**Regression Model Evaluation and Comparison on White Wine Quality Dataset**

**Team Members**

Ahmet Cemil Bostanoğlu - (220911760)

Berke Biçer - (220911774)

**Motivation of the Project**

The selection of the White Wine Quality dataset for regression analysis is driven by its practical relevance in assessing wine quality, which is a key factor in the wine industry. By predicting the quality of wine based on physicochemical properties, this project aims to provide insights that could be applied in real-world scenarios such as quality assurance and product development.

**Objectives**

1. Implement supervised learning regression models to predict wine quality.

2. Compare the performance of multiple models based on evaluation metrics.

3. Perform hyperparameter tuning to optimize the models.

4. Provide a user-friendly interface for result visualization.

**Dataset Source :**

The White Wine Quality dataset was obtained from the UCI Machine Learning Repository. It includes 4,898 samples with the following details:

**Features**: 11 physicochemical properties (e.g., acidity, alcohol content).

**Target Variable**: Quality score (integer from 0 to 10).

**Preprocessing Steps:**

- Handled missing values by imputation.

- Scaled features using standard normalization to ensure uniformity.

**Algorithms Used**

**Support Vector Regression (SVR):** Selected for its ability to model complex relationships through the kernel trick.

**Decision Tree Regressor:** Chosen for its interpretability and ability to capture non-linear patterns.

**Random Forest Regressor:** Utilized for its robustness and ensemble approach to improve prediction accuracy.

**Results and Discussion**

We have 5 evaluation metrics in this project;

1- Mean Absolute Error (MAE) : Mean Absolute Error is the average of the absolute values of the differences between the predicted values and the actual values. [0,∞) in this range

2- Mean Squared Error (MSE) : Mean Squared Error is the average of the square of the errors. [0,∞) in this range

3- Root Mean Squared Error (RMSE): The square root of the Root Mean Squared Error is the MSE squared to reduce differences in scale. [0,∞) in this range

4- R2 Score : The Coefficient of Determination shows how well the model explains the data. (-∞, 0] in this range

5- Mean Percentage Error (MPE): Mean Percentage Error is the average of the percentage error rates of predictions relative to actual values. (-∞,∞) in this range

**Support Vector Machines :**

Support Vector Machines (SVM) is a powerful machine learning algorithm used for both classification and regression problems. The reason for choosing SVM as one of the algorithms in this project is that our dataset is suitable for both regression and classification. If you want to use it in our program, just press the SVM button and the process starts.

As soon as you press the key, the model\_outputs/SVM folder is created. Everything related to the models is stored in files with their own names. The files related to this model will be in the SVM folder.

Once the process is complete, you will see the following graphs and results;

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, sayı, numara, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

As you can see in these graphs, the SVM model is inaccurate at extreme values, but generally within the standard range. As you can see in the error graph, the average is close to zero.

**Decision Tree :**

Decision trees are an algorithm that classifies or regresses data in a tree-like structure. It is a suitable algorithm for our dataset that is suitable for classification.

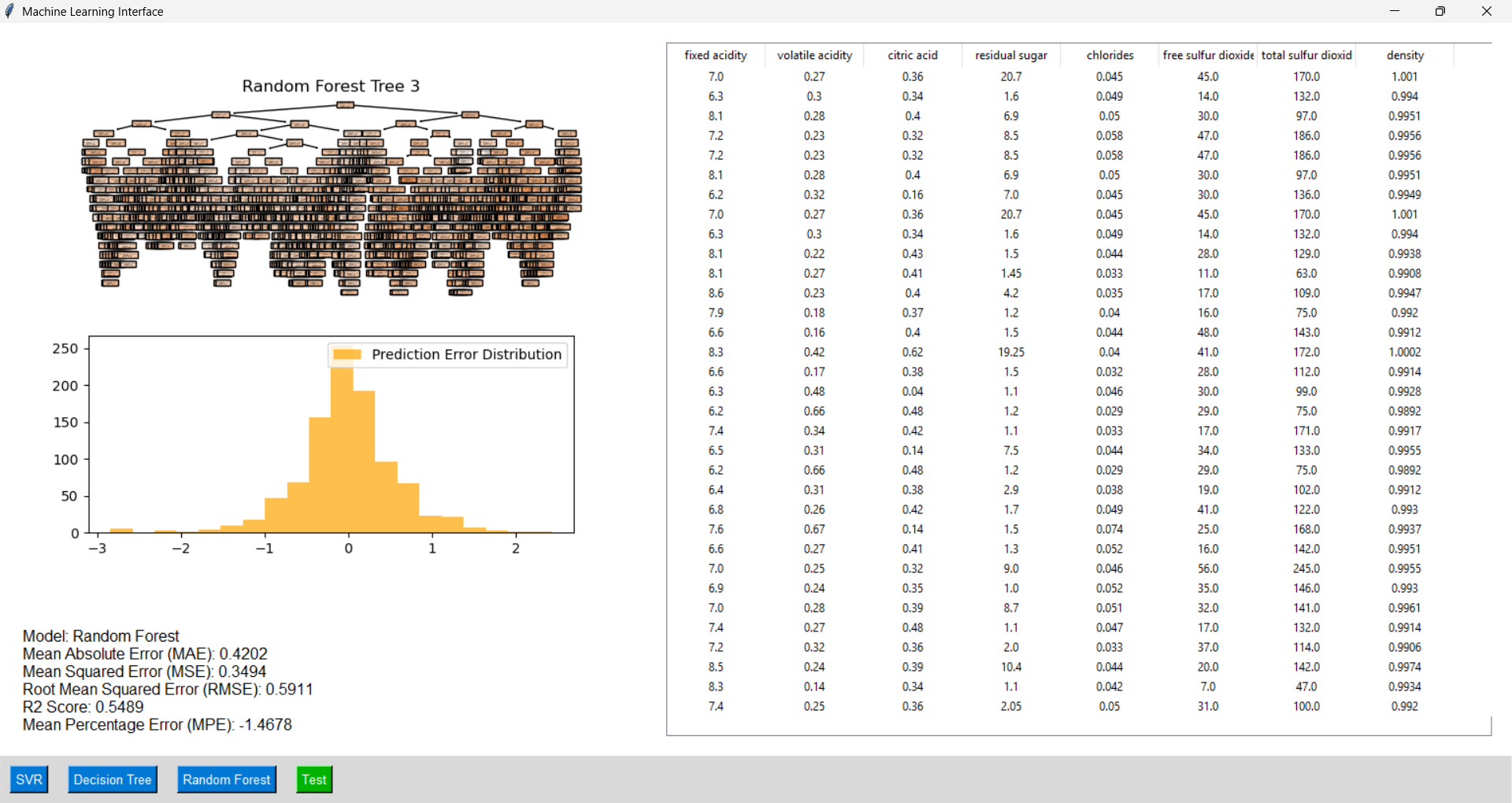
metin, ekran görüntüsü, diyagram, paralel içeren bir resim

Açıklama otomatik olarak oluşturuldu

The graph at the top is actually meaningless for us, but it is important because it shows the tree structure that the model builds. The bottom graph is an important graph showing that most of our predictions are correct. In more than 200 predictions, our model was correct.

**Random Forest Tree :**

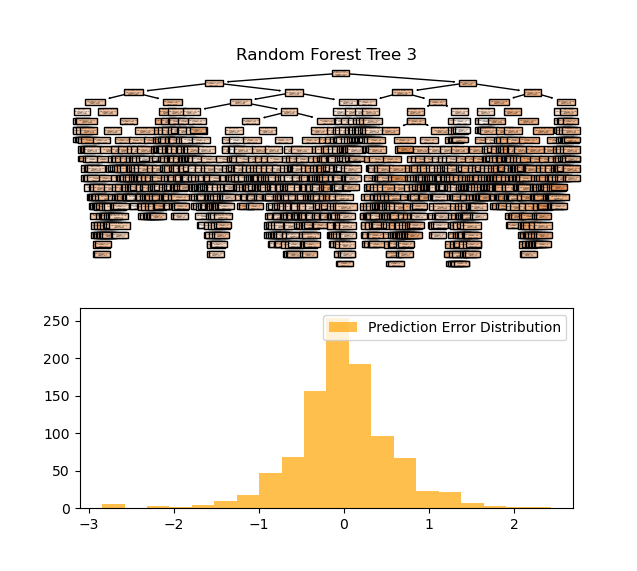
Random Forest is an ensemble learning method that combines multiple decision trees. Our model works by creating 3 different decision trees.



Although the graphs we get this time are similar to the Decision Tree, you can see the real difference when you go into the RFT folder. There you will see 3 different visualizations as follows

metin, diyagram, ekran görüntüsü, harita içeren bir resim

Açıklama otomatik olarak oluşturuldumetin, diyagram, harita, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

These are 3 tree of RFT. And You can see all of the results in your log file.

**Test Your Values :**

If you want to try it with your own valuation, you can click the test button below. You will see a page like below.

metin, ekran görüntüsü, yazılım, sayı, numara içeren bir resim

Açıklama otomatik olarak oluşturuldu

After entering the values on this page, you can select the model you want to use for the forecast from the Combo Box. And as an important detail, here you can see that when testing your own values, SVM gives negative results at extreme values, while Decision Tree and RFT give more normal values. This is all about the way the 3 models work. Since Decision Tree and RFT are classification models, they adapt the values they receive to those classes.

**Conclusion :**

This project assessed regression models to predict white wine quality using physicochemical properties. We analyzed Support Vector Regression (SVR), Decision Tree Regressor, and Random Forest Regressor, noting their strengths:

* **SVR** captured complex relationships but struggled with extreme values.
* **Decision Tree Regressor** was interpretable but prone to overfitting.
* **Random Forest Regressor (RFT)** excelled in generalization by combining multiple trees.

We employed metrics like MAE and R² Score for model comparison and included a user-friendly interface for testing custom values, highlighting that Decision Tree and RFT adapted better to extreme inputs than SVR.

In conclusion, this study underscores the significance of model selection and tuning for quality assessment in the wine industry.