Informe Detallado del Análisis de Calidad del Vino

Introducción

El objetivo principal es predecir la calidad del vino en una escala del 4 al 8 empleando modelos de clasificación como Random Forest, KNN y Regresión Logística.

Datos Utilizados

El dataset utilizado contiene las siguientes características físico-químicas del vino:

Fixed Acidity

Volatile Acidity

Citric Acid

Residual Sugar

Chlorides

Free Sulfur Dioxide

Total Sulfur Dioxide

Density

pH

Sulphates

Alcohol

La etiqueta de calidad (quality) es la variable objetivo.

Metodología

Preprocesamiento de datos:

Escalado de características usando StandardScaler.

Tratamiento de valores atípicos mediante el rango intercuartílico (IQR).

División de datos en conjuntos de entrenamiento (80%) y prueba (20%).

Modelos entrenados:

Random Forest Classifier

K-Nearest Neighbors (KNN)

Regresión Logística

Evaluación:

Métricas: Precisión, Recall, F1-Score, y Exactitud.

Curvas ROC y cálculo de AUC para evaluar el rendimiento multiclase.

Resultados del Análisis

Random Forest:

Precisión promedio ponderada: 64.6%

Mejor rendimiento en las clases 5 y 6.

AUC más alto para la clase 8 (0.88).

KNN:

Precisión promedio ponderada: 57.9%

Desempeño aceptable para la clase 5, pero limitado en clases extremas.

Regresión Logística:

Precisión promedio ponderada: 55.5%

Utilidad como modelo base, pero con bajo rendimiento general.

Curvas ROC

El rendimiento varía según la clase:

Clase 4: AUC = 0.51

Clase 5: AUC = 0.61

Clase 6: AUC = 0.31

Clase 7: AUC = 0.48

Clase 8: AUC = 0.88

El modelo tiene el mejor rendimiento para la Clase 8 y moderado para las Clases 5 y 6.

Conclusión

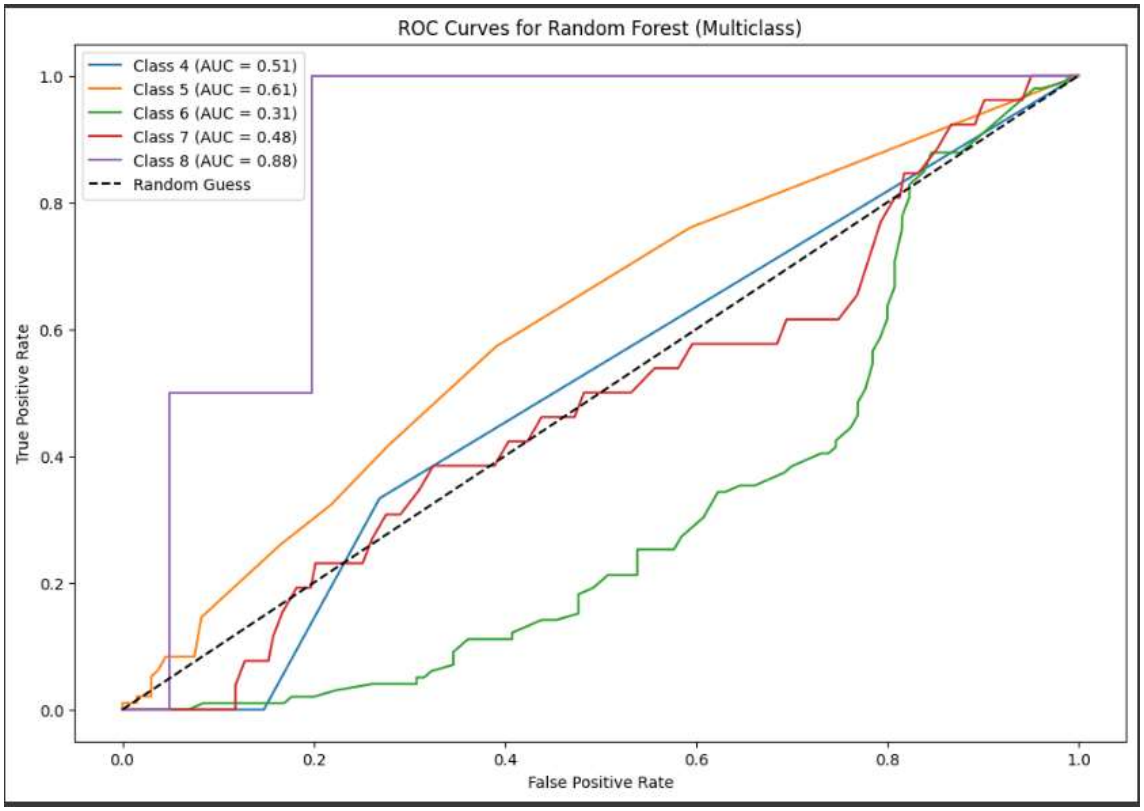
El modelo Random Forest demostró ser la mejor opción para este problema de clasificación multiclase, con un rendimiento superior en la mayoría de las métricas.

KNN y Regresión Logística, aunque útiles como líneas base, mostraron un rendimiento inferior.

Se recomienda optimizar los hiperparámetros y realizar técnicas de balanceo de clases para mejorar aún más el rendimiento en clases minoritarias.

Evidencias Gráficas

```
--- Random Forest ---
              precision    recall  f1-score   support

           4       0.00      0.00      0.00         6
           5       0.71      0.76      0.73        96
           6       0.64      0.67      0.65        99
           7       0.64      0.54      0.58        26
           8       0.00      0.00      0.00         2

    accuracy                           0.67       229
   macro avg       0.40      0.39      0.39       229
weighted avg       0.65      0.67      0.66       229

[[ 0  3  3  0  0]
 [ 1 73 21  1  0]
 [ 0 27 66  6  0]
 [ 0  0 12 14  0]
 [ 0  0  1  1  0]]
--- KNN ---
              precision    recall  f1-score   support

           4       0.00      0.00      0.00         6
           5       0.62      0.79      0.70        96
           6       0.60      0.51      0.55        99
           7       0.50      0.38      0.43        26
           8       0.00      0.00      0.00         2

    accuracy                           0.59       229
   macro avg       0.35      0.34      0.34       229
weighted avg       0.58      0.59      0.58       229

[[ 0  4  2  0  0]
 [ 2 76 18  0  0]
 [ 2 38 50  9  0]
 [ 0  4 12 10  0]
 [ 0  0  1  1  0]]
--- Logistic Regression ---
              precision    recall  f1-score   support

           4       0.00      0.00      0.00         6
           5       0.64      0.77      0.70        96
           6       0.56      0.56      0.56        99
           7       0.40      0.23      0.29        26
           8       0.00      0.00      0.00         2

    accuracy                           0.59       229
   macro avg       0.32      0.31      0.31       229
weighted avg       0.56      0.59      0.57       229

[[ 0  2  4  0  0]
 [ 0 74 20  2  0]
 [ 0 39 55  5  0]
 [ 0  1 19  6  0]
 [ 0  0  0  2  0]]
```


ROC Curves for Random Forest (Multiclass)
Class 4 (AUC = 0.51)
Class 5 (AUC = 0.61)
Class 6 (AUC = 0.31)
Class 7 (AUC = 0.48)
Class 8 (AUC = 0.88)
Random Guess
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0