Technical Report

Machine Learning-Based Classification of Wine Quality Using Physicochemical Properties: A Comprehensive Analysis and Predictive Modeling Approach

Date: June 24th 2024

Authors
Dheemanth Rajakumar
Yaakov Sternberg
Samnang So

Introduction

The objective of this project is to analyze a combined dataset of red and white wine quality, sourced from the UCI Machine Learning Repository. The dataset includes various physicochemical properties of the wines and their quality ratings. The goal is to build a predictive model that can classify the quality of wine based on these properties. This report details the steps taken in data cleaning, exploratory data analysis, model selection, and model evaluation, culminating in conclusions and recommendations. The dataset contains 4,898 instances, each with 11 features representing various physicochemical properties, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, and alcohol content. Understanding the factors that influence wine quality is crucial for wine producers to optimize their production processes and ensure consistent quality standards. By developing a predictive model that can accurately estimate wine quality based on the given physicochemical features, we can provide valuable insights to the wine industry and potentially aid in improving customer satisfaction.

Code Segments

The provided code segment https://github.com/Root18D/AAI-500-Stats-Project/blob/main/Final Project.ipynb is part of a larger machine learning pipeline focused on evaluating multiple classification models for predicting wine quality. The models considered include RandomForest, GradientBoosting, SVC (Support Vector Classifier), and LogisticRegression. The code performs model fitting, prediction, evaluation, and visualization of feature importance and confusion matrices for the best-performing model.

Model Analysis

The `best_models` dictionary stores instances of the best-performing models after hyperparameter tuning. These models are:

- RandomForest (`best_rf`)
- GradientBoosting (`best_gb`)
- SVC (`models['SVC']`)
- LogisticRegression (`models['LogisticRegression']`)

Each model is fitted with the training data (`X_train`, `y_train`) and then used to predict the outcomes (`y_pred`) on the test dataset (`X_test`). The accuracy of each model is printed alongside a detailed classification report that includes precision, recall, f1-score, and support for each class.

Feature Importance Visualization

For the RandomForest model (`best_rf`), the feature importances are extracted and sorted in descending order. A DataFrame `importance_df` is created with two columns: `Feature` and `Importance`, where `Feature` corresponds to the names of the features and `Importance` to their importance scores. A bar plot visualizes these importances, highlighting the most significant features in predicting wine quality.

Confusion Matrix Visualization

The best-performing model is identified by comparing the accuracy scores of all models on the test dataset. The name of this model is stored in `best_model_name`, and the model itself in `best_model`. The predictions of the best model (`y_pred_best`) are used to compute a confusion matrix, which is then visualized using a heatmap. This matrix provides insights into the model's performance, showing the true positives, true negatives, false positives, and false negatives.

Conclusion

The final part of the code prints the name of the best model along with its accuracy on the test dataset. This information is crucial for understanding which model performs best under the given circumstances and can guide further model optimization and deployment decisions.

Recommendations

Based on the analysis, it is recommended to:

- Further investigate the features with the highest importance scores to understand their impact on wine quality.
- Explore model-specific hyperparameter tuning to enhance the performance of the identified best model.
- Consider additional model evaluation metrics such as ROC-AUC score for a more comprehensive assessment.
- Experiment with advanced ensemble techniques that might leverage the strengths of multiple models.

This technical report summarizes the process and findings from the model analysis and evaluation code segment, providing insights into the performance of different classifiers in predicting wine quality.

End of Technical Report