Parcial Random Forest y MLP

Juan Mosquera

Anderson

Bornachera

MLOps

Universidad del Magdalena

Facultad de Ingeniería, Ingeniería de Sistemas

Ing. Andrés Vazquez

Santa Marta, Colombia

18 de septiembre de 2025

Introducción

En el presente informe se documenta el desarrollo de un proceso de evaluación de modelos de Machine Learning *(*ML) y redes neuronales aplicado a un conjunto de datos con múltiples variables. El objetivo principal fue llevar a cabo una exploración detallada del dataset, aplicar procesos de limpieza y preprocesamiento, y posteriormente entrenar diferentes modelos de clasificación para identificar patrones y evaluar su desempeño.

La correcta preparación de los datos es un requisito indispensable en proyectos de ciencia de datos y MLOps, ya que la calidad del dataset afecta directamente la precisión de los modelos. De igual manera, la comparación de distintos algoritmos permite seleccionar la mejor alternativa de acuerdo con métricas de desempeño previamente definidas.

Metodología

1. Exploración y limpieza de datos

El dataset fue cargado utilizando la librería **pandas** de Python. Se aplicaron las siguientes etapas:

- **Inspección inicial** mediante `df.head(50)` para verificar la correcta importación de los registros.
- **Análisis estadístico descriptivo** (`df.describe()`) para evaluar la distribución de las variables numéricas, identificar posibles valores atípicos y determinar rangos relevantes.
- **Validación de la estructura** con `df.size`, `df.shape` y `df.info()`, lo que permitió conocer la cantidad de filas, columnas, tipos de datos y valores no nulos.
- **Detección de valores nulos** (`df.isna().sum()`), identificando aquellas columnas que requerían imputación o eliminación de registros incompletos.
- **Exploración de cardinalidad** a partir del conteo de valores únicos por variable, lo que permitió detectar inconsistencias y redundancias.
- **Tratamiento de inconsistencias**: se eliminaron duplicados, se normalizaron fechas, se limpiaron valores de la columna *Teléfono* y se estandarizaron categorías.

Estas etapas de preprocesamiento aseguraron que los datos tuvieran la calidad necesaria antes de proceder al modelado.

2. Modelado y entrenamiento

Con el dataset depurado, se procedió a entrenar diferentes modelos de **Machine Learning** y redes neuronales. La metodología incluyó:

- **División del dataset** en conjuntos de entrenamiento y prueba, garantizando independencia en la validación.
- **Aplicación de algoritmos de clasificación supervisada**, tales como regresión logística, árboles de decisión, *Random Forest* y redes neuronales multicapa.
- **Ajuste de hiperparámetros** mediante técnicas de validación cruzada para optimizar el rendimiento de los modelos.
- **Evaluación de desempeño** utilizando métricas estándar como exactitud (*accuracy*), precisión, exhaustividad (*recall*) y F1-score.

Resultados

En la evaluación de desempeño se compararon los algoritmos **Random Forest** y **Multilayer Perceptron (MLP)**.

- **Random Forest** alcanzó una **exactitud (accuracy) del 44.94%**, con valores de **precisión = 0.3905**, **recall = 0.4494** y **F1-score = 0.4003**. El reporte de clasificación evidenció un mejor comportamiento en la clase *"si"* (precisión = 0.50, recall = 0.73, F1 = 0.59), mientras que las clases *"No"* y *"no"* tuvieron un desempeño considerablemente menor, reflejando dificultades del modelo en escenarios con clases desbalanceadas.
- **MLP (Red Neuronal Multicapa)** obtuvo una **exactitud (accuracy) del 49.95%**, con métricas globales de **precisión = 0.2495**, **recall = 0.4995** y **F1-score = 0.3328**. El modelo mostró un sesgo fuerte hacia la clase *"si"*, donde alcanzó un recall de 1.00 y un F1 de 0.67, pero sin lograr clasificar correctamente las clases *"No"* y *"no"* (ambas con valores cercanos a 0). Además, se reportaron advertencias de métricas indefinidas debido a la ausencia de predicciones para dichas clases.

Comparativamente, aunque el MLP obtuvo una mayor exactitud global, el **Random Forest logró un mejor equilibrio entre las métricas de precisión, recall y F1-score en todas las clases**, mostrando un desempeño más consistente.

Conclusiones

A partir de los resultados obtenidos se pueden extraer las siguientes conclusiones principales:

1. **Desempeño general**: ninguno de los modelos superó el 50% de exactitud, lo que sugiere que el problema de clasificación presenta un alto grado de complejidad o que existen limitaciones en la representación del dataset, posiblemente asociadas a desbalance de clases.
2. **Random Forest vs MLP**: aunque el **MLP alcanzó una accuracy ligeramente superior (49.95%)**, su bajo nivel de precisión y la incapacidad de clasificar dos de las tres clases correctamente reducen su utilidad práctica. Por el contrario, el **Random Forest**, con 44.94% de exactitud, mostró un rendimiento más balanceado, destacándose en la clase mayoritaria *"si"*, pero sin descuidar completamente a las demás.
3. **Impacto del desbalance**: ambos modelos mostraron dificultades para clasificar correctamente las clases *"No"* y *"no"*. Esto evidencia la necesidad de aplicar técnicas de balanceo de datos, como *oversampling*, *undersampling* o métodos avanzados como **SMOTE**, antes de un nuevo ciclo de entrenamiento.
4. **Recomendación**: para un entorno de producción, el **Random Forest** sería más recomendable en este escenario, debido a su mayor robustez frente a la diversidad de clases y su menor dependencia de un ajuste fino de hiperparámetros, a diferencia del MLP.