Programación

Informe proyecto final 2025-2

Presentado por:

Maria José Arias

Alejandro Mejía

Profesor:

Andrés Quintero Zea

Universidad EIA

Envigado

Junio 2, 2025

2025-1

Informe

Introducción

Este documento resume el proceso completo de desarrollo, implementación y evaluación del proyecto final de la asignatura de Programación. El objetivo principal fue aplicar un enfoque de ciencia de datos para abordar un problema de clasificación binaria en un contexto de salud: la predicción de la presencia de diabetes en pacientes, a partir de un conjunto de características clínicas.

Se emplearon técnicas modernas de exploración, preprocesamiento, modelado y evaluación, priorizando la sensibilidad (recall) del modelo para minimizar los falsos negativos, una decisión clave en aplicaciones donde el costo de no detectar una condición puede ser crítico.

Estrategia de Solución

El pipeline desarrollado siguió una estructura lógica y secuencial dividida en las siguientes etapas:

1. Exploración de Datos

Se inició el proyecto con un análisis exploratorio de datos (EDA), el cual permitió comprender la estructura general del dataset, la naturaleza de las variables y posibles problemas de calidad de datos. Las principales acciones realizadas fueron:

- Análisis descriptivo básico de cada variable (media, desviación estándar, valores máximos y mínimos).
- Visualización de distribuciones mediante histogramas y gráficos de densidad.
- Estudio de correlaciones mediante mapas de calor.
- Detección de valores atípicos (outliers) utilizando diagramas de caja (boxplots) y análisis visual.

Este análisis fue fundamental para guiar decisiones posteriores de limpieza y transformación de los datos.

2. Preprocesamiento de Datos

Con base en los hallazgos de la exploración, se ejecutaron varios pasos de preprocesamiento para garantizar que los datos estuvieran en condiciones óptimas para el entrenamiento de modelos:

- **Imputación de valores nulos:** Se aplicaron técnicas de imputación (como la media para variables numéricas) para completar datos faltantes, manteniendo la integridad del conjunto de datos.
- **Normalización:** Dado que algunos algoritmos (como SVM) son sensibles a la escala, se realizó normalización (MinMaxScaler) para llevar todas las variables a un rango común.
- **Codificación de variables categóricas:** Aunque el dataset no incluía muchas variables categóricas, se consideró la codificación one-hot si fuese necesario.

Modelado

Se entrenaron dos modelos de clasificación, ambos con el objetivo de priorizar el **recall**, es decir, la capacidad del modelo para identificar correctamente los casos positivos (diabetes).

Modelo 1: Random Forest

- Algoritmo basado en un conjunto de árboles de decisión.
- Se utilizó class_weight='balanced' para compensar el desbalance de clases.
- Se ajustaron parámetros como número de árboles, profundidad máxima, etc., mediante validación cruzada.
- Ideal para trabajar con datos tabulares y manejar relaciones no lineales.

Modelo 2: LinearSVC

- Versión lineal del algoritmo Support Vector Classifier.
- También se aplicó class_weight='balanced' para priorizar la clase minoritaria.
- Si bien es menos flexible que Random Forest, ofrece ventajas en velocidad y simplicidad, especialmente en datasets más pequeños o con relaciones lineales entre las variables.

Ambos modelos fueron entrenados con la misma partición de datos y con un enfoque de evaluación consistente.

Evaluación

Para comparar el rendimiento de ambos modelos, se aplicaron múltiples métricas de clasificación:

- **Precisión (Precision):** Proporción de verdaderos positivos entre los elementos clasificados como positivos.
- **Recall (Sensibilidad):** Proporción de verdaderos positivos detectados sobre el total de casos positivos reales.
- **F1-Score:** Promedio armónico entre precisión y recall, útil en casos de desbalance.
- **Matriz de confusión:** Representación visual de los errores de clasificación.
- **Curvas de aprendizaje:** Para verificar el comportamiento del modelo conforme aumenta la cantidad de datos de entrenamiento.

Ambos modelos mostraron un buen desempeño, pero el **Random Forest** se destacó por ofrecer un mejor equilibrio entre las métricas evaluadas, particularmente en el recall, que era el objetivo prioritario.

Comparación de Modelos y Concordancia

Para verificar la similitud entre las predicciones de ambos modelos y el ground truth, se utilizó el **índice de Kappa de Cohen**, una métrica de concordancia ajustada por azar. El valor obtenido fue:

Cohen’s Kappa: 0.8462

Este resultado indica un **nivel alto de acuerdo** entre las predicciones y las etiquetas reales, lo cual respalda la robustez del modelo final.

Conclusiones

Del desarrollo y análisis del proyecto se obtuvieron las siguientes conclusiones principales:

- El análisis exploratorio fue clave para identificar problemas y patrones iniciales en los datos.
- Las técnicas de preprocesamiento aplicadas (imputación, normalización y codificación) permitieron preparar los datos adecuadamente para el modelado.
- Optimizar el modelo para **recall** fue una estrategia acertada en el contexto del problema, al minimizar el riesgo de no identificar pacientes con diabetes.
- El modelo de **Random Forest** demostró ser el más robusto en este caso, superando en rendimiento a LinearSVC.
- La alta concordancia según Cohen’s Kappa respalda la consistencia del modelo y su aplicabilidad.
- El pipeline completo permite ser reutilizado y adaptado para otros conjuntos de datos similares, con potencial uso en entornos clínicos o académicos.