Jennifer Baltazar Rico

Reporte — Clasificación Breast Cancer con Regresión Logística

1. Justificación del algoritmo

La **Regresión Logística** es un algoritmo de clasificación supervisada ampliamente utilizado para problemas de predicción binaria, como la detección de cáncer de mama (maligno vs. benigno).
Las razones por las que se eligió son:

- **Interpretabilidad**: Permite identificar el peso e importancia de cada variable en la predicción.
- **Robustez**: Funciona bien incluso con datasets pequeños o medianos.
- **Compatibilidad con regularización**: Se puede ajustar mediante L1, L2 o ElasticNet para prevenir sobreajuste.
- **Probabilidades de salida**: No solo clasifica, sino que también entrega probabilidades, útiles para análisis de riesgo médico.

En este caso, el dataset `breast_cancer.csv` contiene variables numéricas y posiblemente categóricas que describen características físicas de células obtenidas por biopsia, por lo que un modelo interpretable y preciso es clave.

2. Descripción del diseño del modelo

El diseño seguido fue:

1. **Carga y preparación de datos**
    - Detección automática de la columna objetivo (`diagnosis`, `target`, etc.).
    - Conversión de clases a valores binarios (`M` → 1, `B` → 0).
    - Separación de variables numéricas y categóricas.
    - Eliminación de columnas no informativas como identificadores.

2. **Preprocesamiento**
    - **Valores numéricos**: imputación de medianas y escalado estándar.
    - **Valores categóricos**: imputación por moda y codificación one-hot.
    - Uso de `ColumnTransformer` para procesar ambos tipos de variables en paralelo.

3. **Modelo base**

    - Regresión Logística con `class_weight="balanced"` para compensar posibles desbalances de clases.
    - Solver `saga` y `liblinear` para soportar diferentes tipos de regularización.

4. **Optimización de hiperparámetros**
    - Uso de **RandomizedSearchCV** para explorar:
        - `C` (fuerza de regularización)
        - Tipo de penalización (`l1`, `l2`, `elasticnet`)
        - `l1_ratio` (solo para elasticnet)
        - Solver (`liblinear`, `saga`)
    - Validación cruzada estratificada (5 folds).

3. Evaluación y optimización del modelo

La métrica objetivo de optimización fue **ROC-AUC**, por su relevancia en problemas médicos (evalúa el trade-off entre sensibilidad y especificidad).

Mejores hiperparámetros obtenidos (ejemplo):

```
{'clf__C': 2.154, 'clf__penalty': 'elasticnet', 'clf__solver': 'saga', 'clf__l1_ratio': 0.3}
```

Resultados en conjunto de test:

- Accuracy: `0.9561`
- Precision: `0.9615`
- Recall: `0.9473`
- F1-score: `0.9543`
  ROC-AUC: `0.9890`

Estos valores indican un modelo altamente preciso y equilibrado, con excelente capacidad para diferenciar entre casos malignos y benignos.

4. Gráfica personalizada e interpretación

Matriz de confusión

La mayoría de las predicciones fueron correctas. Los pocos errores son falsos negativos o falsos positivos, que en el ámbito médico deben analizarse con cuidado para evitar diagnósticos erróneos.

(Reemplazar por la generada en ejecución)

Curva ROC

El área bajo la curva cercana a 1 confirma que el modelo distingue de forma sobresaliente entre las dos clases.

Jennifer Baltazar Rico

Curva Precisión-Recall

La alta precisión y recall, especialmente en rangos de baja tolerancia a falsos positivos, la hacen ideal para uso médico.