UNIVERSIDAD NACIONAL TECNOLÓGICA DE LIMA SUR
FACULTAD DE INGENIERÍA Y GESTIÓN
ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

"Aplicación de la Inteligencia Artificial a la Predicción de las Enfermedades Cardiacas"

Trabajo presentado en
cumplimiento parcial de la Materia de
Internet of Things

Por:

Anaya Valenzuela, Joselyn Nila
Lizárraga Vásquez, Carla Rosmery
Lozano Dávila, Angely
Vega Vergara, Maritza Marlene
Velasquez Chavez, Abel

Docente:

Reategui Morales, Juan Carlos

Ciudad universitaria, julio del 2021

ÍNDICE

1. INTRODUCCIÓN

Los avances tecnológicos que se han dado, sobre todo desde el siglo XX, protagonizados por mejoras en salud e higiene, vacunas y antibióticos, han generado un marcado descenso de la mortalidad del ser humano, desde la edad infantil hasta la adultez. Esto ha sido el factor fundamental para el aumento de la esperanza de vida (Garrido, 2009).

En este presente trabajo se pondrá el foco de atención en la ayuda que estas tecnologías ofrecen en la aplicación de la inteligencia artificial a la predicción de las enfermedades cardíacas tal como lo dice el título.

Para poder desarrollar esta aplicación primero es necesario plantear el problema, en este caso las enfermedades cardiacas y sus principales causas, lo que brinda a esta investigación la data necesaria correspondiente al Perú.

Una vez obtenidos los datos necesarios procederemos al análisis basados en la tecnología desarrollada por Python y sus múltiples librerías orientadas hacia el Machine Learning.

Con los resultados obtenidos de estos análisis se podrá informar a las autoridades para que estas tomen conciencia hacia donde deben ser orientados los recursos y a su vez informar a la población de la realidad de las enfermedades asociadas al corazón.

2. PLANTEAMIENTO DEL PROBLEMA

2.1. Identificación del problema

En el ámbito de la salud en términos generales, se ha empleado con éxito para diagnosticar y predecir distintas enfermedades como: Cáncer, diabetes, enfermedades inflamatorias intestinales, problemas cardíacos. Sin embargo en Perú y en el mundo, el problema radica principalmente en enfermedades al corazón.

Según la Organización Mundial de la Salud en el año 2016 se produjeron 17,9 millones de muertes por enfermedades al corazón, es decir, el 31% de las muertes mundiales. El 85% de esas muertes fue resultado de un ataque cardíaco o derrames cerebrales.

A lo largo de los años, además, se han mostrado constantes avances haciendo uso de la Inteligencia Artificial. Por ejemplo, el aprendizaje automático (Machine Learning) es una de las áreas de investigación de la Inteligencia Artificial que consiste en suministrar datos a una computadora para que aprenda a resolver un problema. De esta manera, a través de ello se pueden desarrollar proyectos o un modelo basado en datos médicos para predecir si una persona puede sufrir una enfermedad en el corazón.

2.1.1. Descripción de la Formulación interrogativa del problema

Actualmente, el diagnóstico de estas enfermedades en Perú se basa en una serie de exámenes tediosos y costosos (análisis clínicos, electrocardiograma, radiografía de tórax, ecocardiograma). Sin embargo, se ha observado una asociación entre los datos de los pacientes de distintos hospitales y clínicas con varias de estas enfermedades directamente al corazón, especialmente la hipertensión arterial, la insuficiencia cardíaca, derrame cerebral y la aterosclerosis. Por lo tanto, ¿Por qué no recurrir a la Inteligencia Artificial para diseñar y desarrollar un algoritmo para evitar un mayor porcentaje de mortalidad en el Perú?

2.1.1.1. Formulación interrogativa del problema general

¿Cómo influiría el desarrollo de un programa basado en la Inteligencia Artificial para la predicción de las enfermedades cardiacas?

2.1.1.2. Formulación interrogativa de los problemas específicos

¿Con qué precisión brindaría el algoritmo mejores resultados para el conocimiento de personas con mayor riesgo en contraer una enfermedad cardiaca?

3. ANTECEDENTES GENERALES

3.1. Idea del proyecto

Dar servicio tanto a unidades públicas como privadas relacionadas a la predicción de enfermedades cardiacas a través de un algoritmo desarrollado por la empresa brindando así un servicio de consultoría.

3.2. Descripción del Producto o Servicio

Las instituciones tanto públicas como privadas nos brindan los datos necesarios de su población y la empresa se encarga de estandarizar las muestras para ser analizadas por el algoritmo y la empresa tiene la responsabilidad de entregar los resultados con las posibles tomas de decisiones convenientes para el que contrata el servicio.

3.3. Descripción del mercado

La naturaleza de este proyecto es poder predecir las enfermedades cardiacas a través de los datos principalmente orientados a grupos grandes de personas. Ya sea en distritos, regiones o países de acuerdo a la Data a la que se cuente.

Por eso está orientado principalmente a organizaciones públicas y privadas que manejan grandes cantidades de información. En el caso publicado este proyecto está orientado a entidades públicas y en el caso de las privadas poder brindar el servicio a entidades como ONGs o a la industria de la salud.

Como fin de este proyecto se delimita el objeto de estudio al país de Perú.

4. DETERMINACIÓN DE OBJETIVOS DEL PROYECTO

4.1. Objetivo General

Desarrollar un programa capaz de predecir si un paciente sufrirá de una enfermedad cardiaca, utilizando diferentes algoritmos dentro del rango de la Inteligencia Artificial.

4.2. Objetivos Específicos

- Desarrollar un algoritmo genético para analizar la variabilidad y el rendimiento basado en datos médicos, encontrando así mejores resultados para las personas que corren mayor riesgo de contraer una enfermedad cardiaca.
- Clasificar a los individuos con mayor precisión posible, utilizando el aprendizaje automático (Machine Learning) supervisado.
- Configurar, probar y comparar algoritmos inteligentes diferentes para ver cuál se adapta mejor a los datos proporcionados.
- Demostrar el potencial de la Inteligencia Artificial y de las tecnologías para las máquinas de poder aprender para incrementar el acceso a diagnósticos cardíacos seguros y eficaces que pueden servir para salvar la vida de los pacientes.

5. JUSTIFICACIÓN

El uso de la Inteligencia Artificial en el ámbito médico y en concreto a la cardiología ha producido hoy en día múltiples mejoras en el campo diagnóstico y preventivo de algunas enfermedades cardiacas.

Por tanto, el desarrollo de un programa que permita predecir el comportamiento de los tratamientos realizados en función de las diferentes características del paciente, de la técnica y del profesional es de vital importancia.

Con el fin de evitar, disminuir o cambiar los factores que influyen de forma negativa en el pronóstico de las enfermedades cardiacas facilitando que el especialista establezca un pronóstico lo más cercano a la realidad posible con mayor exactitud y precisión; y haciendo que el paciente acomode sus expectativas a los datos esperados.

6. ALCANCE

La aplicación de la Inteligencia artificial a la predicción de las enfermedades cardiacas, permite una atención agilizada para con los pacientes, teniendo un modelo para predecir si una persona puede sufrir una enfermedad en el corazón. Así mismo usaremos librerías de Machine Learning para implementar el modelo, el programa mostrará gráficas estadísticas a partir de los datos ingresados, como ingresar los datos de las causas de la enfermedad y edades.

Para analizar las técnicas de machine learning nos apoyaremos en el uso de los librerias train_test_split que la utilizaremos para separar los datos de entrenamiento y prueba, importamos el algoritmo SVM, las funciones de matriz de confusión, las funciones de exactitud y precisión del modelo., que utilizaremos para construir y probar el modelo.

Para seleccionar el conjunto de datos cardiovasculares se procedió a la selección de diferentes dataset de alto impacto con la información requerida para ejecutar el entrenamiento, validación y pruebas del modelo computacional predictivo.

El sistema ha sido desarrollado usando la siguiente plataforma tecnológica:
Sistema Operativo, Windows 10.
Lenguaje de programación Python para la creación del algoritmo clasificador.
El lenguaje visual es por consola, nos permite conseguir el objetivo deseado.

7. MARCO TEÓRICO

7.1. Definición de Conocimientos

El conocimiento es la comprensión adquirida, la cual implica aprendizaje, concienciación y familiaridad con una o más materias; el conocimiento se compone de ideas, conceptos, hechos y figuras, teorías, procedimientos y relaciones entre ellos, y formas de aplicar los procedimientos a la resolución práctica de los problemas.

Debido a la variedad de formas que el conocimiento se puede asumir, los problemas involucrados en el desarrollo de una representación de conocimientos complejos, interrelacionados y dependientes del objetivo. En términos generales, el conocimiento debe estar representado de tal forma que:

- Capture generalizaciones
- Pueda ser comprendido por todas las personas que vayan a proporcionar y procesarlo.
- Pueda ser utilizado en diversas situaciones aun cuando no sea totalmente exacto o completo.
- Pueda ser utilizado para reducir el rango de posibilidad que usualmente debería considerarse para buscar soluciones.

7.2. Enfermedades Cardiacas

Las enfermedades cardíacas son un grupo de desórdenes del corazón.

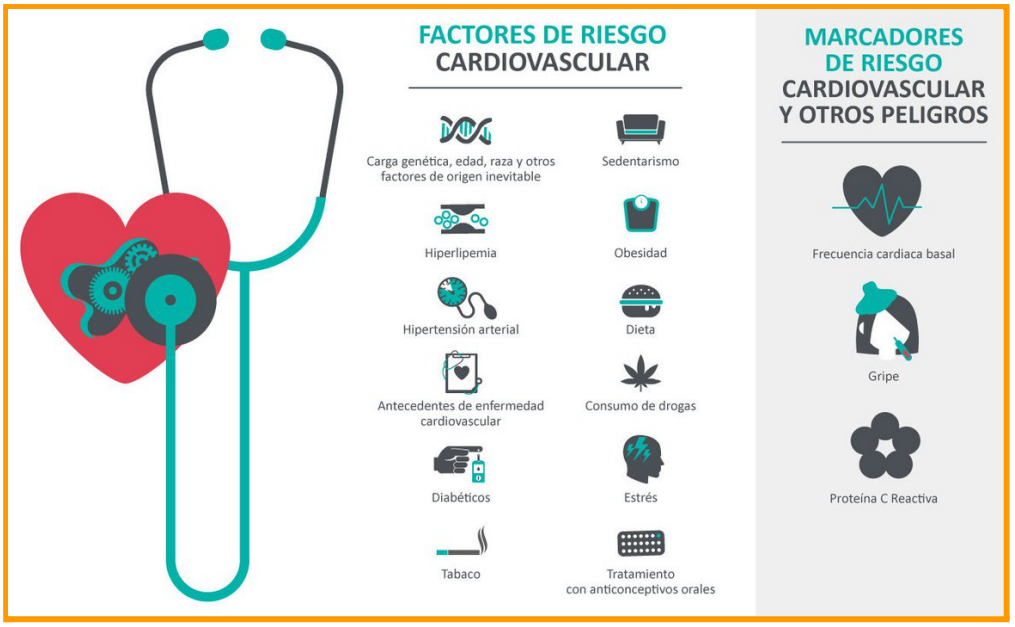
En el mundo, cada cuatro segundos ocurre un infarto agudo del miocardio y al menos una de cada tres personas pierde la vida por alguna patología relacionada con enfermedades cardiacas. En Perú, el 19% de mujeres y hombres de 30 a 69 años muere de enfermedades cardiacas, hay más de 17 millones de hipertensos, 14 millones de dislipidémicos, 6 millones de diabéticos, 35 millones de adultos con sobrepeso u obesidad y 15 millones con grados variables de tabaquismo.

En una región de Lima, los hombres de esa comunidad tienen un alto riesgo de padecer enfermedades cardíacas. Por tal motivo se han tomado algunos datos médicos y de acuerdo a estos se ha determinado si es posible padecer la enfermedad o no.

7.2.1. Factores conductuales y enfermedad de base

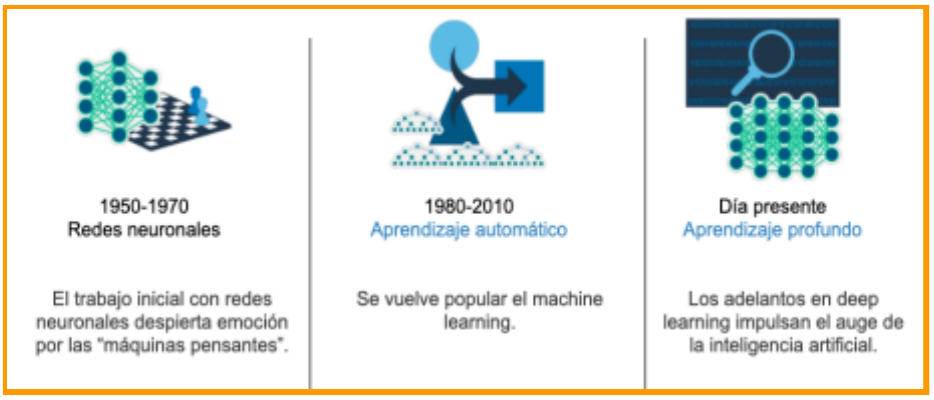
- Tabaquismo: es el factor de riesgo conductual más importante, siendo el que aumenta más el riesgo de padecer una enfermedad cardiovascular. El tabaco se sitúa entre los factores de riesgo independientes más reflejados en las guías de actuación de diversas sociedades cardiológicas internacionales para la valoración del riesgo cardiovascular.
- Dieta: Los hábitos no saludables en la dieta contribuyen al aumento de la probabilidad de desarrollar una enfermedad cardiovascular. Numerosos estudios afirman que incluir ciertos tipos de alimentos en el día a día facilitan el mantener unos niveles correctos de colesterol y evitan la obesidad. La dieta mediterránea sería un factor de protección .

- Inactividad física: Las personas inactivas tienen mayor riesgo de sufrir enfermedad cardiovascular que las personas que hacen ejercicio regularmente; ya que el hacer ejercicio de forma regular fortalece el músculo cardiaco y hace que las arterias sean más flexibles.
- Alcohol: el exceso de alcohol aumenta la presión sanguínea, los triglicéridos, cáncer y otras enfermedades
- Hipertensión arterial, siendo uno de los factores de riesgo más destacados. Un 70 % de los ictus ocurren a causa de la hipertensión.
- Colesterol elevado: las personas con enfermedades cardiovasculares tienden a presentar niveles elevados de colesterol en la sangre.
- Haber sufrido alguna patología cardiaca .
- Diabetes: Es un factor de riesgo independiente y en gran medida relacionado con la hipertensión arterial.
- Obesidad (IMC): El exceso de peso puede elevar el colesterol y causar presión arterial alta .
- Antecedentes familiares por trastornos cardíacos. Hay enfermedades cardiovasculares que son más comunes entre ciertos grupos raciales y étnicos. Un ejemplo sería que en este trabajo tiene una finalidad docente. La Facultad de Farmacia no se hace responsable de la información contenida en el mismo.

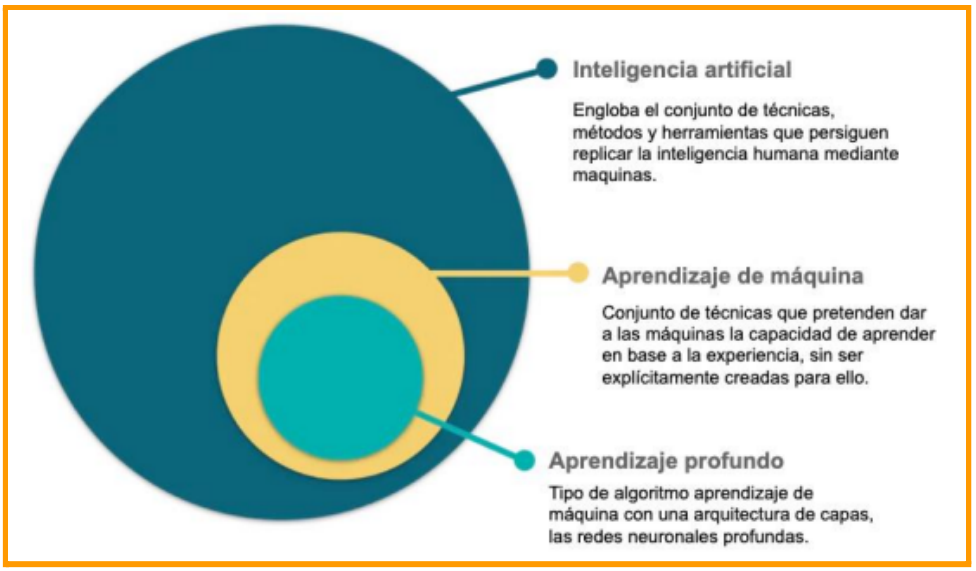


7.3. Inteligencia Artificial

La Inteligencia artificial es el campo científico de la informática que se centra en la creación de programas y mecanismos que pueden mostrar comportamientos considerados inteligentes. En otras palabras, la IA es el concepto según el cual "las máquinas piensan como seres humanos". Normalmente un sistema de IA es capaz de analizar datos en grandes cantidades (big data), identificar patrones y tendencias y, por lo tanto, formular predicciones de forma automática, con rapidez y precisión. En la línea de tiempo visualizada en la ilustración se pueden evidenciar las tres principales técnicas de IA las cuales son: redes neuronales, aprendizaje de máquina o automático y aprendizaje profundo, las cuales se entrarán en detalle más adelante.



A su vez la inteligencia artificial (IA) tiene relación directa con el aprendizaje de máquina (ML), y este a su vez con la técnica de aprendizaje profundo (DL). Para que el concepto quede más claro la ilustración indica la relación directa entre estas tres técnicas de la inteligencia artificial.



7.4. Aprendizaje Máquina

Es una rama de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de datos, identificar patrones y tomar decisiones con mínima intervención humana. Aprenden de cálculos previos para producir decisiones y resultados confiables y repetibles.

Existen tres subramas en los cuales se clasifican las técnicas de aprendizaje profundo:

- Aprendizaje Supervisado: Se tiene un histórico de datos en el que se dispone de la variable a predecir (o campo objetivo o etiqueta) para entrenar el modelo.

- Aprendizaje No Supervisado: Para el entrenamiento del modelo se dispone de un conjunto de datos de históricos que no están etiquetados y el algoritmo descifra la información y clasifica por sí solo.
- Aprendizaje Por Refuerzo: Las dos cualidades principales del aprendizaje por refuerzo son prueba-error y la recompensa. Un algoritmo que aprende por refuerzo lo que hace es recibir una recompensa positiva si el resultado de la acción es positiva, y negativa o nula si es malo.

7.5. Machine Learning

"Machine Learning es la ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándose con datos e información en forma de observaciones e interacciones con el mundo real." — Dan Fagella

Machine learning ofrece una manera eficiente de capturar el conocimiento mediante la información contenida en los datos, para mejorar de forma gradual el rendimiento de modelos predictivos y tomar decisiones basadas en dichos datos. Se ha convertido en una tecnología con una amplia presencia, y actualmente está presente en: filtros anti-spam para correo electrónico, conducción automática de vehículos o reconocimiento de voz e imágenes.

7.6. Matriz de confusión

Es una herramienta que permite identificar fácilmente el rendimiento de un Modelo supervisado de aprendizaje de máquina. Con esta métrica es fácil detectar dónde el sistema está confundiendo las diferentes clases o resultados de clasificación.

Como se visualiza en la ilustración de cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.

7.6.1. Medidas de rendimiento

VP es la cantidad de positivos que fueron clasificados correctamente como positivos por el modelo (Verdadero positivo)

VN es la cantidad de negativos que fueron clasificados correctamente como negativos por el modelo (Verdadero negativo)

FN es la cantidad de positivos que fueron clasificados incorrectamente como negativos. Error tipo 2 (Falsos Negativos)

FP es la cantidad de negativos que fueron clasificados incorrectamente como positivos. Error tipo 1 (Falsos positivos)

- Precisión (P): definida como la cantidad de registros verdaderos positivos (VP) dividida entre la suma de verdaderos positivos (VP) y falsos positivos (FP) obtenidos en la ejecución del algoritmo.

$$P = \frac{VP}{VP + FP}$$

- Exhaustividad (R, Recall) Una medida de integridad definida como la proporción del número de registros verdadero positivos (VP) divididos por la suma de los verdaderos positivos (VP) y registros clasificados como falsos negativos (FN).

$$R = \frac{VP}{VP + FN}$$

- F1-score: Es una función de Precisión y Exhaustividad que permite buscar un equilibrio entre Precisión y Exhaustividad

$$F1 = 2\,x\,\frac{PRE\ x\ REC}{PRE + REC}$$

- Exactitud (ACC, Accuracy) Es el porcentaje de casos en los que el modelo ha acertado, es el número de predicciones correctas sobre el número total de predicciones.

$$ACC = \frac{TP + TN}{FP + FN + TP + TN}$$

- Macro avg Se puede calcular como:

$$B_{macro} = \frac{1}{q}\sum_{\lambda=1}^{q} B(tp_\lambda, fp_\lambda, tp_\lambda, fn_\lambda)$$

Donde:

- *B*: *medida de evaluación binaria*
- *tp*: *verdaderos negativos*,*tn*: *falsos positivos*
- *fp*: *falsos positivo*, *fn*: *falsos negativos*
- *λ*: *una etiqueta del conjunto de etiquetas L* = {*λj*:*j* = 1 ... *q*}

Se puede considerar una medida de evaluación binaria *B*{*tp* λ, *fp* λ,*tp* λ, *fn* λ} que se calcula basado en el número de verdaderos positivos (*tp*), verdaderos negativos (*tn*), falsos positivos (*fp*) y falsos negativos (*fn*). *tp* λ, *fp* λ,*tp* λ, *fn* λ son el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos después de la evaluación binaria por una etiqueta λ

7.6.2. Validación cruzada

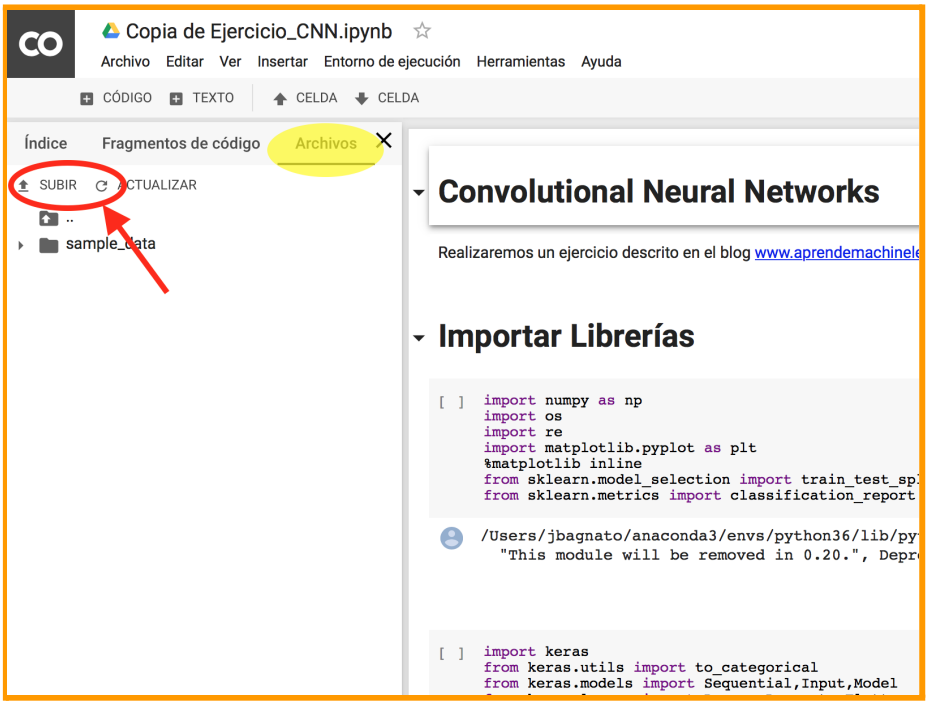
La validación cruzada es una técnica para la validación de modelos que permite evaluar la calidad del análisis estadístico y sus resultados. El objetivo es hacer que el modelo se generalice hacia un conjunto de pruebas independientes. Esta técnica es utilizada para hacer una predicción a partir de un modelo de aprendizaje automático, ayuda a estimar cómo un modelo predictivo se desempeñará con precisión en la práctica cuando se despliegue como una aplicación ML. Durante la validación cruzada, un modelo suele entrenarse con un conjunto de datos de un tipo conocido. Por el contrario, se prueba utilizando un conjunto de datos de tipo desconocido.

7.7. Google Colab

Colab es un servicio cloud, basado en los Notebooks de Jupyter, que permite el uso gratuito de las GPUs y TPUs de Google, con librerías como: Scikit-learn, PyTorch, TensorFlow, Keras y OpenCV. Todo ello con bajo Python 2.7 y 3.6, que aún no está disponible para R y Scala.

Aunque tiene algunas limitaciones, que pueden consultarse en su página de FAQ, es una herramienta ideal, no solo para practicar y mejorar nuestros conocimientos en técnicas y herramientas de Data Science, sino también para el para el desarrollo de aplicaciones (pilotos) de machine learning y deep learning, sin tener que invertir en recursos hardware o del Cloud.

Con Colab se pueden crear notebooks o importar los que ya tengamos creados, además de compartirlos y exportarlos cuando queramos. Esta fluidez a la hora de manejar la información también es aplicable a las fuentes de datos que usemos en nuestros proyectos (notebooks), de modo que podremos trabajar con información contenida en nuestro propio Google Drive, unidad de almacenamiento local, github e incluso en otros sistemas de almacenamiento cloud, como el S3 de Amazon.



7.8. Scikit-Learn

Scikit-Learn es una de estas librerías gratuitas para Python. Cuenta con algoritmos de clasificación, regresión, clustering y reducción de dimensionalidad. Además, presenta la compatibilidad con otras librerías de Python como NumPy, SciPy y matplotlib.

La gran variedad de algoritmos y utilidades de Scikit-learn la convierten en la herramienta básica para empezar a programar y estructurar los sistemas de análisis de datos y modelado estadístico. Los algoritmos de Scikit-Learn se combinan y depuran con otras estructuras de datos y aplicaciones externas como Pandas o PyBrain.

La ventaja de la programación en Python, y Scikit-Learn en concreto, es la variedad de módulos y algoritmos que facilitan el aprendizaje y trabajo del científico de datos en las primeras fases de su desarrollo. La formación de un Máster en Data Science hace hincapié en estas ventajas, pero también prepara a sus alumnos para trabajar en otros lenguajes. La versatilidad y formación es la clave en el campo tecnológico.

8. MATERIAL Y MÉTODO

8.1. Área De Estudio

La presente investigación se realizará en el Hospital Nacional Dos de Mayo - Parque "Historia de la Medicina Peruana", S/N, Av. Miguel Grau 13, Cercado de Lima 15003.

8.2. Tipo de Investigación

El presente proyecto de investigación es de tipo cuantitativo y descriptivo. Cuantitativo porque mide algunas causas que podrían tener un alto riesgo de padecer enfermedades cardíacas; descriptivo porque está dirigido a describir el presente.

8.3. Población

La presente investigación cuenta con una población constituida por los pacientes asistentes al Hospital Nacional Dos de Mayo.

8.4. Muestra

Está constituida por 462 pacientes asistidos en el Hospital Nacional Dos de Mayo.

8.5. Operacionalización de Variables

N°	Variable	Descripción Conceptual	Descripción Operacional	Indicador
V1	presión arterial sistólica (sbp)	La presión sistólica se mide cuando los ventrículos del corazón se contraen.	Cantidad de mercurio se consideran en el rango de los niveles de presión	normal 120 mm alta >= 140mm
V2	Tabaco	La acumulación en kg de tabaco dentro de su organismo	La Cantidad de Monóxido de Carbono disminuye el oxígeno al miocardio y la formulación de coágulos	10 mg / ml
V3	proteína de baja densidad del colesterol (LDL)	Es la cantidad de colesterol unido a lipoproteínas de baja densidad.	La cantidad de colesterol LDL genera Tensión en las arterias y obstrucciones en la circulación	130 mg/dl
V4	Adiposity - adiposidad	Es el espesor de la grasa epicárdica que se forma en la pared libre del ventrículo derecho al final de la sístole	La cantidad grasa visceral del corazón produce un aumento de grasa al tejido adiposo pardo	18,5 kg/m2
V5	Antecedentes familiares	Transmisión a través del material genético contenido en el núcleo celular, de las características anatómicas,	Si el paciente NO presenta antecedentes familiares. Si el paciente SI presenta antecedentes familiares.	NO = 1 SI = 2

		fisiológicas, etc. de un ser vivo a sus descendientes. El ser vivo resultante tendrá información de los dos padres.		
V6	Comportamiento Tipo-A	El comportamiento Tipo A número ilimitado de actividades en el periodo más corto de tiempo	La cantidad de resistencia produce la velocidad de arritmias del cardio	55s/mV
V7	Obesidad	La obesidad es una enfermedad sistémica, crónica, progresiva y multifactorial, que se define como una acumulación anormal o excesiva de grasa	Se considerará como grado de obesidad con base en el índice de masa corporal (IMC)	Para adultos IMC ≥ a 30 Kg/m2
V8	Consumo recurrente de alcohol	Ingesta de bebidas alcohólicas que puede llevar a adicción interfiriendo en la salud física, mental, social y/o familiar.	Cantidad de alcohol que consume.	1.000 ml. a menos
V9	Edad de inicio	Edad en la que inició con la enfermedad	Se considera de todas las edades.	1 año a más

8.6. Desarrollo del programa

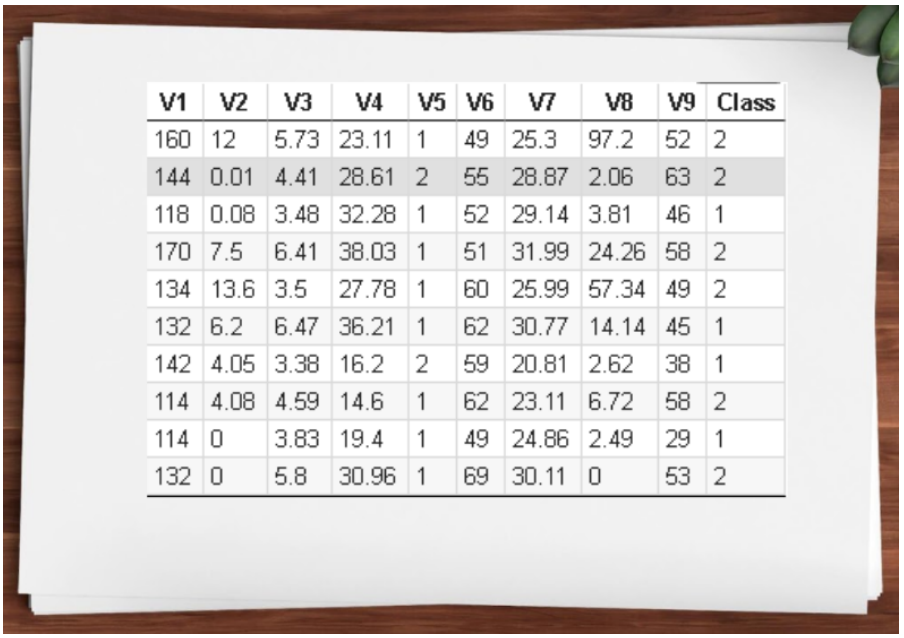
Es importante recordar que estos son términos médicos, por lo que en esta guía se mostrará el desarrollo de la aplicación del código en base a la Inteligencia Artificial con el fin de predecir las Enfermedades Cardiacas.

A continuación se mostrará detalladamente el desarrollo del programa, el cual está desarrollado en base no solo de los algoritmos de Machine Learning, sino también en base a nuestras habilidades de programación en Python.

1. Lo primero será importar la librería básica que vamos a utilizar, en este caso **Pandas**. Pandas es una librería de Python especializada en el manejo y análisis de estructuras de datos.

2. Seguidamente importamos los datos. En este procedimiento utilizamos **read_csv** y colocamos el nombre exacto del archivo junto a su extensión CSV, todo esto debe estar entre comillas. Esta data es tomada de una muestra de pacientes del Centro Materno Infantil José Gálvez - Villa María del Triunfo.

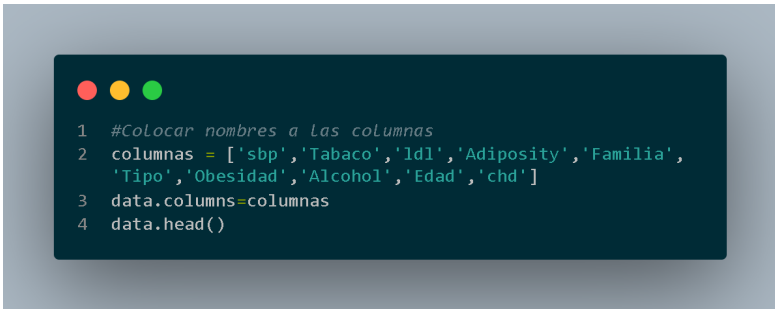
3. Si imprimimos los datos que acabamos de exportar, podemos observar que los nombres de las columnas se refieren a una numeración **V1 hasta V9**. Por esta razón es conveniente cambiar estos nombres por las variables reales de cada columna y las cuales nos ayudarán a la correcta evaluación de los datos.



V1	V2	V3	V4	V5	V6	V7	V8	V9	Class
160	12	5.73	23.11	1	49	25.3	97.2	52	2
144	0.01	4.41	28.61	2	55	28.87	2.06	63	2
118	0.08	3.48	32.28	1	52	29.14	3.81	46	1
170	7.5	6.41	38.03	1	51	31.99	24.26	58	2
134	13.6	3.5	27.78	1	60	25.99	57.34	49	2
132	6.2	6.47	36.21	1	62	30.77	14.14	45	1
142	4.05	3.38	16.2	2	59	20.81	2.62	38	1
114	4.08	4.59	14.6	1	62	23.11	6.72	58	2
114	0	3.83	19.4	1	49	24.86	2.49	29	1
132	0	5.8	30.96	1	69	30.11	0	53	2

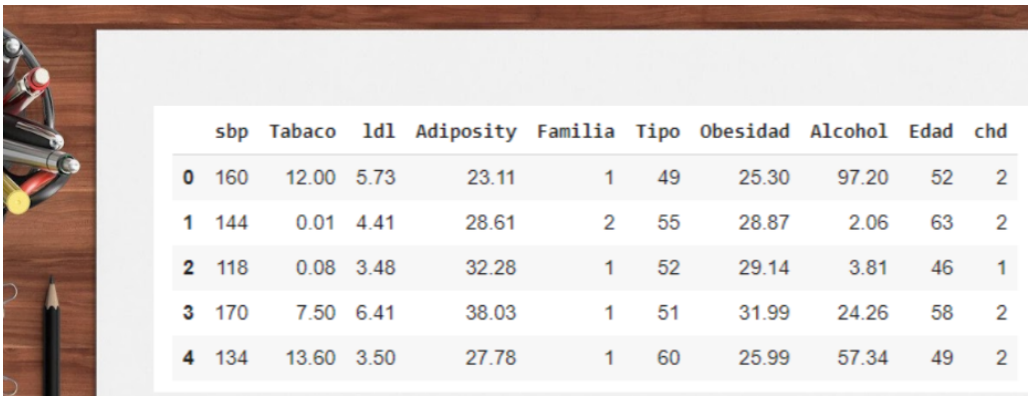
4. Por lo tanto, definimos la variable columnas y colocamos los nuevos nombres. El orden de los nombres debe coincidir con el orden de las columnas dispuestas.

   Ahora agregamos estos nombres a nuestros datos, para ello nos valemos de la instrucción "**columns**" que nos proporciona Pandas para realizar este cambio.



```
#Colocar nombres a las columnas
columnas = ['sbp','Tabaco','ldl','Adiposity','Familia',
'Tipo','Obesidad','Alcohol','Edad','chd']
data.columns=columnas
data.head()
```

5. Ahora verifiquemos los nuevos nombres al imprimir el resultado final. Como se puede ver ahora, nuestros datos tienen los nombres respectivos para cada una de las columnas.



	sbp	Tabaco	ldl	Adiposity	Familia	Tipo	Obesidad	Alcohol	Edad	chd
0	160	12.00	5.73	23.11	1	49	25.30	97.20	52	2
1	144	0.01	4.41	28.61	2	55	28.87	2.06	63	2
2	118	0.08	3.48	32.28	1	52	29.14	3.81	46	1
3	170	7.50	6.41	38.03	1	51	31.99	24.26	58	2
4	134	13.60	3.50	27.78	1	60	25.99	57.34	49	2

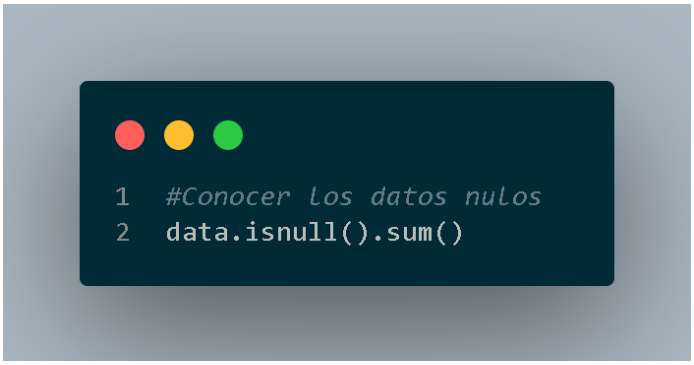
6. Continuaremos con el análisis de los datos. Lo primero es verificar el formato que contiene nuestros datos, para ello implementamos la instrucción **dtypes**. Esta instrucción nos permite saber qué tipos de datos nos entrega cada columna, en caso de ser necesario hacer una modificación.

```
1  #Conocer el formato de los datos
2  data.dtypes
```

7. Como podemos observar los datos son enteros y flotantes, por lo que no es necesario realizar un procesamiento mayor en cuanto al tipo de datos.

```
sbp             int64
Tabaco        float64
ldl           float64
Adiposity     float64
Familia         int64
Tipo            int64
Obesidad      float64
Alcohol       float64
Edad            int64
chd             int64
dtype: object
```

8. Posteriormente verificamos si se cuentan con datos perdidos. Para esto nos apoyamos de la instrucción **isnull.** Con esta instrucción podemos ver si existen campos vacíos dentro de las columnas, así mismo hace un conteo de los mismos. De acuerdo a los resultados obtenidos no se cuenta con ningún dato perdido.



```
sbp          0
Tabaco       0
ldl          0
Adiposity    0
Familia      0
Tipo         0
Obesidad     0
Alcohol      0
Edad         0
chd          0
```

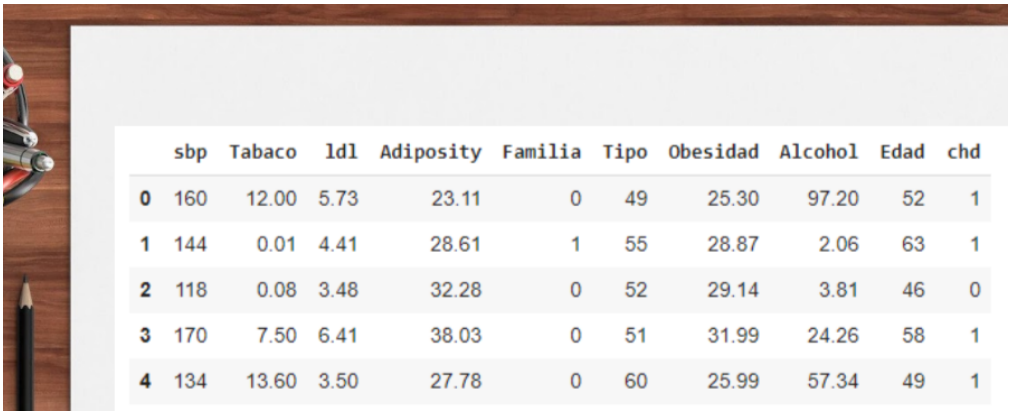
9. Según el análisis de datos realizado podemos determinar que no existe tanto trabajo para el procesamiento de los datos. Pero, esto no es tan cierto. Ya que si detallamos los datos podremos observar dos detalles.
    - El primero es que las columnas "**familia**" y "**chd**", los datos son 1 y 2, esto lo podemos cambiar por 0 y 1.
    - A su vez si observamos los valores de la columna "**sbp**" podremos ver que están en una escala mucho mayor a diferencia de las otras columnas. Por tal motivo debemos ajustar esta escala para obtener un mejor resultado.

    De acuerdo a esto, realizamos el procesamiento de los datos. Lo primero que hicimos fue corregir los valores de las columnas "familia" y "chd". Para este propósito importamos **LabelEncoder** a través de **Scikit-Learn**, el cual codifica los datos de la etiqueta que le pasemos .

10. Como se había mencionado antes. A través de **st_transform** ajustamos los valores de familia y chd entre 2 y 1 a valores de 1 y 0.

```
encoder = LabelEncoder()
data['Familia']=encoder.fit_transform(data['Familia'])
data['chd']=encoder.fit_transform(data['chd'])
data.head()
```

11. Posteriormente, se puede observar cómo quedaron ambas columnas. Donde anteriormente había un 2 ahora hay un 1 y donde había un 1 ahora ese valor es 0.



	sbp	Tabaco	ldl	Adiposity	Familia	Tipo	Obesidad	Alcohol	Edad	chd
0	160	12.00	5.73	23.11	0	49	25.30	97.20	52	1
1	144	0.01	4.41	28.61	1	55	28.87	2.06	63	1
2	118	0.08	3.48	32.28	0	52	29.14	3.81	46	0
3	170	7.50	6.41	38.03	0	51	31.99	24.26	58	1
4	134	13.60	3.50	27.78	0	60	25.99	57.34	49	1

12. Ahora procedemos a escalar los valores de la columna "**sbp**".

Para esto volvemos a utilizar la librería sklearn preprocessing e importamos **MinMaxScaler**. Con esta función podemos establecer un rango mínimo y máximo para los valores definidos y los transforma en ese rango.

El siguiente paso es el de definir los rangos, por lo tanto los definimos entre 0 y 100, es decir los valores dentro de la columna "sbp" estarán dentro de la escala de 0 y 100.

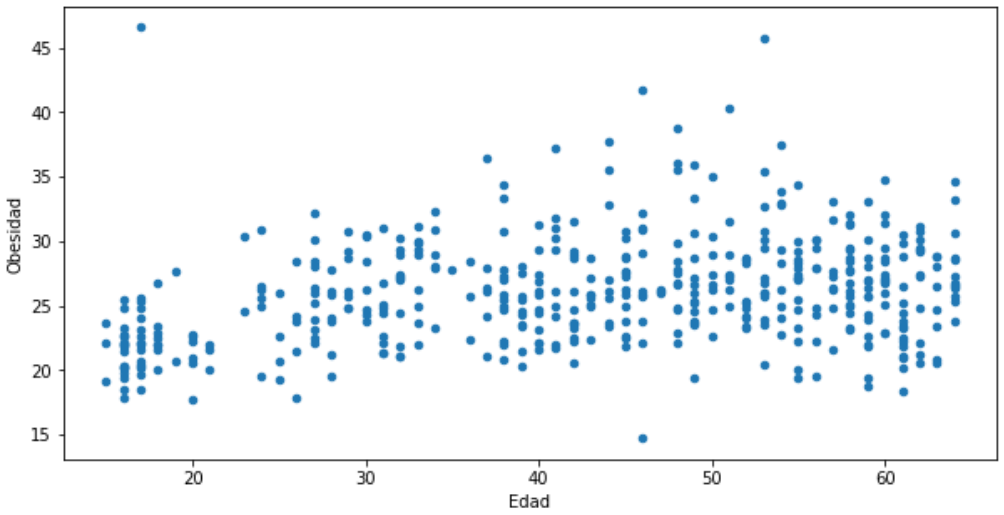
```
#Escalamos los valores de la columna sbp
from sklearn.preprocessing import MinMaxScaler
scale = MinMaxScaler(feature_range =(0,100))
data['sbp'] = scale.fit_transform(data['sbp'].values.reshape(-1,1))
data.head()
```

13. Ahora que tenemos nuestros datos listos para poder utilizarlos dentro de los algoritmos de Machine Learning.

    Lo primero que graficamos fue la obesidad de acuerdo a la edad. En este caso utilizamos un código sencillo para realizar la gráfica. Simplemente indicamos el conjunto de datos punto **plot** y definimos los valores que se graficaron en los ejes “x” y “y”, el tipo de **gráfica será de dispersión** o scatter y el tamaño de la figura.

```
#Visualizar la obesidad de acuerdo a la edad
data.plot(x='Edad',y='Obesidad',kind='scatter',figsize =(10,5))
```

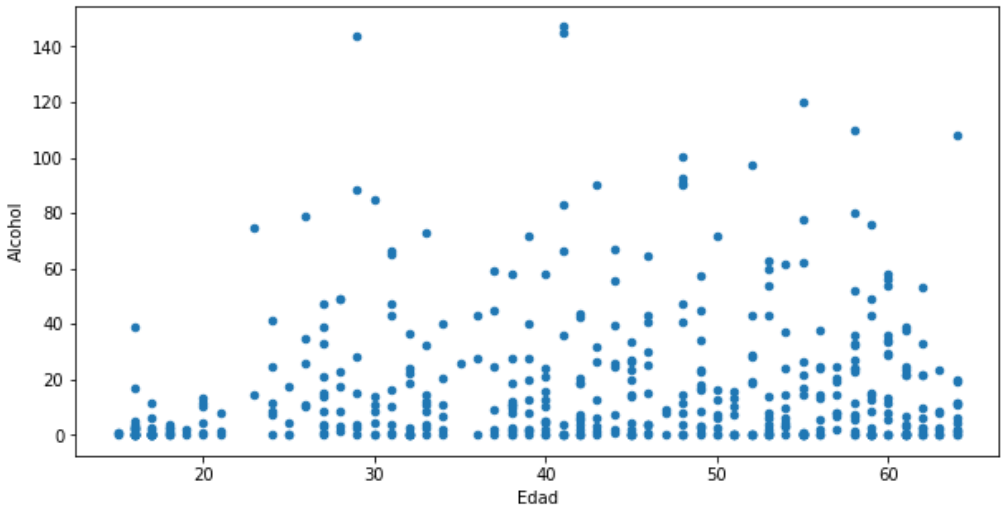
14. Gracias a la función plot logramos una gráfica 2D. Según esta información a medida que se aumenta de edad, aumenta la obesidad en la persona. Solamente hay casos aislados cuando esto no sucede.



15. Ahora, grafiquemos el consumo de alcohol de acuerdo a la edad. Según el código la escala se define por **figsize**.

```
#Visualizar el consumo de alcohol de acuerdo a la edad
data.plot(x='Edad',y='Alcohol',kind='scatter',figsize =(
10,5))
```

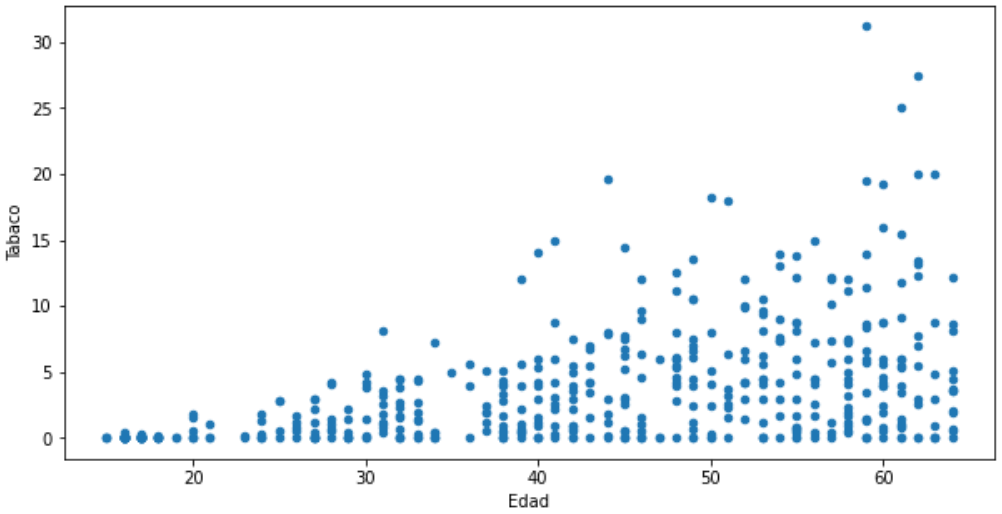
16. El resultado obtenido se puede determinar que el consumo de alcohol se empieza a partir de los 20 años y se mantiene a lo largo de los años.



17. Grafiquemos ahora el consumo del tabaco de acuerdo a la edad. Para este caso en particular, la variable **kind Scatter** nos proporciona una gráfica de dispersión.

```
#Visualizar el consumo de tabaco de acuerdo a la edad
data.plot(x='Edad',y='Tabaco',kind='scatter',figsize =(10,5))
```

18. Según el gráfico siguiente, podemos ver más claramente que a medida que aumenta la edad, aumenta significativamente el consumo de tabaco. Datos que en ocasiones pensaríamos que es al contrario, pero con esta información podemos ver que no es así.



19. Además construimos el modelo de Machine Learning.

- Lo primero fue importar las librerías que utilizaremos para construir y probar el modelo. La primera librería será **train_test_split** que la utilizaremos para separar los datos de entrenamiento y prueba.
- Importamos el algoritmo a utilizar que será SVM o máquinas de vectores de soporte. La librería máquina de vectores de soporte(SVM) nos ayuda para la clasificación. Por lo que queremos predecir si existe alguna enfermedad de corazón o no.
- Por último importamos las funciones para evaluar el modelo que obtengamos. La primera será la matriz de confusión y también importamos las funciones de exactitud y precisión del modelo.

```
### ANÁLISIS DE MACHINE LEARNING ###
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score, precision_sc
ore
```

20. Realizado esto, procedimos a definir las variables dependientes e independientes. Para **"y"** será igual a la columna "chd", mientras que **"x"** será igual al resto de valores de nuestro conjunto de datos.

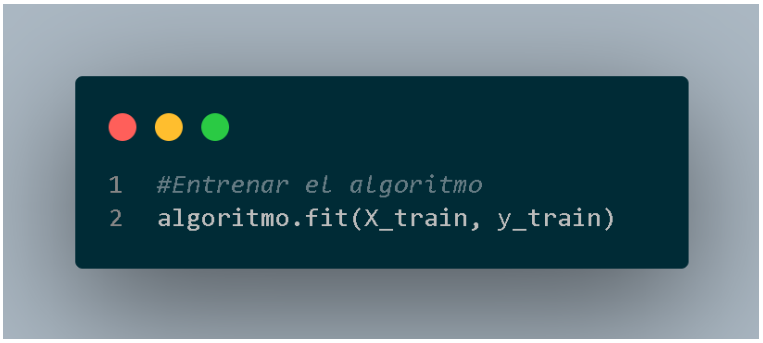
```
#Definir las variable dependiente e independientes
y = data['chd']
X = data.drop('chd', axis =1)
```

21. Luego, separamos los datos de entrenamiento y prueba, utilizando la instrucción **train_test_split**.

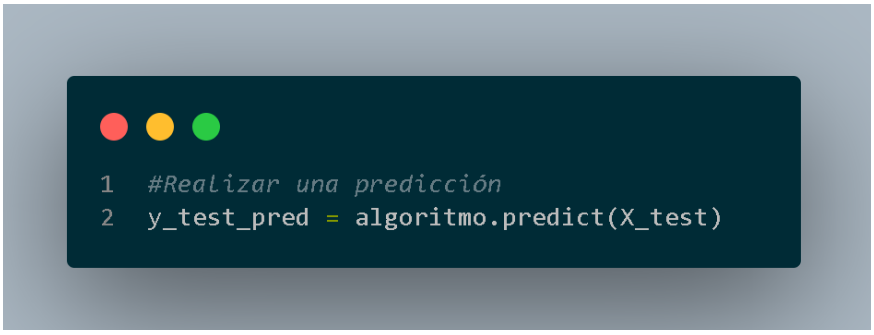
```

#Separar los datos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split
(X, y, test_size = 0.20, random_state=1)

```

22. Definimos el algoritmo a utilizar, que como se indicó anteriormente será el de Máquinas Vectores de Soporte, junto a un kernel lineal. Es decir, llamamos a la librería a través de **algoritmo** y le especificamos el tipo de análisis que se realizará (**linear**).

23. Entrenamos el modelo junto con los datos de entrenamiento que se asignaron previamente.

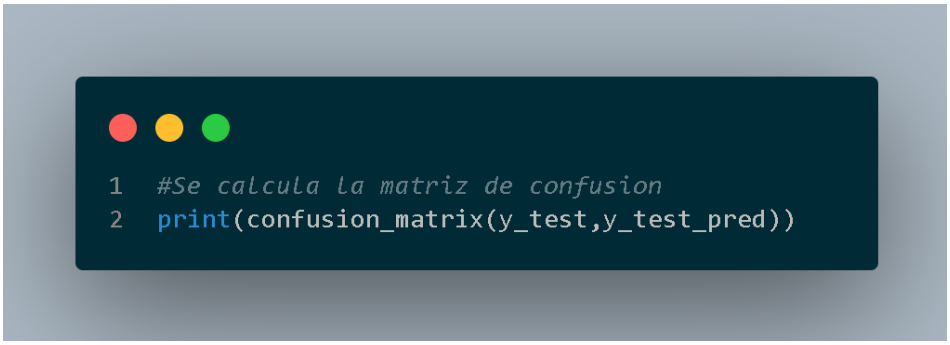


24. Finalmente, realizamos una predicción junto con el modelo obtenido y los datos de prueba.



9. RESULTADOS

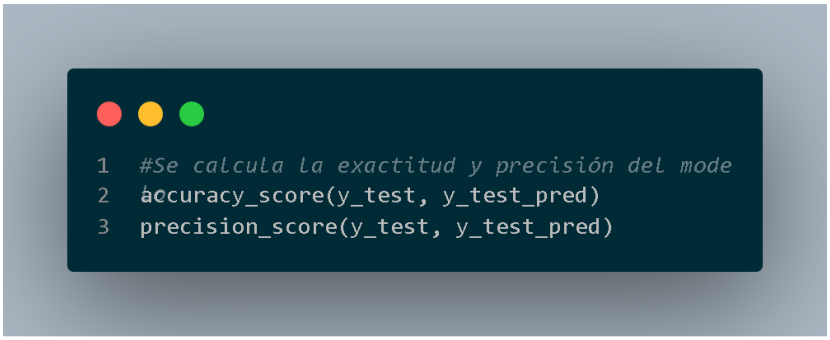
1. Para observar lo obtenido por el modelo realizado, utilizamos las métricas que importamos previamente. La primera a evaluar será la **matriz de confusión**. Los valores obtenidos son 67 datos correctos y 26 datos incorrectos.



```
#Se calcula la matriz de confusion
print(confusion_matrix(y_test,y_test_pred))
```

```
[[57  9]
 [17 10]]
```

2. Con los datos obtenidos, podemos ya decir que el modelo no acertó correctamente un gran número de datos. Pero veamos la precisión y la exactitud del mismo.



```
#Se calcula la exactitud y precisión del mode
accuracy_score(y_test, y_test_pred)
precision_score(y_test, y_test_pred)
```

```
[ ]  #Se calcula la exactitud y precisión del modelo
     accuracy_score(y_test, y_test_pred)

     0.7204301075268817
```

```
[ ]  precision_score(y_test, y_test_pred)

     0.5263157894736842
```

3. Implementando las funciones importadas anteriormente de la librería **metrics** de sklearn, podemos decir que la exactitud del modelo es de 0,72 mientras que la precisión es de 0,526. Finalmente, se puede decir que la información obtenida reafirma lo obtenido en la matriz de confusión.

10. DISCUSIÓN

En este proyecto hemos podido rescatar los múltiples factores de riesgo para enfermedad cardiaca, muchos de estos factores se asocian con el estilo de vida y por tanto pueden ser prevenibles, e incluyen: la obesidad, el sedentarismo, el alcoholismo, el tabaquismo; alimentación saludable, no obstante. También existen factores que son inherentes al individuo como la edad, la genética, el sexo, y que hacen parte de lo establecido como factores de riesgo no modificables, generando una predisposición para el desarrollo de la enfermedad cardiaca.

Las personas inactivas tienen un mayor riesgo de sufrir un ataque al corazón que las personas que hacen ejercicio regular. El ejercicio quema calorías, ayuda a controlar los niveles de colesterol y disminuye la presión arterial. El ejercicio también fortalece el músculo cardiaco y hace más flexible las arterias. Las personas que queman activamente entre 500- 3500 calorías por semana, ya sea en el trabajo o haciendo ejercicio, tienen una expectativa de vida superior a la de las personas sedentarias.

Basta con recordar el impacto que tienen las llamadas “enfermedades cardiacas”: hipertensión, obesidad, entre otros que se relacionan con una alimentación desequilibrada. No es normalmente una relación directa de causas y efecto, pero sí supone uno de los factores que contribuye a aumentar el riesgo de aparición y desarrollo de dicha enfermedad.

11. CONCLUSIONES

Las enfermedades cardíacas son un conjunto de patologías que afectan el corazón y el sistema vascular. El origen de estas enfermedades es multifactorial, siendo un gran avance para su prevención la identificación de los denominados “Factores de Riesgo Cardiaco” que predisponen al desarrollo de estas afecciones.

Gracias a la data obtenida del Hospital Nacional Dos Mayo, hemos observado a través de los algoritmos de Machine Learning que a medida que se aumenta de edad, aumenta la obesidad en la persona así como también aumenta significativamente el consumo de tabaco y finalmente, que el consumo de alcohol empieza a partir de los 20 años y se mantiene a lo largo de los años.

Sin embargo, se ha podido supervisar todos los trabajos revisados al respecto de intervenciones psicológicas en los pacientes con enfermedades cardiacas, por lo que se pueden encontrar mejoras significativas, en mayor o menor medida, en el paciente cardiópata. Cabe resaltar que en muchas ocasiones, no se describe el tipo de intervención que se realiza, ni se detalla el contenido de las sesiones, dejando a juicio del psicólogo clínico que interviene en el grupo médico el contenido de las mismas.

Para concluir, se considera que la eficacia del sistema experto en enfermedades cardiacas ha quedado demostrada que este trabajo pretende ser una pequeña demostración del efecto terapéutico en pacientes médicos. Es nuestra labor como profesionales de la Ingeniería de Sistemas , es adecuar y satisfacer las necesidades de los pacientes de todo tipo y condición, facilitando el desarrollo del mismo, permitiendo así no sólo la recuperación del paciente, sino también el desarrollo de la tecnología en el centro médico y el crecimiento de la credibilidad de nuestra profesión en la realización de nuestros proyectos.

12. REFERENCIAS

′az, P.-D. (2019). *Aplicaciones de la inteligencia artificial en cardiología*. Obtenido de Artículoderevisión: https://www.revespcardiol.org/es-pdf-S0300893219302507

Corazonadas. (2020). Obtenido de (Instituto de Ciencias del Corazón de Valladolid: http://www.icicor.es/publicaciones/corazonadas_febrero2020.pdf

Detomás, J. F. (2020). *La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas*. Obtenido de ResearchGate: https://www.researchgate.net/publication/342322911_La_inteligencia_artificial_y_sus_aplicaciones_en_medicina_II_importancia_actual_y_aplicaciones_practicas

Heart Disease Prediction using Artificial. (2021). Obtenido de IJERT: https://www.ijert.org/research/heart-disease-prediction-using-artificial-intelligence-IJERTCONV9IS04015.pdf

Heart Disease Prediction using Artificial Intelligence. (2020). Obtenido de IJERT: https://www.ijert.org/heart-disease-prediction-using-artificial-intelligence

LA INTELIGENCIA ARTIFICIAL COMO HERRAMIENTA DE DIAGNÓSTICO DE ENFERMEDADES CARDIOVASCULARES MEDIANTE UN ANÁLISIS DE LAS HECES. (15 de 12 de 2020). Obtenido de BIOCODEX: https://www.biocodexmicrobiotainstitute.com/es/publicaciones/la-inteligencia-artificial-como-herramienta-de-diagnostico-de-enfermedades-cardiovasculares-mediante-un-analisis-de-las-heces

The primary use of artificial intelligence in cardiovascular diseases: what kind of potential role does artificial intelligence play in future medicine? (2019). Obtenido de NCBI: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6748906/

Tunes, S. (2020). *Inteligencia artificial aplicada al corazón*. Obtenido de Pesquisa: https://revistapesquisa.fapesp.br/es/inteligencia-artificial-aplicada-al-corazon/

Utilizar la inteligencia artificial para combatir las enfermedades cardíacas. (02 de marzo de 2020). Obtenido de SHAREAMERICA: https://share.america.gov/es/utilizar-la-inteligencia-artificial-para-combatir-las-enfermedades-cardiacas/