Predicción de Diabetes

La diabetes es una enfermedad crónica que requiere diagnóstico temprano para mejorar la calidad de vida. Usamos una base de datos de Diabetes para aplicar aprendizaje supervisado y predecir la diabetes basándonos en características médicas clave. El objetivo es identificar los mejores modelos predictivos y las variables más influyentes.

Base de Datos

+21

ATRIBUTOS

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI
- Pedigree Function
- Age

En base a estos, se utilizaron diversos modelos de machine learning, incluyendo regresión logística, Random Forest, y Support Vector Machines, etc.

Se dividió la BD en:

ENTRENAMIENTO 80%

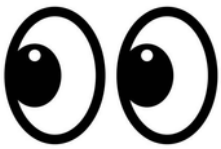
PRUEBA 20%

Además....

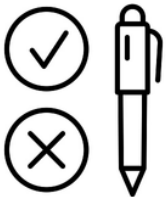
Se realizó:

Limpieza de datos para manejar valores nulos

Normalización de datos



con



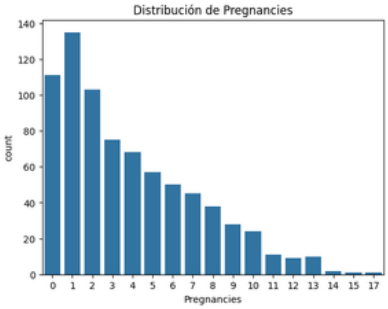
Identificar falsos positivos y negativos, donde es fundamental reducir estos ultimos para no pasar por alto posibles casos de diabetes

ya que aumenta el riesgo de aumentar el grado de la enfermedad o incluso de la muerte.

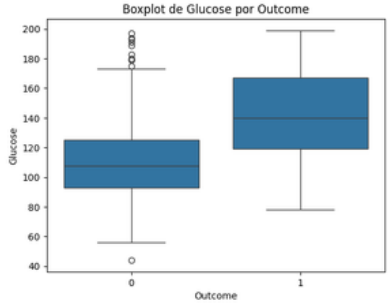
Atributos mas importantes

GLUCOSA
0,49

BMI (IMC)
0,31

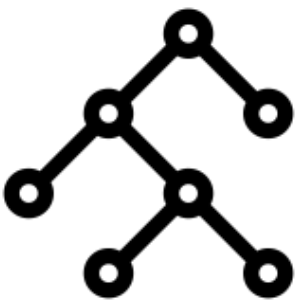


La mayoría de las mujeres de la BD ha tenido de 0 a 2 embarazos



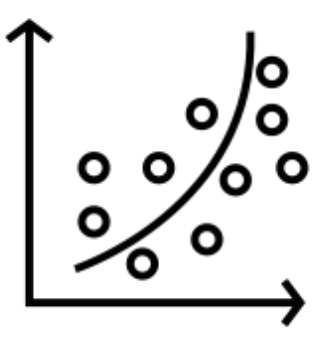
La mayoria de las mujeres con niveles de glucosa mayor a 120 tienen diabetes

Las métricas utilizadas para evaluar los modelos fueron la precisión, el área bajo la curva ROC (AUC), y el recall, con el objetivo de seleccionar el modelo que mejor se ajuste al problema de predicción.



Random Forest

El resultado en la matriz de correlacion fue 0,59 por lo que se afinó el modelo obteniendo una precisión de 0,84



Regresión Logistica

A raíz de la curva ROC-AUC, se obtuvo un valor de 0,67 por lo que se afinó el modelo obteniendo 0,82; mejor valor en términos de discriminación en un problema de clasificación como este.