Project Report

Sushane Dulloo
Indraprastha Institute of Information Technology, Delhi
Okhla Industrial Estate, Phase 3, New Delhi, Delhi, India
sushane21292@iiitd.ac.in

Jatin Kumar
Indraprastha Institute of Information Technology, Delhi
Okhla Industrial Estate, Phase 3, New Delhi, Delhi, India
jatin21153@iiitd.ac.in

Madhav Vyas
Indraprastha Institute of Information Technology, Delhi
Okhla Industrial Estate, Phase 3, New Delhi, Delhi, India
madhav20310@iiitd.ac.in

Arjun Gupta
Indraprastha Institute of Information Technology, Delhi
Okhla Industrial Estate, Phase 3, New Delhi, Delhi, India
arjun21134@iiitd.ac.in

Abstract

This paper is the interim report for our project and highlights in detail our motivation for choosing this project, the literature review of 2 papers of people who had previously worked upon this problem, the description and analysis of the dataset, also the methodology that we pursued, and the results and conclusions we have drawn from the ML models.

1. Abstract

To develop a reliable and accurate machine learning model that can predict the quality of wines based on their chemical properties. Wine quality assessment is crucial for vineyards, wineries, and consumers, as it influences production decisions and purchasing choices. By creating a robust predictive model, we aim to assist winemakers in optimizing their processes.

2. Introduction

Wine quality is influenced by numerous factors, including grape variety, weather conditions, winemaking techniques and chemical properties. Traditional wine assessment methods often rely on subjective human judgment, which can be inconsistent and time-consuming. Machine learning offers an opportunity to create a data-driven approach for predicting wine quality based on objective features and historical data.

The primary objective of this project is to build a predictive model that accurately estimates wine quality based on a set of quantifiable features. This model will serve as a valuable tool for wine producers, distributors and consumers, enabling them to make informed decisions about wine selection, production processes and quality improvement strategies.
We will leverage a dataset containing information about various wines, including attributes such as acidity levels, residual sugar, alcohol content and more. This dataset has been collected from diverse sources and includes both red and white wines.

3. Literature Review

We have read and in this section discussed two research papers. Both of these are previous works on the same problem and work with different models, in different programming languages with datasets similar to ours. A brief description of these papers has been given.

3.1. Prediction of Wine Quality: Comparing Machine Learning Models in R Programming

The paper begins by highlighting the importance of determining wine quality, which is a common interest among researchers and consumers. It discusses the challenges of using both sensory and physicochemical tests to evaluate wine quality.

It mentions the evolution of industries, especially the chemical industry, in analyzing data and collaborating across different fields. The text emphasizes the need for a better approach to wine quality assessment.

It points out the importance of collaboration between data scientists, computer scientists, chemical engineers, and others in improving wine quality research. It highlights the use of object-oriented programming, particularly Python and R, in the industry.

It discusses the role of machine learning algorithms and statistical analysis in wine quality assessment. It mentions

that while Python is commonly used, R is also effective and should not be overlooked.

It mentions that the dataset used for the research was obtained from Kaggle and describes the dataset's features, including input variables related to physicochemical tests and the output variable related to sensory data. Data preprocessing steps are discussed.

It outlines the use of various R packages for building machine learning models, including neural networks, naïve Bayes classification, and random forests. It also mentions the use of packages for data visualization and manipulation.

It discusses the issue of data imbalance, where the majority of wine quality scores fall between 5 and 6. This poses a challenge in analysis.

Concludes that the Random Forest model performed best in predicting wine quality and suggests that alcohol level is a key factor. It also mentions that the data imbalance was a challenge and recommends using a larger dataset for future research.

3.2. Prediction of Wine Quality Using Machine Learning Algorithms

In this study, a diverse array of machine learning algorithms was harnessed to forecast wine quality. The dataset was meticulously divided into distinct training and testing sets, adhering to a pivotal 3:1 ratio. This stratagem was instrumental in averting overfitting, a phenomenon where a model excels in its training domain but falters when applied to unobserved testing data, indicating a deficiency in generalization. The researchers judiciously employed sophisticated methodologies such as Ridge Regression, Support Vector Machine (SVM), Gradient Boosting Regressor, and Artificial Neural Networks. Through a process of meticulous parameter tuning tailored to each algorithm, the research team endeavored to achieve precise wine quality predictions. This meticulous calibration of parameters played a pivotal role in ensuring the models' adeptness at generalizing to unseen data, thereby achieving a delicate equilibrium between underfitting and overfitting. Consequently, this approach significantly bolstered the accuracy and dependability of the predictive outcomes.

4. Dataset Description

4.1. Composition

The dataset we have used has been arrived at by combining 2 datasets, [3] and [4] that have been obtained from Kaggle. The white wines form 75 percent of the datapoints and the red the remaining 25.

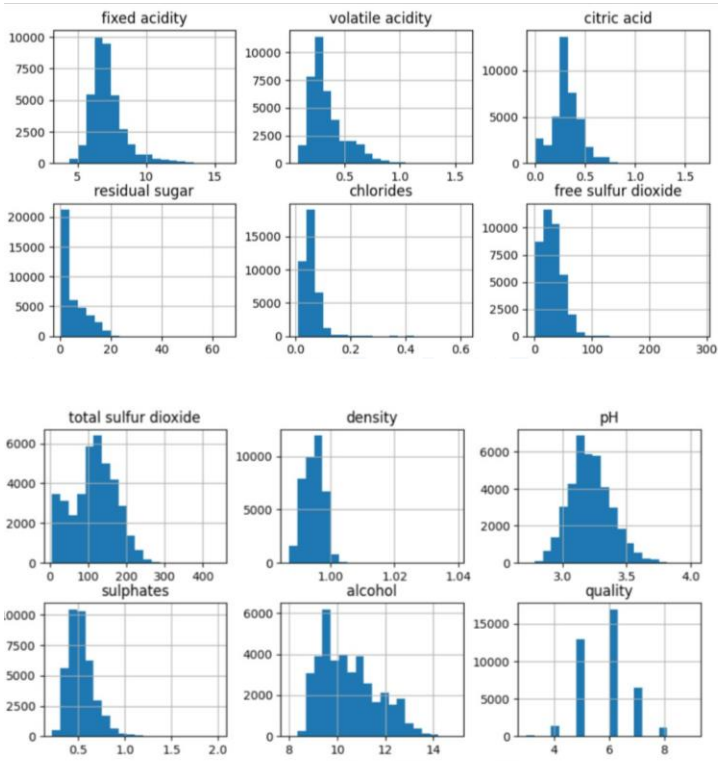
4.2. Dataset Size and Attributes

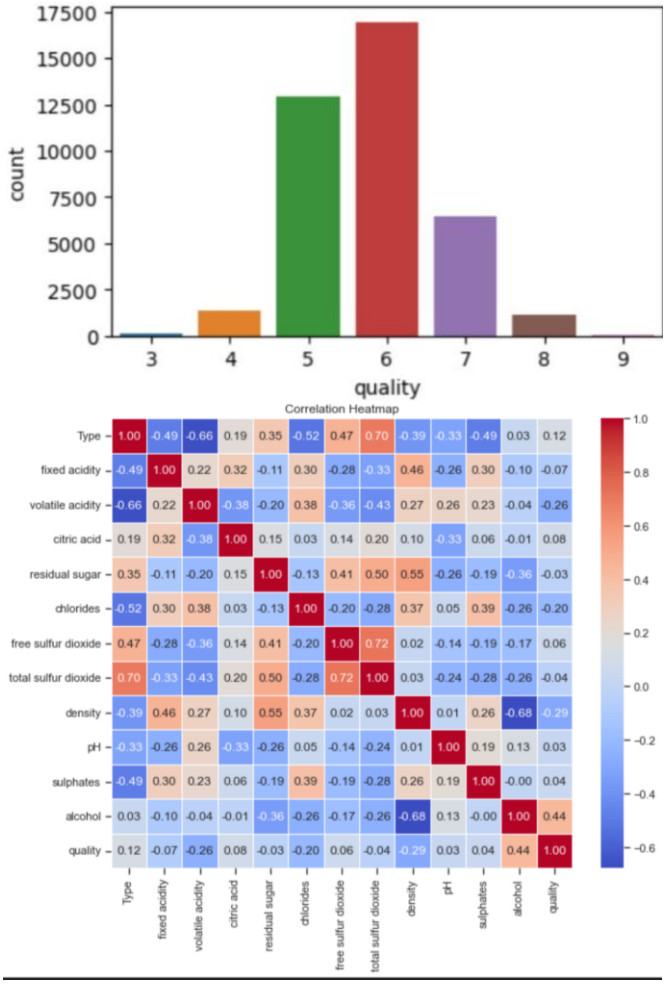
The final dataset has 38983 data points after preprocessing and has 13 attributes namely, type of wine(red or white), fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.

4.3. Data Preprocessing

Both datasets had the same attributes hence combining them was simple. The parameter type of wine as to be encoded (to 0 or 1 value) to get all numeric values. Certain data points had some missing values and hence had to be deleted. Data has been divided into training and testing data in the ratio of 80 : 20.

4.4. Data Visualization





4.5. Stats and Observations

We can clearly observe that the data we have has most of the data points with wine quality between the range of 4 to 8, hence we lack extreme ends of the data and will find it difficult to classify such data points if encountered in the future. From the heatmap, we observe that there is a strong correlation between alcohol content and the quality of wine. And also a negative correlation between the quality of wine and volatile acidity, chlorides and density.

5. Methodology

We first selected the datasets from Kaggle and combined them, completed the preprocessing of the data by deleting certain data points with missing values and encoded attributes. Performed data analysis on the dataset and divided it into training-testing data in the ratio of 80:20. Then modeled logistic regression and SVM on the data and are continuing to consider more models to choose the one with the best performance.

5.1. Logistic Regression Model

In our wine quality prediction project, we implemented a logistic regression model as one of the predictive tools to assess and classify wine quality. Logistic regression, a widely-used statistical method, is traditionally applied to binary classification problems. However, in this project, we have applied it to the task of multiclass classification to predict wine quality levels.

The logistic regression model works by calculating the probability that a given wine sample belongs to a particular class, which is then transformed using the logistic function into a value between 0 and 1. This probability is interpreted as the likelihood of the wine falling into a specific quality category.

Using our dataset the model learned to assign the appropriate quality label to a wine sample based on the provided features.

During the evaluation phase, we assessed the model's performance using classification metrics such as accuracy, precision, recall, etc. These metrics provided insights into the model's ability to correctly classify wings into their respective quality categories.

The logistic regression model served as a valuable component in our wine quality prediction project, contributing to our efforts to provide consumers and producers with a reliable tool for assessing and categorizing wine quality, ultimately enhancing decision-making processes within the wine industry.

5.2. SVM Model

In our project, we employed a Support Vector Machine(SVM) model to enhance our predictive capabilities. SVM is a powerful machine learning algorithm known for its versatility in both classification and regression tasks. In our project, we adapted SVM for multiclass classification to predict wine quality levels accurately.

The SVM model operates by finding the optimal hyperplane that best separates data points belonging to different quality categories while maximizing the margin between these classes. This margin represents the model's ability to generalize well to unseen data.

For our wine quality prediction, we fed the SVM model, our dataset. The model learned to create a decision boundary that effectively classified wine samples into their respective quality groups.

During evaluation, we assessed the SVM model's performance using classification metrics like accuracy, precision, recall, etc. These metrics allowed us to gauge the model's effectiveness in correctly categorizing wines.

The SVM model served as a valuable tool in our wine quality prediction project, providing a robust and accurate means of assessing and classifying wines.

5.3. Decision Tree Model

In our machine learning project focused on wine quality prediction, the Decision Tree model played a pivotal role in unraveling the intricate relationships between various wine characteristics and their ultimate quality ratings. We began by meticulously preprocessing the wine dataset, addressing missing values and encoding categorical variables to ensure a clean and standardized input for the model. Feature selection became a critical step, and the Decision Tree's inherent ability to assess the importance of each feature guided us in identifying key factors influencing wine quality. The model was trained using the CART algorithm, and its performance was evaluated through rigorous cross-validation. The interpretability of the Decision Tree proved invaluable, allowing us to trace decision paths and discern which attributes carried the most weight in determining wine quality. Through this approach, our ML project not only achieved high prediction accuracy but also provided valuable insights into the intricate interplay of factors influencing the quality of wines, offering potential guidance for wine producers and enthusiasts alike.

5.4. Random Forest Model

In our machine learning project focused on predicting wine quality, the Random Forest model served as a cornerstone, leveraging ensemble learning to enhance the accuracy and robustness of our predictions. The Random Forest's strength lies in its ability to mitigate overfitting by constructing multiple decision trees on bootstrapped subsets of the data, introducing variability through both sample and feature selection. The interpretability of the model was enriched by examining feature importance across the ensemble, offering valuable insights into the complex interplay of various wine attributes. Ultimately, the Random Forest model emerged as a robust and effective tool in our wine quality prediction project, contributing to both the accuracy of predictions and the depth of our understanding of the factors influencing wine quality.

6. Results of Models

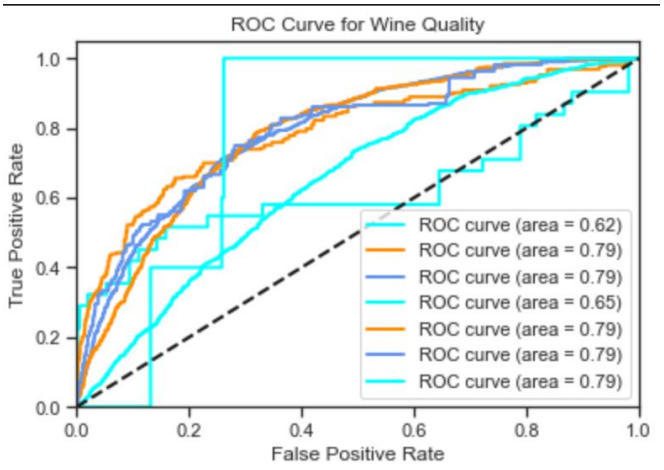
6.1. Results of the Logistic Regression Model

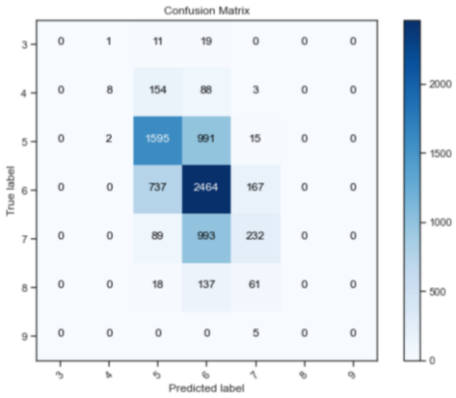
Training MSE: 0.641
Testing MSE: 0.622
Training RMSE: 0.800
Testing RMSE: 0.789
Training R2 Score: 0.161
Testing R2 Score: 0.171
Training MAE: 0.516
Testing MAE: 0.502

Training Accuracy: 0.541
Testing Accuracy: 0.551

Classification Report (Testing Data):

	precision	recall	f1-score	support
3	1.00	0.00	0.00	31
4	0.73	0.03	0.06	253
5	0.61	0.61	0.61	2603
6	0.53	0.73	0.61	3368
7	0.48	0.18	0.26	1314
8	1.00	0.00	0.00	216
9	1.00	0.00	0.00	5
accuracy			0.55	7790
macro avg	0.76	0.22	0.22	7790
weighted avg	0.57	0.55	0.51	7790



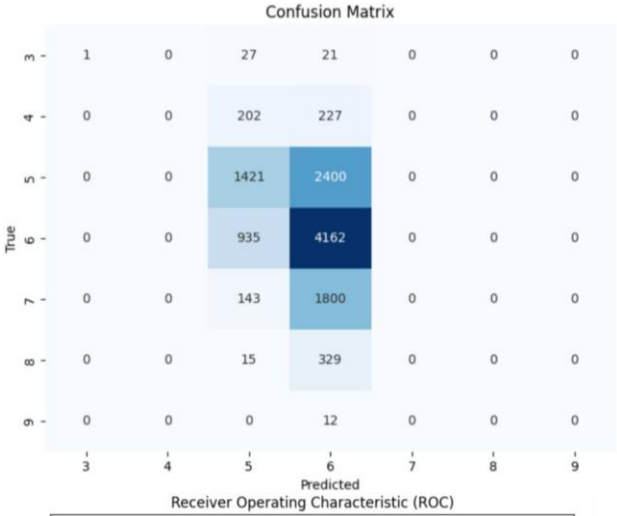


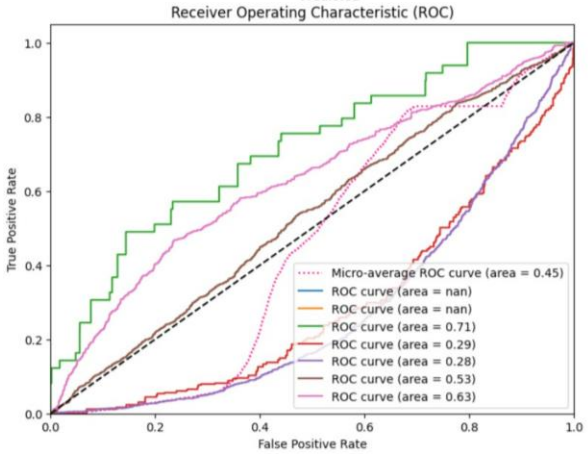
6.2. Results of the SVM Model

Training MSE: 0.73
Testing MSE: 0.74
Training RMSE: 0.85
Testing RMSE: 0.87
Training R2 Score: 0.04

Testing R2 Score: 0.04
Training MAE: 0.59
Testing MAE: 0.59
Training Accuracy: 0.4806
Testing Accuracy: 0.4775

	precision	recall	f1-score	support
3	1.00	0.02	0.04	49
4	0.00	0.00	0.00	429
5	0.52	0.37	0.43	3821
6	0.46	0.82	0.59	5097
7	0.00	0.00	0.00	1943
8	0.00	0.00	0.00	344
9	0.00	0.00	0.00	12
accuracy			0.48	11695
macro avg	0.28	0.17	0.15	11695
weighted avg	0.38	0.48	0.40	11695





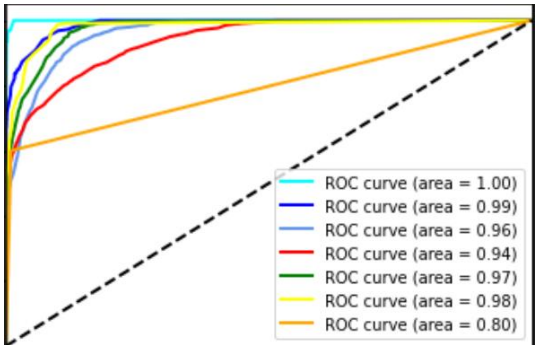
6.3. Results of Decision Tree Model

```
Fitting 5 folds for each of 3024 candidates, totalling 15120 fits
Accuracy: 0.82
Cross-Validation Scores: [0.82747112 0.8156611  0.82400513 0.83207087 0.81640775]
Mean Accuracy: 0.82
Training MSE: 0.22623403299313177
Testing MSE: 0.2590500641848524
Training RMSE: 0.47564065531988725
Testing RMSE: 0.5089696102763429
Training R2 Score: 0.7038860109631078
Testing R2 Score: 0.6550688405597556
Training MAE: 0.18220039797162849
Testing MAE: 0.20462130937098844
Training Accuracy: 0.8376018999935811
Testing Accuracy: 0.8192554557124518
Classification Report (Testing Data):
              precision    recall  f1-score   support

           3       0.88      0.90      0.89        31
           4       0.89      0.73      0.80       253
           5       0.81      0.88      0.84      2603
           6       0.83      0.82      0.82      3368
           7       0.82      0.76      0.79      1314
           8       0.68      0.59      0.63       216
           9       1.00      0.00      0.00         5

    accuracy                           0.82      7790
   macro avg       0.84      0.67      0.68      7790
weighted avg       0.82      0.82      0.82      7790
```

28	1	0	2	0	0	0
0	184	58	11	0	0	0
0	5	2284	288	19	7	0
1	9	427	2759	150	22	0
3	7	42	234	999	29	0
0	0	0	37	51	128	0
0	0	0	0	3	2	0



False Positivity Rate →

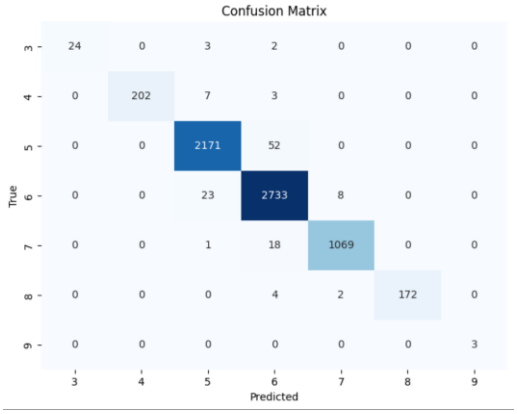
6.4. Results of Random Forest Model

```
Best Hyperparameters: {'max_depth': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
Accuracy(Train): 99.87%
Accuracy(Test): 98.11%
Classification Report:
              precision    recall  f1-score   support

           3       1.00      0.83      0.91        29
           4       1.00      0.95      0.98       212
           5       0.98      0.98      0.98      2223
           6       0.97      0.99      0.98      2764
           7       0.99      0.98      0.99      1088
           8       1.00      0.97      0.98       178
           9       1.00      1.00      1.00         3

    accuracy                           0.98      6497
   macro avg       0.99      0.96      0.97      6497
weighted avg       0.98      0.98      0.98      6497
```

```
TEST:
Mean Squared Error (MSE): 0.03
Root Mean Squared Error (MSE): 0.16
Mean Absolute Error (MAE): 0.02
R-squared (R2) Score: 0.96
TRAIN:
Mean Squared Error (MSE): 0.00
Root Mean Squared Error (MSE): 0.05
Mean Absolute Error (MAE): 0.00
R-squared (R2) Score: 1.00
```



7. Conclusion

In this interim report, we embarked on a comprehensive exploration of the wine quality prediction problem, aiming to leverage machine learning techniques to enhance the wine industry's decision-making processes. Our study was motivated by the need for a reliable and objective method to assess wine quality, considering the multitude of factors influencing it. Through an in-depth literature review, we gleaned insights from previous works, delving into diverse machine-learning models, methodologies, and challenges faced by researchers in similar domains.

Our dataset, meticulously curated from multiple sources, provided a rich foundation for our analysis. After rigorous data preprocessing, where we addressed missing values and encoded attributes, we conducted extensive exploratory data analysis. This analysis illuminated valuable patterns, including strong correlations between alcohol content and wine quality, as well as negative correlations with volatile acidity, chlorides, and density.

We pursued the implementation of four fundamental machine learning models: Logistic Regression, Decision Trees, Random Forests and Support Vector Machine (SVM). Logistic Regression, applied in the context of multiclass classification, allowed us to predict wine quality levels by calculating probabilities and assigning appropriate quality labels.

Our results, indicate promising outcomes. All models showcased their potential in predicting wine quality, setting the stage for further exploration and refinement, at the end , Random Forest Model prevailed and works the best. However, challenges such as the lack of extreme quality data points and the potential impact of data imbalance surfaced during our analysis, emphasizing the need for a more extensive and balanced dataset for future research.

In conclusion, our interim findings lay a solid foundation for our ongoing research. By leveraging machine learning, we are poised to revolutionize wine quality assessment, offering vineyards, wineries, and consumers a dependable tool for informed decision-making. As we progress, further refinements and explorations are anticipated, ensuring our predictive models reach their full potential in accurately estimating wine quality based on objective chemical properties. Future research could focus on models such as XGBoost, etc.

References

[1] https://www.researchgate.net/profile/Olatunde-Akanbi-3/publication/364030507_Prediction_of_Wine_Quality_Comparing_Machine_Learning_Models_in_R_Programming/links/633703989cb4fe44f3ee27e9/Prediction-of-Wine-Quality-Comparing-Machine-Learning-Models-in-R-Programming.pdf

[2] https://www.scirp.org/journal/paperinformation.aspx?paperid=107796

[3] https://www.kaggle.com/datasets/rajyellow46/wine-quality

[4] https://www.kaggle.com/datasets/ankitsen910/wine-quality-dataset