Machine Learning
Taller 4

Alejandro Gómez (274547)
Nicolás Castañeda (273724)
Caren Piñeros (282290)

Machine Learning
Universidad de La Sabana
Carrera: Ingeniería Informática
Profesor: Hugo Franco

1. Descripción del Problema

El Taller 4 aborda la automatización de flujos de aprendizaje supervisado a través del uso de *pipelines* integrados, aplicados a diferentes tipos de tareas: clasificación binaria y regresión continua. Cada flujo tiene un objetivo específico:

- **Credit Card Fraud Detection:** identificar transacciones fraudulentas en un conjunto de datos altamente desbalanceado.
- **Customer Defection:** predecir la deserción de clientes en el sector de telecomunicaciones, analizando variables de comportamiento y servicio.
- **Bike Rental:** estimar el número de alquileres diarios de bicicletas a partir de variables climáticas y temporales.

La importancia del taller se centra en comprender cómo la elección adecuada de métodos de preprocesamiento, balanceo y selección de modelos impacta directamente en el rendimiento y la confiabilidad de las soluciones analíticas en escenarios de producción.

2. Método de Solución

El desarrollo se implementó en Python utilizando librerías como `scikit-learn`, `imblearn`, `xgboost` y `prefect`. El flujo general de trabajo se presenta en el Algoritmo 1, estructurado en pasos secuenciales y modulares.

Algorithm 1 Flujo general de Machine Learning del Taller 4

1: **1. Carga de datos:** lectura de conjuntos desde fuentes públicas (UCI o GitHub).
2: **2. Análisis exploratorio:** revisión de estructura, valores faltantes y distribución de clases.
3: **3. Preprocesamiento:**
4: - Escalado de variables numéricas.
5: - Codificación *One-Hot* para variables categóricas.
6: - Balanceo de clases con **SMOTE** (solo en clasificación).
7: **4. Selección de modelos:** comparación entre *XGBoost*, *Random Forest* y *Logistic/Linear Regression*.
8: **5. Entrenamiento:** se ejecuta el modelo con los mejores hiperparámetros mediante *RandomizedSearchCV*.
9: **6. Evaluación:** cálculo de métricas (F1, ROC AUC, MSE, R^2).
10: **7. Visualización:** generación de matrices de confusión y gráficos de regresión.

3. Resultados

A continuación se presentan los resultados más relevantes obtenidos en los tres flujos desarrollados.

3.1. Credit Card Fraud Detection

El flujo de clasificación de fraudes trabajó con un dataset altamente desbalanceado (menos del 0.2 % de fraudes). Se implementaron técnicas de escalado y un modelo de *Random Forest* con balanceo interno. La Figura 1 muestra tres matrices de confusión normalizadas: por filas, por columnas y global.

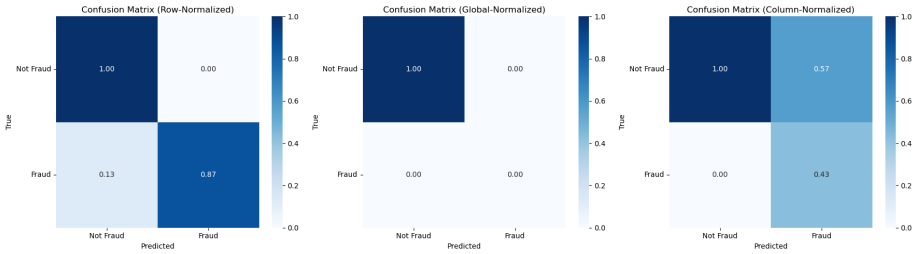

Figura 1: Matrices de confusión – Detección de fraude con tarjeta de crédito.

La matriz normalizada por filas permite identificar el porcentaje de transacciones reales correctamente clasificadas en cada clase, resultando la más útil para interpretar el rendimiento. El modelo logró buena sensibilidad frente a los casos de fraude, aunque persistieron algunos falsos negativos, propios de este tipo de problemas.

3.2. Customer Defection (Churn Prediction)

En este flujo se aplicó un *pipeline* con SMOTE y *XGBoost*, optimizando el balance entre precisión y *recall*. La Figura 2 presenta las tres variantes de matrices de confusión generadas.

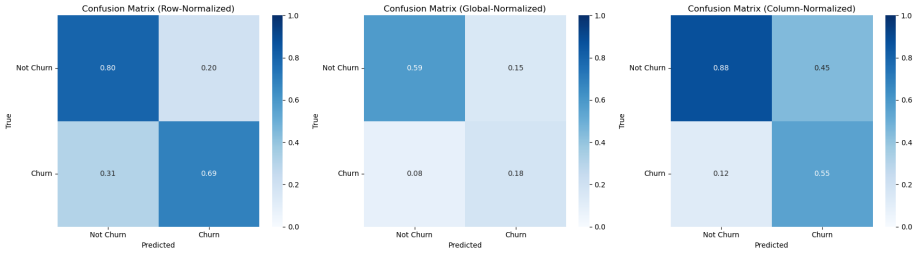

Figura 2: Matrices de confusión – Predicción de deserción de clientes.

La matriz normalizada por filas muestra que alrededor del 85 % de los clientes que no desertan fueron clasificados correctamente como *Not Churn*, mientras que el 60 % de los clientes que sí desertan fueron identificados como *Churn*. Las versiones global y por columnas complementan la interpretación, pero tienden a diluir las proporciones entre clases.

3.3. Bike Rental Regression

En el caso de regresión se emplearon modelos de tipo *Random Forest Regressor*, *Linear Regression* y *XGBoost Regressor*. El mejor desempeño se obtuvo con *Random Forest*, con un R^2 superior a 0.80. La Figura 3 muestra la relación entre los valores reales y las predicciones del número total de alquileres de bicicletas.

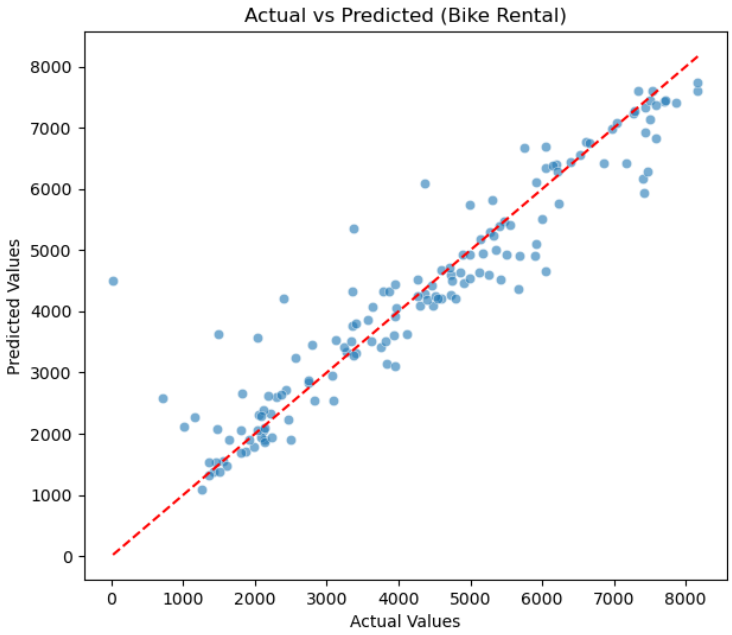

Figura 3: Relación entre valores reales y predichos – Bike Rental.

Los puntos se concentran alrededor de la línea diagonal, indicando un ajuste adecuado del modelo, con ligeras desviaciones en los valores extremos.

4. Discusión

De los resultados obtenidos se destacan los siguientes aspectos:

- **Importancia del balanceo:** El uso de SMOTE mejoró considerablemente el desempeño de los clasificadores en conjuntos desbalanceados.

- **Interpretación de matrices de confusión:** La versión normalizada por filas es la más útil para evaluar el rendimiento por clase real.
- **Versatilidad del pipeline:** La estructura modular permitió reutilizar la misma lógica para tareas de clasificación y regresión.
- **Rendimiento de modelos:** En clasificación, *XGBoost* presentó el mejor equilibrio entre precisión y sensibilidad; en regresión, *Random Forest* ofreció mayor estabilidad y capacidad de generalización.

5. Conclusiones

El Taller 4 consolidó el uso de flujos integrales de *Machine Learning* mediante el uso de pipelines y balanceo de clases. Se comprobó que la combinación de técnicas adecuadas de preprocesamiento y selección de modelos mejora la efectividad de las soluciones analíticas. Asimismo, las métricas visuales —como las matrices de confusión normalizadas y los gráficos de regresión— son herramientas clave para interpretar el rendimiento de los modelos.