Organización de datos - TP 1

Grupo 10 - **Checkpoint 3:**

El trabajo consistió en entrenar varios modelos de tipo cluster. Inicialmente, se realizó un preprocesamiento para transformar las variables categóricas en numéricas, para que los modelos pudieran utilizarlas. Se normalizaron las variables numéricas para que ciertos features no tuvieran mayor impacto que otros a la hora del entrenamiento.

Se entrenó un modelo KNN, y se utilizó Random Search para encontrar los mejores hiperparámetros del modelo. El F1 Score obtenido fue 0.8338679989481987, lo cual nos parece bueno ya que es un modelo más básico que el resto que vamos a realizar. Esto indica que es capaz de realizar predicciones sólidas y puede ser una buena opción inicial para clasificación.

Posteriormente, se entrenó un modelo SVM, variando entre los kernels lineal, polinómico y radial. Para el modelo SVM, se realizó un preprocesamiento adicional debido a que el entrenamiento tomaba mucho tiempo. Se eliminaron varias columnas para reducir el tiempo de entrenamiento y se utilizó PCA para reducir la dimensionalidad del conjunto de datos. El resultado del modelo lineal dio bastante bajo, aproximadamente un F1 Score de 0.72. Esto, creemos que se debe al preprocesamiento agresivo que hicimos para reducir los tiempos de cómputo. No pudimos obtener predicciones del polinómico y radial, aun dejándolos más de varias horas y modificando los parámetros.

Se entrenó también un modelo de Random Forest, utilizando Grid Search para encontrar los mejores hiperparámetros. El F1 Score obtenido fue de 0.8795475556038936. Los mejores parámetros encontrados fueron: {'criterion': 'entropy', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 50}. Fue un gran resultado, el mejor hasta el momento.

Asimismo, se entrenó el modelo XGBoosting, optimizando los hiperparámetros con Cross Validation. Los mejores parámetros encontrados fueron:

> {'subsample': 1, 'n_estimators': 100, 'min_child_weight': 1, 'max_depth': 16, 'learning_rate': 0.1, 'gamma': 0, 'colsample_bytree': 0.7}.

Y el F1 Score obtenido fue de 0.8797610892265201, una muy buena predicción. Esto indica que es un modelo robusto y puede proporcionar predicciones precisas y confiables.

Finalmente, se entrenaron dos ensambles híbridos. El primero, del tipo Stacking, donde el metamodelo fue regresión logística y los modelos base fueron RF, SVM y KNN. Se obtuvo una predicción confiable con un F1 Score de 0.8796311364252781. El segundo, del tipo Voting, donde los clasificadores fueron regresión logística, KNN y RF. El F1 Score de este modelo fue un poco inferior, pero sigue siendo alto: 0.8568961795812647. Los ensambles híbridos, tanto el tipo Stacking como el tipo Voting, demostraron buenos resultados en términos de F1 Score.

En general, los resultados indican que los modelos XGBoosting y Random Forest fueros los más exitosos en términos de rendimiento, seguido de cerca por el ensamble Stacking.