Model Results Documentation

1. Overview

This document summarises and explains the results obtained from applying different machine learning models and algorithms to the dataset. The aim is to compare model performance, explain why the results differ, and clarify the key differences between the models used.

The models evaluated include:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor

Model performance was evaluated using the following metrics:

- R^2 Score
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

2. Summary of Results

Each model produced different performance scores on the test data. These differences are expected and normal in machine learning because each algorithm learns patterns in a different way.

In general:

- Linear Regression provided a baseline performance.
- Decision Tree Regression captured non-linear relationships better than Linear Regression.
- Random Forest Regression achieved more stable and improved performance compared to a single Decision Tree.

3. Why the Results Are Different

The results differ due to several key reasons:

3.1 Model Assumptions

- **Linear Regression** assumes a linear relationship between input features and the target variable.
- **Decision Trees** make no linearity assumptions and can model complex, non-linear relationships.
- **Random Forest** combines multiple decision trees, reducing overfitting and improving generalisation.

3.2 Model Complexity

- Linear Regression is a simple model with low complexity.
- Decision Trees are more complex and can easily fit training data very closely.
- Random Forest increases complexity but controls overfitting by averaging multiple trees.

3.3 Sensitivity to Data

- Linear Regression is sensitive to outliers and multicollinearity.
- Decision Trees are sensitive to small changes in data.
- Random Forest is more robust to noise and data variation.

3.4 Bias–Variance Trade-off

- Linear Regression has **high bias** and **low variance**.
- Decision Trees have **low bias** but **high variance**.
- Random Forest balances bias and variance effectively.

4. Differences Between the Models and Algorithms

4.1 Linear Regression

- Type: Parametric, linear model
- Strengths:
  - Simple and easy to interpret
  - Fast to train
- Weaknesses:
  - Cannot model non-linear relationships well
  - Sensitive to outliers

4.2 Decision Tree Regressor

- Type: Non-parametric, rule-based model
- Strengths:
  - Handles non-linear relationships
  - Easy to visualise and interpret
- Weaknesses:
  - Prone to overfitting
  - Sensitive to noise in data

4.3 Random Forest Regressor

- Type: Ensemble learning model
- Strengths:
  - Higher accuracy and stability
  - Reduces overfitting
  - Handles complex feature interactions
- Weaknesses:
  - Less interpretable than single trees
  - Higher computational cost

5. Interpretation of Evaluation Metrics

- **R^2 Score**: Indicates how well the model explains variance in the target variable. Higher values are better.
- **MSE**: Penalises large errors more heavily. Lower values indicate better performance.
- **RMSE**: Square root of MSE, expressed in the same units as the target variable.
- **MAE**: Average absolute error, easier to interpret and less sensitive to outliers.

6. Conclusion

The differences in results are due to the nature of the algorithms, their assumptions, and how they handle complexity and data patterns. While Linear Regression provides a good baseline, tree-based models perform better when the relationship between variables is non-linear. Random Forest generally offers the best balance between accuracy and generalisation.

Selecting the best model depends on the problem requirements, interpretability needs, and acceptable computational cost.