Predicting Wine Quality Using Machine Learning

AIT 664: Information: Representation, Processing & Visualisation

Aakiff Panjwani
College of Engineering and Computing - DAE
George Mason University
Fairfax, United States
apanjwa@gmu.edu

Sai Kurra
College of Engineering and Computing - DIST
George Mason University
Fairfax Virginia United States
skurra5@gmu.edu

Srija Anasuri
College of Engineering and Computing - DAE
George Mason University
Fairfax, United States
sanasuri@gmu.edu

***Abstract* — Evaluating wine quality has traditionally relied on expert sommeliers and tasters, a process that is subjective, time-intensive, and costly. With the growing diversity of wine varieties and brands, there is a pressing need for automated, scalable methods to predict wine quality. This study leverages data from the Vinho Verde Wine Dataset to investigate correlations between physicochemical properties—such as fixed acidity, volatile acidity, pH, and alcohol content—and wine quality ratings, identifying key factors that influence quality. Various machine learning models are developed and compared to assess their effectiveness in predicting wine quality. Additionally, dimensionality reduction techniques like Principal Component Analysis (PCA) are employed to evaluate how they enhance the model's accuracy and efficiency by addressing multicollinearity, reducing feature dimensionality, and mitigating overfitting. The findings demonstrate the potential of data-driven approaches to replace traditional sensory evaluations, offering an accessible, scalable, and objective solution for wine quality assessment while providing valuable insights into the relationship between physicochemical properties and wine quality.**

Keywords—Machine Learning Models, Principal Component Analysis (PCA), Model Performance, Feature Correlation.

I. INTRODUCTION

Wine quality assessment is a vital aspect of the wine industry, directly affecting consumer preferences, production strategies, and market trends. Traditionally, this evaluation has relied on expert sommeliers, whose subjective and time-consuming assessments often come at a high cost. As the variety of wines continues to expand, there is a growing demand for automated and scalable approaches to predict wine quality more efficiently and objectively.

The physicochemical properties of wine, such as fixed acidity, volatile acidity, pH, and alcohol content, are known to have a significant impact on wine quality. These measurable attributes provide an opportunity to move beyond traditional methods and adopt data-driven approaches to quality prediction. The Vinho Verde Wine Dataset, which includes data on both physicochemical features and quality ratings, serves as an ideal foundation for this exploration.

Machine learning models offer a promising solution for predicting wine quality based on these features by uncovering complex patterns and relationships in the data. In this study, we evaluate the performance of various ML classification models to predict wine quality. Additionally, we incorporate dimensionality reduction techniques such as Principal Component Analysis (PCA) to improve model performance. PCA reduces the number of features, addressing challenges like multicollinearity and overfitting, while focusing on the most important information, which enhances both model accuracy and computational efficiency.

To shape our analysis, we established three fundamental research questions -

1. What relationships exist between the physicochemical characteristics of wine—such as fixed acidity, volatile acidity, pH, and alcohol content—and its overall quality?

2. How well do different classification models perform in predicting wine quality based on the available characteristics?

3. In what ways does the use of PCA (Principal Component Analysis) improve the effectiveness of machine learning models in predicting wine quality?

The primary goal of this research is to move from traditional sensory evaluations to a more accessible, automated, and data-driven approach for predicting wine quality. The results aim to provide valuable insights into the

relationship between physicochemical properties and wine quality, offering a more scalable and objective solution for the wine industry.

II. DATASET OVERVIEW AND PREPARATION

The dataset used in this study consists of red and white wine samples from the Portuguese "Vinho Verde" wine, with the input features derived from chemical tests. The output variable is based on sensory evaluations from wine experts rating the quality of the wines on a scale from 0 to 10. The dataset includes the following -

Number of Instances -

- Red wine - 1,599 samples
- White wine - 4,898 samples

Input Variables -

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

Output/Target Variable -

- Quality (score between 0 and 10)

The dataset is used to explore the relationships between the physicochemical properties of wine and its quality rating and to build machine-learning models that predict wine quality based on these features. Preprocessing steps are applied to prepare the data for analysis, including handling missing values, scaling numerical features, and encoding the target variable for model training.

The red and white wine datasets are first loaded into separate DataFrames, and a new "WineType" column was added to distinguish between the two before merging them into a single DataFrame for analysis. The combined dataset consists of 6,497 rows and 13 columns, reflecting its well-structured nature. Initial data exploration included generating descriptive statistics to gain insights into the distribution and central tendencies of the features.

```
In [388]: # Summary of the Wine dataset
          Wine.describe()
```

Out[388]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	
count	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497.000000	6497
mean	7.215307	0.339666	0.318633	5.443235	0.056034	30.525319	115.744574	0.994697	3.218501	0.531268	10.491801	5
std	1.296434	0.164636	0.145318	4.757804	0.035034	17.749400	56.521855	0.002999	0.160787	0.148806	1.192712	0
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.000000	0.987110	2.720000	0.220000	8.000000	3
25%	6.400000	0.230000	0.250000	1.800000	0.038000	17.000000	77.000000	0.992340	3.110000	0.430000	9.500000	5
50%	7.000000	0.290000	0.310000	3.000000	0.047000	29.000000	118.000000	0.994890	3.210000	0.510000	10.300000	6
75%	7.700000	0.400000	0.390000	8.100000	0.065000	41.000000	156.000000	0.996990	3.320000	0.600000	11.300000	6
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.000000	1.038980	4.010000	2.000000	14.900000	9

TABLE I. Descriptive Statistics

A check for missing values revealed no null entries, ensuring that the data is complete for analysis. Duplicate rows are identified, which is essential for preventing any distortion in analysis and model performance. Duplicates are found in both the red and white wine datasets, where certain entries had identical values across all features. These duplicates are removed, resulting in a clean dataset. Label encoding is then applied, assigning "red" as 0 and "white" as 1. For outlier detection, boxplots are used to visualize the data distribution and identify potential outliers. The Interquartile Range (IQR) method was chosen for handling outliers, as it is more robust to extreme values than methods based on mean and standard deviation, making it suitable for skewed data. Rather than removing the outliers, imputation is used to replace them, which preserves the dataset's integrity while addressing extreme values.

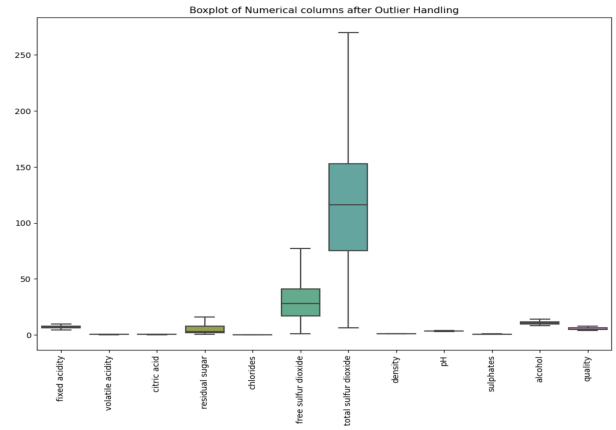

Fig. 1. Boxplot after handling the outliers

Exploratory Data Analysis (EDA) is performed to gain a deeper understanding of the dataset and its underlying

patterns. This process involved a series of steps to visually and statistically analyze the data.

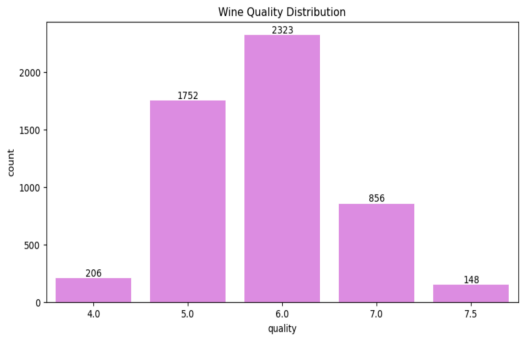

Fig. 2. Wine Quality Distribution

The wine quality distribution graph shows that most of the wine samples are rated with a quality score of 6, while there are significantly fewer samples with quality ratings of 4 and 7.5.

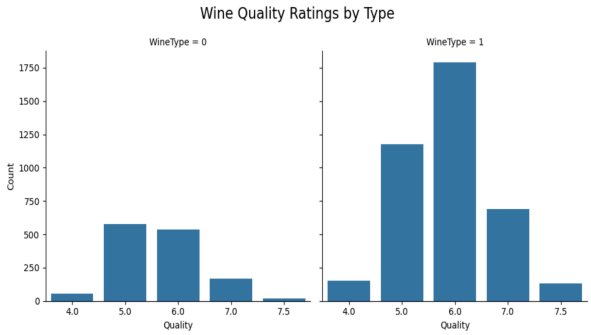

Fig. 3. Wine Quality Ratings by Type

From the above graph, we can see that red wines (0) primarily receive quality ratings of 5 and 6, while most white wines (1) are rated with scores of 5, 6, and 7.

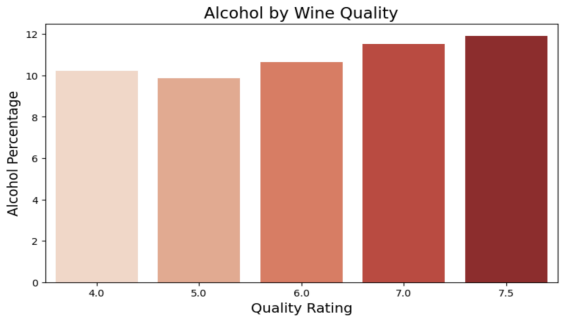

Fig. 4. Alcohol by Wine Quality

The graph clearly illustrates that wines with higher alcohol content tend to receive better quality ratings.

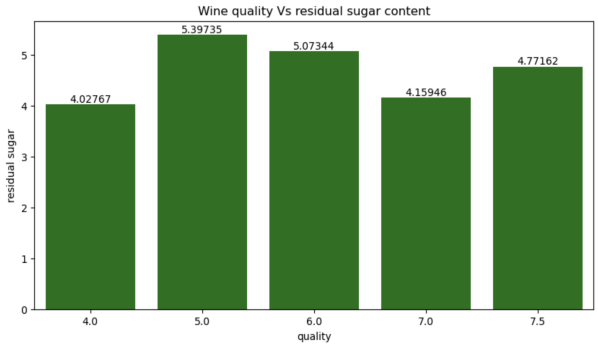

Fig. 5. Wine Quality vs residual sugar content

Wines with a quality rating of 5 have the highest residual sugar, whereas wines with a quality rating of 4 have the lowest. In higher-quality wines (rated 7 and above), residual sugar slightly decreases, indicating a more balanced sweetness. Moderate sugar levels appear to be important for wines with mid-range quality.

A new column, “WineClass”, is created to categorize the wines based on their quality ratings. Wines with a quality score of less than 6 are labeled as 0 (bad), while those with a score of 6 or higher are labeled as 1 (good). Additionally, the dataset is standardized to ensure that all features, particularly those with different scales, have a uniform range. This step is crucial for improving the performance of machine learning models by eliminating scale-related biases during model training.

III. CORRELATION ANALYSIS

The research investigates the relationships between various physicochemical attributes of wine, including fixed acidity, volatile acidity, pH, and alcohol content, and their impact on overall wine quality. A correlation matrix created with Python shows that alcohol content has the most significant positive correlation with quality at 0.48, suggesting that higher alcohol levels improve wine quality. In contrast, volatile acidity and density show a negative correlation with quality. A heatmap accompanying the analysis visually reflects these correlations, emphasizing the important relationship between free and total sulfur dioxide, along with the moderate positive effect of alcohol on quality.

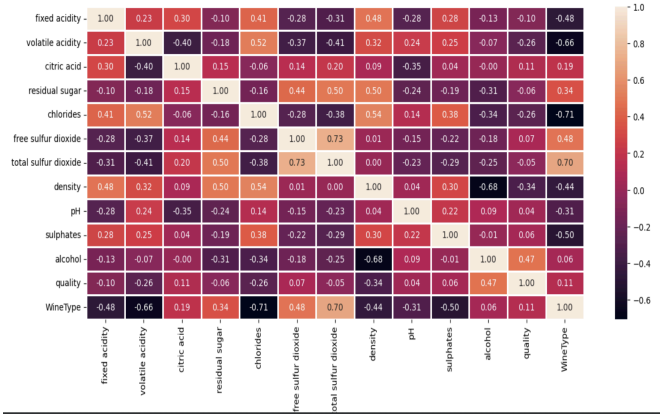

Fig. 6. Heatmap Correlation

Overall, the results highlight the essential role of wine type in affecting volatile acidity and total sulfur dioxide, with implications for enhancing properties to elevate wine quality. Feature selection is employed by correlation analysis, which identifies specific attributes closely linked to wine quality. For example, alcohol content exhibited a moderate positive correlation of 0.47, while density displayed a moderate negative correlation of -0.33. Additionally, volatile acidity had a negative correlation of -0.27, suggesting its significance in identifying lower-quality wines.

Feature selection is based on these findings. We retained features with substantial correlations to quality while removing those with weaker connections. This refined selection allowed us to focus on the most impactful predictors, improving the clarity of our analysis and the interpretability of the model. Tools like correlation heatmaps are instrumental in highlighting these relationships. In conclusion, the dataset uncovered several fascinating patterns that shaped our feature engineering and model selection strategies.

IV. MACHINE LEARNING MODELS

Various classification models can be evaluated to predict wine quality based on the available chemical and other feature data. Logistic Regression, Support Vector Machine (SVM), and Random Forest are tested for their effectiveness. Logistic Regression performs well for linear relationships but may face challenges with more complex, non-linear data. SVM is particularly suited for high-dimensional data with well-separated classes, though it can be computationally expensive. Random Forest, on the other hand, is highly effective at capturing non-linear relationships and interactions between features, and it typically delivers the best performance in terms of accuracy. While Logistic Regression provides a useful baseline, and SVM is beneficial when class separation is clear, Random Forest generally outperforms the others in terms of predictive accuracy.

The dataset is divided into training and testing sets to evaluate the model's performance. For feature selection, a subset of features was chosen, which includes alcohol content, pH, free sulfur dioxide, sulfates, volatile acidity, density, and citric acid, which have a significant influence on wine quality. The dataset is split into training and testing sets, with 80% of the data used for training and 20% reserved for testing, ensuring a fair evaluation of the model.

A. LOGISTIC REGRESSION

Logistic Regression is used as one of the classification models. It is a linear model that predicts the probability of a binary outcome, in this case, whether the wine is of "Good" or "Bad" quality. Logistic Regression works by finding the best-fitting line that separates the two classes. While it is effective for linearly separable data, it may not perform as well on more complex or non-linear relationships within the dataset. It serves as a good baseline model, providing insights into the linear correlation between the selected features and wine quality.

Logistic Regression Classification Report –

	precision	recall	f1-score	support
0	0.65	0.55	0.60	377
1	0.77	0.83	0.80	680
accuracy			0.73	1057
macro avg	0.71	0.69	0.70	1057
weighted avg	0.73	0.73	0.73	1057

Accuracy – 0.7332071901608326
ROC-AUC Score – 0.7841199875175535

TABLE II. Classification Report for Logistic Regression

The Logistic Regression model achieved an accuracy of 73%, demonstrating good performance in classifying wine quality. Precision scores were 0.65 for lower-quality wines (Class 0) and 0.77 for higher-quality wines (Class 1), while recall scores were 0.55 and 0.83, respectively. The F1 scores reflected a similar trend, with Class 0 at 0.60 and Class 1 at 0.80. Additionally, the model's ROC-AUC score of 0.78 indicates a good capability in distinguishing between the two classes. Overall, while the model performs effectively, particularly for higher-quality wines, there is

potential for improvement in identifying lower-quality wines.

B. RANDOM FOREST

Random Forest is an ensemble learning method that builds multiple decision trees on random subsets of the data and features, then averages their predictions to improve accuracy and reduce overfitting. In wine quality prediction, it excels at capturing non-linear relationships and complex feature interactions, unlike models like Logistic Regression that assume linearity. Random Forest is robust to overfitting and provides feature importance, helping identify which variables most influence the predictions. Its ability to handle complex patterns and deliver high accuracy makes it a powerful tool for predicting wine quality.

```
Random Forest Classification Report -
              precision    recall  f1-score   support

           0       0.68      0.65      0.66       377
           1       0.81      0.83      0.82       680

    accuracy                           0.77      1057
   macro avg       0.75      0.74      0.74      1057
weighted avg       0.76      0.77      0.77      1057

Accuracy -  0.7663197729422895
ROC-AUC Score -  0.8213917927913871
```

TABLE III. Classification Report for Random Forest

The Random Forest model achieved an accuracy of 76.3%, indicating good performance in classifying wine quality. It recorded precision scores of 0.68 for lower-quality wines (Class 0) and 0.81 for higher-quality wines (Class 1), with recall scores of 0.65 and 0.83, respectively. The F1 scores were 0.66 for Class 0 and 0.82 for Class 1. Additionally, the model achieved an ROC-AUC score of 0.82, further demonstrating its effectiveness in distinguishing between wine quality classes. Overall, the Random Forest model outperformed the Logistic Regression model and shows a good balance in predicting wine quality, though there is still room for improvement in identifying lower-quality wines.

C. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a supervised learning algorithm used for both classification and regression tasks. It works by finding the optimal hyperplane that best divides the data into distinct classes, aiming to maximize the margin between the support vectors of each class. SVM is highly effective in high-dimensional spaces and when there is a clear separation between classes. When the data is not linearly separable, SVM uses kernel functions to transform the data into a higher-dimensional space where a hyperplane can be found. Although SVM can be computationally expensive, it excels at classifying data with distinct class boundaries, making it suitable for tasks like predicting wine quality.

```
SVM Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.58      0.62       377
           1       0.78      0.84      0.81       680

    accuracy                           0.75      1057
   macro avg       0.73      0.71      0.72      1057
weighted avg       0.74      0.75      0.74      1057

Accuracy: 0.7483443708609272
ROC-AUC Score: 0.8068848494304883
```

TABLE IV. Classification Report for SVM

The SVM model achieved an accuracy of 74.8%, reflecting good performance in classifying wine quality. It recorded precision scores of 0.67 for lower-quality wines (Class 0) and 0.78 for higher-quality wines (Class 1), with recall scores of 0.58 and 0.84, respectively. The F1 scores were 0.62 for Class 0 and 0.81 for Class 1. The model also achieved a ROC-AUC score of 0.81, demonstrating effective differentiation between the two classes. Overall, while the SVM model performs well, particularly in identifying higher-quality wines, there is still room for improvement in detecting lower-quality wines.

To address the class imbalance in the wine quality dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE creates synthetic samples for the minority class, helping to balance the dataset. After applying SMOTE and resampling the data, we re-trained and evaluated the models—Logistic Regression, Support Vector Machine (SVM), and Random Forest. This approach allows the models to better learn from both classes and improve the accuracy of wine quality predictions.

```
Logistic Regression Classification Report -
              precision    recall  f1-score   support

           0       0.74      0.77      0.76       666
           1       0.76      0.73      0.75       665

    accuracy                           0.75      1331
   macro avg       0.75      0.75      0.75      1331
weighted avg       0.75      0.75      0.75      1331

Logistic Regression Accuracy -  0.7513148009015778
Logistic Regression ROC-AUC -  0.818717966086387
```

TABLE V. Classification Report for Logistic Regression

The Logistic Regression model achieved an accuracy of 75%, with precision scores of 0.74 for lower-quality wines (Class 0) and 0.76 for higher-quality wines (Class 1). Recall scores were 0.77 for Class 0 and 0.73 for Class 1, resulting in F1 scores of 0.76 and 0.75, respectively. The model also achieved a ROC-AUC score of 0.81, indicating decent performance in distinguishing between wine quality classes. Overall, while the model performs reasonably well, it improved from the previous logistic model.

Random Forest Classification Report –

	precision	recall	f1-score	support
0	0.78	0.84	0.81	666
1	0.83	0.76	0.79	665
accuracy			0.80	1331
macro avg	0.80	0.80	0.80	1331
weighted avg	0.80	0.80	0.80	1331

Random Forest Accuracy – 0.8016528925619835
Random Forest ROC-AUC – 0.8863340784393416

TABLE VI. Classification Report for Random Forest

The Random Forest model achieved an accuracy of 80%, with precision scores of 0.78 for lower-quality wines (Class 0) and 0.83 for higher-quality wines (Class 1). The recall scores were 0.84 for Class 0 and 0.76 for Class 1, resulting in F1 scores of 0.81 and 0.79, respectively. The model also achieved a ROC-AUC score of 0.89, demonstrating strong performance in distinguishing between wine quality classes. Overall, the Random Forest model shows effective classification, particularly excelling in identifying higher-quality wines. Even this model improved from the previous Random Forest model.

SVM Classification Report –

	precision	recall	f1-score	support
0	0.76	0.81	0.79	666
1	0.80	0.75	0.77	665
accuracy			0.78	1331
macro avg	0.78	0.78	0.78	1331
weighted avg	0.78	0.78	0.78	1331

SVM Accuracy – 0.7798647633358378
SVM ROC-AUC – 0.8411592043170991

TABLE VII. Classification report for SVM

The SVM model achieved an accuracy of 78%, with precision scores of 0.76 for lower-quality wines (Class 0) and 0.80 for higher-quality wines (Class 1). Recall scores are 0.81 for Class 0 and 0.75 for Class 1, resulting in F1 scores of 0.79 for Class 0 and 0.77 for Class 1. The model also achieved an ROC-AUC score of 0.84, indicating decent performance in distinguishing between wine quality classes. Overall, the SVM model performs reasonably well, particularly in identifying lower-quality wines, but has some challenges with higher-quality classifications.

V. RESULTS AND OBSERVATIONS

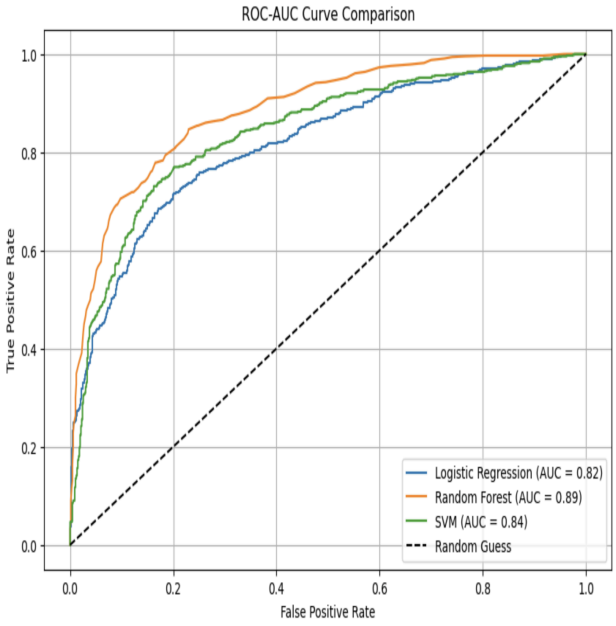

Fig. 7. ROC-AUC Curve Comparision

The ROC-AUC curve comparison highlights the performance of three models - Logistic Regression, Random Forest, and Support Vector Machine (SVM). The Random Forest model achieved the highest AUC score of 0.89, demonstrating excellent classification ability in differentiating between wine quality categories. Logistic Regression followed with an AUC of 0.82 and the SVM with an AUC of 0.84.

In conclusion, the Random Forest model proved to be the most effective for this classification task, offering the best accuracy and F1-Score. Although all models performed well, especially in identifying lower-quality wines, Random Forest emerged as the top choice for optimal wine quality prediction.

VI. PRINCIPAL COMPONENT ANALYSIS

To reduce dimensionality and focus on key features, we applied Principal Component Analysis (PCA). This technique simplifies the dataset by retaining the most

significant variance components while eliminating noise and redundancy. A Random Forest model was developed using the top five features, yielding an accuracy of 71%, indicating a moderate ability to predict wine quality.

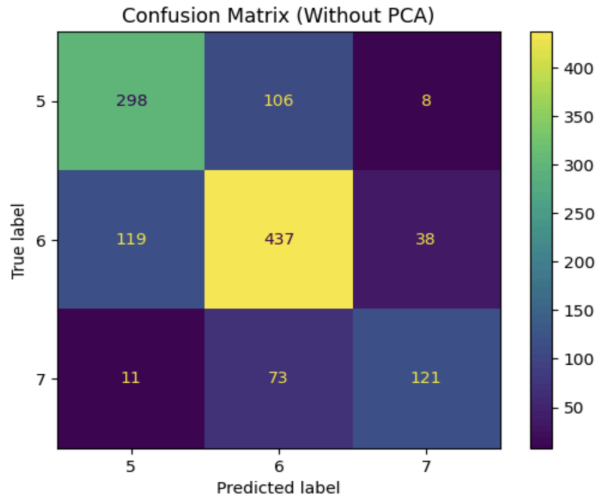

Fig. 8. Confusion Matrix without PCA

Fig. 9. PCA Scatterplot

The PCA scatter plot shows the first two principal components captured significant variance, but when the model was applied solely with these components, accuracy dropped to 67%. This decline can be attributed to information loss, misalignment of nonlinear components, oversimplification of the dataset, scaling issues, or potential amplification of noise. While PCA serves as a useful tool for dimensionality reduction, its effectiveness is context-dependent. In this case, the application of PCA did not yield the expected improvements and diminished model accuracy, highlighting the importance of carefully evaluating the impact of dimensionality reduction techniques on various datasets and predictive models.

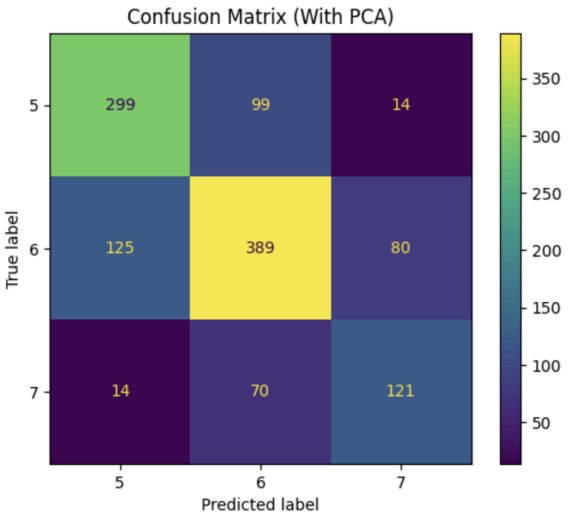

Fig. 10. Confusion Matrix with PCA

VII. CONCLUSION

This project effectively created a predictive model for assessing wine quality through physicochemical characteristics by utilizing machine learning methods. We pinpointed important factors, including alcohol content and volatile acidity, that have a considerable influence on the quality rating. Among the models tested, Random Forest was the most effective, outperforming Logistic Regression and SVM in prediction accuracy. While PCA is a powerful dimensionality reduction technique, its application in this case reduces model accuracy, highlighting the importance of evaluating its impact based on the dataset and specific problem context.

VIII. CONTRIBUTIONS

Everyone on the team contributed to data acquisition and initial exploration. Sai and Srija took charge of data cleaning, exploratory data analysis (EDA), and visualization. Srija and Aakiff worked on modeling, and implementing PCA. We all worked on drafting the final report and presentation.

ACKNOWLEDGMENT

We wish to extend our heartfelt gratitude to Professor Dr. Kazi Lutful Kabir for his invaluable guidance, feedback, and unwavering support during our research journey.

REFERENCES

[1] Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11(2), 278–289. https://doi.org/10.4236/ojs.2021.112015

[2] Santiago, E. (2024a). 18 best types of charts and graphs for data visualization [+guide]. Retrieved from https://blog.hubspot.com/marketing/types-of-graphs-for-data-visualization

[3] 1.4. Support Vector Machines. (2024). Scikit-Learn. https://scikit-learn.org/1.5/modules/svm.html

[4] Brownlee, J. (2020, April 7). 4 Types of Classification Tasks in Machine Learning. Machine Learning Mastery. https://machinelearningmastery.com/types-of-classification-in-machine-learning