Checkpoint 3 - Grupo 07

Introducción

Para este tercer checkpoint, se partió del preprocesamiento realizado durante el checkpoint 2. Los mismos datos fueron utilizados para entrenar a todos los modelos detallados a continuación.

Para encontrar los mejores resultados e hiperparametros, se probó grid y random search CV con 5 y 10 folds.

Por otro lado, en el caso particular del modelo SVM, los datos del set fueron escalados con estandarización y min-max.

Construcción del modelo

Detallar como mínimo los siguientes puntos:

- Parametros optimizados en KNN: **'N_neighbors', ‘'weights'**, **'algorithm'** y **'metric**'.

- Hiperparámetros optimizados para SVN: Se realizó Grid/Random Search para los 3 tipos de kernel:
    - **Linear : ‘C’**
    - **RBF: ‘C’ y ‘gamma’**
    - **Poly: ‘C’, ‘Gamma’, ‘degree’**

- Hiperparámetros optimizados para RF: **"Criterion” ,"min_samples_leaf, "min_samples_split"** y **"n_estimators”.**

- Hiperparámetros optimizados para XGBoost: **'learning_rate', 'n_estimators','max_depth', 'min_child_weight','subsample'** y **'Colsample_bytree'.**

- Modelos usados para el ensamble tipo voting: **KNN, SVM(kernel-bfc), RF** y **XGBoost.**

- Modelos usaron para el ensamble tipo stacking: **KNN, SVM(kernel-bfc), RF** y **XGBoost.**
- Metamodelo: **Regresor logistico.**

Cuadro de Resultados

Modelo	F1-Test	Precision Test	Recall Test	Kaggle
KNN	0.79	0.77	0.80	**0.77347**
SVM	0.84	0.82	0.86	**0.83889**
Random Forest	0.88	0.87	0.88	**0.87537**
XGBoost	0.88	0.88	0.88	**0.86218**
Voting	0.83	0.76	0.93	**0.84581**
Stacking (**El mejor**)	0.89	0.89	0.88	**0.87864**

Nota: indicar brevemente en qué consiste cada modelo de la tabla y detallar el caso del mejor modelo.

RF: es un ensemble "bagging" (Bootstrap Aggregating). Este método de ensamble construye múltiples árboles de decisión y combina las predicciones de estos mediante votación (clasificación) o promedio (regresión).

KNN: Es un algoritmo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. En lugar de crear un modelo durante el proceso de entrenamiento, KNN toma un enfoque de lazy learning donde memoriza el conjunto de datos de entrenamiento. Cuando se realiza una predicción en un nuevo punto de datos, KNN busca los "k" puntos de entrenamiento más cercanos (donde "k" es un número que define el usuario) y hace una predicción basada en la mayoría de las clases (en clasificación) o en la media (en regresión) de los "k" puntos más cercanos.

SVM: construye un hiperplano o conjunto de hiperplanos en el espacio original cuando los conjuntos son linealmente separables o bien en el espacio transformado (espacio de características aumentado) cuando los conjuntos no son linealmente separables.

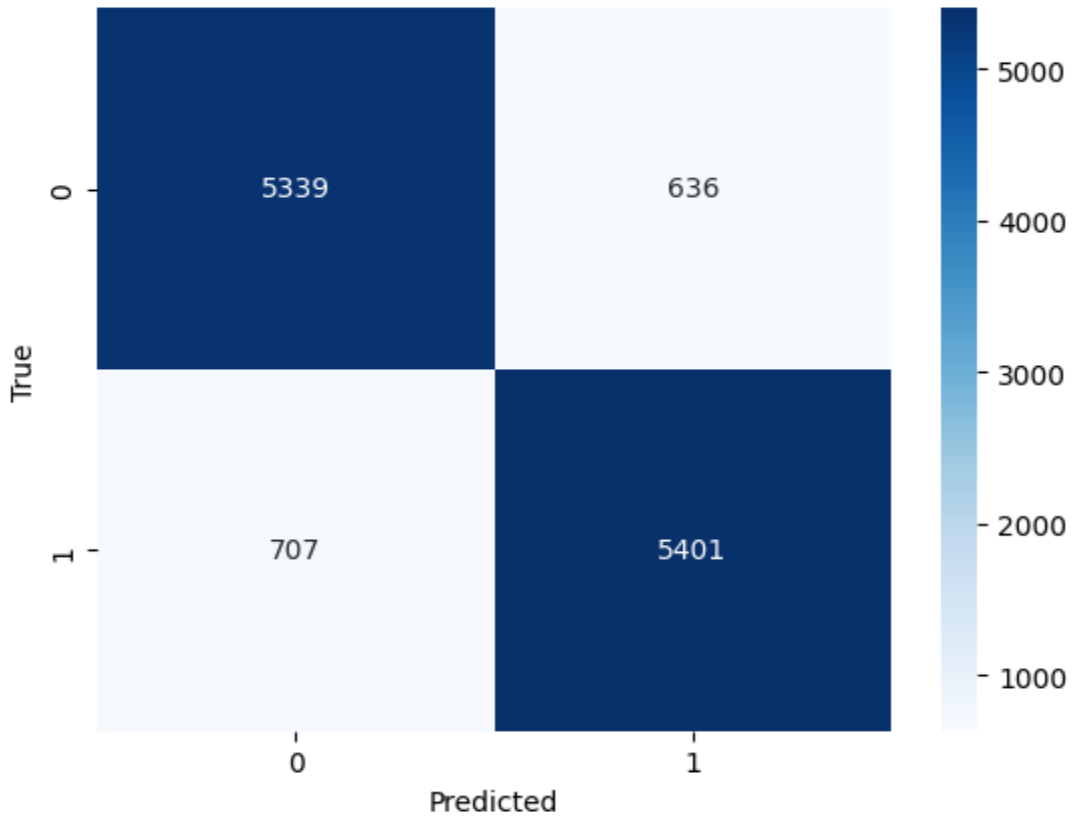
Para esto último se apoya en el uso de funciones kernel, que toma datos de espacio de características original y los mapea a un espacio de características de mayor dimensión.

XGBoost: Es una implementación eficiente y escalable de la técnica de "boosting", que es un enfoque de aprendizaje automático que combina múltiples modelos de aprendizaje débil para crear un modelo más fuerte y preciso.
Voting: La idea principal detrás de un modelo de votación es combinar las predicciones de varios modelos individuales (clasificadores o regresores) para obtener una predicción final más robusta y precisa. Es especialmente útil cuando se tienen varios modelos que pueden aportar diferentes enfoques o perspectivas para un problema de aprendizaje automático.

Stacking: Al igual que el modelo de voting, el stacking combina las predicciones de varios modelos individuales para obtener una predicción final. Sin embargo, el stacking va un paso más allá al utilizar modelos adicionales (llamados meta-modelos o modelos de nivel superior) que toman las predicciones de los modelos base como entrada.

Matriz de Confusion de modelo Stacking

La matriz de confusión mostrada a continuación nos permite ver que el modelo de tipo stacking tiene un balance bastante óptimo entre true positives y true negatives, y lo mismo podría decirse respecto a los false positives y negatives.



Tareas Realizadas

Integrante	Tarea
Rafael Wu	MSV, RF, ensambles híbridos y armado de reporte
Guido Menendez	XGB, ensambles híbridos y armado de reporte
Mateo Riat Sapulia	KNN, ensambles híbridos y armado de reporte