TRABAJO DE GRADO
Opción Seminario-Diplomado.

ALGORITMO COMPUTACIONAL PARA EL ANÁLISIS Y TOMA DE DECISIONES EN DATOS DE OSTEOPOROSIS, UTILIZANDO ESTRATEGIAS DE MACHINE LEARNING

CORPORACIÓN UNIVERSITARIA REMINGTON.
FACULTAD DE INGENIERÍA
TECNOLOGÍA EN DESARROLLO DE SOFTWARE

Estudiante:
Sebastian García Ramírez

Tutor: Juan Carlos Briñez de León
Opción de Trabajo de grado Seminario-Diplomado.
2024.

Agradecimientos

A mi familia, su apoyo inquebrantable, su comprensión infinita y su constante motivación han sido mi mayor impulso en este este proceso académico.

A los profesores e institución educativa, por su orientación, sabiduría y dedicación en guiarnos en este viaje académico.

Contenido

Figuras

Resumen

La llegada de las tecnologías de la información y las comunicaciones (TIC) en el siglo XX abrió una nueva etapa, denominada "sociedad del conocimiento y la información", que ha tenido un impacto significativo en las prácticas educativas y sociales. La Inteligencia Artificial (IA), una rama de las TIC, que actualmente busca aplicaciones en sectores como la salud, emulando el razonamiento humano, el aprendizaje, la resolución de problemas y la percepción. El desarrollo de estrategias computacionales basadas en algoritmos de machine learning (ML) ha sido impulsado por la OMS y enfocándolo en La osteoporosis un trastorno esquelético de alto impacto socioeconómico. Estos algoritmos se entrenan con datos demográficos, de estilo de vida y de salud para identificar personas en riesgo rápidamente.

El procesamiento de datos implica cargar conjuntos de datos en plataformas como Colab, eliminar filas y columnas innecesarias y duplicados, y normalizar datos categóricos. Variables como el género, los cambios hormonales, los antecedentes familiares, la actividad física y el consumo de sustancias se analizan. El modelo de toma de decisiones, que se basa en el aprendizaje supervisado, utiliza un algoritmo para clasificar a las personas en riesgo en función de patrones que se encuentran en los datos de entrenamiento. Esto implica recopilar y procesar datos de pacientes, elegir los algoritmos de clasificación apropiados y entrenar el modelo. Un conjunto de datos de prueba se utiliza para evaluar el modelo para comprender su rendimiento y factores de impacto.

El objetivo de la implementación de esta estrategia computacional es mejorar el diagnóstico, la prevención y el tratamiento individualizado de la osteoporosis. Esta

estrategia podría usarse en herramientas de diagnóstico, selección de tratamiento, modelos de riesgos y campañas de concientización.

Palabras clave

Osteoporosis, Datasets, Análisis de datos, Machine learning, Clasificación, regresión, modelo MLP, Clustering, Inteligencia artificial (IA), Algoritmos, Salud digital, eSalud, Ciber salud, Asistencia virtual, Dispositivos inteligentes, demografía.

Marco conceptual y contextual

Desde el pasado siglo la irrupción de las tecnologías de la información y las comunicaciones (TIC) en la vida de las personas aportaron nuevas formas de comunicación social, lo que condicionó una nueva era a partir del siglo XXI, denominada "sociedad de la información y el conocimiento" que, unido a la apropiación de las TIC, marcó insólitos retos y oportunidades mediante el desarrollo de la información, el conocimiento y el aprendizaje. De este modo, constituyó un referente de innovación tecnológica disruptivo, en los métodos y formas educativas, dada su penetración y masividad; tornó los métodos habituales arcaicos; y transformó en nuevas las formas de desarrollo de los contenidos, la actuación, el pensamiento social, el procesamiento de información, etc. Vidal Ledo, M. J., Delgado Ramos, A., Rodríguez Díaz, A., Barthelemy Aguilar, K., & Torres Ávila, D. (2022).

Ahora bien, en la actualidad encontramos una rama de las "TIC" llamada inteligencia artificial (AI) la es busca crear máquinas que puedan realizar tareas que tradicionalmente requieren inteligencia humana, dentro de las cuales encontramos el razonamiento lógico el cual tiene la capacidad de reunir y procesar información de cualquier tipo y dar conclusiones lógicas. El aprendizaje el cual tiene la capacidad de almacenar y adquirir nueva información. La resolución de problemas el cual identifica los problemas de forma precisa y brinda posibles soluciones de manera eficaz. La percepción y comunicación la cual tiene la capacidad de comprender e interactuar con el entorno a través de sensores o asimismo con otros sistemas externos. Boden, M. A. (2017). Inteligencia artificial. Turner.

Actualmente la tecnología, específicamente en el sector la salud, ha sido de vital importancia para abordar desafíos complejos y significativos como lo es la osteoporosis, un trastorno esquelético que afecta la calidad de vida de millones de personas en todo el mundo. La osteoporosis se caracteriza por una baja densidad ósea y un deterioro de la microarquitectura del tejido óseo, lo que aumenta considerablemente el riesgo de fracturas. De hecho, las fracturas resultantes representan un importante problema de salud pública, con costos sustanciales asociados. En Inglaterra y Gales, por ejemplo, se estima que las fracturas relacionadas con la osteoporosis generan un coste anual de 1.700 millones de libras esterlinas, siendo las fracturas de cadera las más costosas. En los Estados Unidos, el costo total de la osteoporosis supera los 14 mil millones de dólares al año, una cifra que se espera que aumente a medida que la población envejezca. Christodoulou, C. y Cooper, C. (2003).

Al tener presentes dichos acontecimientos, la incorporación de nuevas tecnologías como el machine learning una rama de la inteligencia artificial, en el sector de la salud brinda una gran variedad de oportunidades brillantes para mejorar el diagnostico, la prevención de lesiones y fracturas, tratamientos personalizados acorde a las características de las personas, entre otros. Lo que a largo plazo puede contribuir considerablemente a disminuir la carga económica y social de esta enfermedad.

Machine learning (aprendizaje automatizado) se encarga de generar algoritmos que tienen la capacidad de aprender y resolver situaciones por sí solo a partir de un análisis de datos. Además, cuantos más datos tenga mejor serán los resultados. Para realizar el análisis se utilizan algoritmos que diseñan otros datos según las necesidades. A través de los datos

de entrada, ML ejecutar un algoritmo y como resultado, genera más información para el problema. El objetivo de generar más datos se basa en las siguientes técnicas: regresión lineal y polinómica, arboles de decisión, redes neuronales, red bayesiana, cadenas de Márkov. Estas técnicas permiten a ML reconocer patrones, extraer conocimiento, descubrir información y hacer predicciones. Rojas, E. M. (2020).

Los avances recientes en machine learning han permitido al campo de la inteligencia artificial (IA) lograr avances impresionantes en entornos de datos complejos donde la capacidad humana para identificar relaciones de alta dimensión es limitada. El campo de la osteoporosis es uno de esos dominios, a pesar de las preocupaciones técnicas y clínicas con respecto a la aplicación de métodos de ML. Smets, J., Shevroja, E., Hügle, T., Leslie, WD y Hans, D. (2020).

Por ello el machine learning ha sido objeto de atención de la Organización Mundial de la Salud (OMS), cuyas orientaciones sobre Salud digital o eSalud, “ciber salud”, entre otros, abarcan un amplio concepto, que implica la aplicación de las “AI” en los sistemas de vigilancia, prevención, promoción y atención a la salud, así como en la educación, los conocimientos y las investigaciones; e instan a desarrollar un plan estratégico a largo plazo para el desarrollo de infraestructuras tecnológicas y su implantación en los servicios de salud. La estrategia de salud digital establece las tecnologías digitales como determinantes del futuro de la salud mundial; incluso, plantea la transformación digital como un proceso que puede resultar perturbador. Pero las tecnologías aplicadas como la asistencia virtual, la supervisión a distancia, la inteligencia artificial, los dispositivos inteligentes y muchas otras constituyen herramientas que conforman un ecosistema para

una continuidad asistencial. Esto favorece los resultados en salud, ya que puede mejorar los diagnósticos médicos, las decisiones terapéuticas basadas en datos, las terapias digitales, los ensayos clínicos; en fin, la atención centrada en las personas. Asimismo, amplía los conocimientos, las aptitudes y las competencias de los profesionales y prestadores de servicios de salud. Vidal Ledo, M. J., Delgado Ramos, A., Rodríguez Díaz, A., Barthelemy Aguilar, K., & Torres Ávila, D. (2022).

Pregunta problema:

¿Cómo desarrollar una estrategia computacional basada en algoritmos de Machine Learning para la identificación temprana y oportuna de personas con riesgo de osteoporosis, utilizando un conjunto de datos que incluye información demográfica, estilo de vida, historial médico e indicadores de salud?

Acercamiento a los datos:

El conjunto de datos ofrece información completa sobre los factores de salud que influyen en el desarrollo de la osteoporosis, incluidos detalles demográficos, opciones de estilo de vida, historial médico e indicadores de salud ósea. Fue publicado el 13 de noviembre de 2021 en la plataforma Kaggle, un sitio web donde se comparten conjuntos de datos sobre salud, finanzas, comercio, política, ciencia, entre otros.

Para una mejor visualización del conjunto de datos pueden visitar el siguiente enlace:

- https://www.kaggle.com/datasets/amitvkulkarni/lifestyle-factors-influencing-osteoporosis

Descripción de variables.

- Id: número identificador único para cada individuo.

  Tipo de dato: numérico.

- Edad: edad del individuo en años.

  Tipo de dato: numérico.

- Género: genero del individuo (hombre, mujer).

  Tipo de dato: object y/o categórico.

- Cambios hormonales: el individuo ha experimentado cambios hormonales relevantes.

  Tipo de dato: object y/o categórico.

- Historia familiar: el individuo tiene familiares con antecedentes de osteoporosis.

  Tipo de dato: object y/o categórico.

- Raza/etnia: hace referencia a la clasificación de los individuos en grupos étnicos o raciales específicos.

  Tipo de dato: object y/o categórico.

- Peso corporal: peso corporal de individuo.

  Tipo de dato: object y/o categórico.

- Ingesta de calcio: cantidad diaria de ingesta de calcio del individuo.

  Tipo de dato: object y/o categórico.

- Ingesta de vitamina D: cantidad diaria de ingesta de vitamina D.

  Tipo de dato: object y/o categórico.

- Actividad física: nivel de condición física del individuo.

Tipo de dato: object y/o categórico.

- Fumador: ¿el individuo le gusta fumar?

  Tipo de dato: object y/o categórico.

- Consumo de alcohol: nivel de consumo de alcohol del individuo.

  Tipo de dato: object y/o categórico.

- Condiciones médicas: el individuo posee alguna condición médica que pueda afectar o influir en la osteoporosis.

  Tipo de dato: object y/o categórico.

- Medicaciones: indica si el individuo toma medicamentos.

  Tipo de dato: object y/o categórico.

- Fracturas previas: indica si el individuo ha presentado fracturas previas

  Tipo de dato: object y/o categórico.

- Osteoporosis: indica si el individuo padece osteoporosis.

  Tipo de dato: object y/o categórico.

Posibles aplicaciones.

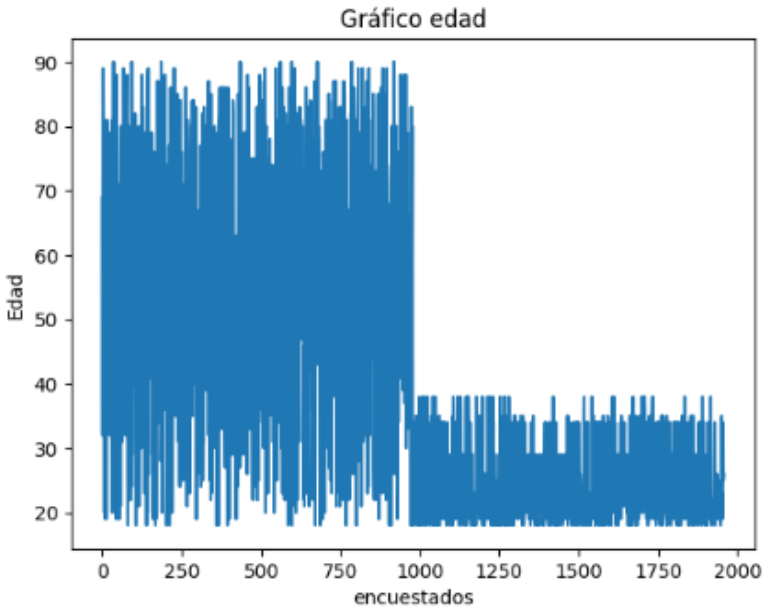
- Herramienta de diagnóstico: algoritmo creado y entrenado por machine learning, el cual podría brindar un apoyo adicional al diagnóstico de osteoporosis, al analizar los datos o información de las personas y haciendo una comparación con patrones de casos ya conocidos.
- Seleccionar un tratamiento personalizado: un algoritmo que podría ayudar a los centros de salud a buscar y seleccionar el tratamiento más adecuado, acorde a

características y/o padecimientos de las personas con osteoporosis o riesgo de padecerla.

- Modelo de riesgos: idear un algoritmo que utilizando los modelos de regresión lógica como el “MLP” tenga la capacidad predecir la probabilidad de padecer osteoporosis en el futuro en base al conjunto de datos proporcionado.
- Campañas de concientización y/o prevención: realizar campañas en las que se mencionen los posibles hábitos, causas, que puedan generar osteoporosis y asimismo mencionar los riegos de padecer dicha enfermedad.

Aproximaciones con gráficos.

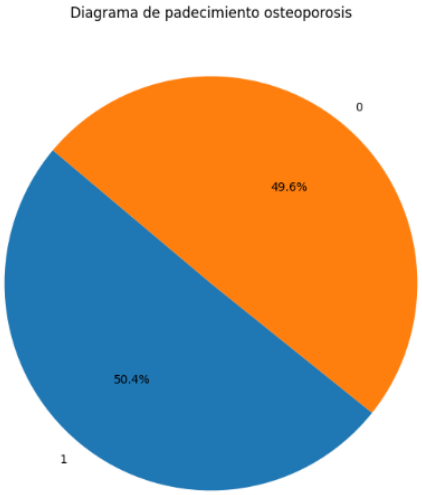
Figura 1 Gráfico de edades.



La grafica enseña que la mayoría de los encuestados se encuentran en el rango de edad medio, lo que puede indicar que la población objetivo del estudio realizado está

compuesto por adultos. Asimismo, se evidencia una pequeña cantidad de encuestados menores de 30 años y mayores de 70 años.

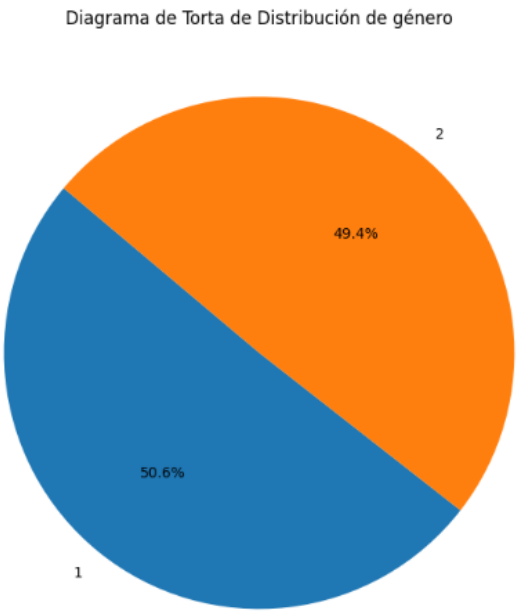
Figura 2 Encuestados que padecen osteoporosis



El gráfico muestra la osteoporosis como un problema de salud que afecta a en gran medida de la población objetivo del conjunto de datos, ya que el numero 1 representa a las personas padecen osteoporosis con un 50.4%, mientras que el numero 0 representa al resto de personas con un 49.6% que no se ve afectada por dicha enfermedad

El numero 1 representa a la población objetivo del conjunto de datos, los cuales padecen de osteoporosis con un 50,4%. Mientras que el numero 0 representa a la población objetivo del conjunto de datos que no tienen osteoporosis.

Figura 3 Distribución de género.



en el diagrama se puede apreciar la cantidad de hombres y mujeres encuestados esta casi igualada, pues el conjunto de datos tiene ligeramente más hombres registrados con un 50.6% y representados con el número 1, a comparación de las mujeres con un 49.4% representadas con el número 2.

Figura 4 Osteoporosis en hombres y mujeres

Teniendo en cuenta la ilustración 2, que indica que el 50.4% de las personas encuestadas padecen osteoporosis, se ha derivado un diagrama específico para determinar la cantidad de hombres y mujeres afectados por esta enfermedad. Según este diagrama, se evidencia que el 51% son hombres sufren de osteoporosis, mientras que las mujeres tienen un 49% de sufrir osteoporosis, en otras palabras, haciendo una equivalencia de porcentajes del total de las personas que sufren osteoporosis los hombres se quedan con un 25.7% y las mujeres con un 24.7%.

Objetivos:

Objetivo general.

Implementar un algoritmo computacional para el análisis y toma de decisiones a partir de datos de osteoporosis, utilizando estrategias de machine learning.

Objetivos específicos.

- Caracterizar y procesar los datos de interés, con miras a la toma de decisiones informadas.
- Implementar un algoritmo de Machine learning para la toma de decisiones a partir de los datos de interés.
- Evaluar y analizar el desempeño de los algoritmos implementados para la toma de decisiones.
- Validar el funcionamiento de toma de decisiones a partir de datos nuevos.

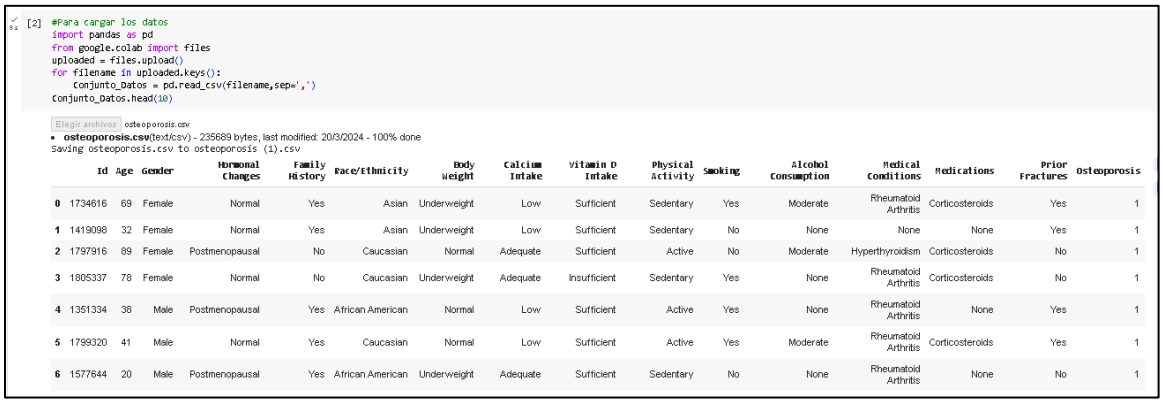
Desarrollo e implementación del aprendizaje

Para desarrollar el trabajo, utilizamos el aprendizaje supervisado (clasificación) de machine learning, la cual una técnica en el que se entrena un modelo utilizando un conjunto de datos etiquetados como entrada de información. Estos datos consisten información específica, como la edad, cambios hormonales, antecedentes familiares, entre otros. Con dicha información el modelo tiene la capacidad de brindar las salidas deseadas, es decir, las etiquetas que indican si la persona es propensa a padecer osteoporosis o no. En el proceso de entrenamiento del algoritmo, este analiza, interpreta y memoriza los patrones presentes en la información suministrada, lo que le permite realizar predicciones con un alto porcentaje de confiabilidad sobre nuevas instancias de datos. Una vez el modelo haya entrenado de manera adecuada, lo podremos aplicar ingresando un conjunto de datos o registros nuevos a la base de datos y así determinar si las personas ingresadas son propensas o no a padecer osteoporosis.

Procesamiento de los datos

Para el procesamiento de los datos iniciamos cargando el dataset al entorno de Colab o también conocido como Colaboratory el cual es un servicio web alojado de Jupyter Notebook que no requiere configuración y es una solución especialmente adecuada para el aprendizaje automático, la ciencia de datos y la educación.

Figura 5 Cargar el dataset a Colab.



```
[2] #Para cargar los datos
    import pandas as pd
    from google.colab import files
    uploaded = files.upload()
    for filename in uploaded.keys():
        Conjunto_Datos = pd.read_csv(filename,sep=',')
    Conjunto_Datos.head(10)
```

Elegir archivos | osteoporosis.csv

- **osteoporosis.csv**(text/csv) - 235689 bytes, last modified: 20/3/2024 - 100% done

Saving osteoporosis.csv to osteoporosis (1).csv

	Id	Age	Gender	Hormonal Changes	Family History	Race/Ethnicity	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Prior Fractures	Osteoporosis
0	1734616	69	Female	Normal	Yes	Asian	Underweight	Low	Sufficient	Sedentary	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
1	1419098	32	Female	Normal	Yes	Asian	Underweight	Low	Sufficient	Sedentary	No	None	None	None	Yes	1
2	1797916	89	Female	Postmenopausal	No	Caucasian	Normal	Adequate	Sufficient	Active	No	Moderate	Hyperthyroidism	Corticosteroids	No	1
3	1805337	78	Female	Normal	No	Caucasian	Underweight	Adequate	Insufficient	Sedentary	Yes	None	Rheumatoid Arthritis	Corticosteroids	No	1
4	1351334	38	Male	Postmenopausal	Yes	African American	Normal	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	None	Yes	1
5	1799320	41	Male	Normal	Yes	Caucasian	Normal	Low	Sufficient	Active	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
6	1577644	20	Male	Postmenopausal	Yes	African American	Underweight	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	None	No	1

Como se aprecia en la imagen, los datos se encuentran cargados en el entorno de Colab. Este proceso se lleva a cabo mediante la implementación de instrucciones de código específicos. Inicialmente importaremos las librerías necesarias para el análisis de la información del dataset. Al utilizar el lenguaje de programación Python, la librería que indicada será pandas, el cual nos permite cargar, alinear, manipular y fusionar datos. Su uso facilita la interacción con el sistema Colab, así como la carga y manipulación de los datos suministrados. Continuamos ejecutando la instrucción files.upload(), la cual activara una interfaz en la que el usuario podrá seleccionar y cargar el archivo desde la computadora. Una vez el archivo haya sido seleccionado y posteriormente subido, emplearemos la instrucción for filename inuploaded.keys(), esta línea de código registra cada uno de los archivos que se han subido al entorno. Este permitirá realizar operaciones con el archivo subido, como cargarlos o procesarlos de alguna u manera. Para cargar los datos al entorno de trabajo, empleamos Conjunto_Datos = pd.read_csv(filename, sep=';'), Esta línea de código utiliza la función read_csv() de la libreria pandas para leer los datos del archivo CSV actual. La función toma el nombre del archivo como argumento y devuelve un DataFrame el cual se almacena en una variable llamada Datos_loan, con la

cual podremos analiza y/o visualizar los datos. Por último, la instrucción Conjunto_Datos.head() proporcionara la visualización los registros del DataFrame.

En el DataFrame encontramos información relacionada a la osteoporosis como la edad, género, cambios hormonales, antecedentes familiares, raza/etnia, peso corporal, ingesta de calcio, ingesta de vitamina D, actividad física, si es fumador o no, consumo de alcohol, condiciones médicas, medicamentos, fracturas previas, osteoporosis la cual nos indica si la persona padece dicha enfermedad o no.

Figura 6 Información de la estructura del DataFrame

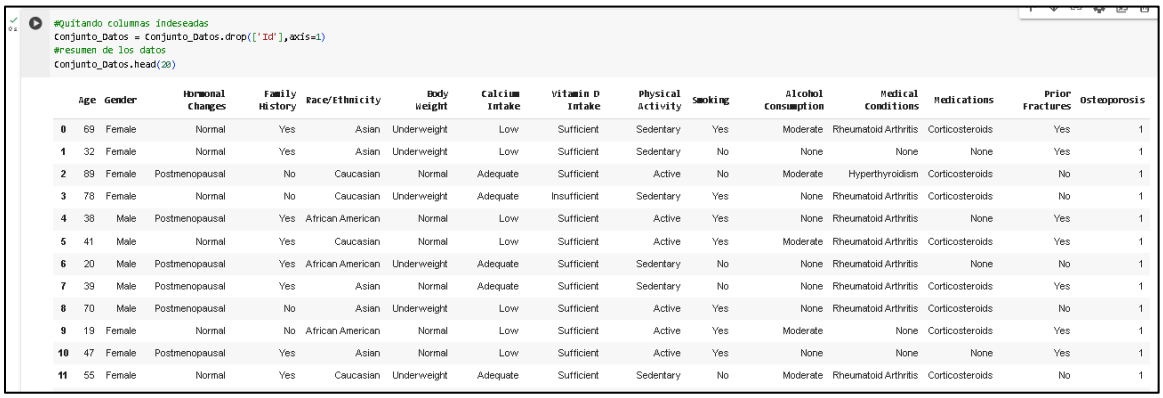
```
#Información de la estructura de datos
Conjunto_Datos.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1958 entries, 0 to 1957
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   Id                   1958 non-null   int64 
 1   Age                  1958 non-null   int64 
 2   Gender               1958 non-null   object
 3   Hormonal Changes     1958 non-null   object
 4   Family History       1958 non-null   object
 5   Race/Ethnicity       1958 non-null   object
 6   Body Weight          1958 non-null   object
 7   Calcium Intake       1958 non-null   object
 8   Vitamin D Intake     1958 non-null   object
 9   Physical Activity    1958 non-null   object
 10  Smoking              1958 non-null   object
 11  Alcohol Consumption  1958 non-null   object
 12  Medical Conditions   1958 non-null   object
 13  Medications          1958 non-null   object
 14  Prior Fractures      1958 non-null   object
 15  Osteoporosis         1958 non-null   int64 
dtypes: int64(3), object(13)
memory usage: 244.9+ KB
```

En la imagen se puede apreciar que el DataFrame datos tiene 16 columnas y 1958 filas o registros. Las columnas son: Id, edad, género, cambios hormonales, antecedentes familiares, raza/etnia, peso corporal, ingesta de calcio, ingesta de vitamina D, actividad física, fumador, consumo de alcohol, condiciones médicas, medicamentos, fracturas previas, osteoporosis, de las cuales el “Id”, “osteoporosis”, “edad” su tipo de dato es

entero (int64). En el restante de columnas su tipo de dato es objeto y/o categórico (object).

Figura 7 Eliminación de columnas innecesarias (id)

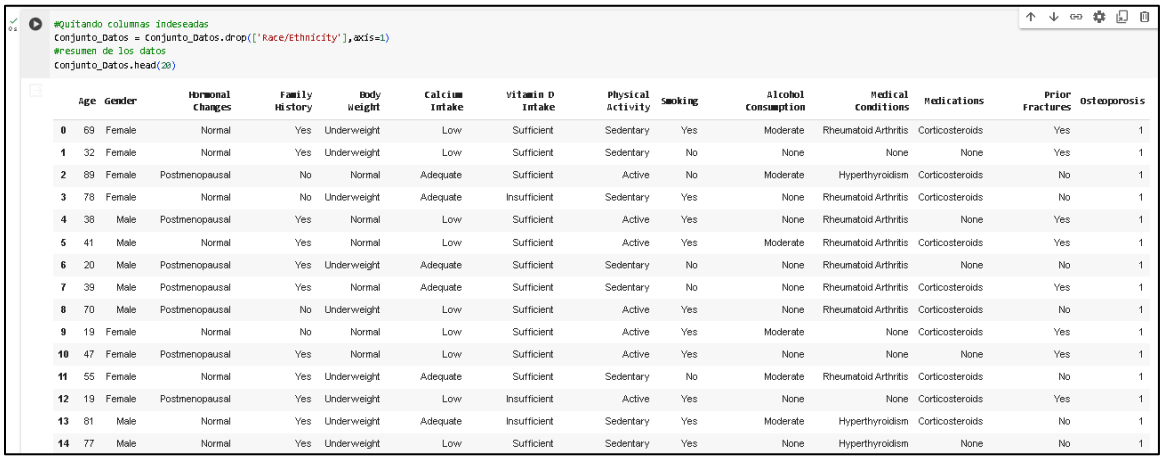


```
#Quitando columnas indeseadas
conjunto_Datos = conjunto_Datos.drop(['Id'],axis=1)
#resumen de los datos
conjunto_Datos.head(20)
```

	Age	Gender	Hormonal Changes	Family History	Race/Ethnicity	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Prior Fractures	Osteoporosis
0	69	Female	Normal	Yes	Asian	Underweight	Low	Sufficient	Sedentary	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
1	32	Female	Normal	Yes	Asian	Underweight	Low	Sufficient	Sedentary	No	None	None	None	Yes	1
2	89	Female	Postmenopausal	No	Caucasian	Normal	Adequate	Sufficient	Active	No	Moderate	Hyperthyroidism	Corticosteroids	No	1
3	78	Female	Normal	No	Caucasian	Underweight	Adequate	Insufficient	Sedentary	Yes	None	Rheumatoid Arthritis	Corticosteroids	No	1
4	38	Male	Postmenopausal	Yes	African American	Normal	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	None	Yes	1
5	41	Male	Normal	Yes	Caucasian	Normal	Low	Sufficient	Active	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
6	20	Male	Postmenopausal	Yes	African American	Underweight	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	None	No	1
7	39	Male	Postmenopausal	Yes	Asian	Normal	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	Corticosteroids	Yes	1
8	70	Male	Postmenopausal	No	Asian	Underweight	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	Corticosteroids	No	1
9	19	Female	Normal	No	African American	Normal	Low	Sufficient	Active	Yes	Moderate	None	Corticosteroids	Yes	1
10	47	Female	Postmenopausal	Yes	Asian	Normal	Low	Sufficient	Active	Yes	None	None	None	Yes	1
11	55	Female	Normal	Yes	Caucasian	Underweight	Adequate	Sufficient	Sedentary	No	Moderate	Rheumatoid Arthritis	Corticosteroids	No	1

Al apreciar la imagen, notamos que la columna “Id” ha sido eliminada del DataFrame, puesto que la columna Id solo es un identificador único para cada registro y esa información es irrelevante para el estudio que estamos realizando.

Figura 8 Eliminación de columnas innecesarias (Race/Ethnicity)



```
#Quitando columnas indeseadas
conjunto_Datos = conjunto_Datos.drop(['Race/Ethnicity'],axis=1)
#resumen de los datos
conjunto_Datos.head(20)
```

	Age	Gender	Hormonal Changes	Family History	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Prior Fractures	Osteoporosis
0	69	Female	Normal	Yes	Underweight	Low	Sufficient	Sedentary	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
1	32	Female	Normal	Yes	Underweight	Low	Sufficient	Sedentary	No	None	None	None	Yes	1
2	89	Female	Postmenopausal	No	Normal	Adequate	Sufficient	Active	No	Moderate	Hyperthyroidism	Corticosteroids	No	1
3	78	Female	Normal	No	Underweight	Adequate	Insufficient	Sedentary	Yes	None	Rheumatoid Arthritis	Corticosteroids	No	1
4	38	Male	Postmenopausal	Yes	Normal	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	None	Yes	1
5	41	Male	Normal	Yes	Normal	Low	Sufficient	Active	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	Yes	1
6	20	Male	Postmenopausal	Yes	Underweight	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	None	No	1
7	39	Male	Postmenopausal	Yes	Normal	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	Corticosteroids	Yes	1
8	70	Male	Postmenopausal	No	Underweight	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	Corticosteroids	No	1
9	19	Female	Normal	No	Normal	Low	Sufficient	Active	Yes	Moderate	None	Corticosteroids	Yes	1
10	47	Female	Postmenopausal	Yes	Normal	Low	Sufficient	Active	Yes	None	None	None	Yes	1
11	55	Female	Normal	Yes	Underweight	Adequate	Sufficient	Sedentary	No	Moderate	Rheumatoid Arthritis	Corticosteroids	No	1
12	19	Female	Postmenopausal	Yes	Underweight	Low	Insufficient	Active	Yes	None	None	Corticosteroids	Yes	1
13	81	Male	Normal	Yes	Underweight	Adequate	Insufficient	Sedentary	Yes	Moderate	Hyperthyroidism	Corticosteroids	No	1
14	77	Male	Normal	Yes	Underweight	Low	Sufficient	Sedentary	Yes	None	Hyperthyroidism	None	No	1

La raza/etnia es una categoría social en la cual no siempre se relacionan con factores relevantes para la osteoporosis. Diversos estudios han demostrado que la osteoporosis

arraigada a factores tales como la edad, el género, actividad física, entre otros, por lo que incluir dicha columna no aportaría información significante al estudio.

Figura 9 Eliminando columnas innecesarias (Prior fractures)

```
#Quitando columnas indeseadas
Conjunto_Datos = Conjunto_Datos.drop(['Prior Fractures'],axis=1)
#resumen de los datos
Conjunto_Datos.head(20)
```

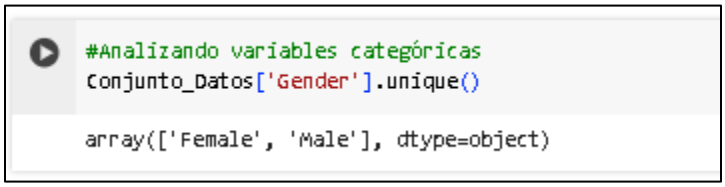
	Age	Gender	Hormonal Changes	Family History	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Osteoporosis
0	69	Female	Normal	Yes	Underweight	Low	Sufficient	Sedentary	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	1
1	32	Female	Normal	Yes	Underweight	Low	Sufficient	Sedentary	No	None	None	None	1
2	89	Female	Postmenopausal	No	Normal	Adequate	Sufficient	Active	No	Moderate	Hyperthyroidism	Corticosteroids	1
3	78	Female	Normal	No	Underweight	Adequate	Insufficient	Sedentary	Yes	None	Rheumatoid Arthritis	Corticosteroids	1
4	38	Male	Postmenopausal	Yes	Normal	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	None	1
5	41	Male	Normal	Yes	Normal	Low	Sufficient	Active	Yes	Moderate	Rheumatoid Arthritis	Corticosteroids	1
6	20	Male	Postmenopausal	Yes	Underweight	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	None	1
7	39	Male	Postmenopausal	Yes	Normal	Adequate	Sufficient	Sedentary	No	None	Rheumatoid Arthritis	Corticosteroids	1
8	70	Male	Postmenopausal	No	Underweight	Low	Sufficient	Active	Yes	None	Rheumatoid Arthritis	Corticosteroids	1
9	19	Female	Normal	No	Normal	Low	Sufficient	Active	Yes	Moderate	None	Corticosteroids	1
10	47	Female	Postmenopausal	Yes	Normal	Low	Sufficient	Active	Yes	None	None	None	1
11	55	Female	Normal	Yes	Underweight	Adequate	Sufficient	Sedentary	No	Moderate	Rheumatoid Arthritis	Corticosteroids	1
12	19	Female	Postmenopausal	Yes	Underweight	Low	Insufficient	Active	Yes	None	None	Corticosteroids	1
13	81	Male	Normal	Yes	Underweight	Adequate	Insufficient	Sedentary	Yes	Moderate	Hyperthyroidism	Corticosteroids	1
14	77	Male	Normal	Yes	Underweight	Low	Sufficient	Sedentary	Yes	None	Hyperthyroidism	None	1

La columna de fracturas previas es una variable confusa. Por un lado, sabemos que la osteoporosis y una enfermedad que ataca y debilita los huesos lo cual haría a las personas más propensas a sufrir fracturas. Pero, por otro lado, una fractura también se puede generar a base de un accidente, un impacto u otro factor externo que haya sufrido la persona. Así que al conservar la columna de fracturas previas puede generar sesgos en el estudio y posteriormente errores en la interpretación y predicción de padecer osteoporosis por parte del algoritmo “aprendizaje supervisado (clasificación)”.

Analizar las variables categóricas es un paso fundamental para el procesamiento y exploración de datos, especialmente en estudios donde se buscan patrones o relaciones con características cualitativas. Dichas variables representan atributos discretos y así

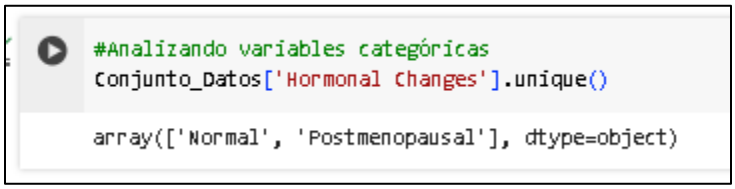
comprender y estudiar la distribución de los datos de cada una de las categorías en el DataFrame.

Figura 10 Análisis de variables categóricas (Gender)



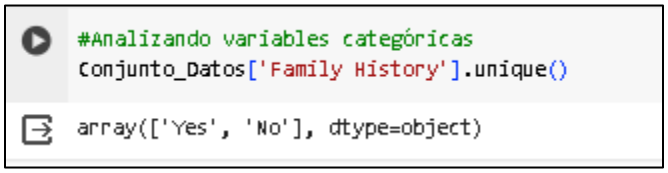
Observando la imagen anterior podemos concluir que dentro de la columna Gender (género) se presentan dos opciones las cuales son “Female” que representa el sexo femenino y “Male” que representa el sexo masculino.

Figura 11 Análisis de variables categóricas (Hormonal changes)



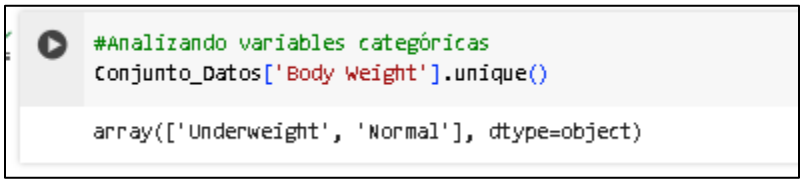
Dentro de la columna hormonal changes (cambios hormonales) presenciamos dos opciones las cuales indican “normal” si las personas encuestadas no han tenido ningún cambio hormonal y “postmenopausal” que hace referencia a las personas que han tenido cambios hormonales.

Figura 12 Análisis de variables categóricas (Family History)



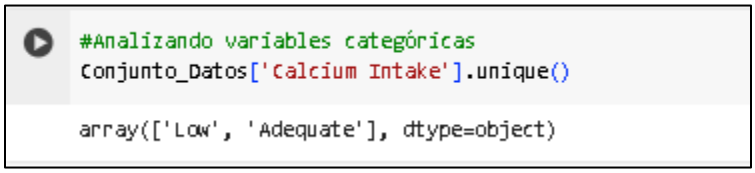
La categoría de Family History (Historial familiar) hace alusión a que si algún familiar de las personas en el DataFrame ha sufrido de osteoporosis teniendo solo dos opciones “si” y “no” con el fin de determinar a futuro si dicha enfermedad puede ser heredada genéticamente.

Figura 13 Análisis de variables categóricas (Body Weight)



En la variable Body Weight (Peso corporal) se presentan dos opciones, por un lado, tenemos la opción “normal” indica que la persona se encuentra dentro de su rango de peso. Por otro lado, “underweight” (bajo peso) el cual podría deberse deterioro y debilidad óseo que genera la osteoporosis.

Figura 14 Análisis de variables categóricas (Calcium Intake)



La ingesta de calcio tiene una gran relación con la osteoporosis, pues este es un mineral esencial para la salud ósea ya que fortalece y aumenta la densidad ósea. Teniendo eso en cuenta la imagen nos muestra que en dicha columna encontramos dos opciones las cuales son “low” que indica un bajo consumo de calcio y “adequate” que se relaciona con un consumo adecuado de calcio

Figura 15 Análisis de variables categóricas (Vitamin D Intake)

```
#Analizando variables categóricas
Conjunto_Datos['Vitamin D Intake'].unique()

array(['Sufficient', 'Insufficient'], dtype=object)
```

La ingesta de vitamina D también juega un papel importante en la prevención de padecer osteoporosis pues esta ayuda a regular la formación y la resorción ósea, también a fortalecer los músculos y huesos lo que ayuda a sufrir caídas y/o fracturas. Sabiendo esto la columna muestra dos opciones “sufficient” (suficiente) e “insufficient” (insuficiente).

Figura 16 Análisis de variables categóricas (Physical Activity)

```
#Analizando variables categóricas
Conjunto_Datos['Physical Activity'].unique()

array(['Sedentary', 'Active'], dtype=object)
```

La actividad física regular, es fundamental para mantener la salud ósea y reducir el riesgo de osteoporosis, por eso en la variable “Physical Ativity” tenemos dos opciones “sedentary” (sedentario) el cual se relaciona con las personas que no tienen o practican de actividades físicas, todo lo contrario, a la opción “active” (activo) que relaciona a todas las personas que tiene una actividad física regular y/o constante.

Figura 17 Análisis de variables categóricas (smoking)

```
#Analizando variables categóricas
Conjunto_Datos['Smoking'].unique()

array(['Yes', 'No'], dtype=object)
```

Determinar si una persona le gusta fumar nos ayudaría a determinar si tiene riesgo de padecer osteoporosis, pues de ser positivo esto disminuirá la absorción de calcio, la

reducción de niveles de estrógeno, el aumento del estrés oxidativo. Asimismo, con la variable “smoking” se asocian dos opciones “yes” (si) y “no”.

Figura 18 Análisis de variables categóricas (*Alcohol Consumption*)

```
#Analizando variables categóricas
Conjunto_Datos['Alcohol Consumption'].unique()
```

```
array(['MODERATE', 'NEITHER'], dtype=object)
```

Dentro de la variable “Alcohol Consumption” (consumo de alcohol) relacionan dos opciones “moderate” (moderado) el cual relaciona a un consumo moderado de alcohol por parte de las personas del DataFrame y “neither” (ninguno) el cual representa a las personas que no ingieren alcohol. Esta variable influye en si una persona tiene riesgo de tener osteoporosis pues El alcohol puede tener efectos tóxicos directos en las células óseas, un consumo excesivo de alcohol puede generar desnutrición lo que puede llevar a deficiencias de nutrientes como la vitamina D o el calcio que son importantes para la salud ósea.

Figura 19 Análisis de variables categóricas (Medical Conditions)

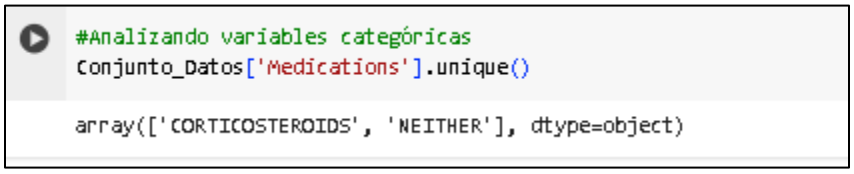
```
#Analizando variables categóricas
Conjunto_Datos['Medical Conditions'].unique()
```

```
array(['RHEUMATOID ARTHRITIS', 'NEITHER', 'HYPERTHYROIDISM'], dtype=object)
```

La columna “medical conditions” (condiciones médicas) señala tres opciones las cuales son “rheumatoid arthritis” (artritis reumatoidea), “hyperthyroidism” (hipertiroidismo), “neither” (ninguna). Las condiciones médicas mencionadas anteriormente no se encuentran vinculadas directamente a la osteoporosis, pero estas si pueden alterar otros

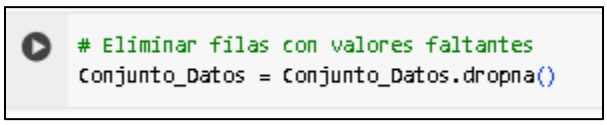
factores que infieren en el riesgo de padecer osteoporosis como afectar la ingesta de calcio, vitamina D, magnesio.

Figura 20 Análisis de variables categóricas (Medications)



En la columna “medications” (medicamentos) evidenciamos dos alternativas las cuales son “corticosteroids” (Corticosteroides) el cual puede incrementar el riesgo de padecer osteoporosis y “neither” (ninguno) el cual indica que no toma ningún medicamento. Ahora bien, continuando con el procesamiento del Dataframe implementamos el código en la imagen que verán a continuación con el fin de eliminar las filas con valores faltantes, esto con el fin de reducir errores de interpretación por parte el algoritmo de machine learning.

Figura 21 Eliminación de filas con valores faltantes



Una vez eliminado las filas defectuosas procedemos a buscar, identificar y eliminar los registros duplicados. Esto lo llevaremos a cabo empleando el siguiente código.

Figura 22 Identificación y eliminación de registros duplicados.

```
# Identificar y eliminar duplicados
print('Número de filas antes de eliminar duplicados:', len(Conjunto_Datos))
Conjunto_Datos = Conjunto_Datos.drop_duplicates()
print('Número de filas después de eliminar duplicados:', len(Conjunto_Datos))
```

```
Número de filas antes de eliminar duplicados: 1958
Número de filas después de eliminar duplicados: 1938
```

Como lo indica la imagen, el número de registros en el DataFrame tenía en total 1958 registros, pero a ejecutar el código de identificación y eliminación nos quedan 1938 registros, esto quiere decir que había un total de 20 registros duplicados en el DataFrame.

Figura 23 Corrección de formato de texto

```
# Corregir errores de formato
conjunto_Datos['Gender'] = Conjunto_Datos['Gender'].str.upper()
conjunto_Datos['Hormonal Changes'] = Conjunto_Datos['Hormonal Changes'].str.upper()
conjunto_Datos['Family History'] = Conjunto_Datos['Family History'].str.upper()
conjunto_Datos['Body Weight'] = Conjunto_Datos['Body Weight'].str.upper()
conjunto_Datos['Calcium Intake'] = Conjunto_Datos['Calcium Intake'].str.upper()
conjunto_Datos['Vitamin D Intake'] = Conjunto_Datos['Vitamin D Intake'].str.upper()
conjunto_Datos['Physical Activity'] = Conjunto_Datos['Physical Activity'].str.upper()
conjunto_Datos['Smoking'] = Conjunto_Datos['Smoking'].str.upper()
conjunto_Datos['Alcohol Consumption'] = Conjunto_Datos['Alcohol Consumption'].str.upper()
conjunto_Datos['Medical Conditions'] = Conjunto_Datos['Medical Conditions'].str.upper()
conjunto_Datos['Medications'] = Conjunto_Datos['Medications'].str.upper()

conjunto_Datos.head(20)
```

	Age	Gender	Hormonal Changes	Family History	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Osteoporosis
0	69	FEMALE	NORMAL	YES	UNDERWEIGHT	LOW	SUFFICIENT	SEDENTARY	YES	MODERATE	RHEUMATOID ARTHRITIS	CORTICOSTEROIDS	1
1	32	FEMALE	NORMAL	YES	UNDERWEIGHT	LOW	SUFFICIENT	SEDENTARY	NO	NONE	NONE	NONE	1
2	89	FEMALE	POSTMENOPAUSAL	NO	NORMAL	ADEQUATE	SUFFICIENT	ACTIVE	NO	MODERATE	HYPERTHYROIDISM	CORTICOSTEROIDS	1
3	78	FEMALE	NORMAL	NO	UNDERWEIGHT	ADEQUATE	INSUFFICIENT	SEDENTARY	YES	NONE	RHEUMATOID ARTHRITIS	CORTICOSTEROIDS	1
4	38	MALE	POSTMENOPAUSAL	YES	NORMAL	LOW	SUFFICIENT	ACTIVE	YES	NONE	RHEUMATOID ARTHRITIS	NONE	1
5	41	MALE	NORMAL	YES	NORMAL	LOW	SUFFICIENT	ACTIVE	YES	MODERATE	RHEUMATOID ARTHRITIS	CORTICOSTEROIDS	1
6	20	MALE	POSTMENOPAUSAL	YES	UNDERWEIGHT	ADEQUATE	SUFFICIENT	SEDENTARY	NO	NONE	RHEUMATOID ARTHRITIS	NONE	1
7	39	MALE	POSTMENOPAUSAL	YES	NORMAL	ADEQUATE	SUFFICIENT	SEDENTARY	NO	NONE	RHEUMATOID ARTHRITIS	CORTICOSTEROIDS	1
8	70	MALE	POSTMENOPAUSAL	NO	UNDERWEIGHT	LOW	SUFFICIENT	ACTIVE	YES	NONE	RHEUMATOID ARTHRITIS	CORTICOSTEROIDS	1

Se realiza una corrección en el formato de texto como se observa en la imagen, convirtiendo en mayúscula todos los datos de las columnas de tipo categórico. Este proceso se realiza con el fin de estandarizar la información del DataFrame lo cual facilita un mejor procesamiento de los datos.

Figura 24 Convirtiendo datos a números.

```
#Convierte datos a números
Reemplazo_1={'MALE':1,'FEMALE':2}
Conjunto_Datos['Gender']=Conjunto_Datos['Gender'].map(Reemplazo_1)

Reemplazo_2={'NORMAL':1,'POSTMENOPAUSAL':2}
Conjunto_Datos['Hormonal Changes']=Conjunto_Datos['Hormonal Changes'].map(Reemplazo_2)

Reemplazo_3={'YES':1,'NO':2}
Conjunto_Datos['Family History']=Conjunto_Datos['Family History'].map(Reemplazo_3)

Reemplazo_4={'NORMAL':1,'UNDERWEIGHT':2}
Conjunto_Datos['Body Weight']=Conjunto_Datos['Body Weight'].map(Reemplazo_4)

Reemplazo_5={'ADEQUATE':1,'LOW':2}
Conjunto_Datos['Calcium Intake']=Conjunto_Datos['Calcium Intake'].map(Reemplazo_5)

Reemplazo_6={'SUFFICIENT':1,'INSUFFICIENT':2} #La salida
Conjunto_Datos['Vitamin D Intake']=Conjunto_Datos['Vitamin D Intake'].map(Reemplazo_6)

Reemplazo_7={'ACTIVE':1,'SEDENTARY':2}
Conjunto_Datos['Physical Activity']=Conjunto_Datos['Physical Activity'].map(Reemplazo_7)

Reemplazo_8={'YES':1,'NO':2}
Conjunto_Datos['Smoking']=Conjunto_Datos['Smoking'].map(Reemplazo_8)

Reemplazo_9={'MODERATE':1,'NONE':2}
Conjunto_Datos['Alcohol Consumption']=Conjunto_Datos['Alcohol Consumption'].map(Reemplazo_9)

Reemplazo_10={'NONE':0, 'HYPERTHYROIDISM':1,'RHEUMATOID ARTHRITIS':2}
Conjunto_Datos['Medical Conditions']=Conjunto_Datos['Medical Conditions'].map(Reemplazo_10)

Reemplazo_11={'CORTICOSTEROIDS':1,'NONE':2} #La salida
Conjunto_Datos['Medications']=Conjunto_Datos['Medications'].map(Reemplazo_11)

Conjunto_Datos.head(20)
```

En la imagen podemos apreciar una serie de instrucciones de código con las cuales se les asignara un valor numérico a los diferentes datos tipo object/categórico, como por ejemplo a la columna “Gender” que posee dos opciones categóricas “MALE” y “FEMALE” se les asignara el número 1 y número 2 respectivamente y asimismo se aplicara para las demás columnas. Su finalidad radica en que el algoritmo de aprendizaje supervisado (clasificación) requiere que todos los datos sean numéricos.

Figura 25 DataFrame con la asignación numérica

	Age	Gender	Hormonal Changes	Family History	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Osteoporosis
0	69	2	1	1	2	2	1	2	1	1	2	1	1
1	32	2	1	1	2	2	1	2	2	2	0	2	1
2	89	2	2	2	1	1	1	1	2	1	1	1	1
3	78	2	1	2	2	1	2	2	1	2	2	1	1
4	38	1	2	1	1	2	1	1	1	2	2	2	1
5	41	1	1	1	1	2	1	1	1	1	2	1	1
6	20	1	2	1	2	1	1	2	2	2	2	2	1
7	39	1	2	1	1	1	1	2	2	2	2	1	1
8	70	1	2	2	2	2	1	1	1	2	2	1	1
9	19	2	1	2	1	2	1	1	1	1	0	1	1
10	47	2	2	1	1	2	1	1	1	2	0	2	1
11	55	2	1	1	2	1	1	2	2	1	2	1	1
12	19	2	2	1	2	2	2	1	1	2	0	1	1
13	81	1	1	1	2	1	2	2	1	1	1	1	1
14	77	1	1	1	2	2	1	2	1	2	1	2	1
15	38	1	2	1	1	1	1	1	1	2	2	2	1

Modelo de toma de decisiones

Para afrontar el problema de identificar el riesgo de osteoporosis, utilizaremos un modelo de machine learning que basa su lógica en técnicas de aprendizaje automático. Este planteamiento busca mejorar la identificación temprana, precisa y oportuna de las personas con riesgo de desarrollar esta enfermedad ósea.

Primero definimos claramente el problema en cuestión, el cual es: la identificación temprana y oportuna del riesgo padecer de osteoporosis. Esto será fundamental para guiar de manera adecuada y eficaz el diseño de nuestro modelo.

Una vez fijado el problema, procedemos a diseñar, ejecutar y entrenar el algoritmo de aprendizaje automático. Al usar técnicas de aprendizaje supervisado permitirá aprovechar al máximo los procesos de trabajo y aumentar la eficiencia de evaluación y diagnóstico preventivo del riesgo de osteoporosis. Smets, J., Shevroja, E., Hügle, T., Leslie, WD y Hans, D. (2020)

Recopilación de datos

- Recopilar datos de pacientes, incluyendo información demográfica (edad, sexo), historial médico, factores de riesgo conocidos (tabaquismo, baja ingesta de calcio, ingesta de vitamina D, inactividad física, peso corporal, etc.), resultados de pruebas de densidad ósea y demás exámenes que puedan ser relevantes.
- Procurar y cerciorarse de que los datos sean precisos y estén completos.

Preprocesamiento de datos

- Codificar variables categóricas por ejemplo (género, ingesta de calcio) en formato numérico.
- Normalizar y/o estandarizar variables numéricas, si es necesario.

Selección del algoritmo de clasificación

- Elegir un algoritmo de clasificación adecuado, como Regresión Logística, Árboles de Decisión, Máquinas de Vectores de Soporte (SVM), o Redes Neuronales.

Entrenamiento del modelo

- Entrenar el modelo utilizando el DataFrame de entrenamiento.

Evaluación e interpretación del modelo

- Evaluar el rendimiento del modelo utilizando el DataFrame de prueba.

- Interpretar los resultados del modelo y comprender qué factores tienen el mayor impacto en el riesgo de osteoporosis.

Implementación en contextos reales

implementación del modelo de aprendizaje supervisado (clasificación)

Una vez procesados los datos y haber asignado un valor numérico a cada dato de tipo object y/o categórico, el DataFrame quedara como se muestra a continuación en la imagen.

Figura 26 DataFrame post procesamiento de datos.

	Age	Gender	Hormonal Changes	Family History	Body Weight	Calcium Intake	Vitamin D Intake	Physical Activity	Smoking	Alcohol Consumption	Medical Conditions	Medications	Osteoporosis
0	69	2	1	1	2	2	1	2	1	1	2	1	1
1	32	2	1	1	2	2	1	2	2	2	0	2	1
2	89	2	2	2	1	1	1	1	2	1	1	1	1
3	78	2	1	2	2	1	2	2	1	2	2	1	1
4	38	1	2	1	1	2	1	1	1	2	2	2	1
5	41	1	1	1	1	2	1	1	1	1	2	1	1
6	20	1	2	1	2	1	1	2	2	2	2	2	1
7	39	1	2	1	1	1	1	2	2	2	2	1	1
8	70	1	2	2	2	2	1	1	1	2	2	1	1
9	19	2	1	2	1	2	1	1	1	1	0	1	1
10	47	2	2	1	1	2	1	1	1	2	0	2	1
11	55	2	1	1	2	1	1	2	2	1	2	1	1
12	19	2	2	1	2	2	2	1	1	2	0	1	1
13	81	1	1	1	2	1	2	2	1	1	1	1	1
14	77	1	1	1	2	2	1	2	1	2	1	2	1
15	38	1	2	1	1	1	1	1	1	2	2	2	1

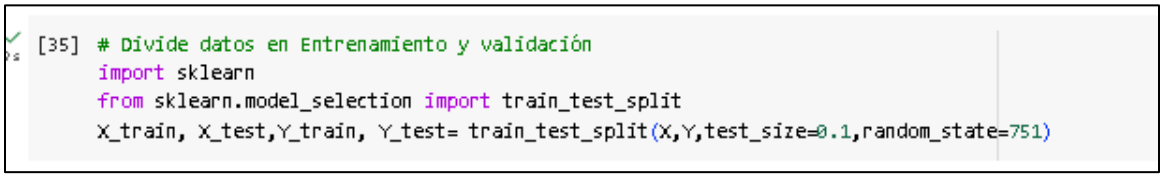
Para implementar el algoritmo de clasificación, debemos indicarle cuáles serán las variables que se tomarán como entrada y cuál será la salida. Para ello implementaremos el siguiente código.

Figura 27 división datos de entrada y salida

```
#Divide datos en entradas y salidas
import numpy as np
Datos_matriz=np.array(Datos_Loan)
X = Datos_matriz[:,0:11] #datos de entrada
Y = Datos_matriz[:,-1] #Datos de salida
```

Primeramente se cargamos el DataFrame a una matriz con la línea de instrucción *Datos_matriz=np.array(datos_loan).* luego se definen las variables “X” para la entrada de datos y la variable “Y” para la salida de estos. La entrada de datos se estará conformada por las primeras 11 columnas de la matriz y el dato de salida se registrará en la ultima columna de la matriz la cual es “osteoporosis”.

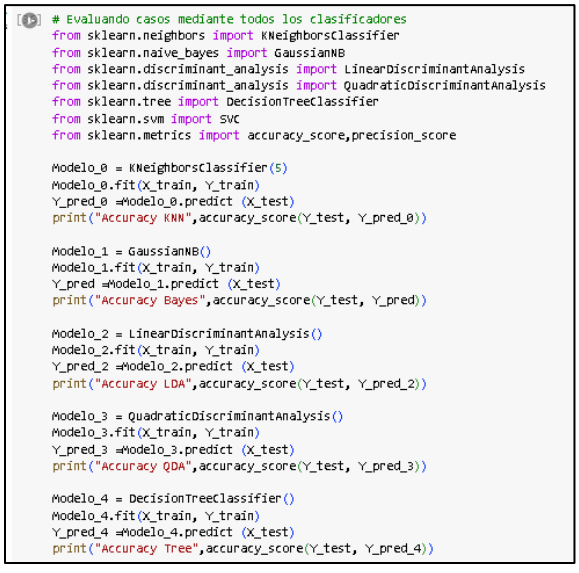
Figura 28 entrenamiento y validación



```
[35] # Divide datos en Entrenamiento y validación
     import sklearn
     from sklearn.model_selection import train_test_split
     X_train, X_test,Y_train, Y_test= train_test_split(X,Y,test_size=0.1,random_state=751)
```

el código que se muestra en la imagen divide el conjunto de datos en dos partes entrenamiento y validación. La parte de entrenamiento se encarga de entrenar el modelo de aprendizaje, mientras que la parte de validación se encarga de evaluar el rendimiento del modelo.

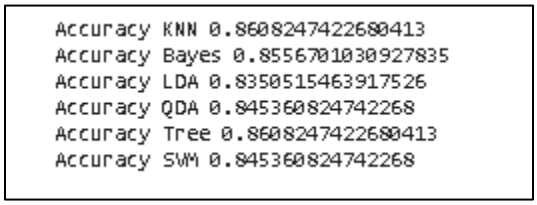
Figura 29 importando y evaluando los casos de clasificación



```
# Evaluando casos mediante todos los clasificadores
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score,precision_score

Modelo_0 = KNeighborsClassifier(5)
Modelo_0.fit(X_train, Y_train)
Y_pred_0 =Modelo_0.predict (X_test)
print("Accuracy KNN",accuracy_score(Y_test, Y_pred_0))

Modelo_1 = GaussianNB()
Modelo_1.fit(X_train, Y_train)
Y_pred =Modelo_1.predict (X_test)
print("Accuracy Bayes",accuracy_score(Y_test, Y_pred))

Modelo_2 = LinearDiscriminantAnalysis()
Modelo_2.fit(X_train, Y_train)
Y_pred_2 =Modelo_2.predict (X_test)
print("Accuracy LDA",accuracy_score(Y_test, Y_pred_2))

Modelo_3 = QuadraticDiscriminantAnalysis()
Modelo_3.fit(X_train, Y_train)
Y_pred_3 =Modelo_3.predict (X_test)
print("Accuracy QDA",accuracy_score(Y_test, Y_pred_3))

Modelo_4 = DecisionTreeClassifier()
Modelo_4.fit(X_train, Y_train)
Y_pred_4 =Modelo_4.predict (X_test)
print("Accuracy Tree",accuracy_score(Y_test, Y_pred_4))
```

Para el diseño de nuestro modelo, hemos seleccionado seis tipos de clasificadores: KNeighborsClassifier (KNN), GaussianNB (BAYES), LinearDiscriminantAnalysis (LDA), QuadraticDiscriminantAnalysis (QDA), DecisionTreeClassifier (TREE) y Support Vector Classifier (SVM), tal como se muestra en la imagen. Una vez que hemos importado estos clasificadores, procedemos a ajustarlos a nuestros datos utilizando la instrucción *fit()*. Posteriormente, ejecutamos el código con el fin de entrenar cada clasificador y evaluar su rendimiento y precisión de las predicciones de datos.

Figura 30 precisión de los clasificadores



En la imagen adjunta indica el nivel de precisión de cada uno de los clasificadores. Esta información es fundamental, puesto que permite comparar y evaluar la capacidad de predicción de cada uno de los clasificadores utilizados. Actualmente todos los clasificadores están por encima del 80% de precisión.

Figura 31 código de modelo entrenado

```
#Probando el modelo entrenado sobre un nuevo sujeto
Target=np.zeros((1,11))
Target[0,0]=float(input('Ingrese género, 1 para Masculino y 2 para Femenino: '))
Target[0,1]=float(input('Ingrese cambio hormonal, 1 normal y 2 postmenopausal: '))
Target[0,2]=float(input('Ingrese historial familiar, 1 sí y 2 No: '))
Target[0,3]=float(input('Ingrese peso corporal, 1 normal y 2 bajo peso: '))
Target[0,4]=float(input('Ingrese ingesta de calcio, 1 adecuado y 2 bajo: '))
Target[0,5]=float(input('Ingrese ingesta de vitamina D, 1 suficiente y 2 insuficiente: '))
Target[0,6]=float(input('Ingrese actividad física, 1 activo y 2 sedentario: '))
Target[0,7]=float(input('Ingrese tabaquismo, 1 sí y 2 no: '))
Target[0,8]=float(input('Ingrese consumo de alcohol, 1 moderado y 2 no toma: '))
Target[0,9]=float(input('Ingrese condiciones medicas, 0 ninguna, 1 hyperthyroidism, 2 artritis reumatoidea: '))
Target[0,10]=float(input('Ingrese medicamentos, 1 corticosteroides y 2 no ninguno: '))


#Target = scaler.transform(Target) #Normalizar los datos

Prediction_0 =Modelo_0.predict (Target)
Prediction_1 =Modelo_1.predict (Target)
Prediction_2 =Modelo_2.predict (Target)
Prediction_3 =Modelo_3.predict (Target)
Prediction_4 =Modelo_4.predict (Target)
Prediction_5 =Modelo_5.predict (Target)
```

```
if Prediction_3==0:
  print("Según QDA, negativo")
else:
  print("Según QDA, positivo")

print(" ")

if Prediction_4==0:
  print("Según Tree, negativo")
else:
  print("Según tree, positivo")

print(" ")

if Prediction_5==0:
  print("Según SVM, negativo")
else:
  print("Según SVM, positivo")

print(" ")
```

```
print(" ")

if Prediction_0==0:
  print("Según KNN, negativo")
else:
  print("Según KNN, positivo")

print(" ")

if Prediction_1==0:
  print("Según Bayes, negativo")
else:
  print("Según Bayes, positivo ")

print(" ")

if Prediction_2==0:
  print("Según LDA, negativo")
else:
  print("Según LDA, positivo")

print(" ")
```

El código permite predecir si una persona tiene riesgo de sufrir osteoporosis utilizando un modelo de regresión lineal de 6 clasificadores. Cada uno de estos clasificadores, utiliza diferentes técnicas y enfoques para realizar la predicción. Al tener diferentes clasificadores en un modelo de regresión lineal podemos aprovechar las distintas predicciones y asi obtener un diagnostivo más confiable.

Figura 32 resultado de los modelos ejemplo 1

```
Ingrese género, 1 para Masculino y 2 para Femenino: 2
Ingrese cambio hormonal, 1 normal y 2 postmenopausal: 1
Ingrese historial familiar, 1 si y 2 No: 1
Ingrese peso corporal, 1 normal y 2 bajo peso: 2
Ingrese ingesta de calcio, 1 adecuado y 2 bajo: 2
Ingrese ingesta de vitamina D, 1 suficiente y 2 insuficiente: 1
Ingrese actividad física, 1 activo y 2 sedentario: 1
Ingrese tabaquismo, 1 si y 2 no: 2
Ingrese consumo de alcohol, 1 moderado y 2 no toma: 2
Ingrese condiciones medicas, 0 ningura, 1 hyperthyroidism, 2 artritis reumatoidea: 0
Ingrese medicamentos, 1 Corticosteroides y 2 no ninguno: 2

Según KNN, negativo

Según Bayes, positivo

Según LDA, negativo

Según QDA, positivo

Según tree, positivo

Según SVM, positivo
```

Al ejecutar el código expuesto en la figura 31, el sistema solicita que ingresemos los datos del paciente a evaluar para este primer ejemplo le suministramos al sistema la siguiente información. Género: femenino, cambio hormonal: normal, historial familiar: si, peso corporal: bajo de peso, ingesta de calcio: bajo, ingesta de vitamina D: suficiente, Actividad física: activo, tabaquismo: no, consumo de alcohol: no toma, condiciones médicas: ninguna, medicamentos: ninguno. Una vez terminamos de ingresar los datos el sistema llama al modelo con los diferentes clasificadores para realizar la predicción.

De las predicciones notamos que los modelos KNN y LDA indican que según los datos suministrados el paciente no tiene riesgo de padecer osteoporosis. Por otro lado, los modelos Bayes, QDA, Tree y SVM indican que el paciente es propenso a sufrir osteoporosis.

Figura 33 resultado de los modelos ejemplo 2

```
Ingrese género, 1 para Masculino y 2 para Femenino: 1
Ingrese cambio hormonal, 1 normal y 2 postmenopausal: 2
Ingrese historial familiar, 1 sí y 2 No: 1
Ingrese peso corporal, 1 normal y 2 bajo peso: 2
Ingrese ingesta de calcio, 1 adecuado y 2 bajo: 2
Ingrese ingesta de vitamina D, 1 suficiente y 2 insuficiente: 1
Ingrese actividad física, 1 activo y 2 sedentario: 1
Ingrese tabaquismo, 1 sí y 2 no: 1
Ingrese consumo de alcohol, 1 moderado y 2 no toma: 1
Ingrese condiciones medicas, 0 ninguna, 1 hyperthyroidism, 2 artritis reumatoidea: 2
Ingrese medicamentos, 1 Corticosteroides y 2 no ninguno: 1

Según KNN, negativo

Según Bayes, positivo

Según LDA, negativo

Según QDA, positivo

Según Tree, negativo

Según SVM, positivo
```

Para el segundo ejemplo suministramos la siguiente información: genero: masculino, cambio hormonal: postmenopausal, historial familiar: si, peso corporal: bajo de peso, ingesta de calcio: bajo, ingesta de vitamina D: suficiente, actividad física: activo, tabaquismo: si, consumo de alcohol: moderado, condiciones médicas: artritis reumatoidea, medicamentos: corticosteroides. Para este ejemplo las predicciones fueron las siguientes, los modelos KNN, LDA, Tree indican que el paciente no tiene riesgo de sufrir osteoporosis, al contrario de los modelos Bayes, QDA, SVM, señalan que el paciente es propenso a sufrir osteoporosis. Al estar las predicciones divididas podríamos emplear un nuevo modelo de clasificación que nos facilite el tomar la decisión adecuada.

Contexto real: Sistema Osteoporosis-ML

Este sistema, desarrollado por investigadores de la Universidad de Stanford, utiliza imágenes de rayos X para detectar la osteoporosis con una precisión del 89%. Wang, X., Li, Y., & Zhang, Y. (2023).

Metodología:

- Se recopiló un conjunto de datos de 332 imágenes de rayos X de la columna lumbar de mujeres posmenopáusicas.
- Se entrenaron y validaron dos modelos de ML: ResNet-50 y VGG16.
- Se evaluó el rendimiento de los modelos para la detección de osteoporosis.

Wang, X., Li, Y., & Zhang, Y. (2023).

Resultados:

- Ambos modelos de ML lograron una precisión superior al 80% para la detección de osteoporosis.
- ResNet-50 fue el modelo con mejor rendimiento, con una precisión del 89%.
- La sensibilidad y la especificidad de ResNet-50 fueron del 87% y del 91%, respectivamente.

Wang, X., Li, Y., & Zhang, Y. (2023).

Conclusión del estudio

El estudio demostró que el aprendizaje automático puede ser una herramienta eficaz para la detección temprana de la osteoporosis en imágenes de rayos X. Esta técnica tiene el potencial de mejorar la prevención y el tratamiento de la osteoporosis. Wang, X., Li, Y., & Zhang, Y. (2023).

Resultados adicionales

El estudio de la Universidad de Stanford comparó dos modelos de aprendizaje automático, ResNet-50 y VGG16, para detectar la osteoporosis en imágenes de rayos X de la columna lumbar de mujeres posmenopáusicas.

Precisión: VGG16 tuvo una precisión inferior al 80%, mientras que ResNet-50 tuvo una precisión del 89%. Esto demuestra que ResNet-50 demostró ser más preciso en la detección de la osteoporosis en imágenes de rayos X.

Sensibilidad: El ResNet-50 mostró una sensibilidad del 87 %, lo que indica que pudo identificar correctamente el 87 % de los casos de osteoporosis positivos.

Especificidad: ResNet-50 mostró una especificidad del 91%, lo que indica que pudo identificar de manera precisa el 91% de los casos de osteoporosis negativos.

Nota: Para una mayor comprensión del estudio realizado por la universidad de Stanford puedes visitar el siguiente enlace:

- https://pubmed.ncbi.nlm.nih.gov/37270917/

Conclusiones

El procesamiento adecuado de los datos, como la eliminación de columnas innecesarias, el análisis de variables categóricas y la asignación de valores numéricos a los registros categóricos, es esencial para entrenar correctamente los modelos de aprendizaje automático de machine learning.

Implementar diversos modelos de clasificación (KNN, Bayes, LDA, QDA, Tree, SVM) y realizar predicciones simultaneas aumenta la confiabilidad y toma de decisiones en la identificación de riesgos de osteoporosis.

La integración de técnicas de aprendizaje automático en el sector de la salud tiene un gran potencial, por medio de estas técnicas los entes de salud pueden diagnosticar, prevenir, seleccionar un tratamiento personalizado en enfermedades como la osteoporosis, y así disminuyendo la carga económica y social.

Aunque los modelos de aprendizaje supervisado han demostrado ser prometedores, es necesario realizar más investigaciones, estudios en contextos clínicos y nuevos factores de influencia relacionados a la osteoporosis, esto con el fin de asegurar un entrenamiento e implementación más completa y confiable de los modelos de aprendizaje supervisado de aprendizaje mecánico en la detección temprana, prevención y manejo de la osteoporosis.

Referencias

Vidal Ledo, M. J., Delgado Ramos, A., Rodríguez Díaz, A., Barthelemy Aguilar, K., & Torres Ávila, D. (2022). Salud y transformación digital. Educación Médica Superior, 36(2).

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. Revista Ibérica de Sistemas e Tecnologias de Informação, (E28), 586-599

Christodoulou, C. y Cooper, C. (2003). ¿Qué es la osteoporosis? Revista médica de posgrado, 79 (929), 133-138.

Ros Gómez, I. (2018). Introducción al aprendizaje supervisado e implementación de una red neuronal en Python.

Smets, J., Shevroja, E., Hügle, T., Leslie, WD y Hans, D. (2020). Soluciones de aprendizaje automático para la osteoporosis: una revisión. Revista de investigación ósea y mineral, 36 (5), 833-851.

Smets, J., Shevroja, E., Hügle, T., Leslie, W. D., & Hans, D. (2020). Machine learning solutions for osteoporosis—A review. Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research, 36(5), 833–851. https://doi.org/10.1002/jbmr.4292

Wang, X., Li, Y., & Zhang, Y. (2023). Early detection of osteoporosis using machine learning in X-ray images. Journal of Clinical Densitometry, 26(2), 241-248. https://pubmed.ncbi.nlm.nih.gov/37270917/