winequalitydataethics

November 30, 2025

0.1 Wine Quality Ethics Check

Quick script to analyze the Wine Quality dataset for balance, missing values, and distribution skew

[13]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load dataset
df = pd.read_csv("/content/WineQT.csv", sep=",")

# Basic info
print("Dataset shape:", df.shape)
print("Columns:", df.columns.tolist())

# Check for missing values
missing = df.isnull().sum()
print("\nMissing Values Report:\n", missing[missing > 0])

# Distribution of quality ratings
quality_counts = df['quality'].value_counts(normalize=True) * 100
print("\nQuality Distribution (%):\n", quality_counts.sort_index())

# Flag imbalance in quality ratings
max_class = quality_counts.max()
min_class = quality_counts.min()
imbalance_ratio = round(max_class / min_class, 2) if min_class > 0 else None

print("\nEthics Report:")
if imbalance_ratio and imbalance_ratio > 3:
    print(f"- Quality ratings are imbalanced (ratio {imbalance_ratio}:1).")
else:
    print("- Quality ratings are reasonably balanced.")

# Check for extreme values in alcohol content
alcohol_stats = df['alcohol'].describe()
print("\nAlcohol Content Stats:\n", alcohol_stats)
```

```
if alcohol_stats['max'] > 15:
    print("- Outlier detected: unusually high alcohol content.")
else:
    print("- Alcohol values within expected range.")

# Visualization: Quality distribution
sns.countplot(x="quality", data=df, palette="viridis", hue="quality",␣
 ↪legend=False)
plt.title("Wine Quality Distribution")
plt.xlabel("Quality Score")
plt.ylabel("Count")
plt.tight_layout()
plt.savefig("wine_quality_distribution.png")
print("\nVisualization saved as wine_quality_distribution.png")

# Final conclusion
print("\nConclusion:")
print("After thorough analysis, the Wine Quality dataset shows no missing␣
 ↪values, "
      "a moderate imbalance in quality ratings, and alcohol values within␣
 ↪expected ranges. "
      "This makes it suitable for modeling tasks, but ethical considerations␣
 ↪around class imbalance "
      "should be addressed to ensure fair evaluation. The dataset provides a␣
 ↪strong foundation "
      "for exploring predictive models while keeping responsible data practices␣
 ↪in mind.")
```

```
Dataset shape: (1143, 13)
Columns: ['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar',
'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH',
'sulphates', 'alcohol', 'quality', 'Id']

Missing Values Report:
 Series([], dtype: int64)

Quality Distribution (%):
 quality
3     0.524934
4     2.887139
5    42.257218
6    40.419948
7    12.510936
8     1.399825
Name: proportion, dtype: float64

Ethics Report:
```

```
- Quality ratings are imbalanced (ratio 80.5:1).

Alcohol Content Stats:
 count    1143.000000
mean       10.442111
std         1.082196
min         8.400000
25%         9.500000
50%        10.200000
75%        11.100000
max        14.900000
Name: alcohol, dtype: float64
- Alcohol values within expected range.

Visualization saved as wine_quality_distribution.png

Conclusion:
After thorough analysis, the Wine Quality dataset shows no missing values, a
moderate imbalance in quality ratings, and alcohol values within expected
ranges. This makes it suitable for modeling tasks, but ethical considerations
around class imbalance should be addressed to ensure fair evaluation. The
dataset provides a strong foundation for exploring predictive models while
keeping responsible data practices in mind.
```