Universidad Nacional de Ingeniería
Facultad de Ciencias
Escuela Profesional de Ciencia de la Computación

Práctica Calificada III

Minería de Datos

CC442 **30/01/2026** **Ciclo: 2025-III**

Puntaje: 20 puntos . **Duración**. 120 minutos

PARTE I (14 puntos)

CASO: Análisis Predictivo de Diagnóstico Oncológico mediante Regresión Logística

Notebooks: *ejercicio-01.ipynb* y *vif.ipynb*

El conjunto de datos *Breast Cancer Wisconsin (Diagnostic)* es un recurso clásico en el ámbito de la minería de datos y la bioestadística. Contiene características computadas a partir de imágenes digitalizadas de aspiraciones con aguja fina (FNA) de masas mamarias. El objetivo principal es clasificar si un tumor es **Benigno (1)** o **Maligno (0)** basándose en sus propiedades morfológicas.

Desarrollar un modelo de **Regresión Logística** utilizando la librería s***tatsmodels*** para realizar un análisis no solo predictivo, sino también inferencial. Se busca identificar qué variables tienen un impacto significativo en el diagnóstico y asegurar la estabilidad estadística del modelo.

Requerimientos del Código

1. **Carga y Selección de Datos:**
   - Cargar el dataset de cáncer de mama directamente desde sklearn.datasets.
   - Seleccionar un subconjunto de variables predictoras (por ejemplo: mean radius, mean texture, mean perimeter, mean area, mean smoothness, mean compactness) para evitar la redundancia excesiva de dimensiones.
2. **Preprocesamiento y Partición:**
   - Dividir los datos en conjuntos de **entrenamiento (80%)** y **prueba (20%)** para validar la capacidad de generalización del modelo.
   - Aplicar **estandarización (Z-score scaling)** a las variables independientes. *Nota: El escalador debe ajustarse solo con los datos de entrenamiento para evitar la fuga de datos (data leakage).*
3. **Diagnóstico de Multicolinealidad:**
   - Calcular el **Factor de Inflación de la Varianza (VIF)** para cada variable.
   - Analizar si existe una correlación excesiva entre los predictores que pueda invalidar la interpretación de los coeficientes.
4. **Modelado Estadístico:**
   - Implementar el modelo de Regresión Logística utilizando la clase Logit de statsmodels.

- Asegurarse de incluir manualmente el **término de intercepto (constante)**.
- Generar y mostrar el resumen estadístico (summary) del modelo.

5. **Evaluación e Interpretación:**
   - Realizar predicciones sobre el conjunto de prueba utilizando un umbral de decisión de **0.5**.
   - Generar un reporte de clasificación (precisión, recall, F1-score) y una matriz de confusión.
   - Calcular e interpretar los **Odds Ratios** (exponencial de los coeficientes) para cuantificar el cambio en la probabilidad de diagnóstico por cada unidad de cambio en los predictores.

Preguntas de Reflexión (Análisis de Resultados)

- ¿Cuáles variables resultaron ser estadísticamente significativas (p-valor < 0.05)?
- ¿Qué indica un VIF elevado en variables como el radio y el perímetro?
- ¿Cómo afecta el signo de los coeficientes a la probabilidad de que el tumor sea clasificado como benigno?

Salida esperada:

```
--- Diagnóstico VIF ---
           Variable          VIF
0             const     1.000000
1       mean radius  1460.924793
2      mean texture     1.144166
3    mean perimeter  1737.092968
4         mean area    44.049377
5   mean smoothness     2.064559
6  mean compactness    11.866719
------------------------------
Optimization terminated successfully.
         Current function value: 0.159583
         Iterations 10
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                  455
Model:                          Logit   Df Residuals:                      448
Method:                           MLE   Df Model:                            6
Date:                Thu, 29 Jan 2026   Pseudo R-squ.:                  0.7581
Time:                        19:33:50   Log-Likelihood:                -72.610
converged:                       True   LL-Null:                       -300.17
Covariance Type:            nonrobust   LLR p-value:                 3.887e-95
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const              -0.2118      0.513     -0.413      0.680      -1.218       0.794
mean radius        22.6977     11.506      1.973      0.049       0.146      45.249
mean texture       -1.5228      0.270     -5.650      0.000      -2.051      -0.995
mean perimeter    -16.9089     11.079     -1.526      0.127     -38.624       4.806
mean area         -12.6785      5.118     -2.477      0.013     -22.709      -2.648
mean smoothness    -1.6515      0.359     -4.599      0.000      -2.355      -0.948
mean compactness    0.2827      0.817      0.346      0.729      -1.319       1.885
==============================================================================
```

```
Possibly complete quasi-separation: A fraction 0.15 of observations can be
perfectly predicted. This might indicate that there is complete
quasi-separation. In this case some parameters will not be identified.

--- Reporte de Clasificación (Scikit-Learn Metrics) ---
              precision    recall  f1-score   support

           0       0.93      0.91      0.92        43
           1       0.94      0.96      0.95        71

    accuracy                           0.94       114
   macro avg       0.94      0.93      0.93       114
weighted avg       0.94      0.94      0.94       114


--- Odds Ratios (Interpretación) ---
const                8.091427e-01
mean radius          7.202470e+09
mean texture         2.180995e-01
mean perimeter       4.534986e-08
mean area            3.117395e-06
mean smoothness      1.917570e-01
mean compactness     1.326692e+00
dtype: float64
```

PARTE II (6 puntos)

Completar códigos.

Notebook: ***ejercicio-02.ipynb***