Milestone #2 Report

Sai Abhishek Gangineni, Robin Godinho, Marta Gonzalez
INST 737

Updates to Dataset

Following our efforts in Milestone 1, our team made several updates and changes to our chosen Greenhouse Gas Emissions (GHG) dataset. The original format of the dataset had the years from 2000-2015 as the columns while the rows were the features for each country and their changing values over the years. This format allowed for some easier visual indications of changes across times for features between the different countries, but the machine learning (ML) classifiers implemented in Milestone 2 required a different setup. We manually, through Excel, inverted our dataset horizontally so that the columns are now the various features that we are studying for each country while the rows are all the values for those features for all countries between 2000-2015. This change allows for the ML classifiers to be able to conduct a feature-dependent analysis to train models that can predict GHG Emissions based on all of our independent variables.

Another update we made to our dataset was to include an additional dataset/feature which is the total population of all the countries between the years 2000-2015. This data feature was added to normalize some of the data columns we have such as GHG Emissions and Energy Consumption because these values are very dependent on the individual population of each country. To allow for direct comparisons between countries for these features, we had to add the population as an additional column before dividing the features by the population to attain a normalized 'per capita' variable that could be directly compared between countries. Unfortunately, we were only able to find the population data for 175 countries which necessitated us to get rid of the additional countries that we had in the original dataset since we didn't have the population data for them. After making all of these changes, our final dataset for Milestone 2 is now populated with data for 175 countries between 2000-2015 for 16 data features.

In this report, we will go through our efforts to utilize this updated dataset to train and test various ML classifiers to be able to predict the dependent variable that is GHG Emissions. Beginning with Linear and Logistic Regression classifiers, we continued with Naive Bayes, Decision Trees, and a Random Forest classifier to create several models which we will cover the results and do a comparative analysis for in this report.

Linear Regressions

The first classifier we created was a Linear Regression model. For each independent variable in our dataset, we computed a linear regression with respect to the dependent feature:

Intercepts
Fossil fuel energy consumption (% of total): **6.947608057743388**
GDP per capita, PPP (current international $): **4.986526566808983**
Agricultural land (% of land area): **10.646425971767691**
Forest area (% of land area): **10.464348457309114**
Urban population (% of total population): **2.8274293850132715**
Access to Clean Fuels and Technologies for cooking (% of total population): **-11.710429859262101**
Access to electricity (% of rural population with access): **-81.6942422031335**
Access to electricity (% of total population): **-350.3896981728974**
Access to electricity (% of urban population with access): **-1434.597319761065**
Energy intensity level of primary energy (MJ/PPP): **6.704224639428859**
Renewable electricity share of total electricity output (%): **10.479251091207262**
Renewable energy share of TFEC (%): **10.59452992625741**
Population, total: **9.210714099364171**
Total electricity output per capita (kWh): **8.749095424495843**
Total final energy consumption per Capita (TFEC) (MJ): **4.133896600783673**

Coefficients
Fossil fuel energy consumption (% of total): **0.02877559**
GDP per capita, PPP (current international $): **0.00013833**
Agricultural land (% of land area): **-0.01840954**
Forest area (% of land area): **-0.01560614**
Urban population (% of total population): **0.08863216**
Access to Clean Fuels and Technologies for cooking (% of total population): **0.2182398**
Access to electricity (% of rural population with access): **0.91751862**
Access to electricity (% of total population): **3.60520098**
Access to electricity (% of urban population with access): **14.44745995**
Energy intensity level of primary energy (MJ/PPP): **0.63793238**
Renewable electricity share of total electricity output (%): **-0.02539886**
Renewable energy share of TFEC (%): **-0.04695938**
Population, total: **1.55332084e-08**
Total electricity output per capita (kWh): **0.00011363**
Total final energy consumption per Capita (TFEC) (MJ): **4.61721605e-05**

Predictive Features

Fossil fuel energy consumption (% of total): **Weak**
GDP per capita, PPP (current international $): **Moderately strong**
Agricultural land (% of land area): **Weak**
Forest area (% of land area): **Weak**
Urban population (% of total population): **Weak**
Access to Clean Fuels and Technologies for cooking (% of total population): **Weak**
Access to electricity (% of rural population with access): **Weak**
Access to electricity (% of total population): **Weak**
Access to electricity (% of urban population with access): **Weak**
Energy intensity level of primary energy (MJ/PPP): **Weak**
Renewable electricity share of total electricity output (%): **Weak**
Renewable energy share of TFEC (%): **Weak**
Population, total: **Moderately Weak**
Total electricity output per capita (kWh): **- Weak**
Total final energy consumption per Capita (TFEC) (MJ): **Moderately Strong**

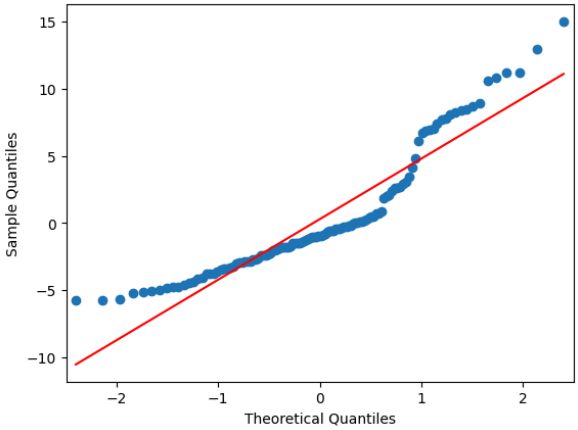
Most predictive features according to training data

GDP per capita, PPP (current international $): **Strong**
Total final energy consumption per Capita (TFEC) (MJ): **Strong**

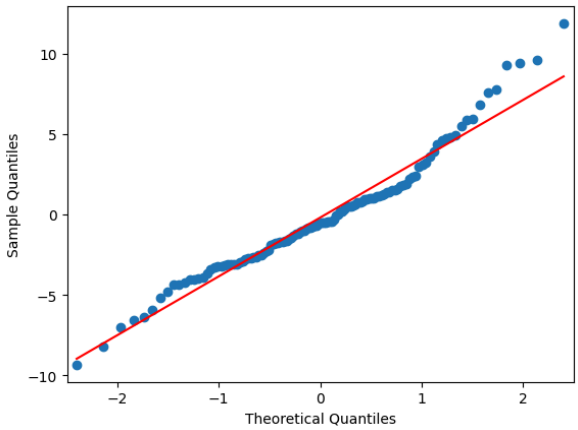
Residuals and Applicability of linear regression applicable to problem

When evaluating the results of the Linear Regression Models for each independent variable on predicting the dependent variable of GHG Emissions (Figures shown below), it was found that most of the variables had a curved plot meaning that a linear regression is not the most suitable model for using the data for those variables. However, the features **GDP per capita**, PPP (current international $) and **Total final energy consumption per Capita** (TFEC) (MJ) have almost an entirely straight residual line plot making a linear regression perfectly applicable, although there are seemingly some slight outliers at the right tail of the plots.

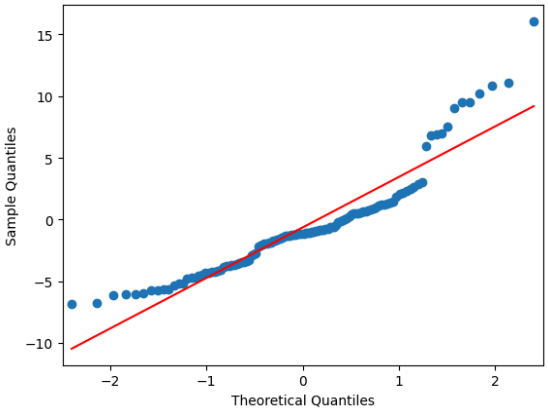
Fossil fuel energy consumption (% of total)



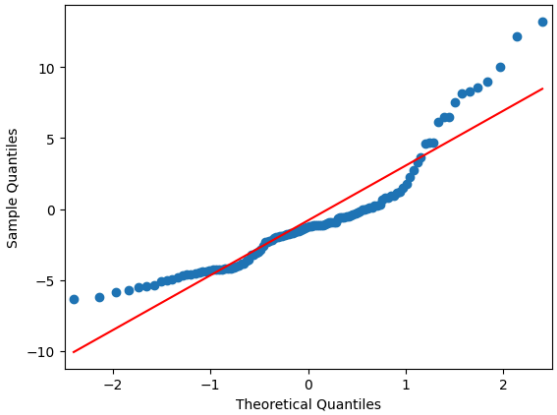
GDP per capita, PPP (current international $)



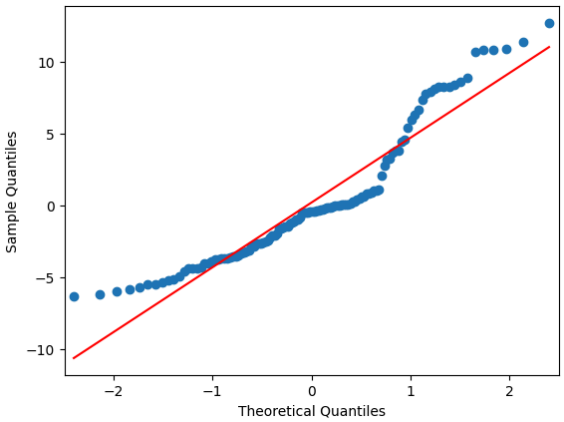
Agricultural land (% of land area)



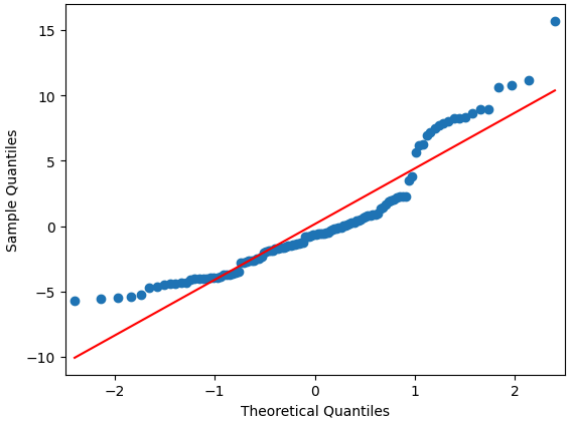
Forest area (% of land area)



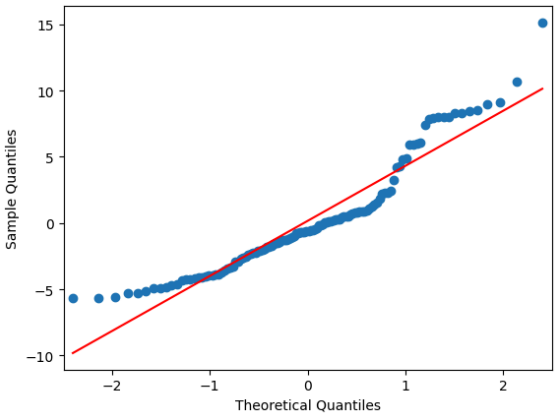
Urban population (% of total population)



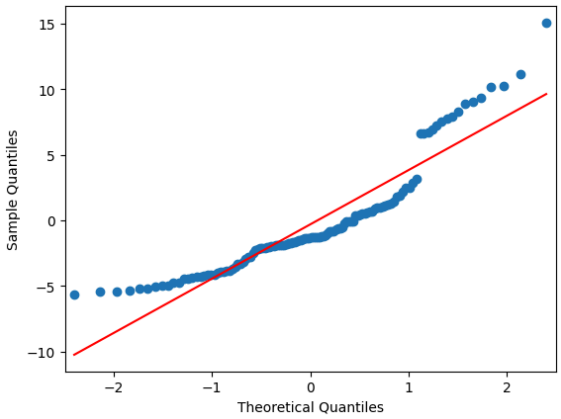
Access to Fuels and Tech for cooking (% of total population)



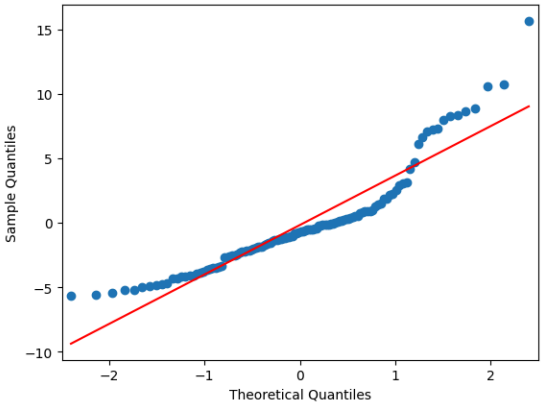
Access to electricity (% of rural population with access)



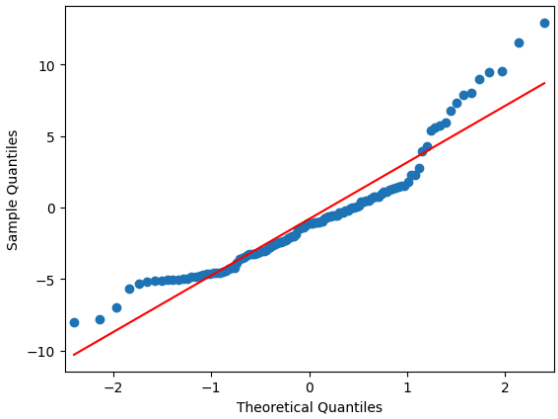
Access to electricity (% of total population)



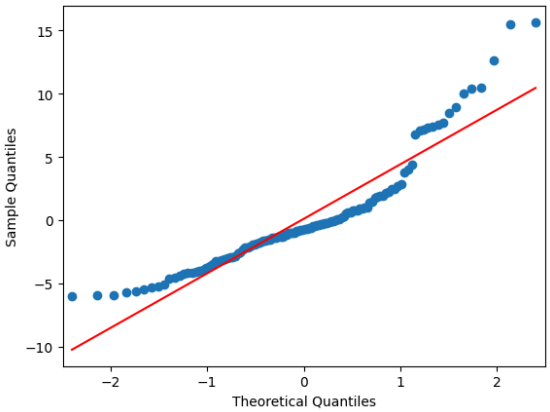
Access to electricity (% of urban population with access)



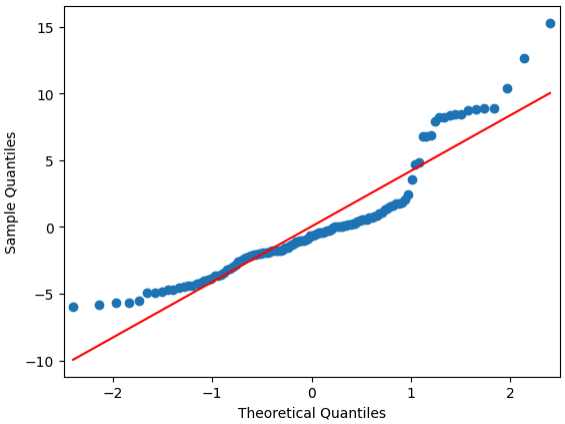
Energy intensity level of primary energy (MJ/PPP)



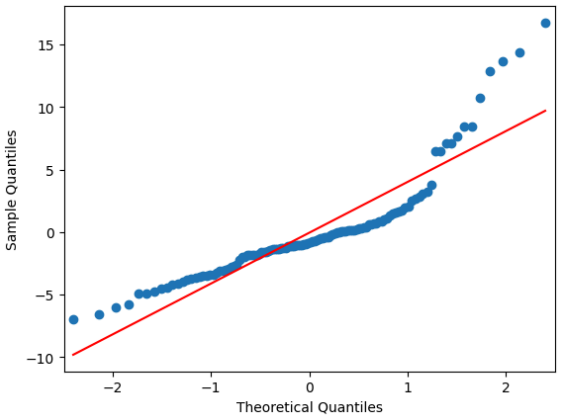
Renewable electricity share of total electricity output (%)



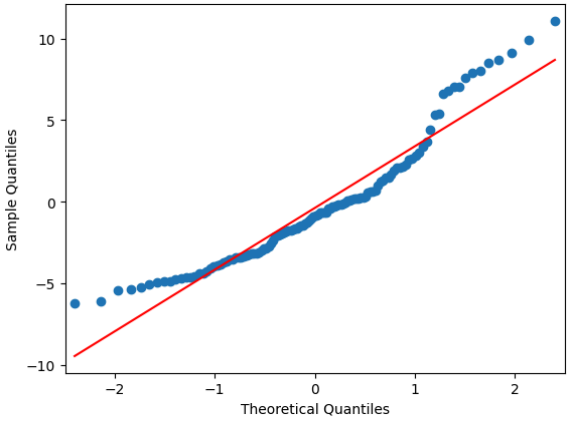
Renewable energy share of TFEC (%)



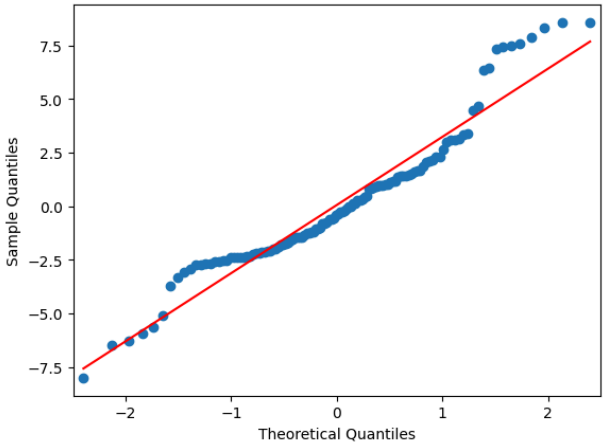
Population, total



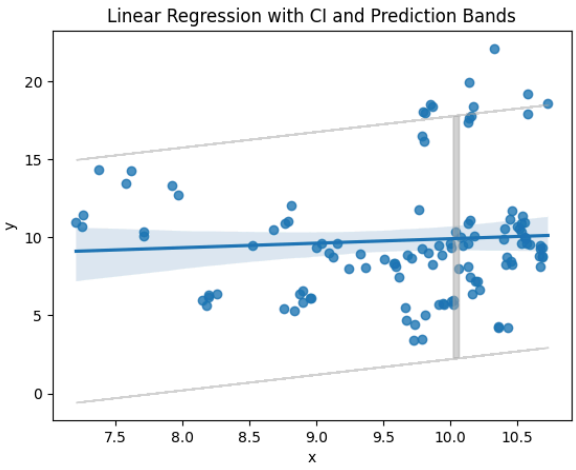
Total electricity output per capita (kWh)



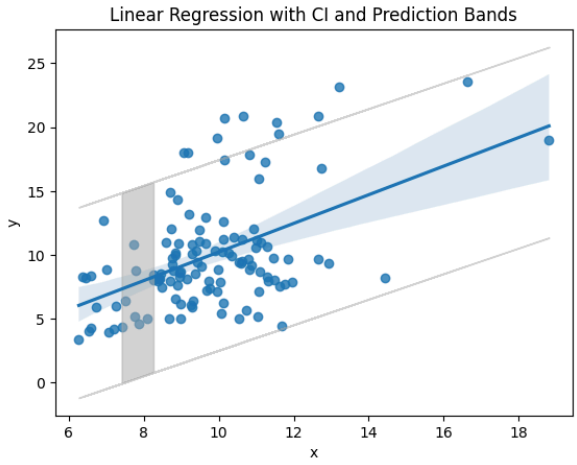
Total final energy consumption per Capita (TFEC) (MJ)



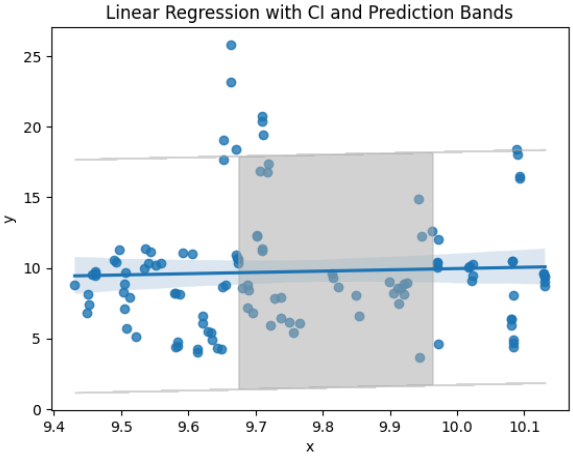
Fossil fuel energy consumption (% of total):



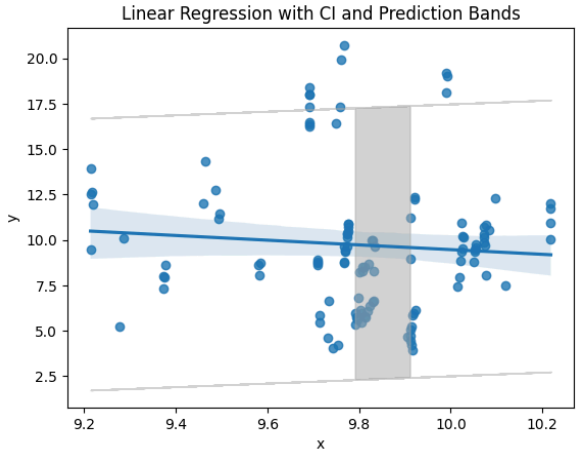
GDP per capita, PPP (current international $)



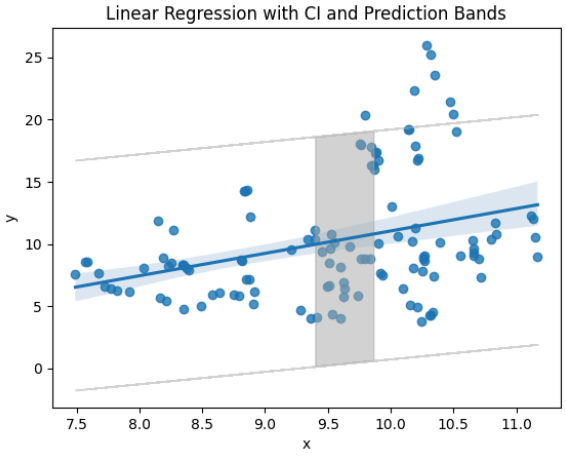
Agricultural land (% of land area)



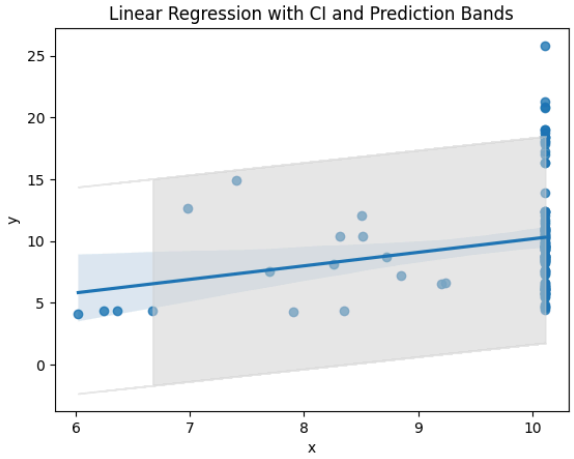
Forest area (% of land area)



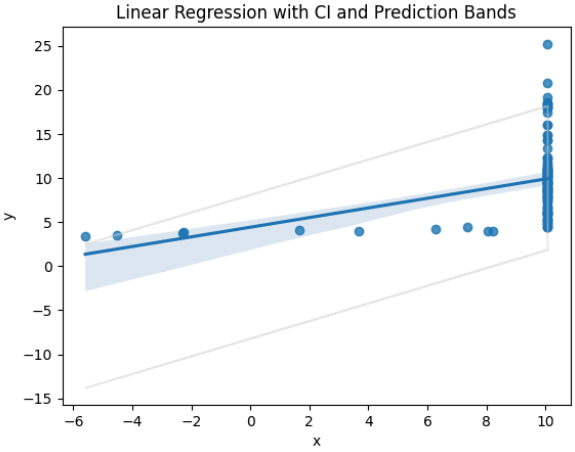
Urban population (% of total population)



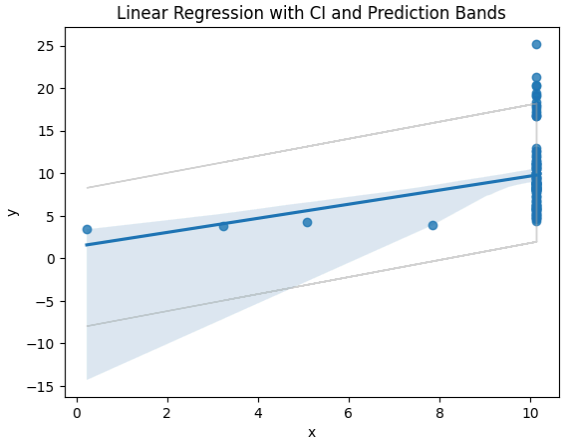
Access to Fuels/Tech for cooking (% of total population)



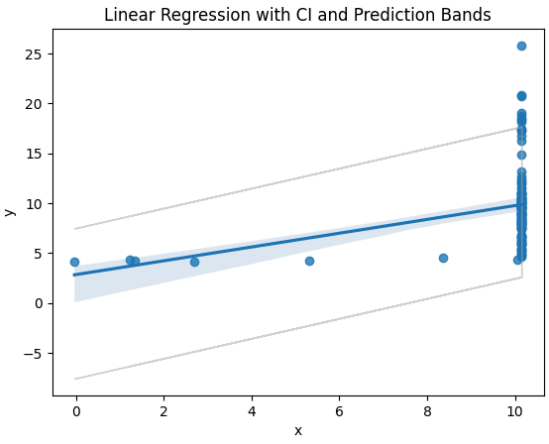
Access to electricity (% of rural population with access)



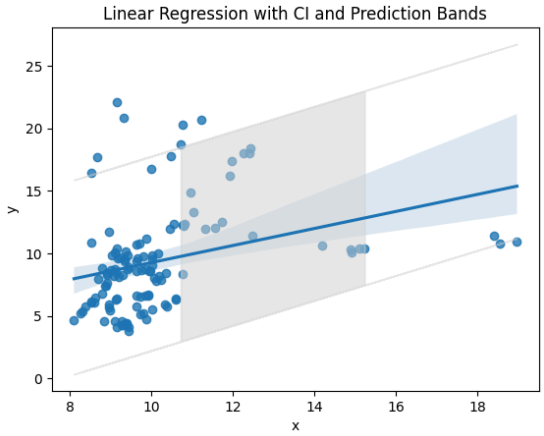
Access to electricity (% of total population)



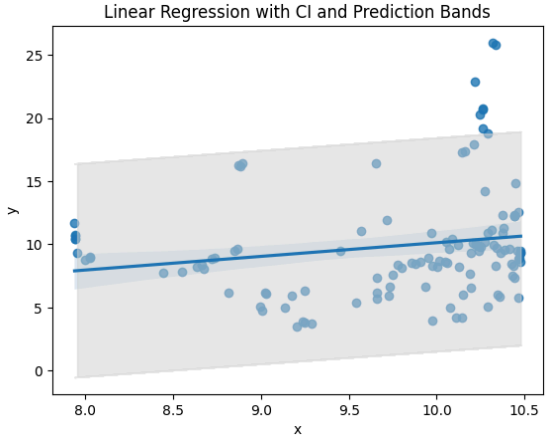
Access to electricity (% of urban population)



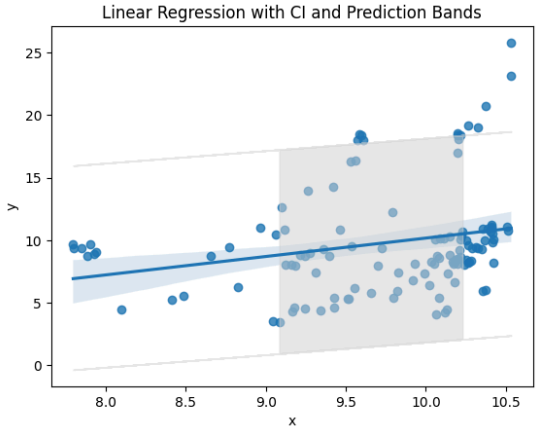
Energy intensity level of primary energy (MJ/PPP)



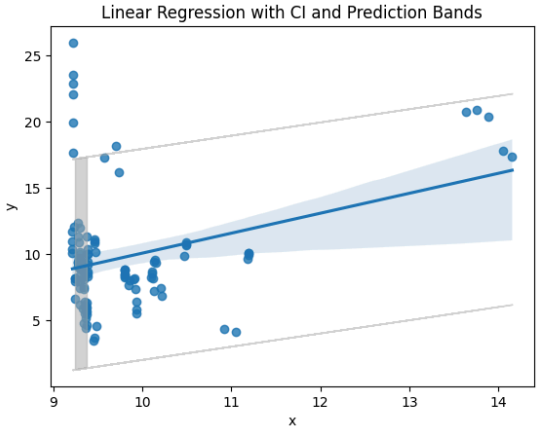
Renewable electricity share of total electricity output (%)



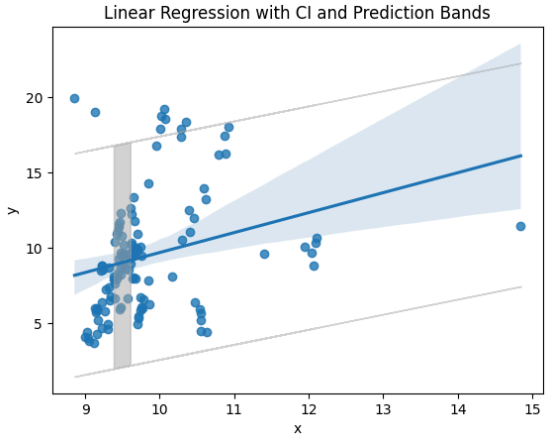
Renewable energy share of TFEC (%)



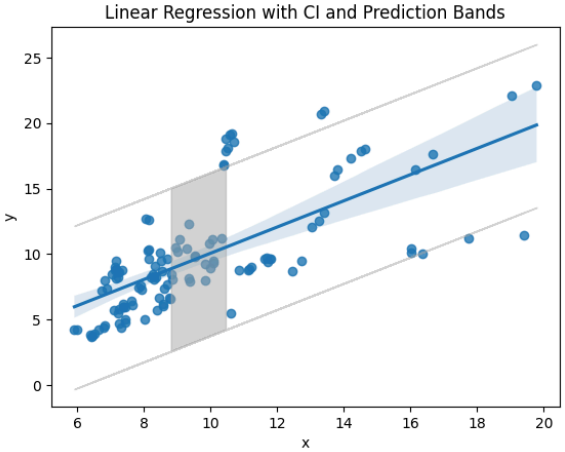
Population, total



Total electricity output per capita (kWh)



Total final energy consumption per Capita (TFEC) (MJ)



Correlation between predicted and real values

Fossil fuel energy consumption (% of total): 0.18979796326508017 - **weak positive correlation**
GDP per capita, PPP (current international $): 0.5008590798516378 - **moderate strong positive correlation**
Agricultural land (% of land area): -0.08139487865439202 - **weak negative correlation**
Forest area (% of land area): -0.09054855598507772 - **weak negative correlation**
Urban population (% of total population): 0.26093337760053437 - **weak positive correlation**
Access to Clean Fuels and Technologies for cooking (% of total population): 0.22022449625079363 - **weak positive correlation**
Access to electricity (% of rural population with access): 0.34327631032027134 - **moderate weak positive correlation**
Access to electricity (% of total population): 0.23244150843987832 - **weak positive correlation**
Access to electricity (% of urban population with access): 0.28780952928312326 - **weak positive correlation**
Energy intensity level of primary energy (MJ/PPP): 0.3189259346053195 - **moderate positive correlation**
Renewable electricity share of total electricity output (%): 0.18403139885181505 - **weak positive correlation**
Renewable energy share of TFEC (%): 0.2401103351612179 - **weak positive correlation**
Population, total: 0.3445304580697669 - **moderate weak positive correlation**
Total electricity output per capita (kWh): 0.27528801816346865 - **weak positive correlation**
Total final energy consumption per Capita (TFEC) (MJ): 0.6881027219789291 - **strong positive correlation**

Mean square error between predicted and real values

Fossil fuel energy consumption (% of total): **20.46162086404233**
GDP per capita, PPP (current international $): **13.41579849389057**
Agricultural land (% of land area): **17.205667633777033**
Forest area (% of land area): **15.563918361289375**
Urban population (% of total population): **20.435601789869786**
Access to Clean Fuels and Technologies for cooking (% of total population): **18.202259689878833**
Access to electricity (% of rural population with access): **17.344070202008073**
Access to electricity (% of total population): **17.283391133741436**
Access to electricity (% of urban population with access): **14.766366283662615**
Energy intensity level of primary energy (MJ/PPP): **16.311233214662124**
Renewable electricity share of total electricity output (%): **18.639446000396415**

Renewable energy share of TFEC (%): **17.35227455212906**
Population, total: **16.564124820505246**
Total electricity output per capita (kWh): **14.46402691610144**
Total final energy consumption per Capita (TFEC) (MJ): **10.10612173243856**

Multivariate Regressions

Then, we started considering combinations of independent features to see if it improves the prediction results. We evaluated different combinations of features as applicable and reported those that improve the results shown from our original models.

Coefficients for each feature

Combo 1:
Total final energy consumption per Capita (TFEC) (MJ): **3.93563041e-05**
GDP per capita, PPP (current international $): **2.40138010e-05**

Combo 2:
Total final energy consumption per Capita (TFEC) (MJ): **6.72205683e-05**
GDP per capita, PPP (current international $): **-6.62537244e-06**
Total electricity output per capita (kWh): **-2.38828318e-04**

Combo 3:
Total final energy consumption per Capita (TFEC) (MJ): **6.90341493e-05**
GDP per capita, PPP (current international $): **-3.81659754e-05**
Total electricity output per capita (kWh): **-2.51605539e-04**
Access to electricity (% of urban population with access): **-2.51605539e-04**

Combo 4:
Fossil fuel energy consumption (% of total): **6.22908715e-02**
GDP per capita, PPP (current international $): **-1.27475017e-05**
Agricultural land (% of land area): **-4.21832749e-02**
Forest area (% of land area): **-1.44882855e-02**
Urban population (% of total population): **-3.88123018e-03**
Access to Clean Fuels and Technologies for cooking (% of total population): **-3.06085476e-01**
Access to electricity (% of rural population with access): **-1.34370801e-01**
Access to electricity (% of total population): **4.08807544e-01**
Access to electricity (% of urban population with access): **1.42061175e+01**
Energy intensity level of primary energy (MJ/PPP): **-8.18714909e-02**
Renewable electricity share of total electricity output (%): **-2.60670963e-02**

Renewable energy share of TFEC (%): **-6.93492504e-02**
Population, total: **8.72600412e-09**
Total electricity output per capita (kWh): **1.01411872e-05**
Total final energy consumption per Capita (TFEC) (MJ): **6.39231550e-05**

Most predictive features according to the training data

Total final energy consumption per Capita (TFEC) (MJ): **Strong**
GDP per capita, PPP (current international $): **Moderate Strong**
Total electricity output per capita (kWh): **Moderate**
Access to electricity (% of urban population with access): **Weak**

Correlation between predicted and real values

Combo 1: 0.7731172197487586 - **strong positive correlation**
Combo 2: 0.7913086785649242 - **strong positive correlation**
Combo 3: 0.7914537361004097 - **strong positive correlation**
Combo 4: 0.9394588900507626 - **extremely strong positive correlation**

Mean square error between predictive and real values

Combo 1: **10.689390736998181**
Combo 2: **7.72605713992491**
Combo 3: **8.484029377883756**
Combo 4: **2.326290166853148**

Regularization

Observed improvements in prediction results

Fossil fuel energy consumption (% of total): Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Access to electricity (% of rural population with access): Ridge Model - **MSE decreased, Correlation coefficient increased**
Access to electricity (% of total population): Ridge Model - **MSE decreased, Correlation coefficient increased**
Access to electricity (% of urban population with access): Ridge Model - **MSE decreased, Correlation coefficient increased**
Energy intensity level of primary energy (MJ/PPP): Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**

Renewable electricity share of total electricity output (%): Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Renewable energy share of TFEC (%): Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Population, total: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Total electricity output per capita (kWh): Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**

Differences or similarities across runs

Combo 1A: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Combo 1B: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Combo 2A: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Combo 2B: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Combo 3A: Ridge Model - **MSE decreased, Correlation coefficient increased**
Combo 3B: Ridge Model - **MSE decreased, Correlation coefficient increased**
Combo 4A: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**
Combo 4B: Lasso and Ridge Models - **MSE decreased, Correlation coefficient increased**

Logistic Regression

The next classifier we created was a Logistic Regression model. First, we created an additional column, ***Binary Class,*** which converted the GHG Emission target variable into a binary categorized variable with values 0 and 1 corresponding to a low and high emission standard of <3.5 tCO2e and >=3.5 tCO2e. For the combination of features that yielded the most accurate results according to our multivariate regression (All Independent Features Combined), we then computed a logistic regression model to predict our new dependent feature, ***Binary Class***:

Intercept : **-2.2557407301886526e-07**

Coefficients
[-1.92289231e-05, 2.53549637e-04, 4.71881796e-06, 2.61081950e-06, -2.78559081e-05, -1.77180474e-05, -1.47183903e-05, -2.13058071e-05, -2.23160337e-05, -1.26999009e-07, -3.15439658e-05, -1.54291766e-05, 1.37848977e-08, 4.36317553e-04, -3.38541553e-05]

Independent Variable	Log-Odds for a Unit Increase	Odds Ratio for a Unit Increase
Fossil fuel energy consumption (% of total)	-1.9228923133422446e-05	0.9999807712617411
GDP per capita, PPP (current international $)	0.0002535496367721389	1.0002535817831981
Agricultural land (% of land area)	4.718817956333046e-06	1.00000471882909
Forest area (% of land area)	2.6108194951520646e-06	1.0000026108229034
Urban population (% of total population)	-2.7855908089574026e-05	0.9999721444798826
Access to Clean Fuels and Technologies for cooking (% of total population)	-1.7718047441750187e-05	0.9999822821095219
Access to electricity (% of rural population with access)	-1.4718390349165552e-05	0.9999852817179659
Access to electricity (% of total population)	-2.130580711001168e-05	0.9999786944198571
Access to electricity (% of urban population with access)	-2.23160336923816e-05	0.9999776842153084
Energy intensity level of primary energy (MJ/PPP)	-1.2699900857066284e-07	0.9999998730009995
Renewable electricity share of total electricity output (%)	-3.154396575573661e-05	0.9999684565317499
Renewable energy share of TFEC (%)	-1.54291765778634e-05	0.9999845709424513
Population, total	1.3784897673540604e-08	1.0000000137848977
Total electricity output per capita (kWh)	0.0004363175525127647	1.0004364127528615
Total final energy consumption per Capita (TFEC) (MJ)	-3.385415526168764e-05	0.9999661464177838

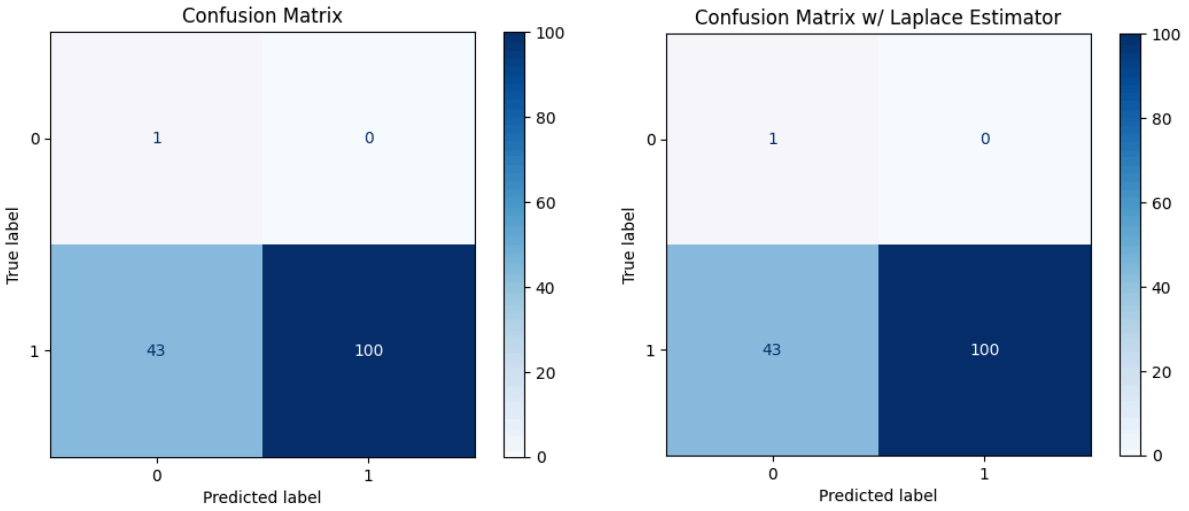
Most Predictive Features

1. Total electricity output per capita (kWh): 0.0004363175525127647
2. GDP per capita, PPP (current international $): 0.0002535496367721389
3. Total final energy consumption per Capita (TFEC) (MJ): 3.385415526168764e-05
4. Renewable electricity share of total electricity output (%): 3.154396575573661e-05
5. Urban population (% of total population): 2.7855908089574026e-05
6. Access to electricity (% of urban population with access): 2.23160336923816e-05
7. Access to electricity (% of total population): 2.130580711001168e-05
8. Fossil fuel energy consumption (% of total): 1.9228923133422446e-05
9. Access to Clean Fuels and Technologies for cooking (% of total population): 1.7718047441750187e-05
10. Renewable energy share of TFEC (%): 1.54291765778634e-05
11. Access to electricity (% of rural population with access): 1.4718390349165552e-05
12. Agricultural land (% of land area): 4.718817956333046e-06
13. Forest area (% of land area): 2.6108194951520646e-06
14. Energy intensity level of primary energy (MJ/PPP): 1.2699900857066284e-07
15. Population, total: 1.3784897673540604e-08

```
# Use the trained logistic regression model to make predictions on the testing dataset
y_pred = logreg.predict(X_test)

# Print the predicted labels
print("Predicted labels:", y_pred)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)

print(f"Accuracy: {accuracy:.2f}")
```

```
Predicted labels: [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
Accuracy: 0.99
```

Naive Bayes Classifier



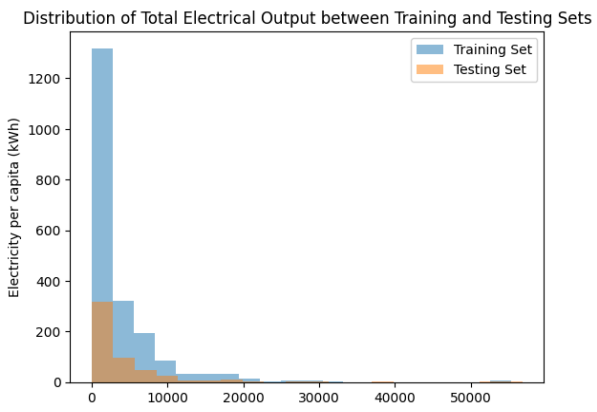
Next, we reported the classification results using a Naïve Bayes classifier. Using the same training and testing set division, we created and trained the classifier before using it to predict the testing data and evaluated the results using the Confusion Matrices above. The first figure is for a normal classifier while the second figure is after adding a Laplace estimator to test if the results would improve. As the matrices show, the Naive Bayes classifier showed the most accurate results in predicting the higher emission class (the countries with emissions above 3.5 tCO2e) and the Laplace estimator did not improve the results as it returned an identical figure.

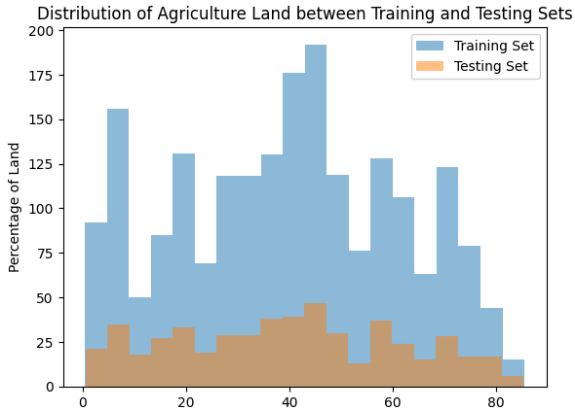
Decision Trees & Random Forests

In this part, we evaluated the efficacy of decision trees and random forests. We created another column to convert our GHG Emission target variable, which is continuous, into a multi-class variable. We created 3 levels (Low, Mid, and High) to categorize different levels of emissions and used all the classes to test the accuracy of the decision tree.

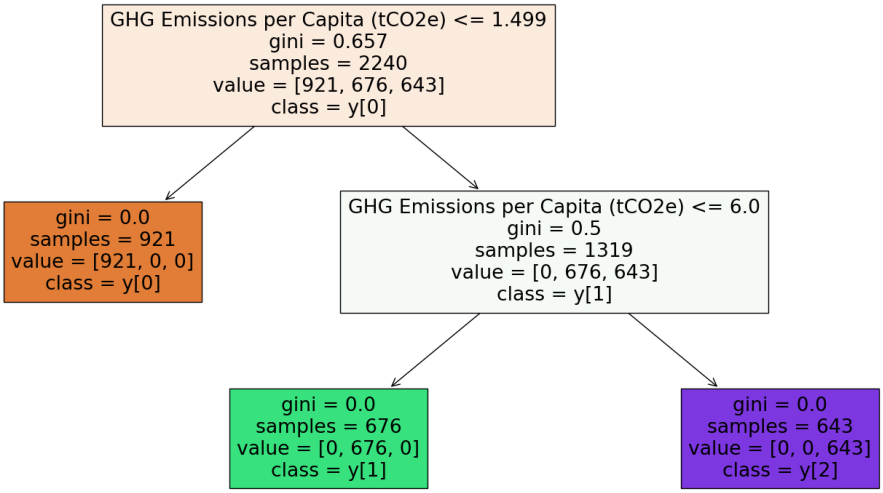
The data set was first split using a 80/20 split with 80 percent of the data assigned for training, and the remaining 20 percent for testing the model.

After splitting the data training and testing sets, it was vital to ensure that the distributions were similar. By creating histogram visualizations, it was clear to see that the distributions of the training and testing sets remained similar after the split.





After training the decision tree, these are the resulting if-then rules generated by the model.



Based on the visualization of the decision tree above, the leaf nodes appear to have a Gini value of 0, which indicates that the node is perfectly homogeneous, and there is no uncertainty or impurity in terms of class labels within that node. This is an ideal scenario for decision trees because it allows for clear decision-making based on the attributes or features at that node.

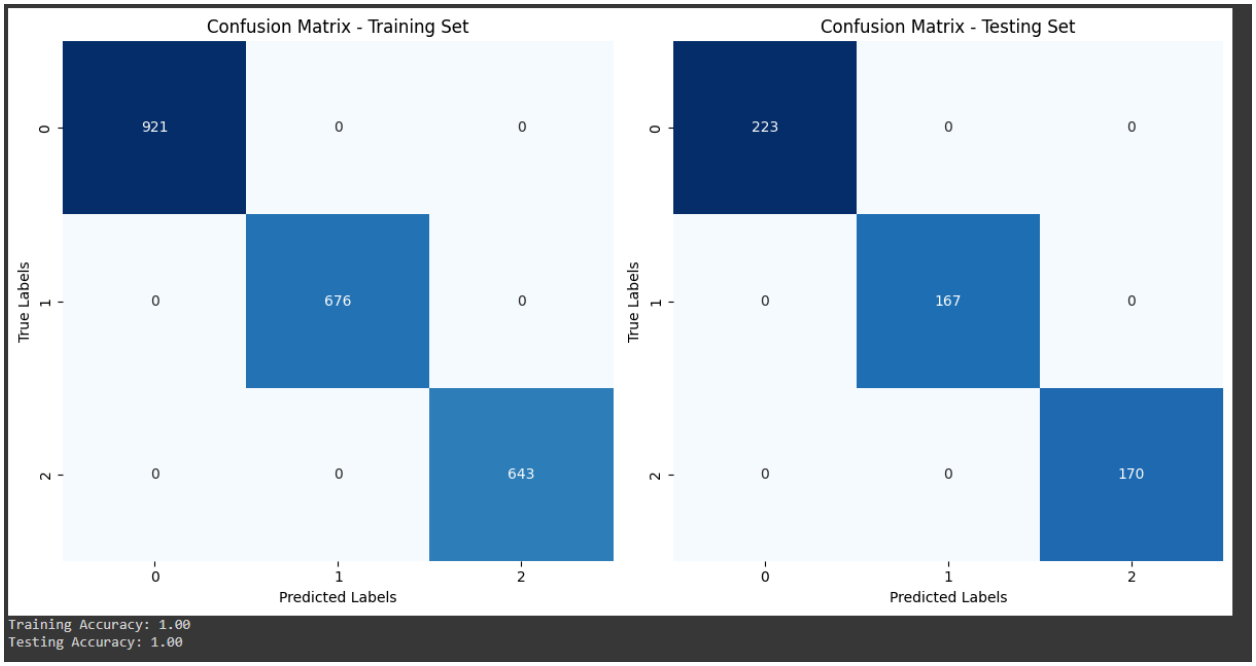


Figure: Confusion matrix for training and testing sets

After modeling the decision tree, it was important to run a confusion matrix on the training and testing sets, to determine how well it was performing in terms of classifying instances into different classes. Based on the results obtained from the confusion matrix in the figure above, both the training and testing sets were performing well and making accurate predictions.

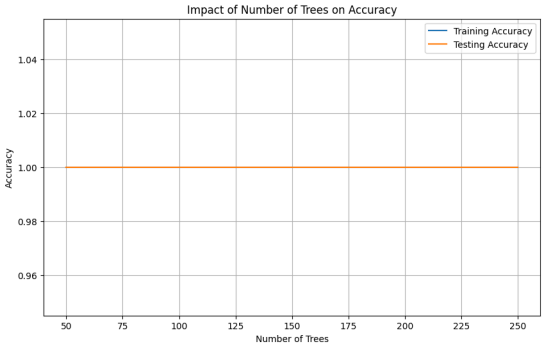
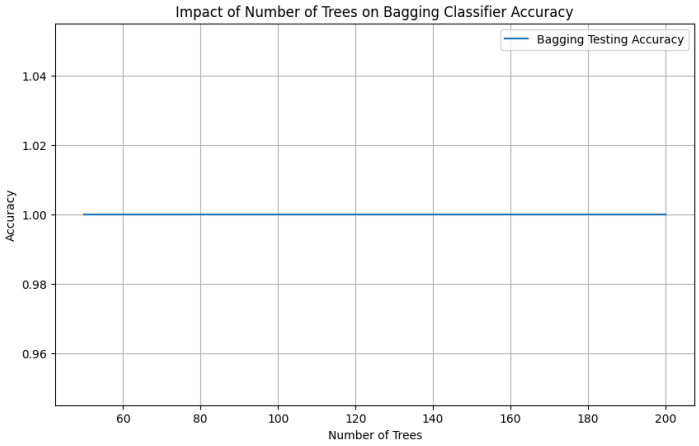


Figure: Gradient Boosting Classifier

We then applied gradient boosting with different numbers of trees to analyze the impact on the prediction results. After performing the Gradient Boosting classifier, the model seems to still be performing well, and making accurate predictions of the testing values with 100 percent accuracy.

We performed the same analysis with bagging and random forests: trained with bagging and then trained with random forests using at least four different numbers of trees. Then, we compared prediction results over the testing sets.

After performing the same analysis using the Bagging the accuracies remain the same at a value of 1, indicating that the model is performing as expected and predicting new data points accurately.



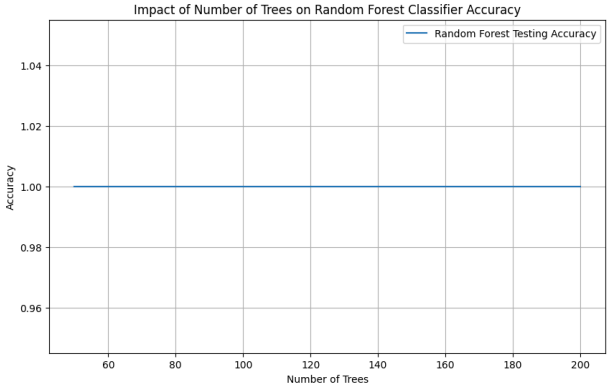


Figure: Random Forest (80/20 Split)

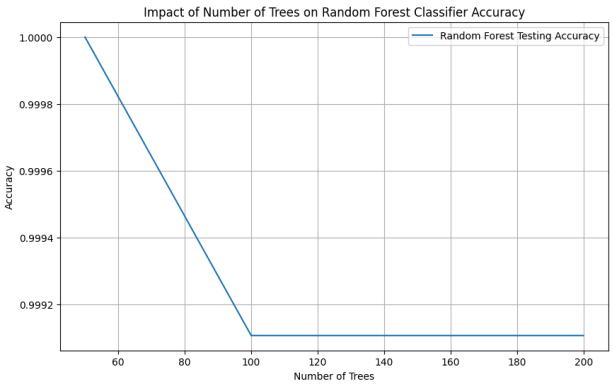


Figure: Random Forest (60/40 Split)

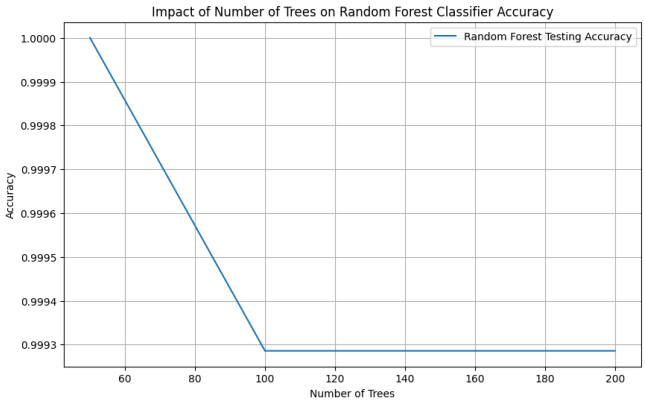


Figure: Random Forest (50/50 Split)

By implementing an increased number of Random Forest trees, it was evident that the more trees the model added, the less accurate the predictions. There appears to be a threshold at approximately 100 trees, indicating that the accuracy would remain the same even if more trees are added to the predictive model.

	Feature	Importance
4	GHG Emissions per Capita (tCO2e)	0.659130
2	Total electricity output per capita (kWh)	0.175586
3	Total final energy consumption per Capita (TFE...	0.139785
1	Population, total	0.013759
0	Agricultural land (% of land area)	0.011739

Table: Importance Features for Forest Trees

After running the various forest trees, the most important feature was the GHG Emission per Capita (tCO2e) at 66%, and subsequently the Total Electricity Output per Capita (kWh) at 18%. This analysis is accurate with what the model aims to achieve, by accurately predicting the correct class of low, medium and high emission ranges.

Comparative Analysis

Linear regression, logistic regression, and decision tree/random forest classifiers are widely used machine learning techniques with distinct characteristics and applications. Linear regression is primarily used for predicting continuous numeric values and works well when there is a linear relationship between the independent and dependent variables. However, it may not be suitable for classification tasks or when the relationship between variables is nonlinear. This kind of linear relationship was mostly absent between our independent variables and GHG emissions except for two variables, but the fact that our target variable is a continuous one makes it the most ideal option on paper. Logistic regression, on the other hand, is specifically designed for binary or categorical classification tasks, predicting the probability of an event occurring. It performs well when the decision boundary between classes is linear or when dealing with categorical data. Again, our target variable is a continuous variable depicting the emission rates per capita for a country so it required us to create a binary standard, with 3.5 tCO2e) as the halfway point, to split those values between low and high in order to create a logistic regression model. Once the binary variable was created though, it was highly accurate at 99% showcasing the strength of a logistic regression. Decision tree and random forest classifiers are versatile algorithms that can handle both regression and classification tasks. They partition the feature space into regions based on feature splits, making them capable of capturing nonlinear relationships and interactions between variables. While decision trees are prone to overfitting and may lack interpretability, random forests mitigate these issues by aggregating multiple decision trees. Overall, the choice between these techniques depends on the nature of the data, the problem at hand, and the desired balance between interpretability and predictive performance.

In our project aimed at predicting greenhouse gas emissions based on socioeconomic indicators, the choice of classifier depends on the specific characteristics of the data and the goals of the analysis. Given that greenhouse gas emissions are a continuous variable, linear regression is the most suitable choice out of the ones we tested in this report as it is designed to predict continuous numeric values. Linear regression models provided insights into the relationship between socioeconomic indicators and greenhouse gas emissions, allowing for interpretation of coefficients and identification of significant predictors. This differs from the logistic regression classifier which, although accurate, failed to provide any additional insights that would be useful to policymakers as the binary standard chosen to convert the continuous variable to a categorial one was an arbitrary choice.

Contributions

Marta Gonzalez - Question 1 of the code (Linear Regressions w/ Multivariate and Regularization), Contributed equally to the report and presentation.

Sai Abhishek Gangineni - Question 2 of the code (Logistic Regressions and Naive Bayes), Contributed equally to the report and presentation

Robin Godinho - Question 3 of the code (Decision Trees, Gradient Boosting, Bagging, Random Forest), Contributed equally to the report and presentation.