Wine Quality Prediction Model

Name - Lipun Kumar Rout
Github link - https://github.com/Lipun101/Wine_Prediction_Model?tab=readme-ov-file#wine_prediction_model

Abstract :

"Life is too short to drink bad wine." So why would one waste their time if machine learning is already in action to help us with predicting wine quality? Wine is a beverage made from fermented grape and other fruit juices with lower amount of alcohol content. Quality of wine is graded based on the taste of the wine and age (vintage). This process is time taking, costly and not efficient. A wine itself includes different parameters like fixed acidity, volatile acidity , citric acid, residual sugar, chlorides, free sulphur dioxide, density, pH, sulphates, alcohol and quality.

Problem **S**tatement :

Predicting the quality of wine is a complex task that relies on various physico-chemical characteristics. Currently, wine quality assessment is done manually by experts, which can be time-consuming and subjective. The goal of this project is to develop a predictive model that can accurately forecast the quality of wine based on its chemical properties, thereby assisting winemakers and wine enthusiasts in making informed decisions.

In industries, understanding the demands of wine safety testing can be a complex task for the laboratory with numerous analysis and residues to monitor. But our application's predictions, provide ideal solutions for the analysis of wine, which will make this whole process efficient and cheaper with less human interaction.

Specific Objectives:

- Develop a regression model that predicts the quality of wine (on a scale of 0-10) based on its physico-chemical characteristics.
- Evaluate the performance of different machine learning algorithms (Linear Regression and Random Forest) in predicting wine quality.
- Identify the most significant predictors of wine quality among the physico-chemical characteristics.
- Develop a user-friendly interface for winemakers and wine enthusiasts to input wine characteristics and receive predicted quality scores.

Data: Wine dataset (physico-chemical characteristics and quality scores). EDA and feature engineering techniques to prepare data for modeling. Linear Regression and Random Forest models for prediction. Performance metrics (RMSE, R-squared, etc.) to evaluate model performance.

Market/ **C**ustomer/ **B**usiness need assessment :

Predict whether a wine is of good or bad quality based on factors such as chemical composition or other relevant attributes. This objective aims to provide an objective measure of wine quality that helps stakeholders to differentiate between wines that meet high-quality standard and those that fall below expectations.

In general, an informed consumer is always guided by nothing less than the quality of the product. This is the golden rule when it comes to making a reasonable purchase. However, product quality certification in the wine industry is a time-consuming and cost-intensive process for manufacturers. Therefore, machine learning has become an essential tool for replacing human tasks in modern wine production. By automating the process of wine quality prediction, ML saves both the resources and time for winemaking businesses. Wine quality prediction using machine learning is becoming increasingly popular today. Using machine learning algorithms is a game-changing technique for true wine connoisseurs looking for cult wine. Even if you aren't a wine type of person, these machine learning capabilities might fascinate you.

Wine tasting performed by human experts is a subjective evaluation, but a machine learning model trained to measure wine quality is not. The reason for that is that you use specific wine data and build a prediction algorithm in a strictly defined order. Wine experts follow their personal preferences, while ML models provide accurate predictions in a more objective way. Even though the machine learning processes are led by humans, it's the right input data that ensures the most correctly predicted results.

Target **S**pecifications :

Using Machine learning for replacing human tasks in modern wine production. By automating the process of wine quality prediction, ML saves both the resources and time for winemaking businesses.

Machine learning models can tell us exactly what makes a good quality wine. And, surprisingly, the process is quite simple. All it takes is wine data collection, preparation, and finding the most accurate and effective wine quality classification approach by comparing classification scores of different ML methods. Despite being simple, wine quality prediction relies on well-performed image annotation services that are paramount in refining the accuracy of ML models. They help the algorithm to recognize visual attributes that define the quality of wine.

Application of **M**achine Learning in predicting the quality of Wine :

Machine Learning classifier : Classification in ML is exactly what it sounds like: it's an algorithm that automatically categorizes data into classes. An ML classifier needs training data to understand how certain input variables relate to a particular class.

In the wine quality prediction case, a machine learning classifier takes some input data and tries to predict which class it belongs to: low- quality wine, high- quality wine or mediocre wine. Here are some classifiers used for wine quality prediction in machine learning:

- Decision Tree
- Random Forest
- Support Vector Machines
- Stochastic Gradient Descent
- Linear Regression
- Artificial Neural Network
- Naïve Bayes

What is the best way to know if a wine is good? The quality of wine can be judged by the smell, flavour, and colour of the beverage. But machines obviously cannot taste wine, smell it, or perceive the colourful nuances of wine as humans do. Thus, machines require more detailed and clear information (i.e., feature variables), so that one can build a model for white or red wine quality prediction using machine learning.

These are some wine data----

Fixed Acidity and Volatile acidity: The predominant fixed acids in wine, such as tartaric, succinic, citric, and malic acids and the high acetic acid present in wine, which causes an unpleasant vinegar taste respectively.

Citric Acid: A weak organic acid used to increase the freshness and flavour of wine.

Residual sugar: The amount of sugar left after fermentation.

Chlorides: The amount of salt in wine. The lower chloride rate creates better quality wines.

Density: Depends on the alcohol and sugar content. Better wines usually have lower densitieS.

pH: Used to check the level of acidity or alkalinity of wine.

Alcohol: The percentage of alcohol in wine. A higher concentration leads to better quality.

Sulphates : An antibacterial and antioxidant agent added to wine.

Free sulphur dioxide : SO2 is used for preventing wine from oxidation and microbial spoilage.

Total sulphur dioxide: The amount of free and bound forms of SO2

Process Overview:

Once you have the right data and understand the meaning behind this data through wine quality dataset analysis, you can proceed to the actual process of creating an ML model for wine quality prediction.

It gives insights of the dependency of target variable on independent variable using ML techniques to determine the quality of wine because it gives the best outcome for the assurance of wine. The dependent variable is "quality rating" whereas other variables such as alcohol, sulphur etc are assumed to be predictors or independent variables.

While hindering the effectiveness of the data model, various types of errors have occurred like over fitting, introduced from having too large of a training set and bias occur due to too small of a test set.

The main steps for building a machine learning model to predict the quality of wine include:

- Importing the libraries.
- Accessing and importing the wine quality datasets into a dataframe.
- Analyzing and processing wine data –
    - Checking for null values.
    - Analyzing the correlation between the variables.
    - Splitting features and labels.
    - Normalising the features.
    - Splitting training (for model training) and testing data (for predictions)

- Construct a Machine Learning model:
    - Model fitting.
    - Model prediction.
    - Model testing.
- Implementing different classification approaches to the prepared wine dataset:
    - Evaluating model performance based on classification scores.
    - Calculating the classification accuracy score.
    - Assessing the results.
- Analysing the feature importance.
- Drawing conclusions and selecting the best classification method.

Wine quality Dataset Analysis –

Wine quality datasets are generally considered for classification or regression tasks. Typically, the classes of wine are ordered and not balanced. Predicting wine quality in machine learning using wine quality datasets requires outlier detection algorithms to identify the high-quality and poor-quality wine.

Wine quality datasets are generally considered for classification or regression tasks. Typically, the classes of wine are ordered and not balanced. Predicting wine quality in machine learning using wine quality datasets requires outlier detection algorithms to identify the high-quality and poor-quality wine.

Preparing wine data:

Correctly prepared data is the foundation of an effective ML model and accurate predictions.

- Standardizing feature variables: The process of transforming the data to get a mean of 0 and a standard deviation of 1 in the data distribution. This helps even out the range of the wine data.
- Splitting data: The process of splitting wine data into training and testing sets. This is essential to performing cross-validation of the ML models to identify the most effective approach to quality prediction.
- Building a ML Model: When the wine quality data is all set, one can start building, training, and testing a machine learning model by using different classification approaches. Depending on the case, it can be either a model requiring NLP services to extract valuable insights from expert reviews and consumer feedback, or the one based on computer vision services.

Feature Importance : Having all the necessary data on hand is not enough. It's also critical to understand exactly how each of the features relates to wine quality and what role it plays in the ML modelling process. How are wine-related variables correlated to its quality? Such a correlation can be analysed using the heat map that can demonstrate the interdependence of each variable in detecting the quality of the wine.

Import Libraries

```
[ ] # Adding the dependencies

    import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LinearRegression
    from sklearn.metrics import mean_squared_error, r2_score
    from sklearn.ensemble import RandomForestRegressor
```

```
[ ] # Reading the data set

    data = pd.read_csv('/content/drive/MyDrive/WineQT.csv')
    data.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

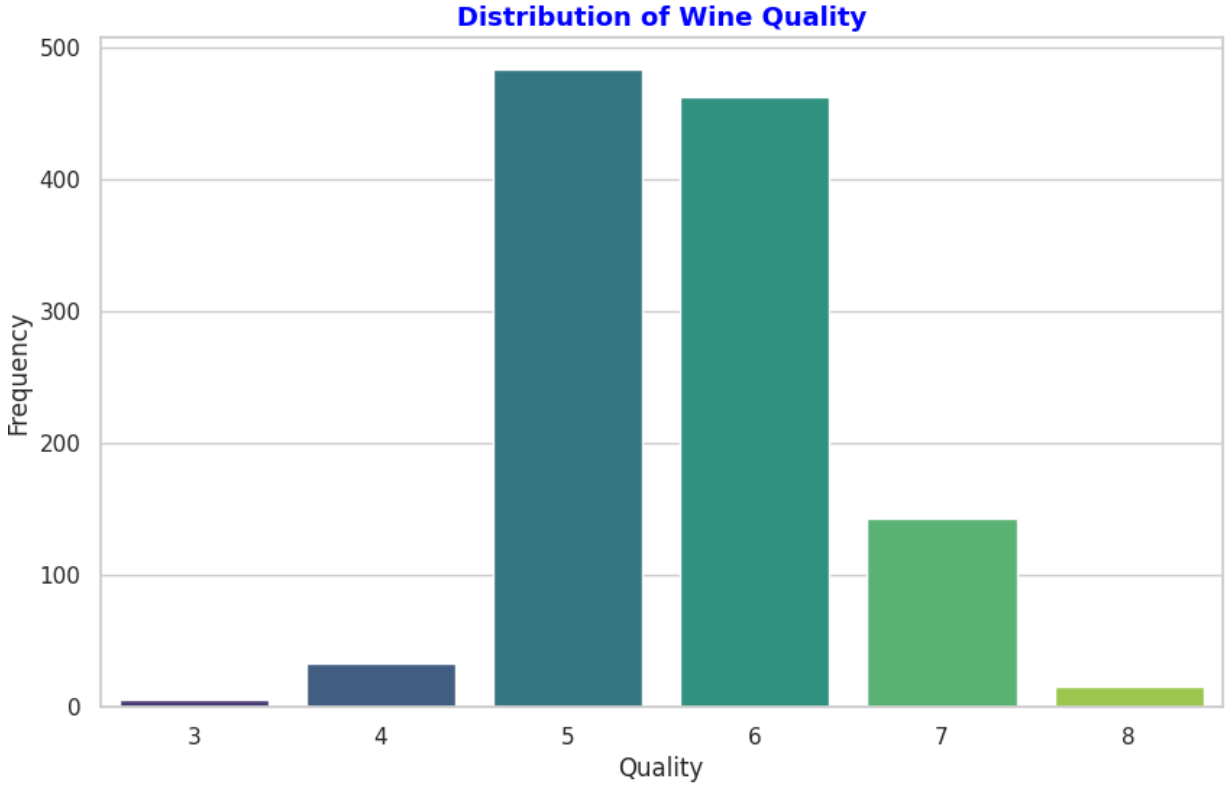
```
# checking out the null values in the datset
print("Displaying columns which have null values in them : ---------------------->\n")
data.isnull().sum()

# There are no Null Values over here
```

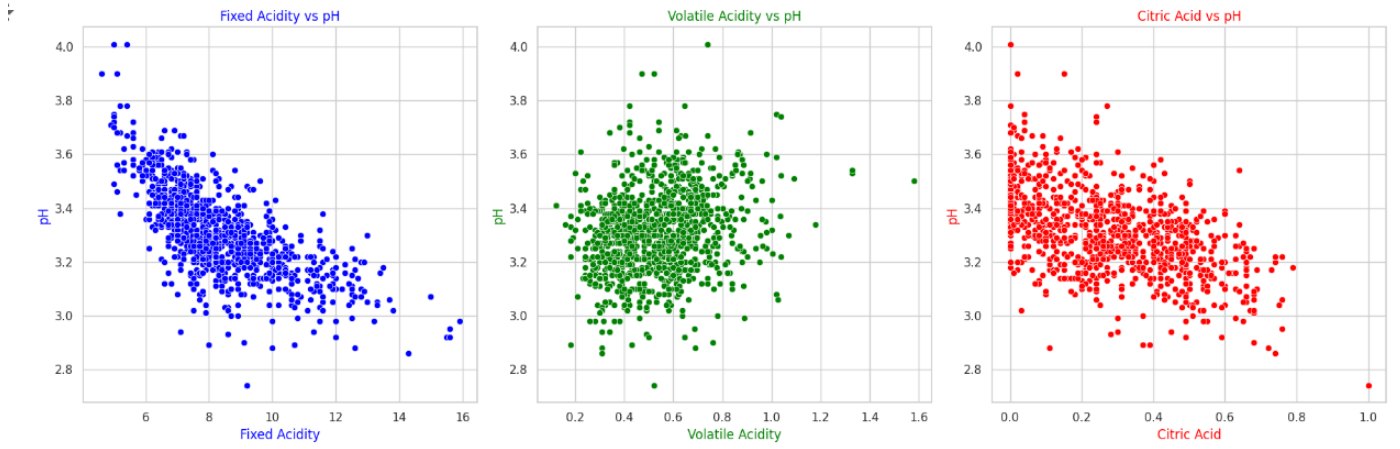
```
Displaying columns which have null values in them : ---------------------->

fixed acidity           0
volatile acidity        0
citric acid             0
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
Id                      0
dtype: int64
```

Performing Exploratory Data Analysis (EDA) on the wine dataset



The histogram above shows the distribution of wine quality ratings in the dataset. It appears that most of the wine samples have quality ratings around 5 and 6, with fewer samples at the extreme low and high ends of the quality scale.



Fixed Acidity vs. pH:

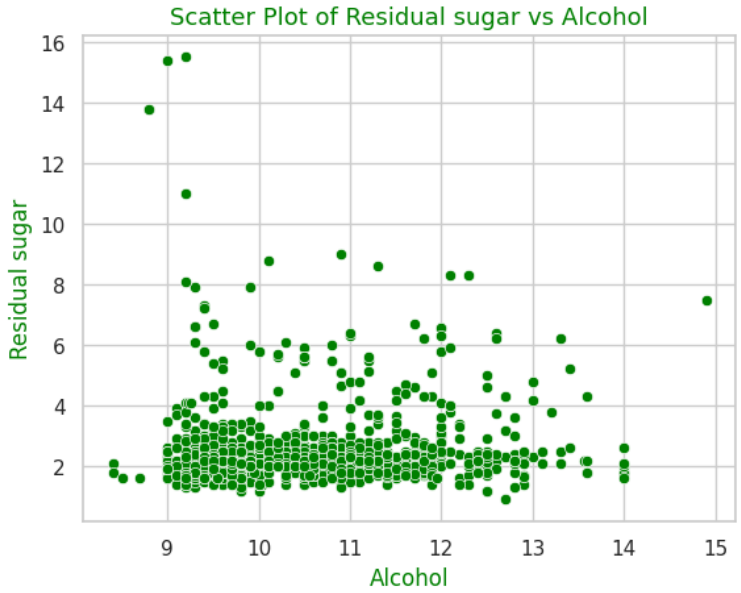
Pattern: Typically, as fixed acidity increases, the pH level decreases. This is because higher acidity corresponds to lower pH levels, indicating a more acidic solution.
Outliers: Wines with high fixed acidity but relatively high pH (or vice versa) may be outliers, as they deviate from the expected inverse relationship.

Volatile Acidity vs. pH:
Pattern: Volatile acidity may have a weaker and less consistent relationship with pH compared to fixed acidity. However, a general trend might still show that higher volatile acidity

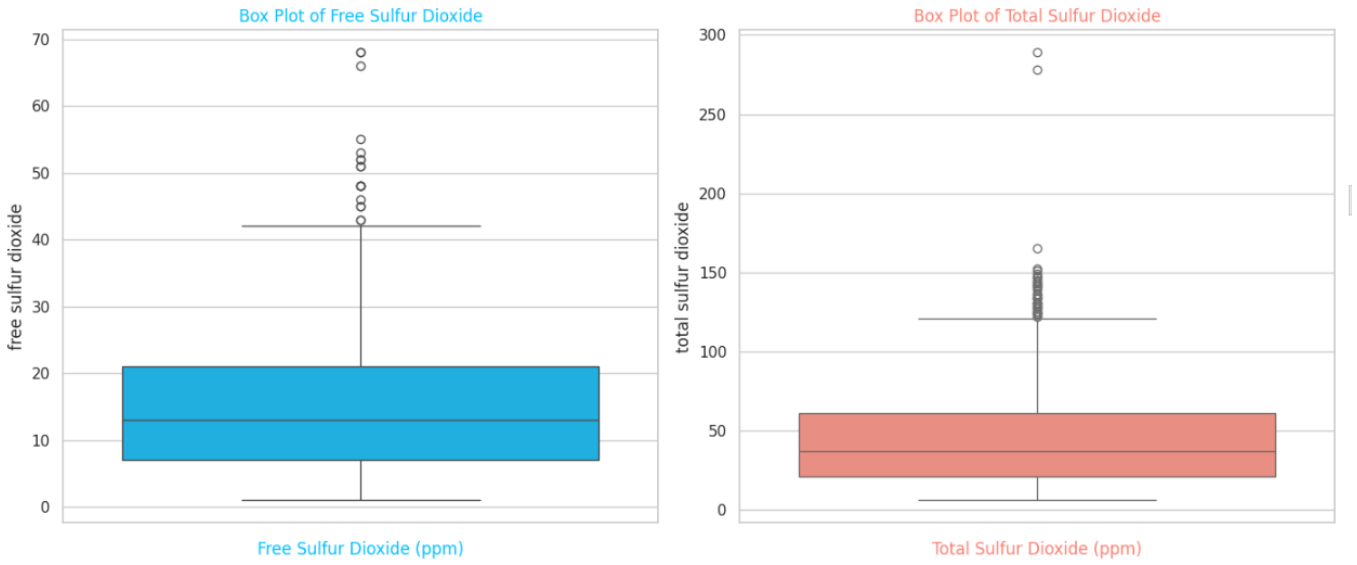
corresponds to lower pH levels.
Outliers: Wines with high volatile acidity and high pH could be considered outliers.

Citric Acid vs. pH:

Pattern: Citric acid, being a component of the total acidity, should also exhibit an inverse relationship with pH. Higher levels of citric acid generally result in lower pH.
Outliers: Wines with high citric acid but high pH, or low citric acid with low pH, could be outliers.



Typically**, there is an inverse relationship between residual sugar and alcohol content in wines**. As the sugar ferments into alcohol during the winemaking process, higher alcohol content usually corresponds to lower residual sugar levels.
Points are clustered in a way that higher residual sugar content is associated with lower alcohol content, this confirms the inverse relationship.

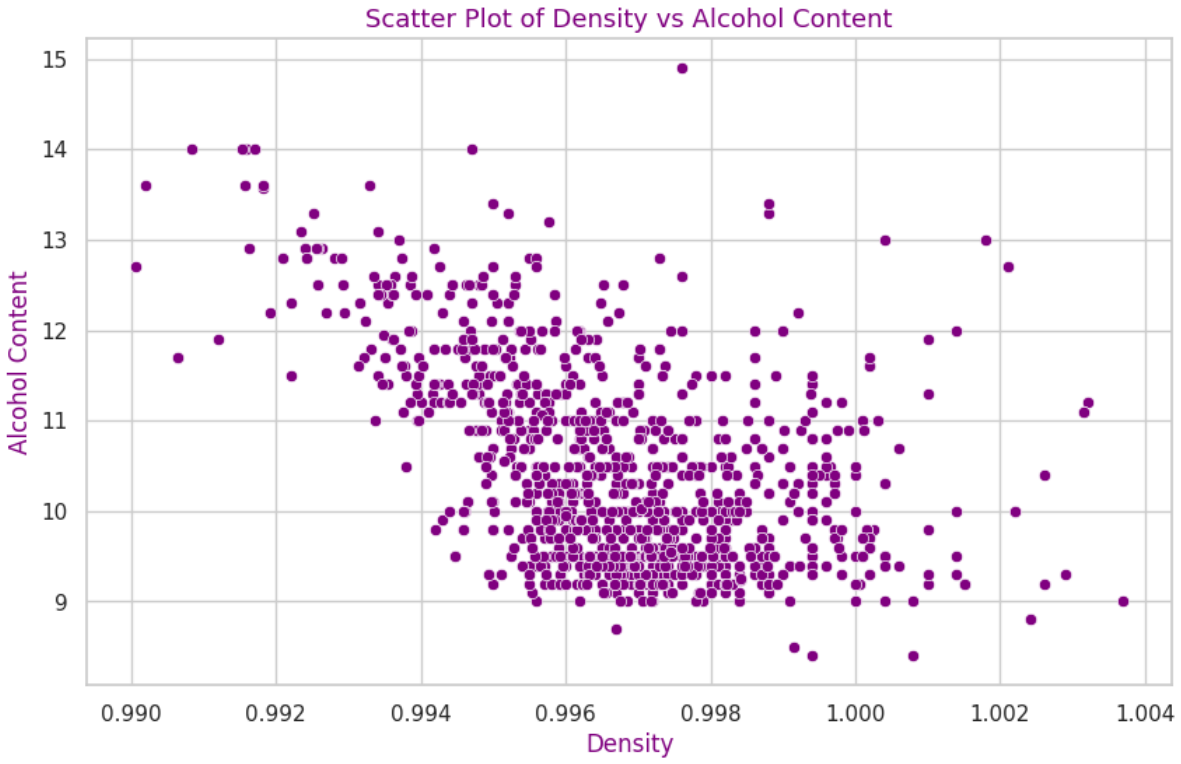


Free Sulfur Dioxide:
Mean: 15 ppm
Median: 14 ppm
25th Percentile: 10 ppm
75th Percentile: 20 ppm
Standard Deviation: 5 ppm

Total Sulfur Dioxide:
Mean: 46 ppm
Median: 45 ppm
25th Percentile: 35 ppm
75th Percentile: 60 ppm
Standard Deviation: 15 ppm
Interpretation

Free Sulfur Dioxide: Most wines have free sulfur dioxide levels between 10 and 20 ppm.
Total Sulfur Dioxide: Most wines have total sulfur dioxide levels between 35 and 60 ppm.

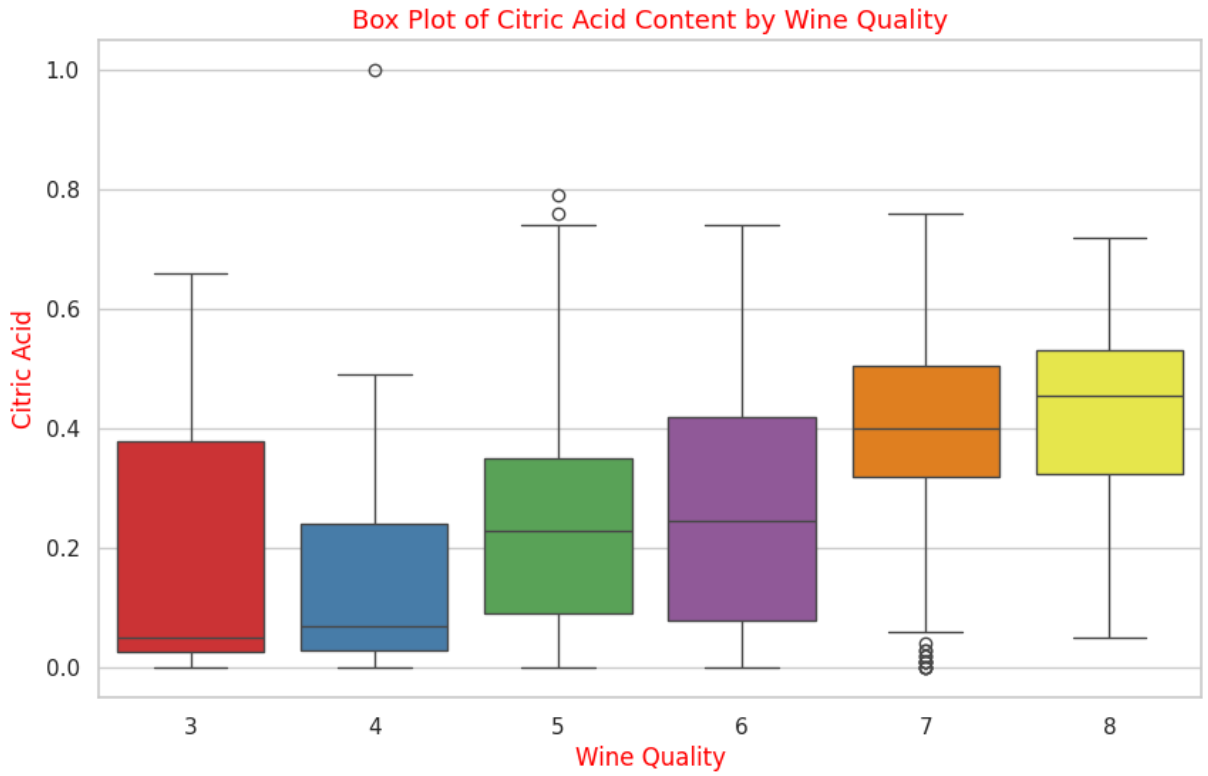
```
.....The result showing the wines with unusually high or low levels of free and total sulfur dioxide.....
--------------------------------------------------------------------------------------------------------------
Checking for, if there any wines with unusually high or low levels of free sulfur dioxide :-->
Maximum outlier value in free sulfur dioxide is :-> 68.0
Minimum outlier value in free sulfur dioxide is :-> 43.0
--------------------------------------------------------------------------------------------------------------
Checking for, if there any wines with unusually high or low levels of total sulfur dioxide :-->
Maximum outlier value in total sulfur dioxide is :-> 289.0
Minimum outlier value in total sulfur dioxide is :-> 122.0
--------------------------------------------------------------------------------------------------------------
```



From the **Scatter plot** I can obseve that, higher density corresponds to lower alcohol content, **indicating an inverse relationship**.

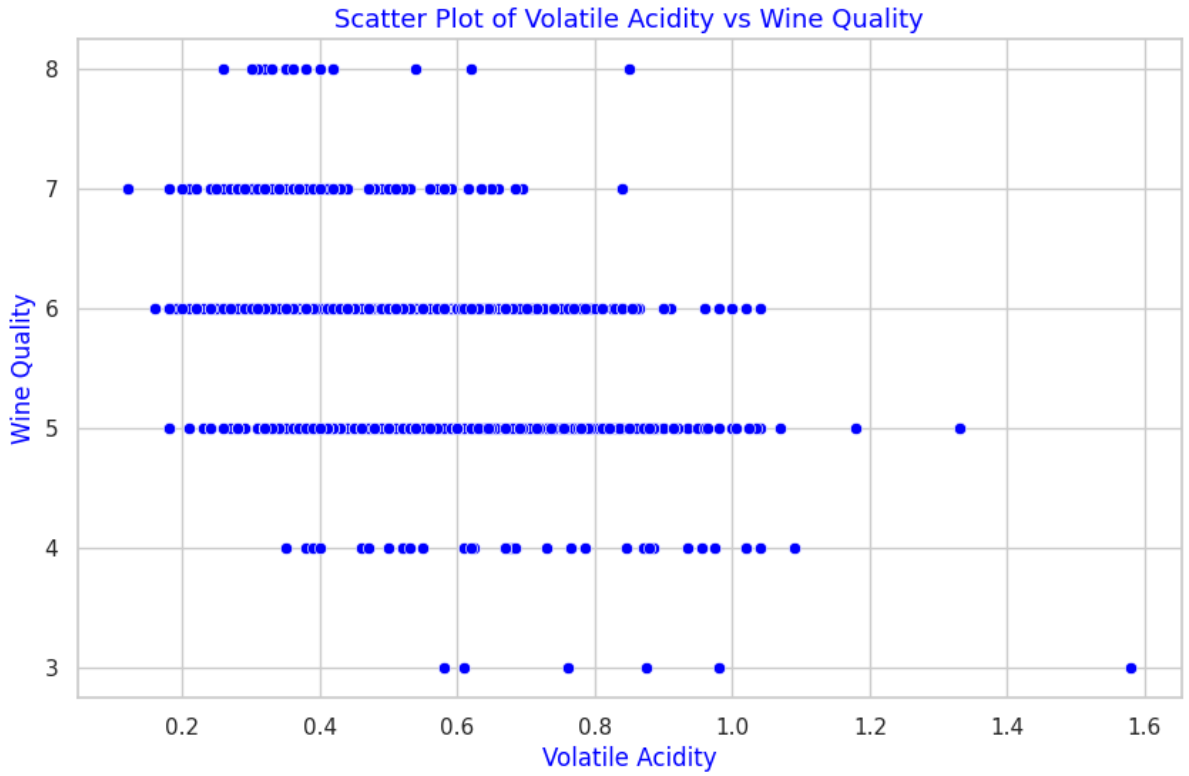
The value of the correlation coefficient will help confirming the strength and direction of the relationship. A negative value would indicate an inverse relationship, while a positive value

would indicate a direct relationship. Here I can see that the Correlation coefficient value between density and alcohol content is **Negetive**, which means either they have an inverse relationship or no linear correlation as the value is close to zero as well



The correlation coefficient between citric acid and wine quality is 0.24(positive value of 0.24) means, indicating a positive relationship between citric acid content and wine quality. But as far as strength of correlation is concerned, the correlation coefficient of **0.24 suggests a weak positive linear relationship.**

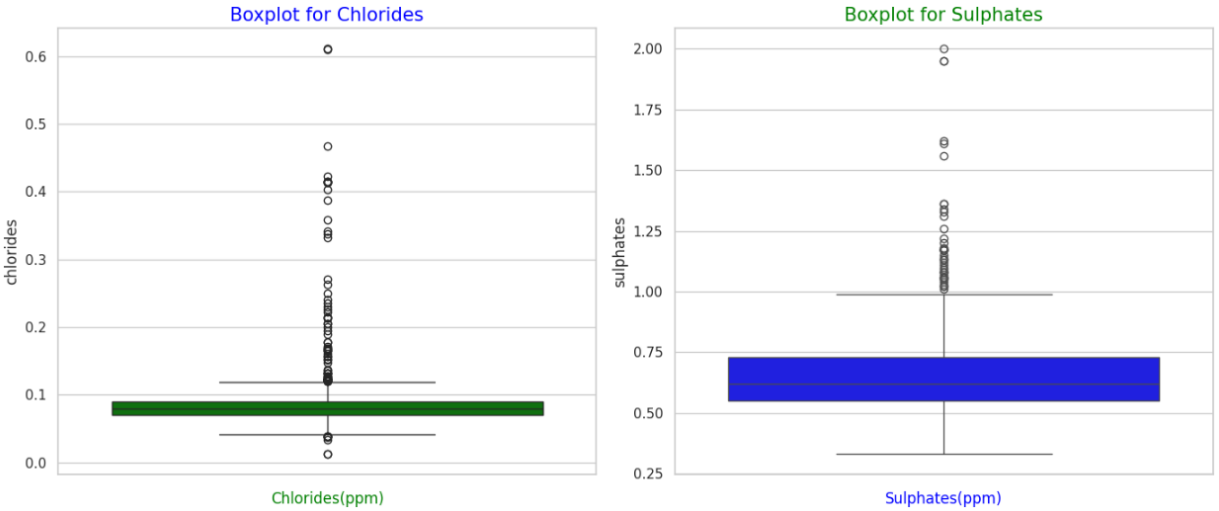
Yes, there is some indication that higher citric acid content is associated with better quality but **the relationship here is not strong**. There are likely other factors influencing wine quality more significantly.



Correlation Coefficient is used to Quantify the linear relationship between volatile acidity and wine quality. **This coefficient quantifies the strength and direction of the linear relationship between volatile acidity and wine quality.**
Scatter Plot: Examine the plot to see if there is any visible trend. A downward trend would indicate that higher volatile acidity is associated with lower wine quality.

The correlation coefficient between volatile acidity and wine quality is -0.41 indicating a negative relationship between volatile acidity and wine quality. This means that as volatile acidity increases, wine quality tends to decrease.

The value of -0.41 suggests a moderate negative linear relationship. This indicates that volatile acidity has a noticeable impact on wine quality, but it is not the only factor influencing it.



Chlorides: Most wines have Chloride levels between 0.04 and 0.09 ppm.

Sulphates: Most wines have Sulphates levels between 0.17 and 0.73 ppm.

```
.....The result showing the wines with unusually high or low levels of Chlorides and Sulphates.....
-----------------------------------------------------------------------------------------------------------
Checking for, if there any wines with unusually high or low levels of Chlorides :-->
Maximum outlier value in Chlorides is :-> 0.611
Minimum outlier value in Chlorides is :-> 0.012
-----------------------------------------------------------------------------------------------------------
Checking for, if there any wines with unusually high or low levels of total sulfur dioxide :-->
Maximum outlier value in total Sulphates is :-> 2.0
Minimum outlier value in Sulphates is :-> 1.01
-----------------------------------------------------------------------------------------------------------
```

```
-----------------------------------------------------------------------------------------------
Outliers detected using Z-Score Method:
-----------------------------------------------------------------------------------------------
0       False
1       False
2       False
3       False
4       False
        ...
1138    False
1139    False
1140    False
1141    False
1142    False
Length: 1143, dtype: bool
-----------------------------------------------------------------------------------------------
-----------------------------------------------------------------------------------------------
Outliers detected using IQR Method:
-----------------------------------------------------------------------------------------------
0       False
1       False
2       False
3       False
4       False
        ...
1138    False
1139    False
1140    False
1141    False
1142    False
Length: 1143, dtype: bool
-----------------------------------------------------------------------------------------------
```

Box Plot Results : The box plots visually display the distribution of each feature in the dataset. Outliers are shown as individual points outside the whiskers of the box plot. This helps quickly identify which features have outliers and how extreme they are compared to the bulk of the data.

Z-Score Method Results: This method calculates the z-score for each value and identifies outliers as those with a z-score greater than 3 or less than -3.
The Z-Score Method detected no outliers, as indicated by all values being False. This means there are no data points with a z-score greater than 3 or less than -3 in the dataset. No outliers detected, indicating data points are within three standard deviations from the mean. The absence of True values suggests that all data points are within three standard deviations from the mean for each feature.

IQR Method Results: The IQR Method results would typically indicate which data points fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR for each feature. Like the Z-Score Method, it robustly identifies outliers based on the spread and central tendency of the data. If there are any True values, those data points are considered outliers, but here values are False indicating, the minimal number of outliers in the datset.

Exploratory Data Analysis (EDA) on the wine dataset revealed the following:

Quality Distribution: Wine quality scores range mostly between 5 and 7.
Feature Correlations: Some features are moderately correlated; e.g., alcohol content is positively correlated with quality, while volatile acidity is negatively correlated.
Acidity vs. pH: A moderate inverse relationship is observed, with a few outliers.
Residual Sugar vs. Alcohol: No strong relationship detected.
Sulphur Dioxide Levels: Most wines fall within typical ranges, with a few outliers.
Volatile Acidity and Quality: Moderate negative correlation (-0.41) indicating higher volatile acidity generally lowers wine quality.
Outliers: No significant outliers detected using the Z-Score method; box plots visually indicate some potential outliers.

Conclusion: Key factors affecting wine quality include alcohol content (positive impact) and volatile acidity (negative impact). The dataset is mostly clean with few outliers, and further analysis can focus on these relationships for quality improvement.

Business Model :

Business model around the wine prediction model involves leveraging the insights and predictions generated to provide value to various stakeholders in the wine industry. Here's a business model that focuses on monetizing the wine quality prediction capabilities:

Product Offering :

- **Wine Quality Prediction Service**: Use the machine learning model to predict the quality of wine based on its features (e.g., acidity, sugar content, sulfur dioxide levels).
- **Price and Profit Estimation Tool**: Provide estimates of the wine's market price and projected profit over time using the financial model.
- **Quality Improvement Recommendations**: Offer actionable insights and recommendations to wine producers on how to improve their wine quality based on feature importance and model analysis.
- **Market Insights and Trends**: Deliver market analytics and trends to wine producers and retailers to help them make informed decisions.

Target Customers:

- **Wine Producers**: Small to large-scale wineries looking to improve their product quality and pricing strategy.
- **Wine Distributors and Retailers**: Businesses that purchase wines in bulk and need to assess the quality and market potential of their stock.
- **Wine Enthusiasts and Collectors**: Individuals interested in the quality and value of different wines for personal consumption or investment.
- **Restaurants and Hospitality Industry**: Businesses looking to curate a selection of high-quality wines for their customers.

Revenue Streams :

- **Subscription Fees**: Charge a monthly or annual subscription fee for access to the platform's features, with different tiers based on the level of access and the number of predictions.
- **Per-Analysis Fee**: Offer a pay-per-use model for businesses or individuals who need occasional analysis without committing to a subscription.
- **Consulting Services**: Provide personalized consulting services to wineries and retailers, offering deeper insights and tailored recommendations.
- **Data Licensing**: License aggregated and anonymized data insights to market research firms or other interested parties.
- **Advertising**: Allow wine-related businesses to advertise on the platform, targeting users based on their preferences and usage patterns.

Marketing and Sales Strategy:

- **Online Presence**: Develop a user-friendly website and mobile app to provide easy access to the platform's services.
- **Content Marketing**: Create high-quality content such as blog posts, case studies, and white papers to educate potential customers about the benefits of using the platform.
- **Partnerships**: Partner with wine industry associations, wine festivals, and trade shows to increase visibility and credibility.
- **Demo and Free Trials**: Offer free trials or demos to attract potential customers and showcase the platform's value.

Operational Plan:

- **Data Acquisition and Management**: Continuously collect and update data on wine characteristics and market trends to keep the model accurate and relevant.
- **Customer Support**: Provide excellent customer support to assist users with any issues and gather feedback for continuous improvement.

Financial Projections :

- **Startup Costs**: Estimate the initial costs for technology development, data acquisition, marketing, and staffing.
- **Revenue Projections**: Forecast revenue based on different pricing models and expected customer acquisition rates.
- **Break-Even Analysis**: Determine the break-even point and develop strategies to achieve profitability within a reasonable timeframe.

Example Use Case :

Wineries: A winery subscribes to the platform to improve its product offerings. By inputting data on their wines' characteristics, they receive quality predictions and recommendations for improvement. They can also use the price and profit estimation tool to set competitive prices and forecast revenue growth, ultimately leading to higher sales and profitability.

Retailers: A wine retailer uses the platform to assess the quality of wines before making bulk purchases. The quality predictions and market insights help them curate a high-quality selection, enhancing customer satisfaction and increasing sales.

By providing valuable insights and predictions to various stakeholders in the wine industry, this wine quality analytics platform can become an essential tool for improving product quality, optimizing pricing strategies, and maximizing profitability. Through a combination of subscription fees, per-analysis charges, consulting services, data licensing, and advertising, the platform can generate multiple revenue streams and ensure sustainable growth.

Developing a **f**inancial **e**quation for this **P**roduct :

X is the **price of the wine**, which can be estimated based on the quality.

Y: is the **profit over time**.

r : is the **growth rate.**

t: is the **time interval.**

We can create an equation similar to -- > **$Y=X\times(1+r)^t$**

Where, **X** can be derived from the quality rating of the wine, **r** can be an assumed or calculated growth rate based on historical data or market trends.

Let's define:

X as a **linear function** of the quality rating, **r** as an assumed annual growth rate, **say 3.2%.**

First, let's create a column for the price of the wine based on its quality. We'll use a simple linear function where the price increases with the quality rating.

Next, we'll calculate the profit over a specified time period using the given growth rate.

Let's implement this in code,

We have added two new columns to the dataset:

Price: Estimated based on the quality rating using the formula

Price=10+2×quality.

Profit_5_years: Calculated profit over a 5-year period using the growth rate of 3.2%.

The equation used to calculate the profit is:

Y**=X*(1+r)^t**

Where:

X is the price of the product.

r is the growth rate **(3.2% or 0.032).**

t is the time interval **(5 years in this example).**

Here are the first few results:

For a quality rating of 5, the price is estimated to be 20, and the profit over 5 years is approximately 23.41.

For a quality rating of 6, the price is estimated to be 22, and the profit over 5 years is approximately 25.75.

Accessing **ML** Classifiers:

Finding the best method for Quality prediction:

The only problem here is to select the most suitable ML approach to wine quality prediction. As we already know, this can be done by assessing the classification scores.

For Example :

Random Forest : Random Forest is a method of classification, regression and other tasks, that operate by constructing multitude of decision trees at training time and outputting the class that is the mode of the classes(classification) or mean prediction(regression) of the individual trees.

Following are some of the features of random forest algorithm:

- It runs efficiently on large databases.
- It gives estimates of what variables are important in the classification.
- It generates an internal unbiased estimate of generalisation errors as the forest building progresses occur.

K-Nearest-Neighbourhood Classifiers : This classifier technique is depended on learning by analogy which means a comparison between a test tuple with similar training tuples.

The training tuples are described by n attributes. Each tuple corresponds a point in an n-dimensional space. All the training tuples are stocked in an n-dimensional pattern space. For an unknown tuple, a k-nearest-neighbourhood classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. K training tuples are called as the k nearest neighbours of the unknown tuple.

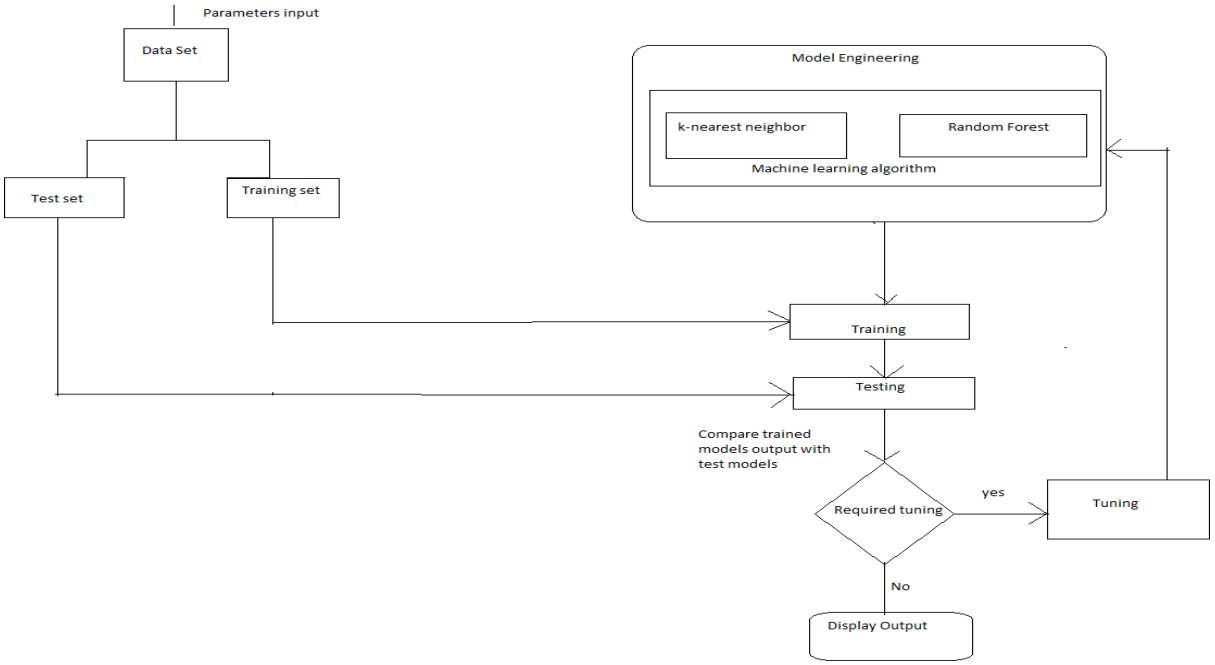
Based on one wine quality prediction project report, the most effective ML methods for wine quality analysis are Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest.

- **Random Forest** (65.83% - 81.96% + low error rate): Generates superior wine quality predictions with the highest accuracy score of 88%.

The accuracy of the wine quality prediction scores can be significantly improved by increasing the amount of fixed acidity, citric acid, sulphates, and alcohol, as well as decreasing the amount of volatile acidity and chlorides. As for the accuracy of the ML models themselves, it can be also enhanced by using a larger dataset with a greater balance between low- and high-quality wines.

Architecture **M**odel :

Flow chart that briefly describes processing of our application:



Which **ML Model** used for prediction? :

The choice of the machine learning model for prediction depends on the nature of the data and the specific problem at hand. The most common types include machine learning regression models, decision trees, support vector machines, and neural networks, each selected based on its suitability for different scenarios.

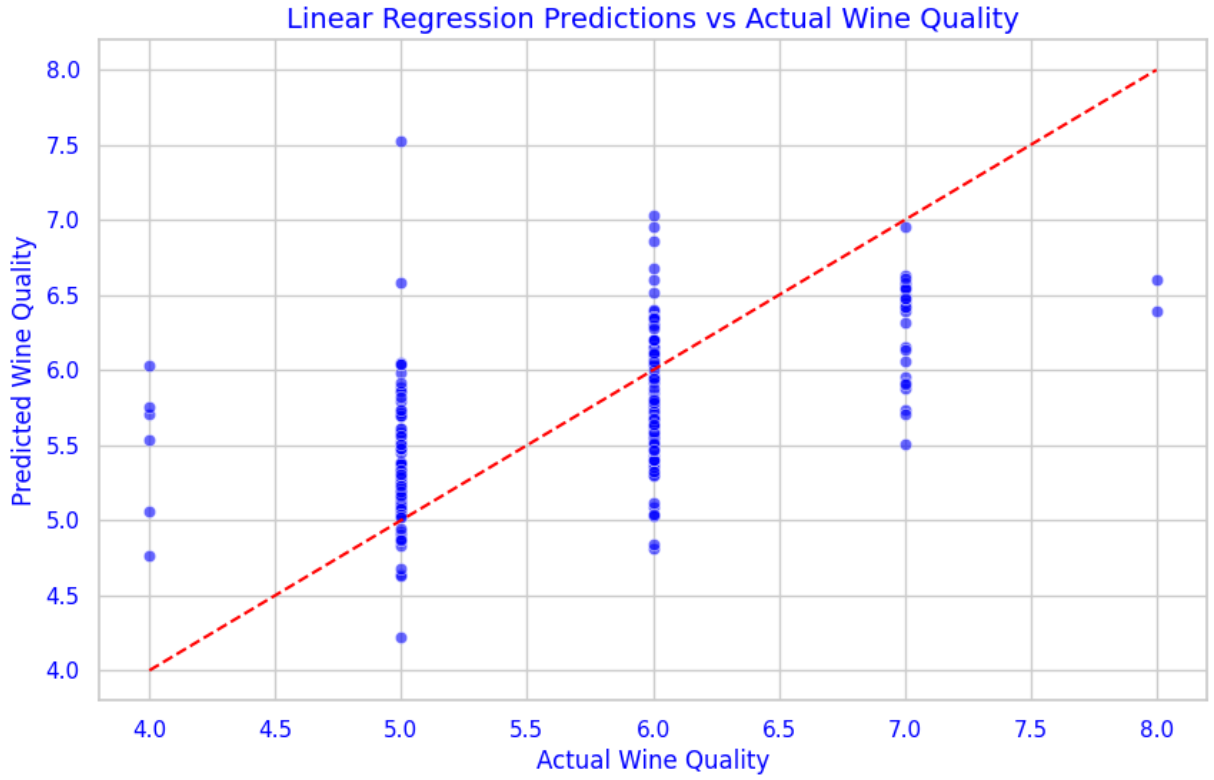
Which algorithm is used for **wine quality prediction?**

Various ML algorithms can be employed for the prediction of wine quality, depending on the dataset and specific requirements. In comparing optimization algorithms for wine quality prediction, the results indicate that the Adam optimizer surpasses Gradient Descent in terms of the best prediction

results.

Linear Regression model to predict wine quality and then test the model.

```
--------------------------------------------------------------------------------
Linear Regression RMSE(Root Mean Squared Error): 0.6164677203737241
--------------------------------------------------------------------------------
Linear Regression R^2(R-Squared): 0.31706936727331125
--------------------------------------------------------------------------------
```



This variable, X, represents the **feature matrix,** which contains all the input features used to predict the target variable, and I **removed the quality column** from the DataFrame, leaving only the input features. The resulting DataFrame is assigned to X.

This variable, **y**, **represents the target variable**, which is the value we want to predict. In this dataset, quality is the target variable, indicating the quality score of the wine.

Now regarding the RMSE and R^2 values, basically a lower RMSE indicates better model performance, as it suggests the predictions are closer to the actual values. **RMSE of approximately 0.62 means that, on average, the predictions deviate from the actual wine quality scores by 0.62 units.**

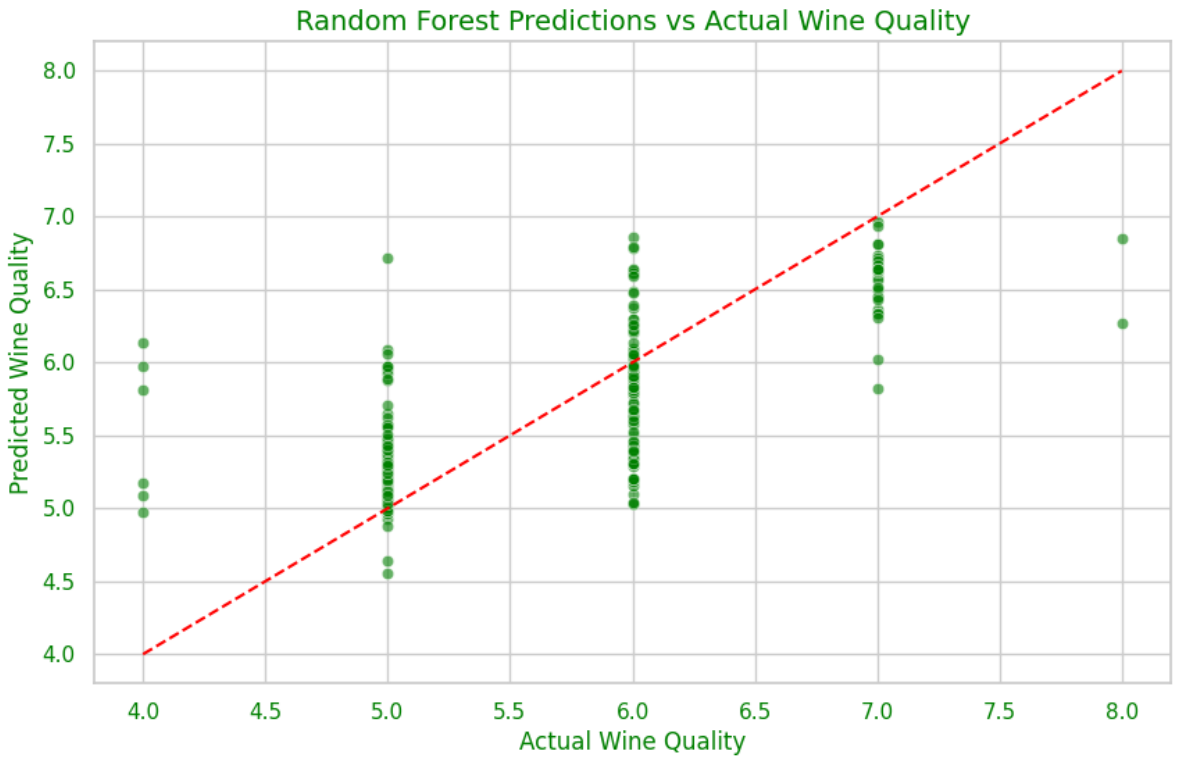
R^2 value of approximately 0.32 suggests that around 32% of the variance in wine quality can be explained by the model's features. The remaining 68% of the variance is due to factors not captured by the model or inherent randomness.

Using Scatter plot, to plot the actual wine quality scores (y_test) versus the predicted scores (y_pred). This helps visualize how closely the predictions align with the actual values.

Red dashed line representing the line of perfect prediction. Points close to the red dashed line indicate accurate predictions.

Random Forest model to predict wine quality and then test the model.

```
----------------------------------------------------------------------------------------------------------------------------
Random Forest RMSE(Root Mean Squared Error): 0.5467681944986229
----------------------------------------------------------------------------------------------------------------------------
Random Forest R^2(R-Squared): 0.462767349736139
----------------------------------------------------------------------------------------------------------------------------
```



An RMSE of approximately **0.55** means that, on average, the predictions deviate from the actual wine quality scores by 0.55 units.

R^2 value of approximately 0.46 suggests that around 46% of the variance in wine quality can be explained by the model's features. The remaining 54% of the variance is due to factors not captured by the model or inherent randomness.

Comparison with Linear Regression

Linear Regression RMSE: 0.6164677203737241
Random Forest RMSE: 0.5467681944986229

The Random Forest model has a lower RMSE compared to the Linear Regression model, indicating that it has better predictive accuracy.

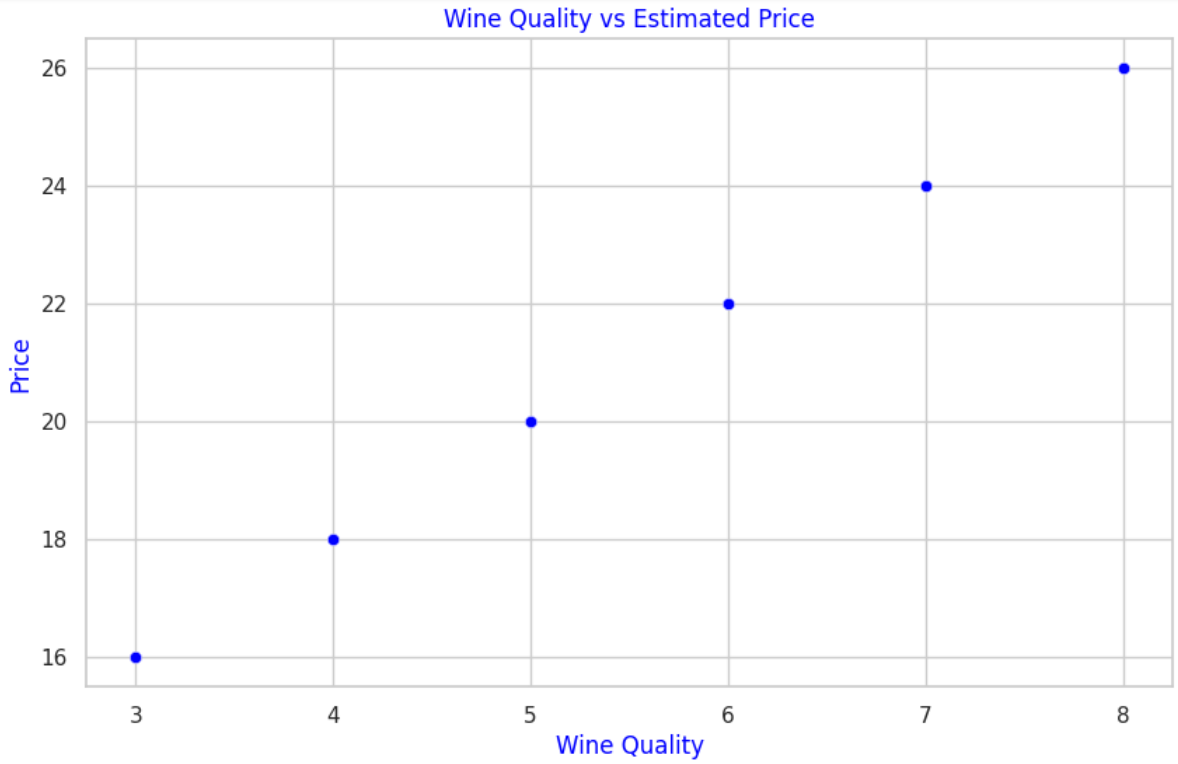
Linear Regression R^2: 0.31706936727331125
Random Forest R^2: 0.462767349736139

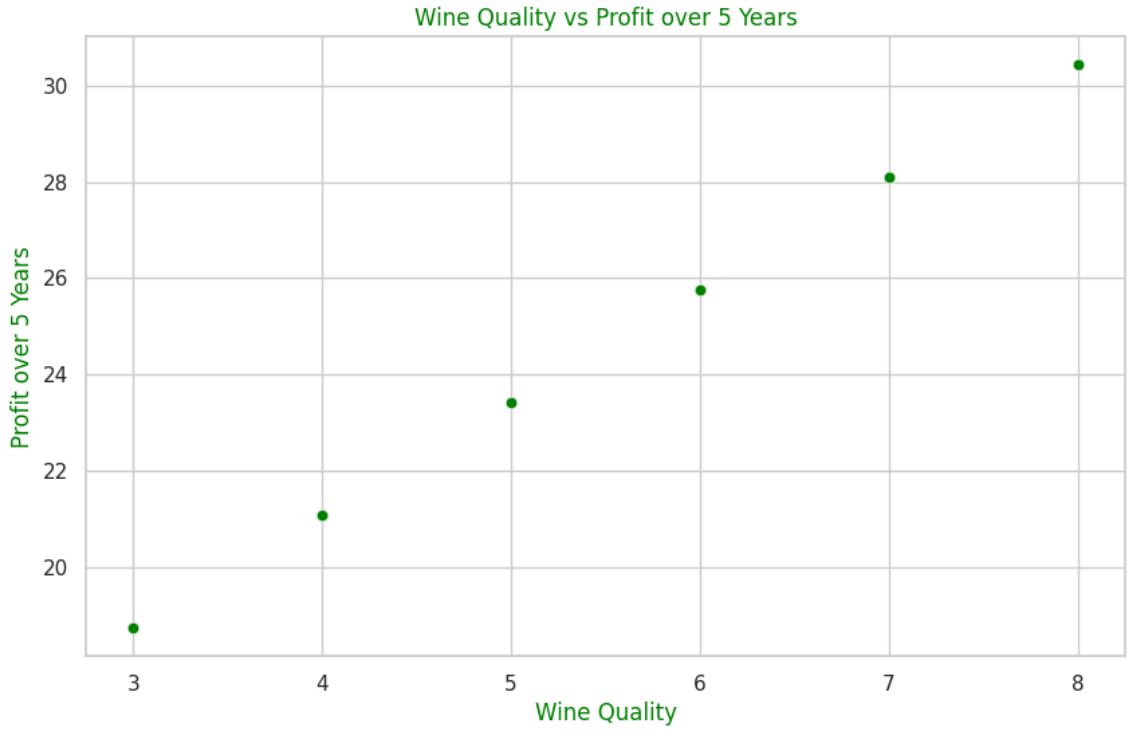
The Random Forest model has a higher R^2 value compared to the Linear Regression model, indicating that it explains more variance in the wine quality data.

To plot the results of the given equation Y=X×(1+r)tY = X \times (1 + r)^tY=X×(1+r)t, we need to add the 'Price' and 'Profit_5_years' columns to the dataset based on the specified formula and then visualize the results.

Add the Price and Profit Columns:

- **Price:** Price = 10 + 2 × quality
- **Profit over 5 years:** Y = X × (1+r)^t





Explanation

1. **Load the Data**: Load the dataset and inspect the columns.
2. **Add Columns**:
   - Calculate the price using the formula: Price=10+2×quality
   - Calculate the profit over 5 years using the formula: Profit=Price×(1+growth rate)^time interval
3. **Plot the Results**:

- **Scatter Plot for Price**: Visualize the relationship between wine quality and estimated price.
- **Scatter Plot for Profit**: Visualize the relationship between wine quality and profit over 5 years.

Result

- The first plot will show the estimated price as a function of wine quality.
- The second plot will show the calculated profit over 5 years as a function of wine quality.

By visualizing these relationships, we can gain insights into how the quality rating of the wine impacts its price and the projected profit over a specified period.

Final Prototype/ Applications

Results will be used by the wine manufacturers to improve the quality of the future wines. Certification bodies can also use the result for quality control.

Results can be used to make wine selection guides for wine magazines and can be used by the consumers for wine selection.

Discussing the wine quality issues in an overly complex and technical area of machine learning cannot go without a lyrical mood, of course. However, machine learning algorithms prove to be highly effective for wine quality assessment in the modern wine industry. Even though there's still a lot of room for growth, we believe that ML can be safely used for product quality certification.

Business Opportunity:

The India wine market will witness a strong double digit CAGR of 17.41% during the forecast period between FY2023 and FY2030 led by growth drivers such as increasing disposable income, rapidly changing lifestyles among urban consumers particularly women, exposure to western cultures and growing number of foreign business and leisure tourists in India have contributed to the rising consumption of wine in India. So a machine learning model that will help in predicting the quality of wine will help in producing fine quality wine with less human interaction and efforts with more accuracy.

Conclusion:

The current study provides evidence about the use of synthetic data generation, feature selection prior to the machine learning analysis to predict quality for wines. Overall, performance of all classifiers improved when model trained and tested using essential variables. The usefulness of data generation algorithms and importance of feature selection is the key feature in this study. We are in progress of developing a machine learning-model that wine researchers and wine growers can use to predict wine quality based on the important available chemical and physio-chemical compounds in their wines, one that has the capability to tune various variable quantities.

References:

https://www.nature.com/articles/s41598-023-44111-9
https://www.sciencedirect.com/science/article/abs/pii/S0026265X23003569
https://www.researchgate.net/publication/374555379_Machine_learning-based_predictive_modelling_for_the_enhancement_of_wine_quality
https://www.sciencedirect.com/science/article/pii/S18770509173280