Tecnológico de Monterrey

Campus:
Monterrey

Inteligencia Artificial Avanzada para la Ciencia de Datos (Gpo 102)

Curso:
TC3006C.102

Módulo 2: Portafolio 2

Daniel Sánchez Villarreal A01197699

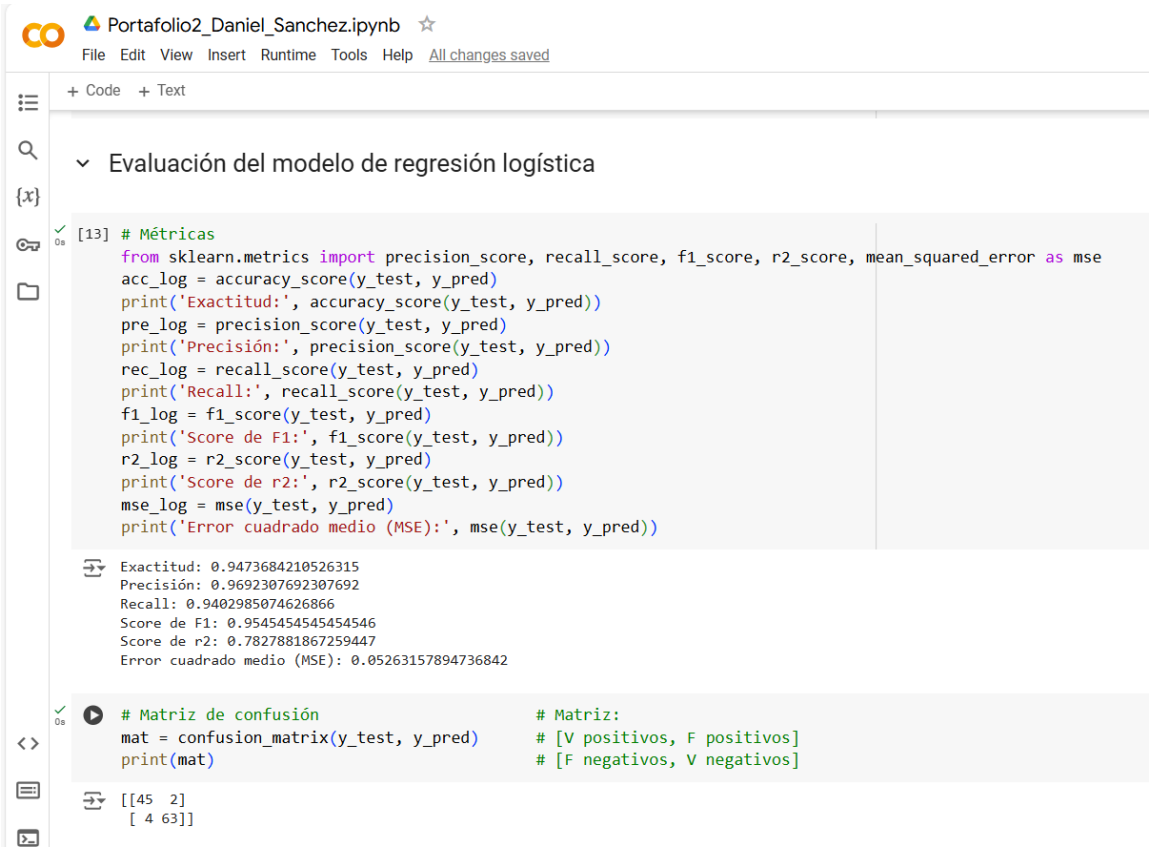
Lugar y Fecha:
Monterrey, Nuevo León
7 de septiembre de 2024

En este Portafolio 2, decidí usar el dataset de breast_cancer y entrenar 6 modelos, los cuales fueron regresión logística, árboles de decisión, random forest, KNN, SVC, y Gaussian Naive Bayes. Esto con el fin de evaluar cuál de todos tiene el mejor desempeño y ese sería el que escogería como mi mejor modelo para después usar GridSearch para buscar e implementarle los mejores hiperparámetros y evaluar su rendimiento para finalmente hacer algunas predicciones con ese modelo mejorado.

Para esto, lo primero que hice fue la preparación de los datos. En este caso, definí la Y como la variable target, y la X para el resto de features en el dataset. Usé target_names para ver la clase de datos para la variable target, los cuales con este comando pude observar que cada instancia (sample) tenía un valor ya sea de benigno o maligno en esa variable target. Entonces lo que hice a continuación fue transformar esos datos para establecer benigno como 0 y maligno como 1 en todas las instancias de dicha variable target en el dataset.

Posteriormente, verifiqué si había o no datos nulos en el dataset y tras ver que no había ningún dato nulo en ninguna de las features, procedí con el entrenamiento de mis 6 modelos para después evaluarlos todos para poder saber cuál de todos tiene el mejor rendimiento y ese será el que voy a elegir para usar GridSearch para implementarle los mejores hiperparámetros y evaluar su rendimiento para después hacer predicciones con dicho modelo.

Comencé con el modelo de regresión logística, el cual tras evaluar sus métricas obtuve los siguientes resultados:



Después, en el de árboles de decisión obtuve lo siguiente: