

**UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS**

**FACULTAD DE INGENIERÍA**

**BUSINESS PREDICTIVE ANALYTICS (SI709)**

**TRABAJO FINAL**

**PROFESOR:**

Caldearon Sulca, Carlos Felipe

**SECCIÓN:** SS92

**AUTORES**: GRUPO 2

Integrantes:

|  |  |
| --- | --- |
| * Bendezú Jiménez, Luis Ángel | U202022364 |
| * Caqui Pizarro, Valeria Milagros | U20211C241 |
| * Silvina Gonzales, Gianella Lucía | U202019652 |
| * Trujillo Neyra Neil Eduardo | U202020118 |
| * Vásquez Castro, André Nicolas | U20211B559 |

Lima, 27 de junio de 2024

**RESUMEN**

El presente informe detalla el desarrollo y la fundamentación de nuestro proyecto enfocado en la Clínica Mayo, con el objetivo de abordar la necesidad de detección y manejo tempranos de las enfermedades cardiovasculares (ECV). Para ello, hemos utilizado un conjunto de datos que incluye información clínica relevante, como factores de riesgo cardiovascular, resultados de pruebas de laboratorio y diagnósticos médicos.

Nuestro análisis exhaustivo de los datos tiene como finalidad proyectar de manera precisa el riesgo de desarrollar enfermedades cardiovasculares en pacientes atendidos en la Clínica Mayo. Este enfoque permitirá identificar a las personas con mayor riesgo de ECV antes de que ocurran eventos graves, como ataques cardíacos o accidentes cerebrovasculares.

**ÍNDICE**

[**OBJETIVO DEL ESTUDIANTE (STUDENT OUTCOME) 4**](#_heading=h.gjdgxs)

[**INTRODUCCIÓN 5**](#_heading=h.30j0zll)

[**CAPÍTULO 1: ENTENDIMIENTO DEL NEGOCIO 6**](#_heading=h.1fob9te)

[1.1. DEFINIR EL PROBLEMA 6](#_heading=h.3znysh7)

[1.2. EVALUAR Y ANALIZAR ESCENARIOS 7](#_heading=h.2et92p0)

[1.3. DEFINIR LOS OBJETIVOS DE ML 7](#_heading=h.tyjcwt)

[1.4. PLAN DE PROYECTO 8](#_heading=h.3dy6vkm)

[**CAPÍTULO 2: PRE-PROCESAMIENTO DE LOS DATOS 12**](#_heading=h.1t3h5sf)

[2.1: COLECTAR LOS DATOS 12](#_heading=h.4d34og8)

[2.2 CALIDAD Y LIMPIEZA DE DATOS 14](#_heading=h.2s8eyo1)

[2.3 EDA 23](#_heading=h.17dp8vu)

[2.4 TRANSFORMACIÓN DE LOS DATOS 31](#_heading=h.843vasph268d)

[**CAPÍTULO 3: RESULTADOS SOBRE ANÁLISIS DE DATOS 37**](#_heading=h.3rdcrjn)

[3.1 INSIGHTS 37](#_heading=h.26in1rg)

[**CAPÍTULO 4: RESULTADOS MODELIZACIÓN Y OPTIMIZACIÓN 42**](#_heading=h.lnxbz9)

[4.1. MODELIZACIÓN 42](#_heading=h.eagt71gsytqp)

[4.2. OPTIMIZACIÓN 57](#_heading=h.8wf1z05yki9r)

[**CAPÍTULO 5: RESULTADOS DE MODELIZACIÓN 59**](#_heading=h.y3gr5a6ovy20)

[5.1. PRESENTACIÓN DE RESULTADOS FINAL 59](#_heading=h.t9fxhfb6i73c)

[**VISUALIZACIÓN DEL DASHBOARD HEART DISEASE POWER BI………………61**](#_VISUALIZACIÓN_DEL_DASHBOARD)

[**APORTES…………………………………………………………………………………...63**](#_APORTES)

[**CONCLUSIONES 64**](#_heading=h.35nkun2)

[**RECOMENDACIONES……………………………………………………………………65**](#_RECOMENDACIONES)

[**GLOSARIO…………………………………………………………………………………66**](#_GLOSARIO)

**[BIBLIOGRAFÍA 67](#_BIBLIOGRAFÍA)**

[**ANEXOS 67**](#_heading=h.44sinio)

### **OBJETIVO DEL ESTUDIANTE (STUDENT OUTCOME)**

**ABET - EAC - Student Outcome 6:** La capacidad para comprender y brindar soporte para el uso, entrega y gestión de sistemas de información dentro de un entorno de sistemas de información.

Para alcanzar este objetivo, entendemos y gestionamos los sistemas de información utilizando el dataset que hemos investigado previamente. De esta manera, organizamos adecuadamente su información. Además, ofreceremos soporte basado en nuestro conocimiento en este ámbito.

**ABET - CAC - Student Outcome 7:** La capacidad de adquirir y aplicar nuevos conocimientos según sea necesario, utilizando estrategias de aprendizaje apropiadas.

Para alcanzar este objetivo, utilizamos machine learning en el desarrollo del proyecto, lo que nos permite gestionar de manera eficaz nuestro dataset y su información. También entendemos las necesidades específicas de la empresa seleccionada, en este caso Netflix, y cómo nuestros conocimientos pueden ser de ayuda.

### **INTRODUCCIÓN**

La Clínica Mayo, reconocida por su excelencia en el cuidado de la salud y la investigación médica, se enfrenta a un desafío crítico en el ámbito de la detección y manejo de las enfermedades cardiovasculares (ECV). Con el contexto global de las ECV como principal causa de mortalidad, cobrando 17,9 millones de vidas cada año y representando el 31% de todas las muertes en todo el mundo, surge una necesidad urgente de abordar la detección y el manejo temprano de estas enfermedades.

El problema radica en la identificación precoz de los factores de riesgo y la implementación de estrategias efectivas para prevenir y tratar las ECV. A pesar de los avances en la investigación y el tratamiento, la carga de morbilidad y mortalidad asociada con las ECV sigue siendo significativa, con un tercio de las muertes prematuras ocurriendo en personas menores de 70 años.

En este contexto, se empleará el análisis de datos y técnicas de machine learning para abordar este desafío. Utilizando un conjunto de datos que proporciona una visión detallada de las características clínicas y biomédicas relevantes para la predicción de enfermedades cardíacas, se buscará identificar patrones y correlaciones significativas. El objetivo es desarrollar modelos predictivos precisos y herramientas de apoyo a la toma de decisiones clínicas que contribuyan al avance de la detección precoz y el manejo efectivo de las enfermedades cardiovasculares; Con el respaldo y la colaboración de la Clínica Mayo, reconocida por su liderazgo en el cuidado de la salud cardiovascular, este proyecto tiene como meta mejorar los resultados para los pacientes y reducir la carga de enfermedad cardiovascular a nivel global.

### **CAPÍTULO 1: ENTENDIMIENTO DEL NEGOCIO**

#### **1.1. DEFINIR EL PROBLEMA**

El principal desafío de negocio para la Clínica Mayo se centra en la necesidad de mejorar la detección y el manejo temprano de las enfermedades cardiovasculares (ECV). A pesar de su reconocida reputación en el cuidado de la salud cardiovascular, la clínica enfrenta dificultades para identificar de manera precoz los factores de riesgo y adoptar medidas preventivas efectivas contra las ECV. Esto se traduce en una limitada capacidad para anticipar y gestionar de manera proactiva las enfermedades cardiovasculares entre su población de pacientes.

Para abordar este desafío, es fundamental desarrollar e implementar soluciones innovadoras que mejoren la detección precoz, la gestión de los factores de riesgo y la adopción de medidas preventivas para las enfermedades cardiovasculares. Estas soluciones deben aprovechar tecnologías avanzadas, como el análisis de datos y el machine learning, para identificar patrones y correlaciones significativas en los datos clínicos de los pacientes.

**Figura 1**

*Imagen sobre fuente de prevenir enfermedades cardiovasculares es una decisión de vida*

*Nota: Se puede ver la explicación de lo importante que es la prevención de las enfermedades cardiovasculares*

#### **1.2. EVALUAR Y ANALIZAR ESCENARIOS**

Para el desarrollo del proyecto en la Clínica Mayo, se empleará un enfoque basado en el análisis predictivo utilizando técnicas de Machine Learning. El objetivo es determinar el trayecto de detección y manejo temprano de las enfermedades cardiovasculares (ECV) en los pacientes, evaluando cómo varía esta detección a lo largo del tiempo.

Para la implementación de Machine Learning, se utilizará Miniconda como gestor de paquetes de Python y Jupyter para llevar a cabo las tareas de extracción, transformación y carga de datos (ETL). Estas herramientas proporcionarán un entorno eficiente y flexible para el análisis de datos y el desarrollo de modelos predictivos en el contexto de la Clínica Mayo.

#### **1.3. DEFINIR LOS OBJETIVOS DE ML**

* Desarrollar un modelo predictivo que pueda identificar tempranamente la probabilidad de desarrollo de enfermedades cardiovasculares en pacientes, utilizando datos clínicos y biomédicos relevantes
* Diseñar algoritmos capaces de predecir la progresión de las enfermedades cardiovasculares en pacientes a lo largo del tiempo, considerando factores de riesgo y variables clínicas.
* Facilitar a los profesionales de la salud en la Clínica Mayo la identificación de patrones y correlaciones significativas entre las variables clínicas y biomédicas, para mejorar la toma de decisiones clínicas y el diseño de intervenciones preventivas.
* Desarrollar herramientas de apoyo a la toma de decisiones que permitan personalizar el manejo y tratamiento de los pacientes con riesgo de enfermedades cardiovasculares, optimizando así los recursos y mejorando los resultados clínicos.

#### **1.4. PLAN DE PROYECTO**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **EDTEDT** | **Nombre de tarea** | **Duración** | **Comienzo** | **Fin** | **Nombre de los recursos** |
| 1 | 1 | Trabajo Final | 48 horas | Viernes  22/03/2024 | Jueves  27/06/2024 | Grupo 02 |
| 2 | 1.1 | Capítulo 1: Entendimiento del negocio | 7 horas | Viernes  22/03/2024 | Jueves  27/06/2024 |  |
| 3 | 1.1.1 | Definir el problema | 2 horas | Viernes  22/03/2024 | Jueves  27/06/2024 | André/Valeria/Luis/Gianella/Neil |
| 4 | 1.2 | Evaluar y analizar escenarios | 1 hora y 30 minutos | Viernes  22/03/2024 | Jueves  27/06/2024 | André/Valeria/Luis/Gianella/Neil |
| 5 | 1.3 | Definir los objetivos de ML | 1 hora y 30 minutos | Viernes  22/03/2024 | Jueves  27/06/2024 | André/Valeria/Luis/Gianella/Neil |
| 6 | 1.4 | Plan de proyecto | 2 horas | Viernes  22/03/2024 | Jueves  27/06/2024 | André/Valeria/Luis/Gianella/Neil |
| 7 | 2 | Capítulo 2: Pre-Procesamiento de los Datos | 16 horas | Miércoles  10/04/2024 | Jueves  27/06/2024 |  |
| 8 | 2.1 | Colocar los datos | 2 horas | Miércoles  10/04/2024 | Jueves  27/06/2024 | André Nicolas Vásquez Castro |
| 9 | 2.2 | Calidad y Limpieza de datos | 6 horas | Miércoles  10/04/2024 | Jueves  27/06/2024 | André Nicolas Vásquez Castro |
| 10 | 2.3 | EDA | 4 horas | Sábado  20/04/2024 | Jueves  27/06/2024 | Luis Bendezú/Neil Trujillo |
| 11 | 2.4 | Transformación de los datos | 4 horas | Viernes  03/05/2024 | Jueves  27/06/2024 | Luis Bendezú/Neil Trujillo |
| 12 | 3 | Capítulo 3: Resultados sobre análisis de datos | 4 horas | Sábado  20/04/2024 | Jueves  27/06/2024 |  |
| 13 | 3.1 | Insights | 4 horas | Sábado  20/04/2024 | Jueves  27/06/2024 | Valeria Caqui/Gianella Silvina/Neil Trujillo/Luis Bendezú |
| 14 | 4 | Capítulo 4: Modelización y Optimización | 15 horas | Sábado  11/05/2024 | Jueves  27/06/2024 |  |
| 15 | 4.1 | Modelización | 8 horas | Sábado  11/05/2024 | Jueves  27/06/2024 | André Nicolas Vásquez Castro |
| 16 | 4.2 | Optimización | 7 horas | Viernes  17/05/2024 | Jueves  27/06/2024 | Valeria Milagros Caqui Pizarro/Gianella Silvina |
| 17 | 5 | Capítulo 5: Resultados de modelización | 6 horas | 10/06/2024 | Jueves  27/06/2024 |  |
| 18 | 5.1 | Presentación de resultados finales | 6 horas | 10/06/2024 | Jueves  27/06/2024 | Luis Ángel Bendezú Jimenez |
| 19 | 6 | Aportes | 2 horas | 15/06/2024 | Jueves  27/06/2024 | André Nicolas Vásquez Castro |
| 20 | 7 | Conclusiones | 2 horas | 18/06/2024 | Jueves  27/06/2024 | Valeria Milagros Caqui Pizarro |
| 21 | 8 | Recomendaciones | 2 horas | 23/06/2024 | Jueves  27/06/2024 | Neil Trujillo |
| 22 | 9 | Glosario | 1 hora | 24/06/2024 | Jueves  27/06/2024 | Luis Ángel Bendezú Jimenez |
| 23 | 10 | Bibliografía | 2 horas | 25/06/2024 | Jueves  27/06/2024 | Gianella Silvina |

### **CAPÍTULO 2: PRE-PROCESAMIENTO DE LOS DATOS**

#### **2.1: COLECTAR LOS DATOS**

**PRE-PROCESAMIENTO DE LOS DATOS**

Para poder recolectar los datos necesarios, hemos tomado un conjunto de datos o también llamado dataset con la información de varias variables necesarias para su análisis del año 2023, estas variables son diagnósticos médicos, factores biomédicos, entre otros.

**Conjunto de datos sobre enfermedades cardiacas**

Durante nuestro proceso de investigación, hemos identificado un valioso conjunto de datos alojado en Kaggle que contiene información detallada sobre enfermedades cardíacas. Este dataset, recopilado en el año 2023, proporciona una amplia gama de variables relevantes para el estudio y la predicción de enfermedades cardiovasculares. Entre los datos incluidos, como habíamos dicho antes, se encuentran diagnósticos médicos, factores biomédicos y clínicos, así como información demográfica de los pacientes. La disponibilidad de este conjunto de datos nos brinda la oportunidad de realizar un análisis profundo y detallado, así como de aplicar técnicas de aprendizaje automático para desarrollar modelos predictivos precisos.

* **CARGA DE DATOS:**

Importamos las bibliotecas y cargamos el drive:

**Texto

Descripción generada automáticamente**

Después hacemos la carga de datos del dataset correspondiente:Captura de pantalla de computadora

Descripción generada automáticamente

#### **2.2 CALIDAD Y LIMPIEZA DE DATOS**

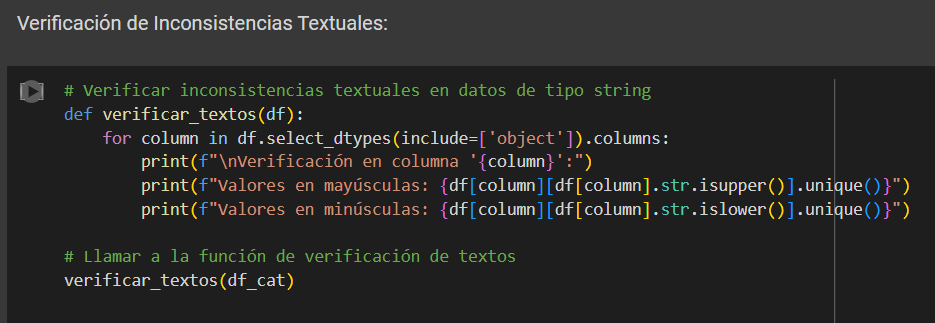
En esta parte del informe, vamos a aplicar la limpieza de datos.

* **CORRECCIÓN DE CABECERAS O HEADERS**

Vamos a empezar con la limpieza de datos, primero la corrección de cabeceras:Captura de pantalla de un celular

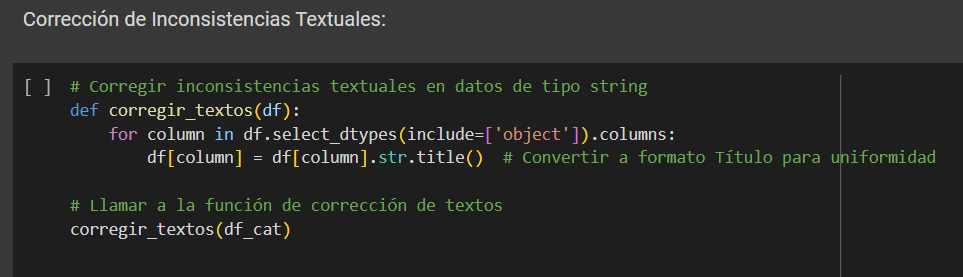
Descripción generada automáticamente

* **VERIFICACIÓN DE INCONSISTENCIAS TEXTUALES**

Primero, verificamos las inconsistencias textuales del dataset:

Texto

Descripción generada automáticamente

Y ahora, hacemos las correcciones de las inconsistencias textuales:

Como se puede ver, ya no hay inconsistencias textuales:Texto

Descripción generada automáticamente

* **IDENTIFICACIÓN DE CELDAS DUPLICADOS**

Texto

Descripción generada automáticamenteAhora vamos a identificar las celdas duplicadas, y borraremos sus filas si se presentan:

Ahora, eliminamos las celdas duplicadas:

**Texto

Descripción generada automáticamente**

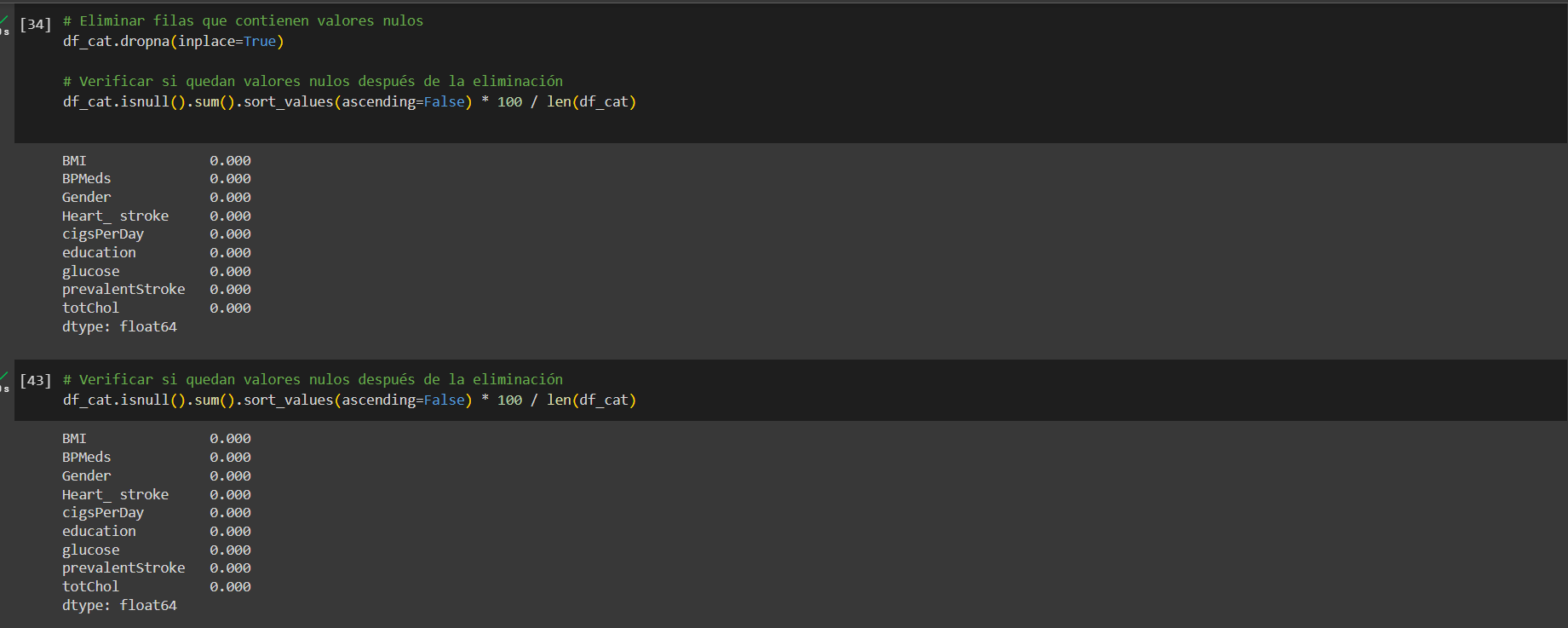
* **VERIFICACIÓN DE VALORES NULOS**

Pantalla de juego de computadora

Descripción generada automáticamente con confianza mediaVamos a verificar y eliminar los valores nulos del dataset, primero, identificamos

Con porcentajes y saber la cantidad de nulos en cada columna de las variables:



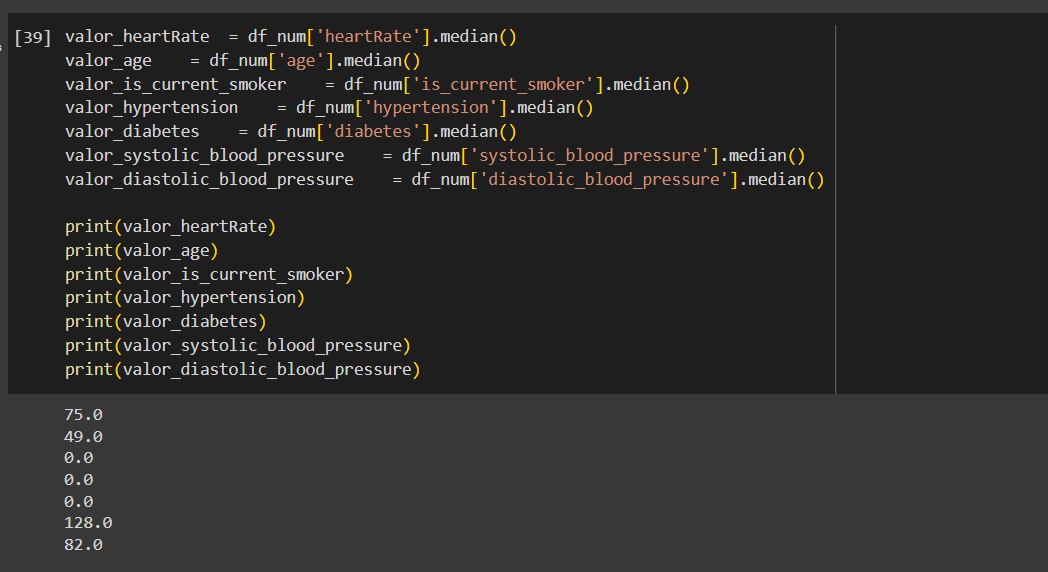
Ahora, eliminamos esos valores nulos y verifica

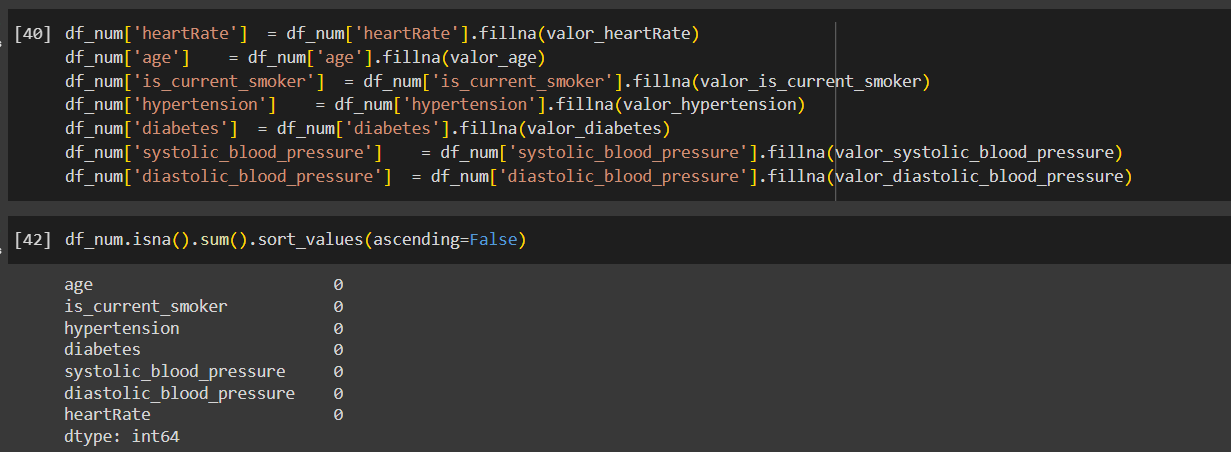
Ahora verificamos y eliminamos de las variables numéricas, primero, identificamos: Texto

Descripción generada automáticamente

Pantalla de computadora con letras

Descripción generada automáticamente con confianza media



Ahora, eliminamos los valores nulos de las variables numéricas:

Texto

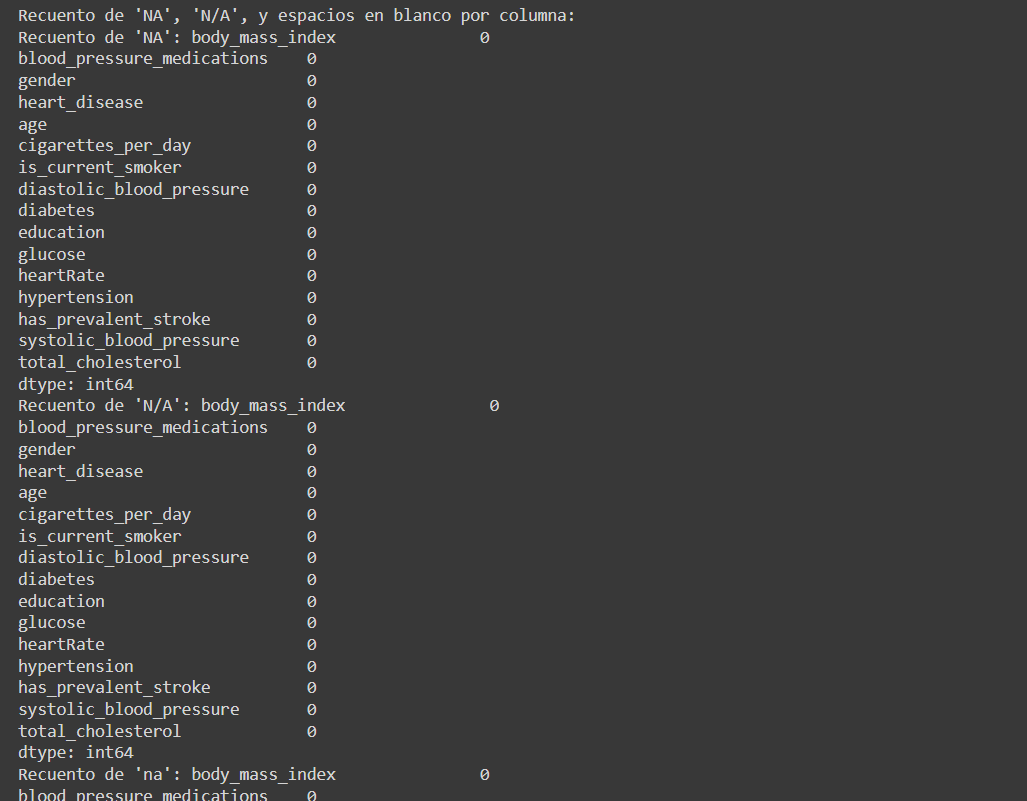
Descripción generada automáticamenteTexto

Descripción generada automáticamenteAquí tenemos otro procedimiento adicional para la verificación de valores nulos y variantes para categóricas y numéricas:

Y con esto se hace la corrección de esos valores nulo y variantes para categóricas y numéricas también:  
Texto

Descripción generada automáticamente

Como se puede ver, ya no hay valores nulos:



Pantalla de un video juego

Descripción generada automáticamente con confianza mediaEjecutamos este código también:

Pantalla de un computador

Descripción generada automáticamente con confianza mediaY por último, grabamos el fin de fase:

#### **2.3 EDA**

Realizamos un análisis exploratorio de datos para examinar los datos obtenidos de la Clínica Mayo con el fin de comprender la estructura y las características que presentan estos mismos. El EDA es importante en este trabajo, ya que nos permite conocer información valiosa que puede guiarnos en la selección de modelos y técnicas de análisis de datos.

1. **¿Cuál es la distribución de la edad de los pacientes en el conjunto de datos?**

La distribución de la edad de los pacientes en nuestro conjunto de datos proporciona una visión general de cómo se distribuyen las edades dentro de nuestra muestra. Este análisis nos permite comprender mejor la composición demográfica de nuestra población de pacientes. Al observar la distribución de la edad, podemos identificar tendencias.

Gráfico, Histograma

Descripción generada automáticamente

1. **¿Cuál es la distribución del índice de masa corporal (BMI) en la población?**

La distribución del índice de masa corporal (BMI) en nuestra población nos brinda información sobre la variabilidad del peso corporal en nuestra muestra. El BMI es una medida importante que nos ayuda a evaluar el estado nutricional y el riesgo de enfermedades relacionadas con el peso. Al examinar su distribución, podemos identificar patrones, como la concentración de valores en rangos específico

Gráfico

Descripción generada automáticamente

1. **¿Cómo se distribuyen los niveles de presión arterial sistólica y diastólica?**

La distribución de los niveles de presión arterial sistólica y diastólica en nuestra población proporciona información crucial sobre la salud cardiovascular de nuestros pacientes. Estos dos valores son fundamentales para evaluar la función cardíaca y el riesgo de enfermedades cardiovasculares. Al analizar su distribución, podemos identificar la gama de valores en los que se encuentran los pacientes, así como posibles anomalías o tendencias

Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

1. **¿Hay alguna correlación entre el género de los pacientes y la prevalencia de enfermedades cardiovasculares?**

La correlación entre el género de los pacientes y la prevalencia de enfermedades cardiovasculares es un área de investigación crucial en la salud cardiovascular. Este análisis busca explorar si existe alguna relación entre el género de los pacientes y la probabilidad de desarrollar enfermedades cardiovasculares, como ataques cardíacos o accidentes cerebrovasculares

Texto

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

1. **¿Qué género tiene mayores niveles de glucosa en la sangre?**

DISTRIBUCIÓN DE LOS NIVELES DE GLUCOSA

En este gráfico de histogramas se aprecia los niveles de glucosa que existe entre hombres y mujeres.

Imagen que contiene Gráfico

Descripción generada automáticamente

1. **¿Cuántos cigarrillos consumidos por día afectan el nivel acelerado del ritmo cardíaco?**

DISTRIBUCIÓN DE RITMOS CARDÍACOS

En este gráfico de histogramas se aprecia la distribución de ritmos cardíacos que tienen los pacientes en donde mayormente predominan los niveles en un intervalo de 60 a 100.

Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

RITMO CARDÍACO SEGÚN INTERVALOS DE CIGARRILLOS POR DÍA

Analizando los datos, observamos que las diferentes capas de color determinan si hay una relación entre la cantidad de cigarrillos fumados y el ritmo cardíaco. Por ejemplo, si las capas de colores más oscuros (que representan un mayor consumo de cigarrillos) tienden a aparecer más en el lado derecho del histograma, esto podría indicar un ritmo cardíaco más alto asociado con un mayor consumo de cigarrillos, lo cual conlleva a una posible situación donde exista una enfermedad cardiovascular.

Texto

Descripción generada automáticamente

**Gráfico, Histograma

Descripción generada automáticamente**

MATRIZ DE CORRELACIÓN

La matriz de correlación revela relaciones significativas entre varias variables de salud. Observamos que la presión arterial sistólica tiene una correlación positiva fuerte con la presión arterial diastólica (0.78) y con la hipertensión (0.69), indicando que a medida que aumenta la presión arterial diastólica, también lo hace la sistólica y la presencia de hipertensión. La edad también muestra una correlación positiva moderada con la hipertensión (0.30) y la presión arterial sistólica (0.39), sugiriendo que la presión arterial tiende a aumentar con la edad. Por otro lado, el hecho de ser fumador actual tiene una correlación negativa con la edad (-0.21), lo que podría indicar que las personas mayores tienden a fumar menos. También se observa que el ritmo cardíaco tiene una correlación baja con la mayoría de las variables, lo que sugiere que puede estar influenciado por factores no incluidos en esta matriz.

Tabla

Descripción generada automáticamente con confianza baja

#### **2.4 TRANSFORMACIÓN DE LOS DATOS**

Primero importamos las librerías, conectamos con el drive, y dividimos las variables categóricas y numéricas, primero con las categóricas.

Texto

Descripción generada automáticamente

Captura de pantalla de computadora

Descripción generada automáticamente

Vamos a convertir las variables categóricas en variables oneHotEncoder, en la imagen de abajo, mencionamos a la columna llamada gender para poder hacer la transformación.

Interfaz de usuario gráfica

Descripción generada automáticamente

Ahora, una vez hecho ese paso, ahora si podemos hacer la transformación de esa variable a oneHotEncoder.

Texto

Descripción generada automáticamente

Repetimos los mismos pasos para los siguientes procedimientos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ahora podemos hacer la transformación de la variable a OneHotEnconder.

Texto

Descripción generada automáticamente

Hacemos la misma transformación de oneHotEncoder a la variable has\_prevalent\_stroke.

Texto

Descripción generada automáticamente

Lo mismo para la variable dependiente heart\_disease.

Texto

Descripción generada automáticamente

Ahora, vamos con las variables numéricas, vamos a identificarlas.

Pantalla de un video juego

Descripción generada automáticamente con confianza media

Ahora, realizamos la transformación de las variables numércias a MinMaxScaler,

Captura de pantalla de computadora

Descripción generada automáticamente

Ahora vamos a hacer la integración de dataframes, que va a imprimir los dataframes que contienen las variables categóricas y númericas.

Texto

Descripción generada automáticamente con confianza media

Ahora, vamos a concatenar los dataframes.



Pantalla de computadora con números

Descripción generada automáticamente con confianza mediaMostramos todas las variables ya transformadas en un nuevo dataframe.

Por último, haremos el paso de serialización, de esa forma permite poder almacenar y recuperar los datos del dataframe en su estado original.



### **CAPÍTULO 3: RESULTADOS SOBRE ANÁLISIS DE DATOS**

#### **3.1 INSIGHTS**

1. **Relación de tendencia de la cantidad de cigarrillos consumidos por día y ritmo cardíaco**

Este gráfico sugiere que hay una relación positiva entre la cantidad de cigarrillos fumados y el ritmo cardíaco, lo que significa que a medida que la cantidad de cigarrillos fumados por día aumenta, también lo hace el ritmo cardíaco promedio.

**Gráfico, Gráfico de líneas

Descripción generada automáticamenteTexto

Descripción generada automáticamente**

1. **Distribución de las edades de las personas afectadas por enfermedades cardiovasculares**

En el gráfico que muestra la distribución de edades de las personas afectadas por enfermedades cardiovasculares en nuestro dataset, se destaca que la edad más común entre los afectados es de 60 años, mientras que la edad menos frecuente es de 36 años. Esta representación visual nos permite identificar claramente las edades con mayor incidencia de enfermedades cardiovasculares en nuestra muestra de datos. Al analizar detenidamente la distribución, se observa que la prevalencia de estas enfermedades tiende a aumentar con la edad, siendo más frecuentes en individuos de mayor edad. Este patrón sugiere una posible correlación entre la edad y la aparición de enfermedades cardiovasculares, destacando la importancia de la prevención y el cuidado de la salud cardiovascular en la población de mayor edad.

Gráfico, Histograma

Descripción generada automáticamente

1. **Relación de los tipos de presiones arteriales y el Índice de Masa**

Explorar la relación entre los diferentes tipos de presión arterial, tanto sistólica como diastólica, y el índice de masa corporal (BMI) nos permite comprender cómo la distribución del peso corporal está asociada con la salud cardiovascular. El BMI es una medida comúnmente utilizada para evaluar el peso relativo en relación con la altura de una persona, mientras que la presión arterial es un indicador crítico de la salud del sistema circulatorio. Al examinar esta relación, podemos identificar si existe alguna correlación entre el peso corporal y los niveles de presión arterial

Gráfico

Descripción generada automáticamente con confianza media

1. **Relación entre Enfermedad Cardíaca y Ritmo Cardíaco**

Se observa una tendencia general de aumento de la presión arterial con el aumento del IMC. Esto significa que las personas con un IMC más alto tienden a tener una presión arterial más alta. La tendencia es más pronunciada en la presión arterial sistólica (la presión cuando el corazón late) que en la presión arterial diastólica (la presión cuando el corazón está en reposo).

**Gráfico, Gráfico de dispersión

Descripción generada automáticamente**

### **CAPÍTULO 4: RESULTADOS MODELIZACIÓN Y OPTIMIZACIÓN**

#### **4.1. MODELIZACIÓN**

Primero, debemos comparar dos modelos para ver quién es el que obtiene los mejores puntajes en las métricas que se obtienen, de ahí, vamos a elegir al modelo que tiene mayores puntajes que el otro modelo, lo optimizamos y presentamos su resultados finales de ese modelo ya optimizado.

Vamos a empezar con la “Modelización Random Forest con todas las variables”.

Pantalla de computadora

Descripción generada automáticamente con confianza media

Se define las variables independientes y la variable dependiente (target) llamado heart\_disease.

Pantalla de juego de computadora

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente

Ahora vamos a importar la librería “train\_test\_split” para poder hacer las variables de entrenamiento y testeo.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Importamos el modelo “RandomForestClassifier” para poder instanciarlo y entrenarlo.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ahora, debemos predecir, importar las métricas, ejecutar la matriz de confusión y ver los resultados de las métricas.

Captura de pantalla de un celular

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Sitio web

Descripción generada automáticamente

Aquí podemos ver los puntajes de las métricas obtenidas tras ejecutar el “Modelo de Random Forest con todas las variables”.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Texto

Descripción generada automáticamenteAhora, vamos a generar los árboles de decisión usando “Random Forest”, nos va a decir que número de árboles de decisión se va a generar, las variables usadas, definimos el índice del árbol, importamos la librería “tree” y vamos a visualizar todos los árboles de decisión generados (La primera imagen de los árboles de decisión no se podrá visualizar bien por la cantidad generada, pero mostraremos más de cerca en las siguientes imágenes después de la primera).

Imagen que contiene mapa, texto, agua, grande

Descripción generada automáticamente

Texto

Descripción generada automáticamente

Diagrama, Escala de tiempo

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Después de hacer todo el procedimiento necesario para este modelo, hemos considerado lo más importante para comenzar con la comparación, este cuadro de métricas con matriz de confusión.

Gráfico

Descripción generada automáticamente con confianza media

Ahora, vamos con el segundo modelo llamado “Random Forest con Feature Selection – Técnica de Random Forest”.

Antes de empezar con el modelo, vamos a hacer el procedimiento de “Feature Selection – Técnica de Random Forest”. Vamos a importar las librerías necesarias, montar el drive y leer el pickle generado de transformación de datos.

Captura de pantalla de computadora

Descripción generada automáticamente

Hacemos lo mismo como en el modelo anterior, separamos las variables independientes con la variable dependiente que es heart\_disease, separamos variables de entrenamiento, testeo y definimos “RandomForest Classifier”.

Texto

Descripción generada automáticamente

Y obtenemos las variables que nos interesa, las más resaltantes, las demás que no son mencionadas se eliminan, todo eso presentado en la segunda imagen.

Texto

Descripción generada automáticamente

En esta imagen, se ve las variables que queremos y las más resaltantes, todo que no está aquí, lo eliminamos con un drop.

Texto

Descripción generada automáticamente

Una vez que ya tenemos las variables que queremos, ahora sí, podemos continuar con el siguiente modelo llamado “Random Forest con Feature Selection – Técnica de Random Forest”.

Importamos las librerías necesarias, cargamos el pickle de la transformación de datos.

Captura de pantalla de computadora

Descripción generada automáticamente

Ahora, borramos las variables que no nos va a servir o las que no salieron en el Feature Selection que hicimos anteriormente, después separamos variables independientes con la variable dependiente heart\_disease.

Una captura de pantalla de un celular

Descripción generada automáticamente con confianza media

Texto

Descripción generada automáticamente

Definimos las variables de entrenamiento, de testeo, importamos el modelo, lo instanciamos, lo entrenamos.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ahora, vamos a predecir usando los datos de testeo, importamos las métricas, la matriz de confusión, y vamos a ver los resultados.

Captura de pantalla de un celular

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Ahora, vamos a ver los resultados de las métricas, se puede observar que se obtuvo un puntaje mayor al anterior modelo que ejecutamos.

Interfaz de usuario gráfica, Aplicación, Sitio web

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza baja

Texto

Descripción generada automáticamenteObservamos los árboles de decisión generados, número de variables usadas (son siete, las más resaltantes), definimos el árbol en índice 5.

Ahora, vamos a ver todos los árboles generados.

Captura de pantalla de computadora

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente con confianza media

Diagrama, Escala de tiempo

Descripción generada automáticamente

Diagrama

Descripción generada automáticamente

Imagen que contiene firmar, mapa, caja, tabla

Descripción generada automáticamente

Escala de tiempo

Descripción generada automáticamente con confianza bajaComparando los modelos, con los puntajes obtenidos y las métricas más importantes que son el F1-score y AUC, se concluye en la comparación que el modelo “Random Forest con Feature Selection – Técnica de Random Forest” es el ganador y el que utilizaremos para los siguientes procedimientos que es la optimización y los nuevos resultados que saldrá después de realizar eso.

Por lo que guardaremos la fase del modelo de “Random Forest con Feature Selection – Técnica de Random Forest”.



#### **4.2. OPTIMIZACIÓN**

Ahora, vamos a optimizar el modelo ganador que es el “Random Forest con Feature Selection – Técnica de Random Forest”, de esa forma obtendremos mejores resultados de las métricas y de la matriz de confusión, importamos librerías, montamos drive y cargamos el pickle del modelo de “Random Forest con Feature Selection – Técnica de Random Forest”.

Captura de pantalla de computadora

Descripción generada automáticamente

Separamos variables independientes con la variable dependiente heart disease, importamos “train\_test\_split”, definimos las variables de entrenamiento y testeo, importamos “RandomForestClassifier” y le damos a una variable algoritmo\_rf.

Texto

Descripción generada automáticamente

Ahora, vamos a usar la técnica de Grid Search para poder buscar sistemáticamente a través de un conjunto predefinido de hiperparámetros para encontrar la combinación óptima que maximiza el rendimiento del modelo ganador que hemos dicho.

Interfaz de usuario gráfica

Descripción generada automáticamente

Texto

Descripción generada automáticamenteUna vez hecho el entrenamiento, nos va a dar los mejores parámetros que tenemos que definir a nuestro modelo ganador.

Y este es el mejor resultado de la métrica AUC.

Captura de pantalla de un celular

Descripción generada automáticamente

### **CAPÍTULO 5: RESULTADOS DE MODELIZACIÓN**

#### **5.1. PRESENTACIÓN DE RESULTADOS FINAL**

Ahora, vamos a entrenar nuestro modelo ganador con los mejores parámetros que nos ha dado la técnica Grid Search.

Captura de pantalla de un celular

Descripción generada automáticamente

Veremos los nuevos resultados de las métricas con la matriz de confusión.

Texto

Descripción generada automáticamente

Aquí están los resultados finales optimizados.

Captura de pantalla de un celular

Descripción generada automáticamente

Imagen que contiene Escala de tiempo

Descripción generada automáticamente

En este trabajo, hemos abordado la necesidad crucial de la detección y manejo temprano de enfermedades cardiovasculares mediante el uso de técnicas avanzadas de análisis de datos. Utilizando un conjunto de datos relevantes, como ultimas optimizaciones aplicamos el algoritmo de Random Forest en tres etapas distintas: inicialmente con todas las variables, luego con una selección de características y, finalmente, optimizando los hiperparámetros a través de Grid Search. En la primera fase, el modelo con todas las variables mostró una precisión (accuracy) de 0.860, con un F1-Score de 0.093, precisión de 0.500, recall de 0.052 y un AUC de 0.719. El algoritmo de Random Forest que funciona creando múltiples árboles de decisión durante el entrenamiento y emitiendo la predicción promedio de estos árboles. Esta técnica es robusta frente al sobreajuste y maneja bien tanto las variables categóricas como las numéricas. Sin embargo, incluir todas las variables puede introducir ruido y características irrelevantes, lo que puede afectar el rendimiento del modelo, como se observó en la baja sensibilidad (recall). Al aplicar la técnica de selección de características en la segunda fase, observamos una ligera mejora en las métricas de rendimiento. La selección de características implica identificar y utilizar solo las variables más relevantes para el modelo, eliminando aquellas que no aportan valor o que pueden introducir ruido. Este proceso puede mejorar el rendimiento del modelo al centrarse en las variables que tienen un impacto significativo en la predicción, lo que se reflejó en la mejora de las métricas observadas. Finalmente, la optimización de hiperparámetros mediante Grid Search llevó el modelo a su mejor desempeño. La precisión alcanzó 0.863, el F1-Score 0.128, la precisión 0.583 y el recall 0.072, con un AUC de 0.721.Optimizar estos parámetros permite encontrar la combinación que maximiza la capacidad predictiva del modelo. Este ajuste fino es crucial para maximizar el rendimiento del modelo, logrando un balance más óptimo entre precisión y recall.

En conclusión se ha demostrado la eficacia en la mejora de la detección temprana de enfermedades cardiovasculares. Cada etapa de mejora ha aportado incrementos significativos en las métricas subrayando la importancia de técnicas avanzadas en el desarrollo de modelos predictivos robustos y precisos. Para futuras investigaciones y aplicaciones prácticas, recomendamos implementar estos modelos en entornos clínicos para validar su utilidad en la detección temprana de enfermedades cardiovasculares, así como la validación de los conjuntos de datos externos para confirmar la generalización y robustez del modelo en diversas poblaciones, garantizando así su eficacia y fiabilidad en diferentes contextos clínicos. La combinación de estas estrategias permitirá desarrollar herramientas predictivas que realmente puedan transformar la atención médica y mejorar los resultados de salud de los pacientes.

### **VISUALIZACIÓN DEL DASHBOARD HEART DISEASE POWER BI**

El dashboard se basa en la búsqueda generalmente de las categorías del género, sea femenino o masculino, en ello veremos lo siguiente:

* **El recuento de cigarros por día:** Nos habla de cuantos cigarros se consume por educación, osea el total de graduados tanto femenino como masculino es de 572, si presionan la categoría de femenino saldrá que la mujeres graduadas que fuman son 378 y hombres de hombres son 194.
* **Top 5: Enfermedades por Edad:** No demuestra el top 5 de las edades con más porcentaje a tener una enfermedad (varia en tanto a mujer o varón).
* **Suma de diabetes:** Es la suma de la cantidad de personas que si tiene diabetes.
* **Fumador actual:** Lo mismo que el anterior, la suma de cuantas personas si son fumadores actuales.
* **Suma de Hipertensión por Heart disease:** Aquí nos explican cuántos si sufren de hipertensión y cuantos no (varia si es mujer o varón).
* **Top 5:** Índice de masa corporal por edad: Nos explica la suma de la masa corporal entre un top 5 de edades con más masa corporal.
* **Suma de medicamentos:** Nos explica la cantidad de personas, ya sea mujer o varón, en base a su educación, que toman medicamentos para su presión arterial.
* **Presión sistólica y diastólica:** Nos da un comparativo entre la suma de las presiones, donde las categorías son por la educación y el heart disease.
* **Suma de ritmo cardiaco:** suma de todos los ritmos, mostrados en porcentajes y separados en mujer y varón para saber quién tiene más probabilidad de heart disease.
* **Interfaz de usuario gráfica, Aplicación

  Descripción generada automáticamenteSuma de glucosa y colesterol:** suma de los totales de las cantidades de glucosa y colesterol, separados por genero al dar click en las categorías.

### **APORTES**

* **André Nicolas Vásquez Castro:** Encargado de la recolección y preprocesamiento de datos, así como de la implementación y optimización del modelo Random Forest.
* **Valeria Milagros Caqui Pizarro:** Responsable de la evaluación y análisis de escenarios, así como de la implementación de modelos predictivos.
* **Luis Ángel Bendezú Jiménez:** Contribuyó en el análisis exploratorio de datos (EDA) y en la transformación de variables categóricas y numéricas.
* **Neil Eduardo Trujillo Neyra:** Participó en la definición del problema, los objetivos de Machine Learning y en la generación de insights a partir de los datos.
* **Gianella Lucía Silvina Gonzales:** Colaboró en la presentación y visualización de los resultados finales, y en la elaboración del informe y conclusiones del proyecto.

Logramos colaborar de manera efectiva para realizar un análisis exhaustivo de los datos de la Clínica Mayo, aplicando técnicas avanzadas de Machine Learning para la predicción de enfermedades cardiovasculares. Juntos, desarrollamos un modelo robusto y preciso, optimizado mediante técnicas de selección de características y ajuste de hiperparámetros. Además, nos aseguramos de validar y comunicar claramente los resultados a través de visualizaciones y presentaciones detalladas, garantizando que los hallazgos fueran comprensibles y útiles para la toma de decisiones clínicas en la prevención y manejo de enfermedades cardiovasculares.

### **CONCLUSIONES**

* Es crucial tener una comprensión clara del problema y de los objetivos del negocio al inicio del proyecto. Esta claridad guiará todas las fases del análisis de datos, asegurando que cada paso tomado contribuya efectivamente hacia la solución del problema identificado.
* La preparación y limpieza de datos son fundamentales para garantizar la calidad del análisis. Esto incluye la corrección de errores, la eliminación de datos duplicados o irrelevantes, y la transformación adecuada de los datos para su análisis. Estos pasos incrementan la precisión de los insights generados y la fiabilidad de los modelos predictivos.
* Las visualizaciones son herramientas poderosas para comunicar hallazgos y tendencias en los datos. Utilizar gráficos adecuados puede facilitar la comprensión de complejidades en los datos y resaltar diferencias y similitudes críticas que pueden ser fundamentales para la toma de decisiones.
* Organizar y gestionar adecuadamente los archivos y las rutas de acceso en plataformas como Google Drive no solo mejora la eficiencia sino también la seguridad del acceso a los datos. Esto es vital en entornos colaborativos donde múltiples usuarios necesitan acceder y manipular los datos de manera concurrente.
* Verificar la veracidad y relevancia de los datasets es esencial. Trabajar con datos que son precisos y pertinentes al problema asegura que las conclusiones y predicciones del análisis sean válidas y aplicables en escenarios reales.

### **RECOMENDACIONES**

* Recomendamos realizar una comparación detallada entre el modelo Random Forest y otros modelos predictivos como Gradient Boosting y K-Nearest Neighbors (KNN). Esta comparación debe incluir métricas como precisión, recall, F1-score y área bajo la curva ROC (AUC). Los resultados deben mostrarse en gráficos para facilitar la visualización de las diferencias y la selección del mejor modelo.
* Se sugiere implementar técnicas de optimización de hiperparámetros, como Grid Search o Random Search, para cada modelo probado. Esto mejorará la precisión y robustez de los modelos predictivos, asegurando que se obtenga el mejor rendimiento posible.
* Es recomendable utilizar técnicas de visualización avanzada para presentar los resultados del análisis de datos y los modelos predictivos. Gráficos de dispersión, diagramas de correlación y mapas de calor pueden mostrar las relaciones entre variables y patrones identificados. Estas visualizaciones facilitarán la interpretación de los resultados y ayudarán en la toma de decisiones informadas sobre la prevención y manejo de enfermedades cardiovasculares.

### **GLOSARIO**

* **Conjunto de Datos (Dataset):** Colección de datos organizada, se utiliza para el análisis y modelado en proyectos de machine learning.
* **Aprendizaje Automático (Machine Learning):** Subcampo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y hacer predicciones o decisiones basadas en datos.
* **Índice de Masa Corporal (IMC):** Medida que relaciona el peso y la altura de una persona, utilizada para evaluar si una persona tiene un peso saludable.
* **Hipertensión:** Condición médica caracterizada por una presión arterial persistentemente alta, que es un factor de riesgo significativo para enfermedades cardiovasculares.
* **Modelo Predictivo:** Algoritmo o conjunto de reglas que se utilizan para predecir un resultado específico basado en datos de entrada. En el contexto de este proyecto, se refiere a modelos que predicen el riesgo de enfermedades cardiovasculares.
* **Random Forest:** Algoritmo de aprendizaje automático basado en la construcción de múltiples árboles de decisión, que utiliza el promedio de las predicciones de estos árboles para mejorar la precisión y evitar el sobreajuste.
* **Grid Search:** Método exhaustivo de búsqueda que evalúa todas las combinaciones posibles de un conjunto predefinido de hiperparámetros.
* **Área Bajo la Curva (AUC):** Métrica que mide la capacidad del modelo para diferenciar entre clases positivas y negativas. Un AUC cercano a 1 indica un modelo con excelente capacidad predictiva.
* **Precisión (Accuracy):** Porcentaje de predicciones correctas realizadas por el modelo.

### **BIBLIOGRAFÍA**

* MedicalNewsToday. (15 de diciembre del 2019). *Todo lo que debes saber acerca de las enfermedades del corazón.* Recuperado de <https://www.medicalnewstoday.com/articles/es/327293>.
* Cigna Healthcare. (12 de abril del 2021). *Síntomas y causas de enfermedad cardíaca.* Recuperado de <https://www.cigna.com/es-us/knowledge-center/heart-health>.
* RadiologyInfo.org. (25 de septiembre del 2022). *Detección temprana de enfermedades cardíacas (del corazón).* Recuperado de <https://www.radiologyinfo.org/es/info/screening-cardiac>.
* Quironsalud. (27 de junio del 2022). *Las pruebas para detectar enfermedades cardíacas, el electrocardiograma y ecocardiograma.* Recuperado de <https://www.quironsalud.com/blogs/es/corazon-salud/pruebas-detectar-enfermedades-cardiacas-electrocardiograma>.

### **ANEXOS**

* **DATASET:**

<https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset>

* **LINK DEL GOOGLE COLABORATE DE NUESTRO TRABAJO FINAL:**

<https://drive.google.com/drive/folders/1z3yPVZUWr6DPqd6k6WUS1kXuztowmX9l?usp=sharing>

* **DASHBOARD HEART DISEASE POWER BI**

