HUMBER INSTITUTE OF TECHNOLOGY
AND ADVANCED LEARNING
(HUMBER COLLEGE)

Enhancing Operational Efficiencies Through Sustainable Development Goals Analysis

Name	Student Number
Taniya Ketan Gadkari	N01579357
Devisha Bhayani	N01578727
Harshit Aggarwal	N01550495
Hira Shaikh	N01452306
Qi Shi	N01578951
Ravneet Singh	N01578955

Submitted to: Professor Lubna Mohammed

Submission Date: August 16, 2024

Introduction

The Sustainable Development Goals (SDGs) were adopted by the 193 UN Member States in 2015 as a universal call to action to end poverty, protect the planet, and ensure peace and prosperity for all by 2030. The Sustainable Development Report (SDR) serves as an annual review of the progress made towards achieving these goals. The SDR 2023, published on the eve of the 2023 Paris Summit for a New Global Financial Pact, specifically emphasizes the urgent need to scale up development finance and reform global financial systems to support the SDGs.

This project leverages the SDG Index data from 2000 to 2023, focusing on various countries to analyze and predict the factors that significantly influence the overall SDG Index scores and individual goal performances. By developing predictive models, this study aims to provide actionable insights that can guide policymakers in their efforts to achieve the SDGs by 2030.

Business Case

Achieving the SDGs is critical for ensuring sustainable development on a global scale. However, the progress has been uneven across countries and regions, and the COVID-19 pandemic has further complicated these efforts. The ability to predict future SDG Index scores based on historical data can help policymakers and stakeholders identify areas that require urgent attention and allocate resources more effectively. This project provides a data-driven approach to understanding the dynamics behind SDG performance and offers predictive tools that can support strategic decision-making and accelerate progress towards the 2030 agenda.

Business Problem

The primary challenge addressed in this project is the identification of key determinants that influence a country's SDG Index score. Understanding these factors is crucial for improving overall performance and ensuring that no country is left behind in the global pursuit of sustainable development. The project also seeks to develop and select the most efficient predictive models that can forecast future SDG Index scores, enabling proactive interventions to meet the SDG targets by 2030.

Objectives of our study

1. To analyze the historical SDG Index data from 2000 to 2023 for various countries.
2. To identify the key factors that significantly influence a country's SDG Index score and individual goal performances.
3. To develop, compare and select predictive models to forecast future SDG Index scores based on historical data.
4. To provide actionable recommendations for policymakers to target interventions more effectively and accelerate progress towards the SDGs.

Literature Review

1. In “Sustainable development goals: A need for relevant indicators”, the United Nations Open Working Group proposed 17 Sustainable Development Goals (SDGs) and 169 targets, supported by a preliminary set of 330 indicators introduced in March 2015. While some SDGs build upon previous Millennium Development Goals, others introduce new concepts. Despite substantial theoretical work on indicator quality, practical challenges remain in ensuring that indicators accurately measure the intended phenomena. This review highlights the need for operationalizing SDG targets and evaluating the relevance of indicators. It argues for the development of a conceptual framework to guide the selection and formulation of appropriate indicators, emphasizing the importance of the “indicator-indicated fact” relationship. Clear, unambiguous indicators are essential for effective communication with decision-makers and the public. Recommendations are provided for indicator providers to support the creation of a robust final indicators framework. (Hák, T., Janoušková, S., & Moldan, B. (2015))

2. In “The sustainable development goals: A case study “, the Sustainable Development Goals (SDGs) represent a comprehensive approach to sustainability, but practical challenges remain in their implementation. This study explores how Tassal, Australia's largest salmon aquaculture company, perceived and engaged with the SDGs. Interviews with leaders, employees, and external partners revealed that Tassal was initially unaware of the SDGs but expressed interest in integrating them into their sustainability practices. Using the Values-Rules-Knowledge (vrk) framework, the research identified that corporate and personal values significantly influenced Tassal's positive response to the SDGs. The company recognized potential benefits in engaging with goals outside their immediate industry focus, such as health and well-being. The findings suggest that businesses can effectively engage with the SDGs by broadening their interpretation of sustainability and reflecting on their core values. The vrk model is proposed as a valuable tool for diagnosing organizational barriers to SDG adoption. The study also highlights the implicit importance of social license in sustainability, a concept not explicitly covered by the SDGs but crucial for social, economic, and environmental sustainability. (Fleming, A., Wise, R. M., Hansen, H., & Sams, L. (2017))

Methodology

Dataset Collection

The dataset for this project is sourced from Kaggle.com, the Sustainable Development Report 2023, which provides comprehensive data on sustainability and progress towards the SDGs for numerous countries. This dataset includes information on sustainability scores, regional classifications, and performance metrics for individual SDGs.

Description of the dataset

The SDG Index dataset comprises annual records from 2000 to 2023, detailing various countries' performances across all 17 SDGs. Each entry in the dataset includes the following attributes:

- Country Name
- Regional Classification
- Year
- SDG Index Score
- Scores for each of the 17 SDGs

This dataset allows for a nuanced assessment of global sustainability efforts and provides a basis for analyzing the factors that drive progress towards sustainable development.

	country_code	country	year	sdg_index_score	goal_1_score	goal_2_score	goal_3_score	goal_4_score	goal_5_score	goal_6_score	...	goal_8_score	goa
0	FIN	Finland	2023	86.760595	99.5750	60.886750	95.386385	97.169333	92.11125	94.3276	...	86.789000	
1	SWE	Sweden	2023	85.981397	98.8885	63.074125	96.904000	99.761667	91.44025	95.0576	...	84.966429	
2	DNK	Denmark	2023	85.683637	99.2155	71.025250	95.398500	99.339667	86.99800	90.7316	...	87.562429	
3	DEU	Germany	2023	83.358447	99.5105	72.366000	93.039357	97.162667	81.92025	88.4434	...	86.967286	
4	AUT	Austria	2023	82.280189	99.4510	73.067500	92.468000	97.914333	84.57925	92.1636	...	83.274143	

5 rows × 21 columns

Figure 1 Snapshot of the dataset

Sustainable Development Goals (SDG) Scores

(What each goal represents)
overall_score: Composite score representing overall performance on SDGs.
goal_1_score: Score for No Poverty.
goal_2_score: Score for Zero Hunger.
goal_3_score: Score for Good Health and Well-being.
goal_4_score: Score for Quality Education.
goal_5_score: Score for Gender Equality.
goal_6_score: Score for Clean Water and Sanitation.
goal_7_score: Score for Affordable and Clean Energy.
goal_8_score: Score for Decent Work and Economic Growth.
goal_9_score: Score for Industry, Innovation, and Infrastructure.
goal_10_score: Score for Reduced Inequalities.
goal_11_score: Score for Sustainable Cities and Communities.
goal_12_score: Score for Responsible Consumption and Production.
goal_13_score: Score for Climate Action.

goal_14_score: Score for Life Below Water.
goal_15_score: Score for Life on Land.
goal_16_score: Score for Peace and Justice Strong Institutions.
goal_17_score: Score for Partnerships to achieve the Goal.

Data preprocessing

Preprocessing of the SDG Index data involves several key steps to prepare it for analysis and modeling:

1.***Loading the Data***: The dataset is loaded from a CSV file into a panda DataFrame, and the first few rows are displayed to understand its structure.

2.***Data Type Verification***: The data types of each column are checked to ensure they are appropriate for analysis.

3.***Missing Values Check***: The code checks for missing values, which is crucial for maintaining data integrity. In this dataset, no missing values are found.

4.***Lagged Feature Creation***: A lagged feature for the SDG Index score is created to analyze trends over time. Any resulting NaN values from this operation are dropped to ensure a clean dataset.

The methodology in our study involves training machine learning models such as linear regression model, to identify the key factors that significantly influence a country's SDG Index score and individual goal performances; further developing models such as Random forest regression and Neural networks to compare and select the best in order to forecast future SDG Index scores based on historical data and provide actionable recommendations for policymakers to target interventions more effectively and accelerate progress towards the SDGs. The notebook utilizes Python and popular libraries like Pandas, NumPy, and Scikit-learn for data preprocessing, exploratory data analysis, and model building.

Solution and Optimization

Inputs and Outputs for Analysis

Inputs (Predictors):

Features: The dataset uses various SDG goal scores as input features for analysis.
These include: goal_1_score to goal_17_score

Output (Target Variable):

Target: The primary output or target variable for analysis is: sdg_index_score

Visualizing the distribution of SDG Index scores using a box plot:

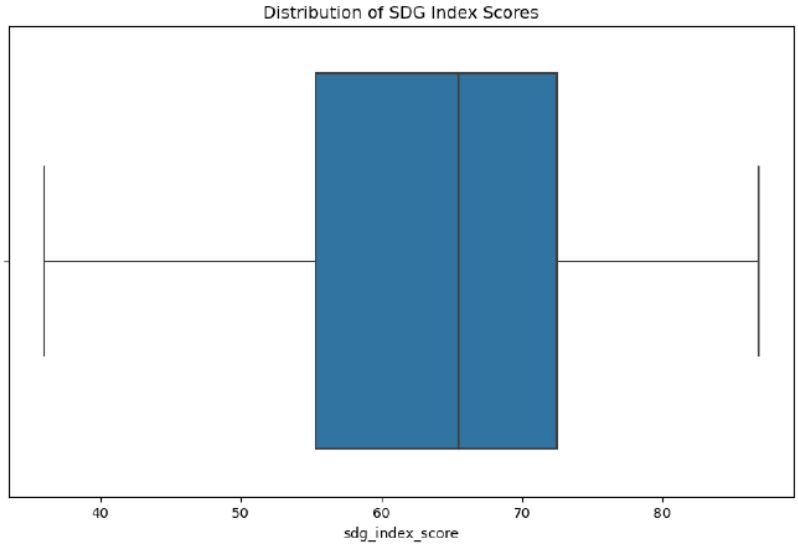

Figure 2 The distribution of SDG Index Scores

The boxplot displays the distribution of SDG Index Scores, showing that the median score is around 67, with most scores ranging between 60 (Q1) and 73 (Q3). The data is symmetrical, though slightly skewed towards the lower end, as indicated by the median being slightly above the center of the interquartile range (IQR). The whiskers extend from approximately 43 to 80, and there are no outliers, suggesting that the SDG scores are generally well-distributed within this range.

Summary statistics for the numerical columns

	year	sdg_index_score	goal_1_score	goal_2_score	goal_3_score	goal_4_score	goal_5_score	goal_6_score	goal_7_score	goal_8_score	goal_9
count	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.000000	3818.
mean	2011.000000	63.850288	64.795233	57.698900	64.306207	72.073808	55.950367	64.523965	57.614013	69.838554	36.
std	6.634118	10.916160	36.836088	11.221834	22.651757	27.008855	17.621801	15.103414	21.871602	10.243085	26.
min	2000.000000	36.000000	0.000000	7.700000	5.900000	0.000000	3.500000	23.300000	0.100000	38.400000	0.
25%	2005.000000	55.100000	30.825000	52.325000	44.900000	55.600000	43.100000	52.600000	41.200000	63.900000	15.
50%	2011.000000	65.400000	81.600000	58.900000	71.150000	81.200000	57.700000	64.900000	65.200000	70.100000	29.
75%	2017.000000	72.300000	98.600000	65.300000	81.700000	94.900000	69.200000	74.700000	72.100000	76.700000	52.
max	2022.000000	86.800000	100.000000	83.400000	97.300000	100.000000	94.000000	95.100000	99.600000	93.600000	99.

Figure 3 Summary Statistics

Histogram of SDG Index Scores

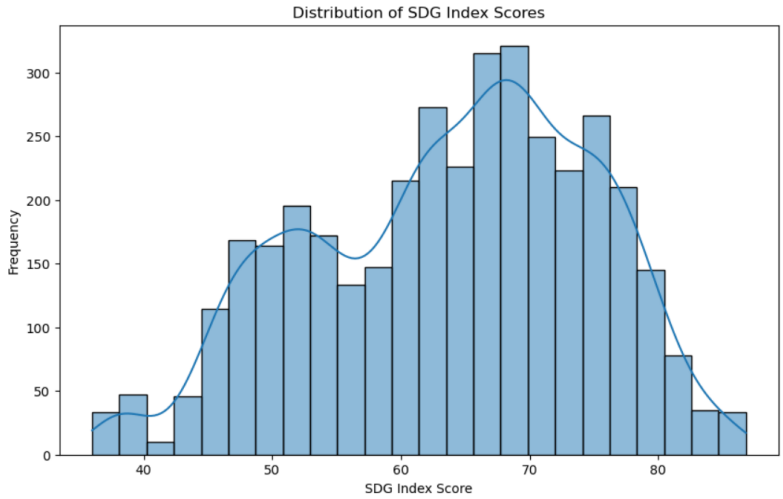

Figure 4 Histogram of SDG index

This histogram displays the frequency distribution of SDG (Sustainable Development Goals) Index Scores, showing how often different score ranges occur. The distribution appears to be roughly normal with a slight positive skew, as evidenced by a higher concentration of scores between 60 and 75. The peak frequency occurs around 70, indicating that many entities have scores in this range. The data tails off on both ends, with fewer scores below 40 and above 80. The smooth curve overlaying the histogram further illustrates the overall distribution trend, confirming the central tendency around the 60-75 range.

Correlation matrix heatmap

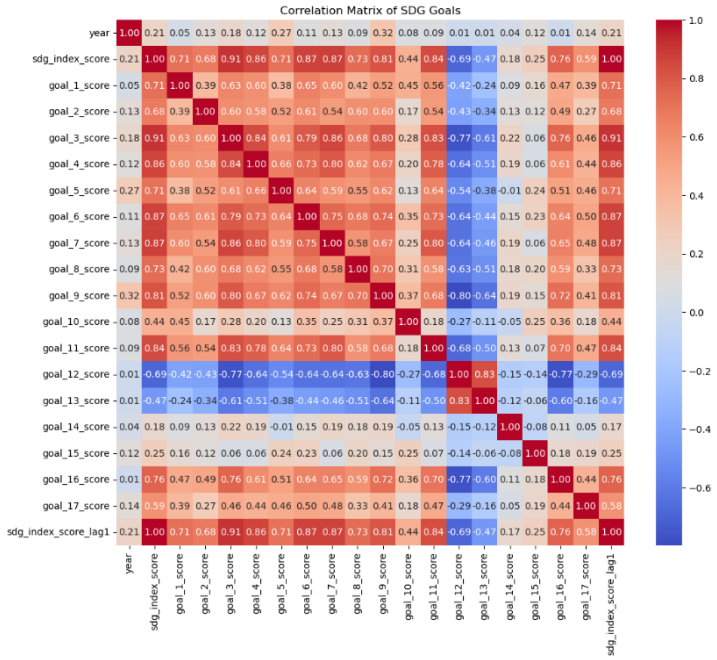

Figure 5 Correlation Matrix heatmap

This is a correlation matrix that displays the relationships between various SDG (Sustainable Development Goals) scores and other related variables. Each cell in the matrix represents the correlation coefficient between two variables, which quantifies the degree to which the two variables are related. The values range from -1 to 1:

- **Positive Correlation** (closer to 1): Indicates that as one variable increases, the other tends to increase as well.
- **Negative Correlation** (closer to -1): Indicates that as one variable increases, the other tends to decrease.
- **No Correlation** (around 0): Indicates no linear relationship between the variables.

Key Insights:

1. **SDG Index Score**: The SDG index score has strong positive correlations with several individual goals, particularly with goal 1 (No Poverty), goal 3 (Good Health and Well-being), and goal 4 (Quality Education). This suggests that higher overall SDG scores are closely associated with better performance in these areas.
2. **Individual Goals**: Many individual SDG goals are positively correlated with each other, indicating that progress in one goal tends to be associated with progress in others. For example, goal 4 (Quality Education) is strongly correlated with goal 1 (No Poverty), goal 3 (Good Health and Well-being), and goal 5 (Gender Equality).
3. **Negative Correlations**: Some goals show negative correlations with others. For example, goal 12 (Responsible Consumption and Production) and goal 13 (Climate Action) have negative correlations with several other goals, indicating that progress in these areas may be associated with less progress in others, possibly due to conflicting priorities.
4. **Year**: The correlation between the year and other variables is generally low, indicating that the scores do not strongly trend over time within the period covered.
5. **Lagged Index Score**: The lagged SDG index score (from a previous period) shows a strong positive correlation with the current SDG index score, indicating that regions or countries with higher SDG performance in the past tend to maintain their performance.

The heatmap colors represent the strength and direction of correlations: dark red indicates strong positive correlations, dark blue indicates strong negative correlations, and lighter colors indicate weaker correlations.

To fulfill our objectives, we have worked on three machine learning algorithms:

1. Linear Regression Model
2. Random Forest Method
3. Neural Networks method

1. Linear Regression Model

Input variables: 'goal_1_score', 'goal_2_score', 'goal_3_score', 'goal_4_score', 'goal_5_score', 'goal_6_score', 'goal_7_score', 'goal_8_score', 'goal_9_score', 'goal_10_score', 'goal_11_score', 'goal_12_score', 'goal_13_score', 'goal_14_score', 'goal_15_score','goal_16_score', 'goal_17_score'

Output variables: sdg_index_score

Giving us:

```
Mean Squared Error: 1.5147351007404093
Predicted SDG Index Score: 87.36920122934518
```

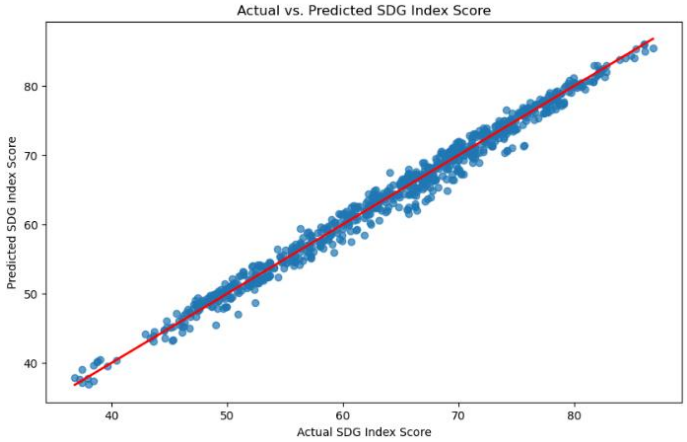

Figure 6 Figure 6 Actual Vs Predicted LRM

This scatter plot compares the actual SDG Index scores (on the x-axis) with the predicted SDG Index scores (on the y-axis). Each point represents a country or region.
The red diagonal line represents a perfect prediction, where the predicted value equals the actual value.
The blue dots are the model's predictions. The closer these points are to the red line, the better the model's performance.
This plot indicates that the model is performing well, as most points are closely aligned with the red line, showing that the predicted scores are very similar to the actual scores.

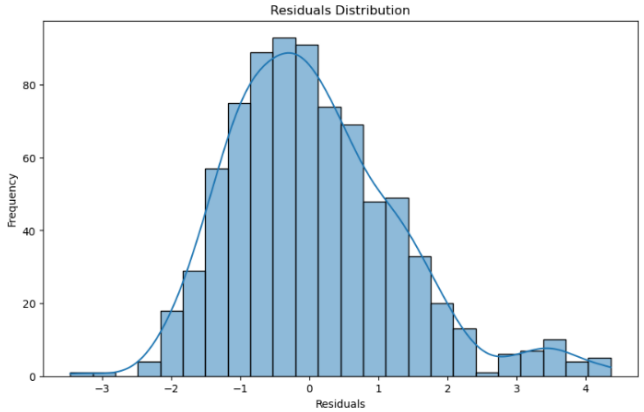

Figure 7 Residuals graph for LRM

This histogram shows the distribution of residuals (the differences between the actual and predicted SDG Index scores).
The residuals are centered around zero, with a roughly normal distribution, indicating that the model's errors are unbiased and normally distributed.
The spread of the residuals shows how much the model's predictions deviate from the actual values. A normal distribution of residuals suggests that the model is well-calibrated and not systematically overestimating or underestimating the index scores.

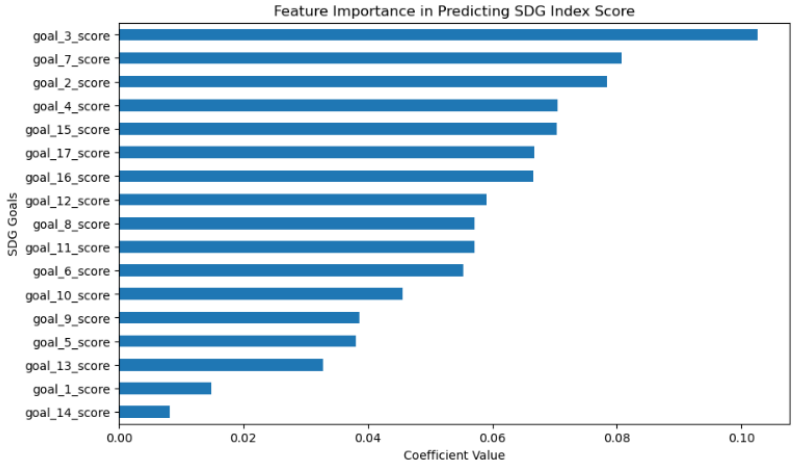

Figure 8 Feature importance for LRM

This bar chart ranks the importance of different Sustainable Development Goals (SDGs) in predicting the overall SDG Index score.

The SDG goals are listed on the y-axis, and their corresponding coefficient values (importance) are on the x-axis.

Higher coefficient values indicate that a particular SDG goal has a stronger influence on predicting the SDG Index score.

In this chart, Goal 3 (Good Health and Well-being) and Goal 7 (Affordable and Clean Energy) are the most influential in predicting the SDG Index score, while Goal 14 (Life Below Water) has the least influence.

2.Random Forest Method

Input variables: 'goal_1_score', 'goal_2_score', 'goal_3_score', 'goal_4_score', 'goal_5_score', 'goal_6_score', 'goal_7_score', 'goal_8_score', 'goal_9_score', 'goal_10_score', 'goal_11_score', 'goal_12_score', 'goal_13_score', 'goal_14_score', 'goal_15_score','goal_16_score', 'goal_17_score'

Output variables: sdg_index_score

```
Mean Squared Error: 0.3916223899924699
```

Giving us:

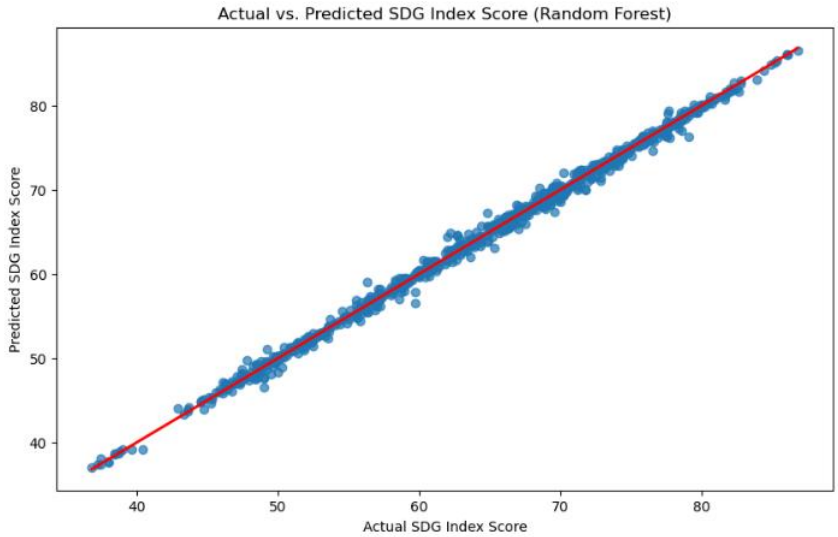

Figure 9

The scatter plot compares the actual SDG Index scores (x-axis) with the predicted scores from the Random Forest model (y-axis).
The red line represents the ideal scenario where the predicted scores perfectly match the actual scores (a 45-degree line).
The points clustered along this line suggest that the model is performing well, with predicted scores closely matching the actual values.

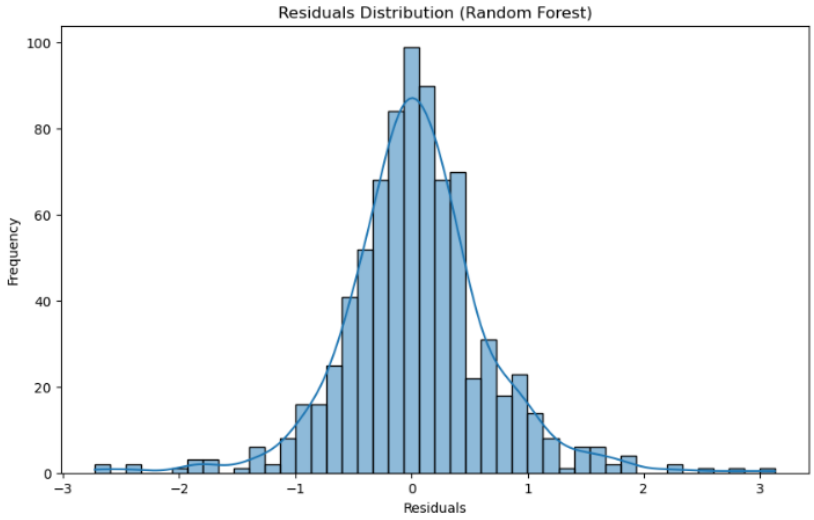

Figure 10

This histogram displays the distribution of residuals, which are the differences between the actual and predicted SDG Index scores.

The residuals are centered around zero and follow a roughly normal distribution, indicating that the model's errors are symmetrically distributed with no significant skewness or bias.

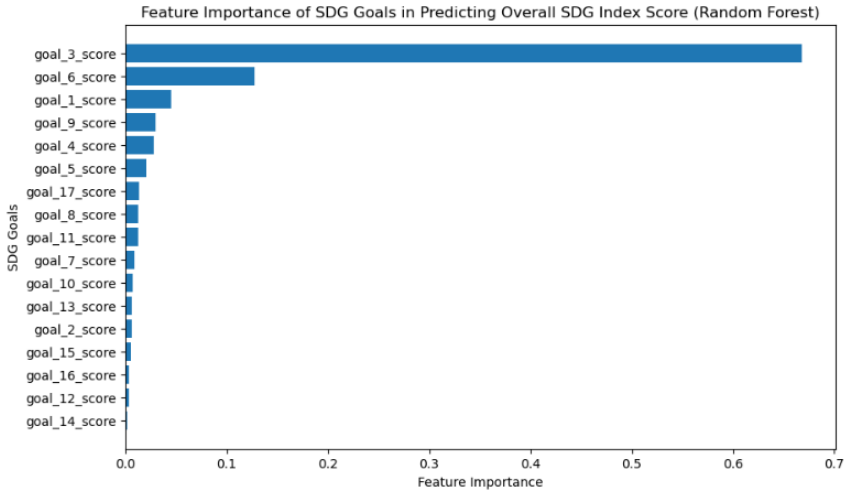

Figure 11

This bar chart ranks the importance of different SDG goals in predicting the overall SDG Index score.

Goal 3 (Good Health and Well-being) is shown to be the most influential feature, followed by Goal 6 (Clean Water and Sanitation).

The other goals have relatively lower importance, indicating that the model considers health and sanitation as key determinants for the SDG Index score.

3. Neural Networks Method

Input variables: 'goal_1_score', 'goal_2_score', 'goal_3_score', 'goal_4_score', 'goal_5_score', 'goal_6_score', 'goal_7_score', 'goal_8_score', 'goal_9_score', 'goal_10_score', 'goal_11_score', 'goal_12_score', 'goal_13_score', 'goal_14_score', 'goal_15_score','goal_16_score', 'goal_17_score'

Output variables: sdg_index_score

Giving us: `Neural Network Mean Squared Error: 0.455196529083763`

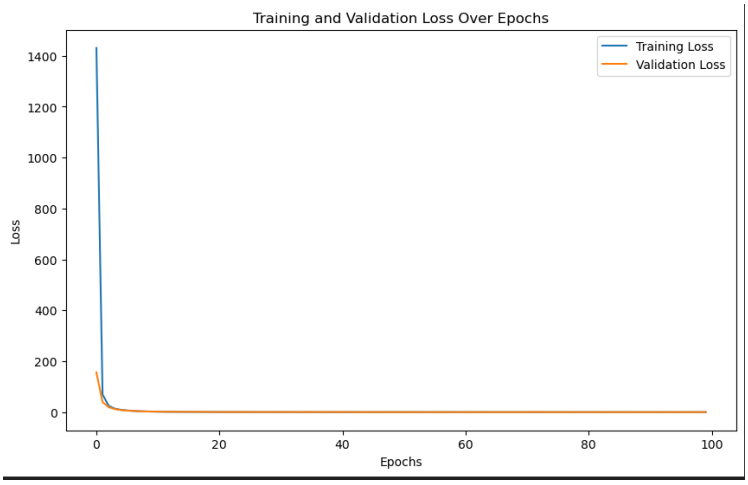

Figure 12 Train-Loss Epochs graph

The training and validation loss plot over 100 epochs shows key trends:

1. **Initial High Loss**: Both losses start high, with training loss being higher.
2. **Rapid Decrease**: The training loss quickly drops, showing fast learning. The validation loss also decreases, indicating improved performance on unseen data.
3. **Stabilization**: Both losses stabilize at low values, suggesting further training isn't reducing the loss.
4. **Convergence**: The losses converge, indicating no overfitting, as training and validation losses are similar.

The rapid initial decline and stabilization indicate effective training with no overfitting.
The close alignment between losses suggests good generalization.
Early stopping could be considered to save resources once validation loss stabilizes.

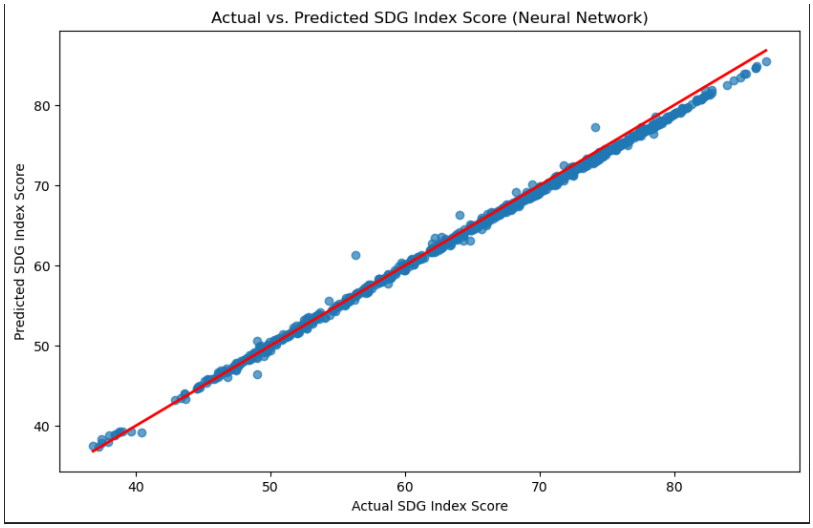

Figure 13 Actual Vs Predicted NNM

Diagonal Line: The red line represents perfect predictions. Points on this line indicate exact matches between predicted and actual scores.
Scatter Points: Most blue points closely align with the red line, showing high prediction accuracy.

Deviations: Some points slightly deviate from the line, indicating minor prediction errors.
High Accuracy: The close clustering around the diagonal line reflects the model's accurate predictions.
Minimal Errors: Minor deviations suggest small, infrequent errors, showing the model has learned well.
Strong Performance: The small spread around the line indicates consistent performance across data points.
Improvement Areas: Analyzing outliers could help fine-tune the model.
Overall, the model accurately predicts SDG Index scores, with minimal errors.

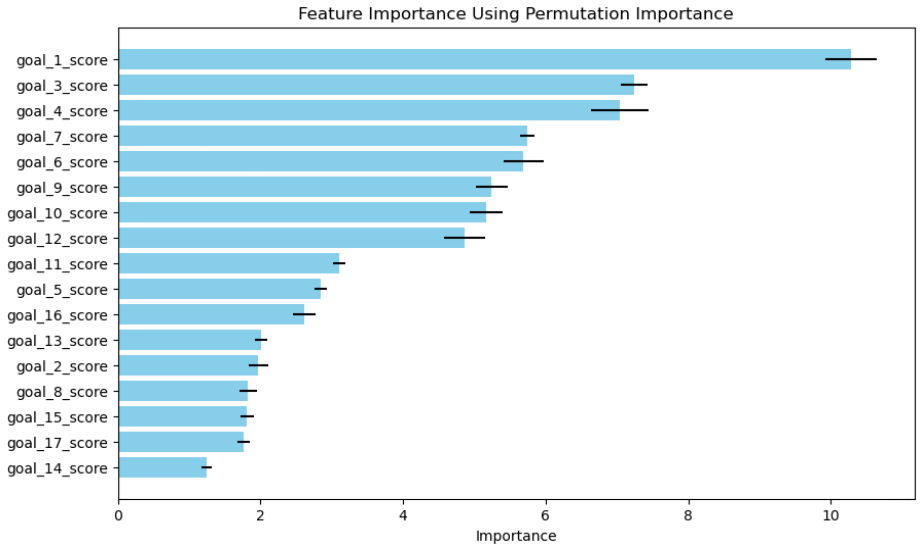

Figure 14 Feature importance for NNM

This bar chart displays the feature importance of variables labeled as goal_X_score using "Permutation Importance."
Feature Importance: Indicates how much each feature contributes to the model's performance. Higher values mean greater importance.
Permutation Importance: Measures the impact on model accuracy when a feature's values are shuffled. Greater performance reduction signifies higher importance.
Y-axis: Lists features (e.g., goal_1_score, goal_3_score).
X-axis: Shows the importance score.
Bars: Represent feature importance; longer bars indicate more importance.
Error Bars: Represent uncertainty in the estimates.
Summary:

- goal_1_score is the most important feature, followed by goal_3_score and goal_4_score.

Conclusion

Based on the evaluation using Mean Squared Error (MSE), the Random Forest Regression model emerged as the most accurate for predicting the SDG Index score. With the lowest MSE of 0.3916, it outperformed both the Neural Network (MSE: 0.4552) and the Regression-Based Forecasting model (MSE: 1.5147). The Random Forest's ability to capture complex relationships within the data makes it the best choice for this task. While the Neural Network showed competitive performance, it did not surpass the Random Forest model, likely due to the latter's robustness in handling feature interactions. The Regression-Based Forecasting model, with the highest MSE, proved less effective, likely due to its linear nature, which may not adequately capture the data's non-linear relationships. Therefore, the Random Forest Regression model is recommended for accurate SDG Index score predictions.

References:

1. **Dataset:** https://www.kaggle.com/datasets/sazidthe1/sustainable-development-report
2. Hák, T., Janoušková, S., & Moldan, B. (2015). Sustainable development goals: A need for relevant indicators. *Environmental Science & Policy*, 55, 54-62. https://doi.org/10.1016/j.envsci.2015.08.002
3. Fleming, A., Wise, R. M., Hansen, H., & Sams, L. (2017). The sustainable development goals: A case study. *Environmental Science & Policy*, 77, 21-30. https://doi.org/10.1016/j.envsci.2017.08.004