Final Project

MDA 620

December 14, 2023

Authored by: Vera Arrebola Granes

Table of Contents

Background

SDG stands for Sustainability Development Goals. They are a collection of 17 interconnected global goals set by the United Nations General Assembly in 2015. The purpose of these goals is to address various social, economic, and environmental challenges faced worldwide and to guide efforts toward a more sustainable future by the year 2030. These goals are interrelated and interconnected, recognizing that addressing one issue often involves addressing multiple factors simultaneously. The SDGs are intended to mobilize governments, businesses, civil society, and individuals to work collectively towards a more prosperous, inclusive, and sustainable world for present and future generations.

The data set I used for this project is extracted from the United Nations website. In order to extract the data, I selected the following criteria, so it fits my requirements: “Data for at least two years since 2015,” “Compare Countries across all goals,” “Country-goal matrix,” and “World (total) by SDG regions.” The link for the data set is the following: https://unstats.un.org/sdgs/dataportal/analytics/DataAvailability.

Problem Scenario and Objective

We live in a world of constant adaptation in which a few decades ago sustainability was not a main concern for the population and its leaders. However, especially in the last decade more questions and concerns have been raised about how we need to make changes if we want to improve the world we live in now. Undoubtedly, moving towards a sustainable society is a complex task and that is why we need to make sure everyone is on the same page and there are frameworks with specific goals. SDGs are the first step towards sustainability. Every country that attended the Assembly agreed on certain rules and goals that they would work to achieve. The data set used for this project shows the progress each country has made so far in each of the 17 SDGs according to the report signed in 2015. It has been eight years since these countries accepted the challenge and it is crucial to take a look at the progression as we are halfway until we reach the deadline.

As we have experienced before, not everyone likes to follow the rules, even if they committed to it. Therefore, checking how each country is progressing is very important. Maybe some countries are not putting as much effort into achieving these goals, or maybe some countries have already reached some of the goals and they can be reformulated to make an even bigger impact. Moreover, it is necessary to understand how each of the 17 SDGs affect each other as they are all interconnected. Hence, the primary aim of this project is to comprehend the interrelations among the Sustainable Development Goals (SDGs) and to identify the leading countries in terms of their performance. These are the 17 SDGs:

1. No Poverty: End poverty in all its forms everywhere.

2. Zero Hunger: End hunger, achieve food security, improve nutrition, and promote sustainable agriculture.

3. Good Health and Well-being: Ensure healthy lives and promote well-being for all at all ages.

4. Quality Education: Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.

5. Gender Equality: Achieve gender equality and empower all women and girls.

6. Clean Water and Sanitation: Ensure availability and sustainable management of water and sanitation for all.

7. Affordable and Clean Energy: Ensure access to affordable, reliable, sustainable, and modern energy for all.

8. Decent Work and Economic Growth: Promote sustained, inclusive, and sustainable economic growth, full and productive employment, and decent work for all.

9. Industry, Innovation, and Infrastructure: Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

10. Reduced Inequality: Reduce inequality within and among countries.

11. Sustainable Cities and Communities: Make cities and human settlements inclusive, safe, resilient, and sustainable.

12. Responsible Consumption and Production: Ensure sustainable consumption and production patterns.

13. Climate Action: Take urgent action to combat climate change and its impacts.

14. Life Below Water: Conserve and sustainably use the oceans, seas, and marine resources for sustainable development.

15. Life on Land: Protect, restore, and promote sustainable use of terrestrial ecosystems, sustainably manage forests, combat desertification, halt and reverse land degradation, and halt biodiversity loss.

16. Peace, Justice, and Strong Institutions: Promote peaceful and inclusive societies for sustainable development, provide access to justice for all, and build effective, accountable, and inclusive institutions at all levels.

17. Partnerships for the Goals: Strengthen the means of implementation and revitalize the global partnership for sustainable development.

All of them are equally important, yet for this project we will be using “Goal 17: Partnerships for the Goals” as the label, as it involves all the goals at the same time.

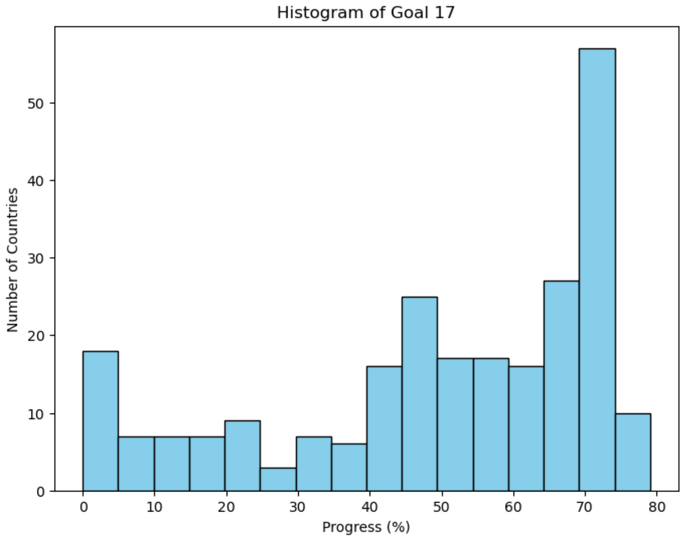
Data Exploration

The data set has 249 rows (each one belonging to a different country) and 19 columns. The first two columns are the Name of the country (GeoAreaName) and an individual country code (GeoAreaCode) unique for each country. From column 3 to column 19 we have all the SDG goals. There is only one categorical variable (GeoAreaName). The following table shows information about each column in the data set, since columns 3 to 19 are very similar I will only include up to column 7 so it is not too repetitive.

Column Name	***Description***	***Data Type***
GeoAreaCode	Unique numerical code	int64
GeoAreaName	Name of each country that agreed on the framework	object
Goal 1	Percentage progress towards the goal	float64
Goal 2	Percentage progress towards the goal	float64
Goal 3	Percentage progress towards the goal	float64
Goal 4	Percentage progress towards the goal	float64
Goal 5	Percentage progress towards the goal	float64

As previously mentioned, Goal 17 will be the label for this project and the other goals will be the features that we will be using to conduct the predictive analysis. Luckily, this data set is very well structured and coherent so there is not much data manipulation needed.

The histogram shows the distribution of the progress made towards Goal 17. As you can see, most of the data lies where the biggest percentages of progress (on the right side), which can be considered as a positive outcome because it means that a large number of countries are more than halfway through achieving their goal.



Data Manipulation

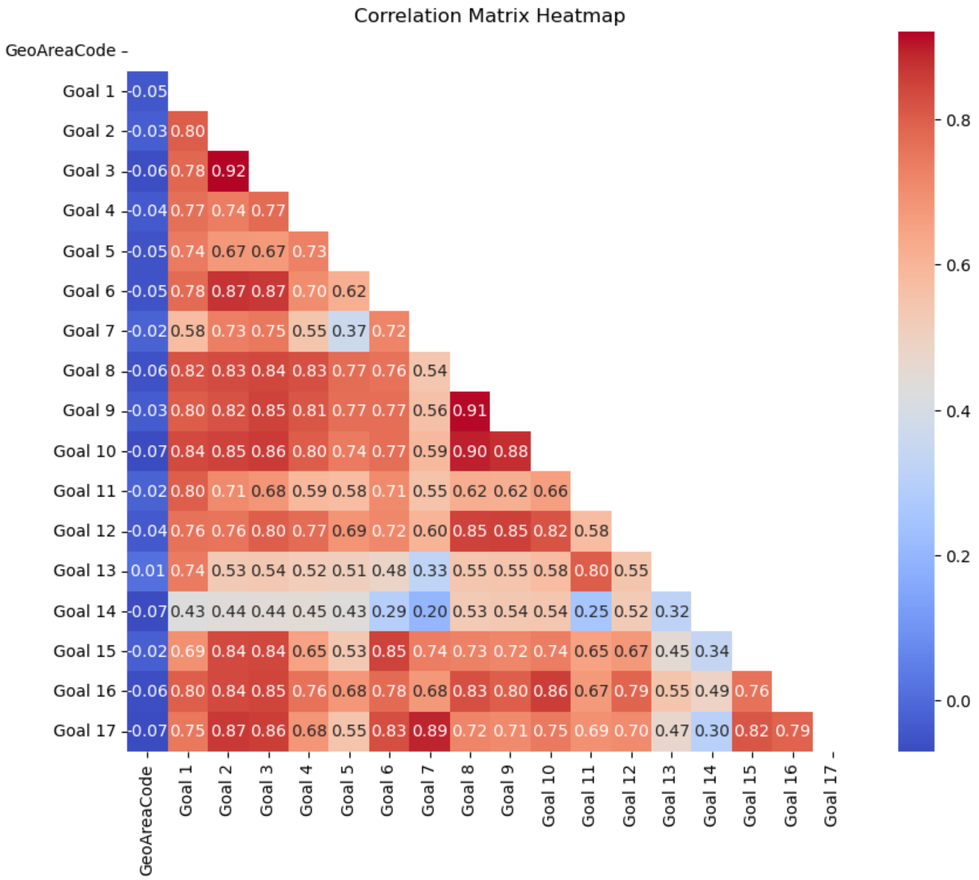
When handling datasets, it's essential to comprehend the content, refine the dataset, and adjust it to suit your research objectives. In the original data set all the goal columns had the percentage sign, yet I decided to remove it so I could work with them as numbers and not objects. The numbers still represent progress as a percentage.

To streamline my investigation, I created a secondary data frame containing solely the Area Code and Area Name. I excluded these columns from the model as they lack useful information for predictive analysis, but it is great to have it for reference. Additionally, I conducted a check for missing values, fortunately finding none. Now we are ready to conduct predictive models.

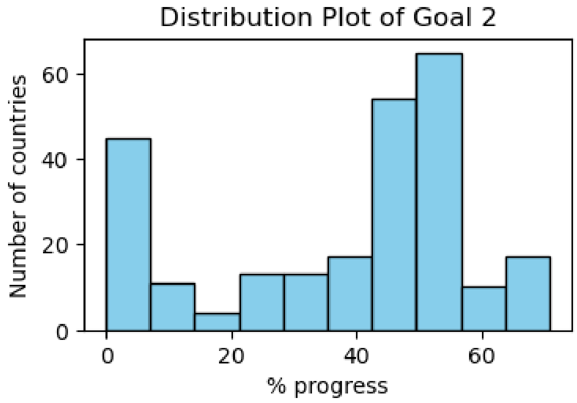
Data Visualization

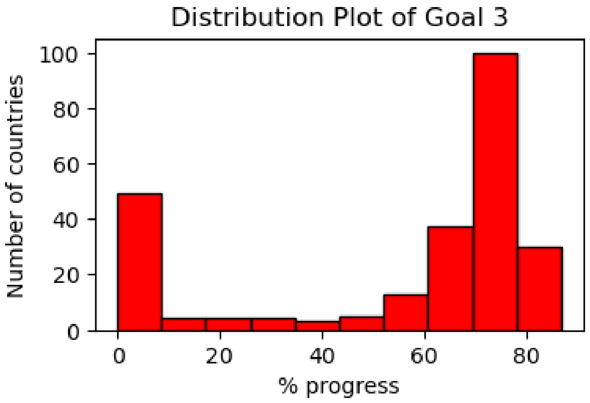
First, I want to check the relationship between all the variables and the label. Correlation coefficients illustrate correlations between predictors and help knowing which variables affect the label the most which can be helpful for the predictive analysis. For this data set, I am looking for those variables that have a high correlation coefficient with the label as well as with some other variables. It is not always the case you want to see high correlation among the features, yet in this case, since we want the progress of each goal to be as high as possible, if there is a strong positive correlation between some of the features, we should also consider it. For this project, we will consider correlations above 0.80 as very strong and above 0.60 as somewhat strong.

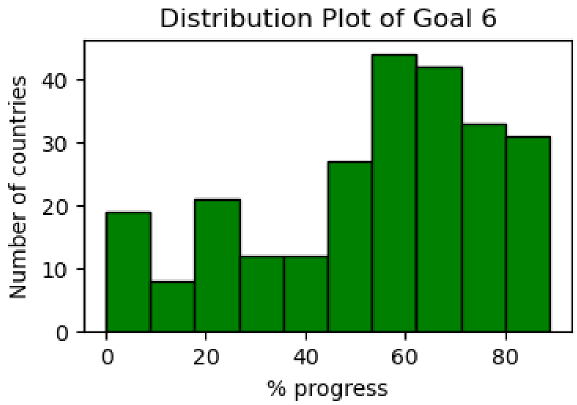
In the correlation heatmap below, we find the strongest correlations (over 0.80) with the label in Goals 2, 3, 6, 7 and 15. Furthermore, if we pay attention to Goals 2, 3, 6 and 15, they also have very high correlations with the rest of the features. Having a strong positive correlation between the goals is important because when, for example, a country moves toward achieving the 100% of Goal 2, the correlation coefficient indicates that it is very probable that Goals 3, 6, 7, 15 and 17 will also increase.

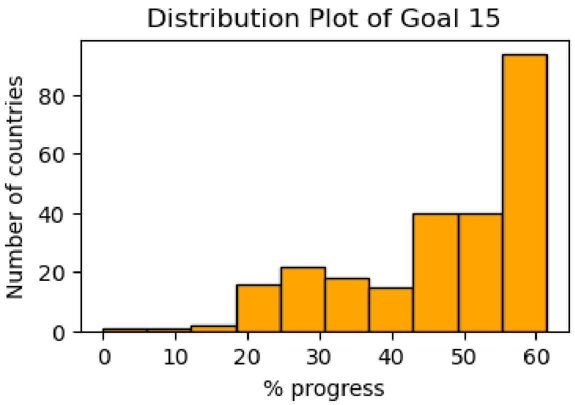


Next, we will be looking at the distribution of the variables with the strongest relationship with the target variable. For Goal 2, most of the countries are concentrated between 40-60% which indicates the progress made is moderate, yet it is also important noticing that many countries have just made 10% progress which is concerning. For Goal 3, the frequency of countries achieving a percentage of progress is highest in the value range of 60-80%, with a frequency of more than 100 countries. The distribution of countries achieving a percentage of progress is positively skewed. For Goal 6, the percentages are more equally distributed. We notice this trend because none of the percentages have a frequency higher than 45 countries. Yet, we can still say it is slightly skewed to the right because more than half of the countries are positioned in the 50-90% progress. Finally, for Goal 15, it is very clear that the majority of countries are making progress towards achieving Goal 15 at intermediate percentages of progress, between 45-65%.



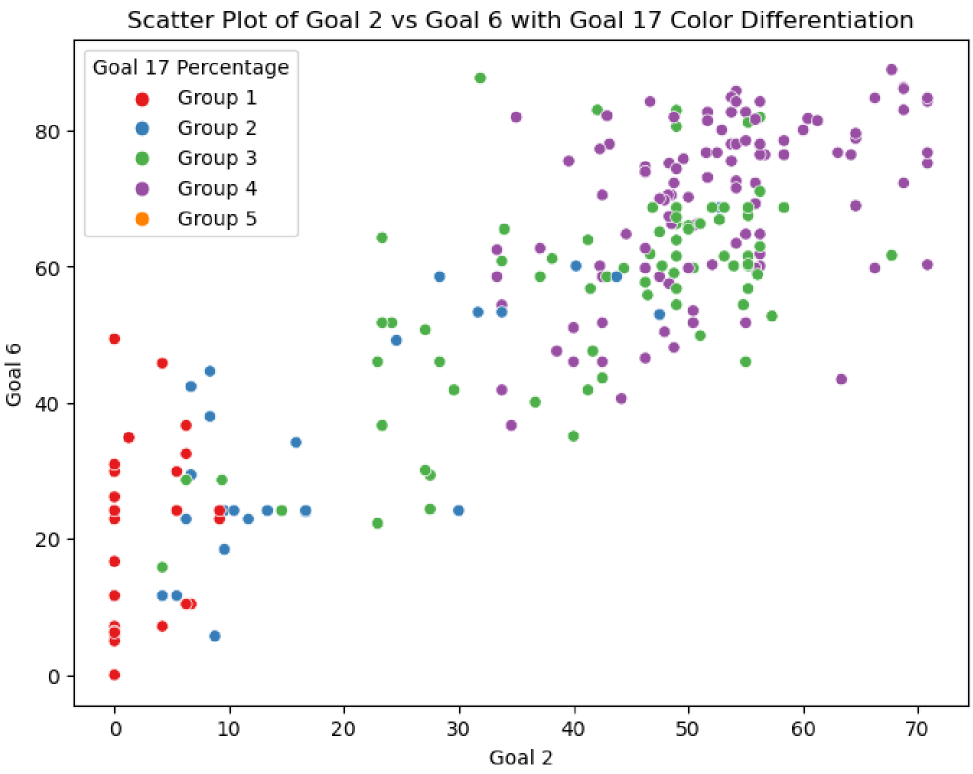






The following scatter plot shows the relationship between Goal 2 and Goal 15, differentiated by Goal 17 Percentage across five different groups. I created five different groups for Goal 17 where Group 1 represents 0-20% of progress towards Goal 17, and consecutively getting to Group 5 that represents 80-100% of progress. Each group is represented by a distinct color.

Based on the scatter plot, it can be inferred that there is a positive correlation between Goal 2 and Goal 15. The scatter plot also shows that the correlation is stronger for higher percentages of Goal 17. The most important insight we notice is that most of the high percentages from Goal 2 and Goal 15 (on the upper right corner) usually belong to Group 3 or Group 4 (which as explained above, represent high percentages of Goal 17). We can conclude that the higher the progress in Goal 2 and Goal 15, the higher the progress in Goal 17. Moreover, we also observe that there are no dots from Group 5, indicating that no country has made more than 80% of progress towards Goal 17.

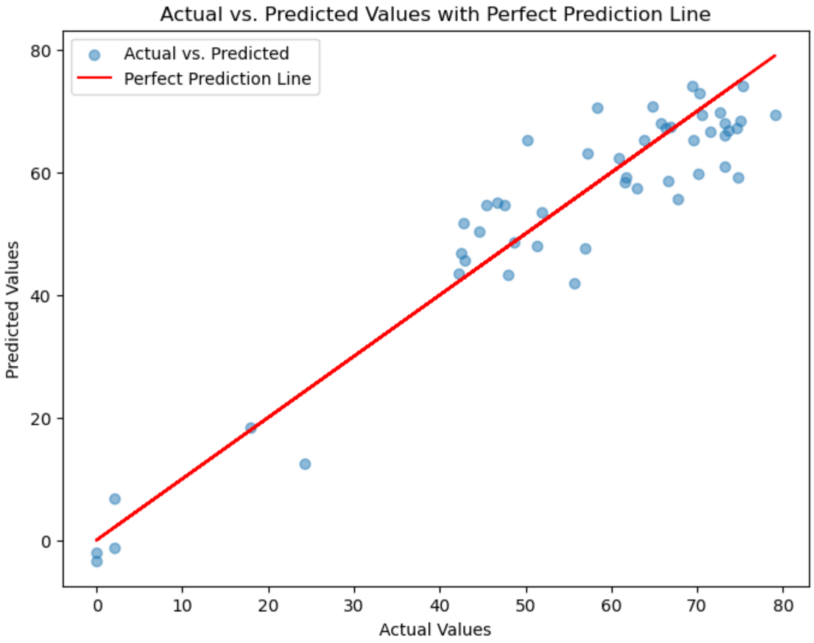


Model Building

For this project I will create three different models: Linear Regression, Random Forest and Decision Tree Regressor. First, Linear Regression provides straightforward interpretations of coefficients. It assumes a linear relationship between the independent variables and the target variable. In addition, Linear Regression is simple to implement, and in most cases, data analysts want to find the best model that also provides the simplest solution. Second, Random Forest is a versatile ensemble learning method capable of capturing non-linear relationships between predictors and the target. It is very different from Linear Regression, so it will give me another perspective of the predictions and data set. Moreover, Random Forest is less prone to overfitting compared to simple linear models. Finally, Decision Trees also capture non-linear relationships between features and the target variable. Furthermore, Decision Trees are less affected by outliers compared to linear models. In the next sections, I will be showing the three models built with Python, feature importance extracted from each of them, as well as the model evaluation showing the R squared and Mean Squared Error (MSE).

Linear Regression

Goal 17 is the label of the model, meaning that with all the independent variables we are trying to predict it. We start by splitting the dataset into training (80%) and testing (20%) sets. Linear Regression provides transparent insights into the relationship between independent and dependent variables, allowing easy interpretation of coefficients. The purpose of making predictions on the test data set is to assess how well the model generalizes to new, unseen data (test set). Plotting the results is crucial because it gives us a visual representation of the model's performance, allowing for a clearer understanding of its strengths, weaknesses, and overall behavior across different scenarios or data points.

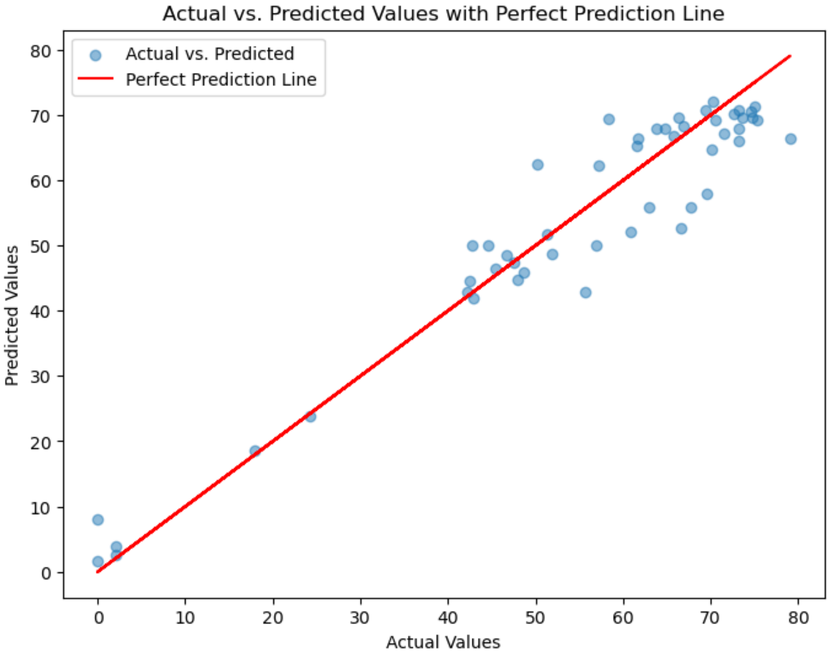


Ideally, all points would fall along the diagonal line, indicating perfect predictions where the actual values perfectly match the predicted values. The closer the points cluster around the diagonal line, the better the model's predictions align with the true values. If the points are scattered far from the diagonal line, it indicates discrepancies between the actual and predicted values. However, visually it looks like most values fall closely around the prediction line. Following the analysis, we check the R squared and MSE (0.885 and 49.204 respectively). We want an R squared as close to 1 as possible and a low MSE because both are interpretation of the performance of the model. I conclude that these values are pretty good, but we need to compare them with the other models.

Random Forest

We begin by dividing the dataset into training (80%) and testing (20%) sets. Random Forest is a powerful algorithm that can handle complex datasets with many features and interactions. It works by creating an ensemble of decision trees, each trained on a random subset of the data. The trees are then combined to make predictions on new data points.

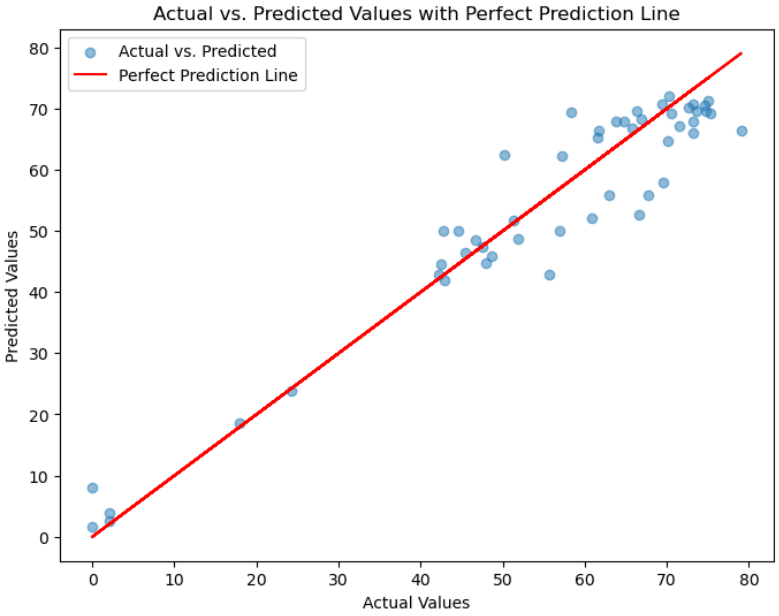
The purpose of making predictions on the test set is to evaluate how well the model generalizes to new, unseen data. To evaluate the performance of the model, we made predictions on the test set and calculated the Mean Squared Error (MSE) and R-squared (R^2). The model performed well on the test set, with an MSE of 35.739 and an R^2 of 0.916. These two metrics together show a very good performance of the model. In addition, we can take a look at the scatter plot comparing Actual Values vs Predicted Values.



Many dots lie on top of the Perfect Prediction Line indicating that the model is performing well, and the predicted values are very close to the actual values. This is a good sign as it suggests that the model is accurately capturing the underlying patterns in the data.

Decision Tree Regressor

For the Decision Tree Regressor Model, we used the same split of the dataset as before, training (80%) and testing (20%) sets. Decision Tree Regressor is a powerful algorithm that can handle complex datasets with many features and interactions. It works by recursively partitioning the data into subsets based on the values of the input features. The partitions are chosen to minimize the variance of the target variable within each subset. Construction of decision trees usually works top-down, by choosing a variable at each step that best splits the set of items. To evaluate the performance of the model, we made predictions on the test set and calculated the Mean Squared Error (MSE) and R-squared (R²). The model performed well on the test set, with an MSE of 63.634 and an R² of 0.851. One more time, pretty good evaluation metrics values.



Similarly, to the other two models, values congregate closely around the Perfect Prediction Line, even having some dots on top of the line showing very close results in predictions.

Model Selection

Mean Squared Error (MSE) and R-squared (R²) are two commonly used metrics for evaluating the performance of predictive models. MSE measures the average squared difference between the predicted and actual values. The smaller the MSE, the closer your model's predictions are to reality. So, when we obtained MSEs for the three models, for example, Decision Tree Regressor had 63.634, it means that predictions deviate on average by approximately 63.634 units from the true values.

R squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale. If the value of R squared is large, you have a better chance of your regression model fitting the observations. Therefore, for example, in the Linear Regression model we obtained an R squared of 0.885, meaning that the model accounts for a substantial portion (88%) of the variance in the SDG data. For a better comparison we look at all the models at the same time:

Model	R squared	MSE
Linear Regression	0.885	49.204
Random Forest	0.916	35.739
Decision Tree Regressor	0.851	63.634

Looking at the evaluation metrics I would conclude that Random Forest has the best performance of the three of them, due to the R squared being the closest to 1 and having the lowest MSE. For this data set, where we measure the percentage of progress towards achieving Goal 17, having a small MSE is essential. We do not want a model whose predictions deviate a lot from real values.

Conclusions

The primary focus was to understand the interrelation between these goals and to find a model capable of predicting the progress of Goal 17 with the information from the rest of the goals. These goals aim to address multifaceted challenges encompassing social, economic, and environmental aspects, guiding efforts towards a more sustainable future by 2030. The analysis delved into the relationships between the SDGs, showcasing their interconnectedness and the necessity of considering them collectively. It was observed that several goals, notably Goals 2, 3, 6, 7, and 15, exhibited strong positive correlations not only with the overarching Goal 17 ("Partnerships for the Goals") but also among themselves. This implies that progress in certain goals influences and correlates with advancements in other related goals.

The results showcased promising performance across all models, with Random Forest demonstrating superior predictive capabilities due to its lower MSE and higher R^2. Its ability to capture non-linear relationships and handle complex interactions between SDG features resulted in the most accurate predictions. The findings underscore the importance of collaborative efforts among nations to achieve sustainable development goals. Additionally, the predictive models can serve as valuable tools for policymakers, enabling them to identify priority areas for intervention and facilitate targeted actions for countries falling behind in goal attainment. Despite the promising results, this analysis has limitations. Future research could explore additional variables or external factors that might impact SDG progress.

Bibliography

Data source: https://unstats.un.org/sdgs/dataportal/analytics/DataAvailability
UN SGDs: https://unstats.un.org/sdgs/
Visualizations: created by author using python.