Predicción de accidentes cerebrovasculares

•••

Proyecto final curso data science

Maximiliano Romano

¿Que es un Accidente cerebrovascular?

Un accidente cerebrovascular sucede cuando el flujo de sangre a una parte del cerebro se detiene. Algunas veces, se denomina "ataque cerebral".

Si el flujo sanguíneo se detiene por más de pocos segundos, el cerebro no puede recibir nutrientes y oxígeno. Las células cerebrales pueden morir, lo que causa daño permanente.

Un accidente cerebrovascular se presenta cuando un vaso sanguíneo en el cerebro se rompe, causando un sangrado dentro de la cabeza.

De acuerdo a datos de la Organización Mundial de la Salud, 15 millones de personas sufren un ACV por año. De éstos, mueren 5 millones y otros 5 millones quedan con una discapacidad permanente.

Es la tercera causa de muerte y la primera causa de discapacidad en el mundo.

Objetivo:
El objetivo del proyecto es predecir la presencia de un accidente cerebrovascular

Preguntas y objetivos

Predecir presencia de infartos cerebrovasculares

¿Cuáles son las variables con más relación con los accidentes cerebrovasculares (stroke)?

¿Cuál modelo de entrenamiento se ajusta mejor al dataset?

¿Existe relación entre el bmi y la edad?

¿Las personas fumadoras tienen más riesgo de tener un infarto?

Dataset

Se utilizó el dataset llamado “Brain stroke prediction dataset” de la web Kaggle

Variables

Gender: "Hombre", "Mujer".

Age: Edad del paciente.

Hypertension: Si el paciente tiene o no hipertensión: 0,1.

Heart_disease: Si el paciente tiene o no una enfermedad cardiaca: 0, 1.

Ever_married: Alguna vez casado: "No" o "Sí".

Work_type: Tipo de trabajo: "niños", "Nunca trabajó", "Privado" o "Autónomo".

Residence_type: Tipo de residencia: "Rural" o "Urbano".

Avg_glucose_level: Nivel medio de glucosa en sangre.

Bmi: Indice de masa corporal.

Smoking_status: "antes fumaba", "nunca ha fumado", "fuma" o "desconocido".

Stroke: Si el paciente tuvo o no un accidente cerebrovascular: 0, 1.

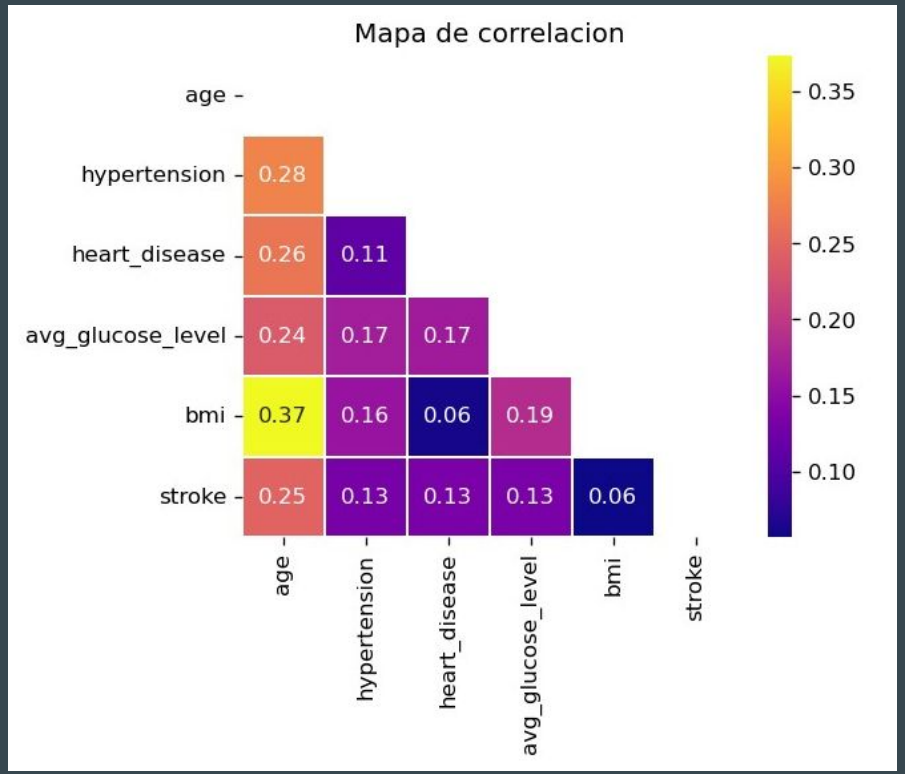
Datos ausentes:

No se encontraron datos datos ausentes

Datos duplicados:

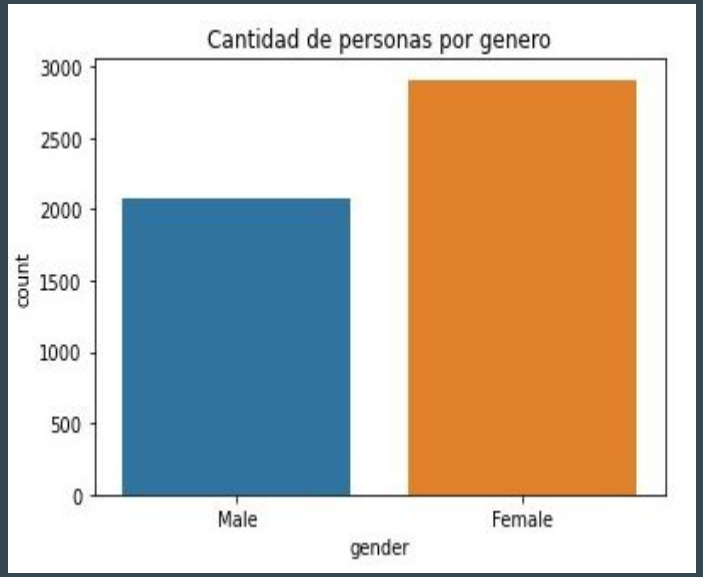
No se encontraron datos duplicados

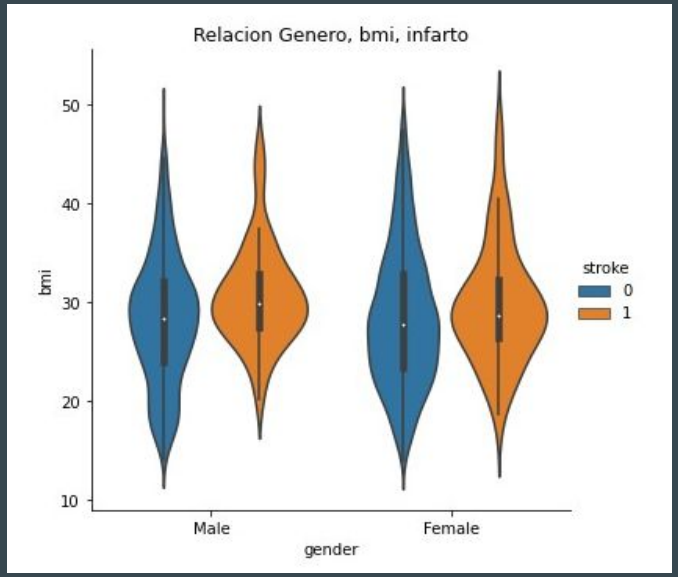
Análisis exploratorio de datos (EDA)



Como podemos ver en el mapa de correlación, la variable Age es muy relevante para stroke

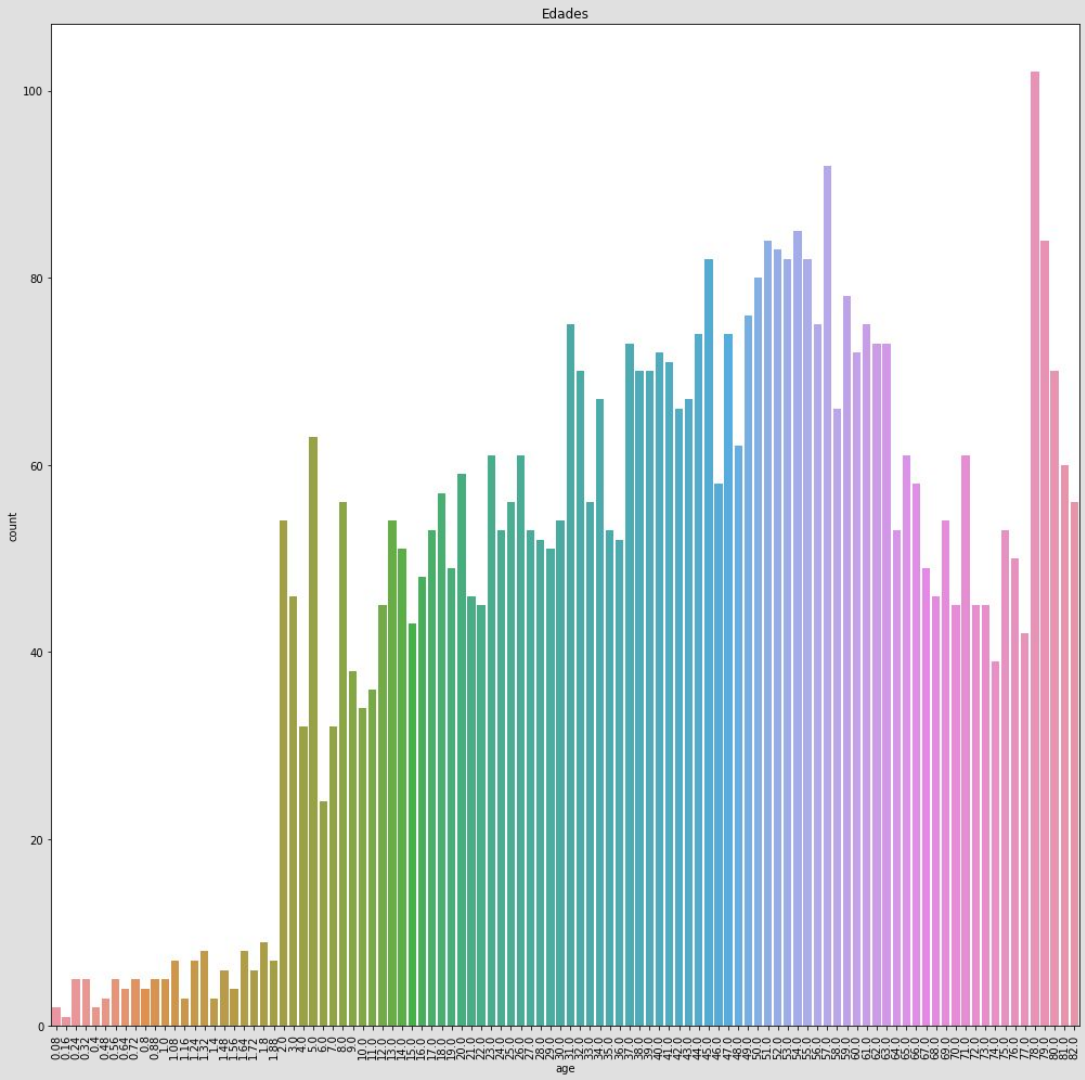
Relación género e infarto



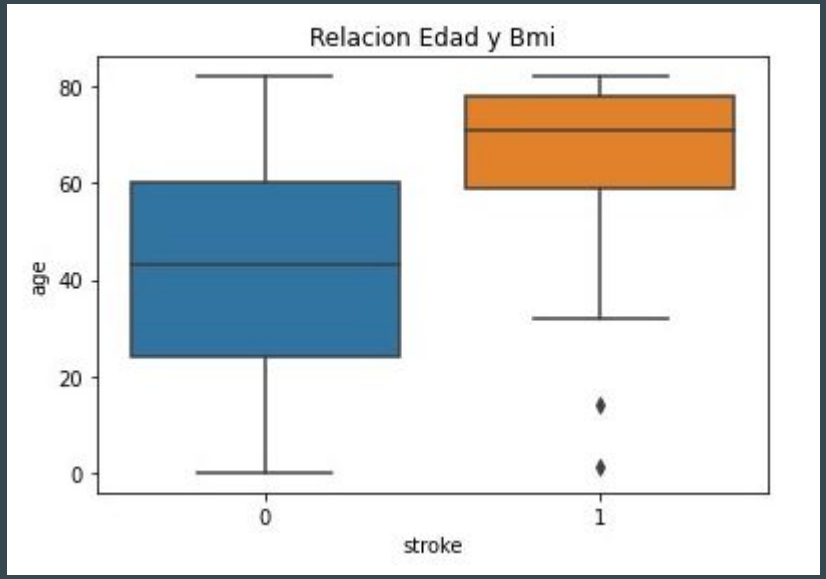


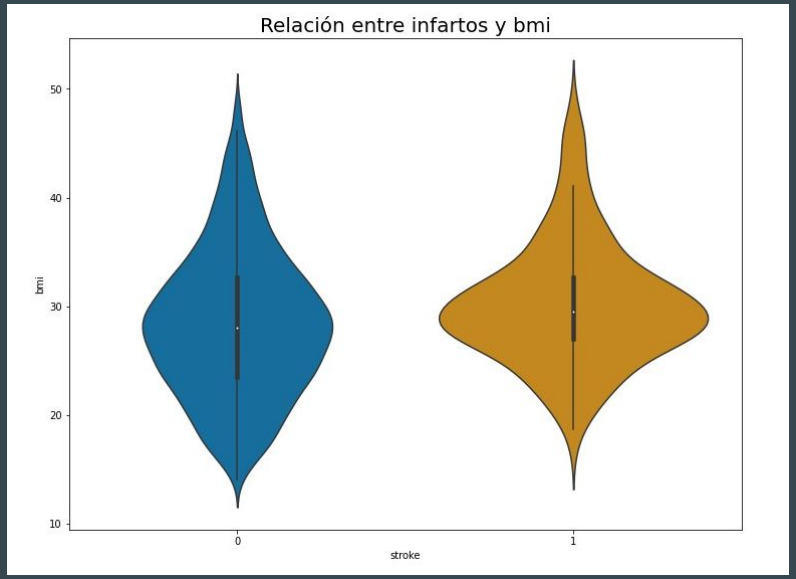
Como se ve en el primer gráfico tenemos mayor cantidad de mujeres en este dataset

No hay relación entre el género de la persona y la probabilidad de infarto. También se puede ver que el bmi en mujeres no infartadas es un poco más alto



La mayor cantidad de datos se encuentra entre los 3 a los 82 años.



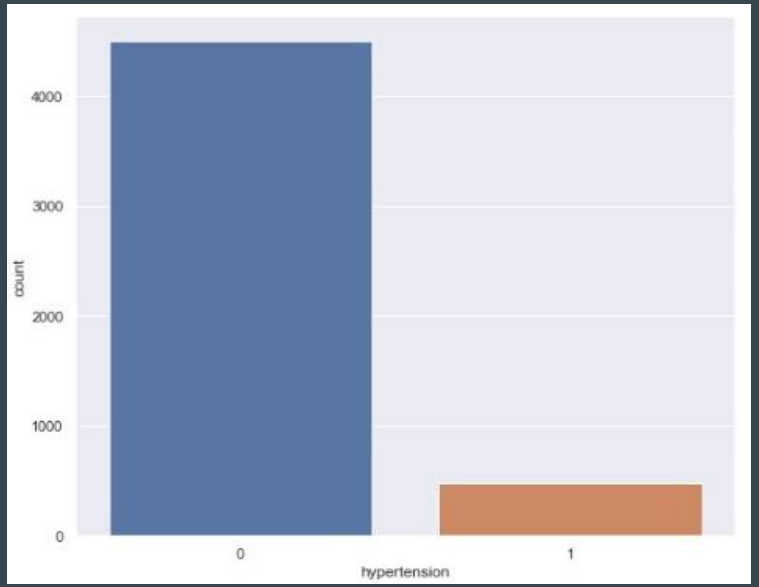


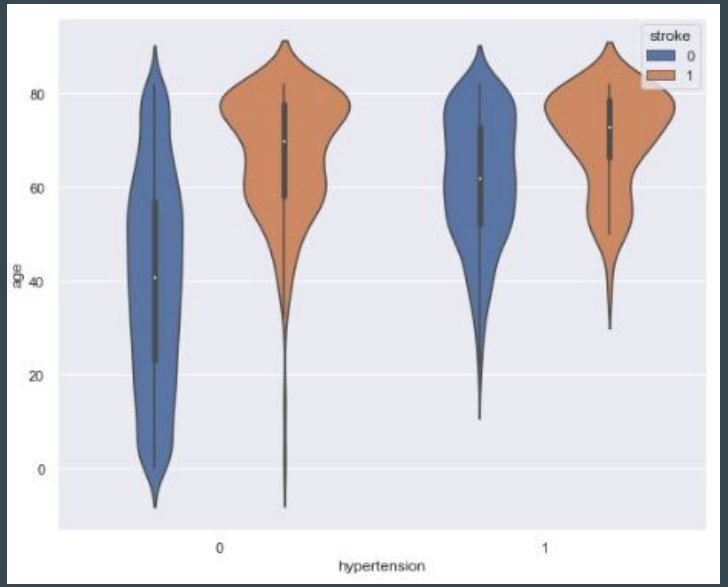
Los infartos se concentran entre las personas de 60 y 80 años (Mucha relación con la edad)

Se pueden ver valores atípicos en la columna de infarto a los 0 y 20 años

Se ve que la mayoría de personas infartadas tenían bmi cercano a 30

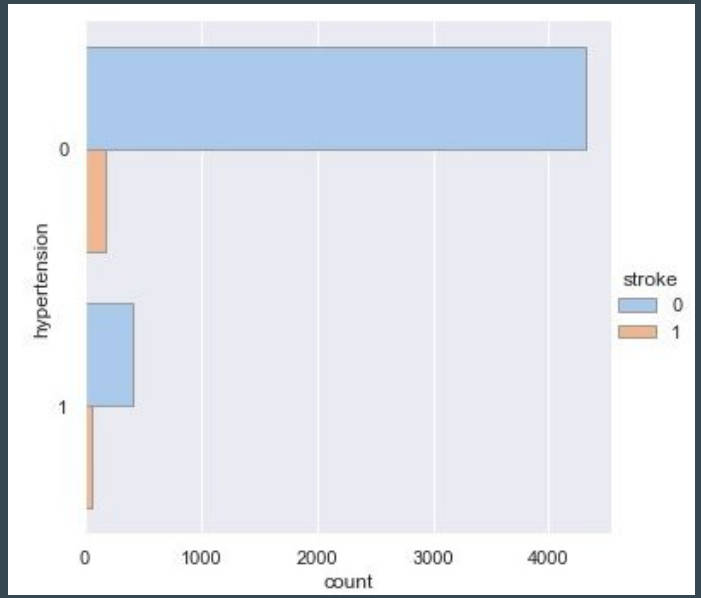
Relación hipertensión y edad





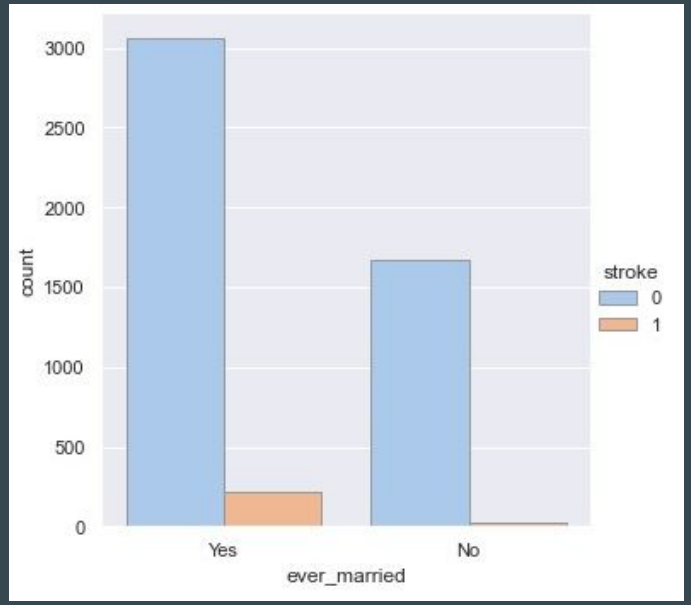
La edad de la gente con hipertensión se encuentra entre los 40 y 80

Relación hipertensión y edad



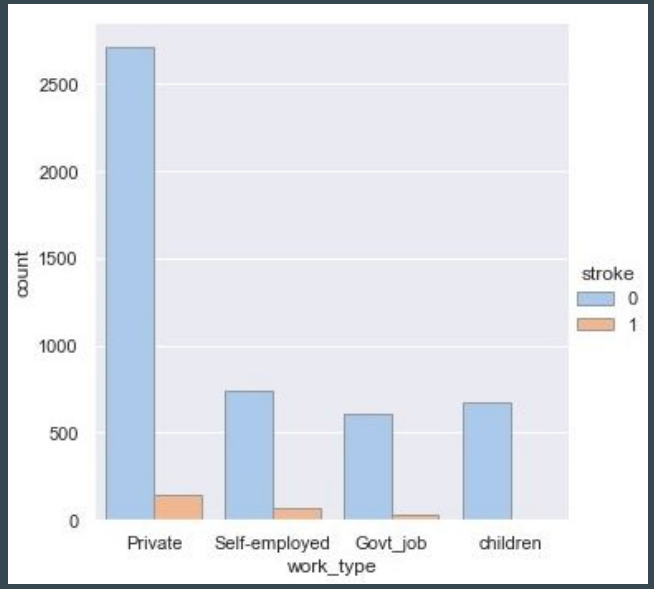
La proporción de gente infartada con hipertensión es más alta.

Relación Estado civil e infarto



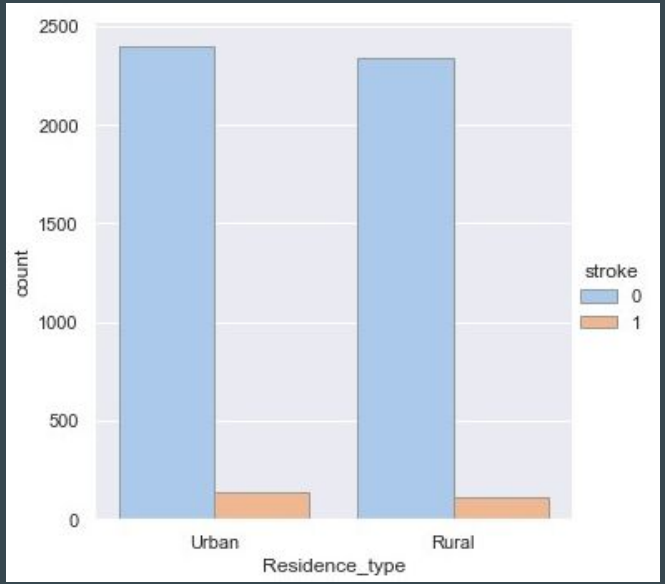
aaaaaaaaaaaaaaaaaaaaaaaaa

Relación tipo de trabajo e infarto



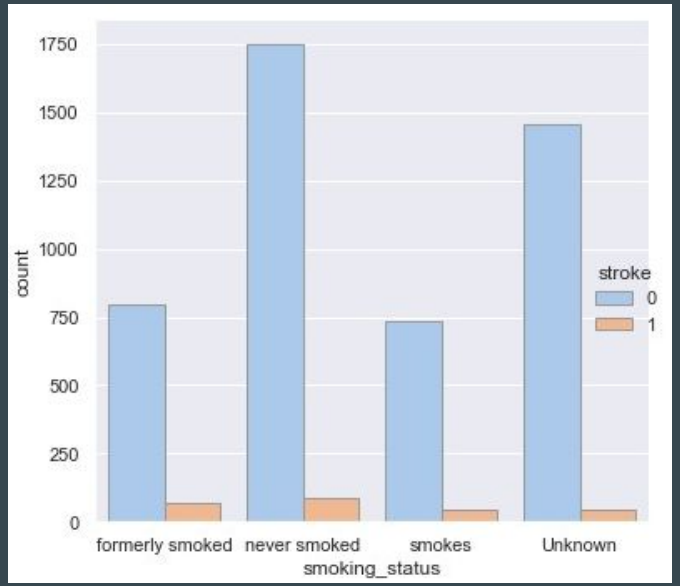
aaaaaaaaaaaaaaaaaaaaaaaaaaa

Relación tipo de residencia e infarto



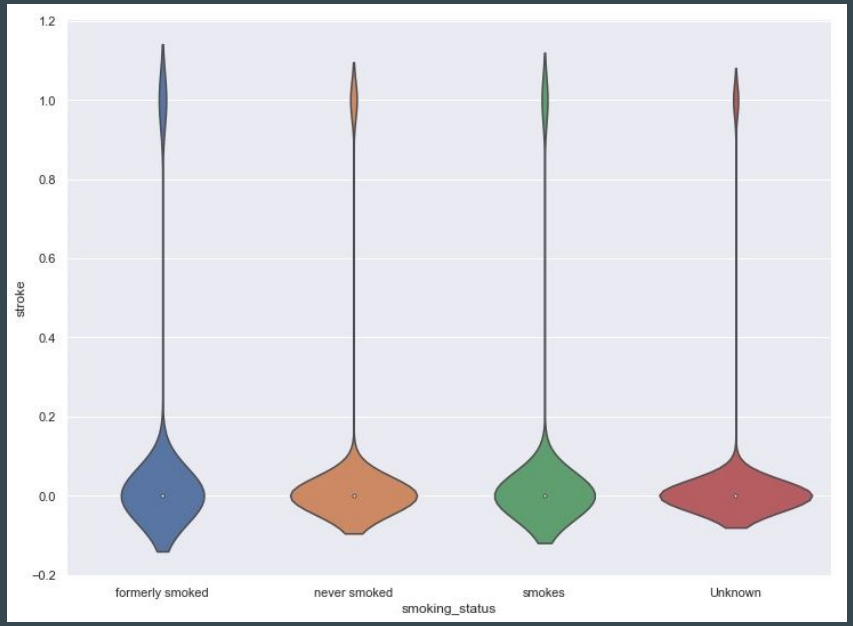
No se ve ninguna diferencia entre las persona que viven en un entorno urbano y rural

Relación fumar e infarto



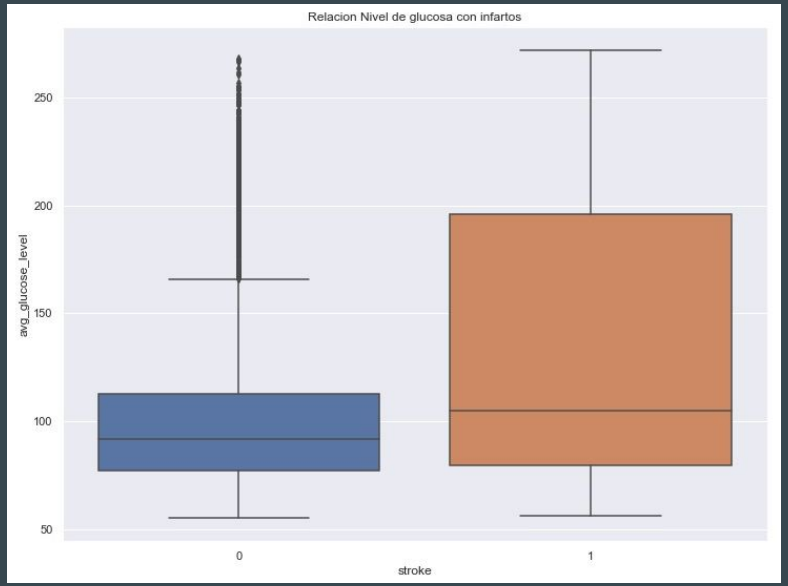
aaaaaaaaaaaaaaaaaaaaaaaaaaa

Relación fumar e infarto



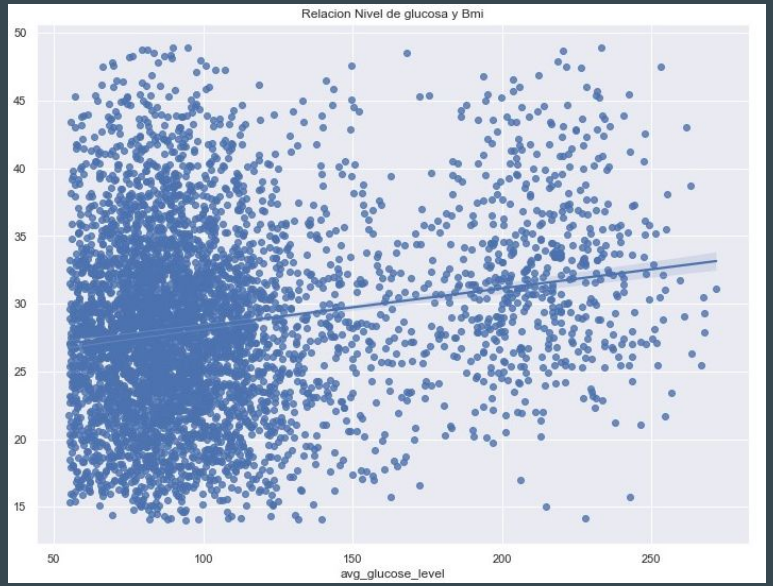
aaaaaaaaaaaaaaaaaaaaaaaaaaa

Relación nivel de glucosa e infarto



Se ve claramente cómo las personas con un alto nivel de glucosa alto tienen más probabilidad de sufrir un infarto

Relación nivel de glucosa y bmi



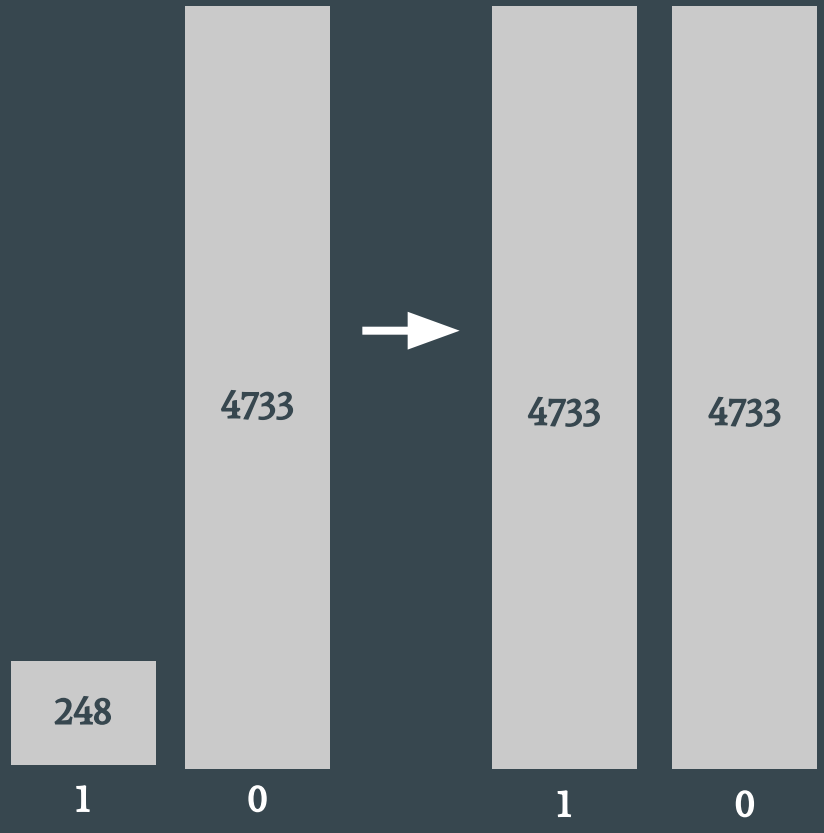
Vemos un gráfico con una línea creciente indicando que a un nivel más alto de glucosa más bmi

Balanceo de datos

Definimos la variable target en stroke

Nos encontramos con que la variable target estaba muy desbalanceada.

Aplicamos RandomOverSampler y Smote para balancear la muestra y que nuestro modelo de predicción funcione correctamente.



Modelos de predicción

Probamos con 5 modelos diferentes:

- Árboles de decisión
- Random forest
- Regresión logística
- KNN
- SVM

Mejores modelos

Árboles de decisión

	Train	Test
Aciertos	81%	82%
Precisión	75%	75%
Recall	94%	94%

Utilizamos un Max_depth(máxima profundidad) de 5

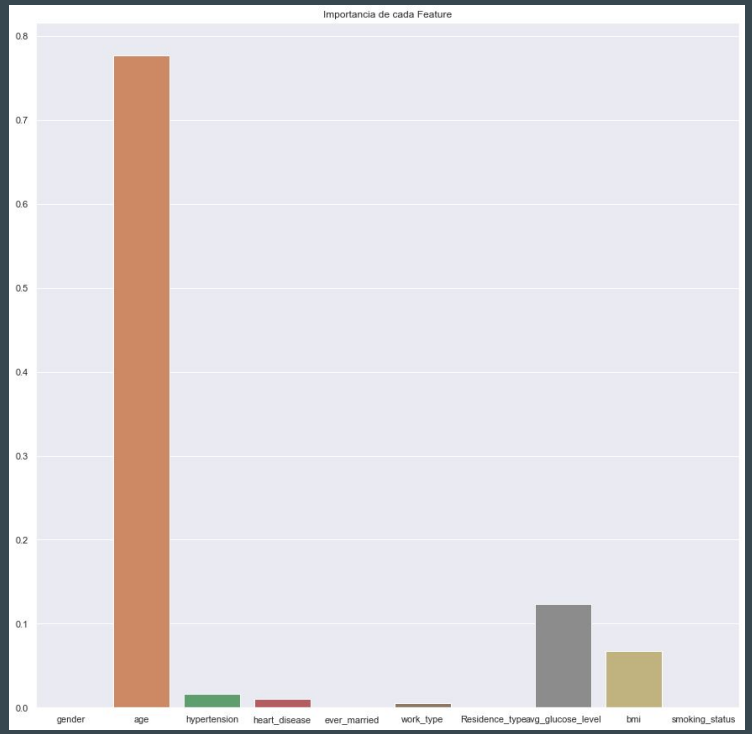
Random forest

	Train	Test
Aciertos	80%	81%
Precisión	75%	76%
Recall	90%	91%

Aquí también utilizamos un Max_depth(máxima profundidad) de 5

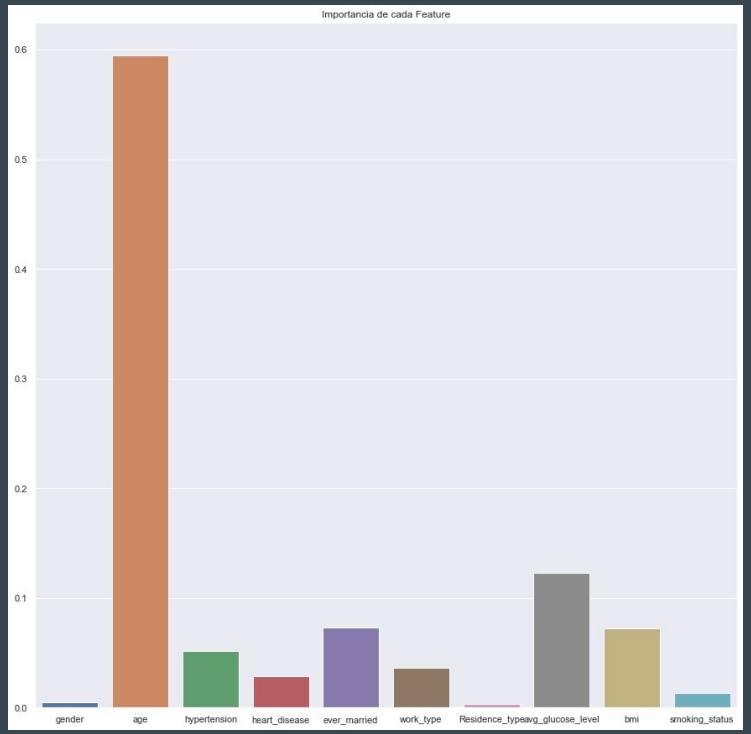
Importancia de cada Feature

Árboles de decisión



Vemos como la variable que más afecta al modelo es la edad seguido del nivel de glucosa y el bmi

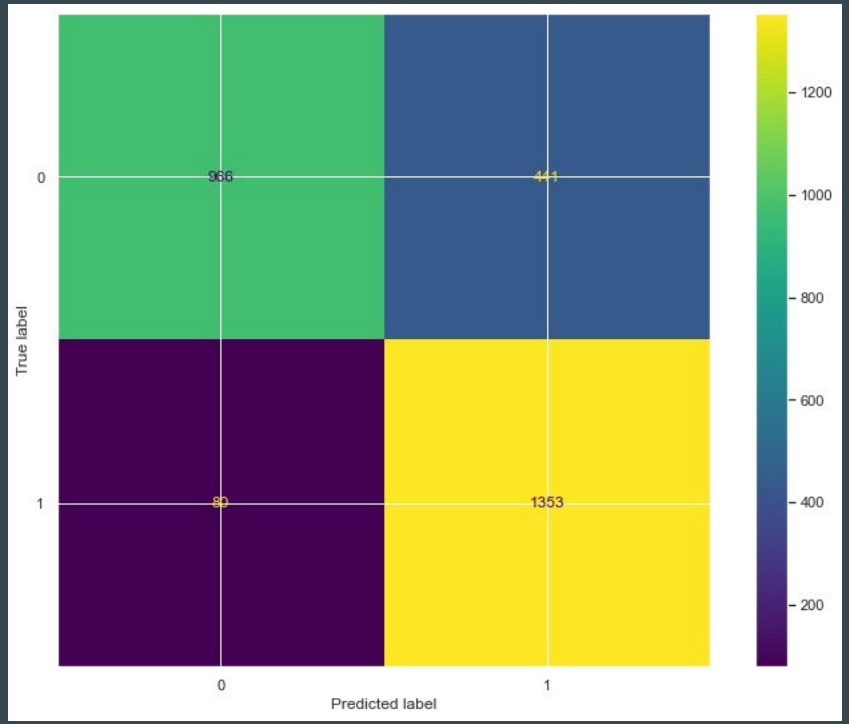
Random forest



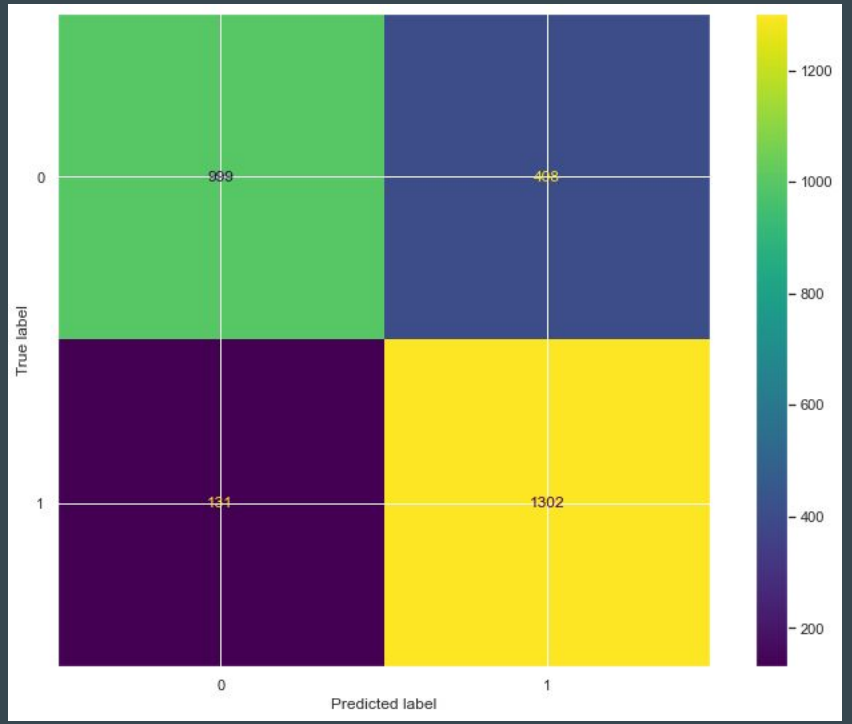
En este caso vemos como otras variables tienen más relevancia

Matriz de confusión

Árboles de decisión



Random forest



Otros modelos

Regresión logística

	Train	Test
Aciertos	76%	77%
Precisión	75%	76%
Recall	78%	80%

Poco rendimiento

KNN

	Train	Test
Aciertos	97%	94%
Precisión	94%	90%
Recall	100%	100%

Overfitting (sobreajuste)

SVM

	Train	Test
Aciertos	77%	78%
Precisión	74%	75%
Recall	83%	84%

Poco rendimiento

Soluciones propuestas

Solución 1	• Escribe aquí tu texto • Escribe aquí tu texto
Solución 2	• Escribe aquí tu texto • Escribe aquí tu texto
Solución 3	• Escribe aquí tu texto • Escribe aquí tu texto
Solución 4	• Escribe aquí tu texto • Escribe aquí tu texto

Muchas gracias!

...