1. **Introduction**

**1.1 Problem Description**

In our rapidly evolving world and ever-changing global landscape 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, a 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 countries.

**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 sensitive or personally identifiable information which guarantees 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 are missing. Any columns with 70% or more data missing were excluded, however the number of such columns were extremely small. 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 library pandas has been used to deal with all these missing values. As the SPI pillars are 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 we could finally move on with our data analysis and machine learning.

Same as was done for the first research question the dataset was loaded for the second question too. However, this time the primary focus was on country, region, income classifications, and SPI.INDEX. To ensure there is 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 was made into numerical representations (income\_encoded and region\_encoded), this allowed 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 pillar 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. The fact that this pillar was more concentrated towards a higher value suggested that it had a stronger overall influence on the SPI score.

For the second question, to understand the variation in SPI score across different regions for the second question, two boxplots were used. The first showed the SPI scores across income classifications. This trend was SPI score increase with higher income levels, showing a strong link between wealth and social progress. 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 shows that income levels plays an extremely important role in deciding statistical maturity.

### 3.4 Model Validation and Cross-Validation

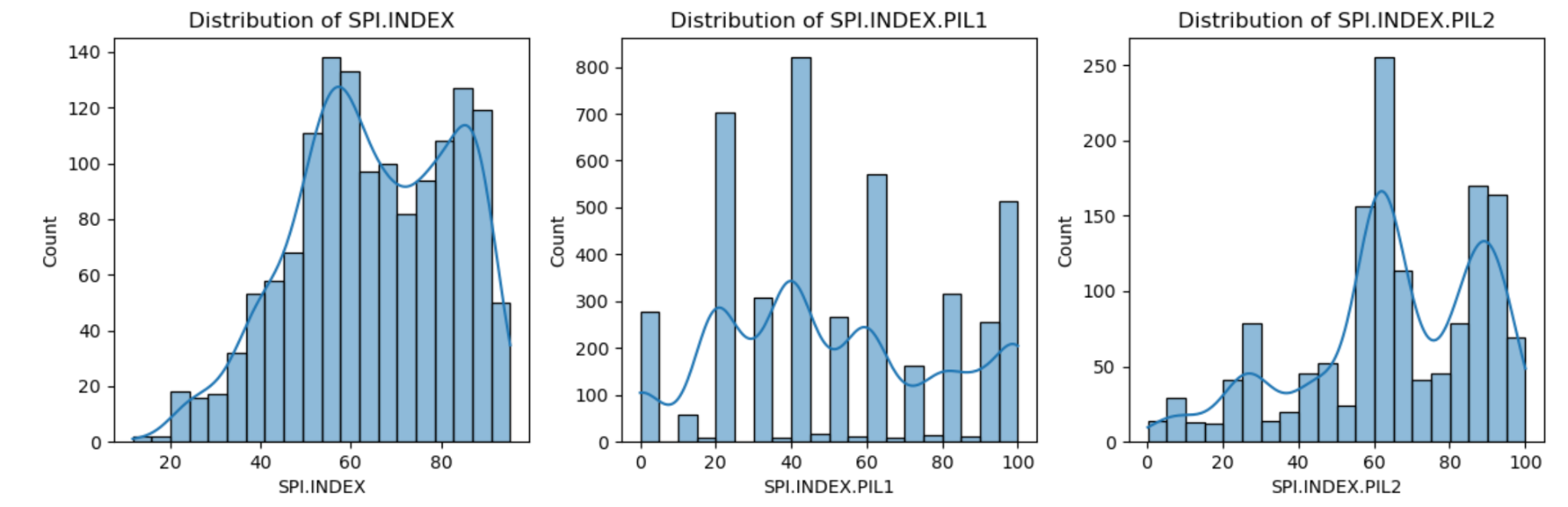
To evaluate the performance of the mode, 5 fold cross validation was applied. The root mean square error (RMSE) was calculated, to assess the accuracy of the predictions.   
RMSE: 0.92  
These indicate that the model performed really well , they were able to capture and analyse how the different SPI pillars affected the overall SPI score. In order 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.

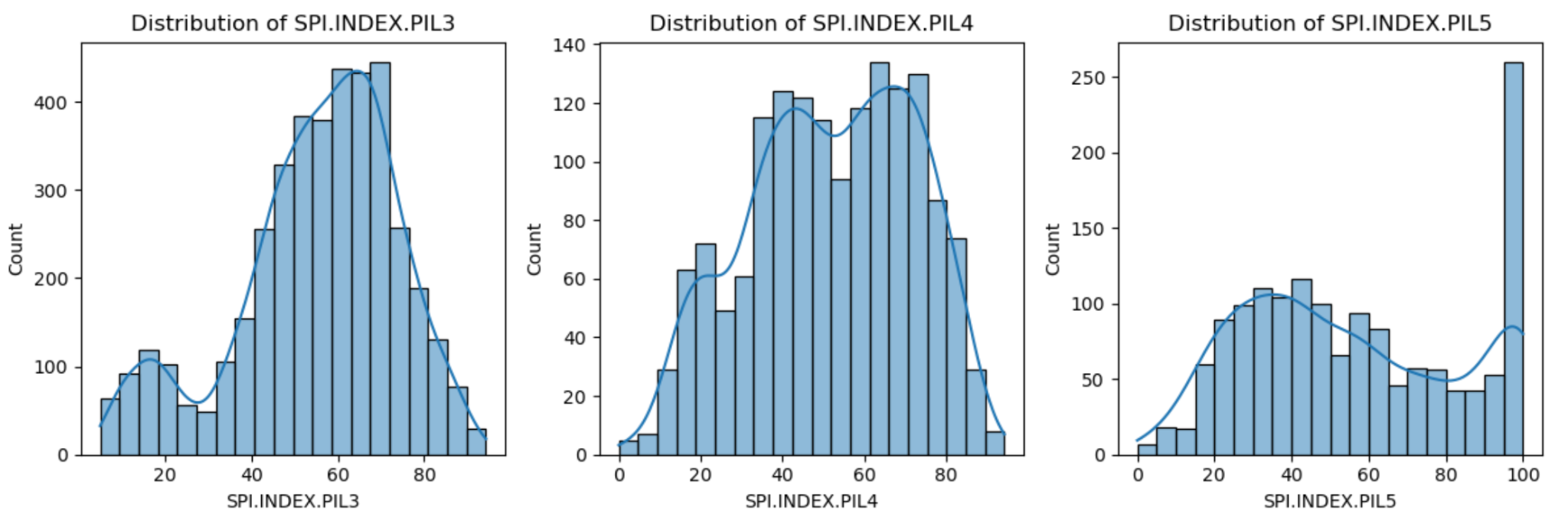
**4.0 Report Presentations (Visualizations and Graphs)**

In order to properly study the effect of each of those factors, several graphs were employed. They will be discussed one by one in the follow paragraphs:-

**4.0 Kernel Density Estimation (First question)**

There are histograms with Kernel Density Estimation, these provide a comprehensive understanding of the distribution of the overall SPI score, and its five pillars. The SPI index is symmetrically distributed and centred around mid to high scores. This suggests that there are many countries that achieve moderate social progress but there are some outliers that have lower scores. However, the SPI.INDEX.PL1 is more polarizing, this shows that countries either excel or struggle in this area.   
For SPI.INDEX.PL2, the distribution is mostly around moderate and high scores; this reflects that most countries do perform well in this pillar. However, SPI.INDEX.PL3, shows a left skewed distribution with scores that are more in the moderate range, the left skew shows that there is steady progress but no exceptional. The SPI.INDEX.PIL4 shows that the pillar has a balanced distribution and is centered with a moderate score. This shows that performance is really consistent, and the last SPI.INDEX.PL5 is right skewed and has a peak , this shows that the countries really excel here .



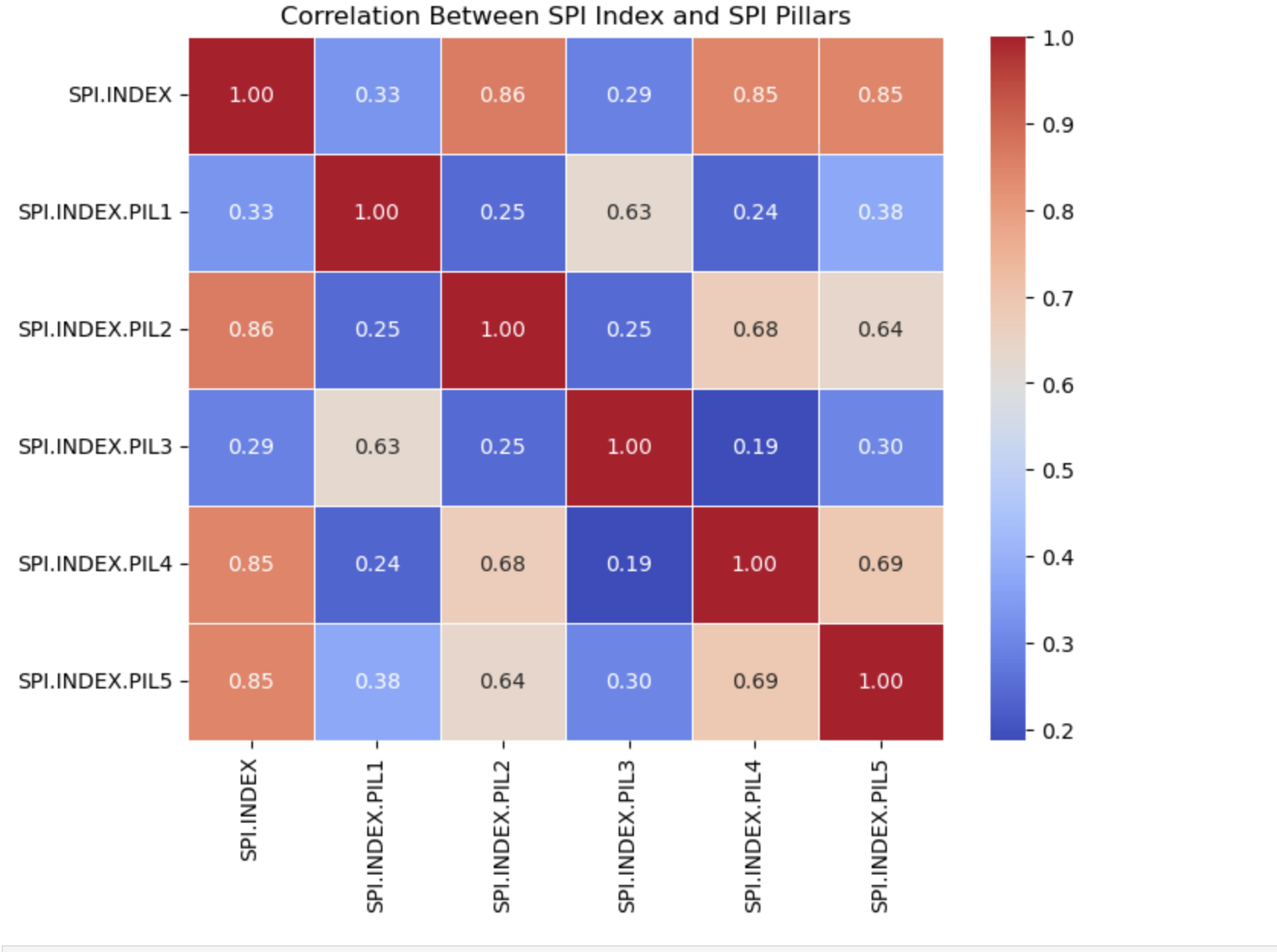
  
  
These insights highlight the varying levels of progress that are there across the different SPI pillars, this allows for the identification of disparities and shows the ares of strength and shows opportunities for improvement.

### 

### 

### 4.2 Correlation Analysis of SPI Index and Pillars(First Question)

The heatmap illustrates the correlation between the overall Social Progress Index (SPI.INDEX) and its five pillars (SPI.INDEX.PIL1 to SPI.INDEX.PIL5). There is a strong observable correlation between **PIL2 (0.86)**, **PIL4 (0.85)**, and **PIL5 (0.85),** this indicates a significant discovery. These pillars are the most significant contributors. In stark contrast, **PIL1 (0.33)** and **PIL3 (0.29)** show weaker correlations, this suggests that they don't factor in much in the overall score. The interpillar correlations , like **PIL1 and PIL3 (0.63),** show the relationships between specific aspects of social progress .These emphasize that the most significant factors in social progress are **PIL2, PIL4, and PIL5** , while **PIL1 and PIL3** , maybe be investigated further to enhance their overall contribution.



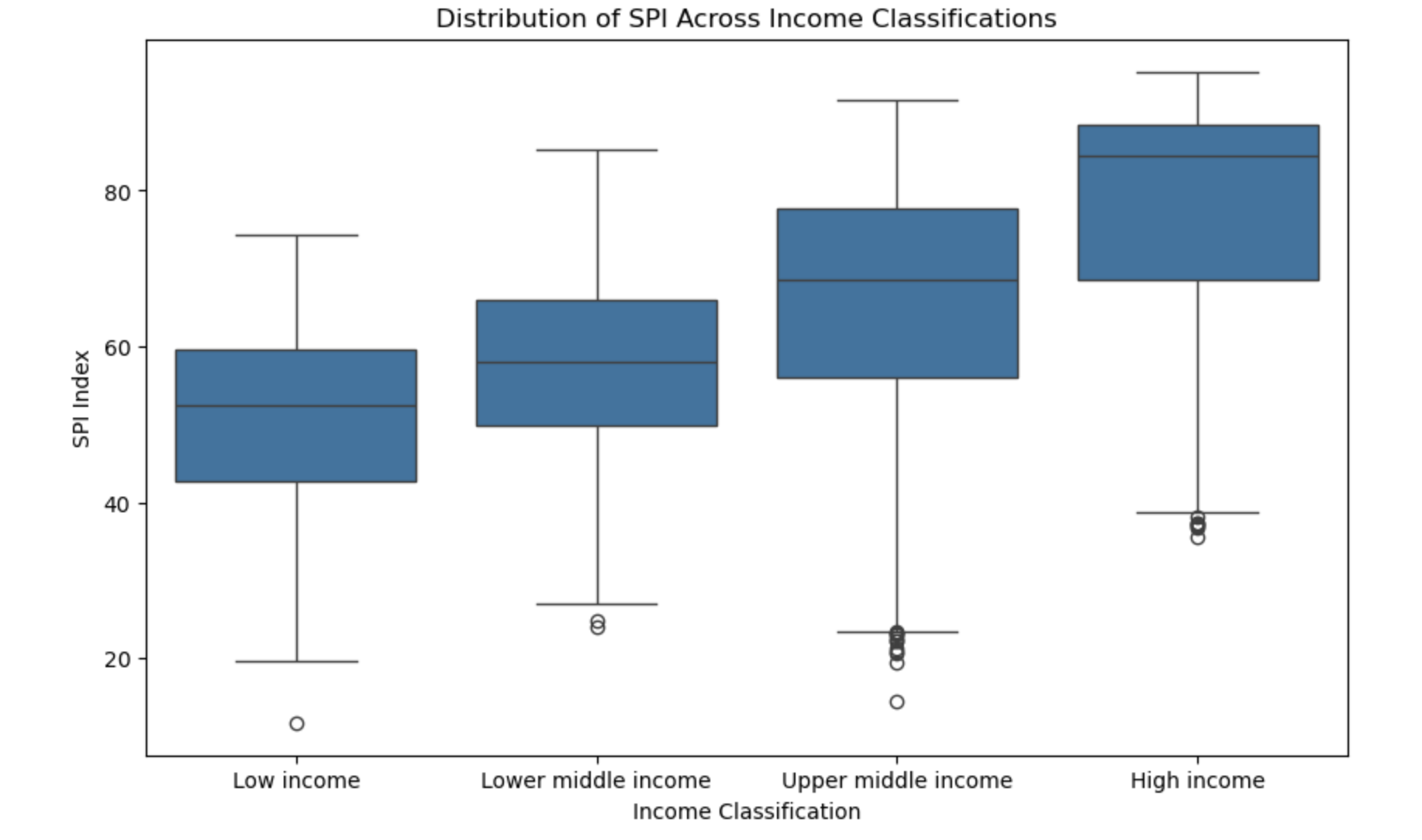
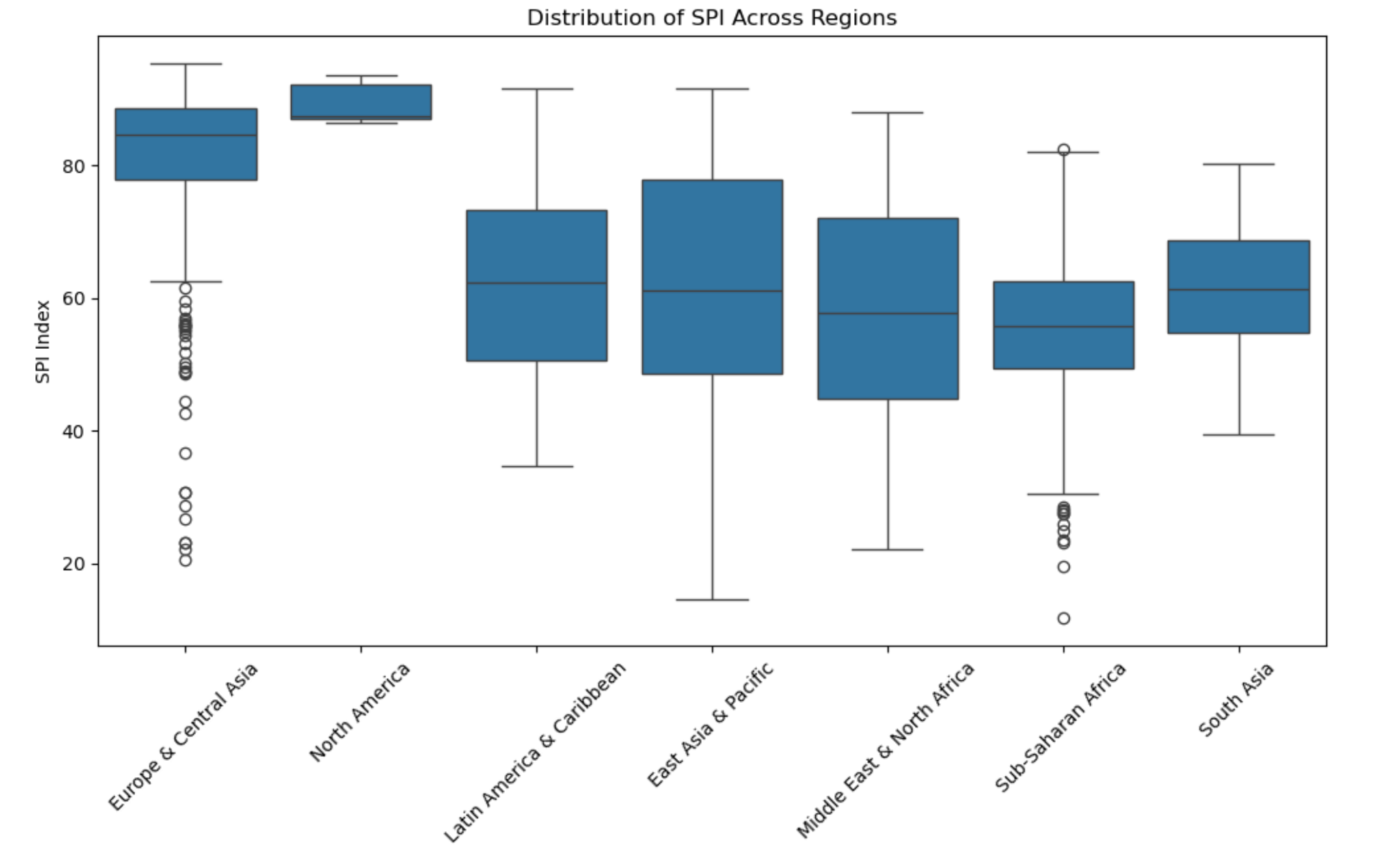
### 4.3 Feature Importance of SPI Pillars (First Question) In order to display the feature importance of the PI pillars, in the overall Social Progress Index, where a Random Forest , a bar chart was used. This shows that the most influential pillar is PIL2 with almost 60% contribution to the score, showing just how dominant rule PIL2 plays.

### 

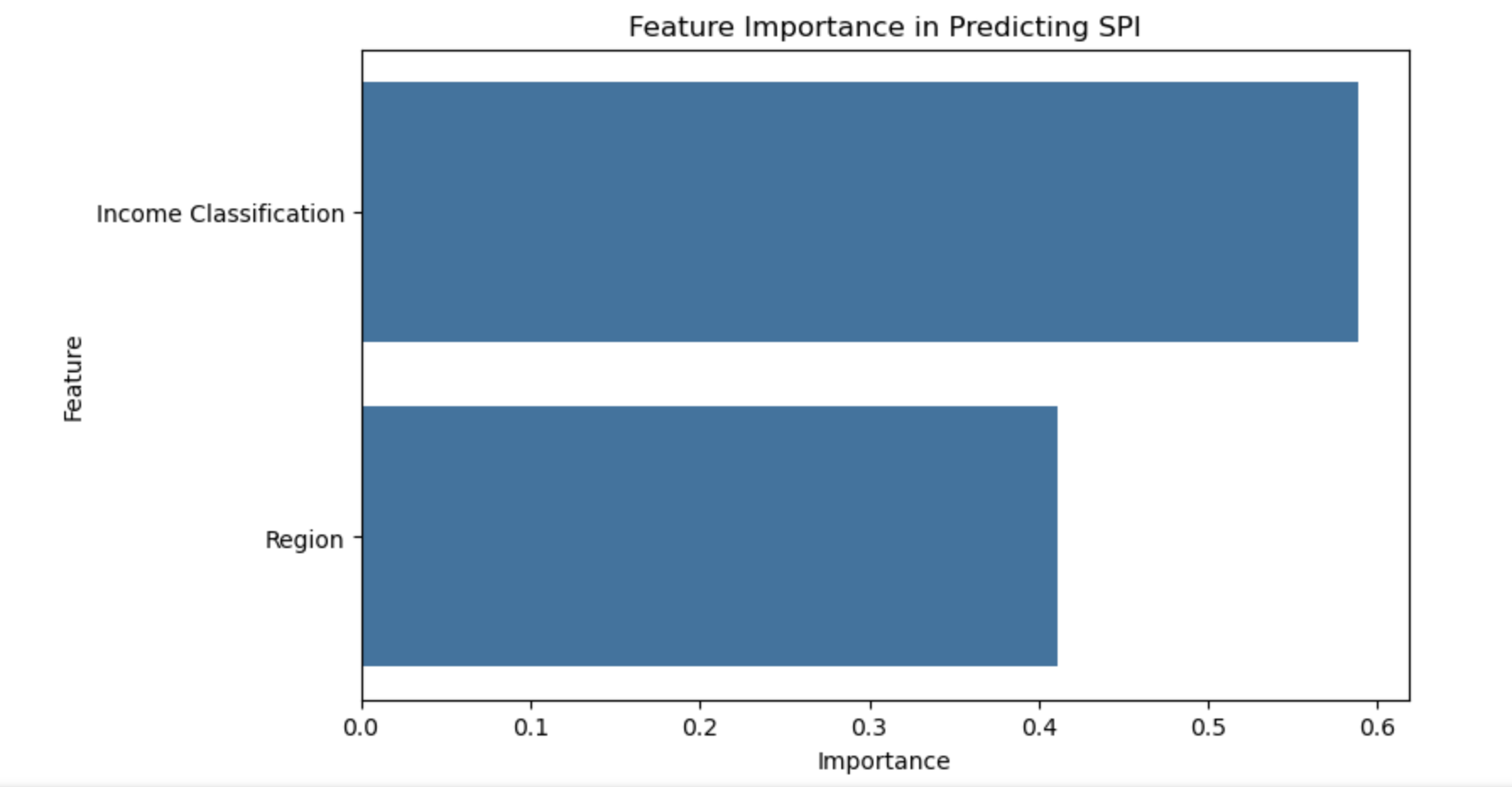
**Analysis of SPI Distribution Across Income Levels and Regions (Second question)**

The first graph shows how SPI varies across income classifications. The boxplot now will show categories across income groups: **Low income**, **Lower middle income**, **Upper middle income**, and **High income**, and plots the SPI scores for each group. In general we see a very clear trend, that countries with higher SPI exhibit higher SPI scores.

The second visualization shows the Distribution of SPI across regions,it compares across various regions, **Europe & Central Asia**, **North America**, and **Sub-Saharan Africa**. There is a clear trend that countries like Africa and Asia display really low median scores. There are outlier though, such as Sub-Saharan Africa, that outperforms relative to regional norms,



### Feature Importance in Predicting SPI

This bar chart shows the importance of various factors when it comes to SPI, this shows that the most contributing factor is income and then the region, which slows that socioeconomic status indeed plays a more vital role than geographical regions when it comes to SPI calculations.  


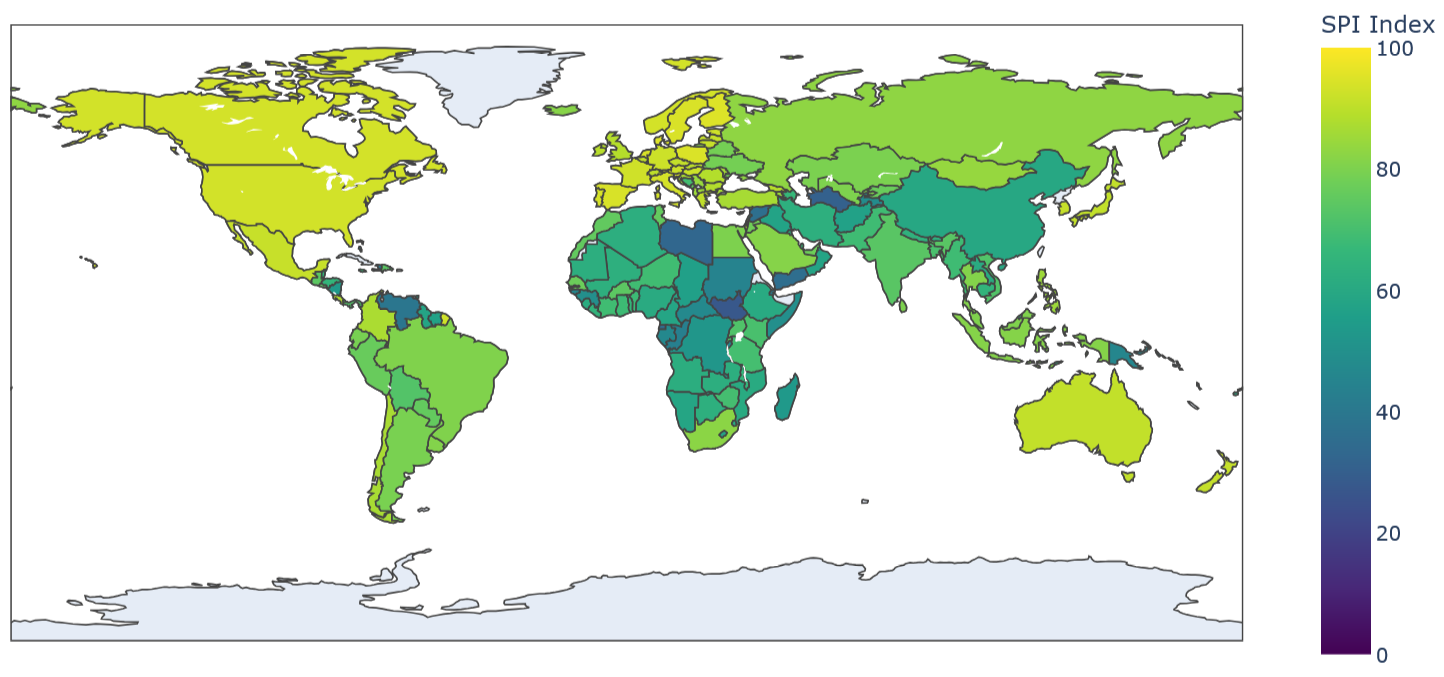
### Conclusion

The analysis of the Statistical Performance Index (SPI) and its contributing pillars has provided significant insights into the global disparities in statistical capacity and the factors influencing overall SPI scores. By examining SPI pillars, we identified that **PIL2** , **PIL5** , and **PIL4** are the most influential contributors to the overall SPI score, emphasizing the critical importance of these areas in enhancing a country's statistical performance. This highlights that investment in wellbeing, opportunity creation, and foundational capabilities can lead to substantial improvements in a nation's statistical systems and social progress.

Further exploration of SPI scores across income classifications and regions revealed stark disparities. High-income countries consistently scored better in SPI, reflecting their robust statistical infrastructures and resource availability. Regions like Europe & Central Asia and North America exhibited significantly higher SPI scores compared to Sub-Saharan Africa and South Asia, underlining the need for targeted interventions in economically and statistically underperforming regions. The analysis also confirmed that **income classification** plays a more critical role than geography in determining SPI, with nearly 59% feature importance in predictive modeling.

The machine learning models, particularly Random Forest Regressor, successfully predicted SPI scores and offered interpretable results through feature importance. The relatively low RMSE values and consistent performance across cross-validation folds 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 **PIL1 (Data Use)** and **PIL3 (Data Services)**.

In conclusion, this project provided actionable insights for policymakers, emphasizing the need to strengthen pillars that contribute significantly to SPI and bridge regional and income-based disparities. It serves as a foundation for future studies aimed at improving statistical capacity globally and fostering equitable development through evidence-based decision-making.



### Reflection

This project has been a comprehensive journey into understanding the Statistical Performance Index (SPI) and the underlying factors contributing to disparities in statistical capacity across countries. The process of data cleaning, exploratory analysis, and predictive modeling has deepened my understanding of not just the technical aspects of data science but also the socioeconomic implications 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 modeling. Additionally, visualizations such as boxplots, correlation heatmaps, and feature importance charts played a pivotal role in communicating complex findings in a digestible format, demonstrating the indispensable value 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 underscored 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. Reflecting on this, I recognize the importance of domain knowledge and contextual understanding in guiding analytical choices and deriving meaningful insights.

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.