Organización de datos, Curso Rodriguez
1er Cuatrimestre 2023

Docente: Rodriguez, Juan Manuel.
Estudiantes:

- Mendoza Hernandez, Sabrina Scarlet (108524)
- Valeriani, Matias Gabriel (108570)
- Jang, Lucas (109151)

Informe final: Reservas de Hotel

A lo largo del trabajo práctico realizamos las tareas de preprocesamiento y transformación de los datos de un dataset de entrenamiento. El objetivo principal del TP1 es aplicar técnicas de análisis exploratorio, preprocesamiento de datos y entrenamiento de modelos de clasificación para predecir si una reserva va a ser cancelada. Comenzamos realizando un análisis exploratorio, el cual con diferentes métodos como, observando correlaciones entre variables, ver datos faltantes y analizar valores atípicos logramos quedarnos con los mejores features (variables) y dejar el dataframe de entrenamiento apto para el análisis correspondiente.
Los clasificadores que utilizamos en el trabajo fueron: Árboles de decisión, KNN, SVM, Random Forest, XGBoost, Voting, Stacking y Red Neuronal.
Mostraremos la performance obtenida por cada uno de estos modelos en nuestro set de entrenamiento, luego mostraremos la calificación obtenida en Kaggle de cada uno de los mismos.

- Árboles de decisión:
  - F1-Score: 0.85897
  - Accuracy: 0.85
  - Precision: 0.84
  - Recall: 0.87

- Clasificador KNN:
  - F1-Score: 0.79
  - Accuracy: 0.78
  - Precision: 0.77
  - Recall: 0.81

- Clasificador SVM (Kernel polinómico):
  - F1-Score: 0.837
  - Accuracy: 0.836
  - Precision: 0.845
  - Recall: 0.83

- Clasificador Random Forest:
  - F1-Score: 0.879
  - Accuracy: 0.877
  - Precision: 0.881

    - Recall: 0.877
- Clasificador XGBoost:
    - F1-Score: 0.873
    - Accuracy: 0.87
    - Precision: 0.869
    - Recall: 0.876

- Stacking:
    - F1-Score: 0.876
    - Accuracy: 0.874
    - Precision: 0.877
    - Recall: 0.875

- Voting:
    - F1-Score: 0.878
    - Accuracy: 0.876
    - Precision: 0.882
    - Recall: 0.873

- Red Neuronal:
    - F1-Score: 0.847
    - Accuracy: 0.85
    - Precision: 0.84
    - Recall: 0.86

En Kaggle los 3 mejores modelos fueron Stacking, Random Forest y XGBoost (la métrica que usa Kaggle es F1-Score). Las respectivas calificaciones obtenidas fueron:

- Stacking: 0.873
- Random Forest: 0.871
- XGBoost: 0.865

Como conclusión final podemos decir que el mejor clasificador obtenido tanto en nuestro entrenamiento, como con los datos del dataset de test, fue el clasificador de Stacking (el clasificador Voting al ser también un ensamble híbrido, dio buenas clasificaciones muy similares a Stacking, pero consideraremos el Stacking ya que nos dió mejores predicciones). Esto se debe a que el algoritmo de Stacking combina múltiples clasificadores para mejorar el rendimiento predictivo. Utiliza un metaclasificador que toma las predicciones de los clasificadores base como características de entrada y realiza una predicción final. Esto puede conducir a un mejor rendimiento en comparación con un solo clasificador. En nuestro caso lo hicimos con clasificadores base como Random Forest, XGBoost y Regresión Logística (utilizamos Random Forest y XGBoost ya que previamente nos habían dado muy buenas matéricas).