CPS844 Project Report

Wine Quality Classification

Student Name: Vrisag Patel

Course: CPS844 - Data Mining

Date: April 5, 2025

Table of Contents

1. Introduction

This project's objective is to use a number of supervised machine learning models to conduct a thorough data mining investigation of the Wine Quality dataset. Predicting the wine's quality score based on a range of physicochemical characteristics, including pH level, alcohol content, residual sugar, and citric acid concentration, is the aim of this multi-class classification job. This is a classification problem rather than a regression problem since the wine quality scores, which vary from 0 to 10, are handled as distinct groups. Red and white wine samples are two separate subsets of the dataset, each with approximately 1,000 instances and 11 numerical features. This makes it appropriate for testing a variety of classification algorithms, methods for preparing data, and feature selection methods.

Finding significant patterns, connections, and dependencies between input variables and wine quality is the project's goal from a data mining standpoint. This could lead to useful insights for the beverage industry. In automated quality control, product standardization, and even customer recommendation systems, precise wine quality classification is useful. Data comprehension, cleansing and preprocessing, exploratory data analysis (EDA), model construction, and model evaluation comprise the traditional data mining pipeline that is used for this investigation. We evaluate five popular classification algorithms: Support Vector Machine (SVM), Random Forest, Decision Tree, K-Nearest Neighbours

(KNN), and Logistic Regression. Both the complete feature set and a smaller set of features chosen by Recursive Feature Elimination (RFE) are used to train and assess each model.

Through this project, we hope to assess how feature selection affects model accuracy and generalization in addition to comparing the predictive performance of different classifiers. The project also aims to determine which characteristics have the greatest influence on wine quality and determine whether dimensionality reduction results in a model that is easier to understand and more computationally efficient. In this study, classification performance is systematically compared, the most informative characteristics are highlighted, and the trade-offs between model complexity, accuracy, and interpretability are discussed.

2. Dataset Description

Through this project, we hope to assess how feature selection affects model accuracy and generalization in addition to comparing the predictive performance of different classifiers. The project also aims to determine which

characteristics have the greatest influence on wine quality and determine whether dimensionality reduction results in a model that is easier to understand and more computationally efficient. In this study, classification performance is systematically compared, the most informative characteristics are highlighted, and the trade-offs between model complexity, accuracy, and interpretability are discussed.

3. Problem Statement

Predicting wine quality using physicochemical traits is the primary issue this study attempts to solve. This is presented as a challenge using supervised binary classification. The objective is to determine the quality of a wine sample by measuring its qualities.

This issue has real-world applications in a number of areas, including automating wine production quality monitoring, directing product development, and helping customers choose premium wines. Additionally, it provides a useful case study for assessing how well machine learning algorithms perform on actual classification tasks.

4. Exploratory Data Analysis (EDA)

An essential component in the data mining process is the exploratory data analysis (EDA) phase, which aids in identifying underlying patterns, spotting

anomalies, and comprehending the connections between variables. EDA was performed on the red and white wine datasets for this project. The disparity in how wine quality ratings were distributed was among the first things noticed. With few samples at the extremes (e.g., ratings of 3 or 8), the majority of wines had ratings of 5 or 6 on a scale of 0 to 10, suggesting a skewed class distribution that may affect model bias and classification performance. To find linear associations between characteristics, a heatmap and correlation matrix were employed. Notably, volatile acidity and wine quality were negatively connected, whereas alcohol content was strongly positively correlated, indicating that better-rated wines typically have lower acidity and higher alcohol content. The distribution and variability of each attribute were examined using visual aids such as boxplots and histograms. These showed that some variables have right-skewed distributions and contain outliers, like residual sugar and total sulphur dioxide. Furthermore, a strong correlation between certain features, such as free and total sulphur dioxide, was discovered; this could induce multicollinearity and impact the interpretability of the model. All things considered, EDA was crucial in directing the preprocessing procedures and providing guidance for feature selection and model-building tactics.

5. Data Preprocessing

Any data mining pipeline must include the preprocessing stage since it guarantees that learning algorithms can function properly and prepares the raw

data for the best model performance. Several preparation techniques were used on the wine quality dataset for this study. To make the modelling process easier, the goal variable, quality, which was initially a multiclass variable with a range of 0 to 10, was first converted into a binary classification task. A wine was classified as "good" (1) if its quality rating was 6 or higher, and "not good" (0) if it was rated lower than 6. This dichotomy is consistent with practical uses, where wine is frequently assessed as acceptable or unsuitable or pass/fail.

After that, a 70-30 ratio was used to divide the dataset into training and testing subsets. This division makes it possible to assess the generalizability of the model appropriately and guarantees that the outcomes of training are not overfit to the particular data. StandardScaler, which rescales features to have a mean of zero and a standard deviation of one, was then used to standardize all input features. Because distance-based algorithms like K-Nearest Neighbours (KNN) and gradient-based techniques like Support Vector Machines (SVM) are sensitive to variations in feature magnitudes, this phase is particularly important.

Explicit outlier elimination was not carried out, despite the fact that visual inspection during EDA showed the presence of outliers in characteristics like residual sugar and sulphur dioxide levels. In order to lessen their possible detrimental effects, models that are either naturally resistant to outliers (such as decision trees and random forests) or that incorporate regularization techniques

(such as logistic regression with penalty terms) were used. The foundation for creating dependable, equitable, and consistent prediction models was established by these preprocessing procedures taken together.

6. Model Selection and Evaluation

Five different machine learning algorithms each selected from a different family to guarantee diversity in modelling approaches were used to tackle the categorization challenge. Standard classification criteria, such as accuracy, precision, recall, and F1-score, were used to assess the models. The red and white wine datasets' performances were evaluated independently. This part describes the rationale for selecting each model, the outcomes, and an analysis of how each model behaves with the data.

Logistic Regression

Because of its ease of use, effectiveness, and interpretability, logistic regression was used as the baseline model. It is a linear classifier that uses the logistic function to model the relationship between the input features and the target class's probability. Given the unbalanced nature of the classes, Logistic Regression did rather well, despite its limited capacity to capture intricate, non-linear correlations in the data. It was useful for comprehending the impact of individual attributes since it provided rapid training and simple coefficient interpretation. However, its linear assumptions limited its performance.

- **Accuracy (Red Wine):** 0.6479
- **Accuracy (White Wine):** 0.6429

K-Nearest Neighbors (KNN)

KNN is a non-parametric, slow technique that uses the majority class of the k-nearest neighbours in the feature space to classify new examples. It is versatile yet computationally costly, particularly when dealing with huge datasets, because it does not assume anything about the distribution of the data. Following hyperparameter adjustment, k=5 was determined to be the ideal value. KNN was sensitive to feature scaling and outliers, even though it was better than logistic regression at capturing local structures in the data. It provided competitive accuracy despite its simplicity, but the curse of dimensionality caused it to struggle in high-dimensional spaces.

- **Accuracy (Red Wine):** 0.6479
- **Accuracy (White Wine):** 0.6330

Decision Tree

Using feature thresholds, decision trees provide a hierarchical and understandable method of making decisions. They can handle both numerical and categorical data without the need for standardization, and they can represent non-linear connections. Until the model achieved full purity—that is, until all of the leaves were pure or a minimal sample split was achieved—it was permitted

to grow. This resulted in overfitting, particularly with the white wine dataset, even though it enabled the model to properly capture the training data structure. Nevertheless, Decision Trees successfully emphasized feature interactions and offered interpretable decision rules.

- **Accuracy (Red Wine):** 0.6104
- **Accuracy (White Wine):** 0.5881

Random Forest

The predictions of several decision trees trained on bootstrapped subsets of the data are combined in Random Forest, an ensemble learning technique. By using majority vote, each tree influences the final choice, minimizing overfitting and enhancing generalization. Random Forest is helpful for feature selection because it also offers feature significance scores. For both the red and white wine datasets, Random Forest produced the best classification accuracy out of all the models that were tested. It was especially successful because of its intrinsic regularization through averaging, resistance to noise, and capacity to describe intricate interactions.

- **Accuracy (Red Wine):** 0.6915
- **Accuracy (White Wine):** 0.6826

Support Vector Machine (SVM)

SVM is a potent classification technique that divides classes by creating the ideal hyperplane in a high-dimensional environment. To deal with non-linear decision boundaries, the radial basis function (RBF) kernel was utilized. SVM performed well on the red wine dataset, but it performed a little worse on the white wine dataset, maybe as a result of the more complicated structure and higher dimensionality. Furthermore, SVMs require careful tuning due to their sensitivity to hyperparameter choices (such as C and gamma) and processing cost. However, SVM showed strong generalization capabilities, especially when used with standardized features.

- **Accuracy (Red Wine):** 0.6677
- **Accuracy (White Wine):** 0.6646

7. Feature Selection Using RFE

Recursive Feature Elimination (RFE) was used as a feature selection method to improve model performance and lessen overfitting. RFE narrows down to a subset of the most relevant attributes by iteratively eliminating the least significant features according to the estimator's feature importances or coefficient weights. In addition to cutting down on training time and model complexity, this technique frequently enhances generalization by getting rid of duplicate or unnecessary features.

In order to take into consideration both linear and non-linear viewpoints when assessing feature importance, RFE was used in this project using two distinct base estimators: Random Forest and Logistic Regression. While Random Forest offers a non-linear viewpoint by providing relevance scores resulting from decision tree splits, Logistic Regression assists in ranking features based on linear correlations.

Selected Features from RFE

After running RFE on the full set of features for each wine dataset, the most predictive attributes identified were:

- **Red Wine Dataset:**
    - Alcohol
    - Sulphates
    - Volatile Acidity
    - Citric Acid
    - Density
- **White Wine Dataset:**
    - Alcohol
    - Residual Sugar
    - Volatile Acidity
    - Sulphates

    - Total Sulfur Dioxide

These qualities are both domain-specific and intuitive. For example, in line with the previous correlation analysis, alcohol consistently showed up as the best predictor of wine quality for both red and white wines. Sulphates and volatile acidity, which affect flavour and preservation, were also frequently noted as significant factors.

Impact on Model Performance

Each dataset's top five features were determined, and then only these features were used to retrain all classification models. This made it possible to directly compare how well models trained on the entire feature set performed to those trained on a smaller, optimized set.

- **Logistic Regression:**

  The accuracy of the model increased slightly, especially when applied to the red wine dataset. This suggests that by removing less useful characteristics, the model was able to concentrate on the most important input variables, lowering noise and enhancing decision bounds.

- **Random Forest:**

  The fact that performance mostly stayed the same indicates that the model is resilient to the addition of new features. Random Forest's efficacy in high-dimensional spaces was confirmed by the fact that its

performance neither declined nor dramatically improved because it automatically handles irrelevant data through ensemble averaging.

- **Support Vector Machine (SVM):**

  Particularly on the white wine dataset, using the smaller feature set resulted in increased computational efficiency and marginally higher performance consistency. Because SVMs can be computationally demanding and sensitive to dimensionality, lowering the feature count improved generalization and training time.

RFE emphasized the most important chemical characteristics associated with wine quality and made the modelling procedure simpler. Feature selection proved especially helpful for interpretable models like Logistic Regression and computationally expensive models like SVM, even though not all models exhibited significant gains.

8. Model Comparison and Results

All five models were evaluated using two distinct configurations to determine the effect of feature selection on classification performance: one using the entire feature set and another using a smaller subset of features found via Recursive Feature Elimination (RFE). Accuracy was the primary evaluation statistic, while individual model evaluations also took precision, recall, and F1-score into account.

Performance Summary

Model	Red Wine Accuracy (Full)	White Wine Accuracy (Full)	Red Wine Accuracy (RFE)	White Wine Accuracy (RFE)
Logistic Regression	0.6479	0.6429	0.6563	0.6479
KNN	0.6479	0.6330	0.6458	0.6454
Decision Tree	0.6104	0.5881	0.6042	0.6126
Random Forest	0.6915	0.6826	0.6906	0.6826
SVM	0.6677	0.6646	0.6563	0.6479

Key Observations and Insights

- Marginal Improvements from Feature Selection:

  Feature selection resulted in slight accuracy gains for the majority of models, especially for Logistic Regression and Decision Tree on the white wine dataset. These results imply that removing characteristics that aren't significant can improve model generalization, particularly for simpler models that might overfit when exposed to high-dimensional data.

- Stability of Random Forest:

  Regardless of whether all characteristics or just RFE-selected ones were employed, the Random Forest classifier continuously produced the best results across the red and white wine datasets. This demonstrates how ensemble decision trees' built-in feature selection algorithms make it resilient to noisy or irrelevant features.

- KNN and SVM Behavior:

  The outcomes with and without feature selection for K-Nearest Neighbours were almost the same. This might be because KNN depends mostly on distance computations, and with the standardized feature scaling, lowering the number of dimensions has no effect. The somewhat lower accuracy of SVM after feature selection, however, would suggest that some of the features that were eliminated still had minimal but complementing predictive value. Nonetheless, a significant advantage of

SVM was its decreased computing cost, particularly when using the RBF kernel.

- Importance of Alcohol as a Feature:

  Across both wine varieties and all models, the attribute "alcohol" consistently showed up as the strongest predictive predictor. The previous EDA results, which demonstrated a high positive association between alcohol and wine quality, reinforce this. Its crucial function in assessing wine quality is further supported by its selection by RFE and notable contribution to model performance.

- Decision Trees Showed Volatility:

  There were larger variations in the Decision Tree model's accuracy with feature selection; it performed better on the white wine dataset and marginally worse on the red. Pruning the input features may have eliminated useful splits for the red wine set, but it may have prevented overfitting in the white wine instance.

For some models, feature selection using RFE was advantageous because it reduced the complexity of the input space and increased model efficiency without compromising accuracy. Simpler or more sensitive models like SVM and Logistic Regression profited from more focused input variables, but models like Random Forest were mainly unaffected. This emphasizes how crucial

model-specific preprocessing techniques are to realistic machine learning workflows.

9. Conclusion

In order to forecast wine quality based on its physicochemical properties, this project investigated the use of data mining techniques. Using a variety of machine learning models and a thorough pipeline of data exploration, preprocessing, model training, feature selection, and evaluation, the main goal was to categorize wine samples—both red and white—into quality categories.

Significant relationships and patterns were found through exploratory data analysis. For example, there was a high positive link between wine quality and alcohol concentration, but a negative correlation between variables like density and volatile acidity. This provided some preliminary information about which features might have the most influence. Both the preprocessing and modelling stages were greatly aided by these discoveries.

A variety of model families (linear, instance-based, tree-based, ensemble, and kernel techniques, respectively) were represented by the five classification algorithms that were put into practice: Support Vector Machine, K-Nearest Neighbours, Decision Tree, Random Forest, and Logistic Regression. Because of its ensemble method, which reduces overfitting and captures intricate feature interactions, the Random Forest classifier outperformed the others in terms of

accuracy on both the red and white wine datasets. Although they were more susceptible to feature scaling and data dispersion, KNN and logistic regression both demonstrated competitive performance.

Recursive Feature Elimination (RFE) was used to assist cut down on the number of input variables while preserving or marginally enhancing the performance of some models. This suggests that not all variables have an equal impact on prediction accuracy and that, without compromising predictive power, effective feature selection can expedite the modelling process, increase efficiency, and facilitate interpretation.

All things considered, the study demonstrates that data mining is an effective strategy for resolving practical classification issues in fields such as food science. It not only makes precise forecasts possible, but it also improves knowledge of the fundamental elements influencing results, in this case the quality of the wine. Alcohol's preponderance as a predictive characteristic highlights the importance of data-driven validation and supports existing domain knowledge.

For future work, the modeling framework can be extended in several promising directions:

- **Ensemble Stacking**: Combining multiple models could lead to improved generalization by leveraging their complementary strengths.

- **Deep Learning Approaches**: Neural networks could uncover non-linear patterns in the data, especially with larger datasets or additional sensory features.
- **Regression Modeling**: Instead of binarizing quality scores, predicting the exact numerical rating using regression could provide more granular insights.
- **Multi-Class Classification**: Exploring a multi-class setup instead of binary labels may more accurately reflect the real-world rating system of wine quality.

From EDA and preprocessing to model selection and evaluation, this study shows how careful use of data mining techniques may produce useful insights and effective predictive systems in real-world situations.

10. References

[1] UCI Machine Learning Repository. *Wine Quality Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Wine+Quality