Data Mining Week 2

March 24, 2024

1 Introduction:

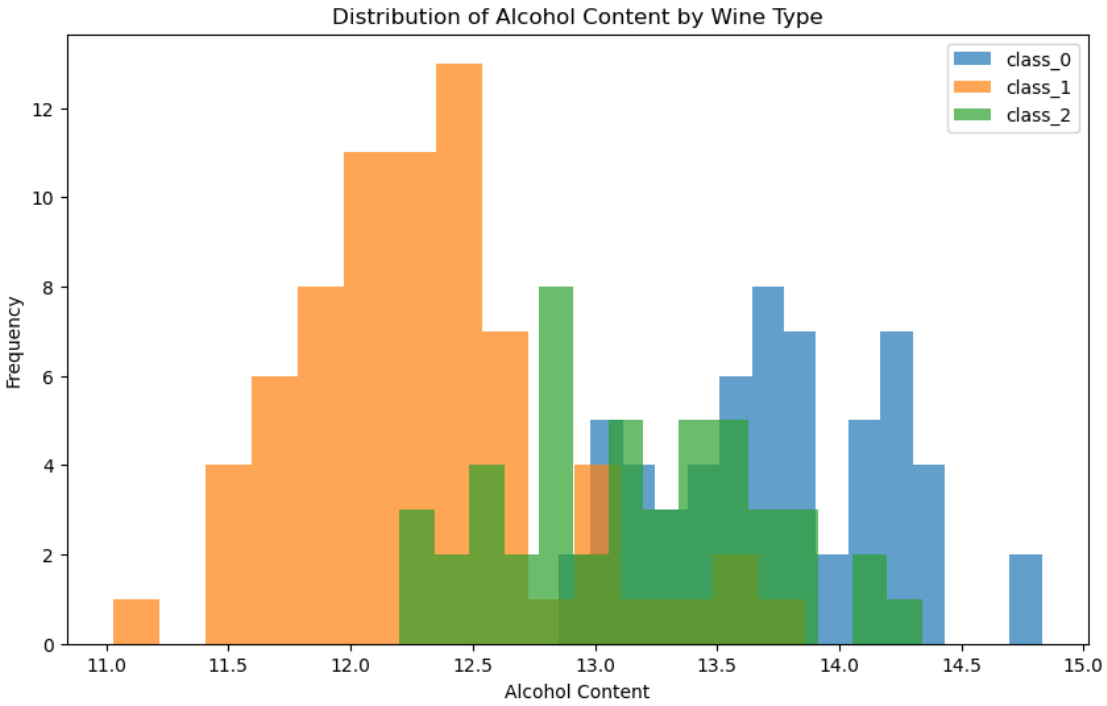
The wine dataset is a well-known dataset in the field of machine learning and is often used for classification tasks. It consists of 178 samples, each with 13 attributes such as alcohol content, color intensity, and flavonoid content. The samples are classified into three types of wine.

2 Questions to Explore

What is the distribution of alcohol content across the different types of wine? How does the average color intensity vary among the different types of wine?

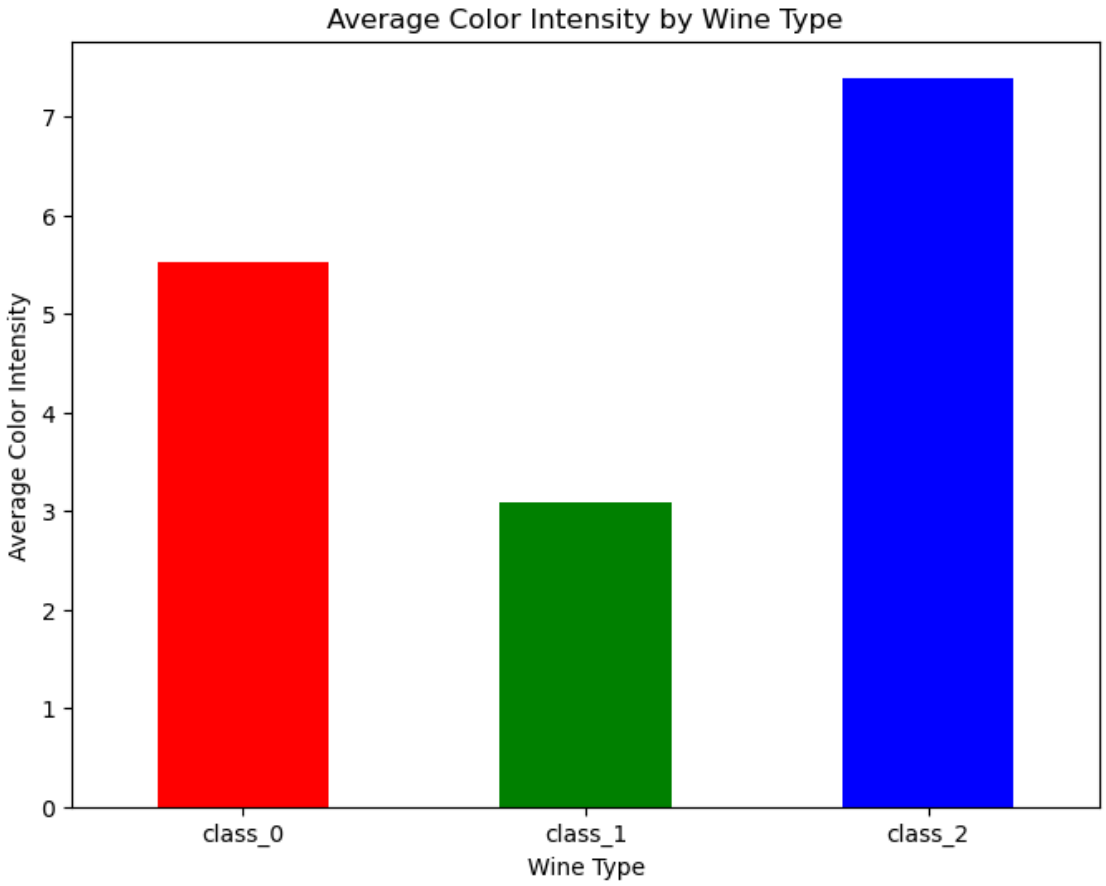
[2]:
```
from sklearn.datasets import load_wine
import pandas as pd

# Load the wine dataset
wine = load_wine()
wine_df = pd.DataFrame(wine.data, columns=wine.feature_names)
wine_df['wine_type'] = pd.Categorical.from_codes(wine.target, wine.target_names)
```

[9]:
```
# Graph 1: Histogram of alcohol content for each type of wine

import matplotlib.pyplot as plt

plt.figure(figsize=(10, 6))
for wine_type, group in wine_df.groupby('wine_type'):
    plt.hist(group['alcohol'], bins=15, label=wine_type, alpha=0.7)
plt.xlabel('Alcohol Content')
plt.ylabel('Frequency')
plt.title('Distribution of Alcohol Content by Wine Type')
plt.legend()
plt.show()
```



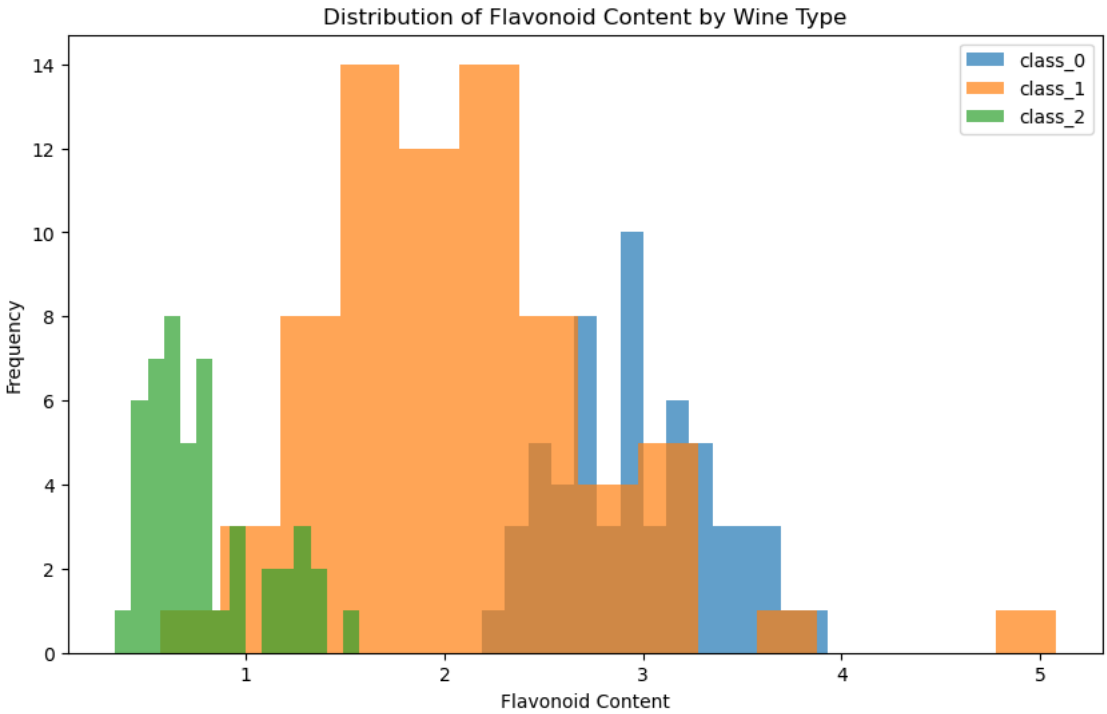
This histogram shows the distribution of alcohol content for each of the three types of wine. We can observe that different types of wine have different distributions of alcohol content, which could be a useful feature for classification.

[10]:
```
# Graph 2: Bar plot of average color intensity for each type of wine
plt.figure(figsize=(8, 6))
avg_color_intensity = wine_df.groupby('wine_type')['color_intensity'].mean()
avg_color_intensity.plot(kind='bar', color=['r', 'g', 'b'])
plt.xlabel('Wine Type')
plt.ylabel('Average Color Intensity')
plt.title('Average Color Intensity by Wine Type')
plt.xticks(rotation=0)
plt.show()
```



The bar plot shows the average color intensity for each type of wine. It appears that the average color intensity varies among the different types of wine, with one type having a noticeably higher average color intensity than the others.

[11]:
```
# Graph 3: Histogram of flavonoid content for each type of wine
plt.figure(figsize=(10, 6))
for wine_type, group in wine_df.groupby('wine_type'):
    plt.hist(group['flavanoids'], bins=15, label=wine_type, alpha=0.7)
plt.xlabel('Flavonoid Content')
plt.ylabel('Frequency')
plt.title('Distribution of Flavonoid Content by Wine Type')
plt.legend()
plt.show()
```



This histogram shows the distribution of flavonoid content for each of the three types of wine. Flavonoids are a type of polyphenol, and their content in wine can affect its taste and health benefits. From the plot, we can observe how the flavonoid content varies among the different types of wine, which could be another useful feature for classifying the wine samples.