Prediction of Wine Quality using Machine Learning

Nikunj Indoriya (22221), Aditya Dandhare (22019)
Department of Electrical Engineering and Computer Science
Indian Institute of Science Education and Research Bhopal

Abstract
Wine quality prediction plays a crucial role in the wine industry by enabling the evaluation and classification of wines based on their physicochemical properties. This paper explores how machine learning (ML) techniques can enhance the accuracy of wine quality assessment, offering a data driven approach to a traditionally subjective process. The challenges of predicting wine quality, such as dealing with diverse datasets, imbalanced classes, and noise in data, are discussed. The study examines various ML methods, including feature selection, regression models, and classification algorithms, to identify optimal approaches for this task. Real-world applications demonstrates the strengths of these methods in automating wine quality evaluation while addressing their limitations and potential improvements.
Keywords: Wine quality, Machine learning, physicochemical properties.

Contents

Introduction

The wine industry has long relied on the expertise of sommeliers and wine critics to evaluate the quality of wine. While traditional methods of quality assessment are valuable, they are often subjective and influenced by individual preferences. In contrast, a data-driven approach offers a more objective and reproducible way to evaluate wine quality based on measurable characteristics.

Wine quality is influenced by a variety of physicochemical properties such as acidity, pH, alcohol content, and residual sugar, which are measurable through laboratory analysis. Predicting wine quality based on these attributes is a complex task, as it requires uncovering intricate relationships between these properties and human perception of quality. Machine learning (ML) has emerged as a powerful and efficient tool to address such challenges, offering the ability to analyze large datasets and uncover patterns that may not be immediately evident to human experts.

This report explores the application of machine learning techniques to predict wine quality. By using available wine quality datasets, various ML models are trained and evaluated for their ability to classify wine into predefined quality categories. The goal of this project is to develop an accurate and robust model that can assist winemakers, retailers, and consumers in making informed decisions. Additionally, this study highlights the challenges in working with real-world datasets, such as handling imbalanced data and optimizing model performance.

In this report, we present a comprehensive pipeline for wine quality prediction, starting from data preprocessing and feature selection to model training and evaluation. The results demonstrate the potential of machine learning in transforming quality assessment processes

within the wine industry.

Data Preprocessing and Visualization

Data Preprocessing

The initial step in the data preprocessing phase involved reading the training dataset, which consisted of feature data and corresponding labels. The `x_train.csv` and `y_train.csv` files were loaded separately and combined into a single DataFrame, `wine_train_df`, using the `pd.concat()` function. The rows from both files were merged along the column axis, ensuring the dataset was structured appropriately for analysis.

The next step involved verifying the data types of the columns within the dataset. This was done using the `dtypes` function, which returned that all the columns were in the correct format. Specifically, 11 columns were of type `float64`, and 1 column (`quality`) was of type `int64`. Since all columns were already in the required format, no additional type conversions were necessary.

To ensure the dataset was free of any missing values, a check was performed using the `isnull().sum()` method. This check confirmed that there were no null values in the dataset, thus eliminating the need for any imputation procedures. This step guaranteed that the dataset was complete and ready for further analysis.

Data Visualization

After preprocessing, several visualizations were created to explore the relationships between the features and wine quality. A covariance matrix was first plotted using a heatmap to display the correlation coefficients between all pairs of features. This provided a comprehensive overview of how different features in the dataset relate to each other, with higher correlations indicating stronger relationships between variables.

Next, the relationship between two key features, `density` and `residual sugar`, was explored through a scatterplot. The Pearson correlation coefficient was calculated for wines of different quality levels. The analysis showed a moderate positive correlation of 0.41 for low-quality wines, 0.36 for medium-quality wines, and a stronger positive correlation of 0.64 for high-quality wines. These values suggest that the association between density and residual sugar becomes stronger as wine quality increases, which may provide useful insights for quality prediction models.

In addition to this, the frequency distribution of wine quality was analyzed. The distribution was visualized using a countplot, which revealed that most of the wines in the dataset had a quality rating of 5 or 6. The distribution showed a relatively small percentage of wines rated at the extremes (3 or 8), indicating that most wines fall within the middle range of quality. This frequency distribution helps to understand the imbalance in wine quality levels and suggests that the model may need to account for this distribution during training.

Further visualizations included a histogram of wine quality, which reinforced the observations from the countplot. This was followed by a density plot that illustrated the distribution of alcohol content across the wines, highlighting the concentration of wines with certain alcohol levels. A box plot was also created to visualize the distribution of multiple features, including `fixed acidity`, `pH`, and `alcohol`. This plot allowed for the identification of outliers and offered insights into the variability of each feature.

Finally, a violin plot of alcohol content by wine quality was created. This plot provided a visual representation of the distribution of alcohol content for wines of different quality levels, using the `Set2` color palette for better aesthetic appeal. The violin plot indicated that wines of higher quality tended to have a wider distribution of alcohol content, which might suggest the presence of more variability in the alcohol levels of higher-quality wines.

Overall, the visualizations conducted in this study provided valuable insights into the relationships between various wine attributes and quality. These analyses will serve as the foundation for subsequent model development, aiding in feature selection and ensuring that the most relevant factors are considered in building predictive models.

Models and Algorithms

In this project, multiple machine learning models were applied to predict wine quality, where the target variable was categorized into two classes: low and high quality. Logistic Regres-

sion was employed to model the relationship between various predictors such as acidity, sugar content, and alcohol level, and the binary target of wine quality. The model's performance was evaluated using accuracy, precision, recall, F1 score, and root mean square error (RMSE) to assess the effectiveness of logistic regression in handling the binary classification.

Random Forest, an ensemble method, was also applied to classify wine quality. The model's performance was analyzed through standard classification metrics, and feature importance was explored using an ExtraTreesClassifier. This helped identify which predictors contributed most to the classification task. The effect of the number of trees on the model's accuracy was investigated, showing how Random Forest scales with increasing estimators.

A Decision Tree model was implemented to further understand the relationship between predictors and wine quality. This method was effective in visually representing the decision-making process and in analyzing how each feature influenced the final prediction. The K-Nearest Neighbors (KNN) algorithm was tested with varying values of k to examine its sensitivity to the number of neighbors considered for classification.

Naive Bayes, with its assumption of feature independence, was also used to model the classification problem. The performance of this model was evaluated in terms of accuracy, precision, recall, F1 score, and RMSE. Finally, Linear Regression was applied, although treating wine quality as a binary classification problem by mapping continuous quality values into two groups, thus enabling a comparative analysis with the other models.

Each of these models contributed to understanding the complexities of predicting wine quality, and their results were analyzed to determine the most effective approach for this classification task.

Results

The classification models applied to predict wine quality showed varied performance.

Logistic Regression achieved an accuracy of 71.68%, with precision and recall of 73.53% and 68.09%, respectively. Its confusion matrix showed 192 true negatives and 175 true positives, with an AUC of 0.80 and an RMSE of 0.53.

Decision Tree had an accuracy of 65.63%, with a precision of 68.00% and recall of 59.53%. Its confusion matrix indicated 183 true negatives and 153 true positives. The AUC was 0.68, and RMSE was 0.59, showing moderate performance.

K-Nearest Neighbors (KNN) had an accuracy of 62.50%, precision of 62.75%, and recall of 62.26%. The confusion matrix revealed 160 true negatives and 160 true positives, with an AUC of 0.69 and RMSE of 0.61.

Naive Bayes models showed differing results. **GaussianNB** performed the best with an accuracy of 73.63%, precision of 73.64%, and recall of 73.93%. **MultinomialNB** and **BernoulliNB** had lower performance with accuracies of 59.77% and 50.20%, respectively.

Random Forest outperformed all other models with an accuracy of 79.30%, precision of 80.08%, and recall of 78.21%. Its confusion matrix showed 205 true negatives and 201 true positives, with an RMSE of 0.46.

Basic Linear showed an accuracy of 87.50%, but its inability to classify high-quality wine led to poor precision, recall, and F1 scores.

Conclusion

This project demonstrated the effectiveness of machine learning in predicting wine quality using measurable physicochemical properties. By training various models, including Logistic Regression, Decision Tree, K-Nearest Neighbors, Naive Bayes, and Random Forest, we were able to classify wines into quality categories with varying degrees of success. The Random Forest model outperformed others in terms of accuracy, precision, and recall, highlighting its potential for practical use in the wine industry.

While machine learning provides a more objective and reproducible approach to quality assessment, challenges such as imbalanced data and performance optimization remain. Future work could explore incorporating additional features or applying advanced techniques to further improve prediction accuracy. Overall, this study illustrates the promise of data-driven methods in transforming wine quality evaluation, offering more reliable insights for winemakers, retailers, and consumers.

References

1. K. R. Dahal1, J. N. Dahal, H. Banjade, S. Gaire, *Prediction of Wine Quality Using Machine Learning Algorithms* , Scientific Research Publishing, 2021. DOI: Link to Paper

2. Muhammad Arief Rachman, *Wine Quality Prediction with Machine Learning Model*, Medium, 2023. Available at: Link to Medium Post

3. Yuliia Kniazieva, *How to Build a Wine Quality Prediction Model Using Machine Learning?*, Label Your Data, 2023. Available at: Link to Post

Github Repository