Metodologia y Proceso de Desarrollo

Este informe presenta el desarrollo y la evaluación de un modelo de clasificación para predecir la estabilidad de la red eléctrica.

El objetivo era maximizar la precisión del modelo, asegurando un equilibrio entre la interpretabilidad y el rendimiento y para eso he decidido realizarlo en tres fases principals:

1️⃣ **Modelo Base:** Se implementó un modelo inicial siguiendo la rúbrica del módulo para establecer un punto de referencia en la clasificación de la estabilidad de la red.

2️⃣ **Feature Engineering:** Se aplicaron mejoras como la creación de nuevas características (polinomiales e interacciones) y la selección de las más relevantes con SHAP. He probado eliminar variables que tenian alta correlacion pero el modelo me bajaba bastante las metricas (por debajo de 0.8 todas) por lo que interprete que incluso con lata correlacion, las variables eran importantes para el modelo.

3️⃣ **Ensemble:** Despues de probar varias combinaciones de modelos, para mejorar el rendimiento entre SVM, XGBoost y Regresion log;istica, el que major resultado me dio en el train fue un **Voting Classifier** que combina XGBoost y Regresión Logística.

Análisis Exploratorio de Datos (EDA)

Se llevó a cabo un análisis exploratorio sobre el conjunto de datos, observando los siguientes puntos clave:

- **Cantidad de Datos:**
  - Conjunto de entrenamiento: 7,500 registros con 13 columnas.
  - Conjunto de prueba: 2,500 registros con 13 columnas.
- **Variables en el dataset:**
  - Se incluyen **4 retardos (tau1, tau2, tau3, tau4)**, **4 potencias (p1, p2, p3, p4)** y **4 conductancias (g1, g2, g3, g4)**.
  - La variable objetivo es **stabf** (estable/inestable).
- **No hay valores nulos.**
- **Distribución de las variables:** Se analizaron los estadísticos descriptivos para identificar sesgos o valores atípicos.
- **Correlaciones:** Se exploraron correlaciones entre las variables, identificando relaciones relevantes.

Feature Engineering realizado en la Segunda Etapa

Para mejorar el rendimiento del modelo, se implementaron varias técnicas de ingeniería de características:

✅ **Escalado de Características:** Se aplicó estandarización para normalizar las variables numéricas.

✅ **Creación de Características Polinómicas:** Se generaron términos cuadráticos para capturar relaciones no lineales.

✅ **Interacciones de Características:** Se crearon multiplicaciones entre pares de variables relevantes.

✅ **Selección de Características con SHAP:** Se analizaron las variables más importantes para el modelo XGBoost y se eliminaron las irrelevantes.

Modelos Evaluados y Comparación de Resultados

Se entrenaron y evaluaron varios modelos de clasificación para identificar el mejor rendimiento:

Modelo	**Precisión**	**ROC-AUC**
Regresión Logística	0.7907	0.8595
SVM	0.8367	-
Árbol de Decisión	0.7900	0.7743
Random Forest	0.8453	0.9160
XGBoost	0.8500	0.9198
Ensamble de Votación	**0.8708**	**0.9387**

📌 **El mejor modelo obtenido es el Ensamble de Votación (XGBoost + Regresión Logística).**

5. Modelo Final: Voting Ensemble

El modelo final combina:

✅ **XGBoost:** n_estimators=300, max_depth=9, learning_rate=0.2, subsample=0.8.
✅ **Regresión Logística:** solver='saga', max_iter=10000.
✅ **Votación Suave (Soft Voting)** para combinar probabilidades y mejorar la predicción.

📌 **Resultados Finales en el Conjunto de Prueba:**

Métrica	Valor
Precisión Final	**0.8708**
ROC-AUC Final	**0.9387**

Este modelo ofrece el mejor equilibrio entre interpretabilidad y rendimiento, logrando una clasificación precisa de la estabilidad de la red eléctrica.

6. Conclusiones y Futuras Mejoras

📌 **Conclusiones:**

- Se exploraron diferentes modelos de clasificación.
- Se aplicaron técnicas avanzadas de ingeniería de características.
- Se evaluaron ensambles para mejorar la generalización del modelo.
- El **Ensamble de Votación (XGBoost + Regresión Logística)** resultó ser la mejor elección.

🚀 **Futuras Mejoras:**

1. Optimizar hiperparámetros de XGBoost y Regresión Logística con Gridsearch (dado el calculo hecho con la version base de colab necesitaria unas 10 horas, por tanto no he hecho la prueba).
2. Implementar un modelo de **Stacking** con un metamodelo más sofisticado.
3. Aplicar técnicas de reducción de dimensionalidad como PCA.

📌 **Modelo Final para Entrega: Ensamble de Votación (XGBoost + Regresión Logística)**

🎯 **Precisión: 0.8708 | ROC-AUC: 0.9387**