INFORME: PREDICCIÓN DE CALIDAD DEL VINO

PORTADA

- **TÍTULO: Predicción de Calidad del Vino Basada en Características Físico-Químicas**
- **AUTOR: Alexis Frutos**
- **FECHA: 14/11/2024**

ÍNDICE DE CONTENIDOS

1. INTRODUCCIÓN

Objetivo: El objetivo de este proyecto es predecir la calidad del vino basándose en variables físico-químicas, aplicando técnicas de clasificación supervisada y optimización de modelos para construir un clasificador robusto.

Motivación: Predecir la calidad del vino es útil para la industria vinícola, ya que permite evaluar rápidamente la calidad del producto sin necesidad de pruebas sensoriales complejas y costosas.

2. DESCRIPCIÓN DEL DATASET

Fuente del Dataset: El conjunto de datos utilizado fue el **Wine Quality Dataset**.

CARACTERÍSTICAS DEL DATASET:

- **Variables**: Se cuenta con varias características físico-químicas, como acidez fija, acidez volátil, cloruros, pH, alcohol, entre otras.

- **Variable Objetivo**: `quality` - Clasificación de la calidad del vino en una escala de 0 a 10 (simplificada a categorías en el análisis).

RESUMEN ESTADÍSTICO:

- **CANTIDAD DE MUESTRAS: 1.143**

- **VARIABLES NUMÉRICAS: 1**
- **VARIABLES FLOTANTES: 11**
- **VALORES NULOS: 0**
- **OUTLIERS: 437**

---

3. METODOLOGÍA

Preprocesamiento de Datos:

- Eliminación de valores nulos y columnas irrelevantes.
- Selección de características y escalado de datos para estandarizar las entradas.

Modelos de Clasificación:

1. **K-Nearest Neighbors (KNN)**
2. **Random Forest**
3. **Regresión Logística**

Optimización de Hiperparámetros:

- Utilizamos `GridSearchCV` para realizar una búsqueda exhaustiva de hiperparámetros y seleccionar los valores óptimos para cada modelo.
- **Resultados de los mejores hiperparámetros**:
  - **KNN**: {'metric': 'manhattan', 'n_neighbors': 7, 'weights': 'distance'}
  - **Random Forest**: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}
  - **Regresión Logística**: {'C': 1, 'solver': 'lbfgs'}

---

4. RESULTADOS Y EVALUACIÓN

Para cada modelo, se midieron las siguientes métricas de rendimiento:

K-Nearest Neighbors (KNN)

- **Exactitud**: 0.64
- **Matriz de Confusión**:

```
[[ 0  4  2  0  0]
 [ 0 67 28  1  0]
 [ 0 27 60 12  0]
 [ 0  0  7 19  0]
 [ 0  0  2  0  0]]
```

- **Reporte de Clasificación**:
  - Precisión, Recall, F1-Score para cada clase

	precision	recall	f1-score	support
4	0.00	0.00	0.00	6
5	0.68	0.70	0.69	96
6	0.61	0.61	0.61	99
7	0.59	0.73	0.66	26
8	0.00	0.00	0.00	2
accuracy			0.64	229
macro avg	0.38	0.41	0.39	229
weighted avg	0.62	0.64	0.63	229

Random Forest

- **Exactitud**: 0.68
- **Matriz de Confusión**:

```
[[ 0  3  3  0  0]
 [ 0 76 19  1  0]
 [ 0 29 65  5  0]
 [ 0  0 11 15  0]
 [ 0  0  2  0  0]]
```

- **Reporte de Clasificación**:
  - Precisión, Recall, F1-Score para cada clase

	precision	recall	f1-score	support
4	0.00	0.00	0.00	6
5	0.68	0.73	0.70	96
6	0.62	0.64	0.63	99
7	0.48	0.42	0.45	26
8	0.00	0.00	0.00	2
accuracy			0.63	229
macro avg	0.36	0.36	0.36	229
weighted avg	0.61	0.63	0.62	229

Regresión Logística

- **Exactitud**: 0.63
- **Matriz de Confusión**:

```
[[ 0  3  3  0  0]
 [ 1 70 23  2  0]
 [ 0 28 63  8  0]
 [ 0  2 13 11  0]
 [ 0  0  0  2  0]]
```

- **Reporte de Clasificación**:
  - Precisión, Recall, F1-Score para cada clase

	precision	recall	f1-score	support
4	0.00	0.00	0.00	6
5	0.68	0.73	0.70	96
6	0.62	0.64	0.63	99
7	0.48	0.42	0.45	26
8	0.00	0.00	0.00	2
accuracy			0.63	229
macro avg	0.36	0.36	0.36	229
weighted avg	0.61	0.63	0.62	229

---

5. ANÁLISIS DE LA CURVA ROC Y AUC

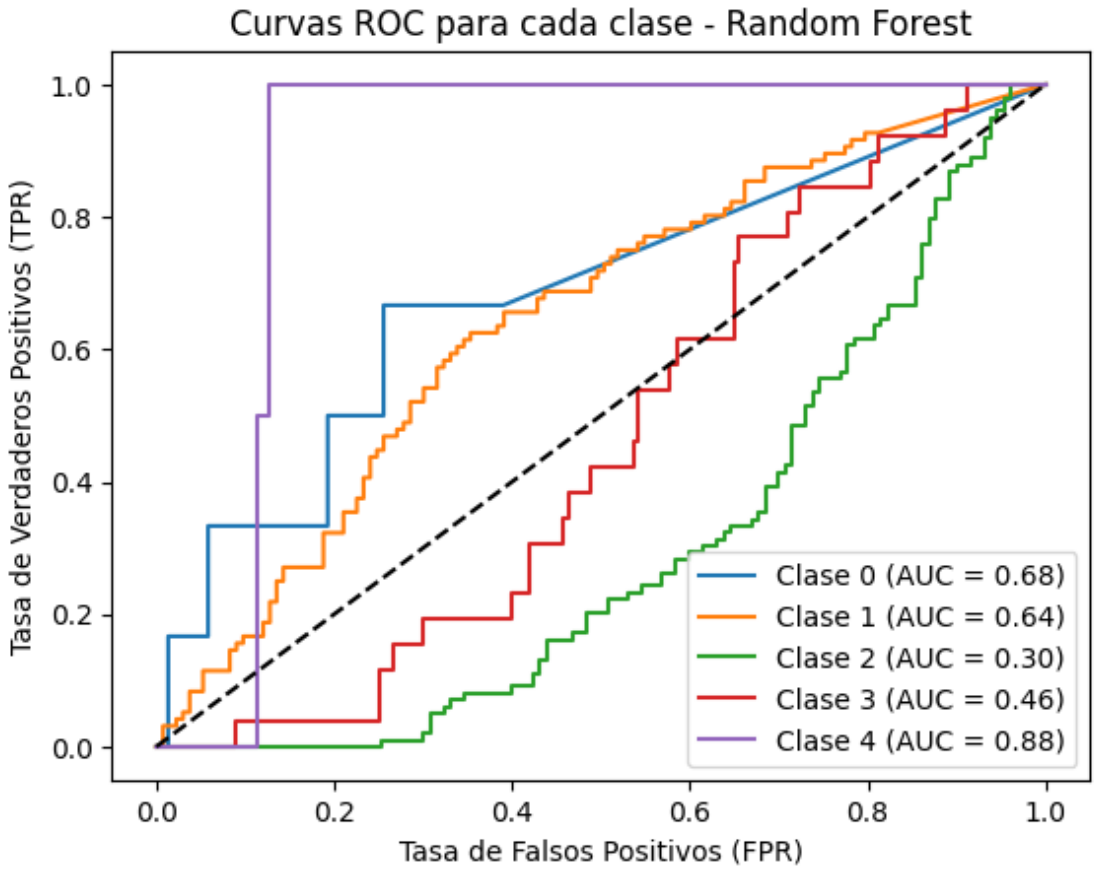
Se generaron las **curvas ROC** y se calcularon los valores de **AUC** para el mejor modelo, **Random Forest**. Estos resultados ofrecen una visión detallada del rendimiento del modelo en cada clase.

- **Resultados del AUC**:
  - Clase 0: AUC = 0.68
  - Clase 1: AUC = 0.64
  - Clase 2: AUC = 0.30
  - Clase 3: AUC = 0.46
  - Clase 4: AUC = 0.88

Interpretación:

- El modelo muestra un buen rendimiento en la Clase 4 (AUC=0.88), mientras que en las Clases 2 y 3 el AUC es bajo, lo que indica dificultad para clasificar estas clases correctamente.
- Las Clases 0 y 1 tienen un AUC moderado, lo que indica un rendimiento aceptable pero mejorable.

Gráfica de la Curva ROC:



6. CONCLUSIONES Y RECOMENDACIONES

Conclusiones:

- De los modelos probados, **Random Forest** presentó el mejor rendimiento en general, con una exactitud de 0.68 y buen AUC en algunas clases.
- El modelo de **KNN** y **Regresión Logística** fueron menos efectivos en comparación, especialmente en las clases más difíciles.