National College of Ireland

Can We Predict Statistical Capacity? A Study on Predicting Statistical Performance Indicators

MSc Research Project
MSc in Data Analytics

Binu Jemima Christopher Nadar
Student ID: x23323493

School of Computing
National College of Ireland

Supervisor: Dr. David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Binu Jemima Christopher Nadar
Student ID:	X23323493
Programme:	MSc in Data Analytics **Year:** 2024 - 2025
Module:	MSc Research Project
Supervisor:	Dr. David Hamill
Submission Due Date:	15th September 2025
Project Title:	Can we Predict Statistical Capacity? A Study on Predicting Statistical Performance Indicators
Word Count:	4538 **Page Count** 18

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Binu Jemima Christopher Nadar

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Can We Predict Statistical Capacity? A Study on Statistical Performance Indicators

Binu Jemima Christopher Nadar
X23323493

Abstract

This study explores the Statistical Performance Indicators (StatPI) developed by the World Bank to evaluate national statistical capacity and support monitoring of Sustainable Development Goals (SDGs). The research addresses two core objectives: identifying data anomalies within the StatPI dataset and assessing the feasibility of predictive modeling to estimate StatPI indicator scores. Through a Data Gap Analysis (DGA), we uncovered inconsistencies between StatPI scores and data availability in the World Development Indicators (WDI), often due to differing source databases. While some discrepancies were explainable, few raised concerns about the reliability and consistency of StatPI scores.
The second part of the study applied predictive modelling, specifically Random Forest regression to assess whether SPI scores could be predicted based on available features. While models using limited binary indicators showed high accuracy, they posed a risk of overfitting. A more complex model using a broader feature set demonstrated more reliable potential, despite a slightly lower performance. Our findings suggest that while predictive analysis for StatPI is currently underutilized, it holds promise; particularly for imputing missing values or signalling declining data capacity.

1 Introduction

Statistical Capacity Development refers to the process of improving a country's ability to produce, manage, and utilize high-quality statistical data for informed decision-making and development. Such capacity is essential for crafting effective policy, tracking national progress, and supporting sustainable development efforts. Strong statistical systems not only benefit national governments but also helps the international organizations to make better-informed decisions. In this context, the World Bank (Dang, et al., 2023) developed a tool known as Statistical Performance Indicators (StatPI) to assess the statistical capacity of National Statistical Systems (NSS) across various countries. This tool plays a key role in monitoring the performance of these systems. Additionally, NSS are also responsible for tracking progress on the Sustainable Development Goals (SDGs), and StatPI supports this monitoring by giving an overview on the data availability to analyze these goals.

While previous researches have primarily examined how StatPI data can be used to analyze national trends, few studies have focused on detecting anomalies or inconsistencies within the data itself. Identifying such anomalies is crucial for ensuring the reliability of this statistical tool. Building on this foundation, there remains a notable gap in research aimed at predicting these indicators. The ability to forecast indicator values can support the imputation of missing data, provide early warnings for systems experiencing a decline in statistical capacity, and enable more effective performance tracking. Accordingly, our research addresses these two dimensions: data quality and data utilization. To assess data quality, we will be comparing the Social Progress Index (SPI) indicators with the StatPI Indicators related to social progress,

examining their alignment and identifying any potential gaps. As for data utilization, we will evaluate the ability of the dataset to forecast social progress outcomes by applying Random Forest.

This study addresses the following two questions -

Research Question 1: Can this research help detect and understand data anomalies within the World Bank StatPI dataset?

Research Question 2: Can we build a reliable predictive model for StatPI (SPI) trends, and what precautions must be taken during the modeling process?

2 Related Work

Since this tool is a relatively recent initiative by the World Bank, there has been limited exploration or research conducted on it so far. It was developed as a more effective alternative to the Statistical Capacity Index (SCI), which had been in use since 2004. The SCI (Fu et al., 2021) played a significant role in emphasizing the importance of strengthening statistical capacity. While there have been few works regarding SCI, there is not much related work with regards to SPI. As a result, much of the related work surrounding with regards to this new tool draws inspiration from methodologies and tools used in other domains.

2.1 Research Motivation and Focus

The paper (Dang, et al., 2023) introduces a new tool called Statistical Performance Indicators (StatPI), designed to measure a country's statistical capacity. Since many national goals and policy decisions depend on the availability and reliability of data, understanding the strengths and weaknesses of statistical systems is crucial. The authors present StatPI as a useful tool to identify areas in need of improvement, but they also highlight several limitations. These include concerns about data accuracy, unclear usage of the tool, inconsistent statistical infrastructure across countries, and gaps in the United Nations Statistics Division's global indicator database.

The study (Rastogi, et al., 2023) emphasizes the critical role of data accuracy in measuring progress. They evaluated whether StatPI accurately reflects performance in areas such as Financial Inclusion (FI), Environmental, Social, and Governance (ESG), and poverty reduction. Interestingly, while ESG performance improved with rising StatPI scores, FI showed a negative correlation. This suggests that increases in StatPI did not always align with improvements in financial inclusion, possibly due to the limited impact poverty had on SPI scores. Although the full methodology was not accessible, the findings informed our project's direction, particularly our focus on examining the relationship between the indicators used for Social Progress Index (SPI) and the indicators used by StatPI.

The paper (Karimzadeh, et al., 2022) highlights how data gaps negatively impact electricity consumption analysis. To address this, they introduced a new dataset incorporating Non-Intrusive Load Monitoring (NILM), enabling detailed tracking of appliance-level consumption. Their results demonstrated more than a 5% improvement in analysis accuracy when using richer datasets. Inspired by this, our project also explores how enhancement of data sources could improve the reliability and precision of StatPI. This (Ul-Durar, et al., 2025) paper

examines the challenges faced by least-developed countries (LDCs), arguing that a disproportionate focus on social progress, without parallel economic development can hinder financial service expansion. This imbalance ultimately affects inclusive growth. Their insights reinforce the need for balanced and context-aware statistical evaluation, which our project seeks to support.

2.2 Relevant Methodologies

The report (Social Progress Imperative & AITi Tiedemann Global, 2025) outlines the detailed methodology used to calculate the Social Progress Index (SPI). Their process involves identifying key components of social progress, selecting appropriate indicators, performing necessary data transformations, determining the sample of countries, and finally computing the index. Since our project utilizes the dataset provided by this report, we will adopt the same indicators used in their SPI calculation. This will allow us to systematically compare whether these SPI indicators are also considered by Statistical Performance Indicators (StatPI) when assessing data availability scores.

The authors (Ariño-Plana, et al., 2016) adapt the concept of Data Gap Analysis (DGA), originally used in the business sector, for application to biodiversity data whilst understanding the difference in type of data. They identify two primary goals of DGA: first, to determine what data is currently available, and second, to identify what data still needs to be collected. This adapted approach to DGA is relevant to our project because, while our focus is on statistical data related to various sustainable development goals rather than business data, the underlying principle of identifying data gaps remains the same. Similarly, this study (Shirey, et al., 2021) compared digitally accessible butterfly records with data from natural history collections to analyze data gaps caused by spatial and taxonomic biases. Their study revealed under-sampling in certain regions and provided a framework to address these deficiencies. Both of these papers employ the concept of DGA, by adapting it to suit their specific types of data. Consequently, we will adopt a similar strategy tailored to our dataset.

The paper (Valencia, et al., 2020) proposes a procedure based on Knowledge Discovery in Data (KDD) to analyze data and extract meaningful insights, which are then structured for training intelligent systems focused on fault diagnosis. Meanwhile, the authors (Plotnikova, et al., 2020) of this paper address a research gap, noting that many adaptations of data mining methodologies suggest that existing approaches do not fully meet the needs across all domains. Their analysis of data mining methods, both in generic and specialized contexts, concludes that refining current methodologies can help bridge these gaps. Similarly, this paper (Vishwakarma & Jain, 2022) discusses various data analysis techniques, including KDD, prediction, clustering, outlier detection, and the analysis of evolution and deviation. Based on these studies, our project will adopt the KDD approach, as it is more suitable for our purposes compared to CRISP-DM, which is primarily designed for business use cases.

The authors (Zhu, et al., 2023) emphasize the importance of including a wider range of indicators to accurately predict pneumonia in symptomatic COVID-19 patients infected with the Omicron variant, which differs significantly from earlier strains. Their study assesses

various hematological indicators to identify those most useful for understanding pneumonia risk. Meanwhile, this author (Zixi, 2021) approaches poverty as a multidimensional issue influenced by various factors, and applies multiple machine learning models to predict poverty outcomes. To develop a more comprehensive prediction framework, the study combines datasets from two different sources, creating an integrated approach. A range of models including linear regression, decision trees, random forests, gradient boosting, and neural networks are used to analyze the data and evaluate both the significance of different factors and the performance of each model. The study identifies key determinants of poverty, such as education, country of residence, urbanization, age, exposure to economic shocks, and access to technology, as significant predictors of whether an individual is likely to experience poverty. In their 2025 study, the authors (Göl, et al., 2025) explore the prediction of anemia in patients by selecting attributes such as malnutrition and physical activity scores, without relying on blood test values. They employed machine learning models including J48 and Random Forest, both of which achieved high accuracy. Based on these results, the authors concluded that anemia can be effectively predicted using indirect health indicators, demonstrating the potential for non-invasive diagnostic approaches. These papers highlight the importance of features while doing predictive analysis.

Collectively, these studies highlight the critical role of selecting relevant features when performing predictive analysis, reinforcing the importance of comprehensive and carefully chosen indicators for prediction.

The study by (Salman, et al., 2024) highlights the effectiveness of the Random Forest algorithm in handling unbalanced datasets and variables with missing values. Similarly, (Putra, et al., 2023) found that Random Forest outperformed the Decision Tree model in predicting car prices. Both studies emphasize Random Forest's strength in reducing bias by generating multiple decision trees and aggregating their outputs. Given that our dataset may contain both bias and missing values, Random Forest is a suitable choice for our initial predictive modelling.

3 Research Methodology

Our research focuses on two key aspects of data analytics: data gap analysis and predictive modeling. We begin by identifying and understanding the gaps within the data, and based on this analysis, we explore the feasibility of applying predictive models to fill those gaps.

For this project, we utilize open datasets from the World Bank, which includes the Statistical Performance Indicators (World Bank Group, 2024) and the World Development Indicators (World Bank Group, n.d.). Although StatPI is composed of indicators drawn from various databases, our focus is limited to the World Bank data to enable a consistent and meaningful comparative analysis. Additionally, since both StatPI and the Social Progress Index (SPI) incorporate data from the World Bank, this approach facilitates a better understanding of their interrelationship. Given the extensive scope of the data and practical constraints, we will select 3–4 countries from each income group, ensuring representation across diverse geographical regions. This sampling strategy aims to establish a baseline understanding of countries' statistical data capacity.

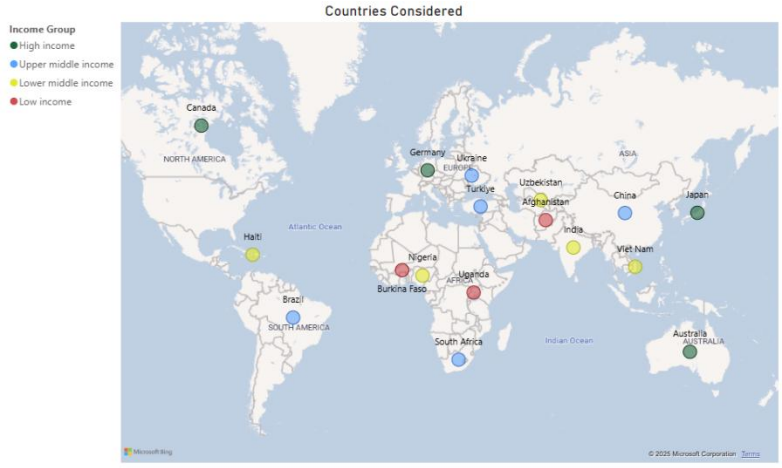

Figure 1: Countries Selected for Project

The countries chosen are -

1. High Income: Australia (Oceania), Canada (North America), Germany (Europe), Japan (Asia)
2. Upper Middle Income: Brazil (South America), China (Asia), South Africa (Southern Africa), Ukraine (Europe)
3. Lower Middle Income: Haiti (North America), India (South Asia), Nigeria (Africa), Uzbekistan (Central Asia), Viet Nam (Southeast Asia)
4. Low Income: Afghanistan (Asia), Burkina Faso (West Africa), Uganda (East Africa)

According to World Bank data, there are no European countries currently classified under the low or lower-middle income groups, as all European nations fall into the upper-middle or high-income categories. Countries such as Ukraine and Moldova have fluctuated between income classifications in recent years due to economic shocks but are currently categorized as upper-middle income. In the Americas, Haiti, which was previously listed under the low-income group, has recently transitioned to the lower-middle income category. As a result, our analysis includes a greater number of Asian and African countries within the low and lower-middle income groups to ensure balanced regional representation.

Data Gap Analysis (DGA):

For this project, we adopt a DGA approach inspired by the modified methodology proposed by (Ariño -Plana, et al., 2016). Their process for identifying data gaps serves as a foundational framework, which we adapt to suit the context of our project

Our DGA will follow the following steps: defining the ideal data, analyzing the existing data, identifying existing gaps, assessing their impact and proposing potential solutions.

- Defining the ideal data – We consider the ideal scenario to be one in which the data availability scores for social progress-related indicators in the StatPI align closely with

the actual availability of those indicators in the WDI dataset. A strong correlation between these sources would indicate accurate representation of data capacity.

- Analyzing the exisiting data - To perform the analysis, we will use a combination of Python and Power BI. Python provides advanced data manipulation capabilities and enables the creation of complex visualizations such as heatmaps and high-computation charts. Power BI, on the other hand, offers interactive visualization features and flexibility for comparative analysis across countries and indicators. The use of slicers in Power BI allows us to dynamically adjust visualizations based on selected countries or indicators.computation and can hang up Power BI.
- Gap Identification - We will begin by identifying gaps between corresponding indicators in both datasets. Since both StatPI and WDI are sourced from the World Bank, we expect a high degree of consistency, making the comparison more robust. Based on these insights we will try to group similar indicators and analyse the discrepancies further.
- Addressing Impact & Solution - Based on the findings, we will explore potential impacts and propose possible improvements.

Predictive Modelling:

The second component of this project focuses on determining whether it is possible to predict the statistical capacity of StatPI indicators. To achieve this, we will apply the KDD methodology, which provides a structured approach for extracting meaningful insights from large datasets. For the predictive modeling, we will employ the Random Forest algorithm. This model is well-suited to our needs, as the dataset includes several categorical features, and it is effective in handling feature interactions, non-linear relationships, and potential outliers. It's robustness and interpretability make it an appropriate choice for this initial stage of predictive analysis.

4 Design Specification

The techniques utilized for this project include:

- Data Gap Analysis (DGA)

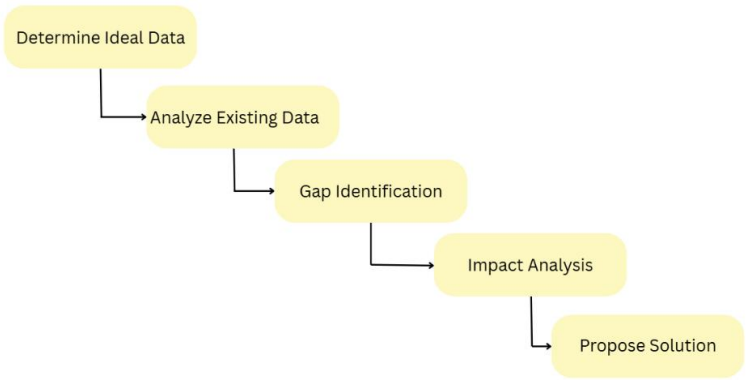

Figure 2: DGA Architecture Diagram

As outlined in the methodology, the Data Gap Analysis (DGA) consists of the following steps:

- Determining Ideal Data: Identifying what data is ideally required for the analysis or objective.
- Analysing Existing Data: Reviewing the current data available to assess its quality, completeness, and relevance.
- Gap Identification: Finding out the discrepancies between the ideal and existing data.
- Impact Analysis *(Not included)*: This step, which involves evaluating the consequences of the identified gaps, will not be conducted due to time constraints, as it requires more in-depth research.
- Proposing Solutions: Recommending ways to address the identified data gaps.

- <u>Knowledge Discovery in Databases (KDD)</u>

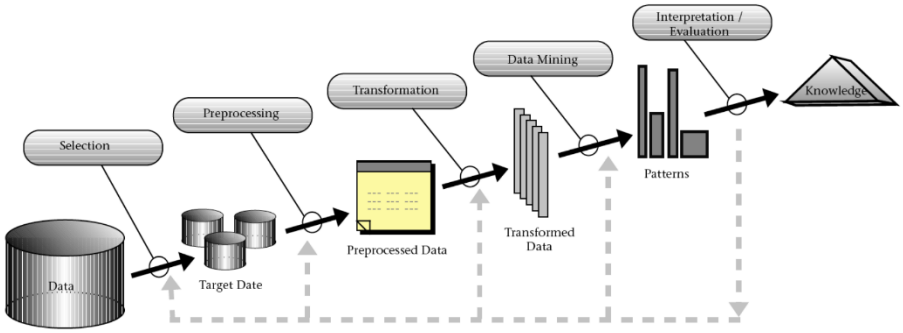

Figure 3: KDD Process Diagram (Fayyad, et al., 1996)

Knowledge Discovery in Databases (KDD) is a structured methodology for deriving insights from datasets. It consists of the following steps:

- Selection of Data: Identifying and collecting relevant data for analysis.
- Preprocessing of Data: Cleaning and preparing the data to improve quality and remove inconsistencies.
- Transformation of Data: Converting data into suitable formats or structures for mining.
- Performing Data Mining Techniques: Applying algorithms to extract patterns from the data.
- Interpretation of Data: Evaluating and understanding the discovered pattern.

- Random Forest

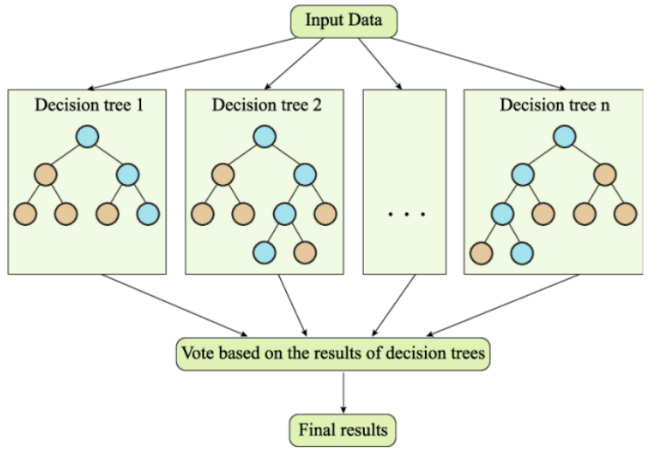

Figure 4: Random Forest Diagram (Zhuang, et al., 2023)

Random Forest is a machine learning algorithm used for both classification and regression tasks. It constructs an ensemble of decision trees by training each tree on a random subset of the input data and selecting a random subset of features at each split to determine the best split. The final prediction is made by aggregating the outputs of all the trees.

5 Implementation

Since the research comprises two parts: Data Gap Analysis (DGA) and Predictive Analysis, our implementation approach differed for each part.

Data Gap Analysis:
For this analysis, we utilized a combination of Power BI and Python**.** Both the WDI and StatPI datasets were imported into Power BI, where various data transformations were performed to create a comprehensive dataset suitable for detailed analysis.

Power BI

1. The datasets 'SPI_index' and 'World_Development_Indicators' were imported into Power BI.
2. Column data types were corrected for columns that initially had incorrect data types.
3. Necessary columns were renamed for clarity.
4. Since the different years were column headers in the World_Development_Indicators dataset, the unpivot function was applied to convert the year columns into rows.
5. This transformation successfully changed the dataset from a wide format to a long format**.**
6. The two datasets were merged into a single dataset named 'SPI_WDI_Merge', joining them based on the 'Country' column to include the 'income' column from SPI_index dataset.

7. The SPI_index dataset was expanded to include the necessary columns required for analysis.
8. A custom column called ‘income_key’ was added to facilitate sorting of the ‘income’ column.
9. Missing values (**NA**) in relevant columns were replaced with **‘0’**.
10. Data types for the necessary columns in the merged dataset were corrected to ensure consistency.

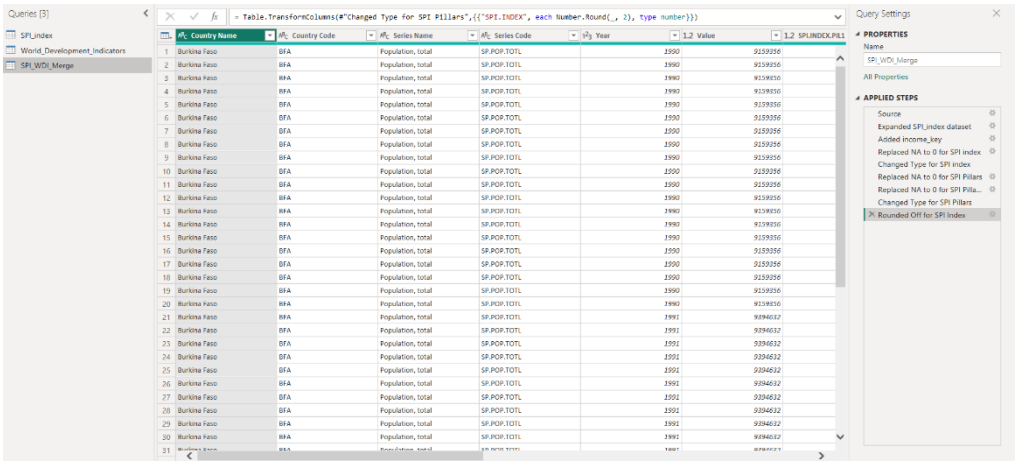

Figure 5: Data Processing and Transformation in Power BI

11. Visualizations are created using the different visualization options in Power BI.

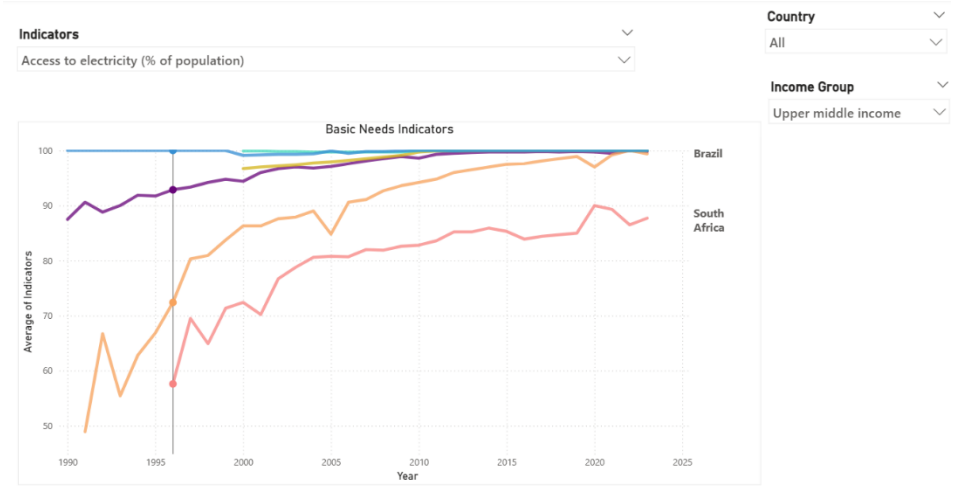

Figure 6: Basic Needs Indicators - Line Chart

Figure 7: SPI Indicators & WDI Indicators – Tables with Data Bars

Python

1. The datasets 'SPI_index' and 'World_Development_Indicators' were imported into Google Colab using Python and converted into dataframes named df_spi (Statistical Performance Indicators) and df_wdi (World Development Indicators).
2. Necessary columns were renamed for clarity.
3. A mapping dictionary for income_group was created and applied to df_wdi based on the 'Country' column.
4. df_wdi was melted to transform the year columns into rows (long format).
5. Appropriate data cleaning techniques were applied to handle missing or inconsistent data.
6. The data was grouped and analysed to identify missing values.
7. A heatmap was created to visualize the number of features with missing values across different years and countries.

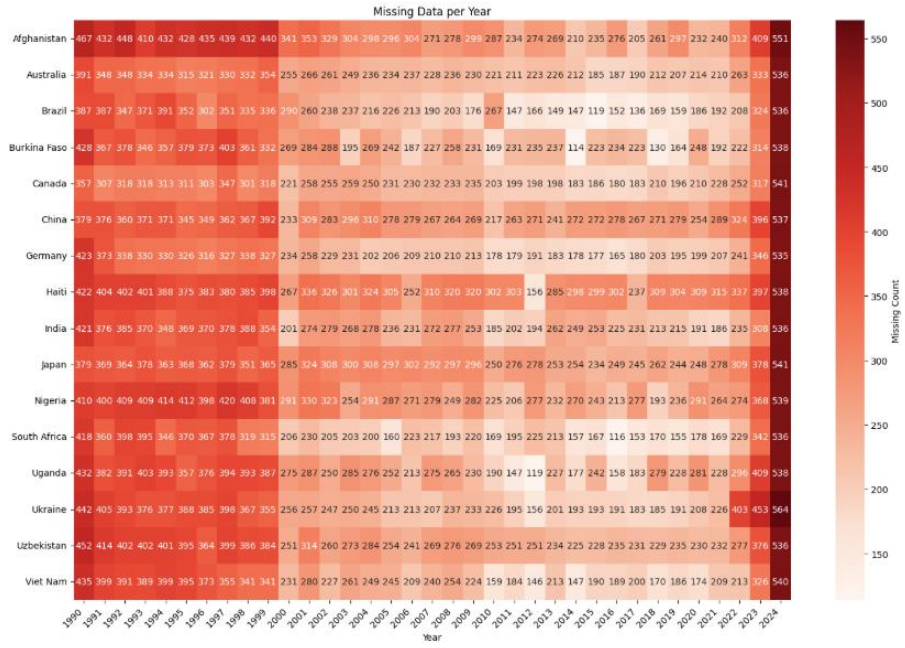

Figure 8: Missing Data Each Year – Heatmap

8. Finally, heatmaps were generated for different country groups to identify which feature groups contained the most missing data.

Predictive Analysis

- To perform predictive analysis on the SPI, we follow the KDD (Knowledge Discovery in Databases) approach.
- Selection: Both datasets (SPI and WDI) are used. Relevant WDI indicators are selected based on the specific statistical indicator we aim to predict.
- Preprocessing: This step involves converting columns to numeric data types, dropping rows where 'Year' is NULL, and filling missing values (NaNs) with empty strings.
- Transformation: The WDI dataset is transformed by converting year columns into rows (long format). Filtering is applied to select only the required WDI indicators. Both datasets are then merged on 'Country' and 'Year', while also including the target statistical indicator. Since the goal is to predict the presence or absence of data, a binary column is created where values are set to '1' if data is present and '0' if absent.
- Data Mining**:** Predictive analysis is performed using a simple Random Forest model. The objective here is to assess the feasibility of prediction rather than to compare different models.
- Evaluation**:** The model's output is evaluated using metrics such as R^2 and RMSE, along with visualizations like Actual vs. Predicted plots and Distribution of Prediction Errors.

6 Evaluation

Data Gap Analysis

For the data gap analysis, a data gap matrix was constructed by evaluating data points across basic needs sector of the Social Progress Index (SPI). To determine which data each statistical indicator in the Statistical Performance Indicators (StatPI) captures, we referred to the World Bank's repository on GitHub (World Bank, n.d.). We then compared the availability of data in the World Development Indicators (WDI) dataset with the scores assigned by StatPI that reflect data availability for the corresponding indicators. This comparison helped identify gaps between the datasets.

Country	World Development Indicator	Domain	Data Available? (Y/N)	Statistical Capacity Indicator	Statistical Indicator Score
Australia	• People using safely managed drinking water services (% of population)	Basic Needs	N	Safely Managed Drinking Water	0.5
Burkina Faso	• Poverty headcount ratio at $3.00 a day (2021 PPP) (% of population) • Proportion of population pushed below the $2.15 poverty line by out-of-pocket health care expenditure	Basic Needs	Y	Availability of Comparable Poverty headcount ratio at $2.15 a day	0
Afghanistan	• Proportion of population pushed below the $2.15 poverty line by out-of-pocket health care expenditure	Basic Needs	Y	Availability of Comparable Poverty headcount ratio at $2.15 a day	0
South Africa	• Poverty headcount ratio at $3.00 a day (2021 PPP) (% of population) • Proportion of population pushed below the $2.15 poverty line by out-of-pocket health care expenditure	Basic Needs	Y	Availability of Comparable Poverty headcount ratio at $2.15 a day	0.5 - 1
High-income countries (Australia, Canada Germany, Japan)	• Average working house of children, working only, ages 7-14 (hours per week) • Chilren in employment, wage workers (% of children in employment, ages 7-14)	Basic Needs	N	Labor force participation rate by sex and age (%)	1
Lower middle income countries (India, Nigeria, Vietnam)	• Average working house of children, working only, ages 7-14 (hours per week) • Chilren in employment, wage workers (% of children in employment, ages 7-14)	Basic Needs	N (Data provided for one year)	Labor force participation rate by sex and age (%)	0.5 - 1

Figure 9: Basic Needs – Data Gap Matrix

Figure 9 illustrates a comparison between the WDI and StatPI for the Basic Needs sector. For example, Australia has a StatPI score of 0.5; however, the WDI dataset lacks data on people's access to safely managed water services. Burkina Faso and Afghanistan both have WDI data related to poverty, yet their StatPI scores for the Poverty indicator are 0, indicating no data availability according to StatPI. South Africa's StatPI scores range from 0.5 to 1 over the years, reflecting data availability improvement over time, but initially showing gaps, despite the WDI dataset containing all necessary data throughout. For the Labour Force indicator, StatPI assigns a score of 1 to higher-income countries, implying data availability, but corresponding WDI data is absent for these countries. Additionally, many lower-income countries have StatPI labour force scores between 0.5 and 1, yet WDI data for these countries is often limited to a single year. This highlights inconsistencies between the two datasets and areas where StatPI scores may not fully align with actual data availability.

Predictive Analysis

Three separate cases, based on varying dependent variables, were analyzed as part of the predictive study.

R^2 (Coefficient of determination) – Measures how well your model's predictions explain the variability of the target variable.
RMSE (Root Mean Squared Error) – Measures the average distance between your predicted and observed values in a regression model.
Residuals – Difference between actual and predicted values.

6.1 Predictive Analysis 1

For the first analysis, we considered the following features and target:

- Target variable: Debt Reporting SPI
- Features: has_value, Year

Results:

- R^2: 0.9764
- RMSE: 0.0738

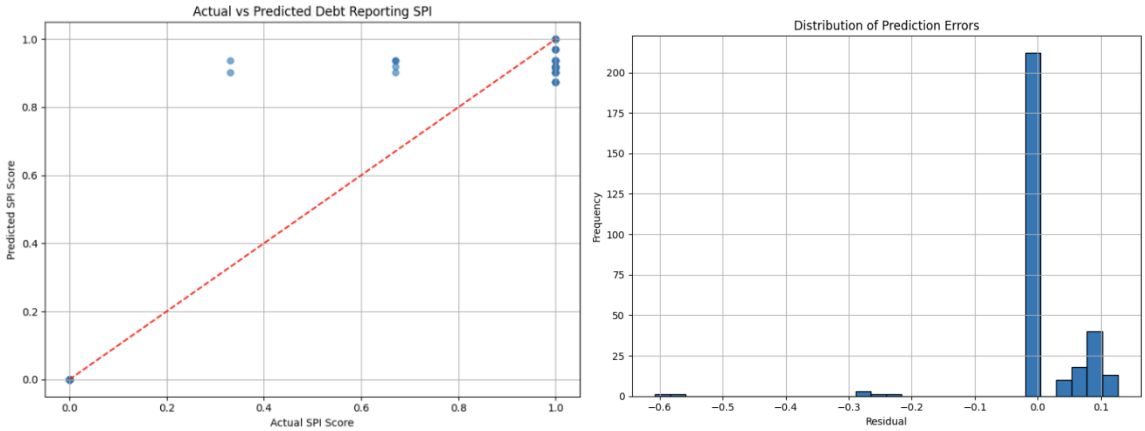

Figure 10: Actual vs Predicted Plot & Residuals Distribution - 1

The model explains approximately 98% of the variance in the SPI scores and shows very low error, indicating a strong fit. The Actual vs Predicted plot showcases most points cluster close to the red line, which means our model is predicting accurately in most cases. A few points are above or below the line, but overall the spread is good. The histogram shows that majority of the residuals are centered tightly around 0, indicating that the model's predictions are generally unbiased. Few residuals with larger but very rare.

6.2 Predictive Analysis 2

For the second analysis, we considered the following features and target:

- Target variable: Debt Reporting SPI
- Features: has_value_ODA, has_value_TDS, has_value_DOD, Year

Results:

- R^2: 0.8843
- RMSE: 0.1532

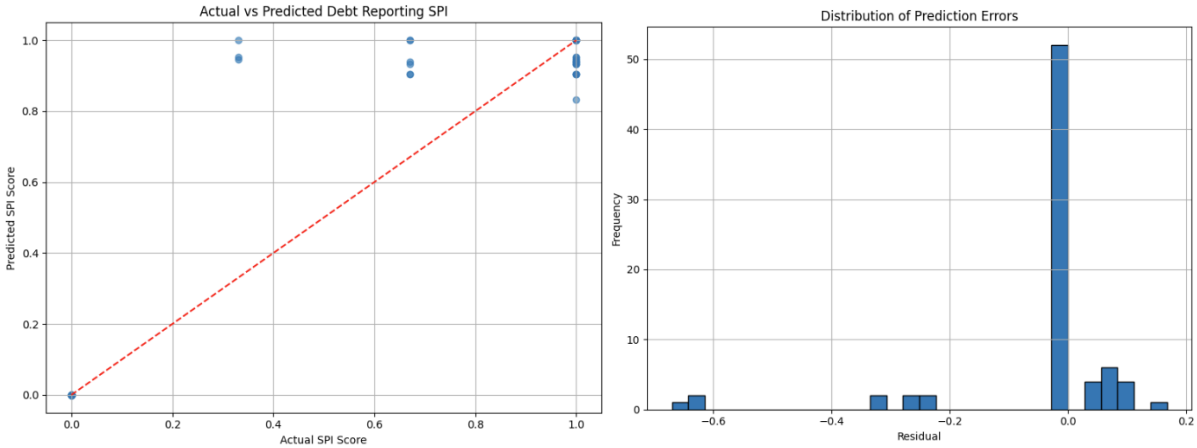

Figure 11: Actual vs Predicted Plot & Residuals Distribution – 2

This model explains approximately 88% of the variance in the SPI scores and shows very low error, indicating a strong fit again. The Actual vs Predicted plot and the histogram showcases similar outputs like the previous predictive analysis.

6.3 Predictive Analysis 3

For the third analysis, we considered the following features and target:

- Target variable: Labour Force SPI
- Features: 17 SL. columns (The SL. columns represent all the columns grouped together related to the labor force.), Year

Results:

- R^2: 0.8014
- RMSE: 0.2029

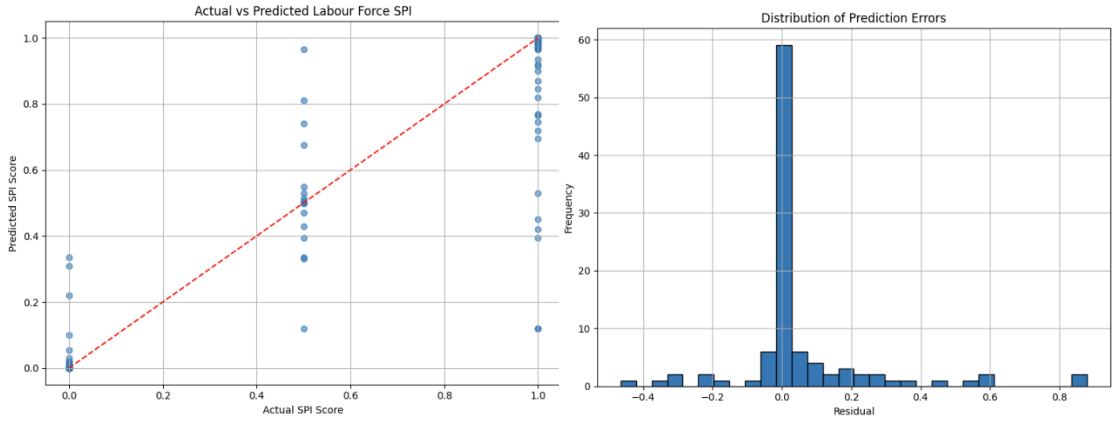

Figure 12: Actual vs Predicted Plot & Residuals Distribution – 3

Approximately 80% of the variation in the target variable is explained by the model, indicating a strong fit. The average prediction error of 0.2029 units further demonstrates the model's good performance. Compared to the previous analysis, the actual vs. predicted values in this case show slightly more scatter, and the distribution of prediction errors is more dispersed, reflecting slightly greater variability in the model's predictions.

6.4 Discussion

Based on the DGA analysis, we observe certain discrepancies related to the SPI indicators. For example, the data source for the Access to Safely Managed Water indicator is the JMP (WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene), which explains why Australia's StatPI score aligns with JMP data availability rather than the World Bank's. Similarly, labour force data is sourced from ILOSTAT (International Labour Organization), which justifies the presence of a StatPI score for this indicator despite missing information in the WDI database. However, the poverty data used to calculate the StatPI score is derived from the WDI, yet the score shows 0 for Afghanistan and 0.5 for South Africa, even though both countries have relevant data available in the World Bank database. Another observation was the disparity in data availability between rural and urban areas, as well as differences in data coverage across gender and age groups, despite the relevant SPI indicator showing a perfect score of 1.

These gaps raise important questions about the reliability of the StatPI scores, given the existence of such mismatches. While not widespread, certain WDI indicators may significantly influence the data availability scores in StatPI, potentially affecting overall confidence in these measurements. Furthermore, this prompts a critical question: are all relevant aspects of these indicators being considered when analyzing their data capacity?

For the predictive analysis, the first two models appear to be good fits. However, because they rely heavily on only a few features, particularly the binary has_value column where there is a risk of overfitting. The high accuracy may be driven mainly by this single feature, suggesting that the model might be capturing patterns specific to the training data rather than generalizable trends. If the test set is not truly independent, this risk of overfitting increases. The third model,

although not performing as well overall, revealed that the features were biased, which is understandable given that the data was not fully cleaned. Despite this, the third model shows promise for predictive analysis because it leverages 17 features, compared to just 2-3 features used in the previous two models.

7 Conclusion and Future Work

Our study successfully addressed both of the research questions. While anomalies were identified in the StatPI data, some were explainable, due to the use of alternate data sources; while others remained unclear. These few inconsistencies largely stem from the reliance on different public databases for different indicators. This fragmented approach is likely to continue producing such discrepancies.

A more effective solution would be to create a unified database that consolidates all the sources used for the various StatPI indicators. Data availability could then be assessed based on this centralized dataset, enabling a clearer understanding of which data points are genuinely missing. This approach would support more targeted improvements in national statistical systems and promote greater data consistency.

Regarding predictive analysis, while it is feasible, its effectiveness largely depends on the number and quality of features available. It holds potential particularly for indicators with rich feature sets. Our research indicated that predictive analysis has not yet been widely applied to StatPI indicators and that could be because there is not much use of it yet. However, if the StatPI framework evolves, predictive modeling could play a valuable role in enhancing data completeness by imputing missing values and providing early signals of declining StatPI indicator scores. Meanwhile, predictive analysis can be explored on present indicators with sufficient number of features to assess their accuracy in forecasting, following the necessary data pre-processing.

References

Ariño-Plana, A., Chavan, V. & Otegui, J., 2016. *Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders.* s.l.:s.n.

Dang, H.-A. H., Pullinger, J., Serajuddin, U. & Stacy, B., 2023. Statistical performance indicators and index—a new tool to measure country statistical capacity. *Scientific Data,* 10(1), p. 146.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* 17(3), pp. 13-95.

Fu, H. et al., 2021. *The Statistical Performance Indicators: A new tool to measure the performance of national statistical systems.* [Online]
Available at: https://blogs.worldbank.org/en/opendata/statistical-performance-indicators-new-tool-measure-performance-national-statistical
[Accessed 1 August 2025].

Göl, M. et al., 2025. Predicting malnutrition-based anemia in geriatric patients using machine learning methods. *Journal of evaluation in clinical practice,* 31(2).

Karimzadeh, M., Zavaliagkos, G. & Panetta, K., 2022. *Feature Extraction and Data Gap Analysis of US Residential Electricity Consumption.* Riga, Latvia, IEEE.

Plotnikova, V., Dumas, M. & Milani, F., 2020. Adaptations of data mining methodologies: a systematic literature review. *PeerJ Computer Science,* Volume 6.

Putra, P. H., Azanuddin, A., Purba, B. & Dalimunthe, Y. A., 2023. Random forest and decision tree algorithms for car price prediction. *Jurnal Matematika Dan Ilmu Pengetahuan Alam LLDikti Wilayah 1 (JUMPA),* 4(1), pp. 81-89.

Rastogi, S., Singh, K. & Kanoujiya, J., 2023. Do countries capture their inclusive growth, sustainability, and poverty correctly? A study on statistical performance indicators defined by World Bank. *Social Responsibility Journal,* 19(10), pp. 1935-1951.

Salman, H. A., Kalakech, A. & Steiti, A., 2024. Random Forest Algorithm Overview. *Babylonian Journal of Machine Learning,* Volume 2024, pp. 69-79.

Shirey, V., Belitz, M. W., Barve, V. & Guralnick, R., 2021. A complete inventory of North American butterfly occurrence data: narrowing data gaps, but increasing bias. *Ecography,* 44(4), pp. 537 - 547.

S. P. I. & A. T. G., 2025. *2025 AlTi Global Social Progress Index Data & Methodology,* Washington, DC: s.n.

Ul-Durar, S. et al., 2025. Sustainable financial inclusion through social progress and regularity quality interaction – Implication for least developed countries. *Research in International Business and Finance,* Volume 76, p. 102811.

Valencia, A. M., Caratar, J., Caicedo, G. & Chamorro, C., 2020. Proposal for a KDD-based procedure to obtain a set of intelligent systems training applied to the identification of failures

in hydroelectric power plants. *Journal of Applied Research and Technology,* 18(6), pp. 376 - 389.

Vishwakarma, P. K. & Jain, N., 2022. *A Review of Data Analysis Methodology.* Greater Noida, India, IEEE.

World Bank Group, 2024. *Statistical Performance Indicators.* [Online]
Available at: https://datacatalog.worldbank.org/search/dataset/0037996/Statistical-Performance-Indicators
[Accessed 01 08 2025].

World Bank Group, n.d. *DataBank.* [Online]
Available at: https://databank.worldbank.org/source/world-development-indicators/preview/on
[Accessed 01 08 2025].

World Bank, n.d. *World Bank on GitHub.* [Online]
Available at: https://github.com/worldbank
[Accessed 01 08 2025].

Zhuang, Y., Sun, X., Li, Y. & Huai, J., 2023. Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches. *Information Fusion,* Volume 95.

Zhu, K. et al., 2023. Value of Laboratory Indicators in Predicting Pneumonia in Symptomatic COVID-19 Patients Infected with the SARS-CoV-2 Omicron Variant. *Infection and Drug Resistance,* Volume 16, pp. 1159 - 1170.

Zixi, H., 2021. *Poverty Prediction Through Machine Learning.* Hangzhou, Cina, IEEE.