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

Berry Team

Siva Sushmitha Meduri

Janani Krishnamurthy

Lakshmi Prasanna Kumar Nalabothu

CONTENTS:

INTRODUCTION

Air pollution and population dynamics are critical issues affecting countries worldwide.

In this project, we aim to classify countries into different income and region categories using decision tree models. We leverage datasets containing air pollution and population data from various countries to build predictive models that provide insights into global income and regional distribution patterns.

GOALS

Income Category Classification:

- Decision tree models to categorize countries into different income groups
- Income Categories: High income, Upper middle income, Lower middle income, Low income

Region Category Classification:

- Decision tree models to categorize countries into different income groups
- **Region Categories:** East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa

Why This Matters:

Understanding the factors that contribute to air pollution disparities can inform policy decisions and targeted interventions to improve air quality and public health.

DATA SOURCES

1) Air Pollution Data

Source: Our World in Data

Link: Air Pollution Data Explorer - Our World in Data

2) Urban Population Data

Source: World Bank

Link: Urban population | Data (worldbank.org)

3) Rural Population Data:

Source: World Bank

Link: Rural population | Data (worldbank.org)

4) Total Population Data:

Source: World Bank

Link: Population, total | Data (worldbank.org)

DATA PREPARATION

Data Merging:

Combined datasets on pollution and population for each country.

Metadata Information:

Number of Countries: 179

Time Period: 1960 to 2022

Variables: Various pollution measures, population figures, Income Categories and Region Categories.

Income Categories of Countries: Low Income Countries(1), Lower Middle Income(2), Upper Middle Income (3) and Higher Income Countries(4)

Region

Region Categories of Countries :East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa

THEORITICAL BACKGROUND

Decision Trees: A versatile supervised learning algorithm that can be used for both classification and regression tasks. It creates a tree-like model of decisions and their possible consequences.

Multiclass Classification: Technique used to categorize instances into one of three or more classes based on their features.

Bagging and Boosting: Ensemble methods that combine multiple models to improve predictive performance.

Cross-Validation: A technique for evaluating model performance and preventing overfitting by partitioning the data into training and validation sets.

WHY DECISION TREES?

Handles Categorical Variables

Interpretability

Non-Linear Relationships

Why not other models?

SVMs: Computationally expensive with large datasets.

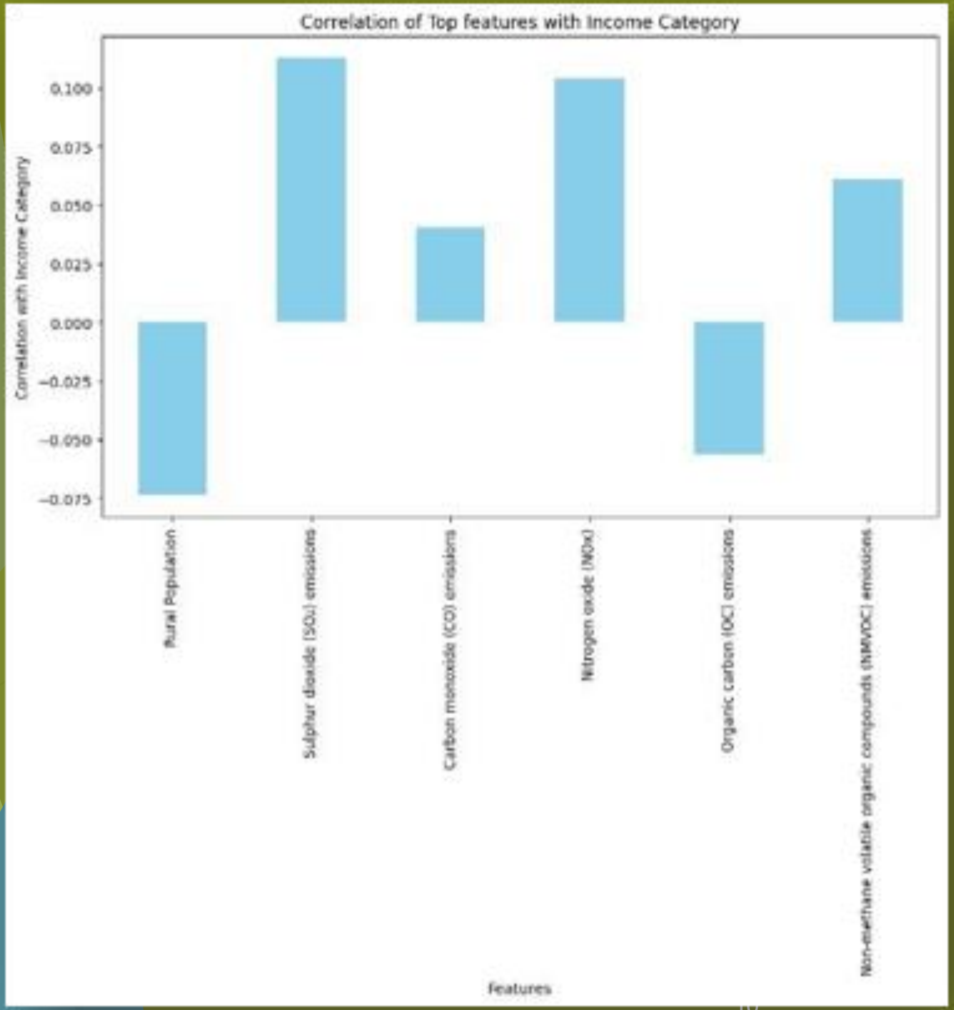
Neural Networks: Simple dataset, doesn't have any images or audio files.

INCOME CLASSIFICATION

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

RELATIONSHIP WITH 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	Non-methane volatile organic compounds(NMVOC)	0.06



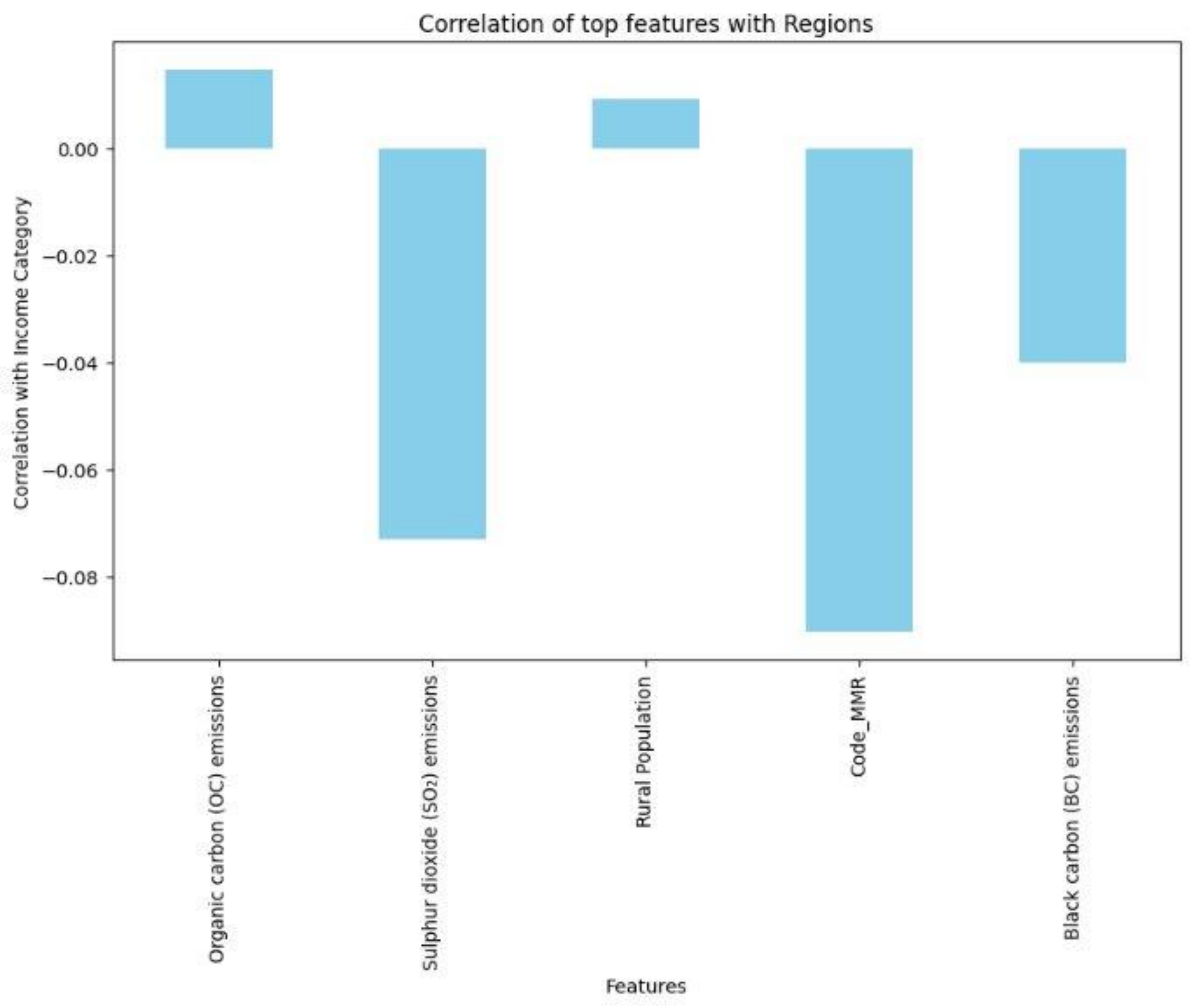
REGION CLASSIFICATION

Predicting the region category of different countries

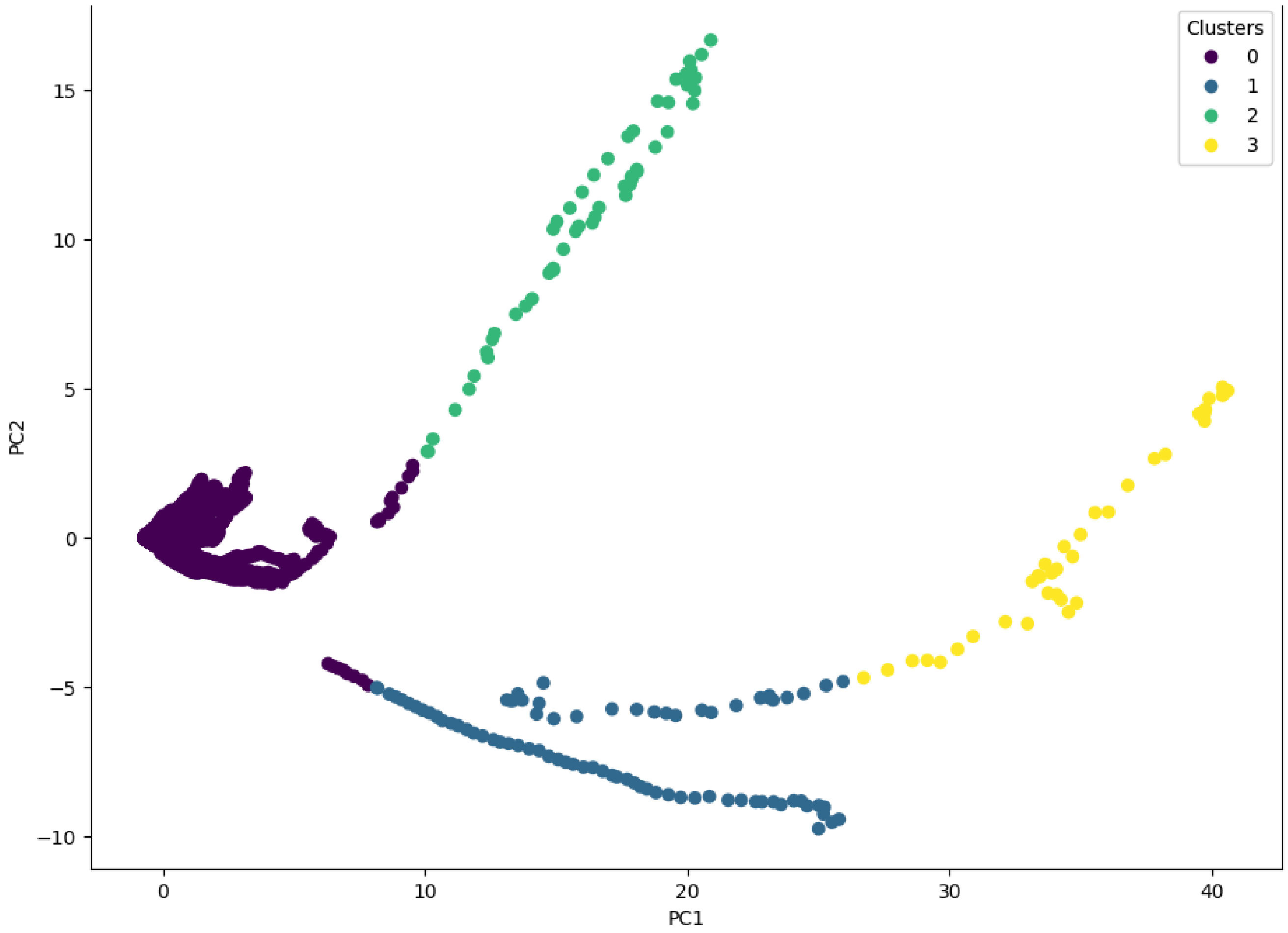
Models	Model Perfomance	
	Accuracy (%)	**Top 5 important Features**
Initial Tree	98.75	1. Org. Carbon 2. Rural Population 3. SO2 4. NO 5. NMVOC
Pruned Tree with max depth - 19 (Used 5 fold CV)	98.04	1. Org. Carbon 2. Rural Population 3. SO2 4. NO 5. NMVOC
Pruned Tree with max leaf node - 118 (Used 5 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(Combodia) 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)

RELATIONSHIP WITH PREDICTORS

ORDER	TOP PREDICTORS	CORRELATION WITH REGION CATEGORY
1	Organic Carbon	0.01
2	Sulphur Dioxide	-0.07
3	Rural Population	0.01
4	Nitrogen Oxide	-0.04
5	Code_MMR (Myanmar)	-0.09
6	Black Carbon	-0.04



EXPLORATORY DATA ANALYSIS



INTERESTING FINDINGS FROM EDA and 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

CONCLUSION

Summary of Findings

- Key predictors identified for income category:
  - Rural Population, SO_2, NO, CO, Organic Carbon, NMVOC
- Key predictors for regional classification:
  - Organic Carbon, SO_2, Rural Population, Black Carbon, CODE_MMR(Myanmar)
- Achieved high accuracies:
  - Income prediction: 94.18% to 99.92%
  - Region prediction: 91.95% to 99.7%

Possible Sources of error:

Weak individual correlations with income levels suggest complexity.

CONCLUSION

Future Improvements:
- Incorporate additional variables to capture broader socioeconomic factors.
- Explore advanced modeling techniques to enhance prediction accuracy.

Impact:
- Emphasizes the importance of environmental and demographic data in creating effective policies.
- Proposes ways to implement specific actions that support sustainable development.

Q&A

THANK YOU