Informe Final Programación

María Ángel Toro Ramírez
Profesor Andrés Quintero Zea
Universidad EIA mayo 2025

Clasificación de Diabetes usando Modelos de Machine Learning

Objetivo

El objetivo de este proyecto es desarrollar un modelo de machine learning que prediga si un paciente tiene diabetes en función de varios atributos médicos. Este modelo puede servir como herramienta de apoyo para aprender sobre el diagnóstico médico para detectar personas con mayor riesgo de diabetes.

Descripción del Dataset

Fuente: UCI Machine Learning Repository (ID 891).

Número de instancias: 768.

Número de características: 8 características.

Variable objetivo: target (0 = no diabetes, 1 = diabetes).

El conjunto de datos contiene 768 instancias con 8 características médicas y un valor objetivo (target) que indica si un paciente tiene diabetes o no.

Pre-procesamiento

Imputación de valores faltantes: Se verificó si había valores faltantes en el dataset y se imputaron utilizando la media de la columna correspondiente.

Estandarización de las características: Se utilizaron técnicas de escalado con StandardScaler para normalizar las características numéricas y mejorar el desempeño de los modelos.

División en conjuntos de entrenamiento y prueba: El dataset se dividió en un 70% para entrenamiento y un 30% para prueba, con una estratificación para garantizar que las proporciones de clases fueran representativas en ambos conjuntos.

Modelos Implementados

Se implementaron y entrenaron dos modelos para la predicción de la diabetes:

Modelo 1: Regresión Logística

Accuracy: 77.92%

F1-score: 0.77

Matriz de Confusión: La matriz de confusión mostró que el modelo tiene un buen rendimiento en la predicción tanto de los pacientes con diabetes como de los que no.

Modelo 2: Random Forest

Accuracy: 77.92%

F1-score: 0.77

Matriz de Confusión: Similar a la regresión logística, el modelo Random Forest mostró resultados equilibrados, con una ligera mejora en la capacidad para manejar casos de clases desbalanceadas.

Ambos modelos se evaluaron utilizando curvas de aprendizaje para ver su rendimiento en función de diferentes tamaños de entrenamiento y matrices de confusión para evaluar la precisión de las predicciones.

Comparación y Conclusión

Los dos modelos mostraron un desempeño similar tanto en términos de **accuracy** como de **F1-score** (alrededor del 77-78%). Dado que los resultados de ambos modelos son casi idénticos, se puede concluir que no hay un modelo significativamente superior entre los dos para este conjunto de datos específico.

Recomendación: Se recomienda usar **Random Forest** por su capacidad de manejar relaciones no lineales y la posibilidad de mejorar con más parámetros y optimización, aunque la diferencia con la regresión logística es mínima.

Referencias

UCI Machine Learning Repository: https://archive.ics.uci.edu/

Scikit-learn: https://scikit-learn.org/

Seaborn & Matplotlib Documentation: https://seaborn.pydata.org/ & https://matplotlib.org/