TABLA 1

Nombre Preprocesamiento	Explicación	Nombre de función
rfe_and_iterative_imputer	Convierte algunas columnas usando hashing_trick, rellena los missings con IterativeImputer, luego reduce la cantidad de columnas del dataset a las halladas con RFE.	preprocesamiento_arbol
hashing_trick_forward_selection_iterative_imputer_scaler	Convierte algunas columnas usando hashing_trick, reduce la cantidad de columnas usando VarianceThreshold y forward selection, y rellena los missings con IterativeImputer. Por último se escalan todos los valores numéricos a un rango entre cero y uno.	preprocessing_knn
continuous_mean_filler	Se queda con los features continuos y rellena los missings con el promedio.	preprocessing_continuos
hashing_trick_mean_scaler	Convierte algunas columnas categóricas a numéricas usando hashing trick, rellena los missings con el promedio y escala los datos	preprocessing_mean_scaled
regularization_iterative_imputer	Reduce la cantidad de columnas del dataset debido al análisis de regularización, luego rellena los missings con IterativeImputer.	preprocessing_imputer_filtered
hashing_trick_simple_imputer	Convierte algunas columnas usando hashing_trick, luego rellena los missings con SimpleImputer.	preprocessing_gb

TABLA 2

Nombre Modelo	Nombre Preprocesamiento	AUC ROC	Accuracy	Precisión	Recall	F1 Score
Árbol de decisión	rfe_and_iterative_imputer	0.849	0.84	0.83	0.84	0.83
KNN	hashing_trick_forward_selection_iterative_imputer_scaler	0.840	0.84	0.83	0.84	0.83
Naive Bayes	continuous_mean_filler	0.828	0.82	0.81	0.82	0.81
SVM	hashing_trick_mean_scaler	0.743	0.74	0.76	0.74	0.75
Redes Neuronales	regularization_iterative_imputer	0.843	0.84	0.82	0.84	0.82
Gradient Boosting	hashing_trick_simple_imputer	0.888	0.86	0.85	0.86	0.85

El modelo que más recomendamos es Gradient Boosting ya que es el que presenta un AUC ROC mayor. Las demás métricas también son significativamente superiores al usar este modelo.

Si quisiéramos tener la menor cantidad de falsos positivos posible, el modelo que más conviene es Gradient Boosting, ya que es el que presenta una mayor precisión (TP / (TP+FP)). Por otra parte, si se quiere tener una lista con todos los días que potencialmente lloverán también recomendamos Gradient Boosting, porque es el que mayor recall tiene (menor cantidad de falsos negativos, es decir, positivos que fueron mal predichos).