



DICIEMBRE 2021

PREVENCIÓN DE ACV

PROYECTO FINAL
CURSO DATA SCIENCE
CODERHOUSE



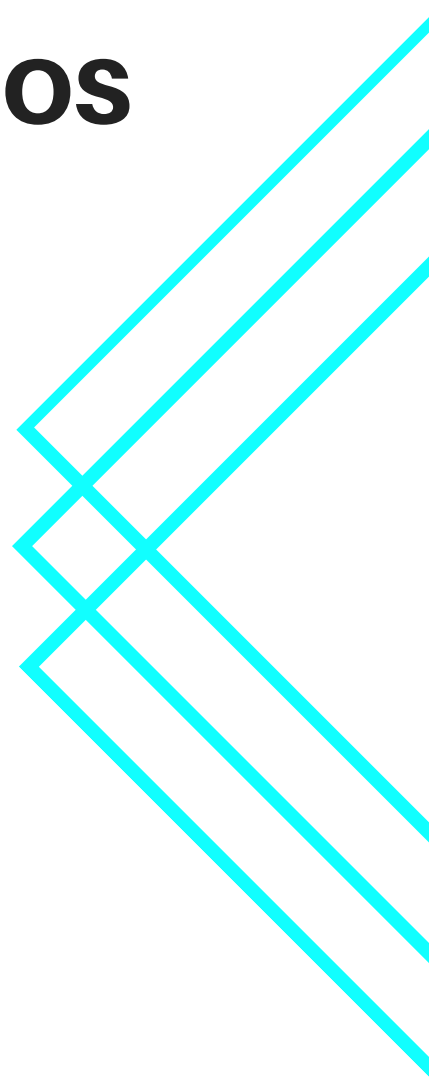
Elaborado por Bautista Pazos
Data Analyst

Tabla de contenidos

- 1.Caso de negocio
- 2.Dataset
- 3.Exploratory Data Analysis (EDA)
- 4.Algoritmo
- 5.Conclusión

Versión

Versión 1.0 (7/12/202



Introducción

Un accidente cerebrovascular (ACV) ocurre cuando algo bloquea el suministro de sangre a una parte del cerebro o cuando un vaso sanguíneo en el cerebro estalla.

En cualquier caso, partes del cerebro se dañan o mueren. Un derrame cerebral puede causar daño cerebral duradero, discapacidad a largo plazo o incluso la muerte.

En la Argentina:

- Se produce un accidente cerebrovascular (ACV) cada nueve minutos.
- 126 mil casos de ACV por año, de los cuales 18 mil terminan en muerte.

En Estados Unidos:

- USD\$ 46 mil millones fueron los costos relacionados a ACVs en 2015. Esto incluye atención médica, medicamentos y días de trabajo perdidos.

Según la OMS, el 80% de los casos son prevenibles

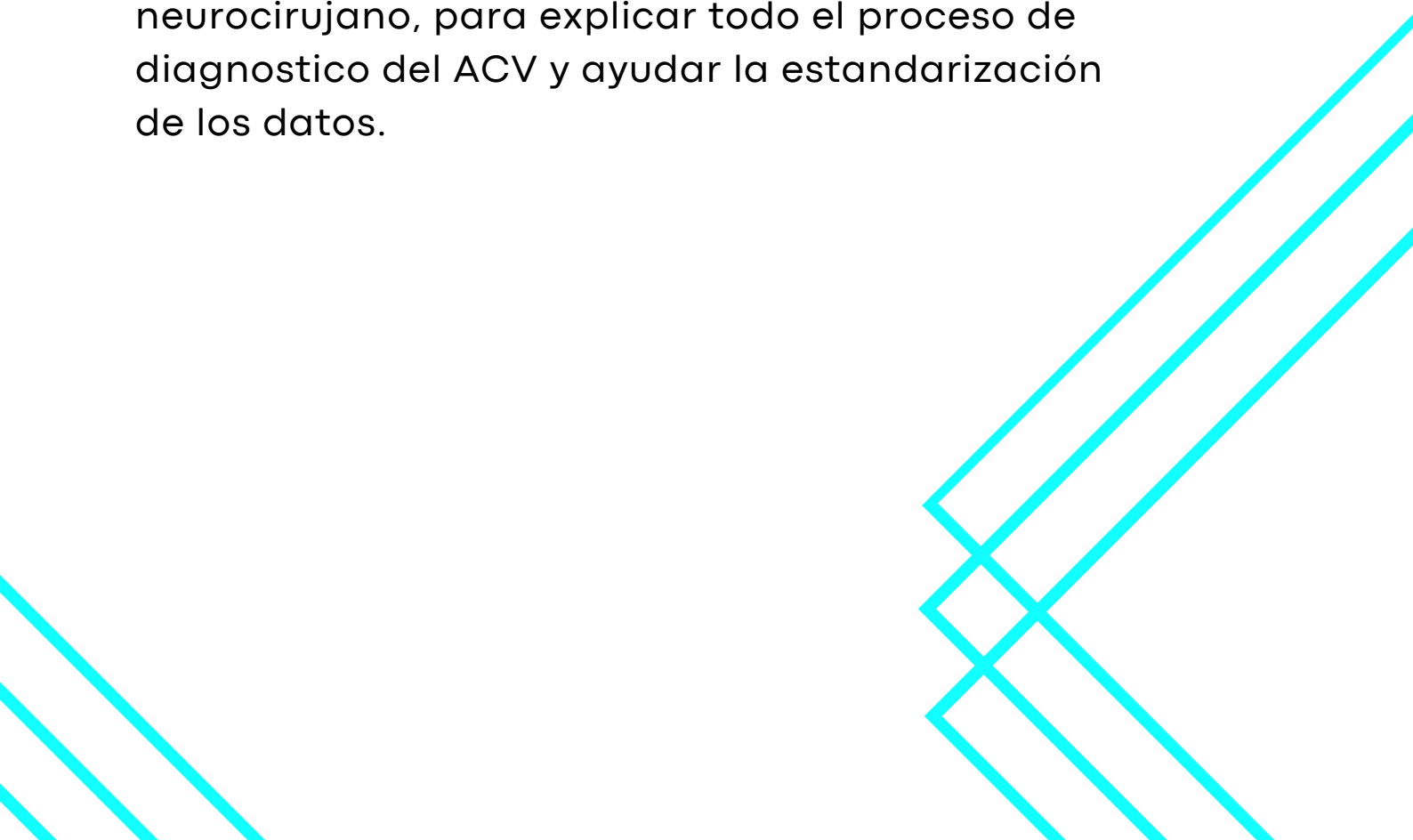


Objetivo

Construir un modelo de Machine Learning utilizando algoritmos de clasificación para que médicos, clínicas, aseguradoras y empresas den un diagnóstico más certero a la hora de prevenir un accidente cerebrovascular.

Apropiación del área investigada

Agradezco la colaboración y la buena predisposición del Dr Esteban Blanco, neurólogo y neurocirujano, para explicar todo el proceso de diagnóstico del ACV y ayudar la estandarización de los datos.

Decorative cyan geometric lines in the bottom right corner, consisting of several overlapping parallel lines forming a series of nested, elongated shapes.

Dataset

Se utilizó un dataset de kaggle.com cuya fuente es anónima.

Cuenta con 5110 observaciones y 12 variables categóricas y continuas.

Variables categóricas:

- gender: Género de individuo. Str
- hypertension: Informa si el paciente tiene hipertensión o no. Int(1,0)
- heart_disease: Informa si el paciente tiene problemas cardíacos o no. Int(1,0)
- ever_married: Informa si el paciente esta casado o no. Str (Yes, No)
- work_type: Distintas categorías de trabajo. Str (children, Govt_job, Never_worked, Private, Selfemployed)
- Residence_type: Tipo de residencia del individuo. Str (Urban, Rural)
- smoking_status: Estatus de fumador del individuo. Str (formerly smoked, never smoked, smokes, unknown)
- stroke: Variable target. Sufrió o no un acv el paciente

Variables continuas:

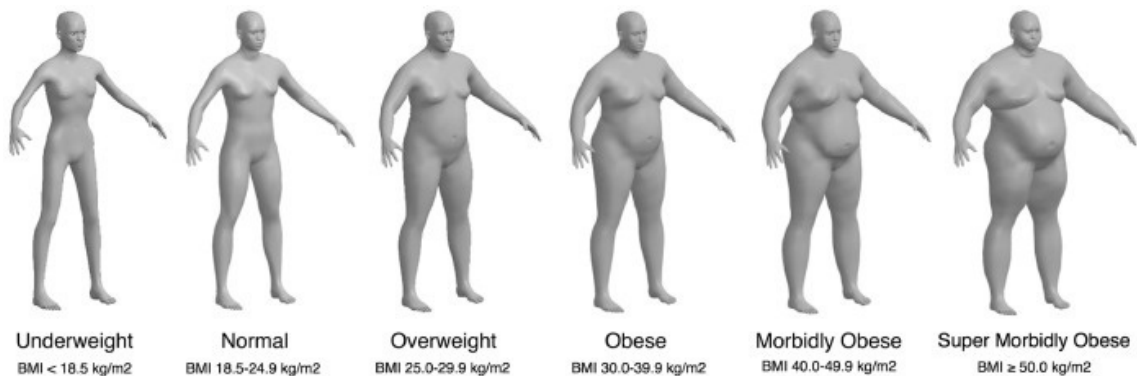
- id: Número de identificación del individuo. Int
- bmi: Índice de masa corporal. Float
- avg_glucose_level: Promedio de glucosa en sangre del individuo. Float

Feature Engineering

En la variable género, había una observación de categoría "Other". Se descartó por no agregar valor.

Se descartó la columna de id porque no agregaba información relevante.

Con ayuda del Dr. Blanco, decidimos que aquellos valores en donde el bmi sea mayor a 50, se fijarían en 50 dado que ya valores mayores a 40 representan casos de extremo riesgo.



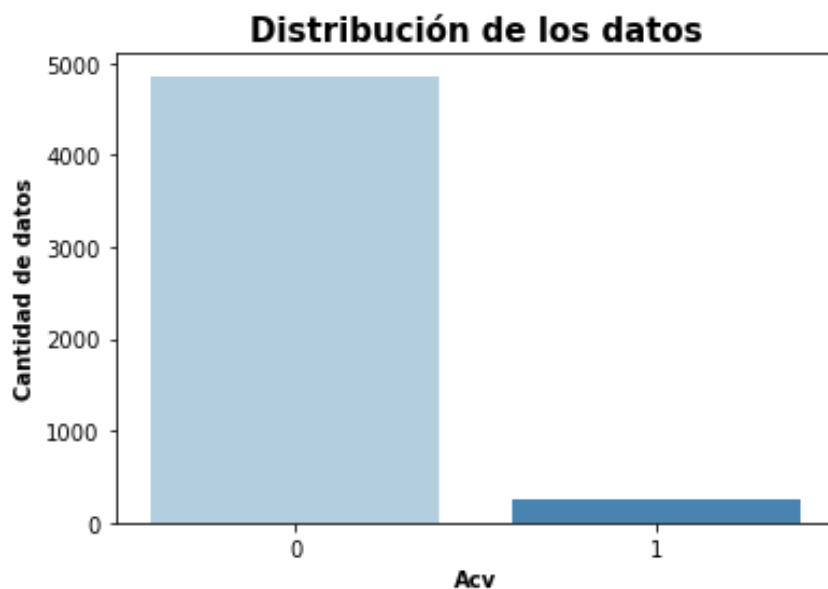
Valores faltantes:

Los únicos nulos se encontraban en la variable de bmi. Fueron reemplazados por el promedio acorde a si sufrió un acv o no.

En smoking_status, el 30% del dataset era desconocido. Se optó para que sea una categoría más.

Feature Engineering

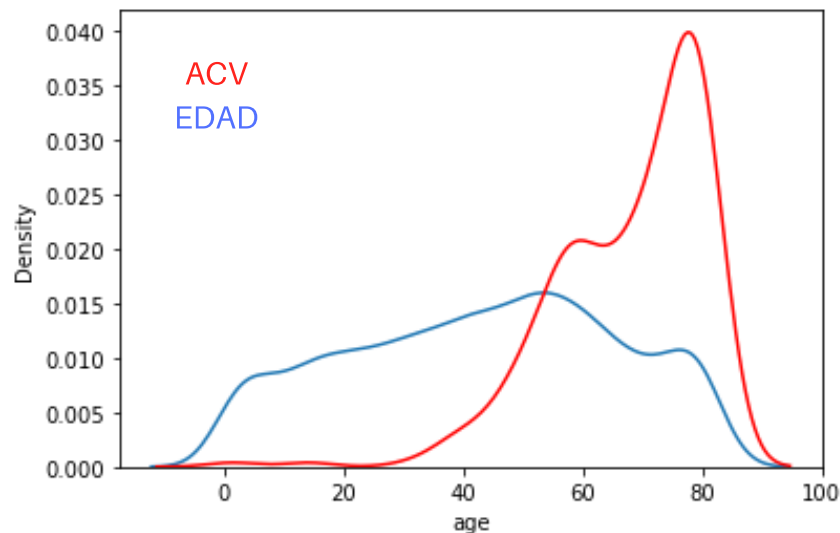
Nos encontramos que la distribución de los datos respecto a la variable target estaba altamente desbalanceada.



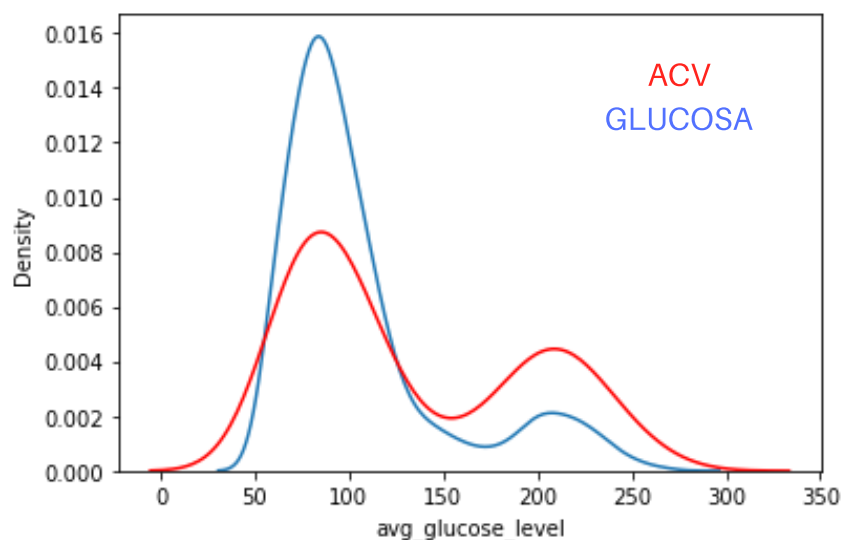
Aplicamos RandomOversampler, StratifiedKFold y SMOTE (funciones frecuentemente utilizadas para problemas de clasificación desbalanceados) para balancear las muestras y tener una división de train y test que sea óptima para el modelo.

Exploratory Data Analysis (EDA)

Variables continuas:



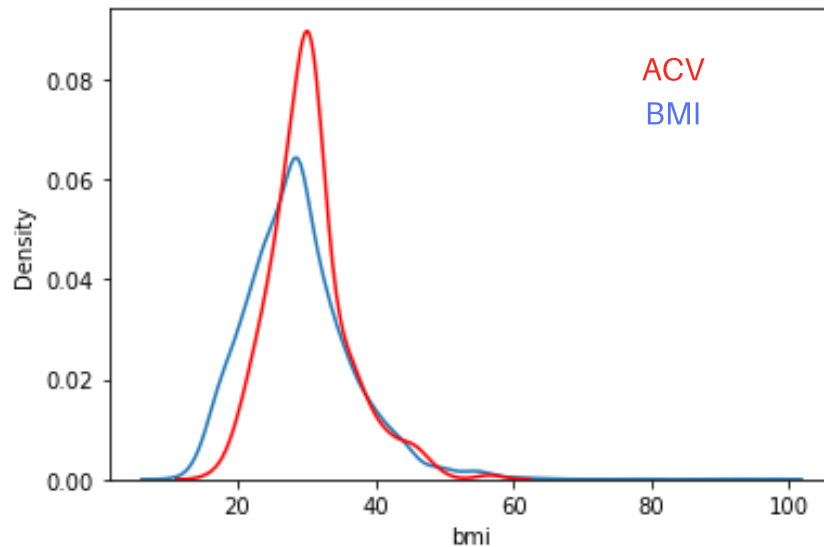
Hay una fuerte implicación de que, a mayor es la edad, mayor es la tendencia a que haya un ACV.



Vemos que hay mas casos de ACVs que la población cuando se ronda los 200, pero la gran mayoría se encuentra rondando los 75, que también son los valores más repetidos de la población.

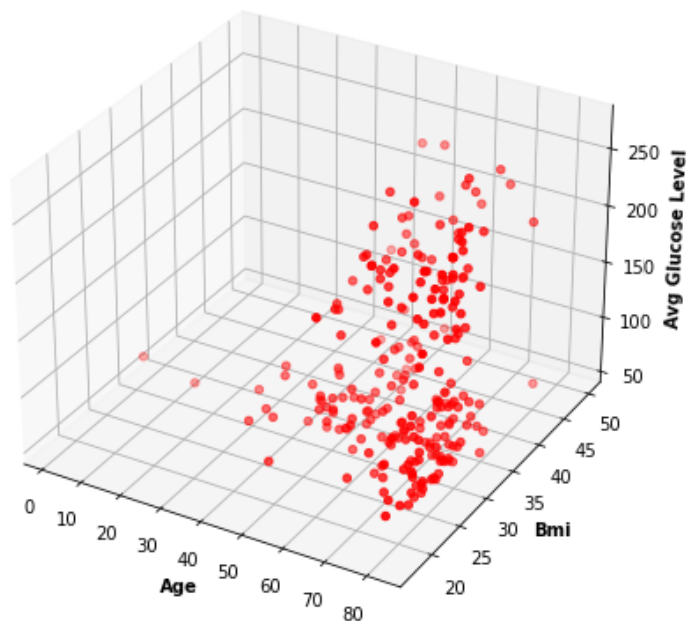
Exploratory Data Analysis (EDA)

Variables continuas:



La gran mayoría de los casos de ACV se producen rondando los 30, aunque también es el valor que más se repite en la población.

Relación Bmi-Age-Avg Glucose en casos de ACV



No habría una relación clara entre las tres variables.

Exploratory Data Analysis (EDA)

Variables categóricas:

41%
Hombres



5% TUVO UN ACV

61%
Mujeres



4% TUVO UN ACV

No hay prácticamente diferencia con respecto al género

10%
Con hipertensión



14% TUVO UN ACV

90%
Sin hipertensión



4% TUVO UN ACV

Tener hipertensión aumentaría la posibilidad de tener un ACV frente a no tenerla.

Exploratory Data Analysis (EDA)

Variables categóricas:



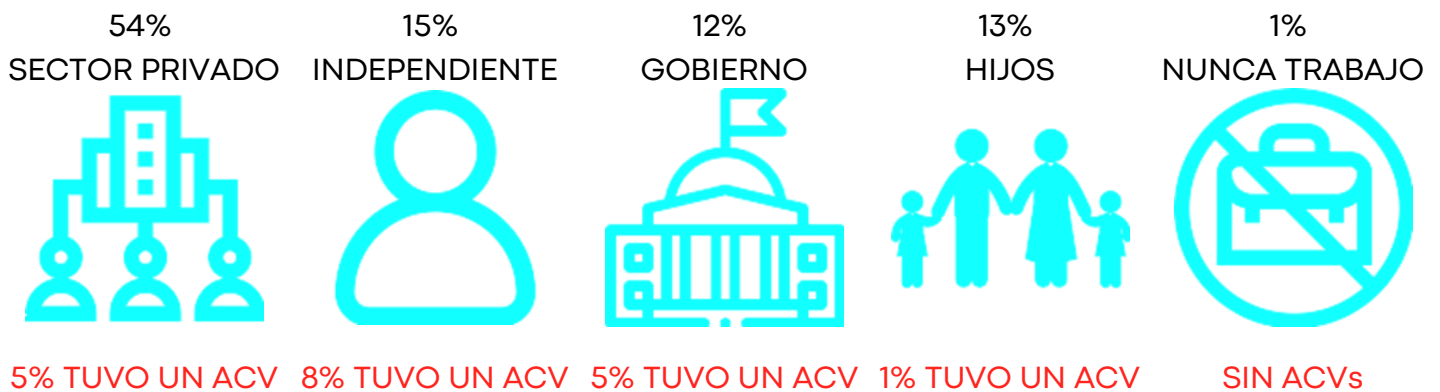
No hay prácticamente diferencia con respecto al estado civil.



Tener una enfermedad cardíaca aumentaría la posibilidad de tener un ACV frente a no tenerla.

Exploratory Data Analysis (EDA)

Variables categóricas:



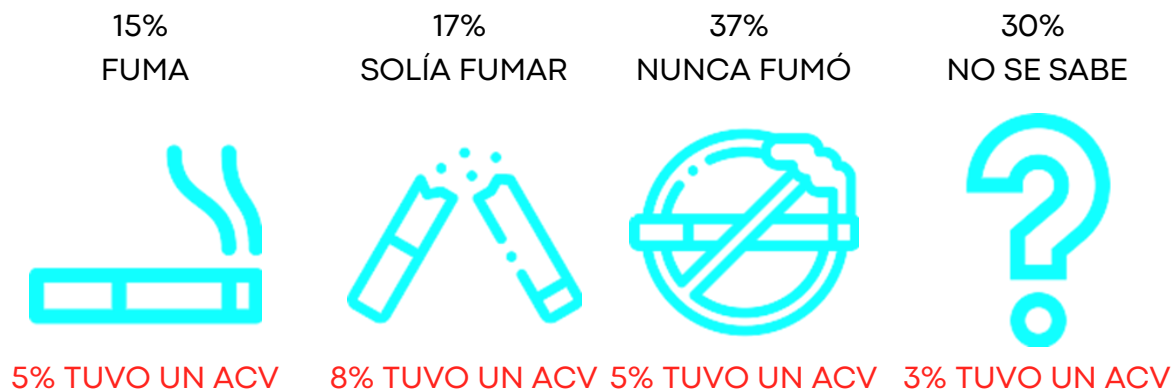
Vemos que el trabajo independiente es el más propenso a sufrir un ACV.



No hay prácticamente diferencia con respecto al tipo de residencia

Exploratory Data Analysis (EDA)

Variables categóricas:



No hay prácticamente diferencia con respecto al tipo de residencia



No hay prácticamente diferencia con respecto al tipo de residencia

Exploratory Data Analysis (EDA)

Matriz de correlaciones



La edad y bmi son las variables con mayor correlación. La variable que más se relaciona con la posibilidad de tener un acv es la edad.

Métricas de desempeño

Valoraremos la performance del modelo acorde a:

1º) Recall: Porcentaje de clasificación correcta de los verdaderos positivos.

Queremos evitar a toda costa los falsos negativos por una cuestión ética.

2º) Precision: Porcentaje de clasificación correcta de los falsos negativos.

Mientras mayor sea, de mayor valor será el modelo.

3º) Accuracy: Porcentaje de clasificación correcta de los falsos negativos.

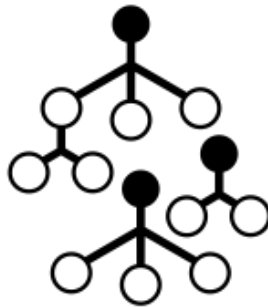
Mide el desempeño general.



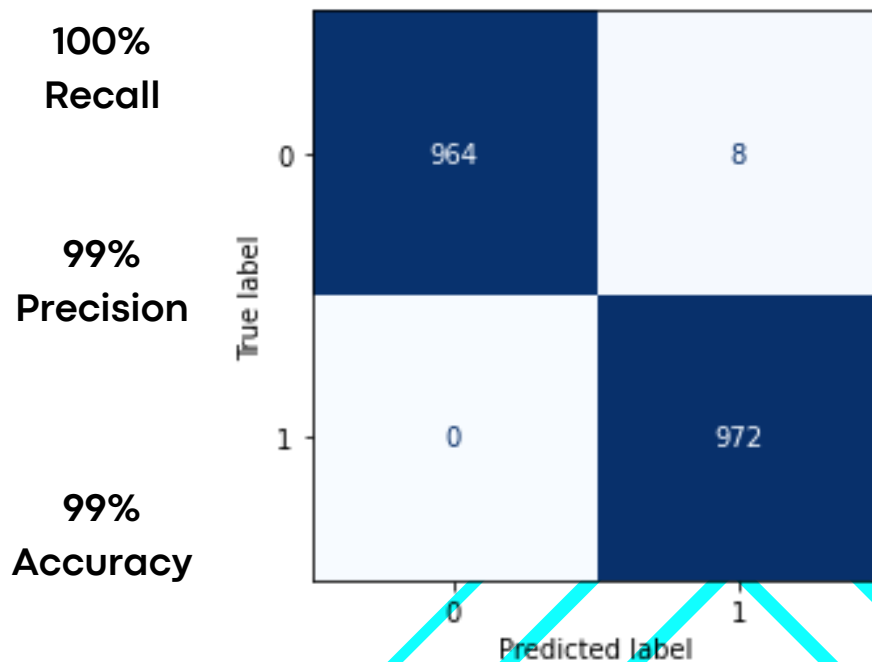
Algoritmo elegido

Se exploraron 5 posibles modelos con los algoritmos de clasificación: Árbol de decisión, Random Forest, SVM, Logistic Regression y KNN.

El de mejor performance fue **Random Forest** y no tuvo que ser optimizado

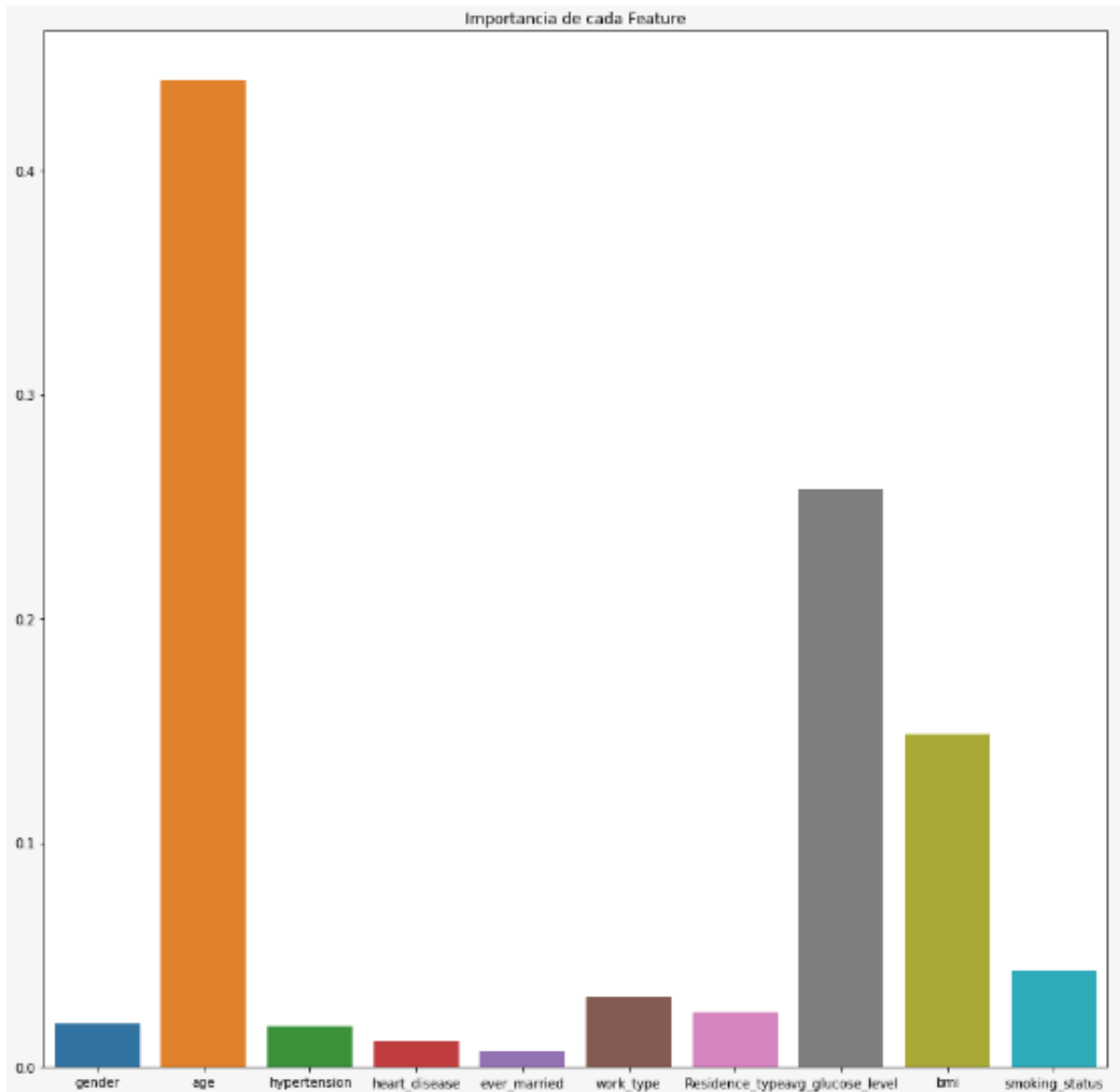


Matriz de confusión



Algoritmo elegido

Importancia de cada variable



Vemos una preponderancia de la edad sobre todas las demás, algo que podíamos intuir con lo visto en el EDA.

Siguientes pasos

1º) Inversión: Buscar inversores interesados en la investigación y desarrollo del proyecto.

2º) App: Desarrollar un front-end para que el usuario final pueda interactuar con el modelo.

3º) Promoción: Llevar la aplicación a los potenciales clientes.

Conclusiones

El modelo superó con creces las expectativas que se tenían con respecto a su performance.

Hubiera sido interesante contar con más variables como el consumo de alcohol, diabetes, colesterol e hipotiroidismo, utilizadas en los diagnósticos médicos.

Muchas gracias!

Contacto:
bautistapazos97@gmail.com

<https://www.linkedin.com/in/bautista-pazos/>

