STATISTICAL PERFORMANCE INDICATORS ANALYSIS USING MACHINE LEARNING PYTHON

1.INTRODUCTION

The Statistical Performance Indicators measure the capacity and maturity of national statistical systems by assessing the use of data, the quality of services, the coverage of topics, the sources of information, and the infrastructure and availability of resources. The goal is to improve development outcomes and track progress toward the Sustainable Development Goals.

To focus on the most informative Dimension and subclassification of the SPI index feature selection techniques for supervised machine learning methods must be developed. This project aims to address this problem by feature engineering of the pillars of which is considered in farming the SPI score over years. This project explores the descriptive and explorative data analysis for the better understanding of the framework. Inferential analysis is done to understand the relationship of the attributes in developing the SPI scores. Machine learning algorithms is used for predictive analysis to deploy the prediction in the frameworks. Performance metrics is used to compare for better selection of predictive algorithms for deployment. Time series analysis is applied for forecasting with the observed scores of Statistical Performance Indicators of a time series to predict future time series scores.

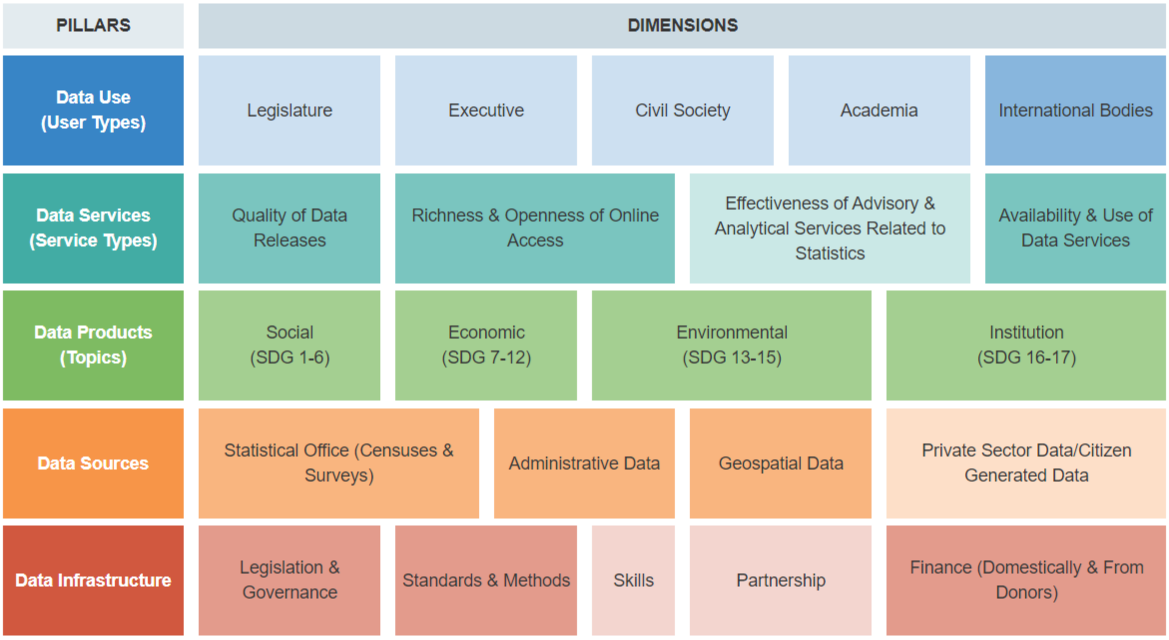
Framework of Statistical Performance Indicators (SPI)

The Statistical Performance Indicators (SPI) is a framework of 5 pillars and 22 dimensions to assess the maturity of national statistical systems. The matrix below provides definitions of each pillar and dimension.

The SPI framework assesses the maturity and performance of national statistical systems in five key areas, called pillars. The five pillars are:

Data Use: Statistics have value only if they are used. So, the first pillar is data use. A successful statistical system produces data that are used widely and frequently.
Data Services: A range of services connects data users to producers and facilitate dialogues between them, thus building trust and a sense of value.
Data Products: The dialogues between users and producers drive the design and range of statistical products and their accuracy, timeliness, frequency, comparability, and levels of disaggregation. The products signal whether countries can produce indicators related to the 17 Sustainable Development Goals.
Data Sources: To create useful products, the statistical system needs to draw on sources inside and outside the government. Data collection thus goes beyond the typical censuses and surveys to include administrative and geospatial data as well as data generated by private firms and citizens.
Data Infrastructure: A mature statistical system has well-developed hard infrastructure (legislation,

governance, standards) and soft infrastructure (skills, partnerships) as well as the financial resources to deliver useful—and widely used—data products and services.



2.DATA EXPLORATION

Data Source: The World Bank's Development Data Group contains various databases that have time series data on a multitude of topics for many countries around the world. This tool allows an individual to extract the specific information they require by choosing a certain database, data series, country or countries and year(s) of interest.

For reliable, usable, high-quality statistics are vital for global prosperity and progress. The Statistical Performance Indicators (SPI) data provide an open-source framework for assessing the performance of statistical systems and the efforts to improve them. This dataset is classified as Public under the Access to Information Classification Policy.

Database Atrributes Decription: Since 2004, the World Bank's Statistical Capacity Indicator (SCI) has been part of this global toolkit. This databases carries 2018-2022 year SPI scores. 186 countries data is explored for this project.

Attributes	Data Type	Description	Varibles as Examples
Attribute 1	Category	country	Finland,Norway,Canada
Attribute 2	Category	iso3c	FIN, NOR, CAN
Attribute 3	Date/time	year	2018 to 2022
Attribute 4	Numerical	Pillar 1 - Data Use - Score	1,2,3,….to.....,90,95,100
Attribute 5	Numerical	Pillar 2 - Data Services - Score	1,2,3,….to.....,90,95,100

Attribute 6	Numerical	Pillar 3 - Data Products - Score	1,2,3,….to.....,90,95,100
Attribute 7	Numerical	Pillar 4 - Data Sources - Score	1,2,3,….to.....,90,95,100
Attribute 8	Numerical	Pillar 5 - Data Infrastructure - Score	1,2,3,….to.....,90,95,100
Attribute 9	Numerical	SPI Overall Score	1,2,3,….to.....,90,95,100
Attribute 10	Numerical	Dimension 1.5: Data use by international organizations	0,0.06,0.004…to...0.09,1
Attribute 11	Numerical	Dimension 2.1: Data Releases	0,0.06,0.004…to...0.09,1
Attribute 12	Numerical	Dimension 2.2: Online access	0,0.06,0.004…to...0.09,1
Attribute 13	Binary	Dimension 2.4: Data services	0,1
Attribute 14	Numerical	Dimension 3.1: Social Statistics	0,0.06,0.004…to...0.09,1
Attribute 15	Numerical	Dimension 3.2: Economic Statistics	0,0.06,0.004…to...0.09,1
Attribute 16	Numerical	Dimension 3.3: Environmental Statistics	0,0.06,0.004…to...0.09,1
Attribute 17	Numerical	Dimension 3.4: Institutional Statistics	0,0.06,0.004…to...0.09,1
Attribute 18	Numerical	Dimension 4.1: Censuses	0,0.06,0.004…to...0.09,1
Attribute 19	Numerical	Dimension 4.1: Surveys	0,0.06,0.004…to...0.09,1
Attribute 20	Numerical	Dimension 4.2: Administrative Data	0,0.06,0.004…to...0.09,1
Attribute 21	Numerical	Dimension 4.3: Geospatial Data	0,0.06,0.004…to...0.09,1
Attribute 22	Numerical	Dimension 5.1: Legislation and governance	0,-99,1
Attribute 23	Numerical	Dimension 5.2: Standards and Methods	0,0.06,0.004…to...0.09,1
Attribute 24	Numerical	Dimension 5.5: Finance	0,-99,1
Attribute 25	Numerical	Dimension 1.5: Data use by international organisations - Availability of Comparable Poverty headcount ratio at $2.15 a day	0,0.5,1
Attribute 26	Binary	Dimension 1.5: Data use by international organisations - Availability of Mortality rate, under-5 (per 1,000 live births) data meeting quality standards according to UN IGME	0,1
Attribute 27	Numerical	Dimension 1.5: Data use by international organisations - Quality of Debt service data according to World Bank	0,0.06,0.004…to...0.09,1
Attribute 28	Numerical	Dimension 1.5: Data use by international organisations - Safely Managed Drinking Water	0,0.5,1
Attribute 29	Numerical	Dimension 1.5: Data use by international organisations - Labor force participation rate by sex and age (%)	0,0.5,1
Attribute 30	Numerical	Dimension 2.1: Data releases - SDDS/e-GDDS subscription	0,0.5,1
Attribute 31	Numerical	Dimension 2.2: Online access - Machine Readability Score	0,0.06,0.004…to...0.09,1
Attribute 32	Numerical	Dimension 2.2: Online access - Non-Proprietary format Score	0,0.06,0.004…to...0.09,1
Attribute 33	Numerical	Dimension 2.2: Online access - Download Options Score	0,0.06,0.004…to...0.09,1
Attribute 34	Numerical	Dimension 2.2: Online access - Metadata Available Score	0,0.06,0.004…to...0.09,1
Attribute 35	Numerical	Dimension 2.2: Online access - Terms of Use Score	0,0.06,0.004…to...0.09,1

Attribute 36	Numerical	Dimension 2.2: Online access - ODIN Open Data Openness score	0,0.06,0.004…to...0.09,1
Attribute 37	Binary	Dimension 2.4: Data access services - NADA metadata	0,1
Attribute 38	Numerical	Dimension 3.1: SDG Goal 1 - GOAL 1: No Poverty	0,0.06,0.004…to...0.09,1
Attribute 39	Numerical	Dimension 3.2: SDG Goal 2 - GOAL 2: Zero Hunger	0,0.06,0.004…to...0.09,1
Attribute 40	Numerical	Dimension 3.3: SDG Goal 3 - GOAL 3: Good Health and Well-being	0,0.06,0.004…to...0.09,1
Attribute 41	Numerical	Dimension 3.4: SDG Goal 4 - GOAL 4: Quality Education	0,0.06,0.004…to...0.09,1
Attribute 42	Numerical	Dimension 3.5: SDG Goal 5 - GOAL 5: Gender Equality	0,0.06,0.004…to...0.09,1
Attribute 43	Numerical	Dimension 3.6: SDG Goal 6 - GOAL 6: Clean Water and Sanitation	0,0.06,0.004…to...0.09,1
Attribute 44	Numerical	Dimension 3.7: SDG Goal 7 - GOAL 7: Affordable and Clean Energy	0,0.06,0.004…to...0.09,1
Attribute 45	Numerical	Dimension 3.8: SDG Goal 8 - GOAL 8: Decent Work and Economic Growth	0,0.06,0.004…to...0.09,1
Attribute 46	Numerical	Dimension 3.9: SDG Goal 9 - GOAL 9: Industry, Innovation and Infrastructure	0,0.06,0.004…to...0.09,1
Attribute 47	Numerical	Dimension 3.10: SDG Goal 10 - GOAL 10: Reduced Inequality	0,0.06,0.004…to...0.09,1
Attribute 48	Numerical	Dimension 3.11: SDG Goal 11 - GOAL 11: Sustainable Cities and Communities	0,0.06,0.004…to...0.09,1
Attribute 49	Numerical	Dimension 3.12: SDG Goal 12 - GOAL 12: Responsible Consumption and Production	0,0.06,0.004…to...0.09,1
Attribute 50	Numerical	Dimension 3.13: SDG Goal 13 - GOAL 13: Climate Action	0,0.06,0.004…to...0.09,1
Attribute 51	Numerical	Dimension 3.15: SDG Goal 15 - GOAL 15: Life on Land	0,0.06,0.004…to...0.09,1
Attribute 52	Numerical	Dimension 3.16: SDG Goal 16 - GOAL 16: Peace and Justice Strong Institutions	0,0.06,0.004…to...0.09,1
Attribute 53	Numerical	Dimension 3.17: SDG Goal 17 - GOAL 17: Partnerships to achieve the Goal	0,0.06,0.004…to...0.09,1
Attribute 54	Numerical	Dimension 4.1: censuses and surveys - Population & Housing census	0,0.5,1
Attribute 55	Numerical	Dimension 4.1: censuses and surveys - Agriculture census	0,0.5,1
Attribute 56	Numerical	Dimension 4.1: censuses and surveys - Business/establishment census	0,0.5,1
Attribute 57	Numerical	Dimension 4.1: censuses and surveys - Household Survey on income, etc	0,0.06,0.004…to...0.09,1
Attribute 58	Numerical	Dimension 4.1: censuses and surveys - Agriculture survey	0,0.5,1
Attribute 59	Numerical	Dimension 4.1: censuses and surveys - Labor Force Survey	0,0.06,0.004…to...0.09,1

Attribute 60	Numerical	Dimension 4.1: censuses and surveys - Health/Demographic survey	0,0.06,0.004…to...0.09,1
Attribute 61	Numerical	Dimension 4.1: censuses and surveys - Business/establishment survey	0,0.06,0.004…to...0.09,1
Attribute 62	Numerical	Dimension 4.2: administrative data - CRVS (WDI)	0,0.5,1
Attribute 63	Numerical	Dimension 4.3: geospatial data - Geospatial data available at 1st Admin Level	0,0.06,0.004…to...0.09,1
Attribute 64	Blank	Dimension 5.1: Legislation and governance - Legislation Indicator based on PARIS21 indicators on SDG 17.18.2	Blank
Attribute 65	Numerical	Dimension 5.2: standards - System of national accounts in use	0,0.5,1
Attribute 66	Numerical	Dimension 5.2: standards - National Accounts base year	0,0.5,1
Attribute 67	Numerical	Dimension 5.2: standards - Classification of national industry	0,0.5,1
Attribute 68	Numerical	Dimension 5.2: standards - CPI base year	0,0.5,1
Attribute 69	Binary	Dimension 5.2: standards - Classification of household consumption	0,1
Attribute 70	Numerical	Dimension 5.2: standards - Classification of status of employment	0,0.5,1
Attribute 71	Numerical	Dimension 5.2: standards - Central government accounting status	0,0.5,1
Attribute 72	Numerical	Dimension 5.2: standards - Compilation of government finance statistics	0,0.5,1
Attribute 73	Binary	Dimension 5.2: standards - Compilation of monetary and financial statistics	0,1
Attribute 74	Binary	Dimension 5.2: standards - Business process	0,1
Attribute 75	Numerical	Dimension 5.5: Finance - Finance Indicator based on PARIS21 indicators on SDG 17.18.3 & SDG 17.19.1	0
Attribute 76	Category	income	Low income, Lower middle income, Higher middle income, high income
Attribute 77	Category	region	Europe & Central Asia, North America, Latin America & Caribbean, East Asia & Pacific, Middle East & North Africa, South Asia
Attribute 78	Numerical	weights	1
Attribute 79	Numerical	population	389299,23332,87500

3.PROJECT OUTCOME

- To understanding the Explorative data analysis of Statistical Performance Indicators over years
- To compare appropriate machine learning algorithms for the factors of Statistical Performance Indicators are used to make a prediction or classification.
- To find the better Performance metrics and Accuracy in machine learning for assessing the effectiveness and reliability of models.
- To analyse the scores of data over time and may also be used to forecast future scores.

4.METHODOLOGY

4.1. Analysis of Variance (ANOVA) in Statistical Analysis

It is a statistical formula used to compare variances across the means (or average) of different groups. A range of scenarios use it to determine if there is any difference between the means of different groups.

$$\mathrm{F} = \frac{\mathrm{MST}}{\mathrm{MSE}}$$

where:

F = ANOVA coefficient

MST = Mean sum of squares due to treatment

MSE = Mean sum of squares due to error

4.2. Spearman's rank correlation in Statistical Analysis

It measures the strength and direction of association between two ranked variables. It basically gives the measure of monotonicity of the relation between two variables i.e. how well the relationship between two variables could be represented using a monotonic function.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i = difference in paired ranks and *n* = number of cases. The formula to use when there are tied ranks is:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

where *i* = paired score.

4.3. Linear Regression in Machine learning

Linear regression is also a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

$$y = b_0 + b_1 * x_1$$

A simple linear regression has an equation of the form **Y = b0+ b1*x1**, where **x1** is the predictor and **Y** is the dependent variable. The slope of the line is **b1**, and **b0** is the intercept (the value of y when x = 0).

4.4. Regularization in Machine learning

It is a set of methods for reducing overfitting in machine learning models. Lasso and Ridge Regression are two popular regularization techniques used to prevent overfitting and improve the accuracy of linear Regression models. Lasso shrink some coefficients to zero, effectively performing feature selection. (features-eliminating). Ridge tends to shrink coefficients but usually doesn't zero them out completely. (features- shrinking)

Mathematical Function of Ridge Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

Lambda (λ) in the equation is a tuning parameter (also referred to as **regularization parameter**)selected using a cross-validation technique that makes the fit small by making squares small (β^2) by adding a shrinkage factor.

The shrinkage factor is lambda times the sum of squares of regression coefficients (The last element in the above equation).

Mathematical Function of LASSO Regression

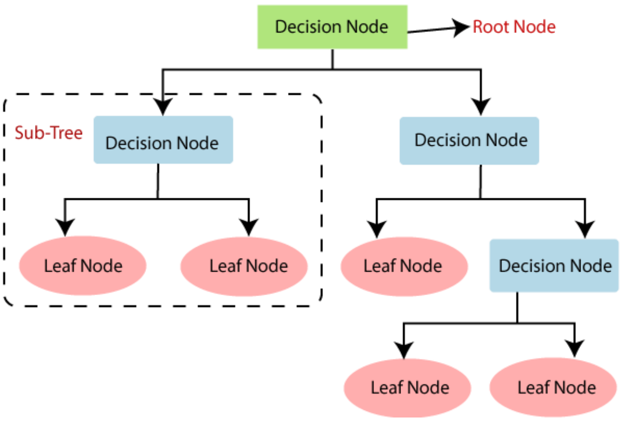
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

The above equation represents the formula for Lasso Regression! where Lambda (λ) is a tuning parameter selected using the before Cross-validation technique. Unlike Ridge Regression, Lasso uses |β| to penalize the high coefficients.

The shrinkage factor is lambda times the sum of Regression coefficients (The last factor in the above equation).

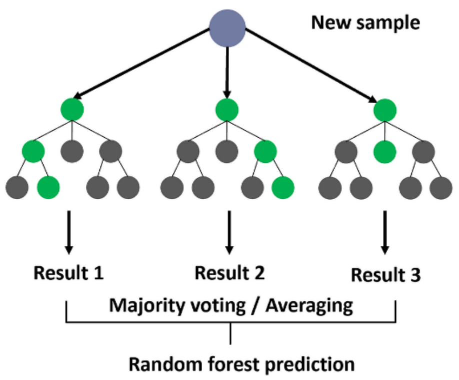
4.5. Decision Tree in Machine learning

A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes. for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.



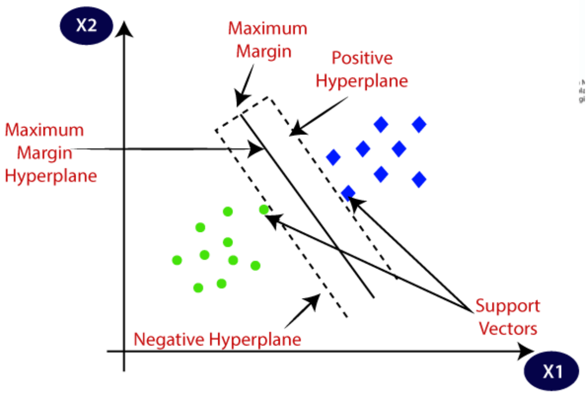
4.6. Random Forest in Machine learning

Random Forest algorithm is a powerful tree learning technique in Machine Learning. It works by creating a number of Decision Trees during the training phase. Each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance.



4.7. Support Vector Machine (SVM) in Machine learning

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.



4.8. Time Series Decomposition Techniques

Time series data consists of observations taken at consecutive points in time. These data can often be decomposed into multiple components to better understand the underlying patterns and trends. Time series decomposition is the process of separating a time series into its constituent components, such as trend, seasonality, and noise. By separating these components, we can gain insights into the behavior of the data and make better forecasts.

Time series decomposition helps us break down a time series dataset into three main components:

1. **Trend:** The trend component represents the long-term movement in the data, representing the underlying pattern.
2. **Seasonality:** The seasonality component represents the repeating, short-term fluctuations caused by factors like seasons or cycles.
3. **Residual (Noise):** The residual component represents random variability that remains after removing the trend and seasonality.

4.9. ARIMA model in Time Series analysis

Time series analysis involves analysing data points collected or recorded at specific time intervals to identify patterns, trends, and relationships over time. Autoregressive modeling and Moving Average modeling are two different approaches to forecasting time series data. ARIMA integrates these two approaches, hence the name. Forecasting is a branch of machine learning using the past behaviour of a time series to predict the one or more future values of that time series.

4.10. Performance Metrics

Mean Squared Error (MSE)

This metric is widely used to evaluate regression models. It represents the average of the squared difference between the original and predicted values. The importance of using MSE in identifying outliers and imbalances in the dataset. MSE will always be non-negative and in simpler terms, the lower the value, the better the fit.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Accuracy Score

The Accuracy score (or just Accuracy) is a Classification metric featuring a fraction of the predictions that a model got right. The metric is prevalent as it is easy to calculate and interpret. Also, it measures the model's performance with a single value. Accuracy is the proportion of all classifications that were correct, whether positive or negative. It is mathematically defined as:

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Confusion Matrix

A confusion matrix is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions.

True Class \ Predicted Class	A	B	C
A	TP	FN	FN
B	FP	TN	FN
C	FP	FN	TN

5. ALGORITHM FRAMEWORK

5.1. Data Preprocessing:

Data Preprocessing can be defined as a process of converting raw data into a format that is understandable and usable for further analysis. It is an important step in the Data Preparation stage. It ensures that the outcome of the analysis is accurate, complete, and consistent. Missing values are replaced using Simple Imputer.

Simple Imputer is a scikit-learn class which is helpful in handling the missing data in the predictive model dataset. It replaces the NaN values with a specified placeholder like ‘mean'(default), ‘median’, ‘most_frequent’ and ‘constant’.

Standardization scales each input variable separately by subtracting the mean (called centering) and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one. We will use the default configuration and scale values to subtract the mean to center them on 0.0 and divide by the standard deviation to give the standard deviation of 1.0.

Label Encoding is a technique that is used to convert categorical columns into numerical ones so that they can be fitted by machine learning models which only take numerical data. It is an important pre-processing step in a machine-learning project.

Feature Engineering: It is a critical step in the data preprocessing phase of machine learning and data analysis. It involves creating, modifying, or selecting features to improve the performance of a model.

One common technique in feature engineering is Principal Component Analysis (PCA). Principal Component Analysis (PCA) is a dimensionality reduction technique used to reduce the number of features in a dataset while retaining most of the variability (information) present in the data. PCA is applied for the dimensions of 5 pillars for dimension reduction. Standard Scalar is used before apply PCA for better dimensionality reduction

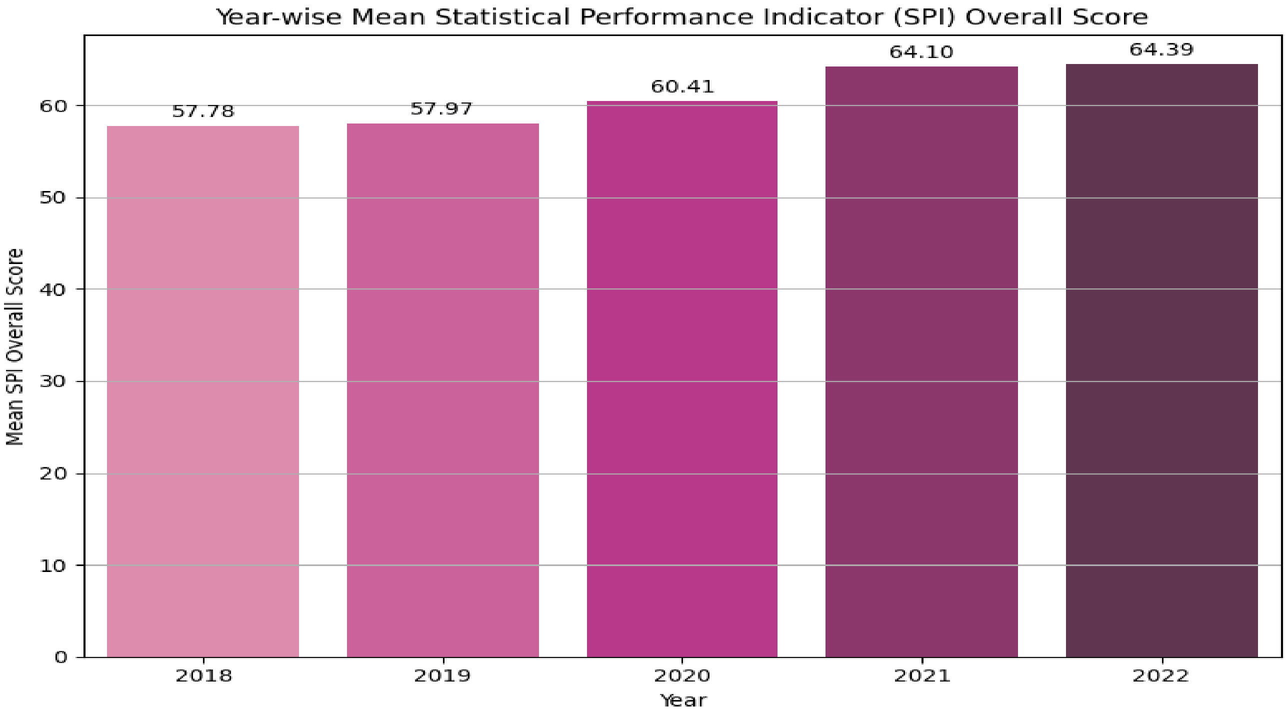
5.2. Exploratory Data Analysis (EDA):

Exploratory Data Analysis (EDA) is an analysis approach that identifies general patterns in the data. These patterns include outliers and features of the data that might be unexpected.

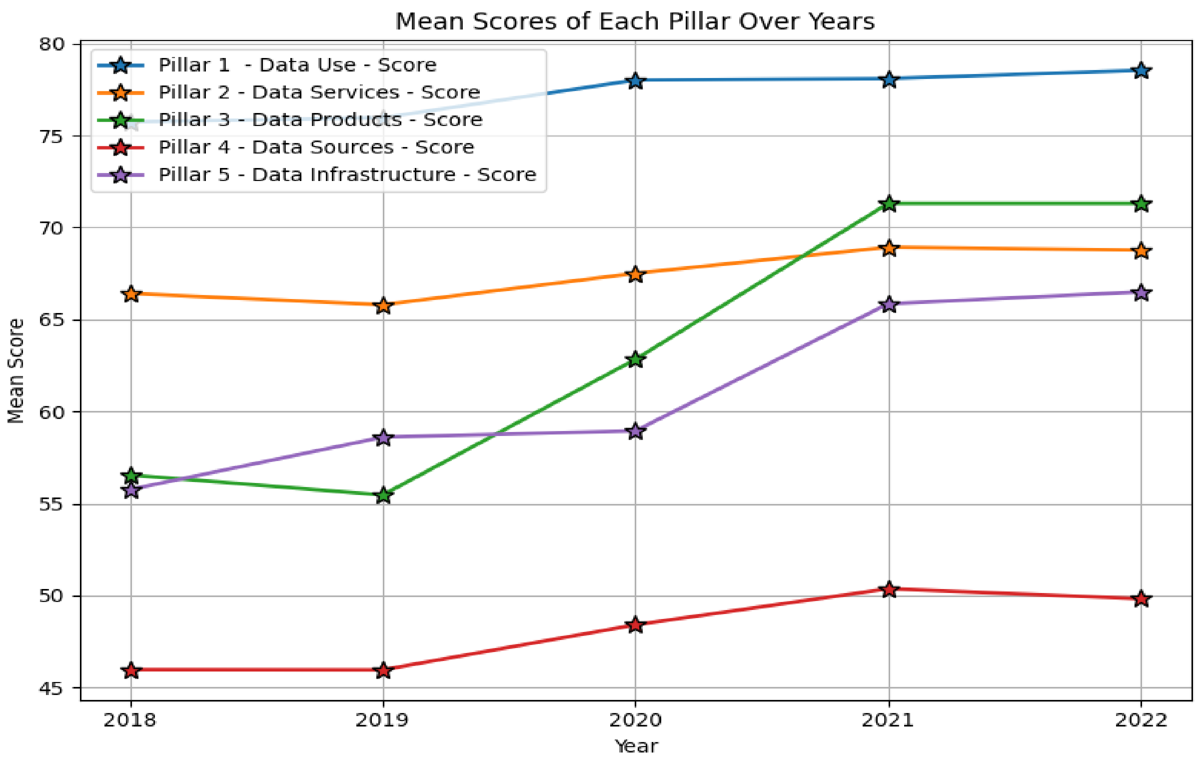
EDA is an important first step in any data analysis. Understanding where outliers occur and how variables are related can help one design statistical analyses that yield meaningful results. In biological monitoring data, sites are likely to be affected by multiple stressors.

Data Visualization:

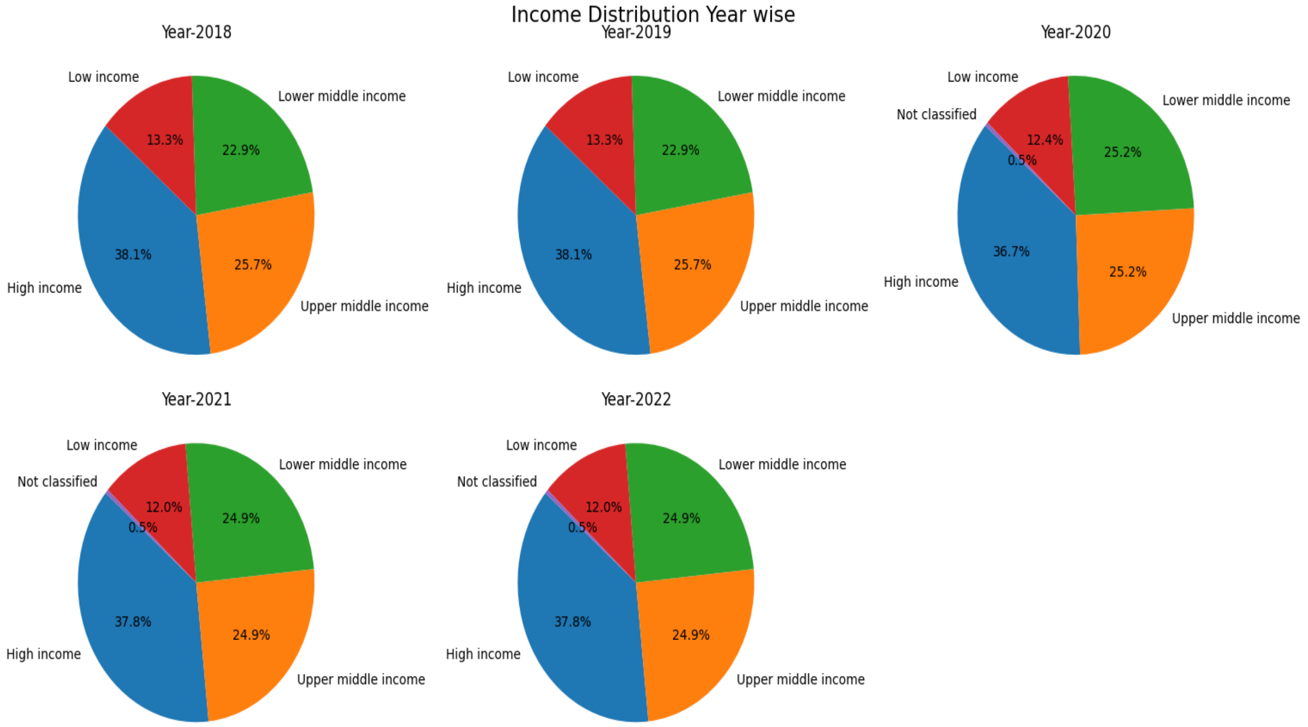
Statistical performance Indicator Overall Score year wise



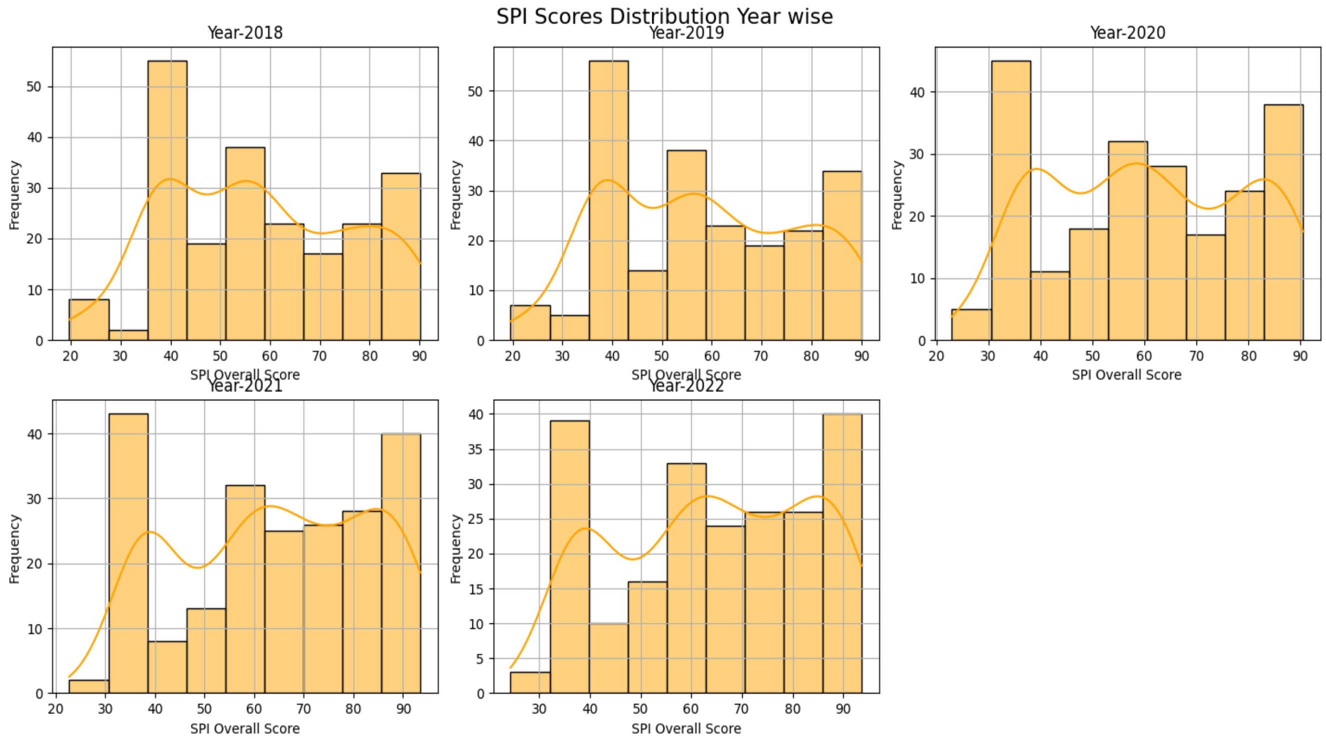
Pillars Score Year wise



Income levels Year wise



Population Year Wise Under Normally Distribution with respect to SPI score



5.3. Model Building and Evaluation:

5.3.1. Inferential Data Analysis

1. Analysis of Variance (ANOVA) For the 5 Pillars of Statistical Performance Indicators Data Use, Data Service, Data Product, Data Sources and Data Infrastructure.

Hypothesis Framing

Null Hypothesis (H0): There are no differences in the mean scores across the different pillars

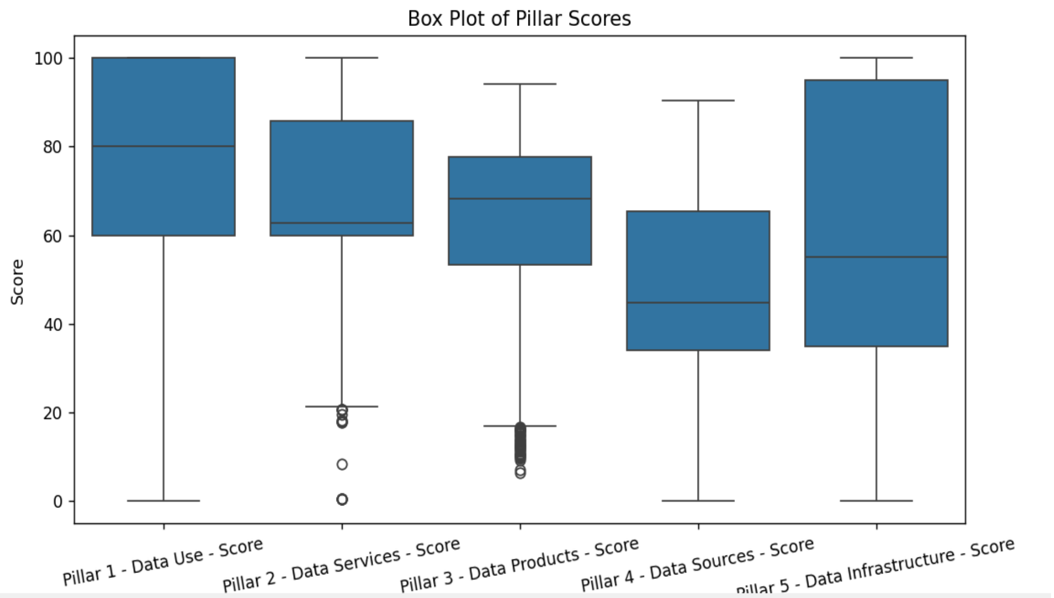
Alternative Hypothesis (H1): There is at least one pillar with a mean score that is significantly different from the others.

Output: F-statistic:237.953, P-value: 0.000

Interpretation of Results:

F-statistic: A higher F-statistic indicates a greater disparity between the group means relative to the variance within the groups.

P-value: If the p-value is less than your significance level (commonly 0.05), you reject the null hypothesis, indicating that **there is a significant difference in mean scores among at least some of the pillars**. If the p-value is greater than 0.05, you do not reject the null hypothesis.



Explanation of the Plot

This above box plot clearly shows that there least one pillar has outliers with a mean score which makes it significantly different from the others.

2. Spearman's rank correlation coefficient for categorical column Income and Numerical Colum Population.

Hypothesis Framing

Null Hypothesis (H_0)

The null hypothesis states that there is no monotonic relationship between the income ranks and the population values. This means that changes in the income category ranks do not systematically relate to changes in the population values.

Alternative Hypothesis (H_1)

The alternative hypothesis states that there is a monotonic relationship between the income ranks and the population values. This means that changes in the income category ranks are systematically related to changes in the population values.

Output: Spearman's rank correlation coefficient: 0.368, p-value: 0.000

Interpretation of Results:

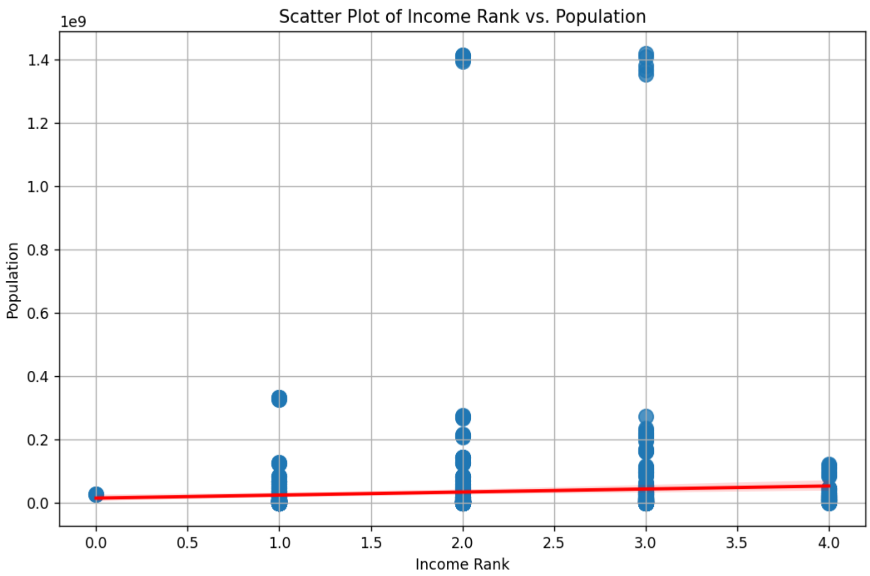
<u>Spearman's Rank Correlation Coefficient (ρ = 0.368):</u>

Magnitude: The coefficient of 0.368 suggests a moderate positive monotonic relationship between the ranked income categories and the numerical population values.

Direction: Since the coefficient is positive, it indicates that higher-ranked income categories are associated with higher population values.

<u>p-value (0.000):</u>

Statistical Significance: The p-value of 0.000 (which is less than 0.05) indicates that the observed correlation is statistically significant.



Explanation of the Plot

Scatter Points: Each point represents an observation in your dataset, plotted according to its income rank and population value.

Regression Line: The red line shows the trend in the data, indicating the general direction of the relationship. This line helps in visualizing the monotonic relationship**.**

5.3.2. Predictive Data Analysis

Supervised Machine Learning Algorithm- Regression

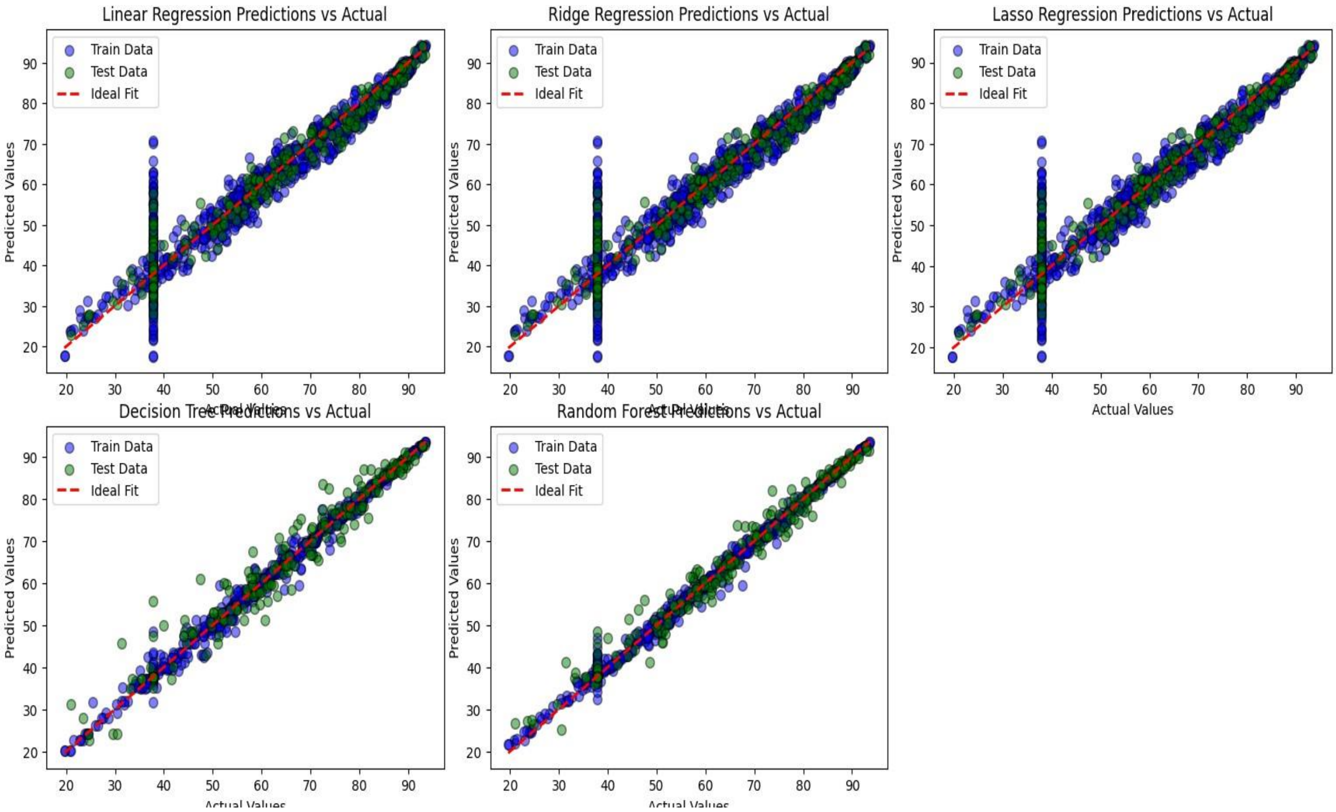
1. **Feature variable as 5 pillars of the Statistical Performance Indicators and target variable as SPI over all scores**

- 80 percent of dataset is used for training and fit the model, and 20 percent dataset is used for testing and evaluate the model.
- Hyperparameter tuning with Grid Search is used for testing a range of multiple hyperparameters for a machine learning model to find the best combination.
- Performance metric Mean Squared Error (MSE) is used to evaluate the performance of regression models. It measures the average squared difference between predicted values and actual values. A lower MSE indicates better model performance.

Machine Learning Algorithm	Training MSE	Testing MSE	Interpretation
Linear Regression- Grid Search CV	27.84	16.54	Overfitting and High Mean Squared Error
Ridge Regression- Grid Search CV	27.84	16.53	Overfitting and High Mean Squared Error
Lasso Regression- Grid Search CV	27.84	16.59	Overfitting and High Mean Squared Error
Decision Tree Regression- Grid Search CV	2.13	14.50	Comparable Better Performance and Overfitting the Training Data
Random Forest Regressor- Grid Search CV	1.47	7.77	Lowest Testing MSE and Avoiding Overfitting

Interpretation

- Linear, Ridge, and Lasso Regression: All show high training and testing MSE, suggesting overfitting and high error rates. They generally perform similarly, with marginal differences in error values.
- Decision Tree Regression: Has a very low training MSE but high testing MSE, indicating severe overfitting.
- Random Forest Regressor: Provides the best performance with the lowest testing MSE, indicating effective generalization and less overfitting compared to other models.



Explanation of the Plot

- Linear, Ridge, and Lasso Regression: Training and testing data is missing the ideal fit line which generally increase the marginal differences in error values.
- Decision Tree Regression: Training fits the line but testing data overfits the model leads to high testing error values.
- Random Forest Regressor: Provides the best performance with best fit line and with less error values.

Supervised Machine Learning Algorithm- Classifier

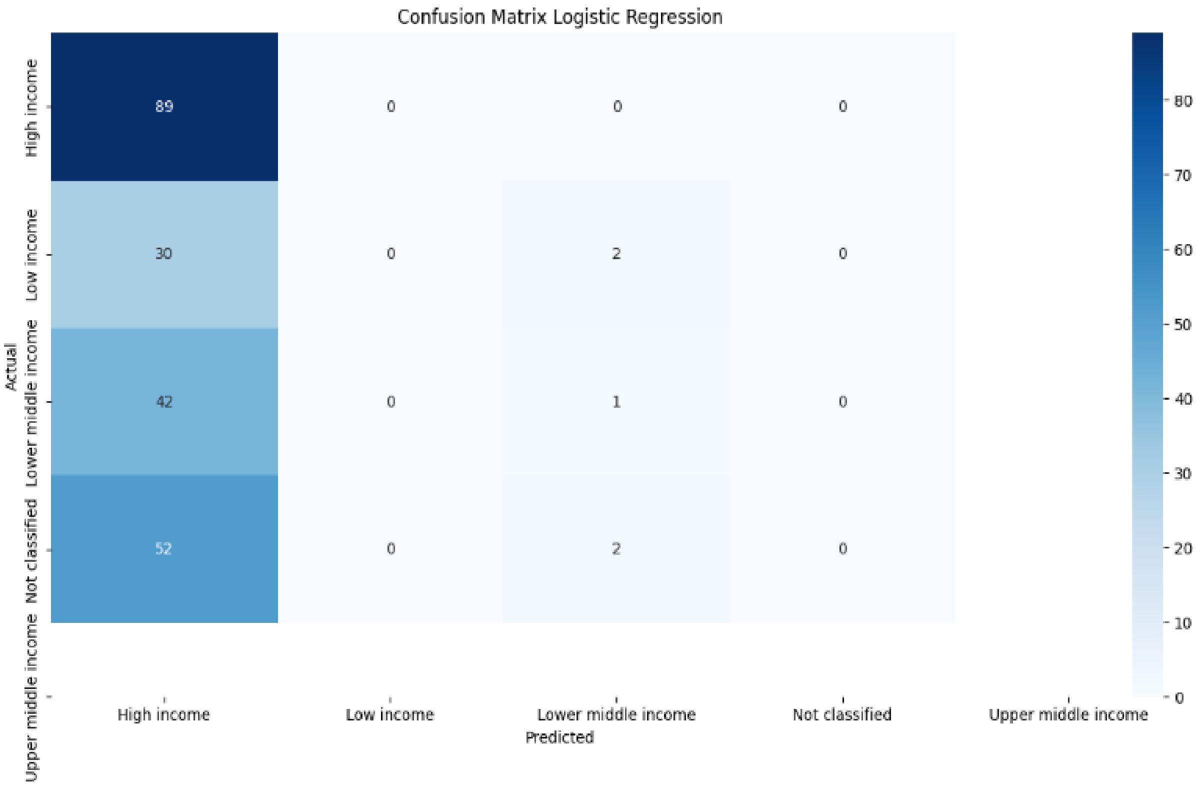
1. **Feature variable as as SPI over all scores and target variable as income groups of Statistical Performance Indicators**

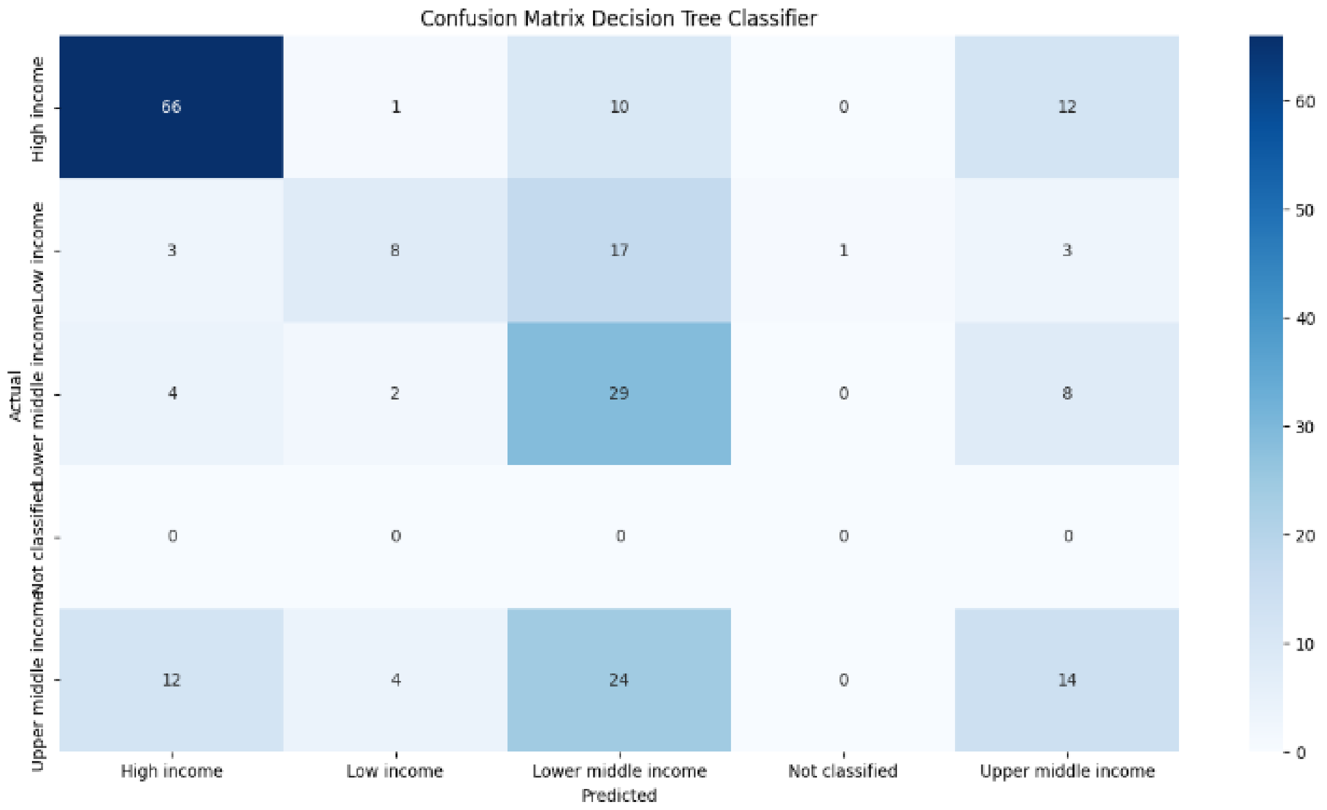
- 80 percent of dataset is used for training and fit the model, and 20 percent dataset is used for testing and evaluate the model.
- Hyperparameter tuning with Grid Search is used for testing a range of multiple hyperparameters for a machine learning model to find the best combination.
- Performance metrics Accuracy is used to measure the proportion of correctly classified instances out of the total instances. It is a simple and effective way to gauge how well a classification model performs.

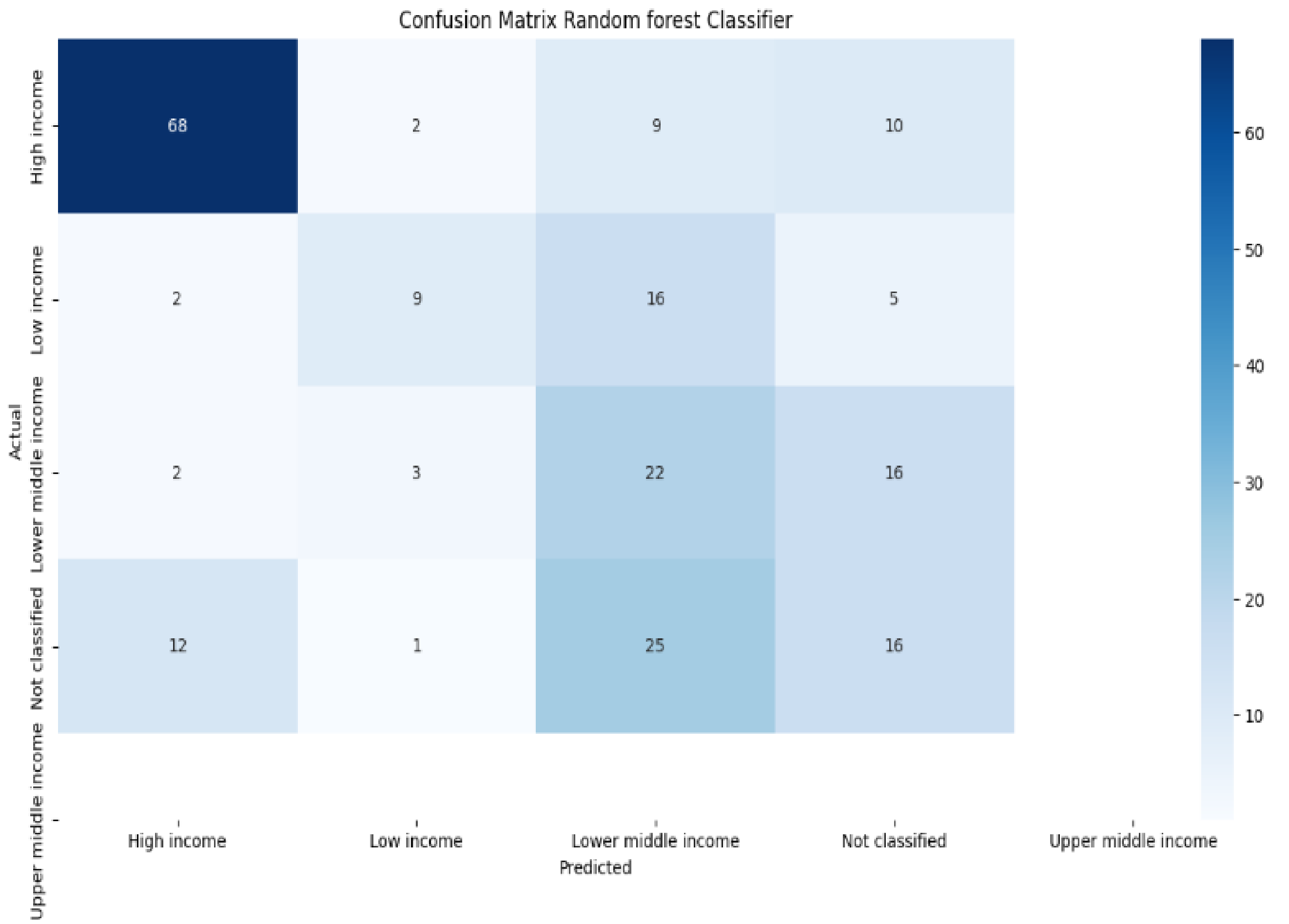
Machine Learning Algorithm	Training Accuracy	Testing Accuracy	Interpretation
Logistic Regression - Grid Search CV	0.37	0.41	Accuracy can be improved for better results
Decision Tree Classifier - Grid Search CV	0.73	0.53	Better Performance and the model might be overfitting
Random Forest Classifier - Grid Search CV	0.68	0.53	Lower Testing Accuracy compared to the Decision Tree
Support Vector Machine	0.52	0.55	Reduced overfitting and Accuracy Increased

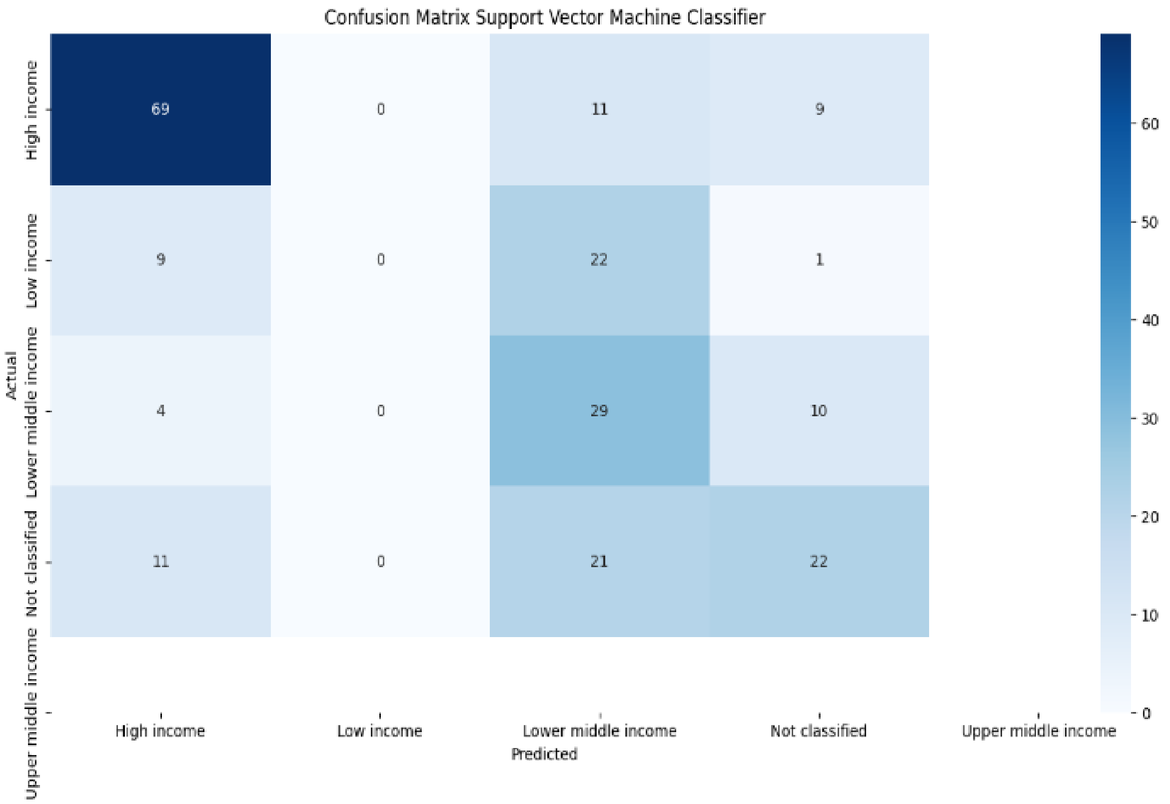
Interpretation

- Logistic Regression: Both training and testing accuracy is very low, consider further hyperparameter tuning, feature engineering, or exploring different algorithms to improve accuracy.
- Decision Tree Classifier: High training score and less testing score Address potential overfitting by pruning the tree, using cross-validation to tune parameters.
- Random Forest Classifier: While it performs better than logistic regression and is less prone to overfitting compared to a single decision tree, further hyperparameter tuning and feature selection might enhance its performance.
- Support Vector Machine (SVM): SVM shows reduced overfitting and improved testing accuracy.




Confusion Matrix Decision Tree Classifier
Actual
High income
Low income
Lower middle income
Not classified
Upper middle income
66
1
10
0
12
3
8
17
1
3
4
2
29
0
8
0
0
0
0
0
12
4
24
0
14
High income
Low income
Lower middle income
Not classified
Upper middle income
Predicted
60
50
40
30
20
10
0


Confusion Matrix Random forest Classifier
Actual
High income
Low income
Lower middle income
Not classified
Upper middle income
68
2
9
10
2
9
16
5
2
3
22
16
12
1
25
16
High income
Low income
Lower middle income
Not classified
Upper middle income
Predicted
60
50
40
30
20
10

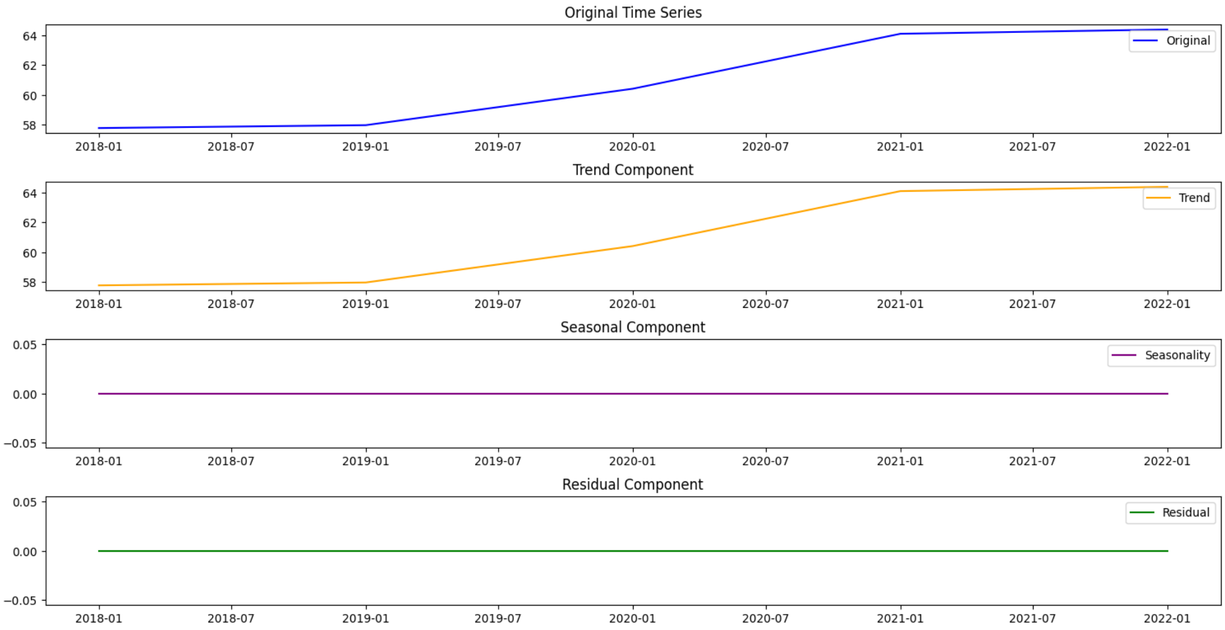


Explanation of the Plot

- Logistic Regression: The diagonal elements are the correctly predicted samples. A total of 90 samples were correctly predicted out of the total 218 samples. Thus, the overall accuracy is 41%.
- Decision Tree Classifier: A total of 117 samples were correctly predicted out of the total 218 samples. Thus, the overall accuracy is 53%.
- Random Forest Classifier: A total of samples 155 were correctly predicted out of the total 218 samples. Thus, the overall accuracy is 53%.
- Support Vector Machine (SVM): A total of 120 samples were correctly predicted out of the total 218 samples. Thus, the overall accuracy is 55%.

5.3.3. Time Series Analysis

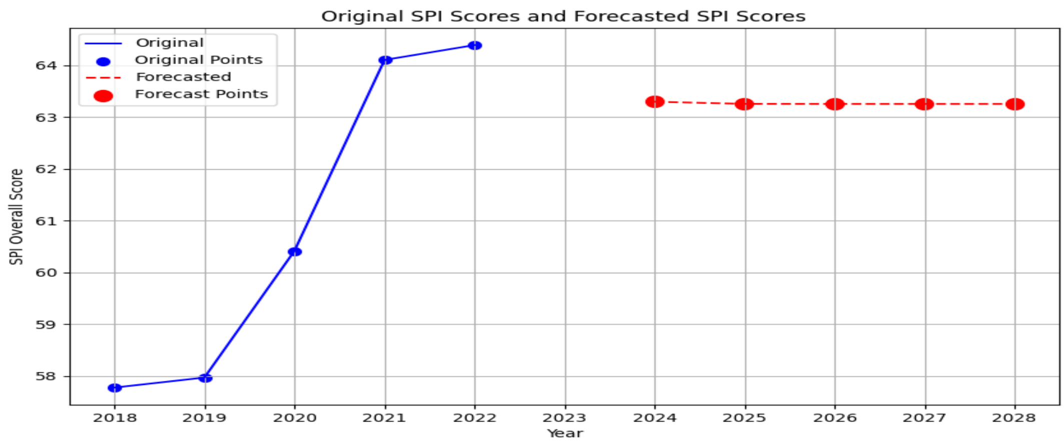
1. **Decomposition of SPI Scores over Years**



Explanation of the Plot

- **Original Time Series:** Decompose the original time series into its constituent components using an additive model.
- **Trend Time Series:** Line plot of the trend component obtained from the additive decomposition of the time series and sets up the plot with specific dimensions.
- **Seasonal Component:** There are no short-term fluctuations caused by factors like seasons or cycles.
- **Residual Component:** There are no random variability that remains after removing the trend and seasonality

2. **SPI Scores Forecasting with ARIMA model**



Explanation of the Plot

- **Blue Line with Dots:** The Actual SPI Scores in the data set used for prediction with ARIMA model
- **Red Line with Dots:** The forecasted SPI Scores with ARIMA model for next 5 years.

Results and conclusion

Overall, this study presents a system for preprocessing, feature extraction with principle component analysis, Explorative data analysis, statistical analysis, Supervised machine learning and Time series analysis on Statistical Performance Indicators dataset for the period of 2018-2022. This study shows that the Statistical Performance Indicators scores is increasing constantly over the years depicting improving development outcomes and track progress toward the Sustainable Development Goals. The mean scores of each pillar are also gradually increasing over years which help to assesses the maturity and performance of national statistical systems of SPI frameworks. The study reveals that there is a immovable tracks of level of incomes over years in different counties based on SPI Scores. Ther is lot of fluctuation in the population when it is framed under normal distribution of SPI Scores.

Statistical analysis ANOVA for the 5 Pillars of Statistical Performance Indicators Data Use, Data Service, Data Product, Data Sources and Data Infrastructure. Infers that there is a greater disparity between the group means relative to the variance within the groups with higher F-statistic and there is a significant difference in mean scores among at least some of the pillars. Spearman's rank correlation coefficient was performed for different levels Income and Population. Spearman's Rank Correlation Coefficient ($\rho = 0.368$) suggests a moderate positive monotonic relationship between the ranked income categories and the population values. In other words, as the rank of the income category increases, there is a tendency for the population values to increase as well, though the relationship is not extremely strong. Since the coefficient is positive, it indicates that higher-ranked income categories are associated with higher population values.

Predictive data analysis computed with Feature variable as 5 pillars of the Statistical Performance Indicators and target variable as SPI over all scores, with hyperparameter GV Grid search. Mean Squared Error of testing and training data of different supervised machine learning algorithms was compared to find the best fit line for future prediction. Random Forest Regressor provides the best performance with the lowest testing Mean Squared Error 7.7, indicating effective generalization and less overfitting compared to other models. A Classification supervised machine learning algorithm was performed with Feature variable as as SPI over all scores and target variable as income groups of Statistical Performance

Indicators. Comparing the accuracy scores of training and testing data with confusion matrix shows Support Vector Machine accuracy 57% which reduced overfitting and improved testing accuracy.

Time Series Analysis with ARIMA model was performed to forecast the SPI scores, decomposition shows the there are no short-term fluctuations caused by factors like seasons or cycles and there are no random variability that remains after removing the trend and seasonality. It would be interesting to see how this analysis would perform a decade down the road