



# **PREDICIR UN ACCIDENTE CEREBROVASCULAR**

Lucas Aguilera, Claudio Bórquez, Josefa Fernández

# Introducción

---

**El problema que va abordar el proyecto es:**

Predecir si es probable que un paciente sufra un accidente cerebrovascular en función de los parámetros de entrada como el sexo, la edad, diversas enfermedades y el tabaquismo.

Optimiza proceso médico.

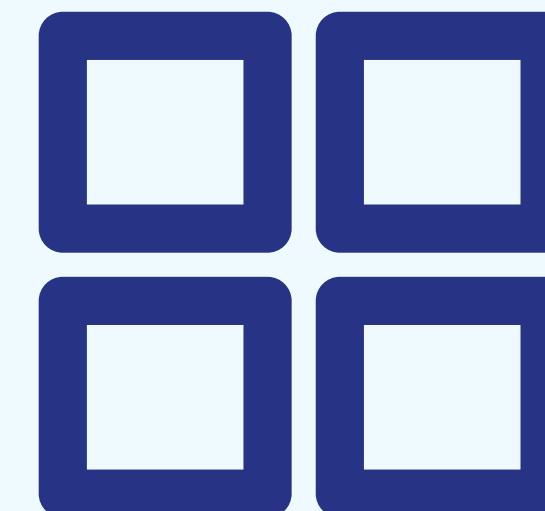


# Datos a utilizar

## Datos clínicos, centrados en accidentes cerebrovasculares



**Datos** abiertos,  
extraídos de  
Kaggle.  
Volumen: (5111,  
12)

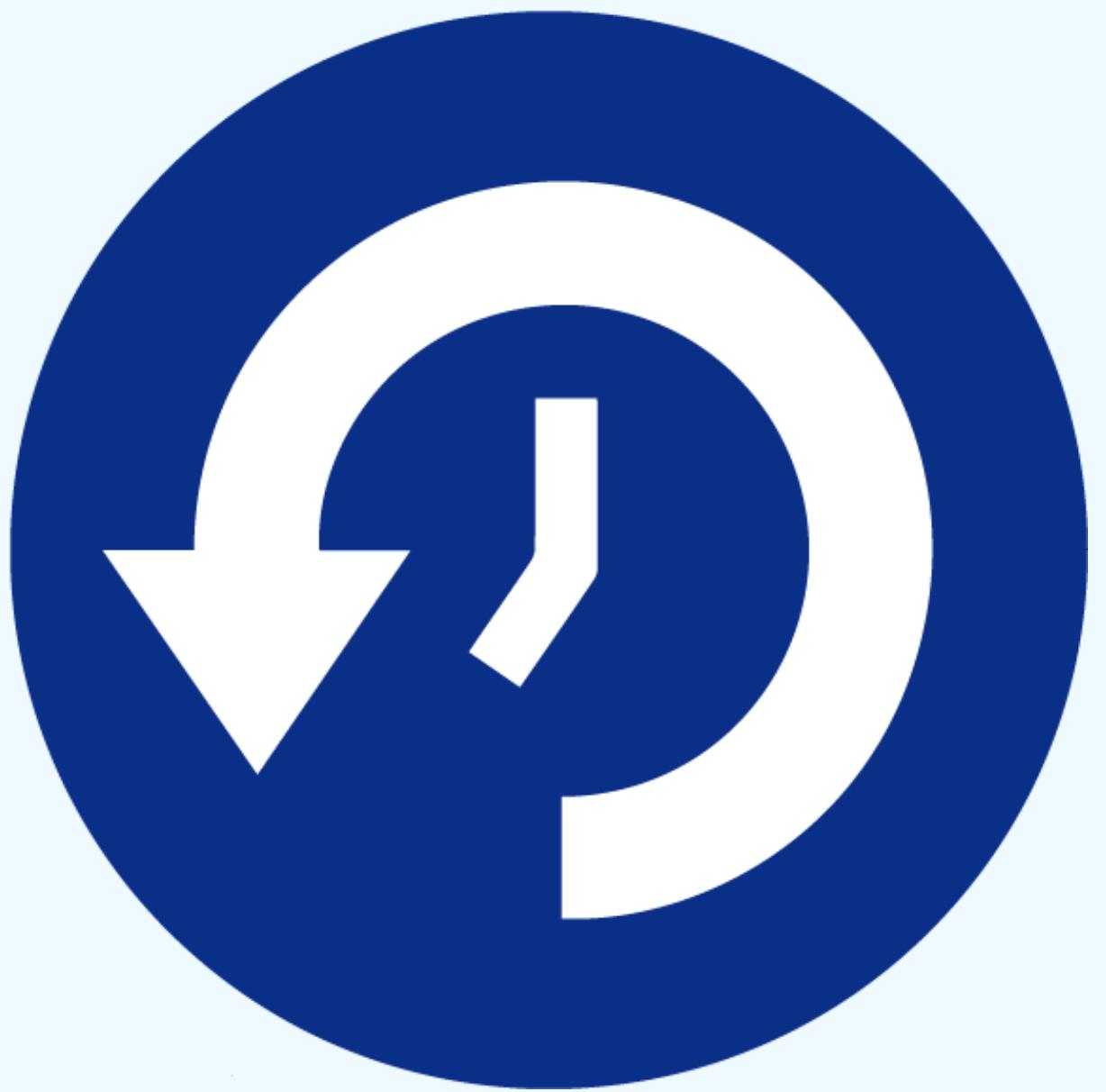


**Categoricos:**  
genre,  
ever\_married,  
work\_type,  
residence\_type,  
smoking\_status

1  
2  
3

**Numericos:**  
Id, age,  
hypertension,  
heart\_disease,  
avg\_glucose\_level, bmi, stroke

# BASELINE





# Análisis de los datos

- 5110 datos y 12 columnas
- Valores nulos en columna BMI,  
valores numéricos y categóricos.
- Parcial escasez de datos

# Pre- procesamiento

Se hace el preprocessamiento adecuado, considerando la limpieza de datos y la posible normalización.

- **Remover columnas**
- **Manejo datos categóricos con One Hot Encoding**
- **Normalizar variables**
- **Manejo de valores nulos**





# Inicio modelaje

- División de datos en train y test
- MLP
- Regresión Logística

# Evaluación del modelo

Entrenamiento de modelos y sus métricas de evaluación.

- **Entrenar modelo para MLP y RL**
- **Métricas de evaluación para MLP y RL**



# Análisis de resultados

- Precisión para los datos 1 en "stroke":

0%  
MLP

vs  
15%  
RL

- Problemas con la clasificación y los modelos que no logran aprender patrones.



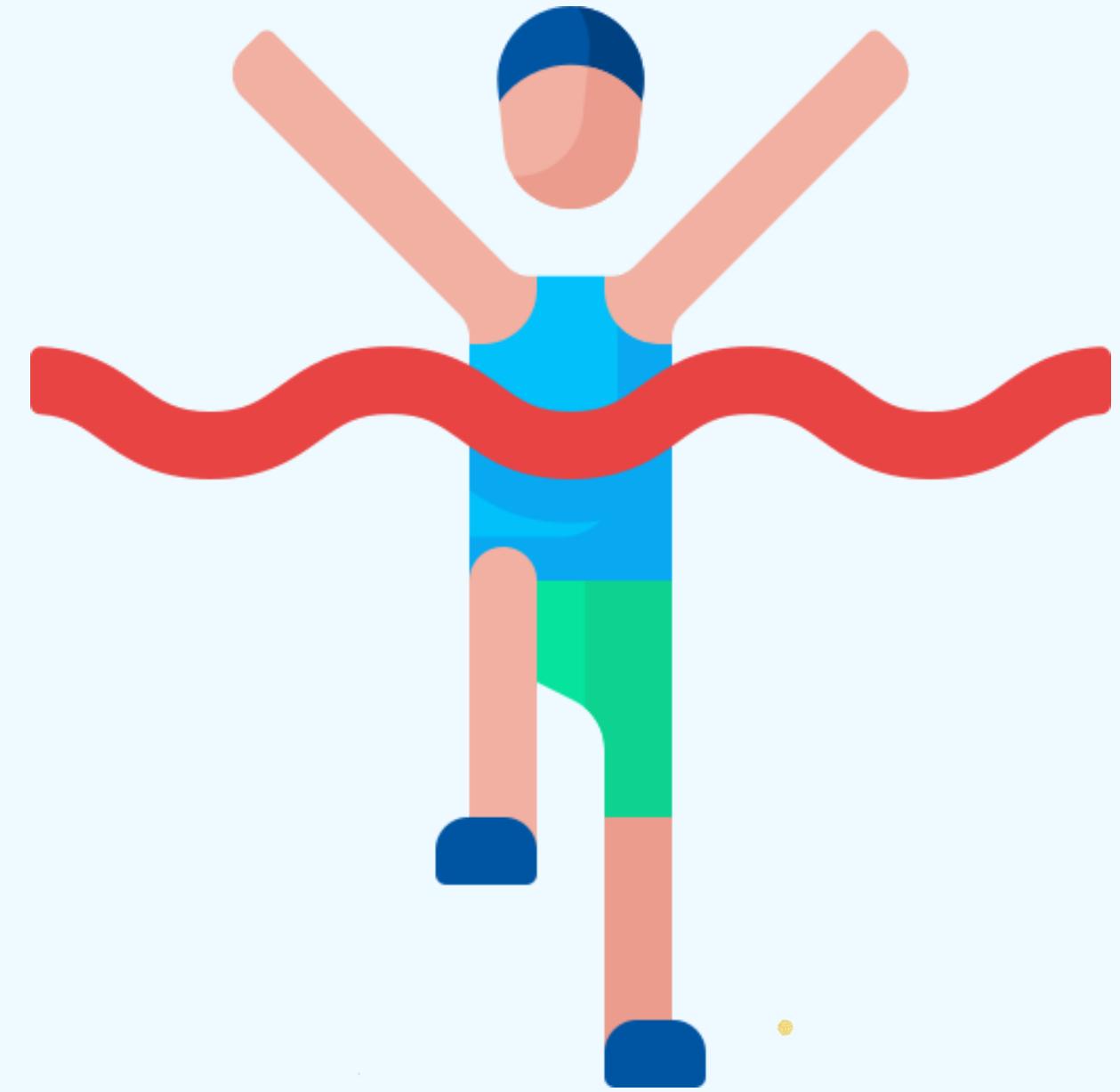
# Feedback

---



- Tienen desbalance de clases.
- Hacer aumento de datos o bien submuestrear la clase mayoritaria hasta lograr desbalance. Construir varios datasets con esta técnica y construir un ensemble.
- El modelo no aprende.
- Van a trabajar con regresión logística, descartando MLP.

# ENTREGA FINAL





# Manejo del desbalance de datos

Se realiza un **submuestreo** de la clase mayoritaria:  $\text{stroke} = 0$ . Tomando muestras de 249 filas de  $\text{stroke\_0}$ , y tomando 22 muestras al azar.

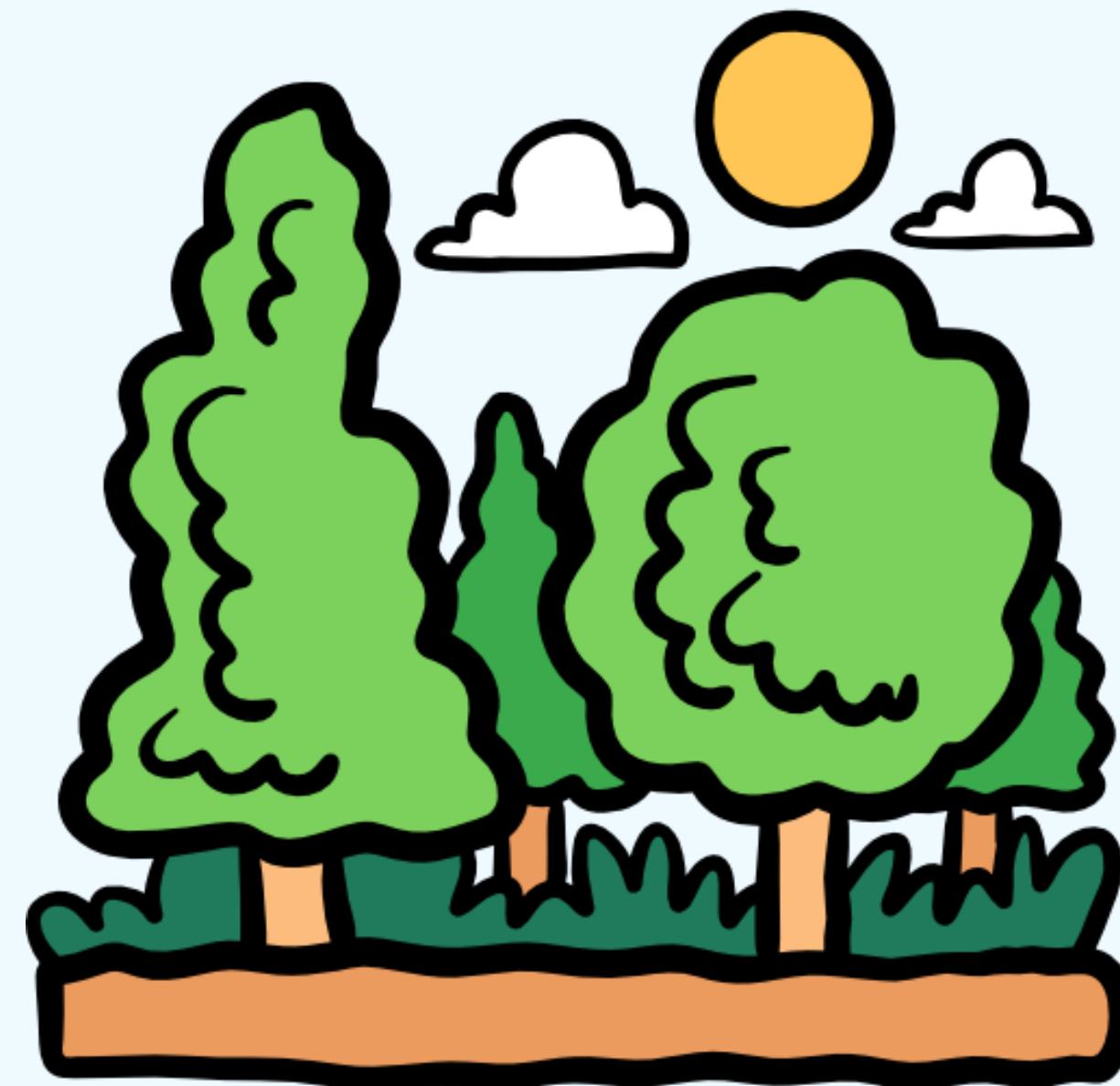
- 249 filas con  $\text{stroke}=1$
- 4861 filas con  $\text{stroke}=0$

# Regresión Logistica

Se entrena distintos modelos de RL en base a cada uno de estos grupos.

- Predicción en cada modelo, usando los mismos datos en todos y se almacena las pred de cada modelo para crear el ensemble.
- **Voting Classifier**
- **Gradient Boosting Classifier**



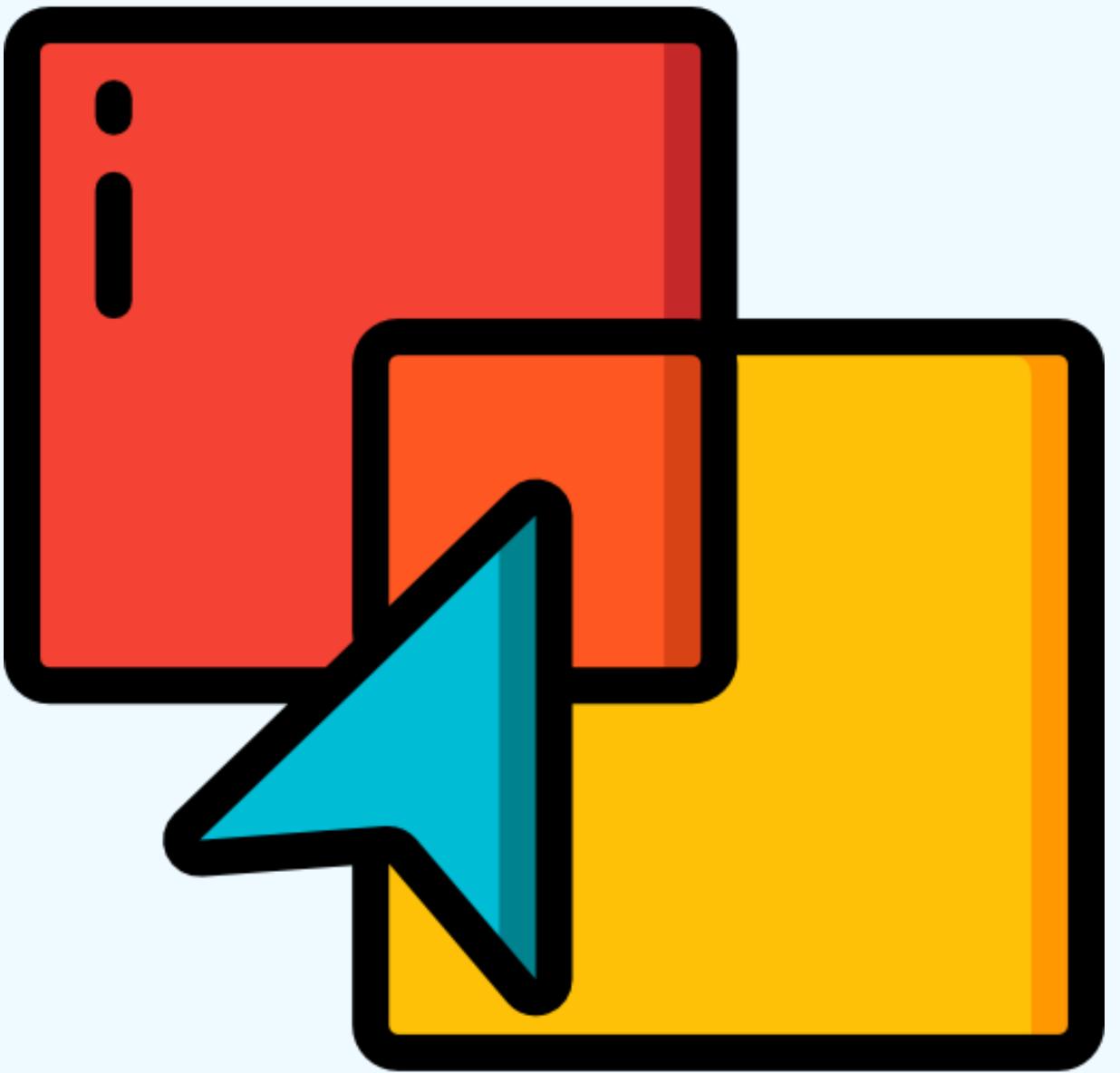


# Random Forest

Se usan modelos de Decision Trees en lugar de RL y se ensamblan en un modelo de Random Forest.

# Ensemble sin aplicar submuestreo

- Se usa AdaBoost.



# Comparación de resultados

|                     | MLP sin<br>submuestreo | RL sin<br>submuestreo | RL con sub. y<br>ensemble Voting<br>Classifier | RL con sub. y<br>ensemble<br>Gradient Boosting<br>Classifier | Decision Trees<br>ensamblado<br>con Random<br>Forest | Ensemble<br>AdaBoost sin<br>sub. |
|---------------------|------------------------|-----------------------|--|--|--|----------------------------------|
| Accuracy            | 0.95                   | 0.75                  | 0.76   | 0.70   | 0.95   | 0.95                             |
| Precision<br>0 vs 1 | 0.95 vs 0.00           | 0.99 vs 0.15          | 0.72 vs 0.82                                   | 0.66 vs 0.75   | 0.95 vs 0.33   | 0.95 vs 0.25                     |

# GRACIAS

Lucas Aguilera, Claudio Bórquez, Josefina Fernández