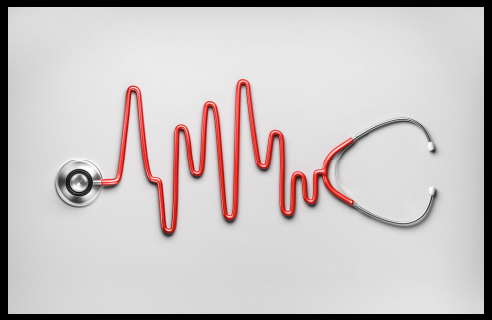

¿Podemos predecir pacientes con ataques cardíacos?
Proyecto Final Data Science
Pablo Pan Veira

Introducción al problema

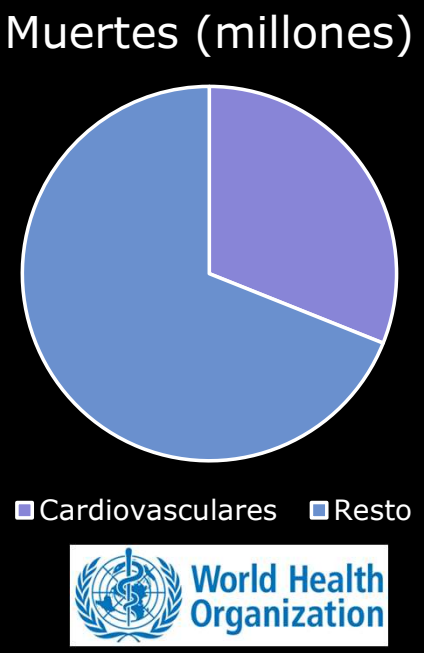


En el campo de la cardiología, la data science es una disciplina que puede ser utilizada para mejorar la detección temprana y el tratamiento de enfermedades cardiovasculares, que son una de las principales causas de muerte en todo el mundo.

En esta presentación, exploraremos el campo de la cardiología, y cómo esta disciplina puede ser utilizada para abordar algunos de los desafíos más importantes en el diagnóstico de afecciones cardíacas.

Contexto

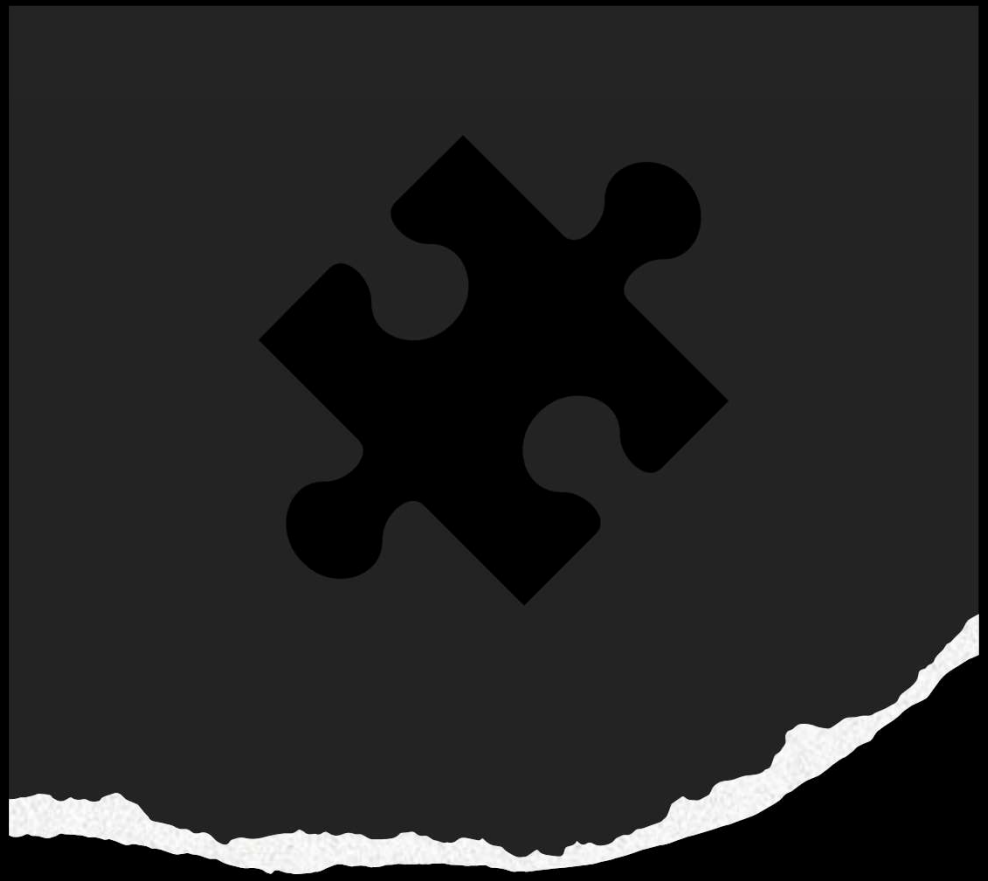
Según la OMS las enfermedades cardiovasculares (ECV) son la principal causa de muerte en todo el mundo. Los últimos datos arrojan que en 2015 fallecieron 17,7 millones de personas, un 31% de todas las muertes mundiales dicho año. Se requiere establecer si es posible determinar las probabilidades de sufrir un ataque cardíaco.



Objetivo

- Analizaremos un dataset e intentaremos dar respuesta a las siguientes preguntas:

- ¿Cuántos pacientes han tenido un ataque cardíaco?
- ¿Hay una diferencia significativa en la incidencia de ataques cardíacos entre hombres y mujeres?
- ¿Hay una relación entre la edad y la incidencia de ataques cardíacos?
- ¿Qué masa corporal tienen los pacientes?
- ¿Hay alguna relación entre la masa corporal, la glucosa y la posibilidad del ataque cardíaco?



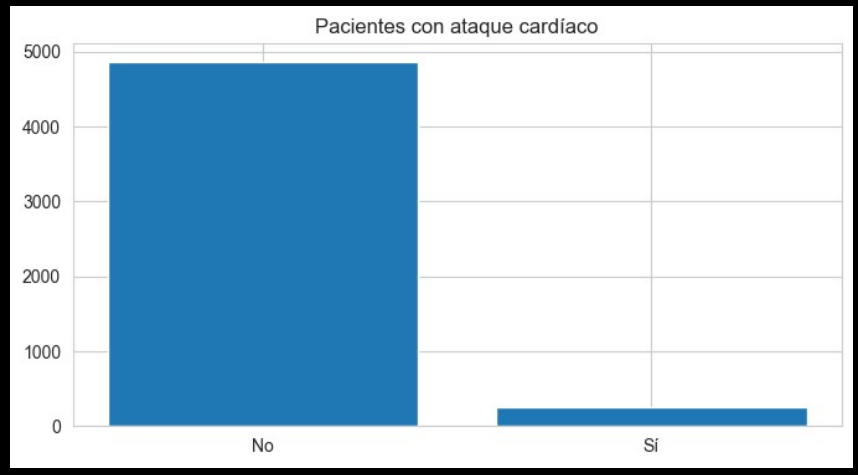
Datos

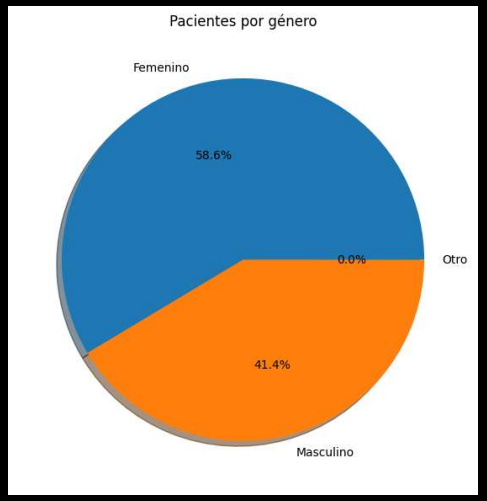
- Los datos se recogen del siguiente csv: healthcare-dataset-stroke-data.csv
- En este dataset se indican una serie de 5110 pacientes con ciertos valores y se indica si han sufrido un ataque cardíaco.

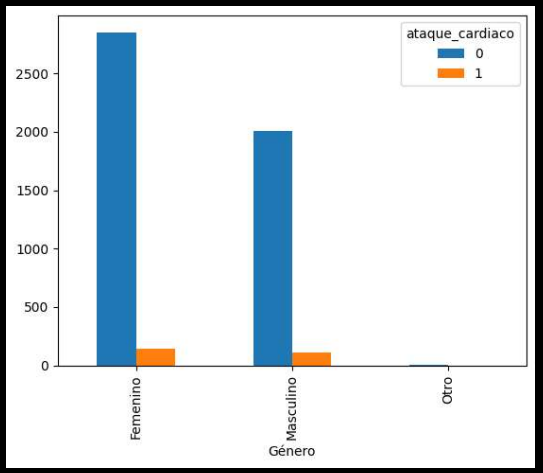
Columna	Tipo de variable	Ejemplo de valor	Descripción
id	Integer	9046	Identificador único del paciente
gender	String	Male	Género del paciente
age	Integer	67	Edad del paciente
hypertension	Integer	0	Indica si el paciente tiene hipertensión o no
heart_disease	Integer	1	Indica si el paciente tiene una enfermedad cardíaca o no
ever_married	String	Yes	Indica si el paciente está casado o no
work_type	String	Private	Indica el tipo de trabajo del paciente
Residence_type	String	Urban	Indica si el paciente vive en una zona urbana o rural
avg_glucose_level	Float	228.69	Nivel promedio de glucosa en sangre del paciente
bmi	Float	36.6	Índice de masa corporal del paciente
smoking_status	String	formerly smoked	Indica el estado de tabaquismo del paciente
stroke	Integer	1	Indica si el paciente tuvo un accidente cerebrovascular o no

Análisis Exploratorio de Datos (EDA)

- ¿Cuántos pacientes han tenido un ataque cardíaco?
- ¿Hay una diferencia significativa en la incidencia de ataques cardíacos entre hombres y mujeres?



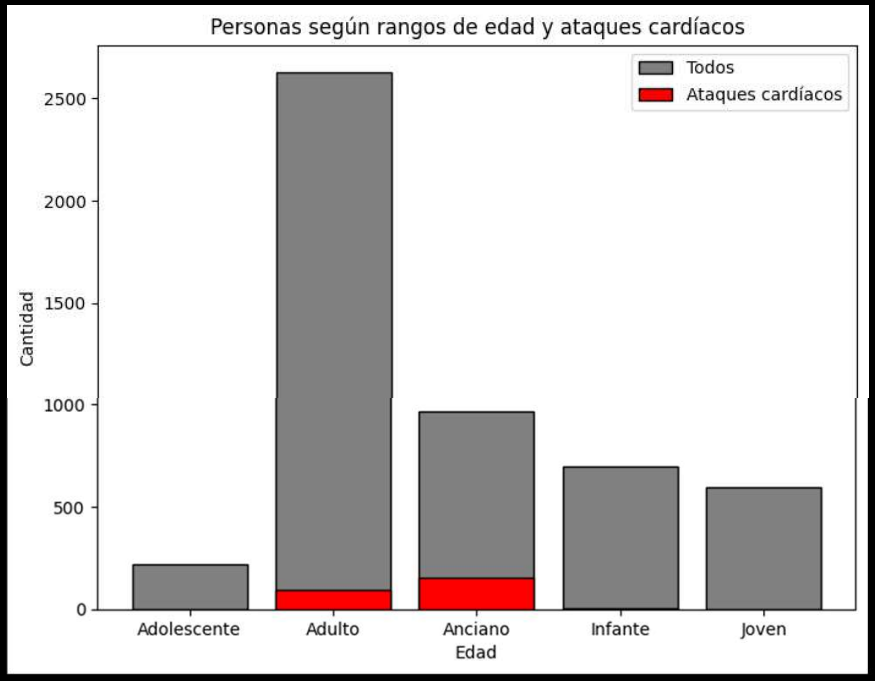




Análisis Exploratorio de Datos (EDA)

- ¿Cuántos pacientes han tenido un ataque cardíaco?
- ¿Hay una diferencia significativa en la incidencia de ataques cardíacos entre hombres y mujeres?

De los gráficos anteriores, observamos a 141 mujeres con ataques cardíacos sobre las 2994 indicadas anteriormente (4,7%), mientras que en el género masculino hay 108 sobre los 2115 (5,10%). Esto hace indicar, aparentemente, que puede haber una propensión, por algún factor a analizar, sobre los hombres a padecer ataques cardíacos.

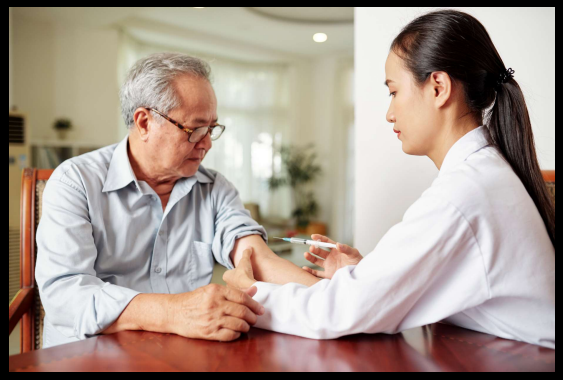


Análisis Exploratorio de Datos (EDA)

¿Hay una relación entre la edad y la incidencia de ataques cardíacos?

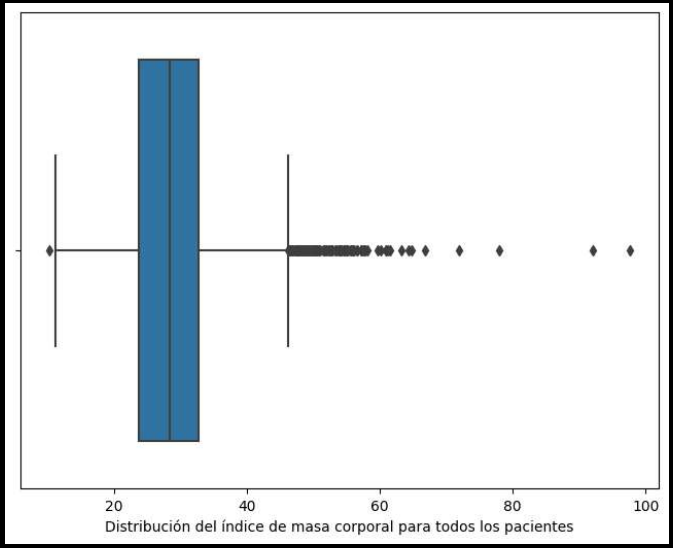
Análisis Exploratorio de Datos (EDA)

- ¿Hay una relación entre la edad y la incidencia de ataques cardíacos?

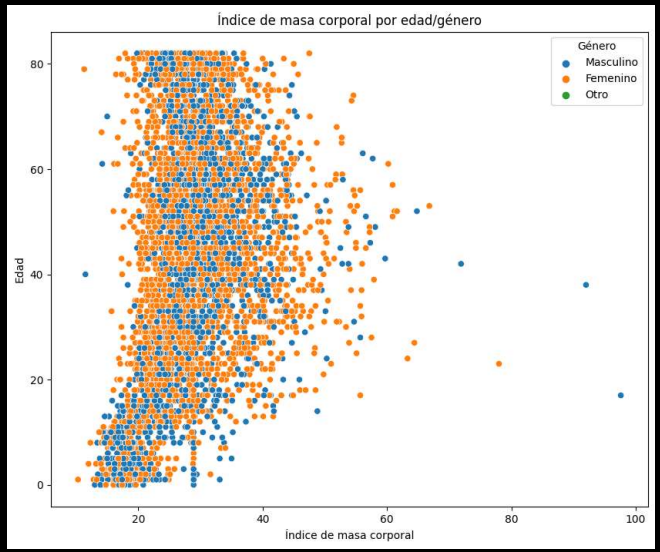


En la gráfica anterior, se puede observar que la mayoría de las personas en el rango de anciano, es decir, mayores de 65 años, tienen un mayor número de ataques cardíacos. Este hecho sugiere que la edad es un factor importante en el desarrollo de enfermedades cardíacas

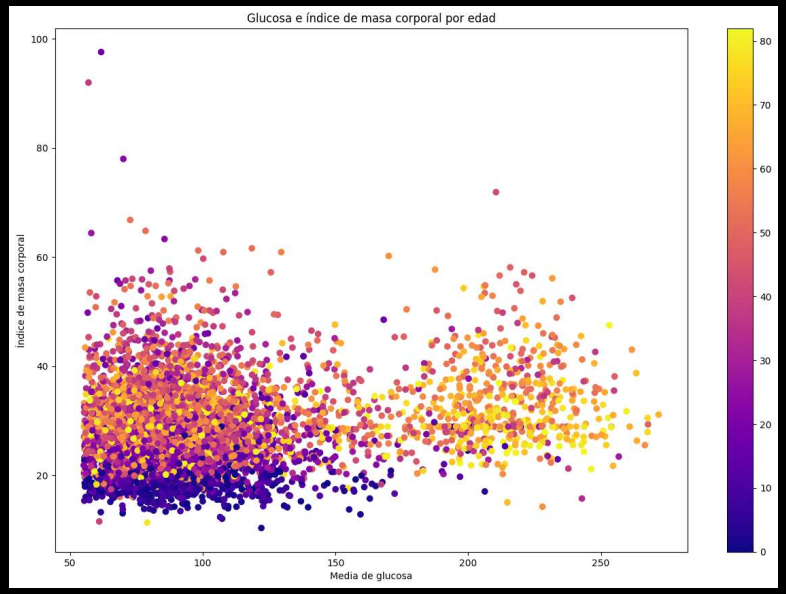
Análisis Exploratorio de Datos (EDA)



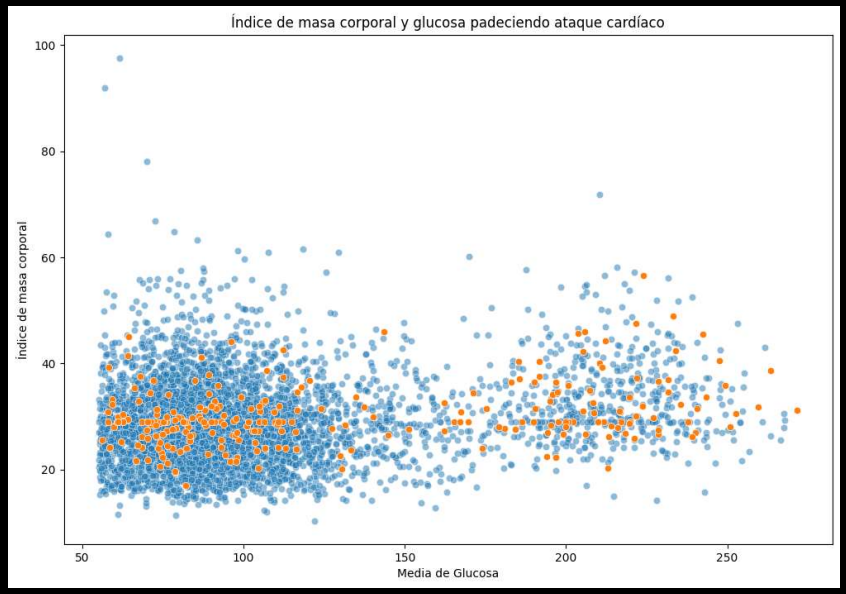
- ¿Qué masa corporal tienen los pacientes?
- ¿Hay alguna relación entre la masa corporal, la glucosa y la posibilidad del ataque cardíaco?



Análisis Exploratorio de Datos (EDA)



- ¿Qué masa corporal tienen los pacientes?
- ¿Hay alguna relación entre la masa corporal, la glucosa y la posibilidad del ataque cardíaco?



Análisis Exploratorio de Datos (EDA)

- ¿Qué masa corporal tienen los pacientes?
- ¿Hay alguna relación entre la masa corporal, la glucosa y la posibilidad del ataque cardíaco?

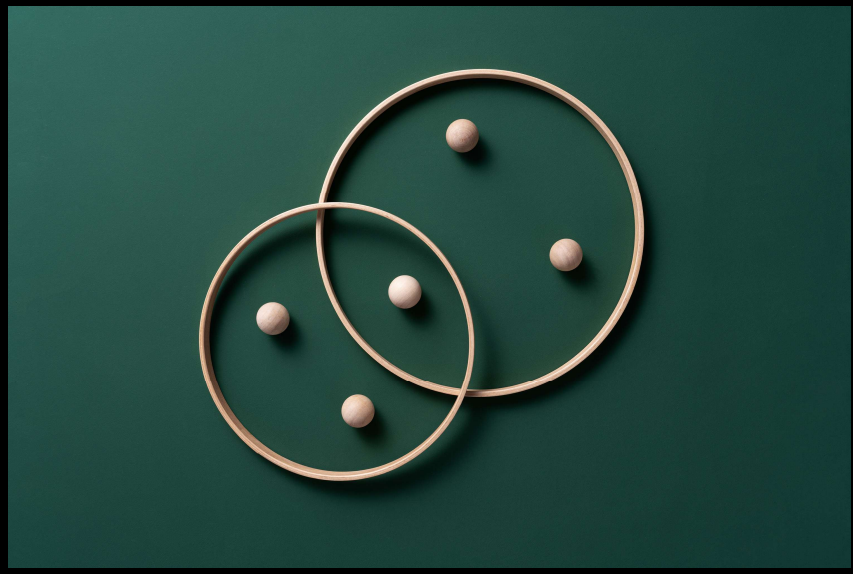
- La masa corporal de los pacientes se queda entre "normal" y "sobrepeso"
- En la última gráfica se observa que la combinación de masa corporal mayor de 30 y la glucosa elevada hacen a estos pacientes más propensos a padecer ataques cardíacos.

Conclusiones EDA

- En nuestro análisis exploratorio de datos (EDA), se identificó una probabilidad significativa de padecer un ataque cardíaco en pacientes con una masa corporal elevada y con niveles medios de glucosa más altos.
- Este análisis es solo una pequeña muestra de los datos disponibles, pero nos brinda una visión general de los factores que podrían estar contribuyendo a los ataques cardíacos.
- Es importante tener en cuenta que muchos otros factores podrían estar influyendo en la probabilidad de sufrir un ataque cardíaco, como la presencia de afecciones cardíacas previas, el hábito de fumar o el estado de salud general del paciente.
- Además, factores sociales como el tipo de residencia o trabajo también podrían tener un impacto en la salud cardíaca de un individuo.
- En conclusión, nuestro análisis es solo un punto de partida para entender mejor las causas de los ataques cardíacos y cómo prevenirlos. Es necesario profundizar en futuras investigaciones para obtener una visión más completa y precisa de la situación.

Machine Learning

- El objetivo principal de nuestro análisis es construir un modelo predictivo capaz de clasificar correctamente los casos de accidente cerebrovascular en base a sus características.
- Para este análisis, hemos seleccionado el algoritmo KNN (K-Nearest Neighbors), un método de aprendizaje supervisado utilizado ampliamente en problemas de clasificación
- Evaluaremos el rendimiento de nuestro modelo utilizando métricas de evaluación estándar, como precisión, recall y f1-score. Además, presentaremos una matriz de confusión para visualizar las predicciones del modelo.
- Basándonos en el rendimiento del modelo KNN y en las métricas de evaluación, extraeremos conclusiones sobre la capacidad del modelo para clasificar adecuadamente los casos de accidente cerebrovascular y su efectividad general en este análisis.



Preprocesamiento de datos

- Limpieza de datos: Se llevó a cabo la limpieza de datos para asegurar que no haya valores faltantes en el conjunto de datos. Los valores nulos fueron reemplazados por el valor promedio de cada columna.
- Creación de nuevas características: Se agregaron nuevas características al conjunto de datos para obtener información adicional. Se creó la columna "comorbidities" que representa la suma de las columnas de hipertensión y enfermedad. También se creó la columna "age_group" para agrupar las edades en rangos específicos.
- Tratamiento de outliers: Se identificaron y eliminaron outliers en las variables "avg_glucose_level" y "bmi" utilizando el rango intercuartílico (IQR). Esto mejora la calidad del análisis al eliminar valores extremos que podrían afectar negativamente el rendimiento del modelo.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke	comorbidities	age_group	activity_level_residence
0	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.600000	formerly smoked	1	1	60-70	3
1	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	28.893237	never smoked	1	0	60-70	2
2	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.500000	never smoked	1	1	70-80	2
3	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.400000	smokes	1	0	40-50	3
4	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24.000000	never smoked	1	1	70-80	2

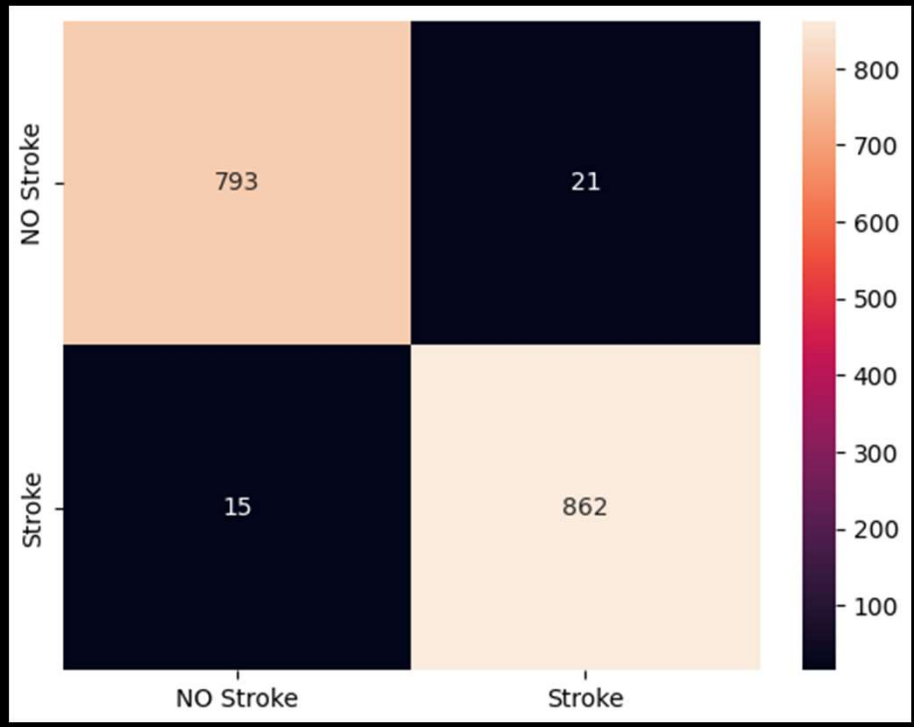
Selección del modelo

- Selección del modelo: Después de evaluar diferentes opciones, hemos seleccionado el algoritmo KNN (K-Nearest Neighbors) para nuestro análisis de clasificación de accidente cerebrovascular. KNN es un algoritmo ampliamente utilizado en problemas de clasificación y es conocido por su simplicidad y efectividad.
- Razonamiento de la selección: Elegimos KNN debido a su capacidad para clasificar nuevos casos basados en la similitud con los casos existentes en el conjunto de entrenamiento. Dado que nuestro objetivo es clasificar correctamente los casos de accidente cerebrovascular, consideramos que KNN es una opción adecuada para este problema.

- El proceso de modelado incluyó la división del conjunto de datos en conjuntos de entrenamiento y prueba, la estandarización de las características y la búsqueda de hiperparámetros óptimos utilizando GridSearchCV. Estos pasos aseguran que nuestro modelo esté entrenado adecuadamente y sea capaz de generalizar a nuevos casos.

Resultados

- Mejor puntuación: La mejor puntuación obtenida durante la validación cruzada fue de 0.9704. Esto indica un buen rendimiento del modelo KNN en la clasificación de casos de accidente cerebrovascular en el conjunto de datos equilibrado.
- Métricas de evaluación: A continuación, se presentan las métricas de evaluación del modelo KNN en el conjunto de prueba:
    - Precisión: La precisión para la clase "No Stroke" fue de 0.9814 y para la clase "Stroke" fue de 0.9762.
    - Recall: El recall para la clase "No Stroke" fue de 0.9742 y para la clase "Stroke" fue de 0.9829.
    - F1-score: El F1-score para la clase "No Stroke" fue de 0.9778 y para la clase "Stroke" fue de 0.9795.
- A continuación, se muestra el mapa de calor de la matriz de confusión, que visualiza las predicciones correctas e incorrectas del modelo en cada clase.



Conclusiones

- Rendimiento del modelo: El modelo KNN con los mejores hiperparámetros encontrados (distancia de Manhattan y 3 vecinos) ha demostrado un buen rendimiento en la clasificación de casos de accidente cerebrovascular. Las métricas de evaluación, como precisión, recall y F1-score, muestran resultados favorables tanto para la clase "No Stroke" como para la clase "Stroke".
- Importancia del preprocesamiento: El preprocesamiento de datos desempeñó un papel crucial en el éxito del modelo. La limpieza de datos, el redondeo de variables, la creación de nuevas características y el tratamiento de outliers permitieron obtener un conjunto de datos más completo y representativo, lo que mejoró el rendimiento del modelo.
- Consideraciones sobre la métrica de distancia: La elección de la métrica de distancia es un factor importante en el rendimiento del modelo KNN. En este análisis, la distancia de Manhattan resultó ser una opción efectiva, capturando la similitud entre las instancias basada en el desplazamiento en líneas verticales y horizontales.
- Importancia de la validación cruzada: La validación cruzada y la búsqueda de hiperparámetros mediante GridSearchCV nos permitieron optimizar el modelo y obtener una estimación más confiable del rendimiento. Estos enfoques garantizan que el modelo sea capaz de generalizar a nuevos casos y reducen el riesgo de sobreajuste.

En general, el modelo KNN aplicado a nuestro conjunto de datos de accidente cerebrovascular ha demostrado ser efectivo en la clasificación y detección de casos. Estas conclusiones respaldan la utilidad de los enfoques de aprendizaje automático en la identificación de factores de riesgo y la toma de decisiones relacionadas con la prevención de accidentes cerebrovasculares.