Exploring Statistical Performance
A Global Perspective on Development Indicators

1.0 Introduction

1.1 Problem Description

In our rapidly evolving world and ever-changing global scenario the availability and quality of statistical data are pivotal for decision-making in societal and economic contexts. To address this the World Bank has developed the Statistical Performance Indicators (SPI) Framework which acts as a comprehensive tool to measure a country's statistical capacity. It covers key areas such as data infrastructure, statistical methodologies, and information dissemination. Significant challenges are posed by differences in statistical performance across different regions and income classifications. Countries with lower SPI scores often lack robust data systems which limit their ability to make informed decisions on economic policies, resource allocation, and sustainable development initiatives.

Understanding the key contributors to SPI and identifying regional and income-based disparities is crucial for policymakers and development agencies. Insights derived from this analysis aim to enable targeted interventions to strengthen statistical systems, reduce disparities, and promote evidence-based decision-making globally.

1.2 Aims and Objectives

This project aims to analyse the SPI Index and its underlying Pillars (Pillar 1 to Pillar 5) to address the following questions:

Question-1: Which SPI pillars (Pillar 1 to Pillar 5) contribute the most to the overall SPI scores?
Question-2: How does the SPI Index vary across income classifications and regions, and what are the key disparities?

Additionally, an interactable global Choropleth Map will be presented to visualize the distribution of SPI scores and its five pillars over the years 2016-2023. This analysis will provide valuable insights for addressing statistical gaps and promoting data-driven growth across different nations.

2.0 Data

2.1 Selection of Data

The dataset has been obtained directly from the World Bank, a highly reliable and reputable source. This ensures credibility and accuracy of the analysis. The SPI dataset includes data covering key variables such as the overall SPI Index and its five pillars: Data Use (Pillar 1), Data Services (Pillar 2), Data Products (Pillar 1), Data Sources (Pillar 4), and Data Infrastructure (Pillar 5). These pillars form the basis for understanding and evaluating statistical performance.

This dataset is publicly available and has been chosen due to its alignment with the project objectives. Community-based platforms such as Kaggle primarily host user-generated datasets. In sharp contrast the World Bank dataset is curated and verified by experts. The focus on publicly available official data ensures transparency and eliminates concerns about accuracy or credibility, which can sometimes arise with platforms like Kaggle.

2.2 Ethical, Privacy, and Security Considerations

This dataset does not contain any sensitive or personally identifiable information which ensures compliance with data ethics standards and data privacy regulations. Using public data ensures that the project avoids ethical concerns related to unauthorized data access or misuse. Additionally, the analysis has been conducted in a responsible manner ensuring that the data is used in its intended context.

2.3 Data Reliability and Integrity

Several measures were undertaken to ensure reliability of the data. The dataset was validated by cross-checking it to confirm its completeness and consistency. Since this is real-world data there are columns where significant chunks of data were missing. Missing values were identified and handled appropriately to maintain the integrity of the analysis. Outliers which could potentially skew results were carefully reviewed and any unjustifiable outliers were excluded to ensure accurate findings.

2.4 Data Nature, Structure, and Preparation

The SPI dataset is structured as a time series dataset containing data for multiple years across various countries. It includes cross-sectional components, allowing analysis across different regions and income classifications.

Key attributes of the dataset include SPI Overall Scores and Pillars, Country and Regions, Income Classification, and Annual data.

The dataset underwent several cleaning and preparation steps such as Handling Missing Data, Outlier Treatment and Aggregation to prepare the dataset for a robust and accurate analysis.

3.0 Analytics Techniques (Models & Approaches)

To provide conclusive answers to the two questions, the following steps were performed: data cleaning, exploratory data analysis (EDA), and predictive modelling.

3.1 Data Cleaning and Preprocessing

For the first question, the first step was to clean the data and make it suitable for analysis. The aim was to deal with the missing data first, which would ensure that the analysis could be performed well. There were missing values in a number of critical columns, including SPI.INDEX and the SPI Pillars (SPI.INDEX.PIL1 to SPI.INDEX.PIL5). The python pandas library was used to deal with all these missing values. As the SPI pillars were integral to understanding the overall SPI score, the missing values were imputed with the median of each of the respective columns. The median value was chosen since it is resistant to outliers and most importantly it preserves the integrity of the dataset. Mean is extremely sensitive to the outliers. If there are unusually high or low values in the dataset, the mean would be pulled towards the direction of the outliers and would prove detrimental to our analysis. Hence, the median was chosen. After the imputation, the dataset was again checked to ensure that there were no more missing values in the columns being used. This ensured that the data was ready for data analysis and machine learning.

Same methods were followed for the second question too. However, this time the primary focus was on country, region, income classifications, and SPI.INDEX. To ensure data integrity, the missing rows were dropped. The cleaned dataset was then again, checked for missing values. After this process of cleaning, Label Encoding was applied to categorical columns, income and the textual data which were converted into numerical representations (income-encoded and region-encoded), allowing the dataset to be compatible with machine learning models.

3.2 Exploratory Data Analysis (EDA)

For the first question, to study the data, histograms with Kernel Density Estimation (KDE) were plotted for SPI.INDEX and its five pillars (SPI.INDEX.PIL1 to SPI.INDEX.PIL5). This served many purposes, allowing insights on the distribution of each variable. This allowed to show patterns such as central tendencies, variability, and skewness. This also provided insights on how each pillar varies and how each affects the overall SPI score. These distributions also demonstrate the differences in the spread of the pillars, showing that some pillars, like SPI.INDEX.PIL2 (Data Services), were more concentrated towards higher values.

For the second question, to understand the variation in SPI score across different regions, two boxplots were plotted. The first showed the SPI scores across income classifications. The trend was that SPI score increase with higher income levels, showing a strong link between nations' wealth and statistical maturity. The second box plot showed the SPI scores across regions like North America, Europe and Central Asia and others. This revealed significant disparities. North America, Europe & Central Asia had higher SPI scores, however Sub-Saharan Africa had low values.

3.3 Predictive Modelling Using Random Forest

For the first question, a Random Forest Regressor algorithm has been used. The model took the five pillars as features and the overall SPI score as its target, after this the data was split into training and test datasets to ensure that the model can be validated.

Random Forest algorithm was chosen as it can handle complex and non-linear relationships, and it can also compute built in feature importance scores. In addition to the Random Forest, other algorithms such as XGBoost, Support Vector Machines can also be used because of their ability to model the non-linear relationship.

A bar chart was plotted to visualize these feature importance scores. The results showed that: 1) SPI.INDEX.PIL2 (Data Services) was the most significant contributor. 2) SPI.INDEX.PIL5 (Data Infrastructure) and SPI.INDEX.PIL4 (Data Sources) followed as the next most impactful pillars. This analysis provided clear, quantitative evidence of the influence of each pillar on the SPI Score.

For the second research question also, a Random Forest Regressor was used. The data was split into 80-20 training and test data respectively. This model successfully achieved Root Mean Squared Error of (RMSE) of 12.88, which showed its predictive accuracy. This research showed that the income classification factor contributed the most towards predicting the SPI scores, at

approximately 59%, whereas region factor at 41% accounted for a major chunk as well. This showed that income levels play an extremely important role in deciding statistical maturity.

3.4 Model Validation and Cross-Validation

To evaluate the performance of the mode, cross validation was applied. The root mean square error (RMSE) was calculated, to assess the accuracy of the predictions. RMSE of 0.93 indicated that the model performed well, and was able to capture and analyse how the different SPI pillars affected the overall SPI score. To ensure that the model's performance was consistent across different sets of data, cross validation was used. This also reduced the risk of overfitting.

4.0 Report Presentations (Visualizations and Graphs)

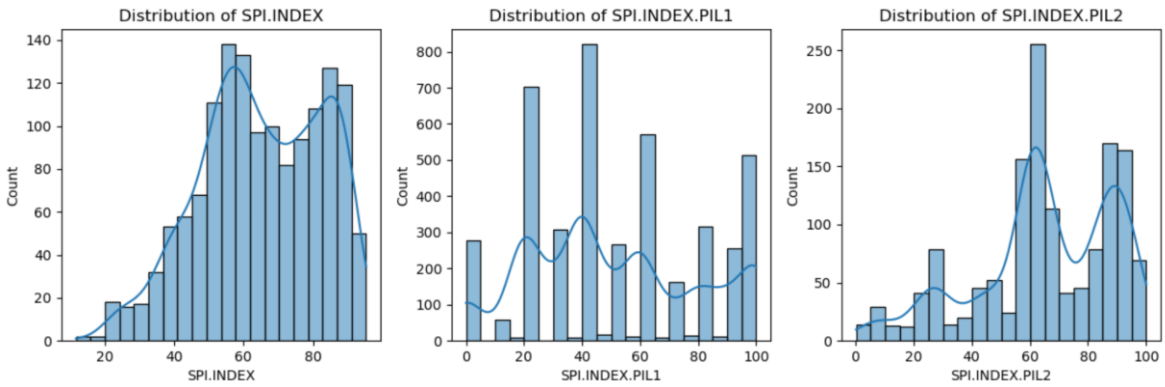
Several graphs were plotted to properly study the effect of each of those factors. They would be discussed one by one in the follow paragraphs:

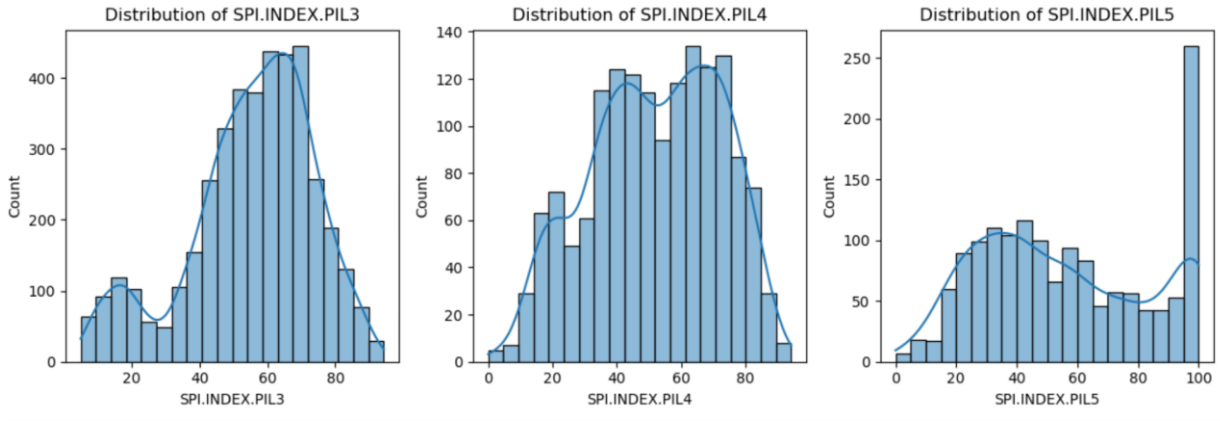
4.1 Question-1

4.1.1 Kernel Density Estimation

The histograms plotted with Kernel Density Estimation provides a comprehensive understanding of the distribution of the overall SPI score, and its five pillars. The SPI Index was centred around mid to high scores. This suggests that there were many countries that achieve moderate or more statistical maturity but there were some outliers that have lower scores. However, the SPI.INDEX.PL1 (Pillar 1 - Data Use) was more polarizing as the countries either excel or struggle in this metric.

For SPI.INDEX.PL2 (Pillar 2 - Data Services), the distribution of values was mostly around moderate and high scores, reflecting that most countries perform well in this pillar. SPI.INDEX.PL3 (Pillar 3 - Data Products) distribution was found to be left skewed with scores that were more in the moderate range, that shows steady progress but not much of exceptional performance. The SPI.INDEX.PIL4 (Pillar 4 - Data Sources) shows that the pillar has a balanced distribution and cantered with a moderate score. This shows that performance was mostly consistent. SPI.INDEX.PL5 (Pillar 5 - Data Infrastructure) distribution was found to be right skewed and has a peak near 100 showing some countries really excel here.

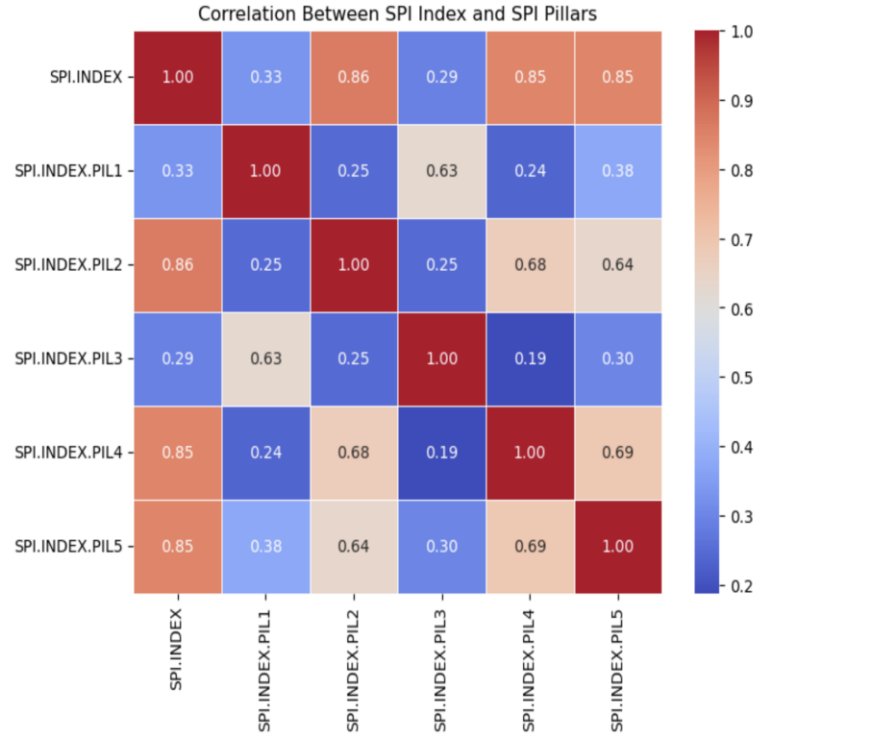




These insights highlight the varying levels of progress found across the different SPI pillars allowing for the identification of disparities. They also show the areas of strength and opportunities for improvement.

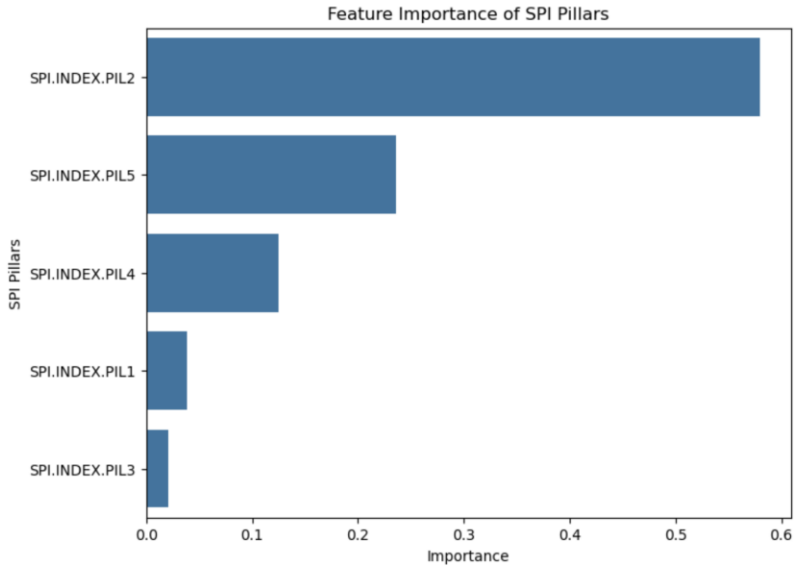
4.1.2 Correlation Analysis of SPI Index and Pillars

The heatmap illustrates the correlation between the overall Statistical Performance Indicators (SPI) Index (SPI.INDEX) and its five pillars (SPI.INDEX.PIL1 to SPI.INDEX.PIL5). There was a strong observable correlation between SPI.INDEX and PIL2 (0.86), PIL4 (0.85), and PIL5 (0.85) indicating that these three pillars were the most significant contributors to SPI Index Scores. In stark contrast, PIL1 (0.33) and PIL3 (0.29) show weaker correlations, suggesting that they do not factor in much in the overall score. The inter-pillar correlations, like PIL1 and PIL3 (0.63), show the relationships between specific aspects of statistical progress. These emphasize that the most significant factors in statistical progress were PIL2, PIL4, and PIL5, while PIL1 and PIL3, might be investigated further to enhance their overall contribution.



4.1.3 Feature Importance of SPI Pillars

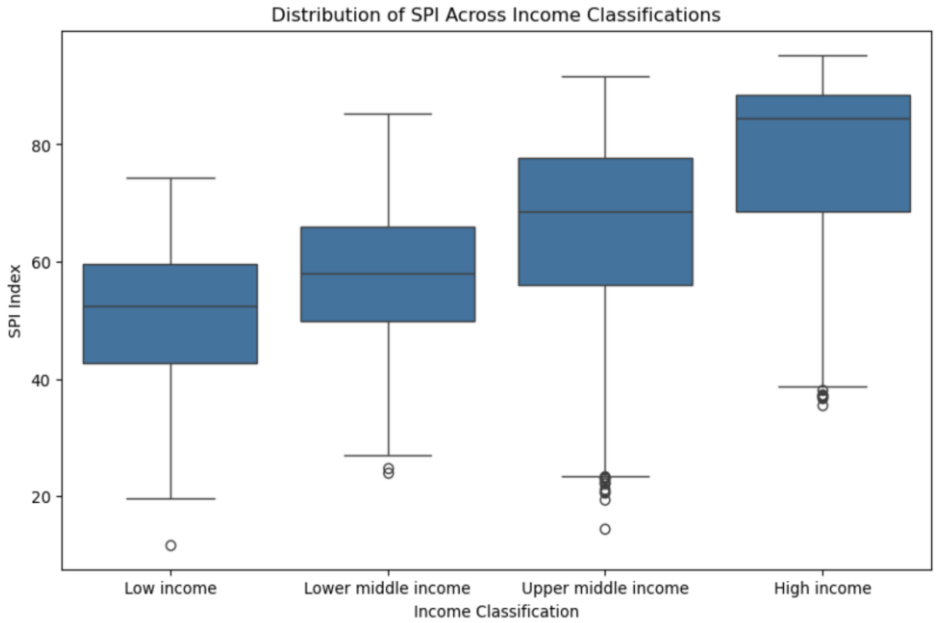
To display the feature importance of the SPI pillars in the overall SPI Index with a Random Forest Regressor a bar chart has been used. Pillar 2, Pillar 5, and Pillar 4 are the most influential contributors to the overall SPI score. Pillar 2 (Data Services) with almost 60% contribution to the score plays the most influential, underscoring the necessity of accessible and open data systems for proper governance and decision-making.



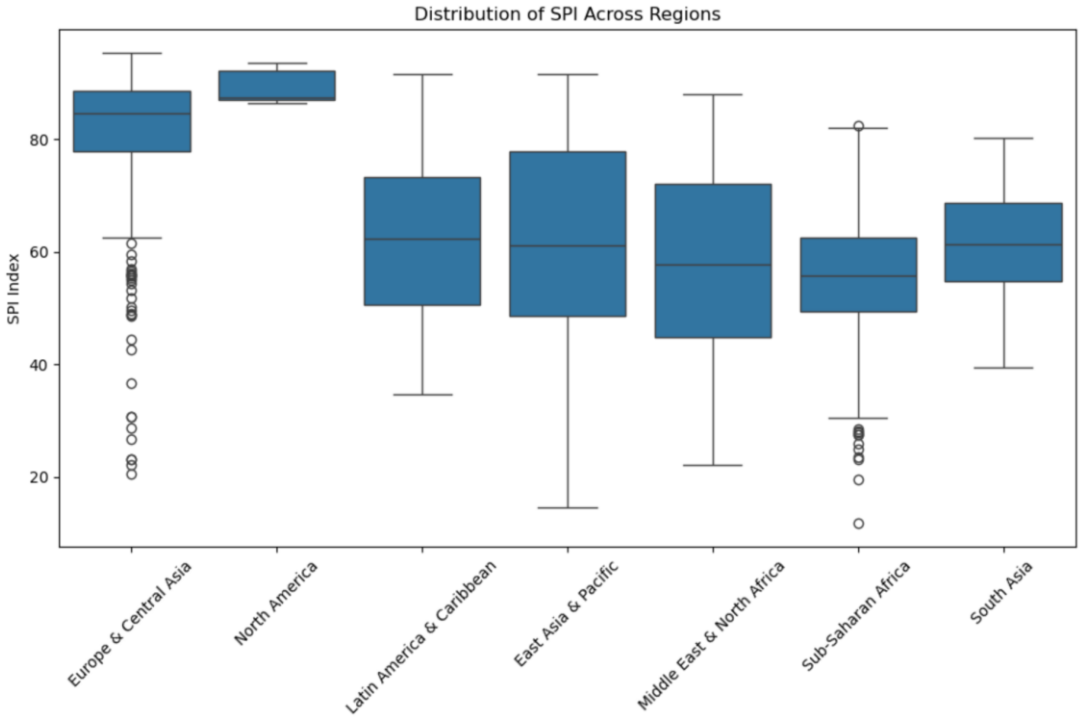
4.2 Question-2

4.2.1 Analysis of SPI Distribution Across Income Levels and Regions

The first graph shows how SPI scores vary across income classifications. The boxplot shows categories across income groups: Low income, Lower middle income, Upper middle income, and High income, and plots the SPI scores for each group. It was observed that countries with higher income exhibit higher SPI scores.

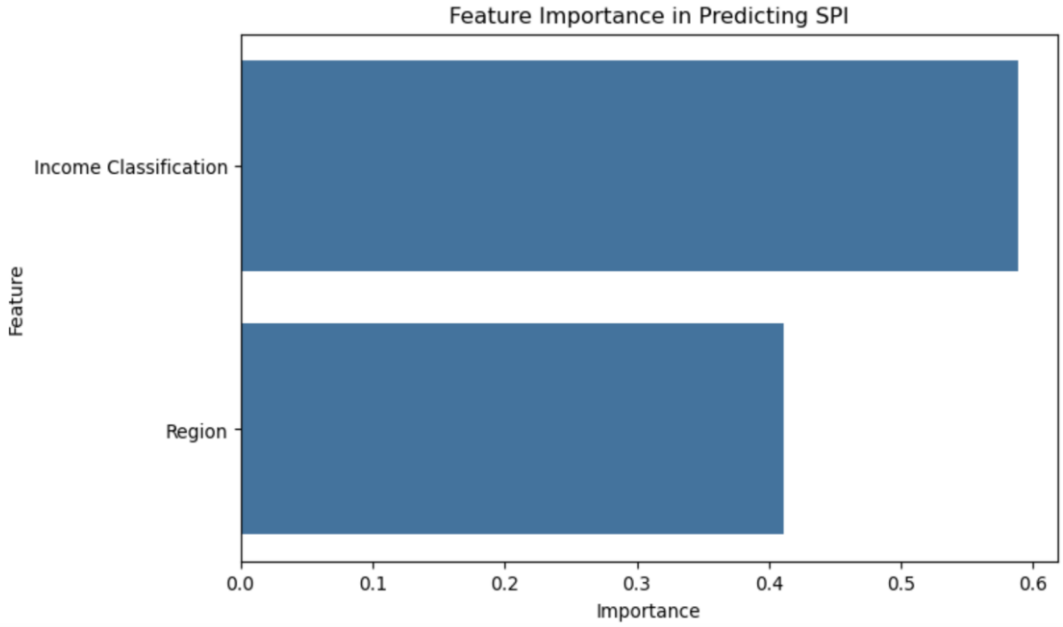


The second visualization shows the distribution of SPI across regions, Europe & Central Asia, North America, Sub-Saharan Africa, etc. A clear trend was found that countries in North America excel, whereas Africa and Asia display low median scores. There were outliers though, such as in Sub-Saharan Africa, that outperform relative to regional performance.



4.2.2 Feature Importance in Predicting SPI

The bar chart shows the importance of various factors when it comes to SPI. The most important contributing factor was income and then the region, which slows that socioeconomic situation of a nation indeed plays a more vital role than its geographical location in determining the maturity of its Statistical Infrastructure.



5.0 Conclusion

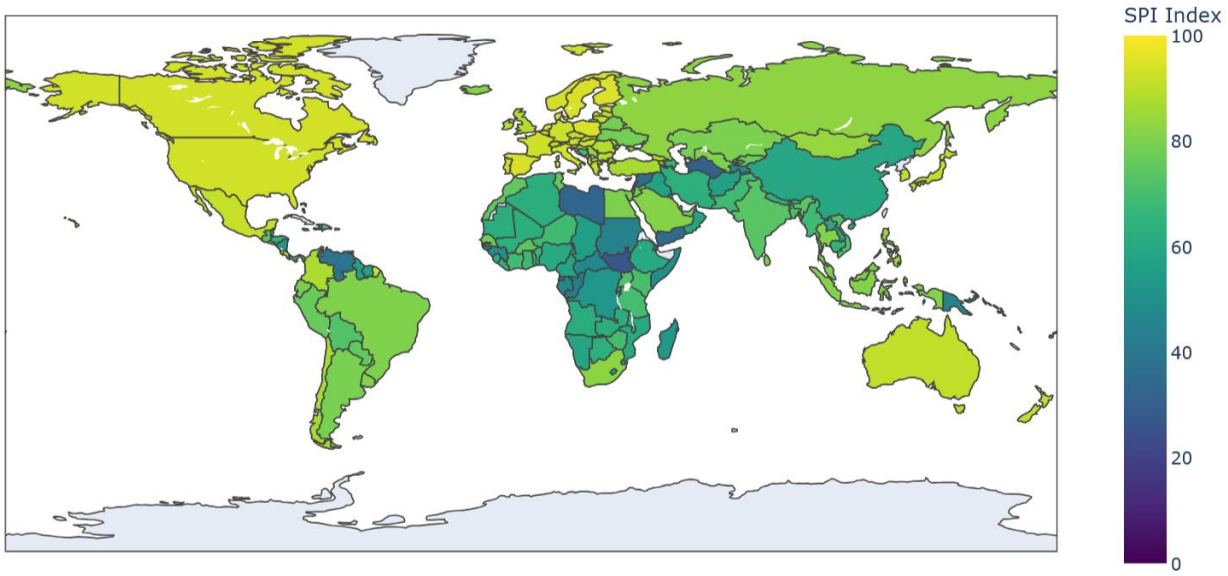
This project examined the Statistical Performance Index (SPI) and its key components, uncovering important insights into the global differences in statistical capacity and the main factors that affect SPI scores. Among the five pillars, Pillar 2 (Data Services), Pillar 5 (Data Infrastructure), and Pillar 4 (Data Sources) stood out as the most significant contributors to overall SPI scores. These results highlight the crucial role of strong data systems in national development. Enhancing these pillars can greatly boost statistical maturity and support evidence-based decision-making around the world.

The analysis of SPI scores across different income classifications and regions revealed significant disparities. High-income countries showed much higher SPI scores, indicating their robust statistical infrastructures. Regions such as Europe & Central Asia and North America excelled in statistical performance, while Sub-Saharan Africa and South Asia fell notably behind, highlighting the necessity for focused support in these underperforming areas. Additionally, it was found that income classification had a more crucial impact than geographical location on SPI scores, as demonstrated by the predictive model, which assigned 59% of the feature importance to income levels.

The global Choropleth Map of SPI scores of 2023 provide a visual understanding of the geographical distribution of SPI, reinforcing the disparities. High-income regions like North America and parts of Europe show consistently high scores, while several African nations faced challenges in achieving statistical maturity.

The Random Forest Regressor ML model successfully offered interpretable results through feature importance. The relatively low RMSE values and consistent performance across cross-validation demonstrated the robustness of the predictive approach. However, the analysis also highlighted areas for further improvement, such as addressing inter-pillar disparities and increasing the contributions of less impactful pillars like Pillar 1 (Data Use) and Pillar 3 (Data Services).

This project provides valuable insights for policymakers to focus their investments on essential SPI pillars and address disparities based on region and income. These findings lay the groundwork for future research that aims to promote equitable development by improving statistical capacity and implementing data-driven governance.



Choropleth Map of Global SPI Score distribution in 2023

6.0 Reflection

This project has been an extensive exploration of the Statistical Performance Indicators (SPI) Framework, its underlying Indexes and the various factors that lead to differences in statistical capacity among countries. Through data cleaning, exploratory analysis, and predictive modelling, valuable insights were gained not only into the technical elements of data science but also into the socioeconomic impacts of statistical systems.

One of the most enlightening aspects was realizing the power of simple yet effective methods like median imputation for handling missing data and label encoding for categorical variables. These preprocessing steps, though fundamental, were critical in ensuring the integrity and compatibility of the data for subsequent analysis and machine learning modelling. Additionally, visualizations such as boxplots, correlation heatmaps, and feature importance charts played a pivotal role in communicating complex findings simply demonstrating the indispensable nature of data visualization.

A key takeaway was the importance of selecting the right machine learning model. The Random Forest Regressor proved to be an excellent choice for this analysis due to its ability to model complex, non-linear relationships and provide interpretable results. However, the exercise also taught me the importance of validating model performance rigorously using techniques like cross-validation to avoid overfitting and ensure generalizability.

This project also highlighted the inherent challenges of working with real-world data, such as handling missing values, managing outliers, and ensuring the ethical use of data. These challenges required careful thought and decision-making, balancing statistical rigor with practical considerations. The importance of domain knowledge and contextual understanding in guiding analytical choices and deriving meaningful insights was understood.

Finally, the project reinforced the importance of storytelling in data analysis. Beyond the technical findings, it became evident that the insights derived from the data need to be communicated effectively to drive action. This experience has equipped me with a holistic

perspective on data-driven problem-solving, blending technical expertise with an appreciation for socioeconomic contexts and policy implications. Moving forward, I aim to leverage these learnings to tackle more complex problems and contribute to impactful, evidence-based solutions.

References

World Bank. (2021). *Measuring the statistical performance of countries: An overview of updates to the World Bank Statistical Capacity Index*. World Bank. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/815721616086786412/measuring-the-statistical-performance-of-countries-an-overview-of-updates-to-the-world-bank-statistical-capacity-index

Appendix

Github Project Link: https://github.com/aniruddhabose/BEMM457-Final-Project