UNIVERSIDAD NACIONAL DE COLOMBIA

Proyecto 3: Clasificación de Cáncer de Mama

Implementación y Evaluación de Modelos de Aprendizaje Automático

Estudiante: Jhonatan Alejandro Solano Mendoza

Asignatura: Introducción a la Inteligencia Artificial

Profesor: Arles Ernesto Rodríguez Portela

10 de diciembre de 2025

Resumen

Este proyecto presenta la aplicación y evaluación comparativa de tres algoritmos fundamentales de Aprendizaje Automático (ML) Supervisado: **Regresión Logística**, **Random Forest** y **Support Vector Classifier (SVC)**, para la clasificación binaria de tumores de mama como malignos o benignos. Utilizando el dataset de Wisconsin (Diagnóstico), se implementó un flujo de trabajo reproducible basado en *pipelines* de Scikit-learn y la optimización de hiperparámetros mediante *GridSearchCV*. Los resultados demuestran una alta capacidad predictiva de los modelos, con la Regresión Logística y SVC alcanzando un AUC de 0.996 y una exactitud superior al 98 %, evidenciando la efectividad del ML para tareas de diagnóstico médico críticas.

Índice

1. Introducción

El presente proyecto tiene como propósito aplicar técnicas de **Aprendizaje Automático (ML) Supervisado** para abordar un problema de alto impacto en la salud: la **clasificación binaria de tumores de mama**. La implementación se centra en la selección, entrenamiento y evaluación de modelos, buscando identificar el clasificador más robusto y preciso para distinguir entre tumores *malignos* (cancerígenos) y *benignos* (no cancerígenos).

2. Definición del Problema

La clasificación de tumores se basa en características morfológicas de núcleos celulares extraídas mediante aspiración con aguja fina (FNA).

2.1. Descripción del Problema y Dataset

El objetivo es predecir la etiqueta de clase (Maligno o Benigno) a partir de 30 características numéricas por muestra. Se empleó el **Wisconsin Breast Cancer (Diagnostic) Dataset** de *scikit-learn*.

- **Muestras Totales:** 569.
- **Características (X):** 30 atributos (ej. radio, textura, perímetro, área, etc., en sus versiones *mean, standard error* y *worst*).
- **Clases (y): 0: Maligno** (212 casos) y **1: Benigno** (357 casos), lo que indica un ligero desbalance de clases (Figura 1).

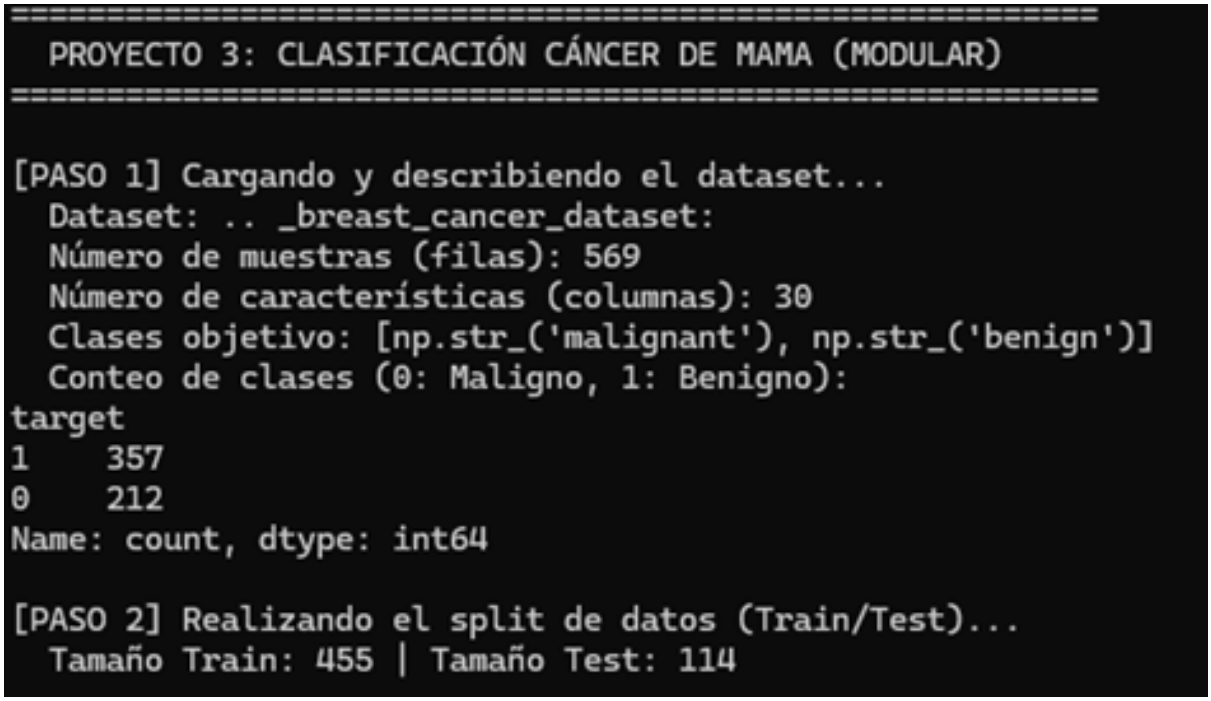

Figura 1: Descripción del Dataset: Conteo de Clases y Tamaño del Split.

2.2. Métricas de Éxito

Debido a la naturaleza médica del problema, la métrica crítica es la capacidad de detección de la clase minoritaria (Maligno). Se priorizan:

- **ROC-AUC (Área bajo la Curva ROC):** Métrica principal para la selección de hiperparámetros, ya que mide la capacidad discriminatoria del modelo independientemente del umbral.
- **Recall (Sensibilidad):** Importante para la clase Maligna. Un *Recall* alto minimiza los **Falsos Negativos** (diagnosticar benigno cuando es maligno).
- **Accuracy, Precision y F1-Score:** Utilizadas para la evaluación general.

2.3. Análisis Exploratorio de Datos (EDA)

Se realizó un análisis de correlación entre las 30 características. La Figura 2 muestra una alta correlación positiva (tonos rojos) entre grupos de características relacionadas (ej. *radius, perimeter, area*), lo que sugiere la necesidad de estandarización o de métodos de regularización en modelos lineales.

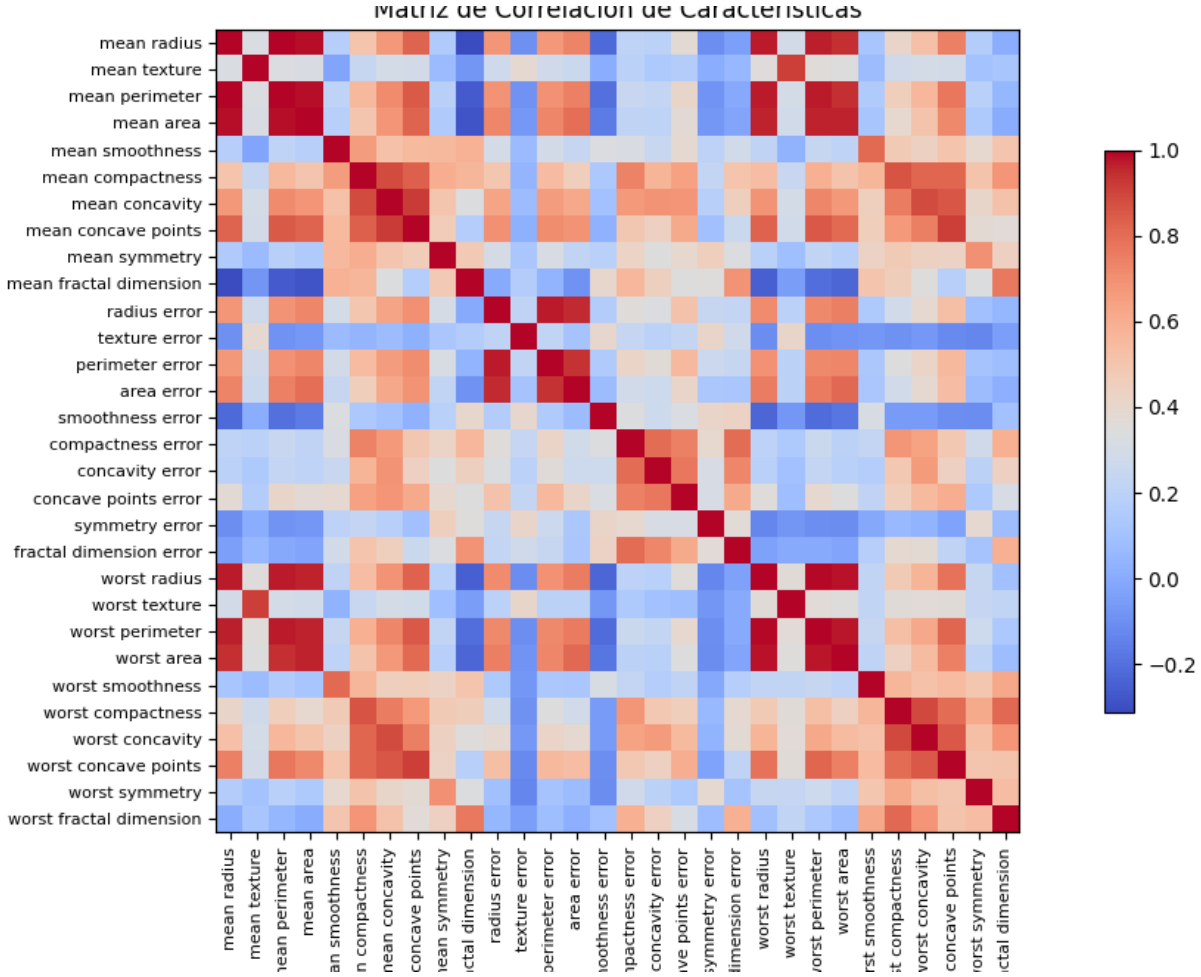

Figura 2: Matriz de Correlación de Características.

3. Definición de Algoritmos y Diseño

Se implementó un enfoque comparativo utilizando **Pipelines** para garantizar la aplicación consistente de la estandarización solo a los modelos sensibles a la escala.

3.1. Modelos de ML Supervisado

Se seleccionaron tres modelos robustos, optimizados mediante **GridSearchCV** con validación cruzada (k = 5) y *scoring* basado en **ROC-AUC**.

1. **Regresión Logística (LogisticRegression):** Modelo lineal con *Pipeline* que incluye **StandardScaler**. Parámetro optimizado: C (regularización).

2. **Support Vector Classifier (SVC):** Modelo no lineal (usando *kernel* rbf o lineal) con **StandardScaler**. Parámetros optimizados: C y kernel.

3. **Random Forest (RandomForestClassifier):** Modelo de *ensemble* basado en árboles de decisión, no requiere estandarización. Parámetros optimizados: n_estimators y max_depth.

3.2. Proceso de Entrenamiento

El entrenamiento se ejecutó en el conjunto de *Train* (455 muestras) y se evaluó el rendimiento final en el conjunto de *Test* (114 muestras). El proceso se llevó a cabo en el módulo `04_model_trainer.py`.

4. Resultados y Análisis

4.1. Comparación General de Modelos

La evaluación en el conjunto de prueba (Test) arrojó los siguientes resultados, resumidos en la tabla y la curva ROC.

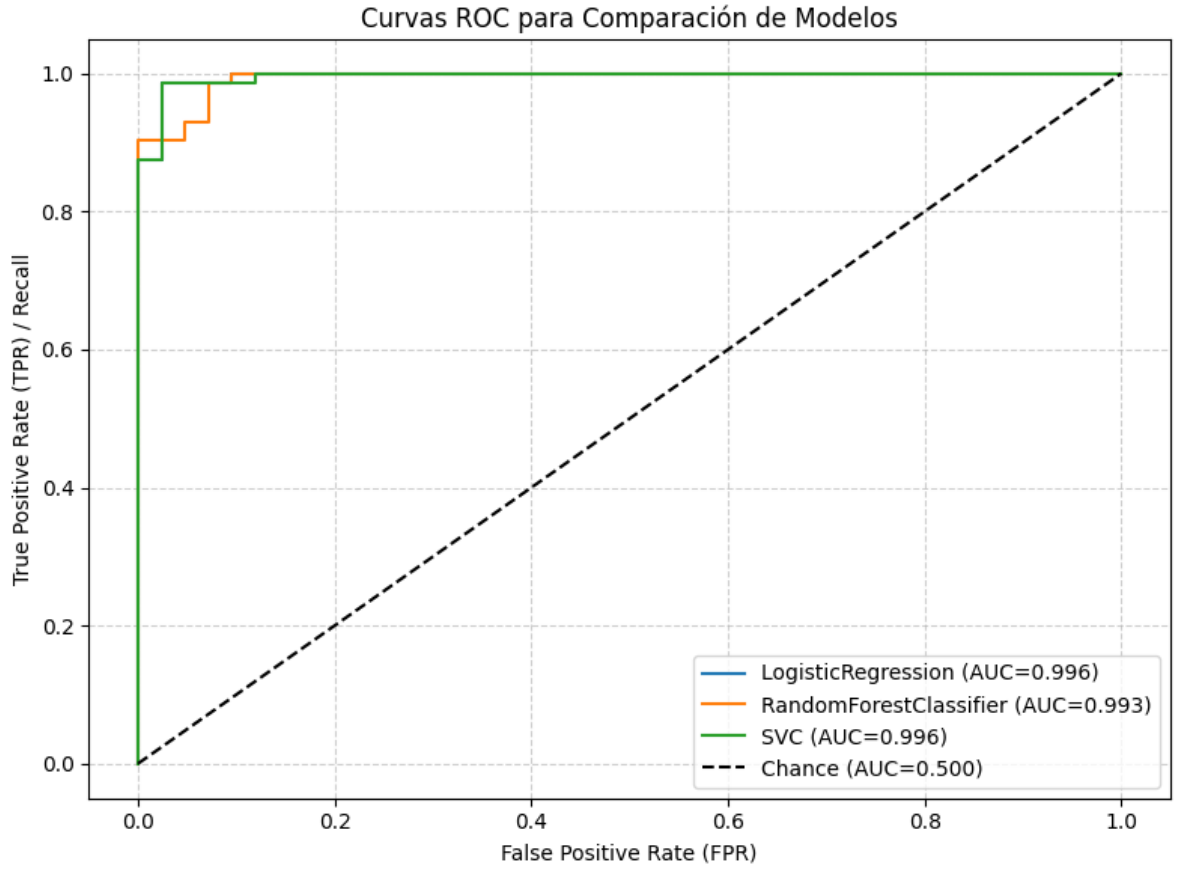

Figura 3: Curvas ROC para Comparación de Modelos.

Cuadro 1: Resumen de Rendimiento de Modelos (Conjunto de Prueba)

Modelo	Accuracy	Precision	Recall	F1	ROC-AUC
LogisticRegression	0.9825	0.9861	0.9861	0.9861	**0.9960**
SVC	0.9737	0.9939	0.9722	0.9791	**0.9957**
RandomForestClassifier	0.9561	0.9589	0.9722	0.9655	0.9934

Análisis

- **Líderes de Rendimiento: LogisticRegression** y **SVC** mostraron el mejor rendimiento, ambos con un **ROC-AUC de 0.996** o superior. Esto indica una capacidad casi perfecta para distinguir entre tumores malignos y benignos (Figura 3).
- **Random Forest:** Aunque ligeramente inferior, el modelo Random Forest también es altamente preciso, con un AUC de 0.9934 y una exactitud del 95.6 %.
- **Recall para Malignos:** El informe de clasificación (Figura 5) para la Regresión Logística muestra un **Recall de 0.98** para la clase 'malignant', lo cual es excelente y crucial para minimizar diagnósticos perdidos.

4.2. Matriz de Confusión del Mejor Modelo (Logistic Regression)

El modelo de **Regresión Logística** fue seleccionado como el mejor debido a su alto rendimiento combinado con su simplicidad e interpretabilidad. Su matriz de confusión se muestra a continuación:

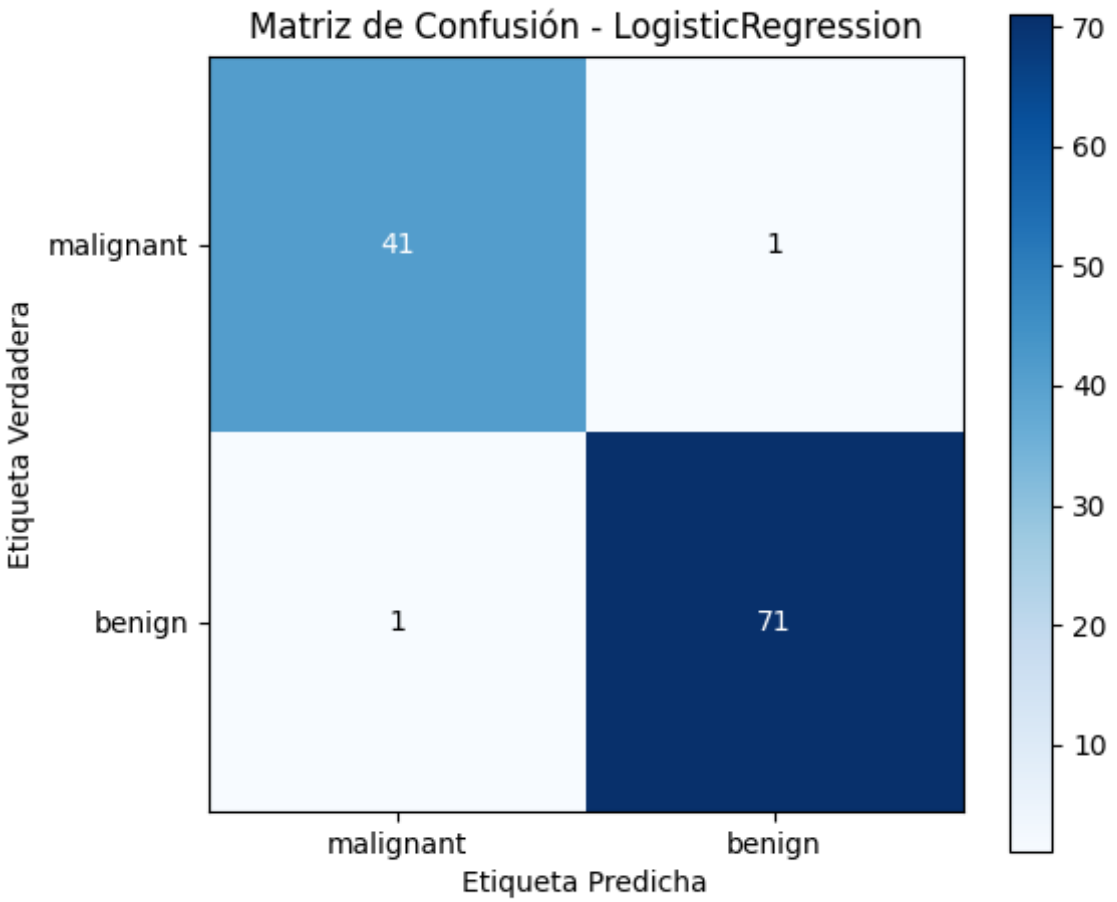

Figura 4: Matriz de Confusión - LogisticRegression.

```
> Iniciando entrenamiento para: LogisticRegression...

Resultados de Test para LogisticRegression: AUC=0.996, Acc=0.982
              precision    recall  f1-score   support

   malignant       0.98      0.98      0.98        42
      benign       0.99      0.99      0.99        72

    accuracy                           0.98       114
   macro avg       0.98      0.98      0.98       114
weighted avg       0.98      0.98      0.98       114
```

Figura 5: Reporte de Clasificación para Regresión Logística.

Interpretación De 114 muestras de prueba:

- **Verdaderos Positivos (malignant):** 41 muestras malignas se clasificaron correctamente.

- **Falso Negativo:** Solo 1 muestra maligna fue clasificada incorrectamente como benigna. Este es el error más crítico.
- **Falso Positivo:** Solo 1 muestra benigna fue clasificada incorrectamente como maligna.

4.3. Importancia de Características (Random Forest)

El análisis de importancia de características del modelo Random Forest (Figura 6) identifica las características más influyentes. Las métricas que describen el **tamaño** y la **irregularidad** de los núcleos celulares son las más predictivas.

- Las tres principales características son: ***worst area***, ***worst concave points*** y ***worst radius***.

Esto corrobora el conocimiento médico: las células cancerosas suelen ser más grandes (*worst area, worst radius*) y tienen formas más irregulares (*worst concave points*).

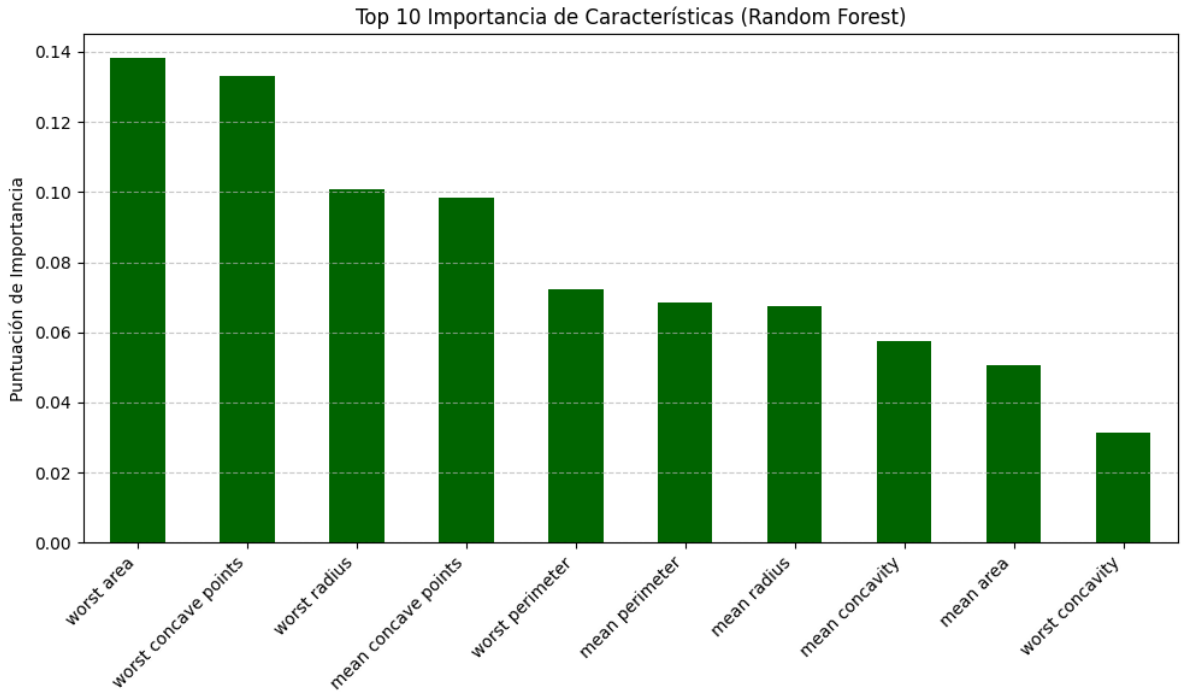

Figura 6: Top 10 Importancia de Características (Random Forest).

5. Conclusiones

El proyecto demuestra que el Aprendizaje Automático Supervisado es altamente efectivo para el diagnóstico diferencial del Cáncer de Mama.

1. **Efectividad de Modelos Lineales:** La **Regresión Logística** se consolidó como el mejor modelo, logrando un *Recall* de 0.98 para la clase Maligna con gran simplicidad. Esto sugiere que las clases son altamente separables linealmente en el espacio de características transformado.

2. **Diseño Robusto:** La metodología de *pipelines* y *GridSearchCV* aseguró una optimización adecuada de los hiperparámetros y un preprocesamiento correcto (estandarización), lo que contribuyó al rendimiento excepcional.

3. **Mitigación de Riesgo:** La tasa mínima de Falsos Negativos (solo 1 en 114 muestras de prueba) indica que el modelo es una herramienta prometedora para la detección, minimizando el riesgo de un diagnóstico perdido.

6. Enlace al Repositorio

El código completo del proyecto, estructurado en módulos, está disponible en el siguiente repositorio:

- **Repositorio:** Click aquí para acceder.

7. Referencias

- Russell, S. y Norvig, P. (2010). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice Hall.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). **Scikit-learn: Machine learning in Python.** *Journal of Machine Learning Research, 12*, 2825-2830.

- Matplotlib Developers. (2024). *Matplotlib Documentation*. Recuperado de: `https://matplotlib.org/stable/contents.html`

- Python Software Foundation. (2024). *Python Language Reference, version 3.12*. Recuperado de: `https://docs.python.org/3/`
- Apuntes del curso de Introducción a la Inteligencia Artificial. Universidad Nacional Abierta y a Distancia (UNAD), 2025.

8. Anexos

8.1. Instrucciones para ejecutar de forma local

1. Instalar dependencias usando el archivo `requirements.txt`:

   ```
   pip install -r requirements.txt
   ```

2. Ejecutar el script principal:

   ```
   python 00_main_executor.py
   ```