Classifying Global Income and Region Categories Using Decision Tree Models: Analysing Air Pollution and Population Data

Abstract

The relationship between environmental factors and economic indicators is a subject of increasing interest and importance. This project investigates the predictive capability of air pollution levels and population size on a country's income level and regional classification using data from numerous countries. Utilizing a comprehensive dataset that includes various countries' air pollutants, population statistics, and economic data, Multi-class Classification Trees were applied to develop predictive models. The predictors include air pollution levels (nitrogen oxide (NOx), sulphur dioxide (SO_2), carbon monoxide (CO), organic carbon (OC), non-methane volatile organic compounds (NMVOC), black carbon (BC), and ammonia (NH_3) emissions)[1], along with demographic metrics such as urban population[2], rural population[3], and total population[4]. The targets for the models are to predict the income category (low, lower middle, upper middle and high income) and region category (demographic regions) of the countries. The initial model for predicting income category achieved an accuracy of 99.25%, which was further improved to 99.92% using the ensembling method Bagging. Across various models, the accuracies ranged from 94.18% to 99.92%. Pruning techniques were implemented on classification trees, adjusting maximum depth and maximum leaf nodes using 5-fold cross-validation. Similarly, for region classification, the initial model achieved an accuracy of 98.75%, with the best model reaching 99.7% accuracy using Bagging. Other models, including pruned trees and gradient boosting, were also explored, with accuracies ranging from 91.95% to 99.7%. The accuracies and the top predictors were noted. In both tasks, the classification trees demonstrated robust performance. The results underscore the importance of considering environmental factors alongside economic metrics in policy formulation.

Introduction and Overview

Air pollution and population dynamics are critical issues affecting countries worldwide. The level of air pollution, often measured by various pollutants such as nitrogen oxides (NOx), sulphur dioxide (SO_2), carbon monoxide (CO), organic carbon (OC), non-methane volatile organic compounds (NMVOC), black carbon (BC), and ammonia (NH_3), has significant health and economic implications. Meanwhile, population size and distribution, including urban and rural populations, impact a country's resources, infrastructure, and overall development. Understanding the interplay between these factors and their influence on a country's economic status and regional characteristics is vital for policymakers and researchers.

This project aims to explore the relationship between air pollution levels, population size, and economic indicators such as income level and regional classification across a diverse set of countries. By leveraging machine learning techniques,

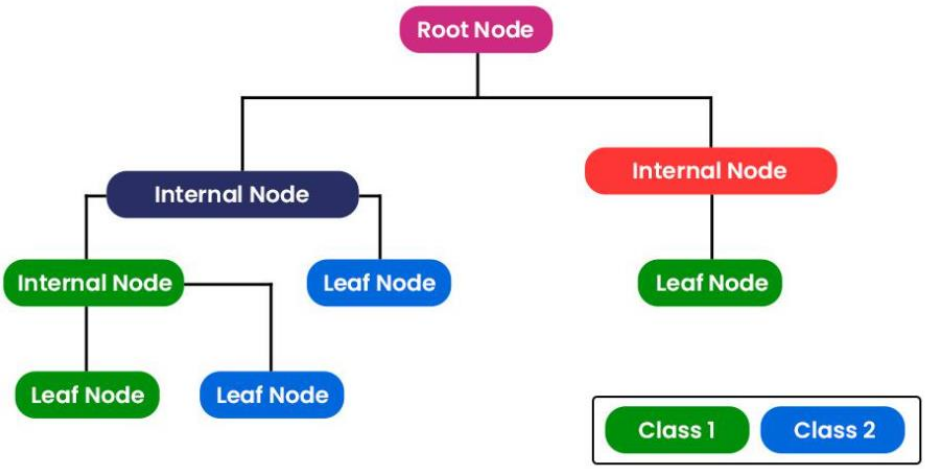
specifically Multi-class Classification Trees, the income level and regional grouping of countries based on their air pollution and population data were predicted.

In this study, various machine learning methods including ensembling techniques like Bagging, Boosting, and Random Forests (RF) were utilised to improve the accuracy and robustness of the models. By employing these advanced techniques, comprehensive analysis is made that can help in formulating targeted policies for economic development and environmental management. Through this project, an endeavour is made to enhance the understanding of the critical factors influencing economic status and regional characteristics, as well as their interplay with environmental metrics.

Theoretical Background:

Decision Trees:

A decision tree is a commonly used algorithm in machine learning that helps with categorizing and forecasting tasks. It has a tree-like layout where each internal node stands for a specific test on a feature, branches show the test results, and leaf nodes display class labels or numerical values.[5]



When creating a decision tree, the goal is to make sure that the resulting groups of data are as pure as possible. This is typically measured using metrics like Gini impurity or entropy. The algorithm goes through each feature and split point in a step-by-step manner to find the best way to divide the data, ultimately creating a tree with branches at each split point.

In this analysis, multi-class trees is used which can classify instances into multiple categories.

Pruning:

When working with decision trees, pruning is the process of cutting away branches or nodes in order to improve the model's accuracy and avoid overfitting. It aims to eliminate unnecessary splits that do not add much value to the tree's predictive abilities. Pruning

plays a key role in finding the right balance between model complexity and performance, leading to a more understandable and adaptable model that works effectively with new data.

Ensemble methods:

Ensemble methods in machine learning are strategies that blend predictions from many separate models to create a stronger and more reliable final prediction. The primary ensemble methods consist of Boosting, Random Forest, and Bagging.

The process of bagging involves running multiple versions of a learning algorithm, such as random forest, where each is trained on a random subset of data and then their predictions are averaged to improve robustness. On the other hand, gradient boosting constructs models sequentially, with each one correcting the errors made by its predecessors, often utilizing decision trees. Unlike bagging, which trains models independently, gradient boosting trains models in a dependent manner, where each model is trained to predict the residual errors of the combined ensemble of all previous models. Random Forests is an ensemble method that combines the principles of bagging and random feature selection to create a robust and powerful predictive model. It consists of a large number of decision trees, each trained on a bootstrap sample of the data and a random subset of features. By averaging predictions from different models, bagging reduces variance, while gradient boosting focuses on reducing bias, resulting in a more precise and reliable predictor for complex datasets. This has proven to be successful in enhancing predictive accuracy, preventing overfitting, and boosting consistency when compared to individual models. Random Forests combine the strengths of bagging and random feature selection to enhance both accuracy and robustness.

Parameter Tuning:

When parameters are tuned for decision trees, adjustments are made to maximize the model's effectiveness. It's similar to tuning a musical instrument to achieve the best possible sound. By tweaking settings such as the tree's maximum depth/leaf nodes or the minimum samples needed to split a node, the accuracy and applicability of the model can be enhanced. Parameter tuning ensures that the decision tree performs at its best for the specific dataset and issue at hand.

Linear Regression:

Linear Regression is a fundamental statistical technique used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It assumes a linear relationship between the variables and aims to minimize the difference between the observed and predicted values.

Methodology

Data clean up or pre-processing:

Air pollution data were obtained from ourworldindata.org and population data (urban, rural, and total) from data.worldbank.org. These sources provided comprehensive air pollution metrics and population distribution details for various countries. The air

pollution data initially contained information from 1750 to 2022, while the population data (urban, rural, and total) spanned from 1960 to 2022. To ensure consistency, data from 1960 to 2022 was considered for both datasets. The datasets were then merged into a unified dataset considering years 1960-2022 for analysis. A metadata sheet attached to the population data was utilized, which included details about the income levels and regions of all countries. Two new columns were created: one for income categories and another for regions. The income category column was populated with values ranging from 1 to 4, corresponding to low income, lower-middle income, upper-middle income, and high income. Similarly, the regions column was assigned values from 1 to 7, representing the regions: East Asia & Pacific (1), Europe & Central Asia (2), Latin America & Caribbean (3), Middle East & North Africa (4), North America (5), South Asia (6), and Sub-Saharan Africa (7). Dummy variables were created for these two target columns, income category and regional category. These dummy variables represented categorical data in a binary format, allowing the incorporation of categorical variables into the predictive models effectively.

Train and Test Data set Splitting:

For all the models, the data was divided into a training set of 75% and test set of 25%.

Exploratory Data Analysis:

The plots and clusters derived from unsupervised learning were analysed to get some insight on the underlying patterns. Other relational bar plots for Income Categories and Region category against their top predictors were visualized.

Models:

Five models were developed for predicting income categories. The initial model employed a Decision Tree classifier fitted on the training data. Subsequently, an effort was made to find the optimal maximum depth, resulting in the selection of a depth of 15 after performing 5-fold cross-validation. The third model utilized a 5-fold cross-validation approach to determine the maximum leaf nodes, with 118 nodes chosen to prune the tree. The fourth model employed the ensembling method Bagging, determining the optimal number of estimators to be 200 through 5-fold cross-validation. Finally, the fifth model utilized Bagging with 100 trees and 190 splits, demonstrating the highest accuracy among all models. Throughout the process, the top five predictors for each model were identified, along with their respective accuracies, facilitating comprehensive comparison and evaluation.

For predicting regional categories, six models were developed. The initial model was established, followed by a pruned model with a maximum depth of 19. Subsequently, cross-validation was employed to prune the model further, selecting a maximum leaf of 118. The fourth model utilized gradient boosting with 5000 trees, while the fifth model utilized bagging with 100 estimators determined through cross-validation. Finally, the sixth model utilized bagging with 100 trees and 190 splits, exhibiting the highest accuracy among all models. Throughout the analysis, the top five predictors for each model were identified, alongside their corresponding accuracies, enabling thorough comparison and assessment.

Linear Regression models were created to predict both income and region classification for comparative analysis.

Computational Results

The computation results are tabulated for each model as below.

1. Income Category Classification Tree

Predicting the income category of different countries		
Models	**Model Performance**	
	Accuracy (%)	**Top 5 important Features**
Initial Tree	99.25	1. Rural Population 2. SO2 3. NO 4. CO 5. Org. Carbon
Pruned Tree with max depth - 15 (Used 5 fold CV)	96.98	1. Rural Population 2. SO2 3. NO 4. CO 5. Org. Carbon
Pruned Tree with max leaf node - 118 (Used 5 fold CV)	94.18	1. Rural Population 2. SO2 3. NO 4. CO 5. Org. Carbon
Ensemble Method - Bagging n_estimator 200 (Used 5 fold CV)	99.35	1. SO2 2. Rural Population 3. NO 4. NMVOC 5. Org. Carbon
Ensemble Method - Bagging (100 Trees, 190 Split)	99.92	1. SO2 2. Rural Population 3. NO 4. NMVOC 5. Org. Carbon

Important Predictors:

Order	Top predictors	Correlation with Income Category
1	Rural Population	-0.07
2	Sulphur Dioxide	0.11
3	Carbon Monoxide	0.04
4	Nitrogen Oxide	0.1
5	Organic Carbon	-0.05
6	NMVOC	0.06

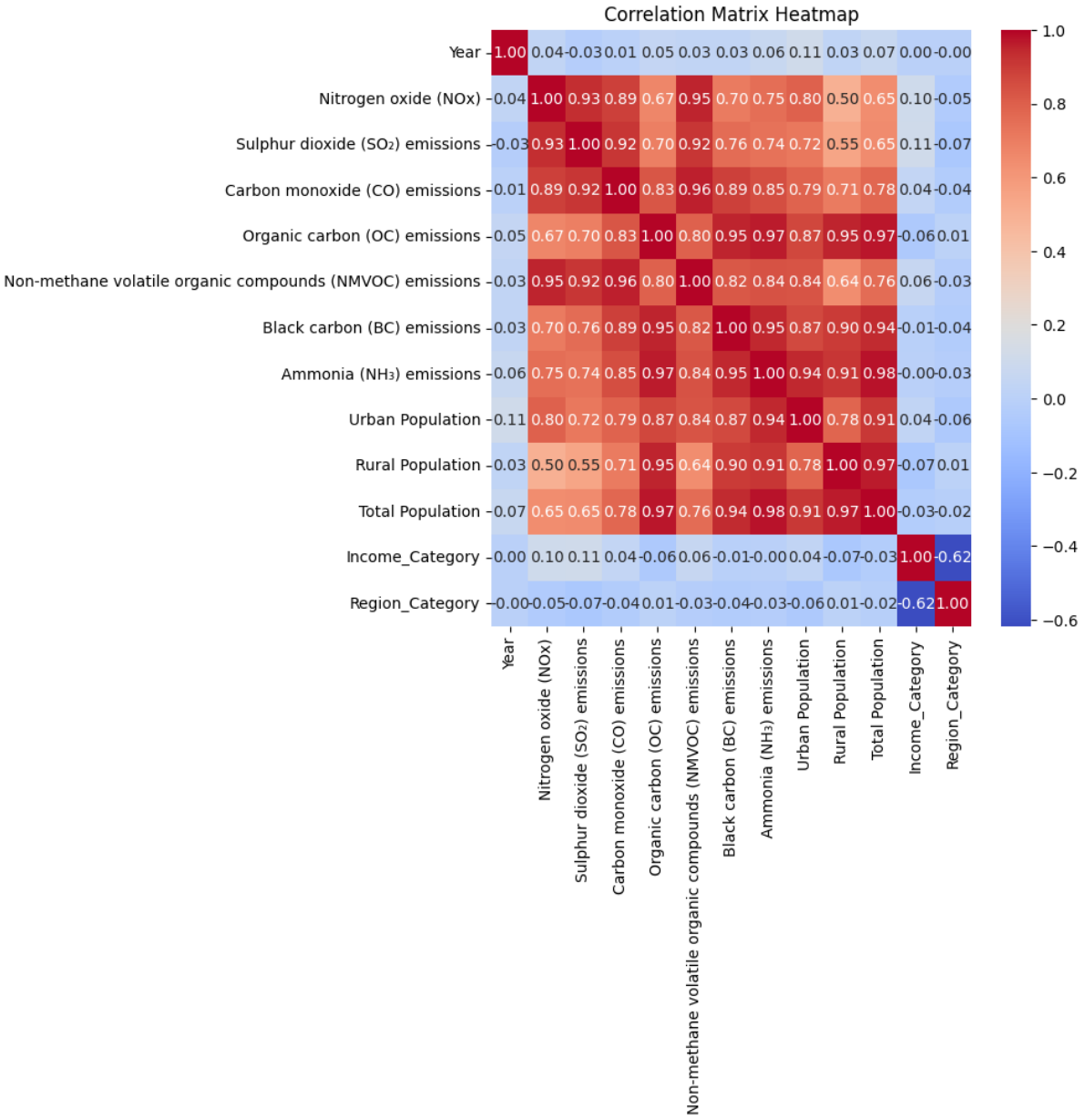
2. Region Category Classification Tree

Predicting the region category of different countries		
Models	**Model Perfomance**	
	Accuracy (%)	**Top 5 important Features**
Initial Tree	98.75	1. Org. Carbon 2. Rural Population 3. SO2 4. NO 5. NMVOC
Pruned Tree with max depth - 19 (Used 5 fold CV)	98.04	1. Org. Carbon 2. Rural Population 3. SO2 4. NO 5. NMVOC
Pruned Tree with max leaf node - 118 (Used 5 fold CV)	91.95	1. Org. Carbon 2. Rural Population 3. SO2 4. NO 5. NMVOC
Ensemble Method - Gradient Boosting (5000 Trees)	97.97	1. Org. Carbon 2. SO2 3. Rural Population 4. Code_KHM (Cambodia) 5. NO
Ensemble Method - Bagging n_estimators -100 Trees (Used 5 fold CV)	98.1	1. Org. Carbon 2. SO2 3. Rural Population 4. Black Carbon 5. CODE_MMR (Myanmar)
Ensemble Method - Bagging (100 Trees, 190 Split)	99.7	1. Org. Carbon 2. SO2 3. Rural Population 4. Black Carbon 5. CODE_MMR (Myanmar)

Important Predictors:

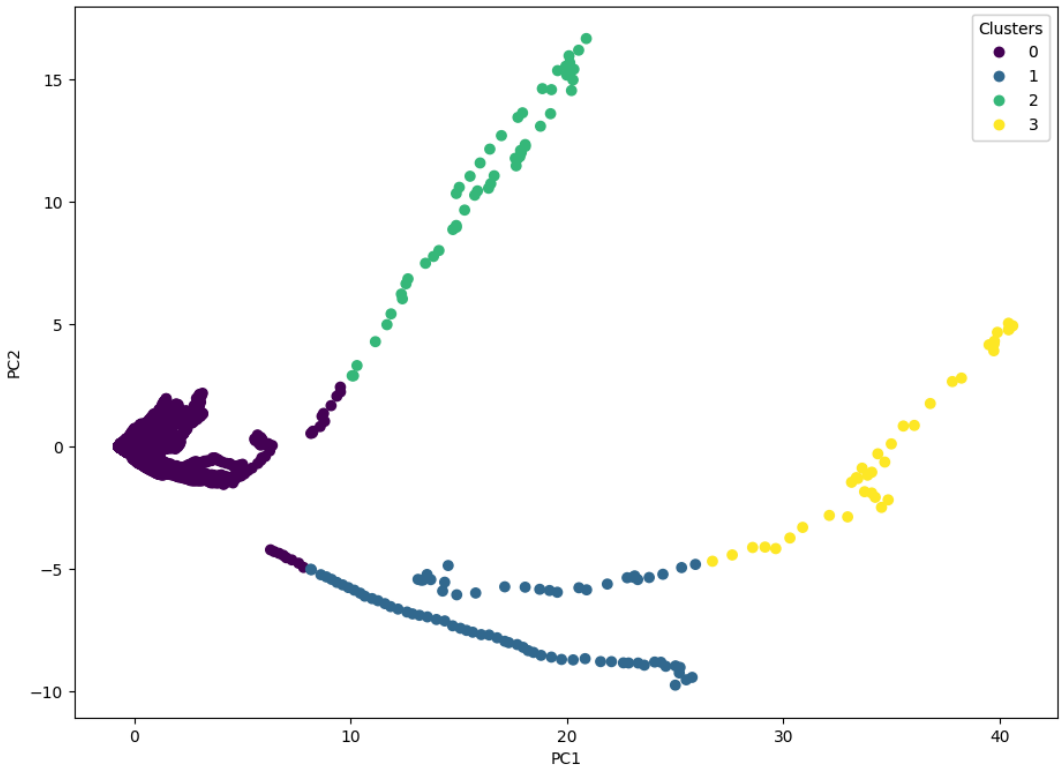
Order	Top predictors	Correlation with Region Category
1	Organic Carbon	0.01
2	Sulphur Dioxide	-0.07
3	Rural Population	0.01

4	Code_KHM (Cambodia)	-0.09
5	Nitrogen Oxide	-0.04
6	CODE_MMR (Myanmar)	-0.09
7	Black Carbon	-0.04



EDA from Unsupervised Learning

Cluster	Countries	Number of Years	Years
Cluster0	India	8	1960 to 1967
	United States	11	2012 to 2022
	All other Countries	63	1960 to 2022
Cluster 1	India	55	1968 to 2022
	China	26	1960 to 1985
Cluster 2	United States	52	1960 to 2011
Cluster 3	China	37	1986 to 2022



Discussion

The project's findings provide important light on the connections between economic indicators, population dynamics, and air pollution. From almost all the models for income category prediction, it was found that certain factors consistently emerge as top predictors, including rural population, Sulphur Dioxide (SO2), nitrogen oxide (NO), carbon monoxide (CO), organic carbon, and non-methane volatile organic compounds (NMVOC). However, upon examining the correlation values of these predictors with income category, it becomes apparent that their relationships with income levels are relatively weak, with correlation coefficients ranging from -0.07 to 0.11. Below are the discussions of the implications of the computational results, significant findings.

Income Category Prediction:

The high accuracy achieved in predicting income categories using air pollution and population data highlights the strong correlation between environmental factors and economic status. The use of Multi-class Classification Trees, especially when enhanced with Bagging, proved effective in capturing these relationships**.** *Rural Population:* The negative correlation between rural population and income category aligns with the trend that higher-income countries tend to have a larger proportion of their population in urban areas. Urbanization is often associated with better economic opportunities and infrastructure. *Sulphur Dioxide (SO_2) Emissions:* SO_2 emissions are indicative of industrial activity predominant in higher-income countries. This suggests that industrial emissions play a significant role in determining a country's economic classification. *Carbon Monoxide (CO):* The positive correlation of CO with income categories indicates that higher levels of CO emissions are associated with more developed industrial activities, typically seen in higher-income countries. This underscores the economic activities tied to industrial emissions and transportation. *Nitrogen Oxide (NO)* and *Non-Methane Volatile Organic Compounds (NMVOC):* These pollutants are commonly linked to transportation and industrial processes, which are more prominent in higher-income countries.

Region Category Prediction:

The high accuracy in predicting regional categories underscores the distinct environmental and demographic characteristics of different regions The models demonstrated the usefulness of Multi-class Classification Trees in regional classification problems, particularly when improved by Bagging, by successfully capturing these subtleties. *Organic Carbon (OC*):As the foremost predictor for regional classification, OC emissions reflect specific regional activities such as biomass burning and industrial processes. This underscores the importance of region-specific environmental policies. The correlation of 0.014577 with regional categories indicates a slight positive association, suggesting that OC emissions are slightly more prevalent in certain regions. This underscores the importance of region-specific environmental policies. *Sulphur Dioxide (SO_2) Emissions*: The negative correlation between SO_2 emissions and regional categories implies that regions with lower levels of sulphur dioxide may belong to different demographic classifications. SO_2 emissions are primarily generated from industrial activities, such as power generation and manufacturing, and their spatial distribution can vary significantly across regions. Regions with lower SO_2 emissions may exhibit cleaner air quality and less industrial pollution, influencing their classification into distinct demographic regions. *Rural Population:* The positive correlation between rural population and regional categories suggests that regions with a higher proportion of rural inhabitants may exhibit specific demographic characteristics. Rural populations often engage in agricultural and traditional livelihood activities, shaping the socio-economic landscape of the region. The presence of a significant rural population may indicate distinct cultural, economic, and infrastructural features, influencing the regional classification. *Black Carbon (BC)* :While less influential than OC and SO_2 emissions, BC emissions still play a role in distinguishing regional categories. BC, primarily originating from incomplete combustion processes, may reflect regional transportation and energy consumption patterns, contributing to regional differentiation. Code_MMR (Myanmar): The presence of specific regional codes, such as Code_MMR representing Myanmar, suggests that certain countries or regions exhibit unique characteristics that influence their classification. These regional identifiers capture geopolitical, cultural, and geographical distinctions that contribute to regional diversity. Understanding the

nuances associated with individual countries or regions allows for tailored policy involvements and targeted development strategies. By accounting for these regional differences, policymakers can address specific challenges and leverage unique opportunities for sustainable growth and development.

For comparison purpose, Linear regression was performed to predict income category and Region category. It was observed that the testing accuracies for both models were exceptionally high. The observation of remarkably high testing accuracies, reaching 99.99%, for both income and regional classification using Linear Regression models raises eyebrows and prompts further scrutiny. While Linear Regression is typically utilized for predicting continuous outcomes like income levels, its application to categorical variables such as regional classification is less conventional. Such extraordinary accuracies suggest the need for a thorough examination of the modelling process and potential overfitting issues. Consequently, additional validation and analysis are necessary to ensure the reliability and generalizability of the findings.

Conclusion

This project delved into the intricate relationship between environmental factors, population dynamics, and economic indicators across diverse countries. By leveraging a comprehensive dataset encompassing air pollution levels, population statistics, and economic data, advanced machine learning techniques were employed, particularly Multi-class Classification Trees, to predict income levels and regional classifications. The results showcased remarkable predictive capabilities, with accuracies ranging from 91.95% to 99.92% across various models. Through the utilization of ensembling methods such as Bagging, significant improvements in predictive performance was achieved, underscoring the robustness of the models. In forecasting income levels, sulphur dioxide (SO2) stood out as the primary predictor, emphasizing the significant impact of air pollution on a country's economic outcomes. Moreover, the substantial presence of rural populations and nitrogen oxide (NO) levels underscored the crucial role of demographic dynamics and air quality in determining income levels. Additionally, non-methane volatile organic compounds (NMVOC) and organic carbon emerged as influential predictors, further accentuating the complex relationship between environmental factors and economic metrics. Regarding regional classification, organic carbon emerged as the foremost predictor, indicating its pivotal role in defining demographic regions. SO2 and rural population also featured prominently, signalling their influence on regional characteristics. Furthermore, black carbon and CODE_MMR (Myanmar) completed the list of top predictors, illustrating the diverse range of variables contributing to regional classifications. These findings underscore the intricate interplay between environmental factors, population dynamics, and economic outcomes. By understanding the significance of these predictors, policymakers can devise targeted interventions aimed at fostering sustainable development and mitigating environmental risks.

Bibliography/References:

1. https://ourworldindata.org/explorers/air-pollution?time=earliest..2022&uniformYAxis=0&country=USA~CHN~IND~GBR~OWID_WRL~OWID_NAM&Pollutant=All+pollutants&Sector=From+all+sectors+%28Total%29&Per+capita=false
2. https://data.worldbank.org/indicator/SP.URB.TOTL

3. https://data.worldbank.org/indicator/SP.RUR.TOTL
4. https://data.worldbank.org/indicator/SP.POP.TOTL
5. https://www.geeksforgeeks.org/cart-classification-and-regression-tree-in-machine-learning/
6. https://medium.com/chinmaygaikwad/hyperparameter-tuning-for-tree-models-f99a66446742
7. Author: Brendan Martin Title: How to Find Decision Tree Depth via Cross-Validation Publication: Towards Data Science Date: June 19, 2019 URL: https://towardsdatascience.com/how-to-find decision-tree-depth-via-cross-validation-2bf143f0f3d6
8. Hastie, T., Tibshirani, R., & Friedman, J. (2009): Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. ISBN: 978-0-387-84857-0
9. https://medium.com/@roshmitadey/bagging-v-s-boosting-be765c970fd1#:~:text=Bagging%3A%20Aims%20to%20create%20diverse,the%20weaknesses%20of%20its%20predecessors.