Informe Final - Grupo 24

Introducción

Haciendo una retrospectiva de todo el trabajo realizado, en el CHP1 realizamos toda la parte de análisis de variables, visualización de datos y detección e imputación de datos atípicos y faltantes. Luego en el CHP2 empezamos con la implementación de los modelos predictores como por ejemplo el DesicionTreeClassifier, optimizando sus hiperparametros con RandomForest y Cross Validation. Con este modelo fue con el que obtuvimos nuestra primer entrega en Kaggle que supera 0.81 (0.81967 en Kaggle), este número se mantuvo tanto en Test como en Kaggle, los que nos indicaba que no estábamos bajo un modelo con overfitting. Pasando al CHP3, se implementaron modelos más robustos como SMV con distintos Kernels, Random Forest, XGBoost y ensambles de Stacking y Voting. De todos los modelos nombrados anteriormente nuestro mejor predictor fue XGBoost superando ampliamente las otras predicciones con 0.8654 en kaggle y un resultado apenas mejor en Test de 0.8740. El resto de los modelos no lograron superar el 0.84 por lo que optamos por no optimizarlos.
Finalizando en Trabajo Práctico implementamos una red neuronal aplicando distintos regularizadores,como L1, L2, L1 Y L2 y Dropout, sin embargo no obtuvimos resultados que superen los anteriores.

Cuadro de Resultados

Modelo	CHPN	F1-Test	Presicion Test	Recall Test	Accuracy	Kaggle
DesicionTree Classifier	2	0.81973	0.8260	0.8135	0.8149	0.81967
XGBoost	3	0.8740	0.8646	0.8835	0.8720	0.8654
Red Neuronal Regularizada con Dropout	4	0.7991	0.6923	0.9292	0.8225	0.8225
Red Neuronal Regularizada con L2	4	0.7844	0.6857	0.9165	0.8137	0.8160

En el CHP1 no se realizó ninguna predicción.
En el CHP2, la mejor predicción obtenida, se logró mediante un DesicionTreeClassifier optimizado con RandomSearch .
En el CHP3 conseguimos nuestra mejor predicción de todo el trabajo práctico utilizando XGBoost con sus parámetros optimizados con RandomSearch.
Finalmente en el CHP4 mediante redes neuronales optimizadas con GridSearch no logramos superar nuestra predicción más alta.

Conclusiones generales

Habiendo finalizado el trabajo práctico pudimos observar la importancia de la detección e imputación de los outliers con las técnicas adecuadas, ya que una vez que logramos encontrar estas técnicas nuestras predicciones mejoraron considerablemente, además de la discretización de distintas clases.
Hubo diferentes opciones que no exploramos ya sea por la falta de conocimiento como por ejemplo: inteligencia artificial, o porque no fueron pedidas en las consignas como pueden ser el ensamble cascading.
Una vez con todos los modelos solicitados a lo largo del trabajo práctico observamos que el DesicionTreeClassifier sin ningún tipo de modificación fue el que resultó ser más rápido y además de ser sencillo. A pesar de esto no resultó un resultado óptimo a comparación de otros modelos, como puede ser el XGBoost que en poco más de 6 minutos obtuvo nuestra predicción más alta. Sin embargo creemos que se puede mejorar esta predicción a través de un análisis del dataset más exhaustivo y optimizar los hiperparametros con GridSeach Cross Validation en el modelo que se utilizó XgBoost, el cual no implementamos debido a su alto tiempo de ejecución.
Para finalizar creemos que nuestro mejor predictor (XGBoost) sería un buen modelo en predicción ya que un f1_score de 0.8654 es un resultado bastante elevado.

Tareas Realizadas

Integrante	Promedio Semanal (hs)
Lautaro Torraca	16
Gianluca Negrotti	14
Marco Tosi	14