Exercise 3: ML-Algorithm Selection

Bimal Pandey

Introduction: This report introduces the different machine learning algorithms to predict the quality of Red Wine based on different features of wine such as acidity, sulfates, pH values. This uses the predictive performance of several regression models to determine which can most accurately predict wine quality.

Dataset Description: The Red Wine dataset consists of 1599 observations and 12 characteristics, out of which 11 are input variables and the remaining one is output variable. Here, the data have only float and integer values (only for the target variable) and there are no null/missing values. The describe() function returns the count, mean, standard deviation, minimum, 25%, 50%, 75%, and maximum values and the qualities of data.

Input Variables:

• fixed Acidity

• volatile acidity

• citric acid

• residual sugar

• chlorides

• free sulfur dioxide

• total sulfur dioxide

• density

• pH

• sulphates

• alcohol

Output variable:

• quality

Experimental setup and Results: To start with, I imported the following libraries and loaded the dataset, and the original data are separated by ";" in the given data set.

• Numpy: It will provide support for efficient numerical computation

• Pandas: It is a convenient library that supports data frames. Working with pandas will bring ease to many crucial data operations.

• Seaborn: It is a visualization library based on Matplotlib which provides a high-level interface for drawing attractive statistical graphics.

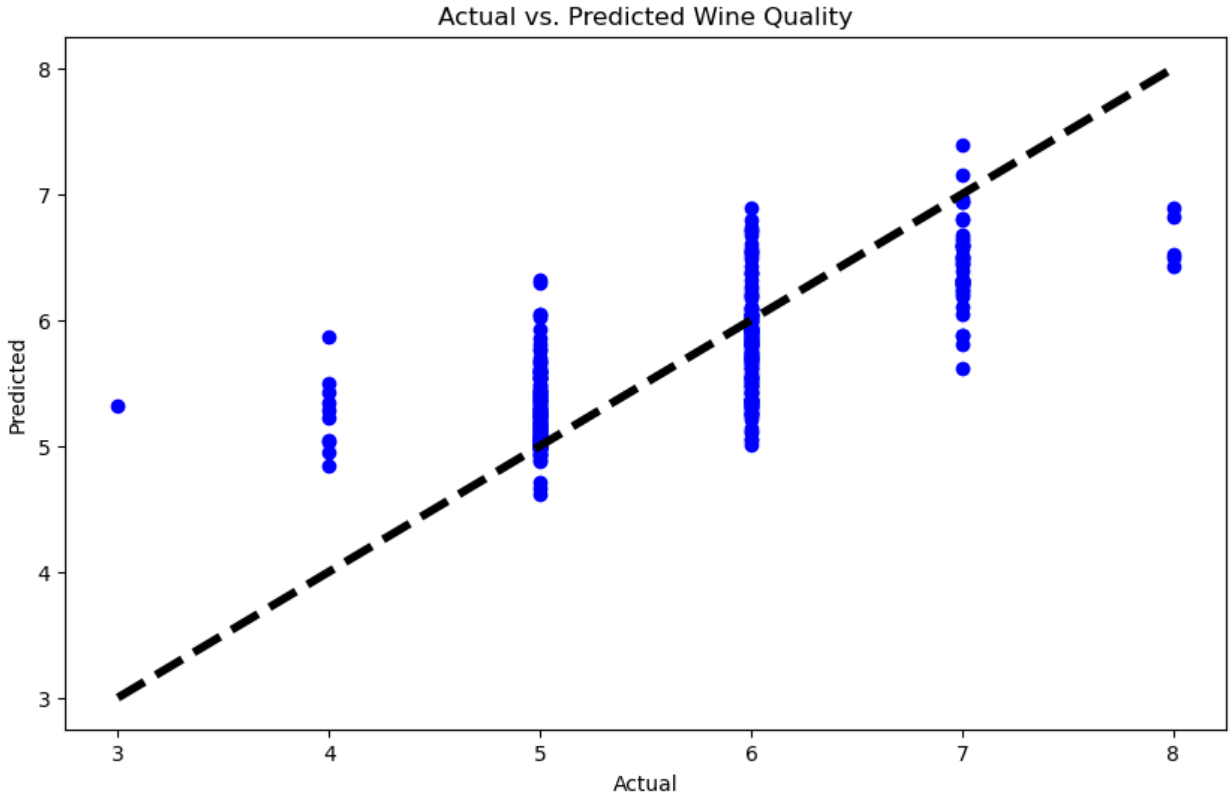
• Sklearn: It is a Python library for data mining, data analysis, and machine learning.

• Matplotlib: It provides a MATLAB-like plotting framework

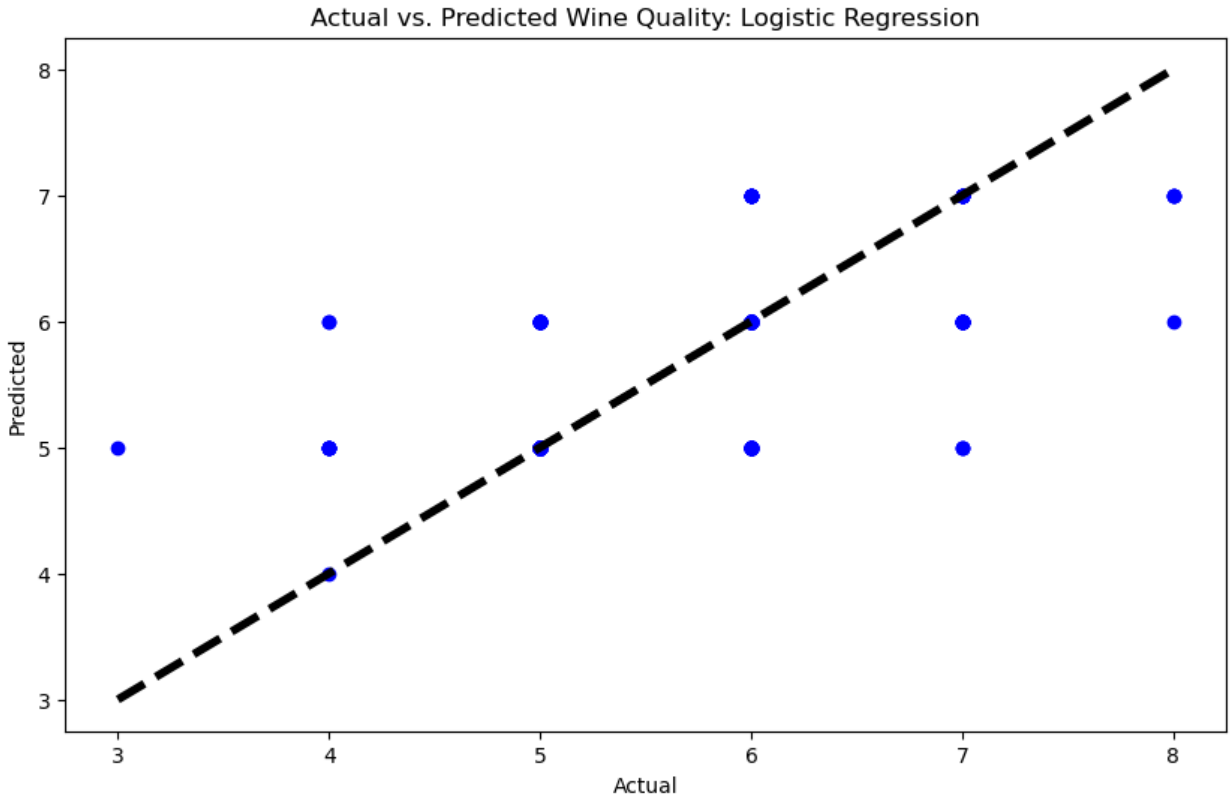
Data Preprocessing: The datasets are divided into training and test sets with 80% training data and the remaining 20% as test data. After dividing the data, 'StandardScaler' from Scikit-learn is used to normalize the features.

The study includes five different machine-learning algorithms for the prediction of wine quality:

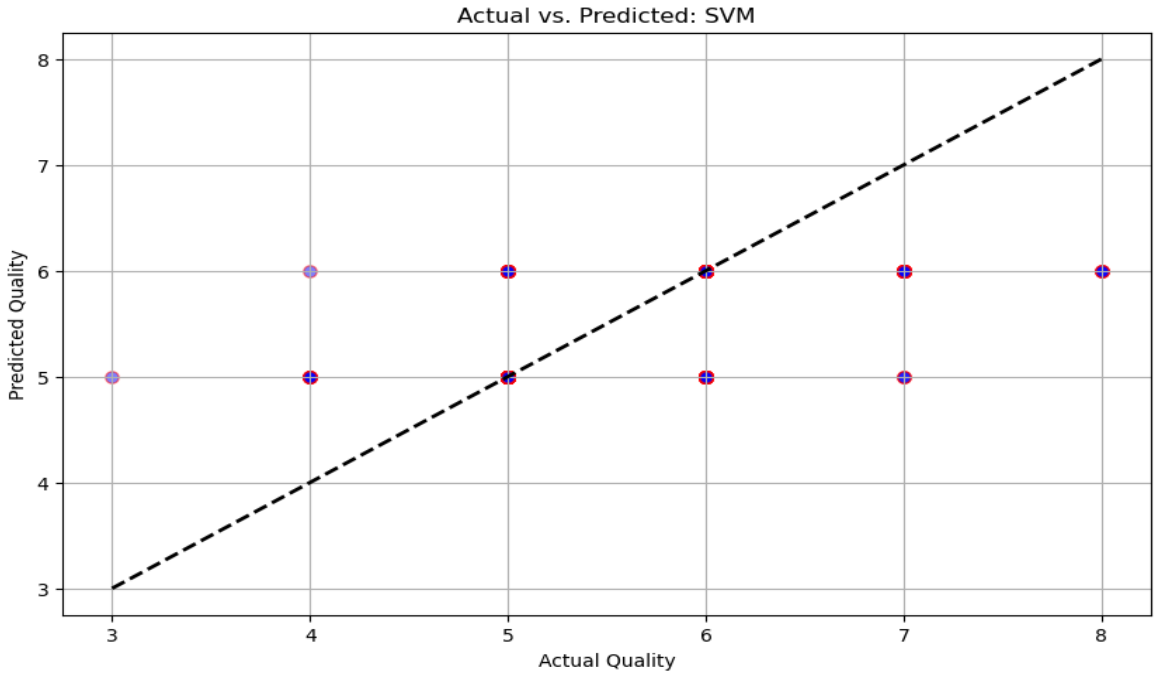
1) ***Random Forest Regression***: In this approach 'RandomForestRegressor' is imported from 'sklearn.ensemble' and MSE and R2 score as evaluated as 0.31 and 0.51 respectively and cross-validation for MSE as 0.42.



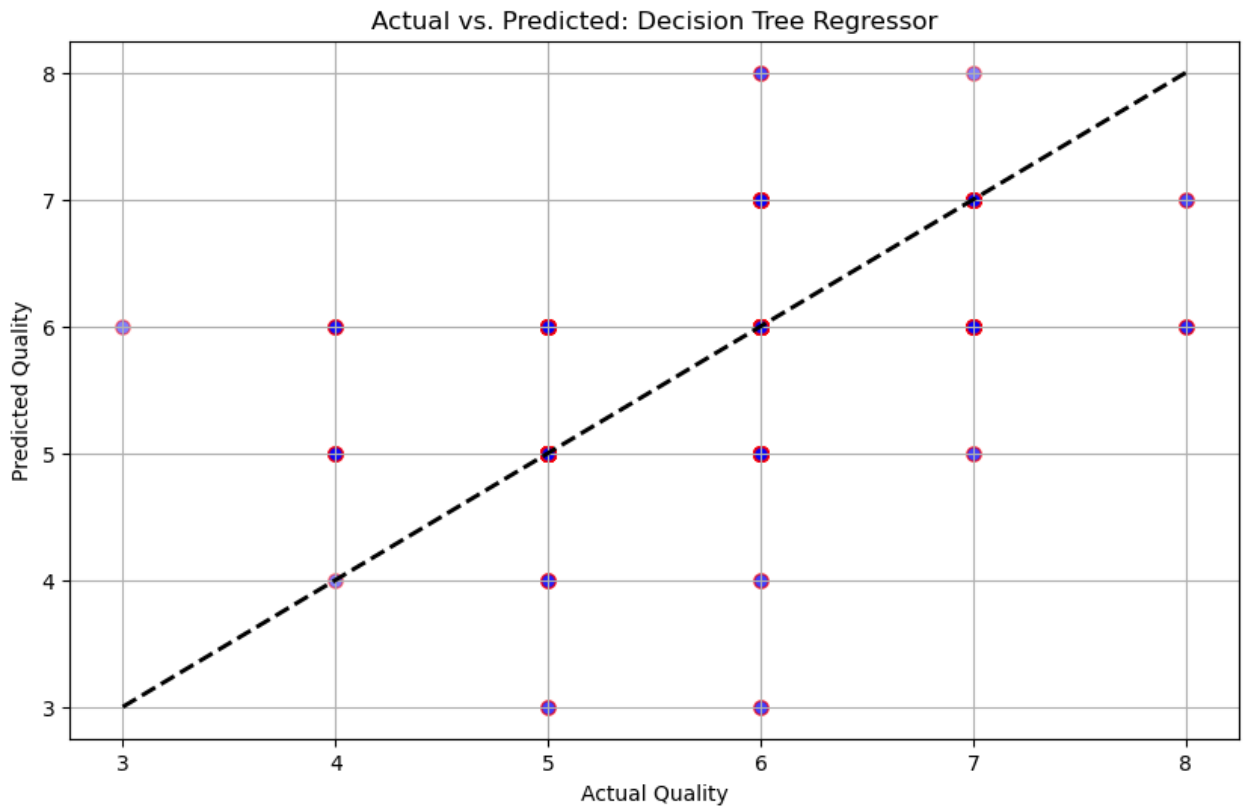
2) ***Logistic Regression:*** Logistic regression model imports LogisticRegression(random state=42, max_iter=10000) and mse and r2 score are obtained as 0.49 and 0.29. Similarly, the cross validation of mse as 0.53.



3. ***Support Vector Machine:*** SVM is the versatile supervised machine learning algorithm. SVM maps the data into high-dimensional feature space for the data point categorization even if the data are not linearly separable. SVC is imported with kernel='linear' and random_state=42 and mse and r2 score are obtained as 0.53 and 0.18 respectively with cross-validation 0.53.

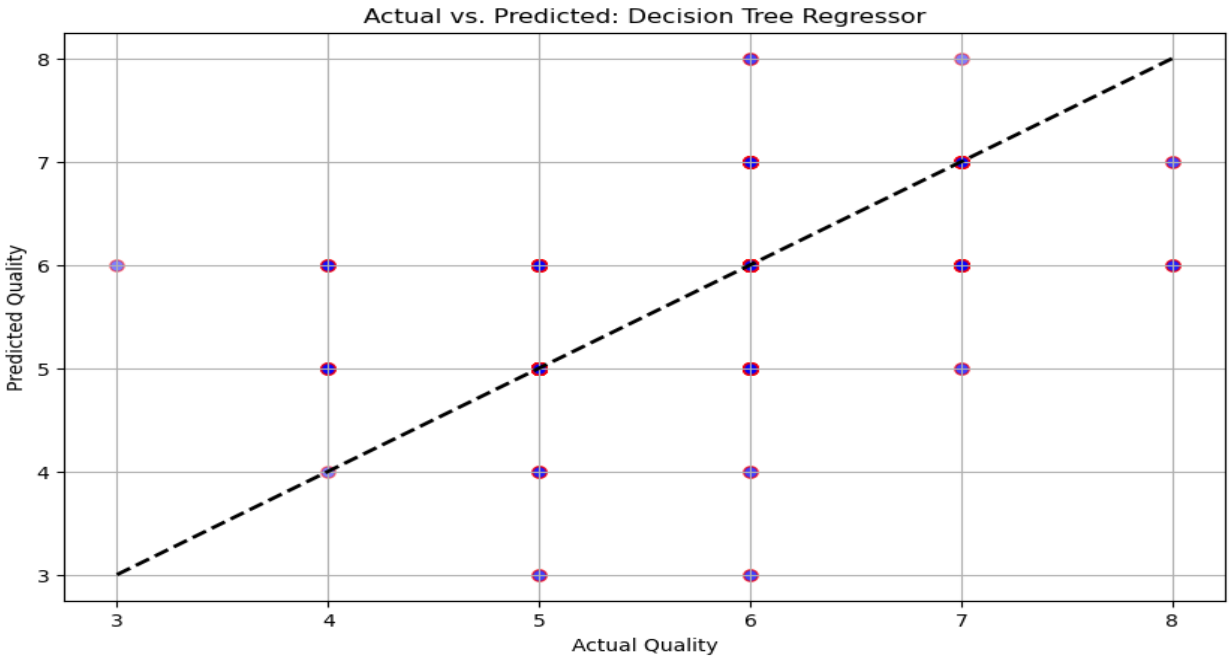


4. ***Decision Tree Regressor*:** DecisionTreeRegressor is imported and fit in the trained data. Mse and r2 score are obtained as 0.61 and 0.062 with cross-validation 0.67.



5. Gradient Boosting: GradientBoostingRegressor is imported with (n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42). The mse and r2 score are obtained as 0.36 and 0.062 with cross-validation for mse as 0.4.

Discussion and Conclusion*:* Random Forest and Gradient Boosting showed superior performance in terms of MSE, suggesting robustness in handling the dataset's nonlinear characteristics. The decision tree exhibits the tendency to overfit, evidenced by its lower performance. SVM showed moderate performance and Logistic Regression was the least effective for this data set.


Actual vs. Predicted: Decision Tree Regressor
Predicted Quality
Actual Quality
3
4
5
6
7
8