Academic Year	Module	Assignment Number	Assessment Type
2024/25	5CS037: Concepts and Technologies of AI	3	Report

Regression Analysis Report

Student Id : 2408286

Student Name : Arjabi Shrestha

Section : L5CG15

Module Leader : Siman Giri

Tutor : Siman Giri

Submitted on : 11-02-2025

Table of Contents

ABSTRACT

The goal is to predict a Continuous target variableoutcome, Gender Inequality Index(GII). The model aims to capture relationships between various socio-economic factors and GII to provide accurate predictions and insights.

The dataset used for this analysis is sourced from Kaggle which was originally developed by the United Nations Development Programme (UNDP). It includes socio-economic indicators for various countries such as GII, maternal mortality, adolescent birth rate,etc. The target variable for this task is GII.

EDA involves handling missing values were by dropping the rows and visualization of data. Correlation between GII and other social economic indicators were are observed.

Model building classification techniques includes Linear Regression and Random Forest Regressor. Hyperparameter were optimization using GridSearchCV for both the models.

The performance of the models was evaluated using R^2, RMSE and MSE.

Model	**Test R^2**	**Training R^2**	**MAE**	**RMSE**
Linear Regression	0.899	0.866	0.052	0.066
Random Forest Regression	0.946	0.988	0.033	0.049
Hypertuned Random forest regression	0.955	0.990	0.031	0.043

The Test R^2 and Training R^2 values are quite similar across all models, indicating that none of the models suffer from overfitting. Among the models, the the Hypertuned Random Forest Model gives the best performance with the highest Test R^2 (0.955), Training R^2 (0.990), and the lowest MAE (0.031) and RMSE (0.043), making it the most suitable choice for predicting the target variable.

1. INTRODUCTION

1.1. Problem Statement

Regression problems involve predicting a continuous output variable based on input features. A model learns to establish a relationship between the input variables and the target variable(continuous). This trained model is then used to predict the target value for new, unseen data by estimating the most likely value based on the learned relationship.

1.2. Dataset

This Regression analysis this dataset was based one was sourced from Kaggle, originally developed by the United Nations Development Programme (UNDP). It contains socio-economic indicators that influence human development and gender inequality.

It contains 196 rows × 11 columns with the target variable being GII. The Independent Variables include Country Male secondary education attainment (%), Female labor force participation (%), Male labor force participation (%), Maternal mortality rate (deaths per 100,000 live births), Adolescent birth rate (births per 1,000 women ages 15–19), Seats held by women in parliament (%), Female secondary education attainment (%).

This project is directly aligned with SDG 5: Gender Equality, which seeks to eliminate gender disparities in education, employment, and political participation, and to empower women globally. By predicting the Gender Inequality Index (GII), this project aims to identify socio-economic factors that contribute to gender inequality. The findings can provide valuable insights to inform policies and actions that promote gender equality and empower women, supporting the broader goal of reducing global gender disparities.

1.3. Objective:

The objective of this analysis is to build a predictive regression model that estimates the **Gender Inequality Index (GII)** based on the socio-economic features provided in the dataset. By using these features, the model aims to understand the relationships between these variables and gender inequality.

2. Methodology

2.1 Data Preprocessing

Before building the model, the data was cleaned by handling missing values. There were no outliers in the dataset.

2.2 Exploratory Data Analysis (EDA)

EDA was performed using better understand the data. Key insights from the EDA include:

Mean GII: 0.3444

Median GII: 0.3630

Standard Deviation of GII: 0.1971

The mean GII of 0.3444 indicates moderate average gender inequality across countries. The median GII of 0.3630 suggests that half of the countries have a GII below this value, with moderate skewness in the data. The standard deviation of 0.1971 shows a moderate level of variability in GII values, indicating differences in gender inequality levels across countries.

The highest and Lowest GII was found to be:

Highest GII: Yemen (0.8200)

Lowest GII: Denmark (0.0130)

Filtering and Sorting:

A new column called GII_Score_Category was created to categorize countries into three categories based on their GII score:

Low: Countries with scores less than the first quartile (Q1).

Medium: Countries with scores between Q1 and Q3 (inclusive).

High: Countries with scores greater than the third quartile (Q3).

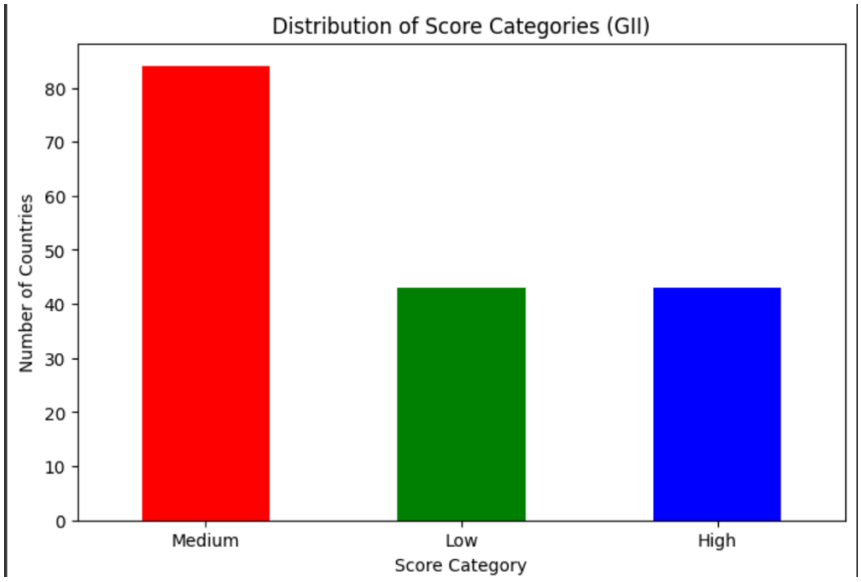


Figure 1: Distribution of Score Categories (GII)

Several plots were made to vizualize the data:

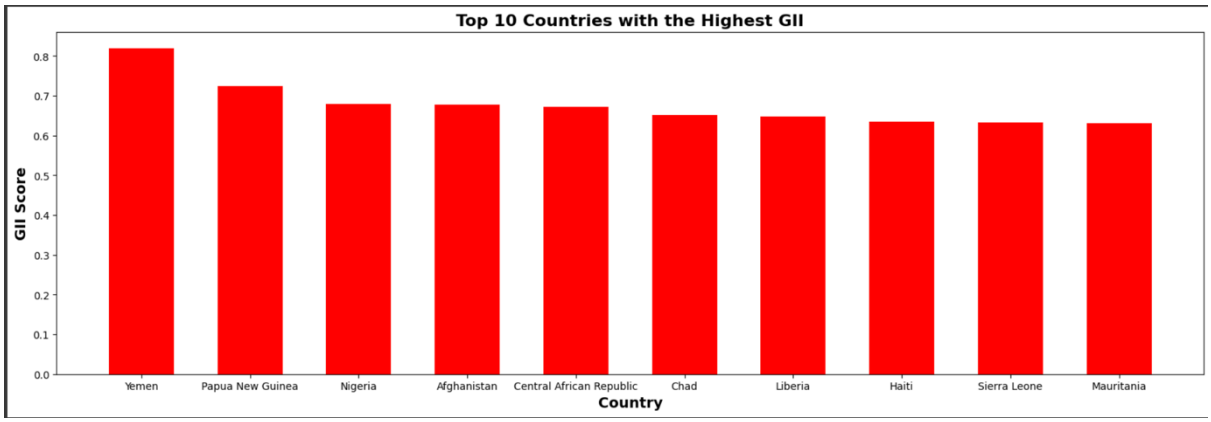
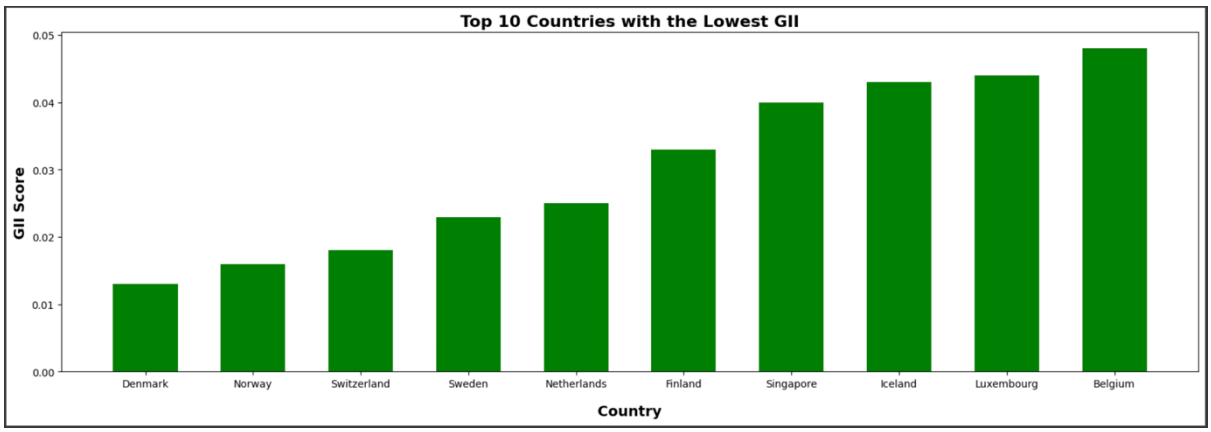


Figure 2:Top 10 Countries with the Highest GII



Histogram:

Figure 3:Top 10 Countries with the Highest GII

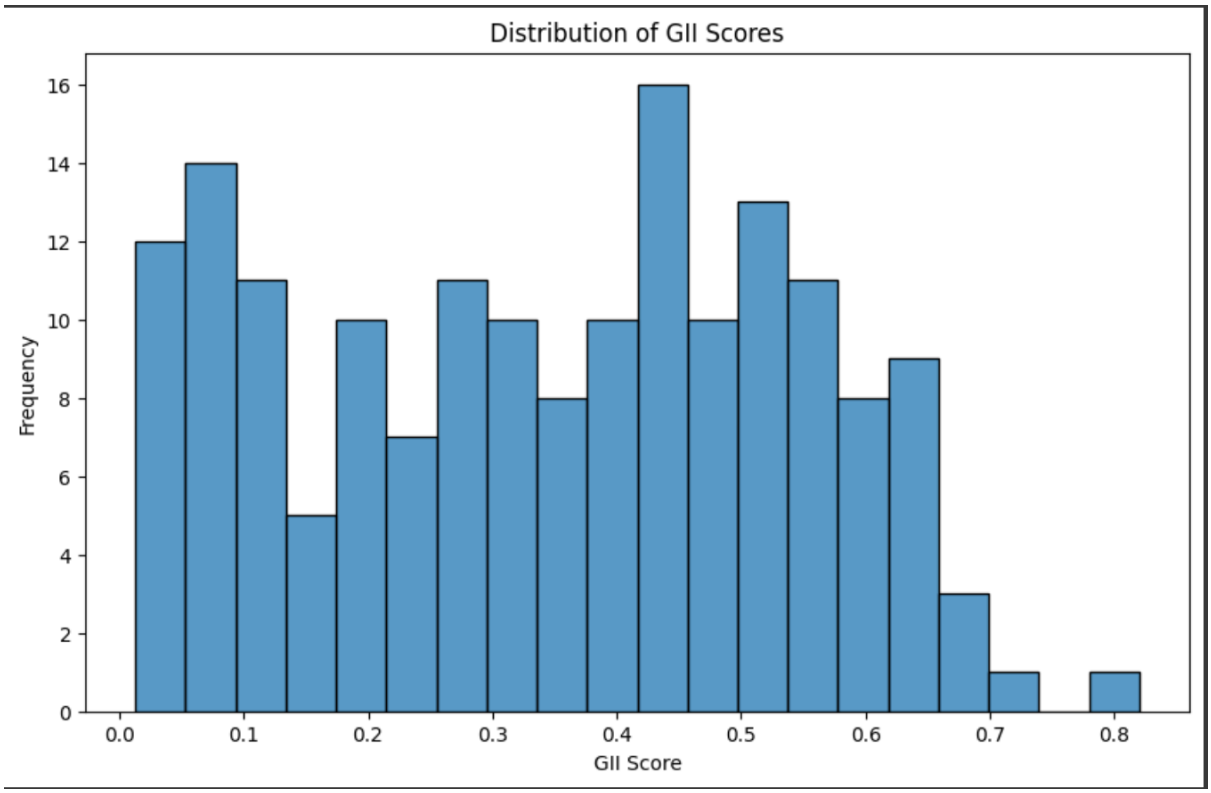


Figure 4:Distribution of GII Scores

This histogram displays the distribution of Gender Inequality Index (GII) scores. The Skewness of GII scores: -0.0165 which indicates a relatively symmetrical distribution

Scatter Plot:

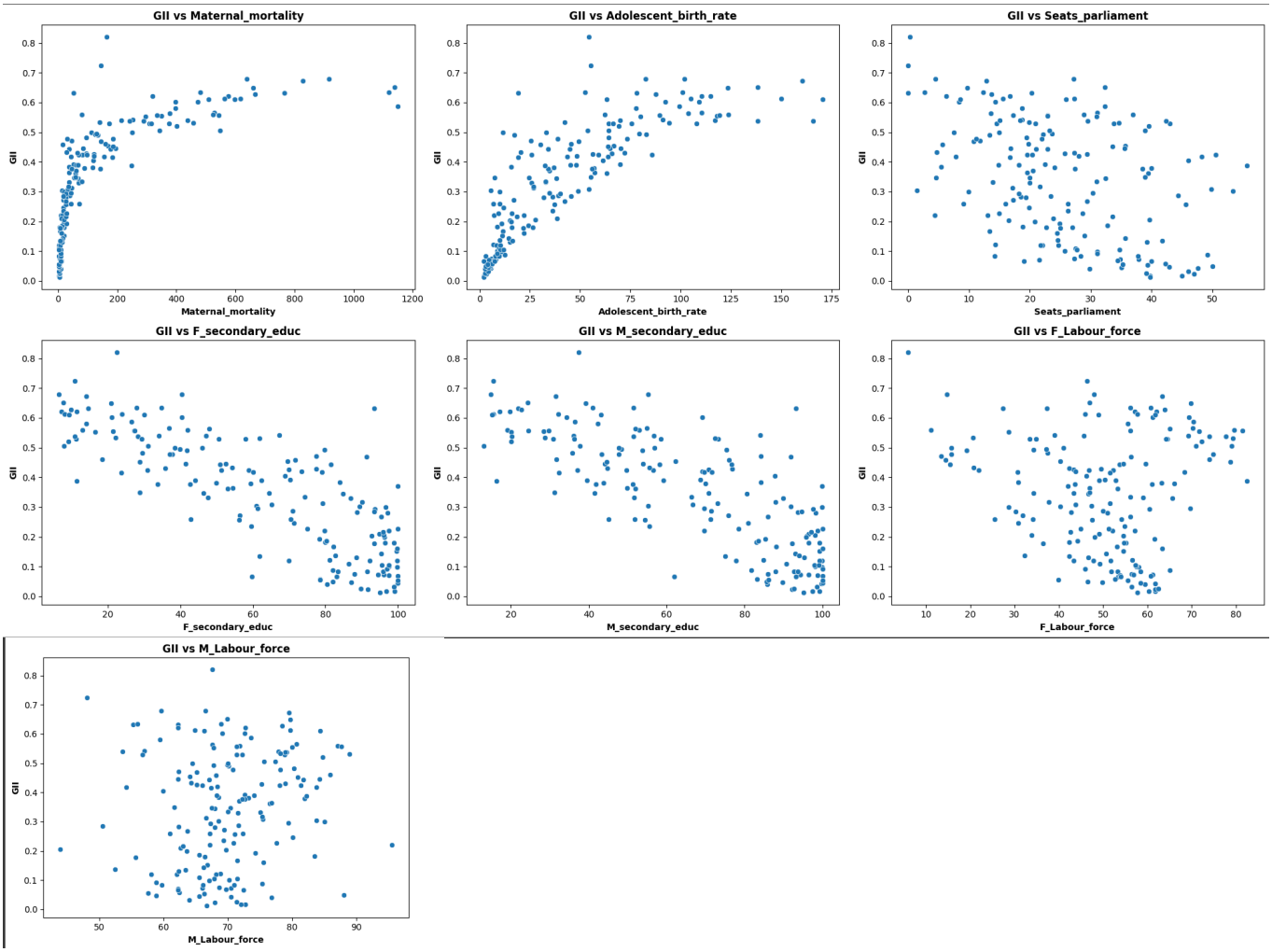


Figure 5:Scatter plot: GII vs Feature Variables

Scatter plot were plotted to help visualize how the target variable (GII) relates to each of the feature variables to better understand the dataset. By plotting GII against other variables, you can visually check for correlation patterns.

The Pearson's Correlation was also calculated to be

Maternal_mortality 0.713515

Adolescent_birth_rate 0.806791

Seats_parliament -0.424116

F_secondary_educ -0.809278

M_secondary_educ -0.782130

F_Labour_force -0.070970

M_Labour_force 0.158270

Box plot:

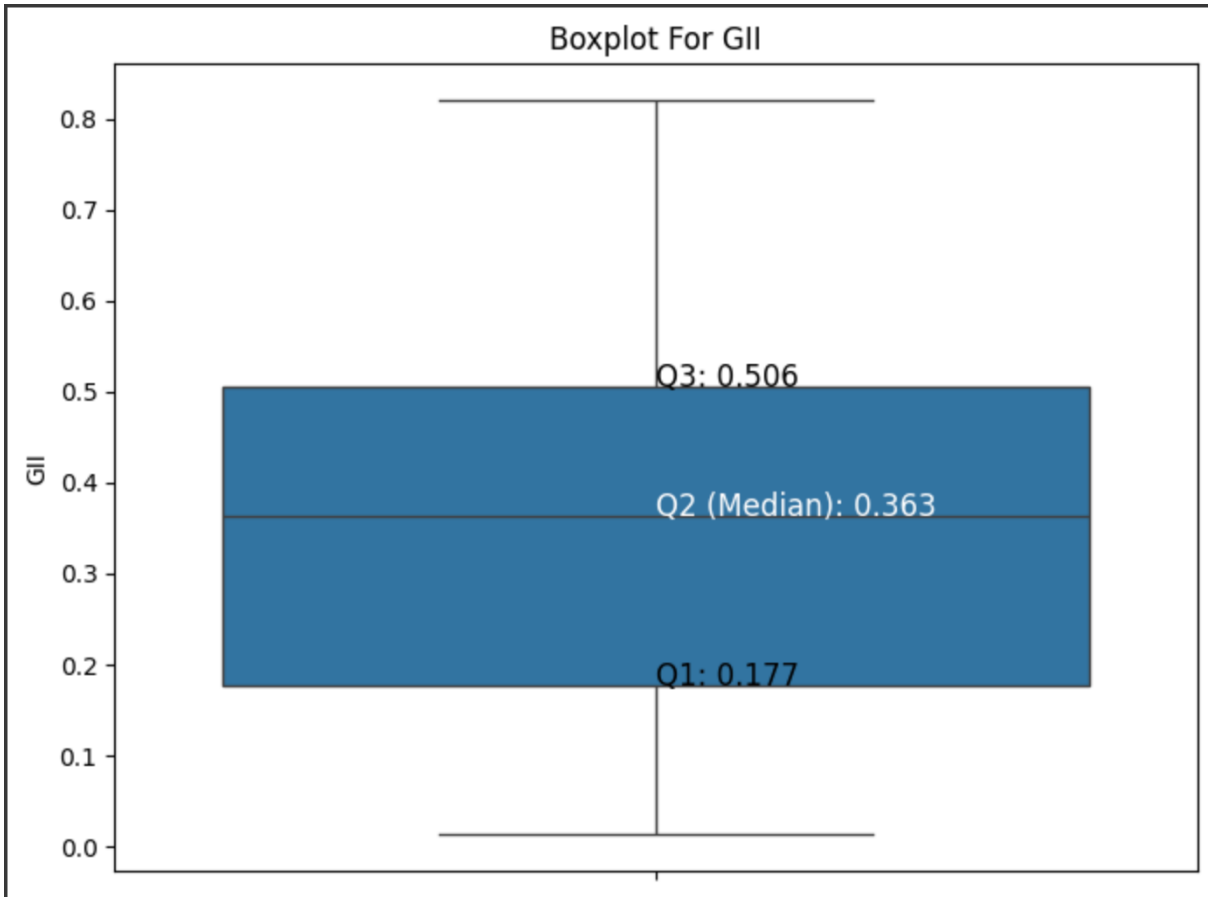


Figure 6: box plot for GII

A box plot was created to visualize the distribution of the Gender Inequality Index (GII). There are **no outliers** outside this range. This indicates that the GII values are relatively consistent and do not contain extreme values.

2.3. Model Building

Two classification models were considered for this task:

1: Linear Regression

2: Random Forest Regressor

The models were splitting the data into 70% training and 30% testing and the StandardScaler was used on the features to ensure that all variables were on the same scale.

The model was trained on the scaled training data using Scikit-learn's LinearRegression() for linear regression and RandomForestRegressor() for random forest regressor

2.4. Model Evaluation

The model's performance was evaluated using several key metrics that are essential for understanding its effectiveness, particularly when dealing with imbalanced classes. The key metrics used were:

R^2: Measures how well the model explains variance in the target variable.

MAE (Mean Absolute Error): Represents the average absolute prediction error.

RMSE (Root Mean Absolute Error): Penalizes larger errors

The following are the observed data for both the model:

Model	**Test R^2**	**Training R^2**	**MAE**	**RMSE**
Linear Regression	0.899	0.866	0.052	0.066
Random Forest Regressor	0.946	0.988	0.033	0.049

The Linear Regression model shows a Test R^2 of 0.899, Training R^2 of 0.866, MAE of 0.052, and RMSE of 0.066, indicating moderate performance with some prediction errors.

Whereas the Random Forest Regressor model shows a Test R^2 of 0.946, Training R^2 of 0.988, MAE of 0.033, and RMSE of 0.049, reflecting higher accuracy and lower prediction errors.

2.5. Hyper-parameter Optimization

GridSearchCV were used on both the models to try to enhance the model's performance.

Optimization for Model 1: Linear Regression

For Linear Regression, the following optimal hyperparameters alpha was found to be 10.0

Optimization for Model 2: Random Forest Regressor

For Random Forest Classifier, Best Parameters were:

'max_depth': 20,

'max_features': None,

'min_samples_leaf': 1,

'min_samples_split': 2,

'n_esn_estimators: 100

2.6. Feature Selection

Feature Selection for model 1: Linear Regression:

Recursive Feature Elimination (RFE) was used to select the most important features. The top 3 selected features were: 'Adolescent_birth_rate', 'Seats_parliament', 'F_secondary_educ'

Feature Selection for model 2: Random Forest Regressor:

Feature Importances were used to determine the key features in random forest. The most important features selected were: ['Maternal_mortality', 'Adolescent_birth_rate', 'Seats_parliament'

3. Conclusion

3.1. Key Findings

The Random Forest Regressor outperforms Linear Regression with a Test R^2 of 0.946 and Training R^2 of 0.988, indicating a better ability to explain variance in both test and training data. The MAE of 0.033 and RMSE of 0.049 are both lower than those of the Linear Regression model, meaning the Random Forest model performs better for this model

3.2. Final Model

Based on Random Forest Regressor, a final model was built using the optimal hyperparameters and selected features. The key findings are as follows:

Optimized Random Forest Regressor:

MAE (Test): 0.031

RMSE (Test): 0.043

Training R^2: 0.990

Test R^2-squared: 0.955

3.3. Challenges

The first challenge included handling a small dataset of 196 rows × 11 columns, which limited model generalization and increased the risk of overfitting. While scikit-learn's feature selection methods were used to identify the most impactful features, and it was difficult to ensure that the selected features captured valuable information.

3.4. Future Work

For future works, searching a dataset with more data would help the model generalize better and reduce overfitting.

4. Discussion

4.1. Model Performance

The Optimized Random Forest Regressor demonstrates strong performance with impressive results across various metrics. The MAE (Test) of 0.031 and RMSE (Test) of 0.043 indicate minimal prediction errors, suggesting that the model provides accurate and reliable predictions. With a Test R^2 of 0.955 the model generalizes well to unseen data.

4.2. Impact of Hyperparameter Tuning and Feature Selection

By adjusting hyperparameters like the number of trees, tree depth, and splitting criteria, the model became more accurate and efficient, significantly reducing errors like MAE and RMSE. This allowed the model to capture more complex patterns in the data while avoiding overfitting.

Through techniques like Feature Importance and Recursive Feature Elimination (RFE), helped identify the most relevant variables affecting GII, by focusing on these important features ensuring that the model wasn't influenced by irrelevant or redundant data.

4.3. Interpretation of Result

The results from the Random Forest Regressor and its tuned version align with the expectations, confirming that the model performs well in predicting the target variable, GII.

After tuning the hyperparameters, the Tuned Random Forest Regressor showed even better performance, with a Test MAE of 0.31->0.033 and Test RMSE of 0.049->0.043, indicating a further reduction in prediction errors. The Training R^2 of 0.990 and Test R^2 of 0.955 reflect the model's strong fit to the data and its ability to generalize effectively.

Overall, the tuned model showed slight improvement in prediction the correct value of GII.

4.4. Limitation

Firstly, there is a risk of overfitting, as indicated by the relatively high R^2, meaning the model might perform well on the training data but not generalize as effectively to new data. The Random Forest Regressor performed well overall for this dataset, but the potential for overfitting persists. Additionally, the dataset is small (196 rows) so it could affect the model's accuracy and ability to generalize.

4.5. Suggestions for Future Research

For future research, expanding the dataset to include more socio-economic variables would improve model generalization. Further fine-tuning hyperparameters using advanced techniques could be explored.