**Predicting Wine Quality: Analysis Using Different Alogrithms**

## Introduction

This assignment is focused on predicting the quality of wine in R programming using different machine learning models like Linear regression, Support vector regression (SVR), Polynomial regression, Decision tree regression, Random forest regression. This report covers data exploration and data preprocessing, model development, visualization, and model evaluation. In this custom dataset, we are taking wine quality as the dependent variables and the rest of variables; type, fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, and alcohol are the independent variables in our dataset.

## Analysis Method

My analysis employs a systematic approach to derive insights from the dataset, emphasizing predictive modeling for wine quality. The analysis process consists of the key machine learning steps; data preprocessing, building the model, validation, and evaluating the performance of the model.

As a part of data preprocessing, I am taking the data set and splitting it with 80:20 split ratio where 80% of the dataset goes into the training set and 20% of the dataset goes into the testing set. The NA values are replaced with the mean values. I am also scaling the dataset so that the dataset contributes equally to maintaining data modeling accuracy. After the data is split and scaled, I am fitting the different machine learning models and also calculating the MSE and RMSE to find out the best model among all for wine quality prediction. Beside this, I am also plotting the scatter plot of Actual Wine Quality vs Predicted Wine Quality to demonstrate the performance of the different machine learning models.

## Data Preprocessing:

Data preprocessing is the crucial step in building a model,  and is a must skill while working as data analysts. In the data preprocessing phase, several essential steps were undertaken to ensure the dataset's readiness for modeling and analysis. This critical stage aimed to enhance data quality and usability.

The Key actions that were taken include:

- **Categorical Data Encoding:** The non-numeric attributes ‘type’ that had values as ‘white’ and ‘red’ were transformed into a numerical format 1 and 2 for effective model integration using the factor() function.
- **Missing Value Handling:** Missing data was replaced with the mean values of respective attributes to maintain data integrity.
- **Data Splitting:** The dataset was divided into training (80%) and validation (20%) subsets to assess model generalization and splitting was done using a sample.split() function.

- **Data Scaling:** Independent variables were scaled using a scale() function to ensure equal feature contribution and enhance model performance.

These data preprocessing steps collectively played a crucial role in preparing the dataset for modeling. By addressing categorical data, handling missing values, partitioning the data, and scaling features, I laid the foundation for robust and accurate predictive modeling, ultimately aiding in the extraction of valuable insights from the data.

**Model Building:**
After the data was formatted and cleaned, I started building the model. In the modeling phase, five distinct regression models were developed to predict wine quality, treating it as the dependent variable. Our independent variables included various attributes such as 'type,' 'fixed acidity,' 'volatile acidity,' 'citric acid,' 'residual sugar,' 'chlorides,' 'free sulfur dioxide,' 'total sulfur dioxide,' 'density,' 'pH,' 'sulphates,' and 'alcohol' drawn from our dataset.
The models built are:

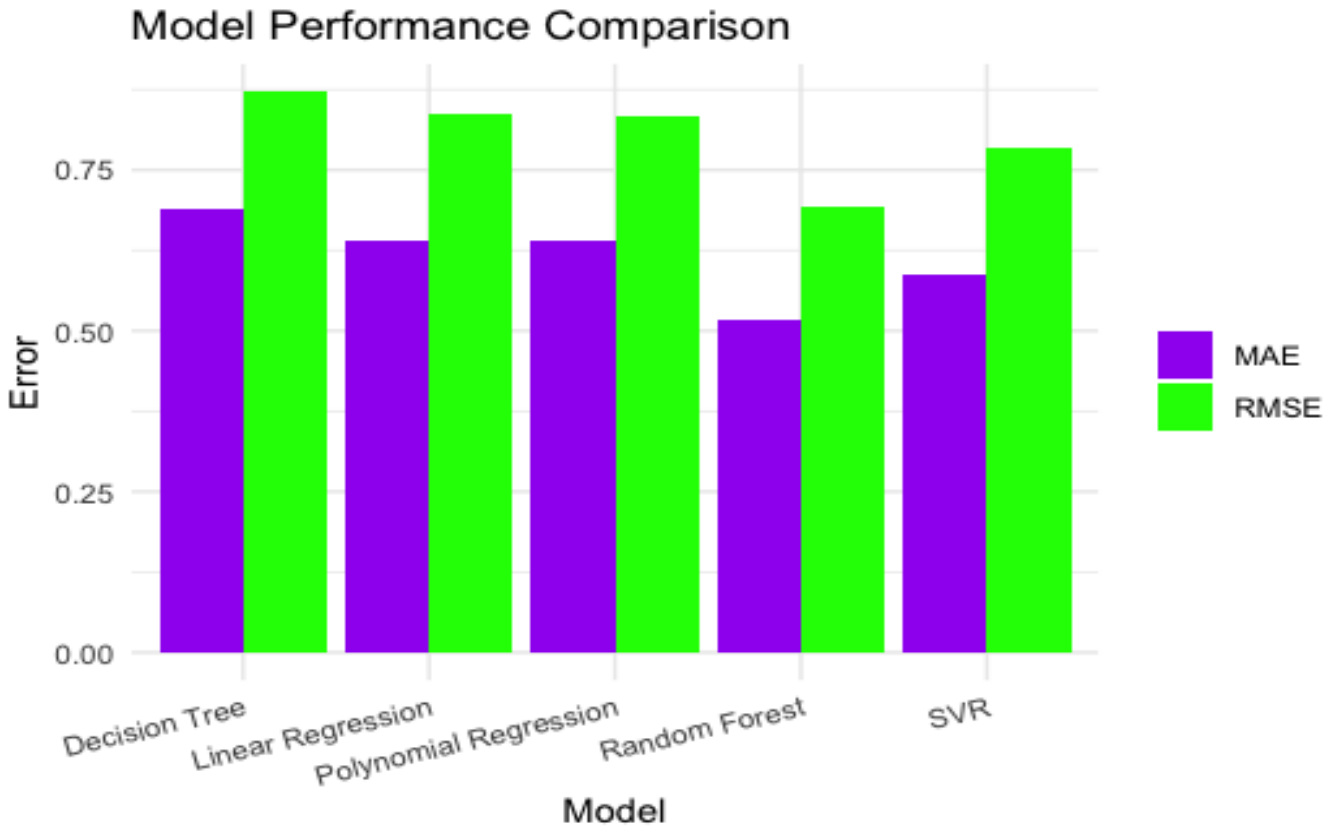
- **Linear Regression model**: Linear regression model was built using the lm() function, allowing us to establish a linear relationship between the independent variables and wine quality.
- **Support Vector Regression (SVR) model:** Support Vector Regression model was implemented using the svm() function, harnessing the power of support vector machines to predict wine quality with a focus on complex relationships within the data.
- **Polynomial Regression model:** Polynomial regression model was constructed through the lm() function, enabling us to capture non-linear associations between attributes and wine quality.
- **Decision Tree Regression model:** Decision tree regression model was developed using the rpart() function, offering a tree-like structure that reveals decision paths for predicting wine quality.
- **Random Forest Regression model:** Random forest regression model was crafted using the randomForest() function, leveraging the strength of ensemble learning to enhance prediction accuracy.

In all cases, the prediction of wine quality was performed using the predict() function.

**Model Evaluation:**
The evaluation of these models was a critical aspect of the analysis, helping me select the most suitable approach for wine quality prediction and gain valuable insights from the data. In the Model Evaluation step, I assessed the performance of five distinct regression models for predicting wine quality. My evaluation was centered on two crucial metrics: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), which serve as dependable indicators of predictive accuracy.

- **Mean Absolute Error (MAE):** With an MAE of approximately 0.63, the model demonstrates reasonable predictive accuracy.
- **Root Mean Squared Error (RMSE):** The RMSE, measuring at 0.69, further validates the model's reliability. These metrics confirm the model's proficiency in predicting wine quality based on the chosen independent variables.

The result of MAE and RMSE for all models is visualized by melting the data frame using the melt() function and then creating a bar chart using the ggplot() library as shown below;

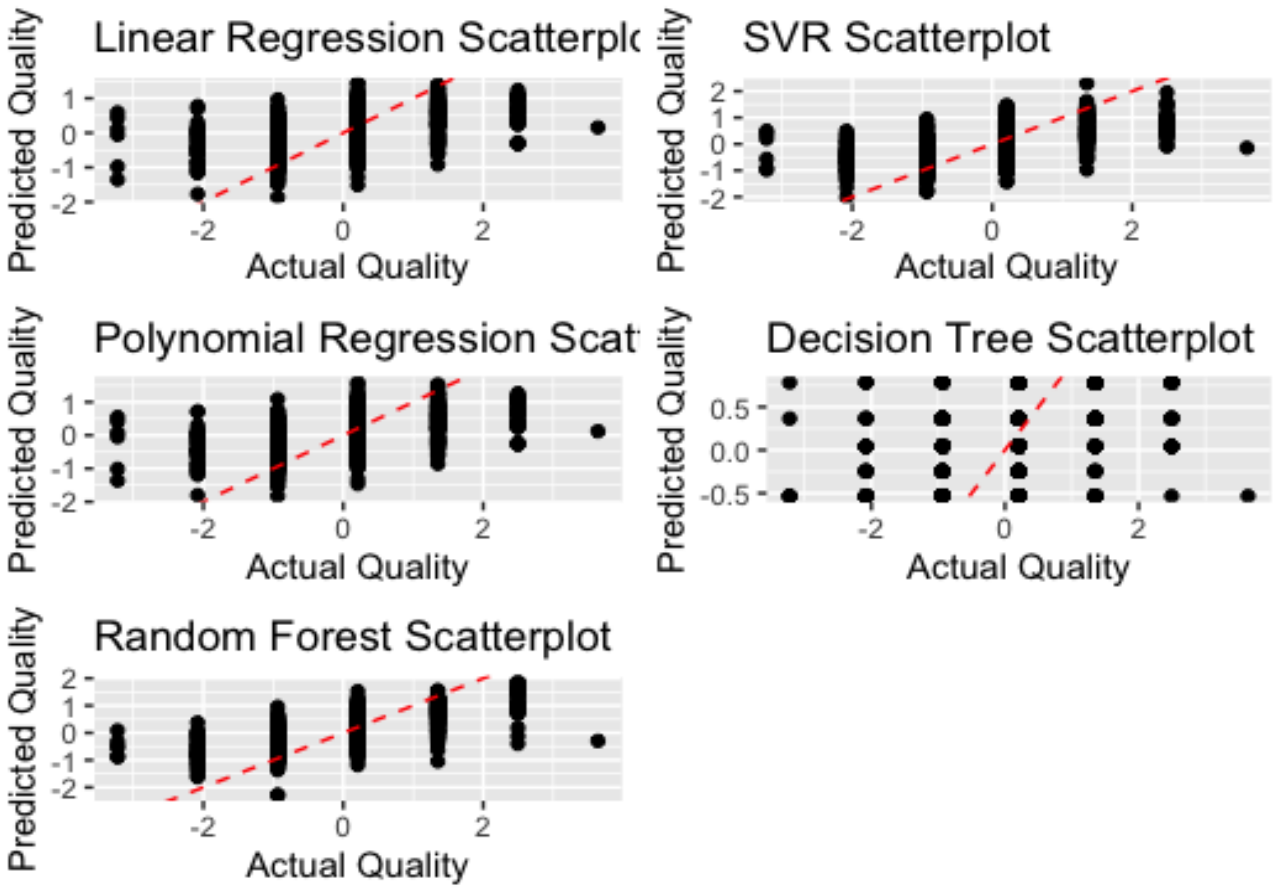



Upon careful examination of these results, the "Random Forest Regression" model emerges as the most effective choice for predicting wine quality. This model attained the lowest MAE (0.51) and RMSE (0.69) values, signifying superior predictive accuracy. Therefore, in the context of my dataset and feature set, the "Random Forest Regression" model stands out as the optimal solution for wine quality prediction.

**Visualization for Evaluating Model Performance:**

Visualizations vividly bring to life the heart of my analysis. The success of the models in predicting wine quality is particularly well-illustrated by the scatterplot, a traditional regression analysis technique. I produced scatter plots using the ggplot2 library that show the correlation between the actual wine quality and the forecasts generated by each model. The grid.arrange() function was used to arrange these plots into a single grid layout.

The scatter plots arranged in one grid for all model is shown below;




The degree to which each model agrees with the real wine quality values may be seen simply and effectively in this scatter plot depiction. It emphasizes the "Random Forest Regression" model's outstanding performance in predicting wine quality, further highlighting that it is the best option for this task.

**Conclusion:**

As shown by the Random Forest Regression model, our investigation of different machine learning algorithms confirms that characteristics like type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol significantly affect wine quality. The best option for predicting wine quality in our dataset is this model, which has the lowest MAE and RMSE values. This project demonstrates how machine learning can be used to address real-world challenges and gain insightful knowledge in the realm of wine quality prediction.