TP2 - Organización de datos

TABLA 1

Nombre	**Detalle**	**Función**
Base	Se encarga de encodear todas las clases categóricas de baja cardinalidad. Además rellena con el promedio la “edad” que tiene un porcentaje de 20% de missing values.	common_preprocessing
Escalador	Se encarga de normalizar y escalar los datos para los modelos sensibles a ruido como KNN. Además llama al preprocesamiento base.	knn_preprocessing
Get Train	Se encarga de separar el dataset en una parte para train y otra (el 20%) para validación	get_train_test_data
Get Dataset	Devuelve el dataset completo	get_dataset

TABLA 2

Modelo	**Preprocesamiento**	**AUC-ROC**	**Accuracy**	**Precision**	**Recall**	**F1 score**
1-DecisionTree	Base	0.8080	0.8199	0.82	0.81	0.79
2-KNN	Escalador	0.7988	0.8198	0.83	0.80	0.81
3-SVM	Escalador	0.8291	0.8447	0.85	0.83	0.84
4-Random Forest	Base	0.8116	0.8323	0.84	0.81	0.82
5-XGBoost	Base	0.8041	0.8261	0.84	0.80	0.81
6-Stacking	Escalador	0.8192	0.8386	0.85	0.82	0.83

Conclusión

Durante la primera etapa del trabajo conseguimos un accuracy de 0.8052 mientras que en las métricas de esta etapa se llegaron a ver accuracy de hasta 0. 8447 siendo una mejora de 4% . Además tomamos como métrica AUC-ROC que nos da más información sobre el comportamiento de los modelos no sólo en un % de acierto sino también sobre qué tipo de errores cometió (false positive, false negative).
Luego de probar una gran variedad de modelos, recomendamos utilizar SVM para este caso no sólo porque es el modelo que mejore métricas generales tiene sino que además es el que tiene mayor cantidad de true positive que son aquellos usuarios a los que vamos a estar atacando con las comunicaciones de marketing y un bajo número de false positive que serían esos usuarios donde vamos a comunicarles la intención de que vuelvan y nos generarían pérdida de dinero.