Red Wine Quality

Yinying Huo
huo1@ualberta.ca

ABSTRACT

This project aims to predict the quality of red wine using machine learning algorithms such as decision trees, logistic regression, and support vector machines. By analyzing chemical properties like fixed acidity, volatile acidity, citric acid, residual sugar, pH, and sulphates, the model classifies wines as "good" or "bad," simplifying the evaluation process for consumers.

KEYWORDS

Red wine quality prediction, machine learning, decision trees, logistic regression, support vector machines, chemical properties, classification.

1 INTRODUCTION

The evaluation of red wine quality is a complex and nuanced task, influenced by chemical properties to each varietal. For consumers discerning between a high-quality wine and a mediocre one can be challenging. This difficulty arises from the need to comprehensively consider numerous factors such as fixed acidity, volatile acidity, citric acid, residual sugar, pH, and sulphates, among others.

This project endeavors to harness the power of machine learning to simplify the evaluation process. By employing algorithms such as decision trees, logistic regression, and support vector machines (both linear and non-linear), the aim is to predict the quality of red wine within a binary classification framework: "good" or "bad."

The motivation behind this project derives from the aspiration to empower consumers with a tool that helps the assessment of red wine quality. By providing a straightforward classification based on objective chemical properties, individuals can make more informed purchasing decisions, even without extensive knowledge of wine tasting or production techniques.

2 PROBLEM SETUP

The objective of this project is to predict the quality of red wine based on a set of chemical properties, framing it as a binary classification task where each wine is categorized as either "good" or "bad." The input features include fixed acidity, volatile acidity, citric acid, residual sugar, pH, sulphates, and others.

2.1 Datasets

The dataset used in this project is sourced from the study titled "Modeling wine preferences by data mining from physicochemical properties" by P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, published in Decision Support Systems, Elsevier, in 2009[1]. The dataset contains observations of various red wines, along with their corresponding physicochemical properties and original quality ratings ranging from 1 to 10.

To redefine the problem as a binary classification task, the quality column will be transformed. Red wines with a quality rating greater than or equal to 6 will be labeled as "good" (assigned a value of 1), while those with a rating below 6 will be labeled as "bad" (assigned a value of 0). This modification enables the framing of the problem in terms of predicting whether a red wine is of high or low quality based on its chemical properties.

2.2 Samples

Each sample in the dataset represents an individual red wine and consists of features such as fixed acidity, volatile acidity, citric acid, residual sugar, pH, sulphates, and the modified quality label. The quality label serves as the target variable for training and evaluating the machine learning models. Samples will be randomly partitioned into training and testing sets to prevent biases in the model development process.

3 METHODOLOGY

For model training, the features are derived from the chemical properties of red wines, with the quality column serving as the target variable.

For experimentation, four machine learning algorithms were chosen: Decision Tree Classifier, Logistic Regression, Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel, and Support Vector Machine with Linear kernel. Each algorithm underwent training using the training data and was subsequently evaluated for accuracy using the testing data.

Baselines were established by assessing the accuracy of each algorithm using cross-validation. This method provided a robust estimate of each algorithm's performance. During cross-validation, the dataset was divided into five folds, with each fold used once as a validation set while the rest were utilized for training. This process was repeated for each algorithm, and the accuracy was calculated as the mean across all folds.

The first algorithm I chose is decision tree because it is the first classification algorithm introduced in CMPUT466, making it a good choice for this project. I use entropy as the criterion, as covered in our lectures. To avoid the risk of overfitting, I set the maximum depth to 3.

The second algorithm is logistic regression, a common choice for binary classification problems.

For the third algorithm, I selected Support Vector Machine (SVM). Given its similarity to logistic regression, I am interested in comparing their performance. SVM aims to maximize the margin between support vectors for class differentiation, while logistic regression aims to maximize the likelihood function.

The decision to utilize a nonlinear SVM stems from the observation that the data, as visualized in Figure 1, does not exhibit clear linear separability between features. This suggests that a linear decision boundary may not be sufficient to accurately classify the data points. In contrast, a nonlinear SVM with a radial basis function (RBF) kernel offers the flexibility to capture complex relationships between features and the target variable.

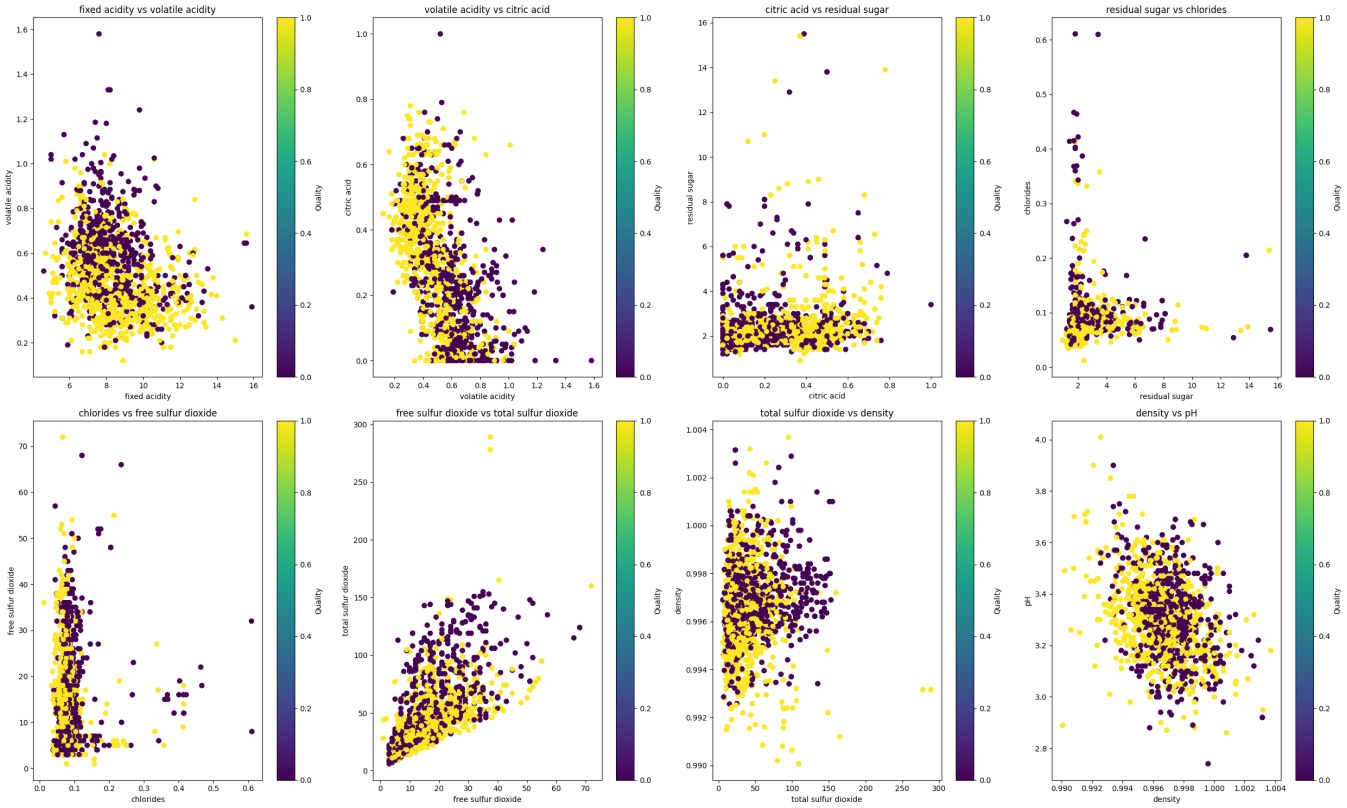

Figure 1: Graph depicting eight combinations of the features in the dataset. Yellow indicates 'good' quantity, while blue represents 'bad' quantity.

Table 1: Accuracy of ML algorithms

Algorithm name	Accuracy
Decision Tree	0.713577
Logistic Regression	0.732345
SVM(RBF)	0.634739
SVM(Linear)	0.732339

Thus, I decided to use both linear SVM and non-linear SVM (RBF kernel).

4 EVALUATION

The chosen metric for evaluating the success of the approach was accuracy, selected for its clarity and intuitiveness in indicating the overall correctness of the models.

5 RESULTS

Table 1 summarize the accuracy of different machine learning algorithm using cross validation.

The linear SVM achieved the highest accuracy among the models. I was surprised to find that its accuracy surpassed non-linear SVM; prior to running the code, I expected the non-linear SVM to perform at least as well as the linear SVM. One possible reason is that I only consider the features and target in 2D cases. Although they appeared non-linearly separable in 2D space, they might be linearly separable in higher-dimensional space, leading to the high accuracy of the linear SVM. Consequently, non-linear kernels will overfit more, resulting in lower accuracy for the non-linear SVM.

The performance of linear SVM is similar to logistic regression. Therefore, the choice between these two models can be based on specific requirements. For example, if we want to understand the confidence level of the model's decisions, we should choose logistic regression because its output is a probability, which can be interpreted as the model's confidence in its prediction (more explainable). SVM, on the other hand, provides the flexibility to choose between non-linear and linear kernels, rendering it adaptable to various scenarios. Additionally, SVM may accelerate training through its dual form. However we sacrifice some explainability.

Decision trees excel in explainability, offering intuitive insights into why a particular red wine is deemed good. Among the three algorithms, decision trees demonstrate the best explainability while maintaining acceptable accuracy.

6 LIMITATIONS

Firstly, the quality of red wine is influenced by various factors beyond the chemical properties included in the dataset. Factors such as aging, storage conditions, and production techniques may

also play significant roles but were not considered in this analysis. Consequently, the predictive performance of the models may be limited by the absence of these factors.

Secondly, the choice of features and the transformation of the quality column into a binary variable may have introduced biases or oversimplifications. For example, the threshold used to classify wines as "good" or "bad" (quality rating ≥ 6) may not align with subjective perceptions of wine quality.

Lastly, the evaluation metric used in this study, accuracy, provides a straightforward measure of model performance but may not capture the full spectrum of model behavior, particularly in imbalanced datasets or when the costs of false positives and false negatives are unequal.

7 ETHICS ANALYSIS

- Dual Use: Red wine producer could exploit vulnerabilities in the models to manipulate wine quality assessments or deceive consumers.
- Exclusion/Underexposure: If historical data is biased towards certain demographics or regions, the model's predictions may reflect and even amplify these biases. It's crucial to evaluate and mitigate bias in the dataset and ensure fairness in model predictions across different demographic groups.

REFERENCES

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 4 (2009), 547–553. https://doi.org/10.1016/j.dss.2009.05.016 Smart Business Networks: Concepts and Empirical Evidence.