Grupo 1:

Si es Bayes es Bueno

ÁGUILA, Pablo SUÁREZ, Daniel DECURGEZ, Rocío

Presentación:
Stroke Prediction Dataset

Stroke **positivo**: 4.9%

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Origen: **CONFIDENCIAL**

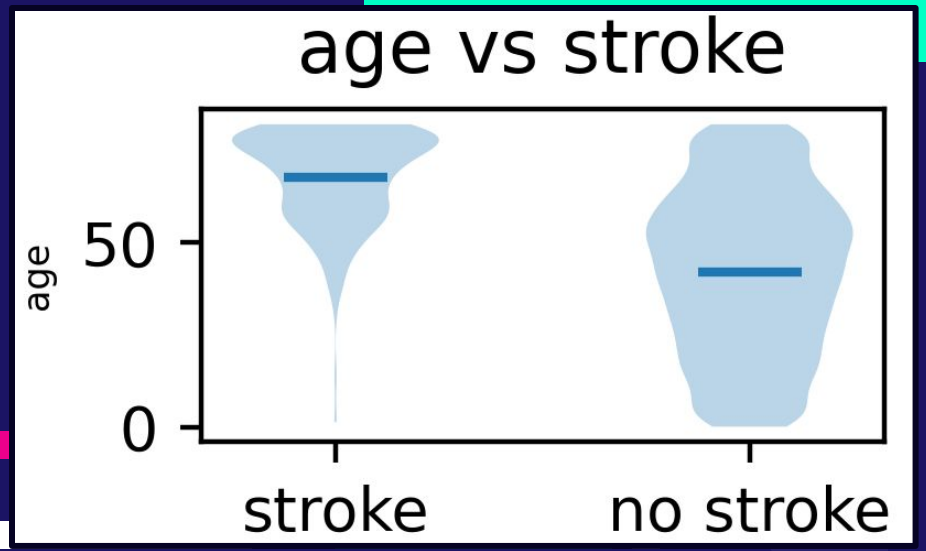
Tipos de variables: Categóricas | Continuas | Target

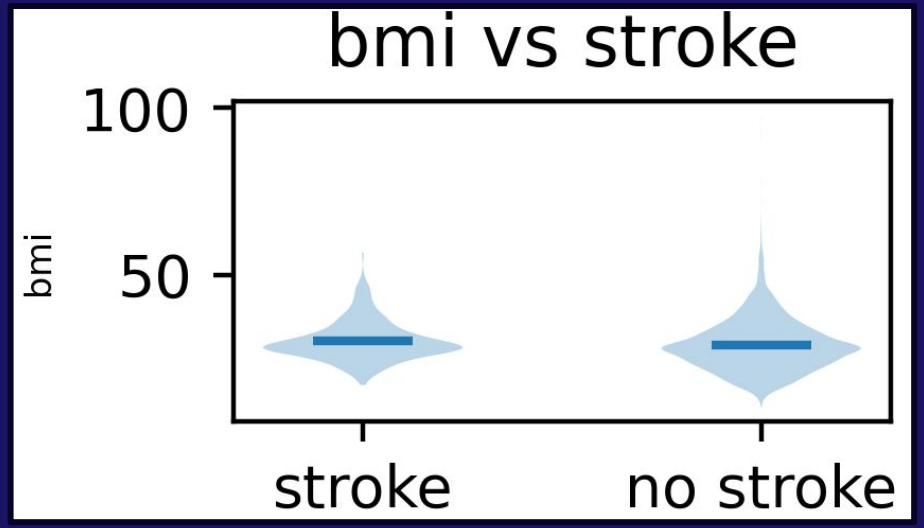
Metadata del Dataset

COLUMNA	TIPO DE VARIABLE	DESCRIPCIÓN
gender	Categórica	Género de la persona
age	Continua	Edad de la persona
hypertension	Categórica	Indicador de hipertensión
heart_disease	Categórica	“0” = No, “1” = Sí
ever_married	Categórica	Estado civil actual y/o pasado
work_type	Categórica	Tipo de trabajo
Residence_type	Categórica	Tipo de residencia actual
avg_glucose_level	Continua	Nivel de glucosa promedio
bmi	Continua	Índice de masa corporal (IMC)
smoking_status	Categórica	Estado de tabaquismo
stroke	Target	“0” = No, “1” = Sí

Análisis Exploratorio de Datos, gráficos y conclusiones

Violinplots





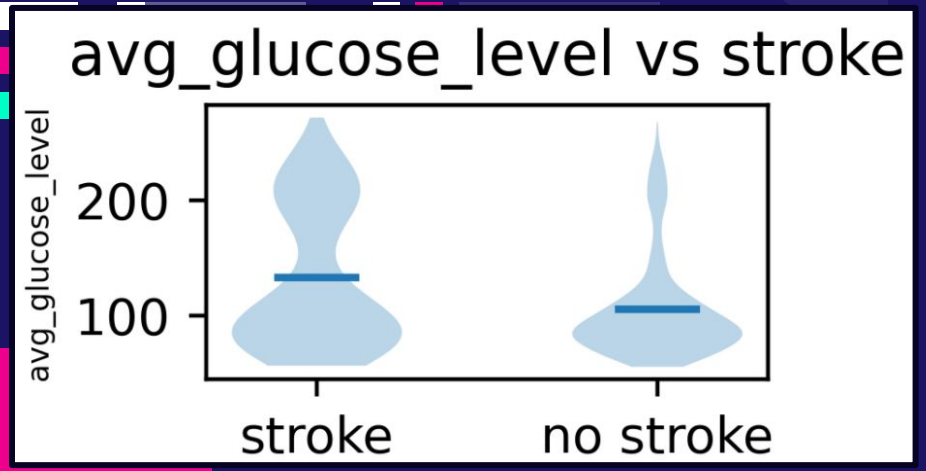


Gráfico de dispersión del análisis tSNE bidimensional clasificado por color según el label.

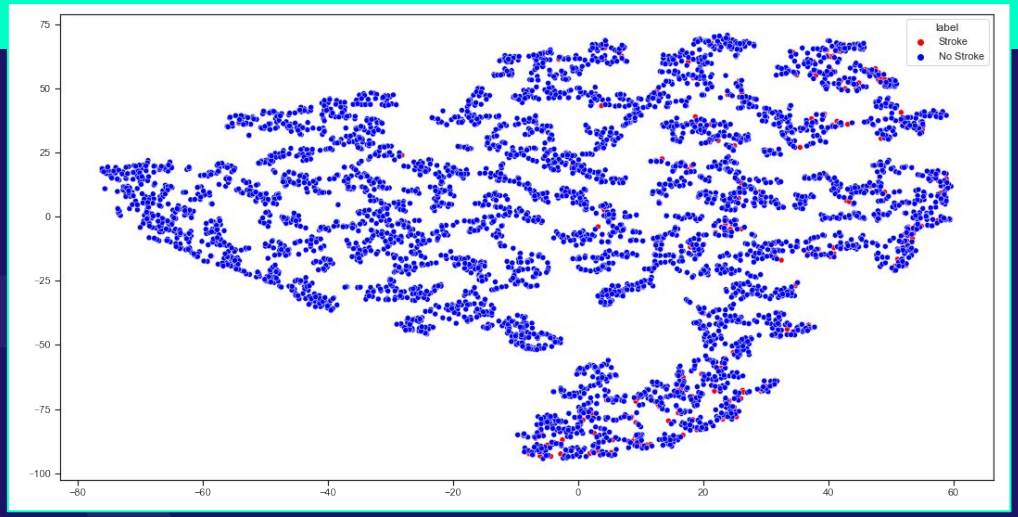
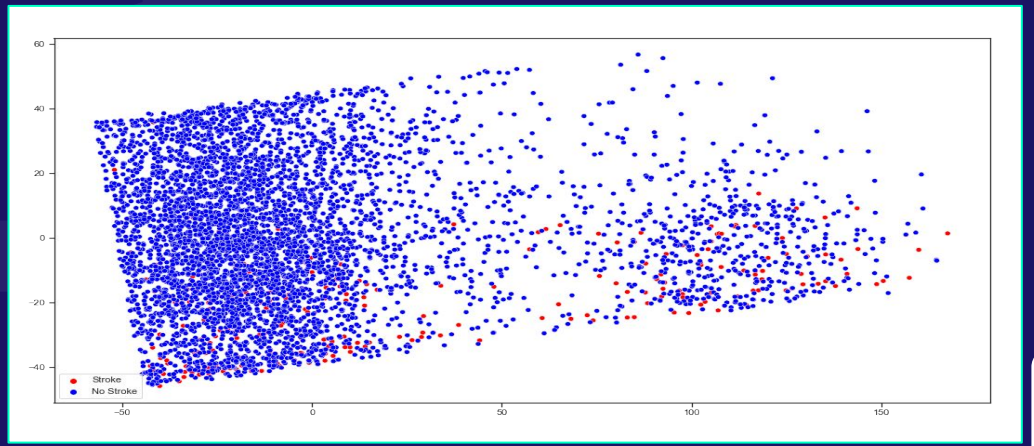
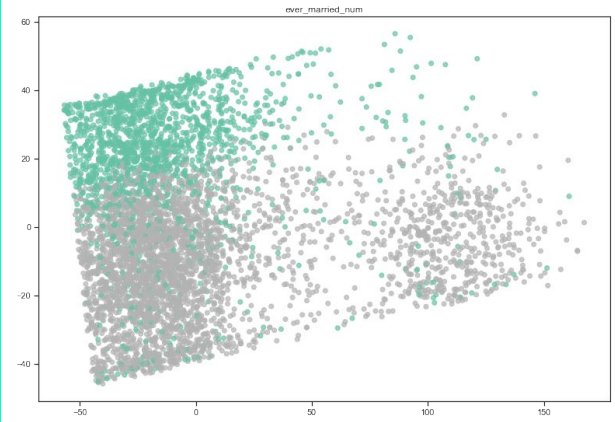
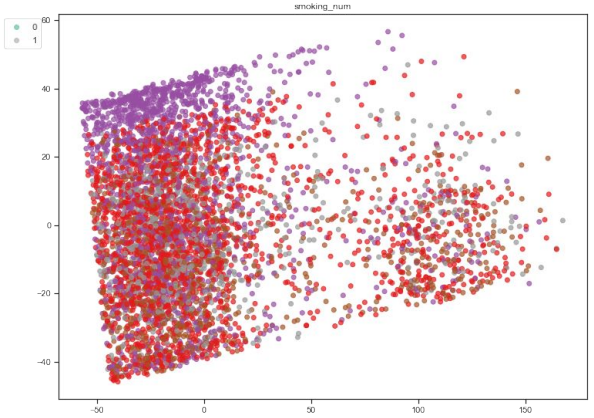
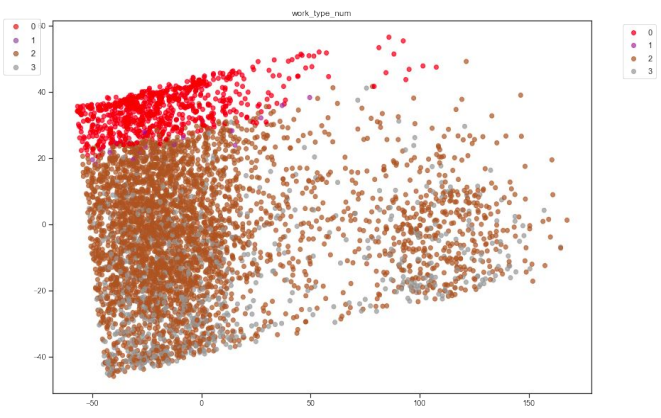


Gráfico de dispersión del análisis PCA bidimensional clasificado por color según el label.









Gráficos de dispersión del análisis PCA bidimensional clasificado por color según las features “ever_married”, “smoking_status” y “work_type”.

Problema a resolver

Buscamos predecir si la persona es propensa a sufrir un accidente cerebrovascular (ACV/*stroke*).

Los ACV matan a **más de 137.000** personas al año. Esto representa aproximadamente 1 de cada 18 muertes. En promedio, alguien muere de un derrame cerebral cada **cuatro minutos**[1].

Algunos factores de riesgo conocidos:

- Presión alta
- Fibrilación auricular
- Diabetes
- Alto colesterol
- Cigarrillo

[1] - American Heart Association

Estrategias posibles

Modelos de clasificación:

- Logistic Regression
- K-nearest neighbors
- Random Forest
- Support vector machine
- Naive Bayes

Tenemos varios features categóricos pero ordinales por lo que probablemente no usemos OneHotEncoder.

¡Muchas gracias!