Predictive Analysis of Wine Quality Using Machine Learning Models

ESE 417 Final Project

Sherlock Zhang, Ice Cui, Beining Wang

**1.Introduction**

Machine learning has revolutionized the field of predictive modeling, enabling data-driven insights across a variety of domains. In this study, we aim to leverage machine learning techniques to predict the quality of red wine based on its physicochemical properties. The dataset, sourced from the UCI Machine Learning Repository, includes 11 input attributes such as fixed acidity, volatile acidity, and alcohol content, along with a quality score ranging from 0 to 10 as the target variable.

The primary objective of this project is to develop, implement, and compare multiple machine learning models to identify the most effective approach for predicting wine quality. Through an exploratory analysis of the dataset, including visualizing distributions and relationships between features, we refine our understanding of the data. Subsequently, we implement and evaluate three advanced machine learning algorithms: Support Vector Machine (SVM), Random Forest, and Artificial Neural Networks (ANN). These models are rigorously tested and fine-tuned to enhance their performance, utilizing methods such as hyperparameter optimization, feature normalization, and data cleaning. Comparative analysis of the results is performed using metrics such as accuracy, precision, and F1-score to determine the best-performing model.

The contributions of this project include a comprehensive evaluation of machine learning techniques for wine quality prediction, insights into data characteristics through exploratory analysis, and practical recommendations for model selection in similar predictive tasks. Through this systematic approach, we aim to showcase the efficacy of machine learning in addressing real-world classification challenges.

**2.Beginning and Exploratory Analysis**

**2.1 OLS Model and Observation**

Linear regression serves as a fundamental starting point in many machine learning tasks due to its simplicity and interpretability. In this project, we utilized a linear regression model as an initial benchmark to understand the relationship between the physicochemical attributes of wine and their quality scores. The results from the linear regression model were as follows: Mean Squared Error (MSE): 0.3900, R-squared (R²): 0.4032. The MSE indicates the average squared difference between the actual and predicted wine quality scores, suggesting a significant level of error. The R² value of 0.4032 implies that the linear model explains only about 40% of the variance in the target variable, leaving a substantial portion unexplained. These results highlight the limitations of a simple linear model in capturing the complex, potentially nonlinear relationships inherent in the data.

**2.2 Data Visualization**

To gain a deeper understanding of the dataset's structure, several visualization techniques were applied to examine the distributions of features, their correlations, and the overall data structure. The distribution plots(figure1) of all input features and the target variable (quality) reveal several key patterns. Most features, such as volatile acidity, chlorides, and sulphates, exhibit skewed distributions with long tails toward higher values. On the other hand, features like pH and density are closer to normal distributions, with well-defined central tendencies. The target variable, quality, shows a discrete distribution, with the majority of wines falling between the scores of 5 and 6, indicating a predominance of medium-quality wines. This imbalance suggests potential challenges in distinguishing between the quality levels and highlights the need for addressing class imbalance during model development.

A screenshot of a graph

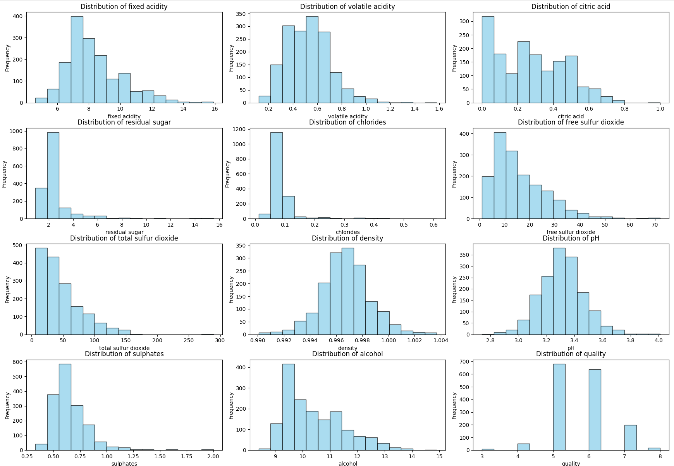
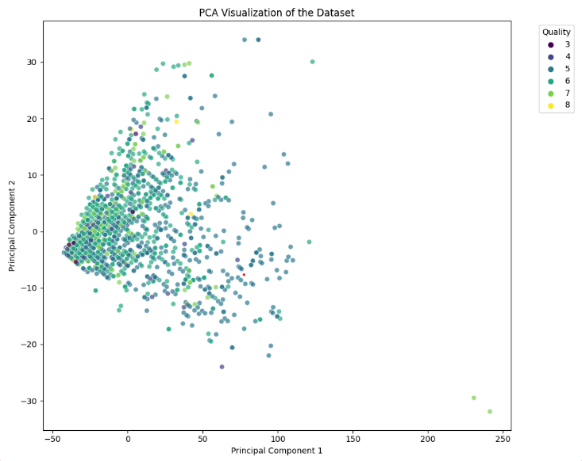
Description automatically generated

Figure 1 Bar Graph for Feature Distribution Figure 2, Heat-Map

Further exploration of the dataset was conducted using a correlation heatmap to assess relationships between features and the target variable. The heatmap(Figure2) reveals that alcohol is positively correlated with quality, suggesting that higher alcohol content is associated with better wine quality. Other features, such as citric acid and sulphates, also show moderate positive correlations with quality. Conversely, volatile acidity demonstrates a strong negative correlation with quality, indicating that higher levels of acidity detract from wine quality. Additionally, certain features, such as fixed acidity and density, as well as free sulfur dioxide and total sulfur dioxide, exhibit strong interdependencies, pointing to potential multicollinearity issues that might affect predictive modeling.

A bar graph with blue rectangular bars

Description automatically generated

Figure 3, PCA Visualization Figure 4. Number of Samples on Quality

Finally, principal component analysis (PCA) was applied to reduce the dataset’s dimensionality and visualize the data in a two-dimensional space(Figure 3). The PCA visualization shows that samples with different quality levels overlap significantly, particularly in the mid-quality range of 5 and 6. This overlap highlights the complexity of the dataset and the challenges associated with distinguishing between these classes using simple models. However, the few samples in the extreme quality ranges (e.g., 3 and 8) are more distinct, suggesting unique physicochemical properties for these wines. At the same time, the bar graph on the right (figure 4) further demonstrates the lack of samples on extreme quality values. This could be a potential “threat” when training the model because of the lack of training sample on these values.

**2.2 Data Processing**

Based on our observations, the dataset exhibits characteristics that necessitate preprocessing to improve model performance and generalizability. First, the distributions of many features, such as volatile acidity, chlorides, and sulphates, are highly skewed, with long tails. To address this, non-linear transformations like the natural logarithm were applied to compress the tails and create more symmetric distributions, reducing the influence of outliers. Correlation analysis further revealed strong dependencies between features such as fixed acidity and density, as well as free sulfur dioxide and total sulfur dioxide. These correlations pose risks of multicollinearity, which can undermine model stability. To mitigate this, Principal Component Analysis (PCA) was used to reduce dimensionality while preserving variance, resolving multicollinearity issues and improving computational efficiency.

Additionally, the dataset’s features vary significantly in range, with some values like volatile acidity being less than 1, while others, such as residual sugar, reaching up to 16. To address this disparity, normalization was applied, scaling all features to a [0, 1] range to ensure balanced contributions to distance-sensitive models like Neural Networks and Support Vector Machines. The preprocessed dataset, now including log-transformed, normalized features and PCA-reduced components, was saved for further modeling.

Finally, it was observed that the dataset has an imbalance in the distribution of quality levels. Extreme quality levels (e.g., 3 and 8) are significantly underrepresented. This imbalance limits the ability of multi-class classification models to accurately distinguish these rare classes, often leading to biased predictions. In future implementations, this limitation could be addressed by reformulating the problem into a binary classification task, such as distinguishing between "low quality" (e.g., 3, 4) and "high quality" (e.g., 7, 8) classes with a standardized criteria, or say threshold. This approach simplifies the task and potentially improves accuracy by grouping rare classes into broader categories, reducing the impact of the dataset imbalance.

**3.Methodology, Implementation, and Result Analysis**

**3.1.Artificial Neuron Network**

Artificial Neural Networks (ANNs) are a class of machine learning models inspired by biological neural networks. They are capable of modeling complex, non-linear relationships in data and have been widely used in classification and regression problems. In this project, ANNs were employed to predict wine quality, and various strategies were implemented to optimize their performance, including hyperparameter tuning, re-sampling, and architectural improvements.

**3.1.1 Hyperparameter Tuning with GridSearchCV**

To identify the optimal ANN configuration, a grid search approach was used with a comprehensive hyperparameter space that included different hidden layer sizes, activation functions, solvers, regularization values, learning rate schemes, and initial learning rates. GridSearchCV with 5-fold cross-validation was applied on the balanced training dataset to ensure robust evaluation. The search identified the best configuration as having two hidden layers with 100 neurons each, the ReLU activation function, the Adam solver, a regularization parameter (alpha) of 0.01, an adaptive learning rate scheme, and an initial learning rate of 0.001. This configuration was subsequently used in the experiments, ensuring efficient training.

|  |  |
| --- | --- |
| **Hyperparameter** | **Best Value** |
| Hidden Layer Sizes | |  |  | | --- | --- | |  | (50, 50) | |
| Activation Function | ReLU |
| Solver | Adam |
| Alpha (Regularization) | |  |  | | --- | --- | |  | 0.0001 | |
| Learning Rate | Constant |
| Initial Learning Rate | 0.001 |

**3.1.2 ANN on Unmodified Dataset**

Initially, the ANN model was trained and tested on the unmodified dataset. This served as a baseline for evaluating the impact of data preprocessing and re-sampling strategies. The results indicated high accuracy for the majority classes, specifically quality levels 5 and 6, while minority classes exhibited poor performance with low precision, recall, and F1-scores. The imbalance in wine quality labels led to biased predictions, highlighting the need for strategies to address class imbalance. The overall accuracy achieved was 56%, with significantly lower metrics for minority classes such as qualities 3 and 8.

**3.1.3 ANN and Re-sampling Strategy**

To address the class imbalance issue, a re-sampling strategy was employed where minority classes were oversampled to match the size of the majority class. This resulted in a balanced dataset and allowed the ANN model to achieve better generalization across all quality levels. Training the model on the balanced dataset showed significant improvements, with overall accuracy increasing to 64%. Metrics for minority classes also showed improvement, although the performance for extreme quality levels (e.g., qualities 3 and 8) remained low due to their inherent complexity and sparsity in the data. The re-sampling strategy effectively mitigated the bias introduced by class imbalance, enabling the model to provide more equitable predictions. However, the improvement also came with increased computational cost due to the larger dataset size and training iterations required for convergence. The results demonstrated that while re-sampling improves the macro-average metrics, further strategies are needed to handle extreme quality levels effectively.

**3.1.4 Challenge and Adjustment**

The low precision, recall, and F1-scores for extreme classes rendered the model inadequate for practical usage. Furthermore, the inherent subjectivity in defining precise quality scores exacerbates the challenge, as minor variations in wine characteristics can lead to adjacent quality labels.

Given these challenges, predicting exact quality scores is both statistically and practically unrealistic. Instead, reframing the problem into a binary classification task was deemed more appropriate, as it aligns with broader quality categorizations (e.g., low-quality vs. high-quality wine), ensuring a more balanced and robust model. To improve the model's validity, we redefined wine quality as a binary classification problem. Wines with quality ≤ 5 were labeled as "low-quality" (class 0), while those with quality > 5 were labeled as "high-quality" (class 1). This binary categorization addresses the sparsity of extreme quality labels by aggregating them into broader categories, simplifying the model's task while still providing actionable insights for decision-making.

Taken all improvement strategies we have implemented, we end up with result as listed below:

A screenshot of a graph

Description automatically generated

(把这张图改成word里的table)

The transition to binary classification significantly improved the model's reliability and interpretability by addressing the challenges of class imbalance and label sparsity. This time, data resampling seems to be a bad strategy for classification problem. But still, an accuracy of 73% is not statistically significant for a prediction model, which leads to further model implementation other than ANN for better prediction improvements.

**3.2 Support Vector Machine**

Support Vector Machines (SVMs) are supervised learning models widely used for classification and regression problems. They operate by finding the optimal hyperplane that maximizes the margin between classes in the feature space, making them particularly effective for high-dimensional and linearly separable data. In this project, SVMs were employed to predict wine quality, both as a score prediction and as a binary classification task.

**3.2.1 SVM on Unmodified Dataset**

Initially, the SVM model was trained and tested on the unmodified dataset to serve as a baseline. The base model showed an accuracy of 51% on the unmodified data, with low performance across minority classes. The results highlighted the limitations of the dataset's inherent imbalance, where the majority classes, such as qualities 5 and 6, dominated the predictions. Minor quality levels, such as 3 and 8, suffered from very low precision and recall, leading to biased outcomes. These observations underscored the necessity of employing data preprocessing techniques and hyperparameter optimization.

**3.2.2 Optimization Strategies and Implementation**

To improve the model's performance, several preprocessing strategies were implemented. First, normalization was applied using MinMaxScaler, which improved the model's accuracy to 59.69%. Normalization scales features to a uniform range, reducing the influence of features with larger magnitudes. Next, standardization was employed using StandardScaler, further improving the accuracy to 60.31%. Standardization normalized features to zero mean and unit variance, which is crucial for SVMs as they are sensitive to feature scaling. Additionally, dimensionality reduction using Principal Component Analysis (PCA) was performed to reduce the dataset to eight principal components. PCA captured most of the variance while addressing potential multicollinearity, yielding an accuracy of 58.75%, which, while slightly below standardized preprocessing, still marked an improvement from the base model.

To further optimize the SVM configuration, hyperparameter tuning was performed using GridSearchCV. A grid search was applied over a parameter space including different kernels (linear, rbf, and polynomial), regularization values (C), and kernel coefficients (gamma). The search identified the optimal parameters as C=10, gamma='scale', and kernel='rbf'. The optimized model achieved an accuracy of 60.94%, the highest among all preprocessing steps implemented individually.

To address the issue of class imbalance, the dataset was balanced using oversampling. Minority classes were upsampled to match the majority class, creating a more balanced representation. The balanced dataset was then standardized and used for training. This approach resulted in the best accuracy of 68.13%, demonstrating the importance of addressing data imbalance for improving model generalization.

**3.2.3 Optimized SVM on Classification Problem**

Given the limitations of predicting exact quality scores, the task was reframed as a binary classification problem. Wines with quality scores ≤ 5 were labeled as "low-quality" (class 0), and those with scores > 5 as "high-quality" (class 1). This binary classification setup simplified the model's task and mitigated issues of extreme sparsity in minority classes. For the binary classification problem, the base model achieved an accuracy of 63.75%. With normalization, the accuracy increased to 75.00%, and with standardization, the accuracy further improved to 77.19%. Using PCA resulted in an accuracy of 74.69%. Hyperparameter tuning with GridSearchCV achieved an accuracy of 76.25%, and training on the balanced dataset yielded an accuracy of 73.68%. The highest accuracy of 77.19% was achieved using a standardized dataset without additional resampling, reflecting the efficiency of scaling features for binary SVM tasks. The binary classification approach provided a more interpretable model, addressing key challenges of class imbalance and label sparsity.

**3.3 Random Forest**

**3.3 Random Forest**

Random Forest is a versatile machine learning algorithm capable of both classification and regression tasks. It builds multiple decision trees during training and outputs either the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Random Forest is particularly effective in handling overfitting and works well with datasets containing complex feature relationships and varying distributions. For this project, Random Forest was employed to predict wine quality, transitioning from score prediction to binary classification to address inherent data challenges.

**3.3.1 Hyperparameter Tuning with GridSearchCV**

To optimize the Random Forest model, we used GridSearchCV to explore a range of hyperparameters, including the number of trees (n\_estimators), maximum tree depth (max\_depth), minimum samples required to split a node (min\_samples\_split), minimum samples required at a leaf node (min\_samples\_leaf), and whether or not to use bootstrap sampling (bootstrap). The optimization process was applied using a standardized version of the dataset to enhance comparability and consistency. The best configuration identified included 300 trees, a maximum depth of 30, minimum samples per leaf set to 2, and the model trained without bootstrap sampling. This optimized configuration achieved high performance when applied to the binary classification problem.

**3.3.2 Random Forest on Unmodified Dataset**

The Random Forest model was initially evaluated on the raw dataset to establish a baseline performance. The base model achieved an accuracy of 65.94%, indicating its robustness in handling the unprocessed data. However, the performance showed room for improvement, particularly for minority class predictions. This reinforced the need for data preprocessing and hyperparameter tuning to enhance the model's predictive power.

**3.3.3 Random Forest and Data Preprocessing**

To improve performance, several preprocessing steps were evaluated, including normalization, standardization, and dimensionality reduction using PCA. Normalization slightly improved accuracy to 66.25%, and standardization yielded a comparable accuracy of 65.94%. PCA, while reducing dimensionality to eight components, resulted in a marginal decrease in accuracy (65.63%), highlighting that the original feature space better preserved the necessary information for prediction.

**3.3.4 Binary Classification Problem**

Recognizing the limitations of score prediction, we redefined the wine quality prediction task as a binary classification problem, categorizing wines with quality ≤ 5 as low quality (class 0) and those with quality > 5 as high quality (class 1). The base binary classification model achieved an accuracy of 79.06%, with normalization and standardization yielding similar performances (78.75% and 78.43%, respectively). Using PCA improved accuracy slightly to 79.37%, while GridSearch optimization maintained performance at 79.06%. The most notable improvement was achieved through balancing the dataset, where oversampling the minority class resulted in an accuracy of 88.01%, highlighting the significance of addressing class imbalance for binary classification tasks.

3.3.5 Summary of Results

The Random Forest model demonstrated strong performance for both score prediction and binary classification tasks. Binary classification consistently achieved higher accuracy, underscoring the benefits of reframing the problem to address label sparsity and imbalance. Balancing the dataset proved particularly effective, significantly boosting the model's performance in binary setting. A screenshot of a computer

Description automatically generated(同样，改成word 里的table)

**4.Conclusion**