Classification of Wine Quality Using Naive Bayes and Decision Tree Models

Andreas Angelides

INM431 Machine Learning · City St George's, University of London · December 2025

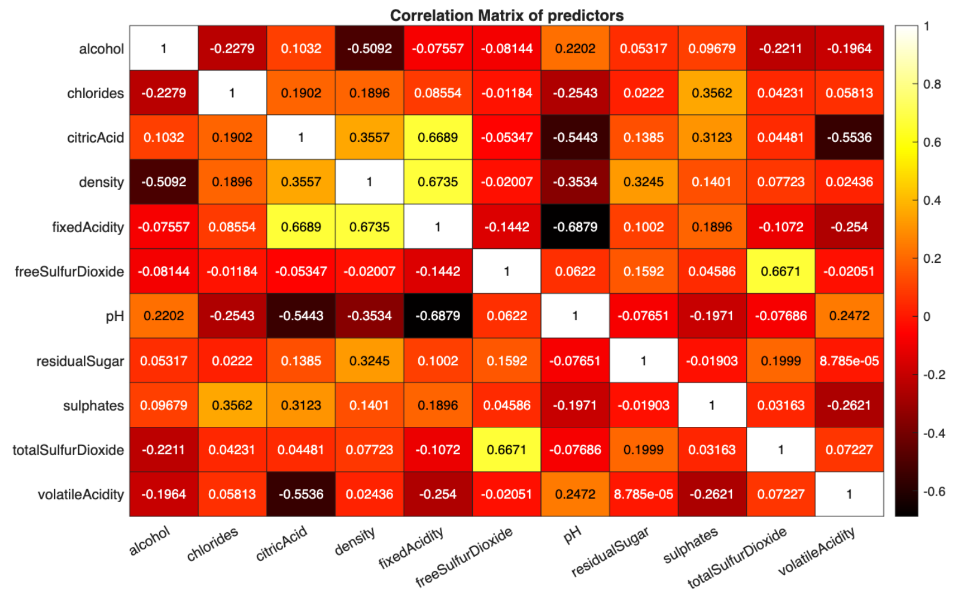
Description and motivation of the problem

The aim of this study is to solve the classification problem of red wine quality prediction, by constructing two supervised learning models: Naive Bayes and Decision Tree. Each wine in this dataset is scored by quality from 0 to 10 and for the purposes of this project we classify the wines into three classes: Low, Medium and High quality. These models are used very often in classification tasks, because of their simplicity and interpretability.**[1]**

Initial analysis of the dataset including basic statistics

The chosen dataset is the Wine Quality Dataset, available on Kaggle, introduced by Paulo Cortez and consists of two datasets related to the red and white wine of the Portuguese wine "Vinho Verde". For the purposes of the study, we are only interested to the red wine dataset.**[2][3]**The dataset consists of 1,599 samples and 12 variables, including eleven physicochemical features such as acidity, sugar, sulphates and alcohol content, along with a quality score assigned by wine experts. The target variable is converted into three classes: **Low (3-4)**, **Medium (5-6)** and **High (7-8)**. We use the range 3-8 for quality, because no values from 0-2 and 9-10 are shown.

A noticeable variation across features can be detected, after interpreting the statistics. The boxplots highlight the spread of each feature and indicate any outliers found. The correlation heatmap indicates that the alcohol content is positively correlated with the quality, while other features such as density and volatile acidity show a negative relationships with the quality. These observations suggest that the physicochemical features contain very important information regarding the prediction of wine quality and motivate the use of our supervised classification models.



Correlation Matrix of predictors

	alcohol	chlorides	citricAcid	density	fixedAcidity	freeSulfurDioxide	pH	residualSugar	sulphates	totalSulfurDioxide	volatileAcidity
alcohol	1	-0.2279	0.1032	-0.5092	-0.07557	-0.08144	0.2202	0.05317	0.09679	-0.2211	-0.1964
chlorides	-0.2279	1	0.1902	0.1896	0.08554	-0.01184	-0.2543	0.0222	0.3562	0.04231	0.05813
citricAcid	0.1032	0.1902	1	0.3557	0.6689	-0.05347	-0.5443	0.1385	0.3123	0.04481	-0.5536
density	-0.5092	0.1896	0.3557	1	0.6735	-0.02007	-0.3534	0.3245	0.1401	0.07723	0.02436
fixedAcidity	-0.07557	0.08554	0.6689	0.6735	1	-0.1442	-0.6879	0.1002	0.1896	-0.1072	-0.254
freeSulfurDioxide	-0.08144	-0.01184	-0.05347	-0.02007	-0.1442	1	0.0622	0.1592	0.04586	0.6671	-0.02051
pH	0.2202	-0.2543	-0.5443	-0.3534	-0.6879	0.0622	1	-0.07651	-0.1971	-0.07686	0.2472
residualSugar	0.05317	0.0222	0.1385	0.3245	0.1002	0.1592	-0.07651	1	-0.01903	0.1999	8.785e-05
sulphates	0.09679	0.3562	0.3123	0.1401	0.1896	0.04586	-0.1971	-0.01903	1	0.03163	-0.2621
totalSulfurDioxide	-0.2211	0.04231	0.04481	0.07723	-0.1072	0.6671	-0.07686	0.1999	0.03163	1	0.07227
volatileAcidity	-0.1964	0.05813	-0.5536	0.02436	-0.254	-0.02051	0.2472	8.785e-05	-0.2621	0.07227	1

Figure: Correlation heatmap of the physicochemical predictors.

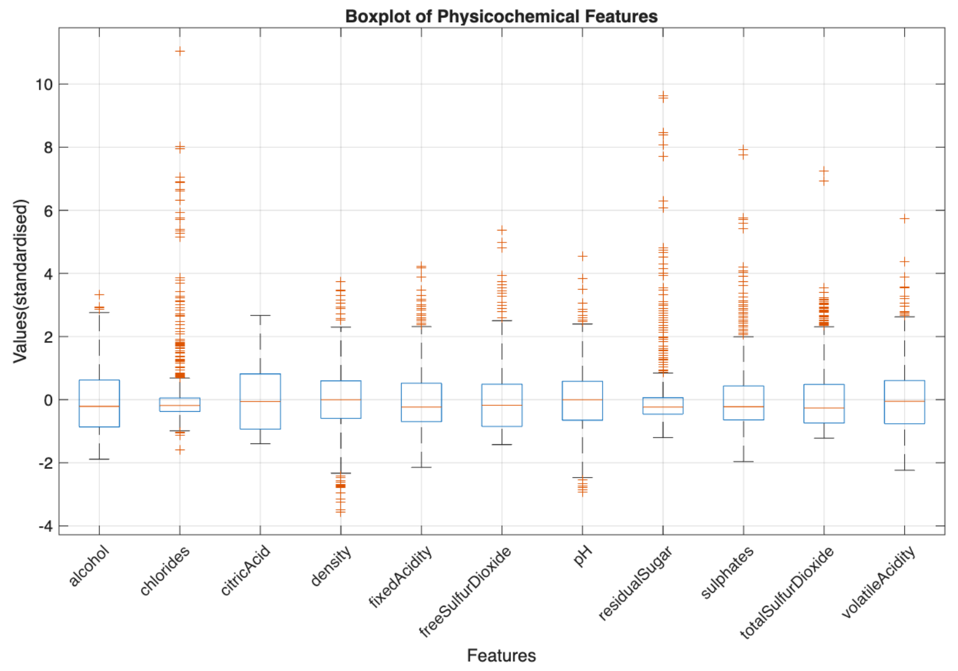


Figure: Boxplots of standardised physicochemical features (outliers visible).

Hypothesis statement

The project assumes that both of the trained models can classify the wine quality into the low, medium and high classes with a pretty accurate result. Moreover, it is expected both models to capture any interesting patterns across the physicochemical features of wine. Although, we use same data between the two models, we expect the Decision Tree to perform slightly better, due to its ability to handle non-linear relationships and interactions between features. The most appropriate data cleaning , data preprocessing and model validation are applied to make sure a fair comparison between the two models is made. The performance of the two trained models is evaluated using the precision, the recall and the F1-score.

Brief summary of Machine Learning Methods

Naive Bayes

Naive Bayes is a supervised probabilistic classification algorithm based on the Bayes' theorem. It assumes that all input features are conditionally independent given the class label. Moreover, Naive Bayes makes the assumption that each feature contributes equally to the result and that no feature variable is more significant than another.**[4]**

Advantages:

- Simple and computationally efficient, even for large datasets with a lot of features.
- Great performance even with limited training data.

Disadvantages:

- Performance can drop when predictors are correlated
- May be impacted by unrelated characteristics.

Decision Tree

Decision Tree is another supervised learning algorithm that can be used for classification and regression applications. As the name says, it displays the predictions that come from a sequence of feature-based splits using a flowchart that creates a tree structure. It begins with a root node and ends with a choice made by leaves.**[5]**

Advantages:

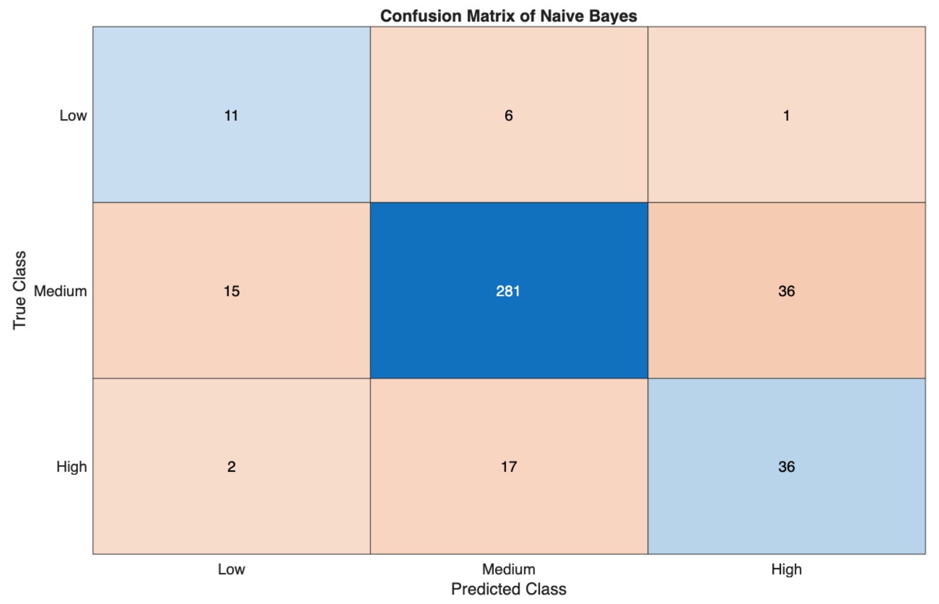
- It is very easy to interpret and visualize.
- It handles both numerical and categorical data.

Disadvantages:

- It can overfit without pruning.
- It is very unstable with small changes and can become very complex.

Description of the choice of training and evaluation methodology

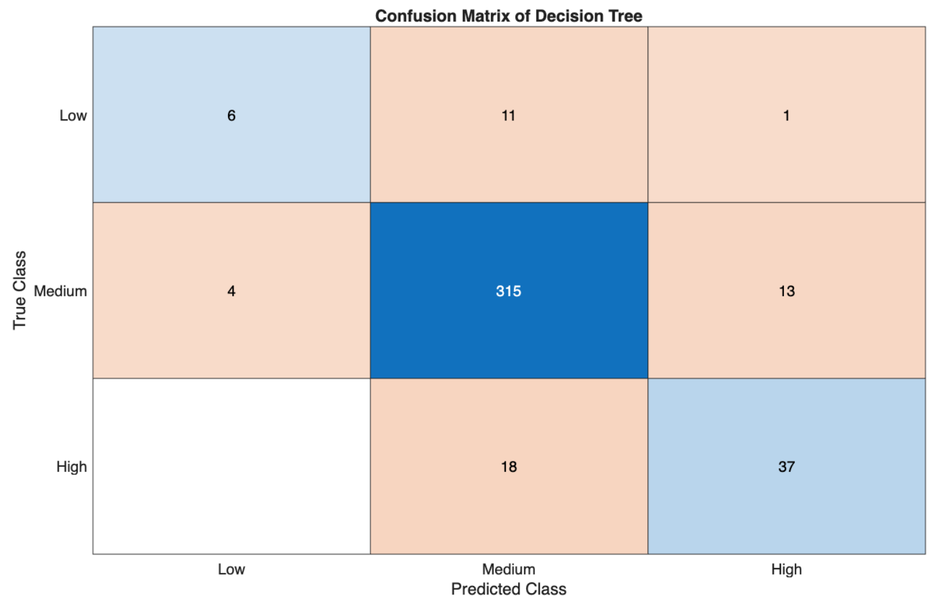
After cleaning the data and doing exploratory analysis (evaluated summary statistics, correlations, and boxplots), the target variable was grouped into the three classes. After that, outliers were detected using simple IQR rule for feature-level and Mahalanobis distance for instance-level, and the strongest suspects were flagged out. The cleaned dataset was then split into **70%** for training data and **30%** for testing data using the holdout method. For the training data, we used the 5-fold cross validation method to train and evaluate the Naive Bayes classifier and the Decision Tree classifier. The model performance was compared using confusion matrix and per-class precision, recall and F1-score and the weighted averages which help us check for class imbalance. The results were visualized using confusion matrices, weighted metrics performance plot and PCA plot to show how well the classes separate.



Confusion Matrix of Naive Bayes

True Class \ Predicted Class	Low	Medium	High
Low	11	6	1
Medium	15	281	36
High	2	17	36

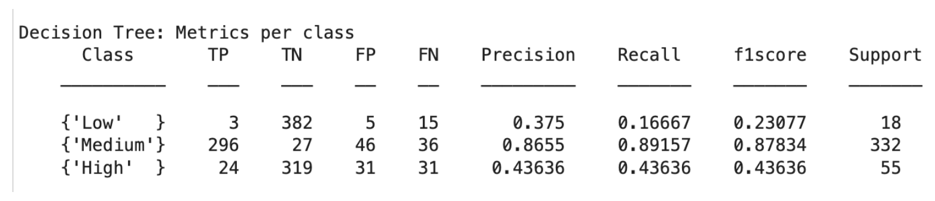
Naive Bayes Confusion Matrix



Confusion Matrix of Decision Tree

True Class \ Predicted Class	Low	Medium	High
Low	6	11	1
Medium	4	315	13
High		18	37

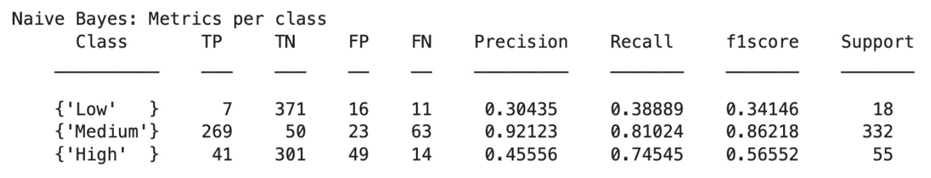
Decision Tree Confusion Matrix



```
Decision Tree: Metrics per class
      Class        TP     TN     FP    FN    Precision    Recall     f1score    Support
    __________    ___    ___    __    __    _________    _______    _______    _______

    {'Low'   }      3    382     5    15        0.375    0.16667    0.23077       18
    {'Medium'}    296     27    46    36       0.8655    0.89157    0.87834      332
    {'High'  }     24    319    31    31      0.43636    0.43636    0.43636       55
```

Decision Tree: per-class metrics table.



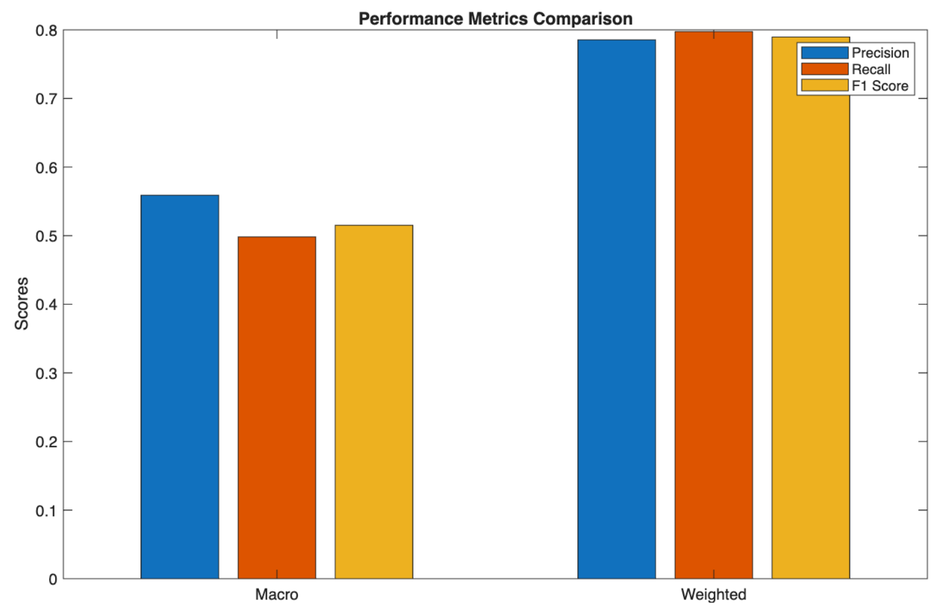
```
Naive Bayes: Metrics per class
      Class        TP     TN     FP    FN    Precision    Recall     f1score    Support
    __________    ___    ___    __    __    _________    _______    _______    _______

    {'Low'   }      7    371    16    11      0.30435    0.38889    0.34146       18
    {'Medium'}    269     50    23    63      0.92123    0.81024    0.86218      332
    {'High'  }     41    301    49    14      0.45556    0.74545    0.56552       55
```

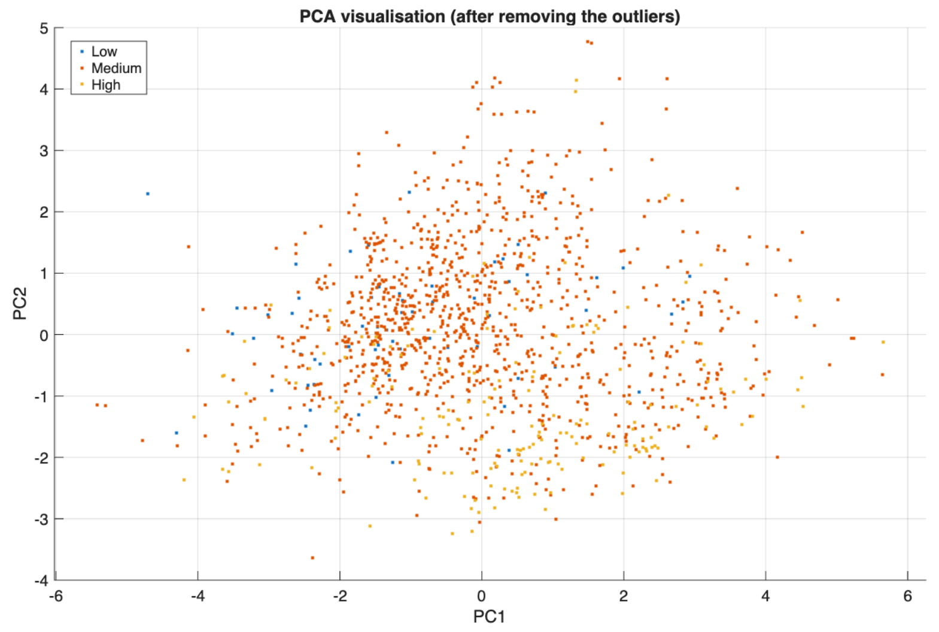
Naive Bayes: per-class metrics table.

Analysis and critical evaluation of results

The results show, that both of our trained supervised learning algorithms learned a lot of meaningful patterns from the physicochemical features of red wine. The task is challenging because of the imbalance between classes (the Medium class dominates). Based on cross-validation loss, the **Decision Tree** performed slightly better than **Naive Bayes**, which indicates better generalization during training. However, after completing the whole test performance using weighted metrics, the **Naive Bayes** achieved higher precision and F1-score, while the **Decision Tree** achieved slightly higher recall. Both models did best on the Medium class but made more mistakes on Low and High. This is expected because the classes are unbalanced and wines with close quality scores often look very similar in the features. The PCA plot also shows a lot of overlap between classes, so more errors are likely due to the data being hard to separate, not just the model choice. Overall, the Naive Bayes is a clear and simple baseline, while the Decision Tree can pick up more complex patterns.



Macro vs weighted precision/recall/F1.



PCA (after outlier removal).

Lessons learned, future work and references

This project showed that data pre-processing and evaluation choices can affect the performance of the models. Removing outliers and handling class imbalance helped both models to perform better and show much more stable results. Overall,the performance is similar. The Decision Tree generalized slightly better in cross-validation loss, while Naive Bayes has slightly higher weighted precision and F1-score on the test set. However, the overlap between wine quality classes suggests that to get the best results is difficult using only physicochemical features. Future work could explore different methods and models, such as the Random Forest, balance better the classes using different techniques, or include production-related features to improve classification performance.

References:

[1]Shin, T. (2020). Predicting Wine Quality with Several Classification Techniques — Towards Data Science. [online] Towards Data Science. Available at: https://towardsdatascience.com/predicting-wine-quality-with-several-classification-techniques-179038ea6434/ [Accessed 1 Dec. 2025].

[2]UCI Machine Learning. "Red Wine Quality." Www.kaggle.com, 2018, www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009. Accessed 15 Dec. 2025.

[3]Cortez, Paulo, et al. "Modeling Wine Preferences by Data Mining from Physicochemical Properties." Decision Support Systems, vol. 47, no. 4, Nov. 2009, pp. 547–553, https://doi.org/10.1016/j.dss.2009.05.016.

[4]GeeksforGeeks. "Naive Bayes Classifiers." GeeksforGeeks, 3 Mar. 2017, www.geeksforgeeks.org/machine-learning/naive-bayes-classifiers/. Accessed 15 Dec. 2025.

[5]Saini, Anshul. "Decision Tree Algorithm - a Complete Guide." Analytics Vidhya, 29 Aug. 2021, www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/. Accessed 15 Dec. 2025.