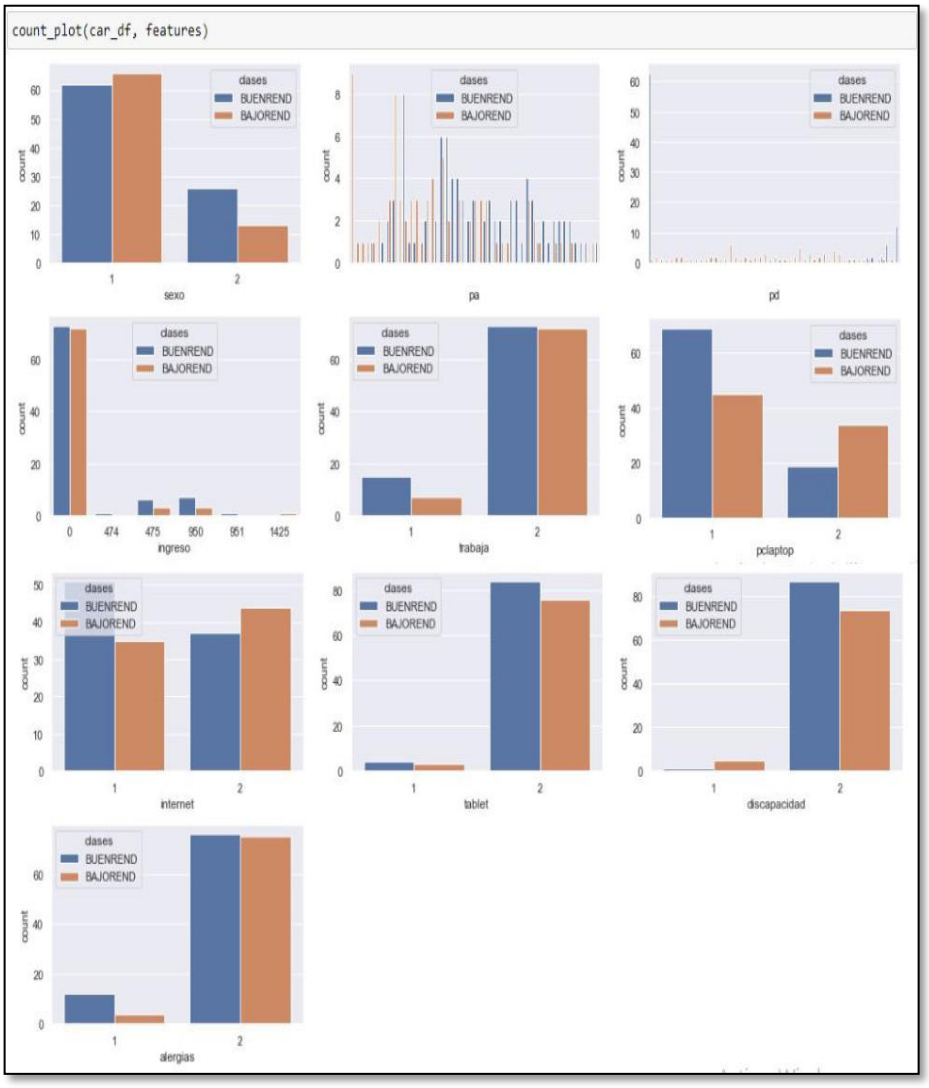
	sexo	pa	pd	ingreso	trabaja	pclaptop	internet	tablet	discapacidad	alergias	clases
0	1	13.0	10.0	475	1	2	2	2	2	2	1
1	2	14.0	0.0	0	2	1	1	2	2	1	1
2	1	13.3	10.0	0	2	1	1	2	2	2	1
3	1	12.9	0.0	0	2	2	2	2	2	2	1
4	1	14.9	0.0	0	2	1	1	2	2	2	1
...	...	...	...	...	...	...	...	...	...	...	...
162	2	11.8	10.0	0	2	1	2	2	2	2	1
163	2	12.3	0.0	951	1	1	2	2	2	2	1
164	2	13.0	0.9	0	2	2	2	2	2	2	0
165	1	12.0	6.0	0	2	1	1	2	2	2	0
166	1	15.3	9.0	475	1	1	1	2	2	2	1

167 rows × 11 columns

Nota. Data procesada para introducirla a los algoritmos para su entrenamiento.

Figura 20

Visualización de los valores de cada feature.



Nota. Visualizamos los datos. Los valores atípicos en las variables categóricas.

4.4 Modelado

Procesamiento de Datos utilizando Algoritmos (Técnicas de Inteligencia artificial)

Los datos que se utilizan son datos históricos de acuerdo a la delimitación espacial se tomó de los años 2019 – 2020, base de datos socioeconómicos, académicos con una muestra de 167 estudiantes de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua.

Se realizó una base de datos procesados de toda la data obtenida, creando matrices de correlación de 10 atributos y 167 instancias denominado “ubdepisi.csv”.

Figura 21

Base de datos caso EPISI – UNAM

sexo	cs	total	cca	ccd	pa	pd	nota	hra_	celu	pc	lapt	inter	tabl	desa	disc	aler
M	1	7	6	1	13	10	12	7	SI	NO	NO	NO	NO	SI	NO	NO
F	1	7	7	0	14	0	14	0	SI	NO	SI	SI	NO	SI	NO	GAT
M	1	7	6	1	13	10	12	0	SI	SI	SI	SI	NO	SI	NO	NO
M	1	7	7	0	13	0	13	0	NO	NO	NO	NO	NO	NO	NO	NO
M	1	7	7	0	15	0	15	0	SI	NO	SI	SI	NO	SI	NO	NO
M	1	7	7	0	15	0	15	0	SI	SI	SI	NO	NO	SI	NO	Ale
M	1	7	7	0	13	0	13	0	SI	NO	SI	SI	NO	SI	NO	NO
F	1	7	7	0	13	0	13	0	SI	SI	NO	NO	NO	SI	NO	NO
M	1	7	7	0	12	0	12	0	SI	SI	NO	SI	NO	SI	SI	NO
M	1	7	7	0	13	0	13	0	SI	NO	SI	SI	SI	SI	NO	NO
M	1	7	4	3	13	8.7	11	0	SI	NO	SI	NO	NO	SI	NO	NO
M	1	7	6	1	13	5	8.9	0	NO	NO	SI	SI	NO	NO	SI	NO
M	1	7	7	0	13	0	13	0	SI	NO	SI	NO	NO	SI	NO	NO
M	1	7	7	0	15	0	15	0	SI	NO	SI	NO	NO	SI	NO	Por
M	1	7	7	0	12	0	12	0	NO	NO	NO	NO	NO	SI	NO	NO
M	1	7	7	0	12	0	12	2	SI	NO	SI	SI	NO	SI	NO	NO
M	1	7	7	0	13	0	13	0	SI	SI	SI	NO	NO	SI	NO	NO
M	1	7	7	0	15	0	15	0	SI	NO	SI	NO	NO	SI	NO	NO
M	1	7	7	0	14	0	14	8	SI	SI	NO	SI	NO	SI	NO	NO
M	1	7	7	0	14	0	14	0	SI	NO	SI	SI	NO	SI	NO	NO
F	1	7	7	0	15	0	15	0	SI	SI	NO	SI	NO	SI	NO	NO
M	1	7	7	0	13	0	13	0	SI	SI	SI	SI	NO	SI	NO	NO
M	1	7	7	0	14	0	14	0	NO	NO	NO	SI	NO	SI	NO	NO
F	1	7	7	0	13	0	13	0	SI	SI	SI	SI	NO	SI	NO	NO
M	1	7	7	0	14	0	14	0	NO	NO	NO	NO	NO	SI	NO	NO
F	1	7	7	0	15	0	15	0	SI	SI	SI	SI	NO	SI	NO	NO
M	1	7	7	0	13	0	13	0	SI	NO	NO	SI	NO	SI	NO	NO
M	1	7	7	0	13	0	13	0	SI	SI	NO	NO	NO	SI	NO	Into
M	1	7	7	0	15	0	15	0	SI	NO	SI	SI	NO	SI	NO	NO

Nota. Base de datos donde necesariamente debe contener la extensión “.csv” para no tener problemas de lectura al momento de procesar la información.

Utilizando la metodología CRISP-DM se desarrolló las diferentes etapas: conocer el negocio, comprensión de la data que se recopiló, el filtrado de las diferentes bases de datos y finalmente la evaluación e interpretación de los resultados.

4.4.1 Datos para entrenamiento

Los datos deben ser divididos, tanto para entrenamiento (*training data*) y datos de prueba, validación (*testing data*). Del dataset realizamos la división 75% para training y 25% Testing se puede variar, pero en este caso estoy considerando este porcentaje para todos los algoritmos.

Los datos de entrenamiento se utilizaron para asegurar que el algoritmo reconozca patrones, con los datos de prueba el algoritmo nos puede predecir nuevas respuestas basadas en su entrenamiento. Los datos de validación cruzada se utilizan para mayor precisión y eficiencia del algoritmo utilizado para entrenar la máquina. Con los datos de validación y prueba debemos tener cuidado ya que la máquina va a botar resultados; por ejemplo, si el grado de precisión ronda por el 50% no tiene mucha utilidad, pero si el resultado alcanza o supera el 90%, nos ofrece un alto nivel de confianza.

Figura 22

División de los datos de training y testing

```
#conjunto de datos en 75% (Entrenamiento) y 25% (Test) utilizando la función train_test_split
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.25, random_state = 0)
print ('Número de filas en el conjunto total: {}'.format(dataset.shape [0]))
print ('Número de filas en el conjunto de entrenamiento: {}'. format (x_train.shape [0]))
print ('Número de filas en el conjunto de prueba: {}'. format (x_test.shape [0]))
```

```
Número de filas en el conjunto total: 167
Número de filas en el conjunto de entrenamiento: 125
Número de filas en el conjunto de prueba: 42
```

Nota. Como podemos observar se ha dividido los datos con el 75% de training y 25% testing

Una vez definido, vamos a proceder a explicar nuestros algoritmos y las técnicas que se va a utilizar:

4.4.2 Algoritmo Decisión tree

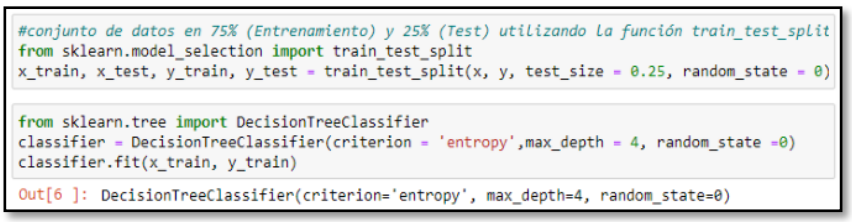
En este algoritmo una vez que las variables están disponibles va a generar un árbol de decisión (Alberto, 2019), según la base de datos de 167 instancias y 10 atributos nos indica el algoritmo:

4.4.2.1 Construcción del modelo Árbol de decision

Para construir el modelo primero debemos entrenar los datos utilizando la librería y sus módulos (from sklearn. tree import DecisionTreeClassifier),

Figura 23

Entrenamiento del árbol



Nota. Tenemos que entrenar los datos para el árbol de decisión

4.4.2.2 Importando árbol

Se tuvo que importar el módulo de graphviz, la cual se observa el valor y la instancia que es la variable dependiente. Es fundamental explicar el resultado que estamos obteniendo, donde se hace la selección de las variables por medio de la entropía.

Figura 24

Estructura del árbol creado

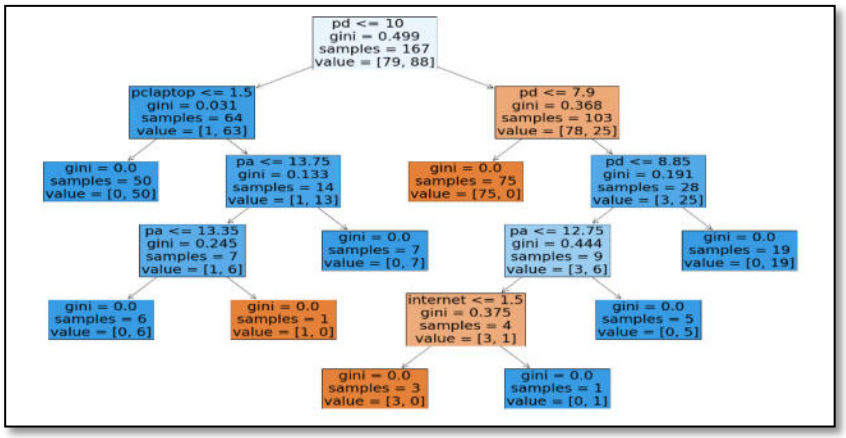
```
print(f"Profundidad del árbol: {modelo.get_depth()}")
print(f"Número de nodos terminales: {modelo.get_n_leaves

plot = plot_tree(
            decision_tree = modelo,
            feature_names = labels.tolist(),
            class_names   = 'clases',
            filled        = True,
            impurity      = False,
            fontsize      = 10,
            ax            = ax
        )
```

Nota. Se muestra la estructura de la creación del árbol

Figura 25

Árbol de decisión



Nota. Se muestra el árbol de decision

4.4.2.3 Predicción del algoritmo

Este algoritmo ha sido entrenado usando el conjunto de datos de entrenamiento, para hacer algunas predicciones en los datos de prueba almacenados para eso se utilizó *sklearn. Metrics*

Figura 26

Predicción de los datos de testing

```
from sklearn.metrics import accuracy_score,confusion_matrix,cl
print("Training accuracy", accuracy_score(y_train,train_predic
#
print("Testing accuracy", accuracy_score(y_test,test_predict))
```

```
Training accuracy 0.947
Testing accuracy 0.947
```

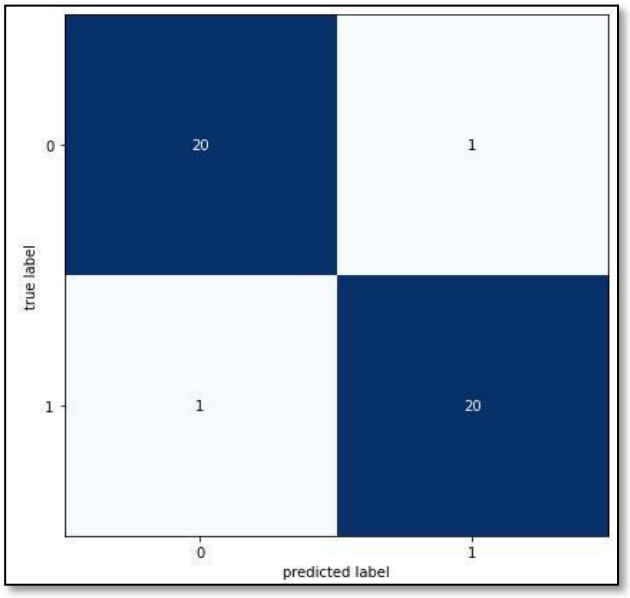
Nota. Predicciones del conjunto de datos de entrenamiento

4.4.2.4 Evaluación del modelo

La matriz de confusión recoge el conjunto de posibilidades entre la clase correcta de un evento, y su predicción. Se trabajo con los módulos de (*from sklearn. metrics import confusion_matrix*) para obtener el resultado.

Figura 27

Matriz de confusión del algoritmo decisión tree



Nota. Se muestra la matriz de confusión

Interpretación de la validación cruzada

Una vez que la matriz realizó la confusión de variables se puede observar que la diagonal está en (20,20) corresponden a los valores estimados de forma correcta por el modelo tanto los verdaderos positivos como los verdaderos negativos. La otra diagonal representa los casos en los que el modelo se ha equivocado. (1 falsos negativos ,1 falsos positivo).

Una forma de validar el modelo, debemos de realizar la exactitud (*accuracy*) de las predicciones sin son correctas, la sensibilidad (*recall*), el análisis de curva (roc), y la puntuación (F1 score).

Figura 28

Métricas del algoritmo decisión tree

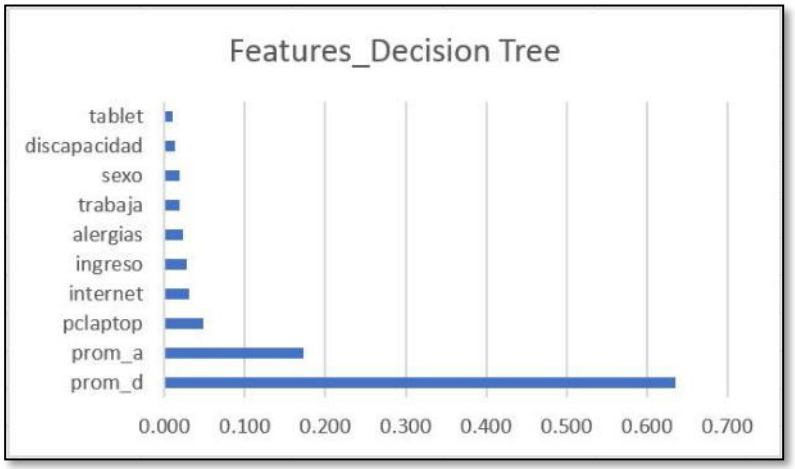
```
#accuracy
from sklearn import metrics
print("Accuracy: ", metrics.accuracy_score(y_test, y_pred
print("F1 Score: ", metrics.f1_score(y_test, y_pred, aver
print("ROC: ", metrics.roc_auc_score(y_test, y_pred))
print("Recall: ", metrics.recall_score(y_test, y_pred, av

Accuracy:  0.9523809523809523
F1 Score:  0.9523809523809523
ROC:  0.9523809523809522
Recall:  0.9523809523809523
```

Nota. Se muestra la métrica del algoritmo Decision Tree accuracy, F1 Score, ROC y Recall

Figura 29

Grupo de variables del algoritmo decisión tree



Nota. Se muestra la variable más importante

4.4.3 Algoritmo RandomForest

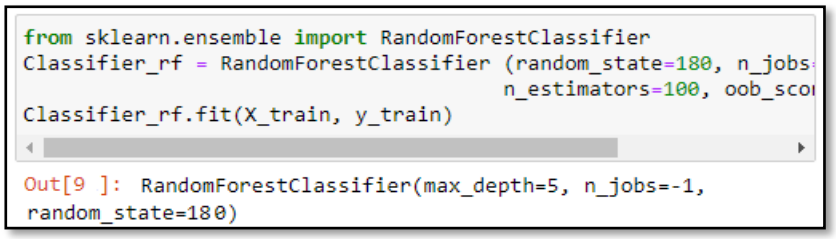
En este algoritmo de árboles aleatorios (Alberto, 2019) se importó la data, los datos categóricos se convirtió en datos binarios, consta de 167 instancias y 10 columnas, para ello se divide los datos en atributos y etiquetas, y luego se divide en conjunto de entrenamiento y prueba.

4.4.3.1 Construcción de RandomForest

Nuestro conjunto de datos, vamos a entrenarlo con el algoritmo de bosque aleatorio. Se va a configurar el (n_estimators), donde se define la cantidad de árboles en el bosque aleatorio donde se colocó 100. En (random_state), se colocó 0 y en (min_samples_leaf). Se considera la cantidad de hojas de muestra que queremos.

Figura 30

Creación del modelo bosque aleatorios



Nota. Se configuró el algoritmo Random Forest

4.4.3.2 Evaluación de RandomForest

La matriz de confusión recoge el conjunto de posibilidades entre la clase correcta de un evento, y su predicción. Se trabajo con los módulos de (*from sklearn. metrics import confusion_matrix*) para obtener el resultado.

Figura 31

Configuración de la matriz de confusión randomforest

```
from mlxtend.plotting import plot_confusion_matrix
from sklearn.metrics import confusion_matrix

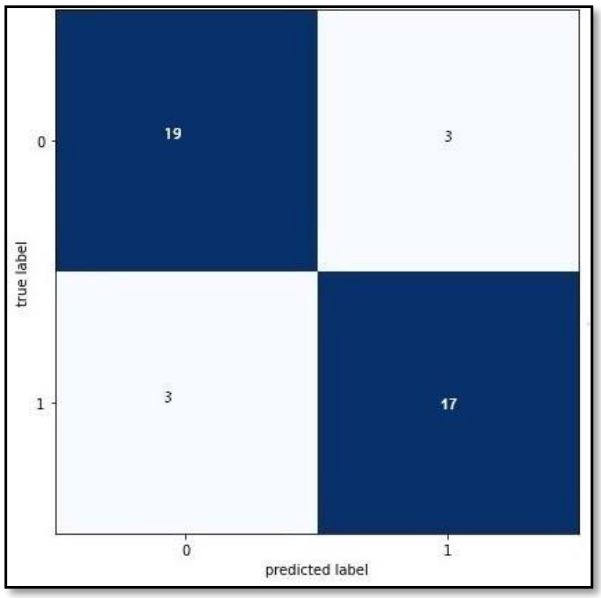
# Predicción del modelo usando los datos de prueba
y_pred = BA_model.predict(X_test)
matriz = confusion_matrix(y_test,y_pred)

plot_confusion_matrix(conf_mat=matriz, figsize=(6,6), show_normed=False)
plt.tight_layout()
```

Nota. Se configuró la matriz de algoritmo Random Forest

Figura 32

Matriz de confusión del algoritmo randomforest



Nota. La matriz de algoritmo Random Forest

Interpretación de la validación cruzada

Una vez que la matriz realizó la confusión de variables se puede observar que la diagonal está en (19,17) se corresponden con los valores estimados de forma correcta por el modelo tanto los verdaderos positivos, como los verdaderos negativos. La otra diagonal representa los casos en los que el modelo «se ha equivocado (3 falso negativos, 3 falso positivo).

Una forma de validar el modelo, debemos de realizar la exactitud (accuracy) de las predicciones sin son correctas, la sensibilidad (recall), el análisis de curva (roc), y la puntuación (F1 score)

Figura 33

Validación del algoritmo randomforest

	precision	recall	f1-score	support
0	0.90	0.74	0.81	19
1	0.79	0.92	0.85	17
accuracy			0.83	42
macro avg	0.84	0.83	0.83	42
weighted avg	0.84	0.83	0.83	42

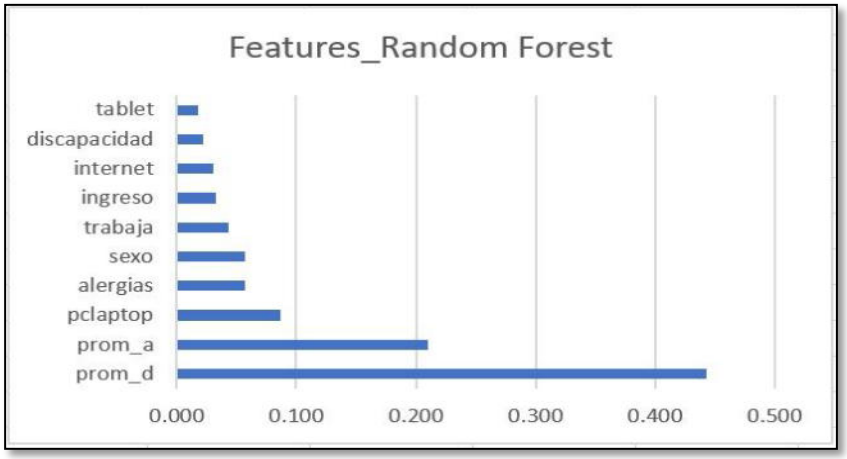
0.8333333333333334

Nota. Se visualiza la accuracy de 0.83, de los 42 datos de prueba.

La precisión alcanzada por nuestro clasificador de bosques aleatorio con 100 árboles es del 83%, cambiar el número de estimadores para este problema no mejoró significativamente los resultados.

Figura 34

Características más influyentes según, randomforest



Nota. Se visualiza la feature_importances, en la cual nos indica la feature_prom_d que es la más influyente y la decisión final se visualiza en la variable dependiente.

4.4.4 Algoritmo Naive Bayes

En este algoritmo de Naive Bayes, (Gutiérrez & Real, 2017) se importó la data, los datos categóricos se convirtió en datos binarios, consta de 167 instancias y 10 columnas, para ello se divide los datos en atributos y etiquetas, y luego se divide en conjunto de entrenamiento y prueba.

4.4.4.1 Construcción de Naive Bayes

Figura 35

Importando librerías y módulos de naive bayes

```
import pandas as pd
import numpy as np
import seaborn as sns
# Diferentes tipos de clasificadores Naive Bayes
from sklearn.naive_bayes import GaussianNB
from sklearn.naive_bayes import MultinomialNB
import matplotlib.pyplot as plt

from matplotlib import colors
%matplotlib inline
plt.rcParams['figure.figsize'] = (16,9)
plt.style.use('ggplot')

from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.preprocessing import OrdinalEncoder
```

Nota. implementación en la librería Sci-sklearn

Figura 36

Creación naive bayes import gaussianNB

```
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(X_train, y_train)

GaussianNB()
```

Nota. Naive Bayes Gaussian para datos continuos

Figura 37

Categorías de codificación, utilizando labelEncoder

	sexo	pa	pd	ingreso	trabaja	pclaptop	internet	tablet	discapacidad	alergias	clases
0	0.0	17.0	50.0	2.0	0.0	1.0	1.0	1.0	1.0	1.0	1
1	1.0	27.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	1
2	0.0	20.0	50.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1
3	0.0	16.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1
4	0.0	36.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	1.0	1

Nota: Utilizando le.fit_transform, se prepara los datos categóricos.

4.4.4.2 Evaluación de Naive Bayes

La matriz de confusión recoge el conjunto de posibilidades entre la clase correcta de un evento, y su predicción. Se trabajó con los módulos de (from sklearn. metrics import confusion_matrix) para obtener el resultado.

Figura 38

Métricas de desempeño, utilizando GaussianNB

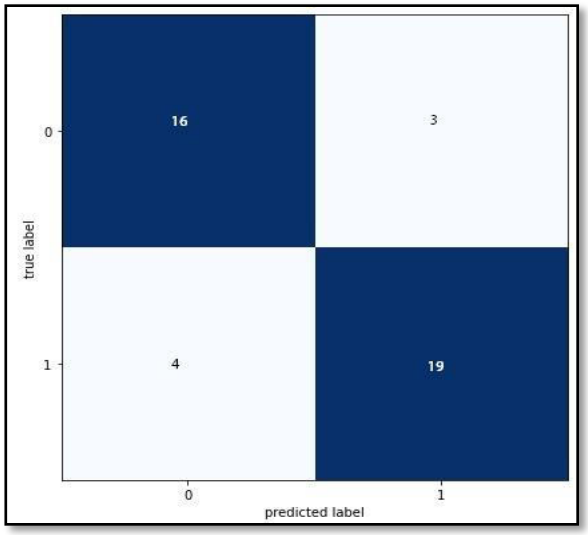
```
print("GaussianNB")
print("=" * 30)
print('muestras mal clasificadas: {}'.format(count_misclassified))
accuracy = accuracy_score(y_test, y_pred_gnb)
print('Accuracy: {:.2f}'.format(accuracy))
```

```
GaussianNB
==============================
muestras mal clasificadas: 7
Accuracy: 0.83
```

Nota: Utilizando GaussianNB, vemos que son 7 muestras mal clasificadas.

Figura 39

Configuración de la matriz de confusión de naive bayes



Nota. La matriz de algoritmo Naive Bayes

Interpretación de la validación cruzada

En la Matriz de confusión, se visualiza la clase verdadero o falso, igualmente en la predicción. Una vez que la matriz realizó la confusión de variables se puede observar que la diagonal está en (16,19) corresponden con los valores estimados de forma correcta por el modelo, los verdaderos positivos como los verdaderos negativos. La otra diagonal representa los casos en los que el modelo se ha equivocado (4 falso negativos, 3 falso positivo).

Figura 40

Métricas del algoritmo de naive bayes

```
from sklearn import metrics
print("Accuracy: ", metrics.accuracy_score(y_test, y_pred))
print("Recall score : ", recall_score(y_test, y_pred_gnb , average='micro'))
print("Precision score : ",precision_score(y_test, y_pred_gnb , average='micro'))
print("F1 score : ",f1_score(y_test, y_pred_gnb , average='micro'))
```

```
Accuracy:  0.8333333333333334
Recall score :  0.8333333333333334
Precision score :  0.8333333333333334
F1 score :  0.8333333333333334
```

Nota. Métrica de recall_score, precision_score y F1_score del Algoritmo Naive Bayes

Figura 41

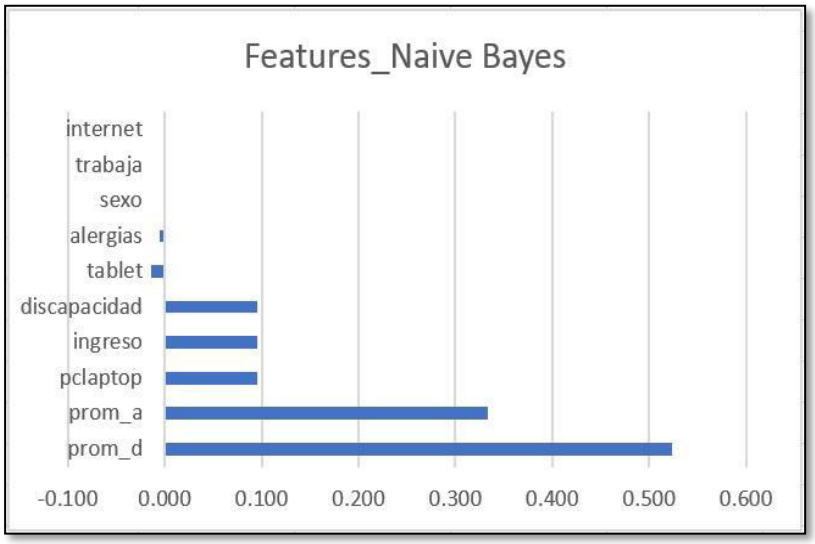
Precisión del algoritmo de naive bayes

	precision	recall	f1-score	support
0	0.80	0.84	0.82	19
1	0.86	0.83	0.84	23
accuracy			0.83	42
macro avg	0.83	0.83	0.83	42
weighted avg	0.83	0.83	0.83	42

Nota. Precisiones del algoritmo Naive Bayes

Figura 42

Características más influyentes de naive bayes



Nota. Se visualiza la feature_importances, en la cual nos indica la feature_prom_d que es la más influyente.

4.4.5 Resultados del Dataset

En estos resultados conoceremos las variables más influyentes (*features_importances*), en los algoritmos utilizados. Para realizar el entrenamiento, el conjunto de datos se divide en dos partes el 75% en training y 25% en testing, como se muestra en la figura 22.

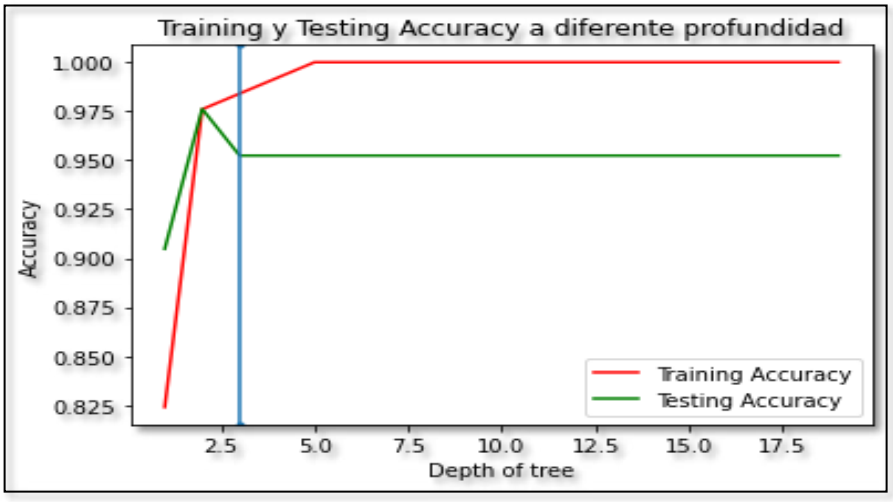
El modelo de mejor desempeño es el algoritmo Decision Tree, con un

promedio de precisión de 95% en el procedimiento de validación cruzada.

El algoritmo a considerado feature_important como prom_d, prom_a, pclaptop. Como se muestra en la figura 42.

Figura 43

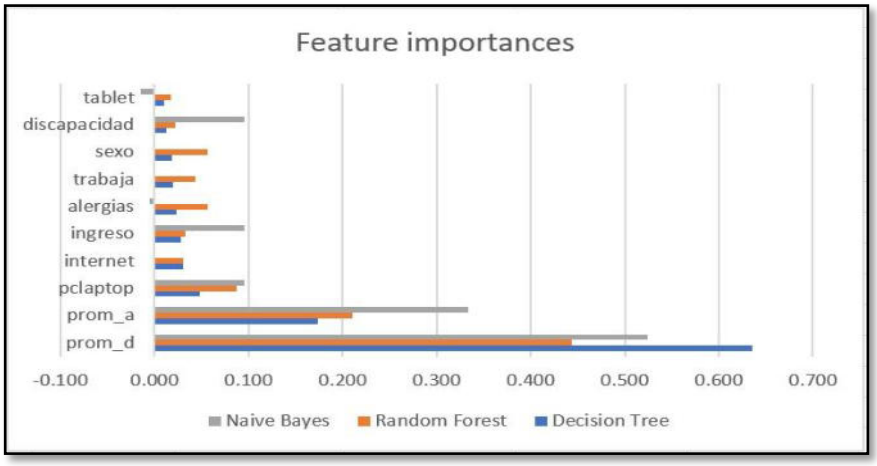
Diferencia entre datos de entrenamiento y prueba



Nota. Se visualiza la diferencia entre la exactitud de datos de entrenamiento y la exactitud de datos de prueba.

Figura 44

Características más importantes



Nota. La característica más influyente es prom_d en factor académico y pclaptop en socioeconómicos.

4.4.6 Evaluación de precisión de las técnicas usadas en la Investigación

En la investigación se utilizó la data ('ubdepisi.csv'), que consta datos de la ficha socioeconómica y datos académicos de los alumnos de Ingeniería de Sistemas e Informática UNAM. Se utilizó los algoritmos Naive Bayes, Random Forest y Decision Tree, de estas técnicas de inteligencia artificial se verá cual tuvo mejor precisión.

Tabla 05

Resultados de las técnicas usadas

Ítem	Técnica	Precisión
1	Algoritmo Decision Tree	95%
2	Redes Bayesianas algoritmo Naive Bayes.	83%
3	Método clasificador Random Forest	83%

Nota. Se observa la precisión de los algoritmos que se utilizaron, en el cual el algoritmo Naive Bayes tuvo la mayor precision.

Tabla 06

Resultados de las características influyentes de los algoritmos

Ítem	Naive Bayes		Random Forest		Algoritmo Decision Tree	
1	Prom_d	5%	Prom_d	4%	Prom_d	6%
2	Prom_a	3%	Prom_a	2%	Prom_a	2%
3	pclaptop	1%	pclaptop	1%	pclaptop	1%
4	ingreso	1%	internet	0.5%	internet	0.5%

Nota. En este cuadro pueden observar la las características más influyentes de los algoritmos usados

Tabla 07

Resultados de las validaciones de las técnicas usadas

Resumen	Decision Tree	Random Forest	Naive Bayes
Instancias correctamente clasificadas	40	36	35
Instancias clasificadas incorrectamente	2	6	7
Accuracy	95%	83%	83%
F1 Score	95%	83%	83%
ROC	95%	83%	83%
Recall	95%	83%	83%
N° total de instancias	167	167	167

Nota. En este cuadro pueden observar la validación de cada algoritmo y método utilizado en el estudio

4.5 Despliegue (Implementación)

En esta investigación de análisis académico caso de estudio Episi – Unam se implementa un modelo de análisis para la Escuela Profesional de Ingeniería de Sistemas e Informática, la implementación y aplicación queda a decisión de las autoridades.

VI. CONCLUSIONES

CONCLUSIÓN GENERAL:

Se logró implementar el modelo de análisis utilizando técnicas de inteligencia artificial para conocer las posibles causantes del bajo rendimiento académico de los estudiantes de la Escuela Profesional de Ingeniería de Sistemas e Informática de la UNAM.

Se utilizó técnicas inteligencia artificial (clasificadores) Random Forest, Decisión tree y Naive Bayes determinando así las causantes del bajo rendimiento académico, donde la variable más resaltante es Promedio desaprobado, pc-laptop.

.

CONCLUSIÓNES ESPECÍFICAS

CE1 Se logró seleccionar, preprocesar y conocer las características de los datos socioeconómicos y académicos de los estudiantes de la Universidad Nacional de Moquegua.

Se seleccionó los datos de la ficha socioeconómica y la base de datos académica de los ciclos 2019-I, 2019-II, 2020-I y 2020-II, luego se procesó los datos de la hoja de cálculo a una data procesada ".csv" para ser leído por Python a través de sus librerías pandas, numpy entre otros. Se analizó las 10 variables independientes que consta de 167 instancias.

CE2 Se logró desarrollar métodos de inteligencia artificial para descubrir patrones de bajo rendimiento académicos de los estudiantes. Caso EPISI UNAM.

Para descubrir los patrones de bajo rendimiento académico, se utilizaron técnicas de aprendizaje supervisado como el árbol de decisión, Naive Bayes, y Random Forest. Se analizaron las diferentes variables de la data ('ubdepisi.csv'), una vez seleccionado la data se entrena el modelo, se divide la data en 75% training y 25% testing. Luego visualizamos el modelo donde muestra las variables más influyentes (*Features_importances*), los datos deben estar limpios sin valores nulos o anómalos. Se calcula la precisión y finalmente se interpreta el modelo.

CE3 Se logró evaluar los resultados de los métodos de inteligencia artificial.

Una vez entrenado y adaptado los modelos según los algoritmos, se realizan las métricas de desempeño, donde se evalúa la accuracy, recall_score, precision_score, F1_score y sobretodo la matriz de confusión. El algoritmo que tuvo mejor precisión es Decision Tree 95%.

VII. RECOMENDACIONES

Tomar en cuenta el presente trabajo de investigación donde se analizó el bajo rendimiento académicos de los alumnos y donde muchas veces se da la deserción estudiantil.

Se utilizó diversos algoritmos con el fin de validar la información es recomendable aprovechar la información de este estudio que hizo a los alumnos matriculados en el 2019 – 2020 y extender el análisis a otros años.

Para ampliar la investigación con otros estudiantes, es recomendable que la data sea uniforme; quiere decir, que los datos deben ser de I ciclo a X ciclo.

VIII. REFERENCIAS

Abel, M. (22 de 01 de 2019). *plandemejora.com*. Obtenido de https://www.plandemejora.com/como-hacer-un-diagrama-de-arbol-de-decisiones/

Adrian, L. (24 de agosto de 2019). *profesionalesreview.com*. Obtenido de https://www.profesionalreview.com/2019/08/24/deep-learning-que-es-y-como-se-relaciona-con-el-machine-learning/

Alberca, a. S. (2020). Manual de Python. En (asalber@ceu.es), *Manual de python* (pág. 102). Paises bajos: Python.

Alberto, R. A.-B. (2019). Artificial neural network analysis of the academic performance. *EL servier*, 8.

Antonio, J. (24 de Diciembre de 2018). *Aprendizaje por refuerzo*. Obtenido de https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/

arimetrics. (2008). *arimetrivcs.com*. Obtenido de https://www.arimetrics.com/glosario-digital/modelo-predictivo

Babbie, E. R. (1996). *Manual para la práctica de la Investigación social - Diseño de Investigación.* Bilboa: Desclée De Brouwer.

Bembibre, C. (abril de 2009). *Tu diccionario hecho facil*. Obtenido de Definición ABC: https://www.definicionabc.com/social/academico.php

Berry., M. J. (2004). Data Mining Techniques: For Marketing, Sales, and Customer Relationship. Management Wiley. *Scielo*.

Breiman, L. (2001). Random Forests. Machine Learning. *DOI: 10.1023/A:1010933404324*, 45.

Breiman, L. F. (1984). Classification and regression trees. *Random Forests. Machine Learning*, 45.

Camus, A. (13 de setiembre de 2020). *Crehana*. Obtenido de https://www.crehana.com/pe/blog/desarrollo-web/que-es-el-machine-learning/

Chadwick, C. (1979). *Teorías del aprndizaje para el docente.* Dubai: Tecla.

Coloma, E. (17 de junio de 2014). *https://udep.edu.pe/.* Obtenido de https://udep.edu.pe/hoy/2014/06/y-si-no-me-gusta-trabajar-en-equipo/

CONEAU. (abril de 2008). *www.sineace.gob.pe/tag/coneau/.* Obtenido de http://www.une.edu.pe/acreditacion/Guia_Acreditacion_CONEAU.pdf: https://www.gob.pe/institucion/sineace/normas-legales/921847-040-2012-sineace-p

Cortina, V. G. (2015). *Aplicación de la metodología CRISP-DM a un proyecto de mineria de datos en el entorno universitario.* Madrid: Universidad Carlos III de Madrid Escuela Politécnica Superior de Ingeniería en Informática.

Covington, M. O. (1979). ¿Are causal attribution causal?: A path analysis of the cognitive model of chievement motivation. Journal of Personality and Social Psychology, 1487-1504. New YORK: Research on Motivation in Education Student Motivation.

Daniel, R. (20 de Mayo de 2019). *Analytics Lane*. Obtenido de Random Forest: https://www.analyticslane.com/2019/05/20/random-forest/

Dean., R. (2000). *Facultad de Ingeniería - U.N.R.C.* Obtenido de La Investigacion Tecnológica en las ciencias de la Ingenieria y la Innovacion Técnológica: https://www.unrc.edu.ar/publicar/23/dossidos.html

Debasish Sena - Naresh Kumar Nagwani. (2015). Application of Time Series Based Prediction Model to Forecast Per Capita Disposable Income. *paper - 2015 IEEE International Advance Computing Conference (IACC)*, https://ieeexplore.ieee.org/.

Denil, M. M. (2014). *Reducir brechas: bosques aleatorios en la teoria y la practica. Actas de la 31° Conferencia Internacional sobre aprendizaje Autónomo.* Beijing, China: JMLR: W y P Vol. 32 9 páginas.

EcuRed. (2016). *Investigacion - Accion*. Obtenido de https://www.ecured.cu/Investigaci%C3%B3n_-_Acci%C3%B3n

Fernandez, S. d. (2011). Series Temporales: Modelo Arima. *Estudio de la Facultad Ciencias Economicas y Empresariales - Departamento de Economia Aplicada*. Madrid: Universidad Autonoma de Madrid .

Gamboa, H. B. (2014). *Inteligncia artificial: principios y aplicaciones.* Quito: Springer.

Garbanzo Vargas, G. M. (2010). Factores Asociados al Rendimiento académico en estudiantes universitarios. *al Rendimiento académico en estudiantes universitarios. Costa Rica.*, 6.

Garcia S., A. (2012). *Inteligencia Artificial - Fundamento práctica y aplicaciones.* Madrid: RC Libros.

Garcia, M. D. (2005). manual weka. En U. d. Waikato, *manual weka* (pág. 44). Nueva Zelanda: web.

Garcia-Cruz, R. G. (2007). Tres aristas de un triangulo: Bajo rendimiento Académico, Fracaso y Deserción escolar y un centro "El Adolescente". *Revista Electrónica de Psicologia, volumen2 n1 Mexico*, 90.

Giovanny, I. A. (2014). Estilos de vida y su relación con las calificaciones escolares: estudio en Ecuador. *Revista Cubama de Investigaciones Biomédicas I-14 REtrieved from*, 36(4). Obtenido de http://scielo.sld.cu

Gonzales, L. E. (2002). CONTROL DE CALIDAD. *Una aproximación basada en los modelos ARIMA con Análisis de Intervención*, 80.

Gutiérrez, E. G., & Real, R. d. (2017). Un modelo basado en el Clasificador Naïve Bayes para la evaluación del desempeño. *RIED. Revista Iberoamericana de Educación a Distancia, vol. 20, núm. 2, 2017, pp. 293-*, 22.

Halpern, R. (1986). “Effects of early childhood intervention on primary schools progress on Latin America”. *Comparative Education Review*, 30.

Hastie, T. F. (2001). The Elements of Statistical Learning. Nueva York, Estados Unidos: Springer. *The Elements of Statistical Learning*, 606.

Herbert, J. (2019). *Ciencia de los datos (La guia definitiva sobre análisis de datos, minerias de datos, almacenamiento de datos, visualizacion de datos, BigData para empresas y aprendizaje automático para principiantes).* Independently published.

IBM. (s.f.). *ibm.com*. Obtenido de https://www.ibm.com/pe-es/cloud/deep-learning

Igarza, E. (13 de 07 de 2018). *repositorio.unfv.edu.pe/handle/UNFV/2234*. Obtenido de repositorio.unfv.edu.pe/handle/UNFV/2234: http://repositorio.unfv.edu.pe/handle/UNFV/2234

Jafari,S. , Pirouz,M. , Pirouz, B. (2013). Application of fuzzy logic for performance evaluation of academic students”,. *IEEE 13th Iranian Conference on Fuzzy Systems*, 13.

Jaimes, C. A. (2015). *Optimizacion del clasificador “Naive Bayes” usando árbol de desicion C4.5* . Lima: Universidad Nacional de Maryor de San Marcos.

Jesus, g. H. (2012). Tecnicas de análisis de datos. *Universidad Carlos III de Madrid*, 266.

Joaquin, A. R. (febrero de 2017). *Cienciadedatos.net*. Obtenido de Árboles de decisión, random forest, gradient boosting y C5.0: https://www.cienciadedatos.net/documentos/33_arboles_de_prediccion_bagging_random_forest_boosting#Random_Forest

Jose, M. H. (18 de 09 de 2020). *Iartificial.net*. Obtenido de https://www.iartificial.net/random-forest-bosque-aleatorio/

Kononenko, I. (1991). *Semi-naive Bayesian classifier. In Y. Kodratoff.* Berlin: Springer-Verlag: (Ed.), Proc Sixth European Working Session on Learning (pp. 206–219.

Lopera, M. E. (2018). Los árboles de decisión como herramienta para el analisis de riesgos de los proyectos. *Maestria en Gerencia de proyectos*, 85.

Lopez Takeyas, I. B. (2006). *algoritmo ID3.* mexico: itnuevolaredo.

Maravall, A. y. (1998). un nuevo método para el control de calidad de los datos en series de datos

en series temporales. *ARIMA*, 28.

Maria, G. C. (2014). Aplicacion de Arboles de decisión para la redicción del rendimiento académico de los estuiantes de los primeros ciclos de la carrera de Ingenieria Civil. *Tesis de la Universidad Continental*. Huancayo, Perú: Universidad continental.

Martinez, L. y. (2016). Transtornos afectivos y su incidencia en el rendimiento académico de los estudiantes de la escuela de educación básica "Paul, Percy, Harry bryan" de la ciudad de machala perteneciente a la ciudadela el bosque. *Tesis para optar el titulo de Licencia do en Ciencias de la Educacion, Mención: Psicologia educativa y Orientacion Vocacional. Universidad Tecnica de Machala Ecuador*, 50.

Matich, D. J. (2001). Obtenido de frro.utn.edu.ar: https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

Mendez, A. (22 de enero de 2019). *plandemejora.com*. Obtenido de Plan de mejora.com: https://www.plandemejora.com/como-hacer-un-diagrama-de-arbol-de-decisiones/

mohmad, B. A.-D. (2011). Suite of decision tree-based classification algorithmson cancer gene expression data. *Egyptian informatics Journal*, 82.

Morales, J. C.-M. (2019). Analysis of the Academic Performance of the Students of Mechanical and Electrical Engineering of the UNTELS. *IEEE EDUNINE*, 12.

Moreno, M. A. (2008). Topicos y Aplicaciones de Inteligencia Artificial. 217.

Naive. (2021). *aprende IA*. Obtenido de Naive Bayes – Teoría: https://aprendeia.com/naive-bayes-teoria-machine-learning/

Nováez, M. (1986). *Psicología de la actividad escolar.* México: Iberoamericana.

Paloma, R. d. (16 de Noviembre de 2017). *Think Big*. Obtenido de https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

Paola, C. B. (2010). Introduccion a las teccnicas de inteligencia artificial aplicadas a la gestión financiera empresarial. *Scielo*, 5.

Peña, A. M. (1992). Los niños con necesidades especiales. Afectividad y conducta. *C.P.E.I.P. Santiago*, 13.

Planck, B. (2014). Factores determinantes del rendimiento académico de los estudiantes de la Universidad de Atacama. *Revista Estudios Pedagógicos Volumen 1*, 39.

Rafael, .. R. (2018). *Los árboles de dcision una herramienta práctica para la toma de decisiones del Happy Manager.* Venezuela: UNERMB.

Ramadhani, N. , Pujianto, U. . (2017). Finding similar clustering pattern between students academic performance and non curricular activities data. *International Conference on Sustainable Information Engineering and Technology*, 10.

Ramadhani,N. , Pujianto,U. . (2017). "Finding similar clustering pattern between students academic performance and non curricular activities data",. *International Conference on Sustainable Information Engineering and Technology,*, 10.

Ratanamahatana, C. y. (September de 2002). 'Selective Bayesian classifier: feature selection for the Naive Bayesian classifier using decision trees', Proceedings of the 3rd International Conference on Data Mining Methods and Databases for Engineering, Finance and Other Fields,. *Bologna*, pág. 623.

Raul, E. L. (13 de Junio de 2017). *Matemáticas, análisis de datos y python*. Obtenido de https://relopezbriega.github.io/blog/2017/06/13/introduccion-al-deep-learning/

Reguero, J. (27 de julio de 2015). *Data Science: qué es y de dónde viene*. Obtenido de Datos Piperlab: https://piperlab.es/2015/07/27/data-science-que-es-y-de-donde-viene/

Rish, I. (2001). An empirical study of the naive Bayes classifier. IBM Research . *Report, Yorktown Heights.* .

Rodriguez, R. D. (2010). Metodologia para el desarrollo de Proyectos en Minería de Datos CRISP-DM.

Rodríguez, S. (2016). *emagister.com*. Obtenido de www.emagister.com/blog/author/sarar/

Romain, G. (2016). Random Forests Teoria y Ejmplos (Departamiento de Ingenería Matemática, Doble titulo Ecole Centrale de Nantes (Francia). *Conferencia 9, GLAM,*, (pág. 62). Francia.

Ronal, A. F. (1962). *statistical methods for research workers.* Reino unido: editorial.

Rudolph, R. (2018). *Machine Learning*. EEUU: CreateSpace Independent Publishing Platform (22 Mayo 2018).

Ruiz, C., Basualdo M. S. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. . *Informática Aplicada a la Ingeniería de Procesos – Orientación I*, 55.

S. Merchan Rubiano, J. D. (2015). "Formulation of a predictive Model for academic performance based on student's academic and demographic data", . *IEEE Frotiers in Education* , 13.

Servente, M. G.-M. (2003). *algoritmos tdidt aplicados a la minería de datos INTELIGENTE* . Buenos aires: UBA.

Silveira, D. P. (20 de 7 de 2016). Obtenido de Ciencia de los datos: https://www.oficinadanet.com.br/post/16919-o-que-e-data-science

SINEACE. (OCTUBRE de 2016). "Modelo de acreditación para programas de estudios de educación superior universitaria", . SINEACE.

SQL Docs. (2017). *Mining Models (Analysis Services - Data Mining)*. Obtenido de https://docs.microsoft.com/en-us/sql/analysis-services/data-mining/mining-models-analysis-services-data-mining?view=sql-server-2017

Sucar, L. E. (2018). Redes Bayesianas. En L. E. Sucar, *Redes Bayesianas* (pág. 28). mexico: INAOE. Obtenido de Redes Bayesianas.

SUNEDU. (2019). Obtenido de "Avances y status del licenciamiento": https://www.sunedu.gob.pe/avances-licenciamiento

SUNEDU. (2021). Obtenido de https://www.sunedu.gob.pe/avances-licenciamiento

Terry, H. (12 de Agosto de 2018). *teachthought.com*. Obtenido de https://www.teachthought.com/the-future-of-learning/a-definition-for-mobile-learning/: https://www.teachthought.com/the-future-of-learning/a-definition-for-mobile-learning/

Thomas, H. (2001). Inteligencia artificial. *Revista Academica Universidad Bolivariana volumen 1 numero 2*, 24.

Topic, F. R. (2019). *Artificial Neural Networks as Models of Neural Information Processing* . Obtenido de https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing

UA, b. (2020). *Universidad de Alcalá*. Obtenido de https://www.master-data-scientist.com/que-es-masters-in-data-science/

UC, B. D. (2018). *Investigación Aplicada"*. Obtenido de http://www.duoc.cl/biblioteca/crai/definicion-y-proposito-de-la-investigacion-aplicada

UNAM. (29 de abril de 2019). *Universidad Nacional de Moquegua*. Obtenido de https://unam.edu.pe/

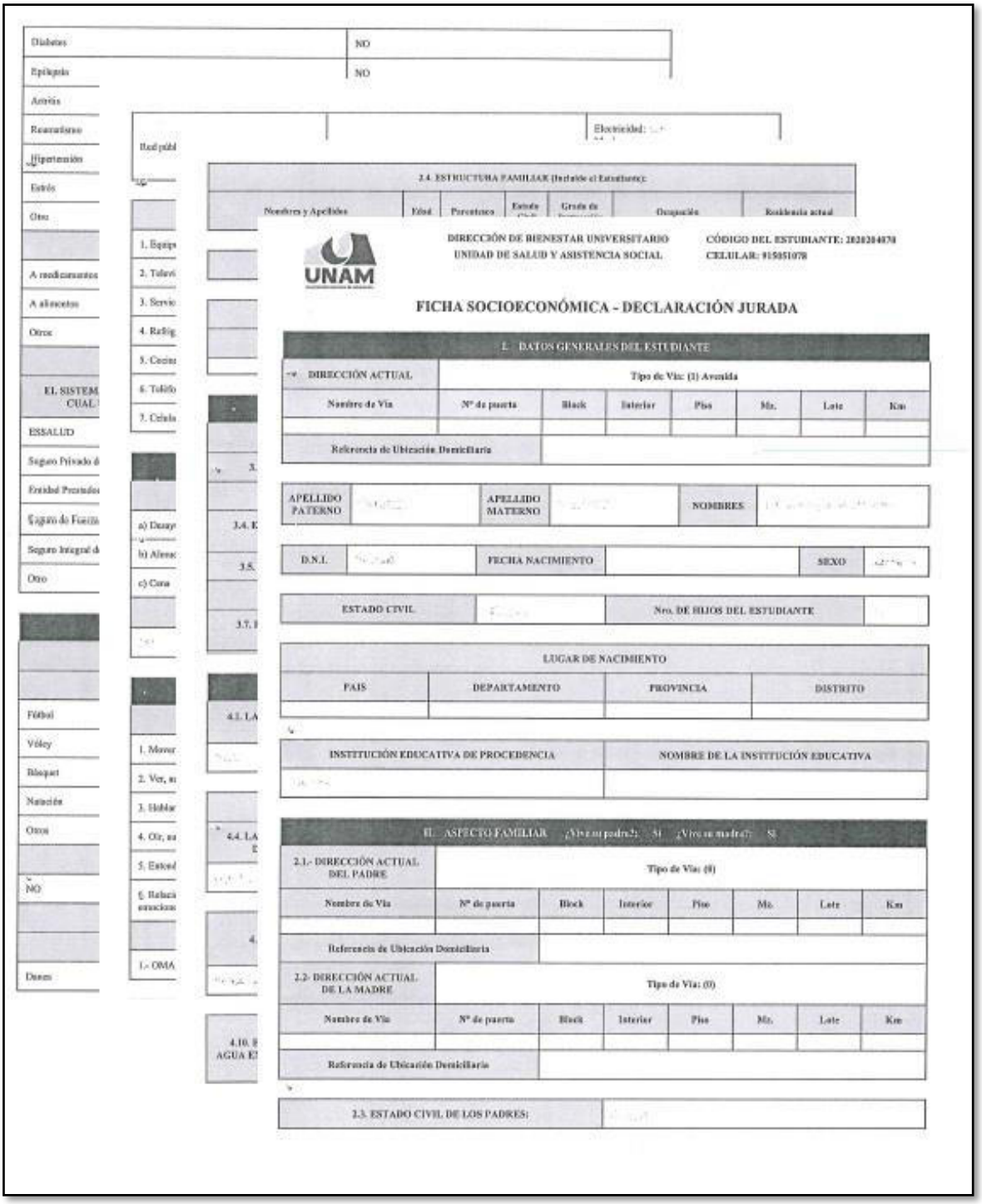
Usama, F. G.-S. (1996). Data Mining to Knowledge Discovery in Databases. *Scielo*, 90.

Wilson Fernando Melo, Francisca Salinas Esteban. (2010). Estudio sobre los factores que influyen en el rendimiento escolar. *BOG Bogota Positiva*, 44. Obtenido de https://evaluacion.educacionbogota.edu.co/files/Factores%20que%20influyen%20en%20el%20rendimiento%20escolar.pdf

Yaritza, G. O. (2014). Estudiantes universitarios con bajo rendimiento académico. *scielo Rev. EDUMECENTRO vol.6 n°2*, 50-62.

ANEXOS

FICHA SOCIOECONÓMICA



DIRECCIÓN DE BIENESTAR UNIVERSITARIO
UNIDAD DE SALUD Y ASISTENCIA SOCIAL

CÓDIGO DEL ESTUDIANTE: 2020204870
CELULAR: 915051070

FICHA SOCIOECONÓMICA - DECLARACIÓN JURADA

I. DATOS GENERALES DEL ESTUDIANTE

DIRECCIÓN ACTUAL	Tipo de Vía: (1) Avenida						
Nombre de Vía	N° de puerta	Block	Interior	Piso	Mz.	Lote	Km
Referencia de Ubicación Domiciliaria							

APELLIDO PATERNO		APELLIDO MATERNO		NOMBRES	
D.N.I.		FECHA NACIMIENTO		SEXO	
ESTADO CIVIL		Nro. DE HIJOS DEL ESTUDIANTE			

LUGAR DE NACIMIENTO			
PAIS	DEPARTAMENTO	PROVINCIA	DISTRITO

INSTITUCIÓN EDUCATIVA DE PROCEDENCIA	NOMBRE DE LA INSTITUCIÓN EDUCATIVA

II. ASPECTO FAMILIAR ¿Vive su padre?: SI ¿Vive su madre?: SI

2.1.- DIRECCIÓN ACTUAL DEL PADRE	Tipo de Vía: (0)						
Nombre de Vía	N° de puerta	Block	Interior	Piso	Mz.	Lote	Km
Referencia de Ubicación Domiciliaria							
2.2- DIRECCIÓN ACTUAL DE LA MADRE	Tipo de Vía: (0)						
Nombre de Vía	N° de puerta	Block	Interior	Piso	Mz.	Lote	Km
Referencia de Ubicación Domiciliaria							

2.3. ESTADO CIVIL DE LOS PADRES:	

Fuente: Entregada por Dirección de Bienestar Universitaria

DATOS ACADEMICOS

Nombres	CodigoCiclo	iCentralCicloAc	Seccion	Plan	Curso	Docente	Cod_curso	Unidad_1	Unidad_2	Unidad_3	Unidad_4	Nota_final	Estado
	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	12.90	13.70	NULL	NULL	13	ESTUDIANTE
	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	13.00	12.50	NULL	NULL	13	ESTUDIANTE
	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	13.50	13.00	11.40	12.50	13	ESTUDIANTE
	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	14.40	8.90	NULL	NULL	12	ESTUDIANTE
	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	15.80	14.80	NULL	NULL	15	ESTUDIANTE
	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	10.30	9.00	NULL	NULL	10	ESTUDIANTE
	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	12.00	12.80	NULL	NULL	12	ESTUDIANTE
	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	16.50	16.90	NULL	NULL	17	ESTUDIANTE
	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	13.50	13.60	NULL	NULL	14	ESTUDIANTE
	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	13.50	12.60	14.00	14.00	14	ESTUDIANTE
	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	14.40	13.50	NULL	NULL	14	ESTUDIANTE
	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	13.60	15.60	NULL	NULL	15	ESTUDIANTE
	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	11.20	11.40	NULL	NULL	11	ESTUDIANTE
	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	14.00	14.90	NULL	NULL	14	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	15.90	15.50	NULL	NULL	16	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	14.50	14.10	NULL	NULL	14	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	14.90	14.50	14.00	14.10	14	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	11.80	11.90	NULL	NULL	12	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	11.80	13.70	NULL	NULL	13	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	10.90	7.60	NULL	NULL	10	ESTUDIANTE
ATHAN DAV	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	12.00	10.50	NULL	NULL	11	ESTUDIANTE
	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	15.40	15.20	NULL	NULL	15	ESTUDIANTE
	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	12.50	13.10	NULL	NULL	13	ESTUDIANTE
	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	12.60	12.50	12.90	13.50	13	ESTUDIANTE
	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	12.70	11.50	NULL	NULL	12	ESTUDIANTE
	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	13.80	9.70	NULL	NULL	12	ESTUDIANTE
	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	7.50	10.60	NULL	NULL	11	ESTUDIANTE
	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	14.00	14.90	NULL	NULL	14	ESTUDIANTE
	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	16.50	16.00	NULL	NULL	16	ESTUDIANTE
	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	14.40	14.00	NULL	NULL	14	ESTUDIANTE
	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	14.50	14.00	13.90	14.50	14	ESTUDIANTE
	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	14.80	15.10	NULL	NULL	15	ESTUDIANTE
	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	16.30	16.40	NULL	NULL	16	ESTUDIANTE
	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	12.30	12.80	NULL	NULL	13	ESTUDIANTE
	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	14.40	14.60	NULL	NULL	15	ESTUDIANTE
S	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	16.70	17.00	NULL	NULL	17	ESTUDIANTE
S	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	14.50	14.10	NULL	NULL	14	ESTUDIANTE
S	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	14.50	15.00	13.50	12.60	14	ESTUDIANTE
S	20202101	20202	A	2016	REDACCION Y COMUNICACION	PERICANAZAS MURILLO, BLANCA	IS-124	15.00	16.00	NULL	NULL	16	ESTUDIANTE
S	20202101	20202	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	QUISPE PACHO, MARCO ANTONIO	IS-123	16.20	16.40	NULL	NULL	16	ESTUDIANTE
S	20202101	20202	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	14.30	14.00	NULL	NULL	14	ESTUDIANTE
S	20202101	20202	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	14.40	15.50	NULL	NULL	15	ESTUDIANTE
OSAID	20202101	20202	A	2016	BIOLOGIA Y MEDIO AMBIENTE	ESPINOZA REYNOSO, ISABEL DEL	IS-127	15.50	15.90	NULL	NULL	16	ESTUDIANTE
OSAID	20202101	20202	A	2016	SOCIOLOGIA	VALDEZ VALDIVIA, JOHN SABINO	IS-126	15.40	13.00	NULL	NULL	14	ESTUDIANTE
OSAID	20202101	20202	A	2016	FILOSOFIA	VALDEZ VALDIVIA, JOHN SABINO	IS-125	15.10	14.50	12.90	12.90	14	ESTUDIANTE
OSAID	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
IO	20192	3	20201	A	2016	PROGRAMACION ORIENTADA A OBJET	ZUÑIGA INCALLA, ALEX PETER	IS-222	14.40	10.00	NULL	NULL	12	ESTUDIANTE
IO	20192	3	20201	A	2016	ESTRUCTURA DE DATOS	SILVA DELGADO, CARLOS ALBERT	IS-221	12.80	12.90	NULL	NULL	13	ESTUDIANTE
IO	20192	2	20192	A	2016	ALGEBRA LINEAL	TAMAYO MAMANI, DAVID	IS-223	11.70	10.00	NULL	NULL	11	ESTUDIANTE
IO	20192	2	20192	A	2016	MATEMATICAS DISCRETAS I	RAMOS RIVERA, SALOMON REY	IS-224	10.30	14.10	NULL	NULL	12	ESTUDIANTE
IO	20192	2	20192	A	2016	MATEMATICA II	TAMAYO MAMANI, DAVID	IS-225	11.80	10.60	NULL	NULL	11	ESTUDIANTE
IO	20192	2	20192	A	2016	ESTADISTICA DESCRIPTIVA	QUISPE PACHO, MARCO ANTONIO	IS-226	15.30	13.90	NULL	NULL	15	ESTUDIANTE
IO	20192	2	20192	A	2016	ORATORIA Y LIDERAZGO	GUTIERREZ CASTRO, FRANKLIN	IS-228	13.80	14.00	NULL	NULL	14	ESTUDIANTE
IO	20192	2	20192	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	12.30	11.60	NULL	NULL	12	ESTUDIANTE
IO	20192	1	20191	A	2016	SOCIOLOGIA	NULL	IS-126	NULL	NULL	NULL	NULL	15	ESTUDIANTE
IO	20192	1	20191	A	2016	FILOSOFIA	NULL	IS-125	NULL	NULL	NULL	NULL	15	ESTUDIANTE
IO	20192	1	20191	A	2016	REDACCION Y COMUNICACION	NULL	IS-124	NULL	NULL	NULL	NULL	14	ESTUDIANTE
IO	20192	1	20191	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	NULL	IS-123	NULL	NULL	NULL	NULL	15	ESTUDIANTE
IO	20192	1	20191	A	2016	MATEMATICA I	NULL	IS-122	NULL	NULL	NULL	NULL	11	ESTUDIANTE
IO	20192	1	20191	A	2016	FUNDAMENTOS DE PROGRAMACION	NULL	IS-121	NULL	NULL	NULL	NULL	10	ESTUDIANTE
IO	20192	1	20191	A	2016	BIOLOGIA Y MEDIO AMBIENTE	NULL	IS-127	NULL	NULL	NULL	NULL	15	ESTUDIANTE
IO	20192	2	20192	A	2016	FISICA ELECTRICA	QUINO VILLANUEVA, GROVERT	IS-327	4.30	0.00	NULL	NULL	2	ESTUDIANTE
ERLY	20192	2	20192	A	2016	MATEMATICAS DISCRETAS I	RAMOS RIVERA, SALOMON REY	IS-224	6.50	0.00	NULL	NULL	3	ESTUDIANTE
ERLY	20192	2	20192	A	2016	ESTADISTICA DESCRIPTIVA	QUISPE PACHO, MARCO ANTONIO	IS-226	4.80	2.70	NULL	NULL	4	ESTUDIANTE
ERLY	20192	2	20192	A	2016	ORATORIA Y LIDERAZGO	GUTIERREZ CASTRO, FRANKLIN	IS-228	12.10	12.00	NULL	NULL	12	ESTUDIANTE
ERLY	20192	2	20192	A	2016	FUNDAMENTOS DE PROGRAMACION	SILVA DELGADO, CARLOS ALBERT	IS-121	0.00	0.00	NULL	NULL	0	ESTUDIANTE
ERLY	20192	2	20192	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	2.00	0.00	NULL	NULL	1	ESTUDIANTE
ERLY	20192	1	20191	A	2016	FUNDAMENTOS DE PROGRAMACION	NULL	IS-121	NULL	NULL	NULL	NULL	9	ESTUDIANTE
ERLY	20192	1	20191	A	2016	SOCIOLOGIA	NULL	IS-126	NULL	NULL	NULL	NULL	13	ESTUDIANTE
ERLY	20192	1	20191	A	2016	FILOSOFIA	NULL	IS-125	NULL	NULL	NULL	NULL	13	ESTUDIANTE
ERLY	20192	1	20191	A	2016	REDACCION Y COMUNICACION	NULL	IS-124	NULL	NULL	NULL	NULL	12	ESTUDIANTE
ERLY	20192	1	20191	A	2016	METODOLOGIA Y TECNICAS DE ESTUDIO	NULL	IS-123	NULL	NULL	NULL	NULL	12	ESTUDIANTE
ERLY	20192	1	20191	A	2016	MATEMATICA I	NULL	IS-122	NULL	NULL	NULL	NULL	6	ESTUDIANTE
ERLY	20192	1	20191	A	2016	BIOLOGIA Y MEDIO AMBIENTE	NULL	IS-127	NULL	NULL	NULL	NULL	12	ESTUDIANTE
NCO	20192	2	20202	A	2016	FISICA ELECTRICA	CALIENES NUÑEZ, RAUL JOSE	IS-327	0.00	0.00	NULL	NULL	0	ESTUDIANTE
NCO	20192	2	20202	A	2016	MATEMATICA II	RAMOS RIVERA, SALOMON REY	IS-225	5.60	0.00	NULL	NULL	3	ESTUDIANTE
NCO	20192	2	20202	A	2016	ALGEBRA LINEAL	RUBIN DE CELIS VIDAL, AMERICO A	IS-223	2.10	0.00	NULL	NULL	1	ESTUDIANTE
NCO	20192	2	20201	A	2016	ANALISIS Y DISEÑO DE SISTEMAS I	GUZMAN VALDIVIA, JOSE ANTONI	IS-422	3.50	1.20	NULL	NULL	5	ESTUDIANTE
NCO	20192	2	20201	A	2016	MATEMATICAS DISCRETAS II	RAMOS RIVERA, SALOMON REY	IS-325	9.50	11.50	NULL	NULL	11	ESTUDIANTE
NCO	20192	2	20201	A	2016	FUNDAMENTOS DE SISTEMAS DE INFOR	GAUNA CHINO, MARIO	IS-323	12.70	13.80	NULL	NULL	13	ESTUDIANTE
NCO	20192	2	20201	A	2016	PROGRAMACION ORIENTADA A OBJET	ZUÑIGA INCALLA, ALEX PETER	IS-222	10.50	4.80	NULL	NULL	8	ESTUDIANTE
NCO	20192	2	20201	A	2016	ESTRUCTURA DE DATOS	SILVA DELGADO, CARLOS ALBERT	IS-221	7.40	1.20	NULL	NULL	4	ESTUDIANTE
NCO	20192	2	20201	A	2016	MATEMATICA I	TAMAYO MAMANI, DAVID	IS-122	11.80	10.70	NULL	NULL	11	ESTUDIANTE
NCO	20192	2	20192	A	2016	FISICA ELECTRICA	QUINO VILLANUEVA, GROVERT	IS-327	8.10	8.00	NULL	NULL	8	ESTUDIANTE
NCO	20192	2	20192	A	2016	MATEMATICAS DISCRETAS I	RAMOS RIVERA, SALOMON REY	IS-224	9.90	11.90	NULL	NULL	11	ESTUDIANTE
NCO	20192	2	20192	A	2016	ESTADISTICA DESCRIPTIVA	QUISPE PACHO, MARCO ANTONIO	IS-226	12.10	11.80	NULL	NULL	12	ESTUDIANTE
NCO	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fuente: Entregada por Dirección de Actividades y Servicios Académicos

MATRIZ DE CONSISTENCIA

PROBLEMA GENERAL	OBJETIVOS GENERAL	HIPÓTESIS GENERAL
¿Es posible determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM?	Determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM	Será factible determinar causantes de bajo rendimiento académico con modelo de Inteligencia Artificial de los estudiantes de la Escuela Profesional de ingeniería de Sistemas e Informática de la UNAM.
PROBLEMA ESPECÍFICOS	**OBJETIVOS ESPECÍFICOS**	**HIPÓTESIS ESPECÍFICOS**
• ¿De qué tipo son los datos que tiene la UNAM sobre los estudiantes? • ¿Será posible descubrir patrones de bajo rendimiento con métodos de inteligencia artificial? • ¿Cuál será el nivel de precisión de los métodos de inteligencia artificial?	• Seleccionar, preprocesar y conocer las características de los datos socioeconómicos y académicos de los estudiantes de la Universidad Nacional de Moquegua. • Desarrollar métodos de inteligencia artificial para descubrir patrones de bajo rendimiento académicos de los estudiantes. Caso EPISI UNAM • Evaluar los resultados de los métodos de inteligencia artificial.	• Se podrá seleccionar, preprocesar y conocer las características de los datos socioeconómicos y académicos de los estudiantes de la Universidad Nacional de Moquegua. • Se podrá descubrir patrones de bajo rendimiento con métodos de inteligencia artificial. • Se podrá evaluar los resultados de los métodos de inteligencia artificial.

"UNIVERSIDAD NACIONAL DE MOQUEGUA"

VICERRECTORADO DE INVESTIGACIÓN

FORMATO DE AUTORIZACIÓN DE PUBLICACIÓN EN EL REPOSITORIO INSTITUCIONAL UNAM

Formulario de autorización para la publicación electrónica en la página web del Repositorio Institucional de la Universidad Nacional de Moquegua, de conformidad con la Ley N° 30035 Ley que regula el Repositorio Nacional Digital de Ciencia, Tecnología e Innovación de Acceso Abierto y Resolución de Comisión Organizadora N° 041-2022-UNAM que aprueba el Reglamento que regula el Repositorio Institucional Digital de la UNAM.

1. DATOS DEL AUTOR

Nombres y Apellidos	: DINA ELISA LEÓN MARÍN
DNI/Carnet de Extranjería/Pasaporte N°	: 40415734
Correo electrónico	: dinaleon002@gmail.com
Teléfono/Celular	: 953521817

2. GRADO, TITULO O ESPECIALIZACIÓN (Marque con una X)

Bachiller	Título Profesional	Magister	Doctor	Segunda Especialidad
	X			

3. TIPO DE PUBLICACIÓN (Marque con una X)

a) Articulo ( )
b) Tesis de pregrado (X)
c) Tesis de Maestría ( )
d) Tesis de Doctorado ( )
e) Monografía ( )
f) Reporte de caso científico o hallazgos ( )
g) Libro ( )
h) Parte de un libro ( )
i) Revisión ( )
j) Conferencia ( )
k) Presentación ( )
l) Artículo preliminar ( )
m) Articulo antes de ser publicado ( )
n) Artículo de prensa ( )
o) Documentos internos ( )
p) Imagen ( )
q) Documentos técnicos ( )
r) Ponencias presentadas en eventos científicos (Jornada, Congreso, Simposio Conferencias, etc.) ( )
s) Informes de suficiencia profesional ( )
t) Patentes producidas por académicos/as ( )

4. TIPO DE TRABAJO DE INVESTIGACIÓN (Marque con una X)

a) Tesis (X)
b) Trabajo de Investigación ( )
c) Trabajo de suficiencia profesional ( )
d) Trabajo académico ( )

5. DATOS DE LA PUBLICACION

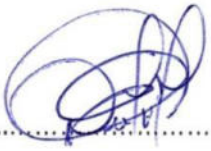
a) Título : "ANÁLISIS DEL RENDIMIENTO ACADÉMICO DE LOS ESTUDIANTES DE INGENIERÍA UTILIZANDO TÉCNICAS DE INTELIGENCIAL ARTIFICIAL - CASO: EPISI-UNAM"

b) Campo del conocimiento OCDE (Indicar el código de clasificación OCDE de la obra : https://catalogos.concytec.gob.pe/vocabulario/ocde_ford.html) : 2.000.00

c) Fecha de publicación : 2022

6. DATOS DEL ASESOR:

a) Nombres y Apellidos : Dr. ANIBAL FLORES GARCÍA
b) DNI/Pasaporte/Cédula : 04743476
c) ORCID : 0000-0003-4638-8242

FIRMA : 

DOCUMENTO DE IDENTIDAD : 40415734
NOMBRES Y APELLIDOS : DINA ELISA LEÓN MARÍN

LICENCIAS

Licencia Estándar

Bajo los siguientes términos, autorizo el depósito de mi trabajo de investigación en el Repositorio Institucional Digital de la UNAM.

Con la autorización de depósito da mi trabajo de investigación, otorgo a la Universidad Nacional de Moquegua, una Licencia No Exclusiva para reproducir. distribuir, comunicar al público, transformar (únicamente mediante su traducción a otros idiomas) y poner a disposición del público mi trabajo de investigación (incluido el resumen), en formato físico o digital, en cualquier medio, conocido o por conocerse, a través de los diversos servicios provistos por la Universidad, creados o por crearse, tales como el Repositorio Institucional Digital de la UNAM, Colección de Tesis, entre otros, en el Perú y en el extranjero, por el tiempo y veces que considere necesarias, y libre de remuneraciones.

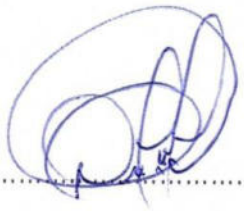
En virtud de dicha Licencia, la UNIVERSIDAD NACIONAL DE MOQUEGUA, podrá reproducir mi trabajo de investigación en cualquier tipo de soporte y en más de un ejemplar, sin modificar su contenido, solo con los propósitos de seguridad, respaldo y preservación.

Declaro que el trabajo de investigación es una creación de mi autoría y exclusiva titularidad, o coautoría con titularidad compartida, y me encuentro facultado a conceder la presente licencia y, asimismo, garantizo que dicho trabajo de investigación no infringe derechos de autor de terceras personas.

La UNIVERSIDAD NACIONAL DE MOQUEGUA consignara el nombre del/los autor/es del trabajo de investigación, y no le hará ninguna modificación más de la permitida en la presente licencia.

Autorizo el depósito (marque con una X).

(__x__) **SI autorizo que se deposite inmediatamente.**



FIRMA : .....



DOCUMENTO DE IDENTIDAD : 40415734

NOMBRES Y APELLIDOS : DINA ELISA LEÓN MARÍN

FECHA : 14 de Setiembre del 2022