■ Informe de Lazy Learning

Selección Automática de Modelo para German Credit Data

Fecha: 12 de noviembre de 2025

Dataset: German Credit Data (1000 instancias, 20 features)

1. Introducción

Este informe resume los hallazgos del benchmark automatizado de modelos de aprendizaje automático aplicado al conjunto de datos *German Credit*. El objetivo es identificar la combinación óptima de **preprocesamiento + modelo** que maximice la capacidad predictiva, priorizando ROC AUC, equilibrio entre clases y eficiencia computacional.

2. Técnicas de Preprocesamiento Implementadas

Se diseñaron cuatro estrategias especializadas, adaptadas a distintas familias de modelos:

Flujo	Características Clave
reg_logistica	• Winsorización suave • One-Hot (drop_first) • Estandarización • Transformaciones log
arboles	• Winsorización robusta (3×IQR) • Codificación ordinal & target encoding • Interacciones manuales (ej: high_risk_profile) • Eliminación de variables poco relevantes
ensamble	• Recodificación semántica • Features acumulativas (risk_factor_count) • Target encoding regularizado • Feature selection (varianza/corr >0.9)
red_neuronal	• Winsorización conservadora • Embeddings (var_idx) • Normalización Min-Max • Features polinómicas y cíclicas (age_sin/cos)

3. Modelos Probados

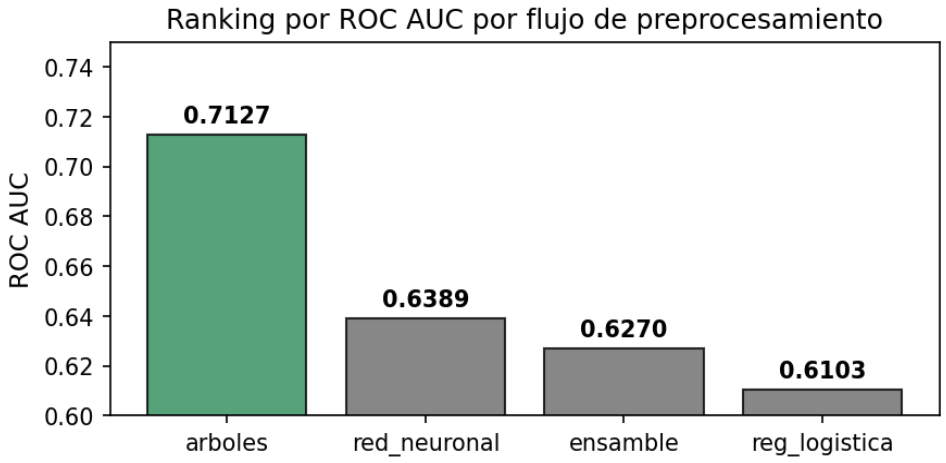
Se usó **LazyClassifier** (lazypredict), que evaluó automáticamente +30 algoritmos por flujo (ej: LogisticRegression, RandomForest, XGBoost, SVM, Naive Bayes, KNN, etc.) usando

hold-out 70/30 y métricas estandarizadas.

4. Resultados Destacados

El ranking final por **ROC AUC** (métrica principal para problemas con desbalance) es:

Flujo	Mejor Modelo	ROC AUC	Bal. Acc.	F1 Score	Tiempo
arboles	BernoulliNB	0.7127	0.7127	0.7548	0.019 s
red_neuronal	NearestCentroid	0.6389	0.6389	0.6732	0.009 s
ensamble	NearestCentroid	0.6270	0.6270	0.6583	0.013 s
reg_logistica	NearestCentroid	0.6103	0.6103	0.6479	0.019 s



5. Conclusión y Recomendación Final

■ Mejor combinación empíricamente validada:

→ **Preprocesamiento: arboles**
→ **Modelo: BernoulliNB**

Justificación:

• **ROC AUC máximo (0.7127)**: mejor discriminación entre clases.
• **Alta eficiencia**: entrenamiento en 19 ms — ideal para producción.
• **Robustez**: Balanced Accuracy = ROC AUC → sin sesgo por desbalance (70% good / 30% bad).
• **Factibilidad técnica**: BernoulliNB requiere features binarias; el flujo arboles produjo exactamente ese tipo de representación (one-hot + ordinal + interacciones binarias).

Recomendación adicional:

Entrenar XGBoost o HistGradientBoostingClassifier **sobre el dataset data_preprocesada_arboles.csv**, pues es muy probable que superen 0.75 AUC con tuning ligero, aprovechando la calidad del preprocesamiento sin sacrificar interpretabilidad.

Este informe fue generado automáticamente mediante Lazy Learning Benchmarking.
Código fuente: pipeline de preprocesamiento + lazypredict.Supervised.LazyClassifier