Informe de Análisis de Clasificación del Conjunto de Datos de Vinos

1. Objetivo del Análisis

El objetivo de este proyecto fue desarrollar un modelo de aprendizaje automático para clasificar muestras de vino en tres clases diferentes basadas en sus características químicas. Este modelo permite predecir la clase de un vino en función de sus atributos, facilitando la categorización rápida y precisa de nuevos datos.

2. Descripción del Conjunto de Datos

El conjunto de datos utilizado proviene del [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data) y contiene las siguientes características:

- Número de muestras: 178

- Características: 13 (Alcohol, Ácido málico, Cenizas, entre otras)

- Clases: Tres clases, ajustadas en este análisis a valores de 0, 1 y 2 para facilitar la clasificación.

3. Procedimiento

3.1. Carga y Preprocesamiento de los Datos:

- Se cargaron los datos en un `DataFrame` de Pandas y se asignaron nombres a las columnas.

- La columna `Clase` se ajustó para utilizar valores de 0 a 2 en vez de 1 a 3, lo cual es una práctica común en modelos de clasificación.

3.2. División de los Datos:

- El conjunto de datos se dividió en conjuntos de entrenamiento y prueba (por ejemplo, 70%-30%).

3.3. Normalización:

- Se aplicó un escalado a las características utilizando `StandardScaler` para mejorar el rendimiento de los algoritmos de clasificación y garantizar que todas las características contribuyan de manera equilibrada.

3.4. Entrenamiento y Selección de Modelos:

- Se probaron varios modelos de clasificación, incluyendo un modelo de Random Forest que se seleccionó como el más efectivo debido a su rendimiento en precisión.

4. Resultados

4.1. Evaluación de la Precisión

El modelo Random Forest alcanzó una precisión alta en la clasificación de las tres clases de vinos, como se muestra en la siguiente matriz de confusión:

```
KNN Reporte de Clasificación:
              precision    recall  f1-score   support

           0       0.93      1.00      0.97        14
           1       1.00      0.86      0.92        14
           2       0.89      1.00      0.94         8

    accuracy                           0.94        36
   macro avg       0.94      0.95      0.94        36
weighted avg       0.95      0.94      0.94        36

[[14  0  0]
 [ 1 12  1]
 [ 0  0  8]]
Random Forest Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        14
           1       1.00      1.00      1.00        14
           2       1.00      1.00      1.00         8

    accuracy                           1.00        36
   macro avg       1.00      1.00      1.00        36
weighted avg       1.00      1.00      1.00        36

[[14  0  0]
 [ 0 14  0]
 [ 0  0  8]]
SVM Reporte de Clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        14
           1       1.00      1.00      1.00        14
           2       1.00      1.00      1.00         8

    accuracy                           1.00        36
   macro avg       1.00      1.00      1.00        36
weighted avg       1.00      1.00      1.00        36

[[14  0  0]
 [ 0 14  0]
 [ 0  0  8]]
```

- Clase 0: 14 muestras correctamente clasificadas.

- Clase 1: 14 muestras correctamente clasificadas.

- Clase 2: 8 muestras correctamente clasificadas.

Esto refleja una excelente precisión en las clases 0 y 1, mientras que la clase 2 presenta una menor cantidad de muestras, lo cual puede influir en su menor precisión.

4.2. Predicción de Nuevas Muestras

Se aplicó el modelo a nuevas muestras de datos, obteniendo las siguientes predicciones:

- Predicción para la primera muestra: Clase 0.

- Predicción para la segunda muestra: Clase 1.

Este resultado muestra que el modelo puede clasificar de manera confiable nuevas muestras basadas en los valores de sus características.

5. Conclusiones

- Desempeño del Modelo: El modelo Random Forest ha demostrado ser eficaz para clasificar vinos en sus respectivas clases. La matriz de confusión muestra que las clases principales se identifican con precisión, aunque hay margen de mejora en la clase 2, posiblemente debido a un número limitado de muestras de esta clase.

- Aplicabilidad: El modelo entrenado permite clasificar muestras de vino con precisión y podría integrarse en un sistema automatizado para la clasificación rápida en una bodega o laboratorio enológico.

- Recomendaciones Futuras: Para mejorar el modelo, se recomienda ampliar la cantidad de datos, especialmente para la clase 2, o considerar el uso de técnicas de balanceo de clases.

6. Anexos

6.1. Código Utilizado

A continuación, se presenta el código principal utilizado para cargar los datos, preprocesarlos, entrenar el modelo y realizar las predicciones.

Código del análisis

```
# Importación de librerías
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier

# Carga del conjunto de datos y preprocesamiento
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"
columnas = ['Clase', 'Alcohol', 'Ácido málico', 'Cenizas', 'Alcalinidad de las cenizas',
        'Magnesio', 'Fenoles totales', 'Flavonoides', 'Fenoles no flavonoides',
        'Proantocianinas', 'Intensidad de color', 'Tono', 'OD280/OD315', 'Prolina']
vinos = pd.read_csv(url, header=None, names=columnas)
vinos['Clase'] = vinos['Clase'].replace({1: 0, 2: 1, 3: 2})

# División de los datos y escalado
X = vinos.drop('Clase', axis=1)
y = vinos['Clase']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Entrenamiento del modelo Random Forest
model = RandomForestClassifier(random_state=42)
model.fit(X_train_scaled, y_train)
```