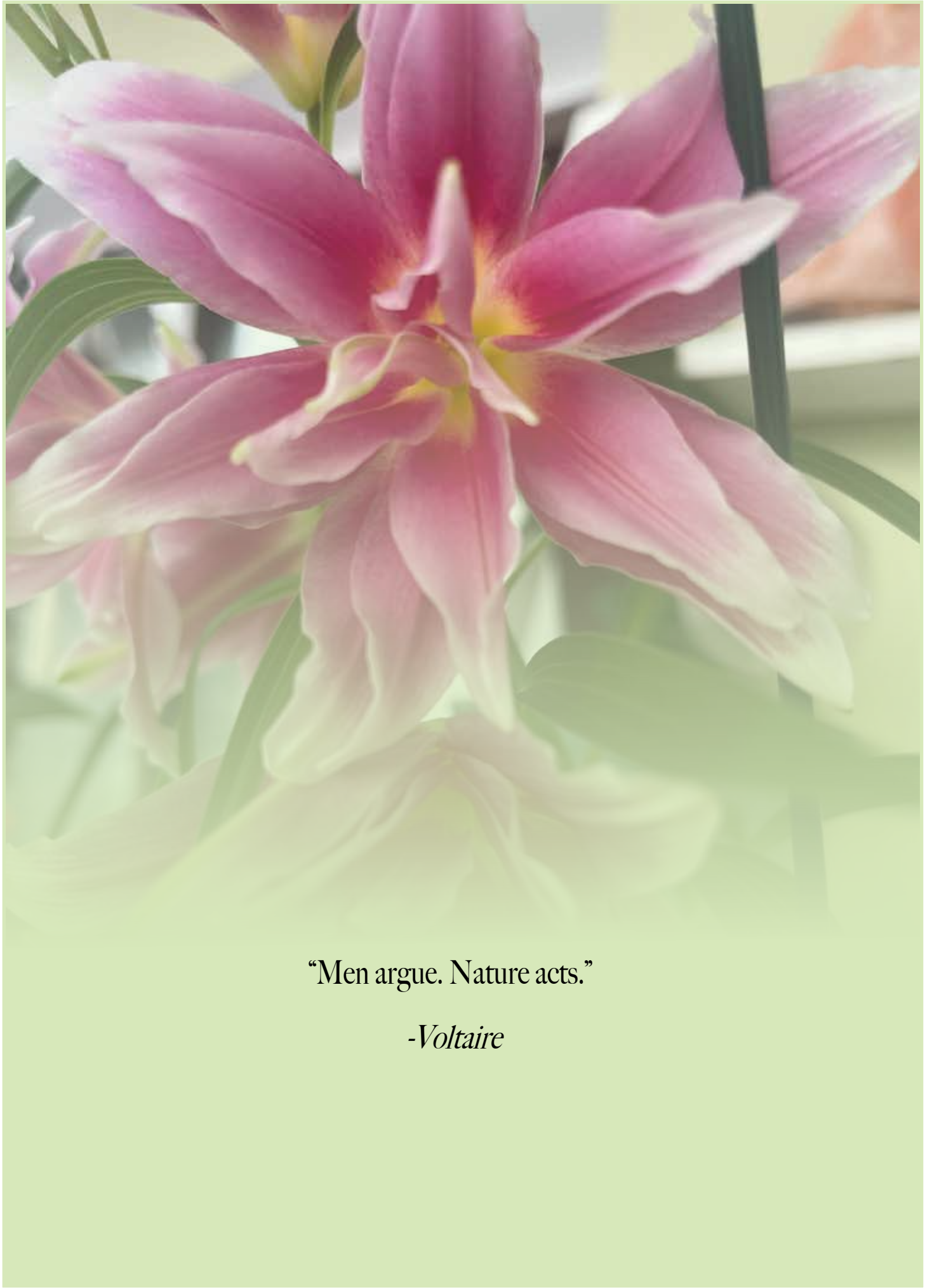
Using Socio-Economic Geopolitical Data to Predict Carbon Dioxide Emissions Around the Globe

A Machine Learning Approach



"Men argue. Nature acts."

-*Voltaire*

Contents

I. Introduction 4
II. Executive Summary 8
III. Stage 1: Discovery 11
IV. Stage 2: Diagnosis 14
V. Stage 3: Predictive Methodology 17
VII. Implications 23
VIII. Conclusion 23
IX. References: 24

I. Introduction

Urban environments are one of the main producers of greenhouse gas emissions. This study takes a macro approach to uncovering unexpected variables that both correlate and/or cause emissions. By examining global data between 1995-2020, this study uncovers unexpected relationships between sociopolitical indicators and emissions rates and variations.

A combination of bad reporting, political realities, and a lack of resources often results in data gaps that can exacerbate inequity, fail to achieve goals, and undermine progress. This work uses machine learning algorithms to fill the gaps and remove blind spots. Fighting climate change is a global issue, but not every actor can tackle the crisis

with the same resources. This model works to solve this issue.

By using data from around the world that goes beyond economics and environmental information, the study works to identify which social, cultural, political, and economic variables have an unexpected relationship with the emissions trend. The goal of the project is to identify such variables, apply weights to them, and then predict emissions outcomes for those countries where the information is under-reported. The data sourced for this project came from Our World Data. The feature data used to train the models includes, but is not limited to, social spending, women in leadership, military expenditure, democracy score, and Cantril ladder score. The data

Countries with Missing Emissions Data

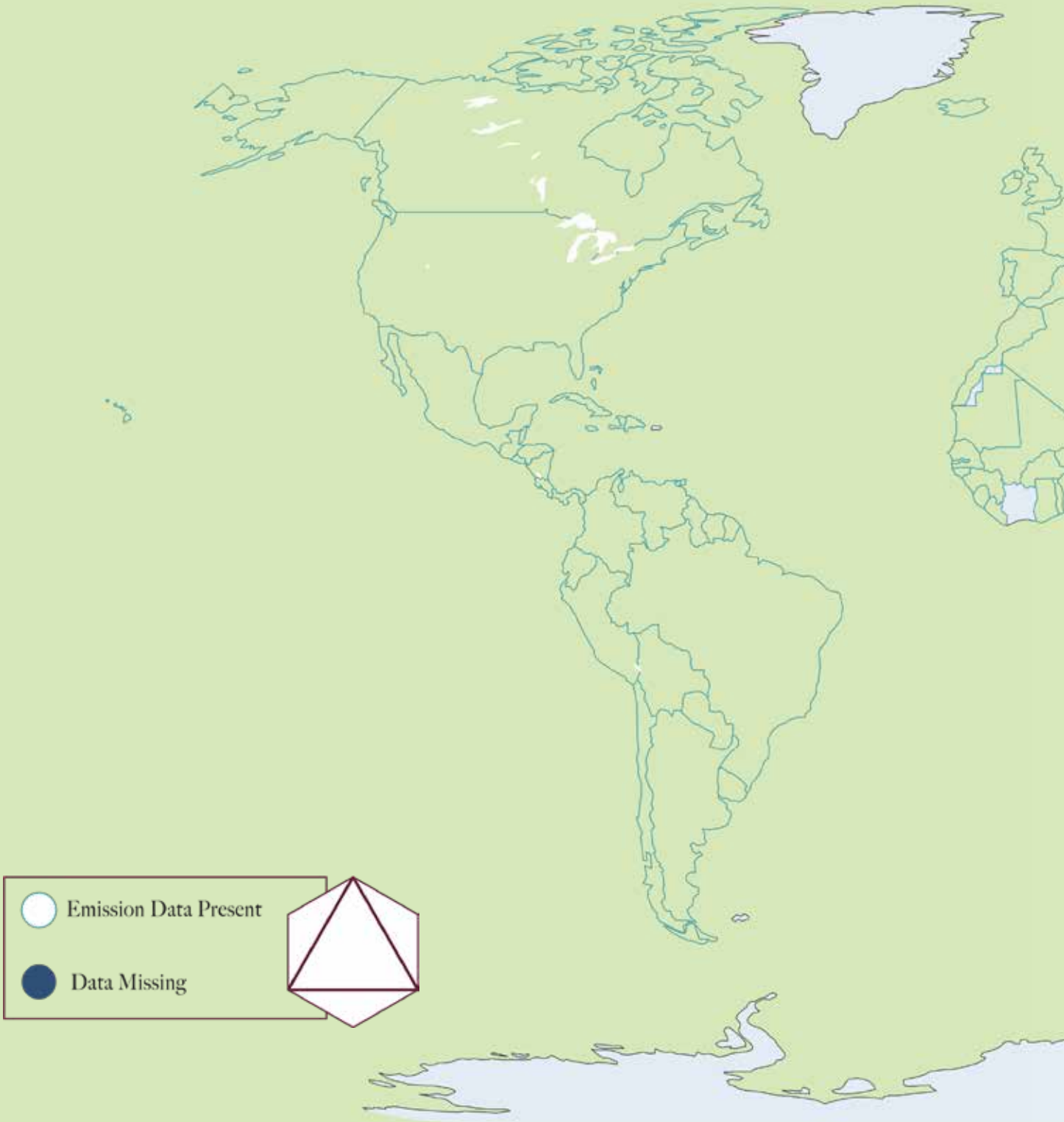
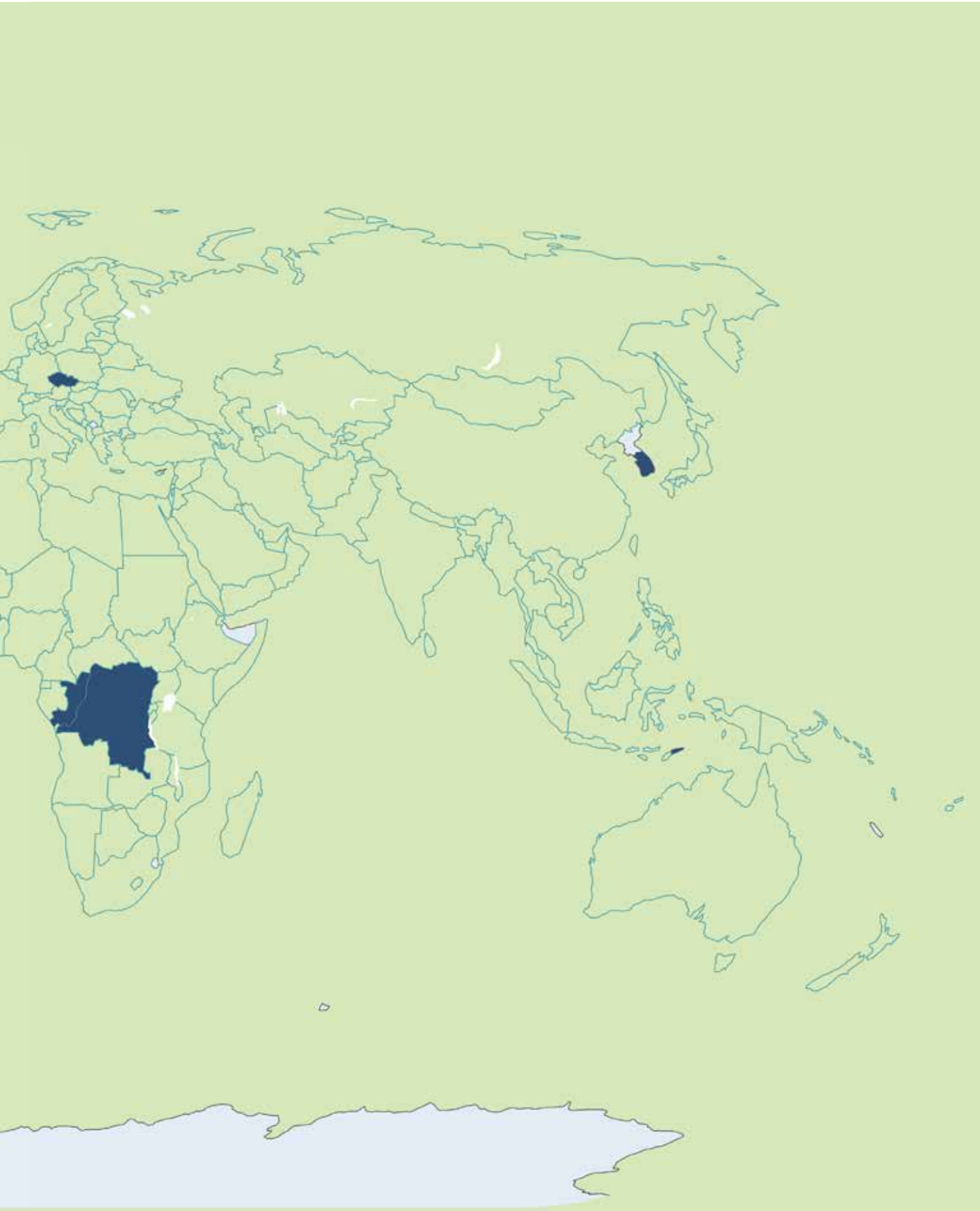


Figure 1.1



2020 Emissions Measurments

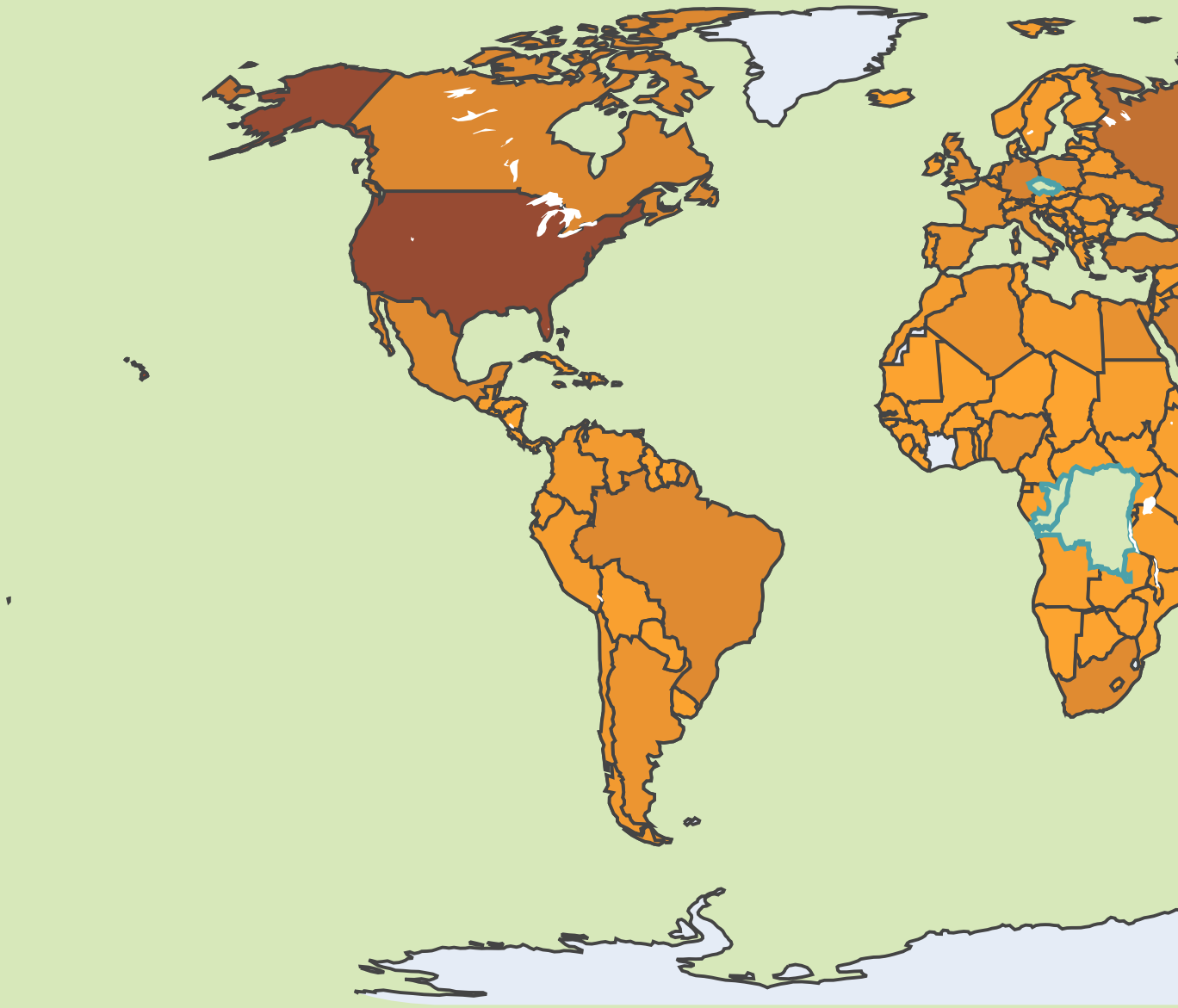
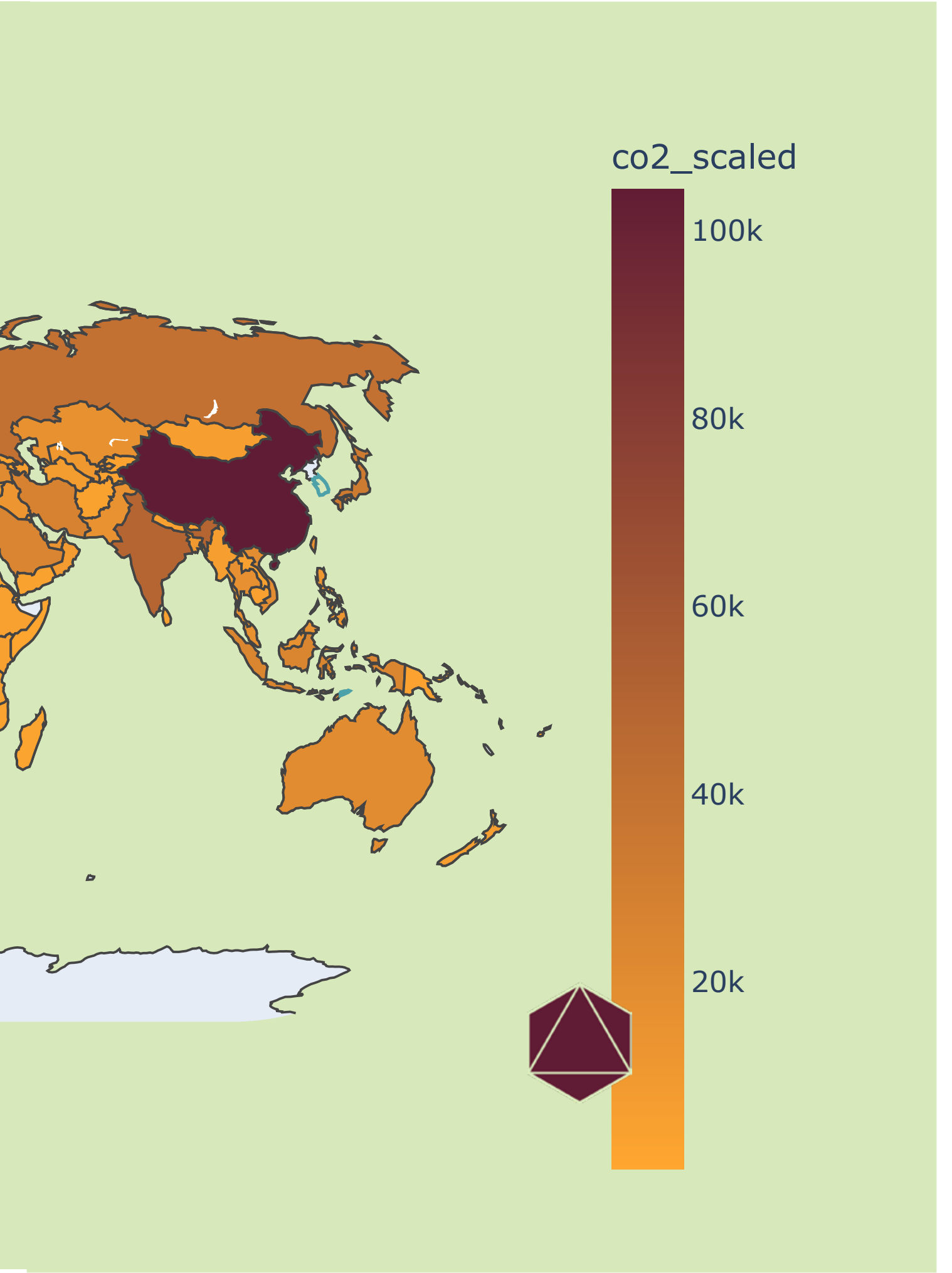


Figure 1.2


co2_scaled
100k
80k
60k
40k
20k

sourced for this project came from Our World Data.

II. Executive Summary

This report includes four stages: discovery, diagnosis, prediction, and analysis. Three models were used. In the first stage, it was crucial to figure out which type of machine learning model would be most helpful in making predictions. In order to do this, I used both a random forest regression model and a linear regression model. This was done to discover which algorithm performed more proficiently when tested. Both of these models were needed to confirm the hypothesis that unconventional features help inform emissions outcomes. Both regression models used in the discovery phase targeted change in emissions over time. Figure 2.1 visualizes the algorithmic strategy of a random forest model.

The random forest regression model was more successful. The second stage of the study
focused on diagnosing the relationships between key variables and the target feature by mapping out the specific impact of each variable on the success of the model. For example, a key feature was "military spending", which likely indicates a causal relationship. While the important variable "proportion of voting rights within the international bank for reconstruction and redevelopment" is likely a correlative one. Mapping out the percentage of impact of these features on the target variable was crucial in the later stages of our work.

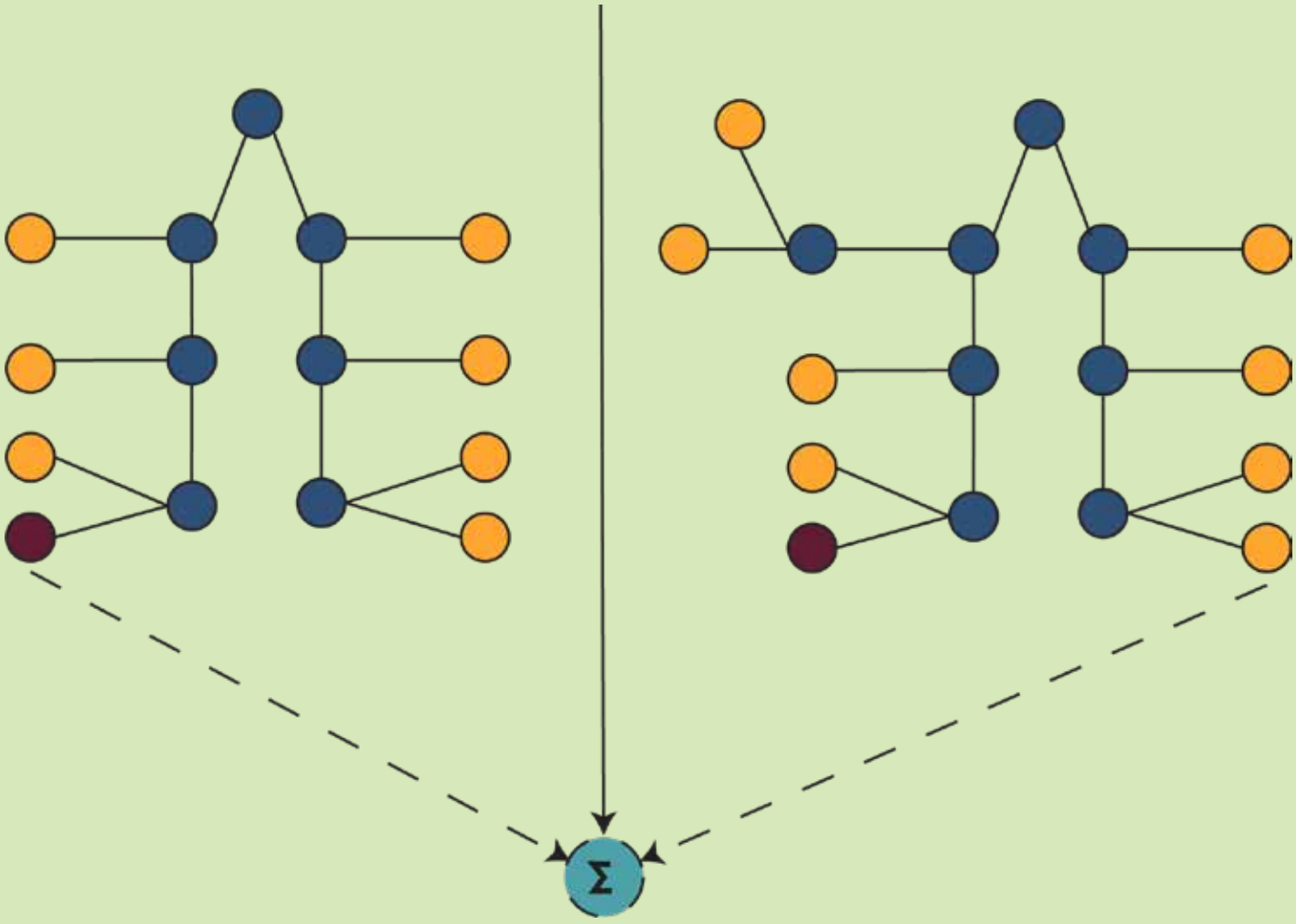


Figure 2.1

In the third stage of the project, I applied my findings to the final model. Using a random forest regression model, I wanted to apply my findings to fill in the gaps for countries missing data. This required careful cleaning and processing, and allowed for a lower, yet still significant, R2 score. During the predictive stage, the target variable was emissions in 2020 rather than the emissions changes over time. This was done because many countries that report to international organizations less frequently do not have as much historical data as others. Because the goal of the third model is to bridge the gaps in the data, the new target feature was selected. Finally, in the fourth stage, I wanted to identify key features once again, to see if there was any variation in feature importance. In addition, I analyzed my findings to better understand where the model succeeded and where it fell short. For example, the model did not successfully predict emissions for larger nations in the second model. That said, as the second model's purpose is to fill in data gaps, larger countries' predictions are not the measure of success in this context.

Figure 1.1 shows the emissions data pulled from ourworlddata.com. As you can see, it has data for the majority of the world, but is missing information from 11 countries, mostly in the global south. The nations missing are Cabo Verde, Democratic Republic of the Congo, the Republic of Congo, the Czech Republic, North and South Korea, Micronesia, Monaco, San Marino, Timor-Leste, and Vatican City. The Countries are in black.

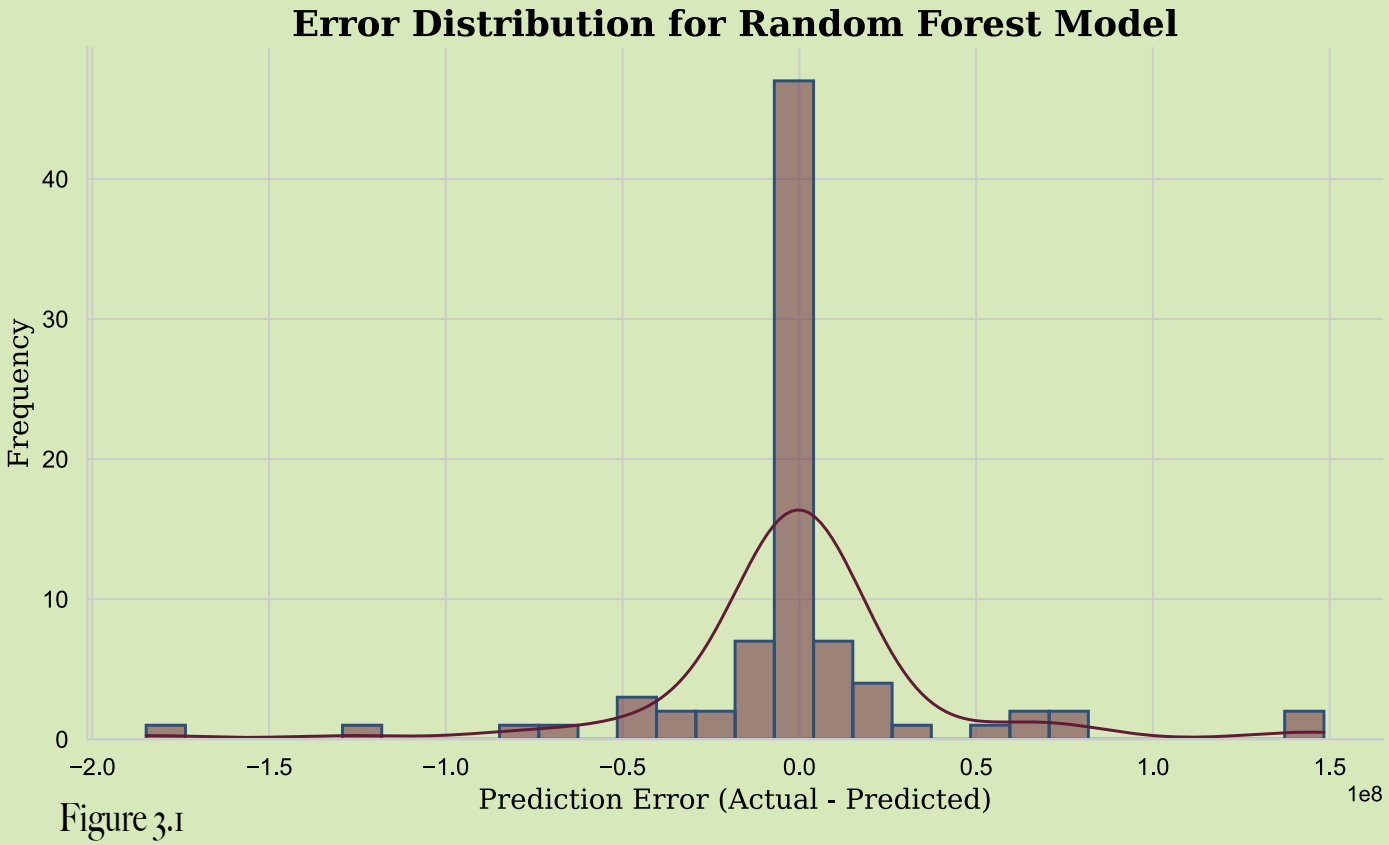


Figure 3.1

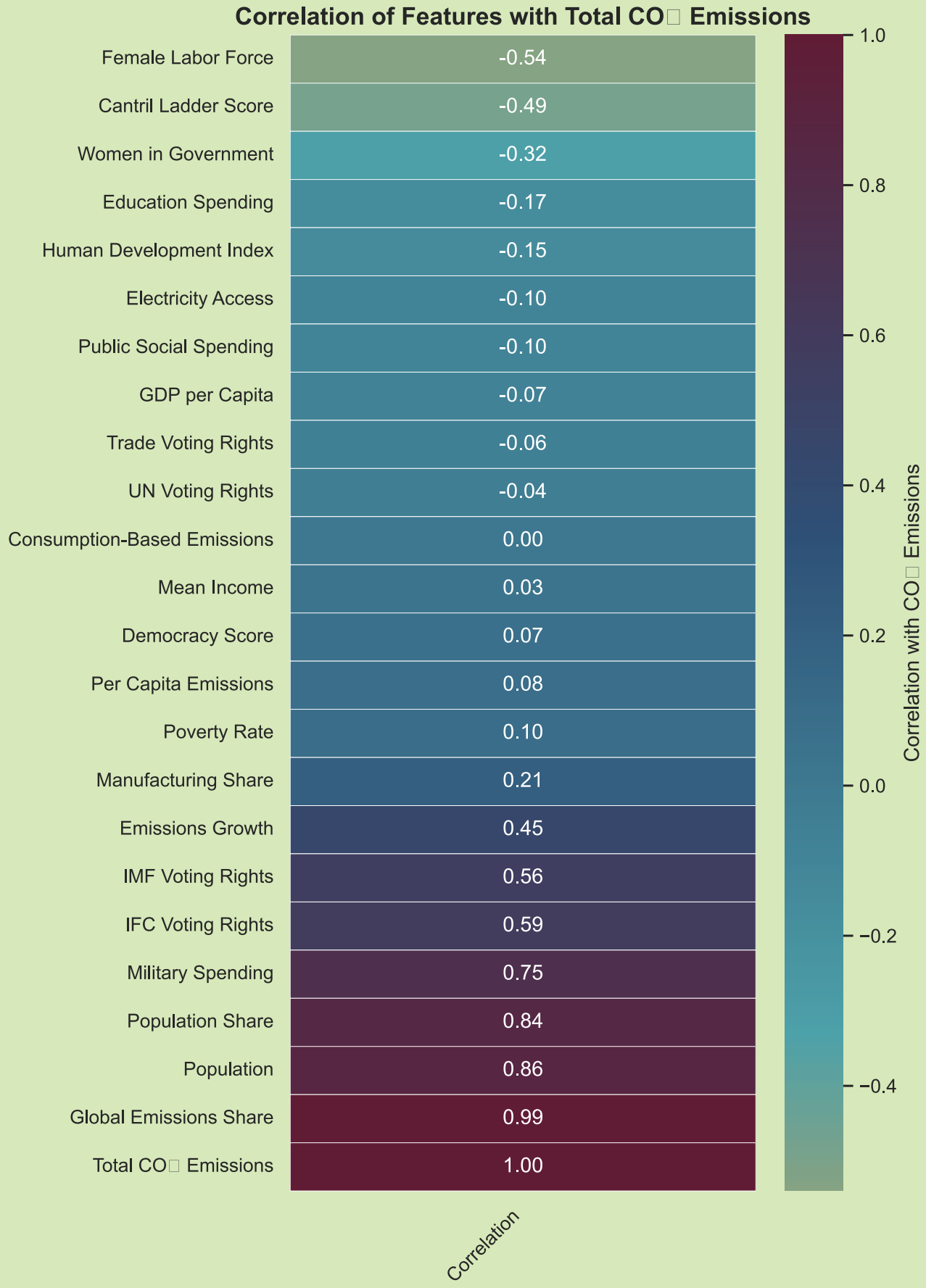

Correlation of Features with Total CO Emissions
Female Labor Force -0.54
Cantril Ladder Score -0.49
Women in Government -0.32
Education Spending -0.17
Human Development Index -0.15
Electricity Access -0.10
Public Social Spending -0.10
GDP per Capita -0.07
Trade Voting Rights -0.06
UN Voting Rights -0.04
Consumption-Based Emissions 0.00
Mean Income 0.03
Democracy Score 0.07
Per Capita Emissions 0.08
Poverty Rate 0.10
Manufacturing Share 0.21
Emissions Growth 0.45
IMF Voting Rights 0.56
IFC Voting Rights 0.59
Military Spending 0.75
Population Share 0.84
Population 0.86
Global Emissions Share 0.99
Total CO Emissions 1.00
Correlation
Correlation with CO Emissions
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Figure 4.1

III. Stage 1: Discovery

As mentioned above, the first stage of the study is to discover which algorithm would be most successful at finding emission changes between 1995 and 2020. Therefore, both a Random Forest and a Regression model are used. Knowing that our goal was to create as successful a model as possible, the worldwide data was filtered to include countries with at least 70% data presence in the dataframe. After cleaning the data and dropping duplicates and overly sparse columns, both models were trained and tested.

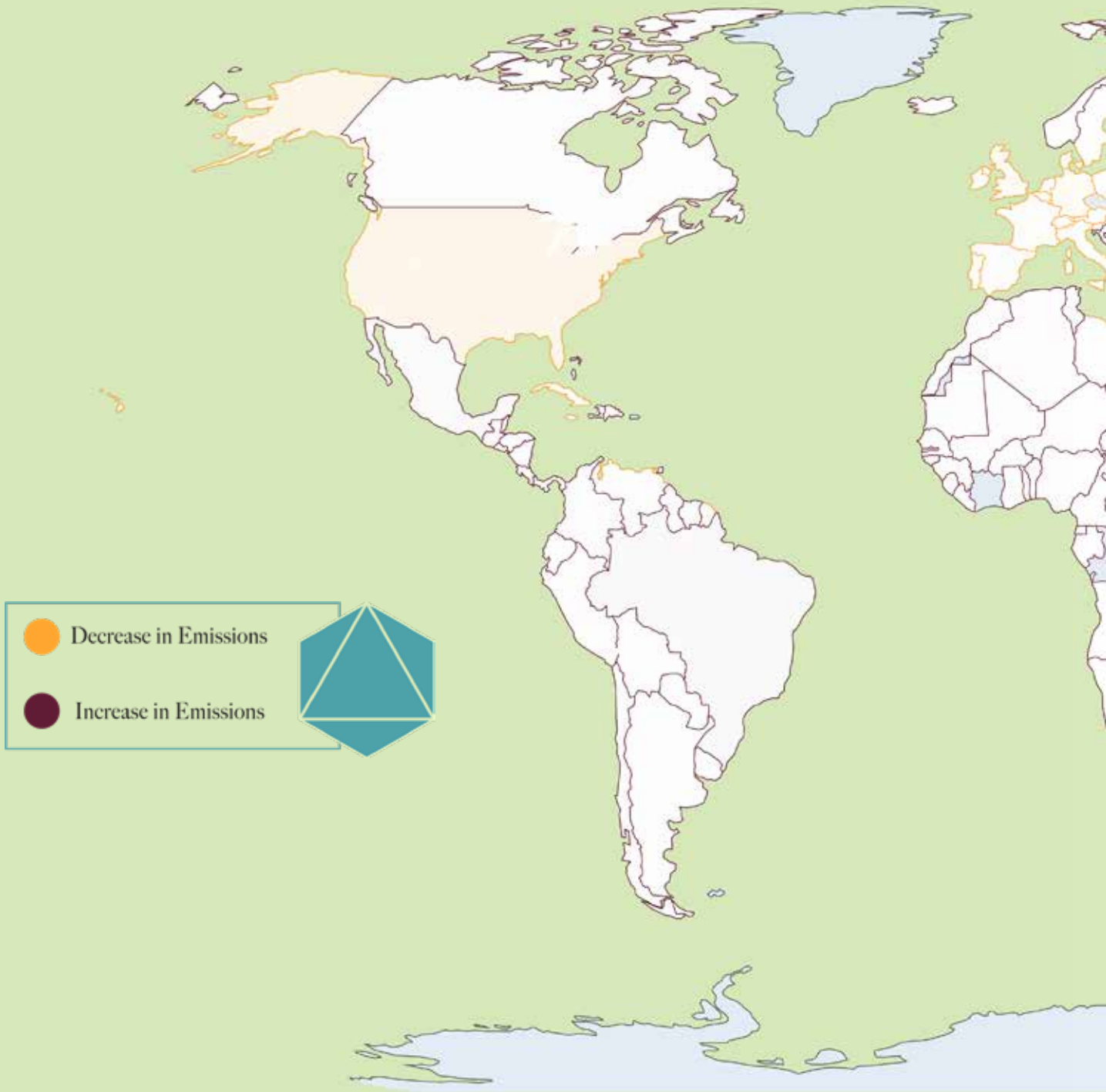
The Linear Regression model achieved an R^2 score of 0.91, while the Random Forest model achieved 0.99, indicating a near-perfect fit. The Random Forest Regressor was especially effective at capturing nonlinear relationships and outperformed the linear model in nearly every case.

A feature importance analysis revealed that population size, military expenditure, Cantril ladder score (national life satisfaction), and voting rights in multilateral institutions were among the strongest predictors. Interestingly, countries with larger populations and growing economies tended to see emissions rise, but nations with stronger democracies and higher education levels often saw decreases over time.

To make sure the model worked, I tested it using standard evaluation metrics: R^2, MAE, and RMSE. The R^2 score tells us how much of the variation in emissions the

Linear Regression Model	Random Forest Model
R^2 = .902	R^2 = = .986
Mean Absolute Error ~85.7 Million	Mean Absolute Error ~19.5 Million

Change in Emissions Between 1995-2020



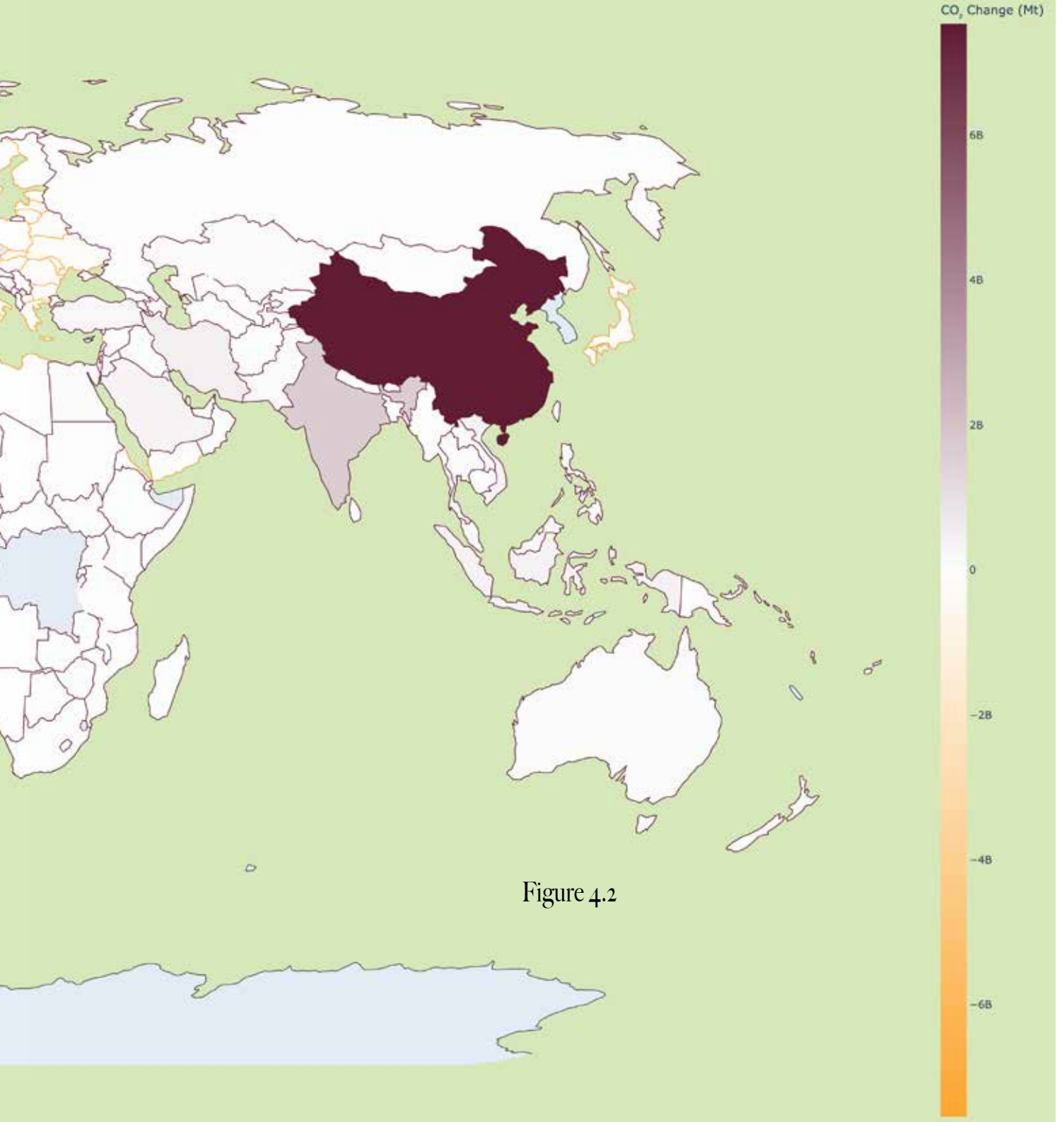
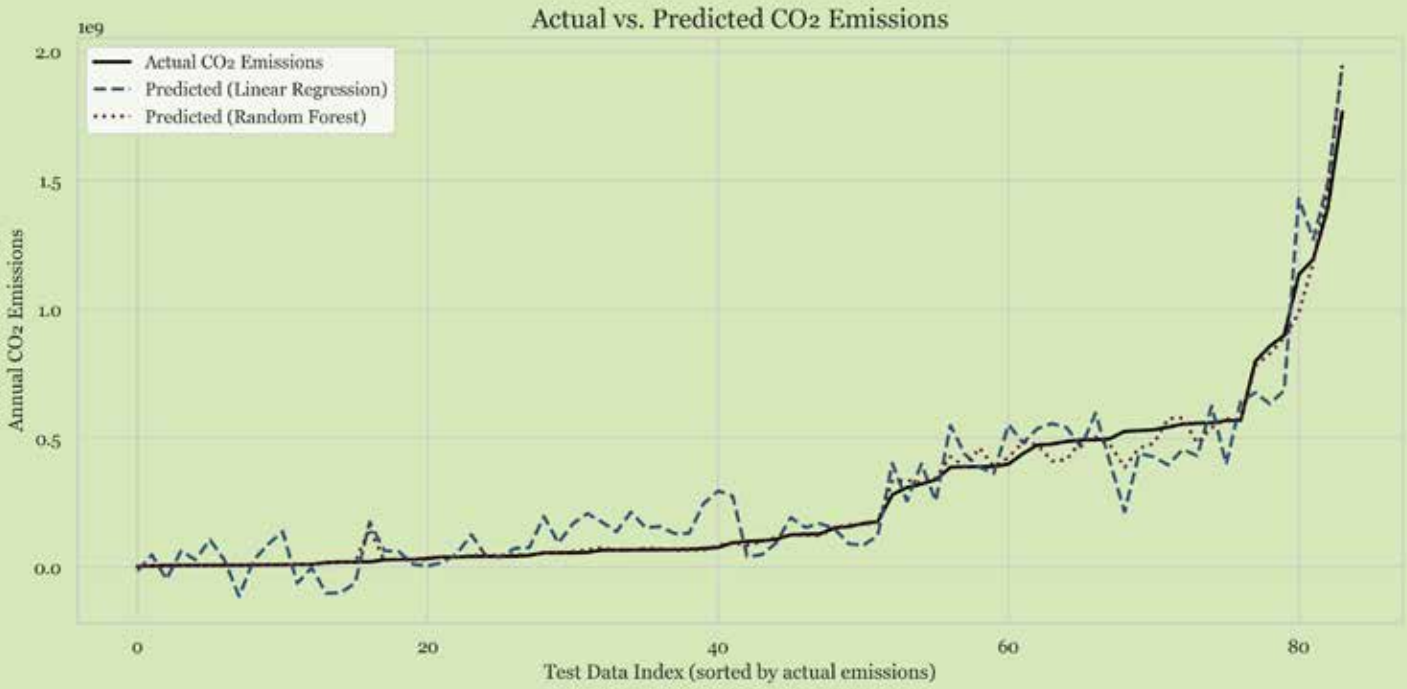


Figure 4.2



Firgure 4.3

model can explain. The Random Forest model scored a 0.986, which means it captured nearly all of the meaningful patterns in the data.

MAE and RMSE measure how far off the predictions were from reality. The lower these numbers are, the better. Random Forest outperformed linear regression across the board, making more accurate predictions and avoiding big errors. In short, the model was both strong and stable, which makes it reliable for helping us estimate emissions in countries where data is missing or incomplete. Figure 3.1 shows the results of MAE.

IV. Stage 2: Diagnosis

Figure 4.1 shows which social, political, and economic variables have the strongest relationships with total CO emissions. As expected, factors like global emissions share, overall population, and military spending show a strong positive correlation with emissions. That said, the socio-political features do play a role. Bigger and more powerful countries tend to emmit more, with India and China seeing some of the largest increases over time. The United States, many countries in Europe, Japan, Jordan, Lebanon, Tunisia, Syria, Zambia, Botswana, Eswatini, Namibia, and Venezuela all had a reduction in emissions. The largest Reduction occurred

in the United States, followed by European Countries. Figure 4.2 shows the results mapped across the world.

Female labor force participation, women in government, education spending, and even subjective well-being (measured through the Cantril ladder) show negative correlations with emissions. This suggests that countries investing in equity, public services, and general quality of life tend to produce fewer emissions, or at least grow them more slowly. On the flip side, military spending and voting power in international finance institutions are strongly linked to emissions growth, hinting at how structural power and resource allocation shape environmental outcomes.

While correlation doesn't equal causation, these results point toward a more interdisciplinary view of climate impact. One that goes beyond GDP and population, and considers how systems of governance, inequality, and development strategy intersect with carbon output.

Figure 4.3 shows the difference between the actual CO_2 emissions for countries in the dataset and the predicted emissions produced by two models: a linear regression model and a random forest model. The solid black line is the real data. The blue dashed line is the linear regression prediction, and the burgundy dotted line is from the random forest model.

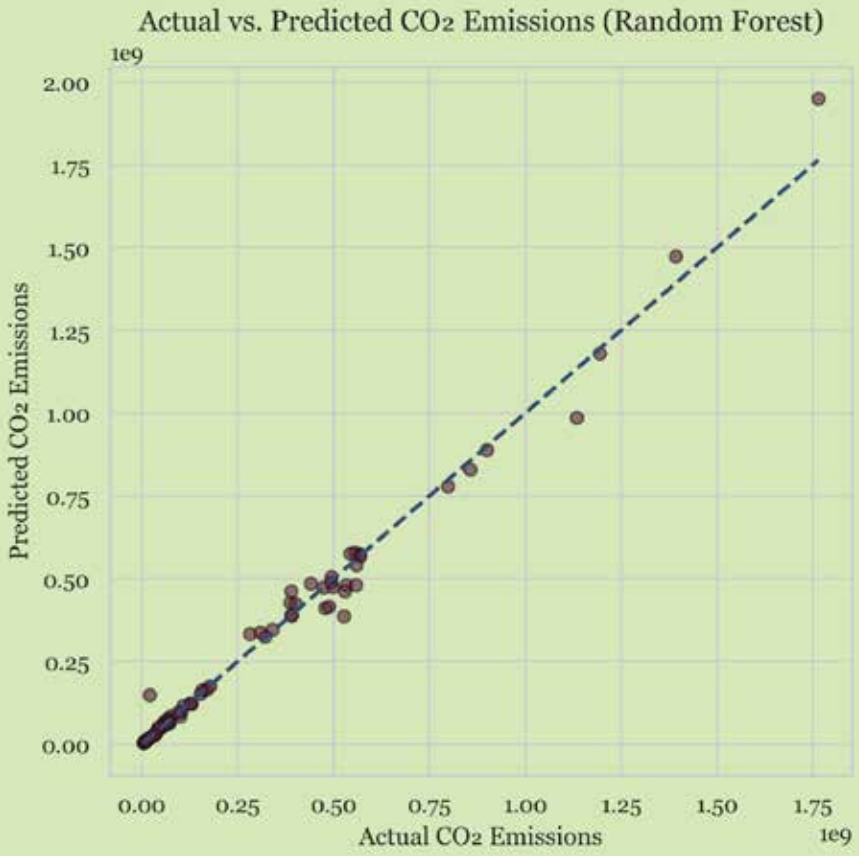


Figure 4.4

What we see here is that the random forest model (burgundy) hugs the real data much more closely than the linear model (blue), which bounces around a lot, especially in countries with lower emissions. This tells us that random forest is much better at picking up the non-linear patterns in the data, especially as emissions start to increase. The biggest deviations happen toward the right side of the chart—these are the highest-emitting countries, and while there's some variation, the model still does a solid job of tracking the general trend.

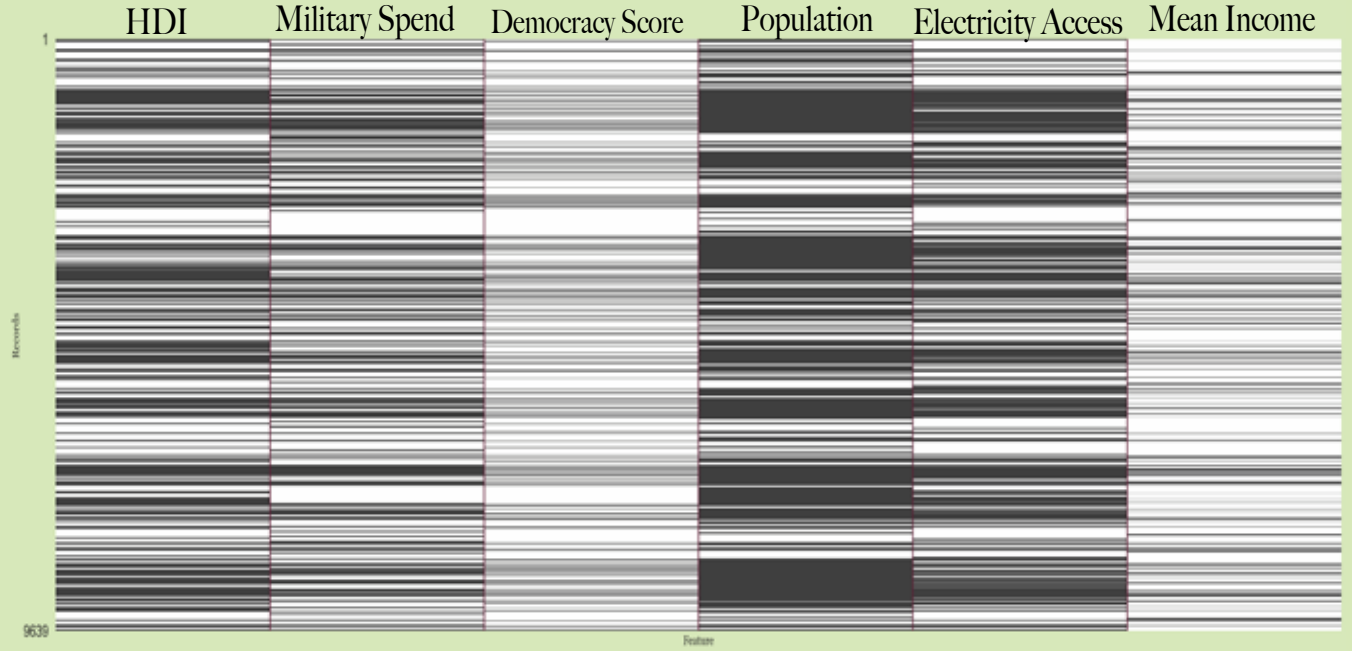
In short, the random forest model gets us pretty close to the real emissions values in most cases, especially for countries with mid-range emissions. It's not perfect at the extremes, but it's a major step forward in using limited socio-political data to make decent emissions estimates.

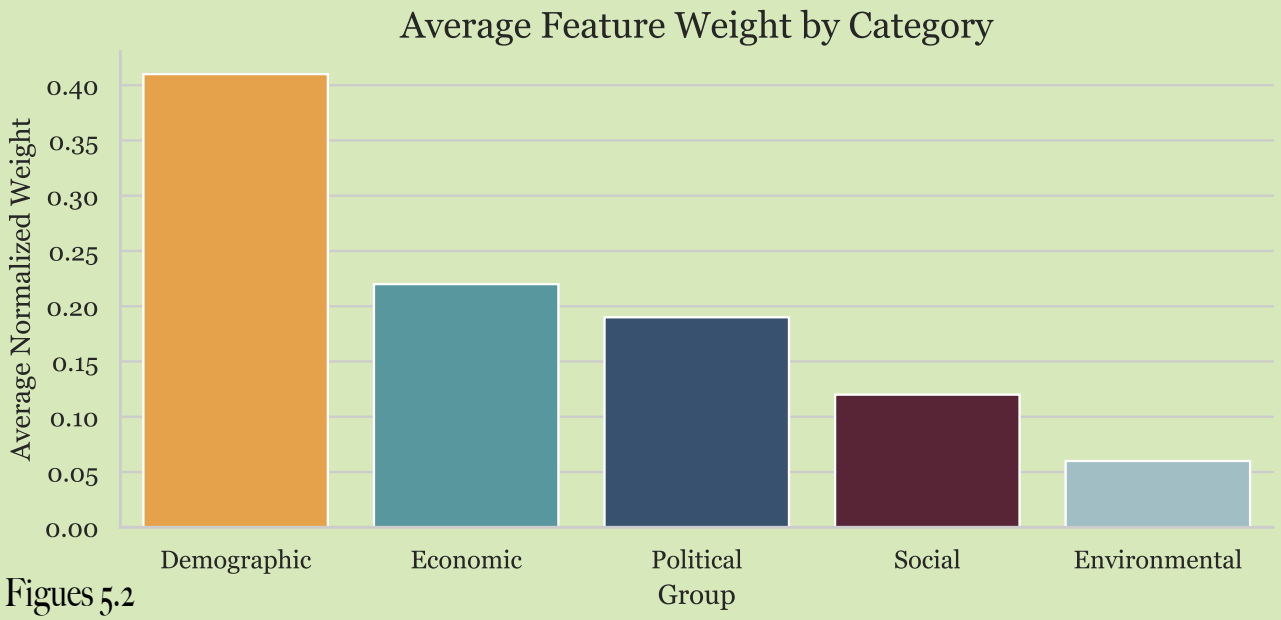
Figure 4.4 compares the actual CO_2 emissions to the emissions predicted by the Random Forest model. Each point represents a country-year observation, with the dashed line serving as a reference for perfect prediction, where the model's output exactly matches observed values. The clustering of points around the line shows that the model performs reasonably well, especially in the low-to-mid emissions range. However, we do see some consistent over- and under-prediction in the higher-emitting countries, where the model tends to slightly overestimate emissions beyond 1.5 billion tons. This suggests that while the model is effective at capturing the overall structure of emissions patterns, it smooths out some of the variance at the extreme ends. Still, the general alignment in the center of the plot indicates that the model is capturing core drivers of emissions fairly reliably.

Model including only countries with emissions data	Model with emissions data missing
$R^2 = .57$	$R^2 = .42$
Mean Squared Error = 2.75×10^{18}	Mean Squared Error = 2.61×10^{17}

Figure 5

Data Completeness by Feature





Figues 5.2

V. Stage 3: Predictive Methodology

To adapt and apply the previous successful random forest model to nations with missing data, the data needed to be re-filtered, cleaned, filled in, and tested. First, the data was filtered so that only countries with up to 70% of the data were missing. Next, the columns were filtered so only columns with more than 30% of the data were present. This was done because columns with less than 30% of the data harmed the model too severely. Figure 5.1 visualizes data presence for certain key features.

In order to apply the random forest model, the empty data needed to be assigned values. Instead of broadly filling in with medians from the entire dataset, I created buckets to find medians that expressed variation in the dataframe more precisely. To ensure a more accurate outcome, the countries were grouped by GDP per capita, region, and population. These categories were chosen because the

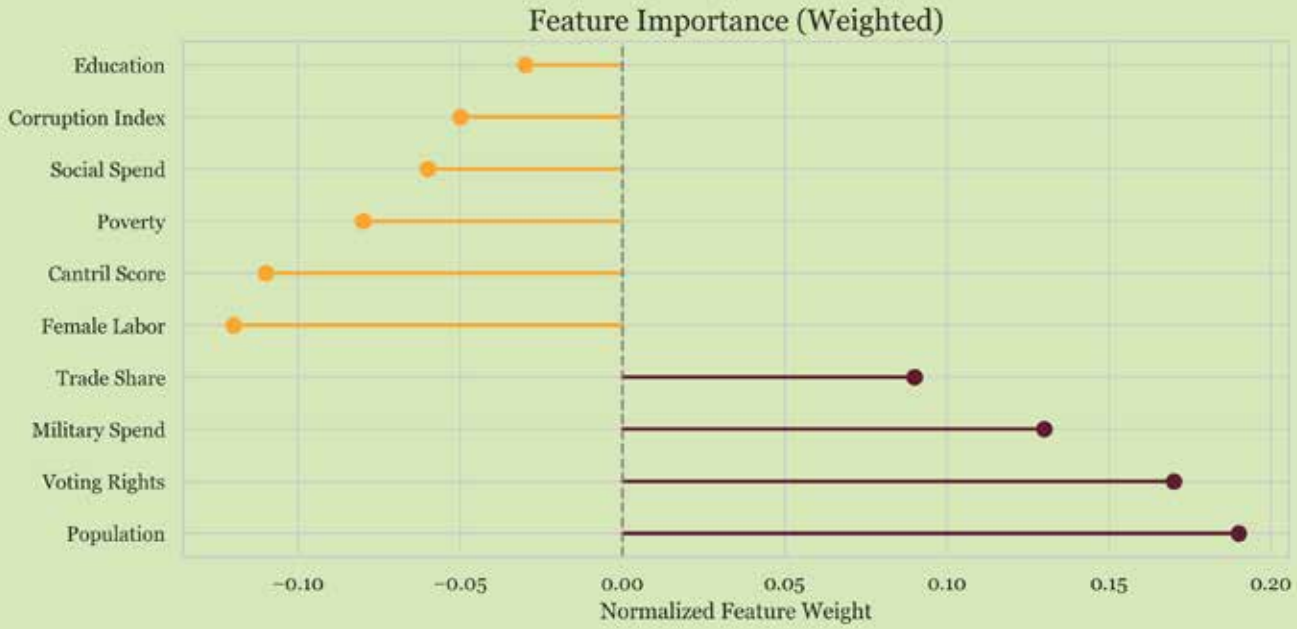
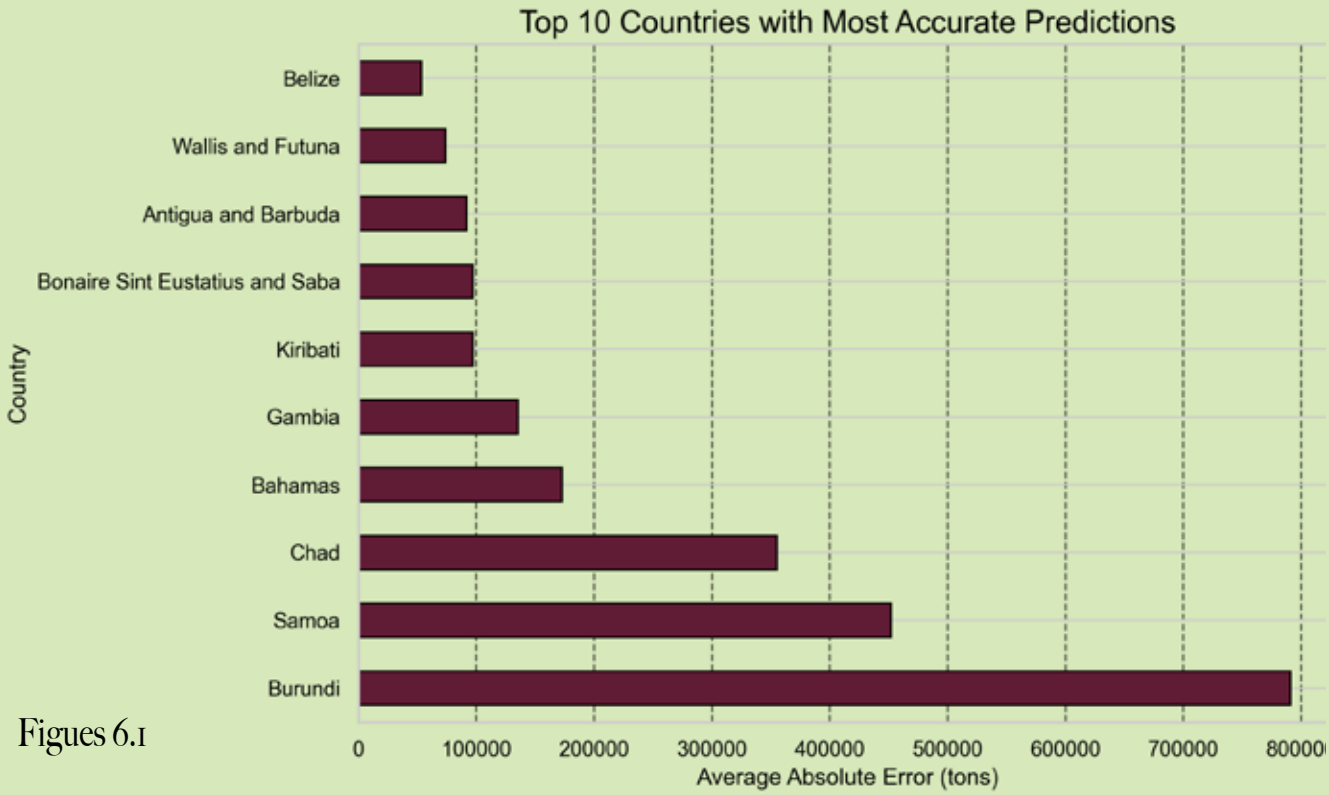


Figure 5.3



Figues 6.1

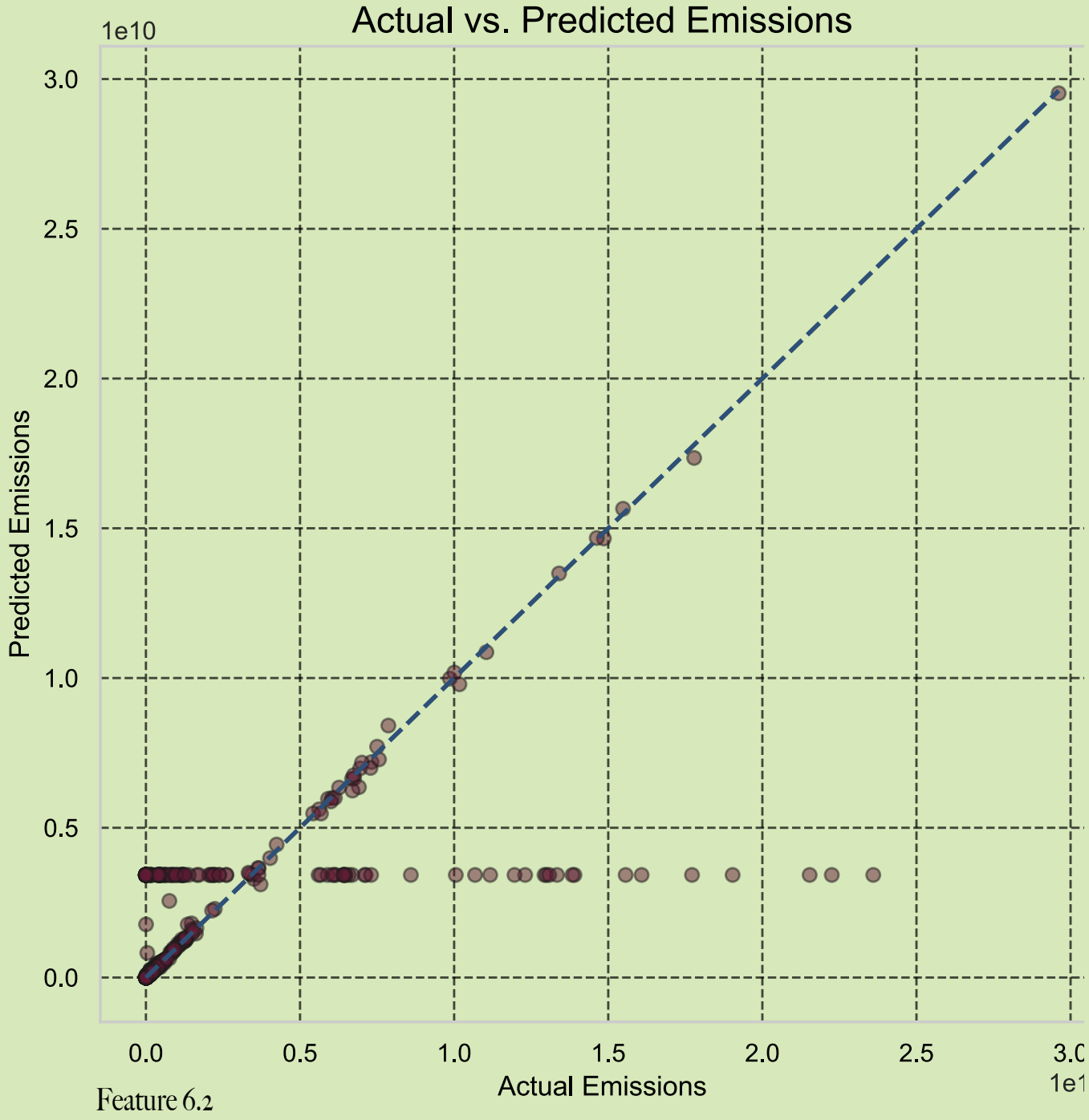
first model demonstrated that these have a significant impact on emissions rates. The median of each feature, once grouped based on the bucket, was applied to fill the missing values.

Next, the study used the findings in stage II to apply weights to features. This way, the model can use each feature according to its importance to predict results. Figure 5.2 highlights the weights applied to different groups of features. Figure 5.3 highlights the top ten specific feature weights that were applied.

I ran this model twice. The first time, I used only countries where emissions rates were known, and in the second, I used a mix of countries where countries with and without emissions data reported were present. The first version of the model scored $R^2 = 0.57$, and the second version, where not all target values were known, scored $R^2 = 0.42$.

VI. Stage Four: Analysis and Findings

While not as precise as the first model, the model in the third stage still captures meaningful trends and confirms that social and



Feature 6.2

political indicators, like access to electricity, women in government, or corruption scores, can hold predictive power when it comes to emissions.

Figure 6.1 highlights the ten countries where the model performed with the highest accuracy, measured by the lowest average absolute error in predicted CO emissions. Unsurprisingly, many of these are small island nations or countries with relatively low emissions levels—places like Belize, Kiribati, and Wallis and Futuna. Their emissions profiles are often more stable and less subject to the dramatic fluctuations seen in industrialized or resource-extractive economies, which helps explain the model's precision in these cases. That said, this also reinforces the importance of applying this model to smaller nations, where reporting gaps are common but the margin for error in climate planning is slim.

Figure 6.2 offers another look at the

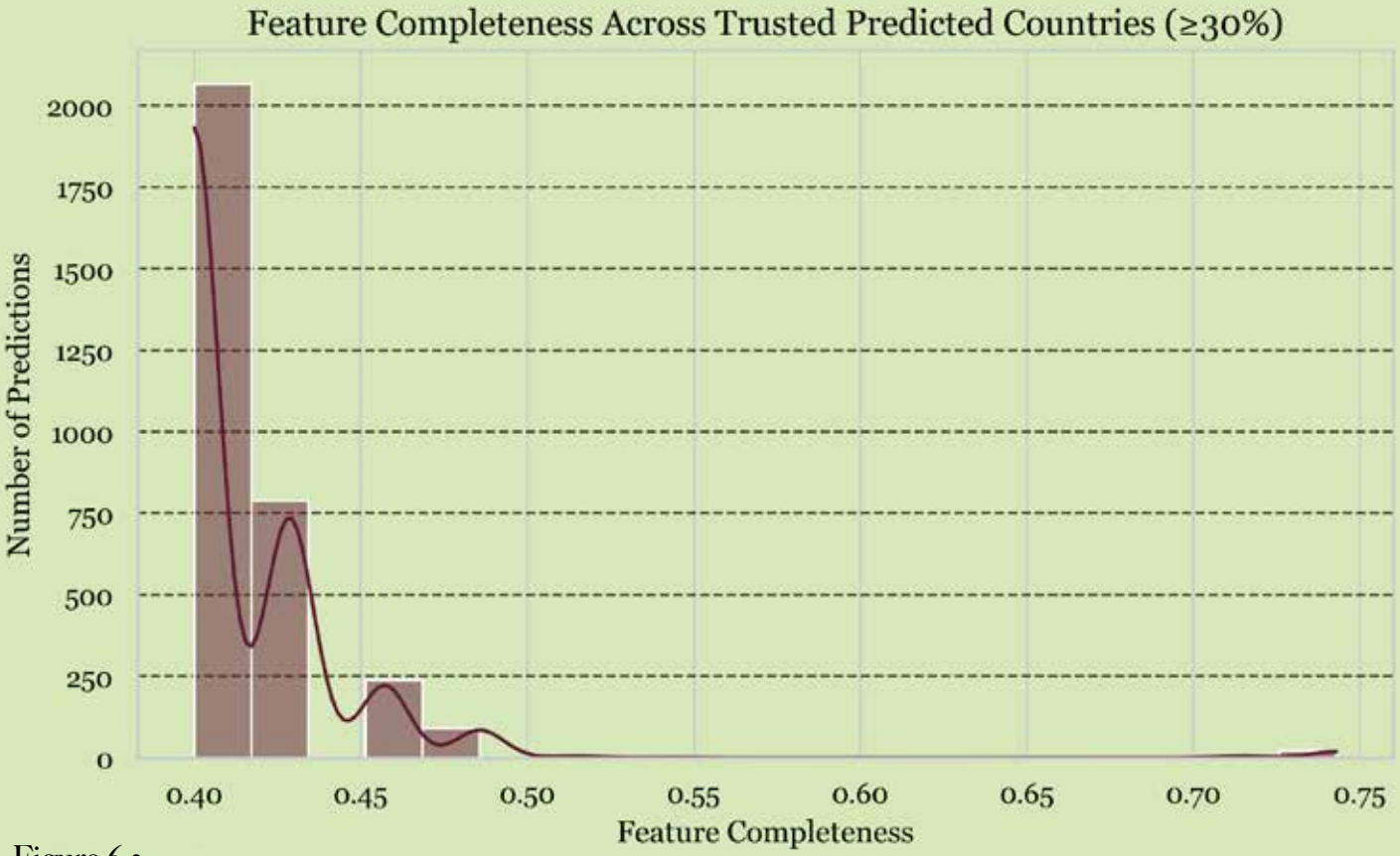
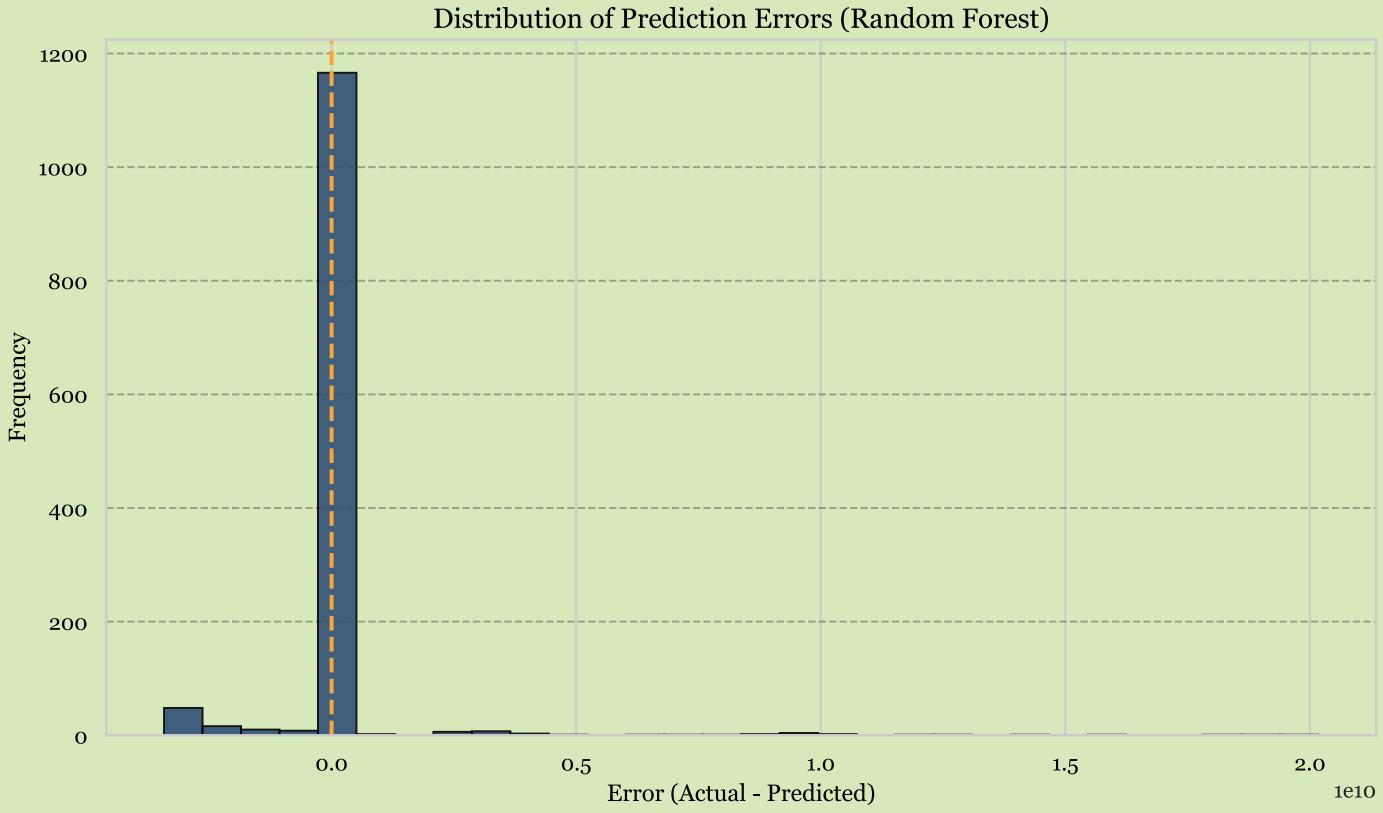
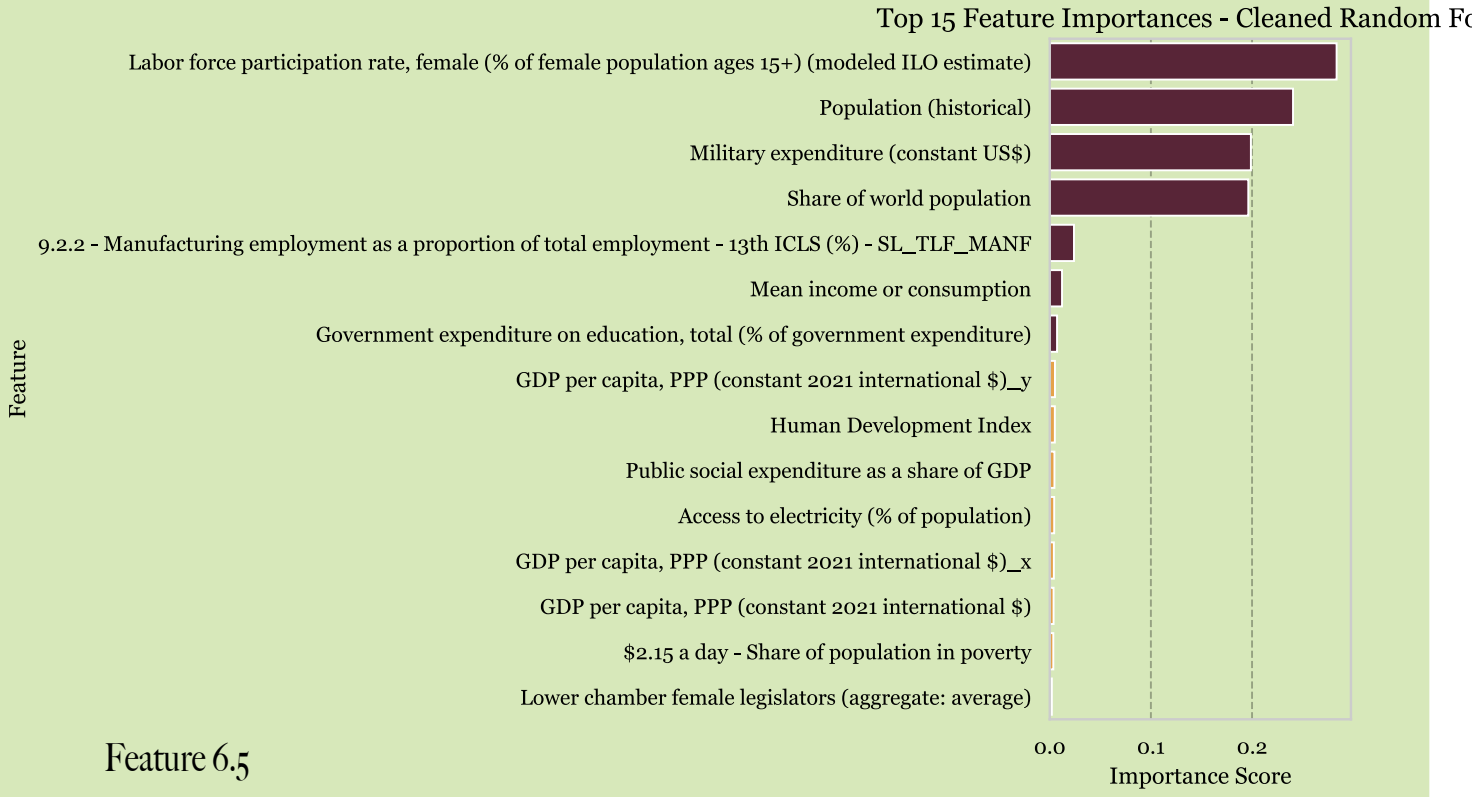


Figure 6.3



Feature 6.4


Top 15 Feature Importances - Cleaned Random Fo
Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate)
Population (historical)
Military expenditure (constant US$)
Share of world population
9.2.2 - Manufacturing employment as a proportion of total employment - 13th ICLS (%) - SL_TLF_MANF
Mean income or consumption
Government expenditure on education, total (% of government expenditure)
GDP per capita, PPP (constant 2021 international $)_y
Human Development Index
Public social expenditure as a share of GDP
Access to electricity (% of population)
GDP per capita, PPP (constant 2021 international $)_x
GDP per capita, PPP (constant 2021 international $)
$2.15 a day - Share of population in poverty
Lower chamber female legislators (aggregate: average)
Feature
0.0
0.1
0.2
Importance Score

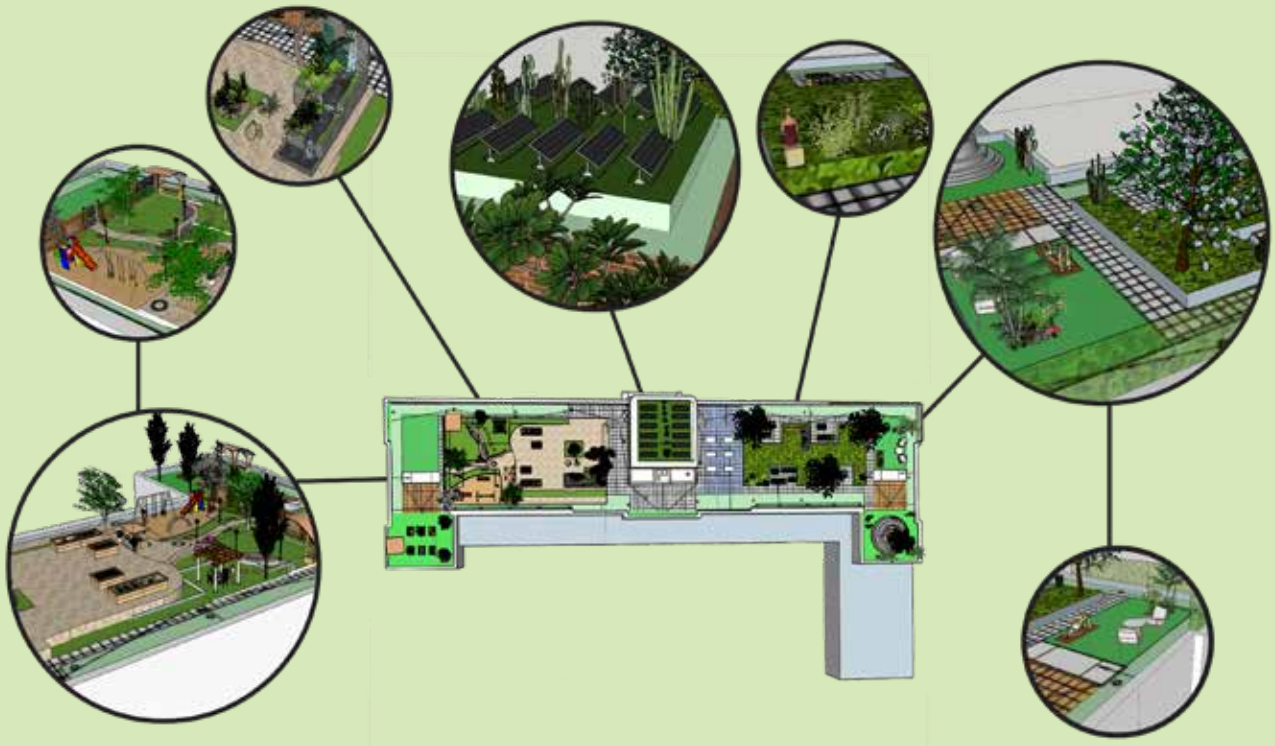
Feature 6.5

model's predictive performance across the full dataset, showing the relationship between actual and predicted emissions. Points clustered tightly along the diagonal indicate strong prediction accuracy. However, we also see horizontal bands at zero predicted emissions, which likely correspond to cases where the model struggled due to extremely sparse input data. These flatlines serve as a visual reminder that while the model is robust overall, certain countries—especially those with poor data coverage—still require additional verification or alternative modeling strategies. Together, these figures point to both the power and the limits of machine learning-based imputation in climate research.

Figure 6.3 shows how much data was available for each country that received a predicted emissions value. All countries in this figure had at least 30% of their data present—a threshold we set to ensure the model had enough context to make meaningful predictions. What's striking is how steeply the distribution drops after that. Most countries barely cross the threshold, hovering between 40% and 45% completeness. This tells us that while the model is helping us cover gaps, those gaps are still wide. In other words, the countries that most need our attention are the ones with the most missing information, making this modeling approach all the more critical.

Figure 6.4 plots the distribution of prediction errors—how far off the model was from actual emissions values. The tight cluster around zero suggests that the model was reasonably calibrated: most of the time, the predictions were close to the real numbers. That said, we do see a long right tail, meaning a few cases had very large errors. These are likely countries with more volatile emissions patterns or poor feature completeness, and they serve as important flags for where model performance should be interpreted with extra care.

Figure 6.5 displays the top 15 features that held the most weight in the final, cleaned Random Forest model. Perhaps the most compelling takeaway here is how dominant social indicators became once we removed direct



emissions leakage. Labor force participation by women, military expenditure, and historical population were among the strongest predictors, surpassing metrics like GDP or energy usage. This underscores one of the core insights of the project: emissions aren't just a matter of industry or wealth. They're deeply embedded in governance, gender equity, and the social fabric of a country.

This model is useful because it fills in important gaps in global datasets. While it may slightly underestimate emissions for larger, high-emitting countries, those countries usually report anyway. What matters is its ability to help us estimate what's happening in smaller or underreported places—and use that to inform international policy and climate strategy.

VII. Implications

In the heat of the Anthropocene, humanity faces a seemingly insurmountable challenge. Armed with the full force of our technological capabilities and the promise of exponential informational growth, we are at a collective crossroads: how can we better use the work we've done to build a brighter future?

With access to an unprecedented informational library, we can now make more informed and impactful decisions. This work takes a first step toward understanding how to harness data to uncover the social and political complexity of an increasingly interdependent, yet divided, world. Climate change is a global issue, but cities play a critical role in shaping the response. Exploring global patterns and their connections to liberalism, feminism, economics, militarism, and environmental shifts helps bring greater clarity to those making decisions. In the next phase, the macro lens used here can be refined to help fill in the gaps at the micro, urban level.

How can governments with limited global influence but strong local political capacity apply these insights? If cities that lack the resources to monitor building emissions or adapt infrastructure are given access to the broader picture, how might governance and planning evolve?

Artificial intelligence will undoubtedly have profound implications, just as technological shifts always have. As it becomes integrated into the power structures shaping our future, those of us at the intersection of policy and technology must be willing to think critically about its risks—and creatively about its potential. AI could be used as both a tool and a weapon. If applied correctly, it may prove remarkably powerful in the fight against the greatest crisis in human history. It is up to us to employ it properly.

VIII. Conclusion

This work demonstrates how machine learning can fill critical gaps in global emissions reporting by drawing from a wide range of social, political, and demographic indicators. While the first model proved highly accurate on countries with full data, the second model extended that predictive power to underreported regions, capturing important patterns in places often excluded from climate assessments. These results don't just improve data coverage; they shift how we understand what drives emissions in the first place. The strength of features like

IX. References:

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Crutzen, P. J., & Stoermer, E. F. (2000). The Anthropocene. International Geosphere-Biosphere Programme (IGBP) Newsletter, 41, 17–18.

C40 Cities & Arup. (2019). The future of urban consumption in a 1.5°C world. C40 Cities Climate Leadership Group. https://www.c40.org/researches/future-of-urban-consumption

Ritchie, H., Roser, M., & Rosado, P. (2020). Our World in Data. Global Change Data Lab. https://ourworldindata.org/

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., ... & Bengio, Y. (2019). Tackling climate change with machine learning. arXiv. https://arxiv.org/abs/1906.05433

Transparency International. (2018). Corruption Perceptions Index 2018. https://www.transparency.org/en/cpi/2018

United Nations Human Settlements Programme (UN-Habitat). (2020). World cities report 2020: The value of sustainable urbanization. https://unhabitat.org/wcr/

United Nations Statistics Division. (n.d.). UNSD SDG Global Database. https://unstats.un.org/sdgs/indicators/database/

World Bank. (n.d.). World Development Indicators. https://databank.worldbank.org/source/world-development-indicators

Freedom House. (n.d.). Freedom in the World Dataset. https://freedomhouse.org/report/freedom-