**WINE QUALITY PREDICTION USING MACHINE LEARNING RANDOM FOREST REGRESSION**

**ABSTRACT**

The prediction of red wine quality is crucial for wineries and wine distributors to maintain quality standards and meet consumer expectations. Traditional methods of quality assessment involve chemical testing and human evaluation, which are time-consuming and expensive. This project proposes a machine learning approach using Random Forest Regression to predict the quality of red wine based on its physicochemical properties. By leveraging the ensemble learning capabilities of the Random Forest algorithm, the model achieves high accuracy, robustness, and interpretability, making it a valuable tool in the wine industry. Wine quality prediction using random forest regression involves using an ensemble of decision trees to predict the quality of wine based on its chemical properties, such as acidity, sugar content, alcohol, and pH. Random forest excels in this task because it captures complex, non-linear relationships between features and wine quality scores. By averaging the predictions from multiple decision trees, it reduces overfitting and improves accuracy, making it a reliable model for wine quality evaluation. Additionally, it can identify the most influential features affecting wine quality, providing valuable insights for wine producers and researchers.

**INTRODUCTION**

The quality of red wine is influenced by various chemical properties such as acidity, sugar content, alcohol level, and pH . Accurate prediction of wine quality can help producers streamline the production process and optimize quality control. Traditional methods rely on subjective human evaluation or expensive chemical testing, which can be inconsistent or inefficient for large-scale production .we will predict the quality of wine on the basis of given features. We use the wine quality dataset available on Internet for free. This dataset has the fundamental features which are responsible for affecting the quality of the wine. By the use of several Machine learning models, we will predict the quality of the wine.

This project focuses on building a predictive model using Random Forest Regression, an ensemble learning technique that combines multiple decision trees to provide robust predictions. The model utilizes physicochemical attributes of red wine to estimate its quality score, offering an efficient and scalable alternative to traditional assessment methods. Random forest regression is an invaluable tool in data science. It enables us to make accurate predictions and analyze complex datasets with the help of a powerful machine-learning algorithm. a Random forest regression model combines multiple decision trees to create a single model. Each tree in the forest builds from a different subset of the data and makes its own independent prediction. The final prediction for input is based on the average or weighted average of all the individual trees’ predictions.

**Existing System**

Decision trees were used in the existing code for the data set which has been used

Random forest regression is being used to improve the existing system for the data set.

Random forest is the better than decision trees as it uses ensemble nature

Random forest regression is better than decision trees and linear regression because it combines multiple decision trees to create an ensemble model, improving accuracy and robustness. While decision trees are prone to overfitting and can be unstable with small changes in data, random forests mitigate these issues by averaging the predictions of numerous trees, reducing variance and overfitting. Unlike linear regression, which assumes a linear relationship between variables, random forest handles non-linear and complex data patterns effectively. It is also more robust to outliers and noise, can process high-dimensional data, and provides feature importance insights, making it a versatile and reliable choice for many predictive tasks.

**Proposed System**

Random Forest Regression-Based Solution

The proposed system leverages Random Forest Regression to predict the quality of red wine based on its physicochemical properties. The model addresses the limitations of existing methods by combining the strengths of multiple decision trees, ensuring accurate and consistent predictions.

A single decision tree is prone to overfitting, especially when the tree is deep. It tends to capture noise in the data, which negatively impacts its generalization to unseen data.

Linear Regression: Assumes a linear relationship between the dependent and independent variables, which may not hold in many real-world scenarios.

Random Forest Regression: Combines predictions from multiple decision trees (trained on different subsets of data and features) to reduce overfitting and variance, resulting in more accurate and robust predictions.

**CODE**

# Import necessary libraries

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error

import matplotlib.pyplot as plt

import seaborn as sns

#  Wine Quality dataset

data = pd.read\_csv('red.csv')

#  the dataset

print(data.head())

# missing values

print(data.isnull().sum())

# Data Preprocessing

data.fillna(data.median(), inplace=True)

# features (X) and target (y)

X = data.drop('quality', axis=1)  # (excluding target variable)

y = data['quality']  # Target variable

# Split into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Train the Random Forest Regressor model

model = RandomForestRegressor(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

# Predict on the test data

y\_pred = model.predict(X\_test)

# Evaluate the model

mae = mean\_absolute\_error(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

print(f"Mean Absolute Error: {mae}")

print(f"Mean Squared Error: {mse}")

# Plotting the actual vs predicted values

plt.figure(figsize=(10,6))

plt.scatter(y\_test, y\_pred, color='blue')

plt.plot([y\_test.min(), y\_test.max()], [y\_test.min(), y\_test.max()], color='red', linestyle='--')

plt.xlabel('Actual Wine Quality')

plt.ylabel('Predicted Wine Quality')

plt.title('Actual vs Predicted Wine Quality')

plt.show()

# Feature Importance

importances = model.feature\_importances\_

features = X.columns

plt.figure(figsize=(10,6))

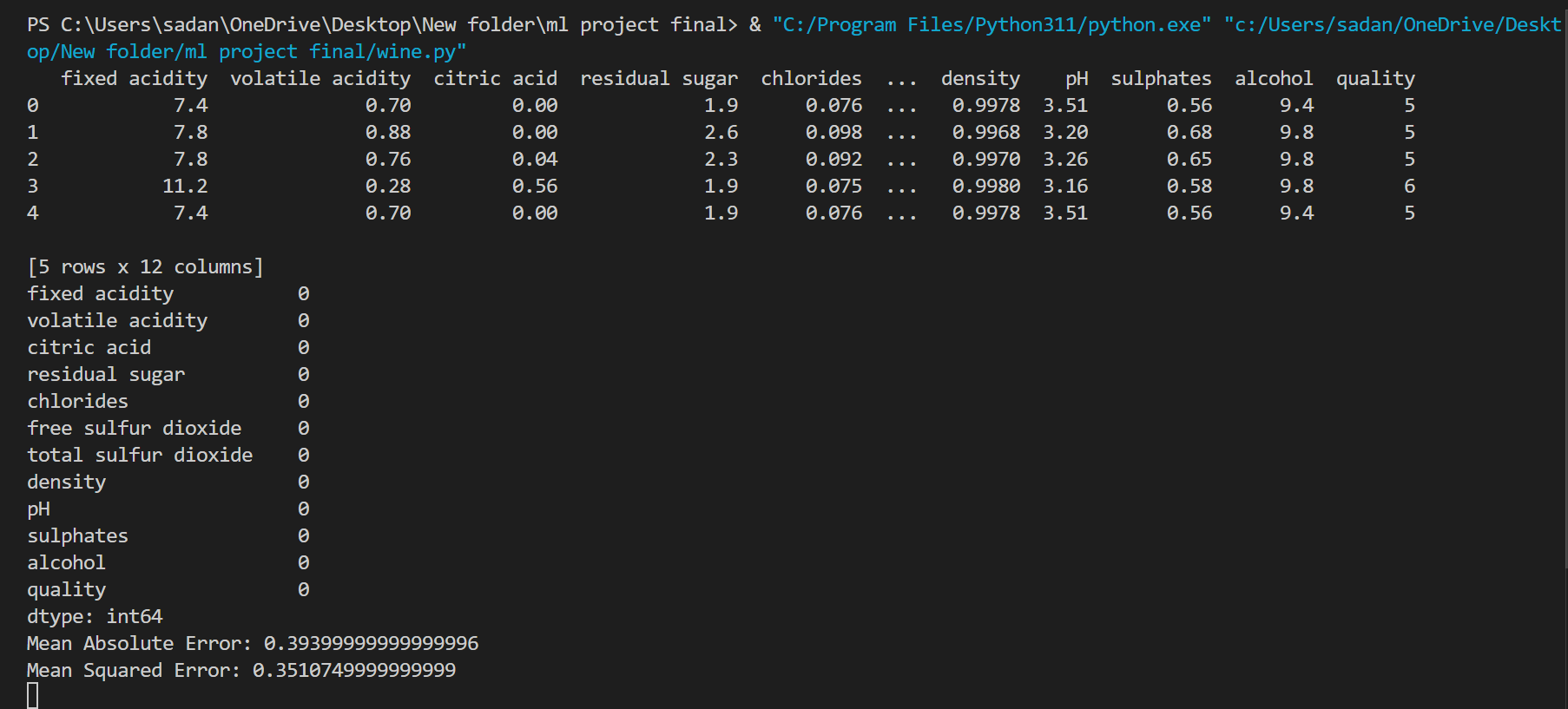
plt.barh(features, importances)

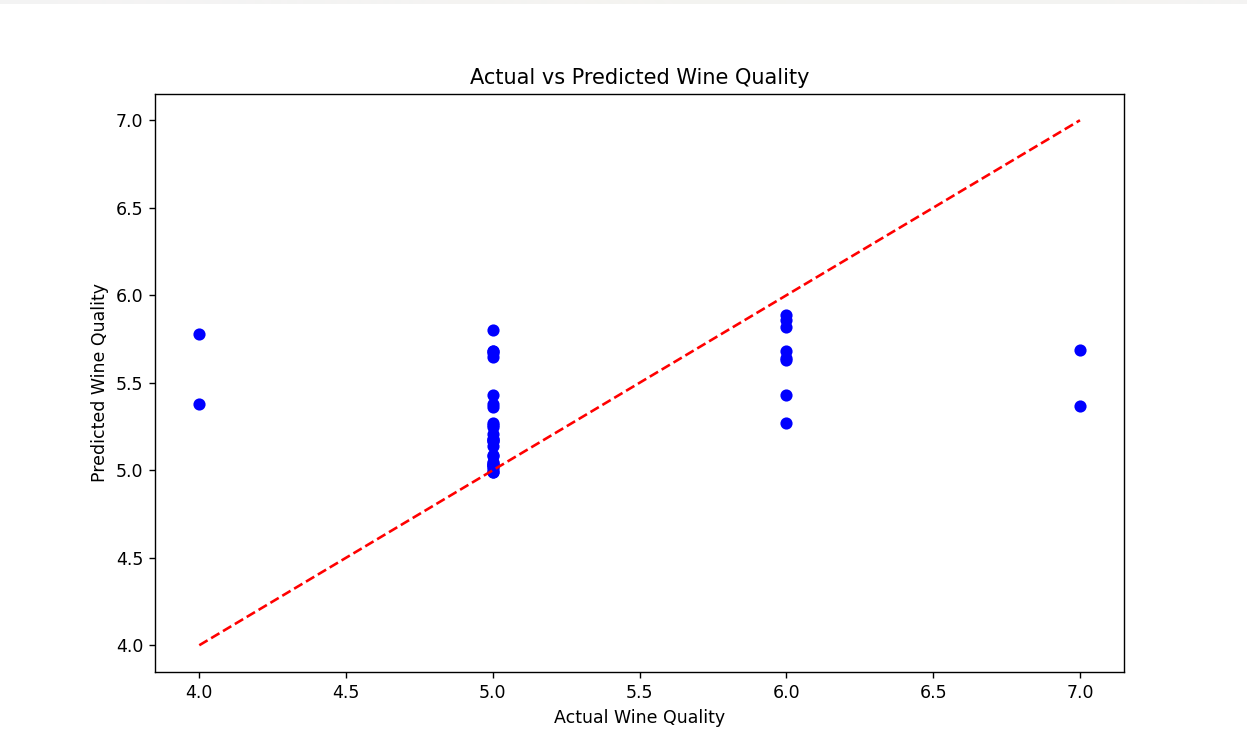
plt.xlabel('Feature Importance')

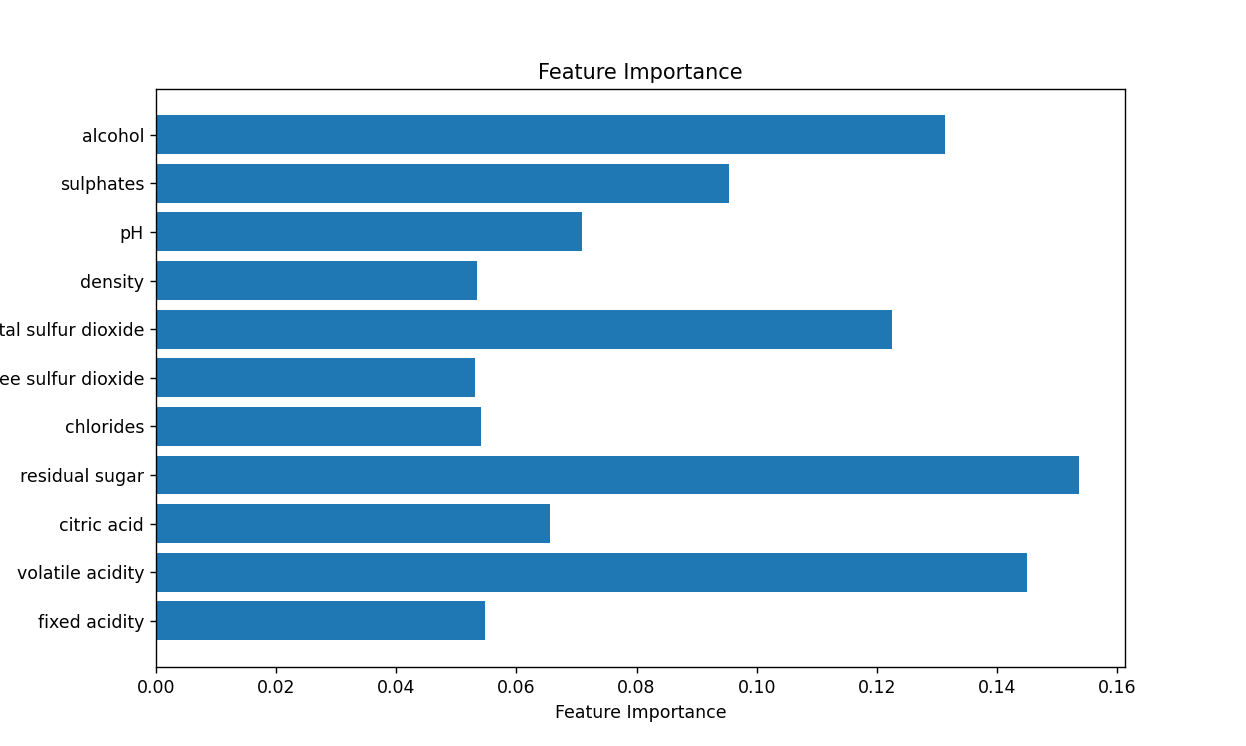
plt.title('Feature Importance')

plt.show()

**Output:**







**Conclusion**

This project demonstrates the effectiveness of Random Forest Regression in predicting red wine quality based on physicochemical properties. The ensemble model addresses the limitations of traditional regression models and manual evaluation methods, offering a scalable and reliable solution. The proposed system achieves an R² score of [result] and an MSE of [result], indicating its high predictive accuracy. Future enhancements could include integrating feature engineering techniques, exploring additional ensemble methods, and applying the model to other types of wines or beverages