Proyecto final

Predicción de precipitaciones (LONDRES)

García Angeles Belén

Hipótesis y Descripción del Problema

El objetivo de este proyecto fue analizar un conjunto de datos meteorológicos históricos para predecir eventos de lluvia. La hipótesis planteada fue que las condiciones atmosféricas como la temperatura, la radiación solar y la cobertura nubosa tienen una correlación significativa con la ocurrencia de lluvia. Utilizando datos como temperaturas máximas, medias y la cobertura de nubes, se buscó desarrollar un modelo predictivo que maximice el área bajo la curva ROC (ROC-AUC) como métrica principal.

Análisis Realizado

Preprocesamiento de los Datos

- **Estandarización de las variables**: Se aplicó **Standard Scaler** para normalizar las características, lo cual es crucial cuando se usan modelos que son sensibles a la escala de las variables, como las redes neuronales. Esto ayuda a mejorar la estabilidad del modelo y facilita la convergencia durante el entrenamiento.
- **División del conjunto de datos**: Los datos fueron divididos en conjuntos de **entrenamiento** y **prueba** para asegurar que los modelos pudieran generalizar bien a nuevos datos no vistos.

Modelos Evaluados

1. **Árboles de decisión**:
    - Los árboles de decisión fueron seleccionados por su capacidad para manejar datos no lineales y proporcionar interpretabilidad en los resultados. Son adecuados para este tipo de problema debido a que pueden modelar interacciones complejas entre las variables.
2. **Regresión logística**:
    - La regresión logística fue considerada por su simplicidad y eficiencia computacional. A pesar de su naturaleza lineal, sigue siendo muy efectiva para muchos problemas de clasificación binaria, especialmente cuando las relaciones entre las variables son lineales o aproximadas.
3. **Redes neuronales**:
    - Las redes neuronales fueron elegidas debido a su capacidad para aprender patrones complejos y no lineales en los datos. En este caso, se utilizó la biblioteca **Optuna** para optimizar los hiper parámetros, como el número de neuronas, funciones de activación, tamaño de lote y tasas de aprendizaje, para maximizar la **ROC-AUC** en el conjunto de pruebas.

Proceso de Optimización

La función objetivo de **Optuna** fue maximizar el **ROC-AUC** en el conjunto de pruebas, dado que esta métrica proporciona una evaluación más completa del desempeño del modelo en clasificación binaria.Se realizaron **50 iteraciones** con diferentes combinaciones de hiper parámetros para encontrar la configuración que mejorará la capacidad predictiva del modelo.

Evaluación de Resultados

Los modelos fueron evaluados utilizando el **conjunto de pruebas** y se usaron métricas adicionales como **precisión** y **pérdida binaria cruzada**. Estas métricas permitieron una evaluación más profunda, más allá del ROC-AUC, para entender mejor el desempeño de cada modelo.

Resultados Obtenidos

El mejor modelo fue una **red neuronal**, que alcanzó un **ROC-AUC de 83%**. Sin embargo, a pesar de la optimización de los hiper parámetros con **Optuna**, las mejoras en el ROC-AUC fueron marginales (sólo décimas adicionales). Este resultado sugiere que, aunque las redes neuronales son potentes, podrían no estar explotando completamente el potencial de los datos en este caso.

La **regresión logística** también demostró un desempeño competitivo, con un **ROC-AUC de 82%**. Esto sugiere que este modelo es igualmente adecuado para el problema, con la ventaja de ser más sencillo y menos costoso computacionalmente.

Conclusiones

- **Confirmación parcial de la hipótesis**: Las variables climáticas seleccionadas (como temperatura, radiación solar y cobertura nubosa) demostraron tener una relación significativa con la ocurrencia de lluvia, lo que valida parcialmente nuestra hipótesis inicial.
- **Rendimiento de los modelos**: Si bien la red neuronal fue el modelo más preciso, la diferencia de rendimiento con la regresión logística fue mínima. Esto podría sugerir que el conjunto de datos tiene límites intrínsecos en su capacidad predictiva, lo que implica que ningún modelo es capaz de mejorar significativamente más allá de un cierto punto.
- **Consideraciones**: Aunque las redes neuronales superaron a otros modelos en términos de precisión, se considera el modelo de **regresión logística** para implementaciones futuras. Este modelo es más eficiente computacionalmente y más fácil de interpretar, lo que podría ser una ventaja en escenarios de producción o cuando se requiera explicar las predicciones de manera clara.