Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

Predicción de la diabetes mediante aprendizaje de maquina con el uso de datos biométricos de estudiantes de pregrado de una universidad privada en la ciudad de Arequipa

Tesis presentada por el Bachiller:

Berrios Zuniga, Alvaro Daniel

ORCID: 0009-0007-0510-0368

para optar el Título Profesional de Ingeniero de Sistemas

Asesor (a):

Dr. Sulla Torres, Jose Alfredo

ORCID: 0000-0001-5129-430X

Arequipa - Perú

2024

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA DE SISTEMAS

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 05 de Junio del 2024

Dictamen: 009164-C-EPIS-2024

Visto el borrador del expediente 009164, presentado por:

2017220641 - BERRIOS ZUNIGA ALVARO DANIEL

Titulado:

PREDICCIÓN DE LA DIABETES MEDIANTE APRENDIZAJE DE MAQUINA CON EL USO DE DATOS BIOMÉTRICOS DE ESTUDIANTES DE PREGRADO DE UNA UNIVERSIDAD PRIVADA EN LA CIUDAD DE AREQUIPA

Nuestro dictamen es:

APROBADO

Titulo Profesional/Titulo de Segunda Especialidad/Grado Académico a optar:

INGENIERO DE SISTEMAS

29217790 - TORRES GAMARRA NESTOR
DICTAMINADOR

29413196 - MONTESINOS MURILLO ANGEL FELIPE
DICTAMINADOR

43635330 - ESQUICHA TEJADA JOSE DAVID
DICTAMINADOR

Predicción de la diabetes mediante aprendizaje de maquina con el uso de datos biométricos de estudiantes de pregrado de una universidad privada en la ciudad de Arequipa

INFORME DE ORIGINALIDAD

14%	3%	5%	11%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad de Jaén Trabajo del estudiante	1%
2	Submitted to Universidad Católica de Santa María Trabajo del estudiante	1%
3	Submitted to Aliat Universidades Trabajo del estudiante	1%
4	Submitted to Instituto Superior de Artes, Ciencias y Comunicación IACC Trabajo del estudiante	1%
5	itespanol.blogspot.com Fuente de Internet	1%
6	Submitted to Pontificia Universidad Catolica del Ecuador - PUCE Trabajo del estudiante	<1%
7	Submitted to BENEMERITA UNIVERSIDAD AUTONOMA DE PUEBLA BIBLIOTECA Trabajo del estudiante	<1%

DEDICATORIA

El presente trabajo de investigación va dedicado en primer lugar a Dios, y a todas las personas que nunca dejaron de creer en mí y me apoyaron en todo este largo camino para ser profesional.

Va dedicado también a mis padres que son el pilar de este logro porque sin ellos no hubiera podido ser posible llegar hasta este objetivo.

A mi hermana la cual nunca dejo que me rindiera y sobre todo que nunca deje de luchar por mis sueños, siempre estaré agradecido y espero sirva como ejemplo para que ella también nunca deje de creer que las cosas si se pueden hacer con esfuerzo.

A mi abuelita que es una segunda madre para mí porque nunca me dejo solo cuando yo quería renunciar a muchas cosas y a mi abuelito que está en el cielo al cual no pude conocer pero que es mi ángel guardián para siempre.

A mi enamorada Alejandra que siempre estuvo pendiente de todo el proceso y nunca dejo de motivarme y me apoyo en el desarrollo de la tesis. Muchas Gracias cariño

A mis amigos incondicionales que siempre me alentaban a poder seguir y no bajar los brazos ante las adversidades. Porque la meta de todos es una sola y es triunfar en la vida con esfuerzo, dedicación y mucha valentía.

A todos los mentores que pudieron ser partícipes de este logro, gracias por sus enseñanzas en las aulas y por todas las lecciones aprendidas porque hoy ese aprendizaje ha dado su fruto.

Muchas Gracias papas por permitirme estudiar la carrera más bonita de todas y permitirme ser Ingeniero de Sistemas.

Con mucho cariño. ¡ Todo este logro va para ustedes ¡

AGRADECIMIENTOS

Agradezco al Ing. José Sulla Torres por su apoyo en el desarrollo de este proyecto de investigación. Al Ing. José Esquicha por la asesoría brindada. Estaré siempre agradecidos pro brindarme las enseñanzas necesarias para cumplir un objetivo profesional y tomar como ejemplo su gran experiencia en el tema de investigación e ingeniería.

RESUMEN

Existen muchas enfermades a nivel mundial que afecta a la población y generan muchas muertes, una de ellas es la Diabetes. La diabetes es una enfermedad crónica que afecta a muchas personas alrededor del mundo y el desarrollo de esta es a causa de la poca producción de la insulina en el cuerpo. Existen diversos tipos de diabetes, pero las más comunes son la diabetes tipo 1 y la diabetes tipo 2. En el Perú según el Ministerio de Salud aproximadamente hasta fines del 2023 1 millón de personas tienen esta enfermedad, lo que involucra según la estadística extraída que 6 de cada 100 personas tienen diabetes y se estima que menos del total sabe su diagnóstico certero de la enfermedad por no tener una detección a tiempo.

Por ello se aplicó la inteligencia artificial para la detección de la diabetes en el alumnado de la una Universidad Privada en la ciudad Arequipa utilizando datos biométricos como el sexo, edad, peso, estatura, glucosa, antecedentes familiares por diabetes, consumo de alcohol, consumo de drogas, consumo de tabaco, y realización de actividad física. .Se usó algoritmos de aprendizaje automático supervisado para poder realizar esta predicción, ya que será capaz de brindar un resultado dependiendo de los datos de la persona.

Los resultados mostraron un nivel de confiabilidad mayor al 90% gracias a la aplicación de la evaluación de métricas en los 7 algoritmos de aprendizaje supervisado usados. La predicción ayudara a evitar más casos de diabetes en el Perú por la falta de detección temprana y tener un seguimiento adecuado de esta enfermedad que cada día avanza más a nivel mundial.

Palabras Claves:

Diabetes, Inteligencia Artificial, Red Neuronal, Metodología CRISP-DM, Python

ABSTRACT

There are many diseases worldwide that affect the population and cause many deaths, one of them is Diabetes. Diabetes is a chronic disease that affects many people around the world and its development is a cause of low insulin production in the body. There are various types of diabetes, but the most common are type 1 diabetes and type 2 diabetes. In Peru, according to the Ministry of Health, approximately until the end of 2023, 1 million people have this disease, which involves, according to the statistics obtained, 6 out of every 100 people have diabetes and it is estimated that less of the total know their accurate diagnosis of the disease due to not having a timely detection.

For this reason, artificial intelligence was applied to detect diabetes in the student at a Private University in the city of Arequipa using biometric data such as sex, age, weight, height, glucose, family history of diabetes, alcohol consumption, consumption. of drugs, tobacco consumption and physical activity. Supervised machine learning algorithms are used to make this prediction, since it will be able to provide a result depending on the person's data.

The results showed a level of reliability greater than 90% thanks to the application of metric evaluation in the 7 supervised learning algorithms used. The prediction will help avoid more cases of diabetes in Peru due to the lack of early detection and adequate monitoring of this disease that is advancing more and more worldwide every day.

Keywords:

Diabetes, Artificial Intelligence, Neural Network, CRISP-DM Methodology, Python

INDICE

INDICE DE TABLAS

INDICE DE FIGURAS

INTRODUCCION

La diabetes es una “enfermedad metabólica crónica” que se caracteriza por tener niveles elevados de glucosa en la sangre (azúcar en sangre), que con el tiempo produce daños graves al corazón, los vasos sanguíneos, los ojos, riñones y los nervios. (Organización Mundial de la Salud, 2016)

La diabetes surge cuando el páncreas no produce suficiente insulina (hormona que regula la concentración de azúcar), o también cuando el organismo no puede utilizar de manera eficaz la insulina que produce el cuerpo. La diabetes afecta mundialmente a muchas personas, desde jóvenes hasta adultos, pero en mayor grado a personas de la tercera edad. (Informe Mundial sobre la diabetes, 2016).

Las muertes por la diabetes han ido en aumento en las últimas décadas y es debido a que las personas en muchos casos no tienen un hábito de vida saludable o no se realizan controles de salud para la detección temprana de esta enfermedad crónica, en los que va del 2021 se estima que murieron 6,7 millones de personas. (Statista, 2023). Y solo en los últimos años el rango ha ido creciendo. Por ejemplo, en el 2010 murieron cerca de 4 millones de personas y en el 2019 4,2 millones de personas, lo que refleja que esto aumenta y preocupa a la salud mundial.

Existen diferentes tipos de diabetes, pero mayormente se presenta en pacientes, la diabetes tipo 1, tipo 2, y la diabetes gestacional. Para la diabetes tipo 1, o también conocida como la diabetes juvenil o diabetes de infancia, no se conoce una causa exacta, se dice que es por genética o factores ambientales que afectan directamente, ya que se caracteriza directamente por la producción

deficiente de insulina. Para el caso de la diabetes tipo 2 conocida como diabetes no insulinodependiente porque no depende de la insulina diaria a comparación de la diabetes tipo 1 que, si la requiere obligatoriamente, este tipo se especifica como un problema al no usar eficazmente la insulina, la gran mayoría de personas tienen ese tipo de diabetes. Este tipo solo se presentaba en adultos, pero estudios demostraron que en niños también se puede contraer este mal. Y, por último, la diabetes gestacional, como su nombre lo dice, se presenta en el momento de la gestación, ya que la madre tiene diabetes y, por lo tanto, el feto también genera complicaciones al nacer y casi siempre no tiene síntomas si no se detecta por pruebas de tamizaje para determinarlo. (Informe Mundial sobre la diabetes, 2016).

Uno de los factores principales para contraer la enfermedad de la diabetes es el exceso de grasa corporal, cuya causa es por falta de alimentación saludable o realizar ejercicio físico constante. Un ejemplo claro es que en países de bajos recursos o de mercados de libre comercio alimentario se encontró mayor IMC (Índice de Masa Corporal) en su población que en países europeos o asiáticos y esto se debe a que se consume más grasas que alimentos nutritivos que contienen fibra como las frutas y verduras, se tiende a consumir más bebida azucarada y contraer la diabetes tipo 2 ya en la edad adulta, pero en realidad este trastorno alimenticio ya viene desde niño. El fumar cigarro constantemente también demostró ser un factor importante para tener diabetes porque dejar de fumar pasado 10 años contrarrestaría este mal, por otro lado, la diabetes gestacional tiene su causa en que si la mujer que va a dar luz posee una edad avanzada tiene más riesgo a contraer este tipo, pero si a eso se le añade que sufre de sobrepeso las probabilidades aumentan sustancialmente, pero también existe un factor genético familiar que puede influir e incluso generar este mal en la descendencia familiar. (Informe Mundial sobre la diabetes, 2016).

Se conoce que en el Perú solo el 69% de las personas diagnosticadas reciben tratamiento para combatirla y solo el 30% mantiene un control adecuado. Del total de casos de diabetes en el país, el 95% es de diabetes tipo 2 y el 5% de otro tipo, que es una de las mayores causas de muerte a nivel nacional. Con el avance de la tecnología y la inteligencia artificial, se puede ayudar a detectar en una etapa temprana la diabetes para que estos casos disminuyan. Es por lo que al ver esta problemática y con el crecimiento de la tecnología, y en más específico, el análisis de datos. Se puede analizar basándonos en un dataset (conjunto de datos ordenados) valores médicos que nos ayuden a poder tratar la información y realizar Machine Learning (aprendizaje automático) que es una rama de la inteligencia artificial (IA) que se utiliza en diferentes campos como la ingeniería, la salud, la contabilidad, mapas espaciales, seguridad, etc. La predicción de poder contraer o no diabetes se basa en diferentes factores como (el cálculo de la masa corporal, el colesterol en sangre, infartos, tendencia a fumar, etc.), con los cuales se puede determinar si una persona tiene o no diabetes y cuál puede ser el tipo que puede contraer por la edad. Actualmente, muchas personas sufren este mal por no saber prevenir a tiempo la enfermedad y un diagnóstico adecuado.

Es por consiguiente que se propone aplicar la inteligencia artificial con el uso de aprendizaje automático para determinar si un alumno tiene diabetes o no. Mediante el uso de algoritmos de aprendizaje supervisado se plantea el entrenamiento de un dataset para determinar cuál es el mejor algoritmo y hacer pruebas correspondientes para obtener un resultado que supere el 90% de confiabilidad en predicción que se obtendrá por las métricas de confiabilidad aplicadas a los algoritmos de entrenamiento, determinando cuál es el mejor para realizar la predicción y determinar si un alumno tiene diabetes o no.

CAPITULO I

1. Planteamiento de la Investigación

1.1 Planteamiento del Problema

En el mundo existe muchas muertes por una enfermedad que ataca principalmente al páncreas la cual es denominada como diabetes, la diabetes es una enfermedad crónica porque es de larga duración y tiene un avance lento y dañino a lo largo del tiempo ya que no tiene una cura completa, se caracteriza por tener niveles elevados de glucosa en la sangre (azúcar en sangre), que con el tiempo produce daños graves al corazón, los vasos sanguíneos, los ojos, riñones y los nervios. (Organización Mundial de la Salud, 2016)

La diabetes surge cuando el páncreas no produce suficiente insulina (hormona que regula la concentración de azúcar [glucosa]) o también cuando el organismo no puede utilizar de manera eficaz la insulina que produce el cuerpo. La diabetes afecta mundialmente a muchas personas desde jóvenes hasta adultos, pero en mayor grado a personas de la tercera edad. (Informe Mundial sobre la diabetes, 2016)

Las muertes por la diabetes han ido en aumento en las últimas décadas y es debido a que las personas en muchos casos no tienen un habito de vida saludable o no se realizan controles de salud para la detección temprana de esta enfermedad crónica, en los que va del 2021 se estima que murieron 6,7 millón de personas. (Statista, 2023)

Y solo en los últimos años el rango ha ido creciendo por ejemplo en el 2010 murieron cerca de 4 millón de personas y en el 2019 4,2 millón de personas lo que refleja que esto aumenta y preocupa a la salud mundial. Existen diferentes tipos de diabetes, pero mayormente se presenta en pacientes

la diabetes tipo 1, tipo 2 y la diabetes gestacional. Para la diabetes tipo 1 o también conocida como la diabetes juvenil o diabetes de infancia no se conoce una causa exacta se dice que es por genética o factores ambientales que afectan directamente ya que se caracteriza directamente por la producción deficiente de insulina. Para el caso de la diabetes tipo 2 conocida como diabetes no insulinodependiente ya que no depende de la insulina diaria a comparación del tipo 1 que, si la requiere obligatoriamente, este tipo se especifica como un problema al no usar eficazmente la insulina, la gran mayoría de personas tienen ese tipo de diabetes. Este tipo solo se presentaba en adultos, pero estudios demostraron que en niños también se puede contraer este mal. Y por último la diabetes Gestacional como su nombre lo dice se presenta en el momento de la gestación ya que la madre tiene diabetes y por lo tanto el feto también generando complicaciones al nacer y casi siempre no tiene síntomas sino se detecta por pruebas de tamizaje para determinarlo. (Informe Mundial sobre la diabetes, 2016)

En el Perú esta enfermedad no es ajena ya que según el Ministerio de Salud aproximadamente hasta fines del 2023 1 millón de personas tienen esta enfermedad lo que involucra según la estadística extraída por el ministerio que 6 de cada 100 personas tienen diabetes y se estima que menos del total sabe su diagnóstico certero de la enfermedad. Solo el 69% de las personas diagnosticadas reciben tratamiento para combatirla y solo el 30% mantiene un control adecuado. Del total de casos de diabetes en el país el 95% es de diabetes tipo 2 y el 5% de otro tipo, pero llega a ser la mayor causa de muerte a nivel nacional. Con el avance de la tecnología y la inteligencia artificial se puede ayudar a detectar en una etapa temprana la diabetes para que estos casos disminuyan. (MINSA,2022)

Uno de los factores principales para contraer la enfermedad de la diabetes es el exceso de grasa corporal, cuya causa es por falta de alimentación saludable o realizar ejercicio físico constante.

Un ejemplo claro es que en países de bajos recursos o de mercados de libre comercio alimentario. Se encontró mayor IMC (Indicé de Masa Corporal) en la población de países europeos, asiáticos y en especial países latinoamericanos, esto se debe a que se consume más grasas que alimentos nutritivos que contienen fibra como las frutas y verduras, se tiende a consumir más bebidas azucaradas y eso ayuda a contraer la diabetes tipo 2 ya en la edad adulta, pero en realidad este trastorno alimenticio ya viene desde niños. El fumar cigarro constantemente también demostró ser un factor importante para tener diabetes por lo que dejar de fumar pasado 10 años contrarrestaría este mal, por otro lado la diabetes gestacional tiene su causa en que si la mujer que va a dar luz posee una edad avanzada tiene más riesgo a contraer este tipo pero si a eso se le añade que sufre de sobrepeso las probabilidades aumentan sustancialmente, pero también existe un factor genético familiar que puede influir e incluso generar este mal en la descendencia familiar. (Informe Mundial sobre la diabetes, 2016)

Es por lo que al ver esta problemática y con el crecimiento de la tecnología y en más específico el análisis de datos. Se puede analizar en base a un dataset (conjunto de datos ordenados) valores médicos que nos ayuden a poder tratar la información y realizar Machine Learning (aprendizaje automático) que es una rama de la inteligencia artificial (IA) que se utiliza en diferentes campos como la ingeniería, la salud, la contabilidad, mapas espaciales, seguridad, etc.

La predicción de poder contraer o no diabetes se basa en diferentes factores como (el cálculo de la masa corporal, el colesterol en sangre, infartos, tendencia a fumar, etc.), con los cuales se puede determinar si una persona tiene o no diabetes y cuál puede ser el tipo que puede contraer por la edad. Actualmente muchas personas sufren este mal por no saber prevenir a tiempo la enfermedad y un diagnóstico adecuado.

Es por consiguiente que se propone aplicar la inteligencia artificial con el uso del aprendizaje automático para determinar si un alumno tiene diabetes o no. Mediante el uso de algoritmos de aprendizaje supervisado se plantea el entrenamiento de un dataset para determinar cuál es el mejor algoritmo, analizado las métricas de confiabilidad hasta encontrar el algoritmo que tenga los mejores resultados y hacer pruebas del entrenamiento para obtener la predicción y determinar si un alumno tiene diabetes o no.

1.2 Objetivos de la Investigación

1.2.1 Objetivo General

Predecir la diabetes con el uso de datos biométricos de los estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa aplicando machine learning.

1.2.2 Objetivos Específicos

1. Analizar y evaluar los datos a requerir de los estudiantes según investigaciones relacionada a la predicción de diabetes que se debe cubrir para el proyecto.
2. Solicitar y extraer información para el uso en el modelo de predicción en base a una opinión médica.
3. Analizar y generar conocimiento de la data extraída de los estudiantes con uso de la estadística.
4. Realizar la transformación y estandarización de los datos para el entrenamiento y predicción.
5. Evaluar el mejor modelo de aprendizaje automático supervisado para la predicción.
6. Entrenar el modelo de aprendizaje automático con los datos extraídos del dataset para realizar la predicción asociada a la diabetes usando casos de prueba.

7. Validar los resultados obtenidos en los casos de prueba de personas de 16 a 34 años de en la ciudad de Arequipa con una opinión de un médico general.

1.3 Preguntas de la investigación

a) ¿Se puede realizar una predicción de la diabetes en base a datos biométricos en los estudiantes una Universidad Privada en la ciudad de Arequipa?

b) ¿Cuáles son los datos que se necesitan para realizar una predicción de este tipo?

c) ¿Qué algoritmos de aprendizaje automático se deben usar para obtener el mejor resultado?

d) ¿Qué métricas de predicción se deben utilizar para medir los algoritmos de aprendizaje automático?

e) ¿Se cumplió con la predicción propuesta?

f) ¿Cuántos casos se encontraron en base a la predicción?

1.4 Línea y Sub-línea de Investigación

1.4.1 Línea de Investigación

Inteligencia Artificial

1.4.2 Sub-Linea de Investigación

Aprendizaje Automático

1.5 Tipo y Nivel de investigación

1.5.1 Tipo de Investigación

a) Según su Finalidad

Aplicada

b) Según la Nota de Datos

Empírica o de Campo

c) Según su Contexto Histórico

Tradicional

1.5.2 Nivel de Investigación

Relacional

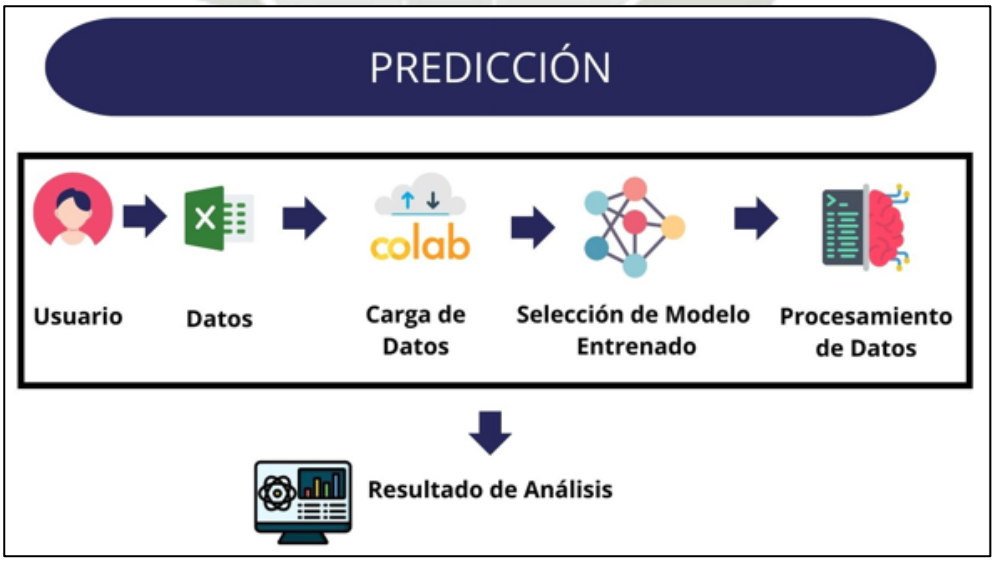
1.6 Palabras Clave

Diabetes, Inteligencia Artificial, Machine Learning, Metodología CRISP-DM, Python.

1.7 Solución Propuesta

En la investigación se entrenará un modelo de aprendizaje automático supervisado el cual permitirá que se pueda determinar si una persona tiene o no diabetes en los estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa. Los datos fueron extraídos de los registros de la Clínica Aliviari en base a los controles médicos que se les hace a los estudiantes anualmente. En la Figura N° 1 se puede apreciar el grafico de la solución propuesta.

***Figura. 1**. Ilustración de Propuesta de Investigación*



Nota: Propia

1.8 Justificación e Importancia

La diabetes es una enfermedad mundial que acaba con la vida de muchas personas en el mundo. Se estima que actualmente alrededor de 463 millones de adultos entre los 20 y 79 años tienen diabetes lo que representa el 9.3% de la población mundial y se estima que este número crezca en 700 millones para el 2045. En Latinoamérica se estima que la prevalencia esta entre 8% y 13% en los adultos de 20 a 79 años. (Archivos de Cardiología Mexico,2023)

Según la Federación Internacional de la Diabetes (FID) a la diabetes la catalogan como una enfermedad altamente mortal que ha matado más de 463 millones de personas en todo el mundo y de las cuales se estima que más de la mitad de ella no fue diagnosticada a tiempo, el alto nivel de glucosa es el principal problema debido a que el páncreas la produce para controlar el azúcar en la sangre, en el Perú hasta septiembre del 2022 se habían registrado 19,842 casos de diabetes en los cuales el 96% representaba la diabetes tipo 2 y solo el 1,4 % la diabetes tipo 1. (Ministerio de Salud, 2022)

Para poder detectar la diabetes se tiene que tener una muestra de sangre del paciente y usarla para determinar el nivel de glucosa, para ello existen varios tipos de análisis por ejemplo la prueba de glucosa en ayunas, mide los niveles de glucosa después de un ayuno nocturno, la prueba de hemoglobina (A1C) en la cual consiste en medir el nivel promedio de glucosa en sangre en los últimos 2-3 meses, el test de O' Sullivan que toma la muestra de valor de glucosa una hora después de haber consumido un líquido dulce y por último la prueba de tolerancia a la glucosa oral (PTGO) esta analiza numerosas medidas de glucosa en sangre a lo largo de un periodo de dos a tres horas en donde se hace en ayunas y luego se toma una glucosa liquida. Las 03 primeras pruebas se toman para determinar la diabetes tipo 1 y 2 y las ultimas para la diabetes gestacional. (Leal,2021)

Es por que se plantea el entrenamiento de un modelo de aprendizaje automático que pueda ayudar a la detección de la diabetes en los estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa que van desde los 16 años hasta los 34 años usando datos biométricos como el sexo, la edad, el peso, la talla, el valor de IMC, el nivel de Glucosa, los antecedentes familiares por diabetes, consumo de alcohol, consumo de tabaco, consumo de drogas y la realización de actividad física.

El propósito del proyecto de investigación en desarrollo es poder realizar el entrenamiento de modelo de aprendizaje automático para predecir la diabetes en los estudiantes de pregrado de una Universidad Privada de la ciudad de Arequipa desde los 16 años hasta los 34 años que servirá de ayuda para realizar una detección temprana de esta enfermedad que afecta a nivel mundial. Para este entrenamiento y predicción se necesitan datos reales de estudiantes en los cuales se puedan determinar que ayudan para la predicción propuesta.

1.9 Aporte

La investigación hará un aporte al campo de la medicina con la ayuda de la inteligencia artificial para combatir esta temible enfermedad denominada diabetes y también para que las personas puedan ser conscientes y acudir a un médico especialista al detectarse en una etapa temprana, para ser tratada y no llegar a niveles extremos donde es muy poco probable una recuperación efectiva. Es por ello por lo que una vez determine los resultados de la investigación a través del entrenamiento y predicción de un modelo de aprendizaje automático esta se puede llevar a una escala mayor usando equipos de IOT para obtener datos precisos y hacer la predicción en tiempo real. Haciendo mejor uso también de las tecnologías emergentes y demostrando que la medicina con la ayuda de la inteligencia artificial puede determinar un diagnóstico rápido de la diabetes para tener un tratamiento, ya que en el Perú es poco usado.

1.10 Enfoque

En la investigación se utilizó el método cuantitativo porque está basado en la recolección de datos y el análisis de valores numéricos que fueron extraídos de un dataset proporcionado por la Clínica Aliviari que ayudara a realizar una predicción en los datos recopilados. Con el aprendizaje automático se quiere lograr el objetivo de predecir la diabetes. El modelo utilizado permitirá una mayor factibilidad de utilización de los datos porque mediante modelos estadísticos podemos comparar resultados y procesarlos hasta llegar a una conclusión definitiva.

1.11 Alcances y Limitaciones

En la presente investigación se aborda el problema de la diabetes y como es una enfermedad crónica a nivel mundial y se contextualiza también la situación en el Perú determinando que es un problema que involucra muchos factores entre los principales tenemos la obesidad, la nutrición y falta de habito de ejercicio. Por lo cual se justifica el desarrollo del proyecto para que por medio del análisis de datos con ayuda de la inteligencia artificial se pueda determinar cuál es el diagnostico en una persona de tener o no diabetes.

Por ello se toman datos biométricos de los controles a los estudiantes de una Universidad Privada en la ciudad de Arequipa como su sexo, edad, peso, talla, valor de IMC, nivel de Glucosa, antecedentes familiares por diabetes, consumo de alcohol, consumo de tabaco, consumo de drogas y realización de actividad física. El campo de estudio principal es la inteligencia artificial que está asociada a la computación y la medicina.

El objetivo principal es entrenar un modelo de aprendizaje en base a los datos extraídos en el dataset de las variables biométricas para que mediante ello se pueda hacer una predicción si un estudiante de pregrado puede tener diabetes y si es así darle recomendaciones de cuidado de su

salud porque esta enfermedad puede afectar a sus órganos y llevarlo hasta la muerte. Previniendo este tipo de enfermedad también se ayudará a las personas a ser conscientes y ayudar que las cifras de mortalidad disminuyan.

La investigación debe cumplir con los siguientes requisitos:

- **Extracción de Información:** Se realiza una captura de datos para el análisis
- **Consolidación de Información:** Cargar los datos en el formato extraído csv.
- **Aprendizaje:** Entrenar un modelo de aprendizaje automático que sea capaz de predecir si un estudiante tiene diabetes o no.

Procesos:

El proceso principal es el entrenamiento en base a algoritmos de aprendizaje que permitan la detección de la diabetes teniendo en cuenta los datos que se usaran para el entrenamiento que se recopilo de alumnos de pregrado de una Universidad Privada en la ciudad de Arequipa en base a su salud.

Áreas Implicadas:

Las áreas que aportan a la investigación son la medicina y la computación las cuales se complementan para formar el análisis respectivo. En la medicina se tiene el área de endocrinología y en el campo de la computación existe la inteligencia artificial.

El modelo de entrenamiento podrá ingresar los datos mencionados y mediante ellos ya tener una respuesta calculando los valores del análisis para determinar si una persona tiene o no diabetes.

En ese sentido se tiene los alcances y limitaciones:

a) Alcances

- Usar datos médicos de los alumnos en base a su salud para determinar si tiene o no diabetes.
- Se usarán herramientas de análisis de datos como Google Colab que es de código abierto lo cual no implica un gasto adicional para la investigación.
- Entrenar un modelo y hacer el análisis más rápido respectivamente para la detección de la diabetes.

b) Limitaciones

- El tiempo de esta investigación es de 10 meses el cual se hará un procesamiento de datos, aprendizaje por medio de un modelo de aprendizaje y la documentación y puesta en marcha para probar su efectividad.
- El tamaño de la muestra es muy pequeña comparado con todos los estudiantes de pregrado que existen actualmente en la Universidad Privada en la ciudad de Arequipa, limitando un estudio de una población más amplia.

1.12 Población Muestra y Universo

Se tuvo acceso a un dataset de 3600 registros de estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa. En la cual con una limpieza de datos se encontró que muchos datos estaban incompletos por lo que al final el conjunto se redujo a 3542 registros.

1.12.1 Universo

Estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa.

1.12.2 Muestra

Muestra no probabilística por conveniencia de estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa entre los 16 y 34 años donde el número total es 3542 alumnos porque se quiere determinar diabetes en población joven.

1.13 Métodos, Técnicas e Instrumentos de Recolección de Datos

1.13.1 Métodos de Investigación

Este proyecto tiene una investigación predictiva debido a que requiere realizar una exploración previa, un análisis, una descripción y unas comparaciones para poder comprender el resultado de este.

1.13.2 Técnicas para la Investigación

Los datos evaluados en el dataset son:

a) Sexo
b) Edad
c) Peso
d) Talla
e) IMC
f) Glucosa
g) Antecedentes Familiares Diabetes
h) Consumo de Alcohol
i) Consumo de Tabaco
j) Consumo de Drogas
k) Actividad Física

Todos los campos son de datos variables dependiendo de los registros que se tengan y pueden ser variables categóricas o numéricas que se basan en si son datos que dan un

análisis en base a el sexo por ejemplo por tipo de edad y se harán los gráficos de análisis respectivos.

En la Tabla 1 se mencionan los instrumentos que se utilizaran para la recolección de investigaciones relacionadas y los datos de los estudiantes de pregrado de la Universidad Privada en la ciudad de Arequipa.

Tabla 1. Definición de Técnicas e Instrumentos

Técnica	Definición	Instrumento
Documental	Documentos relacionados a la investigación para recopilar información útil para el desarrollo del proyecto.	Bases de Datos científicas (Web of Science, IEE Explore y Google Scholar)
Muestra	A través de la recopilación de los datos de los estudiantes de pregrado de la Universidad Privada por medio del acceso de información de la clínica Aliviari se lograr tener datos reales para hacer una predicción y obtener información estadística.	Dataset extraído
Evaluación	Usando la fase de evaluación de la metodología Crisp-DM se podrá determinar que algoritmo de aprendizaje es el mejor para la predicción.	Métricas de aplicación para algoritmos de aprendizaje automático.
Resultados	Se mostrará el número de casos de diabetes encontrados en la predicción de los estudiantes de pregrado.	Algoritmo de IA para predicción.

Nota: Propio

1.13.3 Instrumentos de Tratamiento de Datos

Se utilizará un dataset el cual tendrá todas las columnas con los datos extraídos para lo cual se usará en la etapa de experimentación y posteriormente en la etapa de resultados

de la metodología escogida que pasara por una evaluación para lo cual este debe tener la confiabilidad mayor al 90% para determinar que la predicción fue correcta y ese porcentaje nace de la aplicación de las métricas que se realizada a cada uno de los algoritmos de aprendizaje automático en donde se determinara cual será el mejor algoritmo a usar.

CAPITULO II

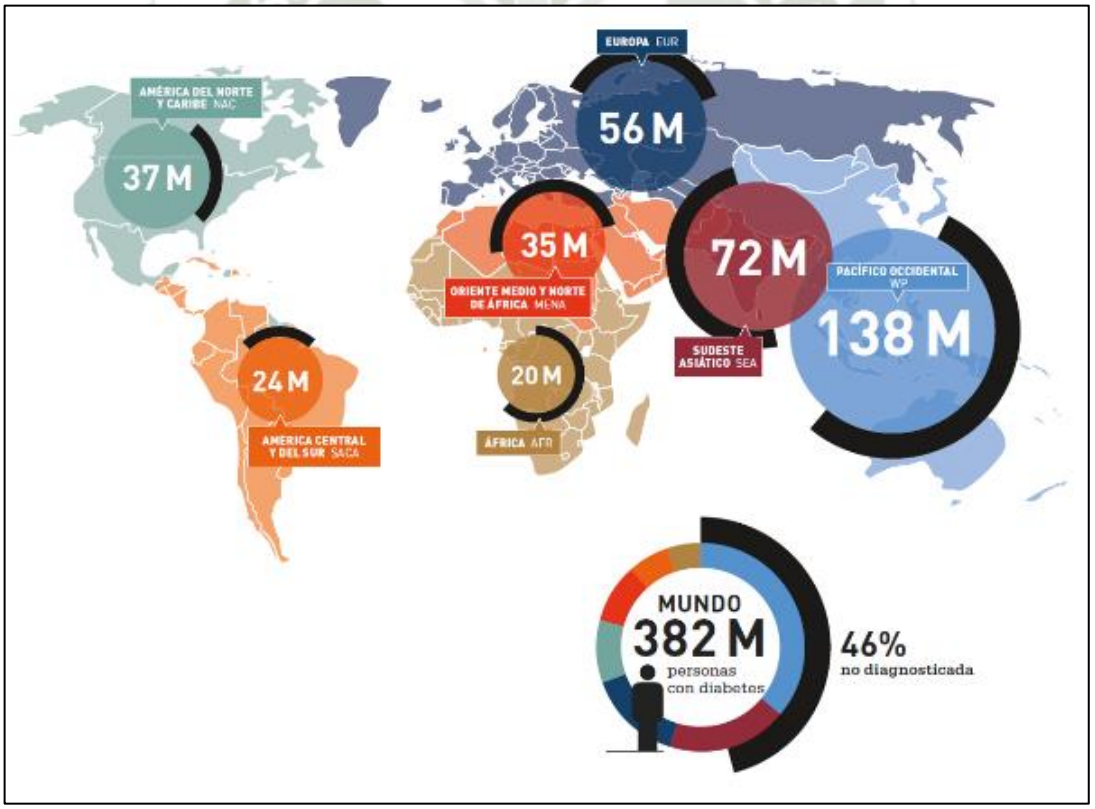
2. Fundamentos Teóricos

2.1 Bases Teóricas de la Investigación

Para una mejor comprensión de la investigación en base a la información general detallada se procederá a explicar más a detalle cada concepto para tener el conocimiento adecuado para comprenderla:

2.2 La Diabetes

Figura. 2.
La Diabetes Alrededor del Mundo desde el año 2021.



Nota: Atlas de la Diabetes del Instituto Internacional de Diabetes, 2021

Como se puede ver en la Figura N° 2**,** la Diabetes es una enfermedad que se da a causa de la disminución de la insulina que es producida por el páncreas para poder controlar los niveles de azúcar en la sangre en el cuerpo al no tener suficiente insulina la glucosa se eleva generando varios problemas de salud y si esta no llega a ser controlada es mortal. Es por ello por lo que los pacientes diagnosticados con esta enfermedad deben tener insulina para poder controlarlo. Según la OMS se estima que 1 de cada 11 adultos en el mundo sufre de diabetes representando un total de 415 millones de personas entre las edades de 20 y 79 años, además se pronosticó que para el año 2040 unos 642 millones de personas entre uno de cada 1 de cada 10 adultos tendrá diabetes. (International Diabetes Federation y Séptima, 2015). La diabetes según Harrison (Sataloff et al., 2015) se puede clasificar en 3 tipos: Diabetes Tipo I que se da más en niños, Diabetes Tipo II que es causa de una disminución de la insulina y la Diabetes Gestacional que se da a causa de una trasmisión de diabetes de una madre gestante a un neonato. (Atlas de la Diabetes del Instituto Internacional de Diabetes, 2021)

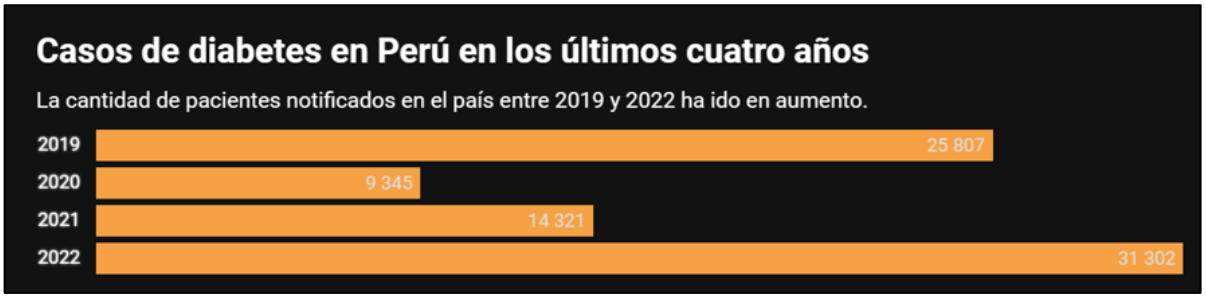
2.2.1 Diabetes en el Perú

El crecimiento de la población y el aumento de la de la obesidad a causa de la "urbanización y los cambios en el estilo de vida" de la población, contribuyen al crecimiento de la diabetes y otras enfermedades crónicas incluso en grupos etarios tales como niños y adolescentes, donde estas enfermedades eran poco frecuentes. Es por eso por lo que para el Perú esta enfermedad no es ajena debido a un análisis realizado en el 2022 se tiene como casos generales detectados de 31.302 la cual ha crecido a comparación del año 2019 que fue de 25.807 y del total se tiene que el 95,7% tiene la diabetes tipo 2 que está vinculada a estilos de vida poco saludables. Según la Asociación de Diabetes del Perú el 7% de la población adulta tiene diabetes y en la costa predomina más aun con el 8.5%,

en la sierra con el 5.5% y en la selva con el 3.5%. Pero también es importante hablar del costo ya que al estado peruano según el MINSA un paciente controlado llega a costar S/.1,392.00 soles anuales y un paciente no controlado llegaría a costar S/. 19,661.00 soles anuales. El costo promedio asciende a S/. 10,526 que multiplicado por el número de casos esperados en el país da un total de 5,297 millones de soles lo que representa el 20% del gasto sanitario del país. Si bien es cierto ya laboratorios farmacéuticos han innovado en ofrecer diferentes tipos de insulina bajo costo para el tratamiento de esta la economía peruana no se abastece de esta es por ello por lo que se recomienda tener buenos hábitos de salud, hacer deporte y revisar los niveles de glucosa en sangre. (Diario Gestión, 2018)

Por otro lado, en Perú, en el tema legal y de políticas los avances para asistir a las personas con esta enfermedad son escandalosamente lentos: todavía no se implementa el Plan Nacional de Prevención y Atención de Pacientes con Diabetes aprobado hace 18 años. El Estado peruano tiene el deber de atender a estos pacientes, que junto a los de otras enfermedades representan casi el 60% de la carga de atención en el sistema de salud. Las atenciones en 759 establecimientos alrededor del territorio nacional para la diabetes demostraron un incremento de 1’011.469 de consultas en 2022 frente a las 165.550 que se registraron en el 2018. . (Diario Gestión, 2018)

Figura. 3.
Casos de Diabetes en el Perú en los últimos 4 años



Nota: Elaboración propia del Diario Ojo Público con datos del Ministerio de Salud – MINSA, 2022

2.2.2 Diabetes en Jóvenes en el Perú

Según el Instituto Nacional de Estadística e Informática (INEI) informó que hasta el año 2019 el 7.5% de la población de 15 años a más fue diagnosticada con diabetes tipo 2 por un profesional de la salud. Solo el 73.9% del total de casos diagnosticados recibió tratamiento en 1 año sin embargo el 26.1% no lo recibió por falta de recursos o acceso ya que se encuentran en zonas rurales. Se determinó que la costa es la mayor población que se detectó casos de diabetes representando en la región que el 4% de toda la población tiene diabetes, en la selva se encontró el 1.9% de casos y en la sierra el 1.6%. De todos los casos encontrados se encontró que el 35.5% de esta población joven tiene sobrepeso y la mayor incidencia está en las zonas urbanas, siendo Tacna, Tumbes e Ica los departamentos con mayor cantidad de población obesa. La obesidad es un factor importante que puede desencadenar en una diabetes de tipo 2 y según el INEI, 9 de cada 100 personas en el Perú consume al menos 5 porciones de frutas o ensaladas y/o verduras al día lo que valida el resultado de casi el 40% de personas con obesidad. (Instituto Nacional de Estadística e Informática, 2019)

2.2.3 Causas de la Diabetes en el Perú

Las principales causas de la diabetes en el Perú están relacionadas a la obesidad, falta de ejercicio físico, vida sedentaria y un mal hábito de vida saludable, los cuales se explican a continuación:

- **Estilo de Vida**

La falta de ejercicio, dieta poco saludable y la obesidad como consecuencia aumentan el riesgo a desarrollar diabetes. Debido a la globalización, ha aumentado la disponibilidad y consumo de comidas rápidas con alto contenido de grasa, sal y calorías. Por ejemplo, se

identifica que el alto consumo de arroz, trigo refinado usado en panadería y carne roja está asociado al aumento de casos de diabetes. (Amelia et. all,2021)

- **Obesidad y Sobre Peso**

La obesidad es un factor de riesgo muy relacionado a la diabetes. Las personas obesas tienen 7 veces más posibilidades de presentar diabetes, en tanto el sobrepeso aumenta el riesgo 3 veces. Las personas que tienen obesidad severa tienen hasta 60 veces mayor riesgo de sufrir diabetes que las que tienen peso normal. Las tasas de obesidad en el Perú varían de 12.6% en los varones hasta el 20.3% en las mujeres. Las tasas crecientes de sobrepeso y obesidad son muy recientes, con porcentaje de personas obesas en Perú que han aumentado de 20% a 35% en un lapso de 10 años desde el 2012. (Organización Mundial de la Salud,2024)

- **Urbanización**

La urbanización ha llevado a un aumento en el transporte mecanizado, aumentando la expansión urbana y limitando las oportunidades para la actividad física diaria y eso conlleva a usar mucho la tecnología en base a la televisión, videojuegos o equipos informáticos conectados a internet. En el Perú, 1 de cada 4 niños entre 5-9 años tuvieron exceso de consumo digital frente a las niñas (10.8% vs 6.8%). La prevalencia del uso de la tecnológica para evitar la realización de actividad física en adolescentes fue de 11% y 13.3% respectivamente, siendo ambos ligeramente mayor en mujeres, lo que presenta un tema preocupante ya que las mujeres por tener un sedentarismo tienen obesidad y esto en edad reproductiva presentan mayor posibilidad de complicaciones serias durante el embarazo lo cual puede desencadenar una diabetes gestacional para el neonato. (Instituto Nacional de Estadística e Informática,2021)

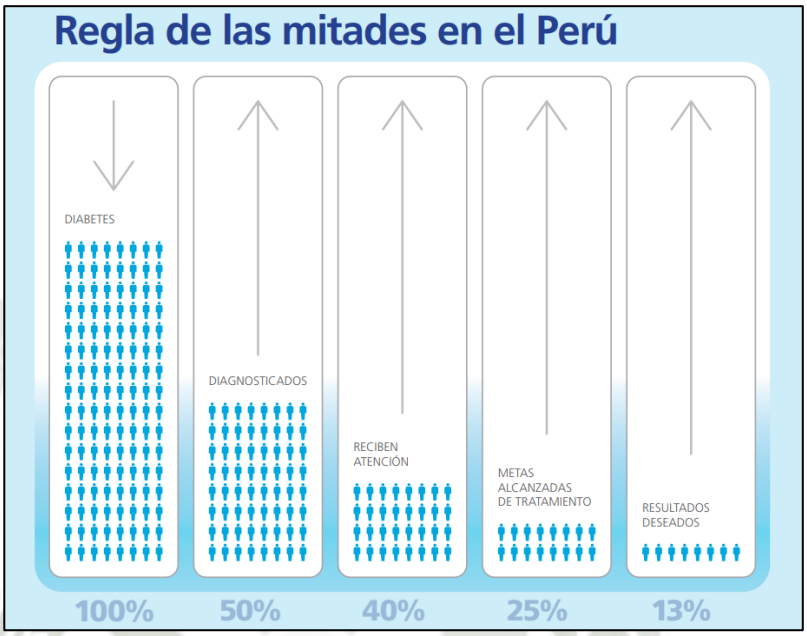
- **Urbanización: Relacionada a la Diabetes**

La urbanización conlleva de cierta forma a contribuir a la obesidad debido a que limita las oportunidades a realizar actividad física para las personas entonces el sobrepeso y la obesidad son predominantes en las ciudades urbanas y en la costa peruana, lo cual se podría explicar por el mayor desarrollo económico y urbanización lo que conlleva a cambios en el estilo de vida y provocan modificaciones en el patrón de alimentación y actividad física. Cabe destacar asimismo el acceso masivo a la televisión con mensajes que invitan al consumo de alimentos con alto contenido energético preparados fuera del hogar en estas regiones y que puede ser un factor de influencia importante. (Instituto de Salud Global de Barcelona, 2018)

2.2.4 Situación Estadística de Diabetes en el Perú

En la actualidad hay 1’996,800 personas con diabetes en el Perú, lo que representa el 6.4% de la población nacional. De todas las personas con diabetes, cerca del 50% son diagnosticadas (998,400 personas). De todas las personas diagnosticadas con diabetes, el 80% de ellas recibe algún tipo de tratamiento (798,720 personas). De todas las personas tratadas por diabetes, el 63% de ellas se adhiere al tratamiento (503,194 personas). De todas las personas diabéticas, diagnosticadas, tratadas y que se adhieren al tratamiento sólo el 50% de ellas vive sin complicaciones aparentes (251,597 personas). Y para el año 2030 se estima que el número de personas con diabetes se habrá incrementado a alrededor de 2’872,000 personas en el Perú. La prevalencia de la intolerancia a la glucosa en el Perú es de 8.11% (2’520,200 personas), este es un importante indicador de potencial desarrollo de diabetes en los siguientes 5 años. (MINSA,2021)

Figura. 4.
Regla de las Mitades Relacionada a la Diabetes en el Perú



Nota: Compendio de la Diabetes en el Perú, 2022

2.2.5 Complicaciones de la Diabetes

- **Desorden cerebrovascular (DCV):** Las personas con diabetes tienen 4 veces más probabilidades de tener un DCV que las personas sin diabetes. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)
- **Enfermedad Renal:** La diabetes es la causa principal de la enfermedad renal crónica y la enfermedad renal en etapa terminal (ERET). La ERET requiere de diálisis o un trasplante de riñón para reemplazar la función renal. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)
- **Amputación:** La diabetes es la causa principal de las amputaciones no traumáticas de los miembros inferiores. El riesgo de cinco años a morir después de una amputación es sustancialmente más alto que para muchos tipos de cáncer. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)

- **Hipoglicemia:** La hipoglicemia es una complicación común del tratamiento de la diabetes y alude a una condición en la que los niveles de glucosa en la sangre son bajos. Los síntomas de hipoglicemia incluyen latidos cardíacos fuertes, temblores, hambre, sudoración, dificultad para concentrarse y/o confusión. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)
- **Ceguera:** El daño a la retina por causa de la diabetes es la causa principal de pérdida de la visión (retinopatía diabética). Un tratamiento eficaz puede reducir el deterioro de la retina en más de un tercio. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)
- **Enfermedad Cardiovascular:** La presión arterial alta, colesterol alto, sobrepeso y obesidad - y la diabetes tipo 2 - son algunos de los principales factores de riesgo biológicos de las enfermedades cardiovasculares. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)
- **Cetoacidis Diabética:** La cetoacidosis diabética es una complicación potencialmente mortal, sobre todo en personas con diabetes tipo 1. Es el resultado de la escasez de insulina en las células del cuerpo, donde el cuerpo metaboliza ácidos grasos y produce cetoácidos. (National Institute of Diabetes and Digestive and Kidney Diseases,2021)

2.2.6 Medidas de Prevención para la Diabetes

Para promover el autocuidado se debe concientizar población en general sobre hábitos de vida saludable para retrasar el inicio de la diabetes, o de ser posible, evitar el desarrollo de esta. Entre las mejores recomendaciones se puede mencionar: La reducción de peso en pacientes con sobrepeso u obesidad es la medida más efectiva para retrasar la

aparición de la diabetes y de otras enfermedades. Se pueden lograr cambios significativos con una reducción del 5-10% del peso. Lo ideal es tener un plan de alimentación disminuyendo un 20% de grasas y azúcares, consumiendo mayor cantidad de fibra y un fraccionamiento en cuatro raciones al día. El segundo mejor componente es el ejercicio físico (150 minutos a la semana de acuerdo con la edad, ocupación y al estado físico) con el objetivo final de tratar de integrarlo a las actividades cotidianas además del desarrollo de hábitos saludables. Cumpliendo con estos tres pilares una persona mejorará su estado nutricional y su calidad de vida por lo tanto tiene un riesgo mucho menor de contraer la diabetes. (Mayo Clinic,2023)

2.3 Inteligencia Artificial

La inteligencia artificial llamada IA se puede definir como un sistema capaz de recabar datos y en base a estos poder tomar decisiones replicando o imitando la inteligencia humana buscando los mejores resultados posibles en base a una acción realizada por el ser humando, puede ser autónomos y tener la posibilidad de autoaprendizaje. Sin embargo, la comisión europea la define como la habilidad de una máquina de replicar capacidades que tienen los humanos para el razonamiento, el aprendizaje o la creatividad. Actualmente la IA está presente en nuestro día a día y su objetivo es mejorarla y se encuentra mayormente en teléfonos móviles en aplicaciones como por ejemplo YouTube al recomendar que poder ver, Netflix al mostrar un video relacionado a lo que estás viendo o Spotify que reproduce una canción asociada al género que estas escuchando. Todas estas funciones son algoritmos que hacen una búsqueda y empiezan a comparar resultados hasta encontrar una similitud y dar el resultado esperado. (Orihuela Martinez,2022)

2.3.1 Aplicación de la IA en diversos campos

La inteligencia artificial en los últimos años ha pasado de convertirse en una rama de la computación a ser aplicada en campos donde antes solo había intervención humana lo cual este concepto ha cambiado y desarrollo muchos algoritmos capaces de superar la mente humana en procesamiento de información. A continuación, se detalla los siguientes:

- **Finanzas:** Es Usado en el sector bancario para organizar operaciones o inversiones con grandes sumas de dinero y para detectar movimientos anormales en las cuentas de los clientes que posteriormente serán evaluados por un especialista. (Russell, 2004)
- **Industria:** La IA es aplicada en la robótica ya que se está utilizando en movilización de grandes cargas, realización de tareas peligrosas de máquinas industriales o en condiciones muy repetitivas en la vida humana. (Russell, 2004)
- **Domótica:** La aplicación de la robótica y la IA a los electrodomésticos ha cambiado el concepto de usar aparatos electrónicos es por eso por lo que actualmente el adjetivo «inteligente» o el prefijo inglés smart lo vamos encontrando cada vez más frecuentemente en lavadoras, relojes, luces, aparatos de aire acondicionado o televisores. De esta manera los electrodomésticos son capaces de aprender patrones de uso por parte de las personas que los utilizan o comportarse dependiendo de las condiciones. (Russell, 2004)
- **Conducción autónoma de vehículos:** La construcción de autos que se manejen solos es una realidad que en los últimos años ha causado mucho interés ya que se han desarrollado sensores en los automóviles que detectan un cambio de carril, agotamiento del conductor, velocidad, distancia, objetos o personas en el trayecto

de impacto o selecciones de la ruta óptima dependiendo de las condiciones del tráfico. Es casi común ya ver en los coches el sistema de aparcamiento automático y eso es gracias a los algoritmos de tomas de decisiones de la IA. (Russell, 2004)

- **Mercadotecnia:** Los mercados a nivel internacional es un campo muy amplio que varía rápidamente es por ello por lo que las tendencias del mercado, el análisis de las características de este, el comportamiento de los compradores puede ser simulado en modelos matemáticos que se aproximan mucho a la realidad. Amazon, Microsoft, Google y otros gigantes tecnológicos utilizan la IA para analizar nuestros gustos y recomendarnos objetos que puedan interesarnos a través de lo que se denomina publicidad personalizada. (Russell, 2004)
- **Reconocimiento facial:** Actualmente el reconocimiento fácil se ha vuelto muy usado en cámaras de vigilancia y esto se da mediante técnicas inteligentes de reconocimiento facial que capturan rasgos faciales y hacen la búsqueda en una base de datos para saber la mejor coincidencia, desarrollan numerosos proyectos también relacionados con publicidad directa personalizada o identificación personal. (Russell, 2004)

2.3.2 Aplicación de la IA en la Medicina

La Inteligencia Artificial recientemente ha empezado a usarse más en el campo de la medicina para revolucionar la detección, análisis y atención a los pacientes de manera automática. Es por ello por lo que se ha empleado en varios softwares en los cuales puede predecir un tumor pequeño solo analizando imágenes o teniendo un sistema para análisis de controles pediátricos para determinar si un niño está sano por lo que es importante que su aplicación crezca, pero siempre de la atención de un especialista en la rama. (IBM,2020)

Existen varios tipos de aplicaciones en la actualidad para la IA en la medicina por lo cual mencionaremos a continuación:

- **Prevención de enfermedades y diagnóstico:** En cuestión a la prevención en el campo de la IA se han desarrollado software con algoritmos para la detección de cáncer por ejemplo en el útero, cabeza y cuello mediante imágenes identificando patrones de repetición. También se desarrollaron programas para la detección de una cardiopatía mediante electrocardiográfico, diabetes y sistemas inteligentes de razonamiento basado en casos. Por otro lado, también mediante aplicativos se pudo realizar mediante una aplicación nuestros movimientos por sensores que pueden identificar patrones que sugieran progresión de la enfermedad del Parkinson lo que demuestra que su aplicación de tipo prevención está creciendo y con ello un diagnóstico rápido. (Ávila,2020)
- **Diagnóstico:** Los diagnósticos que usan IA ha crecido bastante gracias a el entrenamiento que han tenido estos softwares con el paso del tiempo, ya que ahora existen programadas que se pueden aplicar a diversas detecciones como enfermedades infecciosas, oftalmología, enfermedades renales, enfermedades reumatológicas. Por ejemplo investigadores de IBM publicaron una investigación en relación al cáncer de mama maligno los cuales examinaban imágenes de células cancerígenas en personas y los modificaban para que sean de alta calidad encontrando el 87% de casos analizados positivos y 77% de casos no cancerosos pero la idea es ayudar a un radiólogo a realizar un diagnóstico positivos pero en este software también hay falsos positivos que pueden obstaculizar la detección temprana y el tratamiento pero el sistema de IA determino el cáncer de mama en 48

de 71 personas analizadas por lo fue rápido caso contrario hubiera demorado analizar tal cantidad por un radiólogo y no se habrían detectado. (Ávila,2020)

- **Tratamiento:** En el tratamiento a base de la IA se han conocido diverso software que están dentro de equipos electrónicos los cuales ayudan a poder realizar un mejor tratamiento, por ejemplo se aplicó IA después de una cirugía prostática en el cual se aplicó lenguaje natural para realizar predicción en los informes medico detectando frases como por ejemplo (irritado, cansado, en buen estado) logrando realizar un diagnóstico final del 72.32% con lo que se creó un conocimiento clínico que tendrías los pacientes operados. Se plantea en un futuro tener robots quirúrgicos los cuales ayudaran a poder realizar tratamiento personalizados a paciente. (Ávila,2020)
- **Seguimiento, soporte y monitorización:** Con el avance de la IA se van creando los asistentes robóticos en los cuales se pretende recopilar información para dar un seguimiento de los pacientes en un hospital y acompañamiento. Esto permitirá que se muevan en un espacio determinado, puedan escuchar al paciente dar alertas a los especialistas y sobre todo también pueden brindar una medicación ya que estarán programados para ello actualmente hay una predicción beta de Pillo el cual es un robot que reconoce la voz, y da medicación a la hora correcta conjuntamente tiene un sistema de recomendación alimenticia y recopila información lo cual ayuda a determinar si una persona puede curarse en el tiempo correcto. (Ávila,2020)

2.3.3 Técnicas de Inteligencia Artificial

Una técnica de la Inteligencia Artificial es un método que utiliza conocimiento representado de tal forma que desde el punto de vista de la Inteligencia Artificial utiliza

diversas herramientas en la solución de problemas, estas herramientas se presentan en distintas técnicas, entre las técnicas básicas podemos mencionar:

- **Búsqueda de soluciones:** Este concepto se refiere a poder resolver problemas sin un método directo como una estructura o técnica directo entonces la IA indica que se puede encontrar una solución haciendo una búsqueda basada en comportamientos ya que no hay una base en lo cual podemos aplicar arboles de decisión, matrices, búsqueda binaria y demás. (Jones, 2024)
- **Representación del conocimiento:** La IA usa el conocimiento para poder representar una estructura de información con la cual se puede trabajar es por ello por lo que se de reconocer las diferentes formas de conocimiento:
  - **Conocimiento general:** Puede presentarse como fórmulas matemáticas o lógicas, o de manera informal, el lenguaje hablado / escrito.
  - **Conocimiento procedural:** Son secuencias de acciones a seguir, se pueden representar mediante diagramas de flujo, algoritmos, etcétera.
  - **Conocimiento factual:** Son hechos.
  - **Metaconocimiento:** Conocimiento sobre el conocimiento. Puede ser una forma extremadamente importante de conocimiento, sobre todo en sistemas que aprenden. (Anderson, 2023)
- **Reconocimiento de patrones:** El reconocimiento parte de poder clasificar los subgrupos con características comunes en el conjunto que es un grupo con el cual se puede tener muchas conclusiones diferentes y es ampliamente usado en los campos del reconocimiento de lenguaje natural, la visión por computadora,

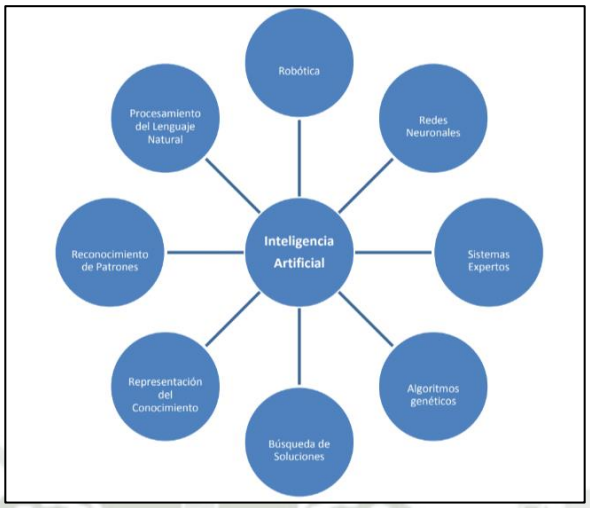
reconocimiento de imágenes, reconocimiento de señales, el diagnóstico de fallos de equipos, el control de procesos, etcétera. (Ochoa,2014)

- **Procesamiento del lenguaje natural:** El lenguaje natural o también llamado ordinario es que se usa para comunicarnos, pero tiene sus reglas y convenciones lingüísticas y sociales por lo cual se transmite conocimiento. Entonces el procesamiento del lenguaje natural con la IA usa las palabras convencionales para poder hacer una búsqueda y dar una solución, pero al usar ese tipo de lenguaje debe estar entrenado para reconocer el habla escrita con una correcta escritura y es más estudios han podido lograr que humanoides interpreten el habla humana y den respuestas similares. (Vásquez, 2009)
- **Robótica:** La robótica no supervisada nació de la mano de la IA ya que se crean robots inteligentes capaces de funcionar sin una supervisión humana. La robótica se entiende como la conexión inteligente entre la percepción y la acción, la robótica tiene por objetivo diseñar y desarrollar máquinas que sean capaces de realizar procesos mecánicos y manuales mediante la interacción de un sistema de control y un sistema sensorial con el que cuentan, permitiendo así, responder a los cambios que surgen en el entorno del mundo real. (Yagüe, 2020)
- **Redes Neuronales:** Las redes neuronales son una técnica de la IA en la cual asemejan ser una neurona porque tienen diversas conexiones y también puede tener varias capas en la cual el procesamiento no es lineal sino transversal el cual realiza una labor de aprendizaje en donde cada salida es un entrenamiento y mientras más veces se efectué tendrá más posibilidad de aprender Las redes neuronales almacenan la información de manera distinta que las computadoras tradicionales.

Una de las ventajas de utilizar las redes neuronales es que pueden seguir funcionando, aunque se destruyan algunas de sus neuronas, esto es atribuido a su estructura de red. Las arquitecturas de las redes neuronales pueden usarse para tareas como la visión; mecanismos de aprendizaje para el reconocimiento de voz, de forma que sus resultados alimenten programas simbólicos de la IA. (García,2019)

- **Algoritmos genéticos:** Estos algoritmos se basan en emular el proceso de selección natural en donde los mejores individuos luchan por sobrevivir y hacen que sus características se mantengan en generaciones posteriores. Se puede considerar a estos algoritmos, como un procedimiento de búsqueda y optimización. Una de las características de los algoritmos genéticos es que tiene la capacidad de "castigar" las malas soluciones, y de "premiar" a las buenas, de forma que estas últimas se propaguen con mayor rapidez. (Davis, 2023)
- **Sistemas expertos:** También son llamados sistemas basados en el conocimiento, dichos sistemas almacenan el conocimiento para un campo determinado y de la solución mediante la deducción lógica o conclusión y esto se da a través de un software que imita el comportamiento experto de un humano dando la solución de un problema. Un sistema experto es un programa de computadora interactivo que incorpora juicios (opiniones), experiencias, reglas de evaluación, intuición y otras habilidades para poder proveer asesoría inteligente sobre diversas tareas, mismas que resuelven problemas complejos empleando modelos de razonamiento humano, para llegar a soluciones idénticas a las que podría llegar un experto humano que se enfrentara al mismo problema. (Collins, 2023)

Figura. 5.
Técnicas de Inteligencia Artificial



Nota: Facultad de Ingeniería UNAM,2020

2.3.4 Tipos de Inteligencia Artificial

La IA esta desarrollada con algoritmos y modelos matemáticos que permiten a las computadoras hacer un entrenamiento y aprender de los datos para tomar decisiones y realizar tareas que se asemejan a la inteligencia humana.

Existen varios tipos de inteligencia artificial, los cuales son:

- **Sistemas que piensan como humanos:** Es el esfuerzo por hacer que las computadoras tengan una mente y un sentido amplio para saber cómo actuar a un determinado problema. (Robinson, 2023)
- **Sistemas que actúan como humanos:** Se basa en cómo hacer que las computadoras hagan las cosas que una persona puede hacer mejor tratando de superarlas en tareas cotidianas. (Robinson, 2023)
- **Sistemas que piensas racionalmente:** Es el estudio de las facultades mentales a través de estudios de modelos computacionales. (Robinson, 2023)

- **Sistemas que actúan racionalmente:** Es un campo en donde busca emular y explicar el comportamiento inteligente en términos de procesos computaciones de una persona. (Robinson, 2023)

2.4 Machine Learning

El machine learning o bien llamado aprendizaje automático es una técnica de la inteligencia artificial para la detección automática de patrones relevantes dentro de un conjunto de datos. En la última década se ha vuelto muy usada para el análisis de datos y aprendizaje de maquina es por ello por lo que se usa grandes bases de datos para ellos y unos ejemplos puede ser el filtro de correos electrónicos como spam, los sistemas de recomendaciones, el reconocimiento facial y últimamente está ya en el campo de la medicina, marketing, logística y la industria 4.0. El ser humano puede crear una herramienta de aprendizaje automático, pero existen diferentes tipos el aprendizaje automático supervisado que involucra la intervención humando y el no supervisado que se da por aprendizaje de máquina, el funcionamiento de cada uno depende del problema a tratar, pero principalmente en ambos se necesita un volumen de datos importante para que el algoritmo asociado a una maquina sea el correcto. (M. Alberto, 2019)

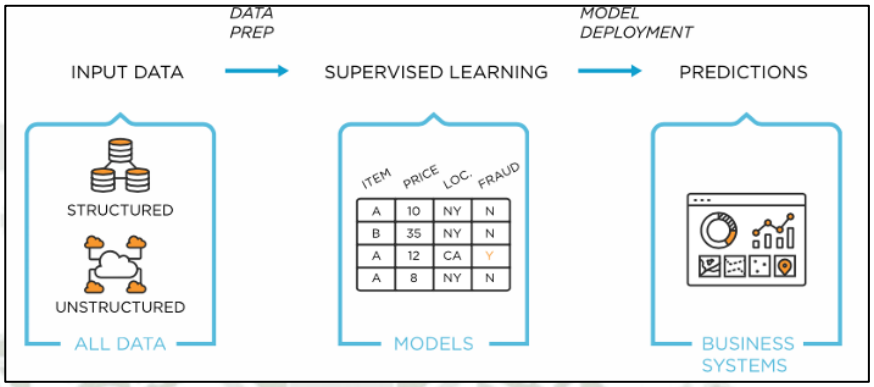
2.4.1 Tipos de Aprendizaje Automático

El aprendizaje automático es una rama de la Inteligencia Artificial que tiene como objetivo lograr que las computadoras aprendan sin haber sido explícitamente programadas para ello.

- **Supervisado:** Un tipo de aprendizaje es el supervisado en donde mientras más se entrene tendrá mejor resultado. Para ello este tipo de aprendizaje puede usar un conjunto de etiquetas con características similares el cual puede ser insertado en un

modelo para poder entrenar y clasificar el conjunto de datos para que el entrenamiento de mejores resultados en el tiempo. (Santana F., 2023)

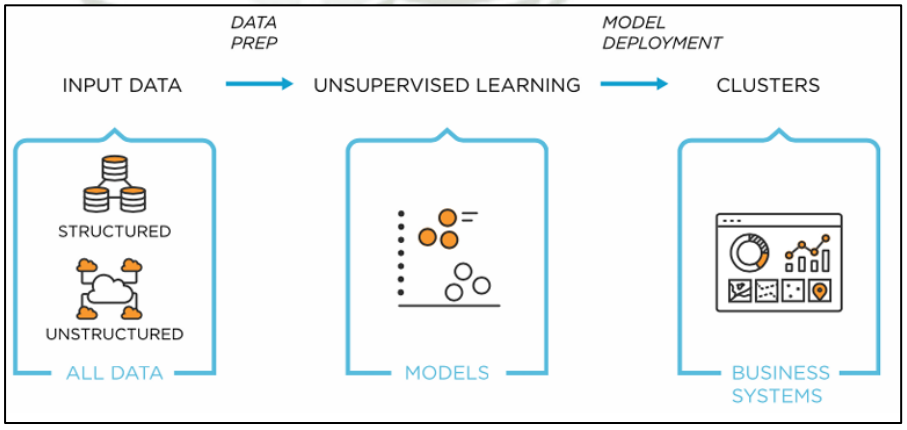
Figura. 6.
Esquema de Aprendizaje Supervisado



Nota: Santana F., 2023

- **No supervisado:** En este tipo de aprendizaje no se necesita tener etiquetas como en el supervisado, sino que se tiene un conjunto multimodal de opciones y se quiere encontrar una relación entre ellos agrupándolos para poder saber su similitud y trabajar con ese conjunto para dar un resultado. (Santana F., 2023)

Figura. 7.
Esquema de Aprendizaje No Supervisado



Nota: Santana F., 2023

- **Por refuerzo:** En el tipo de aprendizaje mencionado se basa en la prueba y el error con recompensa o un castigo antes de poder hacer las comparaciones correspondientes y también se tiene que entrenar para tener mejores resultados a largo plazo. (Santana F., 2023)
- **Profundo:** Existe un nuevo enfoque de aprendizaje profundo en el cual se denomina así por el número de capas ocultas en una red artificial en donde el análisis es transversal y el procesamiento se compara con muchas posibilidades dado que se tiene un conjunto de entrada, luego se procesa y finalmente se obtiene una salida que por más repeticiones que se realice mejor será el aprendizaje. (Santana F., 2023)

2.5 Algoritmos de Aprendizaje Automático Supervisado

2.5.1 Regresión Lineal

La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente que es el resultado y una o más variables independientes que son la predictoras para encontrar el resultado. La idea básica detrás de la regresión lineal es encontrar la mejor línea recta que se ajuste a los datos observados, de manera que pueda utilizarse para predecir valores futuros de la variable dependiente basándose en los valores de las variables independientes. Es una de las herramientas más básicas y ampliamente utilizadas en estadística y aprendizaje automático, y se aplica en una amplia gama de campos, como la economía, la biología, la ingeniería, etc. (Vásquez, C., 2023)

2.5.2 Máquina de Soporte Vectorial (SVM)

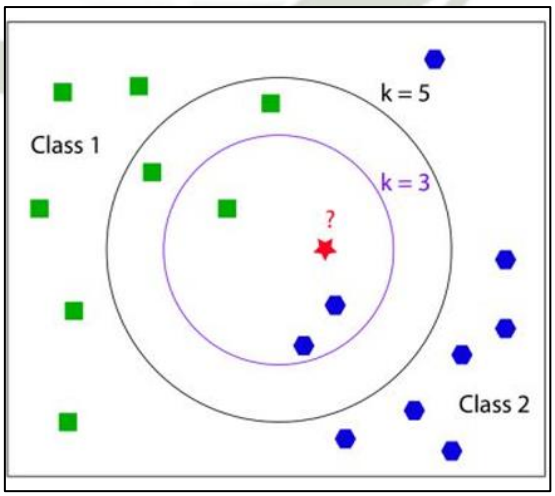
Las máquinas de soporte vectorial es un tipo de aprendizaje supervisado. Son un conjunto de algoritmos que fueron desarrollados por Vladimir Vapnik y su equipo. Estos

algoritmos resuelven problemas de clasificación y regresión. Podemos tener un conjunto de datos y agruparlos por etiquetas para poder predecir por ejemplo una muestra. El SVM construye un hiperplano o un conjunto de hiperplanos dentro de un espacio muestral en donde puede realizar una clasificación o regresión separando las clases se puede hacer una clasificación correcta. (Sarmiento H., 2020)

2.5.3 KNN Vecinos Cercanos

Es un algoritmo no supervisado que se basa en la clasificación de patrones y es un clasificador no paramétrico, es considerado uno de los mejores con desempeño en el ambiento de aprendizaje automático. KNN construye un modelo sencillo de resolver para realizar predicciones y es ampliamente usado en el área logística y de ventas. El algoritmo usa la proximidad de una valor o atributo para realizar clasificaciones o predicciones sobre un conjunto o agrupación de un punto de datos individual. Realizar una normalización a los datos de prueba puede mejorar la exactitud del algoritmo y así realizar una mejor predicción. (Narváez M., 2022)

Figura. 8.
Ejemplo de Clasificación del Algoritmo KNN Vecinos Cercanos

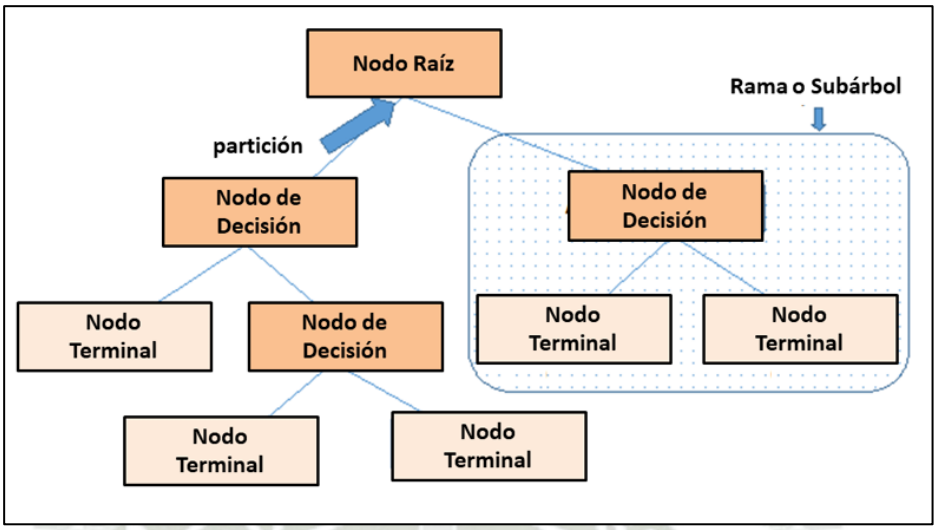


Nota: Revista Perspectivas, 2022

2.5.4 Arboles de Decisión

Un árbol de decisión es un tipo de algoritmo de aprendizaje supervisado el cual es muy usado para la predicción, es usado mayormente en la economía y su estructura se basa en un árbol como su nombre lo detalla ya que se tiene un nodo raíz y nodos hojas los cuales harán una comparación hasta encontrar una similitud, su objetivo es predecir a que clase pertenece un caso o atributos. En cada paso del entrenamiento el algoritmo realiza sucesivas divisiones o particiones de un subconjunto de datos a partir la aplicación de una decisión asociada a una de las variables, por ende, separando a esos datos en dos nuevos subconjuntos (de allí lo de partición binaria). Siguiendo con este proceso en forma recursiva hasta un cierto punto previamente estipulado en el que el proceso de bifurcación se detiene, obtendremos finalmente el clasificador por árbol de decisión. Cada nuevo dato, del que conoceremos el valor de sus atributos, recorrerá las sucesivas ramificaciones del árbol, a partir de las reglas y decisiones generadas mediante el proceso recién descripto. Aunque mediante este tipo de algoritmos de árbol de decisión podemos generar modelos predictivos tanto para variables objetivo de tipo cuantitativa (regresión) como de tipo cualitativa o categórica (clasificación). (Arana C., 2021)

Figura. 9.
Estructura de Funcionamiento de un Árbol de Decisión

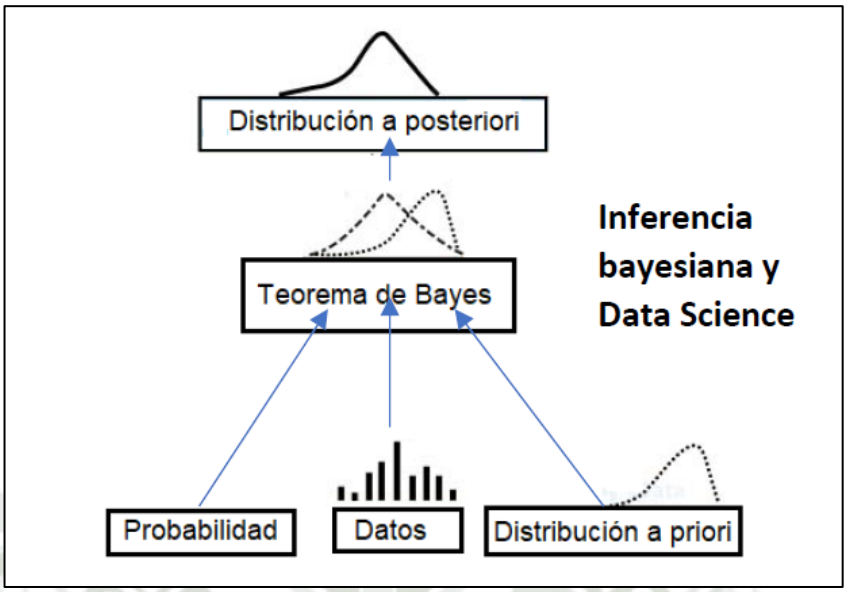


Nota: Arana C., 2021

2.5.5 Análisis Bayesiano

El análisis o inferencia bayesianos es un análisis estadístico que se basa en la hipótesis de que un suceso suceda o no. Este análisis se basa en la probabilidad subjetiva. En lugar de tratar las probabilidades como frecuencias relativas de ocurrencia de eventos, como se hace en el enfoque frecuentista tradicional, el análisis bayesiano considera las probabilidades como representaciones de incertidumbre o creencias subjetivas. El análisis bayesiano tiene varias ventajas, como la capacidad de incorporar información previa de manera explícita, la flexibilidad para manejar conjuntos de datos pequeños o incompletos y la capacidad de proporcionar estimaciones de incertidumbre directamente. Se puede trabajar con un pequeño grupo de datos, pero de calidad. (Hernández R., 2024)

Figura. 10.
Esquema de Funcionamiento del Análisis Bayesiano

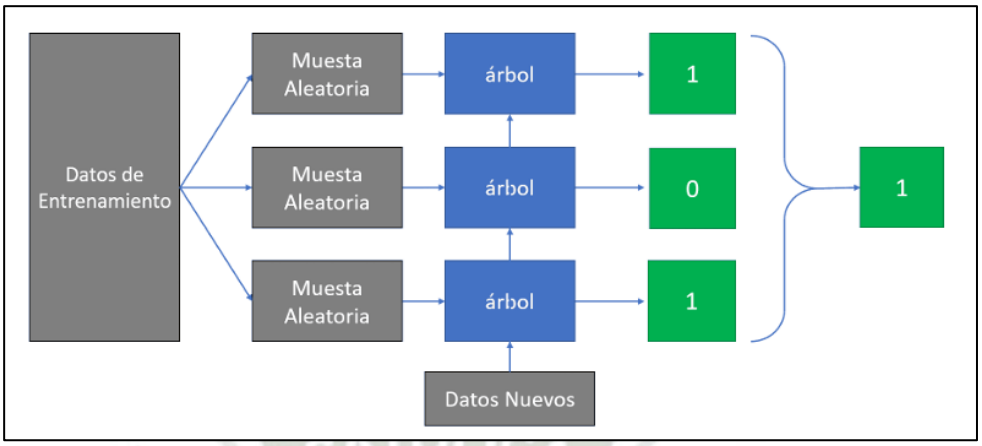


Nota: Hernández R., 2024

2.5.6 Bosques Aleatorios

También conocido como Random Forest, es un método de aprendizaje supervisado que se utiliza para clasificación y regresión, combina múltiples modelos de predicción para mejorar la precisión general del modelo. Un bosque aleatorio está compuesto por una colección de árboles de decisión individuales, donde cada árbol se construye de manera independiente utilizando una muestra aleatoria del conjunto de datos de entrenamiento. Además, durante la construcción de cada árbol, en cada nodo se elige una característica aleatoria para dividir los datos, lo que ayuda a reducir la correlación entre los árboles y mejora la generalización del modelo. Son ampliamente utilizados en una variedad de aplicaciones, incluyendo reconocimiento de patrones, bioinformática, y análisis de datos médicos y financieros, entre otros. (Botana, J., 2021)

Figura. 11.
Esquema de Funcionamiento del Algoritmo de Bosques Aletorios

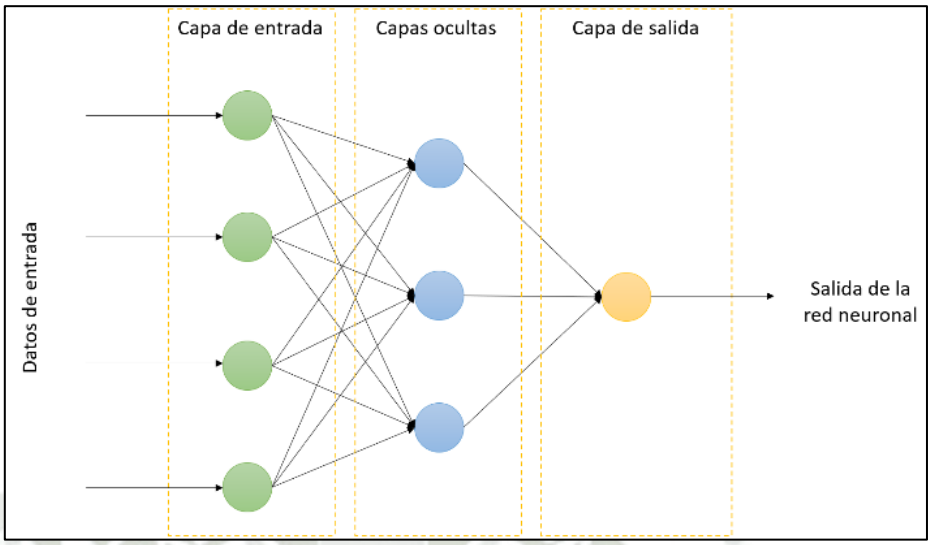


Nota: Botana J., 2021

2.5.7 Red Neuronal

Una red neuronal se entiende como un sistema que procesa información basándose en un método es específico para encontrar un resultado y esto se basó en el funcionamiento de las neuronas que al momento de dar una respuesta es porque la busco, la analizo y encontró una respuesta a una determinada pregunta. Cada punto de una red neuronal es una unidad de procesamiento que tiene una entrada, el propio procesamiento y la salida la cual vendría a ser el resultado, pero para ello se deben sumar el resultado de cada nodo con las salidas determinando un valor resultante y su arquitectura se basa en la de un computador de Von Neuman, pero para este se usa un aprendizaje de máquina. (S. Antonio, S. Emilio, D. Martin, 2010)

Figura. 12.
Esquema de Funcionamiento de una Red Neuronal



Nota: Verrastro Claudio,2012

2.6 Estado del Arte

1. Construcción de un Modelo de predicción para apoyo del diagnóstico de la diabetes (2018)

En el año 2018 Orlando A. Chan presento una investigación realizada sobre un conjunto de datos de 768 pacientes, donde todos los registros son basados en mujeres para la detección de diabetes gestacional, mencionan que dichos atributos en el conjunto de datos son de alta importancia para la detección de DMT2. Algunas de las variables consideradas son: glucosa, insulina, presión sanguínea y edad. El objetivo final de los autores fue crear un sistema experto para detectar diabetes a partir de los atributos seleccionados del conjunto de datos utilizando algoritmos de clasificación proporcionadas por la herramienta WEKA y BigML. Los autores utilizan como clasificador árboles de decisión y obtienen un 70% de precisión en la clasificación de pacientes que no presentan DMT2, un 63% para los que si presentan y un 73.83% de exactitud. En este trabajo realizaremos un tratamiento de datos faltantes a un conjunto de datos con el fin de utilizarlos en mejorar la clasificación usando el algoritmo J48. Posteriormente se eliminan

atributos no relevantes del conjunto utilizando análisis de componentes principales (PCA) y se realiza una comparativa de los resultados, para demostrar que es factible el uso de tratamiento de datos y la reducción de términos con PCA sin perder exactitud en la clasificación. (A. Chan,2018)

2. ***Predicción de la Diabetes mediante el conjunto de diferentes clasificadores de aprendizaje automático (2020)***

El presente artículo analiza el uso de diferentes clasificadores de aprendizaje automático para predecir la diabetes, centrándose en técnicas como redes neuronales, árboles de decisión, bosques aleatorios y métodos de conjunto. Los autores enfatizan la importancia de la validación cruzada, la selección de modelos y las técnicas de preprocesamiento, como el rechazo de valores atípicos y el llenado de valores faltantes, para mejorar la precisión de los modelos de predicción de la diabetes. Proponen un marco sólido para la predicción de la diabetes utilizando un conjunto de clasificadores, logrando un rendimiento superior en comparación con los métodos existentes. El artículo enfatiza la importancia de las técnicas de preprocesamiento, la selección de características y los métodos de conjunto para mejorar la precisión de la predicción de la diabetes. El apéndice proporciona algoritmos para la selección de características y clasificadores de aprendizaje automático para la predicción de la diabetes, selección de características basada en correlación, k-vecino más cercano, árbol de decisión, AdaBoost, bosque aleatorio, Naive Bayes y XGboost. El clasificador de conjunto propuesto demostró un rendimiento superior en comparación con los métodos de última generación existentes, lo que demuestra el potencial del aprendizaje automático en aplicaciones sanitarias. El apéndice proporciona algoritmos para la selección de funciones y clasificadores de aprendizaje automático para la predicción de la diabetes, lo que ofrece un recurso integral para mejorar la precisión de los modelos de predicción de la diabetes. (Kamrul H., 2020)

3. Clasificación de pacientes según su posibilidad de adquirir diabetes mellitus empleando algoritmos de machine Learning (2020)

El presente trabajo tiene como base otra investigación titulada "Uso de Algoritmo de aprendizaje ADAP para pronosticar la aparición de diabetes Mellitus" presentada por J. Smith, J. Everhart, W. Dickson, W. Knowler y R. Johannes, en el Proceedings of the Annual Symposium on Computer Application de 1988. Para ello el algoritmo escogido es el denominado Two- Class Boosted Decision que es aplicado desde Azure Machine Learning Studio para lo cual se basa en un árbol de decisión impulsado que es un filtro para lo cual en el aprendizaje el segundo árbol corrige los errores del primero y así sucesivamente hasta encontrar el mejor resultado posible. El principal objetivo es comparar el rendimiento de los clasificadores y los árboles de decisión en la predicción de la diabetes mellitus o diabetes tipo 2 para lo cual se tuve en el conjunto de variables en su mayoría numéricas y solo una dependiente de tipo texto lo cual representaba el riesgo significativo de contraer la diabetes. Teniendo como resultado final la estimación del 65% usando los árboles de decisión lo cual en la primera vuelta se dio el 70% y 90% de confianza, pero hubo muchos falsos positivos. (J. Pincay et all, 2020)

4. Mecanismos de clasificación de diabetes mellitus en la población de Aguascalientes, México. (2020)

En esta investigación se diseñó, desarrollo y puso a punto un mecanismo integrado por dos mecanismos que fueron ejecutados de forma secuencial, el primero fue utilizando la técnica de selección de variables de testores típicos, con este mecanismo se obtuvieron los sub conjuntos de características interrelacionadas que mejor describen un paciente con la patología de diabetes mellitus, en el segundo mecanismo se empleó la técnica de redes neuronales artificiales con la cual se diseñó, desarrollo, entreno y se validó un clasificador con una precisión de casi el 92%. Cabe mencionar que en esta investigación se contó con información extraída de consultas de pacientes

del sistema expediente clínico electrónico del ISSEA, el cual basa sus diagnósticos en el estándar internacional CIE 10. Como resultados de esta investigación se desarrolló un clasificador con una precisión de casi el 92% para identificación de pacientes con diabetes mellitus en la población de Aguascalientes, y por medio de la selección de variables se identificaron variables que no son comúnmente consideradas en conjunto en la literatura, tales como el estado civil, existencia de complicaciones y antecedentes familiares con diabetes. (P. León et all, 2020)

5. ***Detección y Predicción de la Diabetes mediante Minería de Datos: Una Revisión Integral (2021)***

El presente paper es presentado por Farrukh Aslam en el año 2021 detalla una revisión de las técnicas clave en la extracción y detección para la predicción de la diabetes. Para lo cual explica el uso de Redes Neuronales que es ampliamente usado en la predicción y clasificación logrando una precision en la detección de la diabetes impresionante. Al igual que el uso de Arboles de Decision, K- vecinos más cercanos, Bosques aletorios y Maquinas de Vectores de Soporte que ayudan a los profesionales en el campo del análisis a realizar modelos muy eficientes. Los hallazgos encontrados son en base a las pruebas de los algoritmos para que los profesionales de la salud puedan usarlos en la detección de enfermedades, al utilizar algoritmos avanzados como redes neuronales artificiales y árboles de decisión, los profesionales pueden realizar evaluaciones más precisas de los niveles de riesgo de los pacientes y adaptar los planes de tratamiento en consecuencia. Las técnicas de extracción de datos pueden servir como herramientas valiosas para el apoyo a las decisiones clínicas en el cuidado de la diabetes. Al integrar modelos predictivos y métodos de análisis de datos en los sistemas de atención médica, los profesionales pueden tomar decisiones informadas sobre la atención al paciente, la gestión de medicamentos y las intervenciones en el estilo de vida. En general, los hallazgos presentados en la investigación ofrecen información valiosa e implicaciones prácticas para los profesionales e investigadores de

la salud, permitiéndoles aprovechar las técnicas de extracción de datos para mejorar el control de la diabetes y los resultados de los pacientes. (Aslam, 2021)

6. ***Sistemas de información para la red neuronal convolucional en la detección de diabetes usando imágenes de fondo de ojo (2021)***

En la presente investigación aborda el problema de procesamiento de la detección de la diabetes usando el análisis de fondo de ojo mediante el algoritmo PCA que es usado mayormente para poder determinar patrones que ayuden a saber si una persona tiene o no diabetes. Este algoritmo se basa en poder crear redes neuronales en base a esos patrones pudiendo saber que si una persona tiene una pigmentación media amarilla en el fondo del ojo es que puede tener prediabetes o diabetes ya que por falta de azúcar la retina del ojo empieza a sufrir daños. Los resultados obtenidos fueron validados mediante fórmulas de precisión y exactitud logrando alcanzar que la exactitud fue de 86.92%, con la precisión de 83.87% en un total de 668 registros o imágenes recopiladas del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales, en particular el estudio se basó en mujeres de al menos 21 años en adelante y de ascendencia hindú encontrando 500 registros no diabéticos y 268 diabéticos. (Villegas Cubas,2021)

7. ***Clasificación de Diabetes Mellitus Tipo II detectando factores de riesgo en un conjunto de datos (2021)***

El propósito de este trabajo es la detección de factores de riesgo que pueden originar la DMT2, a partir del análisis de componentes principales en un conjunto de expedientes clínicos. Posteriormente se aplica un algoritmo de aprendizaje automático para corroborar que esos factores contribuyen en la detección de DMT2, al mejorar la precisión en los resultados experimentales. Existen dos variantes de la DM, tipo I y tipo II. La más común es la Diabetes Mellitus tipo II (DMT2), la cual es una enfermedad en la que el organismo no genera suficiente insulina para procesar la glucosa en la sangre dejando mucho de este material circulando en el sistema

sanguíneo. En México se atribuye el 11.8% de muertes desde el 2005 y un 63% como causa de muerte principal de enfermedades crónicas a nivel mundial en el 2015. (Cancino. J., 2021)

8. Una Revisión de las Implementaciones de Sistemas para la Identificación de Tendencias para la Diabetes. (2022)

En esta investigación aborda la diabetes desde el punto de análisis de redes neuronales donde menciona que la diabetes es un trastorno metabólico considerado crónico para la sociedad que puede tener complicaciones a futuro como enfermedades cardiovasculares, accidentes cardiovasculares, insuficiencia renal crónica y otras. Según la Organización Mundial de la Salud para el 2045 el número de pacientes con diabetes mellitus alcanzara los 629 millones de personas en el mundo. Es por ello que esta investigación aborda la necesidad de tener una predicción para saber potenciales pacientes en contraer esta enfermedad por eso el sector salud en el mundo promueve el uso de sistemas de apoyo de diagnóstico y prevención de la diabetes con el objetivo de reducir los efectos, utilizando dispositivos inteligentes y sensores combinando la IoT y la Inteligencia Artificial la cual tiene el potencial de usar Notas de datos de los pacientes que se pueden obtener en entornos clínicos y también fuera de ellos. (Coral, 2022)

9. Predicción de la Diabetes con Aprendizaje Automático Fusionado (2022)

El artículo de investigación presentado propone un sistema de apoyo en la a la toma de decisiones para la diabetes basado en aprendizaje automático. El modelo propuesto logró una precisión del 94.87%, superando a los sistemas existentes. El estudio destaca la importancia del diagnóstico temprano en el control de la tasa de mortalidad de la diabetes. Se compara el rendimiento del modelo propuesto con otras técnicas de aprendizaje automático y se demuestra su superioridad. Para ello utiliza una combinación de Maquina de Vectores, Redes Neuronales y Lógica difusa para predecir la diabetes. El conjunto de datos utilizado tiene 17 atributos

relacionados con los síntomas diabéticos. El modelo se entrena y valida con el 30% de los datos restantes, logrando una precisión de predicción del 92.31%. Los resultados muestran predicciones precisas tanto para casos positivos como negativos de diabetes. El modelo puede ser utilizado para el diagnóstico en tiempo real y la clasificación de la diabetes. Además, propone un Modelo Fusionado para la Predicción de la Diabetes (FMDP) utilizando modelos de Máquina de Vectores de Soporte (SVM) y Redes Neuronales Artificiales (ANN). El conjunto de datos se divide en datos de entrenamiento y prueba, y la salida de estos modelos se utiliza como entrada para un modelo de lógica difusa para determinar el diagnóstico de la diabetes. El modelo propuesto tiene una precisión de predicción del 94.87%. (Ahmed, G., 2022)

10. Desarrollo de un Sistema Inteligente de Control de Diabetes Tipo 1 Basado en Modelos Predictivos. (2022)

El número de pacientes que padecen de diabetes mellitus tipo 1 ha ido incrementando significativamente durante los últimos 40 años. Es estimado que en 1980 había cerca de 108 millones de adultos con diabetes. Hoy en día se conoce que afecta a más de 460 millones de personas en el mundo (9.3% de la población mundial) con una prevalencia global en crecimiento. La diabetes mellitus tipo 1 (T1DM) se trata de la variante más peligrosa, donde el páncreas no produce ninguna insulina de manera autosuficiente. Para remediar el problema planteado, en este trabajo se van a usar técnicas de Deep Learning (DL, por sus siglas en inglés) como forma de optimizar los bolos de insulina inyectados por el paciente, en base a sus valores de glucosa. La mayoría de las técnicas y Desarrollo de un sistema inteligente de control de diabetes de tipo 1 basado en modelos predictivos que tienen la capacidad de predecir resultados futuros con un conjunto de datos previos. El objetivo principal de este trabajo es ampliar la investigación actual sobre la predicción de glucosa en sangre para pacientes T1DM utilizando DL, evaluando modelos basados en redes LSTM y comparando a su vez con modelos matemáticos basados en modelos

estadísticos que utiliza variaciones y regresiones de datos estadísticos con el fin de encontrar patrones para predicción hacia el futuro; además se desarrollará una aplicación para pacientes T1DM en la que podrán obtener gráficas con predicciones de sus valores de BG. (Carabias, 2022)

11. Predicción de la Diabetes basada en el aprendizaje automático mediante datos multisensoriales (2023)

El estudio elaborado por Jari Aditi detalla el uso de algoritmos de aprendizaje automático con el uso de múltiples sensores para predecir la diabetes. La investigación encontró el uso de una serie de combinación de sensores en glucosa como los Electrocardiogramas y Acelerómetros los cuales brindaron una mayor predicción del 98.2% usando el algoritmo XGBoost. Estos sensores usan un método no invasivo para predecir la diabetes utilizando señales fisiológicas. Los hallazgos para la predicción fue usar datos usados por sensores y puesta a prueba del algoritmo de aprendizaje automático que mostró la mayor precisión de predicción en el estudio fue el algoritmo de aumento de gradiente extremo (XGBoost), Este algoritmo superó a otros algoritmos en la predicción de la diabetes utilizando datos multisensor. (Aditi Site, J., 2023)

12. Diabetes, Inteligencia Artificial e Internet de las Cosa Medicas: Nuevas Tendencias en la medicina (2023)

La diabetes Mellitus o llamada generalmente diabetes tipo 2 se caracteriza por tener altos niveles de glucosa en sangre lo que produce complicaciones en el cuerpo humano y según su tratamiento este puede desencadenar hasta la muerte. Por ello el aprendizaje de maquina o Machine Learning ayuda a poder hacer predicciones en donde tiene 2 maneras de clasificarse existe el aprendizaje supervisado que interviene una persona y el no supervisado el cual encuentra patrones en los datos para determinar un resultado a esto se le suma los equipos de IoT para el monitoreo de la glucosa por ejemplo a través de pulseras determinando una alerta si esta se encuentra elevada. Este paper abarco todos los casos de interconexión entre dispositivos inteligente con el análisis de

datos en favor al diagnóstico y se basa en algoritmos de decisión lo cual ayuda a medirlo por 3 métricas: la sensibilidad que mide los resultados verdaderamente positivos, la especificidad que mide los resultados verdaderamente negativos y la precisión que mide la proporción de los resultados positivos y esto ha tenido efecto por ejemplo en modelos para predecir el cáncer. (Sociedad Mexicana de Inteligencia Artificial, 2023)

13. El Internet de las cosas médicas (IoMT): Una revolución tecnológica aplicable a la gestión de la diabetes tipo 1 (2023)

La Diabetes tipo 1 o bien llamada diabetes juvenil es un problema mundial y su inicio se deriva de un mal congénito o es hereditario lo cual hace que el páncreas no produzca la insulina suficiente, la insulina es una hormona esencial para que la glucosa genera energía en el cuerpo y la convierta en ATP que es Adenosin TriFosfato. Por ello el avance de la tecnológica permitió poder hacer que el internet de las cosas incursione en la medicina y este concepto es llamado IoMT ya que involucra dispositivos físicos conectados a internet recopilando datos y permitiendo hacer un análisis y diagnostico que ayude a los especialistas a determinar el avance de la enfermedad en un paciente o mantener el control de este. La base de la recopilación de estos datos se detalla en los sensores que tienen los equipos para el monitoreo de la diabetes tipo 1. Esta investigación ofrece la visión general de como la tecnología IoT puede utilizarse en el monitoreo e interpretación de nos niveles de glucosa gracias a los biosensores que pueden recopilar grandes cantidades de datos y mediante una plataforma ser procesados y tener un resultado estimado del estado de un paciente teniendo en cuenta los riesgos y desafíos de la tecnología en el futuro. (UmaEditorial,2023)

14. Prediagnóstico Medico de la Diabetes Mellitus tipo 2 mediante Machine Learning (2023)

En el presente artículo se aborda la necesidad de aplicar el machine learning o aprendizaje automático para la detección de la diabetes tipo 2 la cual es muy común y la más mortal en el

mundo. Las personas que padecen esta enfermedad reciben un tratamiento que se basa en suministrarse insulina mediante inyecciones por vis subcutánea o por bomba de fusión continua, esta enfermedad no tiene cura, pero se puede manejar teniendo un estilo de vida saludable, todos los efectos secundarios de la diabetes se pueden medir mediante la implementación de sistemas inteligentes de datos basados en redes neuronales. La presente investigación presenta un sistema que realizará el prediagnóstico medico no invasivo que permitirá determinar si el paciente padece diabetes mellitus tipo 2. Se detallo que se usó 3 redes neuronales de aprendizaje supervisado y se usó datos generados por un médico experto con los registros de 50 casos clínicos. La red neuronal seleccionada solo tiene una capa de entrada y de salida en la cual juntando la probabilidad de obtener el mejor resultado se obtuvo 90% de los casos analizados favorables para el prediagnóstico. (Revista Innovación Digital y Desarrollado Sostenible, 2023)

15. Estudio de algoritmos de inteligencia artificial más utilizados para el diagnóstico de diabetes mellitus tipo 2 (2023)

El principal objetivo de este trabajo es realizar una revisión sistémica en la aplicación de técnicas o algoritmos de inteligencia artificial para el diagnóstico de diabetes tipo 2. Se planteo la siguiente pregunta de investigación: ¿Cuáles son las técnicas de inteligencia artificial utilizadas en el diagnóstico de la diabetes tipo 2? Llegaron a la conclusión que los algoritmos identificados como más significativos para la predicción y/o diagnóstico de Diabetes Mellitus Tipo 2 son los siguientes: Artificial Neuronal Network (ANN), Ran-dom Forest (RF) Support Vector Machine (SVM), Árbol de decisión, K-vecinos más cercano (KNN) y Regresión Logística (LR), de los cuales el mejor algoritmo para detección de diabetes mellitus tipo 2 es el algoritmo ANN debido que realiza el procesamiento de las variables de entrada por capas en un gran conjunto de datos. (Guamán, 2023)

16. Análisis Comparativo de Técnicas de Machine Learning sobre el método de muestreo para la predicción de diabetes (2023)

El presente trabajo realizado por Piero Chira y Kevin Rivera egresados de la Universidad Cesar Vallejo abordan la comparación de técnicas de machine learning para el método de muestreo en la predicción de la diabetes. Por ello usan factores de extracción de datos usando la herramienta Kaggle el cual tiene 768 instancias que consideraron como muestra. Usan la metodología KDD para poder hacer el proceso de análisis de datos en donde según la metodología tiene 5 fases, en la fase de evaluación determinación que usaran las métricas de rendimiento exactitud, precisión, especificidad, sensibilidad y F1 score. En los algoritmos de machine learning usado fueron 6, arboles de decisión, bosques aleatorios, máquina de vectores de soporte, K- vecinos más cercanos y Redes neuronales. Para la evaluación de los modelos obtuvieron que hubo 3 modelos que superaron el resultado esperado los cuales son los bosques aleatorios, arboles de decisión y Gradient Boosting Machine, donde respectivamente tuvieron 79,22%, 75,32% y 74,09% para la métrica de exactitud y por otro lado descubrieron que los modelos con menor aprendizaje fueron los de K-vecinos cercanos, redes neuronales y máquina de vectores. Llegando a la conclusión que el uso de algoritmos de aprendizaje automático si ayuda a la predicción y es un factor importante para el factor salud. (Rodríguez, 2023)

2.7 Metodologías de Ciencia de Datos

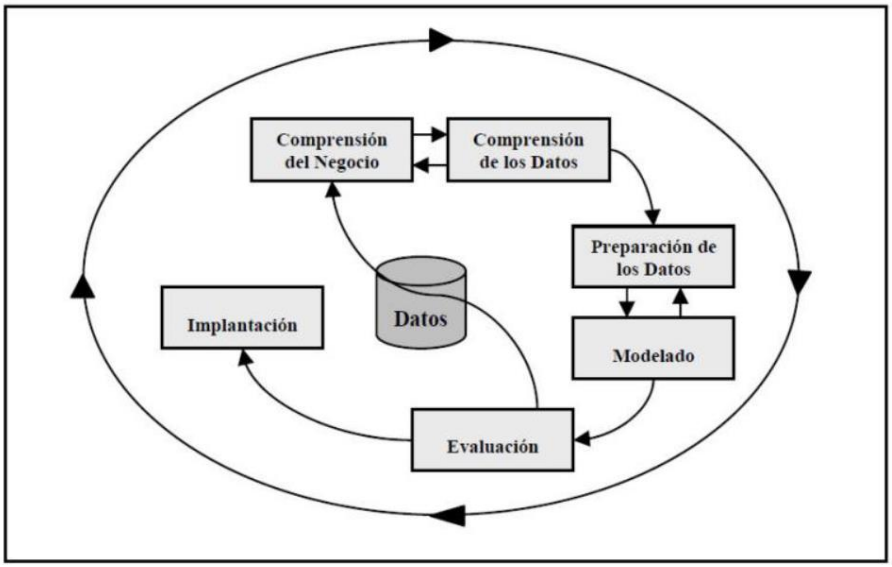
2.7.1 Metodología CRISP-DM

La metodología usada para el presente proyecto es la denominada CRISP-DM que se basa en poder realizar un tratamiento de los datos siguiendo el proceso de ciclo de vida de un proyecto que va desde la limpieza hasta determinar correlaciones entre los datos y tener un valor estándar para definir la funcionalidad que se le dará. La metodología CRISP-DM tiene cuatro niveles de abstracción, que están organizados de forma jerárquica en las

cuales hay tareas que van desde el nivel más general hasta el nivel más específicos haciendo que el conocimiento generado sea mucho más detallado y enriquecedor para en análisis de datos. (Brown, 2023)

La metodología en si está conformada por seis fases, en cada fase existen varias tareas generales que están relacionadas con los niveles posteriores. Las tareas generales se pueden desglosar a tareas específicas, donde se describen las acciones que deben ser desarrolladas. Es por ello por lo que, si en un nivel se tiene la acción o tarea general de "limpieza de datos", en el tercer nivel podemos encontrar un caso específico, como, por ejemplo, "limpieza de datos numéricos. Pare el cuarto nivel se recoge las acciones, decisiones y resultados sobre el proyecto de Data Mining en investigación. (Brown, 2023)

Figura. 13.
Fases de la Metodología CRISP-DM



Nota: R. Montequín, 2003

2.7.1.1 Fases de la Metodología CRISP-DM

La metodología como se mencionó está compuesta por 6 fases en las cuales hay una interacción hasta poder encontrar un resultado que ayuda a obtener

conocimiento y una salida válida para el modelo los cuales están compuestos por el Análisis del Problema, el Análisis de Datos, la Preparación de los Datos, el Modelado de los Datos, la Evaluación del Modelado y la Explotación, pero la interacción nace en las 4 primeras fases de la metodología. A continuación, se detalla cada una de ellas:

a) **Análisis del Problema:** La primera fase análisis del problema, incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación a corto plazo. (R. Montequín, 2003)

b) Análisis de Datos: Esta fase comprende la recolección inicial de datos, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis. (R. Montequín, 2003)

c) Preparación de los Datos: Para esta fase se incluye la selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. La fase de preparación de los datos se encuentra muy relacionada con la fase de modelado, puesto que en función de la técnica de modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas. Por lo tanto, las fases de preparación y modelado interactúan de forma iterativa. (R. Montequín, 2003)

d) Modelado de Datos: Para el modelado se seleccionan las técnicas más apropiadas para el proyecto de Data Mining específico. Las técnicas para utilizar en esta fase se seleccionan en función de los siguientes criterios:

- **Ser apropiada al problema**

- **Disponer de datos adecuados**
- **Cumplir los requerimientos del problema**
- **Tiempo necesario para obtener un modelo**
- **Conocimiento de la técnica**

Antes de proceder al modelado de los datos se debe de establecer un diseño del método de evaluación de los modelos, que permita establecer el grado de bondad de los modelos. Una vez realizadas estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos. (R. Montequín, 2003)

e) **Evaluación de Datos:** Para esta fase se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema. Se debe revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso en el que se hayan podido cometer errores. (R. Montequín, 2003)

f) **Explotación:** En la última fase si el modelo generado es válido en función de los criterios de éxito establecidos en la primera fase, se procede a la explotación del modelo con los datos. En esta fase también se debe de asegurar el mantenimiento del modelo y la posible difusión de los resultados. (R. Montequín, 2003)

2.7.2 Metodología SEMMA

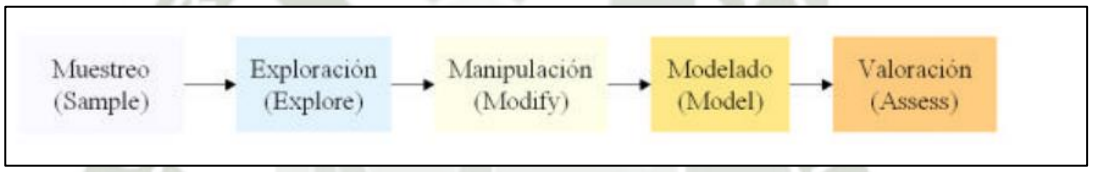
La metodología SEMMA se usa para el análisis de datos y la minería de datos , fue desarrollado por el SAS Institute (Empresa multinacional de software estadístico). La metodología se entiende como un proceso de selección el cual involucra una gran cantidad de datos los cuales se exploran y modelan para encontrar patrones de negocio desconocidos. La palabra en si es un acrónimo de las 5 fases de las cuales está conformada

los cuales están en inglés (Sample, Explore, Modify, Model y Assess). Proporciona un marco estructurado para llevar a cabo análisis predictivos de manera sistemática, desde la preparación de datos hasta la evaluación de modelos. (Jones A. M.,2023)

2.7.2.1 Fases de la Metodología SEMMA

La metodología como se mencionó está compuesta por 5 fases que son el Muestreo, la Exploración, la Manipulación, el Modelado y la Valoración en la cual se puede medir el resultado obtenido. A continuación, podremos verla en la Figura N° 14 y se detallara cada una de ellas:

Figura. 14.
Fases de la Metodología SEMMA



Nota: R. Montequín, 2003

a) **Muestreo (Sample):** Esta fase inicia con la determinación de la muestra con la cual se hará el análisis. El objetivo principal de esta fase es tener una muestra representativa para el análisis si no se cumple invalidaría todo el modelo y resultados, una forma común de selección es la selección al azar dentro de la población si es que no hay parámetros específicos o detallados de estudio y se llamada muestreo aleatorio simple. La metodología SEMMA establece que para cada muestra considerada para el análisis del proceso se debe asociar el nivel de confianza de la muestra. (R. Montequín, 2003)

b) **Exploración (Explore):** Ya habiendo obtenido la muestra escogida la cual puede ser por método de muestreo de conveniencia se pasa a la fase de exploración en

donde se simplifica lo mayor posible el problema a tratar para optimizar la eficiencia del modelo y obtener los mejores resultados. Para cumplir con este requerimiento se pueden usar herramientas estadísticas de visualización o técnicas estadísticas que ayuden a relacionar las variables de la muestra y así tener claro cuáles serán las variables de entrada para el modelo. (R. Montequín, 2003)

c) **Manipulación (Modify):** Para esta fase ya se tiene las variables de la muestra relacionadas y se conoce más datos específicos para el análisis, lo que se hace en esta fase es que los datos tengan el formato adecuado para que puedan ser introducidos en el modelo, una vez ya obtenido todo se procedo recién al modelado en donde ya se ingresan las variables para obtener los resultados. (R. Montequín, 2003)

d) **Modelado (Model):** Para la fase de modelado se quiere encontrar la relación entre las variables cuantitativas (numero) y cualitativas (Variables Descriptivas) para tener un nivel de confianza determinado. Las técnicas utilizadas para el modelado de los datos incluyen métodos estadísticos tradicionales (tales como análisis discriminante, métodos de agrupamiento, y análisis de regresión), así como también técnicas basadas en datos tales como redes neuronales, técnicas adaptativas, lógica fuzzy, árboles de decisión, reglas de asociación y computación evolutiva, etc. (R. Montequín, 2003)

e) **Valoración (Assess):** La última fase como su nombre lo menciona hace la valorización mediante el análisis del modelo o los modelos aplicados para contrastarlos con métodos estadísticos y saber cuál funciono o dio el resultado esperado en base a la muestra con la población obtenida. (R. Montequín, 2003)

2.7.3 Metodología KDD

El término descubrimiento de conocimiento en bases de datos o KDD, fue acuñado en 1989 para referirse al proceso amplio de encontrar conocimiento en datos y para enfatizar la aplicación de "alto nivel" de métodos particulares de minería de datos (Fayyad et al, 1996). Fayyad considera que la fase de minería de datos se refiere, principalmente, a los medios por los cuales se extraen y enumeran los patrones de los datos. (Smith,2023)

2.7.3.1 Fases de Metodología KDD

El proceso KDD, es el proceso de utilizar métodos para extraer lo que se considera conocimiento de acuerdo con la especificación de medidas y umbrales, utilizando una base de datos junto con cualquier preprocesamiento, submuestreo y transformación necesarios. de la base de datos. (Fayyad et al, 1996)

Se consideran cinco etapas, presentadas en la Figura N° 15:

a) **Selección:** Esta etapa consiste en crear un conjunto de datos objetivo, o centrarse en un subconjunto de variables o muestras de datos, sobre las cuales se realizará el descubrimiento. (Fayyad et al, 1996)

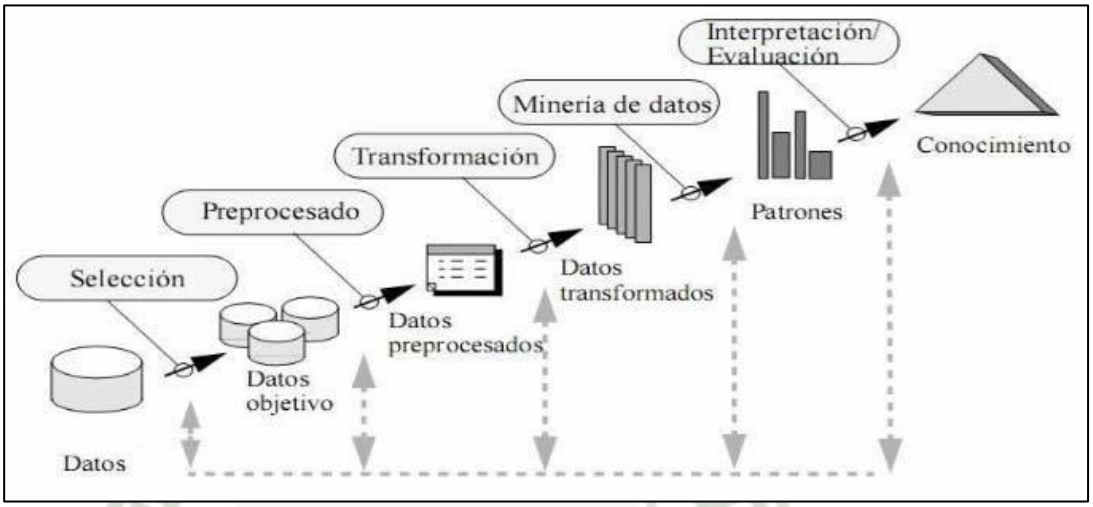
b) **Preprocesamiento:** En esta etapa consiste en la limpieza y el preprocesamiento de los datos de destino para obtener datos consistentes. (Fayyad et al, 1996)

c) **Transformación:** Esta etapa consiste en la transformación de los datos utilizando métodos de transformación o reducción de dimensionalidad. (Fayyad et al, 1996)

d) **Minería de datos:** Esta etapa consiste en la búsqueda de patrones de interés en una forma representacional particular, dependiendo del objetivo de la minería de datos generalmente, predicción. (Fayyad et al, 1996)

e) **Interpretación/Evaluación:** Esta etapa consiste en la interpretación y evaluación de los patrones para generar conocimiento. (Fayyad et al, 1996)

Figura. 15.
Fases de la Metodología KDD



Nota: Fayyad et al, 1996

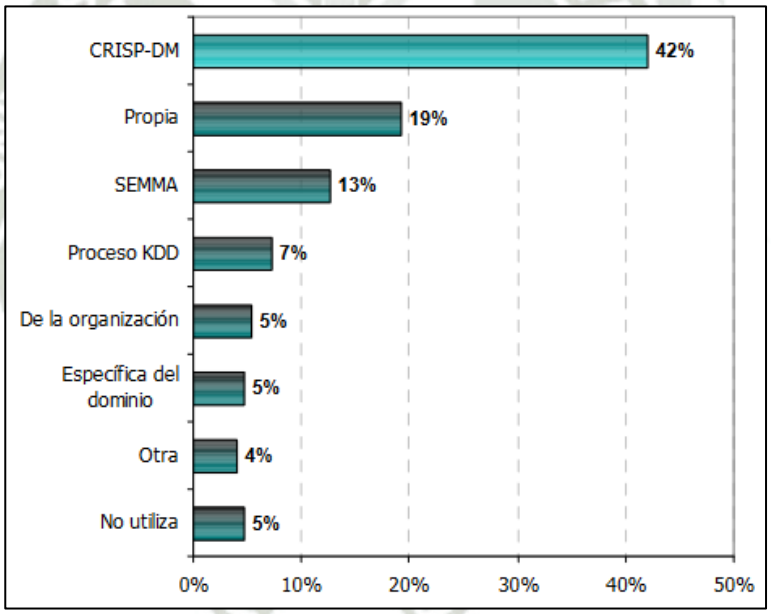
2.8 Metodologías más usadas para el Análisis de Datos

En el área del análisis de datos se centran en su gran mayoría para encontrar mejores técnicas de explotación de información y extracción de datos. Sin embargo, en este campo se ha llegado a una profundización para poder ejecutar este proceso y obtener nuevo conocimiento, es por ello por lo que nace el uso de metodologías, las cuales permiten llevar a cabo de forma sistemática y no trivial el uso de los datos y proporciona una guía para las organizaciones a entender el proceso de descubrimiento con la planificación y ejecución de proyectos. Las metodologías en si son pasos que ejecutan un proceso en base a las tareas de cómo se debe realizar. (KDNuggets, 2010)

Existen diversas metodologías, pero una de las primeras que fue creada en 1996 fue el modelo KDD que fue aceptado por la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. A partir del año 2000 con el crecimiento del área de

análisis de datos nace con un nuevo enfoque la metodología SEMMA y posteriormente la metodología CRISP-DM la cual se ha convertido en la más usada según un estudio realizado por la comunidad KDnuggets (Comunidad de Data Mining). La metodología CRISP-DM es más preferida por los usuarios que la utilizan debido a que profundiza en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de análisis de datos, mientras que las otras metodologías mencionadas solo proveen un guía general del trabajo a realizar en cada fase y no pasos detallados generando confusión en muchos casos si no se tiene un conocimiento previo. (KDNuggets, 2010)

Figura. 16.
Encuesta de Metodología más usada en Ciencia de Datos



Nota: KDNuggets, 2010

Por otro lado, en la investigación desarrollada por el autor S. K. Gupta, denominada "Un estudio comparativo de metodologías de minería de datos para análisis predictivo" compara diferentes metodologías de minería de datos, incluyendo CRISP-DM, para evaluar su efectividad en el análisis predictivo y la minería de datos en diversas industrias. Determinado que CRISP-DM es

una metodología estándar y estructurada que guía el proceso de minería de datos desde la comprensión del negocio hasta la implementación de los resultados. CRISP-DM se destaca por su estructura clara y flexible, que facilita la adaptación a diferentes industrias y problemas. Su enfoque en etapas bien definidas permite una gestión más eficiente de proyectos de minería de datos. CRISP-DM también ha demostrado ser eficaz en una variedad de industrias, proporcionando una metodología estandarizada que ayuda a alcanzar resultados consistentes y reproducibles. Aunque KDD y SEMMA también son efectivas, CRISP-DM mostró una ventaja en términos de integración de etapas y aplicabilidad general, especialmente en la fase de evaluación y despliegue. (S. K. Gupta, 2021)

Por lo que viendo el funcionamiento de la metodología CRIPS-DM en entorno de análisis de datos y su eficacia al momento de encontrar resultados y la encuesta realizada por la comunidad KDNuggets para la presente investigación se usara dicha metodología por servir como guía y por tener muy claro los pasos a seguir.

2.9 Técnicas y Herramientas

2.9.1 Python

Python fue creado en 1991 por Guido van Rossum y nació como una idea de tener un código más simple de usar desde entonces Python es un lenguaje de programación muy versátil y se ha convertido actualmente en el más usado para desarrollo web, creación de software nativo e hibrido y análisis de datos. Python tiene muchas ventajas sobre cualquier otro lenguaje, al igual que tiene variedades de biblioteca que reduce el código a un tercio para el programador y debido a esto Python ha alcanzado el pico más alto en términos de Aprendizaje Automático. (Python Software Fundation,2023)

2.9.1.1 Python en Análisis de Datos

La ciencia de datos es desarrollar un enfoque diferente para registrar, almacenar, analizar los datos y utilizar los datos para obtener información efectiva. Python al ser un lenguaje de código abierto y versátil también puede orientarse al análisis de datos es por eso por lo que proporciona varias bibliotecas para ello como se enumera a continuación:

- **Matplotlib:** Se pueden crear gráficos de trazado 2D.
- **Pandas:** Se usa para el análisis de datos en finanzas, estadísticas, las ciencias sociales y la ingeniería requieren diferentes tipos de estructura de datos y herramientas que son proporcionados por Pandas. (https://pypi.org).
- **NumPy:** Es la biblioteca básica para uso científico. Computación en Python. (https://pypi.org) Los arreglos y matrices multidimensionales pueden ser hechos usando objetos en NumPy, y también Se proporcionan rutinas que permiten a los desarrolladores poder calcular matemáticas avanzadas y funciones estadísticas en matrices con código si es posible. También se utiliza en estructura de datos.
- **SciPy:** Se usa para la manipulación y visualización de datos utilizando un comando de alto nivel proporcionado en gráficos multidimensionales. También tiene funciones para resolver integrales numéricamente, calculando ecuaciones diferenciales, y optimización todo está incluido en el paquete. La biblioteca SciPy también se utiliza para procesamiento de imágenes.
- **Pillow:** Es la biblioteca de imágenes de Python que agrega el soporte para diferentes opciones como apertura, manipulación datos y almacenamiento de imágenes con

diferentes formatos de archivos. Ampliamente usada en procesamiento o entrenamiento de imágenes. (Thompson,2023)

2.9.2 Google Colaboratory

Google Colaboratory (también conocido como Google Colab) es un servicio en la nube basado en Jupyter Notebooks para difundir educación e investigación sobre aprendizaje automático. Proporciona un tiempo de ejecución completamente configurado para aprendizaje profundo y acceso gratuito a una GPU robusta. Este servicio de Google está vinculado a una cuenta de Google Drive y es gratuito. Como Google Colab está basado en Jupyter se debe mencionar que Jupyter es una herramienta de código abierto basada en navegador que integra lenguajes interpretados, bibliotecas y herramientas de visualización. Google Colab proporciona tiempos de ejecución de Python versión 2 y 3 preconfigurados con las bibliotecas esenciales de inteligencia artificial y aprendizaje automático, como TensorFlow, Matplotlib y Keras. La máquina virtual en tiempo de ejecución (VM) se desactiva después de un período de tiempo y se pierden todos los datos y configuraciones del usuario. Sin embargo, la computadora portátil se conserva y también es posible transferir archivos desde el disco duro a la cuenta de Google Drive del usuario. La infraestructura de Google Colaboratory está alojada en la plataforma Google Cloud.

2.9.3 Matriz de Confusión

Para poder medir una métrica en el campo de la inteligencia artificial y el aprendizaje automático tenemos una herramienta denominada matriz de confusión la cual nos permite visualizar cual es el desempeño de un algoritmo de aprendizaje supervisado determinado. (Roberts, 2024)

Se divide en columnas en donde cada una representa el número de predicciones de cada clase sin embargo cada fila representa las instancias en la clase real. Lo que se podría decir es que la matriz de confusión nos permite ver los aciertos y errores que tiene nuestra predicción a la hora de realizar el aprendizaje automático. Y su estructura se puede visualizar en la Figura N° 16.

Figura. 17.
Estructura de una Matriz de Confusión

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Nota: Ivo Düntsch,2019

Para una matriz de confusión se pueden evaluar 4 posibles valores: Los verdaderos positivos, el verdadero negativo, el falso negativo y el falso positivo en donde a continuación podremos ver cada uno:

- **Verdadero Positivo (VP):** Para este caso se tiene que el valor real es positivo y la prueba de predicción también que es positivo. (Ivo Düntsch,2019)
- **Verdadero Negativo (VN):** En estos casos el valor real es negativo y la prueba de predicción también dice que el resultado es negativo. (Ivo Düntsch,2019)

- **Falso Negativo (FN):** Tenemos el caso donde el valor real es positivo, y la prueba de predicción dice que el resultado es negativo. (Ivo Düntsch,2019)
- **Falso Positivo (FP):** Para este caso el valor real es negativo, y la prueba de predicción dice que el resultado es positivo. (Ivo Düntsch,2019)

En base a las 4 opciones mencionadas surgen 4 métricas a medir de la matriz de confusión: Las cuales son la exactitud, la precisión, la sensibilidad y la especificidad.

a) Exactitud: La Exactitud o en inglés, "Accuracy" se refiere a lo cerca que está el resultado de una medición del valor verdadero. La Exactitud es la cantidad de predicciones positivas que fueron correctas para saber el número total de ellas. Se representa como la proporción de resultados verdaderos (tanto verdaderos positivos (VP) como verdaderos negativos (VN)) dividido entre el número total de casos examinados (verdaderos positivos, falsos positivos, verdaderos negativos, falsos negativos), como se puede ver en la formula (1). (Peterson,2023)

$$Exactitud = \frac{(VP+VN)}{(VP+FP+FN+VN)} \quad (1)$$

b) Precisión: La Precisión o en inglés "Precisión" es el apartamiento del conjunto de valores obtenidos a partir de las mediciones repetidas del aprendizaje automático, la regla dice que mientras menor sea la dispersión mayor será la precisión. Es el porcentaje de casos positivos detectados. Se representa por la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos). (Peterson,2023)

$$Precision = \frac{VP}{(VP+FP)} \quad (2)$$

c) **Sensibilidad:** También es conocida como la Tasa de los Verdaderos Positivos o TP y es la proporción de todos los casos positivos que fueron encontrados por el algoritmo con el cual hicimos el aprendizaje automático. (Peterson,2023)

$$Sensibilidad = \frac{\text{VP}}{(\text{VP+FN})} \quad (3)$$

d) **Especificidad:** También es conocida como la tasa de verdaderos negativos o TN en donde se detecta todos los negativos que el algoritmo de clasificación ha detectado correctamente. (Peterson,2023)

$$Especificidad = \frac{\text{VN}}{(\text{VN+FP})} \quad (4)$$

2.9.4 Métricas de Aprendizaje Automático

2.9.4.1 Error Medio Absoluto (MAE)

La métrica de error medio absoluto (MAE por sus siglas en inglés, Mean Absolute Error) es una medida que es muy utilizada para evaluar la precisión de un modelo de aprendizaje automático. Esta métrica cuantifica el promedio de las diferencias absolutas entre las predicciones del modelo y los valores reales en el conjunto de datos de prueba. El MAE mide la magnitud promedio de los errores en las predicciones del modelo, sin considerar su dirección. Cuanto menor sea el MAE, mejor será el rendimiento del modelo, ya que indica que las predicciones están más cercanas a los valores reales. (Zhang,2023)

Por ejemplo, si el MAE es 0, significa que el modelo hace predicciones perfectas para todas las observaciones en el conjunto de datos de prueba. (Zhang,2023)

La fórmula para calcularlo es:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \quad (5)$$

- n = Es el número de observaciones en el conjunto de datos de prueba
- y_i = Son los valores reales de la variable dependiente para la observación i.
- $\widehat{y_i}$ = Son las predicciones del modelo para la observación i.
- |. | = Representa el valor absoluto

2.9.4.2 Spearman RHO (RHO)

Spearman Rho, también conocido como coeficiente de correlación de Spearman o simplemente como Rho de Spearman, es una medida estadística utilizada para evaluar la relación entre dos conjuntos de datos. A diferencia del coeficiente de correlación de Pearson, que evalúa la relación lineal entre dos variables, Spearman Rho evalúa la relación monotónica entre ellas. La correlación de Spearman se calcula a partir de las clasificaciones de las observaciones en lugar de los valores brutos de las variables. Es especialmente útil cuando los datos no siguen una distribución normal o cuando la relación entre las variables no es lineal. (Zhang,2023)

La fórmula para calcularlo es:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (6)$$

- ρ = Es el coeficiente de correlación de Spearman
- d = Es la diferencia entre las clasificaciones de los pares de observaciones en las dos variables.
- n = Es el número de observaciones

2.9.4.3 Error Relativo Absoluto (RAE)

El error relativo es una medida utilizada en estadística y matemáticas para evaluar la precisión de una aproximación o una estimación en relación con el valor

verdadero o esperado de una cantidad. El error relativo se calcula como la relación entre el error absoluto y el valor verdadero. Se expresa comúnmente en términos de porcentaje para facilitar su interpretación. El error relativo proporciona una medida de la precisión relativa de una aproximación en comparación con el valor verdadero. Cuanto más cercano a cero sea el error relativo, más precisa será la aproximación. Por otro lado, un error relativo grande indica que la aproximación es menos precisa en relación con el valor verdadero. (Zhang,2023)

Y la fórmula para calcularlo es:

$$RAE = \frac{|Valor\ Verdadero - Valor\ Aproximado|}{|Valor\ Veradero|} \quad (7)$$

- Valor Verdadero = Es el valor verdadero o esperado de la cantidad
- Valor Aproximado = Es la aproximación o estimación de la cantidad
- |x| = Representa en valor absoluto de x, asegurando que el error relativo siempre sea un valor positivo.

2.9.4.4 Error Cuadrático Medio (MSE)

El Root Mean Squared Error (RMSE), traducido al español como Error Cuadrático Medio de la Raíz, es una medida comúnmente utilizada para evaluar la precisión de un modelo de regresión. El RMSE mide la raíz cuadrada de la media de los errores cuadrados entre las predicciones del modelo y los valores reales de la variable dependiente en un conjunto de datos de prueba. El RMSE es similar al error cuadrático medio (MSE), pero se toma la raíz cuadrada del resultado final para volver a la misma escala que los valores de la variable dependiente. Por lo tanto, el RMSE proporciona una medida de la magnitud promedio de los errores de predicción en las mismas unidades que la variable dependiente. Un RMSE más bajo indica que el

modelo tiene una mejor capacidad para predecir los valores reales, mientras que un RMSE más alto indica una menor precisión. Por lo tanto, el objetivo al utilizar el RMSE es minimizar su valor, lo que implica que el modelo está haciendo predicciones más precisas. (Zhang,2023)

Y la fórmula para calcularlo es:

$$RMSE = \sqrt{\frac{1}{n}(y_i - \widehat{y_i})^2} \quad (8)$$

- n = Es el número de observaciones en el conjunto de datos de prueba.
- y_i = Son los valores reales de la variable dependiente para la observación i.
- $\widehat{y_i}$ = Son las predicciones del modelo para la observación i.

2.9.5 Consideraciones Finales

Para el presente capitulo se abordó temas relacionados a la diabetes y que herramientas se usara para la validación correspondiente del proyecto de investigación. Se detalla el uso de la inteligencia artificial con el aprendizaje automático en al área de la salud para ayudar a la predicción de la diabetes la cual es una enfermedad muy común en la gran parte de la población y una de las más mortales a nivel mundial.

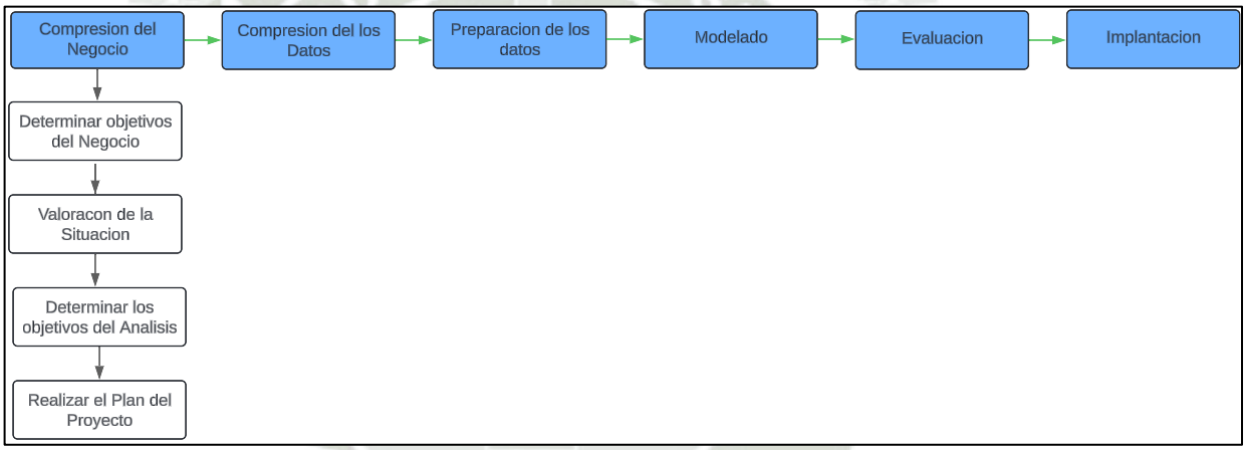
CAPITULO III

3. Desarrollo de la Propuesta de Investigación usando CRISP-DM

3.1 Comprensión del Negocio

En esta primera etapa de la fase del modelo de CRISP-DM, se va a definir el problema del negocio y se establecerán los objetivos y requisitos del proyecto de investigación para poder medir las métricas más adelante y permitan lograr el objetivo.

Figura. 18.
Fase I - Compresión de los requisitos del negocio



Nota: Propia

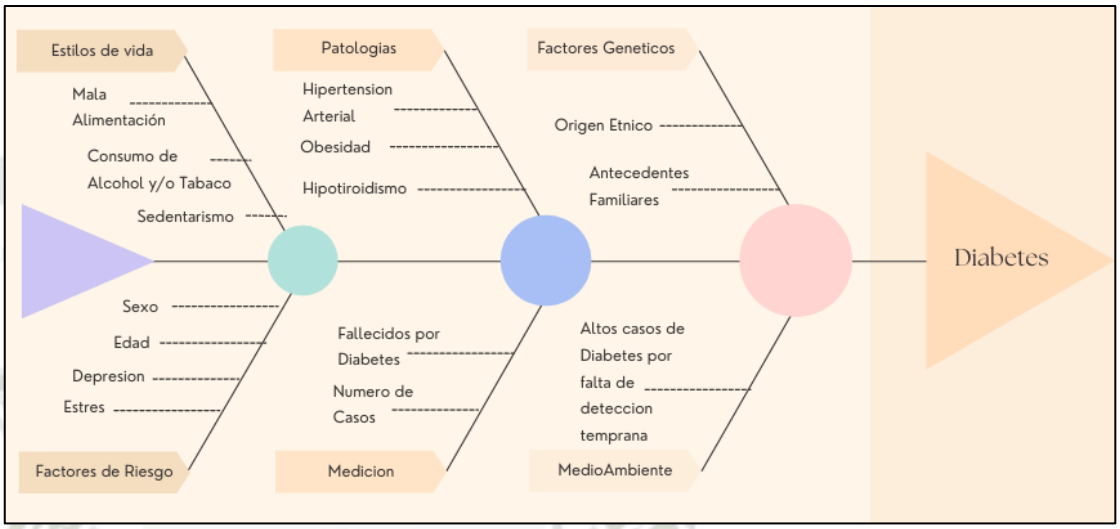
3.1.1 Determinar los Objetivos del Negocio

<u>Análisis del Problema</u>

Para lograr determinar los objetivos que tendrá el proyecto primero se debe entender el problema que se va a abordar en el cual se basa en realizar una predicción de la diabetes donde se consultó al Dr. Harry Calderón Flores las posibles causas de esta enfermedad. Por lo tanto, se usará el diagrama "Causa y Efecto o también llamada diagrama de Ishikawa" para el entendimiento del

problema y sus posibles causas. Se obtuvo el siguiente análisis mostrado en la Figura N° 19.

Figura. 19. *Análisis del problema mediante el diagrama de Ishikawa*



Nota: Propia

- **Situación Actual (Contexto):** En referencia a la situación del negocio que es la predicción de la diabetes se determinó con la referencia del Dr. Harry Calderón Flores que en su condición de médico general indica que para predecir la diabetes es fundamental saber el nivel de glucosa de una persona, su nivel IMC ya que es importante determinar si es alto o bajo porque puede sufrir de obesidad, también si tiene antecedentes familiares en la diabetes porque puede ser hereditario y es muy probable que una persona que tiene familiares cercanos con diabetes también lo tenga y datos relacionados a su salud como actividad física o consumo de alguna sustancia que pueden desencadenar una diabetes. Por lo anteriormente detallado se obtuvo un dataset de 3600 registro de datos de estudiantes de una Universidad Privada entre los 16 a 34 años en la ciudad de Arequipa donde solo se usarán 3542 por tener registros incompletos y este dataset contiene los valores

de sexo, edad, peso, estatura, valor IMC, nivel de glucosa, antecedentes familiares por diabetes, consumo de alcohol, drogas, tabaco y realización de actividad física.

Para comprender mejor el diagnóstico mediante la glucosa en sangre de una persona para contraer diabetes vamos a detallar en la Tabla 2 los valores que representan cada uno de ellos:

Tabla 2.
Interpretación de los Niveles de Glucosa

Interpretación	En Ayunas – Dia siguiente	Después de Comer 2 horas
Hipoglucemia	Menor de 70 mg/dl	-
Normal	Entre 70 – 110 mg/dl	Menor de 140 mg/dl
Pre-Diabetes	Entre 110 – 126 mg/dl	Entre 140 - 200 mg/dl
Diabetes	Mayor a 126 mg/dl	Mayor a 200 mg/dl

Nota: Rojas et al, 2012)

- **Objetivos del Negocio:** El objetivo principal del negocio es realizar la predicción de diabetes en los estudiantes de una Universidad Privada para saber si sus niveles de glucosa son elevados o no y así poder cuidarse o ir a un especialista para recibir tratamiento en caso de que la predicción sea poder positivo para diabetes. Definiendo los siguientes Objetivos:
  - ❖ Procesar y Analizar los datos recopilados para entrenarlos
  - ❖ Realizar una predicción de los datos entrenados de los estudiantes
- **Criterios de éxito del negocio:** En este apartado se va a establecer como criterio de éxito la predicción de diabetes en los estudiantes de una Universidad Privada en la ciudad de Arequipa la cual debe tener una predicción efectiva mayor al 90% de casos puestos a prueba para que sea válida.

3.1.2 Valoración de la Situación

Inventario de recursos:

Los recursos de hardware con los que se cuenta en el presente proyecto de investigación son un monitor, un ordenador portátil con las siguientes características:

- **Marca:** HP
- **Modelo:** HP Victus 15 16-D0XX
- **Procesador:** 11th Gen Intel® Core™ i5-11400H @ 2.70GHz
- **Memoria:** 24.0 GB
- **Capacidad de Almacenamiento:** 500 GB
- **Tarjeta Gráfica:** NVIDIA GeForce GTX 1650
- **Sistema Operativo:** Windows 11

Requisitos funcionales, requisitos no funcionales, supuestos y restricciones:

Para el correcto desarrollo del entrenamiento del modelo de aprendizaje automático del proyecto de investigación se van a establecer los requisitos funcionales y no funcionales, los supuestos y restricciones del proyecto. El análisis de las restricciones que puede ocurrir cuando se realice el entrenamiento va a permitir entender que va a ser necesario para lograr los objetivos del proyecto y que supuestos pueden llegar a conseguir el resultado, en donde para la definición intervino el médico general Dr. Harry Calderón Flores.

Definir los requisitos para proyectos de inteligencia artificial (IA) es un proceso complejo que requiere considerar tanto los aspectos funcionales como los no funcionales del sistema, además de las particularidades de los algoritmos y datos

que serán utilizados. Se usará el estándar IEEE 830-1998 “Especificación de Requisitos (Software Requirements Specification) establece pautas para el análisis y la especificación de requisitos de software. Esta norma es fundamental en el proceso de desarrollo de software, ya que proporciona un marco para definir los requisitos funcionales y no funcionales de un sistema.

1. **Identificación de Stakeholders**

   Para poder determinar a los stakeholder se hizo un analisis inicial a las personas que pueden verse afectada por el proyecto en cual se hizo de la siguiente manera por preguntas:

   - **¿ Quienes son los principales afectados por la diabetes ?**

     Cualquier persona que tenga un alto indice de glucosa.

   - **¿ Quienes Tratan a pacientes que tengan diabetes?**

     Profesionales de la salud

   - **¿ En dónde son diagnosticadas las personas que tienen diabetes?**

     En Hospitales, Clínicas y Postas

   - **A quienes les interesa saber si tienen diabetes o no**

     Cualquier persona o paciente que acuda a un centro médico para descarte.

   - **Quien dar los recursos para poder combatir la diabetes**

     Ministerio de Salud en conjunto con el gobierno.

   Por ello se determinará cuál será su impacto de cada stakeholder en el proyecto.

   - Cualquier persona que tenga un alto índice de glucosa. (Alto)
   - Profesionales de la Salud (Alto)
   - Hospitales, Clínicas y Postas (Medio)
   - Paciente que acuda a un centro médico para descarte. (Alto)
   - Ministerio de Salud y Gobierno (Medio)

Después de la priorización se define cuáles son los stakeholders interno y externos para el proyecto-

b) Internos

- **Propietario del Proyecto:** Responsable de la Investigación
- **Propietarios de los Datos:** Pacientes de data recopilada
- **Profesionales de la salud:** Tratan a pacientes con diabetes

a) Externos

- **Pacientes Nuevos:** Se someten a que sus datos pasen la predicción
- **Centros Médicos:** Seguimiento de casos de diabetes
- **Gobierno Local:** Implementación de equipos en centros de salud
- **Comunidad Local:** Personas que acceden a un centro de salud

2. Definición de Requisitos Funcionales y No Funcionales

- **Requisitos Funcionales**

  En las tablas 3, 4 y 5 se definen los requerimientos funcionales del proyecto de investigación.

Tabla 3. *Requerimiento Funcional RF01 - Captura de Datos*

Código Requerimiento	**RF01**
Nombre	Captura de Datos
Propósito	El prototipo de predicción debe permitir la captura de datos de diferentes Notas en un formato de lectura que tenga la herramienta.
Descripción	El prototipo tiene la conexión a diferentes Notas de datos y formatos ya sea Excel, bases de datos, texto en bloc de notas. El usuario deberá hacer la conexión a cualquier Nota mediante variables de conexión lo cual le permitirá importar los datos para posteriormente hacer su procesamiento.
Entrada	La entrada principal es la solicitud de acceso a la herramienta para realizar la captura de datos.
Salida	La salida del requerimiento son los datos mostrados en formato JSON dentro de la herramienta para hacer su análisis
Prioridad	Alta

Nota: Propia

Tabla 4. *Requerimiento Funcional RF02 - Entrenamiento de Datos*

Código Requerimiento	**RF02**
Nombre	Entrenamiento de Datos
Propósito	El prototipo debe permitir entrenar los datos capturados y guardarlos para usarlos en la predicción.
Descripción	El dispositivo tendrá la opción de cargar múltiples modelos de entrenamiento que son algoritmos de aprendizaje supervisado, lo cual ayudar a poder realizar el entrenamiento con casos de prueba que extraerá de los datos capturados.
Entrada	La entrada para este requerimiento son los datos los cuales fueron capturados y cargados a la herramienta.
Salida	La salida son los casos de prueba de entrenamiento los cuales se podrán ver en una matriz de confusión determinando su efectividad.
Prioridad	Alta

Nota: Propia

Tabla 5. *Requerimiento Funcional RF03 - Predicción de Datos*

Código Requerimiento	**RF03**
Nombre	Predicción de Datos
Propósito	El prototipo de predicción debe permitir cagar el modelo entrenado para usarlo con datos nuevos y realizar la predicción.
Descripción	La herramienta permitirá cargar los datos entrenados para que con el ingreso de nuevos datos se pueda realizar una predicción y esto se generara gracias al algoritmo de aprendizaje supervisado seleccionado.
Entrada	La entrada para este requerimiento es los datos nuevos para realizar una predicción comparado con los datos entrenados.
Salida	La salida es la predicción de los datos nuevos en formato JSON para visualizar el resultado.
Prioridad	Alta

Nota: Propia

- **Requisitos No Funcionales**

  La tabla 6 lista los requerimientos no funcionales del proyecto de investigación.

Tabla 6. *Requerimientos No Funcionales*

Requerimientos	
Usabilidad	**Facilidad de Uso:** El prototipo tendrá un uso fácil ya que solo se basa en la ejecución de códigos ya cargados y se basa en la simplicidad conjuntamente son la salida de los datos con la predicción realizada. **Visualización:** Los resultados mostrados después de la predicción estarán en una tabla en formato JSON mostrando la claridad y entendimiento de este permitiendo tener una respuesta muy rápida.
Confiabilidad	**Funcionamiento Confiable:** El prototipo tiene un funcionamiento web por lo que su funcionamiento es constante y confiable. La velocidad de predicción se mantiene constante para evitar fallas y frustraciones en el usuario. **Precisión:** Se establece que la precisión de la predicción será mayor al 90% lo cual minimiza los errores en el prototipo y se tiene la exactitud en la información ya que es crucial para la utilidad del prototipo.
Portabilidad	La portabilidad del prototipo es amplia ya que puede ser usado desde una Tablet, célula o laptop. Para que se de fácil acceso
Eficiencia	**Procesamiento:** El prototipo de predicción tiene un procesamiento rápido ya que conexión a servidores de Google los cuales hacen un procesamiento que tiene como objetivo no duras mas de 10 segundos. Esto garantiza tener un resultado sin retrasos y en menos tiempo.
Rendimiento	El tiempo y respuesta de la capacidad de procesamiento de la herramienta de IA es rápida ya que tiene un CPU virtual que tiene 12.7GB de RAM y un procesador con ejecuciones en simultaneo que permite que el tiempo de ejecución sea muy eficiente
Espacio	Escalabilidad: La capacidad de almacenamiento del disco del CPU virtual es

	de 107.7GB lo cual permite almacenar una gran cantidad de datos.
Seguridad	El acceso al entorno de ejecución es por medio de una cuenta de Google lo cual lo hace segura porque solo mediante esta se tiene acceso y es personal.

Nota: Propia

3. Validación de Requisitos

Se aplicarán las siguientes métricas para la evaluación de los requisitos funcionales:

- Cobertura de Requisitos: Se debe lograr una cobertura de requisitos del 100% antes del lanzamiento.
- Exactitud: La predicción debe superar el 90% de fiabilidad del modelo utilizado.
- Integridad: Todas las funcionalidades del entorno de predicción deben estar completas y funcionando correctamente
- Consistencia: Los resultados obtenidos por los diferentes registros de predicción obtenidos deben coincidir en un 100%.
- Correctitud: No debe haber más de 1 defecto crítico por cada 100 líneas de código en las funciones principales.
- Rendimiento Funcional: El tiempo de ejecución de cada predicción no debe superar los 2 segundos
- Utilización: La funcionalidad de predicción debe ser utilizada al menos 1 vez al día por un especialista médico.
- Eficiencia: La función de predicción puede utilizar más del 90% de los recursos de CPU durante su ejecución.

Se aplicarán las siguientes métricas para la evaluación de los requisitos no funcionales:

- Rendimiento: El tiempo de respuesta del entorno de predicción no debe superar los 2 segundos en el 100% de registros ingresados.

- Escalabilidad: La carga de información estará soportado hasta el procesamiento de 17.2 GB de RAM y 107.7GB de almacenamiento, es posible aumentar, pero pagando una suscripción lo cual no es necesario para este proyecto.
- Seguridad: Las vulnerabilidades críticas deben ser mitigadas en un plazo de 24 horas desde su detección.
- Usabilidad: Un nuevo usuario debe ser capaz de completar las tareas básicas en menos de 30 minutos después de recibir una breve capacitación.
- Confiabilidad: La disponibilidad del sistema debe ser del 99.9%.
- Mantenibilidad: La actualización de la herramienta a nuevas versiones debe establecer que la tasa de defectos post-implementación no debe exceder de 1 defecto por cada 100 líneas de código.

4. Gestión de Datos

- Recolección de Datos: La recolección de datos se hizo mediante la solicitud de acceso a la Clínica Aliviari.
- Calidad de los Datos: Los datos que se analizaran deben estar completos, deben tener consistencia lógica y no deben tener errores. Es necesario que los datos recopilados sean de una Nota veraz porque una mala calidad de los datos puede afectar significativamente al entrenamiento del modelo.
- Etiquetado de los datos: El etiquetado de datos debe referir a que pertenece cada registro recopilado para determinar qué elementos se usara en el entrenamiento y predicción.
- Privacidad y Ética: Los datos no deben contener ningún elemento que permita identificar algún registro por persona, detallando solo datos biométricos que se usaran para el análisis y predicción.

5. **Iteración y Refinamiento**
    - Uso de Metodologías Agiles: Se tiene que usar una metodología ágil para realizar todo el proyecto de IA y para eso se debe encontrar la mejor para el proyecto como es la metodología CRISP-DM.
    - Pruebas de Usuario: Se debe determinar las pruebas necesarias y correspondientes para validar la predicción.
    - Monitorización y FeedBack: La monitorización se debe basar en el acceso al entorno de predicción y uso de la herramienta para resolver cualquier inconveniente.

Para poder realizar un correcto entrenamiento se tendrá que tomar los siguientes supuestos:

- **Selección de los datos:** Los datos que se analizarán incluirán datos médicos como sexo, edad, peso, talla, IMC, antecedentes familiares por diabetes, consumo de alcohol, drogas, tabaco y realización de actividad física de alumnos de una Universidad Privada en la ciudad de Arequipa en el rango de 16 a 34 años.

Sin embargo, por las características del proyecto de investigación tendremos las siguientes limitaciones:

- **Recursos Limitados:** El acceso a programas con una licencia es un recurso que es limitado ya que la idea del proyecto es usar software libre pero también se puede utilizar una licencia de un programa que puede realizar un análisis más rápido pero el tiempo del proyecto es el correcto para el uso de una herramienta libre.

- **Limitaciones de datos:** Al usar datos del alumnado de pregrado de una Universidad Privada en la ciudad de Arequipa no se está tomando todo el conjunto general ya que no se tiene acceso a esa información y solo se tuvo acceso a un rango de ellos para la recopilación y no es posible tomar una recopilación de datos continua ya que estos varían cada año en cantidad y volumen.

3.1.3 Determinar los objetivos del Análisis

- **Metas del Análisis de Datos:** El objetivo principal del proyecto de investigación es validar el dataset de alumnos de pregrado de una Universidad Privada en la ciudad de Arequipa para determinar en base a sus datos médicos en especial del nivel de glucosa si estos son altos o bajos dependiendo de su valor predecir si tiene o no diabetes para entrenar un modelo con casos de prueba e ingresar nuevos datos a la predicción y determinar si un alumno tiene o no diabetes.

Por ello para asegurar el éxito vamos a comparar las metodologías mencionadas y poder determinar que la metodología CRISP-DM es la mejor para este proyecto. En donde tenemos 3 las cuales son la metodología KDD, CRISP-DM y SEMMA en cual todas sirven para el análisis de datos y su aprobación depende de cuál sea la que mayor porcentaje de predicción nos muestre. Cabe mencionar que existe una gran diferencia entre una metodología y un modelo ya que el modelo indica que hacer al momento de abordar un proyecto de análisis de datos y la metodología sin embargo especifica como hacerlo pasando por diversas fases. Por ello se detalla a continuación cual es la óptima para el proyecto:

Tabla 7.
Comparación de Metodologías de Ciencia de Datos.

	Fases de Metodologías de Minería de Datos		
	KDD	**SEMMA**	**CRISP-DM**
FASES	1. Selección 2. Pre-Procesamiento 3. Transformación 4. Minería de Datos 5. Interpretación/Evaluación	1. Muestreo 2. Exploración 3. Manipulación 4. Modelado 5. Valoración	1. Análisis de Problema 2. Análisis de Datos 3. Preparación de los Datos 4. Modelo de Datos 5. Evaluación de Datos 6. Explotación
ENFOQUE	Orientado a la identificación de patrones más favorables en diversas áreas en la cual se puede aplicar análisis de datos.	Es más específico al momento de desarrollar proyectos de Minería de Datos.	Orientado más a los objetivos empresariales de proyectos grandes o pequeños de Minería de Datos
USO	Su uso se enfoca en encontrar patrones de datos.	Esta más ligado a productos SAS (Statistical Anlysis System)	Es una metodología abierta y gratuita
METODOLOGIA	Metodología de patrones arquitectónicos orientados a los datos	Tiene pasos secuenciales que sirven como guías.	Es una metodología enfocada a la gestión de proyectos
COMPLEJIDAD	Es complejo de implementar ya que tiene una cantidad considerable de pasos a desarrollar y no hay orientación de desarrollo.	Es bastante más simple y ágil, pero esta más orientado a grandes volúmenes de datos.	Es mucho más simple de comprender y aplicar, cuenta con una curva de aprendizaje muy baja para cualquier analista de datos.

Nota: Propia

La comparación demuestra lo siguiente que las metodologías mencionadas KDD, SEMMA y CRISP-DM tienen fases en las cuales se parecen mucho al momento de aplicarlas, pero hay una gran diferencia en la metodología CRISP-DM ya que en las 6 fases que tiene para desarrollar un proyecto profundiza en cada una brindando mayor detalle que las demás en cada etapa, volviéndose una guía por ejemplo para personas que tengan poco conocimiento en análisis de datos. Por otro lado la metodología SEMMA se enfoca especialmente en los aspectos técnicos, en lo cual hay que tener conocimiento previo para encontrar la solución más adecuada lo que excluye las

actividades de análisis y comprensión del problema que se está abordando y adicionalmente esta metodología nace para ser usada para trabajar con el software de minería de datos de la compañía SAS requiriendo un uso adicional de la herramienta y no siendo gratuita como si lo es el uso de la metodología KDD y CRISP-DM. En otro aspecto la metodología KDD es compleja de implementar a comparación de las otras ya que sus fases están enfocadas mayormente a la aplicación de grandes volúmenes de datos por lo que su objetivo es encontrar patrones en la etapa de evaluación e interpretación y así crear conocimiento.

Entonces evaluando la comparación de metodologías, la más fácil de usar y accesible es la metodología CRISP-DM porque es de uso libre, las tareas en cada etapa de su aplicación detallan a grandes rasgos que es lo que se debe seguir sirviendo como guía para alguien con poco conocimiento en el área de análisis de datos y por otra parte está enfocada para hacer proyectos con un gran volumen de datos y un menor volumen de datos por lo que el proyecto no tiene un gran volumen siendo ideal el uso de esta metodología para el presente proyecto.

- **Criterio de Éxito en el Análisis de Datos:** Desde el análisis de datos, se estableció un criterio de éxito en donde hacer las predicciones correctas con una fiabilidad del 90% para saber si tiene o no diabetes un alumno de pregrado de una Universidad Privada en la ciudad de Arequipa.

Sin embargo, la predicción depende mucho de la naturaleza y características de los datos. Por ello se debe tener un dataset con datos de calidad como si fuma o no, si consume alcohol o drogas ya que está involucrado a tener un estilo de vida saludable.

Por ello se estableció una fiabilidad del 90% teniendo en cuenta que se tiene un dataset completo. La fiabilidad de las predicciones será determinada por el algoritmo de modelo de aprendizaje a automático a escoger y este será medido por las métricas.

3.1.4 Realizar el Plan del Proyecto

- **Plan del Proyecto:**

  Para realizar el plan de este proyecto de investigación se va a seguir las etapas de la metodología establecida que es CRISP-DM. Pero es importante mencionar que esta metodología no es lineal, sino es iterativa, por lo que va a depender mucho de las necesidades del modelo repetir las etapas previas para poder determinar un avance en el proyecto.

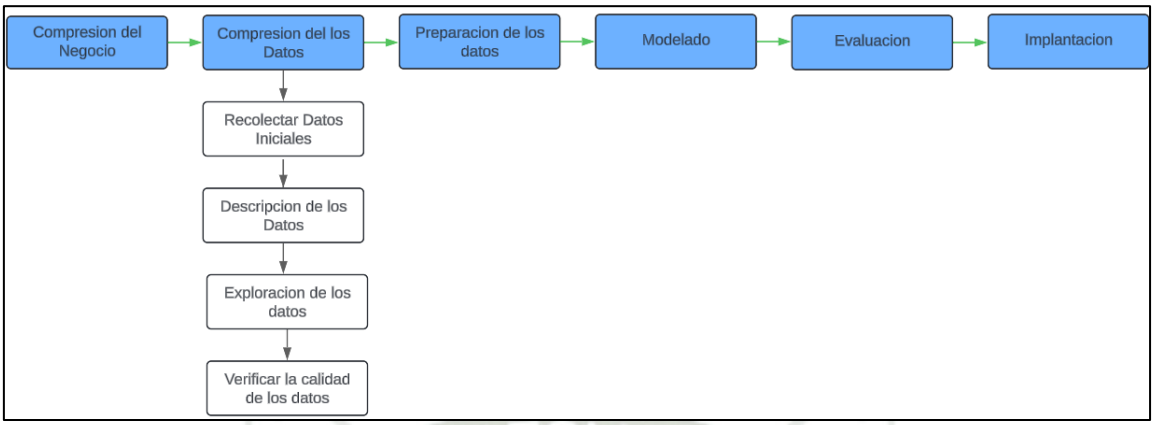
- **Valoración Inicial:**

  Habiendo ya definido de forma clara y concisa el alcance, los objetivos, los supuestos y limitaciones en esta primera fase del desarrollo de la metodología CRISP-DM se puede dar por concluida para pasar a la siguiente fase de Comprensión de datos que está encargada de estudiar las características de los datos y su viabilidad en el proyecto para generar las aportaciones de los objetivos finales.

3.2 Compresión de los Datos

En esta segunda fase se realizará la recolección, análisis, descripción y exploración de la Nota de datos escogida mencionada. La finalidad de la presente etapa es evaluar los datos existentes para poder determinar su calidad y establecer una estrategia de limpieza y formateo de datos.

Figura. 20.
Fase II - Compresión de los Datos



Nota: Propia

3.2.1 Recolectar Datos Iniciales

Los datos utilizados fueron recopilados de la Base de datos de la Clínica Aliviari respecto al alumnado de pregrado de una Universidad Privada en la ciudad de Arequipa hasta el año 2023 en la provincia de Arequipa, región de Arequipa. Se obtuvieron acceso a 3600 estudiantes, pero para la investigación se usó solo 3542 debido a mucha inconsistencia con la información. Los cuales varían desde los 16 años hasta los 34. Se tiene un Excel en cual se obtuvo la información como edad, sexo, talla, peso, IMC, glucosa, antecedentes, consumo de tabaco, consumo de alcohol, consumo de drogas y actividad física.

Tabla 8.
Muestra de Dataset Inicial

Sexo	Edad	Peso	Talla	IMC	Glucosa	Antece Fami. Diabetes	Alcohol	Tabaco	Drogas	Act. Física
F	17	62.0	1.59	24.52	102	NO	NO	NO	NO	NO
F	16	48.5	1.57	19.68	112	NO	NO	NO	NO	NO
M	21	88.7	1.80	27.38	133	NO	NO	NO	NO	SI
M	16	50.5	1.74	16.68	125	NO	NO	NO	NO	NO
F	16	58.0	1.57	23.53	177	NO	NO	NO	NO	NO

Nota: Propia

Los datos inicialmente tienen un formato que no ayuda a la investigación ya que tienen que ser transformados para lo cual se le aplicara un proceso de transformación para estandarizar la información que se aplicara al modelo de aprendizaje y pueda predecir los valores asociados a ellos. Para esta fase se hará la adaptación correspondiente para poder medir los datos de calidad del dataset.

Adaptación de datos y dataset

En el dataset se tiene el IMC el cual es una fórmula matemática la cual tiene como objetivo determinar la masa corporal que va en relación con el peso y la talla de una persona y su resultado da un indicativo de los rangos establecidos para saber si una persona esta corporalmente sana pero no es un factor determinante ya que puede variar en el tiempo y la fórmula es la siguiente:

$$IMC = \frac{\text{Peso}}{\text{Altura} * \text{Altura}} \quad (9)$$

- Peso: Es la Masa Corporal del cuerpo humano
- Altura: Longitud promedio de un ser humano

Tabla 9.
Interpretación del Índice de Masa Corporal

Interpretación	Valores IMC
Peso Bajo < 18.5	
Delgadez Severa	< 16.00
Delgadez Moderada	Entre 16.00 y 16.99
Delgadez Leve	Entre 17.00 y 18.49
Peso Normal > 18.5	
Peso Normal	Entre 18.5 y 24.99
Sobrepeso >= 25.00	
Preobesidad	Entre 25.00 y 29.99
Obesidad >= 30.00	
Obesidad Leve	Entre 30.00 y 34.99
Obesidad Media	Entre 35.00 y 39.99
Obesidad Mórbida	Mayor o igual a 40.00

Nota: Clasificación de la OMS del estado nutricional de acuerdo con el IMC,2024

La conversión “IMC” se determinará de la siguiente manera para entrenar el modelo

Tabla 10.
Interpretación del Atributo IMC en el Dataset

Valores de IMC	Interpretación	Valor atributo índice IMC
< 18.5	Bajo Peso	0
> 18.5 y < 24.99	Peso Normal	1
> 25.0 y < 29.99	Sobrepeso	2
> 30.0 y < 34.9	Obesidad Tipo 1 - Moderada	3
> 35.0 y < 39.0	Obesidad Tipo 2 - Severa	4
>= 40	Obesidad Tipo 3 - Mórbida	5

Nota: Propia

También el dataset se tiene el campo de Glucosa el cual significa que es un valor representativo de la en ayunas y estos niveles no deberían sobrepasar los normales ya que si esto sucede significa que el cuerpo no está produciendo suficiente insulina para reducir los niveles de azúcar en la sangre. Después de comer, la glucosa se eleva moderadamente durante las 2 primeras horas, ante ello el cuerpo aumenta los niveles de insulina y disminuye parcialmente esta cantidad, con la reducción de glucosa y aprovechamiento de esta azúcar para generar energía en el cuerpo la insulina se degrada en el promedio de 2 a 3 horas después de ingerir un alimento y clasificación de los niveles detectados en este análisis es el siguiente:

Tabla 11.
Interpretación de los valores del Atributo Glucosa

Interpretación	Valores Glucosa – miligramo / decilitro
Sin Diabetes	Menos de 140 mg/dl
Pre-Diabetes	Entre 140 y 199 mg/dl
Diabetes	Mayor a 200 mg/dl

Nota: Organización Mundial de la Salud,2024

La interpretación del atributo de glucosa según los parámetros seria de la siguiente manera para adaptarlo al modelo.

Tabla 12.
Interpretación del Atributo Glucosa en el Dataset

Valores Atributo Glucosa	Interpretación	Valor Atributo Glucosa
< 140 mg/dl	Normal	0
<= 140 y >= 199 mg/dl	Pre-Diabetes	1
>= 200 mg/dl	Diabetes	2

Nota: Propia

3.2.2 Descripción de los Datos

Los datos utilizados se encuentran en un Libro de Excel y fueron recopilados en el año 2024 pero la información data del año 2023 en la provincia de Arequipa, región de Arequipa y representa el alumnado de pregrado de una Universidad Privada en la ciudad de Arequipa de los exámenes médicos que se hacen anualmente la Clínica Aliviari. Se obtuvieron 3600 datos, pero para la investigación se usó solo 3542 debido a mucha inconsistencia de información en la cual se encontró 1625 hombres y 1916 mujeres entre los 16 y 34 años. Se tiene un Excel en cual se obtuvo la información como edad, sexo, talla, peso, IMC, glucosa, antecedentes familiares por diabetes, consumo de tabaco, consumo de alcohol, consumo de drogas y actividad física.

Numero de Instancias y Atributos

Para el entrenamiento y pruebas del algoritmo se cuenta con 3542 instancias con 11 atributos:

a) Sexo

b) Edad

c) Peso

d) Talla

e) IMC

f) Glucosa

g) Antecedentes Familiares Diabetes

h) Alcohol

i) Tabaco

j) Drogas

k) Actividad Física

Significado de los atributos

Como se mencionó en el punto 3.2.2, se cuenta con 11 atributos: los cuales se pasarán a describir a continuación:

Tabla 13.
Descripción de los Atributos

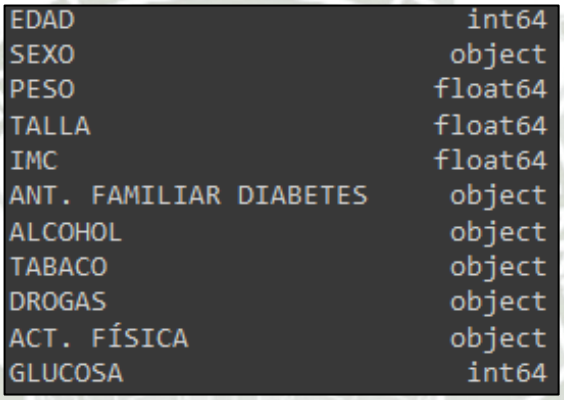
Tipo de Atributo	Nombre Atributo	Tipo de Dato	Descripción
Original	Sexo	Carácter	Genero de la Persona
Original	Edad	Decimal	Edad de la Persona
Original	Peso	Decimal	Peso de la Persona en kilogramos
Original	Talla	Decimal	Talla de la Persona en metros
Original	IMC	Decimal	Índice de Masa Corporal Este campo es producto de la formula: $$IMC = \frac{\textbf{Peso}}{\textbf{Altura} * \textbf{Altura}}$$
Original	Glucosa	Decimal	Glucosa en Sangre
Original	Antecedentes Familiares Diabetes	Carácter	• NO si no tiene antecedentes • SI si tiene antecedentes
Original	Alcohol	Carácter	Si consume Alcohol o no
Original	Tabaco	Carácter	Si consume Tabaco o no
Original	Drogas	Carácter	Si consume Drogas o no
Original	Actividad Física	Carácter	Si realiza actividad física o no

Nota: Propia

Descripción del Formato de Datos

Los campos del dataset para el entrenamiento serán de tipo numérico ya que esto ayudara a que el modelo se entrene de una manera más rápida y todos serán enteros positivos y pueden ser también decimales ya que hay campos que contienen parte decimal y este conjunto ayudara a realizar correctamente la predicción y no se tenga errores al momento de hacer el entrenamiento. Indicando que el tipo object es de tipo carácter.

Figura. 21.
Tipo de Datos en el Dataset



Nota: Google Colab,2024

3.2.3 Exploración de los Datos

En la exploración de los datos se usará la estadística como el campo que nos ayudará a entender mejor las características del dataset para la construcción del modelo posteriormente. Para poder explorar los datos usamos Google Colab usando la librería Seaborn, en donde vamos a generar diagramas de dispersión en el cual vamos a poder determinar la relación entre los conjuntos de datos.

Figura. 22.
Importación de Dataset en Google Colab

```
from google.colab import drive
drive.mount('/content/drive')
```

```
df = pd.read_csv("/content/drive/MyDrive/Dataset_Diabetes_UCSM_2023.csv"
```

Nota: Google Colab,2024

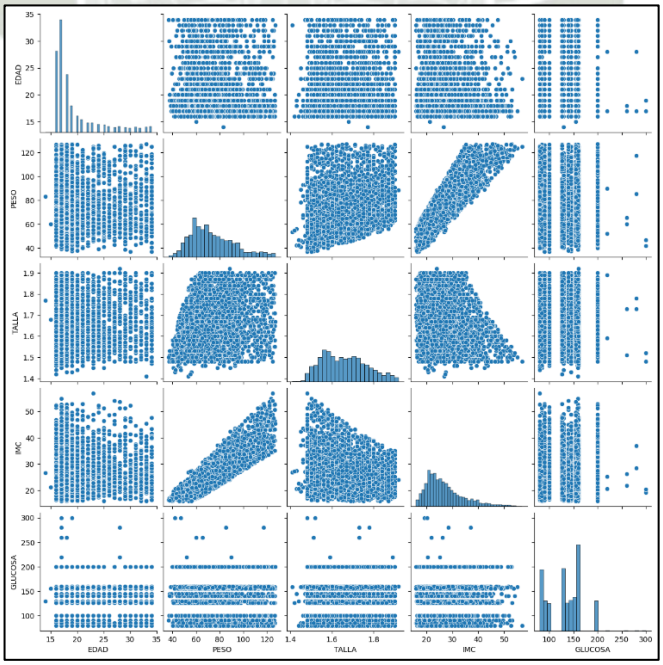
Figura. 23.
Importación de la Librería Seaborn

```
import seaborn as sns
sns.pairplot(df)
```

Nota: Google Colab,2024

El resultado después de la aplicación fue el siguiente:

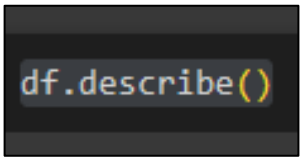
Figura. 24.
Relación de Variables para determinar su dependencia



Nota: Google Colab,2024

También vamos a explorar los datos desde la perspectiva del dataset en donde tendremos las variables según su estadística.

Figura. 25.
Uso de la Función Describe() para Variables Numéricas



Nota: Google Colab,2024

En la Figura N° 26 podremos ver el resultado del análisis, pero en base a la estadística donde nos muestra el conteo general, el valor medio, la desviación, el valor mínimo, el percentil 25, 50% y 75%, por último, el valor máximo que ayuda a entender cada campo numérico respectivamente.

Figura. 26.
Descripción estadística de los atributos numéricos en el dataset

	EDAD	PESO	TALLA	IMC	GLUCOSA
count	3541.000000	3541.000000	3541.000000	3541.000000	3541.000000
mean	19.494493	73.575603	1.660003	26.812087	155.931375
std	4.533589	18.582212	0.104446	7.003533	37.801430
min	16.000000	37.000000	1.410000	16.000000	90.000000
25%	17.000000	59.770000	1.570000	21.670000	124.000000
50%	18.000000	70.000000	1.650000	25.210000	156.000000
75%	20.000000	85.400000	1.730000	30.240000	189.000000
max	34.000000	126.940000	1.920000	56.940000	220.000000

Nota: Google Colab,2024

Identificación de los tipos de datos

En este apartado vamos a identificar el tipo de variable para el análisis respectivamente donde tenemos 2 tipos la variable cualitativa y la variable cuantitativa que la diferencia de

ambas es que la variable cuantitativa aporta valores estadísticos en base a números como cantidades y demás, pero la variable cualitativa no puede hacerlo ya que solo aporta descripción de un determinado atributo.

Tabla 14.
Tipos de Variables del Dataset

Tipo de Atributo	Nombre Atributo	Tipo de Variable
Original	Sexo	Variable Cualitativa
Original	Edad	Variable Cualitativa
Original	Peso	Variable Cuantitativa
Original	Talla	Variable Cuantitativa
Original	IMC	Variable Cuantitativa
Original	Glucosa	Variable Cuantitativa
Creado	Ant. Familiar Diabetes	Variable Cuantitativa
Original	Alcohol	Variable Cualitativa
Original	Tabaco	Variable Cualitativa
Original	Drogas	Variable Cualitativa
Original	Actividad Física	Variable Cualitativa

Nota: Propia

Haciendo el análisis de las variables tenemos lo siguiente ya que de los atributos originales y creados se tiene que 8 variables cualitativas y 14 cuantitativas. Para ello se tiene que las variables cuantitativas tendrán un análisis en base a la librería Pandas y Pyplot que nos determinara la frecuencia de cada apartado de los atributos y llegar a un análisis por variable en el dataset.

Se tiene 2 tipos de análisis uno que es por cada variable a nivel cualitativo en los atributos que si tienen datos contabilizables y las variables cuantitativas que representaran a grandes rasgos los atributos del dataset.

Tabla 15.
Gráficos por tipo de Variable

Variables Cualitativas	Variables Cuantitativas	Análisis Combinado
• Gráfico de Barras • Gráfico Circular • Histograma	• Gráfico Circular	• Diabetes por Sexo y Edad • IMC por Sexo y Edad • Actividad Física por Sexo • Diabetes por IMC y Edad

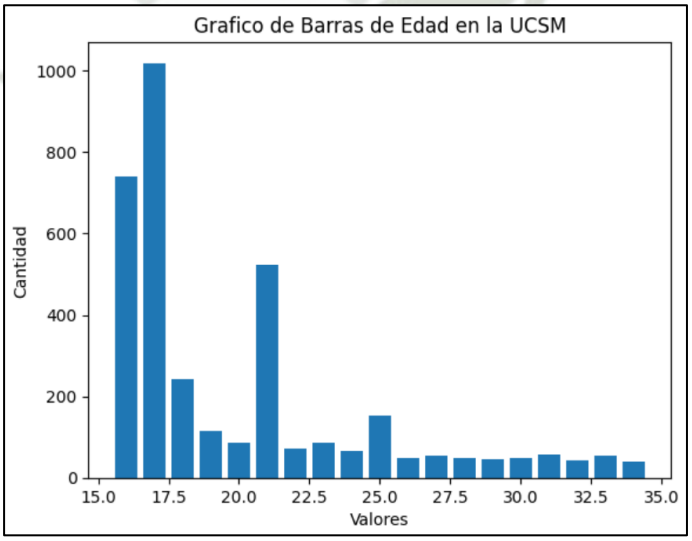
Nota: Propia

Análisis por Atributo

a) Edad

En el siguiente grafico podemos apreciar que la mayor población estudiantil tiene como base los 17 años y la menor población está entre los 25 hasta los 34 años. Entonces podemos afirmar que en los estudiantes de la Universidad Privada de la es una población muy joven. (Ver Figura 27)

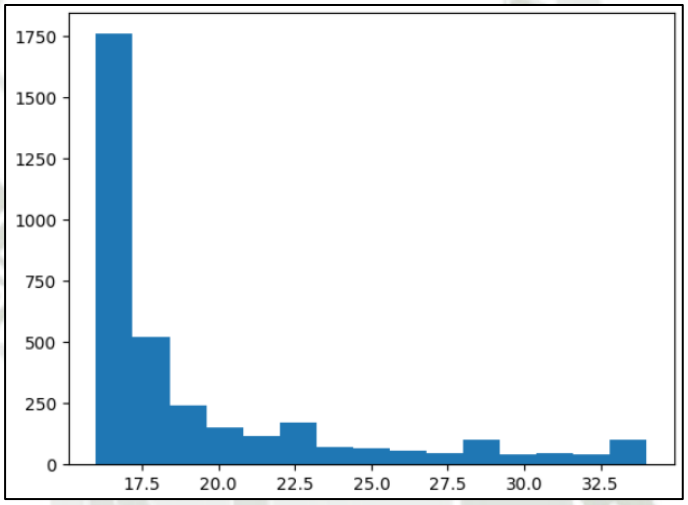
Figura. 27.
Gráfico de Barras por Edad en la UCSM



Nota: Google Colab,2024

Con el histograma de edad podemos comprobar la información mostrada en el gráfico de barras en donde el predominio está en 3 edades fundamentalmente que son 17 hasta 23 años respectivamente. Por lo que está asociada al ingreso a la universidad al terminar la etapa escolar. (Ver Figura 28)

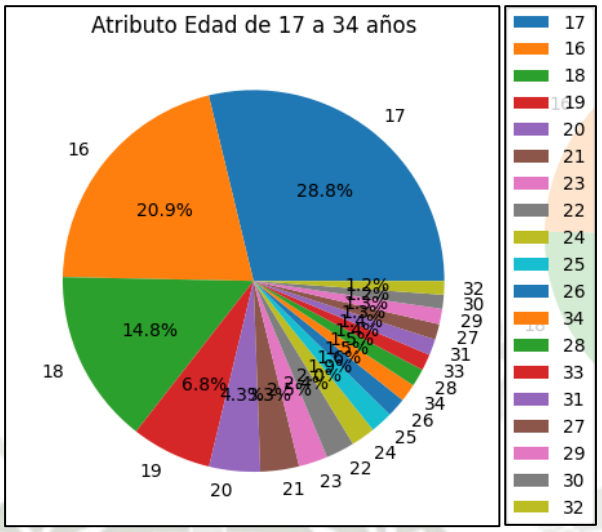
Figura. 28.
Histograma por Edad en la UCSM



Nota: Google Colab,2024

En un análisis por porcentajes se tiene el grafico circular en donde podemos observar que efectivamente la edad mayoritaria es 17 años y la menor de todas es 32 años respectivamente. (Ver Figura 29)

Figura. 29.
Gráfico circular por Edad en la UCSM



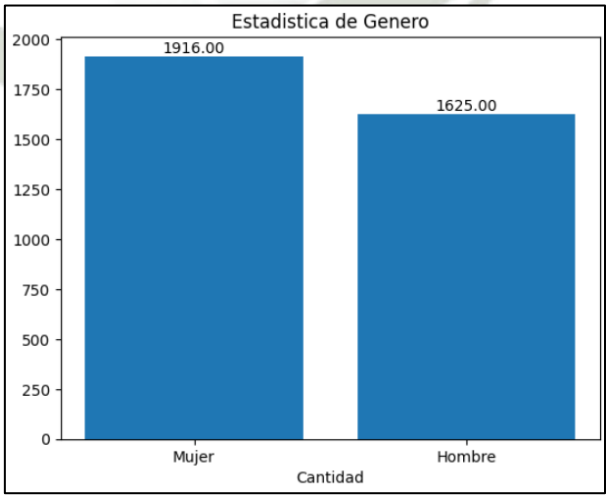
Nota: Google Colab,2024

b) Sexo

En el gráfico de barras se puede observar que la predominancia son el sexo Femenino con 1916 ejemplares y el sexo Masculino con 1625 ejemplares.

(Ver Figura 30)

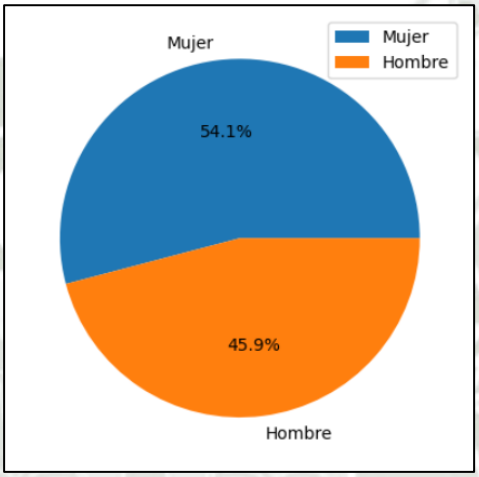
Figura. 30.
Gráfico de Barras por Sexo en la UCSM



Nota: Google Colab,2024

Haciendo un análisis del grafico de barras con el grafico circular podemos comprobar que efectivamente el 54.1% representa el sexo Femenino y el 45.9% representa el sexo Masculino en donde podemos afirmar según este dataset que hay más personas del género femenino estudiando en la Universidad Católica Santa María. (Ver Figura 31)

Figura. 31.
Gráfico Circular por Sexo en la UCSM

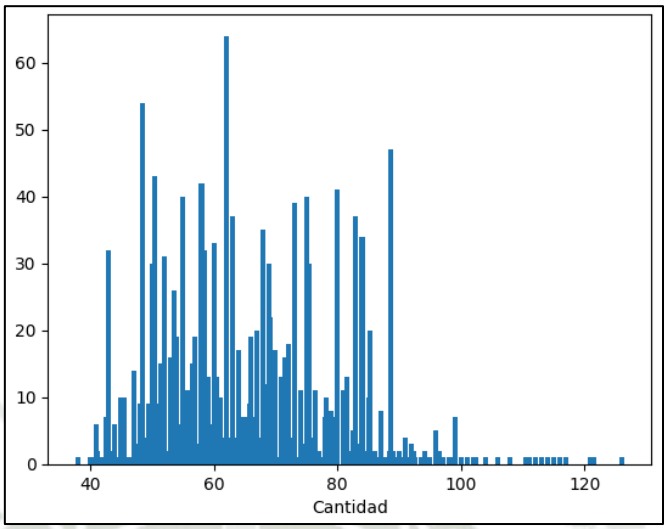


Nota: Google Colab,2024

c) Peso

Para el atributo peso tenemos lo siguiente que este atributo está calculado en kilogramos y el rango que tenemos va desde los 40 kilos hasta los 120 kilos por lo que esto será analizado posteriormente por el IMC que calcula la masa corporal dependiendo de la talla y el peso. (Ver Figura 32)

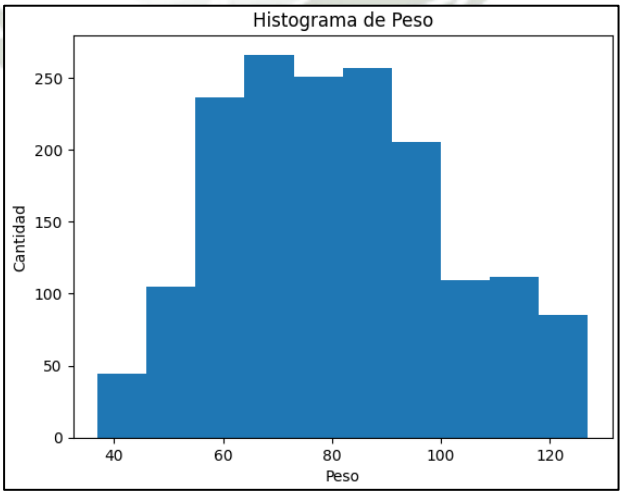
Figura. 32.
Gráfico de Barras por Peso en la UCSM



Nota: Google Colab,2024

En el histograma podremos ver que hay una gran variedad de datos por lo que se tiene que la mayor parte del estudio realizado para el dataset se tiene que el rango de pesos mayoritario esta entre los 60 y 100 kilos respectivamente y que hay muy poca variedad desde los 40 kilos hasta los 60 kilos. (Ver Figura 33)

Figura. 33.
Histograma por Peso en la UCSM

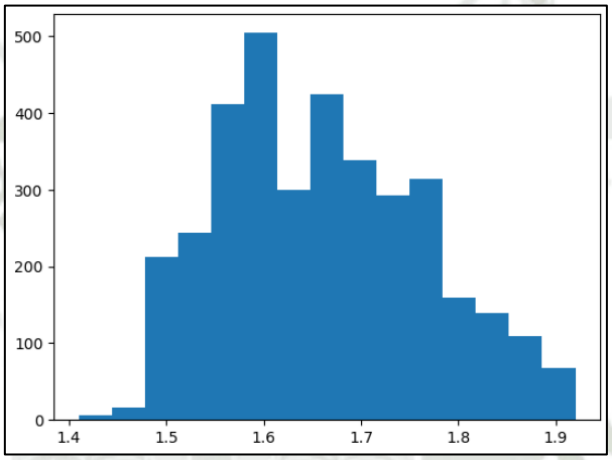


Nota: Google Colab,2024

d) Talla

Respecto a la talla podremos observar en el gráfico de barras que estos valores oscilan desde la talla de 1.50 metros hasta 1.90 metros que es la máxima alcanzada en el dataset para este atributo. Siendo la más común en general la de 1.60 metros. (Ver Figura 34)

Figura. 34.
Histograma por Talla en la UCSM

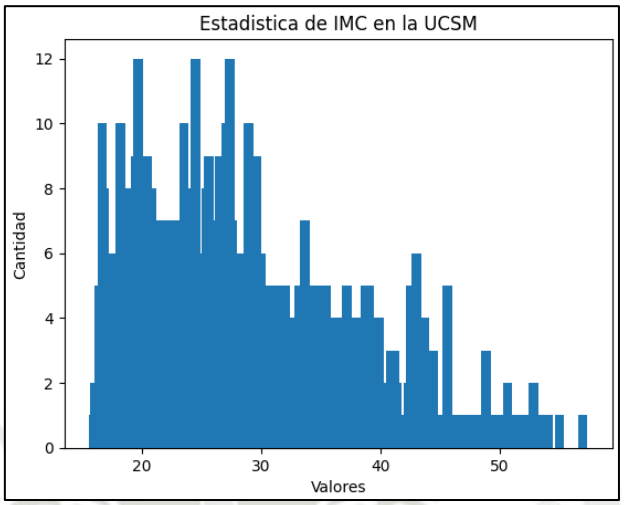


Nota: Google Colab,2024

e) IMC (Índice de Masa Corporal)

Para el análisis del grafico de barras tenemos que la oscilación de valores va desde los 20 hasta los 50 como diagnostico ya que depende del peso y la talla con la fórmula matemática aplicada y podemos ver que los picos más pronunciados esta entre 20 y 30 que más adelante se detallara en el atributo de riesgo de IMC para determinar si tienen sobrepeso o no. (Ver Figura 35)

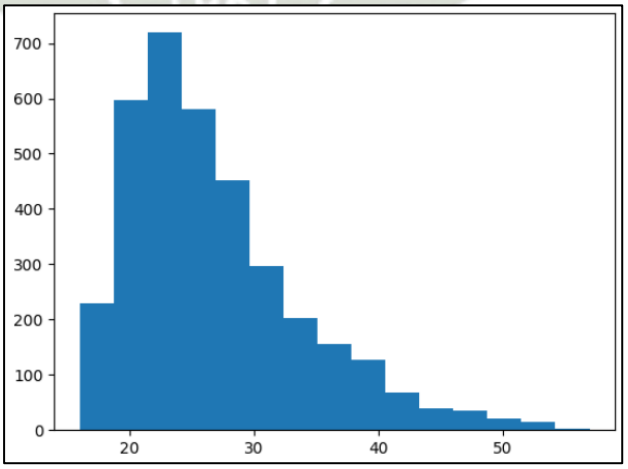
Figura. 35.
Gráfico de Barras por IMC en la UCSM



Nota: Google Colab,2024

Para el atributo IMC se tiene el siguiente histograma en donde podremos observar que los valores oscilan entre 20 y 50 lo que nos hace indican que el mayor rango esta entre 20 y 30 y que el menor número esta entre 30 y 50. Cabe recalcar que el IMC tiene una fórmula matemática y tiene un rango de valores para poder medir si una persona esta correctamente sana. (Ver Figura 36)

Figura. 36.
Histograma por IMC en la UCSM

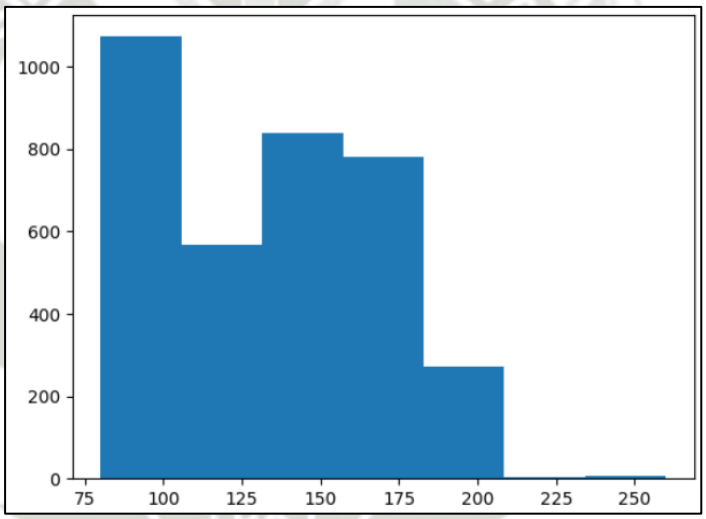


Nota: Google Colab,2024

f) Glucosa

El atributo de glucosa que se tiene en el dataset está enfocado en que la glucosa obtenida es de la glucosa postprandial en donde esta glucosa es tomada después de 2 horas de haber ingerido algún alimento entonces a partir del Grafico podemos analizar que los valores más altos están entre los 100 a 150 mg/dll pero hay un número significativo igual que va desde los 200 hasta los 250 mg/dll que indica una posible diabetes. (Ver Figura 37)

Figura. 37.
Histograma por Glucosa en la UCSM



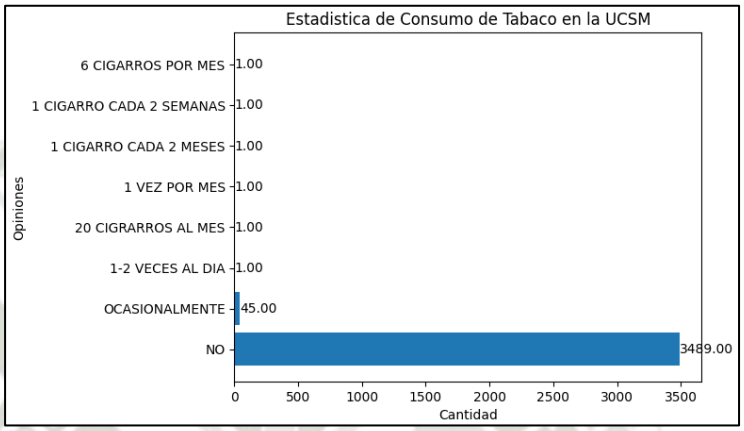
Nota: Google Colab,2024

g) Consumo de Tabaco

Con respecto a el atributo de consumo de tabaco en el dataset se tuvo los siguientes valores donde podemos ver que la mayoría no consume esta sustancia, pero también hay un cierto número que lo consume ocasionalmente. Hay una investigación

relacionada donde se comenta que el consumo de tabaco también ayuda a poder contraer diabetes por lo que su consumo es peligroso. (Ver Figura 38)

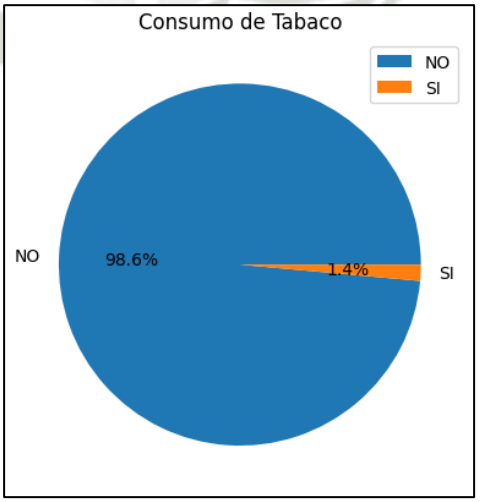
Figura. 38.
Gráfico de Barras por Consumo de Tabaco en la UCSM



Nota: Google Colab,2024

Haciendo en análisis general del consumo de tabaco podemos ver en la Figura N° 39 que la mayoría de los estudiantes no consume alcohol solo el 1.4% lo hace que representa 51 personas del total de 3541 personas analizadas.

Figura. 39.
Gráfico Circular por Consumo de Tabaco en la UCSM

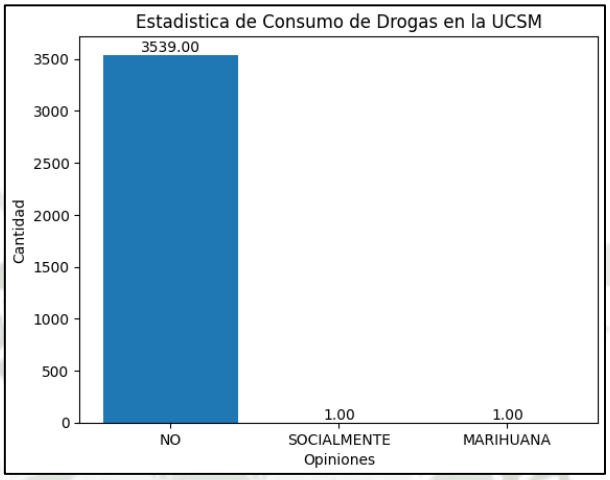


Nota: Google Colab,2024

h) Consumo de Drogas

Respecto al atributo de Drogas se tiene que solo 1 ejemplar refirió que consume marihuana y en cuanto a la demás población negaba esta acción. (Ver Figura 40)

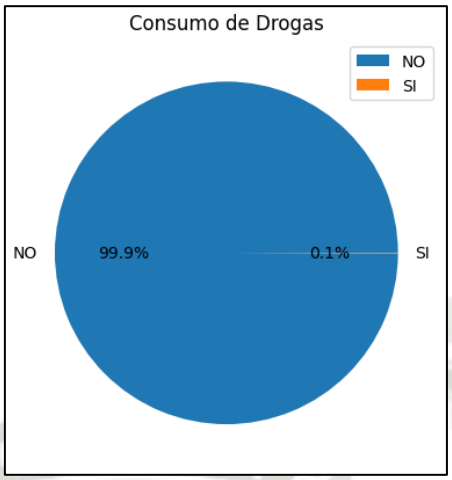
Figura. 40.
Gráfico de Barras por Consumo de Drogas en la UCSM



Nota: Google Colab,2024

Según la estadística recopilada se puede afirmar que del dataset de 3542 individuos solo el 0.1% consume drogas y el 99.9% no lo hace lo que es positivo ya que el consumo de drogas es dañino para la salud y también puede ser un factor importante para contraer la diabetes. (Ver Figura 41)

Figura. 41.
Gráfico Circular por Consumo de Drogas en la UCSM

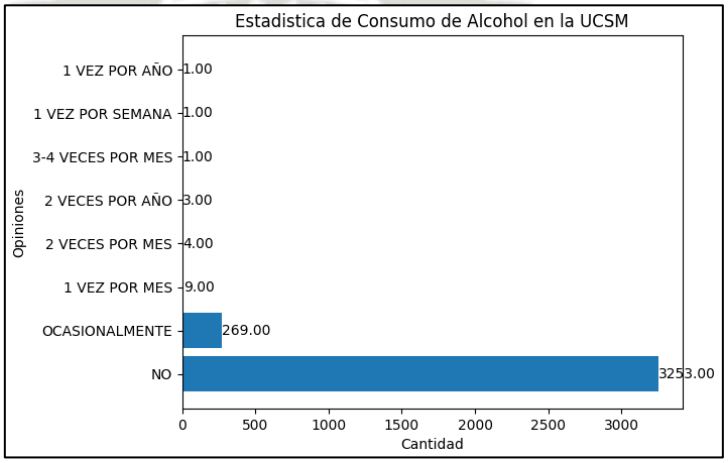


Nota: Google Colab,2024

i) Consumo de Alcohol

Según el dataset para el consumo del atributo alcohol se puede decir que la mayoría de la población no lo consume y hay un cierto número de la población que si lo consume activamente ya sea socialmente o en eventos lo que es positivo debido a que se tiene un consumo bajo de bebidas alcohólicas. (Ver Figura 42)

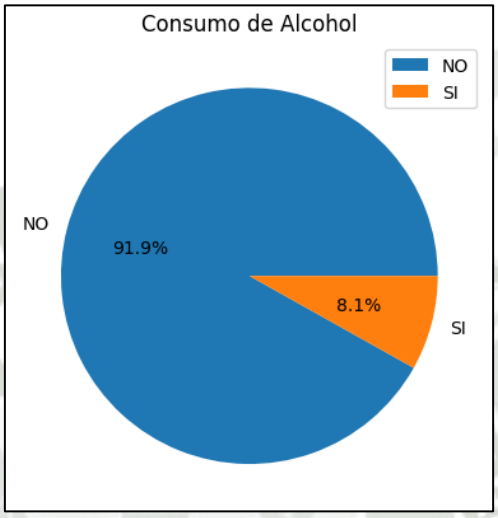
Figura. 42.
Gráfico de Barras por Consumo de Alcohol en la UCSM



Nota: Google Colab,2024

En base a el grafico de barras se realizó la estadística con un gráfico circular en porcentaje y se obtuvo que el 91.9% no consume alcohol, pero el 8.1% si lo hace respectivamente. (Ver Figura 43)

Figura. 43.
Gráfico Circular por Consumo de Alcohol en la UCSM

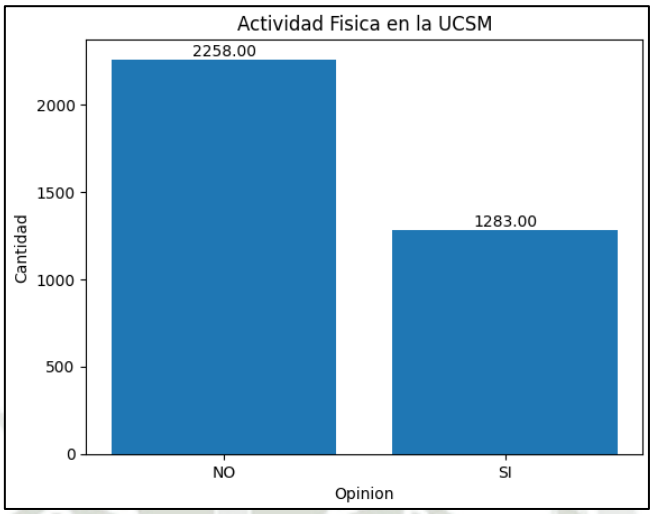


Nota: Google Colab,2024

j) Actividad Física

Con respecto al atributo de actividad física se le consulto a los estudiantes si realizaban algún ejercicio físico constantemente y el resultado fue que solo 1283 estudiantes de 3542 analizados lo hace de alguna forma durante la semana y 2258 no lo hace lo que es un factor importante ya que la diabetes se desencadena por falta de ejercicio y mala alimentación. (Ver Figura 44)

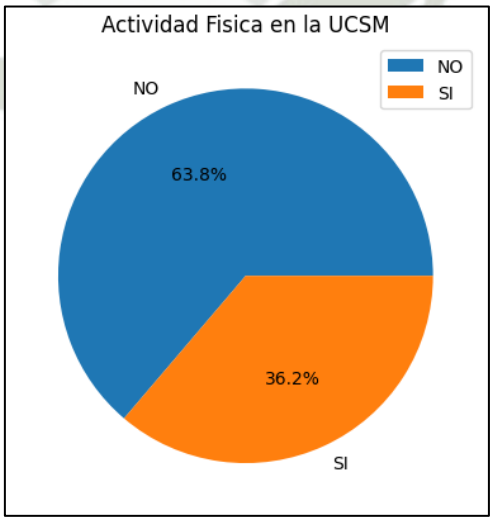
Figura. 44.
Gráfico de Barras por Actividad Fisica en la UCSM



Nota: Google Colab,2024

Haciendo un análisis general podemos afirmar que los 1283 estudiantes que realizan actividad física representan el 36.2% en donde el 63.8% no realiza actividad física representando más de la mitad de la población estudiantil que no tiene un habito de hacer ejercicio. (Ver Figura 45)

Figura. 45.
Gráfico Circular de Actividad Física en la UCSM



Nota: Google Colab,2024

k) Riesgo IMC

Habiendo analizado los valores tenemos lo siguiente para el diagnóstico del IMC en lo respecta al dataset analizado correctamente en el rango de valores establecidos por la OMS.

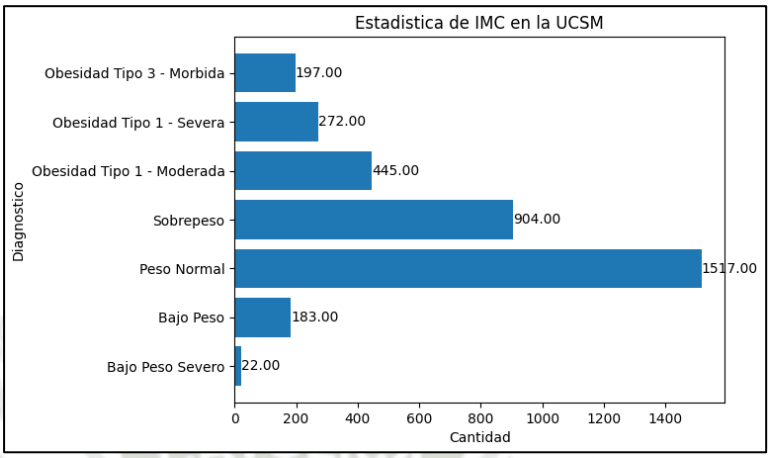
Tabla 16.
Diagnostico por IMC en los estudiantes de la UCSM

Diagnostico	Cantidad
Bajo Peso Severo	22
Bajo Peso	183
Peso Normal	1517
Sobrepeso	904
Obesidad Tipo 1 - Moderada	445
Obesidad Tipo 1 - Severa	272
Obesidad Tipo 3 - Morbida	197

Nota: Propia

En base al análisis extraído del dataset con los rangos establecidos por la OMS tenemos que la mayoría de persona exactamente 1517 tiene un peso normal y hay 904 personas con sobrepeso, 272 personas con obesidad tipo 1 y 192 personas con obesidad tipo 3 mórbida que son pacientes potenciales para contraer la diabetes. Ya que un factor importante es la buena alimentación y ejercicio físico para mantenernos en forma y este número es preocupante ya que pueden crecer los casos de diabetes en la UCSM. (Ver Figura 46)

Figura. 46.
Gráfico de Barras de Diagnostico de IMC en la UCSM

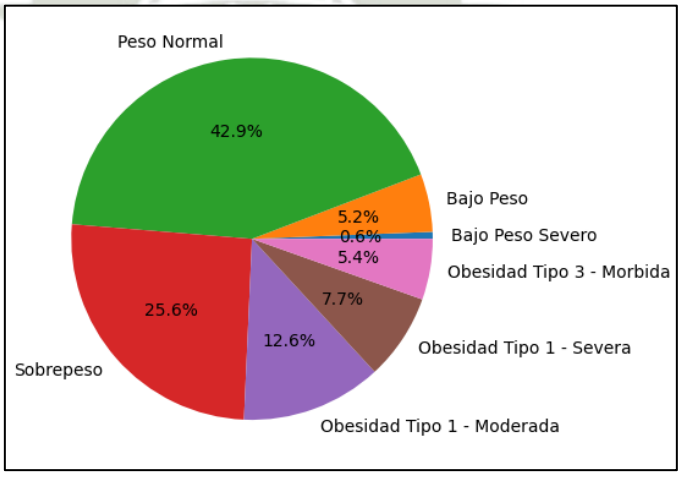


Nota: Google Colab,2024

Con respecto a el análisis del grafico circular podemos determinar que el 62% de todo el dataset analizado tiene un peso normal, mientras que el 23.5% que representa 393 persona según el grafico N° 25 tiene sobrepeso lo que también es un factor importante para que estas personas puedan contraer la diabetes.

(Ver Figura 47)

Figura. 47.
Gráfico Circular de IMC en los alumnos de la UCSM

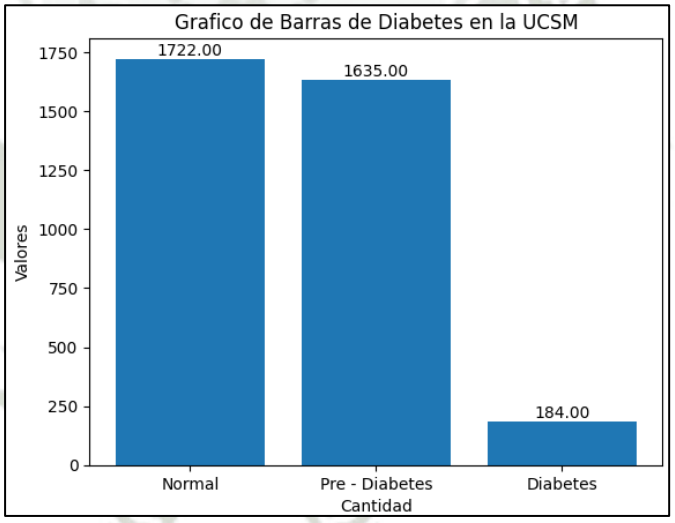


Nota: Google Colab,2024

l) Riesgo de Diabetes

Para el grafico de barras podemos deducir que según los niveles encontrados son 184 personas las cuales tienen diabetes porque sus niveles de glucosa superan los 200 mg/dll indicados por la OMS para la diabetes (Ver Tabla 2) , 1635 personas tienen prediabetes lo cual es una cifra alarmarte porque el número de diabetes positivos podría aumentar si en esta parte de la población no se tiene el cuidado correspondiente y 1722 personas tienen niveles de glucosa normal. Esta estadística se realizó por el campo de glucosa que se obtuvo del dataset extraído de la Clínica Aliviari. (Ver Figura 48)

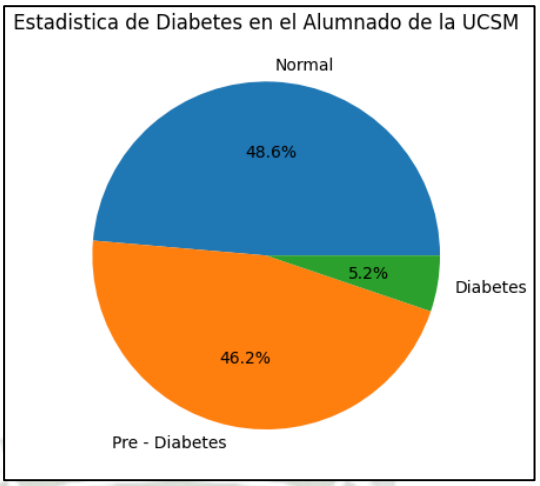
Figura. 48.
Gráfico de Barras de Glucosa en los alumnos de la UCSM



Nota: Google Colab,2024

Analizando el grafico de barras podemos visualizar que hay un buen número de alumnos que tiene un nivel de glucosa normal pero no llega a ser la totalidad, pero si agrupamos a las personas que están entre el nivel de prediabetes es alto, pero no pueden ser establecidos como diabetes. La proporción seria que de cada 100 alumnos 5 tienen una posible diabetes. (Ver Figura 49)

Figura. 49.
Gráfico Circular de Glucosa en los alumnos de la UCSM



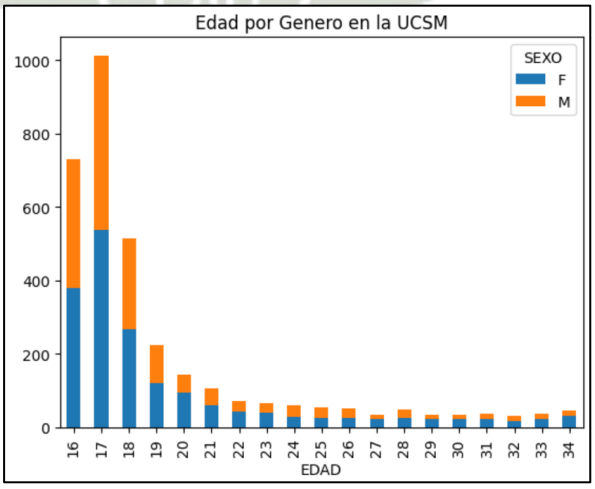
Nota: Google Colab,2024

Análisis Combinado

a) Análisis de Edad por Sexo

Respecto al análisis realizado podemos observar que dentro del dataset la mayor población está entre los 16 a 20 años respectivamente y hay más presencia femenina que masculina. (Ver Figura 50)

Figura. 50.
Gráfico de Barras por Genero y Edad en la UCSM

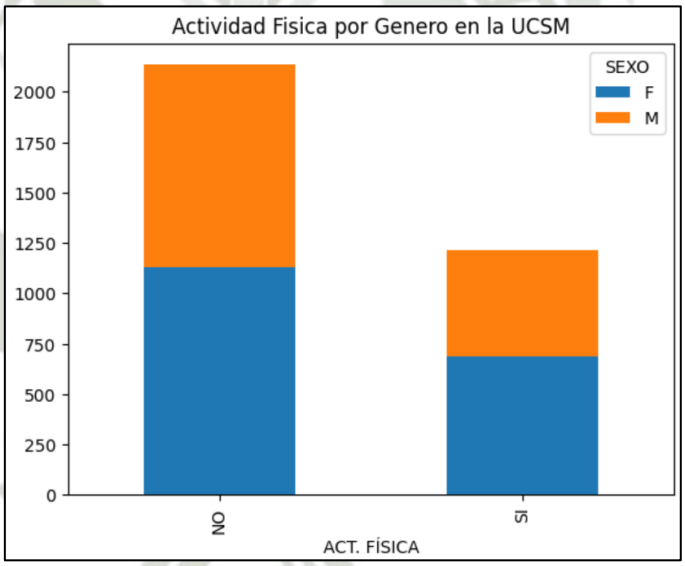


Nota: Google Colab,2024

b) Análisis de Actividad Física por Sexo

Dentro del dataset pudimos encontrar el atributo de actividad física el cual es importante porque una causa principal de la diabetes es la falta de ejercicio y la mala alimentación. Es por ello por lo que realizando un análisis pudimos determinar que hay más persona del sexo femenino que no realiza ejercicio y hay un porcentaje del sexo masculino que si lo hace, esto igual es preocupante porque se deduce que muy pocas personas se dedican a realizar actividad física ya sea en casa u otro establecimiento. (Ver Figura 51)

Figura. 51.
Gráfico de Barras de Actividad Física por Genero en la UCSM



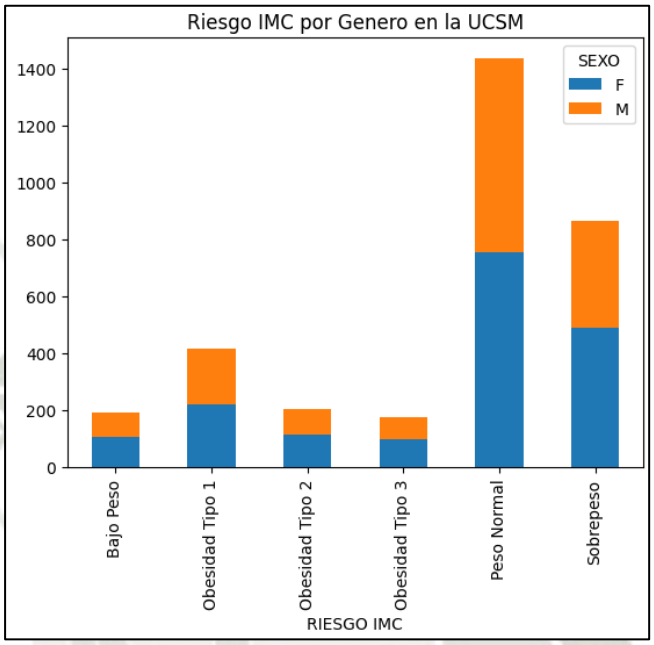
Nota: Google Colab,2024

c) Análisis de Riesgo de IMC por Sexo

El atributo IMC se encuentra en el dataset y este da un análisis respecto a la talla y el peso de una persona por lo cual se realizó un análisis combinatorio para determinar que es la predominancia de este valor en la UCSM. Podemos determinar

que la mayor parte tiene un peso normal entre hombres y mujeres, pero también existe un número considerable de casos con sobre peso. (Ver Figura 52)

Figura. 52.
Gráfico de Barras de Riesgo IMC por Sexo en la UCSM

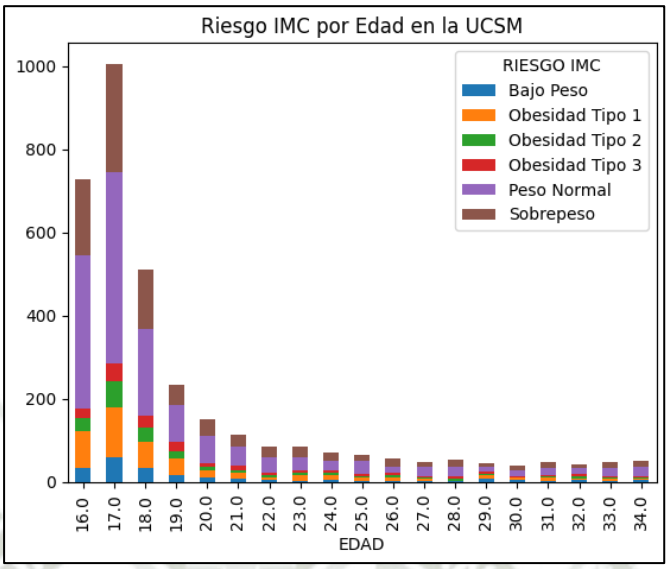


Nota: Google Colab,2024

d) Riesgo de IMC por Edad

Podemos interpretar el siguiente gráfico, mediante el rango de valores de IMC y deducir que hay más personas que sufren de sobrepeso desde los 16 a 18 años y también existen una obesidad de tipo 3 y esto principalmente se debe por una mala alimentación. (Ver Figura 53)

Figura. 53.
Gráfico de Barras de Riesgo IMC por Edad en la UCSM

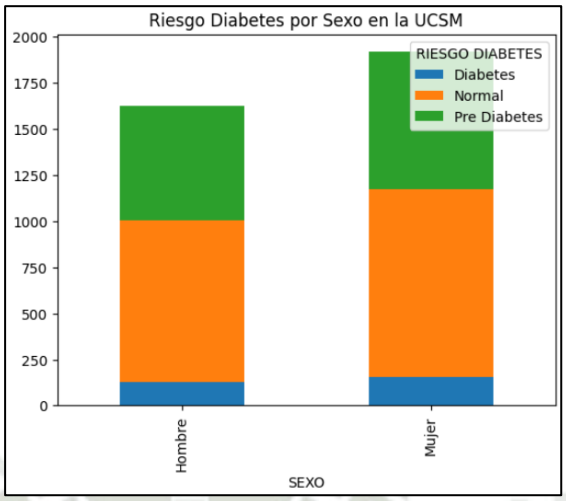


Nota: Google Colab,2024

e) Riesgo de Diabetes por Sexo

En el siguiente grafico se puede observar que hay más predominancia de diabetes en el sexo femenino que en el masculino y de igual forma la prediabetes, pero está asociado a que existen más personas del sexo femenino estudiando en la UCSM, un buen indicador es que hay un gran porcentaje de personas que tienen un diagnóstico normal, pero se quiere que este objetivo crezca. (Ver Figura 54)

Figura. 54.
Gráfico de Barras de Riesgo de Diabetes por Sexo en la UCSM

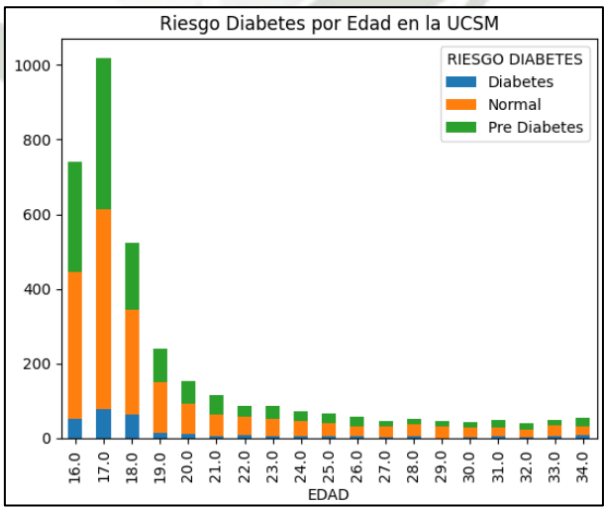


Nota: Google Colab,2024

f) Riesgo de Diabetes por Edad

Se puede observar que el mayor rango de personas con prediabetes esta entre los 16 a 20 años respectiva e igualmente para la diabetes por lo que está asociado a la mala alimentación. (Ver Figura 55)

Figura. 55.
Gráfico de Barras de Riesgo de Diabetes por Edad en la UCSM

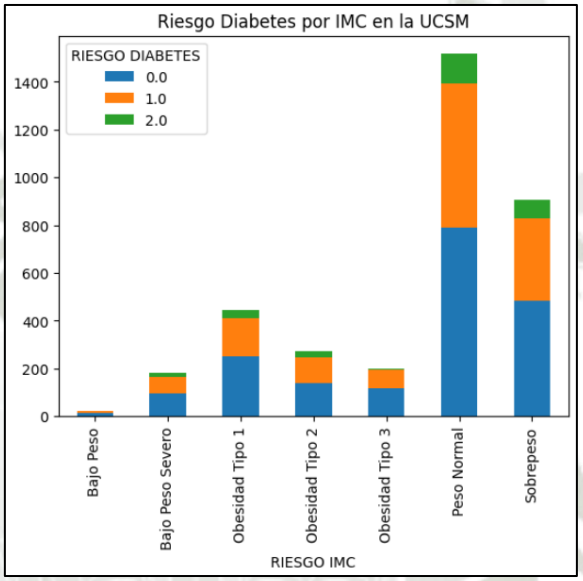


Nota: Google Colab,2024

g) Riesgo de Diabetes por IMC

Para el siguiente grafico se puede visualizar que el mayor número de casos de diabetes están en personas que tienen un peso normal y sobrepeso lo que hace referencia a malos hábitos alimenticios. (Ver Figura 56)

Figura. 56.
Gráfico de Barras por Riesgo de Diabetes e IMC en la UCSM



Nota: Google Colab,2024

3.2.4 Verificación de la Calidad de Datos

Tabla 17.
Verificación de la Calidad de Datos en el Dataset

Punto de Comprobación	Respuesta
Cobertura de los datos	El conjunto de valores lleva a un resultado que está basado en determinar casos de diabetes según su IMC, edad, actividad física, consumo de alcohol y glucosa.
Claves	En el dataset los campos claves están asociado a su IMC y la glucosa respectivamente.
Coherencia entre atributos y valores	Los datos son coherentes porque se relación entre sí para determinar el objetivo por ende son coherentes.
Campos en blanco y atributos	Se determino que en el dataset no se encontró datos vacíos.

Atributos erróneos, valores distintos con significados iguales	Para el dataset se encontró el campo sexo que estaba determinado por texto F y M Por lo que se transformó a: • F - Mujer • M - Hombre
Ortografía y Formato	La ortografía y formato son correctos
Desviaciones y posibles "ruidos"	Los datos para evitar el ruido se transformaron y agruparon en valores de 0 a 5 lo cual no causo esta interferencia en el entrenamiento.
Plausibilidad	No existe plausibilidad
Ruido e inconsistencia entre la Nota de datos	La Nota proviene de un mismo origen lo que no causa ruido e inconsistencias

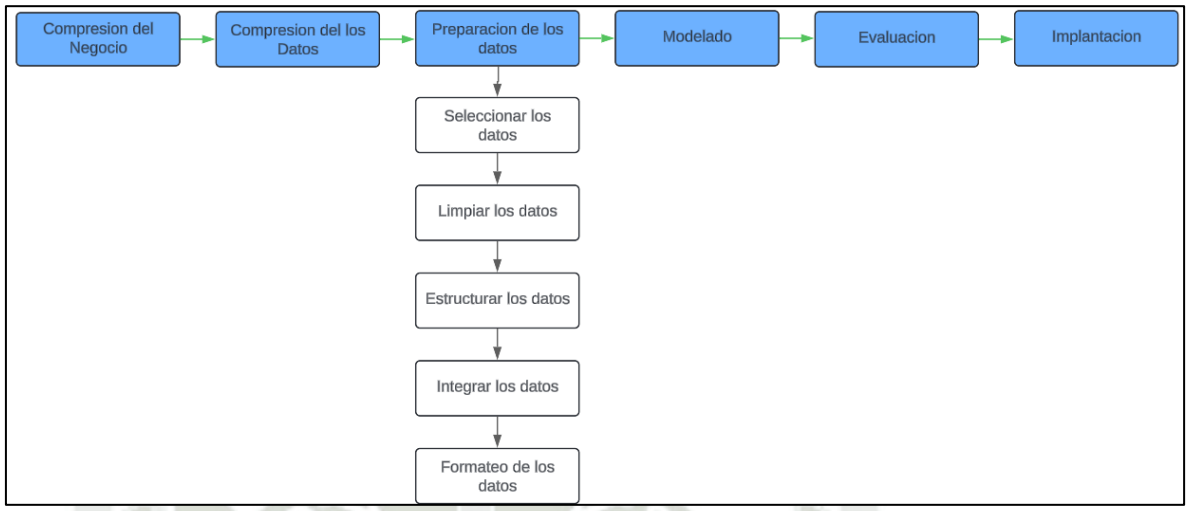
Nota: Google Colab,2024

En base a la comprobación de los datos mediante la exploración se puede determinar que los datos dentro del dataset están completo y cubren la necesidad para comenzar con el entrenamiento adecuado del mismo y las pruebas necesarias para poder conseguir el objetivo de determinar si un alumno tiene o no diabetes. Los datos no tienen errores, es la información extraída de una Nota confiable y tampoco se encontró valores negativos, nulos o inexistentes para que pueda generar un ruido al momento del entrenamiento.

3.3 Preparación de los Datos

En esta fase se procede a realizar una limpieza y preparación previa de los datos para el análisis. Se eliminarán los datos irrelevantes o datos duplicados y se procederán a realizar las transformaciones de los datos necesarias para integrarlos y así estén preparados para el modelado.

Figura. 57.
Fase III - Preparación de los Datos



Nota: Propia

3.3.1 Seleccionar los Datos

El dataset con el que se trabaja tiene 3542 instancias y 11 atributos, pero para el entrenamiento solo tendremos el uso de 4 atributos que es sexo, IMC, Glucosa, Antecedentes Familiares por Diabetes que más adelante haremos el formateo para favorecer el análisis, el dataset se visualiza de la siguiente manera:

Tabla 18.
Presentación del Dataset sin Transformación

Sexo	Edad	Peso	Talla	IMC	Glucosa	Antece Fami. Diabetes	Alcohol	Tabaco	Drogas	Act. Física
Mujer	17	62.0	1.59	24.52	90	NO	NO	NO	NO	NO
Mujer	16	48.5	1.57	19.68	160	NO	NO	NO	NO	NO
Hombre	21	88.7	1.80	27.38	155	NO	NO	NO	NO	SI
Hombre	16	50.5	1.74	16.68	200	NO	NO	NO	NO	NO
Mujer	16	58.0	1.57	23.53	100	NO	NO	NO	NO	NO

Nota: Propia

En la etapa de limpieza se tiene que realizar una conversión para poder obtener que los datos sean numéricos para el entrenamiento por lo cual más adelante lo realizaremos a

fin de evitar el ruido. Se va a tener una selección en base a los atributos e instancias del modelo respectivamente. Por ello tendremos el siguiente análisis:

A nivel de Instancias

A nivel de instancias vamos a usar todos los datos posibles, pero no serán en base a todos los atributos del dataset debido a que el más importante es la glucosa respectivamente. En este se tienen diversos casos que nos ayudaran al momento del entrenamiento y así estos casos ayuden a una mejor predicción que es el objetivo que se quiere lograr.

A nivel de Atributos

A nivel de atributos, no utilizaremos todos ya que solo nos importan los siguientes para realizar el entrenamiento:

a) Sexo
b) Edad
c) IMC
d) Antecedentes Diabetes
e) Glucosa

Pero se le adicionara un nuevo atributo el cual es el análisis del riesgo de glucosa establecido por la OMS, en donde según los niveles de glucosa se obtendrá si una persona tiene o no diabetes y se mencionó que los niveles serán los siguiente:

Tabla 19.
Interpretación de Riesgo de Diabetes en el Dataset

Valor de Glucosa	Diagnostico	Valor en el Dataset
• **Menor de 140 mg/dl**	Normal	0.0
• **Entre 140 mg/dl y 200 mg/dl**	Pre-Diabetes	0.0
• **Mayor a 200 mg/dl**	Diabetes	1.0

Nota: Propia

No se utilizarán los demás atributos como peso, talla, alcohol, drogas, actividad física ya que podría causar ruido en la predicción. Por lo que el dataset quedaría de la siguiente manera:

Tabla 20.
Selección de Datos a Nivel de Atributos para Predicción

Atributo	Tipo de Dato	Se utilizará en el modelo	Motivo
Edad	Numérico	Si	
Sexo	Carácter	Si	
Peso	Numérico	No	Es parte de la formula IMC que con la talla ofrece un mejor entendimiento dependiendo del rango establecido
Talla	Numérico	No	Es parte de la formula IMC que con el peso ofrece un mejor entendimiento dependiendo del rango establecido
IMC	Numérico	Si	
Glucosa	Numérico	Si	
Antece. Fami. Diabetes	Carácter	Si	
Alcohol	Carácter	No	No es relevante para el análisis
Tabaco	Carácter	No	No es relevante para el análisis
Drogas	Carácter	No	No es relevante para el análisis
Actividad Física	Carácter	No	No es relevante para el análisis
Riesgo Diabetes	Numérico	Si	

Nota: Propia

La exclusión de algunos atributos es debido a que no son relevantes para el entrenamiento y por ese motivo es que se tiene solo 5 atributos que ayudaran a ese cumplimiento respectivamente ya planteado en la fase 1 y así evitar el ruido en el entrenamiento.

3.3.2 Limpieza de Datos

El dataset está completo en base a las instancias que son 3542 y los atributos son los correctos para realizar el entrenamiento correctamente y luego proceder a tener las pruebas y así determinar la predicción correcta.

3.3.3 Estructuración de Datos

Los datos se han adecuado en el formato establecido para permitir realizar un modelado para tener una predicción. Es por ello por lo que se siguió los siguientes pasos:

- **Selección de datos de interés:** Se procedió a identificar los datos relevantes para el proyecto y se tuvo en cuenta todos los datos dentro del dataset para en análisis y modelado de predicción.
- **Limpieza de datos:** Se realizo una limpieza a nivel de instancias debido a que se tenía 3600 registros, pero solo se usara 3542 para el entrenamiento ya que la información eliminada tenía información faltante e incongruente.
- **Integrar los datos de cada conjunto de datos:** Los datos están integrados dentro de un mismo dataset por ello no se hizo uso de otra Nota de datos para integra.
- **Conversión de datos:** Se realizaron conversiones de los datos por ejemplo para nombrar las columnas y se tenga un mejor entendimiento, eliminar valores nulos, codificar las variables, poder eliminar ceros y convertir el tipo de datos para facilitar al modelo.

- **Exploración de datos:** Se hizo una exploración de los datos para poder transformarlos y comprender la data con la estadística y así tener una visión más clara del problema a tratar resaltando las características más importantes de los datos antes de proceder al modelado.

3.3.4 Formateo de Datos

Se procederá a realizar una estandarización para que el modelo usado para el entrenamiento sea más rápido y no presente inconsistencias a nivel de información o ruido en el mismo garantizando la calidad de los datos.

a) Formateo Atributo Sexo

El atributo se convertido a tipo carácter convirtiendo a F para Mujer y M para Hombre en el apartado 1.2.3 para la Exploración y tener un mejor entendimiento, pero para el modelo se hará el formateo y conversión de la siguiente manera:

Tabla 21.
Conversión de Atributo Sexo para Predicción

Datos	Conversión
Mujer	1.0
Hombre	2.0

Nota: Propia

Después de la conversión del atributo Sexo este quedara de la siguiente manera:

Tabla 22.
Visualización del Atributo Sexo después del Formateo

Sexo	Edad	IMC	Glucosa	Antece Fami. Diabetes
1.0	17	24.52	102	NO
1.0	16	19.68	112	NO
2.0	21	27.38	133	NO
2.0	16	16.68	125	NO
1.0	16	23.53	177	NO

Nota: Propia

b) Formateo Atributo Edad

El atributo es de tipo numérico. Pero en el modelo a utilizar se tendrá que colocar la parte decimal para que pueda reconocerlo como número y no como carácter de tipo object por lo que se adicionara la parte decimal respectivamente.

Tabla 23.
Atributo Edad antes del Formateo

Sexo	Edad	IMC	Glucosa	Antece Fami. Diabetes
1.0	17	24.52	102	NO
1.0	16	19.68	112	NO
2.0	21	27.38	133	NO
2.0	16	16.68	125	NO
1.0	16	23.53	177	NO

Nota: Propia

El formateo para este atributo consistirá en poder darle la parte decimal para que sea un dato numérico en el modelo y este pueda hacer el entrenamiento más rápido.

Tabla 24.
Visualización de Atributo Edad después del Formateo

Sexo	Edad	IMC	Glucosa	Antece Fami. Diabetes
1.0	17.0	24.52	90	NO
1.0	16.0	19.68	160	NO
2.0	21.0	27.38	155	NO
2.0	16.0	16.68	200	NO
1.0	16.0	23.53	100	NO

Nota: Propia

c) Formateo Atributo Glucosa

Es de tipo numérico, pero para el modelo le agregaremos la parte decimal en la cual nos ayudara a que modelo pueda hacer más rápido la predicción correspondiente.

Tabla 25.
Atributo Glucosa antes del Formateo

Sexo	Edad	IMC	Glucosa	Antece. Fami. Diabetes
1.0	17.0	24.52	102	NO
1.0	16.0	19.68	112	NO
2.0	21.0	27.38	133	NO
2.0	16.0	16.68	125	NO
1.0	16.0	23.53	177	NO

Nota: Propia

La conversión en si solo consiste en adicionar la parte decimal para que modelo lo tome como numérico con parte decimal y se visualizara de la siguiente manera.

Tabla 26.
Visualización del Atributo Glucosa después del Formateo

Sexo	Edad	IMC	Glucosa	Antece Fami. Diabetes
1.0	17.0	24.52	102.0	NO
1.0	16.0	19.68	112.0	NO
2.0	21.0	27.38	133.0	NO
2.0	16.0	16.68	125.0	NO
1.0	16.0	23.53	177.0	NO

Nota: Propia

d) Formateo Atributo Antecedentes Familiares Diabetes

Es de tipo carácter, pero para el modelo le agregaremos una conversión a tipo numérico que nos ayudara a determinar de mejor manera el entrenamiento por ello se realizara lo siguiente.

Tabla 27.
Conversión del Atributo Ant. Familiares por Diabetes en el Dataset

Datos	Conversión para Modelo
NO	0.0
SI	1.0

Nota: Propia

La conversión en si solo consiste en poder convertir los valores establecidos en el atributo de Antecedentes Familiares por Diabetes a “0 y 1” para que el modelo lo tome como numérico y se visualizara de la siguiente manera.

Y el dataset final ya se tendría listo para proceder con el modelado de este ya que todos los atributos seleccionados son de carácter numérico y el modelo puede hacer cálculos matemáticos y estadísticos de forma más rápida.

Tabla 28.
Atributo Antece. Fami. por Diabetes después del Formateo

Sexo	Edad	IMC	Glucosa	Antece Fami. Diabetes
1.0	17.0	24.52	102.0	0.0
1.0	16.0	19.68	112.0	0.0
2.0	21.0	27.38	133.0	0.0
2.0	16.0	16.68	125.0	0.0
1.0	16.0	23.53	177.0	0.0

Nota: Propia

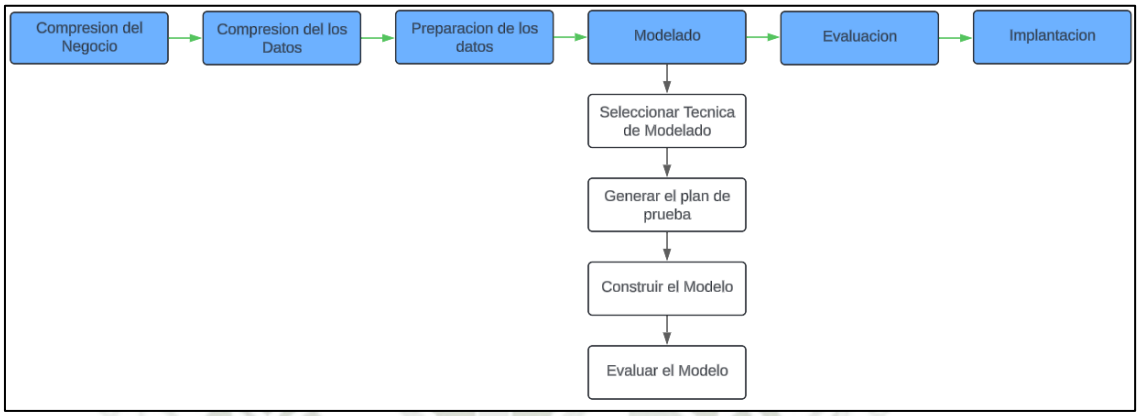
Para la finalización de la exploración en la fase 3 se determinó que no existían valores nulos o incompletos por lo que las 3542 instancian ya están correctamente estandarizadas para poder realizar el entrenamiento correctamente.

3.4 Modelado

La metodología CRISP-DM no se limita a tener un tipo de aprendizaje de forma particular o solo un tipo en específico. Sino más bien puede enfocarse a varios tipos de aprendizaje, pero el modelo de aprendizaje a usar depende mucho de varios factores como es el tipo del problema, la naturaleza de los datos y también los recursos que se tengan a nivel de software y hardware. Por lo tanto, la cuarta fase de la metodología CRISP-DM se tiene que escoger la técnica de modelado más adecuada para el análisis y el cumplimiento del objetivo.

Por lo que una vez escogido se pasara a tener casos de prueba para la predicción y medir ese cumplimiento para validar el funcionamiento.

Figura. 58.
Fase IV – Modelado



Nota: Propia

Una vez que se revisó el objetivo del conjunto para informar el objetivo del negocio, se construirá el modelo para ajustar los parámetros y evaluarlo determinado su calidad y selección adecuada para que sea usado en la predicción.

3.4.1 Selección de la Técnica de Modelado

Para la cuarta etapa se tendrá que poder escoger un modelo el cual servirá para realizar el aprendizaje automático, por ello se usara la herramienta Google Colab para determinar cuál es el mejor. En su versión de actualización desde el 21 de febrero del 2024.

Google Colab es un servicio alojado de Jupyter Notebook que no requiere configuración y que ofrece acceso gratuito a recursos de computación, como GPUs y TPUs. Colab es una solución especialmente adecuada para el aprendizaje automático, la ciencia de datos y la educación. Google Colab es una herramienta de libre acceso por lo tanto no se paga por usar de ella, pero tiene sus limitaciones al ser de versión web. (Google Colab, 2024)

Técnicas de Modelado en Google Colab

Dentro de Google Colab se puede usar varios modelos de análisis que están basados en las técnicas de aprendizaje automático supervisado como son los siguientes:

a) **Arboles de Decisión**

Un árbol de decisión es un algoritmo de aprendizaje supervisado que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. Al poder consultar todo da una respuesta en base a las iteraciones que se realicen. (Smith,2024)

b) **Análisis Bayesiano**

El análisis Bayesiano es una inferencia estadística en la cual consiste en poder determinar en base a un conjunto de datos la probabilidad o la hipótesis de que algo suceda y puede ser cierta o no. (Smith,2024)

c) **Regresión Logística**

La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. Normalmente, la predicción tiene un número finito de resultados, como un sí o un no. (Smith,2024)

d) **Regresión Lineal**

La regresión lineal es una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal. (Smith,2024)

e) K-vecinos más cercanos

El algoritmo de k vecinos más cercanos, también conocido como KNN o k-NN, es un clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual, generalmente se usa como un algoritmo de clasificación. (Smith,2024)

f) Máquina de Soporte Vectorial

Es un algoritmo de aprendizaje automático supervisado utilizado principalmente para clasificación y, en algunos casos, para regresión. Su objetivo principal es encontrar un hiperplano que mejor separe las clases en el espacio de características. (Smith,2024)

g) Red Neuronal

Una red neuronal es un modelo computacional inspirado en el funcionamiento del cerebro humano. Está compuesta por un conjunto de unidades interconectadas llamadas neuronas artificiales o nodos, organizadas en capas. Estas redes son capaces de aprender y reconocer patrones complejos. (Smith,2024)

En conclusión, analizando todo el dataset en la Fase 3 en al apartado 3.1 de la metodología CRISP-DM y considerando el objetivo principal del proyecto de investigación de poder predecir la diabetes en los alumnos de pregrado de una Universidad Privada en la ciudad de Arequipa. El dataset pasará la prueba por todos los algoritmos de aprendizaje supervisado mencionados y se obtendrá cual fue el mejor resultado para poder aplicarlo en la predicción.

Los algoritmos seleccionados tienen sustento en la validación del autor S. Rajaraman, el cual desarrollo la investigación denominada “Modelado predictivo de la diabetes mellitus mediante algoritmos de aprendizaje automático: Un estudio integral”. El estudio tuvo como objetivo evaluar y comparar la eficacia de diversos algoritmos de aprendizaje automático en la predicción de la diabetes mellitus. Se utilizaron conjuntos de datos médicos que incluyen características biométricas y clínicas relevantes para la diabetes, como niveles de glucosa, presión arterial, índice de masa corporal (IMC), entre otros. Comparando 8 algoritmos como regresión lineal, análisis bayesiano, arboles de decisión, KNN Vecinos Cercanos, Maquinas de Soporte de Vectores y Redes Neuronales. Determinando que la regresión lineal mostro un rendimiento razonable, el algoritmo de análisis bayesiano mostro eficacia para manejar la incertidumbre y proporcionar predicciones probabilísticas y en métodos avanzados como redes neuronales o Maquina de soporte vectorial ofrecieron altos niveles de precisión en la predicción, especialmente en escenarios con datos complejos y no lineales. Los métodos basados en árboles de decisión y KNN mostraron buenos resultados en términos de precisión y capacidad para manejar datos con características variadas y la Regresión Logística y Análisis Bayesiano se destacaron por su interpretabilidad y capacidad para manejar datos con distribuciones estadísticas conocidas. Por lo que el uso de técnicas combinadas y que integre varios algoritmos puede aprovechar las fortalezas individuales de los mismos y mejorar la precisión general del resultado esperado. (S. Rajaraman,2017)

Habiendo encontrado esta prueba de algoritmos en la predicción de la diabetes es que se escogen dichos algoritmos para que sean probados en el dataset de estudiantes de pregrado de la Universidad Privada en la ciudad de Arequipa y así obtener un resultado favorable.

3.4.2 Generación del Plan de Prueba

En esta Fase 4 de la metodología CRISP-DM, es fundamental establecer una estrategia para evaluar la precisión y utilidad del modelo. Para que se pueda cumplir con el fin del modelo para hacer la predicción de la diabetes en los alumnos de 16 a 34 años en una Universidad Privada en la ciudad de Arequipa.

Con la herramienta Google Colab se realizarán las evaluaciones por cada técnica de aprendizaje automático mencionada en donde en base a ellos se tomará solo uno en cuenta y se aplicará el entrenamiento correspondiente. En la fase 2 se realizó el entendimiento de los datos por lo que ya se tiene todos los datos completos para hacer las pruebas correspondientes a cada técnica. En Google Colab permite realizar cada uno de ellos y medirlos adecuadamente para ver cual tiene el mejor rendimiento.

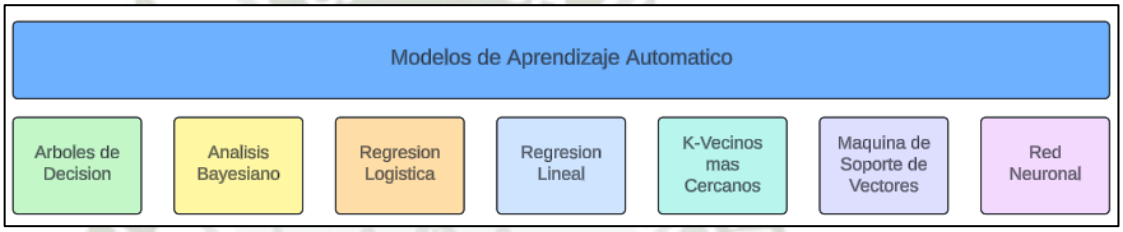
Los pasos establecidos para la evaluación serán:

- Se hará una división de variables las cuales un conjunto de ellas servirá para el entrenamiento y la otra para la validación de la predicción.
- Dependiendo del modelo a utilizar se definirán las métricas más aptas para poder evaluarlo según la matriz de confusión de cada una y la capacidad que tendrá para hacer las predicciones precisas.
- Una vez se tenga el modelo seleccionado se harán las pruebas correspondientes enviándoles datos reales aleatorios para que realice una predicción.
- Finalmente se hará una comparación al final de las métricas de los modelos a fin de escoger la mejor dependiendo de su rendimiento.

3.4.3 Construcción del Modelo

En la presente etapa de la construcción de los modelos de aprendizaje automático, se va a detallar la lógica detrás de la construcción de modelos para el dataset. Se empezará a ejecutar los modelos antes mencionados en Google Colab. Para ello se explicarán como se probó cada uno paso a paso y que salidas tiene respectivamente con la evaluación mediante las métricas.

Figura. 59.
Modelos de Aprendizaje Automático para Predicción



Nota: Propia

Pasos previos a la aplicación de algoritmos

Para los 5 modelos escogidos de aprendizaje automático, se tendrá que importar cada Liberia que ayudará a realizar el análisis, juntamente con todo el dataset donde se tiene 3542 instancias y se usara solo los campos necesarios para usar el modelo.

- **Importación Previa de Dataset**

  Para la importación del dataset tendremos en cuenta que está alojado en Google drive ya que es un servicio en la nube para lo cual se hará de manera más rápida la conexión.

Figura. 60.
Conexión a Google Drive para extracción de Dataset

```
from google.colab import drive
drive.mount('/content/drive')
```

Nota: Google Colab,2024

Como el dataset está en Google drive se tendrá que llamar al archivo por su nombre y así cargar todas las 3542 instancias y 11 atributos correspondientes.

Figura. 61.
Lectura de Dataset Extraido en Google Colab

```
with open("/content/drive/MyDrive/Dataset_Diabetes.csv", "rb") as f:
    encoding = chardet.detect(f.read())["encoding"]

df = pd.read_csv("/content/drive/MyDrive/Dataset_Diabetes_UCSM_2023.csv", encoding=encoding)
```

Nota: Google Colab,2024

- **Selección de datos para probar**

  Como habíamos mencionado usaremos 5 atributos que nos servirá para el análisis, lo cuales son la edad, el sexo, IMC, glucosa, antecedentes familiares de diabetes y riesgo de Diabetes. Por lo cual para hacer las pruebas de técnicas de aprendizaje.

Figura. 62.
Selección de Datos para prueba de Modelos de Aprendizaje Automático

```
#Creamos un nuevo data set para estudio
dataframe = df[['RIESGO DIABETES','SEXO','EDAD','IMC','GLUCOSA','ANT. FAMILIAR DIABETES']]
dataframe.head(5)
```

Nota: Google Colab,2024

Por lo que una vez seleccionado el dataset quedara de la siguiente manera:

Tabla 29.
Dataset Importado para Pruebas

Riesgo Diabetes	Sexo	Edad	IMC	Glucosa	Ant. Fami. Diabetes
0.0	1.0	17.0	24.52	102.0	0.0
0.0	1.0	16.0	19.68	112.0	0.0
0.0	2.0	21.0	27.38	133.0	0.0
0.0	2.0	16.0	16.68	125.0	0.0
0.0	1.0	16.0	23.53	177.0	0.0

Nota: Propia

- **Selección de variables para entrenamiento**

Ya teniendo un dataset para pruebas con todos los atributos numéricos se podrá evaluar correctamente cada técnica de aprendizaje y escoger la mejor respectivamente y usarla en el entrenamiento.

Para el entrenamiento se separará las variables predictoras y las variables condicionantes. Para lo cual tendremos todas las variables que ayudan a la predicción en la variable "X" como Sexo, Edad, IMC, Ant. Familiar Diabetes y en "Y" la variable predictora para la Diabetes que es el Riesgo Diabetes.

Figura. 63.
Creación de Variables Predictoras y Dependientes

```
#Variable a Predecir ( Separa las Variables )
X = dataframe.iloc[:,1:6]

#Variable a Predecir
Y = dataframe.iloc[:,0]
```

Nota: Google Colab,2024

Los parámetros establecidos definidos serán los siguientes:

- **X_train**: A partir del dataset escogido se crea un array con las dimensiones de todas las variables escogidas para el entrenamiento
- **Y_train**: Es un array con dimensiones de todos los resultados posibles a partir de las variables escogidas en el X_train
- **X_test**: Es el número de instancias escogido para realizar las pruebas con todos los datos correspondientes.
- **Y_test**: Es el número de instancias escogido para poder asociarlo al X_test con el cual encontraremos todos los resultados posibles.

a) Arboles de Decisión

En Google Colab se puede desarrollar un árbol de decisión, pero este debe tener una dataset con el cual hacer las pruebas correspondientes.

Figura. 64.
Importación de Variables y Algoritmo de Árbol de Decisión

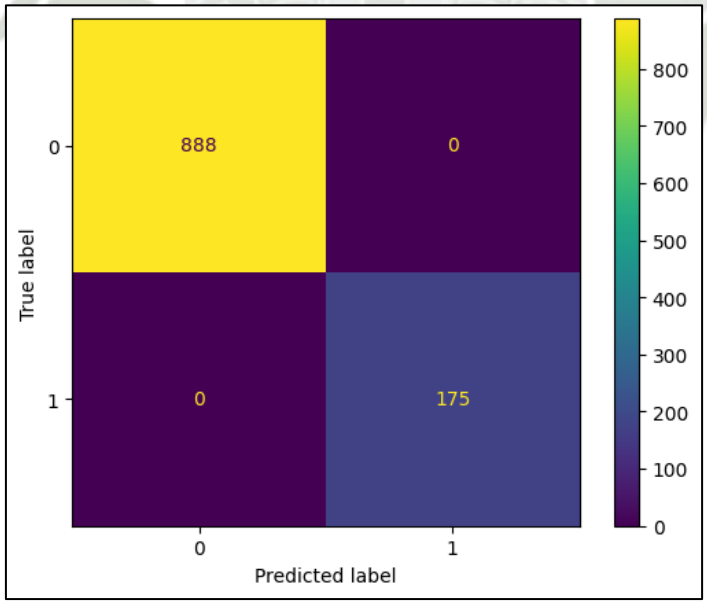
```
from sklearn.model_selection import train_test_split

#X_train y Y_train para entrenamiento
#Y_test y Y_test para prueba
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,train_size=0.7,random_state=0)
Y_test.info()
```

```
from sklearn.tree import DecisionTreeClassifier
#Llamamos al constructor del arbol de decision
arbol = DecisionTreeClassifier(max_depth=4)
arbol_diabetes = arbol.fit(X_train,Y_train)
```

Nota: Google Colab,2024

La Matriz de Confusión resultante es el siguiente:

Figura. 65.
Matriz de Confusión al aplicar el algoritmo de Árbol de Decisión



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 30.
Métricas de Aprendizaje de Árbol de Decisión

Métrica	%
Accuracy (Exactitud)	1.000 %
Precisión	1.000 %
Exhaustividad	1.000 %
Root Mean Square Error (RMSE)	0.000 %
Mean Absolute Error (MAE)	0.000 %
Spearman RHO	1.000 %
Relative Absolute Error (RAE)	0.000 %

Nota: Propia

b) Análisis Bayesiano

En Google Colab se puede desarrollar un análisis bayesiano, pero este debe tener un dataset con el cual hacer las pruebas correspondientes que ya se importó en los pasos previos.

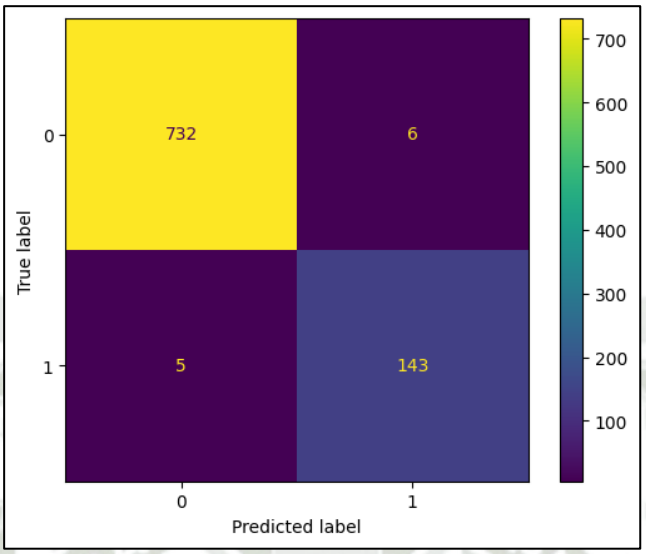
Figura. 66.
Importación de librería para uso de algoritmo Naive Bayes

```
#El Modelo de Naive Bayes
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X_train,Y_train)
y_pred2 = gnb.predict(x_test)
y_pred2
```

Nota: Google Colab,2024

La matriz de confusión resultante es la siguiente:

Figura. 67.
Matriz de Confusión del Aprendizaje de Análisis Bayesiano



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 31.
Métricas de la aplicación del algoritmo de Naive Bayes

Métrica	%
Accuracy (Exactitud)	0.987 %
Precisión	0.959 %
Exhaustividad	0.966 %
Root Mean Square Error (RMSE)	0.111 %
Mean Absolute Error (MAE)	0.012 %
Spearman RHO	0.955 %
Relative Absolute Error (RAE)	0.044 %

Nota: Propia

c) Regresión Logística

La regresión logística consiste en poder medir la probabilidad de una variable dependiente binaria este influenciada por una o más variables en si, por lo que Google Colab permite realizar una predicción ya sea de éxito o fallo.

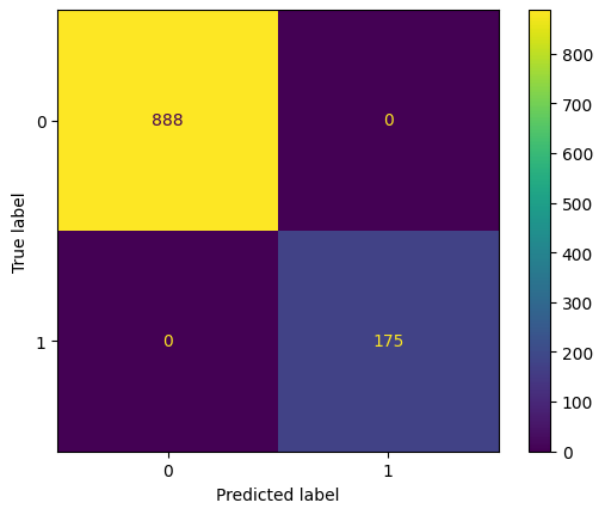
Figura. 68.
Importación de la librería para el uso de Regresión Logística

```
from sklearn.linear_model import LogisticRegression
reg = LogisticRegression(random_state=0)
reg.fit(X_train4,Y_train4)
#Prediccion
y_pred4 = reg.predict(x_test4)
```

Nota: Google Colab,2024

La matriz de confusión resultante después de la predicción es la siguiente:

Figura. 69.
Matriz de Confusión del algoritmo de Regresión Logística



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 32.
Métricas del Aprendizaje de Regresión Logística

Métrica	%
Accuracy (Exactitud)	1.000 %
Precisión	1.000 %
Exhaustividad	1.000 %
Root Mean Square Error (RMSE)	0.000 %
Mean Absolute Error (MAE)	0.000 %
Spearman RHO	1.000 %
Relative Absolute Error (RAE)	0.000 %

Nota: Propia

d) Regresión Lineal

La regresión lineal es un modelo estadístico que se basa en variables dependientes que se les llama variables respuesta y una o más variables independientes que se les llama predictoras. Por lo que también se pueden comparar para hacer comparaciones en Google Colab.

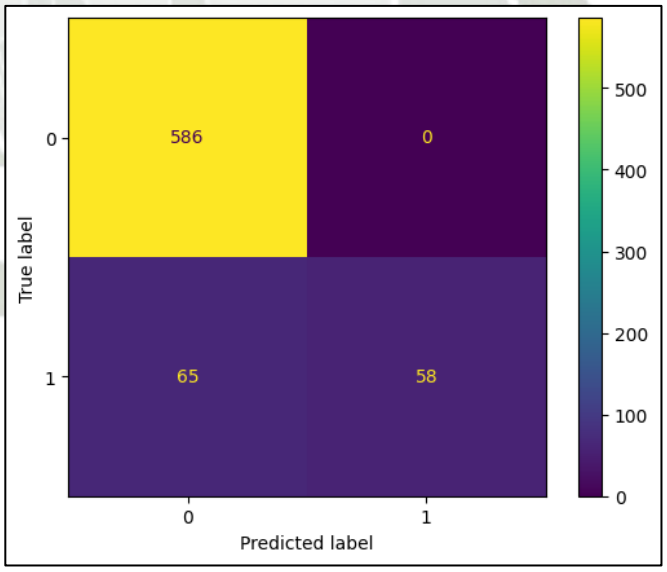
Figura. 70.
Importación de Librería para Regresión Lineal

```
from sklearn.linear_model import LinearRegression
reg_lineal = LinearRegression()
reg_lineal.fit(X_train5,Y_train5)
```

Nota: Google Colab,2024

La matriz de confusión resultante después de la predicción es la siguiente:

Figura. 71.
Matriz de Confusión de Aprendizaje de Regresión Lineal



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 33.
Métricas de Aprendizaje de Regresión Lineal

Métrica	%
Accuracy (Exactitud)	0.908 %
Precisión	1.000 %
Exhaustividad	0.471 %
Root Mean Square Error (RMSE)	0.302 %
Mean Absolute Error (MAE)	0.091 %
Spearman RHO	0.651 %
Relative Absolute Error (RAE)	0.319 %

Nota: Propia

e) KNN- Vecinos Cercanos

En Google Colab se puede desarrollar el algoritmo de KNN – Vecinos cercanos, pero este debe tener un dataset con el cual hacer las pruebas correspondientes que ya se importó en los pasos previos.

Para ello tendremos que separar en 2 variables los datos es coger con el fin de que sea más entendible para el modelo a predecir y se obtendrá de la siguiente manera:

Figura. 72.
Variables para Análisis de algoritmo KNN-Vecinos Cercanos

```
data_x_knn = dt_nb.iloc[:,1:dt_nb.columns.size].values
data_x_knn
```

```
data_y_knn = dt_nb.iloc[:,dt_nb.columns.size-1].values
data_y_knn
```

Nota: Google Colab,2024

Importaremos el algoritmo de vecinos cercanos en donde usaremos la métrica euclidiana para determinar la predicción, por ello tendremos siguiente:

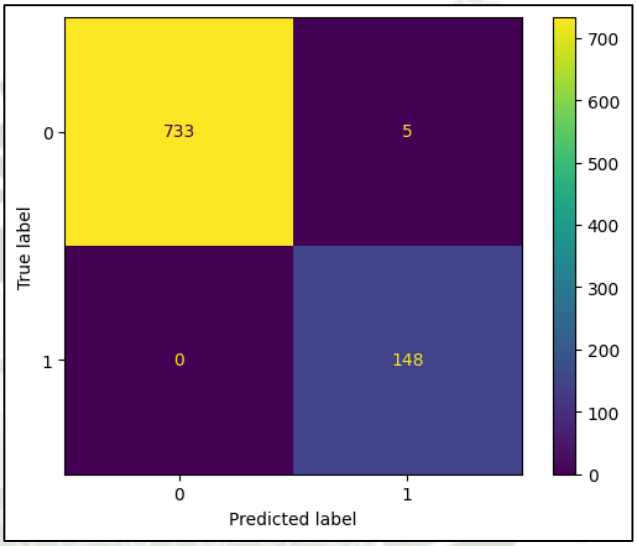
Figura. 73.
Importación de la librería KNN-Vecinos Cercanos

```
#Vecinos mas cercanos
algoritmo = KNeighborsClassifier(n_neighbors=5, metric= 'euclidean')
```

Nota: Google Colab,2024

Después del entrenamiento la matriz de confusión resultante es la siguiente:

Figura. 74.
Matriz de Confusión de Aprendizaje de KNN-Vecinos Cercanos



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 34.
Métricas de Aprendizaje de algoritmo KNN-Vecinos Cercanos

Métrica	%
Accuracy (Exactitud)	0.994 %
Precisión	0.967 %
Exhaustividad	1.000 %
Root Mean Square Error (RMSE)	0.075 %
Mean Absolute Error (MAE)	0.005 %
Spearman RHO	0.980 %
Relative Absolute Error (RAE)	0.020 %

Nota: Propia

f) Máquina de Soporte de Vectores

Este algoritmo de aprendizaje supervisado es usado para la clasificación y regresión ya que su objetivo es poder encontrar el hiperplano optimo que mejor separa los puntos de datos de diferentes clases en un espacio determinado. Es por ello por lo que en Google Colab se puede realizar esta representación.

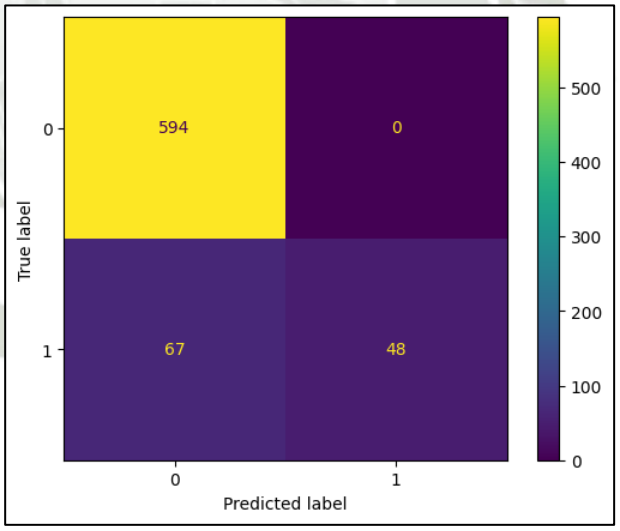
Figura. 75.
Importación de la librería de Maquina de Soporte de Vectores

```
from sklearn import svm
from sklearn.model_selection import cross_val_score
svm = SVC()
svm.fit(X_train6,Y_train6)
```

Nota: Google Colab,2024

Después del entrenamiento la matriz de confusión resultante es la siguiente:

Figura. 76.
Matriz de Confusión de Aprendizaje de Maquina de Soporte de Vectores



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 35.
Métricas de Aprendizaje de Maquina de Soporte de Vectores

Métrica	%
Accuracy (Exactitud)	0.905 %
Precisión	1.000 %
Exhaustividad	0.417 %
Root Mean Square Error (RMSE)	0.307 %
Mean Absolute Error (MAE)	0.094 %
Spearman RHO	0.612 %
Relative Absolute Error (RAE)	0.347 %

Nota: Propia

g) Red Neuronal

Una red neuronal como su nombre lo dice se asemeja mucho al funcionamiento del sistema nervioso por el conjunto de neuronas y la sinapsis que están realizan al procesar la información por lo que para esta prueba de modelo usaremos Google Colab con "MLPRegressor" que es una biblioteca de aprendizaje de Python para crear redes neuronales artificiales permitiendo entrenar modelos.

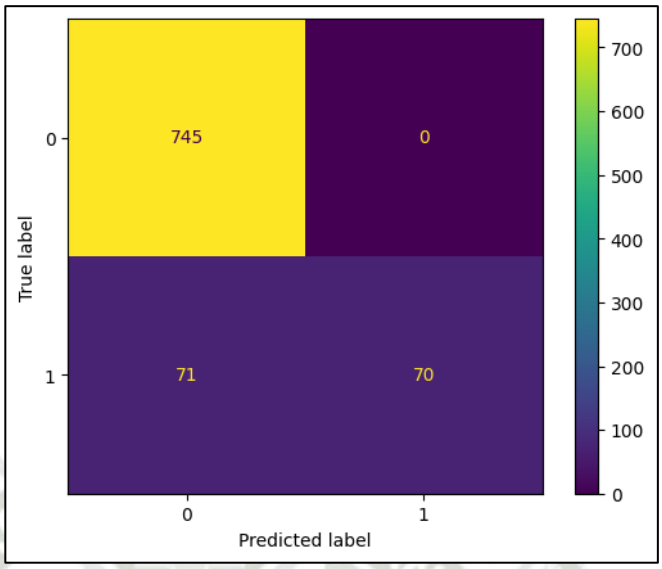
Figura. 77.
Importación de librería MLPRegressor para Red Neuronal

```
from sklearn.neural_network import MLPRegressor
mlr = MLPRegressor(solver='lbfgs',alpha=1e-5,hidden_layer_sizes=(3,3),random_state=1)
mlr.fit(X_train7,Y_train7)
```

Nota: Google Colab,2024

La matriz de confusión resultante después del entrenamiento esta siguiente:

Figura. 78.
Matriz de Confusión de Aprendizaje de Red Neuronal



Nota: Google Colab,2024

Y las métricas evaluadas para este modelo fueron las siguientes:

Tabla 36.
Métricas de Aprendizaje de Red Neuronal

Métrica	%
Accuracy (Exactitud)	0.919 %
Precisión	1.000 %
Exhaustividad	0.496 %
Root Mean Square Error (RMSE)	0.283 %
Mean Absolute Error (MAE)	0.008 %
Spearman RHO	0.673 %
Relative Absolute Error (RAE)	0.299 %

Nota: Propia

3.4.4 Evaluación del Modelo

Una vez realizado el modelado en los algoritmos de aprendizaje automático anteriormente en el apartado 1.4.3, se cumplió con los pasos establecidos en el Plan de Prueba y ya habiendo hecho las predicciones para evaluar los modelos por las métricas de

aprendizaje supervisado tendremos en cuenta lo siguiente consideraciones de las métricas para la evaluación:

- **Exactitud o Accuracy:** Mide la proporción de predicciones correctas sobre el total de predicciones realizadas.
- **Precisión:** Sirve para medir la proporción de los verdaderos positivos sobre el total de instancias clasificadas como positivas con el fin de minimizar a los falsos positivos.
- **Exhaustividad:** Mide la proporción de los verdaderos positivos sobre el total de instancias que son realmente positivas con el objetivo de minimizar a los falsos negativos.

Por otro lado, mediante la matriz de confusión se mostrará los resultados de las predicciones con la finalidad de mostrar la cantidad de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos de acuerdo con los valores obtenidos.

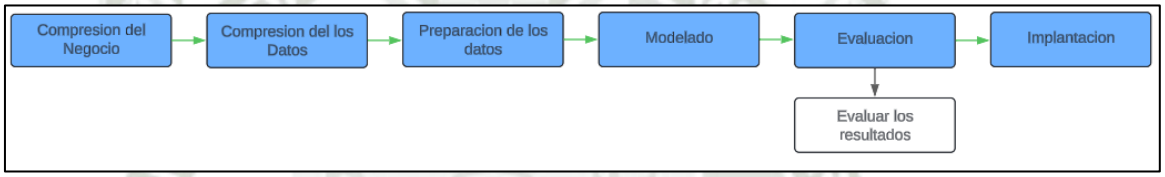
- **Error Medio Absoluto (MSE):** Sirve para medir la media de los errores al cuadrado entre las predicciones de los valores reales con el objetivo de poder penalizar los errores más grandes.
- **Error Cuadrático Medio (RMSE):** Sirve para medir la raíz cuadrada del Error Medio Absoluto y proporcionar así una medida de error que este en la misma escala del valor objetivo.
- **Error Relativo Absoluto (RAE):** Sirve para medir el cálculo entre el valor aproximado que son los verdaderos positivos entre los falsos positivos para así saber su proporción.

- **Spearman RHO:** Indica la relación existente en la proporción de la predicción varia de -1 a 1 si se acerca a 1 es favorable, pero si es cercano a 0 es que no hay ninguna relación entre las variables de predicción.

3.5 Evaluación

Para la fase 5 de evaluación en la metodología CRISP-DM se van a evaluar los resultados obtenidos en la fase 4. Se revisará el proceso con las fases anteriores hasta obtener el resultado más optimo en el cumplimiento del objetivo establecido en la Fase 1.

Figura. 79.
Fase V - Evaluación



Nota: Propia

3.5.1 Evaluación de los resultados

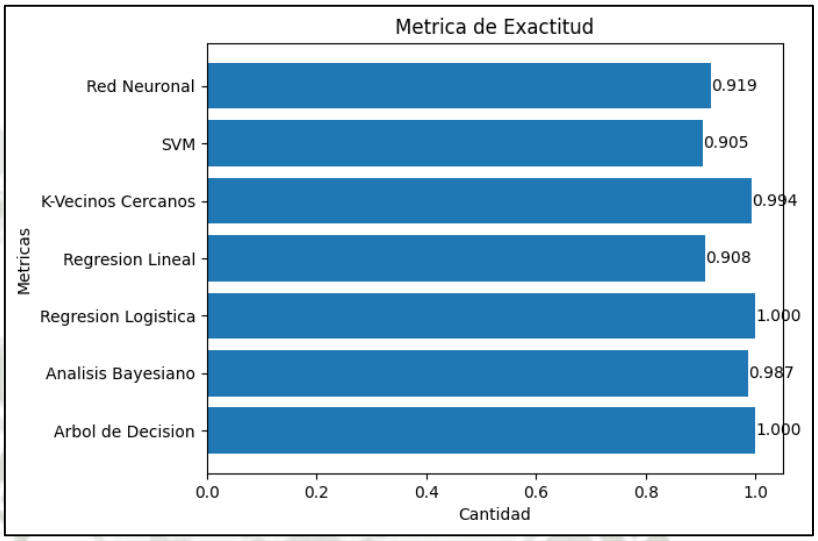
En este apartado evaluaremos los modelos en base a las métricas obtenidas por los 7 algoritmos de aprendizaje automático, por ello tenemos los siguientes resultados:

a) Exactitud

Para la métrica de exactitud (accuracy) se tiene que los valores más exactos son la aplicación de árboles de decisión y regresión logística, pero por otra parte los valores más cercanos a tener una buena predicción también se encuentran en la aplicación de K-vecinos cercanos. Los demás valores de los algoritmos también son buenos ya que superan el 90% pero no llegan a ser los ideales a comparación de los demás. Ya que la métrica mide que tan bien el algoritmo puede clasificar las

instancias en base a los datos de prueba sin tener en cuenta los falsos positivos o falsos negativos de la matriz de confusión. (Ver Figura 80)

Figura. 80.
Análisis de Métrica Accuracy



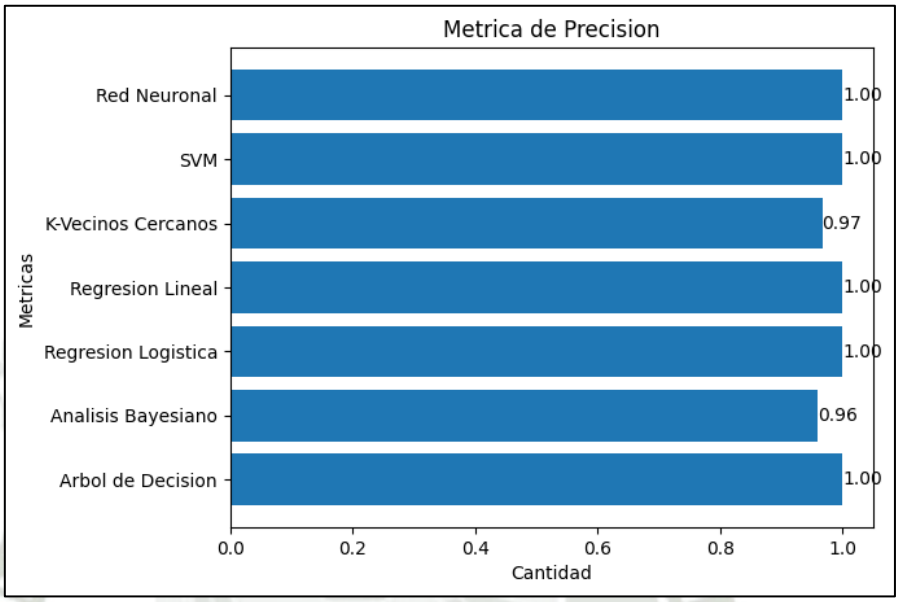
Nota: Propia

b) Precisión

En la métrica de precisión tenemos más opciones posibles para poder aplicar ya que esta métrica nos permite evaluar que tan capaz es el modelo de hacer predicciones sobre los datos de prueba. Y tenemos el análisis de que para hacer una buena predicción se tiene una red neuronal, SVM, regresión lineal, pero esta no tuvo un buen accuracy por lo tanto está en desventaja ya que modelos de aprendizaje con un buen accuracy se complementan con las pruebas, también se tiene la regresión logística pero los demás están muy cerca ya que las predicciones pasan el 95% de asertividad pero se sigue viendo una desventaja respecto a K-vecinos cercanos y el análisis bayesiano debido a que tanto en el accuracy como en la precisión no llegan

a ser ideales para la predicción por estar más debajo de la precisión perfecta de 1.000. (Ver Figura 81)

Figura. 81.
Análisis de Métrica de Precisión



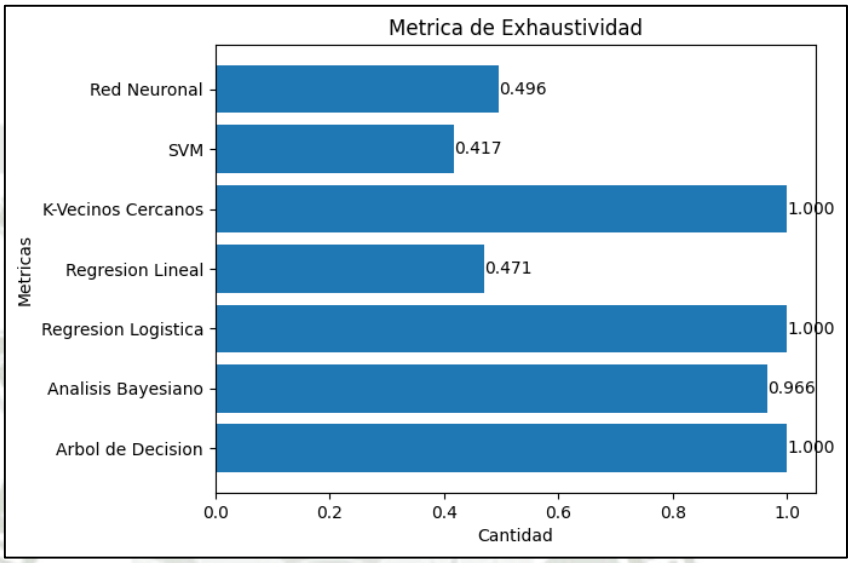
Nota: Propia

c) Exhaustividad

Respecto a la exhaustividad tenemos que la métrica nos permite medir todas las instancias positivas presentes en el dataset por lo que se dice que el modelo puede recordar o capturar los datos de casos positivos de predicción. En ese caso tenemos como mejores modelos usar el K-vecinos cercanos, pero este no llego a 1.000 en el accuracy y la precisión por lo que está en desventaja, sigue el algoritmo de regresión logística que si tuvo un buen rendimiento y el de árbol de decisión que al igual tiene un buen rendimiento. Los demás modelos solo el que se acerca a un valor esperado es el análisis bayesiano pero que anteriormente no llego a ser ideal y los que quedan

al último son la red neuronal, SVM y regresión lineal que para esta métrica no deberían ser considerados al no superar el 50%.

Figura. 82.
Análisis de Métrica Exhaustividad

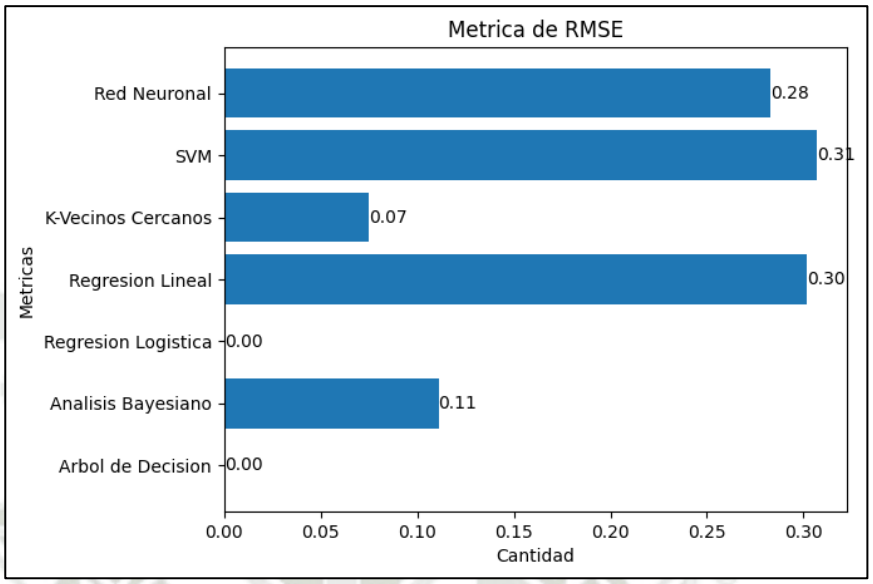


Nota: Propia

d) RMSE

Para la métrica del Error cuadrático medio se tiene lo opuesto al análisis de las métricas ya que este mide las discrepancias que hay entre las predicciones del modelo y los valores de la prueba. Por lo tanto, el valor más bajo representa que no hay alguna discrepancia por lo que usar el modelo de regresión logística y árbol de decisión sería lo ideal y por otro lado los otros 2 algoritmos para usar puede ser K-vecinos cercanos y análisis bayesiano ya que el error es bajo a comparación de los demás modelos que si superan el 20% de error lo que ocasionaría errores en la predicción.

Figura. 83.
Análisis de Métrica RMSE



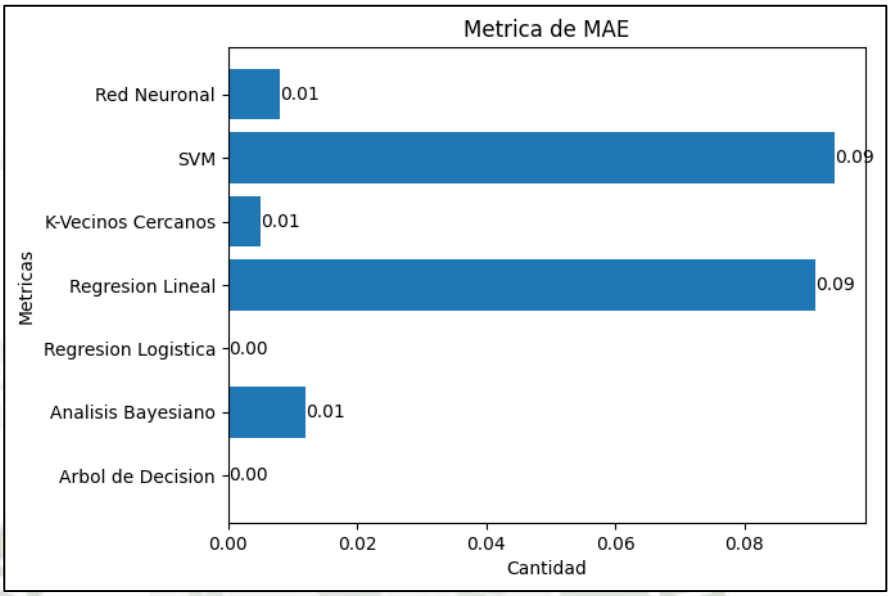
Nota: Propia

e) MAE

La métrica de Error Medio Absoluto ayuda a el modelo a poder medir la precisión, pero en base a las predicciones sin elevarlas al cuadrado como el error cuadrático medio, solo toma los valores con un valor absoluto y los valores menores indican que se está haciendo una mejor predicción para la variable. Por lo tanto, según el análisis se tiene que sigue habiendo los 2 modelos que tienen un buen rendimiento los cuales es la regresión logística y el árbol de decisión. Sin embargo, también se podría usar el modelo de K-vecinos cercanos, análisis bayesiano y red neuronal porque su valor es muy bajo a comparación de los demás modelos, pero el modelo de red neuronal no tuve un buen rendimiento en la métrica de RMSE por lo que las predicciones en algunos factores de prueba pueden ser erradas. En cambio, los modelos de SVM y regresión lineal tienen un porcentaje bajo, pero de igual forma

no son los más ideales para usar por lo tanto no se tomarán en cuenta para la métrica establecida.

Figura. 84.
Análisis de Métrica MAE



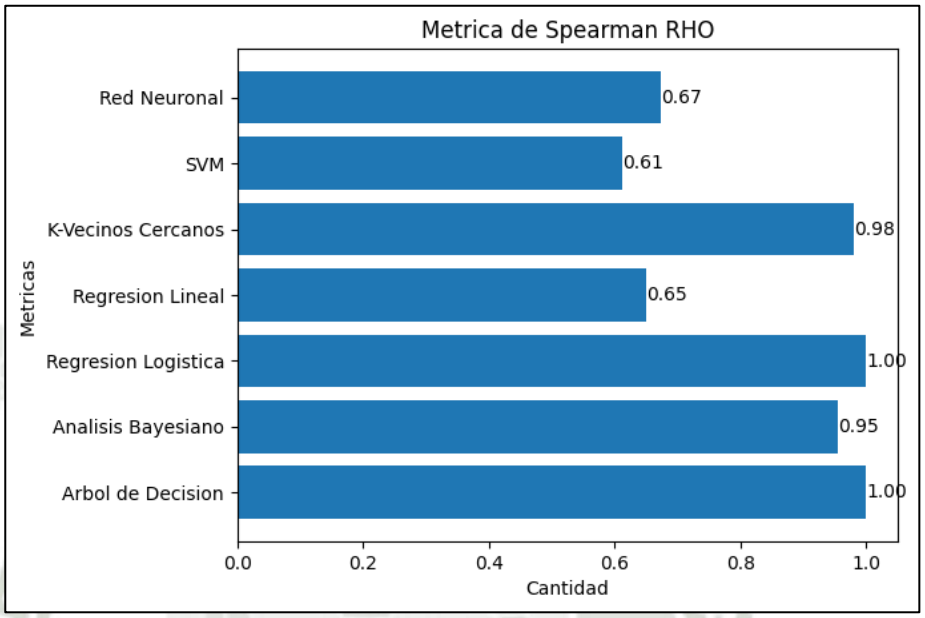
Nota: Propia

f) Spearman RHO

En cuestión de la métrica de spearman ayuda a poder revisar la relación que existe entre los datos reales y los datos de prueba respectivamente para determinar que tan bien se encuentran relacionadas para realizar las predicciones. Por lo que se tiene que la regresión logística y el modelo de árbol de decisión son los más ideales y han tenido hasta el momento un buen rendimiento pero también está el análisis bayesiano y K-vecinos cercanos pero estos en comparación con las otras métricas no tuvieron el rendimiento esperado por lo que aplicarlo puede generar incongruencias al realizar las pruebas y los demás algoritmos como la regresión

lineal, red neuronal y SVM tienen valores muy bajos por lo que para esta métrica no se tendrán en cuenta y también tienen un bajo rendimiento respecto a eso.

Figura. 85.
Análisis de Métrica Spearman RHO



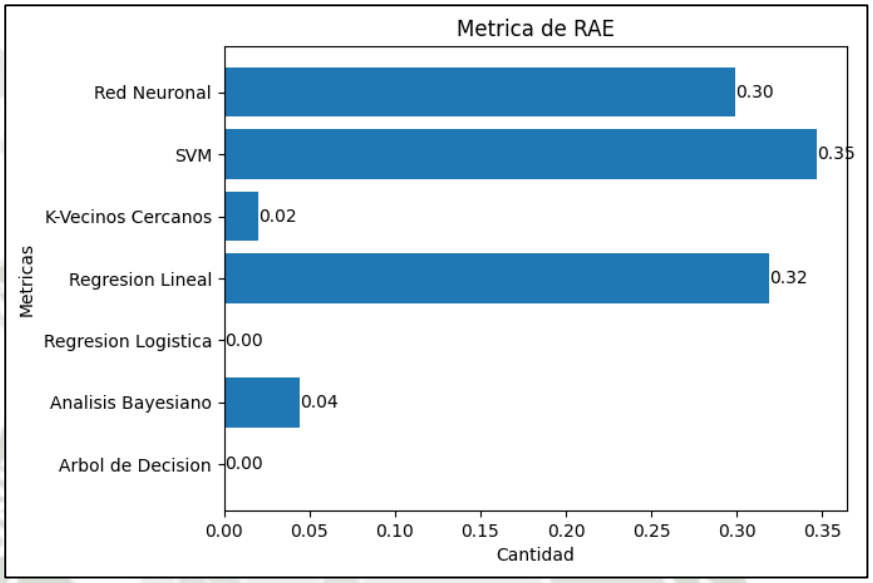
Nota: Propia

g) RAE

La métrica del Error Relativo Absoluto sirve para poder medir un modelo de predicción como es el caso del proyecto de investigación. Los valores más bajos indican que el modelo no tendrá un error al momento de hacer una predicción y en el análisis podemos mencionar que al momento de analizar la última métrica los modelos más adecuados son la regresión logística y el árbol de decisión ya que cumplieron con todas las métricas para tener un buen rendimiento. Sin embargo, también se puede escoger los modelos K-vecinos cercanos y análisis bayesiano porque su margen de error esta entre el 2% y 4% respectivamente, pero se tiene ya modelos ideales lo que no se tomaría en cuenta y por ultimo los modelos de red

neuronal, regresión lineal y SVM superan el 30% y no han tenido un buen rendimiento en las demás métricas por lo que no serán escogidos para realizar la implementación.

Figura. 86.
Análisis de Métrica RAE



Nota: Propia

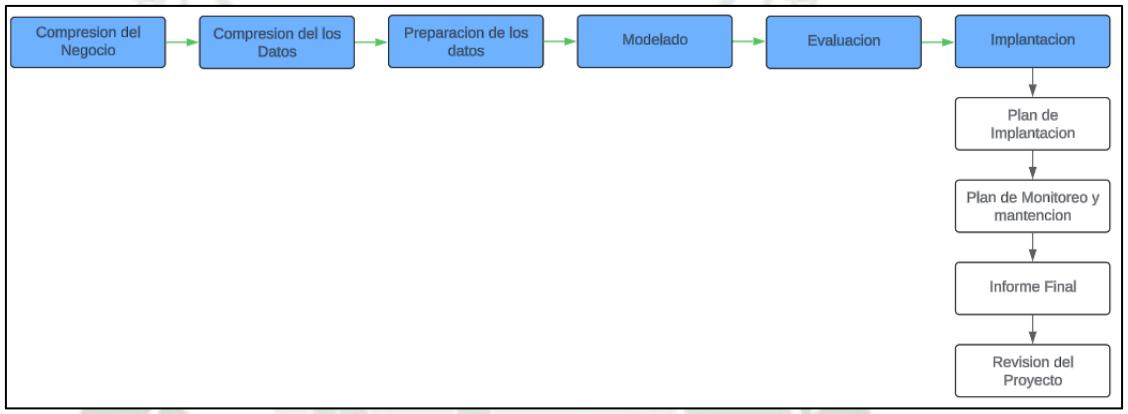
Habiendo ya realizado todo el análisis de los 7 modelos de aprendizaje automático se comprobó que los que tuvieron el mejor rendimiento fue el modelo de regresión logística y el de árbol de decisión por lo que en la implantación se tomara el modelo de árbol de decisión por su buen rendimiento y porque cumplió con todas las métricas siendo el más ideal. Y dentro de sus ventajas esta:

- Fáciles de entender e interpretar por sus decisiones lógicas
- Pueden manejar un conjunto de datos mixto ya sea por el tipo dato
- No tienen una distribución especifica de datos lo hace más versátil
- Pueden ser escalables para pasar de un conjunto pequeño a uno grande
- Manejar naturalmente los datos irrelevantes o ruido ya que pueden ignorarlos

3.6 Implantación

Para esta última etapa de la metodología CRISP-DM se va a establecer un procedimiento para hacer la monitorización y así mantener el modelo escogido, porque el análisis de datos es un proceso que tiene que ser continuo y que se tenga una disponibilidad de datos para lograr el objetivo del negocio.

Figura. 87.
Fase VI - Implantación



Nota: Propia

3.6.1 Plan de Implantación

Para el presente proyecto no se incluye una implantación como tal dentro de ninguna plataforma debido a que se usó una herramienta de software libre, pero se menciona que más adelante en futuras etapas del proyecto se puede colocar dentro de un aplicativo que facilitaría el ingreso y utilidad de los modelos respectivamente.

3.6.2 Plan de Monitoreo

El análisis diseñado en el presente trabajo de investigación, dentro de todos los modelos analizados, se permite poder cargar un dataset con datos actualizados para tener los últimos datos disponibles y poder realizar las mismas predicciones y tener la información de predicción actualizada.

3.6.3 Informe Final

En este apartado se detalla el informe final de todo lo realizado para llegar a realizar cada fase de la metodología:

Tabla 37.
Resumen de Fases y Tareas realizadas durante la metodología CRISP-DM

Fase	Descripción
Fase 1: Comprensión del negocio	• Definición de la situación actual • Establecimiento de los objetivos • Selección de criterios de éxito • Definición de restricciones y supuesto
Fase 2: Comprensión de los datos	• Importación de Nota • Descripción de los datos • Verificación de calidad • Exploración de datos
Fase 3: Preparación de los datos	• Limpieza • Estructuración • Integración
Fase 4: Modelado	• Selección de técnicas de modelado • Generación plan de prueba • Construcción de modelos
Fase 5: Evaluación	• Evaluación de resultados
Fase 6: Implantación	• Definición de plan de monitorización e implementación • Revisión del proyecto

Nota: Propia

En base a los resultados obtenidos por los modelos analizados, se puede observar que todos tienen un aporte distinto al análisis y todos pueden permitir llegar al objetivo final si es que se le da el tiempo adecuado de entrenamiento ya sea por las métricas los modelos pueden ajustarse para obtener los valores esperados.

El tener un dataset con valores de los estudiantes de pregrado de una Universidad Privada nos permitió hacer otros análisis que van relacionados al tema que se quiere tratar como es la diabetes pero que no ayudan al análisis como tal, pero gracias a ellos se descubrió

factores importantes como que no hay un habito de realización de actividad física lo que puede llevar a un mal estilo de vida saludable y tener sedentarismo que a la largo puede ocasionar una diabetes.

Con la validación en la etapa de evaluación del modelo de la metodología CRISP-DM (Ver el apartador 3.4.4) se logró obtener que el algoritmo de aprendizaje automático supervisado que tuvo mayor puntuación en las 8 métricas utilizadas fue el de "Árbol de decisión" el cual se usara para realizar el entrenamiento y predicción de datos reales de estudiantes pregrado entre los 16 hasta los 34 años para validar el proyecto de investigación.

3.6.4 Revisión del Proyecto

La revisión del uso de la metodología en base al proyecto se desarrolló de la siguiente manera:

Tabla 38.
Duración de la Metodología CRISP-DM

		Interaciones										
N°	Fase	1	2	3	4	5	6	7	8	9	10	Duración
1	Comprensión de los requisitos del negocio											3 semanas
2	Comprensión de los datos											3 semanas
3	Preparación de los datos											1 semanas
4	Modelado											4 semanas
5	Evaluación											2 semanas
6	Implantación											1 semanas
TOTAL												14 semanas

Nota: Propia

La descripción de las actividades realizadas son las siguientes:

1. **Comprensión de los requisitos del negocio:** Esta fase inicial se establecio la base para todo el proyecto de investigación ya que aquí se definieron los objetivos del negocio, se identificaron los criterios de éxito y se evalúao la situación actual para alcanzar los objetivos.
2. **Comprensión de los datos:** Se recogieron los datos y se exploraron en detalle para familiarizarse con ellos. También se examinaron características clave, de la calidad y la estructura de estos a través de técnicas de análisis y visualización de datos.
3. **Preparación de los datos:** Los datos son limpiados, son transformados y estructurados en un formato legible para la modelización.
4. **Modelado:** Se seleccionaron 7 modelos de aprendizaje a las cuales se le diviopara el entrenamiento por cada modelo en datos reales y datos de prueba para poder medirlos despues con métricas y saber cual tiene el mejor rendimiento.
5. **Evaluación:** Los modelos son evaluados en base a los resultados de las métricas obtenidas y se escogio uno en particular para poder realizar la prediccion.
6. **Implantación:** El modelo se implanto dentro de un entorno de producción y se monitorean para asegurar que el rendimiento sea el óptimo con cualquier dato que ayude a lograr el objetivo.

Para el desarrollo del proyecto de investigación se tiene el siguiente **Anexo C**

CAPITULO IV

4. Resultados

Para este capítulo vamos a realizar las pruebas correspondientes al modelo escogido para poder determinar los resultados esperados para realizar una predicción de diabetes en base a los atributos de sexo, edad, IMC, glucosa y antecedentes familiares de diabetes. Con las pruebas podremos ver si el modelo es funcional o no.

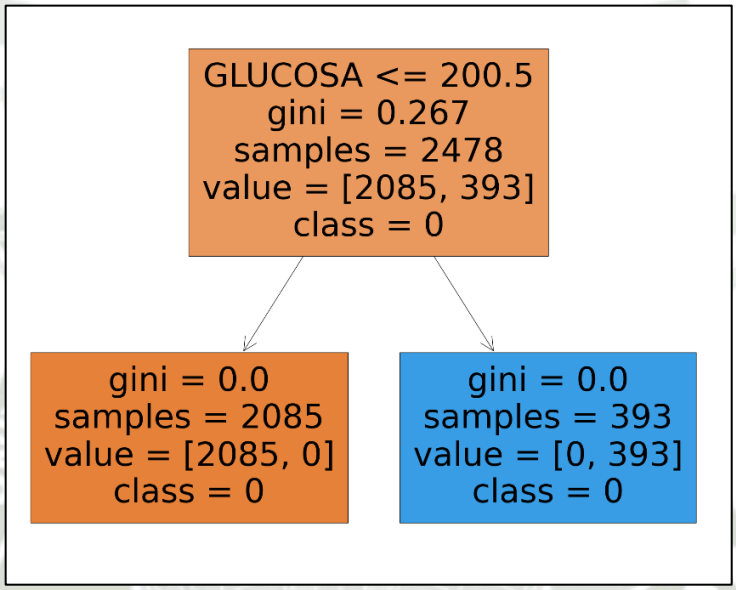
4.1 Resultados del Modelo Implementado

Para poder obtener el resultado del modelo escogido que es arboles de decisión por su buen rendimiento se utilizó el lenguaje Python con la herramienta de Google Colab y usando las librerías de pandas, numpy, seaborn, matplotlib.pyplot, sklearn metrics donde este último ayudo a implementar todos los algoritmos antes mencionados en el apartado de la fase 4 de modelado dentro de la metodología usada. Para poder aplicar el algoritmo de aprendizaje automático supervisado de árboles de decisión se tuvo que primero crear un datsaset son las variables dependientes y las variables resultado que después de tuvo que importar la librería de train_test_split para tener las variables de prueba y fue un 70% variables de prueba contra un 30% de pruebas de datos reales. El Árbol se definió con 4 niveles para el análisis como máximo en casos los requiera, pero para el proyecto solo tomo 2 niveles por el tipo de análisis que realizo.

Durante el entrenamiento se realizaron 4 iteraciones en donde cada uno se evaluado al final con las 7 métricas establecidas en el apartado de metodología del modelo. Los cuales han sido de un valor fundamental para escoger el modelo de aprendizaje ya que sus valores fueron los ideales y en la matriz de confusión realizo un alto número de predicciones a comparación de

los demás modelos mostrados en la evaluación. El accuracy del modelo fue de 1.00 lo que se entiende como una buena predicción y esto se debe a que mediante las comparaciones que hace por sus nodos evita el ruido y solo se enfocada en hallar la variable resultante que es la predicción. Por lo que se obtuvo la siguiente grafica del modelo:

Figura. 88.
Árbol de Decisión Implementado para Predicción



Nota: Google Colab,2024

Para poder medir un modelo de aprendizaje se detalló que se usara una matriz de confusión ya que con ella podremos cuantas predicciones son verdaderas o no el modelo. Para lo cual este modelo dio la siguiente matriz de confusión:

Figura. 89.
Impresión de Matriz de Confusión de Predicción

```
#Prediccion
Y_pred = arbol_diabetes.predict(X_test)
#Calculamos la precision del Modelo por una matriz de confusion
Matriz_de_Confusion = confusion_matrix(Y_test,Y_pred)
Matriz_de_Confusion
```

```
array([[888,   0],
       [  0, 175]])
```

Nota: Google Colab,2024

Podemos observar que en la matriz de confusión el número de acierto o predicciones es alto a comparación de los otros modelos y el error de valores es mínimo porque se puede afirmar que el modelo es ideal para su aplicación y en la fase 4 de modelado se evitó el ruido que se puede ocasionar en el entrenamiento ya que de forma natural el algoritmo lo evita.

4.2 Resultado de Modelo Implementado

En este apartado procederemos a realizar las pruebas correspondientes desde Google Colab para realizar la predicción en base a un dataset de prueba que son datos aleatorios de jóvenes de 17 a 34 años respectivamente simulando un caso de prueba real dentro del entorno del negocio donde se desarrolla este proyecto.

Resultado en Google Colab

Para poder realizar una prueba haremos el guardado del modelo de predicción y posteriormente lo abriremos para poder darle un dataset en el cual tendrá que realizar las predicciones necesarias, por ello tendremos lo siguiente:

Figura. 90.
Guardado del Modelo de Aprendizaje de Árbol de Decisión

```
#Guardado de archivo
pickle5.dump(arbol_diabetes,archivo_modelo)
archivo_modelo.close()
archivo_modelo = open("Modelo Arbol Python.sav", "rb")
#Carga del Modelo
modelo_cargado = pickle5.load(archivo_modelo)
archivo_modelo.close()
print(modelo_cargado)
```

Nota: Google Colab,2024

Luego se enviaron los datos de prueba que están cargados en un Excel para que sea validado correctamente con el modelo guardado y cargado anteriormente, esto permite tener un numero

de registros elevado en un Excel y cargándolo al modelo se puede hacer las predicciones correctas. Por ello tendremos lo siguiente:

Figura. 91.
Carga de Pacientes para Predicción

```
#Pacientes Nuevos
pacientes_nuevos = pd.read_csv("/content/drive/MyDrive/Prediccion_Diabetes.csv")
pacientes_nuevos
```

Nota: Google Colab,2024

El dataset fue convertido previamente y cargado fue la siguiente manera:

Tabla 39.
Dataset cargado para Predicción

Sexo	Edad	IMC	Glucosa	Ant. Familiar Diabetes
1	17	25.38	90	0
2	23	23.98	100	1
1	25	26.30	97	0
1	16	34.14	250	0
2	18	20.10	148	0
1	20	22.50	202	0

Nota: Propia

Por lo que se realizó la predicción y se obtuvo el siguiente resultado:

Figura. 92.
Resultado de Predicción de Árbol de Decisión

```
prediccion_nuevos = modelo_cargado.predict(pacientes_nuevos)
prediccion_nuevos

array([0, 0, 0, 1, 0, 1])
```

Nota: Google Colab,2024

Realizando un formateo en los datos y realizando la interpretación de los resultados obtenidos se puede apreciar lo siguiente:

Figura. 93.
Dataset de Predicción de Pacientes Nuevos con Árbol de Decisión

	SEXO	EDAD	IMC	GLUCOSA	ANT. FAMILIAR DIABETES	Predicción
0	1	19	25.38	90	0	Sin Diabetes
1	2	23	23.98	100	1	Sin Diabetes
2	1	21	26.30	97	0	Sin Diabetes
3	1	25	34.14	250	0	Con Diabetes
4	2	18	20.10	148	0	Sin Diabetes
5	1	27	22.50	202	1	Con Diabetes

Nota: Google Colab,2024

En la Figura N° 93 se puede apreciar el resultado obtenido de aplicar los datos de pacientes nuevos con datos reales que son externos al dataset extraído de la Universidad privada en la ciudad de Arequipa, encontrando que, de 10 personas con los datos extraídos como sexo, edad, el valor de IMC, Nivel de glucosa, antecedentes familiares por diabetes, 2 de ellas tienen diabetes y 8 personas no lo tienen, por lo que la aplicación del algoritmo de árbol de decisión encontró el resultado esperado. Si bien es cierto en toda predicción existe un margen de error porque no es un resultado certero, porque para obtener un diagnóstico más predictivo se usan pruebas estandarizadas por la OMS con lo que respecta a salud. Esta es una predicción en base a los datos extraídos y recopilados por un especialista o encargado que realizo la toma de datos y que al solicitarlo a la Clínica Aliviari se pudo hacer todo este proceso de investigación.

Opinión Medica

En los resultados obtenidos se dio acceso al Dr. Harry Calderón Flores en su condición de Medico General al entorno de predicción que está en la herramienta Google Colab que se encuentra en una versión web donde para acceder solo se necesitan una cuenta de Gmail activa. Dentro del entorno se hizo la ejecución de los resultados obtenidos de la predicción y su opinión es la siguiente: "Usando el Excel y aplicado al entorno de predicción se determina que de los 10 pacientes examinados donde sus edades van desde los 18 años a 27 años, 2 de ellos tienen una posible diabetes por sus altos niveles de glucosa y en especial según los datos del Excel uno de ellos tiene un antecedente cercano familiar para dicha enfermedad que puede ser hereditaria, comprobando que el análisis de predicción es el adecuado pero no es 100% fiable determinarlo por una predicción usando IA ya que se necesitan más exámenes especializados o hacer un seguimiento de los valores constantes de glucosa en los jóvenes de 16 a 34 años para determinar si tiene o no diabetes. Pero si es un avance para detectar la enfermedad que está en una etapa inicial, de igual forma requiere mucha más investigación en el campo de la medicina". Con esta opinión la investigación puede tener una opinión médica en la cual podemos validar que su funcionamiento es el adecuado pero que se necesitan aun mayor investigación y pruebas para determinar al 100% por medio de solo datos en un Excel que una persona efectivamente tiene diabetes. El documento de validación se encuentra en el Anexo F y las pruebas de campo se pueden encontrar en el Anexo G

4.3 Validación de los resultados obtenidos con información según el SIS

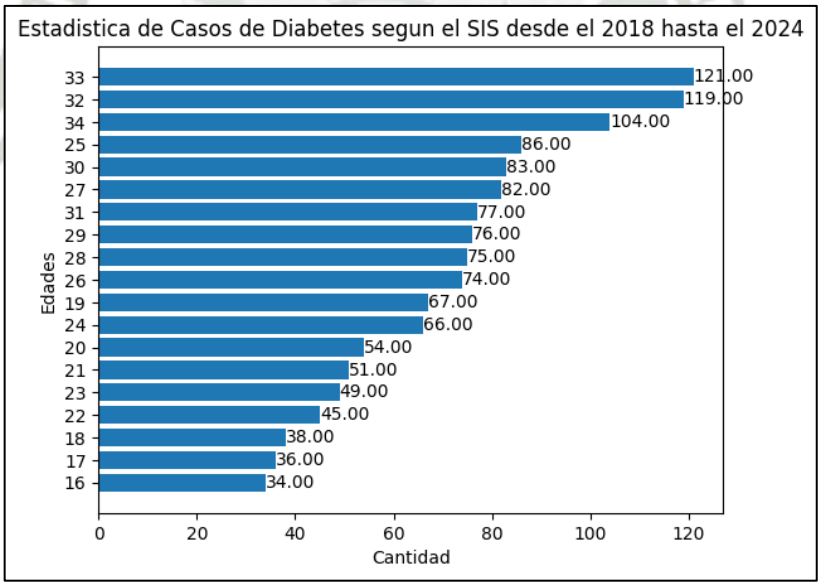
Se obtuvo acceso a el dataset denominado "Afiliados activos en el Seguro Integral de Salud con diagnósticos de Diabetes Mellitus" que brinda la Plataforma Nacional de Datos Abiertos, el cual tiene la información de cada afiliado al SIS (Seguro Integrado de Salud) con diagnóstico de Diabetes Mellitus. La información data de las atenciones desde el 2018 hasta la fecha de corte que es el 31 de abril del 2024 y se encuentra en el Anexo E. (Plataforma Nacional de Datos Abiertos ,2024)

El dataset tiene 10,485.76 registros en la cual se hizo un filtro por provincia y departamento de Arequipa en el rango de 16 a 34 años encontrando que 1337 personas fueron diagnosticadas con diabetes en los últimos 7 años siendo de sexo femenino 894 y de sexo masculino 445 comprobando que el índice de diabetes está más presente en mujeres más que en hombres como el resultado que se extrajo del dataset de la Universidad Privada en la ciudad de Arequipa que mostro más predominancia en el género femenino. No se tiene un registro exacto del avance de la diabetes por año, pero realizando un análisis en los últimos 7 años desde el 2018 hasta el 2024 de los asegurados en el SIS haciendo una división se tendría 191 pacientes por año. En el dataset extraído de la Universidad Privada en la ciudad de Arequipa se encontró 184 personas con una posible diabetes que se aproxima mucho a la deducción de pacientes encontrados por año en el dataset de la Plataforma Nacional de Datos Abiertos.

Arequipa como provincia y ciudad representa el 1.8% de casos de Diabetes en todo el Perú con 19,358 personas diagnosticas según el SIS, encontrando que 1337 son los casos de diabetes en Arequipa entre los 16 y 34 años esto representa del total de casos de 19,358 personas el 6.9% siendo una cifra la cual se próxima el porcentaje encontrado en el dataset extraído de la Universidad Privada en la ciudad de Arequipa siendo 5.2% de casos detectados.

La población más afectada por la diabetes según el SIS entre los 16 hasta los 34 años, son las personas de 25 a 34 años, en el caso de la Universidad Privada en la ciudad de Arequipa las edades con mayor numero va desde 16 hasta los 20 años los cuales tienen un número considerable de casos (Ver Figura 55) , teniendo pocos ejemplares de 25 a 34 años pero que si suman a la estadística general por lo que si se ve un claro aumento en los casos encontrados en la Universidad Privada porque si se demuestra que la diabetes está afectando a la población joven y su detección temprana ayudaría a tener un mejor cuidado de la salud y conciencia con respecto a esta enfermedad. En la Figura N° 94 se puede evidenciar el crecimiento de casos de diabetes con los asegurados en el SIS activos para personas con diagnóstico de un médico especialista desde el 2018 hasta el 2024.

Figura. 94. *Casos de Diabetes en los últimos 7 años en la ciudad de Arequipa*



Nota: Plataforma Nacional de Datos Abiertos, 2024

CONCLUSIONES

1. Se cumplió con el objetivo del proyecto de investigación, se usó los datos biométricos extraídos de la Clínica Aliviari de los estudiantes de pregrado desde los 16 a 34 años para realizar un entrenamiento y predicción en base a los niveles de glucosa. Por lo que usando casos de prueba después del entrenamiento en 10 personas se determinó que 2 de ellas tienen diabetes y 8 no la tienen, para ello se usó el algoritmo de árboles de decisión que paso por la etapa de evaluación de la metodología CRISP-DM y con el cual se determinó el mejor resultado de predicción.
2. Se hizo el análisis y evaluación de los datos que se debían obtener en base a una opinión médica para de los estudiantes de pregrado de la Universidad Privada encontrando que los más importantes fueron el sexo, la edad, el valor de IMC para determinar su masa corporal, el nivel de glucosa y los antecedentes familiares de diabetes, ayudando para ser tomados para la predicción y obtener información sobre los datos relacionados a los estudiantes.
3. Se hizo la solicitud formal de acceso a datos a la Clínica Aliviari indicando que datos se requerían para el proyecto de investigación, el cual fue entregado en un formato Excel y posteriormente siendo colocado en el formato .CSV que sirvió para el análisis y predicción.
4. Se logro analizar estadísticamente las variables extraídas del dataset las cuales fueron 11 el sexo, la edad, el peso, la talla, el IMC, la glucosa, antecedentes familiares de diabetes, consumo de alcohol, consumo de drogas, consumo de tabaco y actividad física generando conocimiento como por ejemplo que del análisis realizado del 100% que representa la muestra de 3542 el 5.2% de estudiantes en el dataset tiene una posible diabetes y que el número de género femenino es de 54.1% y es mayor al género masculino que representa el 45.9% en estudiantes dentro de la Universidad Privada afectando más la diabetes en el género femenino cual se contrasto con la data extraída de las personas activas del SIS diagnosticadas con diabetes siendo esta afirmación verdadera y que también la proporción del filtro realizado mediante el rango de valores de glucosa según la OMS se determinó que 5 de cada 100 alumnos tienen una posible diabetes.
5. Se hizo la transformación de datos en las variables de predicción del dataset extraído para poder usarlo en el algoritmo de aprendizaje automático y esa tarea se realizó en el formateo de datos dentro de la etapa de Preparación de los datos de la metodología CRISP-DM,

colocando a todas las variables como numéricas porque al momento del entrenamiento sea más rápido y es el único tipo de dato aceptado.

6. Los algoritmos usados para la predicción fueron los algoritmos de aprendizaje automático señalados en la tarea de evaluación del modelo en la etapa de modelado de la metodología CRISP-DM los cuales fueron arboles de decisión, regresión lineal, regresión logística, análisis bayesiano, máquina de soporte de vectores, red neuronal y k vecinos más cercanos donde se realizó mediante 8 métricas aplicadas que son la exactitud, precisión, exhaustividad, RMSE, MAE, RHO y RAE encontrando que el algoritmo que mejor rendimiento tuvo para la predicción fue el de "Árbol de Decisión" .
7. Se entreno el algoritmo de aprendizaje automático de "Árbol de Decisión" con los casos de prueba determinando el 60% de casos de prueba y 40% para casos de predicción que se generaron a partir de dataset extraído, determinado con las métricas de evaluación aplicadas para algoritmo de aprendizaje automático que la predicción en base a los datos fue mayor al 90% validando el proyecto de investigación.
8. Se dio acceso a un médico general para poder tomar su opinión en la predicción y utilización del entorno de diagnóstico de una posible diabetes. Encontrando que, de 10 personas tomadas como muestra, 2 de ellas tienen diabetes, pero la opinión media relata que no es 100% fiable enfocarse solo en una predicción porque se necesitan más datos de exámenes especializados para la detección, pero que igual es un avance y esta adecuado.

RECOMENDACIONES Y TRABAJOS FUTUROS

Las recomendaciones son:

1. Una recomendación es poder tener más registros de estudiantes de pregrado ya que para la investigación solo se tomó 3542 registros, esto ayudara a tener más valores de entrenamiento y aumentar el rango de edad poder predecir a un nivel más macro la diabetes y no solo en el alumnado sino también se puede incluir el personal administrativo de la Universidad Privada en la ciudad de Arequipa.
2. Se recomienda obtener los datos mediante un formulario ya que existen muchas herramientas gratuitas como Microsoft Form, Google Form, etc. Porque facilitarían poder tenerlos de una manera más digital y sin errores ya que al inicio del análisis se encontró mucha incongruencia en los datos porque estos se registran en un Excel y no se tiene un formato de datos establecido de ingreso es por ello por lo que se tomó de 3542 registros de los 3600 que se tenía acceso.
3. Se recomienda poder realizar controles de glucosa más seguidos en los alumnos de pregrado de la Universidad Privada para poder alimentar los registros de datos de la herramienta y así tener un control y mejorar la eficiencia del algoritmo usado en la investigación.
4. Se recomienda realizar una validación externa con otro centro de educación superior que tenga asociado datos de estudiantes de pregrado utilizando sus datos biométricos para asegurar que el modelo entrenado cumple su función fuera del entorno específico de la investigación.

Un trabajo futuro es:

1. Mejorar el análisis y la predicción es la integración de Google Colab mediante una API a un aplicativo web o dispositivo de control como un smartwatch para que se permita ingresar los datos en tiempo real a una base de datos y se haga un análisis en base a lo recopilado aplicando el algoritmo de aprendizaje usado en la investigación.

REFERENCIAS

Adite Site, J., Simona Lohan E. (2023). “Machine Learning Based Diabetes Prediction Using Multisensor Data”, in IEEE Sensors Journal, vol. 23, no. 22, pp. 28370-28377, 15 Nov.15, 2023, doi: 10.1109/JSEN.2023.3319360.

Agarwal, P. (2023, August 15). Top 10 machine learning algorithms for beginners. *KDnuggets*.https://www.kdnuggets.com/2023/08/top-10-machine-learning-algorithms-beginners.html

Álvarez Vásquez, C., Coaquira Cuevas, E., Mendoza Hilasaca, E., Pinto Ñaupa, J.” Aplicación de modelo de regresión lineal para predecir el índice de popularidad en la plataforma Spotify”. Revista Innovación y Software. Vol. 4, Nro. 2, pág. 121-135. https://revistas.ulasalle.edu.pe/innosoft

American Diabetes Association Professional Practice Committee.”Glycemic Targets: Standards of Medical Care in Diabetes-2022”. *Diabetes Care*. 2022;45(Suppl 1): S83-S96. PMID: 34964868 pubmed.ncbi.nlm.nih.gov/34964868/.

Anderson, P. R. (2023). *Modelos de representación del conocimiento en inteligencia artificial. Revista de IA y Cognición.* https://www.revistadeiaycognicion.org/modelos-representacion-conocimiento

Arana, C. (2021). “Modelos de aprendizaje automático mediante arboles de decisión”. Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires, Nro. 778.

Arredondo, I. Á. (2018, verano 11). *Día Mundial de la Diabetes: ¿Cuánto gastan los peruanos para tratar esta enfermedad?* Diario Gestión, pp. 1.

Benites, R. y Loja, A. y Ygnacio C. (2022). *Una revisión de las implementaciones de sistemas para la identificación de tendencias de la diabetes* (Vol. 16, Número 2022). Revista de la Carrera de Ingeniería de Sistemas. Universidad de Lima. https://doi.org/10.26439/interfases2022.n016.5957.

Botana., J. (2021). "Bosques Aleatorios". Departamento de psicología y metodología en ciencias del comportamiento. Facultad de psicología. Universidad Complutense de Madrid.

Brown, T. J. (2023). *Metodología CRISP-DM: Un enfoque estructurado para la minería de datos*. *Journal of Data Mining and Analytics*. https://www.journalofdatamining.org/crisp-dm-enfoque-estructurado

Cancino-Gordillo J. y M. Tovar-Vidal, *Clasificación de Diabetes Mellitus tipo II detectando factores de riesgo en un conjunto de datos*, Comput. Sci., 2021, art. n.º 1, pp. 277-286.

Chira Rodriguez, P., Rivera Munive, K. (2023). "Análisis comparativo de técnicas de machine learning sobre el método de muestreo para la predicción de diabetes". Trabajo de Investogacion de Tesis. Universidad Cesar Callejo. Escuela de Ingenieria de Sistemas. Repositorio Institucional–UCV: https://hdl.handle.net/20.500.12692/133747

Chan May, A. y Peña-Koo, J. y Vianne Kinani, J. y Zapata Encalada, M. (2018). *Construcción de un modelo de Predicción para Apoyo al Diagnostico de Diabetes*. (Vol. 40). Pistas Educativas. ttp://itcelaya.edu.mx/ojs/index.php/pistas

Chan, O. y Peña, J. y Vianne, J. y Zapata, M. (2018). *Construcción de un modelo de predicción para apoyo al diagnóstico de diabetes* (Construction of a prediction model to support the diabetes diagnosis). vol. 40, no. 130, pp. 2105–2122.

Collins, R. T. (2023). *Fundamentos de los sistemas expertos: Teoría y práctica*. *Journal of Expert Systems*. https://www.journalofexpertsystems.org/fundamentos-sistemas-expertos

Cuenca-Andrade, G. y Guamán-Quinche, R. y Álvarez-Pineda, F. y Chamba-Zaragocin, W. y Guamán-Quinche, J. (2023). *Estudio de algoritmos de inteligencia artificial más utilizados para el diagnóstico de diabetes mellitus tipo 2*. (Vol. 13, Número 1). CEDAMAZ. https://doi.org/10.54753/cedamaz.v13i1.1804.

Davis, M. R. (2023). Fundamentos de algoritmos genéticos. Journal of Computational Intelligence.https://www.journalofcomputationalintelligence.org/fundamentos-algoritmos-geneticos

De León, M. y Torres Soto, A. y Soto, P. (2020). Mecanismo de clasificación para diabetes mellitus en la población de Aguascalientes, México. Universidad de Aguas Calientes.

Farruhk Aslam, K., Mabrook A., Seyd Ahmad C., (2021). "Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review". in IEEE Access, vol. 9, pp. 43711-43735, 2021, doi: 10.1109/ACCESS.2021.3059343.

Gupta, S. K., Khatri, A. J., & Sharma, V. K. (2021). A comparative study of data mining methodologies for predictive analytics. *Journal of Data Science and Analytics*, 15(3), 245-260. https://doi.org/10.1016/j.jdsa.2021.03.012

Hernández, D. (2007). "Introducción al análisis bayesiano". Instituto de Investigación y Desarrollo Pesquero (INIDEP). Vol. 1, Mar de la Plata, Argentina. http://hdl.handle.net/1834/2560

Herrería, J. y Vladimir-Mosquera, G. y Sánchez, C. (2023). *Prediagnóstico Médico de la Diabetes Mellitus tipo 2 mediante Machine Learning*. (Vol. 3, Número 2). Revista Innovación Digital y Desarrollo Sostenible –IDS, ISSN 2711-3760 (En línea).

Instituto Nacional de Estadística e Informática (INEI) (06 de abril de 2016). Nota de Prensa para el Dia Mundial de la Salud N° 064 para la Diabetes. url: nota-de-prensa-n064-2016-inei.pdf

Jones, L. M. (2024, abril 15). *Resolución de problemas sin métodos directos: Enfoques basados en comportamientos*. Blog de Inteligencia Artificial Avanzada.

Jones, A. M. (2023). *La metodología SEMMA en minería de datos: Un enfoque estructurado. Journal of Data Mining Practices.* https://www.journalofdataminingpractices.org/semma-enfoque-estructurado

Kamrul H., Ashraful D., Eklas H., (2020). "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in *IEEE Access*, vol. 8, pp. 76516-76531, 2020. doi: 10.1109/ACCESS.2020.2989857.

Leal, A. M. (2021). *Diagnóstico de la diabetes mediante el uso de técnicas de aprendizaje automático*, Universidad Politécnica de Valencia. https://riunet.upv.es/handle/10251/174574

Martínez-Cámara, J. J. (2022). *Desarrollo de un Sistema Inteligente de Control de Diabetes de Tipo I: Basado en Modelos Predictivos*. Universidad de Jaén, Escuela Politécnica Superior de Linares.

Martínez, J., (2022). "Técnicas de Inteligencia Artificial en Medicina y Salud". España. Universidad de las Illes Balears. Trabajo de Fin de Grado. Facultad de Ciencias.

Martínez Pérez, J., Ferras Fernández, Y., Bermúdez Cordovi, L. Ortiz Cabrera, Y., Pérez Leyva, E. (2020)." Regresión Logistica y predicción del bajo rendimiento académico de estudiantes de la carrera de medicina". Revista Electrónica Dr. Zollo E. Marinello Vidaurreta, pp. 45(4). http://revzoilomarinello.sld.cu/index.php/zmv/article/view/2230.

Moine, J. M., Haedo, A. S., & Gordillo, S. E. (2011). Estudio comparativo de metodologías para minería de datos. In *XIII Workshop de Investigadores en Ciencias de la Computación*.

Organización Mundial de la Salud. (2023). *Global report on diabetes*. https://www.who.int/publications/i/item/9789240062706

Peterson, H. R. (2023). *Métricas de evaluación de modelos de clasificación: Precisión, recall y F1-score*. *Journal of Machine Learning Metrics*.

Pincay-Ponce, J. y Sánchez-Andrade, D. y Caicedo-Ávila, I. y Macías-Valencia, D. (2020, noviembre). *Clasificación de pacientes según su posibilidad de adquirir Diabetes Mellitus empleando algoritmos de Machine Learning*. https://www.researchgate.net/publication/346445393.

Plataforma Nacional de Datos Abiertos (2024). "Afiliados activos en el Seguro Integral de Salud con diagnóstico de Diabetes Mellitus". Url: Afiliados activos en el Seguro Integral de Salud con diagnóstico de Diabetes Mellitus - [SIS] | Plataforma Nacional de Datos Abiertos

Rajaraman, S., Rajasekaran, K. K. S., & Ravi, V. R. K. (2017). Predictive modeling of diabetes mellitus using machine learning algorithms: A comprehensive study. Journal of Healthcare Engineering. https://doi.org/10.1155/2017/8542376

Rodríguez Montequín M.T. (2002). "Técnicas de análisis de datos. Departamento de Matemáticas". Universidad de Oviedo.

Rodríguez, I. y Campo Valera M. y Rodríguez J. (2023). *El Internet de las Cosas Medicas (IOMT): Una revolución tecnológica aplicable a la gestión de la diabetes mellitus tipo I*. UmaEditorial.

Sarmiento Gómez, H., Barrios Marengo, D., Herrera Acosta, R., Palomino Pacheco, K. (2021) "Comparación de técnicas de aprendizaje automático para la clasificación de pacientes con trastorno mentales y de comportamiento al consumo psicotrópico en la ciudad de barranquilla", Mundo Fesc, vol. 11, pp. 59-69.

Santana, F. (2023)." Análisis de Datos abiertos de Futbol aplicando la metodología CRISP-DM". Trabajo de Fin de Máster. Departamento de Ingeniería de Sistemas Telemáticos. Universidad Politécnica de Madrid. TFM_FernandoSantanaFalcon.pdf (upm.es)

Thompson, R. J. (2023, febrero 25). Uso de bibliotecas de Python para análisis de datos: Pandas, NumPy y Matplotlib. *Tech Trends Today*. https://www.techtrendstoday.com/bibliotecas-python-analisis-datos

Tovar-Vidal, J. M. C.-G. (2021). *Clasificación de Diabetes Mellitus tipo II detectando factores de riesgo en un conjunto de datos*. Benemérita Universidad Autónoma de Puebla.

Usama Ahmed, G., et al. (2022). "Prediction of Diabetes Empowered with Fused Machine Learning," in *IEEE Access*, vol. 10, pp. 8529-8538, 2022, doi: 10.1109/ACCESS.2022.3142097.

Vilema Narváez, M., Moyano Arias, R., Palacios Campana, D., Izurieta Guamán, G. (2022). ” Predicción de clientes potenciales utilizando K vecinos más cercanos en el área de negocios de la Cooperativa Riobamba”. Revista Perspectivas. Vol. 4, N° 1.

Villegas-Cubas, J. y Capuñay-Uceda, O. y Coronado-Navarro, A. y Delgado-Chavarri, A. y Hans Osores-Granda, O. (Ed.). (2022). *Sistemas de información para la Red Neuronal Convolucional en la detección de diabetes usando imágenes de fondo de ojo*. (Vol. E51, Número 2022). Revista Ibérica de Sistemas y Tecnologías de la Información.

Zhang, L. M. (2023). *Métricas de evaluación en aprendizaje automático: Precisión, recall y más*. *Journal of Machine Learning Research*. https://www.jmlr.org/metricas-evaluacion-aprendizaje

ANEXOS

ANEXO A: DOCUMENTO DE ACCESO A DATOS

CARTA DE SOLICITUD DE USO DE INFORMACIÓN DE EMPRESA

Clínica Aliviari

Yo Mg. Angel Montesinos Murillo, en mi calidad de Director de la Escuela Profesional de Ingeniería de Sistemas de la Facultad de Ciencias e Ingenierías Físicas y Formales de la institución de estudios superiores Universidad Catolica De Santa María con R.U.C N° **20141637941** ubicada en la ciudad de Arequipa.

Hago presentación del Bachiller Alvaro Daniel Berrios Zuniga identificado con DNI N° **73938490** egresado de la Escuela Profesional de Ingeniería de Sistemas en la Facultad de Ciencias e Ingenierías Físicas y Formales con código de egresado **2017220641** el cual requiere información para el desarrollo de su proyecto de investigación Tesis denominada; "**Predicción de la Diabetes mediante aprendizaje de maquina en el personal administrativo de la Universidad Catolica de Santa Maria con el desarrollo de un aplicativo para su detección**".

Por lo tanto, realizo la:
SOLICITUD DE ACCESO A INFORMACION CON FINES DE INVESTIGACIÓN,

Al señor Patricio Azalgara Lazo, representante de la empresa o institución Clínica Aliviari perteneciente a la Universidad Catolica de Santa Maria con R.U.C N° **20600633369** para que utilice la siguiente información de la empresa:

Datos médicos relacionados a la Diabetes como sexo, edad, talla, peso, niveles de glucosa, presión arterial, niveles de colesterol si se tuviera la información sobre si es fumador o no, consumo de alcohol o drogas, si realiza actividad física, problemas cardiacos con antecedentes y datos relacionados a la alimentación. Los datos serán recopilados sin el distintivo de nombre e identificación como DNI o cargo, solo se requiere datos reales para la investigación ya que se basa en una predicción y en caso no se pudiera extraer del reporte esos datos el solicitante se compromete a quitarlos para poder realizar la investigación.

La finalidad es que se pueda desarrollar su ( ) Trabajo de Investigación, **(X)Tesis** para optar al grado de ( )Bachiller, ( )Maestro, ( )Doctor o **(X)Título Profesional**.

Si se autoriza la información de la empresa, se mantendrá el nombre o cualquier distintivo de la empresa en relación con los usuarios en reserva. Por lo que el solicitante se compromete a:

- Mantener en Reserva el nombre o cualquier distintivo en relación con sus usuarios.
- Usar los datos netamente para uso de investigación y análisis.

Firma del representante de la escuela profesional de Ing. De Sistemas

DNI: 29413196

El Egresado/Bachiller declara que los datos emitidos en esta carta y en el Trabajo de Investigación, en la Tesis no serán divulgados. En caso de comprobarse dicho acto, el Egresado será sometido al inicio del procedimiento disciplinario correspondiente; asimismo, asumirá toda la responsabilidad ante posibles acciones legales que la empresa o institución, otorgante de información, pueda ejecutar.

Firma del Egresado

DNI: 73938490

Arequipa, 15 de enero de 2024

ANEXO B: GLOSARIO DE TERMINOS

1. **Algoritmo:** Un algoritmo es un conjunto finito de instrucciones o reglas bien definidas y ordenadas que permiten llevar a cabo una tarea o resolver un problema en un número finito de pasos.
2. **Inteligencia Artificial:** La inteligencia artificial (IA) es un campo de la informática que se enfoca en crear sistemas o máquinas capaces de realizar tareas que normalmente requieren inteligencia humana.
3. **Aprendizaje Automático:** Es un subcampo de la inteligencia artificial que se centra en el desarrollo de algoritmos y técnicas que permiten a las computadoras aprender patrones y realizar tareas específicas sin necesidad de ser programadas explícitamente para cada tarea.
4. **Aprendizaje Supervisado:** El aprendizaje supervisado es una técnica dentro del campo del aprendizaje automático donde se proporciona al algoritmo un conjunto de datos de entrenamiento que incluye ejemplos de entrada y la salida esperada correspondiente.
5. **Aprendizaje No Supervisado:** El aprendizaje no supervisado es otra técnica dentro del campo del aprendizaje automático en la que el algoritmo se entrena utilizando conjuntos de datos que no están etiquetados ni categorizados previamente. El algoritmo debe encontrar patrones o estructuras interesantes en los datos por sí mismo.
6. **Métricas de Aprendizaje Automático:** Las métricas de aprendizaje automático son medidas utilizadas para evaluar el rendimiento de un modelo de machine learning en un conjunto de datos dado. Estas métricas permiten entender cuán bien está funcionando el modelo y cómo se compara con otros modelos o con los objetivos específicos del problema que se está abordando.

7. **Análisis Bayesiano:** El análisis bayesiano es un enfoque estadístico que se basa en el teorema de Bayes para actualizar la probabilidad de una hipótesis a medida que se obtiene nueva evidencia.
8. **Regresión Lineal:** La regresión lineal es un método estadístico utilizado para modelar la relación entre una variable independiente (o predictora) y una variable dependiente (o de respuesta) en la que se asume que la relación es lineal.
9. **Regresión Logística:** La regresión logística es un método estadístico utilizado para modelar la probabilidad de que una variable categórica binaria dependiente esté presente como una función de una o más variables independientes.
10. **Arboles de Decisión:** Un árbol de decisión es una técnica de modelado predictivo que se utiliza en el aprendizaje automático y en la minería de datos. Se utiliza para representar y modelar decisiones y su posible consecuencia.
11. **KNN Vecinos Cercanos:** Es un método de aprendizaje supervisado utilizado para clasificación y regresión. En esencia, k-NN clasifica un punto de datos según la mayoría de los votos de sus k vecinos más cercanos en el espacio de características
12. **Red Neuronal:** Es un modelo computacional inspirado en la estructura y funcionamiento del cerebro humano. Consiste en un conjunto de unidades básicas de procesamiento, llamadas neuronas artificiales o nodos, organizadas en capas interconectadas.
13. **Atributo:** Un atributo es una característica o propiedad específica que describe una entidad o un objeto. En otras palabras, un atributo es una dimensión o una variable que proporciona información sobre las características de un objeto en particular.

14. **Instancia**: Una instancia se refiere a un ejemplo individual o una observación en un conjunto de datos. También se conoce como un punto de datos, una observación o una muestra.
15. **Accuracy:** Es una métrica comúnmente utilizada para evaluar el rendimiento de un modelo de clasificación en aprendizaje supervisado. Representa la proporción de predicciones correctas realizadas por el modelo en relación con el total de predicciones realizadas.
16. **Dataset:** Es simplemente una colección de datos organizados en una estructura que facilita su uso y análisis. En el ámbito del análisis de datos, el aprendizaje automático y la minería de datos, un dataset se refiere a un conjunto de datos que comparten alguna característica común o están relacionados de alguna manera.
17. **Python:** Python es un lenguaje de programación de alto nivel, interpretado y de propósito general. Fue creado a principios de los años 90 por Guido van Rossum y ha ganado una gran popularidad en la comunidad de programadores debido a su sintaxis clara y legible, así como a su amplia gama de aplicaciones y bibliotecas disponibles.
18. **Google Colab:** Es una plataforma en la nube desarrollada por Google que proporciona un entorno de ejecución de Jupyter Notebook de forma gratuita. Jupyter Notebook es una aplicación web de código abierto que permite crear y compartir documentos interactivos que contienen código en vivo, visualizaciones y texto explicativo.
19. **MLLPRegressor:** Es una clase proporcionada por la biblioteca scikit-learn en Python. Representa un regresor de redes neuronales artificiales (también conocidas como perceptrones multicapa) en el contexto de aprendizaje automático.

20. **Predicción:** En el contexto del aprendizaje automático y la estadística, una predicción se refiere específicamente a la estimación de un valor desconocido de una variable objetivo (también llamada variable dependiente) utilizando un modelo entrenado.
21. **Modelo de Aprendizaje Automático:** Un modelo de aprendizaje automático es una representación matemática o computacional de la relación entre variables en un conjunto de datos, que se utiliza para hacer predicciones o tomar decisiones basadas en nuevos datos no vistos.
22. **Librería de Inteligencia Artificial:** Es un conjunto de herramientas, algoritmos y funciones diseñadas para facilitar el desarrollo, implementación y experimentación en el campo de la inteligencia artificial (IA). Estas librerías proporcionan una amplia gama de funcionalidades para resolver problemas en áreas como el aprendizaje automático, el procesamiento del lenguaje natural, la visión por computadora, entre otros.

ANEXO C: CRONOGRAMA DE INVESTIGACION

N°	Nombre	Duración	Fecha Inicio	Fecha Fin	Predecesores
1	**Capítulo 1: Planteamiento Teórico**	**30 días**	**17/08/2023**	**17/09/2023**	
2	* 1.1 Introducción	2 días	19/08/2023	20/08/2023	
3	1.1.1 Antecedentes	2 días	21/08/2023	22/08/2023	3
4	1.1.2 Objetivos	2 días	23/08/2023	24/08/2023	3
5	1.1.3. Enfoque	2 días	25/08/2023	26/08/2023	3
6	1.1.4. Alcances y Limitaciones	2 días	27/08/2023	28/08/2023	6
7	1.1.5. Aporte	2 días	29/08/2023	30/08/2023	6
8	1.1.6. Preguntas de Investigación	2 días	31/08/2023	01/09/2023	6
9	1.1.7. Línea, Sublinea, tipo y nivel de investigación	2 días	02/09/2023	03/09/2023	9
10	1.1.8. Cobertura del Estudio	2 días	04/09/2023	05/09/2023	9
11	1.1.9. Métodos, Técnicas e Instrumentos	2 días	06/09/2023	07/09/2023	9
12	1.1.10. Solución Propuesta	2 días	08/09/2023	09/09/2023	12
13	1.1.11. Metodologías, Modelos, Lenguajes	2 días	10/09/2023	11/09/2023	12
14	* 1.2 Fundamentos Teóricos	2 días	12/09/2023	13/09/2023	13
15	1.2.1 Estado del Arte	2 días	14/09/2023	15/09/2023	13
16	* 1.3. Organización de la Tesis	2 días	16/09/2023	17/09/2023	14
17	**Capítulo 2: Marco Teórico**	**26 días**	**18/09/2023**	**13/10/2023**	
18	* 2.1 Definiciones, Acrónimos y Abreviaturas	4 dias	19/09/2023	22/09/2023	17
19	2.1.1 La Diabetes	3 días	23/09/2023	25/09/2023	20
20	2.1.2 Inteligencia Artificial	3 días	26/09/2023	28/09/2023	21
21	2.1.3 Machine Learning	3 días	29/09/2023	01/10/2023	21
22	2.1.4 Redes Neuronales	3 días	02/10/2023	04/10/2023	22
23	2.1.5 Metodología CRISP-DM	5 días	05/10/2023	09/10/2023	23
24	2.1.6 Google Colab	2 días	10/10/2023	11/10/2023	24
25	2.1.7 Python	2 días	12/10/2023	13/10/2023	25
26	**Capítulo 3: Análisis, Construcción y Evaluación**	**195 días**	**14/10/2023**	**10/05/2024**	
27	* 3.1 Compresión de los Requisitos	25 días	15/10/2023	16/11/2023	25
28	3.1.1 Determinar los objetivos	9 días	15/10/2023	24/10/2023	26
29	3.1.2 Evaluación de la situación	8 días	25/10/2023	02/11/2023	27
30	3.1.3 Realización del Plan de Proyecto	8 días	03/11/2023	11/11/2023	28 y 26
31	* 3.2 Comprensión de los datos	**22 días**	12/11/2023	04/12/2023	30
32	3.2.1 Recolección y adaptación de los datos	6 días	12/11/2023	18/11/2023	
33	3.2.2 Descripción formal de los datos	8 días	19/11/2023	27/11/2023	31
34	3.2.3 Exploración de los datos	8 días	28/11/2023	06/12/2023	30
35	* 3.3 Preparación de los datos	**30 días**	07/12/2023	10/01/2024	32
36	3.3.1 Selección de datos	6 días	07/12/2023	13/12/2023	33
37	3.3.2 Limpieza de Datos	6 días	14/12/2023	20/12/2023	34

38	3.3.3 Construcción de datos	6 días	21/12/2023	27/12/2023	35
39	3.3.4 Integración de datos	6 días	28/12/2023	03/01/2024	36
40	3.3.5 Formateado de datos	6 días	04/01/2024	10/01/2024	37
41	* 3.4 Búsqueda y Modelado	**60 días**	11/01/2024	11/03/2024	
42	3.4.1 Selección de la técnica de modelado	10 días	11/01/2024	21/01/2024	39
43	3.4.2 Diseño de la prueba	10 días	22/01/2024	01/02/2024	40
44	3.4.3 Construcción del modelo	25 días	02/02/2024	27/02/2024	41
45	3.4.5 Evaluación del Modelo	15 días	28/02/2024	14/03/2024	42
46	* 3.5 Evaluación	**20 días**	15/03/2024	04/04/2024	
47	3.5.1 Evaluación de los resultados	10 días	15/03/2024	25/03/2024	44
48	3.5.2 Revisión del proceso	5 días	26/03/2024	31/03/2024	45
49	3.5.3 Determinación de los próximos pasos	5 días	01/04/2024	06/04/2024	46
50	* 3.6 Implementación	**30 días**	07/04/2024	07/05/2024	
51	3.6.1 Planeamiento de implementación de modelo	5 días	07/04/2024	12/04/2024	47
52	3.6.2 Planteamiento de monitorización de modelo	5 días	13/04/2024	18/04/2024	48
53	3.6.3 Desarrollo de producto final	15 días	19/04/2024	04/05/2024	49
54	3.6.4 Revisar el proyecto	5 días	05/04/2024	10/05/2024	50
55	**Capítulo 4: Resultados**	**20 días**	11/05/2024	31/05/2024	
56	* 4.1 Verificación del avance del uso de la Metodología CRISP-DM	5 días	11/05/2024	16/05/2024	51
57	* 4.2 Resultados del Modelo Implementado	5 días	17/05/2024	22/05/2024	52
58	4.2.1 Pruebas del Modelo Implementado	5 días	23/05/2024	28/05/2024	53
59	4.2.3 Análisis y discusión del modelo implementado	5 días	29/05/2024	03/06/2024	54
60	Conclusiones	5 días	04/06/2024	09/06/2024	55
61	Recomendaciones y Trabajos Futuros	5 días	10/06/2024	15/06/2024	56
62	Referencias	221 días	18/09/2023	26/04/2024	57
63	**Apéndice**	170 días	21/08/2023	07/02/2024	
64	Apéndice A: Glosario de Terminologías	170 días	21/08/2023	07/02/2024	
65	Apéndice B: Resultados de entrenamiento	15 días	19/07/2024	03/08/2024	

ANEXO D: MUESTRA ORIGINAL DE DATOS

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
	FEC. NACIMIEN	EDAD	SEXO	PA	FC	PESO	TALLA	CIR. CARPO	IMC	HEPATITIS	OSIS HEPATIT	DIFTERIA	TÉTANOS	COVID-19	OSIS COVID-1	EMOGLOBIN	IPO SANGUÍ	FACTOR RH	V CUALITATIV	OMENDACIÓ	OMENDACIÓ	OMENDACIÓ
23-12-23	18-05-06	17	F	90/60	74	62	1,59	17	24,52	SI		SI	SI	SI	3	14,2	O	POSITIVO	NO REACTIVO			
26-12-23	22-03-07	16	F	120/80	70	48,5	1,57	15	19,68	SI		SI	SI	SI	3ra	13,8	O	POSITIVO	NO REACTIVO			
22-12-23	27-11-01	21	M	110/70	89	88,7	1,8	18	27,38	SI		SI	SI	SI	4TA	16,2	O	POSITIVO	NO REACTIVO			
30-12-23	19-02-07	16	M	110/70	80	50,5	1,74	16	16,68	SI		SI	SI	SI	3RA	18	A	POSITIVO	NO REACTIVO			
26-12-23	24-02-07	16	F	90/60	70	58	1,57	17	23,53	NO		NO	NO	NO	2	16	O	POSITIVO	NO REACTIVO			
27-12-23	22-07-06	17	M	100/60	78	80	1,72	18	27,04	SI		SI	SI	SI	2	16	O	POSITIVO	NO REACTIVO			
22-12-23	20-05-07	16	F	120/80	78	75	1,61	16	28,93	NO		NO	NO	SI	3RA	13,6	A	POSITIVO	NO REACTIVO			
29-12-23	14-09-06	17	M	120/80	76	55	1,74	16	18,17	NO		NO	NO	SI	3	18	O	POSITIVO	NO REACTIVO			
27-12-23	10-09-06	17	F	100/60	60	57	1,67	18	20,44	SI		SI	SI	SI		15,1	A	POSITIVO	NO REACTIVO			
22-12-23	23-03-07	16	F	100/60	78	73	1,57	19	29,62	SI		SI	SI	SI	2DA	14,2	O	POSITIVO	NO REACTIVO			
29-12-23	11-12-04	18	M	120/70	80	63	1,8	16	19,44	NO		NO	NO	SI	3	17,4	A	POSITIVO	NO REACTIVO			
26-12-23	16-04-07	16	M	90/60	80	83	1,77	18	26,49	SI		SI	SI	SI	3	17,4	A	POSITIVO	NO REACTIVO			
27-12-23	12-08-06	17	F	90/60	69	68	1,63	16	25,59	si		si	si	si	3	13,6	O	POSITIVO	NO REACTIVO			
28-12-23	24-08-06	17	M	100/70	60	84	1,82	17	25,40	SI		SI	SI	SI	2DA	16,5	O	POSITIVO	NO REACTIVO			
22-12-23	27-12-06	16	M	90/60	74	60	1,71	17	20,52	NO		NO	NO	SI	3RA	16,2	A	POSITIVO	NO REACTIVO			
27-12-23	30-01-07	16	M	100/60	65	58,5	1,56	17	24,04	si		si	si	si		15,4	A	NEGATIVO	NO REACTIV	REALIZAR VARIANTE DE DU		
22-12-23	03-07-06	17	F	90/60	80	43	1,6	16	16,80	SI		SI	SI	SI	3	14,2	O	POSITIVO	NO REACTIVO			
26-12-23	10-10-06	17	F	110/70	77	52	1,58	15	20,83	SI		SI	SI	SI	2	15,7	O	POSITIVO	NO REACTIVO			
28-12-23	14-01-07	16	F	90/60	67	50	1,63	14	18,82	SI		SI	SI	SI	3RA	14,6	O	POSITIVO	NO REACTIVO			
23-12-23	26-07-06	17	M	100/60	81	58,5	1,69	18	20,48	SI		SI	SI	SI	3	17,1	B	POSITIVO	NO REACTIVO			
28-12-23	14-08-06	17	M	100/70	67	69	1,71	17	23,60	SI		SI	SI	SI	2DA	16,2	O	POSITIVO	NO REACTIVO			
29-12-23	06-11-06	17	M	120/75	72	62	1,7	17	21,45							17,4	A	POSITIVO	NO REACTIVO			
23-12-23	05-08-98	25	F	100/60	78	58	1,54	17	24,46	SI		SI	SI	SI	3	15,1	O	POSITIVO	NO REACTIVO			
29-12-23	18-06-04	19	M	120/85	82	75,5	1,82	17	22,79	NO		NO	NO	SI	3	15	O	POSITIVO	NO REACTIVO			
22-12-23	27-04-00	23	F	90/60	84	53,5	1,59	16	21,16	SI		SI	SI	SI	3RA	15,1	O	POSITIVO	NO REACTIVO			
22-12-23	17-09-05	18	F	90/60	66	54,2	1,57	16	21,99	NO		NO	NO	NO		14,6	O	POSITIVO	NO REACTIVO			
28-12-23	18-09-06	17	M	100/60	75	69	1,72	17	23,30	si		si	si	si	2da	16,5	B	POSITIVO	NO REACTIVO			
23-12-23	28-04-07	16	M	100/60	74	55	1,69	16	19,26	SI		SI	SI	XSI		17,1	A	POSITIVO	NO REACTIVO			
29-12-23	12-09-06	17	M	110/70	67	69,1	1,62	18,8	26,33	SI		SI	SI	SI	3RA	16	O	POSITIVO	NO REACTIVO			
26-12-23	17-02-07	16	F	90/60	65	55,5	1,62	15,5	21,15	SI		SI	SI	SI		15,4	O	POSITIVO	NO REACTIVO			
22-12-23	13-11-06	16	M	100/60	78	82	1,79	17	25,59	NO		NO	NO	SI		15,4	O	POSITIVO	NO REACTIVO			
22-12-23	26-07-05	18	M	120/70	85	67	1,69	18	23,46	SI		SI	SI	SI	3RA	16,5	O	POSITIVO	NO REACTIVO			
22-12-23	27-01-07	16	M	100/60	84	67	1,62	18	25,53	SI		SI	SI	SI	3	18	O	POSITIVO	NO REACTIVO			
29-12-23	25-06-04	19	M	110/70	62	85,3	1,76	18	27,54	NO		NO	NO	SI	4TA	16,2	O	POSITIVO	NO REACTIVO			
28-12-23	15-10-06	17	F	110/70	93	48,5	1,59	15	19,18	NO		NO	NO	SI	3RA	15,4	O	POSITIVO	NO REACTIVO			

Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV
29	30	31	32	33	34	35	36	37	38	39	40	#N/A	42	43	44	45	46	47	48	49	50	51
RAYOS X	NT. FAMILIAR	T. PERSONAL	CIRUGÍAS	ALCOHOL	TABACO	DROGAS	ACT. FÍSICA	ALÉRGIAS	NICACIÓN ACT	APACIDAD (C	APACIDAD2 (C	ANAMNESIS	ÁMEN CLÍNI	NÓSTICO1 (C	NÓSTICO2 (C	NÓSTICO3 (C	NÓSTICO4 (C	NÓSTICO4 (C	OMENDACIÓ	OMENDACIÓ	OMENDACIÓ	OMENDACIÓ	OME
RADIOGRAFÍ	MADRE HIPE	ASMA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	RINOPLASTIA	NIEGA	NIEGA	NIEGA	BASQUET 2 V	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NINGUNA DE	NIEGA	NIEGA	NIEGA	BASQUET 1 V	IBUPROFENC	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	BAJO PESO					CONTROL POR NUTRICION, DIETA HIPERCALÓRICA E HIPE				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	BASKET 3 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	DEPORTES	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	MIGRAÑA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NAPROXENC	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	ASMA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	DANZA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	VOLEY 2 VEC	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	PADRE FALLE	NIEGA	FIMOSIS EN I	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	ASMA	NIEGA	NIEGA	NIEGA	NIEGA	GIMNASIO 6	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	PADRE DIABI	NIEGA	RINOPLASTIA	NIEGA	NIEGA	NIEGA	BICICLETA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 6 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 2 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	CRPTORQUIE	NIEGA	NIEGA	NIEGA	FUTBOL	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	BAJO PESO					CONTROL POR NUTRICIÓN, DIETA HIPERCALÓRICA E HIPE				
RADIOGRAFÍ	NIEGA	ITU RECURRE	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	SOP - HIPERI	CIRUGIA DE I	NIEGA	NIEGA	NIEGA	NIEGA	ACEITUNA - I	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 3 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 3 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	MADRE CON	ASMA EN LA	NIEGA	NIEGA	NIEGA	NIEGA	CAMINATA 6	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	CAMINATA 1	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	HIGADO GRA	NIEGA	NIEGA	NIEGA	NIEGA	CAMINATA 3	NIEGA	HIGANATUR	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	RINOPLASTIA	NIEGA	NIEGA	NIEGA	VOLEY 1 VEZ,	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 3 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	TENNIS 5 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	PADRE HIPEF	NIEGA	PO DE FX DE	NIEGA	NIEGA	NIEGA	VOLEY 5 VEC	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	DERMATITIS	NIEGA	NIEGA	NIEGA	NIEGA	BASQUET DIA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL DIAF	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	GIMNASIO D	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	FUTBOL 6 VE	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	SOBREPESO					CONTROL PC	REALIZAR ACTIVIDAD FÍSICA Y CONTROL DE F			
RADIOGRAFÍ	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	NIEGA	#N/A	#N/A	PACIENTE AC	EXAMEN FISI	CLINICAMENTE	SANO				MANTENER PESO SALUDABLE, HACER ACTIVIDAD FÍSICA Y				

ANEXO E: CASOS DE DIABETES SEGÚN SIS

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
FECHA_CORT	FECHA_AFILI	CODIGO_AN	EDAD	UBIGEO	DEPARTAME	PROVINCIA	DISTRITO	SEXO	FECHA_PRIM	TIPO_DIABET	CON_DX_OBI	CON_DX_HIP	CON_DX_SAL	CANT_ATENC	VALOR_NETC	CANT_ATENC	VALOR_NETC	DIAS_HOSP	UBIGEO_ULT	DEPARTAME	PROVINCIA_	DISTRITO_UL	NIVE
20240426	20230331	52BAEBEF1D	65	070106	CALLAO	PROV. CONS	VENTANILLA	FEMENINO	20220620	Diabetes mellitus	no especificada			0	0	0	0	0					
20240426	20230331	7FD6261B4E	13	080907	CUSCO	LA CONVENC	KIMBIRI	MASCULINO	20221001	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	FFBE798BE21	96	150701	LIMA	HUAROCHIRÃ	MATUCANA	FEMENINO	20191216	Diabetes me	SI		SI	0	0	0	0	0					
20240426	20230331	A063A90D38	60	020801	ÃNCASH	CASMA	CASMA	FEMENINO	20191216	Diabetes mellitus tipo 1	SI			0	0	0	0	0					
20240426	20230331	648654B6F9	56	070101	CALLAO	PROV. CONS	CALLAO	MASCULINO	20181029	Diabetes me	SI	SI		1	17.73	0	0	0	070101	CALLAO	PROV. CONS	CALLAO	
20240426	20230331	FD1C342C46	47	110206	ICA	CHINCHA	GROCIO PRA	FEMENINO	20210702	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	3804A5CC15	59	021801	ÃNCASH	SANTA	CHIMBOTE	FEMENINO	20221209	Diabetes me	SI		SI	0	0	0	0	0					
20240426	20230331	2FC44345856	66	150112	LIMA	LIMA	INDEPENDEN	FEMENINO	20230209	Diabetes mellitus tipo 2	SI			2	0	0	0	0	150112	LIMA	LIMA	INDEPENDEN	
20240426	20230331	8E867A3C8B	46	220101	SAN MARTÃ	MOYOBAMB	MOYOBAMB	MASCULINO	20201211	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	2C0F63ED2C	55	130702	LA LIBERTAD	PACASMAYO	GUADALUPE	FEMENINO	20190401	Diabetes me	SI		SI	3	7.2	0	0	0	130702	LA LIBERTAD	PACASMAYO	GUADALUPE	
20240426	20230331	BE4966A6364	44	150103	LIMA	LIMA	ATE	FEMENINO	20220718	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	3BAD999435	29	150103	LIMA	LIMA	ATE	MASCULINO	20180306	Diabetes me	SI			2	77.46	0	0	0	150103	LIMA	LIMA	ATE	
20240426	20230331	E62963645D8	26	020101	ÃNCASH	HUARAZ	HUARAZ	FEMENINO	20220712	Diabetes me	SI			2	9	0	0	0	020101	ÃNCASH	HUARAZ	HUARAZ	
20240426	20230331	31A957450C0	50	150143	LIMA	LIMA	VILLA MARÃA	FEMENINO	20180519	Diabetes me	SI			1	3.6	0	0	0	150133	LIMA	LIMA	SAN JUAN DE	
20240426	20230331	70CCC27812	42	200115	PIURA	PIURA	VEINTISEIS D	FEMENINO	20200219	Diabetes me	SI			0	0	0	0	0					
20240426	20230331	4A40BE08CF	67	150137	LIMA	LIMA	SANTA ANITA	FEMENINO	20180519	Diabetes me	SI	SI		0	0	0	0	0					
20240426	20230331	3C457DB7B6	57	020101	ÃNCASH	HUARAZ	HUARAZ	MASCULINO	20210111	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	17463D8B48	59	110204	ICA	CHINCHA	CHINCHA BA	FEMENINO	20180117	Diabetes me	SI			1	0	0	0	0	110204	ICA	CHINCHA	CHINCHA BA	
20240426	20230331	3F6DACE2E5	44	130401	LA LIBERTAD	CHEPÃ‰N	CHEPEN	FEMENINO	20221102	Diabetes me	SI	SI	SI	0	0	0	0	0					
20240426	20230331	E2C09B4B1F2	67	130704	LA LIBERTAD	PACASMAYO	PACASMAYO	FEMENINO	20200904	Diabetes me	SI	SI	SI	4	132.85	0	0	0	130704	LA LIBERTAD	PACASMAYO	PACASMAYO	
20240426	20230331	8DC991EC40	51	160113	LORETO	MAYNAS	SAN JUAN BA	MASCULINO	20200210	Diabetes mellitus tipo 1				0	0	0	0	0					
20240426	20230331	60D7BB15BF	38	150101	LIMA	LIMA	LIMA	FEMENINO	20190610	Diabetes me	SI			0	0	0	0	0					
20240426	20230331	0D391D2FAB	52	150132	LIMA	LIMA	SAN JUAN DE	MASCULINO	20200107	Diabetes me	SI			0	0	0	0	0					
20240426	20230331	97A08550D8	80	230101	TACNA	TACNA	TACNA	FEMENINO	20201008	Diabetes me	SI		SI	0	0	0	0	0					
20240426	20230331	D896EAA5F5	62	150106	LIMA	LIMA	CARABAYLLO	FEMENINO	20180917	Diabetes mellitus	no especificada		SI	0	0	0	0	0					
20240426	20230331	003FA5540BF	71	200104	PIURA	PIURA	CASTILLA	FEMENINO	20220930	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	FE4C3313A5E	81	040305	AREQUIPA	CARAVELÃ	BELLA UNIÃ“	FEMENINO	20210826	Diabetes me	SI			0	0	0	0	0					
20240426	20230331	B4884BF4C84	47	250105	UCAYALI	CORONEL PO	YARINACOCH	FEMENINO	20180113	Diabetes mellitus	no esp	SI		0	0	0	0	0					
20240426	20230331	AE123E37CAE	76	150135	LIMA	LIMA	SAN MARTÃN	MASCULINO	20180122	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	EB8B94D6C3	50	130101	LA LIBERTAD	TRUJILLO	TRUJILLO	MASCULINO	20180217	Diabetes me	SI		SI	0	0	0	0	0					
20240426	20230331	77F0BCE1E71	58	070101	CALLAO	PROV. CONS	CALLAO	MASCULINO	20200204	Diabetes mellitus tipo 2				0	0	0	0	0					
20240426	20230331	8F3FC559B4D	74	150101	LIMA	LIMA	LIMA	FEMENINO	20180326	Diabetes me	SI	SI	SI	0	0	0	0	0					
20240426	20230331	5998683825A	26	150117	LIMA	LIMA	LOS OLIVOS	FEMENINO	20210119	Diabetes mellitus tipo 2			SI	3	93.62	0	0	0	150135	LIMA	LIMA	SAN MARTÃN	
20240426	20230331	6B2BE53D5F	21	160103	LORETO	MAYNAS	FERNANDO L	FEMENINO	20210915	Diabetes mellitus	no especificada			0	0	0	0	0					
20240426	20230331	BB3941ED18	54	070106	CALLAO	PROV. CONS	VENTANILLA	FEMENINO	20210502	Diabetes me	SI			0	0	0	0	0					
20240426	20230331	E4079F77C0E	59	230101	TACNA	TACNA	TACNA	MASCULINO	20210818	Diabetes me	SI	SI		1	11.4	0	0	0	230101	TACNA	TACNA	TACNA	

Afiliados_activos_DM_SIS

ANEXO F: OPINION MEDICA PARA PREDICCION

Arequipa, 15 de mayo de 2024

VALIDACION DE PROPUESTA PARA LA

PREDICCION DE LA DIABETES

Mediante el presente documento, el Bachiller Berrios Zuniga, Alvaro Daniel egresado de la carrera de Ingeniería de Sistemas de la Universidad Católica de Santa María identificado con N° de DNI 73938490 y el código de estudiante 2017220641.

Brindo acceso al entorno de predicción en Google Colab para la diabetes a el Dr. Calderón Flores, Leirson Harry identificado con el N° de CMP: 065532 el cual tiene un rango de Medico General. Para que pueda aplicarlo en 10 personas dentro del rango de 16 a 34 años y así poder validar el resultado de la tesis denominada:

Predicción de la Diabetes mediante Aprendizaje de Maquina con el uso de Datos Biométricos de estudiantes de pregrado de una Universidad Privada en la ciudad de Arequipa

El resultado de esta evaluación permite dar validez para el uso de la predicción de una posible diabetes usando Machine Learning. La conclusión de la evaluación médica fue la siguiente:

Resultados de la aplicación del entorno de predicción para la diabetes	Usando un Excel y aplicado en el entorno de predicción se determinó que, de los 10 pacientes examinados entre 18 a 27 años, 2 de ellos tienen una posible diabetes por sus altos niveles de glucosa. Validando que, si realiza un análisis adecuado, pero se necesitan más datos de exámenes acorde a la diabetes para tener un diagnóstico al 100% por lo que esta es un avance en la ciudad de Arequipa, pero está en una fase inicial y requiere más investigación en el campo de la medicina.

Leirson Harry Calderón Flores
MÉDICO - CIRUJANO
CMP: 65532

FIRMA
Bachiller en Ingeniería de Sistemas
Alvaro Berrios Zuniga

FIRMA
Médico General
Dr. Harry Calderón Flores

ANEXO G: PRUEBAS DE CAMPO DEL PROTOTIPO