Assignment Report

Abhinav Kumar

EE 769

Wine Quality Prediction

1. Introduction

This project aims to explore and predict wine quality using different regression models. The main objectives are to understand the importance of different variables in predicting wine quality, test model performance across different types of wine, and assess the generalizability of models to out-of-distribution data.

2. Dataset Overview

The datasets used in this project are the **Wine Quality** datasets from the UCI Machine Learning Repository. The datasets include chemical properties and quality ratings for red and white wines.

- **Source:** Wine Quality Dataset
- **Attributes:**
  - **Input Variables:** Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol.
  - **Output Variable:** Quality (score between 0 and 10)

3. Data Exploration and Pre-processing

3.1 Data Exploration

- **Visualization:** Explored the distribution of each feature and quality scores using histograms, box plots, and scatter plots.
- **Correlation Analysis:** Analyzed the correlation between input variables and the quality score using a heatmap.

3.2 Data Pre-processing

- **Handling Missing Values:** Checked for and addressed any missing values.
- **Normalization/Standardization:** Standardized the input features to have a mean of 0 and a standard deviation of 1.
- **Train-Test Split:** Split the data into training, validation, and testing sets.

4. Model Training and Evaluation

4.1 Models Used

- **Random Forest:** A versatile ensemble model that uses multiple decision trees.
- **Support Vector Regression (SVR) with RBF Kernel:** A regression model that uses a non-linear kernel to capture complex relationships.

4.2 Hyperparameter Tuning

- **Random Forest:** Tuned the number of trees and the maximum depth.

- **SVR with RBF Kernel:** Tuned the regularization parameter (C) and gamma.

4.3 Model Evaluation

- **Validation Strategy:** Used cross-validation to optimize model hyperparameters.
- **Performance Metrics:** Evaluated models using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) on both training and test datasets.

5. Feature Importance Analysis

- **Methodology:** Used feature importance scores from the Random Forest model and coefficients from the SVR model to determine the significance of each variable.
- **Observations:** Commented on whether the important variables were consistent across different models and whether certain features were more critical for specific models.

6. Out-of-Distribution Testing

6.1 Cross-Testing Models

- **Red Wine Model on White Wine Data:** Tested the model trained on red wine data using white wine data and vice versa.
- **Generalizability Analysis:** Analyzed the performance drop when applying a model trained on one wine type to another and discussed potential reasons for this behavior.

6.2 Observations

- Commented on whether the models trained on red wines are applicable to white wines and vice versa. Discussed the implications of the performance differences observed.

7. Conclusion

Summarized the findings from the model training, feature importance analysis, and out-of-distribution testing. Provided insights into which models performed best, the transferability of models across wine types, and the importance of different variables in predicting wine quality.

8. References

1. Prof. Amit Sethi, Electrical Engineering, IIT Bombay.

2. Scikit-learn Documentation: https://scikit-learn.org/stable/documentation. html

3. Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Springer.