Modeling Wine Preferences by Data Mining from Physicochemical Properties

Team 17

30 April, 2025

Ananya Jain (22B0301)
Pratyush Ranjan (22B0326)
Yash Kirdak (22B0330)
Vipul Muskan (22B0308)
Mahesh Kumar(22B0433)

Overview

01 Project description

02 Dataset desciption

03 Related works

04 Approach

05 Results

Project description

To predict the quality of white and red wine based on physicochemical properties using various regression models and compare their performance in terms of prediction accuracy, robustness, and interpretability.

- Wine quality is typically judged by expert tasters — a subjective and resource-intensive process.
- Physicochemical lab tests (e.g., acidity, sugar content, alcohol) offer objective, measurable indicators of wine quality.
- Machine learning allows us to learn predictive patterns from this data to assist wine certification and production.

Problem Statement

Given dataset has 11 features:

- Fixed acidity
- Volatile acidity
- Citric acidity
- Residual sugar
- Chlorides
- Free sulphur dioxide
- Total sulphur dioxide
- Density
- pH
- Sulphates
- Alcohol

We have applied machine learning models for identifying the quality of wine

We have used Support Vector Machine, Random Forest and Neural network and Multiple Regression and compared their performance to find the most optimal classification model

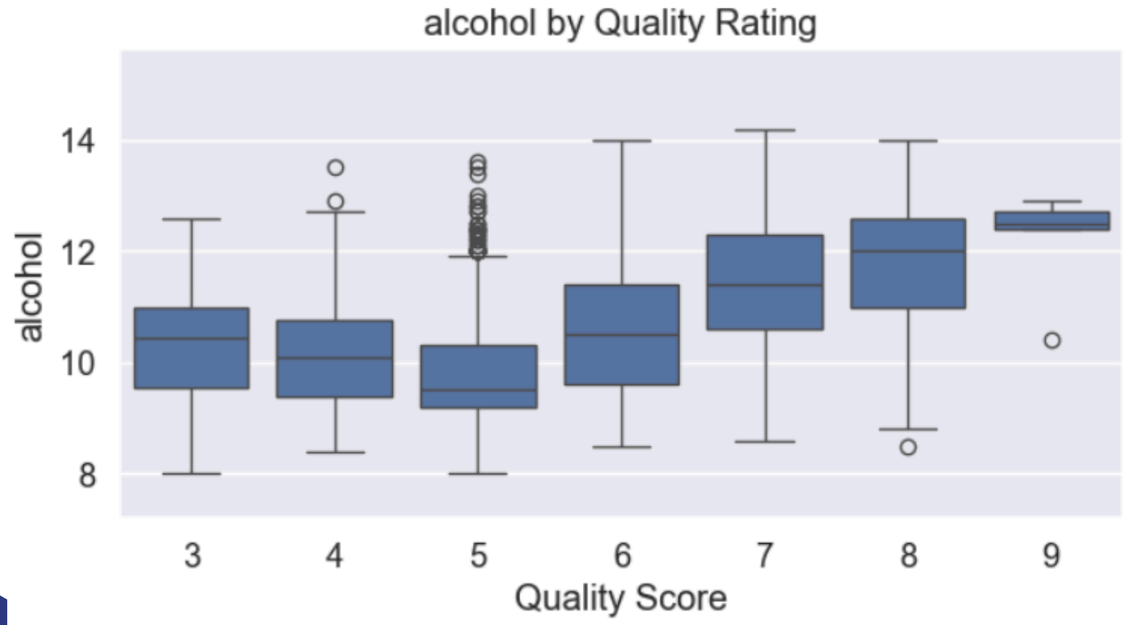
Performance metrics obtained for each model is later described in these presentation

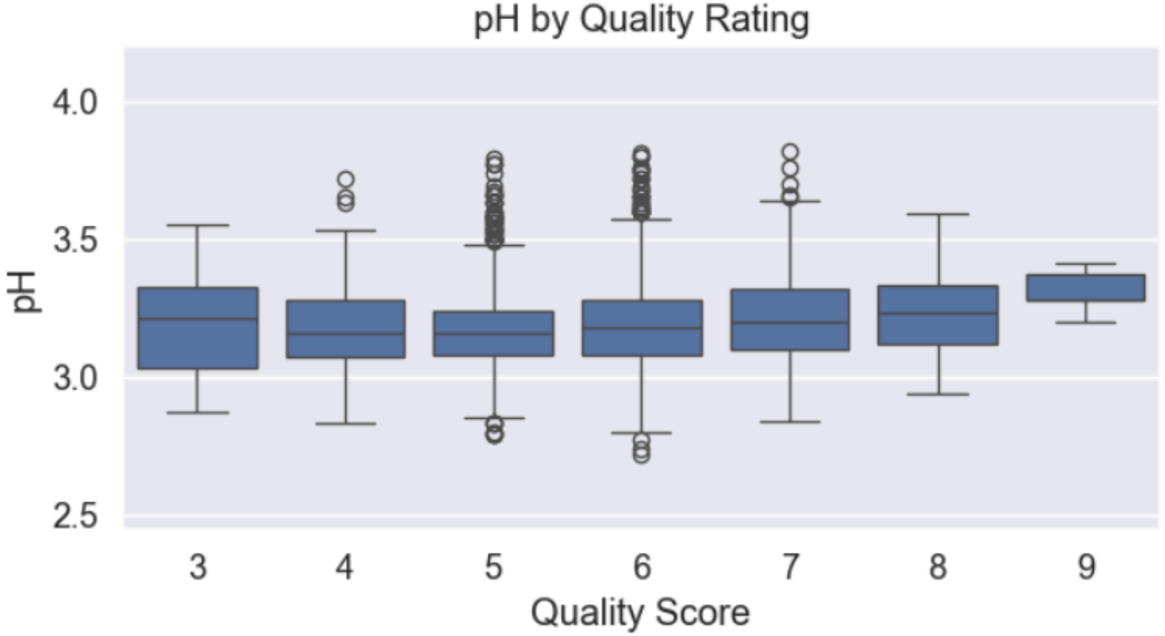
Dataset Description

Origin: The data comes from the Vinho Verde wine region of Portugal, a well-known wine-producing region.

Provided by: The dataset was collected in collaboration with the Viticulture Commission of the Vinho Verde Region (CVRVV).

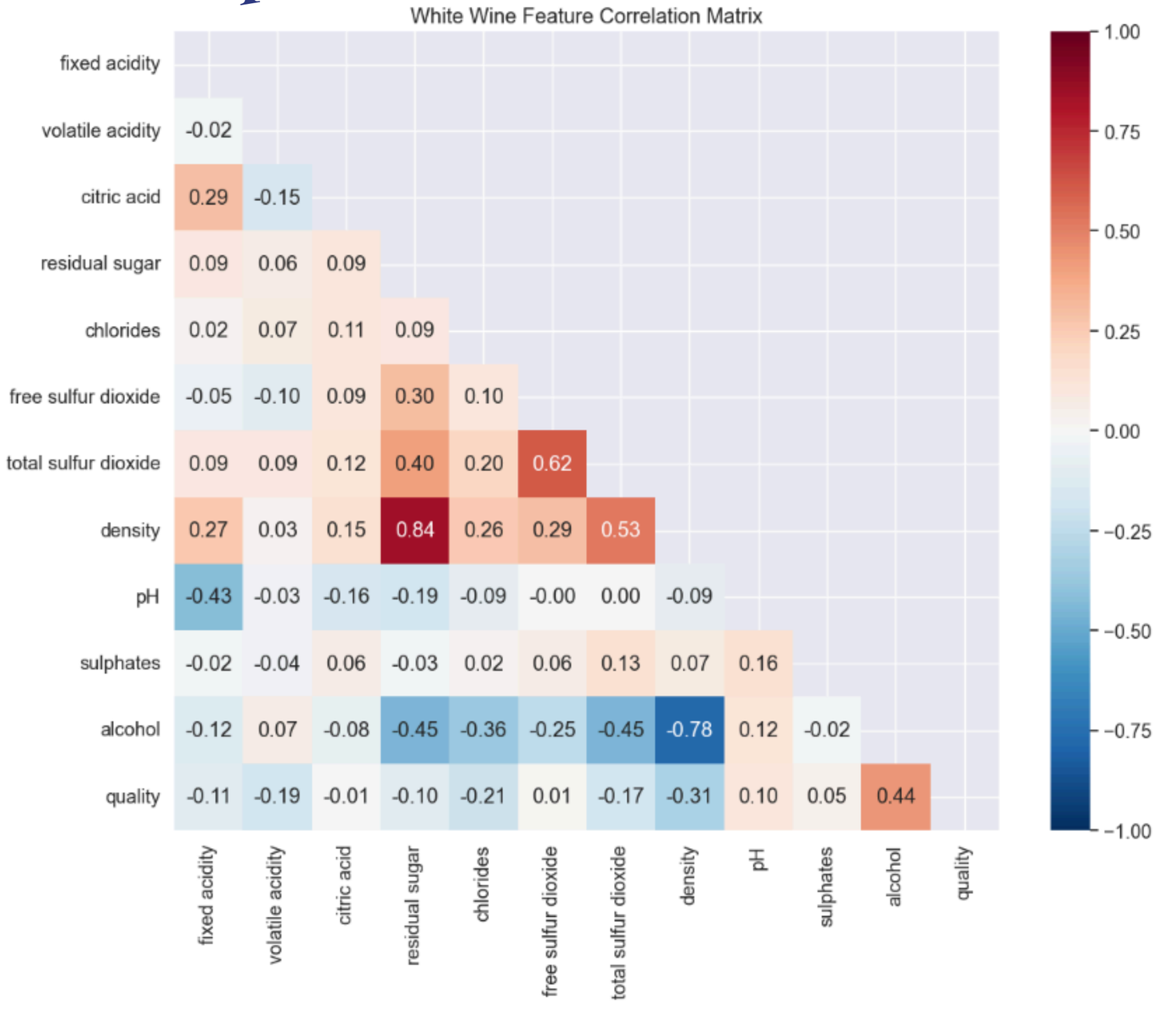
Samples: The dataset includes physicochemical tests on wine samples, using analytical chemical techniques. There are two datasets: winequality-red.csv with 1599 red wine samples and winequality-white.csv with 4898 white wine samples





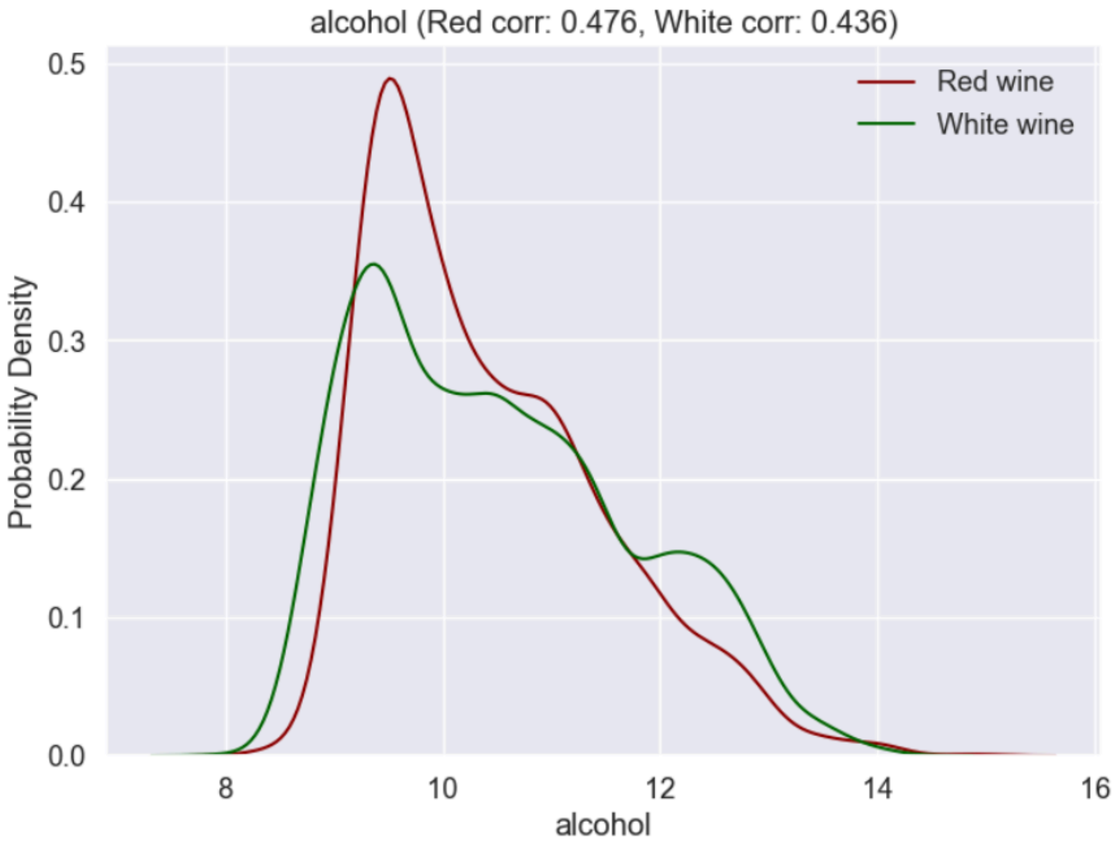
Dataset Description

- Quality Score (Target Variable):
- Each wine was blindly rated by sensory assessors (wine experts).
- Scores range from 0 to 10, based on median of at least three evaluations.
- This score is used as the label or output for modeling.



Dataset Description

Quality Ratings: Both datasets rated on scale of 3-9, with most wines clustering in the 5-6 range



Key Differences Between Wine Types

1. **Chemical Composition**: White wines have higher sugar, free and total sulfur dioxide; red wines have higher fixed acidity, volatile acidity, and sulphates
2. **Quality Drivers**:
    - Red wines: Quality most influenced by alcohol (+), volatile acidity (-), and sulphates (+)
    - White wines: Quality most influenced by alcohol (+), density (-), and chlorides (-)
3. **Feature Variability**: White wines show greater variation in residual sugar and sulfur dioxide; red wines show more variation in acid composition

Related works

Dataset Name	Description & Features	Notable Size/Stats
Spanish Wine Quality Dataset	7,500 Spanish red wine samples, includes price, rating, flavor	7,500 instances
UCI "Wine" Dataset	Italian wines from 3 cultivars, chemical analysis, 13 features	178 instances
Wine Reviews (Kaggle/Wine Enthusiast)	130,000+ expert reviews, text descriptions, points, price, country, variety, etc.	130,000+ reviews
WineSensed (Vivino)	824,000+ reviews, 897,000+ label images, ratings, price, region, grape composition	824,000+ reviews, 897,000+ images

- Most studies focus on the red wine dataset, despite its smaller size, due to higher consumption trends.
- Researchers use models like Random Forest, k-Nearest Neighbor, SVM, Decision Trees, Neural Networks, and ensembles for wine quality classification.
- Random Forest achieved the best accuracy (up to 99.5%), especially with feature selection.
- Machine learning helps automate product quality certification, saving time over manual assessment.
- EDA is widely used to address data imbalance, since medium-quality wines dominate both datasets.

Related Works

Achievements:

- Random Forest consistently outperforms other classifiers (accuracy up to 99.5% with proper feature selection)
- ANN achieved up to 88.28% accuracy on white wine and 85.16% on red wine datasets
- Feature selection techniques significantly improved model performance
- Oversampling methods like SMOTE enhanced results for imbalanced data
- Key features identified: alcohol content, acidity, and sulphates

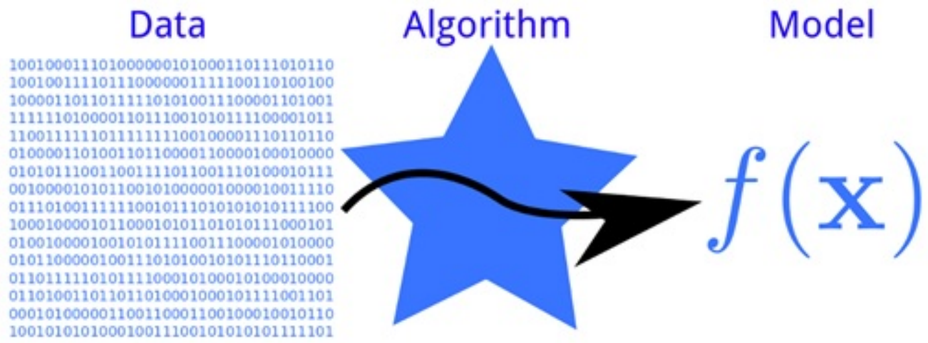
Limitations:

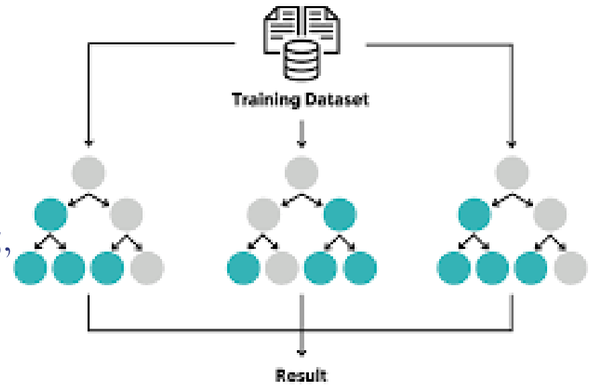
- Most studies focused on either red OR white wine, limiting generalizability
- Data imbalance issues (more medium-quality wines than high/low quality ones)
- Some approaches lacked advanced feature selection methods
- Naive Bayes consistently underperformed (around 46% accuracy)

Related Works

Are We Reproducing or Improving?

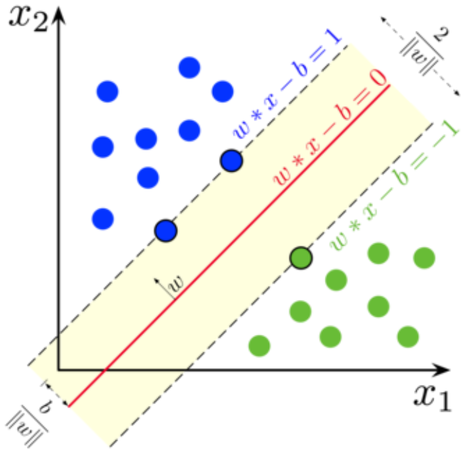
- In our project, we reproduce previous approaches by applying established machine learning models to the wine quality dataset.
- We aim to improve upon past work by:
  - Applied Regression method
  - we performed rigorous hyperparameter tuning using Grid Search with cross-validation for every model (SVM, NN, and RF), leading to better generalization and more reliable results
  - Systematically comparing different models to identify which achieves the best accuracy for wine quality prediction..





Random forest

Vs



SVM

References:

1. Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305–312. https://doi.org/10.1016/j.procs.2017.12.041
2. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016
3. Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 11(02), 278–289. https://doi.org/10.4236/ojs.2021.112015
4. Kaggle platform https://www.kaggle.com/datasets

Approach

Step 1: Data

Collected dataset from Kaggle

Step 2: Approach

Applied the regression models used in the reference paper, i.e. Multiple Linear Regression, Neural Networks, Support Vector Machine as well as tried a new method - Random Forest.

Step 3: Multiple Linear Regression

We fit a linear model to the data and extracted feature coefficients to interpret influence of each feature.

Step 4: (Soft) Support Vector Machine

We used non-linear RBF kernel for better flexibility. Epsilon was manually adjusted to 1, 0.5, and 0.1 to observe effect on tolerance and MAE .GridSearchCV was used to optimize: C, epsilon, gamma. Custom tolerance-based accuracy scorer was used as the GridSearch objective. Selected best model based on accuracy within ±1.

Approach

Step 5: Neural Network

Used MLP Regressor with standard architecture as a baseline. Implemented GridSearchCV wrt hidden layer sizes, activation, alpha, learning rate. Again, optimized using custom tolerance accuracy scorer (±1).

Step 6: Random Forest

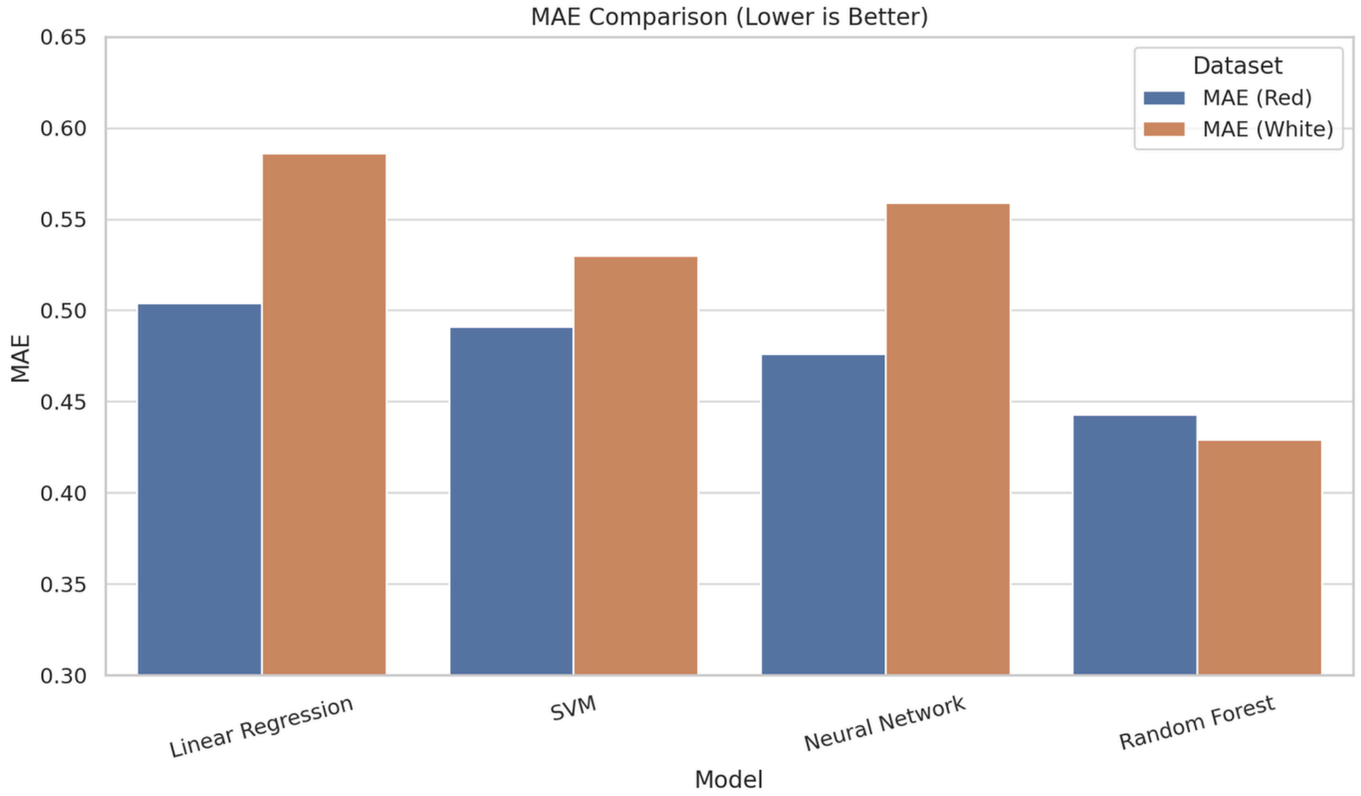
Used Random Forest Regressor as a tree-based model. GridSearchCV used to optimize: n estimators, max depth, min samples split. It was optimized with the same custom accuracy scorer (±1).

Also extracted feature importance to interpret influential variables.

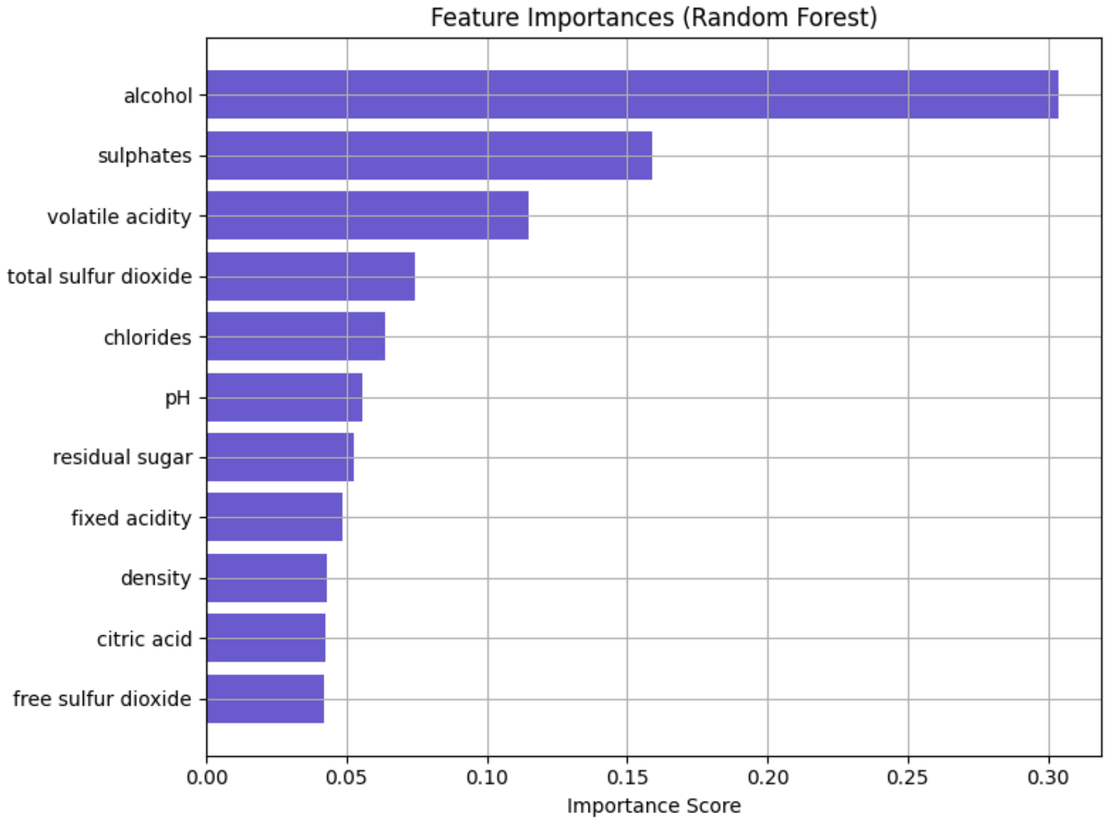
Step 7: Evaluation Metrics Used

MAE, Accuracy +/- 0.5 and Accuracy +/- 1.0

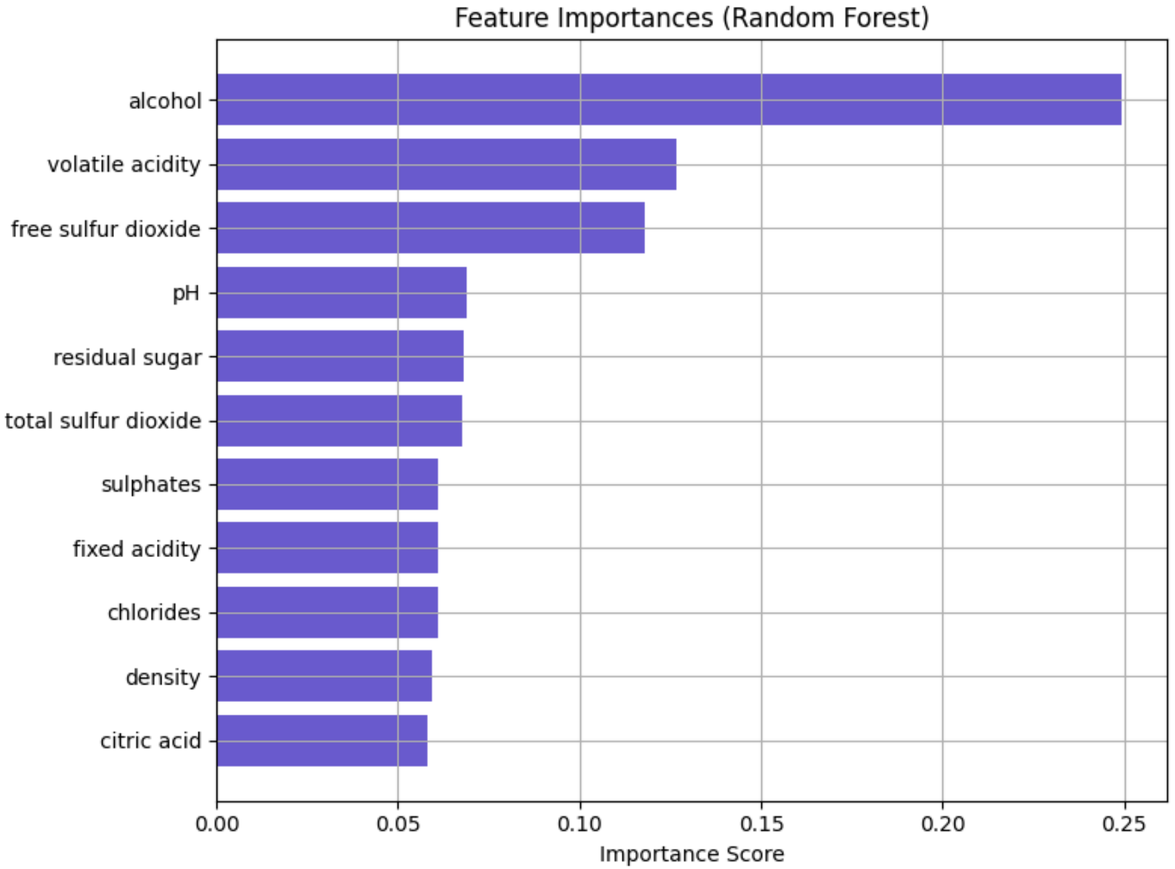
Results


MAE Comparison (Lower is Better)
MAE
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
Dataset
MAE (Red)
MAE (White)
Linear Regression
SVM
Neural Network
Random Forest
Model

Results

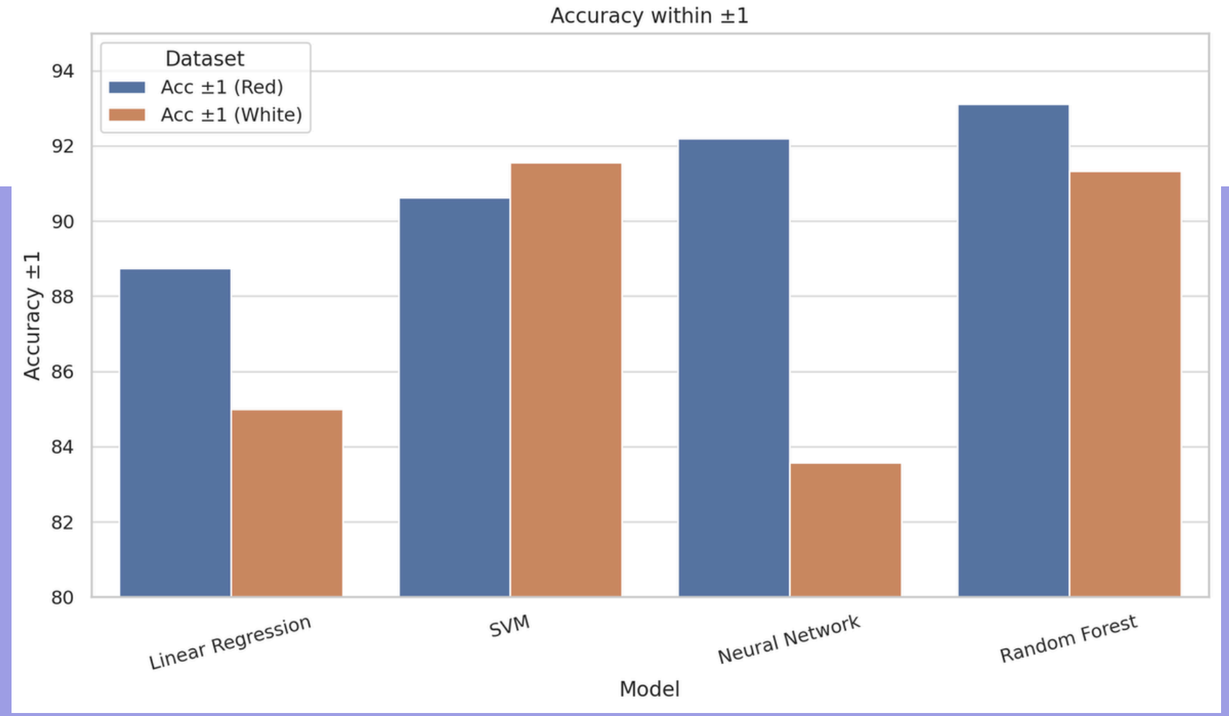


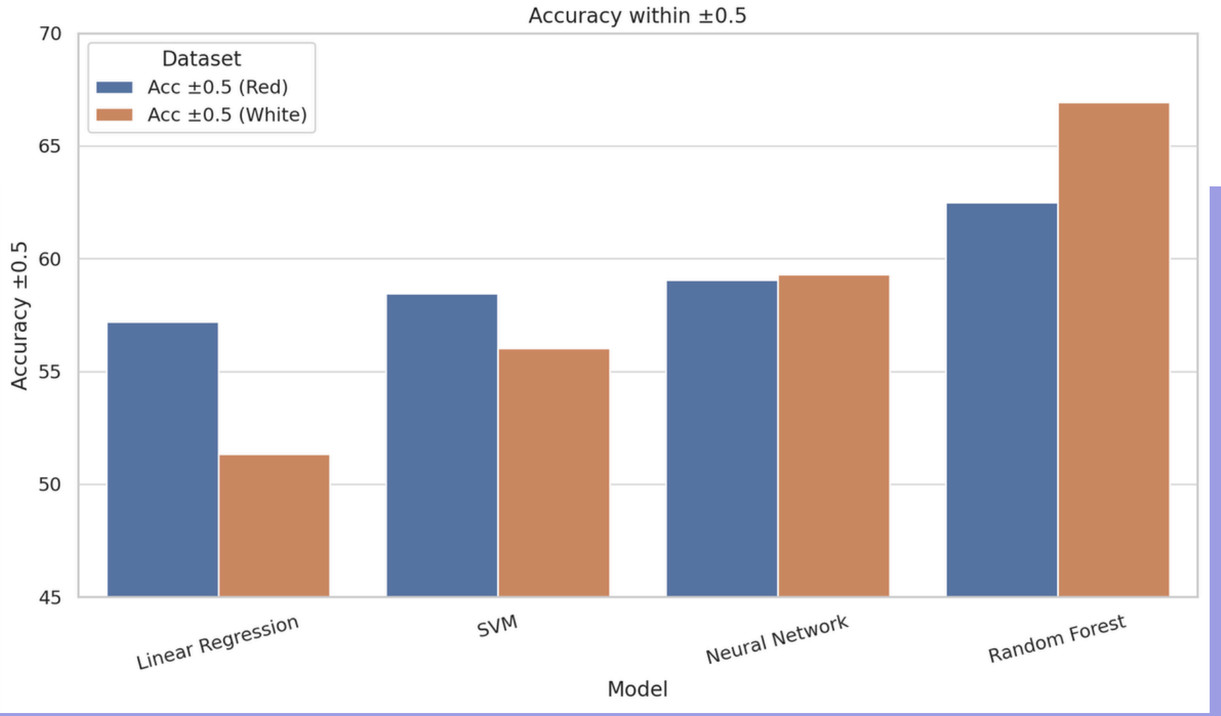
Red Wine



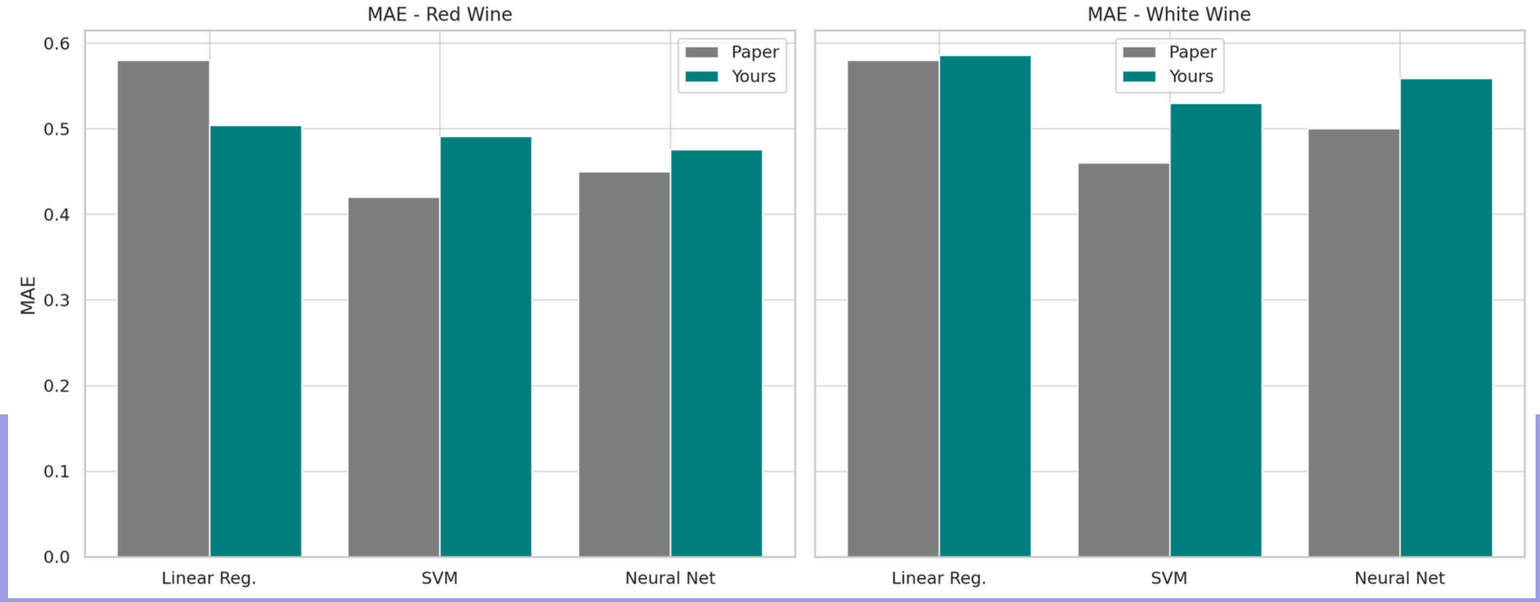
White Wine

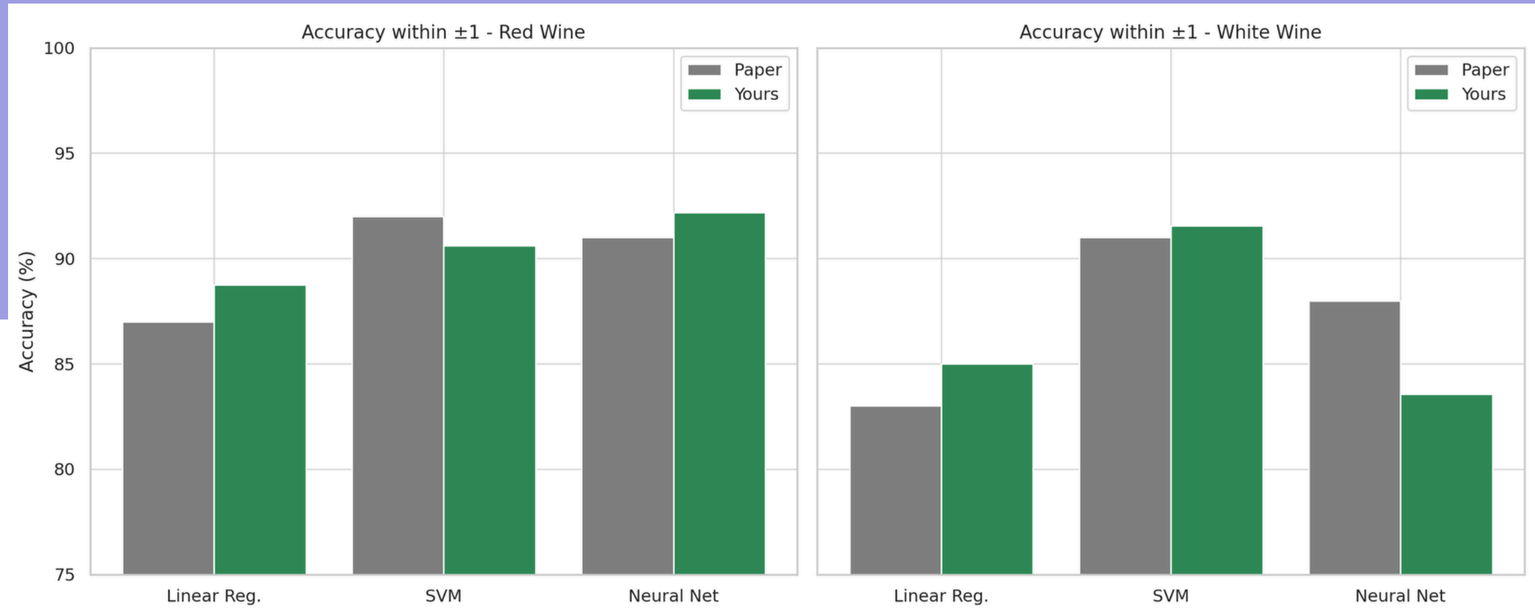
Results


Accuracy within ±1
Dataset
Acc ±1 (Red)
Acc ±1 (White)
Accuracy ±1
94
92
90
88
86
84
82
80
Linear Regression
SVM
Neural Network
Random Forest
Model


Accuracy within ±0.5
Dataset
Acc ±0.5 (Red)
Acc ±0.5 (White)
Accuracy ±0.5
70
65
60
55
50
45
Linear Regression
SVM
Neural Network
Random Forest
Model

Results


MAE - Red Wine
MAE - White Wine
Paper
Yours
MAE
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Linear Reg.
SVM
Neural Net


Accuracy within ±1 - Red Wine
Accuracy within ±1 - White Wine
Paper
Yours
Accuracy (%)
75
80
85
90
95
100
Linear Reg.
SVM
Neural Net

Results: Scope of Improvement

There's still room to push these models even further.
By applying feature selection, we could remove noisy or less relevant variables, which might help the models focus on what really matters.
Some studies have shown that when Random Forest is used with only the most important features, it can reach very high accuracy — even close to 99.5%.

Other areas worth exploring could be:

- Feature selection: Removing weakly predictive or redundant features to reduce noise and improve model clarity.
- Dimensionality reduction: Techniques like PCA can help simplify the feature space.
- Ensemble stacking: Combining predictions from multiple models could enhance robustness.
- Domain-specific feature engineering: Incorporating wine-specific knowledge (e.g., ratios or thresholds) might better capture hidden patterns.
- Other regression methods like adaptive boosting could be tried, as they have reportedly given very high accuracy in some datasets.

Thank You

30 April, 2025