Project Title: Predicting Gender Inequality Index Using Machine Learning Models

1. Introduction

The United Nations Sustainable Development Goals (SDGs) highlight gender equality as one of the critical goals (SDG 5) for sustainable development. Addressing gender inequality remains a global challenge, as it affects multiple areas, including economic growth, healthcare, education, and governance. To make data-driven decisions and foster policies that reduce gender inequality, it is essential to analyze and predict gender inequality levels in different regions. In this project, we use machine learning techniques to predict the **Gender Inequality Index (GII)** based on socio-economic and demographic factors.

2. Problem Definition

The goal is to develop a predictive model for the **Gender Inequality Index (GII)**, which reflects gender-based disparities across countries. By examining factors such as maternal mortality, female education levels, labor force participation rates, and parliamentary representation, we aim to predict the GII for different countries, providing a tool for assessing where improvements might be needed.

3. Dataset

The dataset used in this project provides information on gender inequality across various countries, including factors like:

- **Maternal Mortality Rate**: Number of maternal deaths per 100,000 live births.
- **Adolescent Birth Rate**: Birth rate per 1,000 women aged 15-19.
- **Seats in Parliament**: Percentage of parliamentary seats held by women.
- **Secondary Education**: Percentage of male and female population with secondary education.
- **Labor Force Participation**: Percentage of male and female population in the labor force.

These indicators are crucial for understanding gender inequality levels and predicting the GII.

4. Models Chosen and Justification

In this project, we have chosen two machine learning models to predict GII: **Linear Regression** and **Random Forest Regressor**.

Linear Regression

Linear Regression is a straightforward model that assumes a linear relationship between the predictors and the target variable. Since GII is a continuous value, linear regression can provide a quick benchmark for understanding the dataset's behavior. This model is valuable because:

- It’s interpretable and allows us to understand the direct relationships between the factors and GII.
- It works well with numerical data and can provide insights into which variables contribute most to changes in GII.

Random Forest Regressor

Random Forest Regressor is an ensemble model that combines multiple decision trees to make robust predictions. We chose this model because:

- It can capture complex, non-linear relationships within the data, which may exist among socio-economic factors affecting GII.
- It handles missing values and outliers better than linear regression, providing more accurate and reliable predictions.
- It provides feature importance scores, helping us identify the most influential factors in predicting GII.

5. Methodology

The project follows these key steps:

1. **Data Preprocessing**: Imputed missing values for features and the target variable (GII). Label-encoded categorical variables for compatibility with the models.
2. **Model Training and Validation**: Trained both models using **K-Fold Cross-Validation** to ensure robust evaluation and prevent overfitting. We used **Mean Squared Error (MSE)** and **Mean Absolute Error (MAE)** as performance metrics.
3. **Evaluation and Comparison**: Compared the performance of Linear Regression and Random Forest Regressor to determine which model performs best for predicting GII.

6. Results and Conclusion

The results of both models were evaluated based on the MSE and MAE scores, allowing us to assess the models' predictive accuracy. The **Random Forest Regressor** generally performed better due to its ability to model complex relationships, while **Linear Regression** offered interpretability and provided baseline predictions. The model outcomes can guide further analysis on influential factors in gender inequality and support policy-making efforts to reduce disparities.