Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 102)
F2006B.101

Portafolio Análisis

Estudiante:
Adrián Chávez Morales A01568679

Septiembre 7, 2024

1. Introducción

El propósito de este análisis es diagnosticar y mejorar el desempeño de un modelo de regresión logística aplicado a un dataset de enfermedades del corazón. El modelo ha sido entrenado para predecir si un paciente tiene una enfermedad cardíaca en base a diferentes características, como la edad, el sexo, el nivel de colesterol, entre otras. Utilizando técnicas de regularización, se busca mitigar problemas de varianza y bias, ajustando mejor el modelo a los datos.

El código que acompaña este reporte se puede dividir en etapas, comenzando con la carga y preprocesamiento de datos, seguido por la creación y evaluación del modelo de regresión logística. Posteriormente, se realiza un diagnóstico del grado de bias y varianza para determinar si el modelo está en un estado de *underfitting* u *overfitting*, y finalmente se aplican técnicas de regularización para mejorar el rendimiento del modelo.

2. Funcionamiento del código

2.1. Preprocesamiento de Datos

Primero, los datos son cargados al código. Se seleccionan los features relevantes para el modelo, al igual que el label. Luego, se convierte cualquier columna con valores booleanos a enteros para garantizar que todos los datos sean compatibles con el algoritmo de regresión logística. Posteriormente, los datos se dividen en conjuntos de entrenamiento y prueba, y se normalizan utilizando `StandardScaler` para escalar las características numéricas.

2.2. Creación del Modelo

El modelo de regresión logística es configurado sin regularización inicial y se entrena utilizando el conjunto de datos de entrenamiento. Después del entrenamiento, se realizan predicciones tanto en el conjunto de prueba como en el de entrenamiento. Esto permite evaluar el rendimiento del modelo para diagnosticar bias y varianza.

3. Diagnóstico Inicial del Modelo

3.1. Bias (Sesgo)

El bias o sesgo se refiere a que tan bien el modelo se ajusta a los datos de entrenamiento. En este análisis, el modelo tiene una precisión de `train_accuracy` en el conjunto de entrenamiento. Un bias bajo implica que el modelo captura bien los patrones de los datos de entrenamiento, mientras que un bias alto sugiere que el modelo está subajustado (*underfitting*).

3.2. Varianza

La varianza mide la capacidad del modelo para generalizar a nuevos datos. Comparando la precisión en el conjunto de entrenamiento con la del conjunto de prueba, se observa que el modelo tiene una precisión de `test_accuracy` en el conjunto de prueba. Si existe una gran diferencia entre estos valores, el modelo puede estar sobreajustado (*overfitting*), lo que significa que se ha adaptado demasiado a los datos de entrenamiento y no generaliza bien.

3.3. Ajuste del Modelo

Basándonos en la comparación de la precisión entre el conjunto de entrenamiento y prueba, se puede hacer el siguiente diagnóstico:

- **Underfitting**: Si tanto el entrenamiento como la prueba tienen baja precisión, el modelo no captura suficientemente bien los patrones de los datos.
- **Overfitting**: Si la precisión en el conjunto de entrenamiento es significativamente mayor que en el conjunto de prueba, el modelo está sobreajustado.
- **Ajuste Adecuado**: Si la precisión en ambos conjuntos es alta y similar, el modelo está bien ajustado.

4. Diagnóstico Inicial del Modelo

El modelo fue entrenado sin regularización y arrojó los siguientes resultados en el conjunto de prueba:

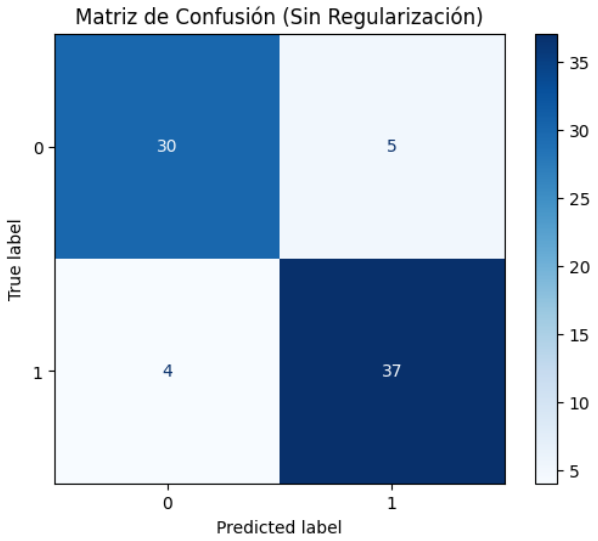
```
Reporte de Clasificación (Sin Regularización):
              precision    recall  f1-score   support

           0       0.88      0.86      0.87        35
           1       0.88      0.90      0.89        41

    accuracy                           0.88        76
   macro avg       0.88      0.88      0.88        76
weighted avg       0.88      0.88      0.88        76
```

El modelo alcanzó un accuracy del 88 %, con un puntaje F1 de 0.87 para la clase 0 (ausencia de enfermedad) y 0.89 para la clase 1 (presencia de enfermedad). Esto sugiere un buen desempeño inicial, pero dado que el modelo puede estar sobreajustado, realizamos un análisis más detallado para diagnosticar posibles problemas de varianza.

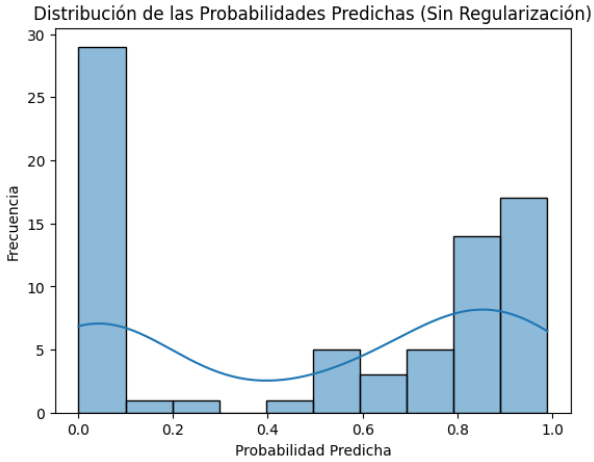
4.1. Bias y Varianza

El sesgo (*bias*) del modelo es bajo, ya que logra un performance/accuracy alto en el conjunto de entrenamiento (.86). Sin embargo, existe una leve diferencia en esta métrica entre el conjunto de entrenamiento y en el conjunto de prueba, lo que sugiere la existencia una varianza moderada. Esto significa que el modelo podría estar sobreajustado (*overfitting*) y que su capacidad de generalización a nuevos datos es algo limitada.

Gráfica 1: Matriz de Confusión (Sin Regularización)



Gráfica 2: Distribución de probabilidades predichas (Sin Regularización)



4.2. Conclusión del Diagnóstico

El análisis inicial sugiere que el modelo está bien ajustado a los datos de entrenamiento, pero presenta una ligera varianza. Por lo tanto, existe la posibilidad de mejorar la generalización del modelo aplicando regularización para reducir el sobreajuste y mantener un desempeño similar o mejor en los datos de prueba.

5. Mejora del Modelo con Regularización

Para abordar el problema de sobreajuste del modelo, se aplicó regularización L2 con el objetivo de penalizar los coeficientes del modelo y reducir su sensibilidad a las variaciones en los datos de entrenamiento. Esto ayuda a mejorar la capacidad de generalización del modelo.

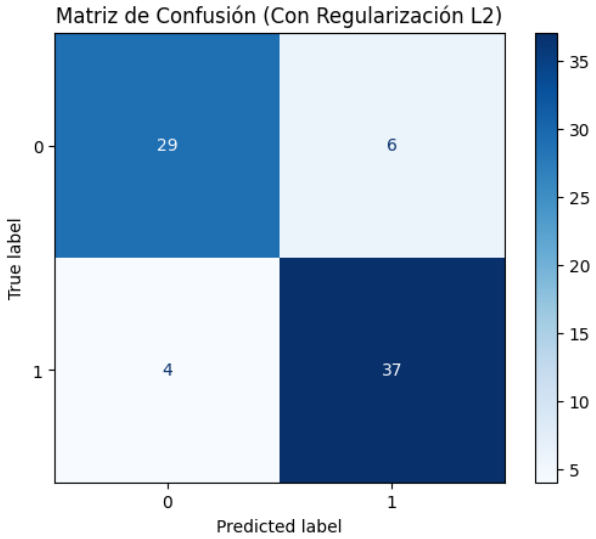
5.1. Resultados con Regularización L2

Después de aplicar regularización, los resultados obtenidos fueron los siguientes:

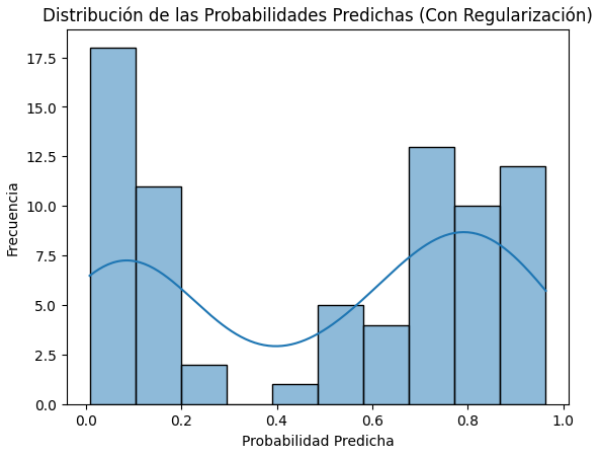
```
Precisión del modelo con regularización L2: 0.87

Reporte de Clasificación (Con Regularización L2):
              precision    recall  f1-score   support

           0       0.88      0.83      0.85        35
           1       0.86      0.90      0.88        41

    accuracy                           0.87        76
   macro avg       0.87      0.87      0.87        76
weighted avg       0.87      0.87      0.87        76
```

El accuracy del modelo con regularización L2 fue de 87 %, esto fue una ligera disminución en comparación con el modelo sin regularización. Sin embargo, el modelo ahora tiene un mejor balance entre bias y varianza, porque al aplicar regularización L2, se introduce una penalización que reduce los coeficientes de los parámetros del modelo, haciendo que el modelo sea más conservador al aprender de los datos. Esto tiene el efecto de reducir la varianza porque el modelo es menos sensible a los cambios en los datos de entrenamiento lo que se traduce en una mayor capacidad de generalización.

Gráfica 3: Matriz de Confusión (Con Regularización L2)



Gráfica 4: Distribución de Probabilidades Predichas (Con Regularización L2)



5.2. Conclusión del Ajuste del Modelo

Después de aplicar la regularización L2, observamos una ligera reducción en la accuracy del modelo, pero este muestra un mejor comportamiento en términos de generalización, lo que indica que el ajuste es más adecuado y no está tan sobreajustado como en el caso inicial.

6. Conclusiones Finales

El modelo de regresión logística entrenado sin regularización mostró buenos resultados iniciales con de del desempeño del 88 %. Sin embargo, presentaba signos de sobreajuste debido a la diferencia entre la precisión en los conjuntos de entrenamiento y prueba. Al aplicar regularización L2, el modelo mantuvo un nivel de desempeño aceptable de 87 %, pero con una mejor capacidad de generalización y menor riesgo de sobreajuste.

Este análisis demuestra cómo la regularización puede mejorar el desempeño de los modelos de aprendizaje automático al reducir la varianza sin sacrificar demasiado la precisión.