Clasificación / Predicción de Calidad del Vino (Core)

Hallazgos y conclusiones del análisis.

Para el análisis requerido se utilizarán técnicas de clasificación aprendidas hasta el momento para predecir la calidad del vino basándose en características físico-químicas. Este análisis permitirá aplicar conceptos como la selección de características, preprocesamiento de datos, entrenamiento y evaluación de modelos de clasificación, y análisis de resultados mediante métricas y visualizaciones.

Descripción del Dataset:

- Este conjunto de datos contiene información sobre distintas características físico-químicas de muestras de vino tinto y su calidad asociada.
- Las características incluyen acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol.
- La calidad del vino está clasificada en una escala del 0 al 10.

Despliegue de los datos

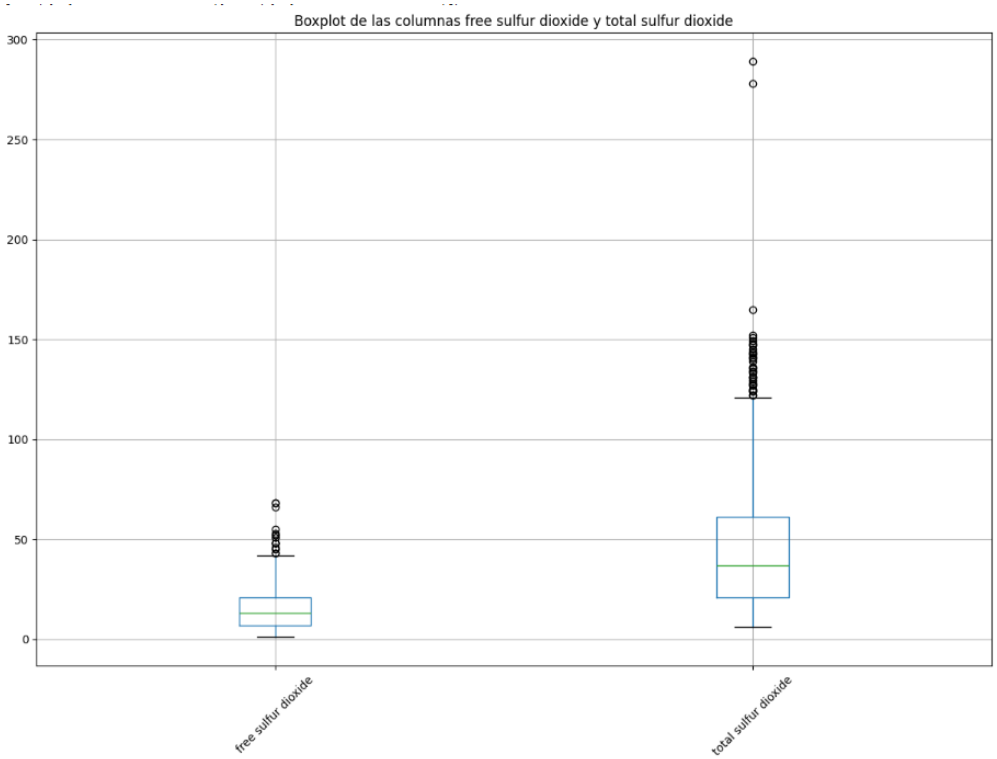
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	0
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	1
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	2
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	3
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	4

Ejecución de pasos

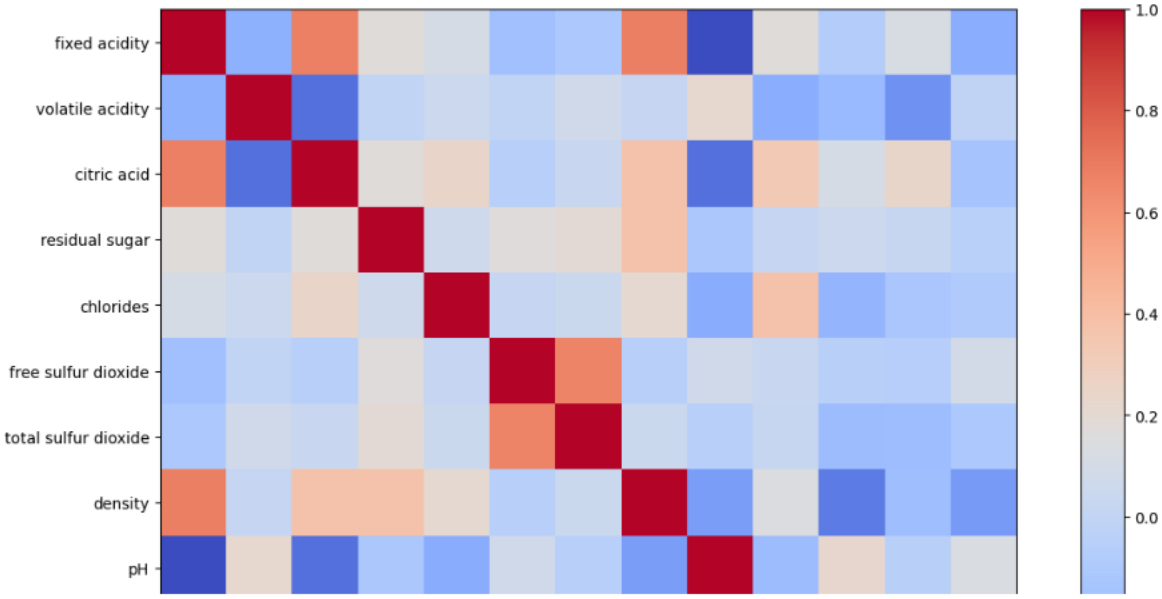
1. Carga y Exploración de Datos:

- Se carga el dataset y se revisa estructura básica.
- Se describen variables y su distribución por medio de un histograma donde se evidencia que los valores de calidad 5 y 6 son los predominantes en la muestra analizada.

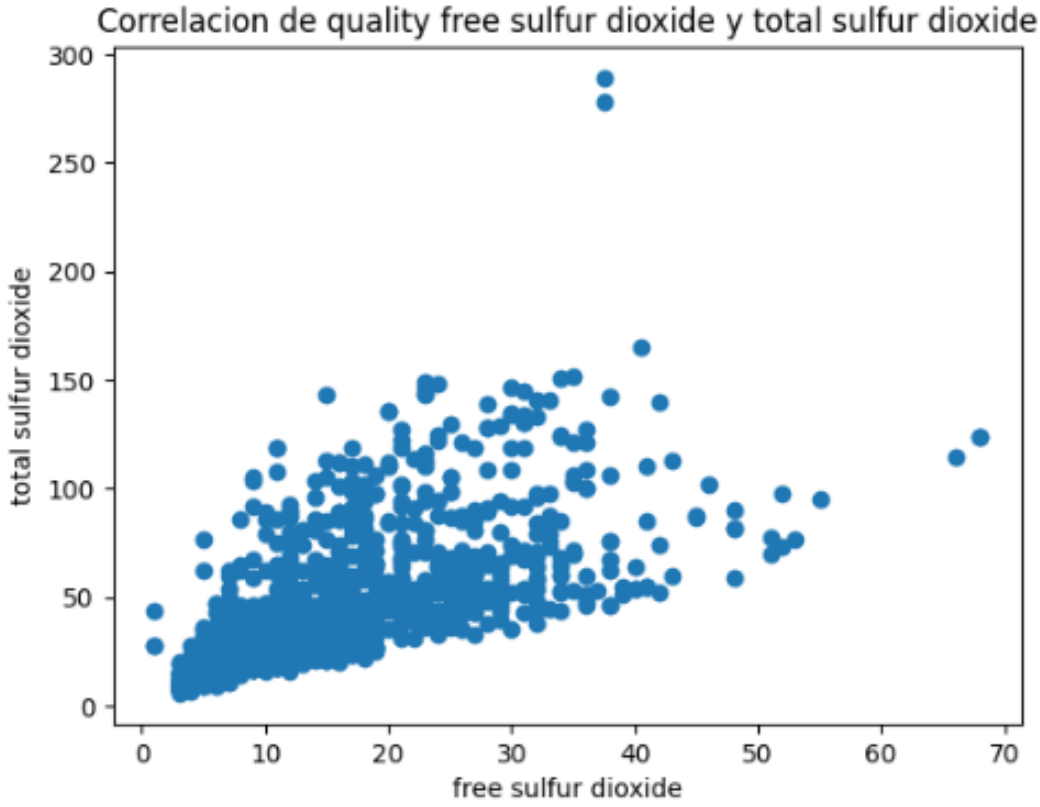
- No se evidencian valores nulos
- Detectan valores outliers.
  - Las columnas "free sulfur dioxide" y "total sulfur dioxide"



- Se evidencia relación entre algunas variables, como fixed acidity con density, como también entre free sulfur dioxide y total sulfur dioxide.



Revisamos le correlación entre free sulfur dioxide versus total sulfur dioxide



2. Preprocesamiento de Datos:

- Seleccionar características importantes para la clasificación.
- Transformar variables categóricas en variables numéricas si es necesario.
- Dividir los datos en conjuntos de entrenamiento y prueba.
- Escalar las características.
- Se obtiene listado de columnas predictoras, todas numéricas, para escalar.
- Se definen transformadores
- Se realiza entrenamiento para los tres modelos. (Creación del pipeline, Visualización de estructura del pipeline y entrenamiento del modelo)
    - Arbol de Decisión
    - KNN

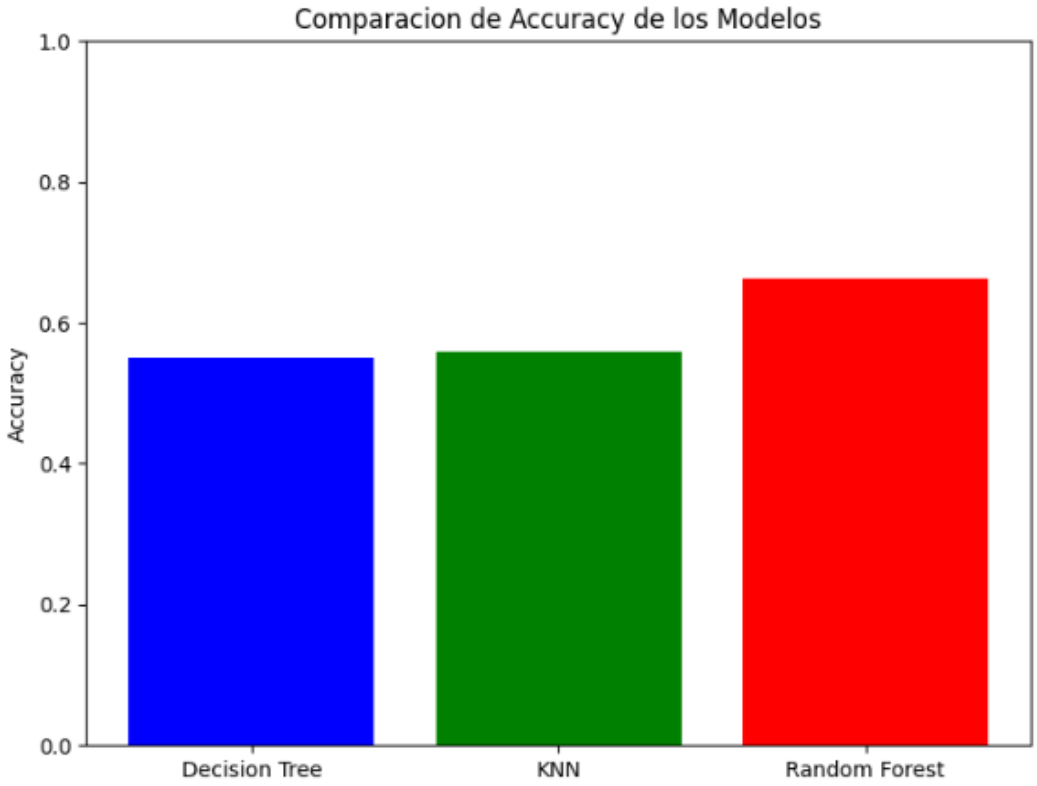
- Random Forest

3. Entrenamiento de Modelos de Clasificación:

- Se realiza entrenamiento para tres modelos, obteniendo los siguientes resultados que se evalúan en la Matriz de confusión que se muestra a continuación:

```
Matriz de confusión Arbol de Decision:
 [[ 0  0  0  0  0  0]
 [ 0  1  2  2  1  0]
 [ 1  7 58 30  0  0]
 [ 2  1 33 50 11  2]
 [ 0  0  0 10 16  0]
 [ 0  0  0  0  1  1]]
Matriz de confusión KNN:
 [[ 0  4  2  0  0]
 [ 2 66 28  0  0]
 [ 1 34 55  9  0]
 [ 0  5 14  7  0]
 [ 0  0  2  0  0]]
Matriz de confusión Random Forest:
 [[ 0  3  3  0  0]
 [ 0 75 20  1  0]
 [ 0 31 61  7  0]
 [ 0  0 10 16  0]
 [ 0  0  1  1  0]]
```

4. Análisis y Comparación de Resultados: Se compara el accuracy de los tres modelos:



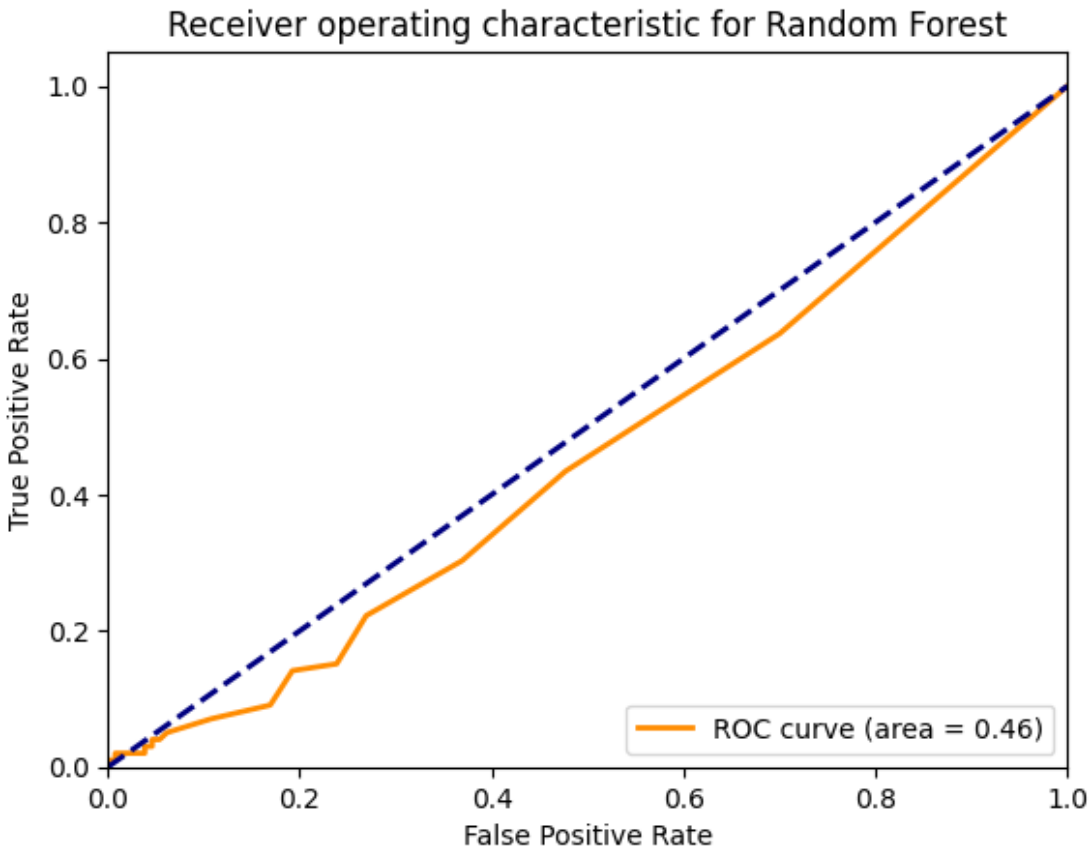
5. Evaluación de Modelos:

- CONCLUSION:

  Random Forest es el modelo de mejor rendimiento.

  Entre los tres, el Random Forest muestra la capacidad más fuerte para clasificar la calidad del vino, especialmente en las clases de calidad media (probablemente las más abundantes en el dataset).

- Se crea y visualiza la curva ROC y cálculo del AUC para el mejor modelo.



En resumen, la curva ROC con un AUC de 0.46 sugiere que el modelo Random Forest tiene problemas para distinguir las clases, que podría ser por un desbalance o definición de clases que el modelo no puede aprender con las características dadas.

Es posible que el EDA no haya tomado las mejores decisiones al trabajar con los datos.