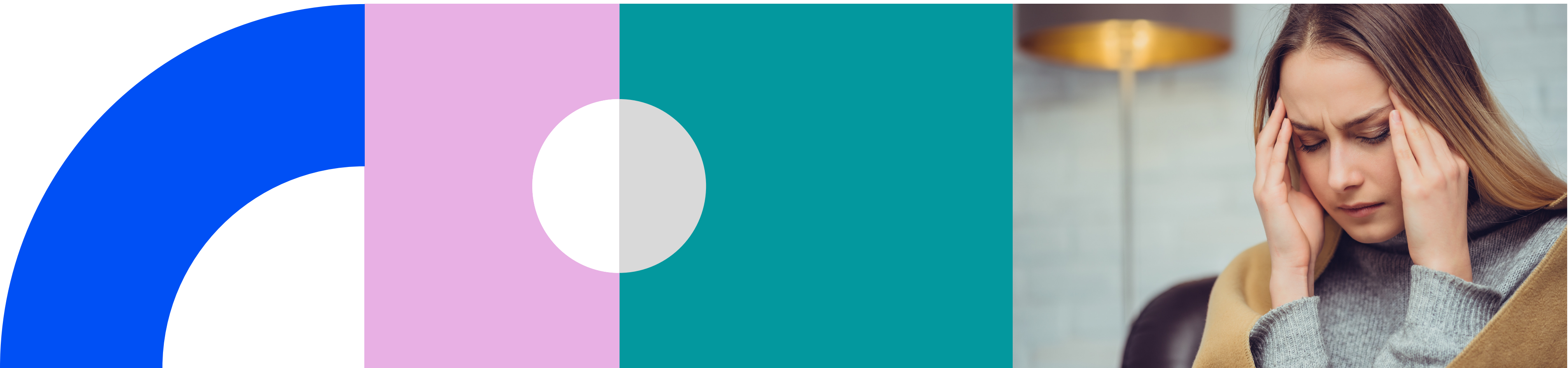


# Stroke Prediction



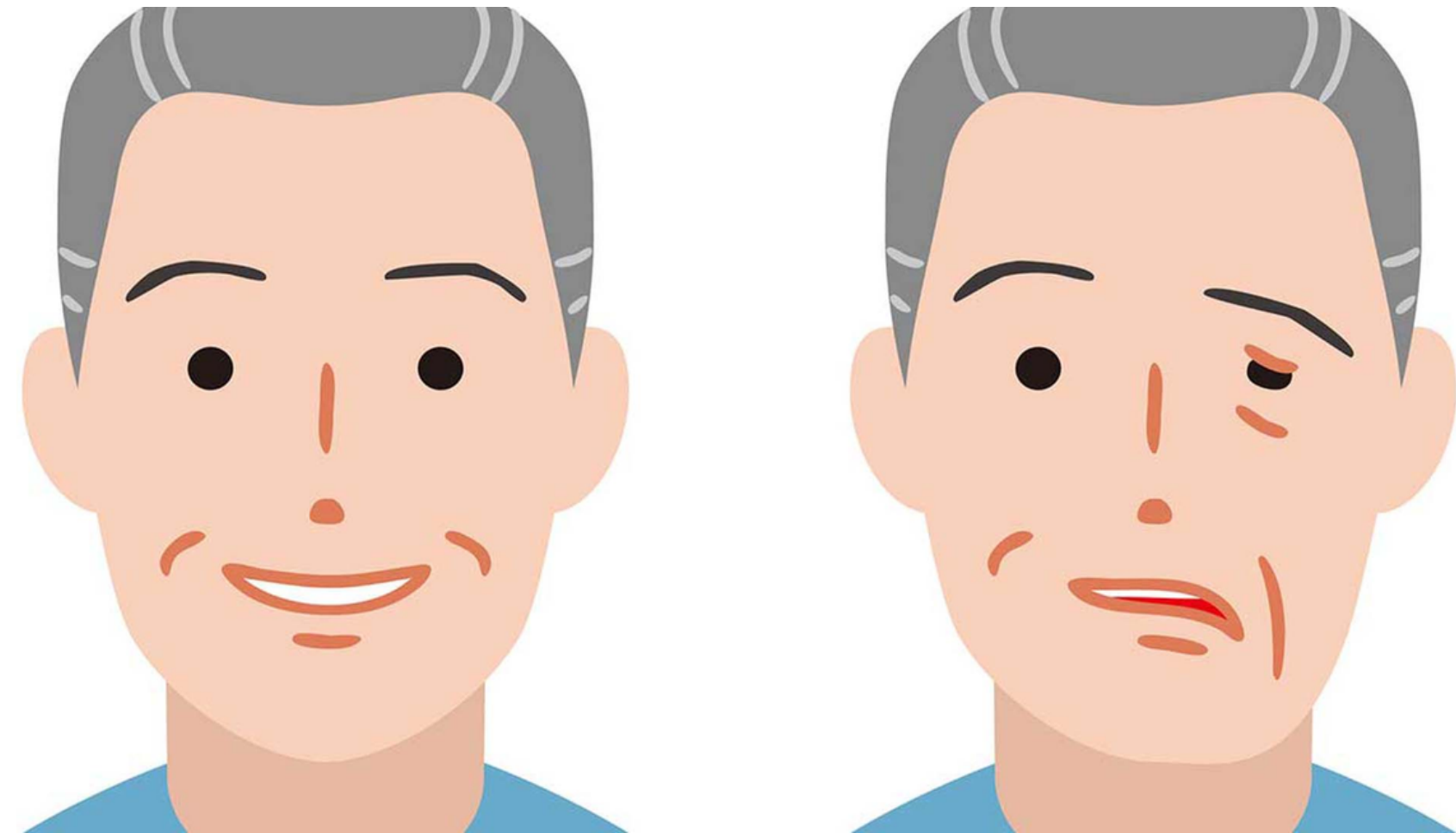
# Infarto cerebral

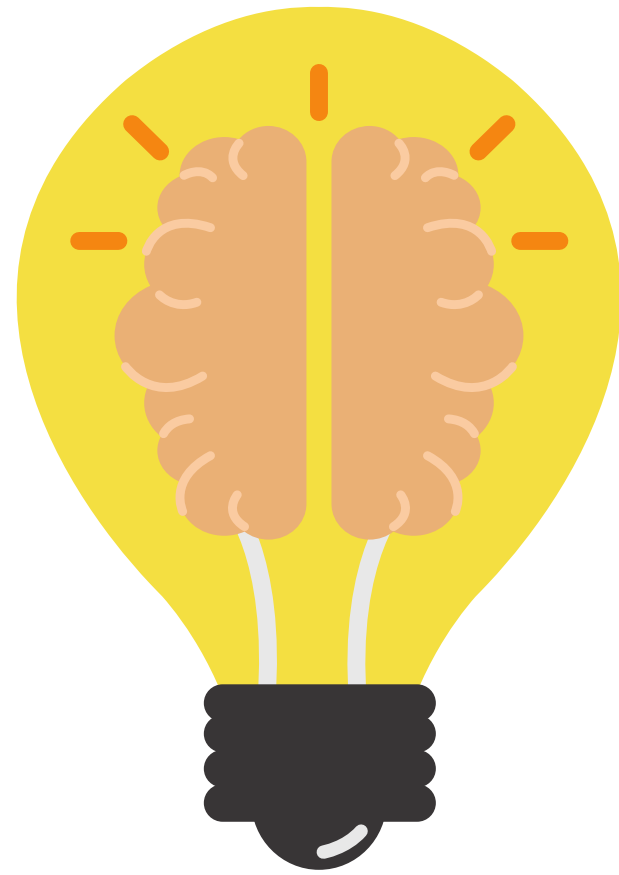
Accidente cerebral, ACV, Apoplejía, Ataque cerebral o derrame cerebral

Sucede cuando alguno de los vasos sanguíneos que lleva sangre al cerebro se bloquea provocando la muerte de las células cerebrales por la falta de oxígeno.

El cerebro coordina todo lo que hacemos y cuando una parte de él no recibe el oxígeno, puede provocar discapacidades e incluso la muerte.

Entre más prolongado sea el tiempo durante el cuál el cerebro esta privado de oxígeno la discapacidad puede ser más severa





# Objetivos

- Analizar la relación del estilo de vida de los pacientes con posibilidad de que tengan un infarto cerebral
- Diseñar un modelo que permita conocer la probabilidad de que los pacientes tengan un infarto cerebral en función de su estilo de vida y estado de salud





# 1 Data Acquisition

## DESCRIPCIÓN DE LAS VARIABLES



### Información del paciente

- Grupo de edad
- Género del paciente



### Estado de Salud

- Padecimiento de hipertensión
- Antecedentes de enfermedades del corazón
- Nivel de glucosa
- Índice de Masa Corporal



### Estilo de vida

- Estado civil
- Tipo de trabajo
- Tipo de residencia
- Hábito de Fumar

# 2 Data Wrangling



## Exploración del dataset

- Estructura
- Tamaño
- Datos Nulos
- Duplicados



## Estructuración y limpieza

- ID
- Género
- Edad
- Datos Nulos



# Insights

- Contamos con información de 5,100 pacientes
- El promedio de edad de los pacientes es de 43 años.
- El promedio de masa corporal de los pacientes es de 29 lo cual corresponde a sobrepeso.
- El promedio de nivel de glucosa de los pacientes es de 109 lo que se considera normal
- Contamos con información de 11 variables que describen el estilo de vida y estado de salud de los pacientes

# 3 Análisis Exploratorio de Datos o EDA



Análisis  
Univariado



Análisis  
Bivariado



Análisis  
Multivariado

# Análisis Univariado

La muestra está compuesta mayormente por mujeres mayores de 50 años.

La mayoría son casados y trabajan en el sector privado.

Tienen sobrepeso, sus niveles de glucosa son normales, no tienen enfermedades del corazón ni hipertensión y no tienen antecedentes de ACV

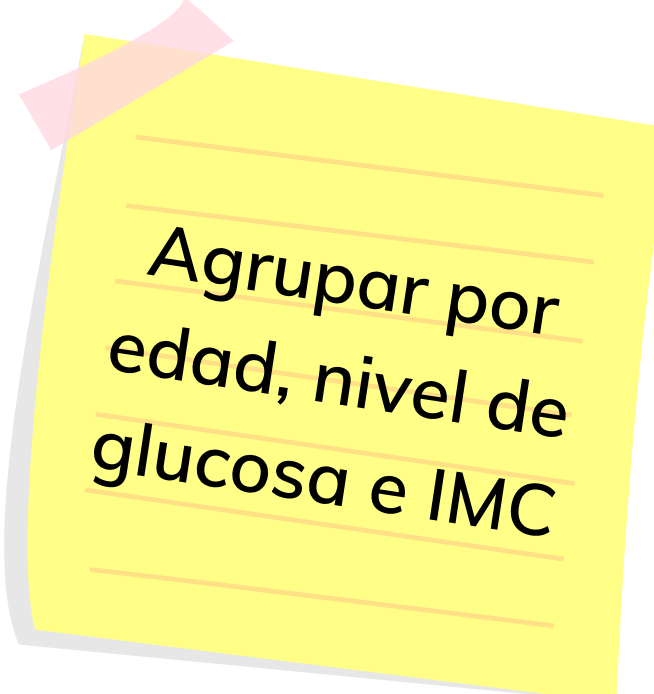
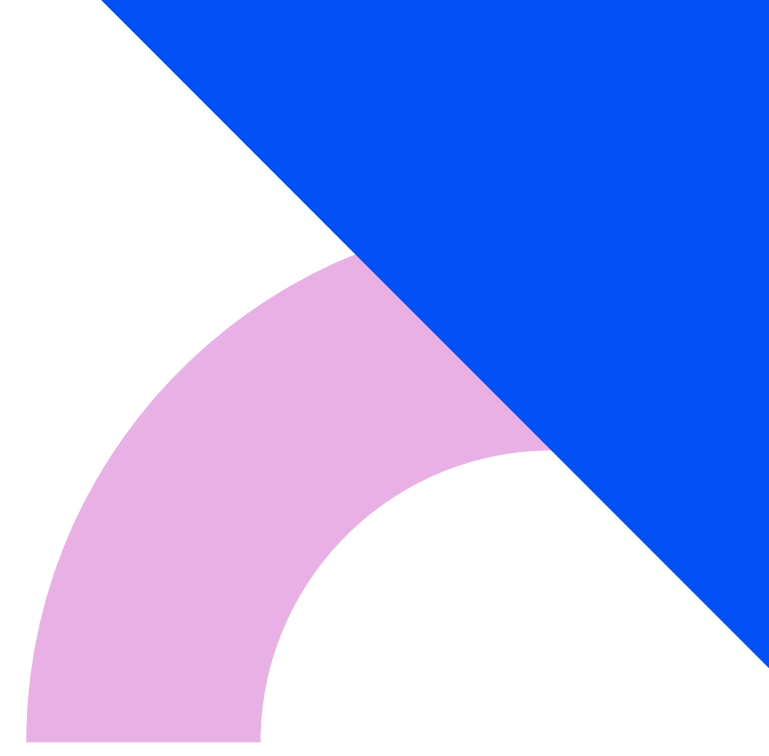
Corrección de  
outliers para  
IMC y nivel de  
glucosa





# Análisis Bivariado

- Los pacientes con problemas cardíacos, diabetes, colesterol alto y el tabaquismo son los que tienen más antecedentes de ACV.
- Uno de los factores por los que se incrementan los casos de infartos cerebrales es por el tabaquismo.
- Algunos estudios sugieren que los hombres presentan mayores casos de infartos cerebrales, sin embargo en nuestra muestra el género no influye en el número de casos



Agrupar por  
edad, nivel de  
glucosa e IMC

# ÁNALISIS MULTIVARIADO

- Uno de los factores de riesgo de tener un infarto cerebral es la edad del paciente, lo cual se confirma en nuestra muestra.
- Cuando los pacientes tienen más de una enfermedad al mismo tiempo los casos de infarto cerebral disminuyen.
- Cuando consideramos más de un aspecto en el estilo de vida de los pacientes los casos de antecedentes de infarto cerebral aumentan. Principalmente para pacientes que están casados, los que nunca han fumado y los que viven en zonas urbanas.



# 4 MODELOS

## PREPARACIÓN DEL DATASET PARA EL MODELADO



Get  
Dummy

Para crear una  
columna de cada  
posible valor de  
nuestras variables



PCA

Para corregir la  
maldición de la  
dimensionalidad



Oversam  
pling

Para compensar  
nuestra base  
desbalanceada

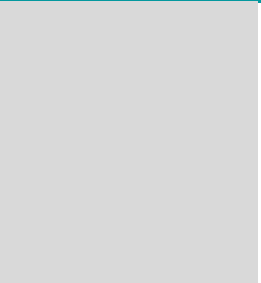
# RESULTADOS DE LOS MODELOS

Modelo	Exactitud	Precisión	Sensibilidad	Score F1	AUC
Árbol de decisión	77.86	73.73	87.63	79.91	0.8216
Regresión Logística	89.34	91.18	87.22	89.15	0.9579
Random Forest	93.90	93.12	94.81	93.96	0.9806
Adaboost	85.92	83.49	89.54	86.41	0.9321
Gradient Boosting	86.17	85.35	87.32	86.33	0.9337
Light GBM	93.99	92.72	95.47	94.07	0.9796
Xgboost	93.95	92.85	95.22	94.02	0.9802





# 5 Conclusiones y elección del modelo



En conclusión, los pacientes con problemas cardíacos, diabetes, colesterol alto y el tabaquismo son los que tienen más antecedentes de ACV.

Las variables que más influencia tienen son la edad del paciente, antecedentes de enfermedades del corazón o hipertensión y su estado civil.

Después de ejecutar todos los modelos y comparar sus métricas nos quedaremos con Light GBM porque es el que tiene un mayor nivel de sensibilidad, la cual es la métrica más importante para el proyecto.

- Certeza en la predicción vs equivocaciones : 95%
- Porcentaje de predicciones correctas 92.72%
- Predicciones correctas de pacientes con probabilidad de tener un ACV 47.73%