Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

PREDICCIÓN DE OBESIDAD EN LA ADOLESCENCIA MEDIANTE APRENDIZAJE DE MÁQUINA A TRAVÉS DE MEDIDAS ANTROPOMÉTRICAS

Tesis presentada por la Bachiller: **Gutiérrez Quintanilla, Andrea Isabel**

Para optar el Título Profesional de **Ingeniero de Sistemas con Especialidad en Ingeniería de Software**

Asesor: **Dr. Esquicha Tejada, Jose David**

Arequipa- Perú

2022

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA DE SISTEMAS

CON ESPECIALIDAD EN INGENIERIA DEL SOFTWARE

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 24 de Octubre del 2022

Dictamen: 004138-C-EPIS-2022

Visto el borrador del expediente 004138, presentado por:

2015201522 - GUTIERREZ QUINTANILLA ANDREA ISABEL

Titulado:

PREDICCIÓN DE OBESIDAD EN LA ADOLESCENCIA MEDIANTE APRENDIZAJE DE MÁQUINA A TRAVÉS DE MEDIDAS ANTROPOMéTRICAS

Nuestro dictamen es:

APROBADO

1568 - ROSAS PAREDES KARINA
DICTAMINADOR

1635 - SULLA TORRES JOSE ALFREDO
DICTAMINADOR

1910 - CASTRO GUTIERREZ EVELING GLORIA
DICTAMINADOR

PRESENTACIÓN

Señor Decano de la Facultad de Ciencias e Ingenierías Físicas y Formales

Señor Director de la Escuela Profesional de Ingeniería de Sistemas.

Sres. Miembros del Jurado.

De conformidad con las disposiciones del Reglamento de Grados y Títulos de la Escuela Profesional de Ingeniería de Sistemas, pongo a vuestra consideración el siguiente trabajo de investigación titulado: "**PREDICCIÓN DE OBESIDAD EN LA ADOLESCENCIA MEDIANTE APRENDIZAJE DE MÁQUINA A TRAVÉS DE MEDIDAS ANTROPOMÉTRICAS**", el cual de ser aprobado me permitirá optar por el Título Profesional de Ingeniero de Sistemas con especialidad en Ingeniería de Software.

GUTIERREZ QUINTANILLA, ANDREA ISABEL.

AGRADECIMIENTOS

Agradezco al Ing. José Sulla Torres y a su grupo de personas que participaron en el proyecto de investigación en Perú recolectando los datos utilizados en este proyecto. Al Ing. José Esquicha por guiarme durante este proceso con sus conocimientos y asesorías brindadas. A ellos mi más amplio agradecimiento, por su confianza, paciencia y valioso apoyo durante todo este proceso.

Mi mayor reconocimiento y gratitud hacia ustedes.

DEDICATORIA

Este trabajo va dedicado en primer lugar a Dios, a la Virgen y a mis ángeles de la guarda; así como a las personas que estuvieron brindándome su apoyo incondicional a lo largo de este proceso: Mi familia y amigos.

A Dios, la Virgen María y mis ángeles de la guarda, que son quienes guiaron y bendijeron cada uno de los pasos que he dado, que me dieron fuerzas para continuar con mis metas trazadas sin rendirme.

A mis padres, quienes son las personas más en mi vida y los que me ha enseñado todo lo que una persona debe poseer, que me enseñaron que incluso la tarea más grande se puede lograr si se hace un paso a la vez. Sin ellos no hubiese podido alcanzar esta gran meta, esto va dedicado a ellos, que desde muy pequeña me inculcaron valores para ser una persona correcta en la vida y que me acompañan en todos mis sueños y metas. A mi hermano, mi eterno cómplice, y la persona por la que me esfuerzo para ser un buen ejemplo de profesional y de vida. ¡Este logro es para ustedes! Y una mención especial a mi fiel compañero de desvelos y tardes de estudio, que con verlo dormido cerca bastaba para sentir su compañía, también es para ti este logro, Odie.

A mis amigos, mis personas especiales, los que a pesar de una pandemia o de estar a más de tres mil kilómetros de distancia estuvieron y siguen ahí. Gracias por todo su apoyo moral, ese que me permitió seguir con empeño, dedicación y cariño este largo camino para poder ser ingeniera; gracias por confiar en mí y en que lo lograría, gracias por el apoyo durante el proceso y por darme una mano cuando lo necesité.

A cada uno de los profesores que a lo largo de mi etapa universitaria han dejado una huella imborrable en mi vida, por haber sido un pilar fundamental para mi aprendizaje.

RESUMEN

La obesidad es considerada por la Organización Mundial de la Salud como una Enfermedad No Transmisible (ENT), que ataca a más de la mitad de la población mundial y es el desencadenante de muchas enfermedades cardiovasculares y diabetes. Actualmente el 53,8% de personas residentes en el Perú mayores de 15 años tienen exceso de peso, ya sea sobrepeso u obesidad. Esto se debe a que, desde muy pequeños, por diferentes circunstancias no tienen una alimentación saludable y balanceada. Este problema se agrava con la llegada del COVID 19 y la cuarentena obligatoria. Por otro lado, el avance de la tecnología y de la Inteligencia Artificial se sigue dando de una manera exponencial, no solamente abarcando en áreas industriales, sino que poco a poco está siendo implementada en el área salud.

Es por ello por lo que se plantea el desarrollo de una red neuronal para la detección de la obesidad en adolescentes a través de medidas antropométricas**.** El sistema será capaz de realizar predicciones como resultado del análisis de los diferentes datos de una persona. El propósito del proyecto es servir de ayuda (por medio de aprendizaje automático) a la detección temprana la obesidad en adolescentes, además de ofrecer un nivel de confiabilidad mayor al 90% en caso la persona posea dicha enfermedad no transmisible y de esta forma prevenir resultados mortales por la falta de seguimiento y detección en las consultas médicas

Palabras claves

Redes neuronales, Obesidad, Aprendizaje de máquina, Medidas antropométricas, IMC, OMS, Metodología CRISP-DM, Inteligencia artificial.

ABSTRACT

Obesity is considered by the World Health Organization as a Non-Communicable Disease (NCD), which attacks more than half of the world's population and is the trigger for many cardiovascular diseases and diabetes. Currently, 53.8% of people residing in Peru over 15 years of age are overweight, either overweight or obese. This is because, from an incredibly early age, due to different circumstances, they do not have a healthy and balanced diet. This problem is aggravated by the arrival of COVID 19 and the mandatory quarantine. On the other hand, the advancement of technology and Artificial Intelligence continues to occur exponentially, not only covering industrial areas, but little by little it is being implemented in the health area.

That is why the development of a neural network for the detection of obesity in adolescents through anthropometric measurements is proposed. The system will be able to make predictions because of the analysis of the different data of a person. The purpose of the project is to help (through machine learning) in the early detection of obesity in adolescents, in addition to offering a reliability level greater than 90% in case the person has said NCD and thus prevent fatal outcomes due to obesity. the lack of follow-up and detection in medical consultations.

Keywords

Neural Networks, Obesity, Machine Learning, Biomarkers, BMI, WHO, CRISP-DM Methodology, Artificial Intelligence.

INTRODUCCIÓN

La obesidad en el mundo es considerada por la Organización Mundial de la Salud como una Enfermedad No Transmisible (ENT), tal y como lo indican en la Declaración Política de la Reunión de Alto Nivel de la Asamblea General sobre la Prevención y el Control de las Enfermedades No Transmisibles. Es considerada una enfermedad que ataca a más de la mitad de la población mundial. Este problema se agrava con la llegada del COVID 19 y la cuarentena obligatoria. Su prevalencia ha aumentado en los últimos 30-40 años y actualmente de cada 10 niños y adolescentes, uno es obeso. Esto se debe a que, desde muy pequeños, por diferentes circunstancias no tienen una alimentación saludable y balanceada. (Naciones Unidas, 2011)

Si esta condición corresponde a un factor de riesgo o enfermedad primaria es un tema ampliamente discutido. Es reconocida como una enfermedad por la Asociación Médica Estadounidense y la Organización Mundial de la Salud, con base en sus características metabólicas y hormonales, como la desregulación del apetito, el equilibrio energético anormal y la disfunción endocrina, entre otras. Sus principales factores de riesgo ambiental son el consumo de alimentos ultra procesados y el sedentarismo. El tratamiento de la obesidad/sobrepeso se basa en la terapia cognitivo-conductual, la intervención dietética y el aumento de la actividad física con disminución del sedentarismo.

La priorización de combatir la obesidad debería mejorar la calidad de vida, evitando la mortalidad temprana, reduciendo el riesgo cardiovascular y a padecer diabetes tipo 2, así como la incidencia de cualquier tipo de cáncer. Si se llega a tener controlado la obesidad en el país, se tendría un gran impacto en nuestra sociedad, mejorando la calidad de vida de los habitantes. Actualmente, con la pandemia del COVID 19, muchas personas han descuidado sus hábitos alimenticios. Según estudios, las personas hemos tenido un aumento de peso en el confinamiento entre 1 y 7 kg. Lo cual está complicando mucho mantener una vida saludable. Paralelamente, el colapso de los hospitales, o el riesgo de ir a un hospital y contagiarse es latente en las personas, lo que ha ocasionado que muchas no se realicen chequeos con frecuencia y que a largo plazo desarrollen enfermedades mortales.

Por otro lado, el avance de la tecnología y de la Inteligencia Artificial específicamente se sigue dando de una manera exponencial, a medida que pasa el tiempo se descubren nuevas cosas que pueden ayudarnos a mejorar nuestro estilo de vida, e incluso facilitarnos las cosas que realizamos a diario. La inteligencia artificial no solamente está abarcando en áreas

industriales, sino que poco a poco está siendo implementada en el área salud, de tal forma que los diagnósticos, por ejemplo, se hacen en tiempo récord y desde la comodidad del hogar. En el ámbito de la salud, se está perfilando con el tiempo como una herramienta capaz de aprender y analizar con rapidez enormes cantidades de información de los historiales de pacientes, de las pruebas de imagen y de los avances científicos para ayudar a los doctores a ofrecer mejores diagnósticos y tratamientos.

Es por ello por lo que se plantea el desarrollo de una red neuronal para la detección de la obesidad en adolescentes a través de marcadores sencillos, como la estatura, peso, grosor de diferentes partes del cuerpo. El propósito del proyecto de desarrollo de una red neuronal para predecir la obesidad en adolescentes es servir de ayuda (por medio de aprendizaje automático) a los médicos en general al momento de realizar pruebas de rutina para la detección temprana la obesidad en adolescentes, además de ofrecer un nivel de confiabilidad mayor al 50% en caso la persona posea dicha ENT y de esta forma prevenir unos resultados mortales por la falta de seguimiento y detección en las consultas. Se requiere proporcionar datos reales de evaluaciones realizadas a adolescentes de diferentes edades y realidades; debido a que está usando aprendizaje automático para poder generar buena respuesta basada en hechos reales, de forma que su uso sea justificable.

Según lo especificado, el documento se organiza en 5 secciones principales: En el primer capítulo se especificará el planteamiento teórico, el cual proporcionará una descripción del propósito, alcance y objetivos del sistema, en el cual se incluirá el modelo de ciclo de vida y la metodología de desarrollo. Como segundo capítulo se tendrá el marco teórico, en el cual se describirá cómo estarán organizados y cómo funcionarán los recursos humanos del sistema. Además, se dará un contexto sobre la planificación estimada, definiendo las fases, hitos y el seguimiento de estas. En el tercer capítulo se tratará el tema de análisis, construcción y evaluación de las técnicas de aprendizaje automático, en el cual se observará la parte más práctica, donde se irá aplicando cada una de las fases de la metodología CRISP-DM al proyecto planteado. En el cuarto capítulo se expondrán los resultados obtenidos, en el cual se realizará un recuento de cada uno de los pasos seguidos a través de la metodología, así como los resultados obtenidos y pruebas realizadas a fin de verificar que el modelo es funcional. Finalmente se expondrá las conclusiones, recomendaciones y trabajos futuros a partir de la investigación realizada.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO I

1. PLANTEAMIENTO DE LA INVESTIGACIÓN

1.1. Planteamiento del Problema

La obesidad y sobrepeso se definen como “acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud”. Ambas son enfermedades crónicas que no tienen distinción con los grupos poblacionales: las padecen personas de bajos y altos recursos; personas muy jóvenes como los niños hasta ancianos. Se diagnostica mediante el examen de Índice de Masa Corporal (IMC), que consiste en dividir el peso de una persona en kilos por el cuadrado de su talla en metros (kg/m^2). (Organización Mundial de la Salud, 2021)

Las causas más comunes que ocasionan esta enfermedad son: dietas no saludables basadas en azúcares y grasas, así como poca actividad física. Hoy en día, durante la pandemia originada por el virus del COVID-19 este problema se ha agravado: A consecuencia del confinamiento, las personas alrededor del mundo se han vuelto más sedentarias y han cambiado sus hábitos alimenticios, ya sea recortando las comidas diarias, como aumentando el azúcar y comidas altas en grasas saturadas en su dieta. Al mismo tiempo, el riesgo al contagio ha disminuido las visitas médicas para chequeos rutinarios, lo que ocasiona que muchas personas no tomen conciencia que están tomando malos hábitos para su salud. Cabe resaltar que esta ENT no tiene género ni estatus social de preferencia, es decir, todas las personas están expuestas por igual a este mal. (Collave Garcia, 2021)

Paralelamente, el crecimiento de la tecnología, y sobre todo de la analítica de datos se da de forma exponencial. Actualmente existen muchas maneras de tratar la información y facilitar el día a día del ser humano. Hablando específicamente de la inteligencia artificial, es un campo muy amplio que hoy en día en el extranjero, sobre todo en países del primer mundo, se utiliza

en diferentes campos: contabilidad, administración, geología, agronomía, medicina, etc.; mientras que en Latinoamérica aún no se utiliza mucho. (Peinado Pineda & Díaz Salas, 2021)

Hoy en día la predicción de obesidad en adolescentes está limitado a realizarse diferentes estudios (sanguíneos, físicos, etc.), los cuales mediante sus resultados demuestran si la persona evaluada sufre de sobrepeso/obesidad y posiblemente qué órganos se están viendo afectados. Muchas veces la detección llega tarde, cuando ya la persona ha desarrollado enfermedades como diabetes, ceguera, cardiopatías isquémicas, cáncer, enfermedades en la vesícula biliar, hígado graso, apnea de sueño y osteoartritis. (Centro Nacional de Alimentación y Nutrición, 2020)

Por lo anteriormente descrito, se propone convertir al aprendizaje de máquina como el mejor aliado en estos momentos tan complicados a causa a la pandemia, creando un sistema web que mediante redes neuronales y el ingreso de diferentes medidas antropométricas se pueda predecir si esa persona tiene o no riesgo a padecer de obesidad, ya que está comprobado que mediante los niveles de diferentes sustancias que se obtienen o medidas del cuerpo humano (llamadas medidas antropométricas) se puede predecir con un nivel de confiabilidad mayor al 90% si es que la persona cuenta con algún tipo de enfermedad, en este caso, obesidad.

1.2. Objetivos de la Investigación

1.2.1. General

Predecir la obesidad en adolescentes utilizando modelos de aprendizaje automático.

1.2.2. Específicos

- Analizar y tomar los requerimientos de las necesidades que debe cubrir el funcionamiento de la predicción.
- Analizar datos referidos de adolescentes de diferentes colegios de Arequipa.
- Realizar la transformación de los datos del dataset analizado previamente.

- Comparación y elección del mejor modelo de aprendizaje automático para predecir la obesidad en adolescentes.
- Entrenar una red neuronal artificial para predecir la obesidad en adolescentes.
- Validación de los resultados en la predicción de obesidad en adolescentes.

1.3. Preguntas de investigación

a) ¿Se puede predecir la obesidad en adolescentes utilizando modelos de aprendizaje automático?

b) ¿Qué necesidades debe cubrir el funcionamiento de la predicción?

c) ¿Qué es lo que se puede concluir a partir del análisis de los datos referidos a casos de obesidad en adolescentes de diferentes colegios de Arequipa?

d) ¿Qué columnas debe tener el dataset personalizado con los datos analizados previamente?

e) ¿Qué criterios se debe tener al momento de la comparación y elección del mejor modelo de aprendizaje automático para predecir la obesidad?

f) ¿Qué consideraciones se debe tener para construir y entrenar una red neuronal artificial capaz de predecir la obesidad?

g) ¿De qué forma se debe evaluar la red neuronal artificial desarrollada para comprobar su correcto funcionamiento?

1.4. Línea y sublínea a la que corresponde el Problema

1.4.1. Línea

Inteligencia Artificial.

1.4.2. Sublínea

Redes Neuronales.

1.5. Tipo y Nivel de la Investigación

1.5.1. Tipo de investigación:

a) Según su finalidad:

Aplicada, ya que este trabajo busca determinar la mejor estrategia para la predicción de obesidad en adolescente según el dataset propuesto. Además, cuenta con una base teórica, que es la que genera el conocimiento práctico.

b) Según la fuente de datos:

Empírica o de campo, ya que se obtendrá resultados a partir de pruebas reales.

1.5.2. Nivel de investigación:

Exploratoria, ya que, si bien es cierto existen trabajos previos en el tema, se considera que esta área es extensa y aún falta por explorar, por lo que se propondrá realizar un análisis entre 5 modelos a fin de encontrar el que mejor se adapte al problema y dataset actual.

1.6. Palabras claves

Redes neuronales, Obesidad, Aprendizaje de máquina, Medidas antropométricas, IMC, OMS, Metodología CRISP-DM, Inteligencia artificial.

1.7. Solución propuesta

Durante la investigación se realizará una red neuronal artificial (RNA), en la cual se procesarán los datos recolectados de adolescentes de la provincia de Arequipa para poder obtener el resultado de si presentan obesidad o no. Además, durante la investigación se estudiarán puntos para tener en cuenta al momento de aplicar dicho algoritmo.

1.8. Justificación e importancia

Los resultados obtenidos en este estudio permitirán comprender la relación que existe entre las medidas antropométricas del cuerpo humano y el diagnóstico de la obesidad, además que permitirá comprender mejor el uso de herramientas de aprendizaje automático en este campo, lo cual demostrará la importancia que tienen dentro de este para el mejor manejo de la información.

1.9. Aporte

Este estudio brindará un aporte científico ya que ayudará a la predicción temprana de la obesidad en adolescentes a través de datos fáciles de obtener como lo son peso, sexo, estatura y edad; ya que, una vez culminado este trabajo, se obtendrá el resultado del análisis empleado en los datos, y a partir de ellos, se mejorará la red neuronal a fin de que el porcentaje de exactitud sea muy cercano al 100%. Por otro lado, también ayudará a hacer un mejor uso de las tecnologías modernas como el aprendizaje automático ya que dentro del área de la medicina son muy pocos los campos donde se utiliza.

1.10. Enfoque

Para este estudio se utilizó un método cuantitativo, por el cual se hizo una investigación minuciosa a partir de las variables que se tienen a través de métodos estadísticos y aplicación de aprendizaje automático para poder obtener resultados que permitan llegar al objetivo planteado. Se determino, además, utilizar este método por la facilidad en la manipulación de las variables y porque permite utilizar métodos estadísticos para comparar resultados o procesarlos a través de métodos estadísticos.

1.11. Alcances y limitaciones

Para el desarrollo de este proyecto se tomó en cuenta diferentes documentos de índole nacional e internacional para poder contextualizar el problema de la obesidad en adolescentes de manera más precisa, los cuales tienen como propósito dar a conocer en qué estado se encuentra la obesidad tanto en el Perú como a nivel mundial, y al mismo tiempo proponer soluciones a nivel gubernamental a fin de disminuir dicha enfermedad no transmisible. Es por ello, por lo que se le consideran importantes ya que proporcionan la información suficiente para justificar el desarrollo del proyecto.

Del mismo modo, se toma en cuenta los datos de personas entre 04 y 21 años de diferentes colegios de la ciudad de Arequipa, Perú, sobre las características clínicas que se observaron y midieron en un registro de 3068 adolescentes, tales como edad, peso, talla, IMC. En sí, el proyecto engloba dos campos de estudio como lo son la computación y la medicina.

El propósito del proyecto es proponer una ayuda (por medio de machine-learning) a los médicos y público en general al momento de realizar pruebas de rutina para la detección temprana de la obesidad en adolescentes, además de ofrecer un nivel de confiabilidad mayor al

50% en caso la persona posea dicha enfermedad no transmisible (ENT) y esta forma prevenir unos resultados mortales por la falta de seguimiento y detección en las consultas.

Se requiere que se le proporcionen datos reales de evaluaciones realizadas a adolescentes de diferentes centros educativos de la ciudad de Arequipa, debido a que está usando machine-learning para poder generar buena respuesta basado en hechos reales, de forma que su uso sea justificable. Los datos deben ser susceptibles al análisis para poder extraer conclusiones científicamente válidas.

Las características requeridas por el sistema, para lograr su objetivo son:

Captura de información

- Se establece la captura de datos a través de los métodos establecidos por las librerías a utilizar.

Consolidación de información

- Analizar los datos, previamente localizados, en un archivo csv.
- Normalizar la información para asegurar la recuperabilidad, clasificación y orden de los datos.

Aprendizaje

- Crear una red neuronal que sea capaz de tomar la información almacenada y aprender de ella, para crear sus propias conclusiones.
- Crear un formato específico que permita plasmar los resultados de la red neuronal.

Procesos

Dentro de los procesos implicados a los que el sistema propuesto va a ayudar es a la detección de posible obesidad en adolescentes, tomando en cuenta los valores antropométricos que presente la persona a ser evaluada.

Áreas implicadas

Las áreas implicadas para la realización de este sistema inteligente son las áreas de biomédicas y computación; específicamente dentro del área de biomédicas el área de la endocrinología; y dentro del área de computación tenemos del área de inteligencia artificial.

El software que se va a desarrollar está diseñado para facilitar el proceso de detección de un posible caso de obesidad en adolescentes tomando en cuenta los valores antropométricos que presente la persona a ser evaluada.

Es por ello, que esta investigación tiene los siguientes puntos de alcance y limitaciones:

a) Se tomarán los datos antropométricos de adolescentes de colegios de la provincia de Arequipa, región Arequipa, país Perú.

b) Se realizará este trabajo de investigación en un ambiente donde se tendrá la posibilidad de usar herramientas para el procesamiento de datos, tales como RapidMiner; así como herramientas para la creación de la red neuronal tales como PyCharm, Google Colab, etc.

c) El tiempo tomado para realizar este trabajo es de un (1) año aproximadamente, tiempo en el cual se trabajará el preprocesamiento de datos, procesamiento de datos, ajuste de red neuronal, así como la documentación de la investigación.

1.12. Población y Muestra o Universo

1.12.1. Universo

Personas de diversos colegios de la ciudad de Arequipa.

1.12.2. Muestra o Universo

Personas de diversos colegios de la ciudad de Arequipa entre 4 y 21 años, de los cuales se cuenta con datos tales como género, estatura, entre otros datos.

1.13. Métodos, Técnicas e Instrumentos de Recolección de datos

1.13.1. Métodos de la investigación

Investigación predictiva, ya que requerirá de una exploración previa, análisis, descripción y comparación, para luego explicar el porqué de lo planteado.

1.13.2. Técnicas para la investigación

De forma inicial, en la tabla de datos tenemos las siguientes columnas:

a) Edad

b) Sexo

c) Peso

d) Estatura_cm

e) Estatu_metr

f) Índice de Masa Corporal (IMC)

Cada una de ellas son consideradas como variables, ya sea categóricas o numéricas. Las variables categóricas se preprocesarán a fin de convertirlas en variables numéricas. Luego de ello, se realizarán los procesamientos en test y en producción para poder obtener resultados.

1.13.3. Instrumentos para tratamiento de datos

Se utilizará una tabla en la que se visualizarán los datos recolectados, los cuales se utilizarán en la etapa experimental del trabajo de investigación. En dicha tabla, se agregarán las columnas con datos que se necesitan para asegurar una exactitud más cercana al 100%.

CAPÍTULO II

2. FUNDAMENTOS TEÓRICOS

2.1. Bases teóricas del Proyecto

Con la finalidad de comprender la información general del tema que se desarrollará, se han escogido una serie de conceptos y temas para tener un mejor dominio del conocimiento.

2.1.1. Malnutrición y obesidad

Figura 1: Obesidad alrededor del mundo

Fuente: (Instituto Internacional de Investigación sobre Políticas Alimentarias, 2016)

Actualmente, se le estima a la malnutrición como el problema más grande que la sociedad encara, ya que es una condición que afecta a uno de cada 3 personas en el mundo. La malnutrición se muestra de distintas modalidades: Retraso de aumento en chicos, personas propensas a infecciones gracias a la falta de vitaminas y minerales de trascendencia en su organismo, personas con exceso de peso o riesgo a sufrir enfermedades crónicas a raíz del sobrepeso o por excesivo consumo de azúcar, sal o grasas. La malnutrición y la alimentación

es una enorme carga mundial de morbilidad (CMM), ya que cada territorio hace frente a una realidad distinta con base a la economía, sociedad que tienen. Por ejemplo, en países con pobreza extrema, la desnutrición en niños y adolescentes es mucho más grave que la que pueden hacer frente países tercermundistas, sin embargo, estos además afrontan el sobrepeso y obesidad en más enorme medida que otros países. El sobrepeso es una de las enfermedades no transmisibles (ENT) más comunes alrededor del mundo. A esta se le define como una "acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud". (Organización Mundial de la Salud, 2021)

En la actualidad, los problemas de salud derivados de la obesidad han penetrado todos los segmentos sociales del Perú. Según los datos del módulo nutricional de la Encuesta Nacional de Hogares del año 2008, indica que el sobrepeso en niños era de un 7.8%. Sobre los adolescentes entre 10 y 19 años, rango de edad donde se manifiestan muchos cambios en tamaño y forma, se sabe que el sobrepeso y obesidad era del 13.5% en varones y 15% de mujeres, además, la Encuesta Global de Salud Escolar, realizada en el 2010, es decir dos años después de la primera, no indica ningún cambio, al contrario, deja a la luz que el problema se agravó, ya que reportó que el 20% de escolares de secundaria presentan sobrepeso y el 3% presentan obesidad. En adultos este problema no mejora, ya que personas mayores entre 35 y 40 años, el 66% de mujeres y el 55% de hombres padecen de obesidad. (Ministerio de Salud del Perú, 2012)

2.1.1.1. Causas de la obesidad en el Perú

Algunas de las causas que exista tanta malnutrición en el mundo es la globalización, ya que ha ocasionado diferentes cambios, se mencionan a continuación:

- **Gran presencia de alimentos ultra procesados**

  Durante este proceso de globalización se ha podido evidenciar una gran presencia de alimentos ultra procesados en la dieta de cada hogar, todo gracias al gran marketing de estas. El consumo de bebidas gaseosas, alcohólicas ha incrementado de forma notable al paso de los años.

- **Reajuste de costos de alimentos básicos**

  El reajuste de ciertos costos de alimentos ha hecho que dichos se vuelvan inaccesibles para varias familias. Ejemplificando, en una época de 18 años, las frutas y vegetales

han aumentado su costo un 118%, el pescado en un 77%, los cereales un 75%, mientras tanto que las grasas sólo un 35%, así como las gaseosas solo un 20%. Esta alteración tan inequitativa es determinante al instante de mercar las provisiones para el hogar.

- **Poca actividad física y sedentarismo**,

  El llevar una vida sedentaria está vinculada a un extenso historial de patologías, donde está la obesidad incluida. Estudios hechos en Latinoamérica demuestran que 2 tercios poblacional cumplen con la actividad física moderada diaria, que son 30 min, y esto lo consiguen en las ocupaciones que hacen para transportarse de un espacio a otro. (Ministerio de Salud del Perú, 2012)

2.1.1.2. Medidas propuestas para combatir la obesidad a nivel mundial

La Comisión para Acabar con la Obesidad Infantil de la Organización Mundial de la Salud (OMS) elaboró recomendaciones para poder combatir la obesidad en niños y adolescente:

a) Fomentar el consumo de alimentos saludables

Utilizar programas integrales que promuevan la alimentación sana y disminuyan la alimentación malsana y bebidas azucaradas.

b) Impulsar la actividad física

Ejercer programas integrales que promuevan la actividad física y que disminuyan los comportamientos sedentarios.

c) Atención pregestacional y natal

Robustecer las orientaciones para la prevención de las patologías no transmisibles, como la obesidad infantil, a través de la atención pregestacional.

d) Dieta y la actividad física en la primera infancia

Ofrecer orientaciones y apoyo al establecimiento de una dieta sana, pautas de sueño y actividad física durante la niñez.

e) La salud, la nutrición y la actividad física para los niños en edad escolar

Utilizar programas integrales que promuevan ámbitos estudiantiles saludables, conocimientos básicos en temas de salud y nutrición y actividad física.

f) Control de peso

Brindar servicios para controlar el peso de las personas a fin de poder modificar el estilo de vida de ser necesario. (Organización Mundial de la Salud, 2016)

2.1.1.3. El COVID 19 y la obesidad en el Perú

La pandemia generada por el COVID-19 ha impactado de manera negativa en la vida de las personas, tanto en la dieta, sueño y actividad. Según diferentes estudios que se han realizado a lo largo del tiempo, se dice que la pandemia del COVID-19 ha sido el causante de que las estadísticas de obesidad se disparen e incrementen alarmantemente dentro del territorio nacional, ya que, al estar sometidos en un confinamiento forzoso, muchas personas han dejado de hacer el ejercicio diario que usualmente hacían cuando se desplazaban de su casa al trabajo, o a diferentes lugares, lo que ha ocasionado un incremento ponderal en hombres y mujeres de entre 1 y 5 kg. El estar sometidos en un encierro forzado también ha causado que la alimentación no sea tan saludable, debido a que al tener la comida poco saludable un precio más bajo que la que sí lo es, se consume más sin medir las consecuencias. Al 2021 se estima que más del 60% de la población del Perú mayores de 15 años sufren de sobrepeso u obesidad. (Collave Garcia, 2021).

Conforme el registro de tele consulta que tiene EsSalud, se pudo observar que los casos de obesidad en niños de entre 7 y 12 años se han duplicado a lo largo de la emergencia por el coronavirus debido a los siguientes puntos: cierre de colegios y confinamiento en sus domicilios, consumo de comida chatarra, dedicación de, inclusive, hasta 12 horas cotidianas a ver televisión y escasa o nula actividad física en el día. En los años 70, la obesidad perjudicaba a menos del 1% de los niños en todo el mundo, en el 2016, la obesidad infantil perjudicaba al 6%. No obstante, actualmente, el 12% de niños en todo el mundo padecen de obesidad. (Centro Nacional de Alimentación y Nutrición, 2020)

Este crecimiento de personas con sobrepeso u obesidad es bastante grave en términos de contagios de coronavirus, debido a que se ha encontrado una interacción entre la obesidad y la tasa de mortalidad por Coronavirus: Territorios como USA o Inglaterra encabezan la lista de territorios con más porcentaje de individuos con sobrepeso/obesidad, y simultáneamente poseen una tasa de mortalidad por Coronavirus 19 alta, y sin embargo, territorios como Japón o Corea del sur, donde se dio prioridad a la salud pública, poseen tasas bastante bajas de mortalidad. Hablando a nivel Latinoamérica, el panorama no es diferente, ya que dentro de la región el problema de sobrepeso y obesidad en las personas se ha incrementado de manera alarmante a partir del inicio del confinamiento, agregando así un problema sanitario más sobre la que actualmente se tiene con el COVID-19. (Organismo Andino de Salud – Convenio Hipólito Unanue, 2021)

2.1.2. Inteligencia artificial

Hoy en día se escucha mucho hablar de cómo es que la tecnología está ayudando y revolucionando en el día a día del ser humano. Desde realizar tareas repetitivas hasta tomar decisiones. Todo esto es gracias al avance de esta dentro del campo de la inteligencia artificial. La inteligencia artificial, también conocida como "IA", es la combinación de diferentes algoritmos que tienen el propósito de crear máquinas que lleguen a tener las mismas capacidades del ser humano al momento de realizar actividades cotidianas y tomar decisiones, es por ello por lo que uno de los objetivos de la IA es que su funcionamiento se acerque al funcionamiento de la mente humana. Aún se le considera una tecnología misteriosa ya que es un campo bastante amplio, pero desde hace unos años ya está presente en el día a día de la sociedad. (Cai et al., 2021)

Esta nueva tecnología hace posible que las máquinas y algoritmos aprendan a partir de la experiencia, que puedan ajustarse a nuevas aportaciones y que resultado de esto puedan realizar tareas como si lo hiciera un ser humano. A partir de lo que propone la IA, las computadoras son entrenadas para realizar actividades específicas procesando una gran cantidad de datos y reconociendo patrones en ellos, los cuales son aplicados en procesamientos iterativos y algoritmos inteligentes, permitiendo que el software aprenda automáticamente de estos patrones. (Peinado Pineda & Díaz Salas, 2021)

En la actualidad, la inteligencia artificial no solo está presente en el mundo empresarial, sino también en el ámbito social, con aplicaciones que van desde la detección de diferentes

enfermedades hasta la lucha contra problemas medioambientales, como la deforestación del Amazonas. (Coca Bergolla et al., 2021)

2.1.2.1. Categorías

En el libro "Inteligencia Artificial: Un Enfoque Moderno", se definen cuatro tipos de inteligencia artificial.

a) **Sistemas que piensan como humanos:**

   Estos son sistemas que intentan imitar el pensamiento humano, como la toma de decisiones, la resolución de problemas y el aprendizaje. Un ejemplo son las redes neuronales.

b) **Sistemas que actúan como humanos:**

   Intentan actuar como humanos. En otros términos, imitan la conducta humana. Un ejemplo de este sistema es la robótica.

c) **Sistemas que piensan racionalmente:**

   Intentan imitar el pensamiento lógico racional humano, es decir, estudian cómo hacer que las máquinas perciban, razonen y actúen en consecuencia. Un ejemplo son los sistemas expertos.

d) **Sistemas que actúan racionalmente:**

   Este sistema intenta imitar lógicamente el comportamiento humano. Se relaciona con comportamientos inteligentes en especie. Un ejemplo son los agentes inteligentes. (Norvig & Russell, 2004)

2.1.2.2. Campos de aplicación actual de la Inteligencia Artificial

A medida que avanza la tecnología y por ende se genera mayor conocimiento científico en el área, se abren más oportunidades para poder utilizarla en diferentes campos. La aplicación de la Inteligencia Artificial en la vida cotidiana no solamente es en el campo tecnológico, sino que puede ir desde el marketing digital, ciencia y medicina, medio ambiente, hasta en disciplinas sociales como antropología y sociología.

- **Medicina**

  Las aplicaciones de inteligencia artificial tienen el potencial de proporcionar lecturas médicas y de rayos X personalizadas. Además, los ChatBot ya nos permiten realizar un autodiagnóstico según nuestra sintomatología, esto gracias a la recopilación de datos, la cual produce patrones que ayudan a detectar componentes genéticos propensos a desarrollar alguna patología.

- **Educación**

  Permite crear actividades eficientes e innovadoras, además de lograr una mejor comprensión de los perfiles de los estudiantes para brindarles un entorno de aprendizaje saludable y seguro.

- **Finanzas**

  En las instituciones financieras, la tecnología de IA se puede utilizar para identificar qué transacciones pueden ser fraudulentas, emplear evaluaciones crediticias rápidas y precisas, automatizar manualmente tareas intensivas de gestión de datos y brindar asesoramiento operativo a los clientes.

- **Agricultura**

  Se puede tener mejores pronósticos meteorológicos, ya que existen muchos desafíos ambientales, y el poder tener la capacidad de predecirlos y poder estar preparados para enfrentarlos ayudaría a tener una mayor precisión al momento que se siembra en la tierra. También ayuda al sistema de monitoreo de suelos y cultivos, en la detección de plagas y en realizar un riego adecuado a los cultivos. (Peinado Pineda & Díaz Salas, 2021)

2.1.2.3. Ramas y subramas la Inteligencia Artificial

Los campos de aplicación de la inteligencia artificial son muchos, y algunos están orientados a satisfacer necesidades muy distintas. A continuación, se explican las 5 ramas que existen de la Inteligencia artificial.

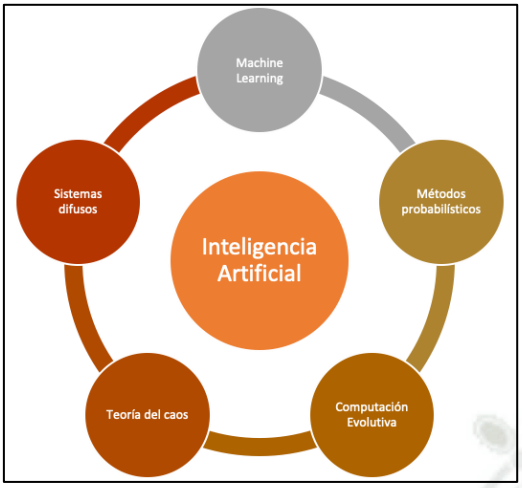


Figura 2: Ramas de la Inteligencia Artificial

Fuente: (Coca Bergolla et al., 2021)

- **Aprendizaje automático (ML)**

  Es la rama más conocida actualmente. El Aprendizaje automático es capaz de diseñar algoritmos que aprenden de sí mismo a través de la experiencia. Son capaces de realizar una tarea en específico sin necesidad de programar explícitamente las instrucciones. Dentro de esta rama se encuentran las Redes Neuronales, modelos basados en KNN y regresión lineal, etc.

- **Métodos probabilísticos**

  Es una rama de la IA definida como "la medida de la incertidumbre asociada con un suceso aleatorio". Hace posible el razonamiento bajo incertidumbre, el cual es necesario ya que los conocimientos son unilaterales. Dentro de esta rama están las redes Bayesianas, cadenas de Márkov, etc.

- **Computación evolutiva**

  Es una rama que involucra problemas de optimización a partir de la evolución biológica. Dentro de esta rama están los árboles genéticos, programación genética, etc.

- **Teoría del caos**

  Es una rama que estudia el comportamiento de los sistemas dinámicos y determinísticos cuyo comportamiento puede predecirse. Es aplicado dentro de la criptografía, robótica, biología.

- **Sistemas difusos**

  Esta rama se basa en la lógica difusa, en la cual la verdad no es exacta por lo que se define por regiones. (Coca Bergolla et al., 2021)

2.1.3. Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser programadas explícitamente, a través del aprendizaje de los datos históricos, ambientales y atributos observados. Está presente en aplicaciones como Netflix, Spotify, Gmail, Siri y Alexa. (BBVA, 2019)

Es un procedimiento analítico que permite que un sistema aprenda por sí mismo, sin asistencia humana, para descubrir patrones, tendencias e interrelaciones en los datos de manera automatizada y, como resultado de la comprensión anterior, la nueva información de los superiores se presenta en cada relación, prospecto. Se estima como una herramienta diseñada para mejorar la investigación de datos, para facilitar las predicciones futuras, ya sea utilizando nuevos sistemas o simplemente mejorando los existentes, utilizando algoritmos basados en información antigua o sacando el máximo rendimiento del sistema. (Ramirez Hinestroza, 2018)

Es por ello por lo que muchas organizaciones públicas como privadas están optando por utilizar este tipo de algoritmos para los servicios y productos que brinda, ya que de esta forma pueden mejorar los procesos a fin de mejorar la vivencia y entretenimiento de los consumidores, lo que aumentaría su competitividad y profesionalidad en el mercado en el que se encuentran. (Banu, 2022)

2.1.3.1 Tipos de aprendizaje automático

Existen 3 tipos de aprendizaje automático: Aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo.

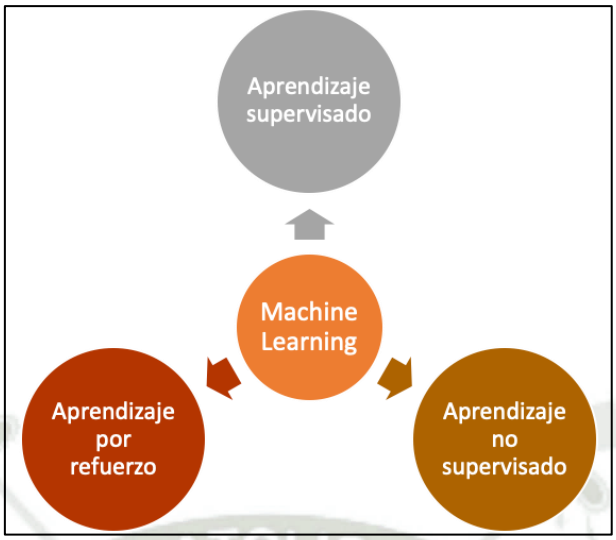


Figura 3: Tipo de aprendizaje automático

Fuente: (Banu, 2022)

- **Aprendizaje supervisado**

  El sistema se entrena proporcionando una cierta cantidad de datos que define en detalle con etiquetas. Una vez que se proporciona una cantidad suficiente de dichos datos, se pueden ingresar nuevos datos sin etiquetas en función de diferentes patrones registrados durante el entrenamiento. Este sistema se llama clasificación. Lo característico de este tipo de ML es que utiliza distintos tipos ejemplos para aprender cómo tratar nuevos casos.

- **Aprendizaje no supervisado**

  No se utilizan verdades básicas ni etiquetas en este tipo de aprendizaje. Estos sistemas están diseñados para comprender directamente y abstraer patrones de información. Este es un modelo de problema llamado agrupamiento. Este es un método de entrenamiento que es más similar a la forma en que los humanos procesan la información.

- **Aprendizaje por refuerzo**

  El sistema aprende de la experiencia. Esta es una técnica basada en prueba y error y utilizando una función de recompensa que optimiza el comportamiento del sistema. Es una de las formas más interesantes para que un sistema de IA aprenda ya que no requiere introducir mucha información. (Banu, 2022)

2.1.4. Redes neuronales

Las redes neuronales son un área del aprendizaje automático que trata de imitar el funcionamiento de las neuronas del cerebro humano. En la última década, los avances en hardware han permitido que la capacidad computacional crezca considerablemente, lo que ha permitido evaluar este tipo de sistemas con gran velocidad. (Kumar et al., 2022)

Entrenar una red neuronal consiste en ajustar todos los pesos de las entradas de cada una de las neuronas que son parte de la red neuronal, para que las respuestas de la capa de salida se ajusten lo más viable a los datos que conocemos. Inicialmente, cada una de las ponderaciones son aleatorias y las respuestas que resultan de la red son, probablemente, inexactas, empero la red aprende por medio del entrenamiento. Siempre se muestran a la red ejemplos para los que se sabe el resultado, y las respuestas que da se comparan con los resultados conocidos. La información procedente de esta comparación se pasa hacia atrás por medio de la red, cambiando las ponderaciones gradualmente. Mientras progresa el entrenamiento, la red se va realizando cada vez más rigurosa en la replicación de resultados conocidos. Una vez entrenada, la red se puede utilizar en casos futuros en los cuales se desconoce el resultado. (IBM, 2020)

2.1.4.1. Estructura básica de una red neuronal

En una neurona artificial, la suma de las entradas multiplicadas por sus pesos asociados establece el “impulso nervioso” que obtiene la neurona. Este costo, se procesa en el centro de la célula por medio de una función de activación que regresa un costo que se envía como salida de la neurona. (García-Olalla Olivera, 2019)

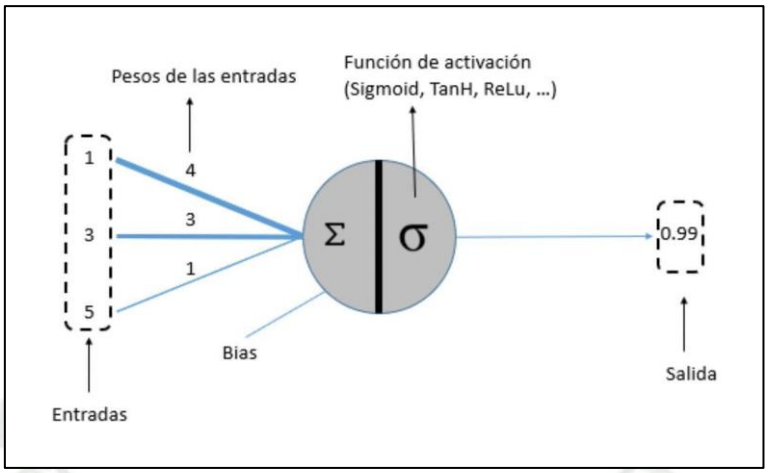


Figura 4: Estructura básica de una neurona artificial

Fuente: (García-Olalla Olivera, 2019)

Al mismo tiempo, las unidades de procesamiento de una red neuronal se organizan en capas.

a) Una capa de entrada

Cuenta con unidades que representan los campos de entrada. Las neuronas de la primera capa reciben como entrada los datos reales que alimentan a la red neuronal.

b) Una o varias capas ocultas

Se llaman así ya que no se conocen los datos de entrada y salida de cada una de ellas.

c) Una capa de salida

La salida de la última capa es el resultado visible de la red, con una unidad o unidades que representan el campo o los campos de destino. (García-Olalla Olivera, 2019)

2.1.5. Medidas antropométricas

Las medidas antropométricas, según la UNESCO, se definen como "el estudio del tamaño, proporción, maduración, forma y composición corporal, y funciones generales del organismo, con el objetivo de describir las características físicas, evaluar y monitorizar el crecimiento, nutrición y los efectos de la actividad física". (Programa Salud, Trabajo y Ambiente en América Central (SALTRA), Instituto Regional de Estudios en Sustancias Tóxicas, Universidad Nacional de Heredia, 2014)

Estas medidas tienen como pilares cuatro puntos: medidas corporales, estudio de los tipos de cuerpos humanos con tipos de temperamentos, o somatotipos; la proporcionalidad y la composición corporal. Dentro del área de la salud son utilizadas para evaluar el estado nutricional de las personas, a fin de detectar anormalidades a tiempo, es por ello que utilizan las siguientes medidas corporales: Peso del cuerpo, Altura del cuerpo (estatura, talla), Altura al ojo, Altura al hombro, Altura al codo, Altura a la cadera, Altura al glúteo, Altura a la muñeca, Altura al tercer dedo (medio), Anchura lateral de brazos, Anchura de codos, Largura de brazos, Largura de puño, Anchura de hombros, Anchura de pecho, Anchura de cadera, Largura de brazo, Circunferencia de cuello, Circunferencia de pecho, Circunferencia de cintura, Circunferencia de cadera, Circunferencia de cabeza, Anchura de oídos, Anchura de cara, Anchura de cabeza, Altura de cabeza, Largura de cabeza, Largura de mano, Largura de palma de mano, Anchura de palma de mano, Diámetro agarre de mano, Anchura de muslos, Altura de cabeza sentado, Altura al ojo sentado, Altura al hombro sentado, Altura al codo sentado, Anchura del muslo sentado, Altura a los dedos sentado, Altura al puño sentado, Largura del muslo sentado, Largura de rodilla sentado, Altura del cuerpo sentado, Altura al glúteo sentado, Altura a la rodilla sentado, Altura al muslo sentado, Largura de brazo y mano, Anchura de espalda, Anchura de cadera sentado, Largura del pie, Altura del pie, Anchura del pie, Pliegue bicipital, Pliegue tricipital, Pliegue subescapular, Pliegue suprailíaco, Pliegue del muslo, Pliegue abdominal, Pliegue del pecho, Pliegue axilar, Pliegue de la pierna. (Programa Salud, Trabajo y Ambiente en América Central (SALTRA), Instituto Regional de Estudios en Sustancias Tóxicas, Universidad Nacional de Heredia, 2014)

2.2. Estado del arte

La obesidad es una de las enfermedades que ha tenido mayor crecimiento a lo largo de los últimos años: En las últimas décadas se ha triplicado el número de personas que la padecen alrededor del mundo; y a raíz de eso se le considera como “la epidemia del siglo XXI”. Lo más peligroso de esta enfermedad son los efectos adversos a los que se exponen estas personas: enfermedades coronarias, problemas de colesterol y triglicéridos; diabetes, accidentes cerebrovasculares e incluso diferentes tipos de cáncer (próstata, mama, colon, etc.). (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades, 2020)

La obesidad tiene una etiología compleja, que incluye tanto factores desarrollados en la etapa del embarazo, así como cambios antropométricos, metabólicos y hormonales que se dan en esta etapa, razón por la cual la incidencia en niños, niñas y jóvenes está aumentando a un

ritmo alarmante en muchos países, tomando más fuerza a partir de la llegada de la COVID-19. Esta condición es una amenaza para los sistemas de salud de muchos países, ya que la obesidad está asociada a diversas comorbilidades, tales como enfermedades cardiovasculares, diabetes, síndrome metabólico, etc. (Organización Mundial de la Salud, 2021)

Es esta etapa, los modelos de aprendizaje automático están tomando más fuerza como un método extremadamente útil en el campo de la medicina, ya que tienen con un excelente poder predictivo, la capacidad de modelar relaciones no lineales y complejas entre variable, además de poder lidiar con datos dimensionales, que son típicos en este campo. El uso de estos modelos en la medicina cotidiana reemplaza y facilita el análisis de modelos estadísticos que hoy por hoy, sin usarlos, son difíciles de manejar en algunos casos. En la mayoría de los casos, los modelos estadísticos tradicionales utilizan un conjunto reducido de factores de riesgo, mientras que con aprendizaje automático se puede utilizar otro tipo de variables más complejas. Cuando se comparan las regresiones en el mismo trabajo con los modelos de aprendizaje, es aquí donde se puede visualizar la mejora en los resultados obtenidos con estos modelos, ya que pueden generar mejores porcentajes de predicción que no solo ajusta el conjunto de entrenamiento, sino que también dan mejores resultados en las validaciones realizadas. Enfocándose en la predicción de enfermedades, como la obesidad, se debe considerar que la prevención de modelos de aprendizaje automático ha mostrado buenos resultados, no solo en el proceso de identificación de poblaciones en riesgo, sino también en la búsqueda formas de lograr las metas de prevención. (Peinado Pineda & Díaz Salas, 2021)

El uso de modelos de aprendizaje automático en el campo médico también ofrece nuevas ventajas y conocimientos para la predicción y prevención de enfermedades, así como dar la posibilidad de ser una herramienta de simulación, para que se puedan obtener nuevos conocimientos con enfoques terapéuticos. Por ejemplo, el uso de aprendizaje automático ha aumentado la precisión de la predicción en comparación con los modelos estadísticos de uso frecuente, ya que tienen la capacidad de modelar relaciones no lineales complejas entre variables; asimismo, permite modelar automáticamente datos dimensionales, además de ampliar el conjunto de variables predictoras en los modelos a uno mucho más amplio y de multidominio, lo que también permite utilizar nuevas fuentes de datos complejas distintas a numéricas, tales como texto, imágenes, etc. (Banco Interamericano de Desarrollo, 2020)

Hoy por hoy existen diferentes trabajos y estudios enfocándose en el uso de aprendizaje automático en el campo de la medicina, cada uno diferente, ya que, al ser un campo bastante

amplio, permite descubrir cosas nuevas cada día. El uso de biomarcadores y medidas antropométricas para predicción de enfermedades es un tema que hoy en día está tomando más fuerza en el momento de hablar de predicción de enfermedades. Existen varios artículos en los que mencionan que el uso de estos será el futuro que seguirán los nuevos métodos empleados para la predicción de enfermedades. Hablando específicamente de la obesidad, se dice que el IMC hoy en día no es una fuente confiable para detectarla en una persona, ya que esta solamente es una medida imperfecta de acumulación anormal o excesiva de grasa corporal. Existen biomarcadores y medidas antropométricas que ayudan a la detección, tales como la insulina, circunferencia de la cintura, entre otros. Se considera que los biomarcadores y las medidas antropométricas a la larga podrían ayudar a detectar la obesidad de una manera más rápida y menos costosa, a comparación de las resonancias magnéticas. (Nimptsch et al., 2018)

En el año 2018, en la conferencia LACCEI International Multi-Conference for Engineering, Education, and Technology "Innovation in Education and Inclusion" se present el paper "Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents", en donde se expuso que, a partir de un modelo neuro-difuso ANFIS, el método backpropagation y lógica difusa sobre atributos tales como edad, peso, estatura e IMC, se puede lograr clasificar la obesidad de niños y adolescentes varones con un error de aproximadamente 3%. (Sulla Torres et al., 2018)

Dentro del trabajo "Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Aprendizaje automático Models", se observa que los autores realizaron una comparativa de los modelos de aprendizaje automático y de regresión, tanto lineal como Lasso; observando que hay una mejor predicción con aprendizaje automático a comparación de métodos tradicionales. Se demostró, además, la importancia del entrenamiento para las predicciones, ya que mejora la exactitud de las pruebas. Para la interpretabilidad, indican que se deben considerar 3 factores: los modelos de regresión multivariable pueden parecer más interpretables, los algoritmos de aprendizaje automático ofrecen resultados parcialmente interpretables y algunos modelos de aprendizaje automático ofrecen un rendimiento e interpretabilidad superiores a la regresión lineal. Consideran también que los factores demográficos, socioeconómicos, médicos y ambientales dentro de las muestras que utilizaron fueron la causa de la prevalencia de la obesidad. Esto se pudo observar a partir de los modelos de aprendizaje empleados ya que, dependiendo de los factores demográficos y

socioeconómicos, se evidenció una variación en la obesidad de cada una de las muestras. (Scheinker et al., 2019)

El aprendizaje automático a través de árbol de decisiones puede ayudar a predecir el riesgo de desarrollar enfermedades. Los árboles de decisión identifican las relaciones entre variables y asociaciones con riesgos que pueden no ser identificados por las técnicas tradicionales de análisis epidemiológico, por lo que establecer relaciones entre diferentes polimorfismos se considera una fortaleza de este método. Además, hace hincapié en que la edad y género son determinantes al momento de predecir algún tipo de enfermedad ya que el tamaño y forma corporal de los adultos difieren notoriamente entre hombres y mujeres, además que cambian con el tiempo. Otros factores que influyen en los resultados son etnia y edad. Este trabajo coincide en que los métodos tradicionales son menos efectivos que los métodos de predicción por aprendizaje de máquina. Asimismo, se le considera como demostrativo piloto para las nuevas tecnologías de análisis de datos para la predicción de enfermedades, ya que el uso de estas permitirá crear estrategias tempranas para prevenir enfermedades, además de ser uno de los primeros trabajos que utilizan arboles de decisión con información genética. (Rodríguez-Pardo et al., 2019)

El trabajo titulado “Predicting nationwide obesity from food sales using aprendizaje automático” habla sobre que se puede predecir la obesidad en una ciudad o país determinado a partir de la cantidad que consume de ciertos productos, tales como harina, lácteos y bebidas gasificadas; para ello utilizaron como modelo árboles aleatorios y librerías como Extreme Gradient Boosting. La evaluación se hizo en productos consumidos por niños menores de dos años con datos de varios países con buena economía en base a la venta de 5 alimentos. Luego de la aplicación de RF pudieron observar que el uso de aprendizaje automático les permitió estimar la obesidad a nivel nacional a partir de las categorías de venta de cinco alimentos, lo cual fue notoriamente menos costoso que las encuestas nacionales. Consideran que utilizar métodos de aprendizaje automático es más cómodo y menos costoso que realizar encuestas, las cuales son muy caras, asimismo, refieren a que los enfoques tradicionales de regresión limitan el análisis a un pequeño conjunto de predictores e imponen suposiciones de independencia y linealidad. (Dunstan et al., 2019)

Algunos factores sociales como efectos adversos de la globalización, crecimiento de los supermercados, urbanización no planificada, sedentarismo, entre otros, desarrollan lentamente factores de riesgo conductual en las personas, que a la larga derivan a tener afecciones en la

salud. Es por ello por lo que en 2020 se realizó una revisión de varios métodos de aprendizaje automático y su ejecución utilizando datos de salud de muestra de personas, entre 20 y 60 años, disponibles en repositorios públicos relacionados con enfermedades a consecuencia del estilo de vida, tales como la obesidad. Este estudio utilizado diferentes técnicas para predecir si esa persona es obesa o no, pero, antes que nada, prepararon la data según el modelo que iban a emplear (árbol de decisiones, regresión, etc.). Con esto, pudieron demostrar qué factores son los que influyen para que una persona sea obesa o no. Asimismo, sugieren hacer un estudio similar a este, pero empleando otras técnicas, como redes neuronales. (Chatterjee et al., 2020)

La obesidad es una enfermedad cada vez más frecuente entre los jóvenes, alcanzando niveles pandémicos en la actualidad. Padecer esta enfermedad a edades tempranas supone un grave riesgo para la salud a corto y medio plazo y se relaciona con otro tipo de patologías como problemas cardiovasculares, hipertensión o diabetes, entre otras. la situación de esta enfermedad en los Estados Unidos es particularmente preocupante, de tal forma que hemos centrado nuestro estudio en la población joven de ese país, como se pudo ver reflejado en el estudio titulado “Aplicación de Técnicas de Aprendizaje automático para la Predicción de la Obesidad en Jóvenes de Estados Unidos”, en donde se realizó un análisis de las causas que dan lugar a esta enfermedad no transmisible para posteriormente desarrollar un modelo predictivo, modelo BIC4, de la obesidad mediante técnicas de aprendizaje automático. (Gutiérrez Contreras, 2022)

Un estudio realizado en India en el año 2021 propone una CNN para que a través de imágenes térmicas se pueda detectar si una persona es obesa o no, esto gracias a que detectaron que en la zona del abdomen hay una diferencia de temperatura entre ambos casos de personas. De los dos modelos que proponen, se ve un mejor resultado en el modelo personalizado, ya que se le entreno para que evalúe zonas específicas en las imágenes térmicas. Hicieron, además, un pequeño estudio o de los trabajos hasta el momento que realizaron este y concluyeron que son pocos los que existen, por lo que consideran que es un campo aun sin explorar: “La aplicación del aprendizaje profundo en la detección de la condición de obesidad a partir de imágenes térmicas no se investiga en la literatura reciente”. Asimismo, menciona que los modelos entrenados tienen una buena precisión de entrenamiento, pero mala al momento de las pruebas, y esto se debe al sobre entrenamiento. (Snekhalatha et al., 2021)

Desde otro ángulo, no solamente se ha estudiado la predicción de obesidad a partir de variables antropométricas, sino que también de valores que se obtengan de estudios en

diferentes órganos del cuerpo, tal es así el caso de un estudio que propone que, a partir de electrorretinografías se puede predecir si una persona es obesa o no. La metodología que se utilizó fueron redes neuronales y modelos de enjambre de partículas basados en redes neuronales. Este estudio demostró que la obesidad está relacionada con los resultados que se obtienen de los electrorretinograma. A través de la optimización del PSO se obtuvo mejores resultados. Tuvieron limitaciones por la cantidad de data para el modelo. (Senyer Yapici et al., 2021)

Paralelamente a considerar solamente variables o datos de la persona a evaluar, hay algunos estudios que consideran a más de una persona para poder predecir enfermedades, tal es el caso del estudio “Ranking of a wide multidomain set of predictor variables of children obesity by aprendizaje automático variable importance techniques", el cual utiliza muchas variables, además de ser uno de los pocos que incluye a los padres como variables que definen al niño, ya que, en este trabajo consideran que lo que los padres tienen como costumbres, el niño también lo hará, El objetivo de este estudio principalmente fue mostrar que hay más de una variable de la que puede depender la predicción, para demostrar esto utilizaron Random Forest y Gradient Boosting Machine. (Marcos-Pasero et al., 2021)

Se sabe que, a través del útero de la madre los bebes absorben todo lo que la madre come, por lo que en esta etapa se debe cuidar mucho la madre ya que hay evidencia que el entorno uterino puede causar una influencia permanente en la salud futura del feto y puede conducir a una mayor susceptibilidad a enfermedades más adelante. Partiendo de ese conocimiento, se desarrolló un estudio en el 2021 que se basa en ver si un niño es propenso a ser obeso o no a partir de rasgos de la madre. Proponen que este factor viene incluso desde antes de nacer, ya que puede depender de cuanto sea la glucosa de la madre durante el embarazo o de la ascendencia. Este trabajo usa valores antropométricos, tales como IMC; datos demográficos, medicamentos, diagnósticos y pruebas de laboratorio de niños y sus familias. Se evaluó el rendimiento de múltiples modelos de predicción para la obesidad infantil a los 5 y 6 años en diferentes edades del niño. A partir de esto, se pudo concluir que los predictores más influyentes al nacer son las mediciones antropométricas de los hermanos, madre y padre. A pesar de que existieron limitaciones con la data, se pudieron obtener datos bastantes preciso. (Rossman et al., 2021)

Aproximadamente 370 millones de niños y adolescentes en todo el mundo presentaron sobrepeso u obesidad en 2016. El riesgo de desarrollar comorbilidades graves depende de la

edad de inicio y la duración de la obesidad, es por eso que en el trabajo de investigación titulado "Predicting of excess body fat in children", se realiza una revisión exhaustiva a los modelos que hoy por hoy existen para detectar el exceso de grasa corporal en los niños, así como los factores de la vida temprana que predicen el exceso de grasa corporal y su desarrollo, en donde se pudo concluir que la detección temprana del exceso de grasa corporal y las intervenciones efectivas para normalizar el peso corporal en niños y adolescentes con riesgo de obesidad son cuestiones clave para prevenir el riesgo de enfermedad más adelante en el ciclo de vida. (Córdoba-Rodríguez et al., 2022)

2.3. Técnicas y Herramientas

2.3.1. Metodología para el procesamiento y análisis de datos

2.3.1.1. Metodología CRISP-DM

CRISP-DM es la abreviación de "Cross Industry Standard Process of Data Mining", esta metodología es capaz de transformar los datos en conocimiento e información para la gestión. Fue creada hace 20 años aproximadamente, a partir de la necesidad que tenían los profesionales de Minería de datos para atender los proyectos que estaban directamente relacionados con el procesamiento y análisis de un gran volumen de datos. (Espinosa-Zúñiga, 2020)

2.3.1.1.1. Fases de CRISP-DM

Esta metodología está compuesta por seis fases, las cuales dependen entre sí tanto secuencial como cíclicamente, lo que permite encontrar mejores interacciones para la mejora de resultados en las fases.

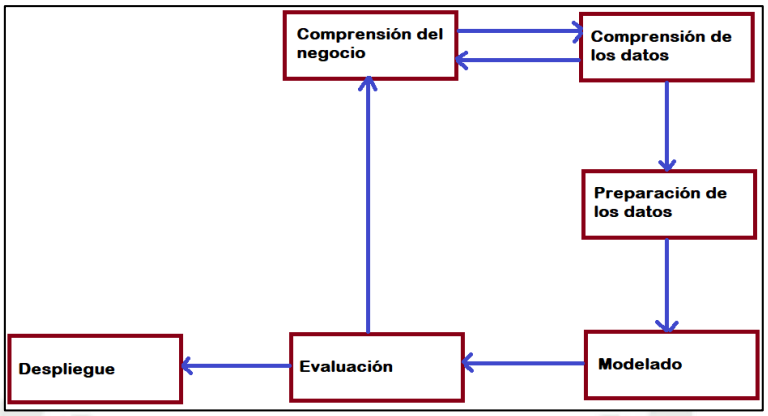


Figura 5: Fases de la metodología CRISP-DM

Fuente: (Espinosa-Zúñiga, 2020)

a) Comprensión del negocio

En esta primera fase se desarrolla una comprensión profunda del problema a resolver, identificando las necesidades y objetivos del proyecto, para luego traducirlos en objetivos técnicos y desarrollar un plan para el proyecto. Primero, se establece una medida de éxito, ya sea cualitativa o cuantitativa, Como segundo paso, se evalúa la situación actual para determinar el contexto y necesidades del problema. Finalmente, se desarrolla un plan de proyecto que considera qué pasos se deben seguir y qué procedimientos se utilizarán.

b) Comprensión de los datos

En esta segunda fase se lleva a cabo la recopilación y exploración de datos preliminares. Esta fase es crítica, ya que, si se analiza mal, se puede llegar a un aumento de tiempo y costes de proyecto. Para poder realizar el análisis se debe desarrollar una serie de pasos:

- **Recopilación de datos** iniciales y adaptación de estos a las necesidades del proyecto.
- **Descripción de los datos** obtenidos en el paso inicial de manera formal. Número de instancias (filas) y atributos (columnas), cada uno con su significado y formato de datos.
- **Exploración de datos** aplicando técnicas básicas de estadística descriptiva.

- **Validación de datos para** determinar su consistencia, número y distribución de valores nulos o fuera de rango. Esto se realiza a fin de evitar ruido en el modelado.

c) Preparación de los datos

Esta fase de preparación consiste en seleccionar, limpiar y generar conjuntos de datos correctos, organizados y listos para utilizar en la siguiente fase, los cuales corresponden al 75% del tiempo total del proyecto. Al igual que la anterior, es considerada una fase crítica, ya que los errores en los datos que se ignoren y no se resuelva en esta fase, se transferirán a la fase de modelado, lo cual ocasionará una reducción de precisión del modelo y posiblemente proporcione resultados basados en datos incorrectos.

d) Modelado

En la fase de modelado se crean diferentes tipos de modelos de conocimiento que se basan en los datos proporcionados anteriormente. Al igual que las anteriores fases, para poder obtener resultados óptimos, se deben seguir una serie de pasos:

- **Elección de algoritmo** de modelado más apropiado según el problema.
- **Generación de plan de pruebas** donde se configura los valores de los parámetros que se utilizará en el modelo, además de las métricas de evaluación.
- **Construcción de modelos**, ejecutando los algoritmos seleccionados para generar uno o más modelos y calcular métricas.
- **Evaluación del modelo** con las métricas establecidas para asegurar que se cumple con los criterios de éxito.

e) Evaluación

Esta fase se efectuará una vez el modelo obtenido sea el esperado. En caso no lo es, se debe repetir las fases anteriores hasta llegar al modelo. En esta fase se evalúa formalmente los resultados obtenidos, teniendo en cuenta los criterios de éxito preestablecidos y revisando todo el proceso con el objetivo, a fin de identificar errores.

f) Despliegue

En esta última fase se definen las estrategias de implementación, seguimiento y mantenimiento de modelo, los cuales ayudarán a observar cualquier comportamiento anormal del sistema y corregirlo de manera que no ocasione defectos. Finalmente se realiza una revisión final de todos los pasos realizados anteriormente. (Espinosa-Zúñiga, 2020)

2.3.1.2. KDD

Es una metodología en el que se combinan descubrimiento y análisis que consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente preprocesar los datos, hacer minería de datos y presentar resultados. (Timarán Pereira et al., 2016)

2.3.1.2.1. Fases de KDD

Esta metodología está compuesta por cinco fases interactivas e iterativas donde dentro de cada una se requiere la toma de decisiones a partir de la intervención del usuario

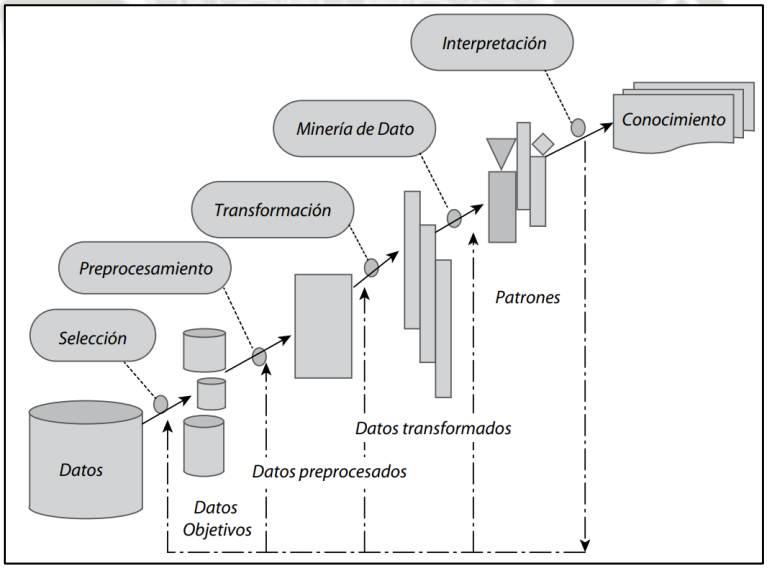


Figura 6: Etapas de metodología KDD

Fuente: (Timarán Pereira et al., 2016)

a) **Selección**

En esta fase se crea un conjunto de datos objetivo, ya sea con todo el conjunto de datos o una muestra de este. La selección en esta varía según los objetivos del negocio.

b) **Preprocesamiento y limpieza**

En esta fase se analiza la calidad de los datos, y sobre eso se aplican operaciones tales como remoción de datos ruidos, así como la selección de estrategias para el manejo de datos desconocidos, vacíos y duplicados.

c) **Transformación y reducción**

En esta fase se busca características en los datos para representarlos dependiendo del objetivo del proceso. Se utilizan métodos para reducción de dimensiones o transformación de los datos.

d) **Minería de datos**

En esta fase se busca descubrir patrones en los datos, tanto insospechados como de interés.

e) **Interpretación y evaluación**

En esta fase se interpretan los patrones descubiertos y de ser necesario se puede retornar a etapas anteriores para crear nuevas iteraciones. (Timarán Pereira et al., 2016)

2.3.1.3. SEMMA

Es definida como el proceso de selección, exploración y modelado de grandes cantidades de datos para descubrir patrones. El nombre de esta metodología es el acrónimo de las cinco fases de esta. Sample, Explore, Modify, Model, Assess. (Rodríguez Montequín et al., 2003)

2.3.1.3.1. Fases de SEMMA

Esta metodología está compuesta por cinco fases interactivas e iterativas donde dentro de cada una se requiere la toma de decisiones a partir de la intervención del usuario.

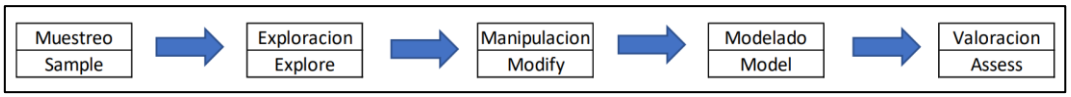


Figura 7: Metodología SEMMA

Fuente: (Rodríguez Montequín et al., 2003)

a) **Muestreo**

En esta fase se selecciona la muestra del problema en estudio. Esta muestra es sobre la que se aplicará el análisis.

b) **Exploración**

En esta fase se realiza la exploración de la muestra a fin de simplificar el problema lo más posible para poder optimizar la eficiencia del modelo.

c) **Manipulación**

En esta fase se realiza la estandarización de los datos a partir de la exploración previa.

d) **Modelado**

En esta fase se busca la relación entre variables explicativas y objeto. Se puede utilizar métodos estadísticos, técnicas basadas en datos (aprendizaje automático), etc.

e) **Valoración**

En esta fase se realiza el análisis de bondad del modelo, realizando una comparación con estudios previos. (Rodríguez Montequín et al., 2003)

2.3.4.1. Lenguaje Python

Considerado como el lenguaje de programación más popular del mundo según el índice PYPL, Python es un lenguaje de programación multiplataforma, de código abierto, multiparadigma y multinivel, que es utilizado para desarrollo web, creación de software nativo e híbrido y procesamiento de datos. (Carbonnelle, 2022)

2.3.4.1.1. Ventajas de Python

Además de ser considerado como un lenguaje de fácil aprendizaje, por su sintaxis sencilla de interpretar, estas son algunas ventajas que Python ofrece:

- Poderoso y versátil, por lo que puede ser utilizado en cualquier área.
- Fácil manejo de la estructura de datos.
- Depuración más veloz.
- Permite utilizarlo en programación orientada a objetos, estructural o funcional.
- Permite empaquetar código.
- Mejor gestión de memoria.
- Fácil interpretación.
- Permite la automatización de tareas y procesos. (Challenger-Pérez et al., 2013)

2.3.2. RapidMiner

RapidMiner es una plataforma cruzada que combina minería de datos con minería de texto, aprendizaje automático, inteligencia y análisis comerciales. Cuenta con más de 500 operadores que ayudan a tener diferentes enfoques sobre los datos, además, es capaz de integrarse con otras aplicaciones tales como Weka y R. Este software impulsa los procesos comerciales y ayuda a aprovechar al máximo sus datos mediante la manipulación de variables y gráficos numéricos y no numéricos en la respuesta rápida y el soporte de múltiples fuentes que proporciona. Gracias a esta herramienta se puede tomar medidas rápidas para acelerar la adquisición de datos y permite a los analistas de datos crear nuevos procesos de extracción de datos, configurar análisis predictivos, etc. (Aranda, 2019)

2.3.2.1. Ventajas de RapidMiner

RapidMiner es una herramienta sencilla que cuenta con muchas ventajas, algunas de ellas se enumeran a continuación:

- Permite ser usada desde diferentes perfiles profesionales.
- Acelera el proceso de desarrollo respecto a herramientas habituales.
- Permite ser usada en entornos de Big data y Small data.

- Cuenta con una interfaz muy intuitiva y ágil.
- Fácil creación de batches o pipelines automatizados.
- Transacción con terceras herramientas a tiempo real.
- Permite incluir los resultados obtenidos en otras herramientas.
- Cuenta con un gran listado de capacidades para las fases de la limpieza de datos, modelado y publicación. (Aranda, 2019)

2.3.3. Matriz de confusión

La matriz de confusión es la base para las métricas de evaluación de modelos de clasificación, no es considerada como una medida de rendimiento, pero sí una métrica muy intuitiva y sencilla con la que se puede hallar la exactitud y precisión de un modelo. Esta herramienta consiste en una tabla donde se tienen dos dimensiones: Actual, que es lo que se espera; y Predicción, que es lo que el modelo devuelve al ser ejecutado. Al mismo tiempo, estas dos dimensiones se subdividen en: Positivo, que son los aciertos; y Negativo, que son los errores del modelo. (Flores Jara, 2015)

Clasificación		Predicción		Total observado
		Positivo	Negativo	
Actual	Positivo	TP	FN	TP+FN
	Negativo	FP	TN	FP+TN
Total predecido		TP+FP	FN+TN	N

Figura 8: Matriz de confusión

Fuente: (Flores Jara, 2015)

La matriz de confusión permite evaluar 4 valores del modelo: Los verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos:

a) Verdaderos Positivos (True Positives – TP)

Son los casos en los cuales los datos reales son verdaderos y la predicción además es verdadera. En otras palabras, es el número de predicciones positivas correctas que hizo el clasificador.

b) Verdaderos Negativos (True Negatives – TN)

Son los casos en los cuales los datos reales con falsos y la predicción además es falsa. En otras palabras, es el número de predicciones negativas correctas que hizo el clasificador.

c) Falsos Positivos (False Positives – FP)

Son los casos en que los datos reales indica que es falso y la predicción devuelve que es verdadero. En otras palabras, es el número de predicciones positivas incorrectas que hizo el clasificador.

d) Falsos Negativos (False Negatives – FN)

Son los casos en que los datos reales indica que es verdadero y el pronóstico es falso. En otras palabras, es el número de predicciones negativas incorrectas que hizo el clasificador. (Arias Zuluaga, 2020)

2.3.4. Métricas para modelos de aprendizaje automático

La evaluación del algoritmo de clasificación supervisada tiene como objetivo medir el poder predictivo de los modelos durante el entrenamiento al comparar las clasificaciones hechas en pruebas con las etiquetas obtenidas

2.3.4.1. Accuracy

Es el porcentaje o número relativo de ejemplos correctamente calificados, en otras palabras, de predicciones correctas. Es una métrica para evaluar modelos de clasificación a partir de la fracción de predicciones que el modelo acertó, mas no es una técnica que funcione correctamente cuando se cuenta con clases desbalanceadas, es decir, cuando se utiliza un dataset con una disparidad significativa entre las etiquetas positivas y negativas. (Flores Jara, 2015)

Esta métrica tiene dos fórmulas según sea el caso:

Clasificación

$$Accuracy = \frac{Número\ de\ predicciones\ correctas}{Número\ total\ de\ predicciones} \quad (1)$$

Clasificación binaria

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

2.3.4.2. Classification Error

Es una métrica que se mide a partir del número de ejemplos mal clasificados, es decir, es el porcentaje de predicciones erróneas. (RapidMiner, 2022)

La fórmula con la que se representa el error de clasificación es el siguiente:

$$classification\_error = 1 - (TP + TN) \quad (3)$$

2.3.4.3. Kappa (k)

También llamado estadísticas Kappa para la clasificación, permite medir la concordancia entre los resultados de variables cualitativas de un modelo, lo cual se ve reflejado en el índice k, el cual a partir de la matriz de confusión permite evaluar si la clasificación que realizó el modelo es similar con la clasificación previa del dataset. Es considerada como una medida más robusta ya que tiene en cuenta las predicciones que se realizan de por casualidad. (Flores Jara, 2015)

La fórmula que representa el índice k es el siguiente:

$$k = \frac{P_0 - P_e}{1 - P_e} \tag{4}$$

Donde:

$$P_0 = \frac{TP + TN}{N} \tag{5}$$

$$P_e = \frac{(TP + FP) * (TP + FN) + (FN + TN) * (FP + TN)}{N^2} \tag{6}$$

Teniendo en cuenta que:

$$0 \leq k \leq 1 \tag{7}$$

Y que la escala con su concordancia es la siguiente:

Tabla 1: Equivalencia de valores de métrica Kappa

Valor de k	Concordancia
< 0.20	Pobre
$0.21 - 0.40$	Débil
$0.41 - 0.60$	Moderada
$0.61 - 0.80$	Buena
$0.81 - 1.00$	Muy buena

Fuente: (CERDA & VILLARROEL, 2008)

2.3.4.4. Spearman Rho (r_s)

Es un coeficiente de correlación basado en rangos, conocido como correlación no paramétrica, que se utiliza para medir la asociación entre las dos variables cuando la asociación es no lineal. Esta métrica evalúa qué tan bien se puede describir la relación entre dos variables (ya sean lineales o no) usando una función monótona. (Mondragón Barrera, 2014)

La fórmula para hallar este valor es la siguiente:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{8}$$

Donde:

- r_s es el coeficiente de correlación Sperman Rho.
- d_i es la diferencia en los rangos de los valores de las variables.
- n es el número de observaciones.

La escala con su concordancia es la siguiente:

Tabla 2: Equivalencia de valores de métrica Spearman Rho

Rango	Relación
-1.00 a -0.91	Correlación negativa perfecta
-0.90 a -0.76	Correlación negativa muy fuerte
-0.75 a -0.51	Correlación negativa considerable
-0.50 a -0.11	Correlación negativa media
-0.10 a -0.01	Correlación negativa débil
0.00	No existe correlación
+0.01 a +0.10	Correlación positiva débil
+0.11 a +0.50	Correlación positiva media
+0.51 a +0.75	Correlación positiva considerable
+0.76 a +0.90	Correlación positiva muy fuerte
+0.91 a +1.00	Correlación positiva perfecta

Fuente: (Mondragón Barrera, 2014)

2.3.4.5. Absolute Error (*MAE*)

Es el promedio de las diferencias absolutas entre los valores actuales y predichos. Esta toma la media de este error de cada muestra en el conjunto de datos y da la salida. (Willmot & Matsuura, 2006)

La fórmula para hallar este valor es la siguiente:

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (9)$$

Donde:

- y_i es el valor predicho.
- x_i es el valor real.
- n es el número de valores predichos.

Dentro de esta métrica existen dos conceptos importantes:

- **Underfitting:**

  Este escenario se da cuando el modelo coincide casi exactamente con los datos de entrenamiento, pero tiene un rendimiento deficiente cuando se enfrenta a nuevos.

- **Overfitting:**

  Este escenario se da cuando el modelo no logra capturar patrones e información importantes en los datos, lo que hace que el modelo tenga un rendimiento deficiente en los datos de entrenamiento. (Tripathi, 2020)

2.3.4.6. Relative Error (MRE)

Es una métrica que permite medir el rendimiento de modelos predictivos, la cual consiste en el promedio de la desviación absoluta de la predicción dividida por el valor verdadero. (De la Fuente Carmona, 2022)

La fórmula para hallar este valor es la siguiente:

$$MRE = mean\left(\frac{|y_i - \hat{y}_i|}{y_i}\right) \quad (10)$$

Donde:

- y_i es el valor predicho.
- $\hat{y}_i$ es el valor real.

2.3.4.7. Root Mean Squared Error ($RMSE$)

Es la métrica más utilizada en modelos de regresión, particularmente en modelos de aprendizaje automático supervisado. Ayuda a visualizar el ajuste absoluto que tiene el modelo

con los datos, es decir, cuán cerca están los puntos de los datos de dataset con los valores predichos por el modelo. A menor valor se obtenga de la métrica, mejor será el ajuste. Debe ser considerado un criterio importante si el modelo es de predicción. (De la Fuente Carmona, 2022)

La fórmula para hallar este valor es la siguiente:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}} \tag{11}$$

Donde:

- y_i es el valor predicho.
- x_i es el valor real.
- n es el número de valores predichos.

2.3.4.8. Correlation (r_{xy})

El coeficiente de correlación muestra la intensidad con la que el valor cuantitativo predicho y el original están relacionados linealmente. Este coeficiente es sensible a valores atípicos (outliers) (De la Fuente Carmona, 2022).

La fórmula para hallar este valor es la siguiente:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(x_i - \bar{y})^2}} \tag{12}$$

Donde:

- y_i es el valor predicho.
- x_i es el valor real esperado.
- $\bar{y}$ es el promedio de los valores predichos.
- $\bar{x}$ es el promedio de los valores predichos.
- n es el número de valores predichos.

Este coeficiente puede clasificar al modelo de 5 formas distintas:

- Si $r = 1$, entonces es una asociación perfecta positiva, con lo que ambas varían en el mismo sentido.

- Si $r = -1$, entonces es una asociación perfecta negativa, con lo que ambas varían en sentido opuesto.

- Si $r = 0$, entonces no existe una asociación entre ambas variables.

- Si $0 < r < 1$, entonces es una asociación positiva, pero el grado de asociación va a depender de cuán cerca se encuentre a 1: Será mayor si está más cerca de 1 y menor si se acerca a 0.

Si $-1 < r < 0$, entonces es una asociación negativa, pero el grado de asociación va a depender de cuán cerca se encuentre a 1: Será mayor si está más cerca de -1 y menor si se acerca a 0.

2.4. Consideraciones finales

En este capítulo se presentó algunas definiciones relacionadas a la obesidad, aprendizaje automático e inteligencia artificial; asimismo de las herramientas y técnicas que se utilizarán a lo largo del proyecto. Las técnicas de inteligencia artificial están siendo utilizadas con mayor frecuencia en diferentes áreas de la vida cotidiana del ser humano, puntualmente, se está viendo una mayor investigación en el área de la salud, en la cual se está buscando mayormente predecir enfermedades a partir de datos, ya sean numéricos, características o imágenes.

CAPÍTULO III

3. ANÁLISIS, CONSTRUCCIÓN Y EVALUACIÓN DE LAS TÉCNICAS DE APRENDIZAJE AUTOMÁTICO SUPERVISADO

En este tercer capítulo del documento se observará la parte más práctica, es por ello por lo que se realizará un análisis previo de las metodologías para el proceso de minería de datos: KDD, CRISP-DM y SEMMA, el cual se detalla en la siguiente tabla 03:

Tabla 3: Análisis previo de las metodologías para el proceso de minería de datos

	Fases	**Características**	**Etapas iterativas**	**Evaluación del resultado**
KDD	5 fases: Pre KDD * Selección * Preprocesamiento * Transformación * Minería de Datos * Evaluación e Implantación. Post KDD	Es un proceso iterativo, interactivo y lineal. No describe las tareas y actividades específicas que se deben realizar en cada una de sus etapas	Si	Objetivos del proyecto
CRISP-DM	6 fases: * Comprensión del negocio * Comprensión de los datos * Preparación de Datos * Modelamiento * Evaluación * Despliegue	Hace énfasis en los detalles de cada fase, es decir, cada etapa se divide en diferentes tareas y actividades; asimismo, hace que los proyectos, grandes y pequeños, de minería de datos sean más rápidos, económicos, fiables y manejables.	Si	Modelo Objetivos del proyecto

SEMMA	5 fases: * Muestra * Exploración * Modificación * Modelo * Evaluación	Permite aplicar estadística exploratoria y técnicas de visualización de manera fácil, así como la selección y transformación de variables más significativa, con el objetivo de crear modelos para predecir resultados y evaluarlos de manera que sirva de apoyo para la toma de decisiones	No	Modelo

Fuente: (Aquino, 2016)

Luego de realizar la comparación de las herramientas disponibles, se decide utilizar la metodología CRISP-DM ya que realiza la evaluación a nivel de modelo y objetivo del negocio; permite la disminución de costos y la reducción de tiempo en proyectos. Asimismo, consta de 6 fases no unidireccionales que son de gran utilidad ya que permite crear modelos que se adaptan a las necesidades que se tiene.

3.1. Comprensión de los requisitos del negocio

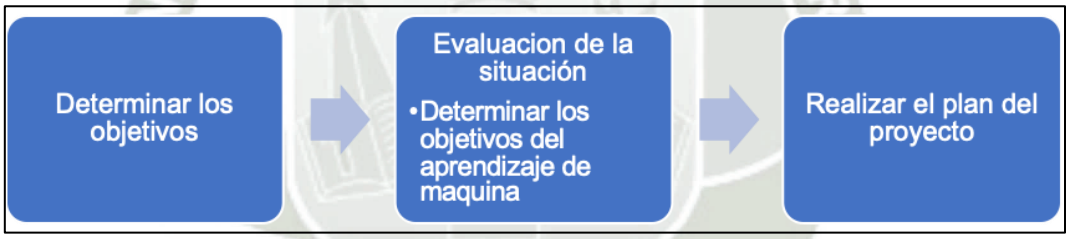


Figura 9: Fase 1: Comprensión de los requisitos del negocio

Fuente: Propia

Dentro de esta fase se determinarán los objetivos que se quieren alcanzar con el desarrollo del proyecto, luego se evaluará la situación en la que se encuentra, determinando, además, los objetivos del aprendizaje de máquina, para finalmente definir cuál es el plan que se seguirá para llegar al objetivo planteado.

3.1.1. Determinar los objetivos

Las técnicas de aprendizaje automático que se aplicarán en el proyecto tienen como objetivo realizar predicciones de obesidad en adolescentes con ayuda de datos y medidas antropométricas, tales como edad, sexo, talla y peso; a fin de poder mejorar el diagnostico de obesidad en adolescentes a futuro.

Contexto

Al principio del proyecto se cuenta con una base de datos de personas entre 4 y 21 años, a quienes se les ha diagnosticado o descartado obesidad a partir del valor de IMC a partir de sus medidas antropométricas.

En la Tabla 4 se visualizará las ventajas de las medidas antropométricas y los biomarcadores que se tuvo en cuenta para elegir utilizar medidas antropométricas.

Tabla 4: Ventajas de medidas antropométricas y biomarcadores

	Medidas antropométricas	**Biomarcadores**
Ventajas	• Métodos no invasivos. • Accesibles y fáciles de realizar.	• Determinación en muestras que tienen un carácter invasivo mínimo. • Fácil aplicabilidad.
Desventajas	• Poca especificidad. • Requieren de personal capacitado y con experiencia.	• Sus resultados pueden ser influidos por condiciones comórbidas. • Requieren de una interpretación crítica de los mismos.

Fuente: (Cequera & García de León Méndez, 2014), (Castillo Hernández & Zenteno Cuevas, 2004)

A partir de las ventajas y desventajas mencionadas anteriormente, se decidió plantear el uso de medidas antropométricas ya que se obtienen a partir de métodos no invasivos y cuenta con medidas sencillas, como peso y talla, que no se necesita de un especialista en el tema.

En la Tabla 5 se visualizará las etapas de vida del ser humano y en qué rango de edades es que ocurren.

Tabla 5: Etapas en la vida del ser humano

Etapa		Rango de edad
Recién nacido		De 0 a 12 meses
Etapa de la infancia	Primera infancia o infancia temprana	De 1 a 3 años
	Segunda infancia o etapa preescolar	De 3 a 6 años
	Tercera infancia	De 6 a 9 años
Madurez infantil		De 9 a 12 años
Preadolescencia		De 12 a 14 años
Adolescencia		De 14 a 18 años
Juventud		De 18 a 25 años
Adultez		De 25 a 60 años
Vejez		Desde los 60 años

Fuente: (Ministerio de Educación, 2013)

Se decidió centralizar la predicción de obesidad en adolescentes ya que, del 100% de datos que se tiene al inicio de este trabajo, más del 60% son registros de adolescentes, lo cual es un dato importante al momento de entrenar el modelo, ya que se puede garantizar con ello que el sistema tendrá variedad de datos de dónde aprender en ese rango de edad.

Por otro lado, en la Tabla 6 se muestra la interpretación el IMC que será útil al momento de analizar los datos. Dentro de esta tabla se pueden ver los diferentes grados de obesidad que existen.

Tabla 6: Interpretación de IMC

Interpretación	Valor en atributo IMC
Bajo peso	$IMC < 18.5$
Peso normal	$IMC > 18.5\ y\ < 24.9$
Sobrepeso grado 1	$IMC > 25\ y\ < 26.9$
Sobrepeso grado 2	$IMC > 27\ y\ < 29.9$
Obesidad tipo 1	$IMC > 30\ y\ < 34.9$
Obesidad tipo 2	$IMC > 35\ y\ < 39.9$
Obesidad mórbida o tipo 3	$IMC > 40\ y\ < 49.9$
Obesidad extrema o tipo 4	$IMC \geq 50$

Fuente (Suárez-Carmona & Sánchez-Oliver, 2014)

Existen trabajos en los cuales se ha logrado predecir obesidad mediante redes bayesianas o árboles de decisión, mas no por redes neuronales, que es lo que se plantea realizar en este proyecto, además de tener en cuenta de las ventajas que ofrece el modelo: Alta

tolerancia a fallos, facilidad al momento de manejar características complejas en los datos, adaptación de pesos a manera que se utiliza, así como reconocimiento de patrones en la etapa de producción.

Objetivos del negocio

El objetivo principal del proyecto, como ya se ha mencionado anteriormente, es lograr predecir si un adolescente presenta obesidad o no a partir de datos básicos como sexo, estatura y peso.

Esta predicción resulta útil ya que facilitará el diagnóstico de obesidad en adolescentes de una manera más sencilla y rápida.

Criterios de éxito del negocio:

Desde el punto de vista del negocio se establece como criterio de éxito predecir si un adolescente tiene obesidad en base a datos como sexo, peso y estatura con un elevado porcentaje de fiabilidad, de tal forma que se puedan dar consejos útiles a los adolescentes acerca de si deben seguir un mejor estilo de vida, visitar a un especialista de la salud o se si encuentran bien de salud.

3.1.2. Evaluación de la situación

Se cuenta con las medidas antropométricas recolectadas en un registro de 3068 personas entre 4 y 21 años, de los cuales se cuenta con la edad, sexo, peso, estatura e IMC de cada uno de ellos. Esta cantidad de datos con los que se cuenta se puede decir que se tienen el potencial para implementar el modelo planteado.

Hoy en día se tiene el problema de la sobrepeso u obesidad infantil se encuentra dentro de todos los estatus sociales en el Perú. Hablando a nivel geográfico, los departamentos del Perú que tienen mayor porcentaje de personas con obesidad son Tacna (36.5%), Ica (31,9%), Moquegua (31,7%), Madre de Dios (29,3%), Región Lima (28.8%), Provincia Constitucional del Callao (26,8%). Actualmente, las regiones con personas con sobrepeso son Lima Metropolitana que es de 64.7%, Moquegua (40.9%), Tumbes (40.1%), Arequipa (39,7%), La Libertad y Madre de Dios (ambos con 39.5%). (Instituto Nacional de Salud, 2020)

Estas estadísticas significan un problema, ya que al no ser tratada a tiempo la obesidad puede desencadenar problemas de salud, tales como diabetes, problemas cardíacos, nivel alto de colesterol y triglicéridos en la sangre, problemas cardiovasculares, osteoartritis, apnea del sueño, mala atención y problemas en el trabajo, cálculos biliares y problemas del hígado. El sobrepeso y la obesidad también aumentan el riesgo de fallecer de cáncer.

3.1.2.1. Elección del objetivo del aprendizaje de máquina

Antes de plantear los objetivos del modelo, en la Tabla 7 se analizará las ventajas de los dos grandes tipos de modelos que existen: Aprendizaje automático y aprendizaje profundo, a fin de escoger qué tipo de modelo es el que se ajusta al objetivo planteado.

Tabla 7: Ventajas de Aprendizaje automático y Aprendizaje profundo

Aprendizaje automático	Aprendizaje profundo
• Requiere una cantidad no tan grande de datos • Utiliza algoritmos para analizar, aprender y generar resultados • Utiliza datos estructurados y etiquetados • Es necesario darle reglas • Requiere mayor intervención humana • Simples y no requieren equipos especializados • Menor tiempo de identificación y clasificación • Base de datos manejable	• Requiere muchos datos iniciales • Estructura los algoritmos en capas para aprender y generar resultados • Utiliza datos no estructurados y extrae características automática e independientemente • Genera sus propias reglas • No requiere mucha intervención humana, logra autonomía • Modelos complejos y requiere equipos especializados (robustos y potentes) • Mayor tiempo de identificación y clasificación • Base de datos no manejable (millones de datos)

Fuente: (Samaniego, 2021)

Se escogió utilizar **modelos de aprendizaje automático**, ya que según el dataset que se tiene, se debía realizar un procesamiento y selección de características manual previo antes de entrenar el modelo. Además, a través del modelo se buscaba analizar los datos procesados, aprender y tomar decisiones (predecir) en base a lo aprendido. Se escogió aprendizaje automático, también, por la simplicidad que se tiene para crear los modelos.

En esta etapa se optará por utilizar una técnica que ayudará a mejorar el aprendizaje de máquina a fin de que pueda predecir la obesidad de adolescentes a partir de medidas antropométricas.

Los objetivos en términos de aprendizaje de máquina son:

a) Predecir la obesidad en adolescentes a partir de medidas antropométricas.

b) Desarrollar una red neuronal capaz de aprender a partir de datos ingresados y devolver si el adolescente tiene obesidad o no.

Teniendo en este punto claro la problemática y los objetivos principales y secundarios, se puede desarrollar la siguiente fase de la metodología CRISP-DM.

Criterios de éxito del aprendizaje de máquina:

Desde el punto de vista del aprendizaje de máquina se establece como criterio de éxito la posibilidad de realizar predicciones de obesidad de adolescentes con un alto porcentaje de precisión, exactamente con el 85% de precisión.

El grado de fiabilidad lo determinará el área de aprendizaje automático que se emplee al momento de desarrollar el proyecto, por lo que este tema se volverá a mencionar en el punto 3.5. Evaluación.

3.1.3. Realizar el plan de proyecto

El proyecto se dividirá en las siguientes fases a fin de garantizar la organización y tiempo de este.

Tabla 8: Fases del proyecto

Fases	Tiempo de ejecución
Fase 1: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar el uso de ellos.	06 semanas
Fase 2: Elección de las técnicas de modelado y ejecución de estas sobre los datos.	06 semanas
Fase 3: Análisis de los resultados obtenidos en la fase anterior, si fuera necesario repetir la fase 2.	12 semanas
Fase 4: Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos.	04 semanas
Fase 5: Presentación de los resultados finales.	06 semanas

Fuente: Propia

En paralelo a la realización de cada una de las fases se irá construyendo el glosario de terminología de aprendizaje de máquina. (ver ANEXO A: GLOSARIO DE TERMINOLOGÍAS DE APRENDIZAJE DE MÁQUINA)

Evaluación previa de herramientas y técnicas

La técnica que se va a emplear para solucionar el problema planteado será el de redes neuronales. Tal como se explicó en el punto 2.1.4, las redes neuronales son un área del aprendizaje automático que trata de imitar el funcionamiento de las neuronas del cerebro humano. Se basa en que, dados ciertos parámetros, a partir de una combinación, se pueden predecir ciertos resultados. Al necesitar parámetros de entrada, el primer paso será estandarizar y limpiar los datos en la fase de preparación a fin de minimizar el porcentaje de error que se pueda tener en el momento del entrenamiento. Según las necesidades del sistema, se implementará cada una de las capas y nodos a la red, a fin de que se pueda garantizar una tasa de precisión no menor del 85%.

Por otro lado, a lo largo de este proyecto se utilizarán diferentes herramientas en las distintas fases de la metodología, desde el análisis hasta la evaluación, las cuales se mencionan a continuación:

- RapidMiner

  RapidMiner es una plataforma cruzada que combina minería de datos con minería de texto, aprendizaje automático, inteligencia y análisis comerciales. Esta herramienta

permitirá analizar la data a fin de definir el subcampo de aprendizaje automático que se utilizará.

- Python

  Python es un lenguaje de programación multiplataforma, de código abierto, multiparadigma y multinivel, que es utilizado para desarrollo web, creación de software nativo e híbrido y procesamiento de datos. Se utilizará este lenguaje en la fase de implementación, en donde se desarrollará el modelo. Cuenta, además, con librerías que permitirán el fácil desarrollo de esta.

3.2. Comprensión de los datos

En esta segunda fase se realizará la recolección, análisis y verificación de datos con el propósito de conocerlos para que en la fase 3 sean preparados de la mejor manera. Esta fase comienza con la recopilación de datos, seguido de la descripción formal de los mismos, así como su exploración y verificación. El propósito de estas actividades es poder familiarizarse con los datos, identificando el grado de calidad que tienen y descubriendo conocimientos preliminares necesarios, así como subconjuntos que puedan ayudar para llegar al objetivo.

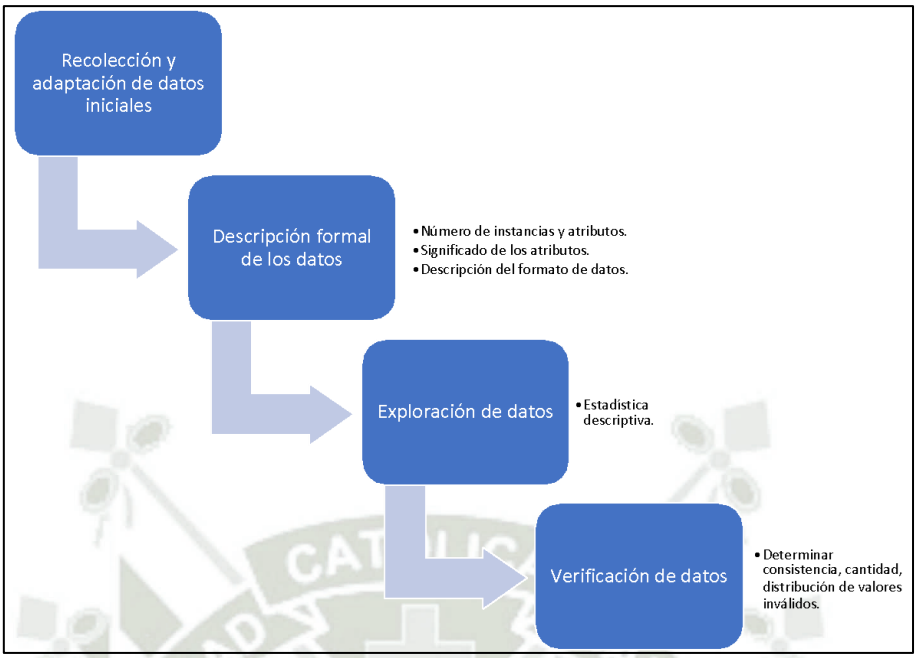


Figura 10: Fase 2: Comprensión de los datos

Fuente: Propia

3.2.1. Recolección y adaptación de datos iniciales

Los datos utilizados fueron recopilados en el año 2018 en la provincia de Arequipa, región Arequipa. Se cuenta con datos de 3068 personas entre 4 y 20 años, de las cuales se la siguiente información en un archivo Excel: edad, sexo, peso, estatura en centímetros y estatura en metros. En la siguiente tabla se muestran el formato en el que se tiene inicialmente los datos.

Tabla 9: Hoja de cálculo - Vista preliminar de dataset

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
5.96	1	18	114	1.14
6.75	1	20	118	1.18
7.95	1	24	132.5	1.33

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.52	mujeres	21.2	128.5	1.29
6.96	mujeres	19	116	1.16
7.15	mujeres	19	122	1.22
7.17	mujeres	19	118	1.18

Fuente: Propia

Los datos vienen en diferentes formatos que más adelante se tendrán que estandarizar y adaptar a fin de que el sistema pueda evaluarlos y realizar una predicción más precisa. En esta fase únicamente se harán las adaptaciones y estandarizaciones para poder evaluar el estado del dataset, así como la pureza de esta.

Adaptación de datos y dataset

En el dataset que se tiene para el preprocesamiento, se agregó tres atributos adicionales, las cuales ayudarán a llegar al objetivo:

- Un atributo "IMC", la cual será el resultado de dividir el peso en kilogramos con la altura en metros elevada al cuadrado $\left(\frac{peso\ en\ kg}{(altura\ en\ m)^2}\right)$. La cual se podrá interpretar de la siguiente manera:

Tabla 10: Interpretación de atributo IMC

Interpretación	Valor en atributo IMC
Una persona tiene bajo peso	$IMC < 18.5$
Una persona tiene peso normal	$IMC > 18.5\ y\ < 24.9$
Una persona tiene sobrepeso	$IMC > 25\ y\ < 29.9$
Una persona tiene obesidad tipo 1	$IMC > 30\ y\ < 34.9$
Una persona tiene obesidad tipo 2	$IMC > 35\ y\ < 39.9$
Una persona tiene obesidad tipo 3	$IMC \geq 40$

Fuente: (Suárez-Carmona & Sánchez-Oliver, 2014)

- Un atributo "Índice", la cual se determinará según el resultado del atributo "IMC".

Tabla 11: Interpretación de atributo Índice

Valor en atributo IMC	Valor en atributo "Índice"
< 18.5	0
$> 18.5\ y\ < 24.9$	1
$> 25\ y\ < 29.9$	2
$> 30\ y\ < 34.9$	3
$> 35\ y\ < 39.9$	4
≥ 40	5

Fuente: Propia

- Un atributo "Riesgo", siguiendo la lógica expuesta en la Tabla 12.

Tabla 12: Interpretación de atributo Riesgo

Valor en atributo Índice	Valor en atributo Riesgo
0,1	0
2,3,4,5	1

Fuente: Propia

Quedando el dataset de la siguiente forma:

Tabla 13: Hoja de cálculo - Vista previa de dataset adaptado

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC	Índice	Riesgo
6.66	1	20.5	122.5	1.23	13.550136	0	0
6.66	1	20.1	120.3	1.2	13.958333	0	0
5.95	1	20.1	120	1.2	13.958333	0	0
5.11	1	20	119.2	1.19	14.123296	0	0
5.96	1	18	114	1.14	13.850416	0	0
6.75	1	20	118	1.18	14.363689	0	0
7.95	1	24	132.5	1.33	13.567754	0	0
6.52	mujeres	21.2	128.5	1.29	12.739619	0	0
6.97	mujeres	21.7	126.1	1.26	13.668430	0	0
6.4	mujeres	19.6	118.3	1.18	14.076415	0	0
6.52	mujeres	18.5	115.2	1.15	13.988658	0	0
6.96	mujeres	19	116	1.16	14.120095	0	0
7.15	mujeres	19	122	1.22	12.765386	0	0
7.17	mujeres	19	118	1.18	13.645504	0	0

Fuente: Propia

En cuanto a la transformación de datos, dado que el atributo sexo contiene caracteres alfanuméricos, se ha codificado para que, a cada género, masculino y femenino, se le asigne un valor numérico, 1 y 2 respectivamente, para facilitar su integración con el algoritmo.

Tabla 14: Hoja de cálculo - Vista previa de dataset transformado

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC	Índice	Riesgo
6.52	2	21.2	128.5	1.29	12.739619	0	0
6.97	2	21.7	126.1	1.26	13.668430	0	0
6.4	2	19.6	118.3	1.18	14.076415	0	0
6.52	2	18.5	115.2	1.15	13.988658	0	0
6.96	2	19	116	1.16	14.120095	0	0
7.15	2	19	122	1.22	12.765386	0	0
7.17	2	19	118	1.18	13.645504	0	0

Fuente: Propia

Finalmente, luego de todas las transformaciones y adaptaciones, se ha procedido a unificar la información en una sola hoja de cálculo, ya que, originalmente la información se encontraba en dos hojas.

Tabla 15: Hoja de cálculo - Vista previa de dataset unificado

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC	Índice	Riesgo
6.66	1	20.5	122.5	1.23	13.550136	0	0
6.66	1	20.1	120.3	1.2	13.958333	0	0
7.95	1	24	132.5	1.33	13.567754	0	0
6.52	2	21.2	128.5	1.29	12.739619	0	0
6.97	2	21.7	126.1	1.26	13.668430	0	0
6.4	2	19.6	118.3	1.18	14.076415	0	0

Fuente: Propia

A continuación, se listan los datos con los que se cuentan actualmente:

a) Edad

Edad de la persona de la que se recolectó los datos. Es un valor numérico.

b) Sexo

Sexo de la persona de la que se recolectó los datos. Actualmente es un valor numérico, más inicialmente era un valor mixto, es decir, en algunos casos alfanumérico y otros numérico.

c) Peso

Peso en kilogramos de la persona de la que se recolectó los datos. Es un valor numérico.

d) Estatura_cm

Estatura en centímetros de la persona de la que se recolectó los datos. Es un valor numérico.

e) Estatu_metr

Estatura en metros de la persona de la que se recolectó los datos. Es un valor numérico.

f) IMC

índice de masa corporal (IMC) de la persona de la que se recolectó los datos. Este valor fue generado luego de la recogida de datos. Es un valor numérico.

g) Índice

Índice de obesidad según IMC de la persona de la que se recolectó los datos. Este valor fue generado luego de la recogida de datos. Es un valor numérico.

h) Riesgo

Riesgo de obesidad de la persona de la que se recolectó los datos. Este valor fue generado luego de la recogida de datos. Es un valor numérico.

Los atributos específicos que serán útiles al momento del entrenamiento del modelo son:

a) Edad

b) Sexo

c) Peso

d) Estatu_metr

e) Riesgo

Como se mencionó anteriormente, se tuvo que hacer una transformación de datos en el campo “sexo”, a fin de unificar los valores ya que se encontraban en tipo numérico y alfanumérico, de tal forma que se facilite el entrenamiento del modelo y se evite ambigüedades.

3.2.2. Descripción formal de los datos

Los datos utilizados actualmente se encuentran en un libro Excel y fueron recopilados en el año 2018 en la provincia de Arequipa, región Arequipa. Se cuenta con datos de 3068 personas: 1733 hombres y 1335 mujeres entre 4 y 20 años, de las cuales se la siguiente información: edad, sexo, peso, estatura en centímetros y estatura en metros.

3.2.2.1. Número de instancias y atributos

Para el entrenamiento y pruebas del algoritmo se cuenta con 3068 instancias con 8 atributos: 5 originales y 3 generados:

a) Edad

b) Sexo

c) Peso

d) Estatura_cm

e) Estatu_metr

f) IMC

g) Índice

h) Riesgo

3.2.2.2. Significado de los atributos

Como se mencionó en el punto 3.2.2.1, se cuentan con 8 atributos: 5 originales y tres generados, los cuales se detallará a continuación:

Tabla 16: Descripción formal de los atributos

Tipo de atributo	Nombre de atributo	Tipo de dato	Descripción
Original	Edad	Decimal	Edad de la persona
Original	Sexo	Entero	Sexo de la persona
Original	Peso	Decimal	Peso en kilogramos de la persona
Original	Estatura_cm	Decimal	Estatura en centímetros de la persona
Original	Estatu_metr	Decimal	Estatura en metros de la persona
Generado	IMC	Decimal	IMC de la persona. Este campo es generado a partir de los atributos **Peso** y **Estatu_metr**, siguiendo la fórmula para hallar el IMC: $\left(\frac{peso\ en\ kg}{(altura\ en\ m)^2}\right)$
Generado	Índice	Entero	Codificación del IMC de la persona. En el punto 3.2. se explica la equivalencia a detalle: • **0** = < 18.5 • **1** = ≥ 18.5 y < 24.9 • **2** = ≥ 25 y < 29.9 • **3** = ≥ 30 y < 34.9 • **4** = ≥ 35 y < 39.9 • **5** = ≥ 40
Generado	Riesgo	Entero	Codificación del riesgo de obesidad a partir del IMC de la persona. En el punto 3.2. se explica la equivalencia a detalle: • **0** = No hay riesgo/No presenta obesidad • **1** = Si hay riesgo/Presenta obesidad

Fuente: Propia

3.2.2.3. Descripción del formato de datos

Dentro del dataset se ha considerado que todos los campos sean numéricos, ya sean enteros (números positivos y negativos, incluido el cero, que no tienen parte fraccionaria en su estructura) o decimales (es un número no entero que consta de una parte entera y una parte fraccionaria), a fin de facilitar la integración con el algoritmo y así evitar errores al momento del entrenamiento.

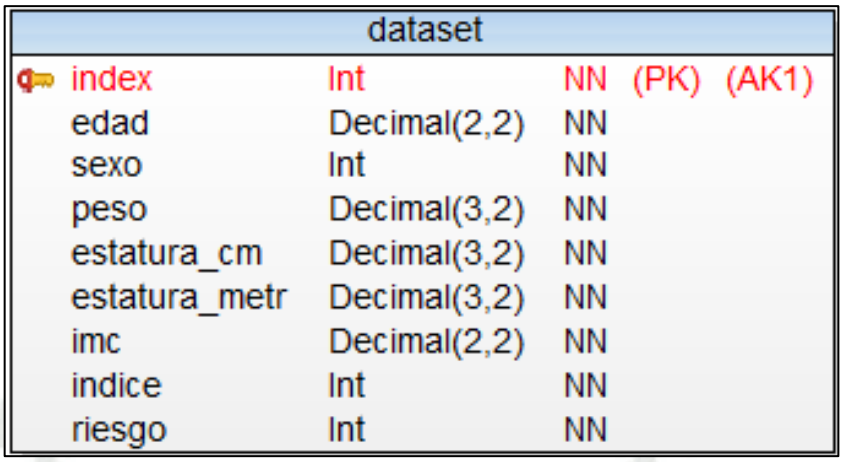


Figura 11: Diagrama de clases de dataset

Fuente: Toad Data Modeler 7.1

3.2.3. Exploración de datos

Para la exploración de los datos, se utilizará estadística descriptiva, ya que lo que se quiere lograr es explorar los datos a fin de identificar características que se deban tener en cuenta para la construcción del algoritmo.

En esta etapa se utilizará la librería de visualización de datos de Python llamada Seaborn, con la cual se podrá ver la relación entre cada una de las variables con las que se cuenta en el dataset a partir de la creación de una cuadrícula de ejes.

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

Figura 12: Importación de librería Seaborn en Colab

Fuente: Propia

En la siguiente figura se puede visualizar la relación del conjunto de datos entre cada variable del dataset.

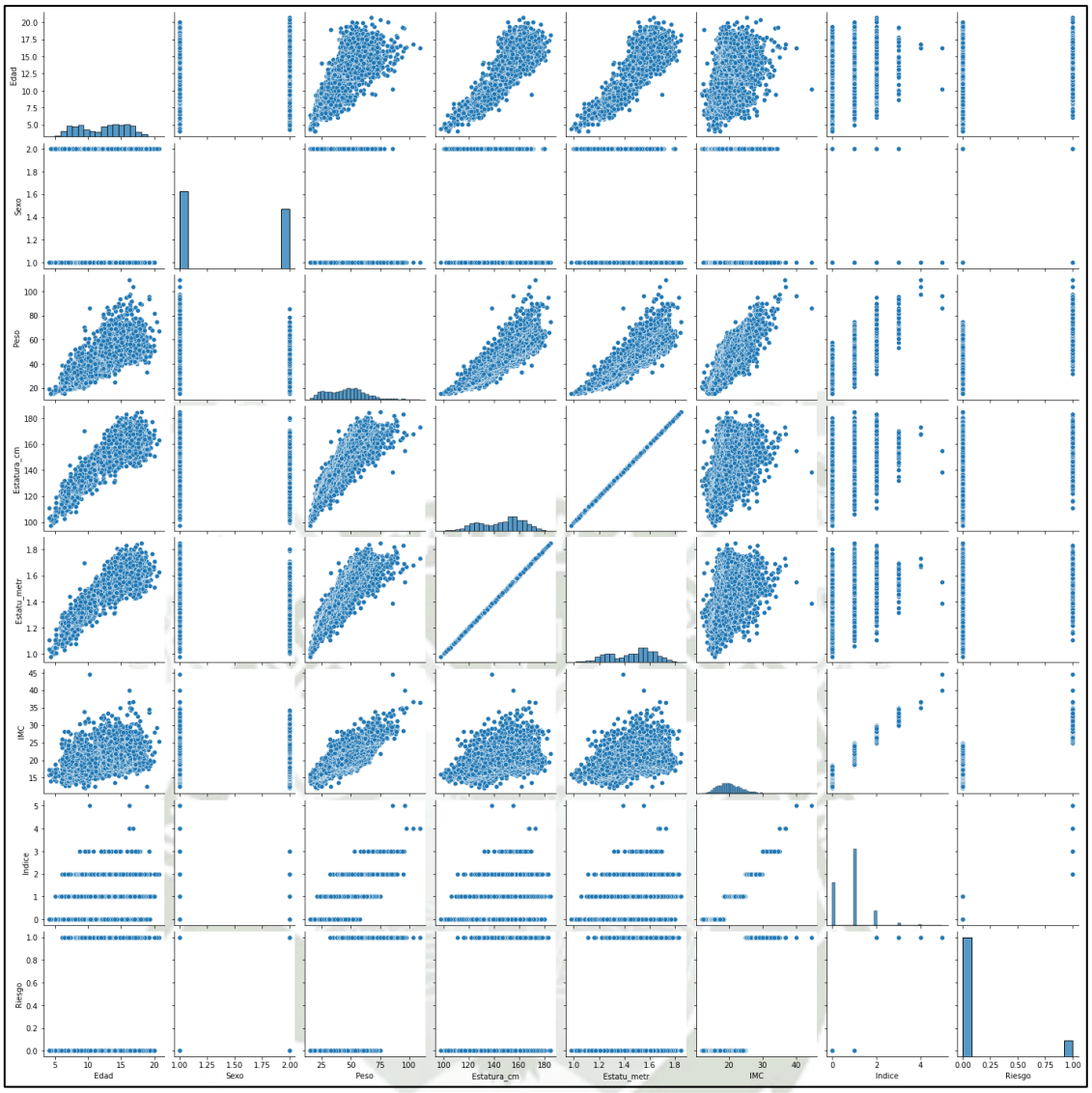


Figura 13: Relación del conjunto de datos entre sí

Fuente: Propia

Al mismo tiempo, se verificará las estadísticas generales de cada una de las variables del dataset.

```
dataset_original.describe().T
```

Figura 14: Uso de método describe ()

Fuente: Propia

En la siguiente tabla se observa el resultado del análisis estadístico aplicado al dataset, en donde se puede visualizar los valores para el número de muestras, el valor medio, la desviación

estándar, el mínimo, el máximo, la mediana y los valores bajo los percentiles 25%, 50 y 75% de cada uno.

Tabla 17: Análisis estadístico a dataset

	count	mean	std	min	25%	50%	75%	max
Edad	3068.0	12.316786	3.585704	4.100000	8.900000	12.880000	15.242500	20.680000
Sexo	3068.0	1.435137	0.495856	1.000000	1.000000	1.000000	2.000000	2.000000
Peso	3068.0	45.255887	15.079973	15.100000	32.500000	46.000000	55.700000	109.300000
Estatura_cm	3068.0	146.308149	17.020440	97.700000	131.500000	150.000000	159.600000	185.000000
Estatu_metr	3068.0	1.463367	0.170121	0.980000	1.320000	1.500000	1.600000	1.850000
IMC	3068.0	20.488369	3.688799	12.052342	17.849215	20.070056	22.684321	44.511154
Índice	3068.0	0.821708	0.682224	0.000000	0.000000	1.000000	1.000000	5.000000
Riesgo	3068.0	0.119948	0.324954	0.000000	0.000000	0.000000	0.000000	1.000000

Fuente: Propia

3.2.3.1. Identificación de los tipos de datos

La identificación del tipo de datos es importante ya que permite decidir qué análisis estadístico y qué método de análisis estadístico sería el más apropiado de utilizar para cada tipo de datos. A continuación, se hará el análisis de cada atributo que se tiene en el dataset, el cual se informará en la siguiente tabla sobre la columna “Tipo de variable”.

Tabla 18: Tipo de variable de los atributos

Tipo de atributo	Nombre de atributo	Tipo de Variable
Original	Edad	**Variable cualitativa**
Original	Sexo	**Variable cualitativa**
Original	Peso	**Variable cuantitativa continua**
Original	Estatura_cm	**Variable cuantitativa continua**
Original	Estatu_metr	**Variable cuantitativa continua**
Generado	IMC	**Variable cuantitativa continua**
Generado	Índice	**Variable cualitativa**
Generado	Riesgo	**Variable cualitativa**

Fuente: Propia

Luego de realizar la categorización, se puede visualizar que, de los 8 atributos, 4 son numéricos o cualitativos y 4 son categóricos o cuantitativos continuos. Es por ello por lo que el análisis se realizará a través de los siguientes gráficos a través del lenguaje de programación Python, mediante sus librerías Plotly y Pandas.

Se realizarán dos tipos de análisis: Análisis individual de cada una de las variables, y análisis combinados. Para el análisis individual se utilizarán los siguientes gráficos según el tipo de variable.

- Variable cualitativa

  a) Gráfico de barras

  b) Gráfico de tortas

  c) Histograma

- Variable cuantitativa continua

  a) Histograma

Luego del análisis individual de cada atributo, se realizará los siguientes análisis combinados:

- Edad y sexo
- Riesgo de obesidad por sexo
- Riesgo de obesidad por edad
- Riesgo de obesidad según altura

Análisis por atributo

a) Edad

En el siguiente gráfico de barras se puede observar que en la mayoría de los datos que se tiene son de personas de 14 años y la minoría son personas de 20 años. Evaluando los rangos, se puede ver que la mayoría de los datos está entre 14.42 y 15.28 y la minoría está en el rango de 20.44 y 21.30 (ver Figura 15).

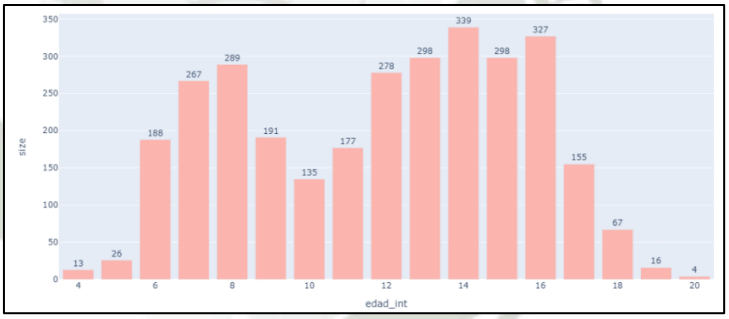


Figura 15: Gráfico de barras de atributo edad

Fuente: Propia

La información visualizada por el anterior gráfico se puede comprobar en el siguiente histograma, mayoría de los datos está entre 14.00 y 14.49 y la minoría está en el rango de 20.50 y 20.99 (ver Figura 16).

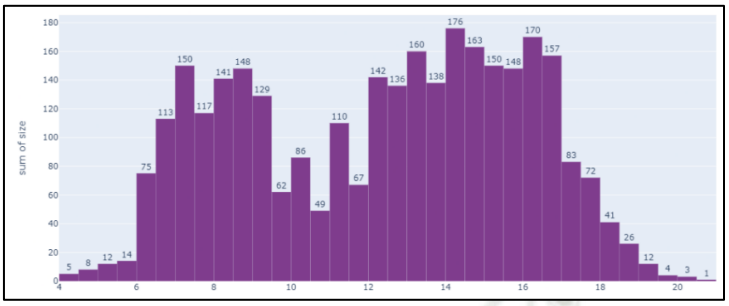


Figura 16: Histograma de atributo edad

Fuente: Propia

Finalmente, con el gráfico circular 2D, o gráfico de torta, se puede visualizar que la cantidad de datos que se tiene de cada categoría cuando se aproxima los valores a cero decimales es equitativa, con lo que se puede concluir que es preferible utilizar decimales, tal como se propuso, para tener una mejor precisión en los resultados (ver Figura 17).

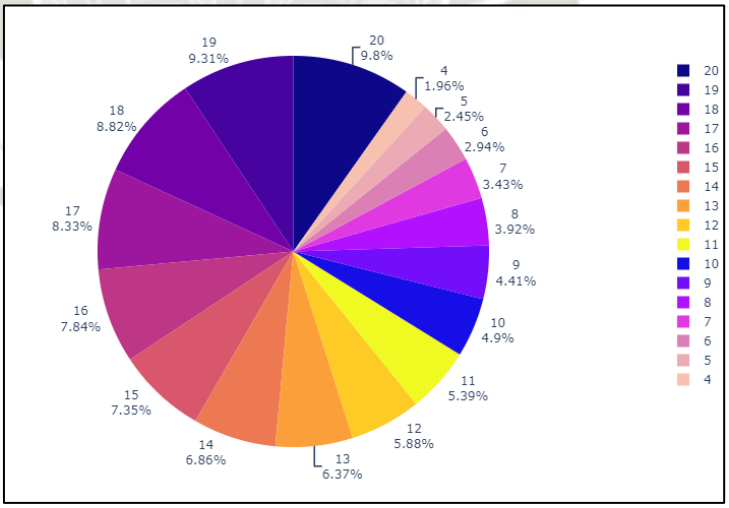


Figura 17: Gráfico de torta de atributo edad

Fuente: Propia

b) Sexo

En el gráfico se puede observar que la mayoría de los datos con los que se cuenta son de hombres (ver Figura 18).

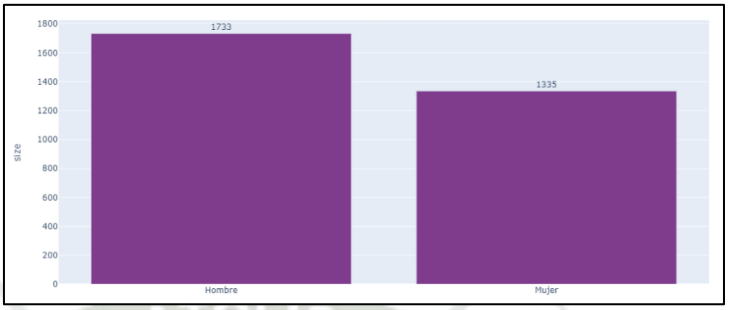


Figura 18: Gráfico de barras de atributo sexo

Fuente: Propia

Con el gráfico circular 2D, o gráfico de torta, se puede visualizar en efecto, un 56% de los datos representan al sexo masculino, mientras que el 44% representan al sexo femenino (ver Figura 19).

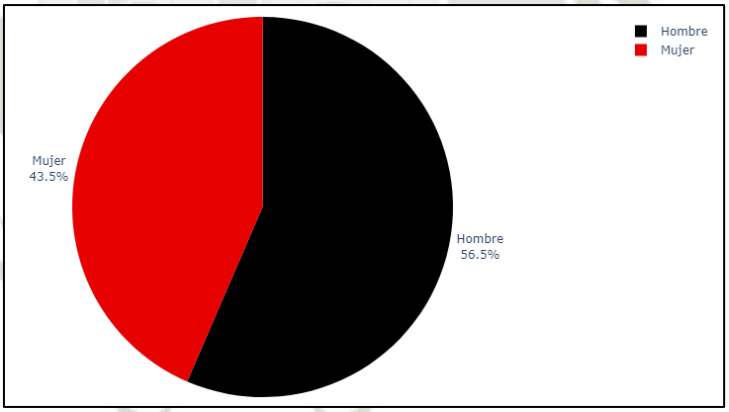


Figura 19: Gráfico de torta de atributo sexo

Fuente: Propia

c) Peso

En el siguiente gráfico de barras se puede observar los datos que se tienen en el atributo Peso son variados, que pueden ir desde los 15 kg hasta los 105 kg. (ver Figura 20).

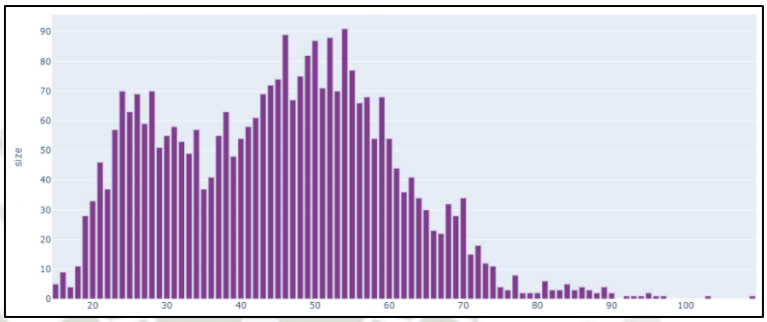


Figura 20: Grafico de barras de atributo Peso

Fuente: Propia

Al mismo tiempo, en el siguiente histograma se puede observar que los datos que se tienen son dispersos, pero que la mayoría de los datos son de personas que tienen un peso entre 50.0 y 54.9. Además, se puede observar que no se cuentan con muchos datos a partir de 90.0 kg hasta 109.9 kg (ver Figura 21).

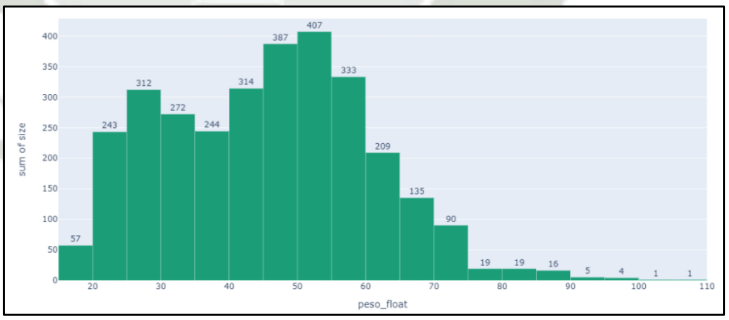


Figura 21: Histograma de atributo peso

Fuente: Propia

d) Estatu_metr y Estatura_cm

Al ser equivalentes Estatu_mts y Estatura_cm, se evaluará únicamente Estatu_mts, que es el atributo que más adelante se utilizará.

En el siguiente gráfico de barras se puede observar los datos que se tienen en el atributo Estatu_metr son variados, que pueden ir desde los 0.98 mts hasta los 1.85 mts (ver Figura 22).

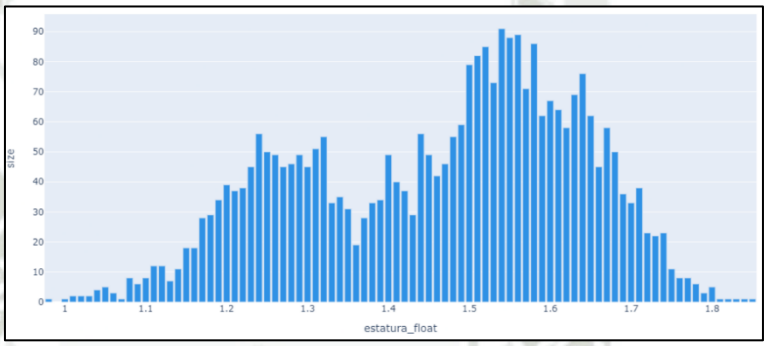


Figura 22: Gráfico de barras Estatura_mts

Fuente: Propia

Al ser la equivalencia del grafico anterior, se puede apreciar que la mayoría de los datos se encuentran en el rango de 1.50 m - 1.59 m. así como se tiene una carencia de datos en los tangos de 0.90 m – 0.99 m, y 1.80 m - 1.89 m (ver Figura 23).

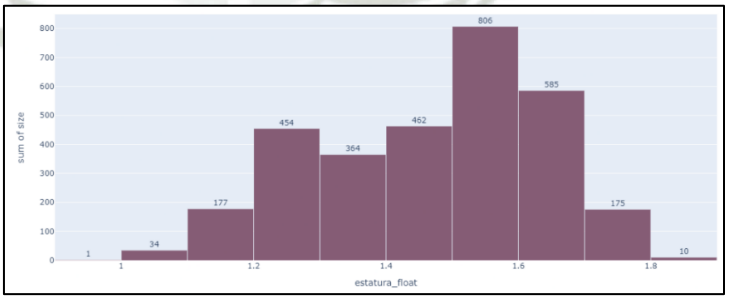


Figura 23: Histograma de atributo estatura_metr

Fuente: Propia

e) IMC

En el siguiente histograma se puede visualizar que, a partir de la unión de todos los datos para hallar el IMC de cada registro, se obtuvo en su mayoría un IMC dentro del rango 19.00 – 20.00, además que, por la dispersión de los pesos, se tienen pocos casos de IMC entre el rango 35.00 y 45.00 (ver Figura 24).

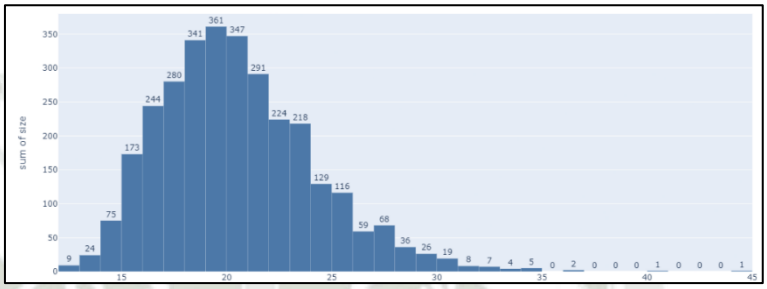


Figura 24: Histograma de atributo IMC

Fuente: Propia

f) Índice

En el siguiente gráfico de barras se puede apreciar que, a partir del IMC obtenido, la mayoría de las personas cuentan con un peso normal, mientras que se tienen pocos casos con obesidad tipo 1, 2 y 3 (ver Figura 25).

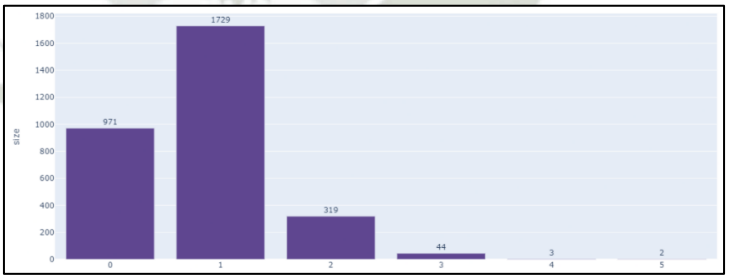


Figura 25: Grafico de barras de atributo índice

Fuente: Propia

Con el gráfico circular 2D, o gráfico de torta, se puede visualizar que, en efecto, la mayoría de los datos del dataset indican tener un peso normal (código 1) a partir del IMC obtenido, seguido del bajo peso y sobrepeso, código 0 y 2 respectivamente. (ver Figura 26).

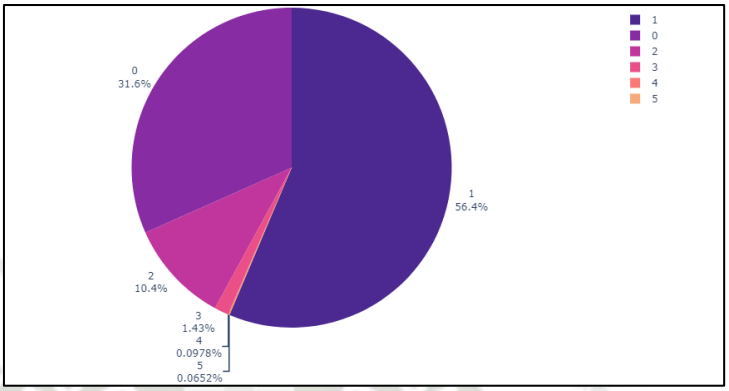


Figura 26: Gráfico de torta de atributo índice

Fuente: Propia

g) Riesgo

Siguiendo el análisis de IMC y lo que conlleva, se puede observar en el siguiente gráfico de barras que la mayoría de los registros no tienen riesgo de padecer obesidad (ver Figura 27).

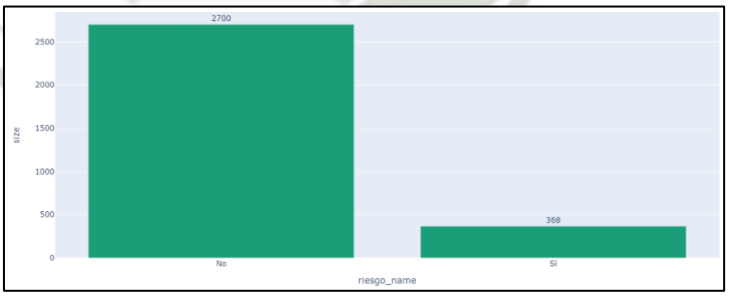


Figura 27: Gráfico de barras de atributo riesgo

Fuente: Propia

Con el gráfico circular 2D, o gráfico de torta, se puede visualizar que efectivamente la mayoría de los datos no tienen riesgo a obesidad (ver Figura 28).

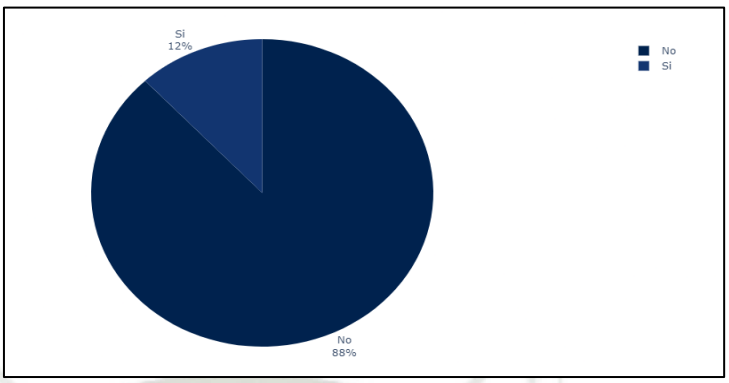


Figura 28: Gráfico de torta de atributo riesgo

Fuente: Propia

Análisis combinado

a) Edad y sexo

Según los datos con los que se cuenta actualmente, se puede observar que dentro del dataset, la mayor cantidad de personas de sexo masculino "hombre" está en el rango (ver Figura 29).

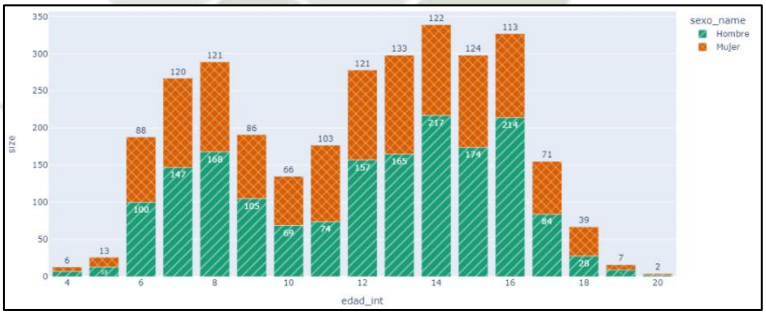


Figura 29: Grafico análisis edad y sexo

Fuente: Propia

b) Riesgo de obesidad por sexo

Según los datos con los que se cuenta actualmente, se puede observar que el mayor riesgo a sufrir obesidad lo tienen las personas del sexo masculino, a comparación de las personas de sexo femenino (ver Figura 30).

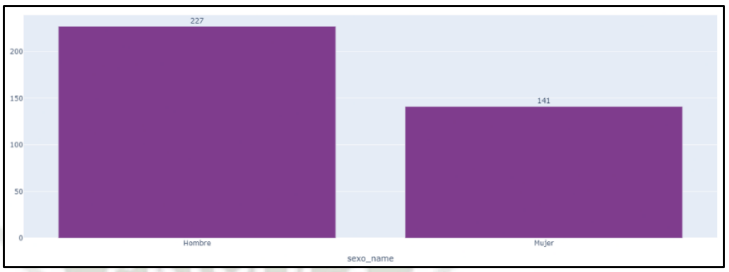


Figura 30: Gráfico de barras de análisis combinado Riesgo de obesidad por sexo

Fuente: Propia

c) Riesgo de obesidad por edad

Interpretando el siguiente gráfico, se puede observar que, según los datos que se tiene, el mayor riesgo de tener obesidad es a los 14 años. Mientras que el riesgo mínimo es entre los 6 y 7 años, así como a los 19 y 20 años (ver Figura 31).

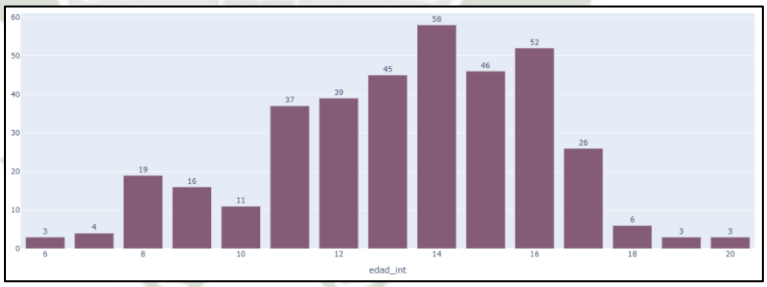


Figura 31: Gráfico de barras de análisis combinado Riesgo de obesidad por edad

Fuente: Propia

d) Riesgo de obesidad según altura

En el siguiente gráfico podemos observar que el riesgo de sufrir obesidad según los datos que se tiene se da en personas sin distinción de sexo o edad que miden 1.54 mts. En este caso, son casos muy dispersos como para definir en qué medidas hay minoría de casos (ver Figura 32).

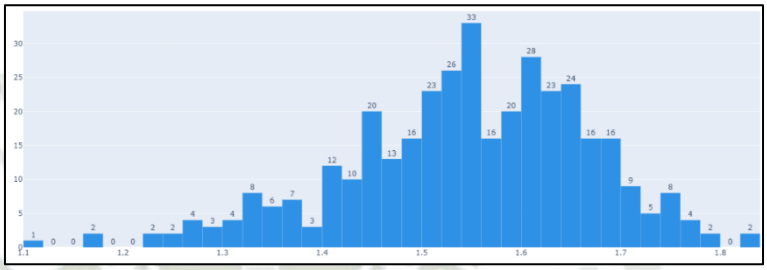


Figura 32: Gráfico de barras de análisis combinado Riesgo de obesidad según altura

Fuente: Propia

3.2.4. Verificación de datos

Para este cuarto paso, se responderá una lista de comprobación a fin de poder realizar una evaluación profunda de los datos:

Tabla 19: Verificación de datos

Punto de comprobación	Respuesta
Cobertura de los datos	Todos los valores representan todos los casos posibles que se necesitan: Casos en los que la persona tenga riesgo de obesidad y casos en los que no, según edad, peso, talla.
Claves	En el dataset todos los atributos arman la clave necesaria para llegar al resultado.
Coherencia entre atributos y valores	Los atributos y valores de cada uno se satisfacen simultáneamente, es decir, son coherentes.
Campos en blanco y atributos omitidos	Se pudo observar dentro del dataset que no hay campos en blanco Para el entrenamiento se omitirán los atributos “Índice” e “IMC”
Atributos erróneos, valores distintos con significados iguales	Dentro del dataset se estandarizó el campo “sexo” a numérico, ya que estaban algunos valores en formato texto, por lo cual se transformó a que: **-** 1 es sexo masculino **-** 2 es sexo femenino

Ortografía y formato	Al ser todos los campos numéricos, no se tiene problema de ambigüedad.
Desviaciones y posibles "ruidos" (outliers)	No se encontraron posibles valores que amenacen con causar "ruido" al momento del entrenamiento.
Plausibilidad	Tal como se pudo apreciar en la exploración de los datos, los valores de este son muy variados, por lo que se puede concluir que no exista plausibilidad en el dataset.
Ruido e inconsistencia entre las fuentes de datos	Los datos son de una sola fuente inicial por lo que no se encontró ruido o inconsistencias.

Fuente: Propia

A partir del análisis de cada punto de comprobación a partir de la exploración, se puede concluir que los datos con los que se cuentan preliminarmente se pueden afirmar que se encuentran completos, es decir, los datos cubren los casos necesarios para el entrenamiento y posteriores pruebas del sistema a fin de conseguir el objetivo principal.

Los datos no contienen errores, todos los campos son formato numérico. Tampoco se ha podido visualizar valores fuera de los rangos, valores nulos o valores negativos, por lo que no se considera que exista un riesgo de ruido en el proceso de entrenamiento, o no se le puede considerar al dataset "contaminado".

3.3. Preparación de los datos

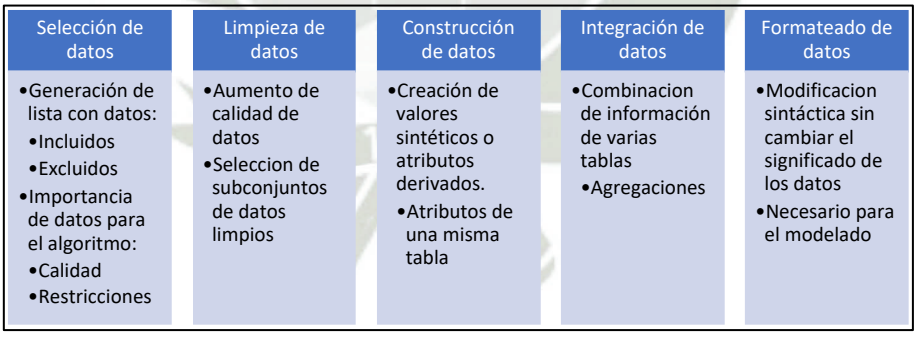


Figura 33: Fase 3 Preparación de los datos

Fuente: Propia

Dentro de esta fase se preparará los datos de cara a la fase 4 de modelado. Implicará seleccionar los datos a utilizar, limpiarlos para garantizar una mejor consistencia de estos, construir datos sintéticos de ser necesarios, así como integración de atributos de diferente tabla, para finalmente formatearlos.

Dentro de esta fase se realizarán algunos pasos que se realizaron en la fase anterior, pero de una manera más estandarizada.

Presentación de datos iniciales

El fichero con el que se trabajará es un fichero que cuenta con 3068 instancias y 5 atributos, los cuales son:

Tabla 20: Hoja de cálculo - Presentación de datos iniciales

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.2
7.95	1	24	132.5	1.33

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

a) Edad

b) Sexo

c) Peso

d) Estatura_cm

e) Estatu_metr

En la fase 2, se pudo visualizar que los datos eran de diferentes tipos: numéricos y alfanuméricos, por lo que se tendrá que estandarizar más adelante a fin de evitar inconsistencias y ruido al momento del entrenamiento.

3.3.1. Selección de datos

La selección de datos se realizará tanto a nivel de atributos como a nivel de instancias.

A nivel de instancias

En términos de instancias, se utilizarán todas las que están presente en el dataset, ya que han sido recolectados de acuerdo con el objetivo que se quiere lograr. En total se tienen instancias que representan diferentes casos que se pueden presentar en el uso del

algoritmo, lo que va a favorecer que en el momento del entrenamiento se cuente con variedad de casos.

A nivel de atributos

En términos de atributos, no se utilizarán todos. Del dataset se utilizará los atributos:

a) Edad

b) Sexo

c) Estatura_metr

d) Peso

No se utilizará el atributo estatura_cm ya que puede causar ambigüedad por lo que es el equivalente a la variable estatura_metr. Entonces, el dataset por el momento estaría quedando de la siguiente manera:

Tabla 21: Selección de datos a nivel de atributos

Atributo	Tipo de dato	Se utilizará en el modelo	Motivo
Edad	Numérico	Si	
Sexo	Numérico/Alfanumérico	Si	
Peso	Numérico	Si	
Estatura_metr	Numérico	Si	
Estatura_cm	Numérico	No	Es equivalente a variable estatura_metr. El incluirlo en el entrenamiento podría causar ruido.

Fuente: Propia

La razón por la que se está incluyendo o excluyendo atributos es por la importancia que cumplirán en relación con poder cumplir el objetivo planteado definido en la fase 1 y el evitar la ambigüedad entre atributos que a futuro puedan generar ruido en el entrenamiento.

3.3.2. Limpieza de datos

El dataset con el que se entrenará el algoritmo contiene todos los datos, tanto instancias como atributos, necesarios para cumplir el objetivo de entrenamiento y posteriormente hacer las pruebas a fin de ver que la predicción es la correcta.

A nivel de estandarización, se hará una evaluación atributo por atributo a fin de visualizar inconsistencias que se puedan presentar actualmente y puedan ser propensas a generar ruido en el modelo más adelante, con lo que se logrará aumentar la calidad de los datos.

a) Edad

El atributo es de tipo numérico.

Haciendo análisis más a profundidad se puede visualizar que se tienen datos dentro del rango de 4.1 y 20, algunos hasta con dos decimales.

Tabla 22: Hoja de cálculo - Visualización de atributo edad previo a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.6	1	20.1	120.3	1.2
7.95	1	24	132.5	1.33

Fuente: Propia

La limpieza que se le hará a este atributo es estandarizar el número de decimales que tienen a 2. Quedando el dataset de la siguiente manera:

Tabla 23: Hoja de cálculo - Visualización de atributo edad posterior a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.60	1	20.1	120.3	1.2
7.95	1	24	132.5	1.33

Fuente: Propia

b) Sexo

El atributo es numérico y alfanumérico.

Tabla 24: Hoja de cálculo - Visualización de atributo sexo

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.2
7.95	1	24	132.5	1.33
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

Al ser diferentes, el campo se tiene que normalizar a un solo tipo de valores, por lo que en el apartado 3.3.3. se realizará una transformación de ese atributo a valor numérico, de tal forma que el dataset quede unificado y el atributo sea numérico únicamente.

c) Estatura_metr

El atributo es de tipo numérico.

Haciendo análisis más a profundidad se puede visualizar que se tienen datos dentro del rango de 0.98 y 1.85, algunos hasta con dos decimales.

Tabla 25: Hoja de cálculo - Visualización de atributo estatura_metr previo a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.2
7.95	1	24	132.5	1.3
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

La limpieza que se le hará a este atributo es estandarizar el número de decimales que tienen a 2. Quedando el dataset de la siguiente manera:

Tabla 26: Hoja de cálculo - Visualización de atributo estatura_metr posterior a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.20
7.95	1	24	132.5	1.30
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

d) Peso

El atributo es de tipo numérico.

Haciendo análisis más a profundidad se puede visualizar que se tienen datos dentro del rango de 15.1 y 75, algunos hasta con dos decimales.

Tabla 27: Hoja de cálculo - Visualización de atributo peso previo a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.20
7.95	1	24	132.5	1.30
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

La limpieza que se le hará a este atributo es estandarizar el número de decimales que tienen a 2. Quedando el dataset de la siguiente manera:

Tabla 28: Visualización de atributo peso posterior a limpieza

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.20
7.95	1	24.0	132.5	1.30
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19.0	118	1.18

Fuente: Propia

En la exploración que se realizó en la fase 2, también se pudo visualizar que no se cuentan con variables vacías, nulas o incorrectas, todos los valores con los que se encuentran están dentro de los rangos permitidos.

3.3.3. Construcción de datos

En esta tarea se realizarán operaciones de preparación tales como el desarrollo de atributos derivados, ingreso de nuevos registros de ser necesarios y transformación de valores para atributos existentes.

Atributos derivados

Los atributos derivados son los atributos nuevos que se construyen con uno o más atributos existentes en el mismo dataset.

Para poder seguir preparando el dataset, se tendrá que crear tres atributos derivados: IMC, Índice y Riesgo.

a) IMC

Este atributo será el resultado de dividir el atributo **Peso** entre el atributo **Estatu_metr** elevado al cuadrado $\left(\frac{peso\ en\ kg}{(altura\ en\ m)^2}\right)$. Este atributo se creará para todos los valores a través de la siguiente fórmula de Excel:

$$= (Cxx)/((Dxx) * (Dxx))$$

Donde C es el atributo peso y D el atributo Estatu_metr.

Tabla 29: Hoja de cálculo - Visualización de atributo derivado IMC

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC
6.66	1	20.5	122.5	1.23	13.550136
6.66	1	20.1	120.3	1.2	13.958333
7.95	1	24	132.5	1.33	13.567754
6.52	2	21.2	128.5	1.29	12.739619
6.97	2	21.7	126.1	1.26	13.668430
6.4	2	19.6	118.3	1.18	14.076415

Fuente: Propia

El atributo IMC será de tipo de dato numérico y tendrá 6 decimales en todos los campos.

b) Índice

Este atributo será el resultado de codificar el IMC de la persona, tal y como se explica en el punto 3.2. La transformación en el dataset se realizó a través de la siguiente fórmula de Excel:

$$
\begin{aligned}
&= SI(Exx < 18.5; 0;\\
&SI(Y(Exx >= 18.5; Exx < 24.9); 1;\\
&SI(Y(Exx >= 24.9; Exx < 29.9); 2;\\
&SI(Y(Exx >= 29.9; Exx < 34.9); 3;\\
&SI(Y(Exx >= 34.9; Exx < 39.9); 4;\\
&SI(Exx >= 39.9; 5; "\#ERROR"))))))
\end{aligned}
$$

Donde E es el atributo IMC.

Tabla 30: Hoja de cálculo - Visualización de atributo índice

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC	Índice
6.66	1	20.5	122.5	1.23	13.550136	0
6.66	1	20.1	120.3	1.2	13.958333	0
7.95	1	24	132.5	1.33	13.567754	0
6.52	2	21.2	128.5	1.29	12.739619	0
6.97	2	21.7	126.1	1.26	13.668430	0
6.4	2	19.6	118.3	1.18	14.076415	0

Fuente: Propia

El nuevo atributo Índice es de tipo numérico.

c) Riesgo

Este atributo será el resultado de codificar el atributo derivado Riesgo, tal y como se explicar en el punto 3.2. La transformación en el dataset se realizó a través de la siguiente fórmula de Excel:

$$= SI(O(Fxx = 0; Fxx = 1); 0; 1)$$

Donde F es el atributo Índice.

Tabla 31: Hoja de cálculo - Visualización de atributo riesgo

Edad	Sexo	Peso	Estatura_cm	Estatu_metr	IMC	Índice	Riesgo
6.66	1	20.5	122.5	1.23	13.550136	0	0
6.66	1	20.1	120.3	1.2	13.958333	0	0
7.95	1	24	132.5	1.33	13.567754	0	0
6.52	2	21.2	128.5	1.29	12.739619	0	0
6.97	2	21.7	126.1	1.26	13.668430	0	0
6.4	2	19.6	118.3	1.18	14.076415	0	0

Fuente: Propia

El nuevo atributo Riesgo es de tipo numérico.

Ingreso de nuevos registros

Dentro del dataset no se necesita ingresar nuevos registros, ya que con la cantidad que se tiene, se podrá entrenar y probar el modelo sin problema.

Transformación de valores para atributos existentes

Tal como se pudo visualizar en el apartado 3.2.2., el atributo "sexo" es numérico y alfanumérico, lo cual a futuro podría causar ruido en el entrenamiento. En esta fase se transformará los valores alfanuméricos a numérico.

a) Sexo

Se puede visualizar que el campo es alfanumérico y numérico.

Tabla 32: Hoja de cálculo - Visualización de atributo sexo previo a transformación

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.20
7.95	1	24	132.5	1.30
6.52	mujeres	21.2	128.5	1.29
6.97	mujeres	21.7	126.1	1.26
7.17	mujeres	19	118	1.18

Fuente: Propia

En esta fase se transformará la palabra “mujeres” al valor numérico “2”, quedando el dataset de la siguiente manera.

Tabla 33: Hoja de cálculo - Visualización de atributo sexo posterior a transformación

Edad	Sexo	Peso	Estatura_cm	Estatu_metr
6.66	1	20.5	122.5	1.23
6.66	1	20.1	120.3	1.20
7.95	1	24	132.5	1.30
6.52	2	21.2	128.5	1.29
6.97	2	21.7	126.1	1.26
7.17	2	19	118	1.18

Fuente: Propia

Siguiendo con un análisis más profundo, al ser un número entero, no se necesita mayor limpieza, quedando de esa forma el dataset. Con lo que el atributo quedaría listo para la etapa de entrenamiento.

3.3.4. Integración de datos

En esta etapa se comprobó si es posible integrar fuentes de entrada, tablas del dataset. En este caso, no ha sido necesario la creación de nuevas estructuras a partir de dos tablas, ya que el dataset con el que se cuenta solo forma una tabla.

3.3.5. Formateado de datos

A lo largo de la fase 3 de la metodología se estuvo realizando estandarizaciones y transformaciones en los datos, a fin de evitar posibles ruidos al momento del entrenamiento del modelo. Se ha preferido colocar a todos los atributos en tipo numérico ya que en tipo alfanumérico se podía llegar a tener ambigüedades en los datos. A continuación, se detallará campo a campo los cambios que se hicieron en cada uno de los atributos.

a) Edad

El atributo es de tipo numérico desde el dataset original, se ha mantenido eso, pero se ha estandarizado el número de decimales a 2.

b) Sexo

El atributo originalmente era de tipo numérico y alfanumérico, por lo que se procedió a realizar una modificación sintáctica a los datos alfanuméricos, a fin de que el atributo quede únicamente de tipo numérico. Al hacer el cambio de valor, la variable no cambió ni su valor ni su significado, la equivalencia es la siguiente:

Tabla 34: Equivalencia de valores alfanúmeros y numéricos en atributo sexo

Valor alfanumérico original	Valor numérico actual
-	1
mujeres	2

Fuente: Propia

c) Estatura_metr

El atributo es de tipo numérico desde el dataset original, se ha mantenido eso, pero se ha estandarizado el número de decimales a 2.

d) Peso

El atributo es de tipo numérico desde el dataset original, se ha mantenido eso, pero se ha estandarizado el número de decimales a 2.

e) IMC

Este es un atributo generado o atributo derivado, resultado de la división del peso entre la estatura elevada al cuadrado. Es de tipo numérico y está estandarizado a 6 decimales.

f) Índice

Este es un atributo generado o atributo derivado, resultado de la codificación del atributo IMC a partir de las siguientes equivalencias:

Tabla 35: Equivalencia de valores en atributo índice

Valor en atributo IMC	Valor en atributo "Índice"
< 18.5	0
$> 18.5\ y\ < 24.9$	1
$> 25\ y\ < 29.9$	2
$> 30\ y\ < 34.9$	3
$> 35\ y\ < 39.9$	4
≥ 40	5

Fuente: (Suárez-Carmona & Sánchez-Oliver, 2014)

Es de tipo numérico y no necesitó ser estandarizado a nivel de decimales ya que es un número entero.

g) Riesgo

Este es un atributo generado o atributo derivado, resultado de la codificación del atributo Índice a partir de las siguientes equivalencias:

Tabla 36: Equivalencia de valores en atributo riego

Valor en atributo Índice	Valor en atributo Riesgo
0,1	0
2,3,4,5	1

Fuente: Propia

Es de tipo numérico y no necesitó ser estandarizado a nivel de decimales ya que es un número entero.

Finalmente, no es necesario reordenar a los atributos ni los registros, ya que en ese orden se podrá realizar el entrenamiento del algoritmo correctamente.

Luego de seguir todos los pasos de preprocesamiento, el dataset quedaría de la siguiente manera:

Tabla 37: Estado de dataset luego de preprocesamiento de datos

Atributo	Tipo de dato	Se utilizará en el modelo	Motivo
Edad	Numérico	Si	
Sexo	Numérico	Si	
Peso	Numérico	Si	
Estatura_metr	Numérico	Si	
Estatura_cm	Numérico	No	Es equivalente a variable estatura_metr. El incluirlo en el entrenamiento podría causar ruido.
IMC	Numérico	No	El incluirlo en el entrenamiento podría causar ruido.
Índice	Numérico	No	El incluirlo en el entrenamiento podría causar ruido.
Riesgo	Numérico	Si	

Fuente: Propia

3.4. Búsqueda/Modelado

En esta cuarta fase de la metodología se escogerá la técnica más adecuada para cumplir el objetivo planteado. Luego de eso se diseñará el plan de pruebas (test) a aplicar en el modelo que se construirá en un tercer paso y finalmente se tendrá que evaluar a fin de ver si cumple con los criterios planteados o no.

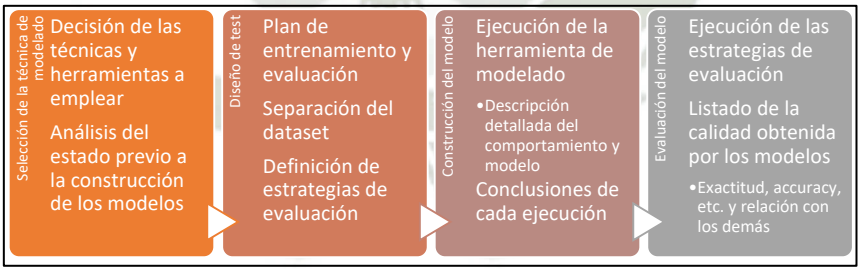


Figura 34: Fase 4 Búsqueda/Modelado

Fuente: Propia

3.4.1. Selección de la técnica de modelado

En esta cuarta fase de la metodología se realizarán utilizarán diferentes modelos de análisis basados en las técnicas de aprendizaje automático a fin de poder decidir cuál es la que

mejor se adapta al dataset con el que se cuenta. Para esto, se utilizará la herramienta RapidMiner Studio en la versión 9.10.001.

Figura 35: Versión de RapidMiner

Fuente: RapidMiner

RapidMiner es una plataforma cruzada que combina minería de datos con minería de texto, aprendizaje automático, inteligencia y análisis comerciales. Esta herramienta permite desarrollar procesos de análisis, asi como acelerar la creación de analíticas predictivas a través de los más de 500 operadores orientados al análisis de datos con los que cuenta, tanto de entrada y salida, preprocesamiento de datos y visualización de estos.

Dentro de RapidMiner se construirán los modelos de análisis basados en técnicas de aprendizaje automático:

a) Árboles de decisión

Trata de encontrar una variable que divida el dataset en grupos lógicos hasta llegar a la solución de problema. Es de gran ayuda a la hora de determinar decisiones durante un proceso.

b) Redes neuronales

Trabaja en base a reconocimiento de patrones que imita las neuronas humanas. Es capaz de modelar relaciones y complejas y se utiliza cuando no se conoce la relación exacta que hay entre los valores en cuestión.

c) Máquinas de Vectores de Soporte (SVM)

Se basa en reconocimiento de patrones y es más utilizado en problemas de clasificación o regresión.

d) Análisis bayesiano

Utiliza las evidencias obtenidas para concluir si las suposiciones planteadas sean ciertas.

e) Regresión logística

Normalmente utilizada para predecir variables categóricas. Es útil en el momento de diseñar la eventualidad de un caso planteado.

f) Regresión lineal

También llamado método de los mínimos numéricos trata de representar el "mejor encaje" entre todos los puntos a través del cálculo de la suma de las distancias al cuadrado entre los puntos.

g) K-Vecinos más cercanos

Trabaja en base a reconocer patrones para conocer la probabilidad con que un elemento pertenezca a una clase según su cercanía en el espacio.

A primera instancia, el dataset está adaptado para todos los modelos a analizar, a nivel de calidad, distribución y formato, según lo planteado en la fase 3.

3.4.2. Diseño del test

A través de la herramienta RapidMiner también se realizarán las evaluaciones a los modelos que se mencionaron anteriormente, a fin de poder probar la calidad y validez de los modelos. Para poder entender las diferentes evaluaciones, primero se debe conocer qué es la matriz de confusión y sus componentes; así como las métricas a utilizar, los cual se mencionó en el capítulo II.

RapidMiner permite utilizar la estrategia de evaluación cross-validation, o validación cruzada, que básicamente consiste en dividir el dataset aleatoriamente en n grupos del mismo

tamaño, donde n-1 grupos se usan para el entrenamiento y el sobrante para la validación. Esto se repite n veces usando siempre diferentes grupos para la validación en cada iteración. Esta estrategia es también considerada como un procedimiento de re-sampling o re-muestreo.

3.4.3. Construcción del modelo

A continuación, se empezará a ejecutar cada uno de los modelos desarrollados en RapidMiner. En este apartado se explicará cada uno de los ajustes que se realizaron a los parámetros de los modelos, así como la salida y resultados de las métricas aplicadas en cross-validation.

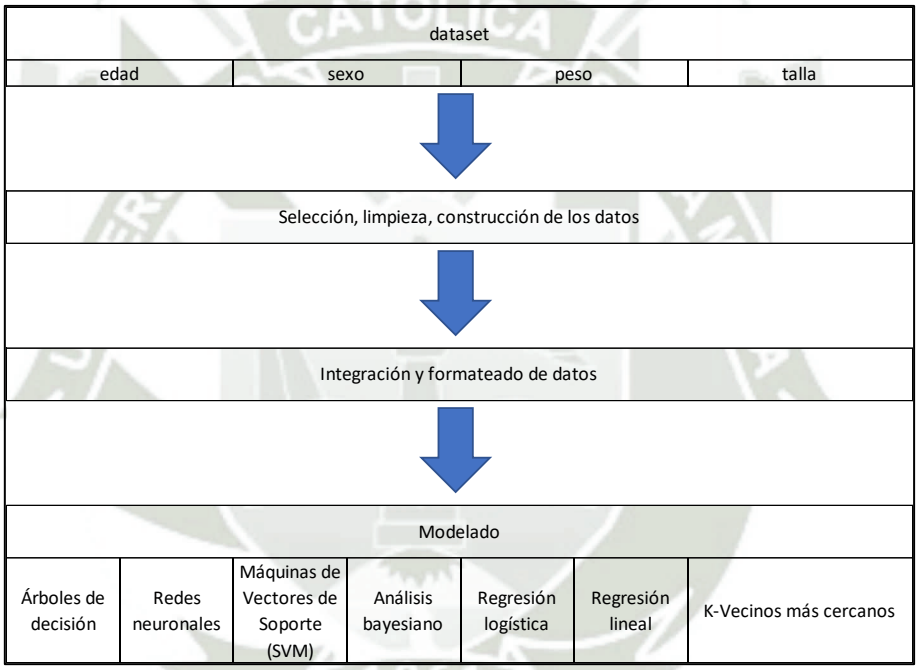


Figura 36: Proceso de modelado en herramienta

Fuente: Propia

Ajuste de pasos previos a cross-validation

Para los 7 modelos que se ejecutarán, y posteriormente evaluarán, se tienen dos pasos en común a implementar dentro de RapidMiner: Importar los datos, seleccionar los atributos que se considerarán y las que no se considerarán dentro del modelo; y finalmente realizar la configuración general del proceso de cross-validation.

- **Importación del dataset**

  La importación del dataset a la herramienta se realiza a través de la opción de importar datos, donde, en cada uno de los pasos, se va configurando el tipo de datos de cada uno de los atributos.

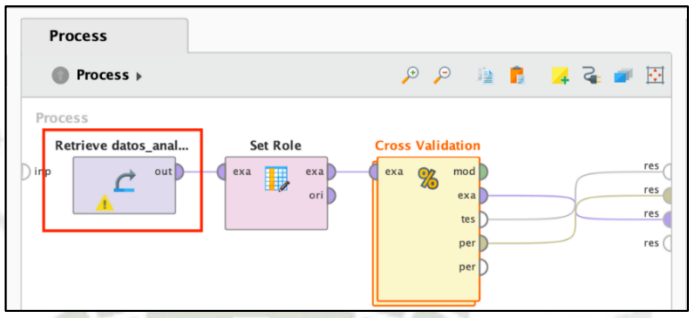
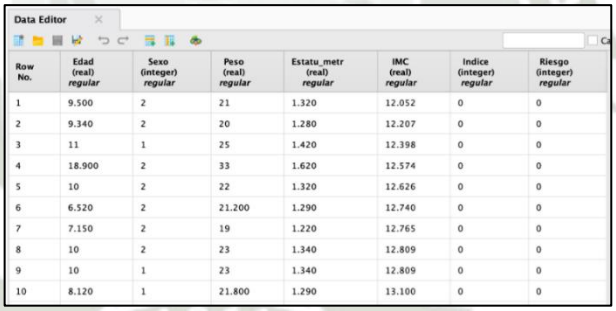


Figura 37: Importación de dataset

Fuente: RapidMiner

Quedando tal y como se explicó en el apartado 3.3.5. Formateado de datos.



Row No.	Edad (real) regular	Sexo (integer) regular	Peso (real) regular	Estatu_metr (real) regular	IMC (real) regular	Indice (integer) regular	Riesgo (integer) regular
1	9.500	2	21	1.320	12.052	0	0
2	9.340	2	20	1.280	12.207	0	0
3	11	1	25	1.420	12.398	0	0
4	18.900	2	33	1.620	12.574	0	0
5	10	2	22	1.320	12.626	0	0
6	6.520	2	21.200	1.290	12.740	0	0
7	7.150	2	19	1.220	12.765	0	0
8	10	2	23	1.340	12.809	0	0
9	10	1	23	1.340	12.809	0	0
10	8.120	1	21.800	1.290	13.100	0	0

Figura 38: RapidMiner - Visualización de datos importados

Fuente: RapidMiner

- **Selección de atributos a incluir y excluir en el modelo**

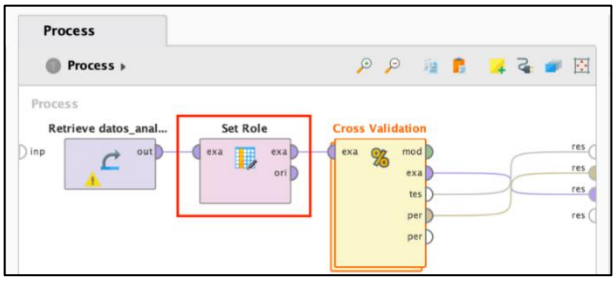


Figura 39: Configuración de atributos

Fuente: RapidMiner

En el operador "Set Role" es donde se configura el atributo de salida y los atributos que serán ignorados al momento de crear el modelo. En este proyecto se seleccionó como atributo de salida "Riesgo"

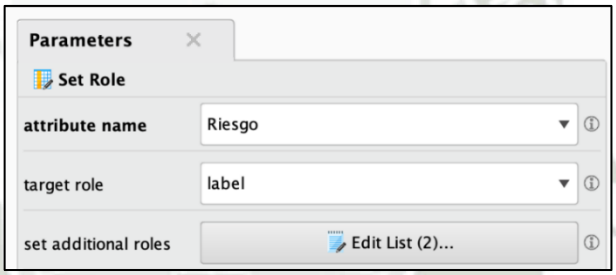


Figura 40: Configuración de atributo de salida

Fuente: RapidMiner

Y como atributos excluidos al momento de crear el modelo

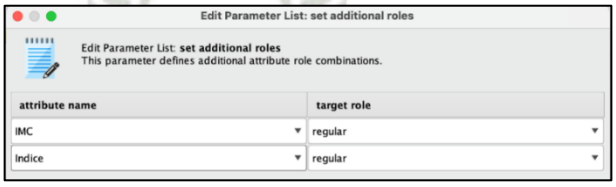


Figura 41: Configuración de atributos excluidos

Fuente: RapidMiner

Quedando tal y como se explicó en el apartado 3.3.5. Formateado de datos.

- **Configuración general de cross-validation**

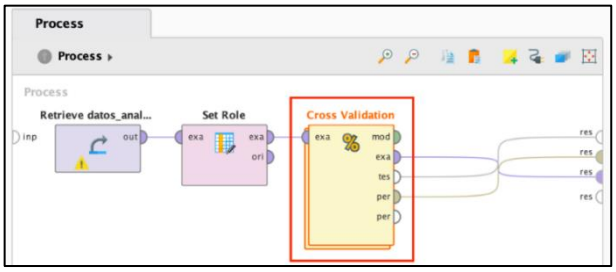


Figura 42: Configuración general de cross-validation

Fuente: RapidMiner

El operador cross-validation es un operador anidado, ya que cuenta con los dos subprocesos característicos de este método: un subproceso de entrenamiento y un subproceso de pruebas.

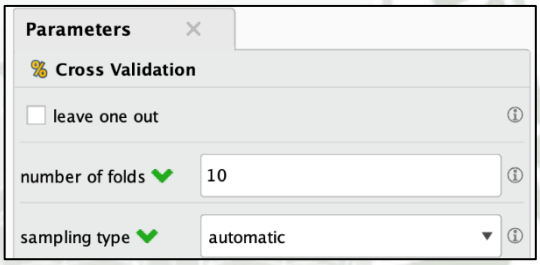


Figura 43: Configuración establecida de cross-validation

Fuente: RapidMiner

Como configuraciones generales para los 7 modelos a evaluar se tendrá el número de particiones del dataset (number of folds) y el tipo de muestreo (sampling type). Ambos se eligieron por defecto.

3.4.3.1. Árboles de decisión

El modelo de árboles de decisión es una colección de nodos que buscan crear una decisión sobre subclases que se crean a partir de la información que contiene el dataset en función a la mayoría de los ejemplos que tiene cuando los atributos son nominales y el promedio cuando los atributos son numéricos. Cada nodo representa una regla de división y estos dejan de crearse cuando se cumplen los criterios de parada.

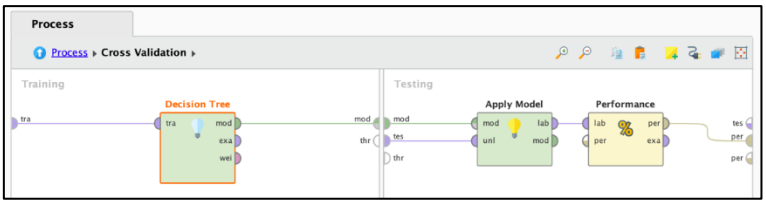


Figura 44: Árboles de decisión

Fuente: RapidMiner

Para la construcción de este modelo, la herramienta RapidMiner permite gestionar algunos parámetros:

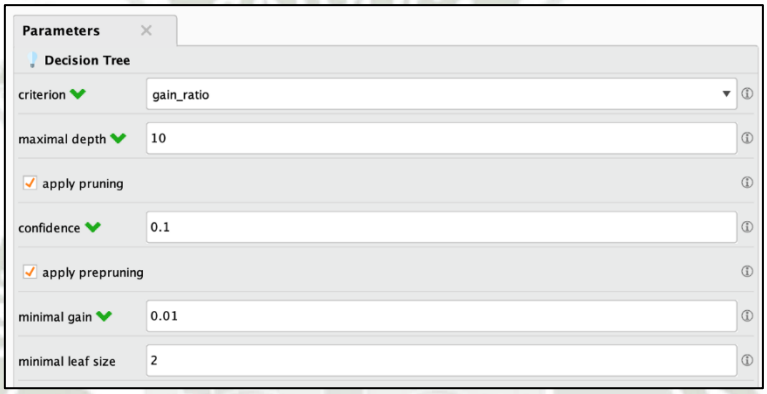


Figura 45: Configuración de parámetros de modelo árboles de decisión

Fuente: RapidMiner

- **Criterion**

  Hace referencia al criterio por el que se dividirán los atributos. Para este modelo se utilizó "gain_ratio", que consiste en ir ajustando el árbol a partir de las ganancias de informacion a fin de permitir una amplitud y uniformidad en los valores.

- **Maximal depth**

  Es el parámetro que indicará la profundidad máxima del árbol. En este modelo se indicó que la profundidad máxima será 10.

- **Apply pruning**

  Especifica si se aplicará una "poda" después de la generación del modelo según el valor de confidence. Se indicó que sí se aplique.

- **Confidence**

  Este parámetro indica el nivel de confianza para que se utiliza en el cálculo del error pesimista de poda. Se utilizó el predeterminado: 0.1.

- **Apply prepruning**

  Permite tener más criterios de para en el árbol. A fin de tenerlos, este valor se tiene activo. Con esto, se activa automáticamente el parámetro "minimal gain" y "minimal leaf size".

- **Minimal gain**

  Este parámetro indica la ganancia mínima de los nodos. Esto se utiliza al momento de saber si un nodo necesita ser dividido o no: Los nodos se dividen cuando su ganancia es mayor a este parámetro. A más alto el valor de la ganancia mínima, menor será la cantidad de divisiones en el árbol. El valor que se colocó fue 0.01.

- **Minimal leaf size**

  Indica el tamaño mínimo de cada hoja, que por defecto se tiene que sea 2.

Al momento de ejecutar el proceso, se obtienen los siguientes resultados indicados en la Tabla 38:

Tabla 38: Resultado de métricas aplicadas a modelo Arboles de decisión

Métricas	Decision tree	
	%	micro average
Accuracy	100.00%+/-0.00%	1
classification_error	0.00%+/-0.00%	0
kappa	1.000+/-0.000	1
spearman_rho	1.000+/-0.000	10
absolute_error	0.000+/-0.000	0.000+/-0.000
relative_error	0.00%+/-0.00%	0.00%+/-0.00%
root_mean_squared_error	0.000+/-0.000	0.000+/-0.000
correlation	1.000+/-0.000	1

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.2. Redes neuronales

Las redes neuronales artificiales (RNA) son un modelo matemático o computacional inspirado en la estructura de las redes neuronales biológicas. Consta de un grupo de neuronas artificiales interconectadas que procesan información a fin de predecir datos. Es también considerado un sistema adaptativo que cambia a manera que recibe información durante la fase de aprendizaje.

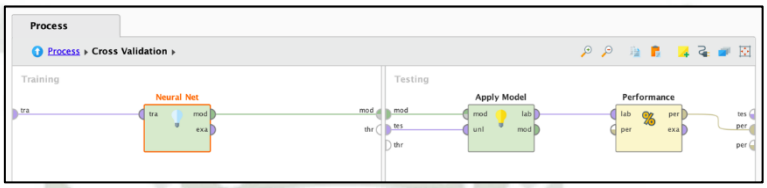


Figura 46: Redes neuronales

Fuente: RapidMiner

Para la creación de una red neuronal artificial feed-forward (un solo sentido), la herramienta RapidMiner permite gestionar algunos parámetros que se detallarán a continuación:

- **Hidden layers**

  En este parámetro se indica el nombre y tamaño de las capas ocultas de la red neuronal. RapidMiner incluye siempre un nodo en cada capa, el cual no está conectado a la capa anterior.

  Para el modelo se aplicó una capa oculta con 5 nodos.

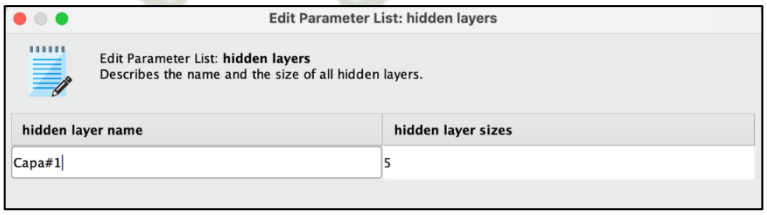


Figura 47: Configuración de capas ocultas

Fuente: RapidMiner

- **Training cycles**

  Especifica el número de entrenamientos que realizará la red. Para el modelo se indicó que realizaría 2 entrenamientos.

- **Learning rate**

  Este parámetro determina cuánto cambia el peso en cada paso, se indicó que cambiaría 0.3.

- **Momentum**

  Consiste en agregar una fracción de la actualización del peso anterior al actual, lo que evita máximos locales. Se indicó que el valor de este parámetro sería 0.6.

- **Error epsilon**

  Es el valor que indica que debe parar la optimización del modelo cuando el error de entrenamiento está por debajo de este valor. Se utilizó el valor por defecto de RapidMiner, que es 1.0E-4.

Quedando la configuración de la siguiente manera:

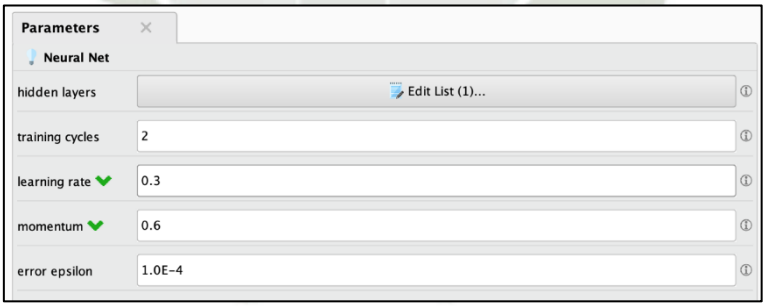


Figura 48: Configuración de parámetros de modelo redes neuronales

Fuente: RapidMiner

Al ejecutar el modelo, se obtuvieron los siguientes resultados:

Tabla 39: Resultado de métricas aplicadas a modelo redes neuronales

Métricas	Neural Network	
	%	micro average
Accuracy	95.70%+/-0.98%	95.70%
classification_error	4.30%+/-0.98%	4.30%
kappa	0.765+/-0.059	0.767
spearman_rho	0.781+/-0.052	7.813
absolute_error	0.109+/-0.011	0.109+/-0.151
relative_error	10.95%+/-1.14%	10.95%+/-15.13%
root_mean_squared_error	0.187+/-0.009	0.187+/-0.000
correlation	0.781+/-0.052	0.781

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.3. Máquinas de Vectores de Soporte (SVM)

Las máquinas de vectores de soporte, o SVM por sus siglas en inglés, proporciona un algoritmo rápido y de buenos resultados ya que trabaja con funciones de pérdida lineales, cuánticas o asimétricas, que se representan como puntos en el espacio clasificados por categorías, en otras palabras, este modelo construye hiperplanos, que son un conjunto de puntos que tienen dentro un vector constante, en un espacio infinito. El proceso que sigue es el siguiente: El modelo selecciona un conjunto de datos de entrada y predice para cada uno cuál de las dos clases posibles comprende la entrada. Este modelo de aprendizaje está basado en Java y es utilizado tanto en problemas de regresión como de clasificación.

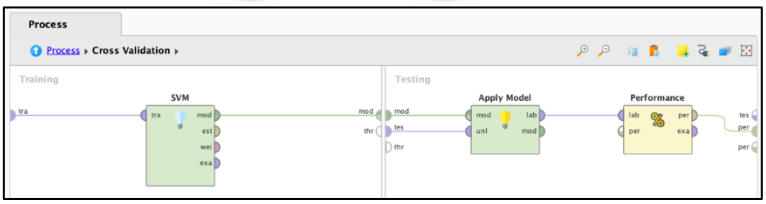


Figura 49: Máquinas de Vectores de Soporte (SVM)

Fuente: RapidMiner

La herramienta RapidMiner permite gestionar algunos parámetros que se detallarán a continuación:

- **Kernel type**

  En este parámetro se ha seleccionado la opción dot, que consiste en que el núcleo de puntos estará definido por $k(x, y) = x * y$.

- **C**

  Hace referencia a la constante de complejidad del modelo, el cual indica cuánta tolerancia al error tendrá el modelo. A más alto el valor de este parámetro, tendrá límites más suaves. Para la construcción del modelo en la herramienta, se ha definido con valor 0.0.

- **Convergence epsilon**

  Es el parámetro de optimización del modelo. Se definió como 0.001.

- **L pos**

  Este parámetro es parte de la función de pérdida del modelo, que indica el factor para la constante de complejidad en positivos. Por defecto se colocó valor 1.0.

- **L neg**

  Este parámetro es parte de la función de pérdida del modelo, que indica el factor para la constante de complejidad en negativos. Por defecto se colocó valor 1.0.

- **Epsilon**

  Este parámetro es parte de la función de pérdida del modelo, indica la constante de insensibilidad. Por defecto se colocó el valor 0.0.

- **Epsilon plus**

  Este parámetro es parte de la función de pérdida del modelo, indica la épsilon para los positivos. Por defecto se colocó el valor 0.0.

- **Epsilon minus**

  Este parámetro es parte de la función de pérdida del modelo, indica la épsilon para los negativos. Por defecto se colocó el valor 0.0.

Quedando la configuración de la siguiente manera:

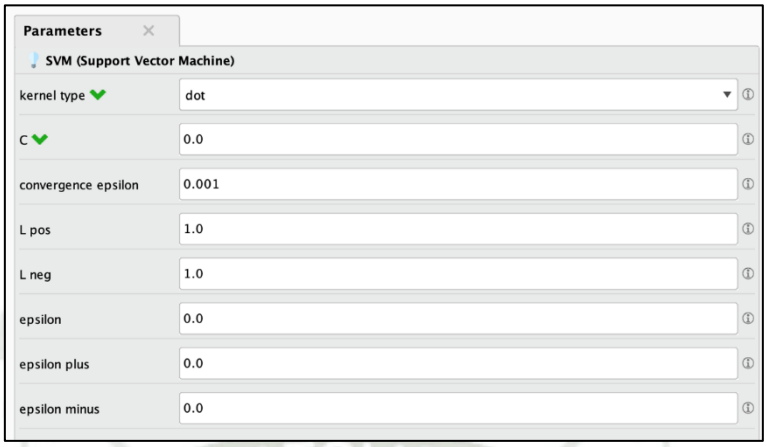


Figura 50: Configuración de parámetros de modelo Máquinas de Vectores de Soporte (SVM)

Fuente: RapidMiner

Al ejecutar el modelo en la herramienta, se obtuvieron los siguientes resultados:

Tabla 40: Resultado de métricas aplicadas a modelo Máquinas de Vectores de Soporte (SVM)

Métricas	**SVM**	
	%	**micro average**
Accuracy	100.00%+/-0.00%	100.00%
classification_error	0.00%+/-0.00%	0.00%
kappa	1.000+/-0.000	1
spearman_rho	1.000+/-0.000	10
absolute_error	0.189+/-0.006	0.189+/-0.103
relative_error	8.92%+/-0.61%	18.92%+/-10.29%
root_mean_squared_error	0.215+/-0.004	0.215+/-0.000
correlation	1.000+/-0.000	1

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.4. Análisis bayesiano

Es un modelo utilizado con mayor frecuencia en detecciones de spam y análisis de recomendaciones, ya que es capaz de construir un buen modelo con pocos datos a través de la probabilidad gaussiana.

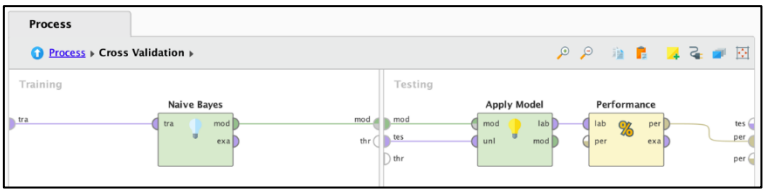


Figura 51: Análisis bayesiano

Fuente: RapidMiner

La herramienta RapidMiner en este modelo no cuenta con parámetros editables. Por lo que luego de crear el modelo, y ejecutarlo, se obtuvieron los siguientes resultados indicados en la Tabla 41:

Tabla 41: Resultado de métricas aplicadas a modelo Análisis Bayesiano

Métricas	**Naive Bayes**	
	%	**micro average**
accuracy	99.80%+/-0.28%	99.80%
classification_error	0.20%+/-0.28%	0.20%
kappa	0.990+/-0.013	0.991
spearman_rho	0.991+/-0.013	9.906
absolute_error	0.010+/-0.003	0.010+/-0.051
relative_error	0.99%+/-0.28%	0.99%+/-5.08%
root_mean_squared_error	0.048+/-0.020	0.052+/-0.000
correlation	0.991+/-0.013	0.991

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.5. Regresión logística

Este modelo de forma predeterminada utiliza un número recomendado de subprocesos y e inicia con un clúster. Puede utilizar datos de entrenamiento con etiquetas binomiales y atributos de características nominales o numéricas.

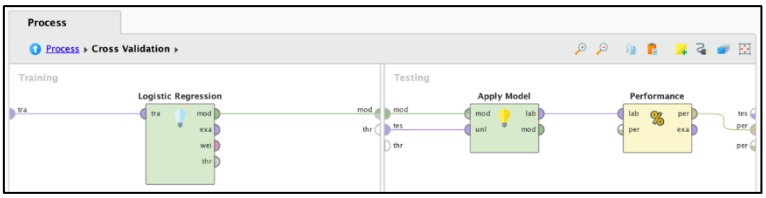


Figura 52: Regresión logística

Fuente: RapidMiner

La herramienta RapidMiner permite gestionar algunos parámetros:

Figura 53: Configuración de parámetros de modelo Regresión logística

Fuente: RapidMiner

- **Solver**

  Hace referencia al solver que utilizará el modelo. Para este caso se seleccionó IRLSM, que tiene la particularidad de ser rápido en problemas con pocos predictores.

- **Reproducible**

  Este parámetro permite que la construcción del modelo sea reproducible. Por defecto se seleccionó falso.

- **Use regularization**

  Este parámetro indica si se debe utilizar regularización. Para este caso no se necesita, así que se seleccionó falso.

- **Standardize**

  Este parámetro indica si se debe estandarizar los atributos del dataset a fin de que tengan una media cero y varianza unitaria. Para este caso no se necesita, así que se seleccionó falso.

- **Non-negative coefficients**

  Este parámetro indica si se debe restringir los coeficientes para que no sean negativos. Para este caso se necesita, así que se seleccionó verdadero.

Al momento de ejecutar el proceso, se obtienen los siguientes resultados indicados en la Tabla 42:

Tabla 42: Resultado de métricas aplicadas a modelo Regresión Logística

Métricas	Logistic Regression	
	%	micro average
Accuracy	100.00%+/-0.00%	100.00%
classification_error	0.00%+/-0.00%	0.00%
kappa	1.000+/-0.000	1
spearman_rho	1.000+/-0.000	10
absolute_error	0.189+/-0.006	0.189+/-0.103
relative_error	8.92%+/-0.61%	18.92%+/-10.29%
root_mean_squared_error	0.215+/-0.004	0.215+/-0.000
correlation	1.000+/-0.000	1

Fuente: Propia

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.6. Regresión lineal

Es un modelo que busca modelar la relación entre una variable escalar y varias explicativas ajustando una ecuación lineal a los datos. Utiliza el criterio de Akaike, que es una medida de ajuste de modelos estadísticos, para la selección del modelo.

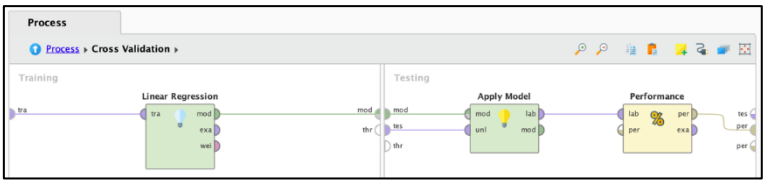


Figura 54: Regresión lineal

Fuente: RapidMiner

La herramienta RapidMiner permite gestionar algunos parámetros que se detallarán a continuación:

- **Min tolerance**

  Este parámetro indica la tolerancia mínima para eliminar características. Para la construcción del modelo en RapidMiner, se colocó el valor de 0.05.

- **Ridge**

  Indica el parámetro de ridge a utilizar en la regresión. Por defecto el valor es 1.0E-8.

Quedando la configuración de la siguiente manera:

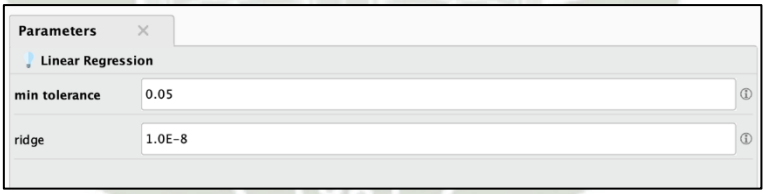


Figura 55: Configuración de parámetros de modelo Regresión Lineal

Fuente: RapidMiner

Al momento de ejecutar el proceso, se obtienen los siguientes resultados indicados en la Tabla 43:

Tabla 43: Resultado de métricas aplicadas a modelo Regresión Lineal

Métricas	Linear regression	
	%	micro average
accuracy	98.86%+/-1.00%	98.86%
classification_error	1.14%+/-1.00%	1.14%
kappa	0.942+/-0.054	0.944
spearman_rho	0.944+/-0.050	9.445
absolute_error	0.399+/-0.002	0.399+/-0.045
relative_error	39.89%+/-0.24%	39.89%+/-4.50%
root_mean_squared_error	0.401+/-0.002	0.401+/-0.000
correlation	0.944+/-0.050	0.945

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.3.7. K-Vecinos más cercanos

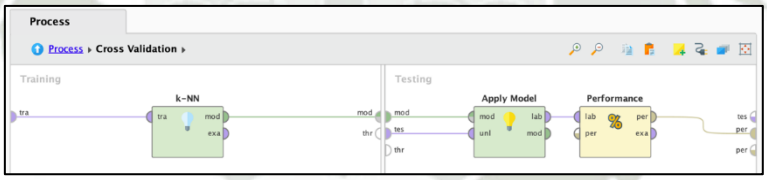


Figura 56: K-Vecinos más cercanos

Fuente: RapidMiner

Este modelo está basado en comparaciones con los k ejemplos de entrenamiento que se tiene a fin de buscar los más cercanos en términos de un espacio n-dimensional.

La herramienta RapidMiner permite gestionar algunos parámetros que se detallarán a continuación:

- **K**

  Este parámetro indica el valor k para el modelo, usualmente es un número pequeño e impar. Para el modelo se utilizó el valor 5.

- **Weighted vote**

  Este parámetro indica si las distancias influirán en la predicción. Para el modelo se seleccionó que sí se tendrán en cuenta.

- **Measure types**

  Este parámetro indica qué tipo de medida se utilizará para encontrar a los k-n vecinos. Para el modelo se escogió MixedMeasures, que indica que se utilizarán valores nominales y numéricos.

- **Mixed Measure**

  En este parámetro se indica qué medida se utilizará. Por defecto se utilizará mixedEuclideanDistance.

Quedando la configuración de la siguiente manera:

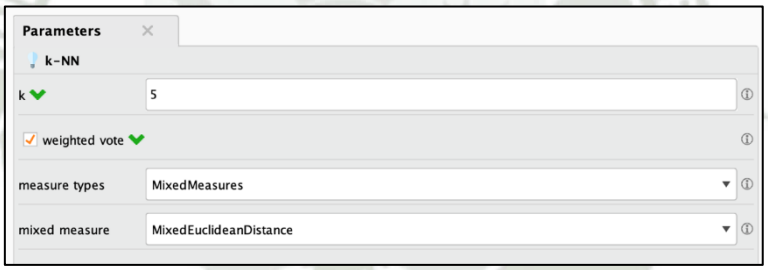


Figura 57: Configuración de parámetros de modelo K-Vecinos más cercanos

Fuente: RapidMiner

Al momento de ejecutar el proceso, se obtienen los siguientes resultados indicados en la Tabla 44:

Tabla 44: Resultado de métricas aplicadas a modelo K-Vecinos más cercanos

Métricas	**KNN**	
	%	**micro average**
Accuracy	99.38% +/- 0.50%	99.38%
classification_error	0.62% +/- 0.50%	0.62%
kappa	0.970 +/- 0.025	0.97
spearman_rho	0.970 +/- 0.024	9.703
absolute_error	0.012 +/- 0.003	0.012 +/- 0.071
relative_error	1.16% +/- 0.28%	1.16% +/- 7.12%
root_mean_squared_error	0.071 +/- 0.015	0.072 +/- 0.000
correlation	0.970 +/- 0.024	0.97

Fuente: RapidMiner

Los valores obtenidos se evaluarán en el punto 3.4.4. Evaluación del modelo.

3.4.4. Evaluación del modelo

A pesar de que en la fase 3.5. Evaluación de la metodología CRISP-DM se realizará un análisis más profundo de los modelos generados anteriormente, en esta subsección el análisis está más orientado a los resultados obtenidos en cada una de las ejecuciones de los modelos, mientras que en la siguiente fase estará más orientado a los objetivos planteados.

En términos de resultados, se tendrá en cuenta los resultados de las siguientes métricas:

a) Accuracy

Porcentaje o número de ejemplos correctos.

b) classification_error

Porcentaje o número de ejemplos incorrectos.

c) Kappa

Evalúa si concuerda la clasificación obtenida con el dataset.

d) spearman_rho

Mide la asignación entre dos variables cuando la asociación es no lineal.

e) absolute_error

Promedio de las diferencias absolutas entre los valores.

f) relative_error

Rendimiento de modelos predictivos.

g) root_mean_squared_error

Cálculo de la función de error.

h) Correlation

Intensidad con la que se relacionan el valor predicho y original.

En la siguiente tabla se observa los resultados para los 7 modelos creados en la herramienta RapidMiner, donde se evaluarán las métricas en el siguiente orden:

- Decision tree
- Neural Network
- SVM
- Naive Bayes
- Logistic regression
- Linear regression
- KNN

Tabla 45: Resultados de métricas aplicadas

Métricas	1	2	3	4	5	6	7
Accuracy	100.00%+/-0.00%	95.70% +/- 0.98%	100.00% +/- 0.00%	99.80%+/-0.28%	100.00%+/-0.00%	98.86%+/-1.00%	99.38% +/- 0.50%
classification_error	0.00%+/-0.00%	4.30% +/- 0.98%	0.00% +/- 0.00%	0.20%+/-0.28%	0.00%+/-0.00%	1.14%+/-1.00%	0.62% +/- 0.50%
kappa	1.000+/-0.000	0.765 +/- 0.059	1.000 +/- 0.000	0.990+/-0.013	1.000+/-0.000	0.942+/-0.054	0.970 +/- 0.025
spearman_rho	1.000+/-0.000	0.781 +/- 0.052	1.000 +/- 0.000	0.991+/-0.013	1.000+/-0.000	0.944+/-0.050	0.970 +/- 0.024
absolute_error	0.000+/-0.000	0.109 +/- 0.011	0.189 +/- 0.006	0.010+/-0.003	0.000+/-0.000	0.399+/-0.002	0.012 +/- 0.003
relative_error	0.00%+/-0.00%	10.95% +/- 1.14%	8.92% +/- 0.61%	0.99%+/-0.28%	0.00%+/-0.00%	39.89%+/-0.24%	1.16% +/- 0.28%
root_mean_squared_error	0.000+/-0.000	0.187 +/- 0.009	0.215 +/- 0.004	0.048+/-0.020	0.000+/-0.000	0.401+/-0.002	0.071 +/- 0.015
correlation	1.000+/-0.000	0.781 +/- 0.052	1.000 +/- 0.000	0.991+/-0.013	1.000+/-0.000	0.944+/-0.050	0.970 +/- 0.024

Fuente: RapidMiner

Utilizando los valores de la tabla se procederá a hacer la primera evaluación de cada uno de los modelos.

El primer modelo y quinto modelo creados en la herramienta fue Decision Tree y Logistic Regression, los cuales mostraron tener valores perfectos en cada una de las métricas. Es decir, tuvieron un valor de 100% de accuracy, valor 0% en classification error y relative error; valor 1 en kappa, spearman rho y correlation; y valor 0 en absolute error y root mean squared error. Esto podría interpretarse como valores perfectos, pero también demostraría que estos modelos pueden llegar a un estado de underfitting, es decir, que coinciden con todos los ejemplos de entrenamiento, pero al momento de ingresar valores nuevos, no tendrían un buen rendimiento.

El segundo modelo creado en la herramienta fue Neural Network, el cual mostró tener valores considerados aceptables y será el escogido para solucionar el objetivo principal planteado. Tiene un valor de accuracy de 95.70% y classification error 4.30%, considerados aceptables; un valor kappa de 0.765, que dentro de la interpretación de este valor es considerado un modelo bueno; un valor de spearman rho de 0.781, considerado bueno; según el error absoluto que es 0.109 se puede visualizar que no existirá overfitting ni underfitting; el relative

error es 10.95%, el cual según la interpretación es un modelo considerado bueno; el valor de correlation es de 0.781, el cual indica que la asociación entre los valores predichos y reales es positiva; el valor de la métrica root mean squared error es 0.187, considerado bueno para el modelo.

EL tercer modelo creado fue SVM, el cual si bien es cierto presenta buenos valores, es muy costos en tiempo de ejecución. Los valores de accuracy, classification error, kappa y spearman rho son perfectos según sus interpretaciones. Lo mismo pasa con el séptimo modelo creado, KNN, si bien los valores obtenidos de las métricas son óptimos, el modelo es muy costoso ya que constantemente se está ejecutando.

El cuarto modelo creado en la herramienta fue Naive Bayes, el cual tiene un valor de accuracy y classification error de 99.80% y 0.20% respectivamente; valor kappa de 0.990, considerado según la interpretación como muy bueno; el valor de relative error es 0.99%, considerado razonable; y un valor de correlation de 0.991, que según la interpretación de la métrica se puede considerar que el modelo tiene una asociación positiva. El modelo demostró tener un buen comportamiento con los datos, pero la desventaja de este modelo es que puede ocurrir lo que se denomina frecuencia cero, que es cuando el modelo reconoce una nueva categoría y no puede predecir, por lo que entrega una probabilidad 0.

El sexto modelo creado en la herramienta fue linear regression, el cual mostró tener valores considerados normales. Tiene un valor de accuracy de 98.86% y classification error 1.14%, considerados aceptables; un valor kappa de 0.942, que dentro de la interpretación de este valor es considerado un modelo muy bueno; un valor de spearman rho de 0.970, considerado bueno; según el error absoluto que es 0.399 se puede visualizar que no existirá overfitting ni underfitting; el relative error es 39.89%; el valor de correlation es de 0.994, el cual indica que la asociación entre los valores predichos y reales es positiva; el valor de la métrica root mean squared error es de 0.401, todos considerados buenos para el modelo.

3.5. Evaluación

En esta fase de la metodología CRISP-DM se evaluará a más profundidad los modelos generados en la fase anterior, pero con un enfoque de los objetivos planteados para este trabajo. Una vez realizada esta evaluación se revisará el proceso a seguir y se determinarán los pasos siguientes para la fase de implementación.

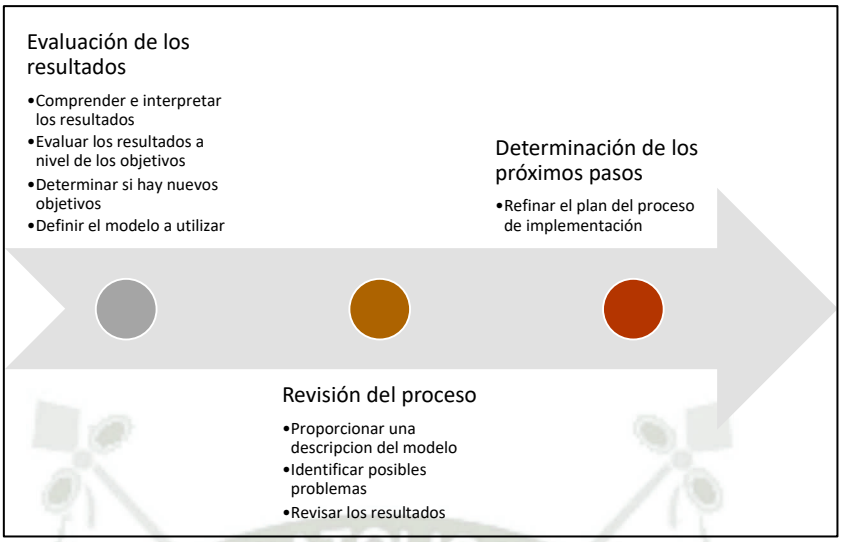


Figura 58: Fase 5 Evaluación

Fuente: Propia

3.5.1. Evaluación de los resultados

Desde el punto de vista del problema, se había indicado en el apartado 1.1.2. como objetivo principal "Predecir la obesidad en adolescentes utilizando una aplicación basada en aprendizaje automático", para poder lograr esto, es necesario tener en cuenta las diferentes métricas del aprendizaje automático que se ejecutan al modelo del entrenamiento y pruebas a fin de poder obtener un modelo con una fiabilidad "aceptable". Es por eso por lo que, en base a lo obtenido en el apartado anterior, se realizará una evaluación de todos los modelos para así descartar los que no cumplan los valores aceptables en las métricas.

Modelo aprobado

Por el análisis realizado en el apartado 3.4.4. Evaluación del modelo, el modelo que se utilizará en la fase 6 es Neural Network, ya que según las métricas con las que se ha evaluado cada uno de los modelos, este es considerado como un modelo óptimo para llegar al objetivo planteado, además de tener en cuenta de las ventajas que ofrece el modelo: Alta tolerancia a fallos, facilidad al momento de manejar características complejas en los datos, adaptación de pesos a manera que se utiliza, así como reconocimiento de patrones en la etapa de producción. Los modelos Decision Tree y Logistic Regression son descartados por posible underfitting, el modelo SVM por sus altos costos de ejecución, al igual que el modelo KNN; el modelo Naive Bayes por posible

"frecuencia cero" a futuro y el modelo Linear Regression sería considerado como una segunda opción para la solución del objetivo, ya que esta propuesta no obliga a utilizar redes neuronales en este tipo de soluciones.

3.5.2. Revisión del proceso

El desarrollo de cada una de las fases de la metodología CRISP-DM para lograr el objetivo planteado se ha realizado tal y como estaba previsto. En esta subsección se realizará un resumen de todo lo desarrollado, a fin de revisar y detectar en caso se esté omitiendo algún factor importante.

La primera elección que se hizo fue entre utilizar modelos de aprendizaje automático o aprendizaje profundo para llegar al objetivo. Se escogió utilizar **modelos de aprendizaje automático**, ya que según el dataset que se tiene, se debía realizar un procesamiento y selección de características manual previo antes de entrenar el modelo. Además, a través del modelo se buscaba analizar los datos procesados, aprender y tomar decisiones (predecir) en base a lo aprendido. Se escogió aprendizaje automático, también, por la simplicidad que se tiene para crear los modelos.

Luego de escoger qué tipo de modelo se utilizará, se realizó el preprocesamiento de los datos: Cambio de etiquetas a tipo numérico y análisis de cada uno de los atributos a fin de ver cuáles son esenciales y cuáles deben ser omitidos. Se realizó el cambio de etiquetas a datos numéricos para poder encontrar una homogeneidad en los datos que se tenía. A partir de esto, se realizó la segunda elección, que fue si utilizar modelos de regresión o de clasificación según el objetivo que se planteó y el dataset que se tiene, en donde se decidió utilizar **modelos de clasificación**, ya que el resultado que se busca obtener del modelo es una etiqueta discreta, es decir, el resultado será parte de un conjunto finito de resultados.

A lo largo de la fase 3 de la metodología CRISP-DM, se realizó la preparación de los datos, donde a partir de los 5 atributos iniciales que se tenía en el dataset, se obtuvieron 3 atributos derivados y se omitieron para el modelo 3 atributos, los cuales eran equivalentes y de mantenerlos, se tenía el riesgo de que el modelo al momento de entrenarlo tuviera ruido.

Finalmente se realizó el modelado dentro de la herramienta RapidMiner de 7 modelos de aprendizaje automático: Arboles de decisión, redes neuronales, máquinas de vectores de soporte (SVM), análisis bayesiano, regresión logística, regresión lineal y KNN. De los mencionados

anteriormente, a partir de las métricas aplicadas en cada una de las ejecuciones de los modelos y de los resultados obtenidos, se decidió utilizar redes neuronales por los resultados aceptables obtenidos y por las ventajas tiene utilizar este modelo.

El modelo por implementar en la fase 6 será una red neuronal que constará 3 capas: Una capa de entrada, una capa oculta y una capa de salida.

La capa de entrada constará de 4 nodos: uno por cada atributo del dataset, mientras que la capa de salida constará de 2 nodos: uno por cada clase del atributo de salida. Dentro de la capa oculta constará de 5 nodos, los cuales se encargarán de procesar los datos a fin de dar una predicción.

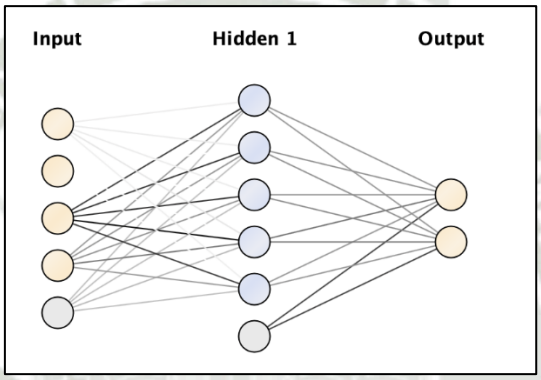


Figura 59: Esquema de la red neuronal propuesta

Fuente: RapidMiner

Cada uno de los nodos dentro de la capa oculta tendrán función de activación sigmoide $\left(f(x) = \frac{1}{1-e^{x}}\right)$, el cual transforma los valores ingresados por el nodo en valores entre 0 y 1 y es considerada como una función de activación ideal para la última capa de los modelos de predicción.

3.5.3. Determinación de los próximos pasos

La siguiente fase por realizar será el desarrollar el modelo elegido para lograr los objetivos planteados.

3.6. Implementación

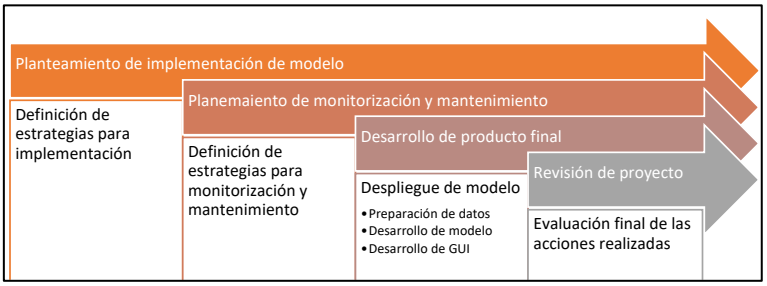


Figura 60: Fase 6 Implementación

Fuente: Propia

En esta última fase de la metodología CRISP-DM se explicará cómo se realizará la implementación del modelo, se realizará la implementación del modelo, para finalmente exponer los resultados en el siguiente capítulo del documento.

3.6.1. Planeamiento de implementación de modelo

Para el desarrollo del modelo elegido (redes neuronales) se utilizará el lenguaje de programación **Python**, que es considerado como uno de los mejores lenguajes para inteligencia artificial gracias a su versatilidad, dinamismo y facilidad para programar. Este lenguaje cuenta con una amplia biblioteca de librerías, tales como Numpy, Scipy, Matplotlib, Pandas y Scikit-learn, entre otros, para el desarrollo de modelos de aprendizaje automático.

Para el desarrollo del modelo, se utilizará una herramienta en la nube, a fin de poder aprovechar las ventajas que ofrece el trabajar oncloud. Para ello, se ha creado una tabla comparativa con 6 herramientas que permiten realizar este trabajo.

Tabla 46: Comparativa de herramientas oncloud

Plataforma	Características	Lenguajes soportados	Precio por uso
Google Colab	• Permite escribir y documentar código • Permite trabajo colaborativo • Permite el uso de GPU de Google	• Python	Gratuito
Notebooks Azure	• Acceso a entorno en Linux • Permite enlace con GitHub	• Python 2 • Python 3 • F# • R	Gratuito
Kaggle	• Permite construir bases de datos • Permite explorar y construir modelos • Permite realizar competencias de programación	• Python	Gratuito
Amazon SageMaker	• Permite crear modelos de aprendizaje automático	• Python	Pago
IBM Data Platform Notebooks (Watson Studio)	• Permite crear modelos de aprendizaje automático	• Python • Scala • R	Pago
Jupyter org	• Permite crear y compartir modelos de aprendizaje automático	• Python • Scala • R • Ruby • Go	Gratuito

Fuente: Propia

Luego de realizar la comparación de las herramientas disponibles, se decide utilizar **Google Colab**, por el procesador que ofrece (GPU), además de la facilidad del trabajo colaborativo y que es una herramienta oncloud gratuita.

Respecto al dataset, a fin de evitar un overfitting (sobre entrenamiento) o underfitting (sobajaste) por las características que presenta, se realizará la partición de la data de la siguiente manera: **70% entrenamiento (60% entrenamiento y 10% validación), 30% pruebas**.

3.6.2. Planeamiento de la monitorización y mantenimiento

Es importante para el modelo definir de qué manera es que se va a realizar la supervisión y mantenimiento del modelo durante la etapa de entrenamiento y pruebas. Esto se implementará

con el modelo para que, cada vez que se ejecute, se tenga los valores de: Accuracy, root mean squared error, mean squared error, AUC, mean absolute error y recall; de cara a ver que sigue siendo optimo.

3.6.3. Desarrollo de producto final

A lo largo de este apartado se desarrollará el modelo, como se mencionó anteriormente, en Google Colab.

3.6.3.1. Despliegue del modelo

La primera parte por desarrollar es la red neuronal, la cual se realizará en Google Colab. Este desarrollo estará dividido en 3 secciones principales: Preparación de datos, desarrollo del modelo y prueba manual del modelo.

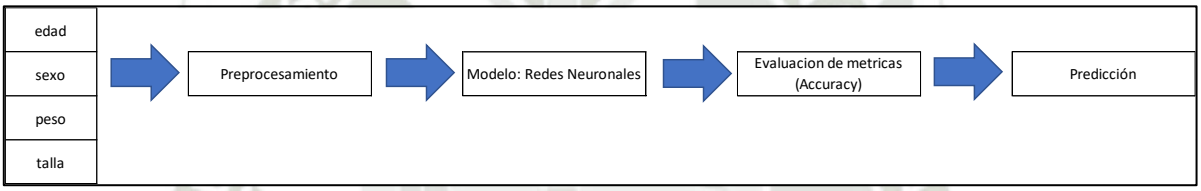


Figura 61: Predicción de obesidad a partir de dataset

Fuente: Propia

Como pasos previos se realizará la importación de las librerías necesarias para desarrollar el modelo.

```
[ ]  # Importing the general libraries
     import numpy as np
     import matplotlib.pyplot as plt
     import pandas as pd

[ ]  # Importing the Keras libraries and packages
     import keras
     from keras.models import Sequential
     from keras.layers import Dense
     from keras.layers import Dropout
     from keras.layers.merge import concatenate
```

Figura 62: Importación de librerías

Fuente: Google Colab

En esta etapa se harán dos análisis: El primero con 3068 registros de personas entre 4 y 20 años; y el segundo con 1763 registros de personas entre 12 a 18 años

3.6.3.1.1. Análisis con registros de personas entre 4 a 20 años

Dentro de este apartado se importarán y normalizarán los datos del dataset que fueron previamente analizados y clasificados en la fase 2 y preparados en la fase 3 de la metodología.

3.6.3.1.1.1. Preparación de datos

Lectura de datos y asignación de columnas

El paso previo a la lectura a través de código fue la importación del archivo.csv a Google Colab, viéndose los datos de la siguiente manera:

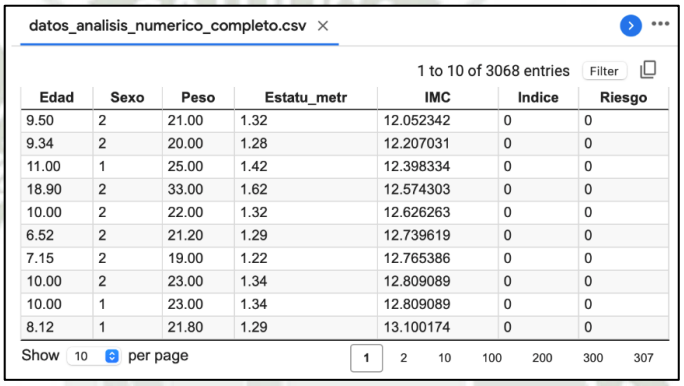


Figura 63: Visualización de datos en Google Colab

Fuente: Google Colab

Los datos se importarán a través del método read_csv perteneciente a la librería Pandas, importada previamente.

```
# Importing the dataset
dataset = pd.read_csv('/content/datos_analisis_numerico_completo.csv')
X = dataset.iloc[:, 0:4].values
y = dataset.iloc[:, 6].values
```

Figura 64: Importación de dataset

Fuente: Google Colab

Además, a través del método iloc, también de la librería Pandas, se asignará las columnas que corresponden a X (atributos) y (resultado).

Limpieza de datos

En el análisis de los datos durante la fase 2, se pudo visualizar que no hay registros con valores nulos o vacíos, por lo que se considera que el dataset está limpio. Además, es necesario rellenar todos los atributos para poder realizar la predicción.

```
[33] print(dataset.head(1))

   Edad  Sexo  Peso  Estatu_metr        IMC  Indice  Riesgo
0   9.5     2  21.0         1.32  12.052342       0       0
```

Figura 65: Visualización de dataset importado

Fuente: Google Colab

División de datos

En este punto, se realizará la división del dataset para poder tener data de entrenamiento y data de test. Esto se realizará a través de método train_test_split del paquete Sklearn. A este método se le indicará que la cantidad de datos para test debe ser el 30% del total del dataset, además, se le indica que el valor de aleatoriedad en los datos, el cual es 0.

```
# Splitting the dataset into the Training set and Test set (1697)
from sklearn.model_selection import train_test_split

X_rest, X_test, y_rest, y_test = train_test_split(X, y,test_size=0.30)
X_train, X_valid, y_train, y_valid = train_test_split(X_rest, y_rest, test_size=0.143)
```

Figura 66: Preparación de data de entrenamiento y prueba

Fuente: Google Colab

Actualmente, en el test se tienen 3068 registros, por lo que la cantidad de datos para el entrenamiento y validación que se tiene es 2147, que equivale al 70%, de los cuales el 60% fue para entrenamiento, es decir 1840 registros, y 10% para validación, es decir 307 registros; y la cantidad de datos para test es 921, equivalente al 30% del total. A través del método shape se puede comprobar cuántos datos tiene cada array o matriz, según corresponda.

Normalización

A través del método fit_transform de la librería Sklearn se estandarizarán los datos tanto de entrenamiento como de prueba.

```
[10] # Feature Scaling
     from sklearn.preprocessing import StandardScaler
     sc = StandardScaler()
     X_train = sc.fit_transform(X_train)
     X_test = sc.fit_transform(X_test)
```

Figura 67: Normalización de datos

Fuente: Google Colab

3.6.3.1.1.2. Desarrollo del modelo

En este punto ya se tiene los datos listos para ser procesados a través del modelo para entrenamiento. A través de la librería Graphviz se graficó el esquema de cómo es que será la distribución de la red neuronal.

```
[ ] from graphviz import Digraph

    dot = Digraph(comment='NeuralNetwork')

    dot.node('A', 'Edad')
    dot.node('B', 'Sexo')
    dot.node('C', 'Peso')
    dot.node('D', 'Estatura_metr')
    dot.node('E', 'IMC')
    dot.node('F', 'Indice')

    dot.node('H','Nodo01')
    dot.node('I','Nodo02')
    dot.node('J','Nodo03')
    dot.node('K','Nodo04')
    dot.node('L','Nodo05')

    dot.node('G', 'Riesgo')

    dot.edges(['AH', 'AI', 'AJ', 'AK', 'AL',
               'BH', 'BI', 'BJ', 'BK', 'BL',
               'CH', 'CI', 'CJ', 'CK', 'CL',
               'DH', 'DI', 'DJ', 'DK', 'DL',
                'HG','IG','JG','KG','LG'])

    dot.format = 'png'
    dot.render('RedNeuronal', view = True)
```

Figura 68: Gráfico de la red neuronal a implementar

Fuente: Google Colab

Al inicio, se tiene 6 nodos de entrada, de los cuales solo 4 serán los que se conecten con la capa oculta, la cual tendrá 5 nodos por los que procesará los datos, para finalmente conectarse al único nodo de la capa, el cual tiene 2 clases: 0 y 1.

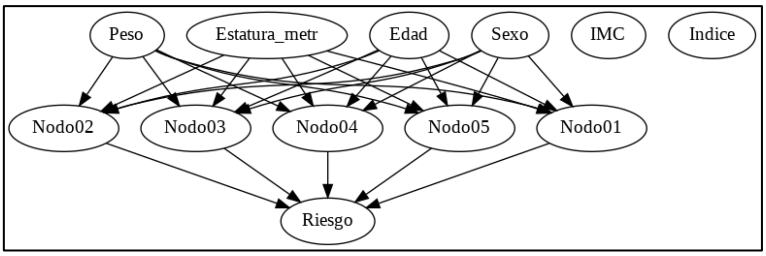


Figura 69: Diagrama de la red neuronal a desarrollar

Fuente: Google Colab

Declaración del modelo

Primero, se creará un modelo vacío, indicando a partir del método Sequential que será un modelo secuencia, es decir, cada capa que se indique irá después de la anterior.

```
[ ]  # Initialising the ANN
     classifier = Sequential()
```

Figura 70: Inicialización de modelo

Fuente: Google Colab

Como se indicó anteriormente, la red neuronal tendrá: 1 capa de entrada con 4 nodos, 1 capa oculta y 1 capa de salida. El siguiente paso será crear las 3 capas en la red, para ello, se utilizará el método add, en el cual, para la primera capa oculta se indicará que tendrá 5 nodos, y al mismo tiempo, a través del valor de imput_dim, se estará indicando que la capa de entrada tiene 4 nodos y que esta capa oculta tendrá una función de activación sigmoid. Además, para evitar un overfitting, se le incluirá el método Dropout.

```
[17] # Adding the input layer and the first output layer
     classifier.add(Dense(5, input_dim=4, activation = 'sigmoid'))
     classifier.add(Dropout(0.95))
```

Figura 71: Creación de capa oculta

Fuente: Google Colab

A continuación, se creará la capa de salida, la cual tendrá un solo nodo y una función de activación sigmoid.

```
[ ]  # Adding the input layer and the first output layer
     classifier.add(Dense(1, activation = 'sigmoid'))
```

Figura 72: Creación de capa de salida

Fuente: Google Colab

Para la compilación de la red, se indicará que el tipo de pérdida será a través de binary_crossentropy, el optimizador para la compilación se dará a través del algoritmo Adam y que, al momento de entrenar la red, la evaluación se hará a través de las siguientes métricas: Accuracy, root mean squared error, mean squared error, AUC, mean absolute error y recall. La mayoría de ellas fueron parte de la evaluación en la herramienta RapidMiner de los modelos planteados previo a la implementación en la fase 4.

```
[ ]  # Compiling the ANN
     classifier.compile(loss='binary_crossentropy', optimizer='adam',
                        metrics=['accuracy','RootMeanSquaredError',
                                 'MeanSquaredError','AUC','MeanAbsoluteError','Recall'
                                 ]
                        )
```

Figura 73: Compilación del modelo

Fuente: Google Colab

Entrenamiento del modelo

Para la fase de entrenamiento de la red neuronal, se utilizará el método fit, al cual se le indicará los valores de entrenamiento, tanto entrada como salida, el número de iteraciones que tendrá que hacer (epochs), el número de muestras por actualización de gradiente (batch_size) y los datos que se utilizarán para la validación (validation_data).

```
[ ]  # Fitting the ANN to the Training set
     classifier.fit(np.array(X_train), np.array(y_train), epochs=25,
                    batch_size=10, validation_data=(X_test,y_test), verbose=1)
```

Figura 74: Entrenamiento de red neuronal

Fuente: Google Colab

Predicción del modelo

Finalmente, para evaluar la red neuronal, se creará una matriz de confusión y se imprimirá para poder visualizarla.

```
[ ]  # Making the Confusion Matrix
     from sklearn.metrics import confusion_matrix
     cm = confusion_matrix(y_test, y_pred)

[ ]  print(cm)
```

Figura 75: Creación de matriz de confusión

Fuente: Google Colab

3.6.3.1.1.3. Prueba manual del modelo

Además de la matriz de confusión que se generará luego del entrenamiento, se realizarán pruebas manuales a fin de ver que estén realizando las predicciones correctamente.

```
[ ]  # Predicting a single new observation
     new_prediction = classifier.predict(sc.transform(np.array([[24,2,58,1.58]])))
     new_prediction = (new_prediction > 0.5)

     new_prediction1 = classifier.predict(sc.transform(np.array([[54,2,900,1.38]])))
     new_prediction1 = (new_prediction1 > 0.5)
```

Figura 76: Predicción de casos manuales

Fuente: Google Colab

Y para poder visualizar el resultado, se imprimirá por consola el resultado.

```
[ ]  print(new_prediction)
     print(new_prediction1)

     [[False]]
     [[ True]]
```

Figura 77: Impresión de resultados de pruebas

Fuente: Google Colab

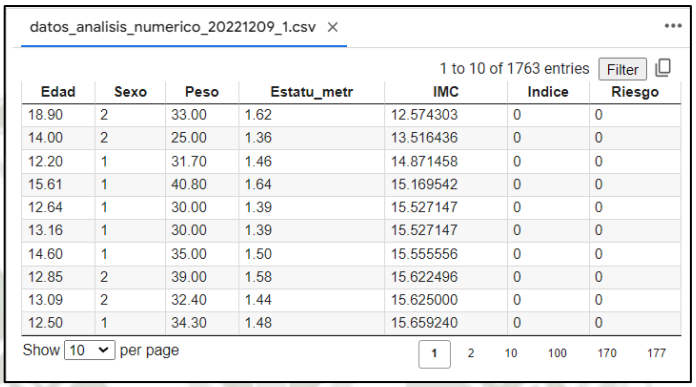
3.6.3.1.2. Análisis con registros de personas entre 12 a 18 años

Dentro de este apartado se importarán y normalizarán los datos del dataset que fueron previamente clasificados en la fase 2 y preparados en la fase 3 de la metodología.

3.6.3.1.2.1. Preparación de datos

Lectura de datos y asignación de columnas

El paso previo a la lectura a través de código fue la importación del archivo.csv a Google Colab, viéndose los datos de la siguiente manera:



datos_analisis_numerico_20221209_1.csv

1 to 10 of 1763 entries Filter

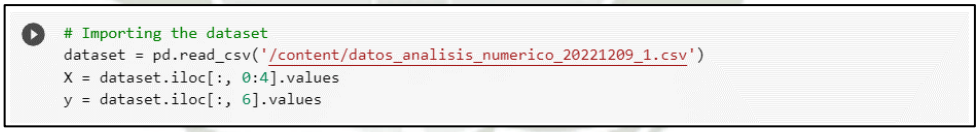
Edad	Sexo	Peso	Estatu_metr	IMC	Indice	Riesgo
18.90	2	33.00	1.62	12.574303	0	0
14.00	2	25.00	1.36	13.516436	0	0
12.20	1	31.70	1.46	14.871458	0	0
15.61	1	40.80	1.64	15.169542	0	0
12.64	1	30.00	1.39	15.527147	0	0
13.16	1	30.00	1.39	15.527147	0	0
14.60	1	35.00	1.50	15.555556	0	0
12.85	2	39.00	1.58	15.622496	0	0
13.09	2	32.40	1.44	15.625000	0	0
12.50	1	34.30	1.48	15.659240	0	0

Show 10 per page 1 2 10 100 170 177

Figura 78: Visualización de datos en Google Colab

Fuente: Google Colab

Los datos se importarán a través del método read_csv perteneciente a la librería Pandas, importada previamente.



```
# Importing the dataset
dataset = pd.read_csv('/content/datos_analisis_numerico_20221209_1.csv')
X = dataset.iloc[:, 0:4].values
y = dataset.iloc[:, 6].values
```

Figura 79: Importación de dataset

Fuente: Google Colab

Además, a través del método iloc, también de la librería Pandas, se asignará las columnas que corresponden a X (atributos) y (resultado).

Limpieza de datos

En el análisis de los datos durante la fase 2, se pudo visualizar que no hay registros con valores nulos o vacíos, por lo que se considera que el dataset está limpio. Además, es necesario rellenar todos los atributos para poder realizar la predicción.

```
[13] print(dataset.head(1))

   Edad  Sexo  Peso  Estatu_metr        IMC  Indice  Riesgo
0  18.9     2  33.0         1.62  12.574303       0       0
```

Figura 80: Visualización de dataset importado

Fuente: Google Colab

División de datos

En este punto, se realizará la división del dataset para poder tener data de entrenamiento y data de test. Esto se realizará a través de método train_test_split del paquete Sklearn. A este método se le indicará que la cantidad de datos para test debe ser el 30% del total del dataset, además, se le indica que el valor de aleatoriedad en los datos, el cual es 0.

```
# Splitting the dataset into the Training set and Test set (1697)
from sklearn.model_selection import train_test_split

X_rest, X_test, y_rest, y_test = train_test_split(X, y,test_size=0.30)
X_train, X_valid, y_train, y_valid = train_test_split(X_rest, y_rest, test_size=0.143)
```

Figura 81: Preparación de data de entrenamiento y prueba

Fuente: Google Colab

Actualmente, en el test se tienen 3068 registros, por lo que la cantidad de datos para el entrenamiento y validación que se tiene es 2147, que equivale al 70%, de los cuales el 60% fue para entrenamiento, es decir 1840 registros, y 10% para validación, es decir 307 registros; y la cantidad de datos para test es 921, equivalente al 30% del total. A través del método shape se puede comprobar cuántos datos tiene cada array o matriz, según corresponda.

Normalización

A través del método fit_transform de la librería Sklearn se estandarizarán los datos tanto de entrenamiento como de prueba.

```
[10] # Feature Scaling
     from sklearn.preprocessing import StandardScaler
     sc = StandardScaler()
     X_train = sc.fit_transform(X_train)
     X_test = sc.fit_transform(X_test)
```

Figura 82: Normalización de datos

Fuente: Google Colab

3.6.3.1.2.2. Desarrollo del modelo

En este punto ya se tiene los datos listos para ser procesados a través del modelo para entrenamiento. A través de la librería Graphviz se graficó el esquema de cómo es que será la distribución de la red neuronal.

```
[ ] from graphviz import Digraph

    dot = Digraph(comment='NeuralNetwork')

    dot.node('A', 'Edad')
    dot.node('B', 'Sexo')
    dot.node('C', 'Peso')
    dot.node('D', 'Estatura_metr')
    dot.node('E', 'IMC')
    dot.node('F', 'Indice')

    dot.node('H','Nodo01')
    dot.node('I','Nodo02')
    dot.node('J','Nodo03')
    dot.node('K','Nodo04')
    dot.node('L','Nodo05')

    dot.node('G', 'Riesgo')

    dot.edges(['AH', 'AI', 'AJ', 'AK', 'AL',
               'BH', 'BI', 'BJ', 'BK', 'BL',
               'CH', 'CI', 'CJ', 'CK', 'CL',
               'DH', 'DI', 'DJ', 'DK', 'DL',
                'HG','IG','JG','KG','LG'])

    dot.format = 'png'
    dot.render('RedNeuronal', view = True)
```

Figura 83: Gráfico de la red neuronal a implementar

Fuente: Google Colab

Al inicio, se tiene 6 nodos de entrada, de los cuales solo 4 serán los que se conecten con la capa oculta, la cual tendrá 5 nodos por los que procesará los datos, para finalmente conectarse al único nodo de la capa, el cual tiene 2 clases: 0 y 1.

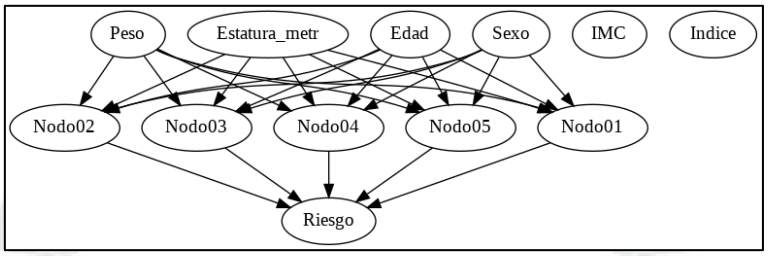


Figura 84: Diagrama de la red neuronal a desarrollar

Fuente: Google Colab

Declaración del modelo

Primero, se creará un modelo vacío, indicando a partir del método Sequential que será un modelo secuencia, es decir, cada capa que se indique irá después de la anterior.

```
[ ]  # Initialising the ANN
     classifier = Sequential()
```

Figura 85: Inicialización de modelo

Fuente: Google Colab

Como se indicó anteriormente, la red neuronal tendrá: 1 capa de entrada con 4 nodos, 1 capa oculta y 1 capa de salida. El siguiente paso será crear las 3 capas en la red, para ello, se utilizará el método add, en el cual, para la primera capa oculta se indicará que tendrá 5 nodos, y al mismo tiempo, a través del valor de imput_dim, se estará indicando que la capa de entrada tiene 4 nodos y que esta capa oculta tendrá una función de activación sigmoid. Además, para evitar un overfitting, se le incluirá el método Dropout.

```
[17] # Adding the input layer and the first output layer
     classifier.add(Dense(5, input_dim=4, activation = 'sigmoid'))
     classifier.add(Dropout(0.95))
```

Figura 86: Creación de capa oculta

Fuente: Google Colab

A continuación, se creará la capa de salida, la cual tendrá un solo nodo y una función de activación sigmoid.

```
[ ]  # Adding the input layer and the first output layer
     classifier.add(Dense(1, activation = 'sigmoid'))
```

Figura 87: Creación de capa de salida

Fuente: Google Colab

Para la compilación de la red, se indicará que el tipo de pérdida será a través de binary_crossentropy, el optimizador para la compilación se dará a través del algoritmo Adam y que, al momento de entrenar la red, la evaluación se hará a través de las siguientes métricas: Accuracy, root mean squared error, mean squared error, AUC, mean absolute error y recall. La mayoría de ellas fueron parte de la evaluación en la herramienta RapidMiner de los modelos planteados previo a la implementación en la fase 4.

```
[ ]  # Compiling the ANN
     classifier.compile(loss='binary_crossentropy', optimizer='adam',
                        metrics=['accuracy','RootMeanSquaredError',
                                 'MeanSquaredError','AUC','MeanAbsoluteError','Recall'
                                 ]
                        )
```

Figura 88: Compilación del modelo

Fuente: Google Colab

Entrenamiento del modelo

Para la fase de entrenamiento de la red neuronal, se utilizará el método fit, al cual se le indicará los valores de entrenamiento, tanto entrada como salida, el número de iteraciones que tendrá que hacer (epochs), el número de muestras por actualización de gradiente (batch_size) y los datos que se utilizarán para la validación (validation_data).

```
[ ] # Fitting the ANN to the Training set
    classifier.fit(np.array(X_train), np.array(y_train), epochs=25,
                   batch_size=10, validation_data=(X_test,y_test), verbose=1)
```

Figura 89: Entrenamiento de red neuronal

Fuente: Google Colab

Predicción del modelo

Finalmente, para evaluar la red neuronal, se creará una matriz de confusión y se imprimirá para poder visualizarla.

```
[ ] # Making the Confusion Matrix
    from sklearn.metrics import confusion_matrix
    cm = confusion_matrix(y_test, y_pred)

[ ] print(cm)
```

Figura 90: Creación de matriz de confusión

Fuente: Google Colab

3.6.3.1.2.3. Prueba manual del modelo

Además de la matriz de confusión que se generará luego del entrenamiento, se realizarán pruebas manuales a fin de ver que estén realizando las predicciones correctamente.

```
[ ] # Predicting a single new observation
    new_prediction = classifier.predict(sc.transform(np.array([[24,2,58,1.58]])))
    new_prediction = (new_prediction > 0.5)

    new_prediction1 = classifier.predict(sc.transform(np.array([[54,2,900,1.38]])))
    new_prediction1 = (new_prediction1 > 0.5)
```

Figura 91: Predicción de casos manuales

Fuente: Google Colab

Y para poder visualizar el resultado, se imprimirá por consola el resultado.

```
[ ]  print(new_prediction)
     print(new_prediction1)

     [[False]]
     [[ True]]
```

Figura 92: Impresión de resultados de pruebas

Fuente: Google Colab

3.6.4. Revisar el proyecto

En este último apartado de la fase 6 de la metodología CRISP-DM se hará una evaluación de todo lo desarrollado durante la implementación.

A lo largo de la fase 6 se ha desarrollado la red neuronal, con la cual se ha llegado a los valores esperados en cada una de las métricas con las que se ha evaluado el producto. Se debe hacer el hincapié se podría ampliar la predicción a diferentes márgenes de edad, tales como ancianos o adultos. Asimismo, se evaluará ambos dataset y el funcionamiento que se ha obtenido en ambos casos.

CAPÍTULO IV

4. RESULTADOS

En este capítulo se realizará un recuento de cada uno de los pasos seguidos a través de la metodología, así como se explicarán los resultados obtenidos y se realizarán pruebas a fin de verificar que el modelo es funcional.

4.1. Recuento de la metodología CRISP-DM durante el proyecto

En la primera parte del documento se decidió utilizar la metodología CRISP-DM ya que esta metodología permite crear modelos de minería de datos que resuelven objetivos concretos. El uso de esta para el desarrollo de este proyecto ha permitido lograr al objetivo que se planteó al inicio de este, que es analizar datos referidos a casos de obesidad en adolescentes de diferentes colegios de Arequipa. En cada una de las fases de esta se ha logrado adecuar el dataset con el que se contaba a fin de evitar ruido al momento del entrenamiento del modelo, asimismo, se han evaluado diferentes modelos y comparado los resultados que tenían para lograr obtener el modelo óptimo para solucionar el objetivo.

A continuación, se hará un repaso de cada una de las fases que se siguieron para lograr el objetivo:

Fase 1: Comprensión de los requisitos del negocio

En esta fase se determinaron los objetivos que se debían alcanzar al finalizar las fases, para ello, se analizó el contexto en el que se encontraba, así como los objetivos y criterios del negocio, es decir, qué es lo que se deseaba lograr y cómo. Como segunda parte de esta fase se tuvo la evaluación de la situación, en donde se explicó el contexto en el que se encontraba la obesidad en el Perú y se definió lo que se deseaba lograr aplicando aprendizaje de máquina en los datos. Finalmente, se estableció un plan de acción donde

se determinó el tiempo para desarrollar el modelo, así como las herramientas que se utilizarían a lo largo de las fases.

Fase 2: Comprensión de los datos

Una vez se tuvo claro el contexto y las herramientas a utilizar, se empezó con el análisis del dataset: viendo las columnas que se tenían, las que se debían agregar y quitar. Además, se evaluó cómo es que se debía transformar los datos a fin de lograr una estandarización entre los registros de cada uno de los atributos, para finalmente realizar un análisis gráfico de los registros que se tiene en los atributos para luego realizar la verificación de estos.

Fase 3: Preparación de los datos

En esta fase se realizó la presentación oficial de los atributos del dataset, se realizó cada una de las modificaciones que fueron detectadas en la fase 2, generando así un nuevo dataset que ya estaría listo para ser aplicado en el modelado.

Fase 4: Búsqueda/Modelado

En vista que ya se disponía del dataset con el que se podía generar un modelo de aprendizaje automático, se procedió a determinar la herramienta que se iba a utilizar, así como los modelos que se iban a generar a fin de encontrar el óptimo para lograr el objetivo planteado y las métricas con las que se iban a evaluar. Una vez se tuvo claro lo que se iba a realizar, se procedió a diseñar los modelos en la herramienta RapidMiner y a ejecutarlos, obteniendo así los resultados de cada una de las métricas, con las que en la siguiente fase se evaluarían.

Fase 5: Evaluación

Con los resultados de las métricas obtenidos anteriormente, se procedió a evaluar los modelos y a tomar la decisión de cuál era el óptimo para llegar al objetivo, que fueron las redes neuronales.

Fase 6: Implementación

Finalmente, en esta fase se explicó el desarrollo paso a paso de la red neuronal, con lo cual se puede considerar que se logró llegar al objetivo.

4.2. Resultados del modelo implementado

Como se mencionó a lo largo del desarrollo de la fase 6, se utilizó el lenguaje de programación Python, así como las diferentes librerías y frameworks que brinda (keras, sklearn, numpy, matploit, pandas, etc.), para crear la red neuronal. Este modelo consistió en tres capas: una capa de entrada con 4 nodos, que eran datos que se encontraban en el dataset previamente analizado; una capa oculta con 5 nodos con función de activación sigmoid, además, para evitar overfitting (sobreajuste), se utilizó la función Dropout con valor 0.95, para que, si el modelo sobrepasara el 95% de ajuste, detuviera las iteraciones; y una capa de salida con un nodo.

Durante el entrenamiento del modelo se realizaron 25 iteraciones (epochs), y en cada uno de ellos se evaluó el comportamiento con 6 métricas, las cuales habían sido fundamentales en la fase 4 para escoger qué modelo de aprendizaje automático era ideal para solucionar el problema: Accuracy, Root Mean Squared Error, Mean Squared Error, AUC, Mean Absolute Error, Recall. Al finalizar el entrenamiento, se puede observar que el accuracy del modelo es de 95,7%, el cual es un porcentaje aceptable. Ya que en el momento de la implementación de la red neuronal se indicó que se utilizara la técnica de cross-validation, el modelo pudo desarrollar una matriz de confusión con las pruebas que se hicieron, la cual es la siguiente:

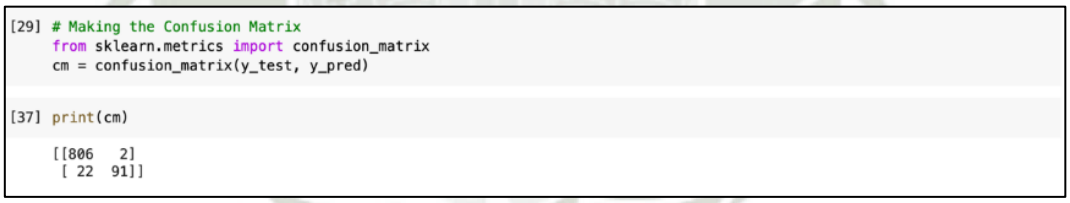


Figura 93: Impresión de matriz de confusión de test

Fuente: Google Colab

En esta matriz se puede observar que el número de aciertos es muy alto, y los errores mínimos, por lo que se puede confirmar que el accuracy es ideal para el modelo. También se puede observar que, el preprocesamiento del dataset en la fase 3 evitó que el modelo tuviera datos que podrían ocasionar ruido al momento del entrenamiento.

4.2.1. Pruebas del modelo implementado

Las pruebas en el modelo implementado se realizarán directamente en la red neuronal en Google Colab. Los datos utilizados en estas pruebas fueron tanto del dataset como datos nuevos

previamente validados a través del cálculo del IMC, el cual es el determinante de si una persona tiene o es propensa a tener obesidad o no.

Pruebas en Google Colab

Se realizaron 4 pruebas a través del método predict, con datos tanto nuevos como existentes en el dataset. Estos datos se ingresaron en un array siguiendo el siguiente orden: Edad, sexo, peso, altura.

```
new_prediction2 = classifier.predict(sc.transform(np.array([[4,1,10,0.90]])))
new_prediction2 = (new_prediction2 > 0.5)

new_prediction3 = classifier.predict(sc.transform(np.array([[4,1,40,0.90]])))
new_prediction3 = (new_prediction3 > 0.5)
```

Figura 94: Pruebas manuales de modelo por consola

Fuente: Google Colab

Cada uno de los resultados fueron obtenidos fueron almacenados en variables, las cuales luego fueron mostradas por consola, confirmando así que las predicciones eran correctas.

```
print('Edad: 04 | Sexo: 1 | Peso:  10 | Altura: 0.90 | ¿Obesidad?: ',new_prediction2)
print('Edad: 04 | Sexo: 1 | Peso:  40 | Altura: 0.90 | ¿Obesidad?: ',new_prediction3)
```

```
Edad: 04 | Sexo: 1 | Peso:  10 | Altura: 0.90 | ¿Obesidad?:  [[False]]
Edad: 04 | Sexo: 1 | Peso:  40 | Altura: 0.90 | ¿Obesidad?:  [[ True]]
```

Figura 95: Impresión de resultados de pruebas manuales por consola

Fuente: Google Colab

4.3. Análisis y discusión del modelo implementado

Es importante evaluar el modelo de una manera profunda para evitar concluir erróneamente, es por ello por lo que, para poder apreciar mejor los valores de cada una de las métricas a lo largo de las iteraciones, tanto en el entrenamiento como en las pruebas, con ayuda de la librería Matplotlib de Python se realizaron gráficos comparativos de cada una de las métricas aplicadas al momento del entrenamiento del modelo, tanto para el modelo con el dataset completo como para el dataset únicamente con edades de 12 a 18 años.

4.3.1. Análisis con registros de personas entre 4 a 20 años

Valores conseguidos a la iteración número 1:

Epoch 1/25

215/215 [==============================] - 9s 18ms/step - loss: 0.7082 - accuracy: 0.5114 - root_mean_squared_error: 0.5071 - mean_squared_error: 0.2571 - auc: 0.5200 - mean_absolute_error: 0.4998 - recall: 0.5216 - val_loss: 0.5703 - val_accuracy: 0.8849 - val_root_mean_squared_error: 0.4359 - val_mean_squared_error: 0.1900 - val_auc: 0.6890 - val_mean_absolute_error: 0.4288 - val_recall: 0.0885

Valores conseguidos a la iteración número 25:

Epoch 25/25

215/215 [==============================] - 1s 3ms/step - loss: 0.1154 - accuracy: 0.9576 - root_mean_squared_error: 0.1740 - mean_squared_error: 0.0303 - auc: 0.9945 - mean_absolute_error: 0.0907 - recall: 0.6549 - val_loss: 0.1008 - val_accuracy: 0.9739 - val_root_mean_squared_error: 0.1553 - val_mean_squared_error: 0.0241 - val_auc: 0.9975 - val_mean_absolute_error: 0.0825 - val_recall: 0.8053

A continuación, se presentarán 6 gráficas de cada una de las métricas empleadas para evaluar el modelo al momento de la construcción.

Métrica AUC

En la Figura 96 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica del área bajo la curva, o conocida por sus siglas en inglés AUC. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un AUC de 0.5200 en una escala del 0 al 1; y finalizó el entrenamiento con un AUC de 0.9945. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica no fue tan bajo como en el entrenamiento, sino que fue de 0.6890 y finalizó con un valor de 0.9975, considerado óptimo.

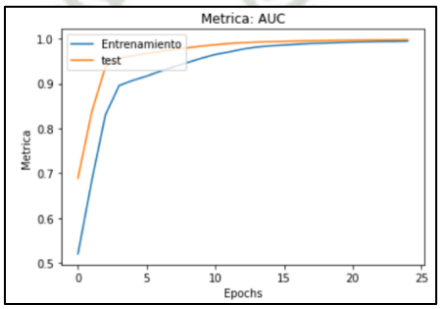


Figura 96: Visualización gráfica de métrica AUC

Fuente: Google Colab

Métrica Accuracy

En la Figura 97 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica de precisión, o accuracy. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un accuracy de 0.5114 en una escala del 0 al 1; y finalizó el entrenamiento con un accuracy de 0.9576. Lo que también se puede observar es que, al igual que la métrica anterior, al momento de hacer las pruebas, el valor de la métrica no fue tan bajo como en el entrenamiento, sino que fue de 0.8849 y finalizó con un valor de 0.9739, considerado óptimo ya que es mayor al 90%.

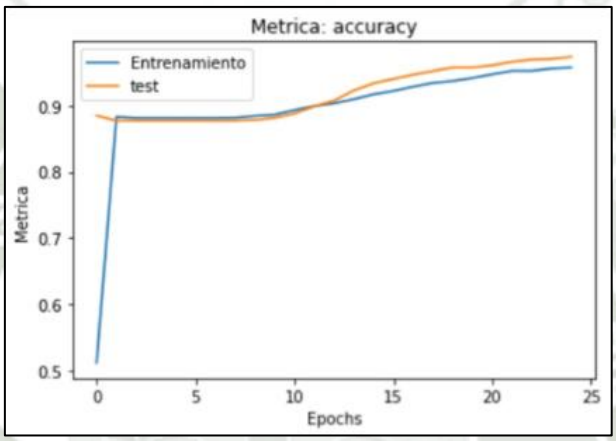


Figura 97: Visualización gráfica de métrica Accuracy

Fuente: Google Colab

Métrica Mean Absolute Error

En la Figura 98 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Mean Absolute Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Mean Absolute Error de 0.4998 en una escala del 0 al 1; y finalizó el entrenamiento con un Mean Absolute Error de 0.0907. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue más bajo que el valor obtenido en la etapa de entrenamiento, 0.4288 y finalizó con un valor de 0.0825, considerado óptimo ya que está muy próximo a cero.

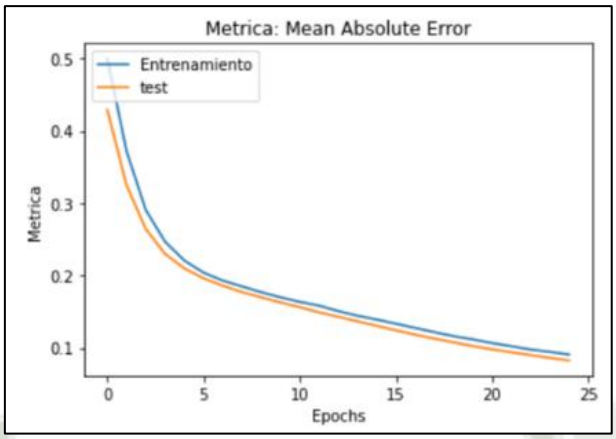


Figura 98: Mean Absolute Error

Fuente: Google Colab

Métrica Root Mean Squared Error

En la Figura 99 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Root Mean Squared Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Root Mean Squared Error de 0.5071 en una escala del 0 al 1; y finalizó el entrenamiento con un Root Mean Squared Error de 0.1740. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue bajo a comparación del valor obtenido en la etapa de entrenamiento, siendo el valor 0.4359 y finalizó con un valor de 0.1553, considerado óptimo ya que está muy próximo a cero.

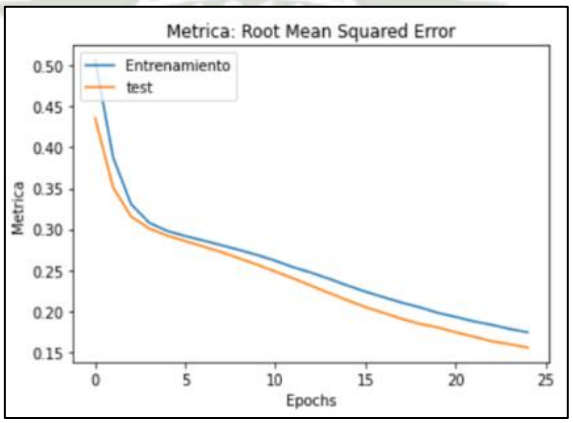


Figura 99: Visualización gráfica de métrica Root Mean Squared Error

Fuente: Google Colab

Métrica Recall

En la Figura 100 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Recall. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Recall de 0.5216 en una escala del 0 al 1; y finalizó el entrenamiento con un Recall de 0.6549. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue más bajo que el obtenido en el entrenamiento, siendo este 0.0885 y finalizó con un valor de 0.8053.

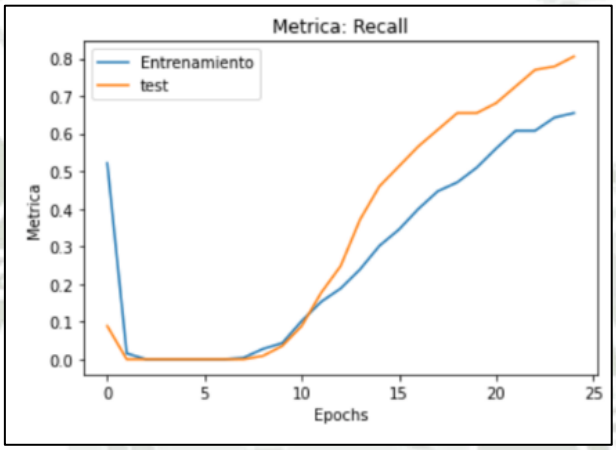


Figura 100: Visualización gráfica de métrica Recall

Fuente: Google Colab

Métrica Mean Squared Error

En la Figura 101 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Mean Squared Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Mean Squared Error de 0.2571 en una escala del 0 al 1; y finalizó el entrenamiento con un Mean Squared Error de 0.0303. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica no fue más bajo que el entrenamiento, siendo de 0.0241 y finalizó con un valor de 0.1900, considerado óptimo ya que está muy próximo a cero.

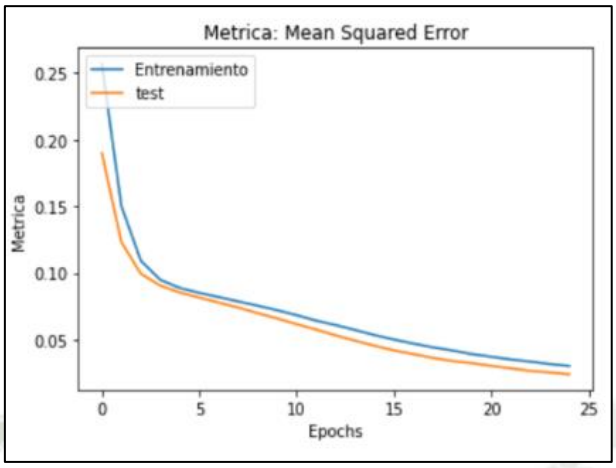


Figura 101: Visualización gráfica de métrica Mean Squared Error

Fuente: Google Colab

Con estas gráficas se puede concluir que 25 iteraciones en el entrenamiento fue suficiente para lograr una precisión aceptable (>90%) sin llegar a overfitting (sobreajuste). También se puede apreciar que, tanto entrenamiento como test, a manera que avanza cada iteración, el modelo mejora, incluso, en la etapa de test el modelo afina su precisión y rendimiento en cada una de las métricas.

Al mismo tiempo, si se evalúan los resultados obtenidos con los trabajos previos a esta investigación, se puede observar que en el trabajo Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents (Sulla Torres et al., 2018) se propone el desarrollo de un modelo neuro difuso para la clasificación de obesidad en niños y adolescentes de sexo masculino a partir de un dataset de 2938 registros. Dentro de esta clasificación, se puede observar que el modelo obtuvo un accuracy de 96.96% y tasa de error de 3.04% luego de haber realizado 500 iteraciones. Si bien es cierto que la precisión es más alta, se debe tener en cuenta que se realizan más iteraciones sobre una base más pequeña. Se podría implementar dicho modelo para el dataset con el que se cuenta actualmente y ver el comportamiento de este.

Por otro lado, en el trabajo Computer aided diagnosis of obesity based on thermal imaging using various convolutional neural networks (Snekhalatha et al., 2021), se puede observar que se propone crear una CNN para que, a partir de imágenes térmicas de diferentes partes del cuerpo, se detecte si una persona es obesa o no. Evaluando los resultados, se puede concluir que lograron un 92% de precisión (accuracy), un valor de AUC de 0.948 en el primer modelo,

y en el segundo lograron 79% de precisión y un valor de AUC de 0.90. El modelo CNN que proponen devolvió resultados más bajos a comparación de la red neuronal propuesta en las etapas anteriores.

4.3.2. Análisis con registros de personas entre 12 a 18 años

Valores conseguidos a la iteración número 1:

Epoch 1/25

106/106 [==============================] - 4s 9ms/step - loss: 1.0946 - accuracy: 0.7701 - root_mean_squared_error: 0.5306 - mean_squared_error: 0.2816 - auc: 0.4929 - mean_absolute_error: 0.4989 - recall: 0.1267 - val_loss: 0.7464 - val_accuracy: 0.2892 - val_root_mean_squared_error: 0.5258 - val_mean_squared_error: 0.2764 - val_auc: 0.3508 - val_mean_absolute_error: 0.5243 - val_recall: 0.5918

Valores conseguidos a la iteración número 25:

Epoch 25/25

106/106 [==============================] - 0s 3ms/step - loss: 0.4162 - accuracy: 0.8581 - root_mean_squared_error: 0.3531 - mean_squared_error: 0.1247 - auc: 0.4887 - mean_absolute_error: 0.2701 - recall: 0.0000e+00 - val_loss: 0.4636 - val_accuracy: 0.8147 - val_root_mean_squared_error: 0.3828 - val_mean_squared_error: 0.1465 - val_auc: 0.9563 - val_mean_absolute_error: 0.2846 - val_recall: 0.0000e+00

A continuación, se presentarán 6 gráficas de cada una de las métricas empleadas para evaluar el modelo al momento de la construcción.

Métrica AUC

En la Figura 102, se puede observar el comportamiento del modelo a partir de la evaluación con la métrica del área bajo la curva, o conocida por sus siglas en inglés AUC se puede observar el comportamiento del modelo a partir de la evaluación con la métrica del área bajo la curva. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un AUC de 0.4929 en una escala del 0 al 1; y finalizó el entrenamiento con un AUC de 0.4887. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica no fue tan bajo como en el entrenamiento, sino que fue de 0.3508 y finalizó con un valor de 0.9563, observando así un comportamiento inusual ya que los valores de la métrica en entrenamiento y test están totalmente alejados, pero al mismo tiempo considerado correcto.

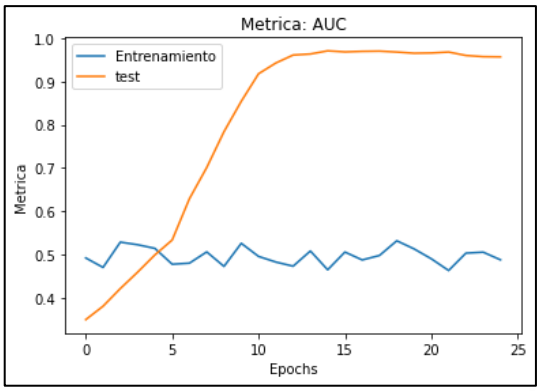


Figura 102: Visualización gráfica de métrica AUC

Fuente: Google Colab

Métrica Accuracy

En la Figura 103 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica de precisión, o accuracy. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un accuracy de 0.7701 en una escala del 0 al 1; y finalizó el entrenamiento con un accuracy de 0.8581. Lo que también se puede observar es que, al igual que la métrica anterior, al momento de hacer las pruebas, el valor de la métrica no fue tan bajo como en el entrenamiento, sino que fue de 0.2892 y finalizó con un valor de 0.8147, considerado bueno ya que es mayor al 80%. Esta variación de la métrica en entrenamiento y test se debe a los pocos datos que se manejó en la construcción del modelo.

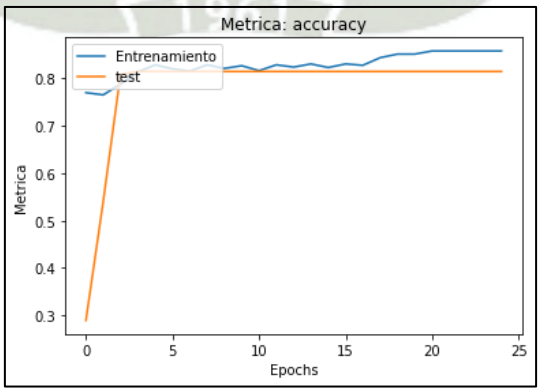


Figura 103: Visualización gráfica de métrica Accuracy

Fuente: Google Colab

Métrica Mean Absolute Error

En la Figura 104 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Mean Absolute Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Mean Absolute Error de 0.4989 en una escala del 0 al 1; y finalizó el entrenamiento con un Mean Absolute Error de 0.2701. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue más bajo que el valor obtenido en la etapa de entrenamiento, 0.2764 y finalizó con un valor de 0.1465, observando así un comportamiento inusual ya que los valores de la métrica en entrenamiento y test están totalmente alejados, pero al mismo tiempo considerado correcto ya que está muy próximo a cero.

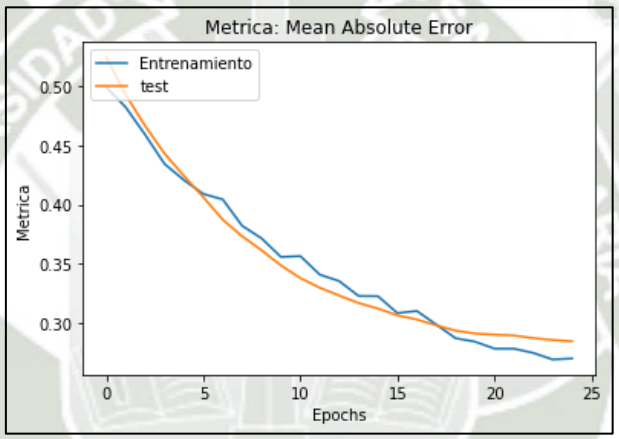


Figura 104: Mean Absolute Error

Fuente: Google Colab

Métrica Root Mean Squared Error

En la Figura 105 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Root Mean Squared Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Root Mean Squared Error de 0.5306 en una escala del 0 al 1; y finalizó el entrenamiento con un Root Mean Squared Error de 0.3531. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue bajo a comparación del valor obtenido en la etapa de entrenamiento, siendo el valor 0.5258 y finalizó con un valor de 0.3828, observando así un comportamiento inusual ya que los valores de la métrica en entrenamiento y test están totalmente alejados, pero al mismo tiempo considerado correcto ya que está muy próximo a cero.

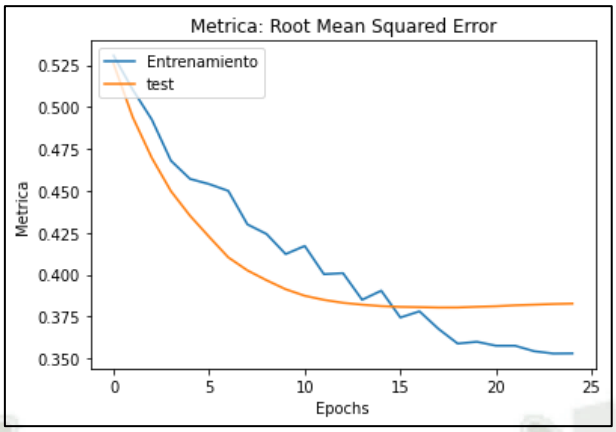


Figura 105: Visualización gráfica de métrica Root Mean Squared Error

Fuente: Google Colab

Métrica Recall

En la Figura 106 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Recall. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Recall de 0.1267 en una escala del 0 al 1; y finalizó el entrenamiento con un Recall de 0.0000e+00. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica fue más bajo que el obtenido en el entrenamiento, siendo este 0.1267 y finalizó con un valor de 0.0000e+00.

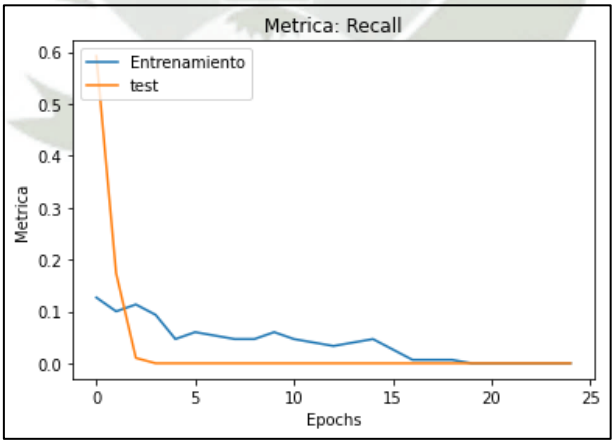


Figura 106: Visualización gráfica de métrica Recall

Fuente: Google Colab

Métrica Mean Squared Error

En la Figura 107 se puede observar el comportamiento del modelo a partir de la evaluación con la métrica Mean Squared Error. Se puede observar que, en la etapa del entrenamiento, el modelo empezó con un Mean Squared Error de 0.2816 en una escala del 0 al 1; y finalizó el entrenamiento con un Mean Squared Error de 0.1247. Lo que también se puede observar es que, al momento de hacer las pruebas, el valor de la métrica no fue más bajo que el entrenamiento, siendo de 0.2764 y finalizó con un valor de 0.1465, observando así un comportamiento inusual ya que los valores de la métrica en entrenamiento y test están totalmente alejados, pero al mismo tiempo considerado correcto ya que está muy próximo a cero.

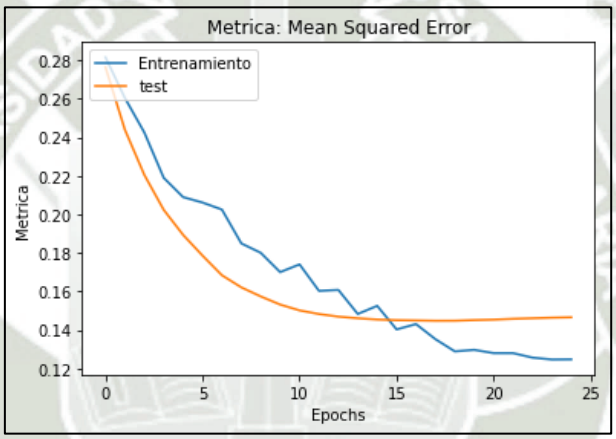


Figura 107: Visualización gráfica de métrica Mean Squared Error

Fuente: Google Colab

Con estas gráficas se puede concluir que 25 iteraciones en el entrenamiento fue suficiente para lograr una precisión buena (>80%) sin llegar a overfitting (sobreajuste). También se puede apreciar que, tanto entrenamiento como test, existe una diferencia notable entre la métrica aplicada en la etapa de entrenamiento, así como la etapa de test, esto debido a que el modelo tiene muy pocos datos para predecir resultados.

Al mismo tiempo, si se evalúan los resultados obtenidos con los trabajos previos a esta investigación, se puede observar que en el trabajo Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents (Sulla Torres et al., 2018) se propone el desarrollo de un modelo neuro difuso para la clasificación de obesidad en niños y adolescentes de sexo masculino a partir de un dataset de 2938 registros. Dentro de esta

clasificación, se puede observar que el modelo obtuvo un accuracy de 96.96% y tasa de error de 3.04% luego de haber realizado 500 iteraciones. Si bien es cierto que la precisión es más alta, se debe tener en cuenta que se realizan más iteraciones sobre una base más pequeña. A comparación del dataset con el que se cuenta en este estudio, se puede observar que en la red neuronal planteada en este proyecto así tenga más iteraciones el accuracy no aumenta, por lo cual como se mencionó anteriormente, este problema viene a raíz que el dataset luego de hacer la limpieza de registros con edades que se encuentran fuera del rango considerado para la adolescencia es muy pequeño para entrenar la red, por lo cual se recomendará aumentar el número de registros dentro de este.

CONCLUSIONES

1. Se cumplió con la predicción de obesidad en adolescentes mediante aprendizaje automático a partir de medidas antropométricas, tales como sexo, edad, peso y talla. A lo largo del desarrollo de la metodología CRISP-DM se estableció que la técnica de aprendizaje automático más adecuada es redes neuronales con un promedio de predicción entre el 85.81% y 95.76%. Además, fue la que más se adecuó al objetivo planteado en el proyecto de investigación dado que esta técnica brinda ventajas como alta tolerancia a fallos, facilidad al momento de manejar características complejas en los datos, adaptación de pesos a manera que se utiliza, así como reconocimiento de patrones en la etapa de producción al modelo.

2. Se estableció dos necesidades a cubrir con la aplicación: primero, poder realizar predicciones a adolescentes en base a datos como edad, sexo, peso y estatura con un elevado porcentaje de fiabilidad, de tal forma que se puedan dar consejos útiles a los adolescentes acerca de si deben seguir un mejor estilo de vida, visitar a un especialista de la salud o se si encuentran bien de salud; y al mismo tiempo, facilitar el diagnóstico de obesidad en adolescentes de una manera más sencilla y rápida. La segunda necesidad planteada fue ayudar a hacer un mejor uso de las tecnologías modernas como el aprendizaje automático.

3. Se obtuvo la recolección de datos como edad, sexo, talla en metros y peso en kilogramos de 3068 personas entre 4 y 21 años en colegios de la provincia de Arequipa para entrenar un modelo de aprendizaje automático para predecir la obesidad en adolescentes, de los cuales 1763 registros son de adolescentes entre 12 y 18 años.

4. A partir del análisis previo al dataset, se determinó que se debía realizar una transformación al mismo, lo que incluyó transformar datos, crear y quitar columnas para la etapa del modelado. Se identificó que se que se debían agregar tres atributos nuevos al dataset: IMC e índice, las cuales serían atributos temporales; y riesgo; asimismo, se tuvo que transformar los datos que contenía el atributo sexo, ya que tenía tantos datos alfanuméricos como numéricos, lo que a futuro podía causar ruido en el entrenamiento. Finalmente, se realizó dos construcciones del modelo con diferente origen de datos: Primero, un dataset con la data completa, es decir, con rango de edades de 4 a 21 años;

y un segundo dataset con datos únicamente dentro del rango considerado adolescencia, que es de 12 a 18 años.

5. Se realizó la comparación de 7 modelos de aprendizaje automático dentro de la herramienta RapidMiner: Arboles de decisión, redes neuronales, máquinas de vectores de soporte (SVM), análisis bayesiano, regresión logística, regresión lineal y KNN. De los mencionados anteriormente, a partir de las métricas aplicadas en cada una de las ejecuciones de los modelos y de los resultados obtenidos, se decidió utilizar redes neuronales, ya que según las métricas con las que se ha evaluado cada uno de los modelos, este es considerado como un modelo óptimo para llegar al objetivo planteado, además de tener en cuenta de las ventajas que ofrece el modelo: Alta tolerancia a fallos, facilidad al momento de manejar características complejas en los datos, adaptación de pesos a manera que se utiliza, así como reconocimiento de patrones en la etapa de producción.

6. Se creó una red neuronal que constó de 3 capas: una capa de entrada con 4 nodos, que eran datos que se encontraban en el dataset previamente analizado; una capa oculta con 5 nodos con función de activación sigmoid, y una capa de salida con un nodo, que es el resultado de la predicción del modelo. Además, se utilizó la función Dropout con valor 0.95, para que si el modelo sobrepasara el 95% de ajuste detuviera las iteraciones.

7. Se comprobó que el primer modelo es óptimo a partir de las métricas implementadas al momento de la compilación de la red neuronal: Accuracy, Root Mean Squared Error, Mean Squared Error, AUC, Mean Absolute Error Y Recall. Estos resultados se graficaron para una mejor evaluación a fin de evitar concluir erróneamente sobre estos valores. Con estas gráficas se concluyó que 25 iteraciones en el entrenamiento fue suficiente para lograr una precisión aceptable (>90%) sin llegar a overfitting (sobreajuste). También se apreció que, tanto entrenamiento como test, a manera que avanza cada iteración, el modelo mejora, incluso, en la etapa de test el modelo afinó su precisión y rendimiento en cada una de las métricas. Al mismo tiempo, se realizaron pruebas al modelo directamente con datos del dataset y nuevos, con lo que se verificó el buen funcionamiento del modelo.

RECOMENDACIONES Y TRABAJOS FUTUROS

Al finalizar la investigación, se proponen las siguientes recomendaciones:

1. Se recomienda ampliar la cantidad de registros con las características antropométricas mencionadas de personas en el modelo planteado anteriormente, a fin de tener un dataset más amplio y conseguir mejores resultados.
2. Para identificar si el modelo es óptimo o está presentando problemas tales como underfitting o el overfitting se recomienda dividir el conjunto de datos de entrada en dos subconjuntos, siendo la repartición un 70 % para el entrenamiento, en donde 60% sea para entrenamiento y 10% para validación; y un 30 % para evaluación.
3. Desarrollar modelos con valores biométricos, es decir, valores que se obtienen a partir de una muestra de fluido, ya sea sangre, saliva, etc. A fin de poder desarrollar una comparación sobre la exactitud de los modelos cuando se utilizan medidas antropométricas y cuando se utilizan valores biométricos.
4. Obtener más registros para el dataset planteado, a fin de poder utilizar otros métodos de aprendizaje automático en los que se necesitan más datos. Asimismo, se propone obtener datos de edades variadas a fin de que el modelo no solo sea para predicción de obesidad en adolescentes, sino sea para todas las edades.
5. A nivel de interfaz gráfica de usuario, se puede implementar una interfaz para la toma de datos, los cuales serían enviados a la red neuronal y posteriormente mostrar el resultado en la interfaz. Además, se puede considerar la implementación de reportes para visualizar el estado de las predicciones, asimismo, se propone implementar más ventanas donde se pueda brindar diferentes predicciones a partir de otros datos y se pueda ofrecer calculadora de índices importantes dentro de este tema, tales como el IP o índice ponderal para recién nacidos.

REFERENCIAS

Aquino, A. A. (2016). *Proceso de minería de datos centrado en el usuario con base en la norma ISO 9241-210:2010.* Universidad Veracruzana, FACULTAD DE ESTADÍSTICA E INFORMÁTICA. Veracruz: Universidad Veracruzana.

Aranda, M. (08 de Mayo de 2019). *RapidMiner, la democratización del Data Science*. Mind Analytics: https://bit.ly/3QWIFWK

Arias Zuluaga, E. T. (2020). *Desarrollo de un modelo predictivo con inteligencia artificial para establecer clasificación ASA a pacientes en una consulta preanestésica.* Medellín: Universidad de Antioquia.

Banco Interamericano de Desarrollo. (2020). *La Inteligencia Artificial en el Sector Salud.* Washington D.C.: Banco Interamericano de Desarrollo.

Banu, A. (04 de Marzo de 2022). *Artificial Intelligence for Sustainable Health Care Advancements.* Nueva York: CRC Press. https://doi.org/10.1201/9871003153405-2

BBVA. (08 de Noviembre de 2019). *'Machine learning': ¿qué es y cómo funciona?* 'Machine learning': ¿qué es y cómo funciona?: https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/

Cai, Z., Liu, L., Chen, B., & Wang, Y. (2021). *Artificial Intelligence: From Beginning to Date.* Singapur: World Scientific Publishing Co. Pte. Ltd.

Carbonnelle, P. (2022). *PYPL PopularitY of Programming Language.* GitHub: https://pypl.github.io/PYPL.html

Castillo Hernández, J. L., & Zenteno Cuevas, R. (2004). Valoración del Estado Nutricional. En U. Veracruzana, *Revista Medica de la Universidad Veracruzana.*

Centro Nacional de Alimentación y Nutrición. (2020). *Sobrepeso y obesidad en la población peruana.* Lima: Ministerio de Salud.

Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. (2020). Editorial: La obesidad como problema de salud pública. En P. y. Centro Nacional de Epidemiología, *Boletín Epidemiológico del Perú 2020* (pp. 293-294). Lima: Ministerio de Salud.

Cequera, A., & García de León Méndez, M. C. (2014). Biomarcadores para fibrosis hepática, avances, ventajas y desventajas. En R. d. Mexico, *Revista de Gastroenterología de Mexico* (Vol. 79, pp. 187-199). México D.F. https://doi.org/10.1016/j.rgmx.2014.05.003

CERDA, J., & VILLARROEL, L. (2008). *Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa.* Santiago de Chile: Revista Chilena de Pediatría.

Challenger-Pérez, I., Díaz-Ricardo, Y., & Becerra-García, R. A. (2013). *El lenguaje de programación Python.* Centro de Información y Gestión Tecnológica de Santiago de Cuba. Holguín: Redalyc.

Chatterjee, A., Gerdes, M. W., & Martinez , S. G. (2020). *Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview.* Noruega: Sensors. https://doi.org/10.3390/s20092734

Coca Bergolla, Y., Cuza Soca, C., & Llivina Lavigne, M. (2021). *Aplicaciones de la Inteligencia Artificial.* La Habana: UNESCO. https://doi.org/978-959-18-1343-5

Collave Garcia, Y. (05 de Marzo de 2021). "El 75% de muertes por COVID-19 se relaciona con sobrepeso u obesidad en Perú". *Diario El Comercio*, p. Virtual.

Comisión para acabar con la obesidad infantil. (2016). *Informe de la Comisión para acabar con la obesidad infantil.* Ginebra: Catalogación por la Biblioteca de la OMS.

Córdoba-Rodríguez, D. P., Rodriguez, G., & Moreno, L. A. (2022). Predicting of excess body fat in children. *Current Opinion in Clinical Nutrition and Metabolic Care* , 304-310. https://doi.org/10.1097/MCO.0000000000000848

De la Fuente Carmona, A. (2022). *Diseño de Soluciones Avanzadas Basadas en Técnicas de Machine Learning para la Toma de Decisiones en Gestión de Activos.* Sevilla: Universidad de Sevilla.

Dunstan, J., Aguirre, M., Bastías, M., Nau and Thomas A Glass, C., & Tobar, F. (2019). *Predicting nationwide obesity from food sales using machine learning.* Chile: Health Informatics Journal. https://doi.org/10.1177/1460458219845959

Espinosa-Zúñiga, J. (2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública.* México: Ingeniería Investigación y Tecnología. https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Flores Jara, C. A. (2015). *Extracción de Información en EMR para la Identificación de Obesidad mediante el Estudio de Comorbilidades Asociadas.* Concepción: UNIVERSIDAD DE CONCEPCIÓN.

García-Olalla Olivera, O. (16 de Septiembre de 2019). *Redes Neuronales artificiales: Qué son y cómo se entrenan*. Xeridia en español: https://www.xeridia.com/blog/redes-neuronales-artificiales-que-son-y-como-se-entrenan-parte-i

Gutiérrez Contreras, A. (2022). *Aplicación de Técnicas de Machine Learning para la Predicción de la Obesidad en Jóvenes de Estados Unidos.* Universidad Complutense de Madrid, Facultad de Estudios Estadisticos. Madrid: Universidad Complutense de Madrid.

IBM. (2020). *Redes Neuronales*. IBM: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=nodes-neural-networks

Instituto Internacional de Investigación sobre Políticas Alimentarias. (2016). *Informe de la Nutrición Mundial 2016: De la promesa al impacto: terminar con la malnutrición de aquí a 2030.* Washington, DC.: International Food Policy Research Institute. https://doi.org/10.2499/9780896295865

Instituto Nacional de Salud. (2006). *Encuesta Nacional de Indicadores Nutricionales, Bioquímicos, Socioeconómicos y Culturales Relacionados con las Enfermedades Crónico Degenerativas.* Lima: Ministerio de Salud.

Instituto Nacional de Salud. (2020). *Más del 60% de peruanos mayores de 15 años sufre de sobrepeso u obesidad y podría hacer formas graves de COVID-19.* Lima: INS. https://web.ins.gob.pe/es/prensa/noticia/mas-del-60-de-peruanos-mayores-de-15-anos-sufre-de-sobrepeso-u-obesidad-y-podria

Kumar, R., Sehgal, K., & Lal Meena, A. (2022). *Artificial Intelligence Encouraging students in higher education to Seize its Potential.* Nueva York. https://doi.org/10.1201/9871003153405-8

Marcos-Pasero, H., Colmenarejo, G., Aguilar-Aguilar, E., Ramírez de Molina, A., Reglero, G., & Loria-Kohen, V. (2021). *Ranking of a wide multidomain set of predictor variables of children obesity by machine learning obesity by machine learning.* Nature - Scientifics Reports. https://doi.org/10.1038/s41598-021-81205-8

Ministerio de Educación. (2013). *Desarrollo y crecimiento humano, Comprensión de la discapacidad.* La Paz: Viceministerio de Educación Superior de Formación Profesional/Dirección General de Formación de Maestros.

Ministerio de Salud del Perú. (2012). *Un gordo problema: sobrepeso y obesidad en el Perú.* Lima: Ministerio de Salud del Perú.

Mondragón Barrera, M. A. (2014). USO DE LA CORRELACIÓN DE SPEARMAN EN UN ESTUDIO DE INTERVENCIÓN EN FISIOTERAPIA. En M. A. Mondragón Barrera, *Movimiento científico* (pp. 98-104). Medellin: Corporacion Universitaria Iberoamerica.

Naciones Unidas. (2011). Declaración Política de la Reunión de Alto Nivel de la Asamblea General sobre la Prevención y el Control de las Enfermedades No Transmisibles. *Seguimiento de los resultados de la Cumbre del Milenio* (p. 14). Nueva York: Naciones Unidas. Retrieved 12 de 04 de 2021, from https://undocs.org/es/A/66/L.1

Nimptsch, K., Konigorski, S., & Pischon, T. (2018). *Diagnosis of obesity and use of obesity biomarkers in science and clinical medicine.* Alemania: Metabolism Clinical and Experimental. https://doi.org/10.1016/j.metabol.2018.12.006

Norvig, P., & Russell, S. J. (2004). *INTELIGENCIA ARTIFICIAL. UN ENFOQUE MODERNO* (Segunda ed.). Madrid: PEARSON PRENTICE HALL. https://doi.org/978-84-205-4003-0

Ordovás, J. M., & Corella, D. (2015). *Biomarcadores: antecedentes, clasificación y guía para su aplicación en epidemiología nutricional.* Valencia: Revista Española de Nutrición Comuinitaria.

Organismo Andino de Salud – Convenio Hipólito Unanue. (2021). *SITUACIÓN DEL SOBREPESO Y OBESIDAD Y EL IMPACTO DE LA ENFERMEDAD POR COVID-19 EN PAÍSES ANDINOS.* Lima: Organismo Andino de Salud – Convenio Hipólito Unanue.

Organización Mundial de la Salud. (2016). *Informe de la Comisión para acabar con la obesidad infantil.* Ginebra: Organización Mundial de la Salud.

Organización Mundial de la Salud. (2016). *Informe de la Comisión para acabar con la obesidad infantil.* Ginebra: OMS. https://doi.org/978 92 4 351006 4

Organización Mundial de la Salud. (09 de Junio de 2021). *Obesidad y Sobrepeso: Organización Mundial de la Salud*. Organización Mundial de la Salud Web Site: https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

Peinado Pineda, I. S., & Díaz Salas, I. (2021). *Inteligencia Artificial Aplicada a la Cadena de Suministro Globales.* Monteria. SAS.

Programa Salud, Trabajo y Ambiente en América Central (SALTRA), Instituto Regional de Estudios en Sustancias Tóxicas, Universidad Nacional de Heredia. (2014). *MANUAL DE MEDIDAS ANTROPOMÉTRICAS.* Heredia: SALTRA.

Ramirez Hinestroza, D. (2018). *EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD.* Pereira: UNIVERSIDAD LIBRE SECCIONAL PEREIRA.

RapidMiner. (2022). *Performance (Classification)*. RapidMiner Documentation: https://docs.rapidminer.com/latest/studio/operators/validation/performance/predictive/performance_classification.html

Rivera Dommarco, J. A., Hernandez Avila, M., Aguilar Salinas, C. A., Vadillo Ortega, F., & Murayana Rendon, C. (2013). *Obesidad en Mexico: Recomendaciones para una politica de estado.* Mexico CD: Universidad Nacional Autonoma de Mexico.

Rodríguez Montequín, M. T., Álvarez Cabal, J. V., Mesa Fernández, J. M., & González Valdés, A. (2003). METODOLOGÍAS PARA LA REALIZACIÓN DE PROYECTOS DE DATA MINING. 257-265.

Rodríguez-Pardo, C., Segura, A., Zamorano-León, J. J., Martínez-Santos, C., Martínez, D., Collado-Yurrita, L., . . . López-Farre, A. (2019). *Decision tree learning to predict overweight/obesity based on body mass index and gene polymporphisms.* Madrid: Gene. https://doi.org/10.1016/j.gene.2019.03.011

Rossman, H., Shilo, S., Barbash-Hazan, S., Shalom Artzi, N., Hadar, E., Balicer, R. D., . . . Segal, E. (2021). *Proponen que este factor viene incluso desde antes de nacer, ya que depende de cuanto sea la glucosa de la madre durante el embarazo.* Canada: The Journal Of Pediatrics. https://doi.org/10.1016/j.jpeds.2021.02.010

Samaniego, J. F. (2021). *'Deep learning' y 'machine learning': ¿en qué se diferencian?* Madrid: Orange en Español.

Scheinker, D., Valencia, A., & Rodriguez, F. (2019). *Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Machine Learning Models.* Estados Unidos: JAMA Network. https://doi.org/10.1001/jamanetworkopen.2019.2884 (Re

Senyer Yapici, I., Erkaymaz, O., & Uzun Arslan, R. (2021). *A hybrid intelligent classifier to estimate obesity levels based on ERG signals.* Turquía: Physics Letters A. https://doi.org/10.1016/j.physleta.2021.127281

Snekhalatha, U., Palani Thanaraj, K., & Sangamithirai, K. (2021). *Computer aided diagnosis of obesity based on thermal imaging using various convolutional neural networks.* India: Biomedical Signal Processing and Control. https://doi.org/10.1016/j.bspc.2020.102233

Suárez-Carmona, W., & Sánchez-Oliver, A. J. (2014). Índice de masa corporal: ventajas y desventajas de su uso en la obesidad. Relación con la fuerza y la actividad física. En Varios, *Nutrición Clínica en Medicina* (Vol. XII, pp. 128-139). Sevilla: Nutrición Clínica en Medicina. https://doi.org/10.7400/NCM.2018.12.3.5067

Sulla Torres, J., Soto Paredes, C., Cardenas Soria, R., Huancco Coila, L., & Alfaro Casas, L. (2018). *Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents.* Lima: LACCEI.

Timarán Pereira, S. R., Hernández Arteaga, I., Caicedo Zambrano, S. J., Hidalgo Troya, A., & Alvarado Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En S. R. Timarán Pereira , I. Hernández Arteaga, S. J. Caicedo Zambrano, A. Hidalgo Troya, & J. C. Alvarado Pérez, *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesiona* (pp. 63-86). Bogota: Ediciones Universidad Cooperativa de Colombia. https://doi.org/10.16925/9789587600490

Tripathi, M. (2020). *Underfitting and Overfitting in Machine Learning.* Altrincham: Data Science Foundation.

Willmot, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. En T. &. Group, *International Journal of Geographical Information Science* (Vol. 20, pp. 89-102). Newark: Taylor & Francis Group. https://doi.org/10.1080/13658810500286976

ANEXO(S)

ANEXO A: GLOSARIO DE TERMINOLOGÍAS DE APRENDIZAJE DE MÁQUINA

1. **Agregación (Clustering):** Implica determinar cómo dividimos un conjunto de objetos en grupos a partir de la combinación actual de propiedades y valores.
2. **Algoritmo (Algorithm):** Es una secuencia lógica de instrucciones que describen paso a paso cómo resolver un problema.
3. **Análisis Bayesiano:** Las evidencias y observaciones se emplean para actualizar o inferir la probabilidad de que una hipótesis sea cierta.
4. **Aprendizaje de refuerzo (Reinforcement Learning):** Los algoritmos deben aprender cómo lograr objetivos complejos o de largo plazo en unos pocos pasos.
5. **Aprendizaje o entrenamiento (learning, training):** El proceso de detección de patrones en un conjunto de datos.
6. **Aprendizaje profundo (Deep Learning):** Un conjunto de algoritmos diseñados para reproducir los mismos resultados que el cerebro humano, siguiendo la lógica de un proceso en capas, simula las funciones básicas del cerebro a través de las neuronas.
7. **Aprendizaje supervisado y no supervisado (supervised and unsupervised machine learning):** El aprendizaje supervisado implica hacer predicciones futuras basadas en el comportamiento o las características, y el aprendizaje no supervisado utiliza datos históricos no etiquetados.
8. **Árbol de decisión (decision tree):** Particiona las muestras según la profundidad del árbol, y promedia los valores de la variable objetivo en los nodos hoja. Su interpretación es sencilla y se aplica tanto a variables de componentes cualitativos como cuantitativos.
9. **Auto aprendizaje automático (AutoML):** El sistema, creado por Google, se encarga de diseñar redes neuronales artificiales y especializadas que se pueden aplicar al desarrollo de diversas funciones.
10. **Característica, atributo, factor, propiedad o campo (feature, attribute, property, field):** Atributos que describen cada instancia en el conjunto de datos.
11. **Clasificación y regresión (classification and regression):** La clasificación predice una clase, la regresión predice un número.
12. **Confianza (confidence):** Es la probabilidad de éxito calculada por el sistema para cada predicción.
13. **Conjunto de datos (dataset):** Materia prima para sistemas de predicción.

14. **Django:** Ofrece formularios basados en modelos, tiene su propio lenguaje de plantillas y tiene una excelente documentación que está disponible gratuitamente.
15. **Ensemble Models:** Crea un nuevo modelo entrenando varios modelos similares combinando resultados.
16. **Estadística descriptiva:** Recopilar, presentar y caracterizar conjuntos de datos.
17. **Frameworks:** Un conjunto estandarizado de conceptos, prácticas y estándares enfocados en un tipo específico de problema como referencia para enfrentar y resolver nuevos problemas de naturaleza similar.
18. **Google Colab:** Permite a cualquier usuario escribir y ejecutar código Python arbitrario en el navegador. Es particularmente adecuado para tareas de aprendizaje automático, análisis de datos y educación.
19. **Ingeniería de factores (feature engineering):** El proceso de eliminar el ruido de una señal.
20. **Instancia, ejemplo o registro (instance, sample, record):** Una instancia es cada dato disponible para el análisis, y consta de las características que lo describen.
21. **Inteligencia artificial:** Intento de imitar la inteligencia humana realizado por un software que ha sido explícita o implícitamente desarrollado para completar una tarea.
22. **K-Vecinos más Cercanos:** Reconoce patrones para conocer la probabilidad de que un elemento pertenezca a una clase según su cercanía.
23. **Aprendizaje automático:** Inteligencia artificial sin necesidad de programar todas las reglas necesarias. Aprende a partir de datos históricos.
24. **Métricas:** Permiten evaluar modelos de aprendizaje automático.
25. **Minería de datos (data mining):** Descubre patrones previamente desconocidos.
26. **MLOps:** Un conjunto de técnicas enfocadas en asegurar la robustez de los modelos de aprendizaje automático tanto en el momento de la implementación como en la operación.
27. **Modelo (model):** Un filtro toma nuevos datos y su salida es una clasificación.
28. **Modelos de respuesta incremental:** Modela el cambio de probabilidad causado por una acción.
29. **NASNet (Neural Architecture Search Network):** Fue desarrollado por AutoML para clasificar e identificar objetos en fotos.
30. **Objetivo (objective):** El atributo o factor a predecir.
31. **Potenciación de gradiente (Sampling):** Forma una media ponderada.

32. **Procesado de Lenguaje Natural (NLP):** Abarca todas las técnicas relacionadas con el procesamiento de la comunicación humana, tanto hablada como escrita, basadas en las reglas del diccionario.
33. **Python:** Es un lenguaje de programación interpretado de alto nivel cuya filosofía enfatiza la legibilidad de su código para desarrollar diversas aplicaciones.
34. **Random Forest:** Se pueden usar en regresión para ajustarse mejor a la variable objetivo, aunque es difícil interpretar el modelo.
35. **Recommender Systems:** Algoritmos diseñados para recomendar artículos relevantes a los usuarios.
36. **Redes neuronales (Neural Networks):** Permite modelar la variable objetivo mediante el cálculo de los coeficientes mediante retro propagación. El uso de múltiples capas intermedias aumentará la complejidad del modelo y posiblemente mejorará el ajuste de las predicciones.
37. **Regresión (Regression):** Intentan modelar el comportamiento de variables cuantitativas en términos de otros predictores, que pueden ser cuantitativos o cualitativos, con el objetivo habitual de realizar predicciones o estimaciones.
38. **Regresión Lineal:** Permite generar una aproximación de la variable objetivo. Requiere el uso de variables cuantitativas, y destaca por su facilidad de interpretación.
39. **Regresión Logística:** Predice el resultado de una variable categórica útil para modelar la probabilidad de un evento.
40. **Series temporales:** Mejora las predicciones sobre datos recopilados. Mezcla de dará mining tradicional y forecasting.
41. **Support Vector Machines:** Métodos basados en el aprendizaje para resolver problemas de clasificación y regresión. Son muy populares en el procesamiento del lenguaje natural, el habla, el reconocimiento de imágenes y las aplicaciones de visión artificial.
42. **Tensorflow:** Es una biblioteca de código abierto desarrollada por Google para llevar a cabo proyectos de aprendizaje automático.

ANEXO B: PLAN DE TESIS

1. Planteamiento de la investigación

1.1. Planteamiento del problema

La obesidad y sobrepeso se definen como “acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud”. (Organización Mundial de la Salud, 2021)

Ambas son enfermedades crónicas que no tiene distinción con los grupos poblacionales: Las padecen personas de bajos y altos recursos; personas muy jóvenes como los niños hasta ancianos. Se diagnostica mediante el examen de Índice de Masa Corporal (IMC), que consiste en dividir el peso de una persona en kilos por el cuadrado de su talla en metros (kg/m^2).

La obesidad es una de las enfermedades que ha tenido mayor crecimiento a lo largo de los últimos años: En las últimas décadas se ha triplicado el número de personas que la padecen alrededor del mundo; y a raíz de eso se le considera como “la epidemia del siglo XXI”. Lo más peligroso de esta enfermedad son los efectos adversos a los que se exponen estas personas: enfermedades coronarias, problemas de colesterol y triglicéridos; diabetes, accidentes cerebrovasculares e incluso diferentes tipos de cáncer (próstata, mama, colon, etc.). (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades, 2020)

Las causas más comunes que ocasionan esta enfermedad son: Dietas no saludables basadas en azúcares y grasas, así como poca actividad física. Hoy en día, durante la pandemia originada por el virus del COVID-19 este problema se ha agravado: A consecuencia del confinamiento, las personas alrededor del mundo se han vuelto más sedentarias y han cambiado sus hábitos alimenticios, ya sea recortando las comidas diarias, como aumentando el azúcar y comidas altas en grasas saturadas en su dieta. Al mismo tiempo, el riesgo al contagio ha disminuido las visitas médicas para chequeos rutinarios, lo que ocasiona que muchas personas no tomen conciencia que están tomando malos hábitos para su salud.

Paralelamente, el crecimiento de la tecnología, y sobre todo de la analítica de datos se da de forma exponencial. Actualmente existen muchas maneras de tratar la información y facilitar el día a día del ser humano. Hablando específicamente de la inteligencia artificial, es un campo muy amplio que hoy en día en el extranjero, sobre todo en países del primer mundo, se utiliza en diferentes campos: Contabilidad, administración, geología, agronomía, medicina, etc.; mientras que en Latinoamérica aún no se utiliza mucho.

Hoy en día la predicción de obesidad en adolescentes está limitado a realizarse diferentes estudios (sanguíneos, físicos, etc.), los cuales mediante sus resultados demuestran si el paciente sufre de sobrepeso/obesidad y posiblemente qué órganos se estén viendo afectados. Muchas veces la detección llega tarde, cuando ya el paciente ha desarrollado enfermedades como diabetes, o incluso algún tipo de cáncer.

Es por eso por lo que se propone convertir al aprendizaje de máquina como el mejor aliado en estos momentos tan complicados a causa a la pandemia, creando un sistema web que mediante redes neuronales y el ingreso de diferentes biomarcadores se pueda predecir si esa persona tiene o no riesgo a padecer de obesidad, ya que está comprobado que mediante los niveles de diferentes sustancias que se obtienen o medidas del cuerpo humano (llamados biomarcadores) se puede predecir con un nivel de confiabilidad mayor al 50% si es que la persona cuenta con algún tipo de enfermedad, en este caso, obesidad.

1.2. Objetivos de la investigación

1.2.1. General

Predecir la obesidad en adolescentes utilizando modelos de aprendizaje automático.

1.2.2. Específicos

- Analizar y tomar los requerimientos de las necesidades que debe cubrir el funcionamiento de la predicción.
- Analizar datos referidos de adolescentes de diferentes colegios de Arequipa.
- Realizar la transformación de los datos del dataset analizado previamente.
- Comparación y elección del mejor modelo de aprendizaje automático para predecir la obesidad en adolescentes.
- Entrenar una red neuronal artificial para predecir la obesidad en adolescentes.
- Validación de los resultados en la predicción de obesidad en adolescentes.

1.3. Preguntas de investigación

a) ¿Se puede predecir la obesidad en adolescentes utilizando modelos de aprendizaje automático?

b) ¿Qué necesidades debe cubrir el funcionamiento de la predicción?

c) ¿Qué es lo que se puede concluir a partir del análisis de los datos referidos a casos de obesidad en adolescentes de diferentes colegios de Arequipa?

d) ¿Qué columnas debe tener el dataset personalizado con los datos analizados previamente?

e) ¿Qué criterios se debe tener al momento de la comparación y elección del mejor modelo de aprendizaje automático para predecir la obesidad?

f) ¿Qué consideraciones se debe tener para construir y entrenar una red neuronal artificial capaz de predecir la obesidad?

g) ¿De qué forma se debe evaluar la red neuronal artificial desarrollada para comprobar su correcto funcionamiento?

1.4. Línea y sublínea de investigación a la que corresponde el problema

1.4.1. Línea

Inteligencia Artificial.

1.4.2. Sublínea

Aprendizaje Automático.

1.5. Palabras clave

1. Redes neuronales.
2. Obesidad.
3. Aprendizaje automático.
4. Medidas antropométricas.
5. IMC.
6. OMS.

7. Metodología RUP.
8. Stakeholders.
9. Software.
10. Mockups.
11. Inteligencia artificial.

1.6. Solución propuesta

1.6.1. Justificación e importancia

La obesidad en el mundo es considerada por la Organización Mundial de la Salud como una Enfermedad No Transmisible (ENT), tal y como lo indican en la Declaración Política de la Reunión de Alto Nivel de la Asamblea General sobre la Prevención y el Control de las Enfermedades No Transmisibles. Es considerada una enfermedad que ataca a más de la mitad de la población mundial. Este problema se agrava con la llegada del COVID 19 y la cuarentena obligatoria. Su prevalencia ha aumentado en los últimos 30-40 años y actualmente de cada 10 niños y adolescentes, uno es obeso. Esto se debe a que, desde muy pequeños, por diferentes circunstancias no tienen una alimentación saludable y balanceada. (Naciones Unidas, 2011)

Si esta condición corresponde a un factor de riesgo o enfermedad primaria es un tema ampliamente discutido. Es reconocida como una enfermedad por la Asociación Médica Estadounidense y la Organización Mundial de la Salud, con base en sus características metabólicas y hormonales, como la desregulación del apetito, el equilibrio energético anormal y la disfunción endocrina, entre otras. Sus principales factores de riesgo ambiental son el consumo de alimentos ultra procesados y el sedentarismo. El tratamiento de la obesidad/sobrepeso se basa en la terapia cognitivo-conductual, la intervención dietética y el aumento de la actividad física con disminución del sedentarismo.

La priorización de combatir la obesidad debería mejorar la calidad de vida, evitando la mortalidad temprana, reduciendo el riesgo cardiovascular y a padecer diabetes tipo 2, así como la incidencia de cualquier tipo de cáncer. Si se llega a tener controlado la obesidad en el país, se tendría un gran impacto en nuestra sociedad, mejorando la calidad de vida de los habitantes. Actualmente, con la pandemia del COVID 19, muchas personas han descuidado sus hábitos alimenticios. Según estudios, las personas hemos tenido un aumento de peso en el confinamiento entre 1 y 7 kg. Lo cual está complicando mucho mantener una vida saludable. Paralelamente, el colapso de los hospitales, o el riesgo de ir a un hospital y contagiarse es latente

en las personas, lo que ha ocasionado que muchas no se realicen chequeos con frecuencia y que a largo plazo desarrollen enfermedades mortales.

Por otro lado, el avance de la tecnología y de la Inteligencia Artificial específicamente se sigue dando de una manera exponencial, a medida que pasa el tiempo se descubren nuevas cosas que pueden ayudarnos a mejorar nuestro estilo de vida, e incluso facilitarnos las cosas que realizamos a diario. La inteligencia artificial no solamente está abarcando en áreas industriales, sino que poco a poco está siendo implementada en el área salud, de tal forma que los diagnósticos, por ejemplo, se hacen en tiempo récord y desde la comodidad del hogar. En el ámbito de la salud, se está perfilando con el tiempo como una herramienta capaz de aprender y analizar con rapidez enormes cantidades de información de los historiales de pacientes, de las pruebas de imagen y de los avances científicos para ayudar a los doctores a ofrecer mejores diagnósticos y tratamientos.

Es por ello por lo que se plantea el desarrollo de una red neuronal para la detección de la obesidad en adolescentes a través de marcadores sencillos, como la estatura, peso, grosor de diferentes partes del cuerpo, etc. El sistema dará predicciones como resultado del análisis de los diferentes datos de un paciente durante una sesión en forma de un reporte, pero de la misma manera ofrece una serie de tratamientos basados en los resultados de las predicciones.

El propósito del proyecto de desarrollo de una red neuronal para predecir la obesidad en adolescentes es servir de ayuda (por medio de aprendizaje automático) a los médicos en general al momento de realizar pruebas de rutina para la detección temprana la obesidad en adolescentes, además de ofrecer un nivel de confiabilidad mayor al 50% en caso la persona posea dicha ENT y de esta forma prevenir unos resultados mortales por la falta de seguimiento y detección en las consultas. Se requiere proporcionar datos reales de evaluaciones realizadas a adolescentes de diferentes edades y realidades; debido a que está usando aprendizaje automático para poder generar buena respuesta basada en hechos reales, de forma que su uso sea justificable.

1.6.1.1. Motivación. beneficios y beneficiarios

El avance de la tecnología y de los métodos de aprendizaje automático dándose con un crecimiento de una manera exponencial, a medida que pasa el tiempo se descubren nuevas formas de prevenir, diagnosticar y tratar diferentes enfermedades a partir de su capacidad de

convertirse en una herramienta capaz de aprender y analizar rápidamente grandes cantidades de información de personas.

A comparación de todos los trabajos que existen hasta el momento, este trabajo de investigación se diferencia en que se propone una red neuronal para la predicción de la obesidad en niños y adolescentes, además que los trabajos existentes no utilizan los mismos parámetros para entrenar los métodos de predicción a partir de datos antropométricos que no tienen mucha complejidad para la obtención (edad, peso, altura, etc.).

Los beneficiarios de este proyecto a largo plazo son todas las personas sin distinción de sexo, edad, etc.; ya que el proyecto puede escalar a no solo ser predicción de obesidad en adolescentes, sino que puede ser predicción de público en general.

1.6.1.2. Factibilidad

1.6.1.2.1. Metodologías

La metodología por utilizar dentro del proyecto de investigación será CRISP-DM, el cual proporciona una descripción estandarizada del ciclo de vida de un proyecto de análisis de datos estándar, análoga a la ingeniería de software con modelos de ciclo de vida de desarrollo de software.

1.6.1.2.2. Lenguajes Aplicables

Uno de los lenguajes aplicables a nuestro proyecto es principalmente Python, ya que será el lenguaje en el que se lleve a cabo la implementación de la red neuronal.

Para el procesamiento de los datos se utilizará la herramienta RapidMiner. Y para la construcción del modelo se utilizará Google Colab.

1.6.2. Descripción de la solución

Durante la investigación se realizará una red neuronal artificial (RNA), en la cual se procesarán los datos recolectados de adolescentes de la provincia de Arequipa para poder obtener el resultado de si presentan obesidad o no. Además, durante la investigación se estudiarán puntos para tener en cuenta al momento de aplicar dicho algoritmo.

2. Fundamentos teóricos

2.1. Estado del arte

La obesidad es una de las enfermedades que ha tenido mayor crecimiento a lo largo de los últimos años: En las últimas décadas se ha triplicado el número de personas que la padecen alrededor del mundo; y a raíz de eso se le considera como "la epidemia del siglo XXI". Lo más peligroso de esta enfermedad son los efectos adversos a los que se exponen estas personas: enfermedades coronarias, problemas de colesterol y triglicéridos; diabetes, accidentes cerebrovasculares e incluso diferentes tipos de cáncer (próstata, mama, colon, etc.). (Centro Nacional de Epidemiología, Prevención y Control de Enfermedades, 2020)

La obesidad tiene una etiología compleja, que incluye tanto factores desarrollados en la etapa del embarazo, así como cambios antropométricos, metabólicos y hormonales que se dan en esta etapa, razón por la cual la incidencia en niños, niñas y jóvenes está aumentando a un ritmo alarmante en muchos países, tomando más fuerza a partir de la llegada de la COVID-19. Esta condición es una amenaza para los sistemas de salud de muchos países, ya que la obesidad está asociada a diversas comorbilidades, tales como enfermedades cardiovasculares, diabetes, síndrome metabólico, etc. (Organización Mundial de la Salud, 2021)

Es esta etapa, los modelos de aprendizaje automático están tomando más fuerza como un método extremadamente útil en el campo de la medicina, ya que tienen con un excelente poder predictivo, la capacidad de modelar relaciones no lineales y complejas entre variable, además de poder lidiar con datos dimensionales, que son típicos en este campo. El uso de estos modelos en la medicina cotidiana reemplaza y facilita el análisis de modelos estadísticos que hoy por hoy, sin usarlos, son difíciles de manejar en algunos casos. En la mayoría de los casos, los modelos estadísticos tradicionales utilizan un conjunto reducido de factores de riesgo, mientras que con aprendizaje automático se puede utilizar otro tipo de variables más complejas. Cuando se comparan las regresiones en el mismo trabajo con los modelos de aprendizaje, es aquí donde se puede visualizar la mejora en los resultados obtenidos con estos modelos, ya que pueden generar mejores porcentajes de predicción que no solo ajusta el conjunto de entrenamiento, sino que también dan mejores resultados en las validaciones realizadas. Enfocándose en la predicción de enfermedades, como la obesidad, se debe considerar que la prevención de modelos de aprendizaje automático han mostrado buenos resultados, no solo en el proceso de

identificación de poblaciones en riesgo, sino también en la búsqueda formas de lograr las metas de prevención. (Peinado Pineda & Díaz Salas, 2021)

El uso de modelos de aprendizaje automático en el campo médico también ofrece nuevas ventajas y conocimientos para la predicción y prevención de enfermedades, así como dar la posibilidad de ser una herramienta de simulación, para que se puedan obtener nuevos conocimientos con enfoques terapéuticos. Por ejemplo, el uso de aprendizaje automático ha aumentado la precisión de la predicción en comparación con los modelos estadísticos de uso frecuente, ya que tienen la capacidad de modelar relaciones no lineales complejas entre variables; asimismo, permite modelar automáticamente datos dimensionales, además de ampliar el conjunto de variables predictoras en los modelos a uno mucho más amplio y de multidominio, lo que también permite utilizar nuevas fuentes de datos complejas distintas a numéricas, tales como texto, imágenes, etc. (Banco Interamericano de Desarrollo, 2020)

Hoy por hoy existen diferentes trabajos y estudios enfocándose en el uso de aprendizaje automático en el campo de la medicina, cada uno diferente, ya que, al ser un campo bastante amplio, permite descubrir cosas nuevas cada día. El uso de biomarcadores y medidas antropométricas para predicción de enfermedades es un tema que hoy en día está tomando más fuerza en el momento de hablar de predicción de enfermedades. Existen varios artículos en los que mencionan que el uso de estos será el futuro que seguirán los nuevos métodos empleados para la predicción de enfermedades. Hablando específicamente de la obesidad, se dice que el IMC hoy en día no es una fuente confiable para detectarla en una persona, ya que esta solamente es una medida imperfecta de acumulación anormal o excesiva de grasa corporal. Existen biomarcadores y medidas antropométricas que ayudan a la detección, tales como la insulina, circunferencia de la cintura, entre otros. Se considera que los biomarcadores y las medidas antropométricas a la larga podrían ayudar a detectar la obesidad de una manera más rápida y menos costosa, a comparación de las resonancias magnéticas. (Nimptsch et al., 2018)

En el año 2018, en la conferencia LACCEI International Multi-Conference for Engineering, Education, and Technology “Innovation in Education and Inclusion” se present el paper “Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents”, en donde se expuso que, a partir de un modelo neuro-difuso ANFIS, el método backpropagation y lógica difusa sobre atributos tales como edad, peso, estatura e IMC, se puede lograr clasificar la obesidad de niños y adolescentes varones con un error de aproximadamente 3%. (Sulla Torres et al., 2018)

Dentro del trabajo “Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Aprendizaje automático Models”, se observa que los autores realizaron una comparativa de los modelos de aprendizaje automático y de regresión, tanto lineal como Lasso; observando que hay una mejor predicción con aprendizaje automático a comparación de métodos tradicionales. Se demostró, además, la importancia del entrenamiento para las predicciones, ya que mejora la exactitud de las pruebas. Para la interpretabilidad, indican que se deben considerar 3 factores: los modelos de regresión multivariable pueden parecer más interpretables, los algoritmos de aprendizaje automático ofrecen resultados parcialmente interpretables y algunos modelos de aprendizaje automático ofrecen un rendimiento e interpretabilidad superiores a la regresión lineal. Consideran también que los factores demográficos, socioeconómicos, médicos y ambientales dentro de las muestras que utilizaron fueron la causa de la prevalencia de la obesidad. Esto se pudo observar a partir de los modelos de aprendizaje empleados ya que, dependiendo de los factores demográficos y socioeconómicos, se evidenció una variación en la obesidad de cada una de las muestras. (Scheinker et al., 2019)

El aprendizaje automático a través de árbol de decisiones puede ayudar a predecir el riesgo de desarrollar enfermedades. Los árboles de decisión identifican las relaciones entre variables y asociaciones con riesgos que pueden no ser identificados por las técnicas tradicionales de análisis epidemiológico, por lo que establecer relaciones entre diferentes polimorfismos se considera una fortaleza de este método. Además, hace hincapié en que la edad y género son determinantes al momento de predecir algún tipo de enfermedad ya que el tamaño y forma corporal de los adultos difieren notoriamente entre hombres y mujeres, además que cambian con el tiempo. Otros factores que influyen en los resultados son etnia y edad. Este trabajo coincide en que los métodos tradicionales son menos efectivos que los métodos de predicción por aprendizaje de máquina. Asimismo, se le considera como demostrativo piloto para las nuevas tecnologías de análisis de datos para la predicción de enfermedades, ya que el uso de estas permitirá crear estrategias tempranas para prevenir enfermedades, además de ser uno de los primeros trabajos que utilizan arboles de decisión con información genética. (Rodríguez-Pardo et al., 2019)

El trabajo titulado “Predicting nationwide obesity from food sales using aprendizaje automático” habla sobre que se puede predecir la obesidad en una ciudad o país determinado a partir de la cantidad que consume de ciertos productos, tales como harina, lácteos y bebidas

gasificadas; para ello utilizaron como modelo árboles aleatorios y librerías como Extreme Gradient Boosting. La evaluación se hizo en productos consumidos por niños menores de dos años con datos de varios países con buena economía en base a la venta de 5 alimentos. Luego de la aplicación de RF pudieron observar que el uso de aprendizaje automático les permitió estimar la obesidad a nivel nacional a partir de las categorías de venta de cinco alimentos, lo cual fue notoriamente menos costoso que las encuestas nacionales. Consideran que utilizar métodos de aprendizaje automático es más cómodo y menos costoso que realizar encuestas, las cuales son muy caras, asimismo, refieren a que los enfoques tradicionales de regresión limitan el análisis a un pequeño conjunto de predictores e imponen suposiciones de independencia y linealidad. (Dunstan et al., 2019)

Algunos factores sociales como efectos adversos de la globalización, crecimiento de los supermercados, urbanización no planificada, sedentarismo, entre otros, desarrollan lentamente factores de riesgo conductual en las personas, que a la larga derivan a tener afecciones en la salud. Es por ello por lo que en 2020 se realizó una revisión de varios métodos de aprendizaje automático y su ejecución utilizando datos de salud de muestra de personas, entre 20 y 60 años, disponibles en repositorios públicos relacionados con enfermedades a consecuencia del estilo de vida, tales como la obesidad. Este estudio utilizado diferentes técnicas para predecir si esa persona es obesa o no, pero, antes que nada, prepararon la data según el modelo que iban a emplear (árbol de decisiones, regresión, etc.). Con esto, pudieron demostrar qué factores son los que influyen para que una persona sea obesa o no. Asimismo, sugieren hacer un estudio similar a este, pero empleando otras técnicas, como redes neuronales. (Chatterjee et al., 2020)

Un estudio realizado en India en el año 2021 propone una CNN para que a través de imágenes térmicas se pueda detectar si una persona es obesa o no, esto gracias a que detectaron que en la zona del abdomen hay una diferencia de temperatura entre ambos casos de personas. De los dos modelos que proponen, se ve un mejor resultado en el modelo personalizado, ya que se le entreno para que evalúe zonas específicas en las imágenes térmicas. Hicieron, además, un pequeño estudio o de los trabajos hasta el momento que realizaron este y concluyeron que son pocos los que existen, por lo que consideran que es un campo aun sin explorar: “La aplicación del aprendizaje profundo en la detección de la condición de obesidad a partir de imágenes térmicas no se investiga en la literatura reciente”. Asimismo, menciona que los modelos entrenados tienen una buena precisión de entrenamiento, pero mala al momento de las pruebas, y esto se debe al sobre entrenamiento. (Snekhalatha et al., 2021)

Desde otro ángulo, no solamente se ha estudiado la predicción de obesidad a partir de variables antropométricas, sino que también de valores que se obtengan de estudios en diferentes órganos del cuerpo, tal es así el caso de un estudio que propone que, a partir de electrorretinografías se puede predecir si una persona es obesa o no. La metodología que se utilizó fueron redes neuronales y modelos de enjambre de partículas basados en redes neuronales. Este estudio demostró que la obesidad está relacionada con los resultados que se obtienen de los electrorretinograma. A través de la optimización del PSO se obtuvo mejores resultados. Tuvieron limitaciones por la cantidad de data para el modelo. (Senyer Yapici et al., 2021)

Paralelamente a considerar solamente variables o datos de la persona a evaluar, hay algunos estudios que consideran a más de una persona para poder predecir enfermedades, tal es el caso del estudio “Ranking of a wide multidomain set of predictor variables of children obesity by aprendizaje automático variable importance techniques", el cual utiliza muchas variables, además de ser uno de los pocos que incluye a los padres como variables que definen al niño, ya que, en este trabajo consideran que lo que los padres tienen como costumbres, el niño también lo hará, El objetivo de este estudio principalmente fue mostrar que hay más de una variable de la que puede depender la predicción, para demostrar esto utilizaron Random Forest y Gradient Boosting Machine. (Marcos-Pasero et al., 2021)

Se sabe que, a través del útero de la madre los bebes absorben todo lo que la madre come, por lo que en esta etapa se debe cuidar mucho la madre ya que hay evidencia que el entorno uterino puede causar una influencia permanente en la salud futura del feto y puede conducir a una mayor susceptibilidad a enfermedades más adelante. Partiendo de ese conocimiento, se desarrolló un estudio en el 2021 que se basa en ver si un niño es propenso a ser obeso o no a partir de rasgos de la madre. Proponen que este factor viene incluso desde antes de nacer, ya que puede depender de cuanto sea la glucosa de la madre durante el embarazo o de la ascendencia. Este trabajo usa valores antropométricos, tales como IMC; datos demográficos, medicamentos, diagnósticos y pruebas de laboratorio de niños y sus familias. Se evaluó el rendimiento de múltiples modelos de predicción para la obesidad infantil a los 5 y 6 años en diferentes edades del niño. A partir de esto, se pudo concluir que los predictores más influyentes al nacer son las mediciones antropométricas de los hermanos, madre y padre. A pesar de que existieron limitaciones con la data, se pudieron obtener datos bastantes preciso. (Rossman et al., 2021)

2.2. Bases teóricas de la investigación

Con la finalidad de comprender la información general del tema que se desarrollará se han escogido una serie de conceptos y temas para tener un mejor dominio del conocimiento:

2.2.1. Inteligencia artificial

La Inteligencia Artificial, también conocida como "IA", es la combinación de algoritmos planteados con el propósito de crear máquinas que presenten las mismas capacidades que el ser humano, para ser más precisos, trata de imitar la mente humana a través de un software que ha sido explícita o implícitamente desarrollado para completar una tarea. Se pretende que la IA se acerque al funcionamiento de la mente humana. Y se recurre a ella cuando se considera útil incorporar a un sistema de ordenadores un conocimiento o comportamiento ante los eventos que serían más propios de un ser humano Una tecnología que todavía es misteriosa al momento de descifrar, pero que desde hace unos años está presente en el día a día de la sociedad. (Cai et al., 2021)

La IA hace posible que las máquinas aprendan de la experiencia, se ajusten a nuevas aportaciones y realicen tareas como seres humanos. Empleando estas tecnologías que se proponen dentro de este, las computadoras pueden ser entrenadas para realizar tareas específicas procesando grandes cantidades de datos y reconociendo patrones en los datos. La inteligencia artificial funciona combinando grandes cantidades de datos con procesamiento rápido e iterativo y algoritmos inteligentes, permitiendo al software aprender automáticamente de patrones o características en los datos a partir de algoritmos. Son capacidades matemáticas de aprendizaje, y de datos que hacen falta para entrenar dichos algoritmos, estos son datos observables, disponibles públicamente o datos generados en algunas empresas, los mismos que repiten el proceso para aprender a partir de ellos. En la actualidad, la inteligencia artificial no solo ha revolucionado el mundo empresarial, sino también el ámbito social, con aplicaciones que van desde la rápida detección del cáncer hasta la lucha contra la deforestación del Amazonas. (Peinado Pineda & Díaz Salas, 2021)

2.2.2. Aprendizaje automático

El 'machine learning', aprendizaje automático, es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Aprende de datos históricos de un producto en cuestión, del contexto y atributos de observación. El

sistema aprende a partir de lo que se le da. Esta tecnología está presente en un sinfín de aplicaciones como las recomendaciones de Netflix o Spotify, las respuestas inteligentes de Gmail o el habla de Siri y Alexa. (BBVA, 2019)

Es un método analítico que permite que un sistema por sí mismo, sin ayuda de un ser humano y en forma automatizada, aprenda a descubrir patrones, tendencias y relaciones en los datos, y gracias a dicho conocimiento, en cada interacción con información nueva se ofrecen mejores perspectivas. Se le considera como una herramienta que buscan mejorar el análisis de datos, en pro de una predicción futura, ya sea por la implementación de nuevos sistemas o simplemente el mejoramiento de los ya existentes, mediante el uso de algoritmos basados en información antigua o reciente que permita el funcionamiento óptimo del sistema a trabajar. (Ramirez Hinestroza, 2018)

La tecnología del Aprendizaje Automático está sirviendo para recopilar y modelar el conocimiento, con el fin de proporcionar información más específica y elaborar mejores herramientas de trabajo para las personas. El uso de algoritmos marcará la competitividad y la profesionalidad durante los próximos años. Por ello, no son pocas las empresas que utilizan el aprendizaje automático en sus servicios y productos, aprovechando los beneficios que puede reportar su aplicación, tanto para los procesos de sus organizaciones como para mejorar la experiencia de trabajo y entretenimiento de sus clientes. (Banu, 2022)

2.2.3. Redes neuronales

Las redes neuronales son una de las familias de algoritmos de aprendizaje automático. Se trata de una técnica que se inspira en el funcionamiento de las neuronas de nuestro cerebro. Se basan en una idea sencilla: dados unos parámetros hay una forma de combinarlos para predecir un cierto resultado. Los datos de entrada van pasando secuencialmente por distintas " capas" en las que se aplican una serie de reglas de aprendizaje moduladas por una función peso. Tras pasar por la última capa, los resultados se comparan con el resultado "correcto", y se van ajustando los parámetros dados por las funciones "peso". El sistema de redes neuronales artificiales es una respuesta al enorme incremento en las dimensiones y números de datos disponibles. Durante la última década, los avances en hardware han producido una explosión de capacidad computacional, lo que ha permitido generar y evaluar el trabajo del sistema con gran velocidad. (Kumar et al., 2022)

2.2.4. Malnutrición y obesidad

Figura 1: Obesidad a nivel mundial

Actualmente, se le estima a la malnutrición como el problema más grande que la sociedad encara, ya que es una condición que afecta a uno de cada 3 personas en el mundo. La malnutrición se muestra de distintas modalidades: Retraso de aumento en chicos, personas propensas a infecciones gracias a la falta de vitaminas y minerales de trascendencia en su organismo, personas con exceso de peso o riesgo a sufrir enfermedades crónicas a raíz del sobrepeso o por excesivo consumo de azúcar, sal o grasas. La malnutrición y la alimentación es una enorme carga mundial de morbilidad (CMM), ya que cada territorio hace frente a una realidad distinta con base a la economía, sociedad que tienen. Por ejemplo, en países con pobreza extrema, la desnutrición en niños y adolescentes es mucho más grave que la que pueden hacer frente países tercermundistas, sin embargo, estos además afrontan el sobrepeso y obesidad en más enorme medida que otros países. El sobrepeso es una de las enfermedades no transmisibles (ENT) más comunes alrededor del mundo. A esta se le define como una "acumulación anormal o excesiva de grasa que puede ser perjudicial para la salud". (Organización Mundial de la Salud, 2021)

En la actualidad, los problemas de salud derivados de la obesidad han penetrado todos los segmentos sociales del Perú. Según los datos del módulo nutricional de la Encuesta Nacional de Hogares del año 2008, indica que el sobrepeso en niños era de un 7.8%. Sobre los adolescentes entre 10 y 19 años, rango de edad donde se manifiestan muchos cambios en tamaño y forma, se sabe que el sobrepeso y obesidad era del 13.5% en varones y 15% de mujeres, además, la Encuesta Global de Salud Escolar, realizada en el 2010, es decir dos años después de la primera, no indica ningún cambio, al contrario, deja a la luz que el problema se agravó, ya que reportó que el 20% de escolares de secundaria presentan sobrepeso y el 3% presentan obesidad. En adultos este problema no mejora, ya que personas mayores entre 35 y 40 años, el 66% de mujeres y el 55% de hombres padecen de obesidad. (Ministerio de Salud del Perú, 2012)

3. Marco metodológico

3.1. Alcances y limitaciones

Para el desarrollo de este proyecto se tomó en cuenta diferentes documentos de índole nacional e internacional para poder contextualizar el problema de la obesidad en adolescentes de manera más precisa que tienen como propósito dar a conocer en qué estado se encuentra la obesidad tanto en el Perú como a nivel mundial, y al mismo tiempo proponer soluciones a nivel gubernamental a fin de disminuir dicha enfermedad no transmisible. Es por ello, por lo que se le consideran importantes ya que proporcionan la información suficiente para justificar el desarrollo del proyecto.

Del mismo modo, se toma en cuenta los datos de personas entre 12 y 21 años de diferentes colegios de la ciudad de Arequipa, Perú, sobre las características clínicas que se observaron y midieron en un registro de 3068 adolescentes, tales como edad, peso, talla, IMC. En sí, el proyecto engloba dos campos de estudio como lo son la computación y la medicina.

El propósito del proyecto es proponer una ayuda (por medio de machine-learning) a los médicos y público en general al momento de realizar pruebas de rutina para la detección temprana de la obesidad en adolescentes, además de ofrecer un nivel de confiabilidad mayor al 50% en caso la persona posea dicha enfermedad no transmisible (ENT) y esta forma prevenir unos resultados mortales por la falta de seguimiento y detección en las consultas.

Se requiere que se le proporcionen datos reales de evaluaciones realizadas a adolescentes de diferentes centros educativos de la ciudad de Arequipa, debido a que está usando machine-learning para poder generar buena respuesta basado en hechos reales, de forma que su uso sea justificable. Los datos deben ser susceptibles al análisis para poder extraer conclusiones científicamente válidas.

Las características requeridas por el sistema, para lograr su objetivo son:

Captura de información

- Se establece la captura de datos.

Consolidación de información

- Almacenar los datos, previamente localizados, en un archivo csv.

- Normalizar la información para asegurar la recuperabilidad, clasificación y orden de los datos.

Aprendizaje

- Crear una red neuronal que sea capaz de tomar la información almacenada y aprender de ella, para crear sus propias conclusiones

- Crear un formato específico que permita plasmar los resultados de la red neuronal

Procesos

Dentro de los procesos implicados a los que el sistema propuesto va a ayudar es a la detección de posible obesidad en adolescentes, tomando en cuenta los valores antropométricos que presente la persona a ser evaluada.

Áreas implicadas

Las áreas implicadas para la realización de este sistema inteligente son las áreas de biomédicas y computación; específicamente dentro del área de biomédicas el área de la endocrinología; y dentro del área de computación tenemos del área de inteligencia artificial.

El software que se va a desarrollar está diseñado para facilitar el proceso de detección de un posible caso de obesidad en adolescentes tomando en cuenta los valores antropométricos que presente la persona a ser evaluada.

Es por ello, que esta investigación tiene los siguientes puntos de alcance y limitaciones:

a) Se tomarán los datos antropométricos de adolescentes de colegios de la provincia de Arequipa, región Arequipa, país Perú.

b) Se realizará este trabajo de investigación en un ambiente donde se tendrá la posibilidad de usar herramientas para el procesamiento de datos, tales como RapidMiner; así como herramientas para la creación de la red neuronal tales como PyCharm, Google Colab, etc.

c) El tiempo tomado para realizar este trabajo es de un (1) año aproximadamente, tiempo en el cual se trabajará el preprocesamiento de datos, procesamiento de datos, ajuste de red neuronal, así como la documentación de la investigación.

3.2. Aporte

Este estudio brindará un aporte científico ya que ayudará a la predicción temprana de la obesidad en adolescentes a través de datos fáciles de obtener (peso, sexo, etc.), ya que, una vez culminado este trabajo, se obtendrá el resultado del análisis empleado en los datos, y a partir de ellos, se mejorará la red neuronal a fin de que el porcentaje de exactitud sea muy cercano al 100%. Por otro lado, también ayudará a hacer un mejor uso de las tecnologías modernas como el aprendizaje automático ya que dentro del área de la medicina son muy pocos los campos donde se utiliza.

3.3. Tipo y nivel de investigación

3.3.1. Según su finalidad:

Aplicada.

3.3.2. Según la fuente de datos:

Empírica o de campo.

3.3.3. Según el contexto histórico:

Tradicional.

3.3.4. Nivel de investigación:

Exploratoria.

3.4. Población y muestra o universo

Se obtuvieron 3068 registros personas entre 4 y 21 años.

3.5. Métodos, técnicas e instrumentos de recolección de datos

3.5.1. Métodos de la investigación

Investigación predictiva.

3.5.2. Técnicas para la investigación

De forma inicial, en la tabla de datos tenemos las siguientes columnas:

a) Edad

b) Sexo

c) Peso

d) Estatura_cm

e) Estatu_metr

f) Índice de Masa Corporal (IMC)

Cada una de ellas son consideradas como variables, ya sea categóricas o numéricas. Las variables categóricas se preprocesarán a fin de convertirlas en variables numéricas. Luego de ello, se realizarán los procesamientos en test y en producción para poder obtener resultados.

3.5.3. Instrumentos para tratamiento de datos

Se utilizará una tabla en la que se visualizarán los datos recolectados, los cuales se utilizarán en la etapa experimental del trabajo de investigación. En dicha tabla, se agregarán las columnas con datos que se necesitan para asegurar una exactitud más cercana al 100%.

4. Plan de trabajo

El desarrollo se llevará a cabo en base a fases con una o más iteraciones en cada una de ellas. La siguiente tabla muestra una la distribución de tiempos y el número de iteraciones de cada fase.

Fase	Cantidad de iteraciones	Duración
Planteamiento teórico/ Marco Teórico	1	13 semanas
Comprensión de los requisitos del negocio	1	05 semanas
Comprensión de los datos	1	06 semanas
Preparación de los datos	1	06 semanas
Búsqueda/Modelado	6	12 semanas
Evaluación	1	04 semanas
Implementación	1	06 semanas
Resultados/Conclusiones	1	04 semanas
	TOTAL	56 semanas

Las actividades por realizar se mencionan en la siguiente tabla:

Descripción	Hito
Planteamiento teórico/ Marco Teórico	En estos capítulos se explicará el contexto en el que se desarrollará la investigación, así como se expondrá el estado del arte y el marco teórico necesario para entenderlo.
Comprensión de los requisitos del negocio	Dentro de esta fase se determinarán los objetivos que se quieren alcanzar con el desarrollo del proyecto, luego se evaluará la situación en la que se encuentra, determinando, además, los objetivos del aprendizaje de máquina, para finalmente definir cuál es el plan que se seguirá para llegar al objetivo planteado.
Comprensión de los datos	En esta segunda fase se realizará la recolección, análisis y verificación de datos con el propósito de conocerlos para que en la fase 3 sean preparados de la mejor manera. Esta fase comienza con la recopilación de datos, seguido de la descripción formal de los mismos, así como su exploración y verificación. El propósito de estas actividades es poder familiarizarse con los datos, identificando el grado de calidad que tienen y descubriendo conocimientos preliminares necesarios, así como subconjuntos que puedan ayudar para llegar al objetivo.
Preparación de los datos	Dentro de esta fase se preparará los datos de cara a la fase 4 de modelado. Implicará seleccionar los datos a utilizar, limpiarlos para garantizar una mejor consistencia de estos, construir datos sintéticos de ser necesarios, así como integración de atributos de diferente tabla, para finalmente formatearlos.
Búsqueda/Modelado	En esta cuarta fase de la metodología se escogerá la técnica más adecuada para cumplir el objetivo planteado. Luego de eso se diseñará el plan de pruebas (test) a aplicar en el modelo que se construirá en un tercer paso y finalmente se tendrá que evaluar a fin de ver si cumple con los criterios planteados o no.
Evaluación	En esta fase de la metodología CRISP-DM se evaluará a más profundidad los modelos generados en la fase anterior, pero con un enfoque de los objetivos planteados para este trabajo. Una vez realizada esta evaluación se revisará el proceso a seguir y se determinarán los pasos siguientes para la fase de implementación.
Implementación	En esta última fase de la metodología CRISP-DM se explicará cómo se realizará la implementación del modelo, se realizará la implementación tanto del modelo, para finalmente exponer los resultados en el siguiente capítulo del documento.
Resultados/Conclusiones	En este capítulo se realizará un recuento de cada uno de los pasos seguidos a través de la metodología, así como se explicarán los resultados obtenidos y se realizarán pruebas a fin de verificar que el modelo es funcional. Finalmente, se desarrollarán las conclusiones obtenidas a partir del modelo creado.

El cronograma del desarrollo del siguiente proyecto se adjunta en el Anexo A.

5. Referencias

BBVA. (08 de noviembre de 2019). 'Machine learning': ¿qué es y cómo funciona? Obtenido de 'Machine learning': ¿qué es y cómo funciona?: https://www.bbva.com/es/machine-learning-que-es-y-como-funciona/

Centro Nacional de Epidemiología, Prevención y Control de Enfermedades. (2020). Editorial: La obesidad como problema de salud pública. En P. y. Centro Nacional de Epidemiología, Boletín Epidemiológico del Perú 2020 (págs. 293-294). Lima: Ministerio de Salud.

Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. Noruega: Sensors. doi:10.3390/s20092734

Collave Garcia, Y. (05 de marzo de 2021). “El 75% de muertes por COVID-19 se relaciona con sobrepeso u obesidad en Perú”. Diario El Comercio, pág. Virtual.

Comisión para acabar con la obesidad infantil. (2016). Informe de la Comisión para acabar con la obesidad infantil. Ginebra: Catalogación por la Biblioteca de la OMS.

Dunstan, J., Aguirre, M., Bastías, M., Nau and Thomas A Glass, C., & Tobar, F. (2019). Predicting nationwide obesity from food sales using machine learning. Chile: Health Informatics Journal. doi:10.1177/1460458219845959

IBM. (2020). Redes Neuronales. Obtenido de IBM: https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=nodes-neural-networks

Marcos Pasero, H., Colmenarejo, G., Aguilar, E., Ramírez de Molina, A., Reglero, G., & Loria Kohen, V. (2021). Ranking of a wide multidomain set of predictor variables of children obesity by machine learning obesity by machine learning. Nature - Scientifics Reports. doi:10.1038/s41598-021-81205-8

Naciones Unidas. (2011). Declaración Política de la Reunión de Alto Nivel de la Asamblea General sobre la Prevención y el Control de las Enfermedades No Transmisibles. Seguimiento de los resultados de la Cumbre del Milenio (pág. 14). Nueva York: Naciones Unidas. Recuperado el 12 de 04 de 2021, de https://undocs.org/es/A/66/L.1

Nimptsch, K., Konigorski, S., & Pischon, T. (2018). Diagnosis of obesity and use of obesity biomarkers in science and clinical medicine. Alemania: Metabolism Clinical and Experimental. doi: 10.1016/j.metabol.2018.12.006

Ramirez Hinestroza, D. (2018). EL MACHINE LEARNING A TRAVÉS DE LOS TIEMPOS, Y LOS APORTES A LA HUMANIDAD. Pereira: UNIVERSIDAD LIBRE SECCIONAL PEREIRA.

Rodríguez-Pardo, C., Segura, A., Zamorano-León, J. J., Martínez-Santos, C., Martínez, D., Collado-Yurrita, L.... López-Farre, A. (2019). Decision tree learning to predict overweight/obesity based on body mass index and gene polymporphisms. Madrid: Gene. doi:10.1016/j.gene.2019.03.011

Rossman, H., Shilo, S., Barbash-Hazan, S., Shalom Artzi, N., Hadar, E., Balicer, R. D.,... Segal, E. (2021). Proponen que este factor viene incluso desde antes de nacer, ya que depende de cuanto sea la glucosa de la madre durante el embarazo. Canada: The Journal Of Pediatrics. doi:10.1016/j.jpeds.2021.02.010

Scheinker, D., Valencia, A., & Rodriguez, F. (2019). Identification of Factors Associated With Variation in US County-Level Obesity Prevalence Rates Using Epidemiologic vs Machine Learning Models. Estados Unidos: JAMA Network. doi:10.1001/jamanetworkopen.2019.2884 (Re

Senyer Yapici, I., Erkaymaz, O., & Uzun Arslan, R. (2021). A hybrid intelligent classifier to estimate obesity levels based on ERG signals. Turquía: Physics Letters A. doi:10.1016/j.physleta.2021.127281

Snekhalatha, U., Palani Thanaraj, K., & Sangamithirai, K. (2021). Computer aided diagnosis of obesity based on thermal imaging using various convolutional neural networks. India: Biomedical Signal Processing and Control. doi:10.1016/j.bspc.2020.102233

6. Posible temario de informe final

Índice o Tabla de Contenidos

Dictamen Aprobatorio del Borrador de Tesis

Presentación

Agradecimientos

Dedicatoria

Epígrafe

Índice o Tabla de Contenidos

Índice de Tablas

Índice de Figuras

Resumen

Palabras claves

Abstract

Keywords

Capítulo 1: Planteamiento teórico

1.1. Introducción

1.1.1. Antecedentes

1.1.2. Objetivos

1.1.3. Enfoque

1.1.4. Alcances y limitaciones

1.1.5. Aporte

1.1.6. Preguntas de investigación

1.1.7. Línea, sublínea, tipo y nivel de investigación

1.1.8. Cobertura del estudio

1.1.9. Métodos, Técnicas e Instrumentos para la investigación y tratamiento de datos

1.1.10. Solución propuesta

1.1.11. Metodologías, Modelos, Lenguajes Aplicables

1.2. Fundamentos teóricos

1.2.1. Estado del arte

1.3. Organización de la tesis

7. Anexo A: Cronograma de desarrollo de proyecto

	Nombre	Duración	Inicio	Terminado	Predecesores
2	**Capítulo 1: Planteamiento teórico**	**26 días**	**19/07/21 08:00 AM**	**23/08/21 05:00 PM**	
3	1.1. Introducción	2 días	19/07/21 08:00 AM	20/07/21 05:00 PM	
4	1.1.1. Antecedentes	2 días	21/07/21 08:00 AM	22/07/21 05:00 PM	3
5	1.1.2. Objetivos	2 días	21/07/21 08:00 AM	22/07/21 05:00 PM	3
6	1.1.3. Enfoque	2 días	21/07/21 08:00 AM	22/07/21 05:00 PM	3
7	1.1.4. Alcances y limitaciones	2 días	23/07/21 08:00 AM	26/07/21 05:00 PM	6
8	1.1.5. Aporte	2 días	23/07/21 08:00 AM	26/07/21 05:00 PM	6
9	1.1.6. Preguntas de investigación	2 días	23/07/21 08:00 AM	26/07/21 05:00 PM	6
10	1.1.7. Línea, sublínea, tipo y nivel de investiga...	2 días	27/07/21 08:00 AM	28/07/21 05:00 PM	9
11	1.1.8. Cobertura del estudio	2 días	27/07/21 08:00 AM	28/07/21 05:00 PM	9
12	1.1.9. Métodos, Técnicas e Instrumentos para...	2 días	27/07/21 08:00 AM	28/07/21 05:00 PM	9
13	1.1.10. Solución propuesta	2 días	29/07/21 08:00 AM	30/07/21 05:00 PM	12
14	1.1.11. Metodologías, Modelos, Lenguajes Apli...	2 días	29/07/21 08:00 AM	30/07/21 05:00 PM	12
15	**1.2. Fundamentos teóricos**	**16 días**	**02/08/21 08:00 AM**	**23/08/21 05:00 PM**	
16	1.2.1. Estado del arte	16 días	02/08/21 08:00 AM	23/08/21 05:00 PM	14
17	1.3. Organización de la tesis	3 días	24/08/21 08:00 AM	26/08/21 05:00 PM	16
18	**Capítulo 2: Marco Teórico**	**36 días**	**27/08/21 08:00 AM**	**15/10/21 05:00 PM**	
19	**2.1. Definiciones, Acrónimos, y Abreviaci...**	**36 días**	**27/08/21 08:00 AM**	**15/10/21 05:00 PM**	

20	2.1.1. Malnutrición y obesidad	4 días	27/08/21 08:00 AM	01/09/21 05:00 PM	17
21	2.1.2. Inteligencia artificial	4 días	02/09/21 08:00 AM	07/09/21 05:00 PM	20
22	2.1.3. Machine learning	4 días	08/09/21 08:00 AM	13/09/21 05:00 PM	21
23	2.1.4. Redes neuronales	4 días	14/09/21 08:00 AM	17/09/21 05:00 PM	22
24	2.1.5. Metodología CRISP-DM	4 días	20/09/21 08:00 AM	23/09/21 05:00 PM	23
25	2.1.6. Biomarcadores	4 días	24/09/21 08:00 AM	29/09/21 05:00 PM	24
26	2.1.7. Medidas antropométricas	4 días	30/09/21 08:00 AM	05/10/21 05:00 PM	25
27	2.1.8. RapidMiner	4 días	06/10/21 08:00 AM	11/10/21 05:00 PM	26
28	2.1.9. Lenguaje Python	4 días	12/10/21 08:00 AM	15/10/21 05:00 PM	27
29	**Capítulo 3: Análisis, construcción y evalua...**	**195 días**	**18/10/21 08:00 AM**	**15/07/22 05:00 PM**	
30	**3.1. Comprensión de los requisitos del ne...**	**25 días**	**18/10/21 08:00 AM**	**19/11/21 05:00 PM**	
31	3.1.1. Determinar los objetivos	9 días	18/10/21 08:00 AM	28/10/21 05:00 PM	28
32	3.1.2. Evaluación de la situación	8 días	29/10/21 08:00 AM	09/11/21 05:00 PM	31
33	3.1.3. Realizar el plan de proyecto	8 días	10/11/21 08:00 AM	19/11/21 05:00 PM	32
34	**3.2. Comprensión de los datos**	**30 días**	**22/11/21 08:00 AM**	**31/12/21 05:00 PM**	
35	3.2.1. Recolección y adaptación de datos inic...	6 días	22/11/21 08:00 AM	29/11/21 05:00 PM	33
36	3.2.2. Descripción formal de los datos	8 días	30/11/21 08:00 AM	09/12/21 05:00 PM	35
37	3.2.3. Exploración de datos	8 días	10/12/21 08:00 AM	21/12/21 05:00 PM	36
38	3.2.4. Verificación de datos	8 días	22/12/21 08:00 AM	31/12/21 05:00 PM	37
39	**3.3. Preparación de los datos**	**30 días**	**03/01/22 08:00 AM**	**11/02/22 05:00 PM**	

40	3.3.1. Selección de datos	6 días	03/01/22 08:00 AM	10/01/22 05:00 PM	38
41	3.3.2. Limpieza de datos	6 días	11/01/22 08:00 AM	18/01/22 05:00 PM	40
42	3.3.3. Construcción de datos	6 días	19/01/22 08:00 AM	26/01/22 05:00 PM	41
43	3.3.4. Integración de datos	6 días	27/01/22 08:00 AM	03/02/22 05:00 PM	42
44	3.3.5. Formateado de datos	6 días	04/02/22 08:00 AM	11/02/22 05:00 PM	43
45	**3.4. Búsqueda/Modelado**	**60 días**	**14/02/22 08:00 AM**	**06/05/22 05:00 PM**	
46	3.4.1. Selección de la técnica de modelado	10 días	14/02/22 08:00 AM	25/02/22 05:00 PM	44
47	3.4.2. Diseño del test	10 días	28/02/22 08:00 AM	11/03/22 05:00 PM	46
48	3.4.3. Construcción del modelo	25 días	14/03/22 08:00 AM	15/04/22 05:00 PM	47
49	3.4.4. Evaluación del modelo	15 días	18/04/22 08:00 AM	06/05/22 05:00 PM	48
50	**3.5. Evaluación**	**20 días**	**09/05/22 08:00 AM**	**03/06/22 05:00 PM**	
51	3.5.1. Evaluación de los resultados	10 días	09/05/22 08:00 AM	20/05/22 05:00 PM	49
52	3.5.2. Revisión del proceso	5 días	23/05/22 08:00 AM	27/05/22 05:00 PM	51
53	3.5.3. Determinación de los próximos pasos	5 días	30/05/22 08:00 AM	03/06/22 05:00 PM	52
54	**3.6. Implementación**	**30 días**	**06/06/22 08:00 AM**	**15/07/22 05:00 PM**	
55	3.6.1. Planeamiento de implementación de m..	5 días	06/06/22 08:00 AM	10/06/22 05:00 PM	53
56	3.6.2. Planeamiento de la monitorización y m..	5 días	13/06/22 08:00 AM	17/06/22 05:00 PM	55
57	3.6.3. Desarrollo de producto final	15 días	20/06/22 08:00 AM	08/07/22 05:00 PM	56
58	3.6.4. Revisar el proyecto	5 días	11/07/22 08:00 AM	15/07/22 05:00 PM	57
59	**Capítulo 4: Resultados**	**15 días**	**18/07/22 08:00 AM**	**05/08/22 05:00 PM**	

60	4.1. Recuento de la metodología CRISP-DM dur...	5 días	18/07/22 08:00 AM	22/07/22 05:00 PM	58
61	**4.2. Resultados del modelo implementado**	**5 días**	**25/07/22 08:00 AM**	**29/07/22 05:00 PM**	
62	4.2.1. Pruebas del modelo implementado	5 días	25/07/22 08:00 AM	29/07/22 05:00 PM	60
63	4.3. Análisis y discusión del modelo implementad	5 días	01/08/22 08:00 AM	05/08/22 05:00 PM	61
64	Conclusiones	5 días	08/08/22 08:00 AM	12/08/22 05:00 PM	63
65	Recomendaciones y Trabajos Futuros	5 días	08/08/22 08:00 AM	12/08/22 05:00 PM	63
66	Referencias	254 días	19/07/21 08:00 AM	07/07/22 05:00 PM	
67	**Apéndice(s)**	**170 días**	**22/11/21 08:00 AM**	**15/07/22 05:00 PM**	
68	Apéndice A: Glosario de terminologías de apren...	170 días	22/11/21 08:00 AM	15/07/22 05:00 PM	33
69	Apéndice B: Resultados de entrenamiento de r..	15 días	20/06/22 08:00 AM	08/07/22 05:00 PM	56