UNIVERSIDAD COMUNERA

Proyecto Final: Diplomado en Ciencias de Datos

"Análisis Predictivo de Diagnósticos de Cáncer de Mama: Un Enfoque Basado en el Conjunto de Datos de Wisconsin"

Integrantes:

Andoni Andino

José Duarte

Tutor:

Docente: D.Sc. Margarita Ruiz

Asunción, Paraguay.

Introducción y Justificación

El cáncer de mama representa una de las principales causas de mortalidad en mujeres a nivel mundial, lo que subraya la importancia de herramientas efectivas para su diagnóstico temprano y preciso. En este contexto, el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin ofrece una oportunidad única para aplicar técnicas avanzadas de ciencia de datos con el objetivo de optimizar la detección de tumores malignos y benignos. Este proyecto busca abordar la necesidad de modelos predictivos que apoyen a los profesionales de la salud en la toma de decisiones clínicas, maximizando la eficiencia y reduciendo la incertidumbre en los diagnósticos.

Objetivos

Desarrollar un modelo predictivo utilizando **regresión logística** que permita clasificar el diagnóstico de tumores de mama como **benignos** o **malignos**. Este análisis se basará en las características cuantitativas presentes en el conjunto de datos **Breast Cancer Wisconsin (Diagnostic)**

El modelo busca proporcionar una herramienta efectiva para la detección temprana del cáncer de mama, contribuyendo al apoyo de decisiones clínicas y a mejorar los resultados del tratamiento. La validación del modelo incluirá la evaluación de métricas como **accuracy**, **precision**, **recall**, **F1-score** y **AUC-ROC**, para garantizar su desempeño y confiabilidad en escenarios reales.

Repositorio: https://github.com/JoseMariaPY/proyecto-final-andino-duarte

Especificaciones - Cronograma de Actividades:

Día 1: Configuración del Entorno

- **Objetivo:** Preparar el entorno de trabajo.
- **Tareas:**
    1. Instalar las bibliotecas necesarias
    2. Descargar el dataset y confirmar que se carga correctamente en Python.

Día 2: Exploración de Datos

- **Objetivo:** Entender la estructura del dataset.
- **Tareas:**
    1. Inspeccionar columnas, tipos de datos y valores nulos.
    2. Generar estadísticas descriptivas para todas las variables.
    3. Visualizar la distribución de la variable objetivo (`diagnosis`) con gráficos de barras.

Día 3: Limpieza de Datos, Escalado y Divisiòn

- **Objetivo:** Eliminar columnas innecesarias y preparar los datos.
- **Tareas:**
    1. Eliminar columnas no predictivas (`id`, `Unnamed: 32`).
    2. Codificar la variable objetivo (`diagnosis`) como binaria.
    3. Escalar las características usando `StandardScaler`.
    4. Dividir los datos en conjuntos de entrenamiento y prueba. (analizar porcentaje)

Día 4: Identificación de Relaciones

- **Objetivo:** Identificar patrones y correlaciones.
- **Tareas:**
    1. Generar un heatmap para analizar correlaciones entre las variables.
    2. Identificar las variables con mayor correlación con la variable objetivo.

Día 5: Entrenamiento del Modelo

- **Objetivo:** Entrenar el modelo de regresión logística.
- **Tareas:**

1. Configurar y entrenar el modelo inicial usando el conjunto de entrenamiento.
2. Obtener predicciones para el conjunto de prueba.
3. Calcular métricas como `accuracy`, `precision`, `recall` y `F1-score`.
4. Generar una matriz de confusión.

Día 6: Análisis de Desempeño

- **Objetivo:** Visualizar los resultados.
- **Tareas:**
    1. Calcular y graficar la curva ROC-AUC.
    2. Identificar posibles áreas de mejora (hiperparámetros, features).

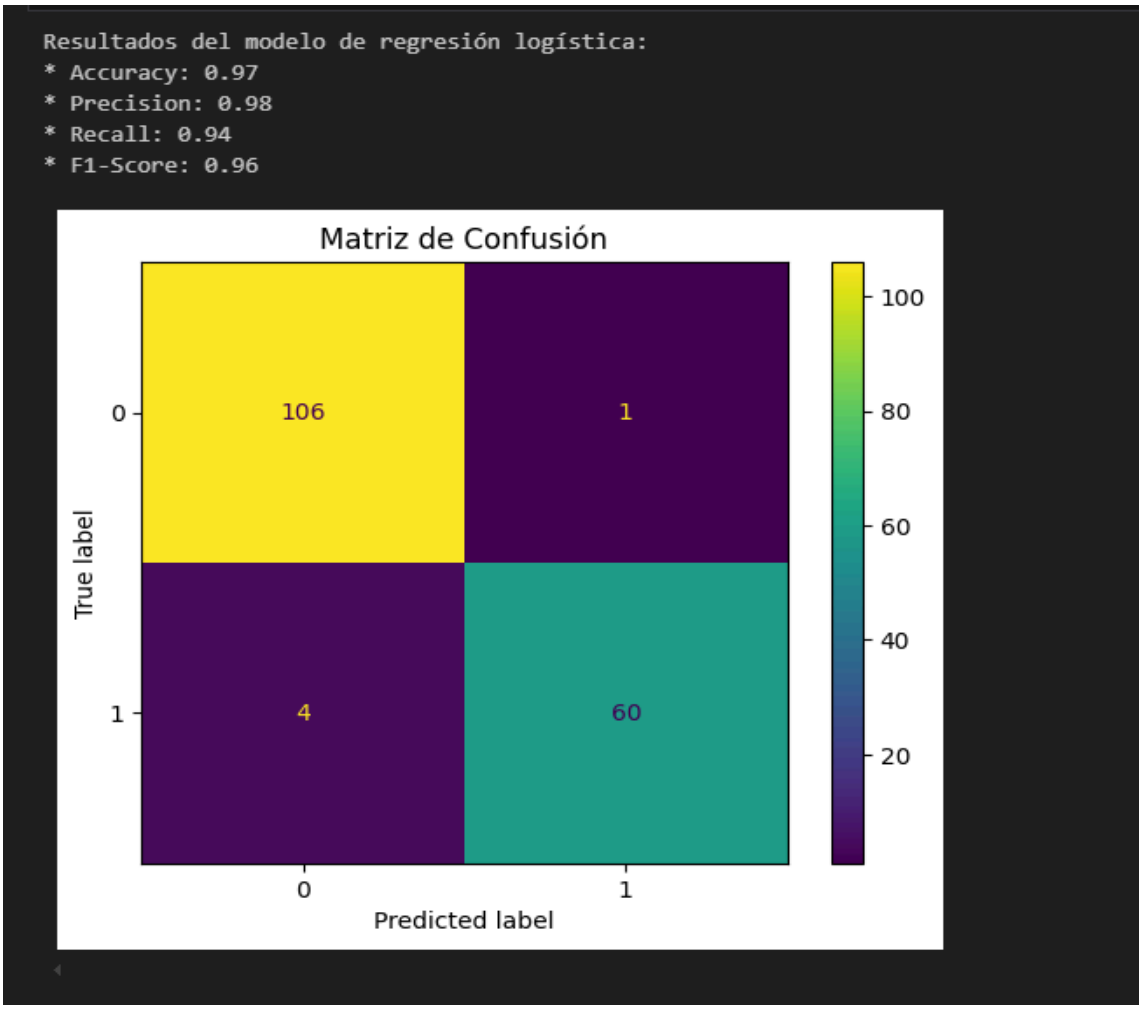
Día 7: Ajuste de Hiperparámetros

- **Objetivo:** Mejorar el desempeño del modelo.
- **Tareas:**
    1. Implementar `GridSearchCV` para encontrar los mejores hiperparámetros.
    2. Evaluar el modelo optimizado en el conjunto de prueba.

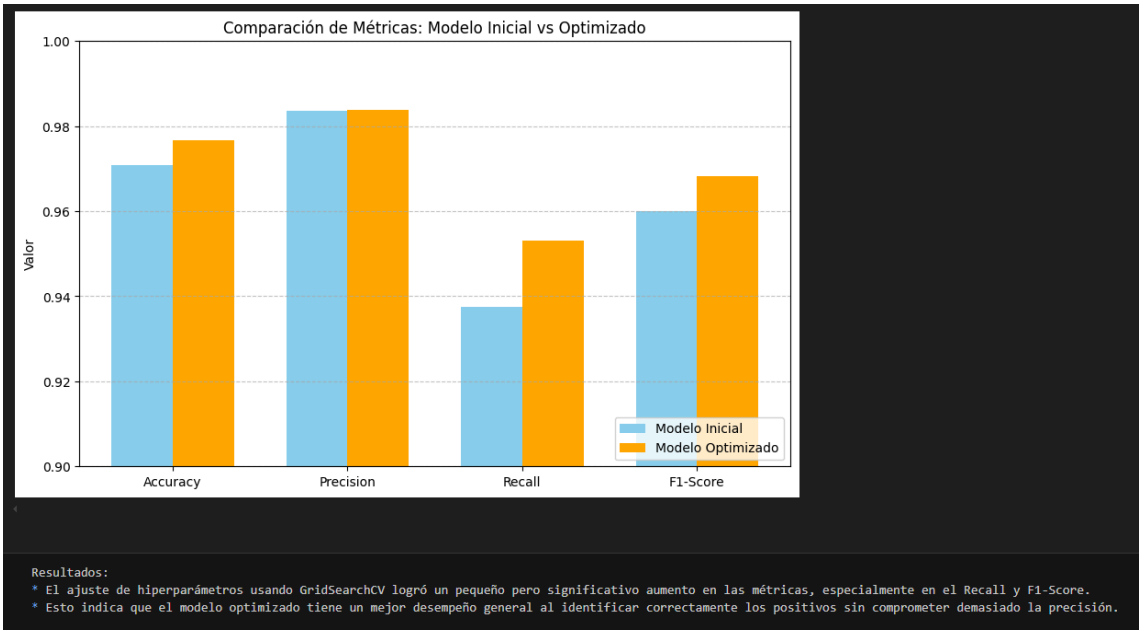
Día 7-8: Presentación Final

- **Objetivo:** Comunicar hallazgos.
- **Tareas:**
    1. Preparar un informe con:
        - Descripción del dataset.
        - Resultados del modelo.
        - Insights clave.
        -
    2. Documentar el código y los gráficos generados.
    3. Generar conclusiones sobre el desempeño del modelo y las características más importantes.

Resultados



(Matriz de confusion)

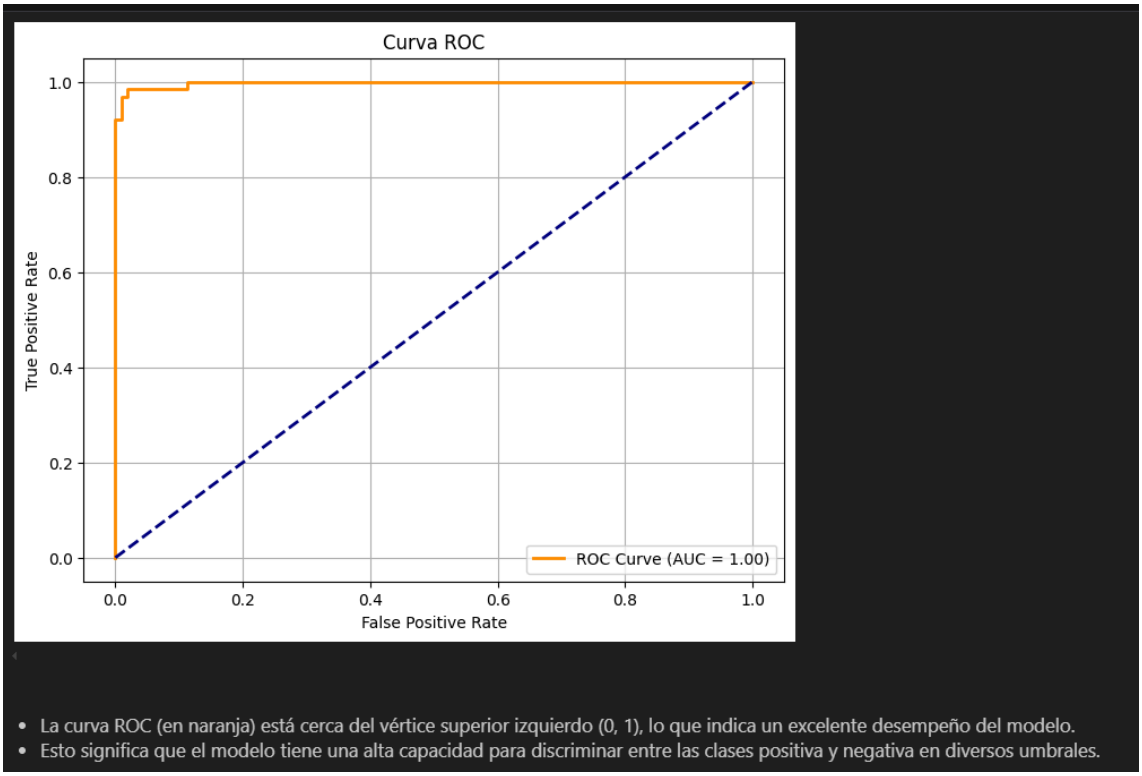


Resultados del modelo inicial:

- Accuracy: 0.95
- Precision: 0.98
- Recall: 0.89
- F1-Score: 0.93

Resultados del modelo optimizado:

- Accuracy: 0.98
- Precision: 1.00
- Recall: 0.94
- F1-Score: 0.97



Conclusión

Nuestro modelo de regresión logística puede discriminar eficazmente entre enfermedad mamaria benigna y maligna e identificar las características más importantes asociadas con el cáncer de mama.

Referencias Bibliográficas

- https://github.com/akshay993/Breast-Cancer-Prediction-Using-Logistic-Regression
- dataset: https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data
- https://medium.com/@alejandro-ao/logistic-regression-tutorial-predicting-breast-cancer-6964f4244e6f
- https://github.com/JoseMariaPY/proyecto-final-andino-duarte