Introducción a la Regresión Logística:

La regresión logística es un método de aprendizaje supervisado que se utiliza para problemas de clasificación. En lugar de predecir valores numéricos como en la regresión lineal, la regresión logística predice la probabilidad de que una observación pertenezca a una clase específica. Es ampliamente utilizada en machine learning y estadísticas para resolver problemas de clasificación binaria o multiclase.

Elección de Implementación:

Opté por utilizar la implementación de regresión logística de scikit-learn en lugar de crear una desde cero debido a su eficiencia y capacidad para obtener resultados sólidos en menos tiempo. Los resultados obtenidos con esta implementación fueron prometedores, alcanzando una precisión del 0.88. La curva de aprendizaje muestra un buen rendimiento, los coeficientes son coherentes y significativos. Dado que este modelo es relativamente simple, pudimos dedicar tiempo y esfuerzo para ajustar los hiperparámetros y mejorar aún más los resultados.

Separación y Evaluación de Datos:

Dividimos nuestros datos en conjuntos de entrenamiento (train) y prueba (test) utilizando una división de 90/10, utilizando un valor de random state igual a 42 para asegurar la reproducibilidad. Esta división nos permitió evaluar el modelo de manera efectiva y ajustar los hiperparámetros.

Diagnóstico del Sesgo:

Observamos que los datos muestran un sesgo hacia la derecha en la variable 'Age'. Aunque su distribución parece normal, es importante investigar más a fondo para determinar si estos valores son válidos o si se deben eliminar.

Diagnóstico de la Varianza:

Calculamos la varianza de los datos y encontramos que era significativamente alta en relación con la escala de los datos. Este alto valor de varianza se relaciona con el sesgo observado en la variable 'Age', lo que sugiere una distribución desigual. Por lo tanto, calificamos esto como alta varianza.

Diagnóstico del Ajuste del Modelo:

Para evaluar el ajuste del modelo, analizamos la curva de aprendizaje. Observamos que los conjuntos de entrenamiento y validación comienzan a separarse y luego convergen, lo que es un buen indicador de que el modelo se está ajustando correctamente. No se observa subajuste (underfitting) ni sobreajuste (overfitting).

Optimización de Hiperparámetros:

Utilizamos el framework Optuna para mejorar iterativamente la métrica de precisión (accuracy) del modelo. A través de esta optimización, logramos mejorar la métrica de precisión hasta alcanzar un valor de 0.88, lo que indica una mejora sustancial en el rendimiento del modelo.

En resumen, la implementación de regresión logística utilizando scikit-learn ha demostrado ser efectiva en este conjunto de datos. Identificamos y abordamos problemas de sesgo y alta varianza, y logramos un ajuste adecuado del modelo mediante la optimización de hiperparámetros. Los resultados son prometedores y respaldados por indicadores claros y gráficas comparativas en nuestro informe.