Machine Learning-Based Quality Prediction for Red Wine Using Physicochemical Properties

Submitted For

Statistical Machine Learning CSET-211

Submitted By:

Harsh Bhardwaj (E23CSEU0427)

Submitted To:

Dr. Susmita Das (Assistant Professor)

July-Dec 2024

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

BENNETT UNIVERSITY
THE TIMES GROUP

Candidate’s Declaration:

I hereby declare that the project report entitled "Machine Learning-Based Quality Prediction for Red Wine Using Physicochemical Properties" submitted for partial fulfillment of the requirements for the degree/diploma of [Degree/Program Name] is my original work. This work has been carried out under the guidance of Dr. Susmita Das, Assistant Professor.

I affirm that this report has not been submitted for any other degree or diploma at any other university or institution. All sources of information and references used in this project are duly acknowledged.

I understand that any act of plagiarism, if detected, will result in disqualification and disciplinary action as per the regulations.

Date: 17th November, 2024
Candidate’s Name: Harsh Bhardwaj
Roll Number/ID: E23CSEU0427

Acknowledgment:

I would like to express my sincere gratitude to Dr. Susmita Das, Assistant Professor, for her invaluable guidance, encouragement, and support throughout the duration of this project. Her expertise and insights have been instrumental in shaping the direction of my work and enhancing my understanding of machine learning applications in wine quality prediction. I am deeply thankful for her constructive feedback, which has helped me improve and refine my project. This project would not have been possible without her mentorship, and I am grateful for the opportunity to work under her supervision.

Project Report Index

1. **Abstract**
   Brief overview of the project objectives, methodology, and key results.
2. **Introduction**
   - Background on wine quality prediction.
   - Importance of physicochemical factors in determining wine quality.
   - Project objectives and goals.
3. **Related Work**
   Overview of previous studies or projects on wine quality prediction (if applicable), and comparison with your approach.
4. **Methodology**
   - **Data Collection**: Description of the dataset used and its features.
   - **Data Pre-Processing**: Handling missing values, scaling, and SMOTE for class balance.
   - **Model Selection**: Overview of machine learning models (Logistic Regression, SVM, Random Forest).
   - **Model Training and Tuning**: Cross-validation, hyperparameter tuning with GridSearchCV.
   - **Feature Importance**: Discussion on the significance of each feature (for Random Forest).
5. **Hardware and Software Requirements**
   - Hardware specifications for model training and testing.
   - List of software libraries and frameworks (Python, Flask, SCIKit-Learn, etc.).
6. **Experimental Results**
   - **Model Evaluation**: Accuracy, classification report, and confusion matrix for each model.
   - **Comparison of Models**: Cross-validation scores and test accuracies.
   - **Best Model Selection**: Final choice of model based on performance.
7. **Conclusion**
   Summary of findings, effectiveness of the selected model, and its potential application in real-world wine quality assessment.
8. **Future Scope**
   Suggestions for improving model accuracy, expanding dataset, incorporating additional features, or deploying the model in different domains.
9. **GitHub Repository**
   Link To The Complete Project Code: GitHub Repository

1. Abstract

This project focuses on predicting the quality of red wine using various machine learning algorithms. By leveraging features like acidity, residual sugar, pH, and alcohol levels, we aim to classify wine quality into specific categories. This Flask-based web application integrates an optimized Support Vector Machine (SVC) model, allowing users to input wine characteristics and receive quality predictions. The project emphasizes data preprocessing, hyperparameter tuning, and visualization to enhance model accuracy.

2. Introduction

Wine quality is influenced by numerous chemical properties, making it a complex task to determine. This project addresses the challenge by developing a machine learning model that predicts red wine quality based on physicochemical properties. The model allows wineries, sommeliers, and enthusiasts to assess wine quality quickly, aiding decision-making and improving product standards. We trained the model on the UCI Wine Quality dataset, aiming to make an accurate prediction of wine quality scores.

3. Related Work

Numerous studies have focused on wine quality prediction using machine learning algorithms like Decision Trees, k-Nearest Neighbors, and Support Vector Machines. In this project, we compare models like Logistic Regression, Random Forest, and SVC to find the best-performing model, using cross-validation for reliable results and SMOTE for addressing class imbalance. Through a Flask web interface, we also deploy our model for practical usage.

4. Methodology

Data Collection

The dataset used for this project is sourced from the UCI Machine Learning Repository, which includes 12 features related to wine characteristics and a target feature, *Quality*.

Data Pre-Processing

1. **Splitting and Balancing**: The data was split into training and testing sets, with 20% reserved for testing. SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the class distribution of the target variable.
2. **Scaling**: A Standard Scaler was applied to normalize the features, which is particularly essential for algorithms sensitive to feature scaling, such as SVC.

Model Development

Multiple machine learning models were tested:

- **Logistic Regression**: A baseline classifier with simplicity but limitations for nonlinear relationships.
- **Support Vector Classifier (SVC)**: Chosen for its ability to manage high-dimensional data effectively. Hyperparameter tuning was conducted with a grid search to find the best settings.
- **Random Forest Classifier**: Known for robustness in handling structured data. The grid search was used for tuning tree depth, number of estimators, and minimum sample splits.

Model Evaluation

Accuracy, precision, recall, and confusion matrices were utilized to evaluate model performance. Cross-validation was employed to ensure reliability across folds, and the SVC model was selected based on its superior performance.

Deployment

A Flask web application was built to allow users to input wine characteristics and get a quality prediction. Dark and light themes were integrated, and a user-friendly interface was created with HTML, CSS, and JavaScript.

5. Hardware/Software Required

Software

- **Python 3.x**: For scripting and development.
- **Flask**: For creating the web application.
- **Libraries**:
    - **SCIKit-Learn**: For machine learning model development.
    - **Imbalanced-Learn (SMOTE)**: For handling class imbalance.
    - **Matplotlib and Seaborn**: For visualizations.
    - **Joblib**: For saving the trained model.

Hardware

- A standard PC or laptop with at least 8GB RAM and a dual-core processor is sufficient to run the model and application.

6. Experimental Results

The results indicate that the Support Vector Classifier (SVC) model performed best:

- **Cross-Validation Score**: The SVC achieved a cross-validation accuracy score of approximately 0.80 across 5 folds.
- **Accuracy on Test Set**: SVC achieved a test accuracy of around 0.78.
- **Confusion Matrix**: Visualized confusion matrices highlighted the model's strengths in distinguishing between certain quality levels.
- **Feature Importance**: Random Forest feature importance highlighted the significance of variables like *alcohol*, *density*, and *volatile acidity*.

Results demonstrated the effectiveness of SVC in handling this classification problem, with its hyperparameter-tuned version being saved for deployment.

7. Conclusions

This project successfully implemented a wine quality prediction model using machine learning techniques. The SVC model, after tuning, provided the best performance among tested models. The Flask web application allows real-time predictions, offering potential utility in the wine industry for preliminary quality assessments based on physicochemical properties.

8. Future Scope

1. **Incorporating More Features**: Additional sensory features could improve prediction accuracy.
2. **Expanding to White Wine**: A separate model could be developed for white wine, expanding the application's scope.
3. **Further Hyperparameter Tuning**: More complex tuning could improve model accuracy.
4. **Real-time Data Integration**: Incorporating real-time sensor data from wine production processes could enable dynamic predictions.

9. GitHub Link

For The Complete Project, Including Code, Data, And Model Files, Please Visit The GitHub Repository: [Red Wine Quality Project]