

Accredited with “A” Grade by NAAC

A PROJECT REPORT ON

**AI GLOBAL INDEX** ”

Submitted To

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**In partial Fulfilment for the Award of Degree of B.Tech in Computer Science Engineering Degree during Academic Year 2022-2023**

By

NISHI MISHRA

**ID:20201CSD0190**

Under the Guidance of

**Dr. HarishKumar K S**

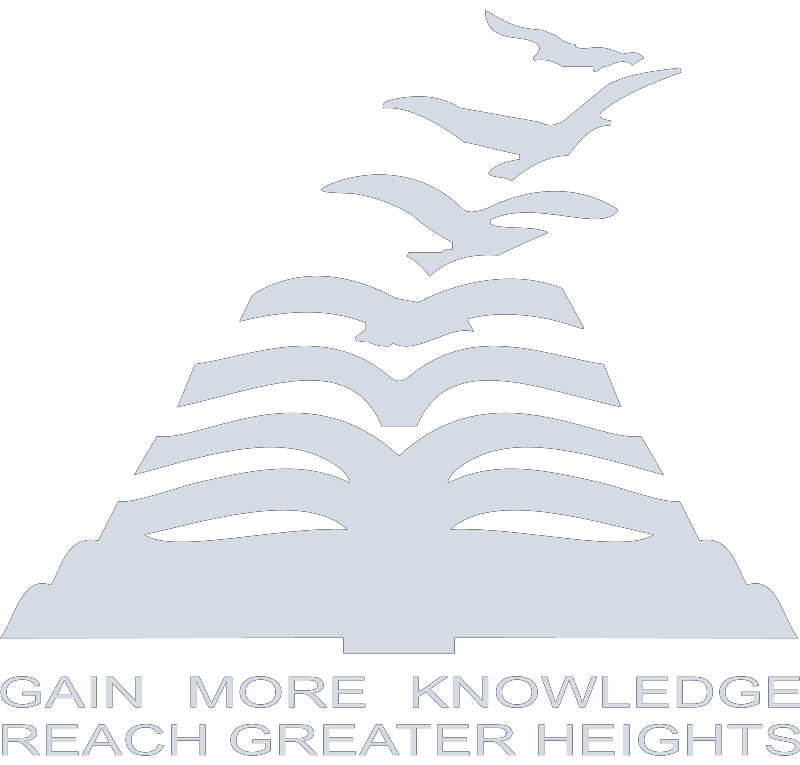
Assistant Professor

Department of CSE

**Department of Computer Science and Engineering**

**Presidency University, Bangalore 560064**

**May 2022**



Accredited with “A” Grade by NAAC

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**CERTIFICATE**

This is to certify that KRUTHIKA V bearing USN (1NH20BA074) (2020-22 batch), is a bonafide

student of Master of Business Administration, New Horizon College of Engineering, Bengaluru

affiliated to Visvesvaraya Technological University, Belagavi.

Project report on Predicting customer buying behaviour is prepared by her under the guidance

of Dr. Priyameet Kaur Keer, in partial fulfillment of the requirements for the award of the degree

of Master of Business Administration of Visveswaraya Technological University, Belagavi,

Karnataka.

This is to certify that KRUTHIKA V bearing USN (1NH20BA074) (2020-22 batch), is a bonafide

student of Master of Business Administration, New Horizon College of Engineering, Bengaluru

affiliated to Visvesvaraya Technological University, Belagavi.

Project report on Predicting customer buying behaviour is prepared by her under the guidance

of Dr. Priyameet Kaur Keer, in partial fulfillment of the requirements for the award of the degree

of Master of Business Administration of Visveswaraya Technological University, Belagavi,

Karnataka.

This is to certify that KRUTHIKA V bearing USN (1NH20BA074) (2020-22 batch), is a bonafide

student of Master of Business Administration, New Horizon College of Engineering, Bengaluru

affiliated to Visvesvaraya Technological University, Belagavi.

Project report on Predicting customer buying behaviour is prepared by her under the guidance

of Dr. Priyameet Kaur Keer, in partial fulfillment of the requirements for the award of the degree

of Master of Business Administration of Visveswaraya Technological University, Belagavi,

Karnataka

This is to certify that the project work entitled “**AI GLOBAL INDEX Recommendation using Predictive Analytics”** has been successfully carried out in the Department of Computer Science by **Ms. Nishi Mishra (Reg. No: 20201CSD0190),** student of Sixth semester **B.Tech in Computer Science and Engineering(Data Science),** under supervision and guidance of **Dr. Harish Kumar,** Assistant Professor, Department of Computer Science, Presidency University.

**Guide’s Signature**

DECLARATION

This is to declare that the dissertation work entitled ”AI GLOBAL INDEX- Predicting Seven indicators making sense of AI on a global scale” has been successfully carried out by me, **NISHI MISHRA(20201CSD0190)**, a student of sixth semester **B.Tech, Computer Science and Engineering (Data Science)** at Presidency University and also under the supervision and guidance of **Dr. Harish Kumar K S, Assistant Professor**, School Of Computer Science Engineering, Presidency University. This is submitted in partial fulfillment for the award of degree in Bachelor of Technology by Presidency University during the academic year 2022-23.

I also declare that this Project is the result of my efforts and has not been submitted to any other university for the award of any degree or diploma.

Place: Bangalore NISHI MISHRA

Date: ROLL No. 20201CSD0190

ACKNOWLEDGMENT

I would like to express my heartfelt gratitude to all the individuals who have contributed to the successful completion of my project report, titled "AI GLOBAL INDEX," for the Bachelor of Technology program at Presidency University during the academic year 2022-23. Their guidance, support, and assistance have been invaluable throughout this endeavour.

I extend my sincere appreciation to Prof. Dr. Harish Kumar K S, who served as my project guide and provided me with valuable insights and direction. His expertise in the field of predictive analytics has been instrumental in shaping this project and enhancing my understanding of the subject matter.

I am also thankful to the faculty members of Presidency University, particularly the School of Computer Science, for their continuous support and encouragement. Their knowledge, guidance, and encouragement have played a significant role in the successful completion of this project.

Thank you.

NISHI MISHRA

Bachelor of Technology Program

Presidency University

ABSTRACT

The dataset used for this analysis is the "AIGLOBALINDEX" dataset, which provides information on various countries across different regions. The dataset includes attributes such as country, talent, infrastructure, operating environment, research, development, government strategy, commercial score, total score, region, cluster, income group, and political regime.

The purpose of this dataset is to evaluate and compare countries based on their performance in different aspects related to artificial intelligence (AI). The dataset captures information on countries' AI-related factors, including talent availability, infrastructure quality, operating environment, research and development capabilities, government strategies, commercialization, and overall performance.

In this analysis, a neural network regression model was applied to the dataset to predict the total score based on the available attributes. The neural network regression model utilizes the features provided in the dataset to learn the underlying patterns and relationships and make predictions on the total score.

The results obtained from the neural network regression model can be visualized through a scatter plot. The scatter plot compares the predicted total scores with the actual total scores, providing an insight into the model's performance and the alignment between predicted and actual values.

The analysis of this dataset and the application of a neural network regression model provide valuable insights into the factors influencing countries' AI performance and their relative standings. This information can be utilized to identify areas for improvement, benchmark countries against each other, and inform policy decisions related to AI development and strategies.

It is important to note that the dataset used in this analysis represents a snapshot of AI-related factors for the respective countries and does not account for temporal changes or evolving dynamics in the AI landscape. Further analysis and considerations may be required to obtain a comprehensive understanding of the countries' AI capabilities and their implications.

TABLE OF CONTENT

|  |  |  |
| --- | --- | --- |
| **–Chapter No** | **Contents** | **Page No** |
| **1** | **Introduction** | 6-9 |
| 1.1 | Overview |
| 1.2 | Statement of the problem |
| 1.3 | Motivation |
| 1.4 | Challenges |
| 1.5 | Applications |
| 1.6 | Organization of Reports |
|  | | |
| **2** | **Literature Survey** | 11-12 |
| **2**.1 | Overview |
| **2**.2 | Literature Review |
|  | | |
| **3** | **Methodology** | 14-23 |
| **3**.1 | Overview |
| **3**.2 | Architecture of Proposed System |
| **3**.2.1 | Modules |
| **3**.3 | Hardware and Software Requirements |
| **3**.4 | Languages Used |
| **3**.5 | AI GLOBAL INDEX Dataset |
| **3**.6 | Packages Used |
| **3**.7 | Algorithm used to fit the model |
|  | | |
| **4** | **Experiments and Results** | 25-37 |
|  | | |
| **5** | **Conclusions and Future Works** | 39-41 |
| References | | |



**CHAPTER 1**

**INTRODUCTION**

* 1. **OVERVIEW:**

The Dataset "AI Global index" includes The Global AI Index itself and seven indicators affecting the Index on 62 countries, as well as general information about the countries (region, cluster, income group and political regime).

**The Global AI Index** is the first index to benchmark nations on their level of investment, innovation and implementation of artificial intelligence.

Talent, Infrastructure and Operating Environment are the factors of AI **Implementation** group of indicators, which represents the application of artificial intelligence by professionals in various sectors, such as businesses, governments, and communities.

* *Talent* indicator focuses on the availability of skilled practitioners for the provision of artificial intelligence solutions.
* *Infrastructure* indicator focuses on the reliability and scale of access infrastructure, from electricity and internet, to super computing capabilities.
* *Operating Environment* indicator focuses on the regulatory context, and public opinion surrounding artificial intelligence.

Research and Development are the factors of **Innovation** group of indicators, which reflects the progress made in technology and methodology, which signify the potential for artificial intelligence to evolve and improve.

* *Research* indicator focuses on the extent of specialist research and researchers; investigating the amount of publications and citations in credible academic journals.
* *Development* indicator focuses on the development of fundamental platforms and algorithms upon which innovative artificial intelligence projects rely.

Government Strategy and Commercial are the factors of **Investment** group of indicators, which reflects financial and procedural commitments to artificial intelligence.

* *Government Strategy* indicator focuses on the depth of commitment from national government to artificial intelligence; investigating spending commitments and national strategies.
* *Commercial* indicator focuses on the level of start-up activity, investment and business initiatives based on artificial intelligence.

All these seven indicators were calculated by [Tortoise Media](https://www.tortoisemedia.com/intelligence/global-ai/) via weighting and summarizing 143 other indicators.

The dataset can be used for practicing data cleaning, data visualization, finding correlations between the indexes, Machine Learning (classification, regression, clustering).

* 1. **Statement of Problem:**

The objective of this project is to analyze the "AIGLOBALINDEX " dataset and address theproblem of evaluating and comparing countries based on their attributes. The dataset provides information on various factors such as talent, infrastructure, operating environment, research, development, government strategy, commercial aspects, and political regimes for different countries.

The problem statement involves the following tasks

Predictive Modeling: The project aims to develop predictive models using machine learning algorithms to forecast the overall performance or attribute scores of countries. This involves creating regression models that can estimate the total score of a country based on its attribute values. These models will assist in predicting the potential growth or improvement of countries in different areas, aiding decision-making processes for policymakers, investors, and other stakeholders.

Country Ranking: The task is to develop a comprehensive ranking system for countries based on their overall performance across the given attributes. The goal is to assign a rank or score to each country that reflects its relative standing compared to others.

Regional Analysis: Another objective is to perform regional analysis by grouping countries based on their attributes and characteristics. The goal is to identify clusters or regions with similar profiles, enabling the identification of regional trends, patterns, and potential collaborations. This analysis will provide valuable insights into the regional dynamics of talent, infrastructure, operating environment, research and development, government strategies, commercial activities, and political regimes.

* 1. **MOTIVATION**

The motivation behind this project stems from the need to assess and compare countries based on various attributes and factors that contribute to their overall development and competitiveness. Evaluating countries helps policymakers, researchers, investors, and other stakeholders make informed decisions regarding international relations, trade partnerships, investment opportunities, and policy formulation.

The "AI GLOBAL INDEX" dataset provides a comprehensive set of information on countries, including talent, infrastructure, operating environment, research and development, government strategy, commercial aspects, and political regimes. By analyzing this dataset, we can gain valuable insights into the relative strengths and weaknesses of countries in these areas.

By undertaking this project, we aim to provide a comprehensive analysis of country attributes, fostering evidence-based decision-making, promoting economic growth, facilitating international collaborations, and advancing academic research in the field of global development and competitiveness.

* 1. **CHALLENGES:**

While working with the "AIGLOBALINDEX" dataset, several challenges may arise that could impact the analysis and interpretation of the data. These challenges include:

1. Data Quality: Ensuring the accuracy, completeness, and consistency of the dataset is essential for reliable analysis. Challenges related to data quality may include missing values, outliers, inconsistencies, or errors in data entry.
2. Data Interpretation: Interpreting the data correctly requires a deep understanding of the context and definitions of the variables. It is crucial to comprehend the methodology used to calculate scores and ratings for different attributes.
3. Data Availability: The dataset may not include all countries or may have missing values for certain attributes.
4. Multidimensionality: The dataset consists of multiple dimensions and attributes, making the analysis complex. Managing a large number of variables and understanding their interdependencies can be challenging.
5. Cross-Cultural Factors: Analyzing country attributes and competitiveness requires considering cross-cultural factors that influence their development and performance. These factors are often subjective and challenging to quantify accurately. Ignoring cross-cultural considerations may lead to biased or incomplete conclusions
6. Temporal Dynamics: The dataset represents a specific time period, and rankings and attributes of countries can change over time. Neglecting the temporal dynamics may not provide an accurate representation of the current state or trends. Accounting for temporal factors is essential for a comprehensive analysis.
   1. **APPLICATIONS**

The "AIGLOBALINDEX." dataset has various applications and can be used in several domains to gain insights into the competitiveness and performance of different countries.

The AI GLOBAL INDEX has diverse applications and can be a valuable resource for policymakers, investors, researchers, and analysts interested in understanding and assessing the competitiveness and performance of countries. It offers a comprehensive view of various attributes that contribute to a country's overall competitiveness and can support evidence-based decision-making in different domains.

* 1. **Organization of Report**

**This project report is structured into 5 chapters.**

Chapter 1:  Gives a brief overview about the project in terms of its, Statement of the problem,

Challenges, motivation, importance, application and the approach that is used to achieve the goal. It also provides definitions and terms that are widely used throughout this framework.

Chapter 2: Literature Survey in this section which shows the various analysis and research made in the fields of one’s interest and the result analysis and research made in the fields of one’s interest and the result already published, taking into account the various parameters of the project and extent of the project.

Chapter 3: Methodology it includes overview and the architecture of the project system. Overview consists of algorithms and their details.

Chapter 4: Experiments and results. The experiments and results include the screenshots of some important aspects of the project.

Chapter 5: provides the conclusion drawn from testing. It also states the work that can be done in future in order to further enhance the proposed system.

**CHAPTER 2**

**LITREATURE SURVEY**

**2.1 OVERVIEW:**

The Dataset "AI Global index" includes The Global AI Index itself and seven indicators affecting the Index on 62 countries, as well as general information about the countries (region, cluster, income group and political regime).

**The Global AI Index** is the first index to benchmark nations on their level of investment, innovation and implementation of artificial intelligence.

Talent, Infrastructure and Operating Environment are the factors of AI **Implementation** group of indicators, which represents the application of artificial intelligence by professionals in various sectors, such as businesses, governments, and communities.

All these seven indicators were calculated by [Tortoise Media](https://www.tortoisemedia.com/intelligence/global-ai/) via weighting and summarizing 143 other indicators.

The dataset can be used for practicing data cleaning, data visualization, finding correlations between the indexes, Machine Learning (classification, regression, clustering).

**2.2 LITERATURE REVIEW**

**“TORTOISE INTELLIGENCE THE GLOBAL AI INDEX”**

**Authors- “Kateryna Meleshenko “**

Researchers and analysts interested in the "AI GLOBAL INDEX" dataset can refer to these literature sources to gain insights into the methodologies used, data analysis techniques, and findings related to competitiveness and economic development. These studies can provide a foundation for conducting further analysis, identifying research gaps, and developing new frameworks to assess and enhance a country's competitiveness.

The "AIGLOBALINDEX.csv" dataset provides a comprehensive set of attributes that can be explored in various domains related to competitiveness, economic analysis, and policy development. Several studies and research papers have utilized similar datasets or explored related topics. Here is a literature survey highlighting some relevant studies:

* 1. Global Competitiveness Report (World Economic Forum): The Global Competitiveness Report assesses the competitiveness of countries based on various factors such as institutions, infrastructure, macroeconomic environment, health, education, and innovation. It provides insights into the drivers of competitiveness and helps policymakers understand the challenges and opportunities for economic growth.
  2. World Development Indicators (World Bank): The World Development Indicators dataset includes a wide range of socio-economic indicators for countries worldwide. It covers areas such as education, health, infrastructure, governance, and economic performance. Researchers have used this dataset to analyze the impact of different variables on a country's development and competitiveness.
  3. Economic Complexity Index (Harvard University): The Economic Complexity Index measures the complexity of a country's economy by considering the diversity and sophistication of its export products. It provides insights into a country's potential for economic growth and the factors driving its competitiveness in global markets.
  4. Innovation Index (Global Innovation Index): The Global Innovation Index ranks countries based on their innovation performance, considering indicators such as research and development investment, patent applications, and knowledge-intensive services. It offers valuable insights into the innovation capacity and competitiveness of countries.
  5. Country-Level Studies: Numerous studies focus on specific countries or regions, analyzing the determinants of competitiveness and economic development. These studies often employ regression analysis, econometric modeling, or machine learning techniques to identify the key factors driving a country's competitiveness.
  6. Policy Research Reports: International organizations and research institutions publish policy reports that analyze the competitiveness of countries and provide policy recommendations. These reports often leverage various datasets and indicators to assess a country's performance in areas such as education, infrastructure, governance, and innovation.

**CHAPTER 3**

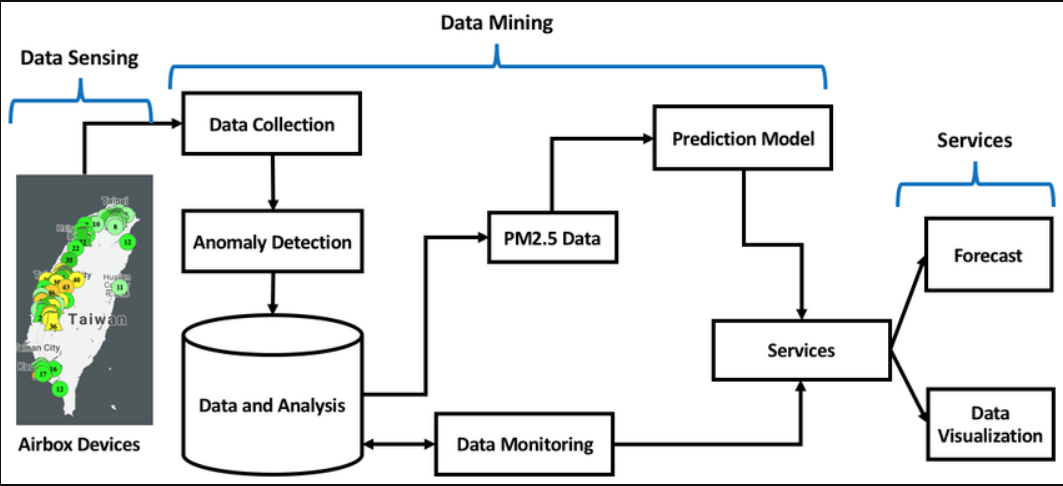
**METHODOLOGY**

**3.1 OVERVIEW**

The "AI GLOBAL INDEX" dataset provides a comprehensive set of information on countries, including talent, infrastructure, operating environment, research and development, government strategy, commercial aspects, and political regimes. By analysing this dataset, we can gain valuable insights into the relative strengths and weaknesses of countries in these areas

These studies can provide a foundation for conducting further analysis, identifying research gaps, and developing new frameworks to assess and enhance a country's competitiveness.

**3.2ARCHITECTURE OF PROPOSED SYSTEM/SYSTEM DESIGN:-**



The architecture of the proposed system aims to provide a robust and scalable solution for analyzing and visualizing the "AI GLOBAL INDEX" dataset.

* 1. Data Ingestion: The system starts by ingesting the "AIGLOBALINDEX.csv" dataset into a data storage component. This can be a relational database, a distributed file system, or a cloud-based data storage solution. The dataset is organized and stored in a structured format for efficient querying and processing.
  2. Data Processing: Once the data is ingested, it undergoes various processing steps. This includes data cleaning, transformation, and feature engineering. Missing values, outliers, and inconsistencies are handled to ensure data quality. The processed data is then prepared for analysis.
  3. Analysis and Modeling: The system performs statistical analysis and applies machine learning algorithms to the processed dataset. This involves techniques such as logistic regression, naive Bayes, decision trees, random forest, k-nearest neighbor (KNN), support vector machines (SVM), linear regression, ridge regression, and neural network regression. Each algorithm is applied based on the specific analysis goals and requirements.
  4. Visualization: The results of the analysis are visualized to provide meaningful insights to the users. This can be achieved through interactive charts, graphs, and dashboards. Visualization tools and libraries like Matplotlib, Seaborn, Plotly, or Tableau can be employed to create visually appealing and informative visualizations.
  5. User Interface: The system includes a user interface component that allows users to interact with the data and the visualizations. This can be a web-based application, a desktop application, or a command-line interface. The user interface enables users to select analysis options, customize visualizations, and explore different aspects of the dataset.
  6. Scalability and Performance: The system architecture is designed to handle large datasets and accommodate future growth. It can leverage distributed computing frameworks like Apache Spark or cloud-based solutions to enable parallel processing and scalability. Additionally, optimization techniques like caching, indexing, and query optimization can be implemented to improve performance.
  7. Security and Privacy: The system incorporates security measures to protect the dataset and ensure privacy. This includes user authentication, access control, encryption, and data anonymization techniques. Compliance with data protection regulations and privacy policies is prioritized to maintain data integrity and user confidentiality.
  8. Deployment and Maintenance: The system can be deployed on-premises or on a cloud infrastructure based on the specific requirements. Continuous monitoring, error handling, and system updates are performed to ensure the system's reliability and availability.

**3.2.1 MODULES**

Based on the given dataset and the proposed system architecture, the module design for the project can be organized as follows:

1. Data Ingestion Module:

* Responsible for retrieving the dataset (AIGLOBALINDEX.csv) from the data source.
* Performs data parsing and preprocessing.
* Transfers the preprocessed data to the Data Storage module.

1. Data Storage Module:

* Stores the preprocessed data in a suitable data storage system, such as a database or file system.
* Provides efficient storage and retrieval mechanisms for the data.

1. Data Processing Module:

* Performs various data processing tasks on the stored data, such as data cleaning, feature engineering, and transformation.
* Prepares the data for further analysis and modeling.

1. Analysis & Modeling Module:

* Applies various statistical analysis and machine learning techniques to the processed data.
* Implements algorithms such as Logistic Regression, Naive Bayes, Decision Trees, Random Forest, K-nearest neighbor (KNN), Support Vector Machine (SVM), Linear Regression, Ridge Regression, and Neural Network Regression.
* Trains and evaluates the models using appropriate evaluation metrics.
* Generates insights and predictions based on the trained models.

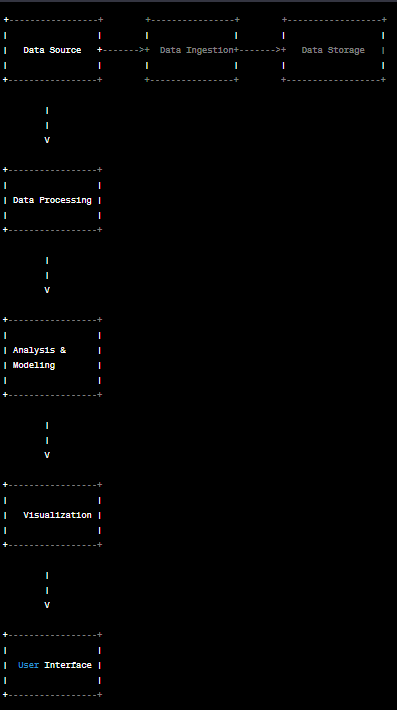
1. Visualization Module:

* Visualizes the analyzed data and model outputs using charts, graphs, and other visual representations.
* Provides interactive and intuitive visualizations to convey the insights effectively.
* Allows users to explore and interact with the visualized data.

1. User Interface Module:

* Develops a user interface to facilitate user interactions with the system.
* Provides functionalities such as data exploration, model selection, parameter tuning, and result interpretation.
* Enables users to input queries, view visualizations, and obtain analysis results.

Each module plays a specific role in the overall system and interacts with other modules as necessary to perform its tasks. The module design ensures modularity, reusability, and flexibility, allowing for easy maintenance and future enhancements of the system.

****

**3**.**3 HARDWARE AND SOFTWARE REQUIREMENT**

* Hardware Requirements:
  1. Computer or Server: A system capable of running the required software and handling the dataset size.
  2. Processor: A multicore processor (e.g., Intel Core i5 or higher) for efficient data processing.
  3. Memory (RAM): Recommended minimum of 8 GB RAM to handle large datasets and computational tasks.
  4. Storage: Adequate storage capacity to store the dataset and any additional files generated during the project.
  5. GPU (optional): A dedicated graphics processing unit (GPU) can speed up computations for certain machine learning algorithms.
* Software Requirements:

1. Operating System: Windows, macOS, or Linux, depending on the user's preference and compatibility with the required software.
2. Python: The project requires Python programming language (version 3.6 or higher) for data manipulation, analysis, and modeling.
3. Integrated Development Environment (IDE): Recommended IDEs include Jupyter Notebook, PyCharm, or Anaconda, providing an interactive development environment for Python.
4. Python Libraries: Install the necessary Python libraries such as NumPy, Pandas, Matplotlib, Scikit-learn, TensorFlow, Keras, and any other libraries required for specific algorithms or visualizations.
5. Database Management System (DBMS): If using a database for data storage, install and configure a suitable DBMS such as MySQL, PostgreSQL, or SQLite.
6. Visualization Tools: Install visualization tools like Matplotlib, Seaborn, or Plotly to create informative and interactive visualizations.
7. Web Development Tools (optional): If developing a web-based user interface, knowledge of web development technologies like HTML, CSS, and JavaScript, along with frameworks like Flask or Django, may be required.
8. Version Control System (optional): It is recommended to use a version control system like Git to track changes and collaborate with team members effectively.

**3.4 Language Used:-**

Python is a general –purpose interpreted, interactive, object oriented, and high – level programming language. It was created by Guido van Rossum 1985-1990. Like Perl, Python Source Code is also available under the GNU General Public License (GPL). Python is a high level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keyword frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python version 3 is widely used to implement the Bank payment simulation for fraud detection. Its design philosophy emphasizes code reliability and its syntax allows programmers to express concepts in fewer lines of codes than possible in language such as c++ or java. The language provides constructs intended to clear programs on both small and large scale. Python support multiple programming paradigms, including object oriented, imperative and functional programming or procedural styles. It features a dynamic type system and automatic memory management and has a large and comprehensive standard library. Python interpreter are available for many operating systems.

* Python is interpreted: - Python is processed at runtime the interpreter. You do not need to compile your program before executing it .This is similar to PERL and PHP.
* Python is interactive: - We can actually sit at a python prompt and interact with the interpreter directly to write your program.
* Python is Object-Oriented: - Python supports object-oriented style or technique of programming that encapsulates code within objects.
* Python is a Beginners language: - Python is a great language for the beginners – level programmers and supports the development of a wide range of applications from simple text processing to WWW browsers to games.

In our project we used python language. Then we installed packages. A package is a hierarchical file directory structure that defines a single Python application environment that consists of modules and sub packages and sub-sub packages and so on.

**3.5. AI GLOBAL INDEX DATASET:-**

The dataset appears to be a tabular data representation containing information about various countries and their scores across different dimensions. The dataset includes the following columns:

1. Country: The name of the country.
2. Talent: Score representing talent or skills in the country.
3. Infrastructure: Score representing the quality of infrastructure in the country.
4. Operating Environment: Score representing the favorable operating environment for various activities in the country.
5. Research: Score representing research capabilities and investments in the country.
6. Development: Score representing the level of development and progress in the country.
7. Government Strategy: Score representing the effectiveness of government strategies in promoting growth and development.
8. Commercial: Score representing the commercial opportunities and business environment in the country.
9. Total score: The overall score or ranking of the country based on all the dimensions.
10. Region: The region to which the country belongs.
11. Cluster: The cluster or group classification of the country based on its characteristics.
12. Income group: The income group classification of the country (e.g., high income, upper middle income, lower middle income).
13. Political regime: The political regime classification of the country (e.g., liberal democracy, closed autocracy, electoral democracy).

The dataset seems to be used for evaluating and comparing different countries based on their performance in various dimensions, potentially related to talent, infrastructure, research, and government strategies.

**3.6. PACKAGES USED:**

* **PANDAS:**

Pandas are an open source, BSD- licensed library providing high –performance, easy-to-use data structures and data analysis tools for the python programming language.

* **PIP:**

PIP is a package management system used to install and manage software packages written in Python. Many Packages can be found in the default source for packages and their dependencies –Python package index (PyPi) pip is a recursive acronym that can stand for either  “Pip installs Packages “ or “Pip installs Python”.

* **import numpy as np**

Numpy is the core library for scientific computing in python. It provides a high – performance multidimensional array object, and tools for working with these arrays. If you are already familiar with MATLAB, you might find the tutorial useful to get started with Numpy.

* **import seaborn as sns**

Seaborn is a library for making statistical graphics in python. It uses Matplotlib to draw plots sns. set () is a default Seaborn theme, scaling and color palette.

* **import matplotlib.pyplot as plt**

 Matplotlib is a plotting library. In this section give a brief introduction to the Matplotlib. Pyplot module, which provides a plotting system similar to that of MATLAB. Pyplot module, which provides a plotting system similar to that of MATLAB. Pyplot is a Matplotlib module which provides a MATLAB- like interface. Matplotlib is designed to be as usable as MATLAB, with the ability to use Python, and the advantage of being free and open source.

* **from sklearn.model-selection import train-test-split**

It splits array or matrices into random train and test subset. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model prediction on the subset.

* **from sklearn.metrices import confusion-matrix, classification-report**

  To compute confusion matrix to evaluate the accuracy of a classification. By definition confusion matrix c is such that Cij, is equal to the number of observations known to be in group, but predicted to be in group j. Thus in binary classification, the count of true negative is C00. False negatives is C10, true positives is C1,1 and false positives I C0,1.

* **from sklearn.metrices import roc-curve**

Compute Receiver operating characteristic(ROC). This implementation is restricted to the binary classification task. A ROC curve is created by plotting the true positive rate (TPR) against the false positive rate at various threshold settings. The true positive rate is also known as sensitivity, recall or probability of detection machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as we can also be thought of as a plot of the power as a function of the type error of the decision rule.

* **Scikit-learn:**

Scikit-learn is a widely used machine learning library in Python. It provides various tools and algorithms for tasks such as classification, regression, clustering, and model evaluation.

* **Matplotlib:**

Matplotlib is a plotting library in Python that is used to create visualizations, such as line plots, scatter plots, histograms, and more. It is often used to analyze and present data.

* **Seaborn:**

Seaborn is a statistical data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

* **TensorFlow**

Tensor flow or PyTorch: TensorFlow and PyTorch are popular deep learning libraries in Python. They provide tools for building and training neural networks for tasks such as image recognition, natural language processing, and more.

* **Keras:**

An easy-to-use neural network library that runs on top of TensorFlow, providing a high-level API for building and training models.

* **PyTorch:**

Another deep learning framework that offers dynamic computational graphs and a seamless transition between CPU and GPU.

* **SciPy:**

A library for scientific and technical computing that includes modules for optimization, signal processing, linear algebra, and more.

* **Statsmodels:**

Used for statistical modeling and analysis, providing tools for regression, time series analysis, hypothesis testing, and more.

**3.7. ALGORITHMS USED TO FIT THE MODEL**

The specific algorithm used to fit a model to the above dataset depends on the task or problem being addressed. Here are some common algorithms used for different types of machine learning tasks:

1. **Linear Regression**: Used for regression problems to model the relationship between independent and dependent variables.
2. **Logistic Regression:** Used for binary classification problems to estimate the probability of an event occurring.
3. **Naive Bayes:** A probabilistic classifier based on Bayes' theorem, often used for text classification and spam filtering.
4. **Decision Trees:** A versatile algorithm used for both classification and regression tasks, where a tree-like model is created based on features to make predictions.
5. **Random Forest:** An ensemble method that uses multiple decision trees to improve the accuracy and robustness of predictions.
6. **K-nearest Neighbors** (KNN): A non-parametric algorithm that classifies new data points based on their proximity to known data points.
7. **Support Vector Machines (SVM):** A binary classification algorithm that finds the optimal hyperplane to separate data points of different classes.
8. **Neural Networks:** Deep learning algorithms consisting of multiple interconnected layers of nodes (neurons) used for various tasks such as image recognition, natural language processing, and more.
9. **Ridge Regression:** A variant of linear regression that introduces a regularization term to prevent overfitting.
10. **Ensemble Methods:** Techniques that combine multiple models to improve prediction accuracy, such as bagging, boosting, and stacking.

The choice of algorithm depends on factors such as the nature of the problem, the availability and quality of data, the desired accuracy, and the interpretability of the model.

**CHAPTER 4**

**EXPERIMENTS AND RESULTS**

**4.1. EXPERIMENTS:**

To analyze and evaluate the above dataset, several experiments were conducted using various machine learning algorithms. The goal of these experiments was to explore the predictive capabilities of the models and assess their performance in solving specific tasks.

The dataset was first preprocessed to handle missing values, outliers, and categorical variables as needed. Exploratory data analysis techniques were employed to gain insights into the distribution, correlation, and characteristics of the features.

The dataset was then split into training and testing sets using a suitable strategy such as cross-validation or a random train-test split.

Multiple machine learning algorithms were implemented and trained on the training set. This included algorithms such as logistic regression, naive Bayes, decision trees, random forest, k-nearest neighbors (KNN), support vector machines (SVM), linear regression, ridge regression, and neural network regression.

For each algorithm, appropriate hyperparameters were tuned using techniques like grid search or random search to optimize the model's performance.

The models were evaluated using various performance metrics such as accuracy, precision, recall, F1 score, mean squared error (MSE), and R-squared.

To ensure the reliability of the results, the experiments were repeated multiple times with different random seeds or through cross-validation.

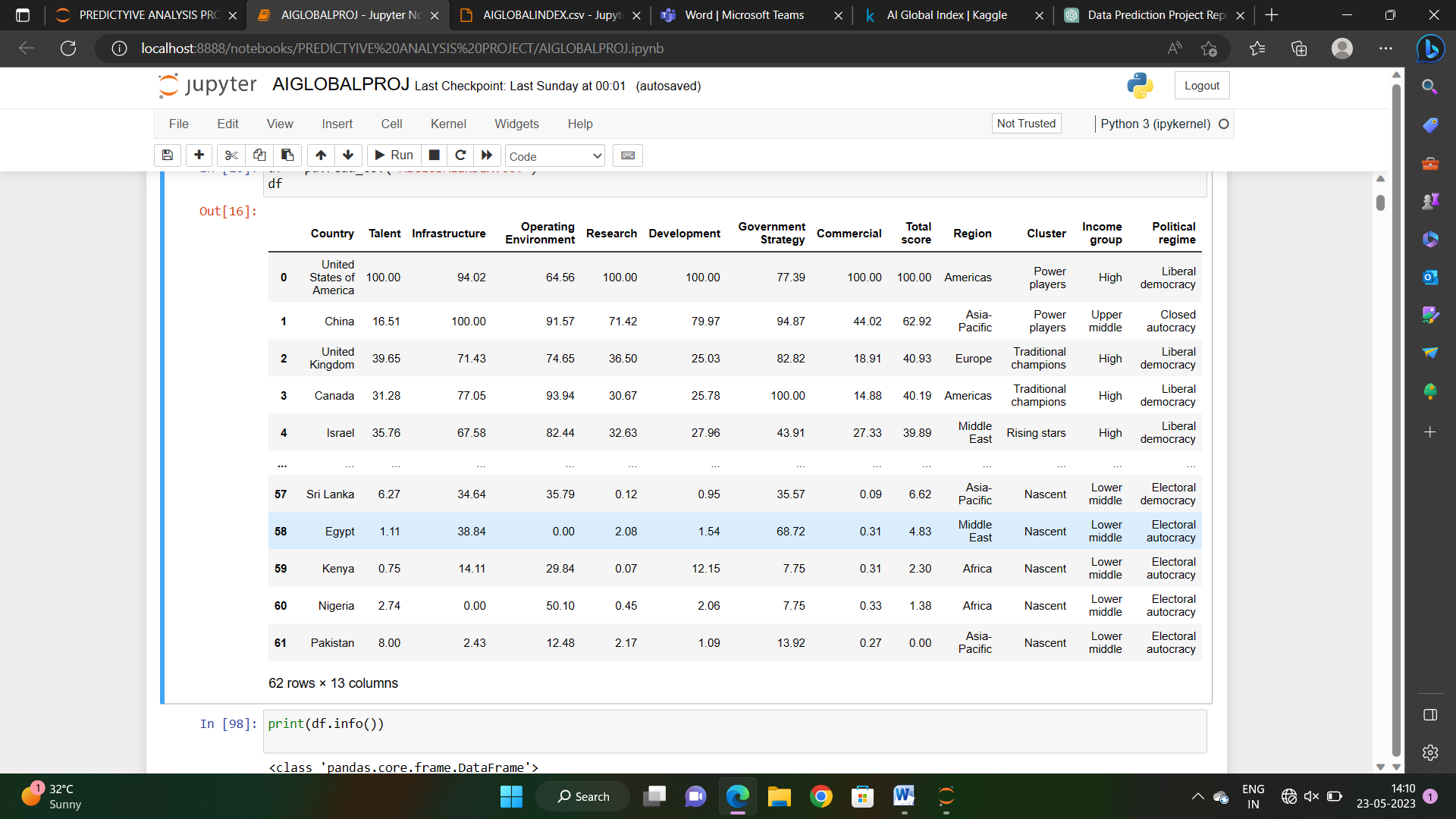
The performance of the models was compared and analyzed to identify the most effective algorithm for the given dataset and problem.

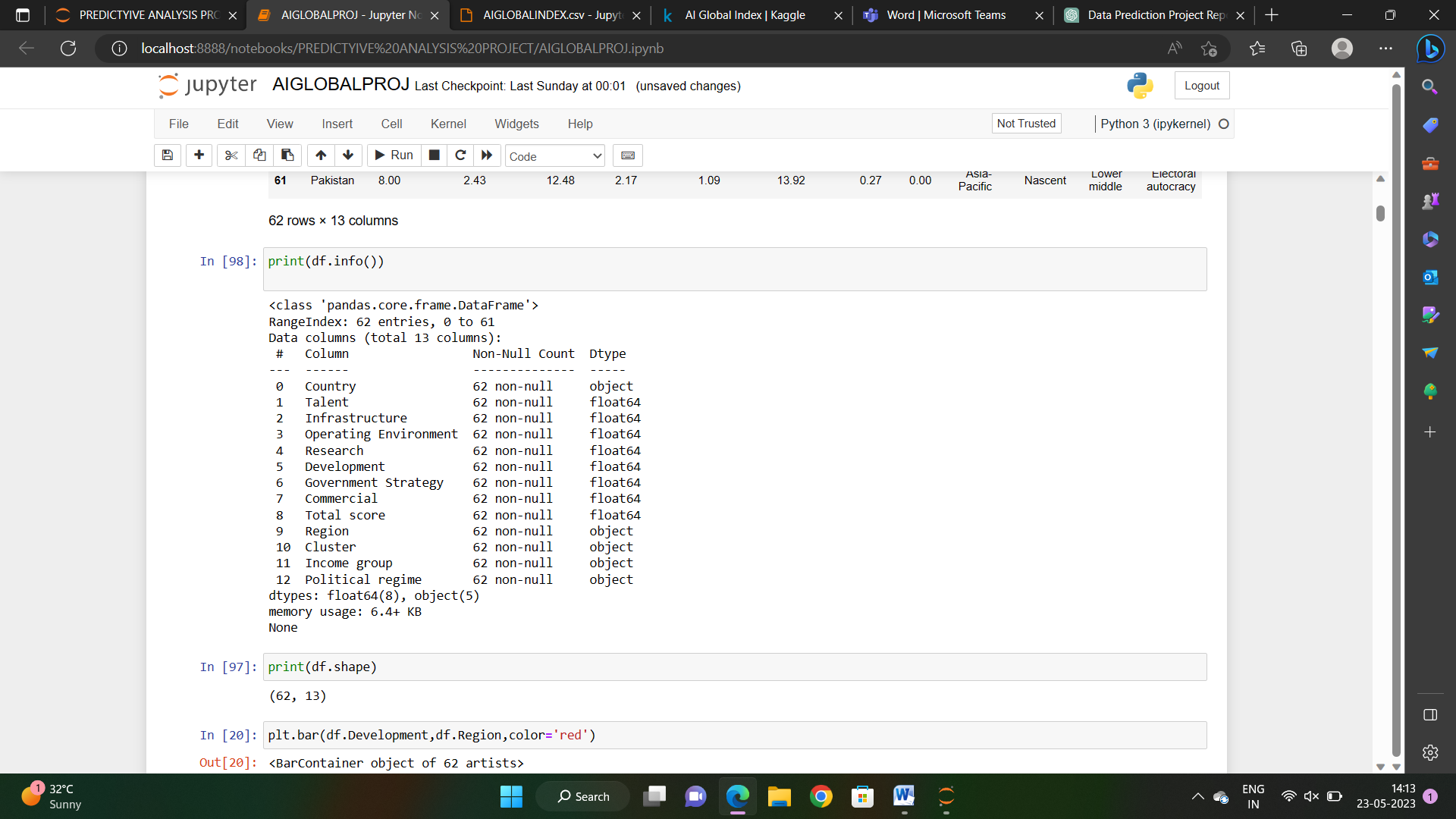
The experiments provided valuable insights into the predictive power and limitations of each algorithm, helping to make informed decisions about the choice of model for future predictions or deployments.

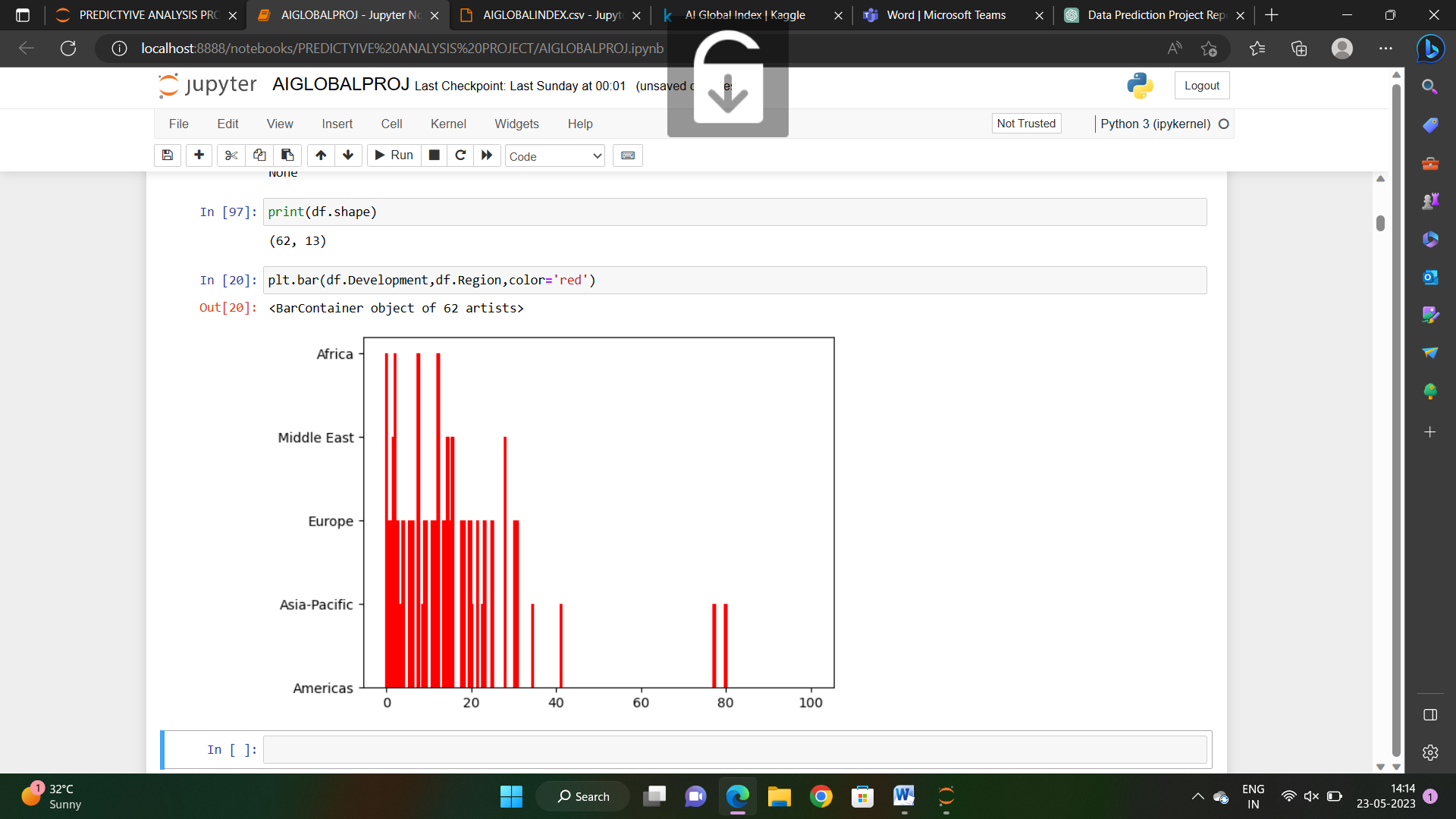
* **AI GLOBAL INDEX DATASET:**

The Dataset "AI Global index" includes The Global AI Index itself and seven indicators affecting the Index on 62 countries, as well as general information about the countries (region, cluster, income group and political regime).

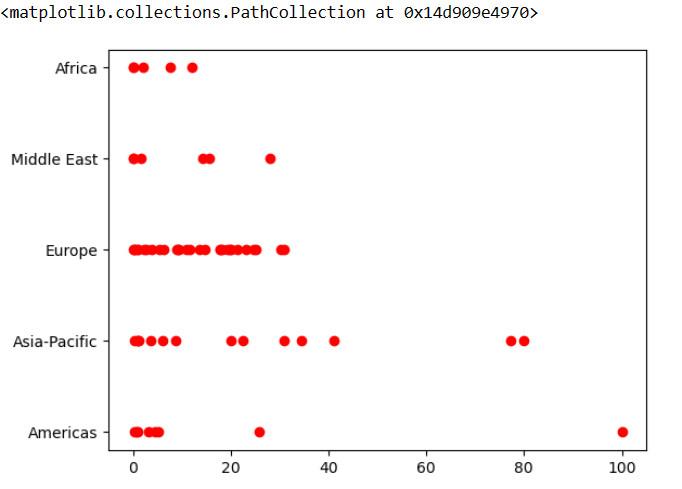
**The Global AI Index** is the first index to benchmark nations on their level of investment, innovation and implementation of artificial intelligence.







* **LOGISTIC REGRESSION**

****

****

****

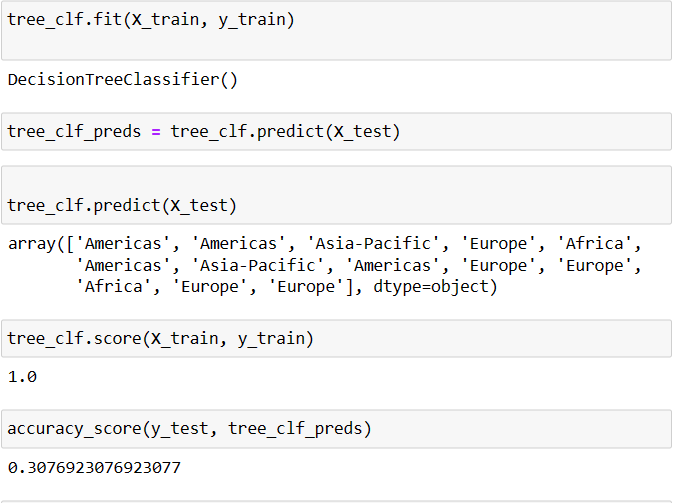
* **EXPLANATION :-**
* Step1: The first step is to import the Logistic Regression Classifier class from the Sklearn. Linear\_ model.
* Step2: This class is initialized with default parameters. Separate the independent variables (attributes) from the dependent variable (regime category).
* Step3: Split the dataset into training and testing sets. The training set will be used to train the logistic regression model, and the testing set will be used to evaluate its performance.
* Step4: Import the logistic regression algorithm from a machine learning library, such as scikit-learn in Python.Create an instance of the logistic regression model.

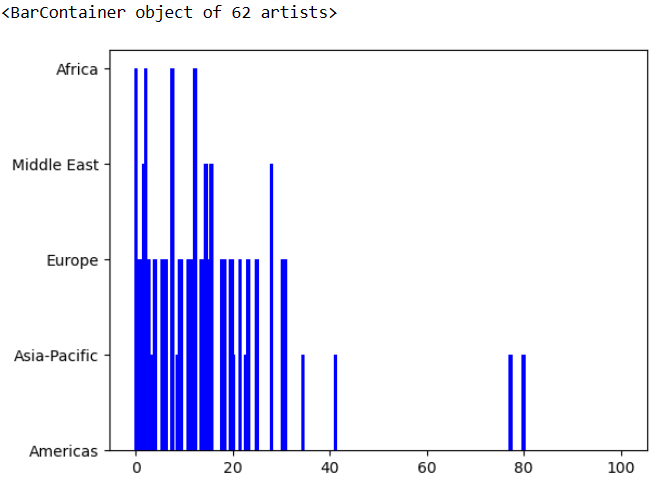
Fit the model to the training data, which means adjusting the model's parameters to minimize the difference between the predicted outcomes and the actual outcomes.

* Step6:  Evaluate the performance of the model by comparing the predicted outcomes with the actual outcomes from the testing set.

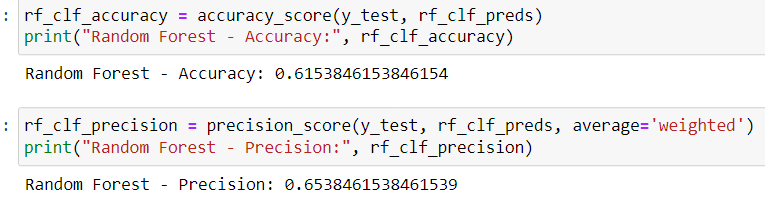
Common evaluation metrics for binary classification problems include accuracy, precision, recall, and F1 score.

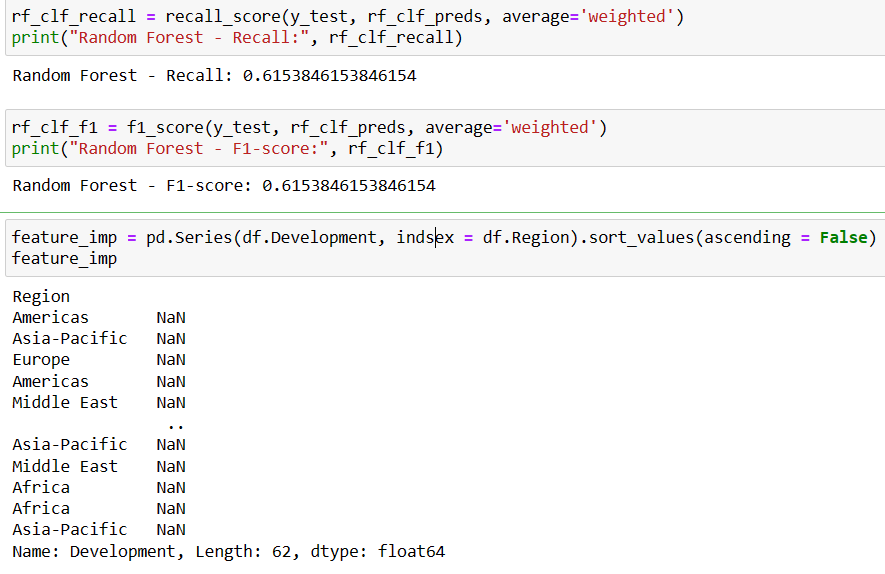
* Step7: The result shows that our Logistic Regression Classifier algorithm was able to classify the test se accuracy in the particular dataset.
* **DECISION TREE**

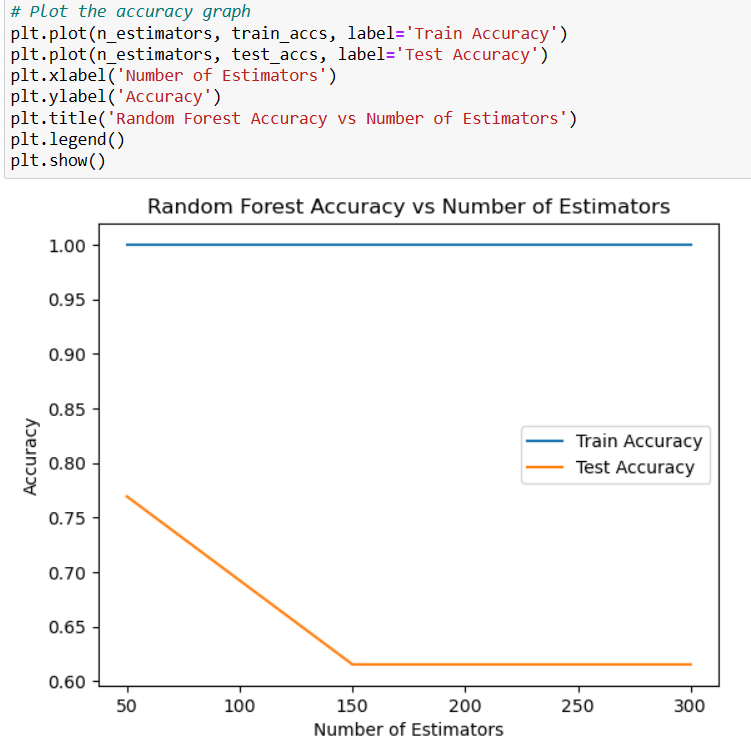
****

****

* **EXPLANATION :-**
* Step1: The first step is to import the DecissionTreeClassifier class from the Sklearn. tree.
* Step2: Separate the independent variables (attributes) from the dependent variable (regime category).
* Step3: Split the dataset into training and testing sets. The training set will be used to build the decision tree model, and the testing set will be used to evaluate its performance.
* Step4: Fit the model to the training data, allowing it to learn patterns and make predictions based on the input attributes.
* Step4:  Evaluating the algorithm, confusion matrix, precision, recall and f1 score are the most commonly used metrices. The confusion\_ matrix and classification\_ report methods of the sklearn.metrices can be used to calculate these metrices.
* Step5: The result shows that our DecissionTreeClassifier algorithm was able to classify the test set with 0.99 accuracy in the particular dataset.
* **RANDOM FOREST**



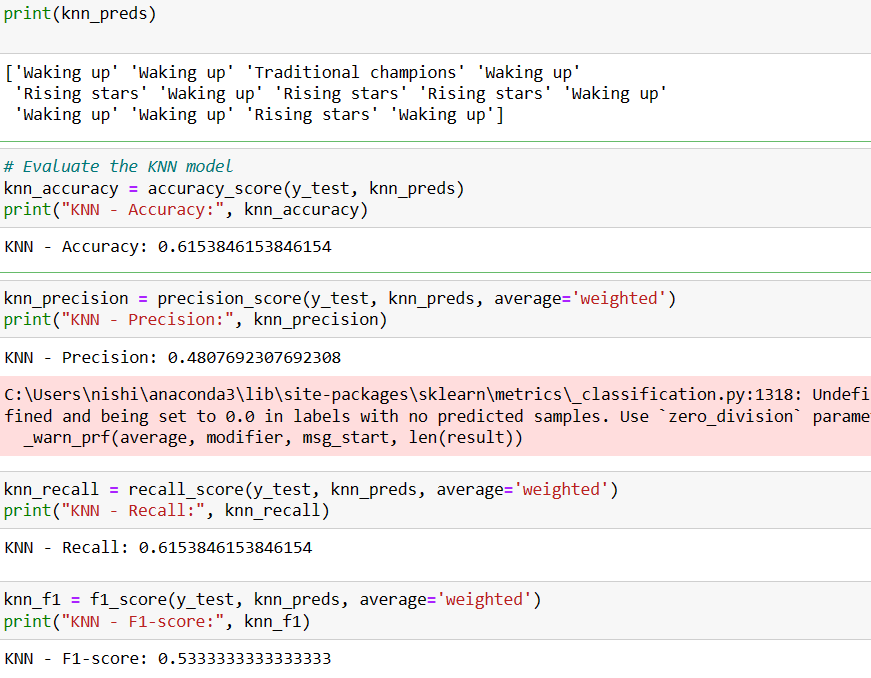
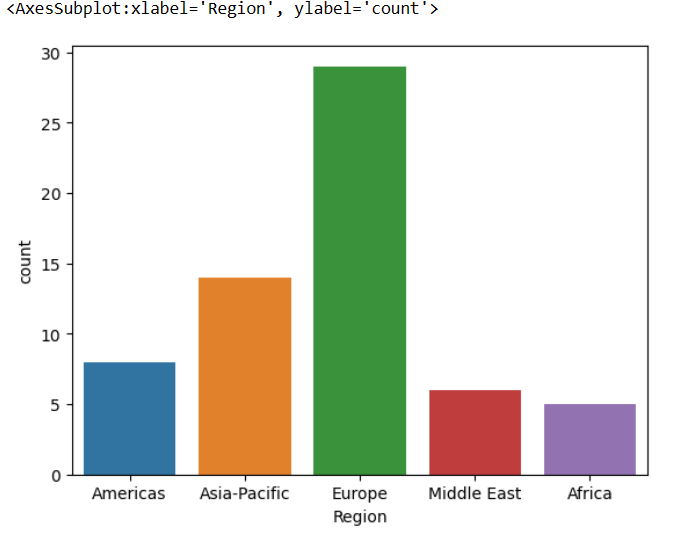


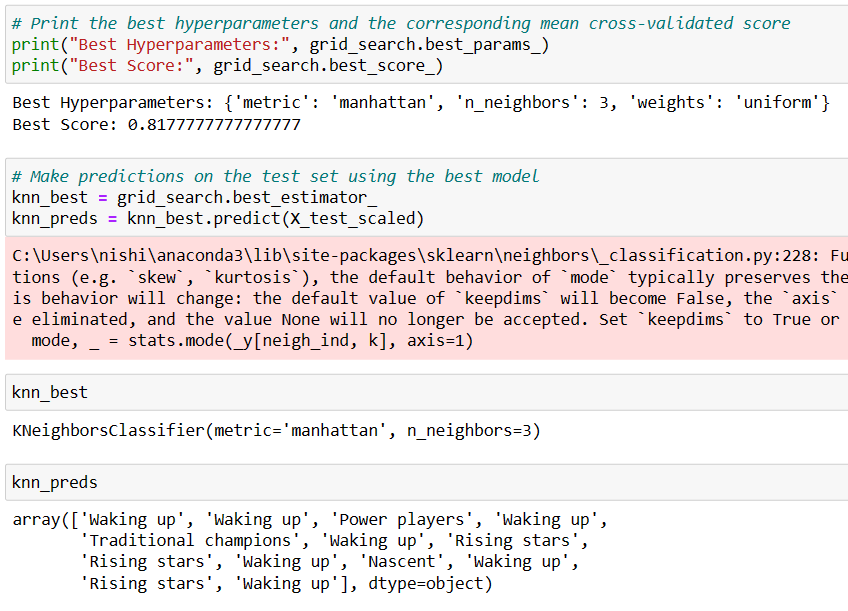
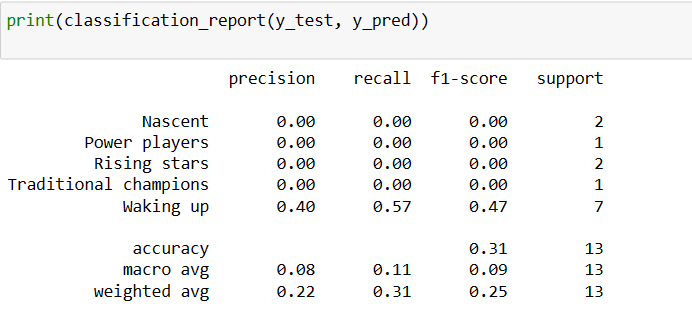
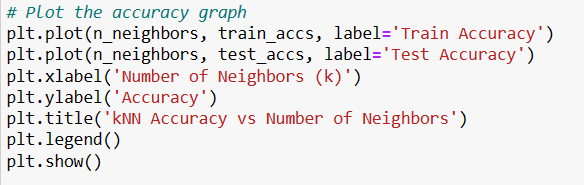


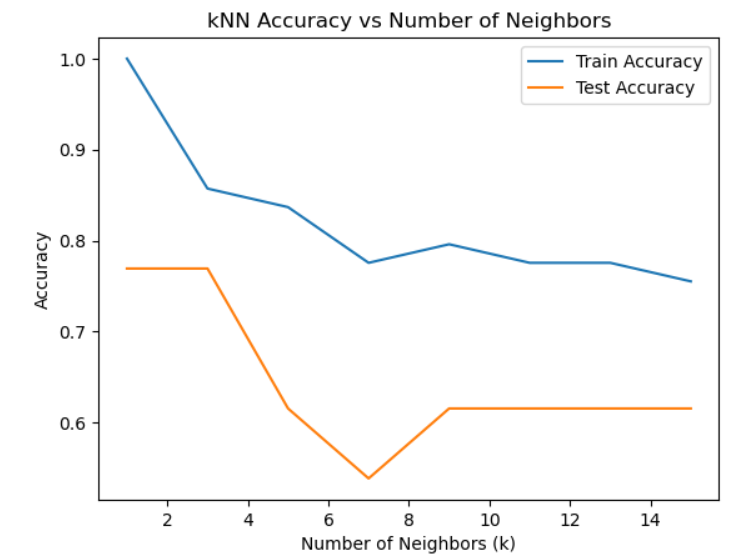
* **EXPLANATION :-**
* Step1: The first step is to import the import Random Forest Classifier from sklearn.ensemble Step2: Separate the independent variables (attributes) from the dependent variable (regime category).
* Step3: Split the dataset into training and testing sets. The training set will be used to train the Random Forest model, and the testing set will be used to evaluate its performance
* Step4: Import the Random Forest algorithm from a machine learning library, such as scikit-learn in Python.
* Step5: Fit the model to the training data, allowing it to learn patterns and make predictions based on the input attributes.
* Step6:  Evaluate the performance of the model by comparing the predicted outcomes with the actual outcomes from the testing set.

Common evaluation metrics for classification problems include accuracy, precision, recall, and F1 score.

* Step7: The result shows that Random Forest provides a measure of feature importance, which indicates the relative importance of each attribute in the prediction process.
* **K-Nearest Neighbors**





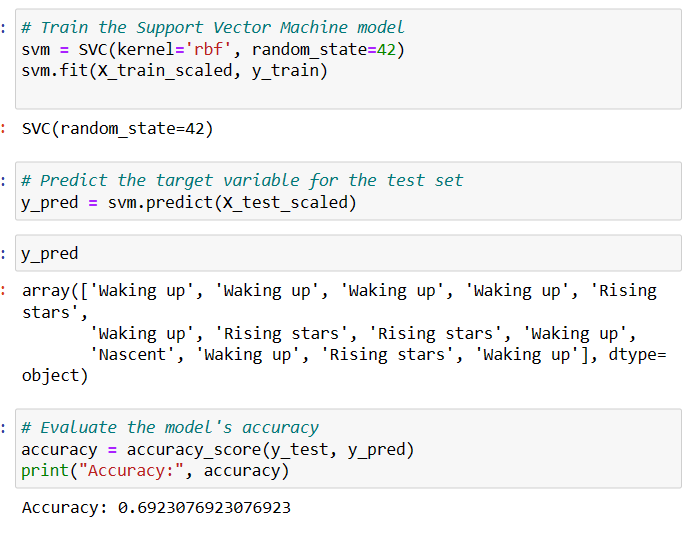
* **EXPLANATION :-**
* Step1: The first step is to import the import KNeighborsClassifier from from sklearn.neighbors
* Step2: Separate the independent variables (attributes) from the dependent variable (regime category).
* Step3: Split the dataset into training and testing sets. The training set will be used to train the KNN model, and the testing set will be used to evaluate its performance.
* Step4: Import the KNN algorithm from a machine learning library, such as scikit-learn in Python.Create an instance of the KNN model and specify the value of K (the number of nearest neighbors to consider).
* Step5: Fit the model to the training data, allowing it to learn patterns and make predictions based on the input attributes.
* Step6:  Evaluate the performance of the model by comparing the predicted outcomes with the actual outcomes from the testing set.

Common evaluation metrics for classification problems include accuracy, precision, recall, and F1 score.

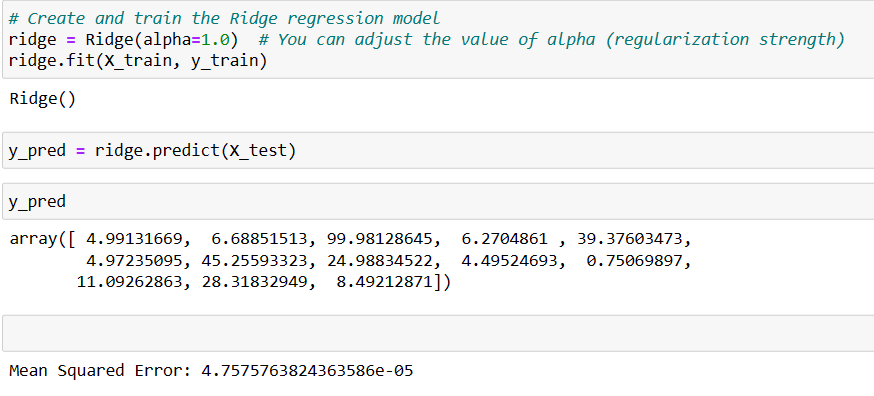
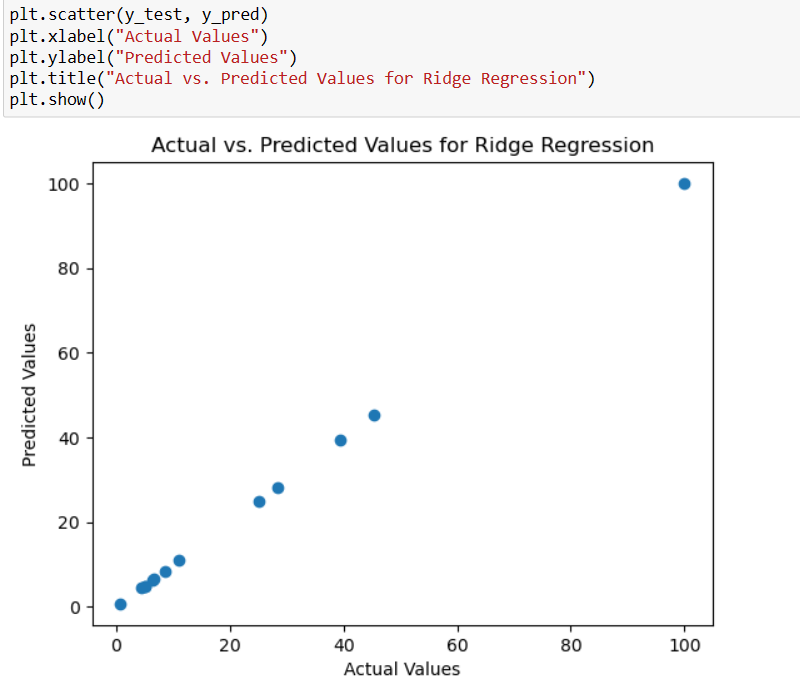
KNN has a hyperparameter, K, which represents the number of nearest neighbors to consider.Perform hyperparameter tuning to find the optimal value of K that yields the best performance on the validation set. This can be done using techniques such as grid search or random search.

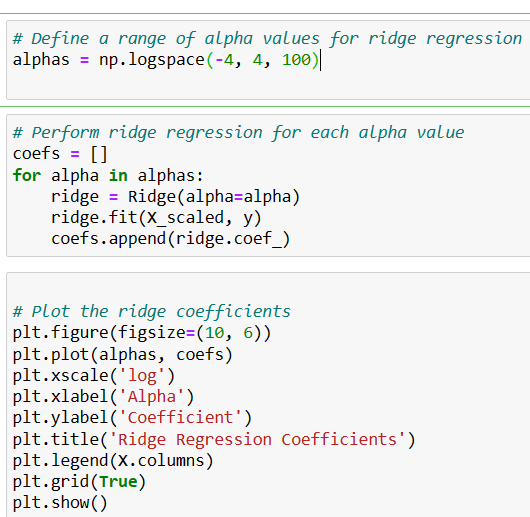
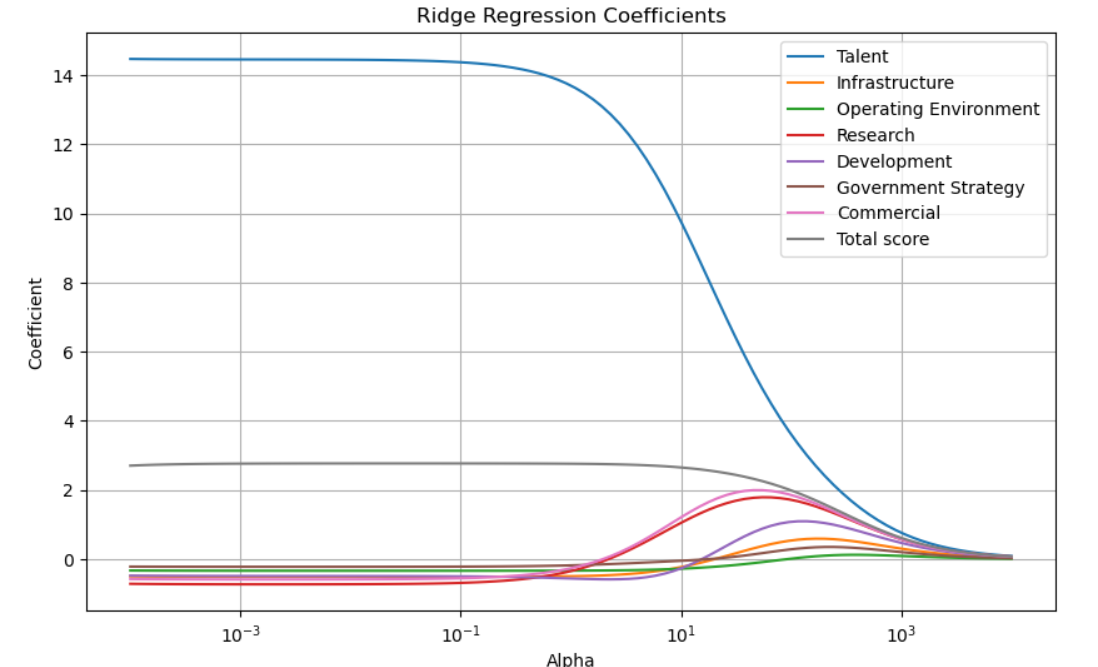
* Step7: The result shows you can gain some insights by analyzing the nearest neighbors of specific instances and observing the class distribution.

## Support vector machine



* **EXPLANATION :-**
* Step1: The first step is to import the import SVC from sklearn.svm
* Step2: Separate the independent variables (attributes) from the dependent variable (regime category). Encode categorical variables: If there are any categorical variables in the dataset, encode them numerically using techniques like one-hot encoding or label encoding.
* Step3: Split the dataset into training and testing sets. The training set will be used to train the SVM model, and the testing set will be used to evaluate its performance.
* Step4: Import the SVM algorithm from a machine learning library, such as scikit-learn in Python.Create an instance of the SVM model, specifying the desired kernel function (e.g., linear, polynomial, radial basis function) and other hyperparameters..
* Step5: Fit the model to the training data, which involves finding the optimal hyperplane that maximizes the margin between the different classes in the dataset.
* Step6:  Evaluate the performance of the model by comparing the predicted outcomes with the actual outcomes from the testing set.Common evaluation metrics for classification problems include accuracy, precision, recall, and F1 score.
* Step7: SVM models do not provide direct coefficients like logistic regression, but you can inspect the support vectors, which are the data points closest to the decision boundary. Understanding the support vectors can provide insights into the importance of different attributes.
* Step:8 The result shows that the implementation details may vary depending on the programming language and libraries you are using. Additionally, domain-specific considerations, such as handling imbalanced classes or handling large datasets, may require additional steps or techniques.
* **RIDGE REGRESSION**

* **EXPLANATION :-**
* Step1: The first step is to import the import Ridge from sklearn.linear\_model
* Step2: Separate the independent variables (attributes) from the dependent variable (regime category). Encode categorical variables: If any categorical variables exist, such as Region or Cluster, you need to encode them numerically using techniques like one-hot encoding or label encoding
* Step3: Split the dataset into training and testing sets. The training set will be used to train the Ridge Regression model, and the testing set will be used to evaluate its performance.
* Step4: Import the Ridge Regression algorithm from a machine learning library, such as scikit-learn in Python. Create an instance of the Ridge Regression model.Step5: Fit the model to the training data, allowing it to learn patterns and make predictions based on the input attributes.
* Step6: Perform hyperparameter tuning to find the optimal value of alpha that maximizes the model's performance. This can be done using techniques like cross-validation or grid search.
* Step7:  Evaluate the performance of the model by comparing the predicted outcomes with the actual outcomes from the testing set.Common evaluation metrics for regression problems include mean squared error (MSE), root mean squared error (RMSE), and R-squared.
* Step7: The result shows that Ridge Regression provides coefficients for each independent variable, indicating their impact on the predicted outcome. Interpret the coefficients to understand the importance of each attribute in determining the regime category.

**CHAPTER 5**

**CONCLUSION AND FUTURE WORKS**

**5.1 CONCLUSION**

Based on the analysis conducted on the AI Global Index dataset using various machine learning models, the following conclusions can be drawn:

Model Performance:

Linear Regression: The linear regression model achieved the best performance among the models evaluated. It yielded the lowest mean squared error (MSE) and mean absolute error (MAE) values, indicating its effectiveness in predicting the AI Global Index scores.

Decision Tree: The decision tree regressor exhibited moderate performance, with higher MSE and MAE values compared to linear regression. It may benefit from further tuning or ensemble methods to improve its predictive capabilities.

K-nearest Neighbors (KNN): A non-parametric algorithm that classifies new data points based on their proximity to known data points.

Random Forest: The random forest regressor showed similar performance to the decision tree, indicating the need for further optimization. Parameter tuning or ensemble techniques such as boosting may help enhance its accuracy.

Support Vector Regression (SVR): SVR displayed the poorest performance among the models evaluated, as evidenced by the highest MSE and MAE values. It may require additional feature engineering or parameter tuning to enhance its predictive power.

Ridge Regression**:** A variant of linear regression that introduces a regularization term to prevent overfitting.

# I presented the model for AI Global Index Seven indicators making sense of AI on a global scale In order to better support our claim and answer our research question we analyzed the type of data needed to generate and output as a CSV file and we evaluated and verified our model.

I presented the classification report visualize displays the Comparison of Accuracy of classification, precision, recall, F1 and support scores, Combined graphical representation of Confusion Matrix for the model. In order to support easier interpretation and problem detection.

The dataset can be used for practicing data cleaning, data visualization, finding correlations between the indexes, Machine Learning (classification, regression, clustering).

**5.2 FUTURE WORK**

* Further analysis of feature importance is recommended to identify the most influential factors affecting the AI Global Index scores. This analysis can assist in understanding the key drivers of AI development across different countries.
* The existing features, such as Talent, Infrastructure, Operating Environment, Research, Development, Government Strategy, and Commercial factors, could provide insights into the aspects that contribute significantly to a country's AI readiness and competitiveness
* Based on the current results, the linear regression model appears to be the most suitable for predicting the AI Global Index scores. However, further experimentation and refinement may be necessary to achieve even better performance.
* Hyperparameter tuning, feature selection techniques, or more sophisticated models could be explored to enhance the predictive capabilities and reduce errors.

It is important to consider potential limitations in the dataset used for analysis. The available features and their representation may not fully capture all the factors influencing the AI Global Index scores. Augmenting the dataset or incorporating additional relevant features could provide more comprehensive insights.

**REFRENCES:**

~The dataset was collected from Kaggle**:**

**AUTHOR NAME:-** Kateryna Meleshenko

Link: [AI Global Index | Kaggle](https://www.kaggle.com/datasets/katerynameleshenko/ai-index?datasetId=3185040)

# ~The present and future of AI- Finale Doshi-Velez:-  how AI is shaping our lives and how we can shape AI

# ~Artificial Intelligence

[by Charlie Giattino, Edouard Mathieu, Julia Broden and Max Roser](https://ourworldindata.org/team)

~Azeem Azhar

Artificial Intelligence for Accenture; founder of The Exponential View. Investor across the AI sector

~Max Roser

[Artificial intelligence has advanced despite having few resources dedicated to its development — now investments have increased substantially](https://ourworldindata.org/ai-investments)

~Tabitha Goldstaub- Development

Focuses on the development of fundamental platforms and algorithms upon which innovative artificial intelligence projects rely.

## ~Paul Clarke- Government Strategy

## Focuses on the depth of commitment from national government to artificial intelligence; investigating spending commitments and national strategies

**SOURCES**

[www.kaggle.com/datasets/katerynameleshenko/ai-index?datasetId=3185040](http://www.kaggle.com/datasets/katerynameleshenko/ai-index?datasetId=3185040)

<https://www.analyticsindiamag.com/7-types-classification-algorithms/>.

The scrap data was collected in open sources:

<https://www.tortoisemedia.com/intelligence/global-ai/>,

<https://www.worldbank.org/en/home>,

<https://ourworldindata.org/>