Milestone 2 Report

Feature Engineering/Selection, Data Modeling Feb 24 – Apr 7

Student: Tzu-Chieh Chao

1. Project Objective

This project utilizes machine learning and interactive dashboards to investigate the interconnections between carbon emissions, energy sustainability, and economic growth among G20 countries from 2000 to 2020. By employing extensive datasets sourced from Kaggle, the research implements predictive modeling techniques to estimate per capita CO_2 emissions (CO2_per_capita) and GDP growth rates (gdp_growth). Additionally, classification models are utilized to categorize nations into high and low carbon emitters as well as different tiers of energy efficiency, thereby providing valuable insights into global sustainability trends. The machine learning framework incorporates both regression and classification methodologies, which are refined through cross-validation and feature selection processes. The findings are presented through an interactive dashboard, developed using Plotly Dash or Streamlit, which facilitates the exploration of dynamic correlations—such as the relationship between GDP per capita and emissions—and allows for the simulation of scenario-based outcomes. By integrating data science with environmental economics, this project establishes a scalable framework for evaluating climate policies and their socio-economic implications, highlighting the importance of data-driven approaches to decarbonization.

2. Tools

- Programming Language: Python
- Libraries:
    - Data Manipulation: Pandas, NumPy, re
    - Visualization: Matplotlib, Seaborn, Plotly, mplcursors
    - Statistical Analysis: SciPy
    - Machine Learning : Scikit-learn
- Version Control: Git & GitHub
- Computing Resource: HiperGator(University of Florida's supercomputer)

3. Data Sources

To ensure a comprehensive analysis, I've selected the following datasets from Kaggle:

1. Global Data on Sustainable Energy(2000-2020)
    - Description: This dataset presents an analysis of sustainable energy indicators and various relevant factors across all countries from the year 2000 to 2020. It encompasses critical dimensions such as access to electricity, the utilization of renewable energy sources, carbon emissions, energy intensity, financial flows, and economic growth. Researchers can engage in comparative analyses among

nations, monitor advancements towards Sustainable Development Goal 7, and derive significant insights into the trends of global energy consumption over the specified period.

- Source: global-data-on-sustainable-energy

2. CO2 Emission by countries Year wise (1750-2022)
    - Description: This dataset will facilitate predictions regarding global warming.
    - Source: co2-emission-by-countries-

3. G20 Countries' Currency Exchange Rates against USD
    - Description: This dataset includes the currency exchange rates of G20 countries in relation to the US dollar, thereby facilitating a comprehensive understanding of the macroeconomic landscape.
    - Source: g20-countries-currency-exchange-rates-against-usd
    -

4. Data Preprocessing

Initially, I perform data cleaning and analysis on three separate datasets. I remove any countries that are not part of the G20 from all three datasets. Furthermore, after relocating the columns with a significant amount of missing data, all missing values have been resolved.

In the Global Data on Sustainable Energy dataset . I changed the column name "Entity" to "Country" and renamed "Density. (P/Km2)" to "Density (Square kilometre)." Additionally, I converted the data type of "Density (Square kilometre)" from object to float.

```
# Rename 'Entity' column to 'Country / Density

df = df.rename(columns={'Entity': 'Country'})
df.rename(columns={r'Density\n(P/Km2)': 'Density(Square kilometre)'}, inplace=True)
```

✓ 0.0s

```
#Convert to numeric
df['Density(Square kilometre)'] = pd.to_numeric(df['Density(Square kilometre)'], errors='coerce')
```

✓ 0.0s

In addition, I will remove the columns ['renewable-electricity-generating-capacity-per-capita', 'Financial flows to developing countries (US $)', 'Renewable energy share in total final energy consumption (%)', 'Electricity from nuclear (TWh)', 'Energy intensity level of primary energy (MJ/$2017 PPP GDP)', 'Value_co2_emissions_kt_by_country'] due to their excessive missing values.

```
#

df = df.drop(columns=['Renewable-electricity-generating-capacity-per-capita',
                      'Financial flows to developing countries (US $)',
                      'Renewable energy share in the total final energy consumption (%)',
                      'Electricity from nuclear (TWh)',
                      'Energy intensity level of primary energy (MJ/$2017 PPP GDP)',
                      'Value_co2_emissions_kt_by_country',
                     ],
             errors='ignore')
```

In the CO2 emission by countries(1750-2022) datasets. I eliminated the "Code" (and "Calling Code" columns as they are not meaningful.

```
#There is no meaning in Code and Calling Code
df = df.drop(columns=['Code','Calling Code'])
```
✓ 0.0s

In addition, I updated the column name "% of World" to "proportion of global land area," changed "Density(Km2)" to "Density (Square kilometre)," and renamed "Area" to "Area (Square kilometre)." I also converted the data types of both "Area (Square kilometre)" and "Density (Square kilometre)" from object to numeric.

```
# Clean the '% of World' column - convert from string '1.80%' to float 0.018
df['% of World'] = df['% of World'].apply(lambda x: float(x.replace('%', '')) / 100 if isinstance(x, str) else x)
df.rename(columns={'% of World': 'proportion of global land area'}, inplace=True)


# Clean the 'Density(km2)' column - extract numeric value and rename
def extract_density(value):
    if isinstance(value, str):
        match = re.search(r'(\d+)', value)
        if match:
            return int(match.group(1))
    return value


# Extract numeric values first
df['Density(km2)'] = df['Density(km2)'].apply(extract_density)

# Rename the column to the requested format
df.rename(columns={'Density(km2)': 'Density(Square kilometre)'}, inplace=True)
df.rename(columns={'Area': 'Area(Square kilometre)'}, inplace=True)
```

In the G20 Countries' Currency Exchange rates against USD dataset.

I modified the column name "COUNTRY" to "Country" and changed "YEAR" to

"Year." I I also replace the country code with its full name.

```
def clean_g20_exchange_rates(input_file, output_file):
    """
    Clean the G20 Exchange Rates CSV file by converting country codes to full names.

    Parameters:
    input_file (str): Path to the input CSV file with country codes
    output_file (str): Path to save the cleaned CSV file with full country names
    """
```

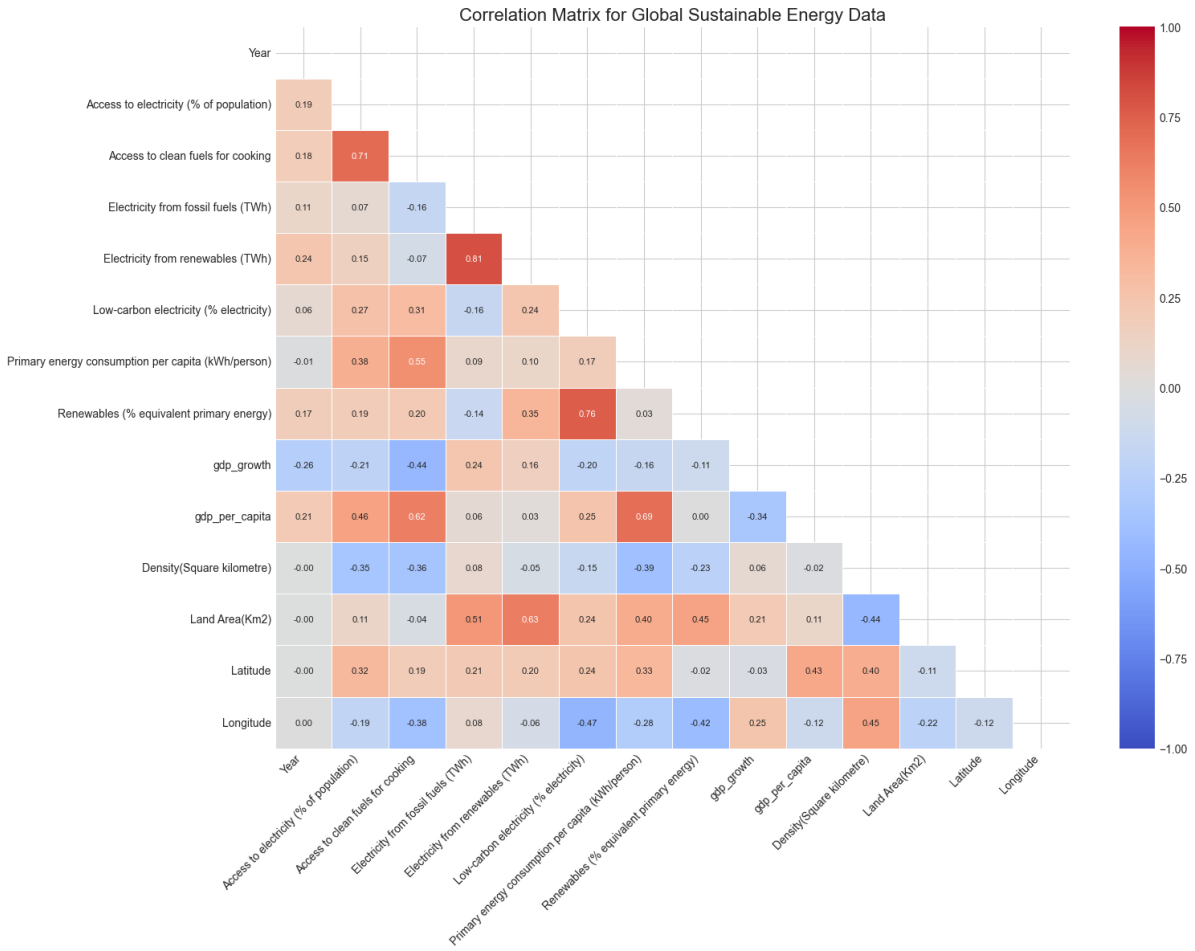
```
        "GBR": "United Kingdom",
        "IDN": "Indonesia",
        "IND": "India",
        "ITA": "Italy",
        "JPN": "Japan",
        "KOR": "South Korea",
        "MEX": "Mexico",
        "RUS": "Russia",
        "SAU": "Saudi Arabia",
        "TUR": "Turkey",
        "USA": "United States",
        "ZAF": "South Africa",
        "EU27_2020": "European Union"
    }

    # Replace country codes with full names
    df['COUNTRY'] = df['COUNTRY'].map(country_mapping)
    df.rename(columns={'COUNTRY': 'Country'}, inplace=True)
    df.rename(columns={'YEAR': 'Year'}, inplace=True)

    # Save the cleaned data to a new CSV file
    df.to_csv(output_file, index=False)
```

Finally, merging three distinct datasets.

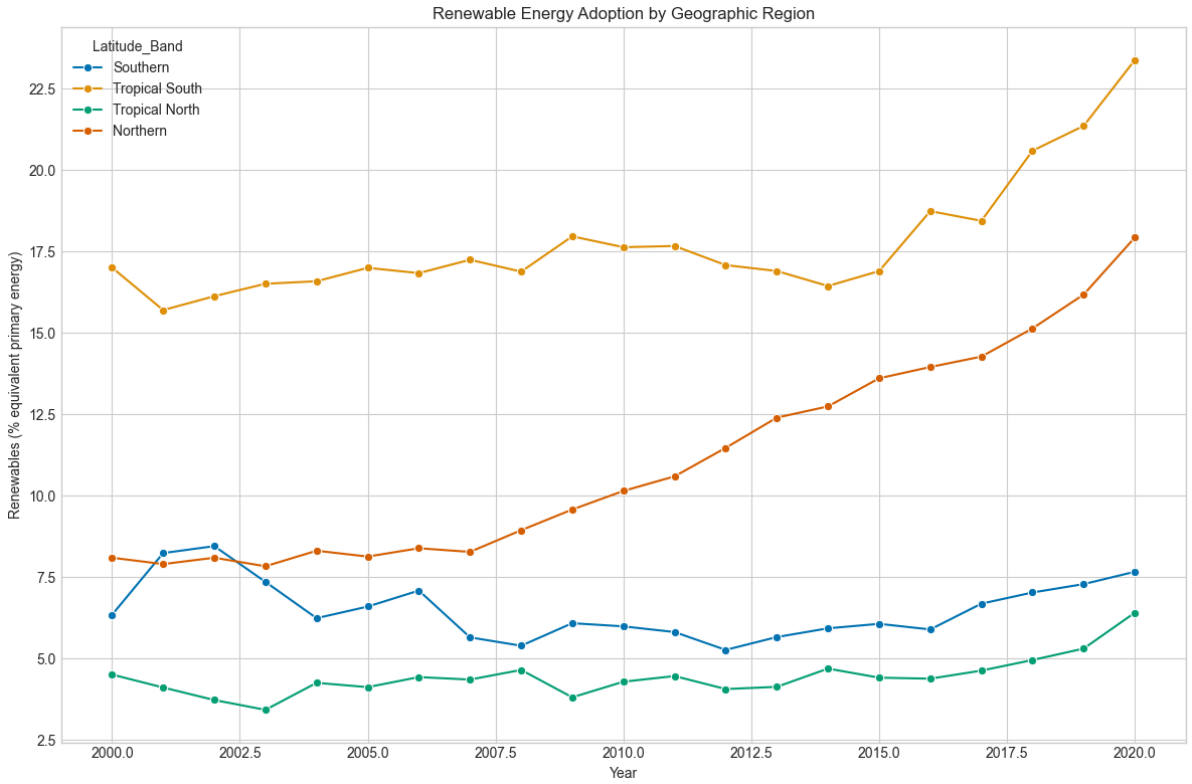
```
co2_df = pd.read_csv('New_CO2_emission_by_countries.csv')
energy_df = pd.read_csv('New_global_data_on_sustainable_energy.csv')
exchange_df = pd.read_csv('New_G20_Exchange_Rates.csv')

# Merge datasets using inner join
# First merge CO2 and energy data
merged_df = pd.merge(
    co2_df,
    energy_df,
    how='inner',
    on=['Country', 'Year']
)

# Then merge with exchange rate data
final_df = pd.merge(
    merged_df,
    exchange_df.rename(columns={'Value': 'Exchange_Rate'}),
    how='inner',
    on=['Country', 'Year']
)
```

5. EDA

Correlation Matrix on Sustainable energy data



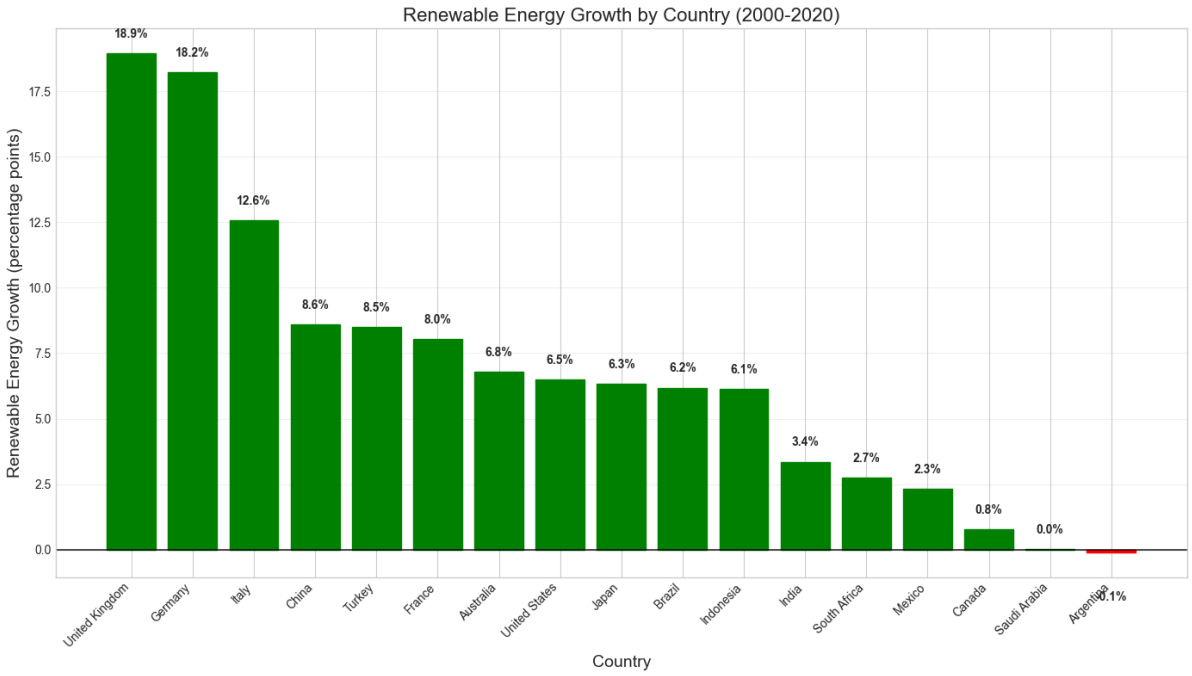
Renewable Energy Regional Analysis(Latitude)



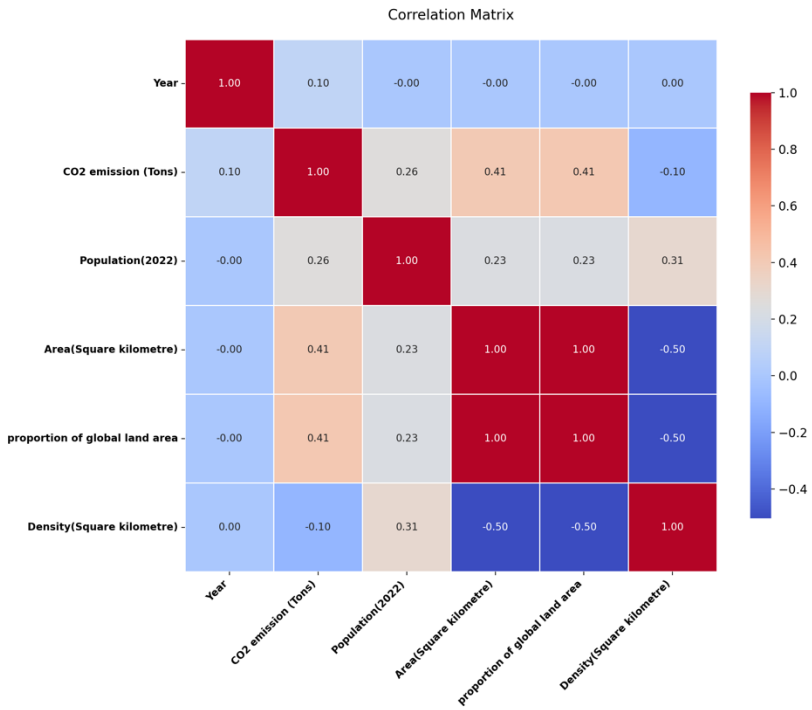
Renewables (% equivalent primary energy): Equivalent primary energy that is derived from renewable sources. Higher values may indicate:

- Higher levels of industrialization Libraries:
- Generally, climatic conditions that demand more energy (such as very cold or hot areas)
- Higher standard of living, using more energy-intensive products and services
- Energy efficiency may be low

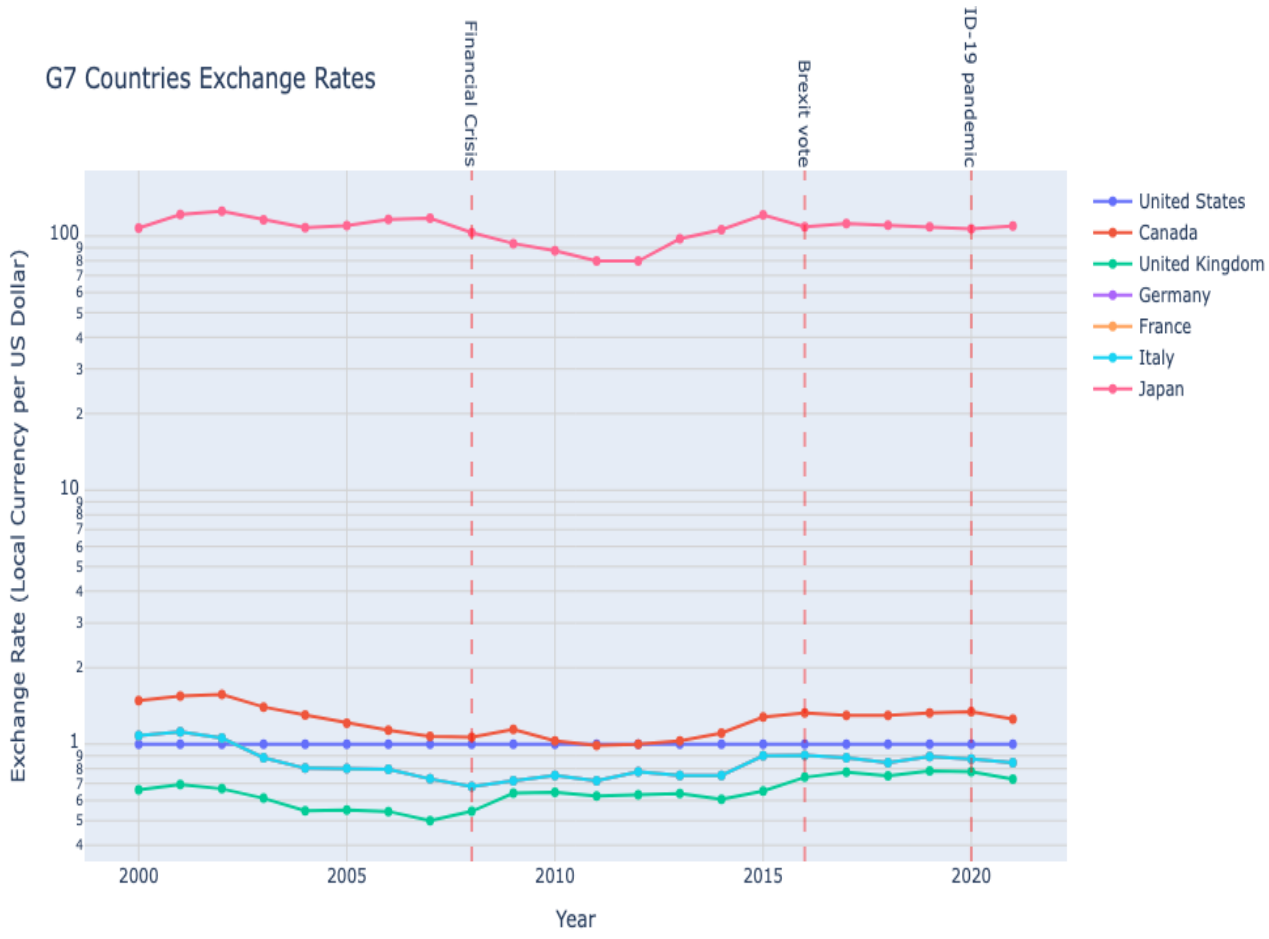
Renewable Energy Growth(2000-2020)



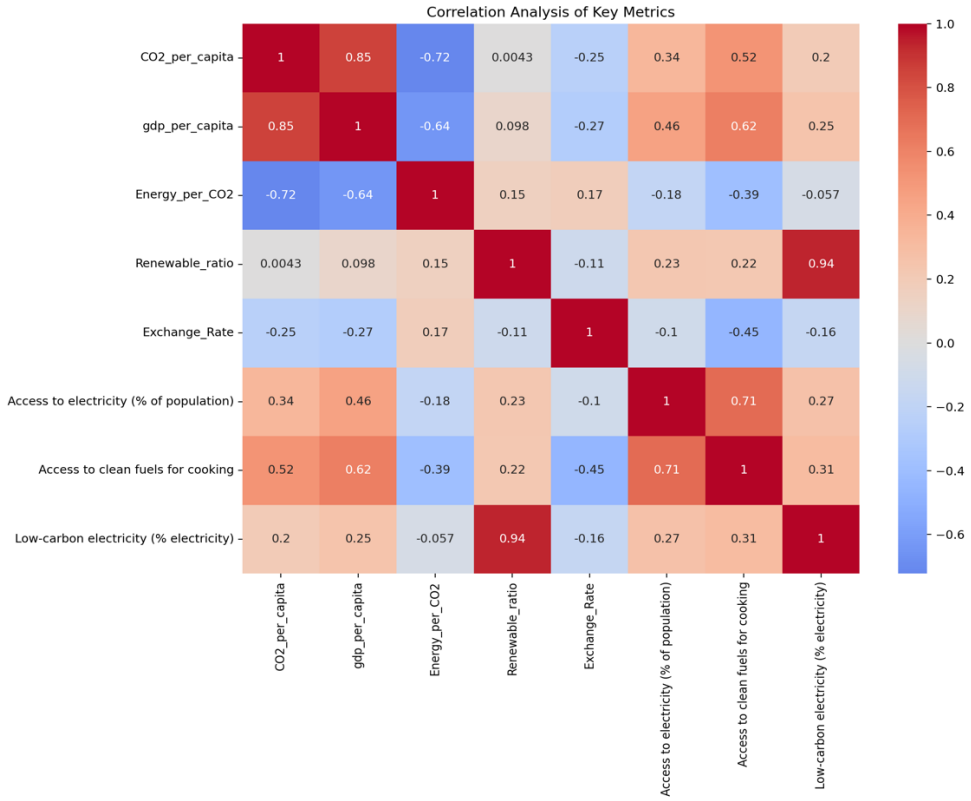
Correlation Matrix on CO2 data



Trend of Exchange Rates

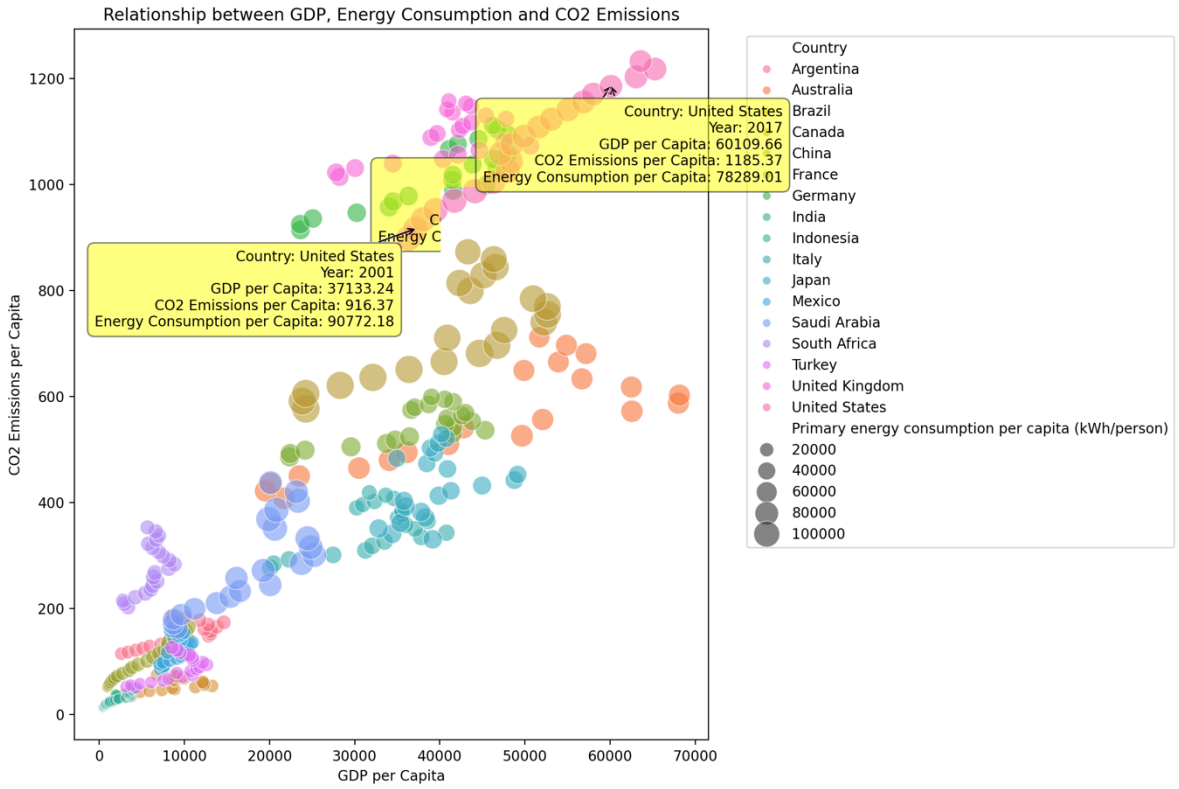


Total Correlation Matrix



Comparison(GDP/CO2/Energy)

Primary energy consumption per capita (kWh/person): Energy consumption per person in kilowatt-hours.



6. Feature Engineering

A. Create New Features

1. **CO2_per_capita**

- Calculation: Total CO2 emissions divided by population
- Meaning: Measures the average carbon footprint per individual in a country
- Interpretation:
    - Higher values indicate more carbon-intensive lifestyles
    - Allows direct comparison of emission levels across countries with different population sizes
    - Reflects individual contribution to global carbon emissions

2. **Energy_per_CO2**

- Calculation: Primary energy consumption per capita divided by CO2 per capita
- Meaning: Efficiency metric of energy usage relative to carbon output
- Interpretation:
    - Higher values indicate more efficient energy use with lower carbon emissions
    - Reflects how effectively a country converts energy consumption into economic output
    - Lower carbon intensity per unit of energy consumed
    - Provides insight into a country's energy efficiency and environmental performance

3. **Renewable_ratio**

- Calculation: Electricity from renewables divided by total electricity generation
- Meaning: Proportion of electricity generated from renewable sources
- Interpretation:
    - Ranges from 0 to 1 (0% to 100% renewable electricity)
    - Higher values indicate greater renewable energy adoption
    - Measures progress in clean energy transition
    - Reflects a country's commitment to sustainable energy production
    - Lower values suggest heavy reliance on fossil fuel-based electricity generation

4. **CO2_per_GDP**

- Calculation: CO2 emissions / (GDP per capita * Population)
- Meaning: Carbon intensity of economic output
- Interpretation:
    - Measures how much carbon is emitted per unit of economic value
    - Higher values indicate less efficient, more carbon-intensive economies

- Helps compare environmental efficiency across different economic scales
- Reveals the carbon cost of economic production

5. **GDP_per_energy**

- Calculation: GDP per capita / Primary energy consumption per capita
- Meaning: Economic value generated per unit of energy
- Interpretation:
  - Higher values indicate more economic productivity with less energy input
  - Reflects energy efficiency and economic innovation
  - Shows how effectively a country converts energy into economic output
  - Lower values suggest energy-intensive economic models

6. **CO2_per_area**

- Calculation: CO2 emissions / Country area
- Meaning: Emission density across geographic space
- Interpretation:
  - Higher values indicate more concentrated carbon emissions
  - Reflects local environmental pressure

- Helps understand emissions in context of land use and industrial concentration
- Useful for comparing countries with different geographic sizes

7. **Real_Purchasing_Power_GDP**

- Calculation: GDP per capita / Exchange Rate
- Meaning: Adjusted economic value accounting for currency strength
- Interpretation:
  - Provides more accurate comparison of living standards
  - Accounts for differences in international purchasing power
  - Lower values might indicate higher real purchasing power
  - Helps normalize economic comparisons across different currency contexts

8. **Exchange_Rate_Volatility**

- Calculation: Standard deviation of exchange rates over a 3-year rolling window
- Meaning: Currency stability indicator
- Interpretation:
  - Higher values suggest more unstable currency
  - Reflects economic uncertainty and market volatility
  - Indicates potential economic risk

- Helps understand financial market dynamics

9. **Economic_External_Sensitivity**

- Calculation: Exchange Rate Volatility * GDP per capita
- Meaning: Economic vulnerability to external currency shocks
- Interpretation:
  - Combines currency instability with economic size
  - Higher values indicate larger economies more exposed to international market fluctuations
  - Measures potential economic disruption risk
  - Helps assess economic resilience

10. **Exchange_Adjusted_GDP_Growth**

- Calculation: GDP growth rate - Exchange rate percentage change
- Meaning: Real economic growth isolated from currency effects
- Interpretation:
  - Removes currency fluctuation impact from growth figures
  - Provides a more accurate view of actual economic productivity
  - Helps distinguish between real economic improvement and currency-driven growth
  - Useful for understanding genuine economic development

11. **CO2_growth_rate**

- Calculation: Year-over-year percentage change in CO2 emissions
- Meaning: Emission trend and change rate
- Interpretation:
    - Positive values indicate increasing emissions
    - Negative values suggest emissions reduction
    - Helps track environmental progress
    - Reveals trajectory of carbon output

12. **GDP_growth_per_capita**

- Calculation: Year-over-year percentage change in GDP per capita
- Meaning: Economic development pace
- Interpretation:
    - Measures individual economic advancement
    - Reflects economic growth at personal level
    - Helps understand economic dynamism
    - Provides insight into living standard improvements

13. **CO2_trend**

- Calculation: Normalized emission change since first dataset year
- Meaning: Long-term emission trajectory
- Interpretation:
    - Shows how emissions have changed relative to baseline

- Provides context for emission patterns
- Helps understand long-term environmental changes
- Normalizes comparison across different starting points

14. **Renewable_adoption_rate**

- Calculation: Year-over-year percentage change in renewable energy
- Meaning: Speed of renewable energy transition
- Interpretation:
  - Positive values indicate accelerating renewable adoption
  - Reflects commitment to clean energy
  - Helps track progress in energy transformation
  - Provides insight into sustainability efforts

B. Imputation

The methodology employed for addressing missing data in a time series context involves a systematic imputation strategy that begins with the identification of columns containing missing values. This process is conducted on a country-specific basis. For each distinct country, the approach entails a thorough search for the earliest valid observation within each column that exhibits missing data. This initial data point is then utilized to fill all corresponding missing entries for that particular country. Such a strategy

ensures a localized and temporally sensitive imputation process, thereby preserving the unique characteristics inherent to each country's dataset. Consequently, this results in a complete and consistent data structure that is primed for subsequent analysis. By implementing this technique, the algorithm upholds the integrity of the original dataset while effectively addressing the challenges posed by incomplete time-series information.

7. Data Modeling

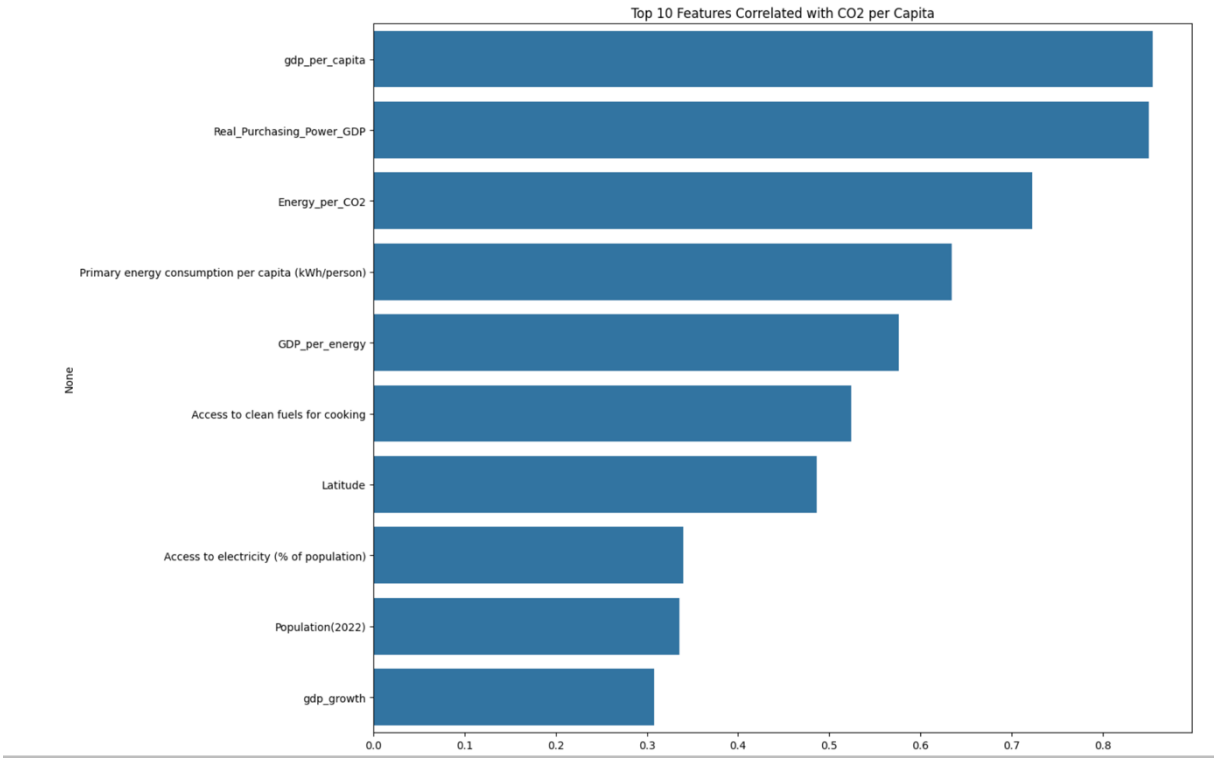
A. CO2 Per Capita Prediction Analysis

A comprehensive model was established to accurately predict per capita CO2 emissions. The dependent variable, "CO2_per_capita," was forecasted utilizing 25 features derived from a refined dataset comprising 357 samples. This dataset was partitioned into training (267 samples) and testing (90 samples) subsets.

Key Findings

The analysis of feature correlation indicated that economic indicators exert a substantial impact on CO2 emissions, with GDP per capita and Real Purchasing Power GDP exhibiting the most pronounced correlation coefficients,

approximately 0.8. Additionally, energy-related metrics, such as Energy per CO2 and primary energy consumption, displayed noteworthy correlations of 0.7 and 0.6, respectively.




```
Top 10 selected features:
1. Real_Purchasing_Power_GDP: 757.88
2. gdp_per_capita: 739.00
3. Energy_per_CO2: 305.81
4. Primary energy consumption per capita (kWh/person): 171.28
5. GDP_per_energy: 142.32
6. Access to clean fuels for cooking: 100.76
7. Latitude: 78.39
8. Population(2022): 34.94
9. Access to electricity (% of population): 34.31
10. gdp_growth: 28.14
```

The model training employed both direct feature selection and PCA approaches.

Feature selection identified 10 key predictors, while PCA reduced dimensionality from 25 to 12 components while preserving 95.6% of variance. Multiple regression algorithms were evaluated using 5-fold cross-validation, with Gradient Boosting consistently outperforming other models in both approaches.

The Gradient Boosting model using selected features achieved impressive results with a cross-validation RMSE of 39.58 and test RMSE of 29.52. Its R^2 value of 0.9936 indicates the model explains over 99% of the variance in CO2 per capita emissions.

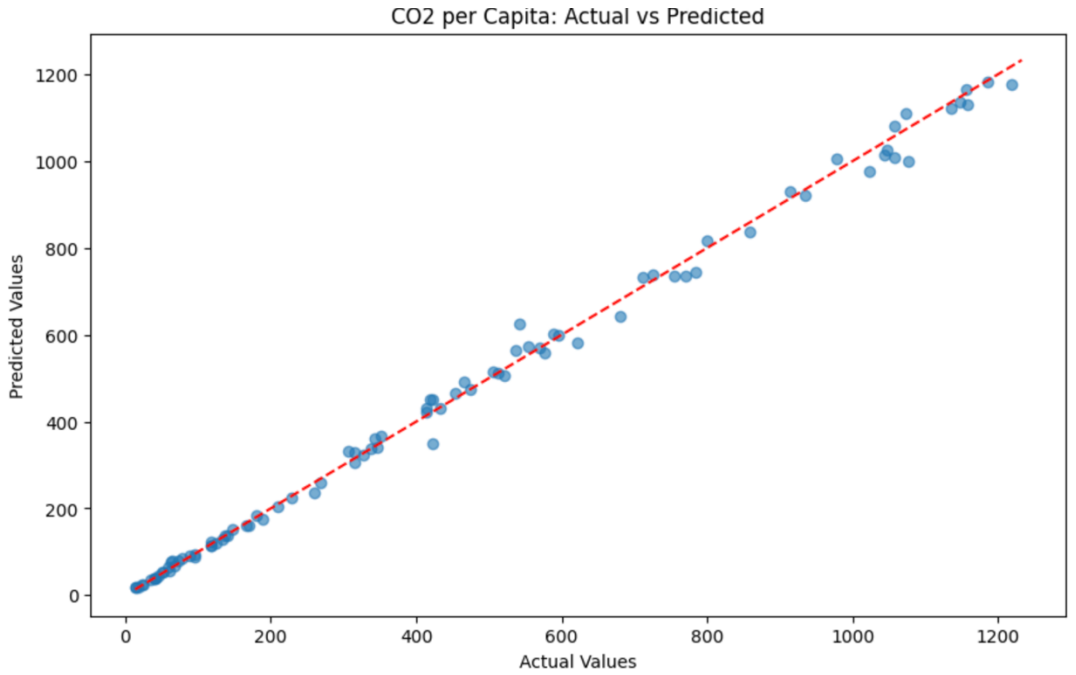
```
PCA reduced dimensions from 25 to 12 components
Variance explained: 0.9560

Training with selected features and 5-fold cross-validation:
Linear Regression    CV RMSE: 134.1485, Test RMSE: 146.1105, Test R²: 0.8428
Error training Ridge Regression with selected features: solve() got an unexpected keyword argument 'sym_pos'
Lasso Regression     CV RMSE: 134.1459, Test RMSE: 146.1013, Test R²: 0.8428
ElasticNet           CV RMSE: 133.7524, Test RMSE: 145.6563, Test R²: 0.8438
Random Forest        CV RMSE: 59.7599, Test RMSE: 64.8589, Test R²: 0.9690
Gradient Boosting    CV RMSE: 39.5776, Test RMSE: 29.5247, Test R²: 0.9936
SVR                  CV RMSE: 358.0126, Test RMSE: 376.2217, Test R²: -0.0421

Training with PCA features and 5-fold cross-validation:
Linear Regression    CV RMSE: 134.9084, Test RMSE: 142.2547, Test R²: 0.8510
Error training Ridge Regression with PCA features: solve() got an unexpected keyword argument 'sym_pos'
Lasso Regression     CV RMSE: 134.9061, Test RMSE: 142.2518, Test R²: 0.8510
ElasticNet           CV RMSE: 134.8636, Test RMSE: 142.0726, Test R²: 0.8514
Random Forest        CV RMSE: 135.6201, Test RMSE: 142.8470, Test R²: 0.8498
Gradient Boosting    CV RMSE: 103.6641, Test RMSE: 114.8925, Test R²: 0.9028
SVR                  CV RMSE: 366.4628, Test RMSE: 385.3521, Test R²: -0.0933

Best performing model (based on CV RMSE): Gradient Boosting (Selected)
CV RMSE: 39.5776
Test RMSE: 29.5247
Test R²: 0.9936
```

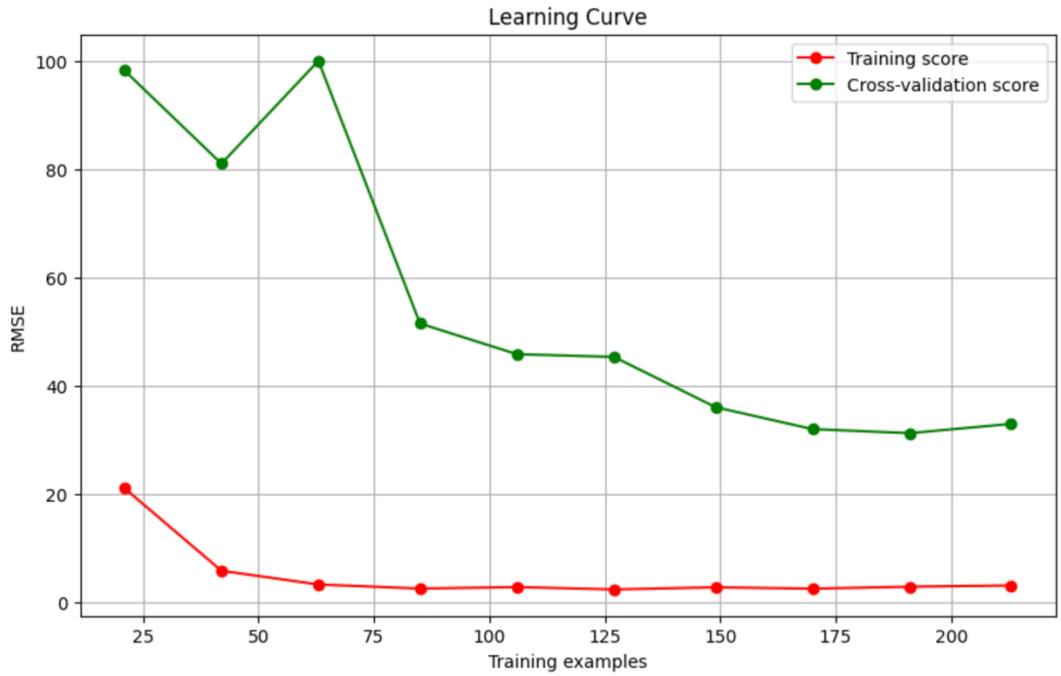
The Gradient Boosting model was tuned to prevent overfitting, achieving optimal parameters including a learning rate of 0.1, max depth of 4, and 150 estimators. Regularization techniques (min_samples_leaf=5, min_samples_split=5, subsample=0.9) enhanced generalization. Cross-validation RMSE was 30.94, while test performance improved significantly (Test RMSE: 22.49, MAE: 15.18), indicating robust out-of-sample prediction. The high R^2 score (0.9963) confirms

exceptional model fit. These results suggest the tuned model balances complexity and generalization effectively, mitigating overfitting while maintaining strong predictive accuracy on unseen data.

```
Tuning Gradient Boosting to prevent overfitting...
Optimal parameters: {'model__learning_rate': 0.1, 'model__max_depth': 4, 'model__min_samples_leaf': 5, 'model__min_samples_spl
it': 5, 'model__n_estimators': 150, 'model__subsample': 0.9}
Final CV RMSE: 30.9414
Final Test RMSE: 22.4883
Final Test MAE: 15.1813
Final Test R²: 0.9963
```

The actual versus predicted visualization demonstrates exceptional prediction accuracy across the entire range of CO2 values. The learning curve shows stable performance after approximately 150 training examples, with consistently low training error and convergence of validation error, indicating a well-balanced model that avoids both underfitting and overfitting.





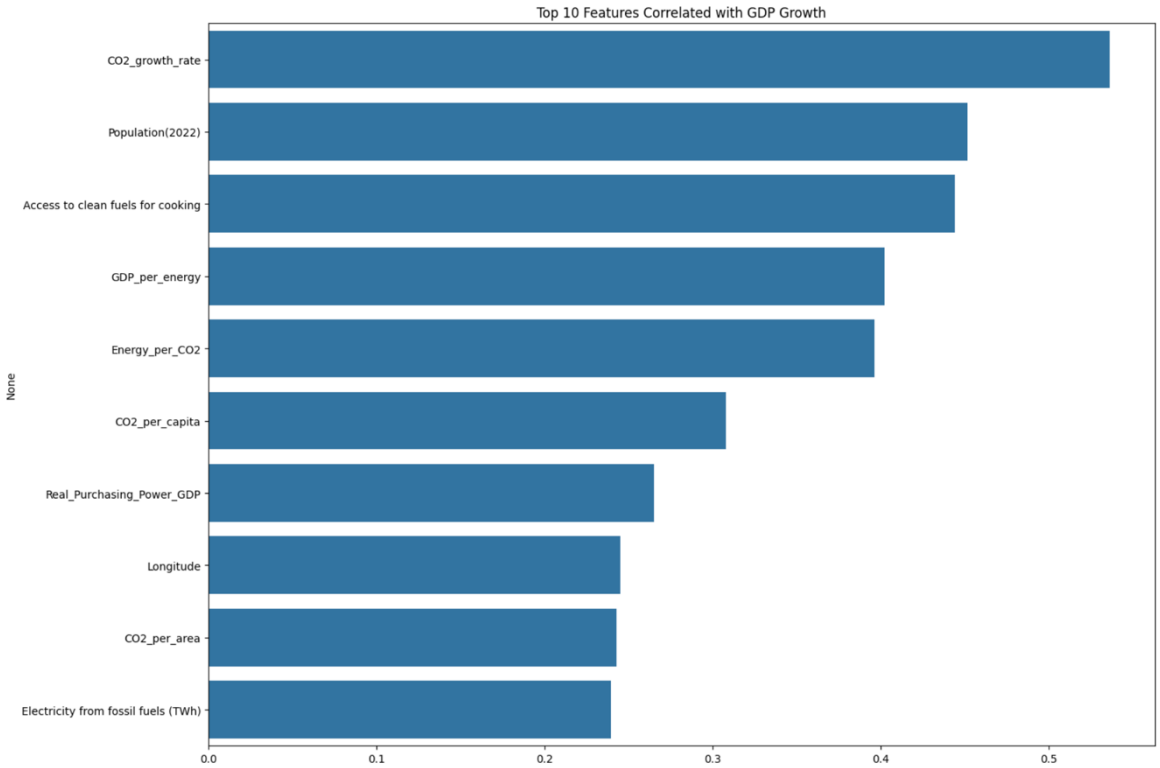
B. GDP Growth Rate Prediction Analysis

This analysis aimed to develop a machine learning model for predicting GDP growth rates using socioeconomic and environmental indicators. The model

excluded direct GDP-related features to avoid data leakage while focusing on identifying the most influential predictors.

Key Findings

The correlation analysis revealed strong relationships between GDP growth and several environmental and socioeconomic factors. CO2 growth rate emerged as the strongest predictor (correlation coefficient 0.54), followed by population size (0.45) and access to clean fuels for cooking (0.44). This suggests significant interconnections between environmental sustainability, population demographics, and economic growth.



Feature selection was performed using multiple techniques (correlation analysis, f_regression, and RFE), consistently identifying CO2_growth_rate as the most significant predictor. The model retained 10 key features while PCA reduced dimensionality from 27 to 12 components while preserving 95% of variance.

```
Top 10 features selected by f_regression:
                                Feature       Score
24                      CO2_growth_rate  121.619004
1                      Population(2022)   67.484630
6     Access to clean fuels for cooking   66.227168
16                       Energy_per_CO2   60.920979
19                       GDP_per_energy   50.331968
15                       CO2_per_capita   29.333779
21           Real_Purchasing_Power_GDP   21.465532
13                             Longitude   17.428232
7   Electricity from fossil fuels (TWh)   17.356893
20                         CO2_per_area   17.208627


Top 10 correlated features with GDP growth:
CO2_growth_rate                          0.536653
Population(2022)                         0.451729
Access to clean fuels for cooking        0.444328
GDP_per_energy                           0.402354
Energy_per_CO2                           0.396343
CO2_per_capita                           0.308024
Real_Purchasing_Power_GDP                0.265308
Longitude                                0.245327
CO2_per_area                             0.242916
Electricity from fossil fuels (TWh)      0.239722
Name: gdp_growth, dtype: float64
```

Multiple regression models were evaluated with Linear Regression achieving the best overall performance ($R^2 = 0.29$), outperforming more complex models like

Random Forest ($R^2 = 0.11$). This indicates that GDP growth has a relatively linear relationship with the selected predictors. The final tuned SVR model with RBF kernel achieved a cross-validation RMSE of 2.99 and test RMSE of 3.46, with modest predictive power ($R^2 = 0.23$).

```
Performing Recursive Feature Elimination...
Features selected by RFE: ['Area(Square kilometre)', 'Density(Square kilometre)', 'Access to electricity (% of population)',
'Renewables (% equivalent primary energy)', 'Latitude', 'Renewable_ratio', 'CO2_per_area', 'Exchange_Rate_Volatility', 'Econom
ic_External_Sensitivity', 'CO2_growth_rate']

Original features: 27
PCA components retained (95% variance): 12

Training models with selected features...
SVR: RMSE=3.3902, MAE=2.1818, R²=0.2589
Linear Regression: RMSE=3.3153, MAE=2.2730, R²=0.2912
ElasticNet: RMSE=3.3800, MAE=2.3459, R²=0.2633
Random Forest: RMSE=3.7089, MAE=2.3717, R²=0.1129
SVR with PCA: RMSE=3.3800, MAE=2.1059, R²=0.2633

Performing hyperparameter tuning for SVR...
Fitting 5 folds for each of 24 candidates, totalling 120 fits
Best parameters: {'model__C': 10, 'model__epsilon': 0.2, 'model__gamma': 'scale', 'model__kernel': 'rbf'}
Best cross-validation RMSE: 2.9872
Test set performance with tuned SVR:
RMSE: 3.4573
MAE: 2.1842
R²: 0.2292
```

Challenges and Limitations

The error analysis shows the model struggles with extreme values, particularly underestimating significant negative growth and overestimating high positive growth. The prediction errors exhibit a mostly normal distribution around zero, suggesting unbiased predictions for moderate growth rates.

Practical Applications

Simulation testing demonstrated that increasing Energy_per_CO2 by 20% could potentially improve GDP growth by approximately 0.09 percentage points,

highlighting the model's utility for scenario analysis and policy planning. The interplay between environmental efficiency and economic growth suggests that sustainable practices can positively impact economic development.

This model provides valuable insights for policymakers and economists seeking to understand growth drivers and simulate the potential economic impacts of environmental and energy policy changes.

C. High Carbon Emission Country Classification Model Analysis

This report analyzes the development and performance of a machine learning model designed to classify countries as high or low carbon emitters based on socioeconomic and geographic indicators.

Data Preparation and Target Definition

The analysis began by establishing a binary classification problem using CO2 per capita emissions as the determining factor. Countries exceeding the global mean of 406.91 tons per capita were classified as high emitters (class 1), representing 40.3% of observations, while those below were classified as low emitters (class 0), accounting for 59.7%. This slight class imbalance was addressed through balanced class weighting during model training.

Data quality assessment identified and removed 9 infinity values and 3 highly correlated features to improve model robustness. The cleaned dataset was split using stratified sampling (70% training, 30% testing) to preserve class proportions.

Feature Selection

Feature selection using mutual information identified the most discriminative predictors for high-emission classification. Geographic and demographic factors emerged as surprisingly powerful indicators, with Population, Longitude, Density, Area, and Latitude showing the highest association scores (0.58-0.56). Economic factors like GDP per capita and energy consumption indicators also proved significant predictors.

Dimensionality reduction through PCA retained 10 components, capturing 90% of variance while reducing the original 22 features.

Model Development and Evaluation

Three classification algorithms were trained and evaluated:

1. Logistic Regression: Achieved solid baseline performance (F1 score: 0.87)

2. Random Forest: Demonstrated superior performance (F1 score: 0.98)
3. Support Vector Machine: Showed excellent discrimination capability (F1 score: 0.93)

The Random Forest classifier emerged as the best-performing model with exceptional metrics:

- Cross-validation F1 score: 0.93 (±0.05)
- Test accuracy: 0.98
- Test precision: 1.00 (perfect precision - no false positives)
- Test recall: 0.95
- Area under ROC curve: 1.00 (perfect discrimination)

Hyperparameter tuning focused on preventing overfitting while maintaining performance. The optimized Random Forest used a moderate tree depth (5), required multiple samples per split/leaf, and implemented feature subsampling (log2). The tuned model achieved a slightly improved cross-validation F1 score of 0.94 and test F1 score of 0.95, with a small training-CV gap of 0.02 indicating minimal overfitting.

The confusion matrix revealed that the model correctly classified all low-emission countries and made only a few errors misclassifying high-emission countries as low-emission ones.

Insights and Implications

The model's exceptional performance suggests strong underlying patterns differentiating high and low carbon-emitting countries. Geographic factors like population density and spatial location emerged as surprisingly powerful predictors, potentially indicating regional development patterns and resource availability.

The model could support multi-faceted climate policy applications:

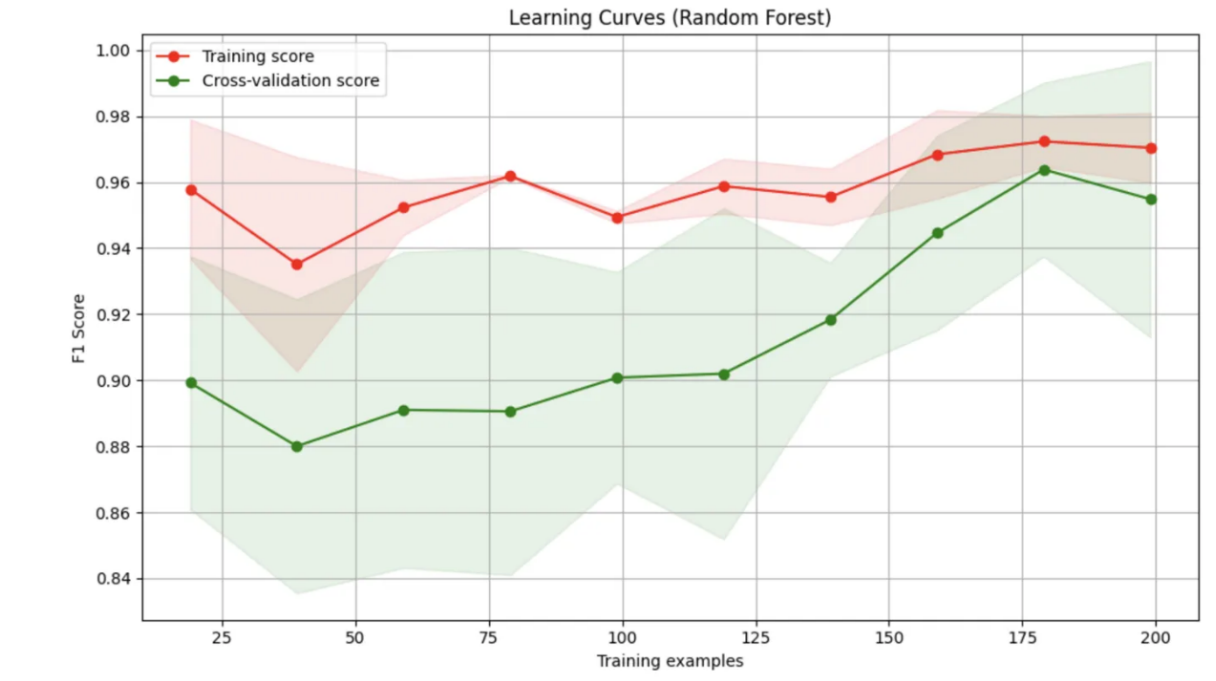
1. Identifying countries at risk of becoming high emitters
2. Simulating policy impacts on emission classification probabilities
3. Providing data-driven insights for international climate agreements
4. Targeting assistance to countries with similar profiles to high emitters

This classification model represents a valuable tool for understanding the structural factors driving high carbon emissions and could inform more effective and targeted climate change mitigation strategies.

Model Robustness and Learning Dynamics

The learning curve for the Random Forest classifier demonstrates excellent model stability as more training examples are introduced. The visualization reveals several important characteristics:

1. Both training and cross-validation F1 scores remain consistently high (above 0.88), indicating strong predictive performance throughout the training process.
2. The gap between training and cross-validation scores narrows as the training set size increases beyond 150 examples, suggesting the model generalizes well once sufficient data is available.
3. Cross-validation performance shows steady improvement from approximately 0.88 to 0.96 as training examples increase, while training performance remains relatively stable around 0.96-0.97.
4. The narrow shaded confidence intervals indicate consistent performance across different data subsets, confirming the model's reliability.



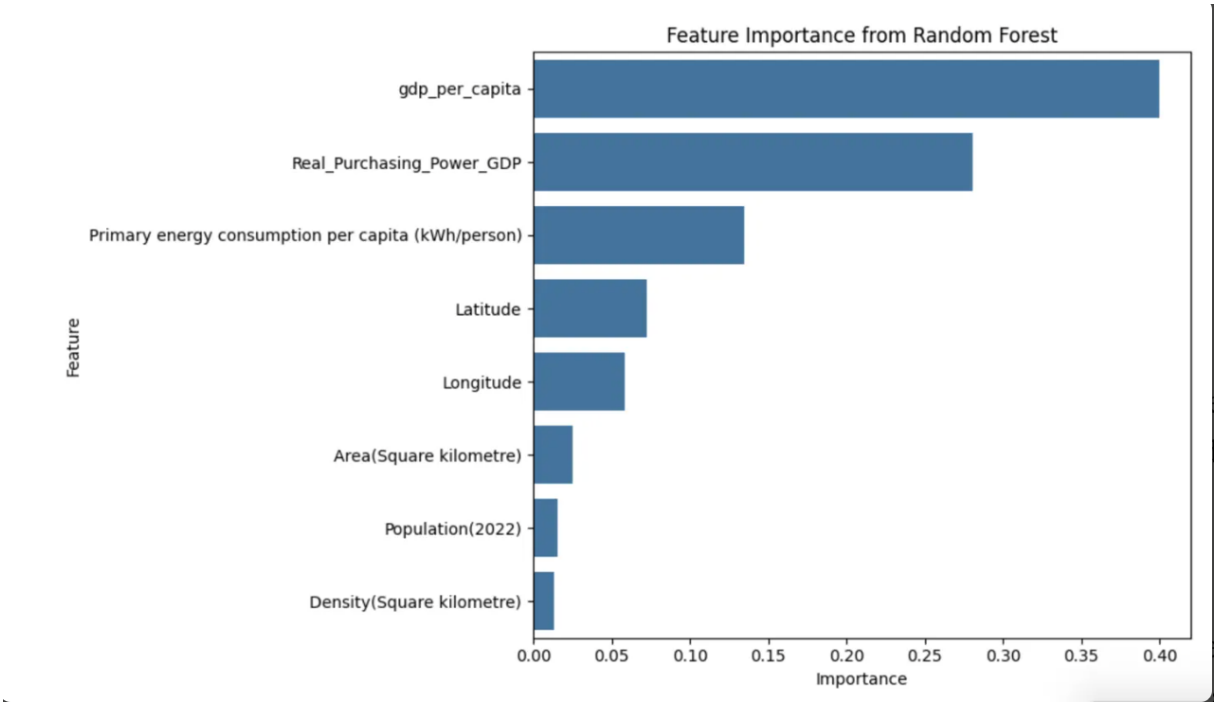
Feature Importance Analysis

The feature importance visualization and detailed ranking reveal critical insights about the drivers of high carbon emissions:

1. Economic factors dominate classification decisions:
    - GDP per capita (0.399) is by far the strongest predictor
    - Real Purchasing Power GDP (0.281) ranks second
    - Together, these economic indicators account for approximately 68% of the model's predictive power
2. Energy consumption patterns are significant:
    - Primary energy consumption per capita (0.135) shows moderate importance
    - This suggests that not just wealth but how energy is consumed plays a key role
3. Geographic factors show limited but measurable influence:
    - Latitude (0.072) and Longitude (0.059) contribute modestly
    - This likely captures regional development patterns and climate effects
    - Population size, density and land area contribute minimally (each <3%)

This feature importance hierarchy contrasts somewhat with the initial mutual information analysis that showed geographic factors as strong predictors,

suggesting the Random Forest identified more complex economic relationships during training.



D. Energy Efficiency Classification Model Analysis

This analysis explores a multi-class classification model designed to categorize countries into three energy efficiency tiers (High, Medium, and Low) based on various socioeconomic and environmental indicators. The model's purpose is to identify the key factors that distinguish between different levels of energy efficiency and provide insights for policy development aimed at improving energy utilization.

Data Preparation and Target Definition

The classification target was created by dividing countries into three approximately equal groups based on their Energy_per_CO2 ratios, yielding a balanced distribution of energy efficiency classes:

- High efficiency
- Medium efficiency
- Low efficiency

Feature Selection a

Feature selection using f_classif identified the most discriminative predictors for energy efficiency classification, revealing a significantly different set of influential factors compared to the high emission classification model:

1. **CO2_growth_rate** (score: 390.65) emerged as the dominant predictor with a score more than twice that of any other feature
2. **CO2_per_capita** (score: 172.61) ranked second
3. **gdp_per_capita** (score: 110.28) and **GDP_per_energy** (score: 104.29) showed strong influence

The prominence of CO2_growth_rate as the leading predictor suggests that the trajectory of emissions, rather than absolute levels, is most indicative of energy efficiency. This highlights a critical insight: countries actively improving their emission profiles (even if current levels are high) are likely to demonstrate better energy efficiency practices.

```
Top 10 features for energy efficiency classification:
                              Feature       Score
21                    CO2_growth_rate  390.646781
14                     CO2_per_capita  172.608987
10                     gdp_per_capita  110.276539
16                     GDP_per_energy  104.285990
18          Real_Purchasing_Power_GDP  103.668708
25                      high_emission  101.569503
17                      CO2_per_area   94.593112
5   Access to clean fuels for cooking   44.369246
9                           gdp_growth   34.014375
0                 CO2 emission (Tons)   33.481032
```

Model Development and Performance

Three classification models were evaluated:

1. **Decision Tree (Initial)**: Achieved solid baseline performance with cross-validation F1 score of 0.86 and test F1 score of 0.93

```
Training Decision Tree Classifier with cross-validation...
Cross-validation F1 scores: [0.79084249 0.81058201 0.92554657 0.88617869 0.88644197]
Mean CV F1 score: 0.8599 (±0.0508)
Test Accuracy: 0.9333
Test Precision: 0.9343
Test Recall: 0.9333
Test F1 Score: 0.9327
```

2. **Decision Tree (Tuned)**: Hyperparameter optimization with anti-overfitting focus (entropy criterion, max_depth=5, min_samples_leaf=6, min_samples_split=5) produced a model with slightly improved robustness but lower test performance (F1 score: 0.90)

```
Performing hyperparameter tuning for Decision Tree...
Best parameters: {'criterion': 'entropy', 'max_depth': 5, 'min_samples_leaf': 6, 'min_samples_split': 5}
Best cross-validation F1 Score: 0.8799
Training F1 Score: 0.9358
Training-CV gap: 0.0559

Test set performance with tuned Decision Tree:
Accuracy: 0.9000
Precision: 0.9008
Recall: 0.9000
F1 Score: 0.8999
```

3. **Random Forest**: Provided comparable performance (F1 score: 0.89) with better generalization characteristics

```
Training Random Forest for comparison...
RF Cross-validation F1 scores: [0.88696402 0.83157895 0.85195144 0.98114665 0.90380476]
RF Mean CV F1 score: 0.8911 (±0.0517)
Random Forest - Test Metrics:
Accuracy: 0.8889
Precision: 0.8924
Recall: 0.8889
F1 Score: 0.8900
```

The per-class performance metrics revealed interesting patterns:

- High efficiency class: Precision 0.97, Recall 0.94, F1 score 0.95
- Low efficiency class: Precision 0.91, Recall 1.00, F1 score 0.95
- Medium efficiency class: Precision 0.93, Recall 0.86, F1 score 0.89

The perfect recall for low efficiency countries indicates the model never misses identifying countries with poor energy practices, while the exceptional precision for high efficiency countries shows high confidence when assigning the high efficiency label.

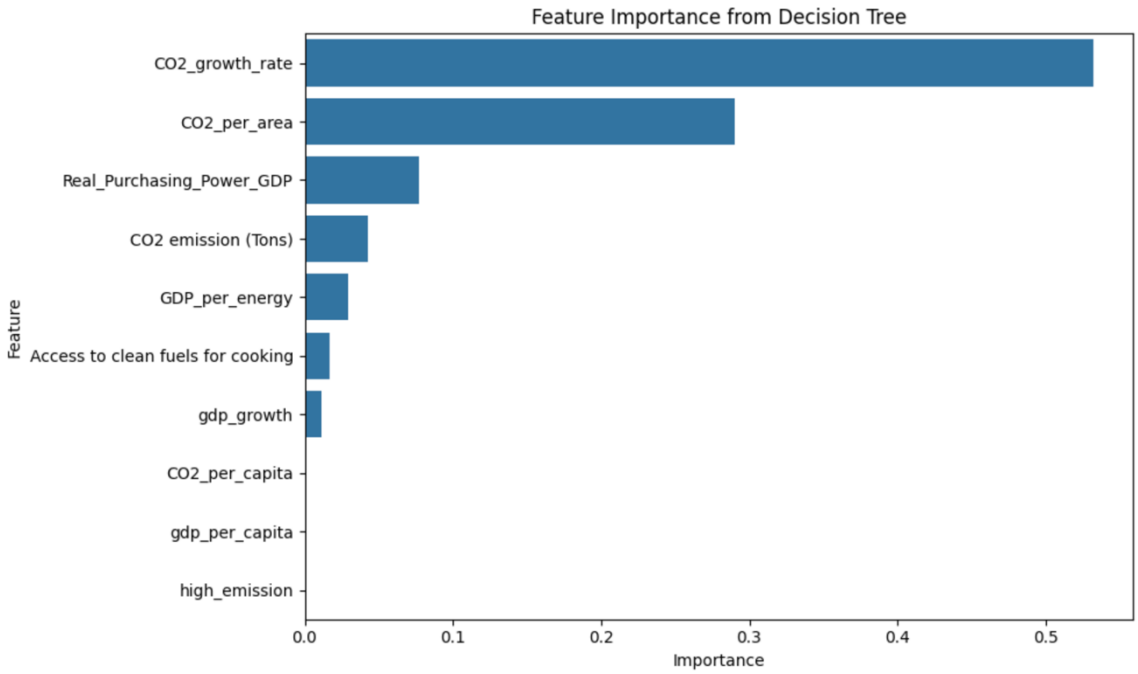
```
Classification Report:
              precision    recall  f1-score   support

        High       0.97      0.94      0.95        31
         Low       0.91      1.00      0.95        30
      Medium       0.93      0.86      0.89        29

    accuracy                           0.93        90
   macro avg       0.93      0.93      0.93        90
weighted avg       0.93      0.93      0.93        90
```

Feature Importance Analysis

The Decision Tree's feature importance analysis revealed a different pattern than the initial feature selection:

1. **CO2_growth_rate** remained dominant (importance: 0.53)
2. **CO2_per_area** showed unexpectedly high importance (0.29)
3. **Real_Purchasing_Power_GDP** (0.08) had modest influence
4. Several features including gdp_per_capita showed zero importance in the final model

This discrepancy between initial feature selection and model-derived importance suggests complex interactions between predictors that aren't captured by univariate statistical tests. The tree-based model effectively identified CO2_growth_rate and CO2_per_area as the critical decision factors for classification.



Economic and Environmental Relationships

The visualization of GDP per capita versus energy consumption by efficiency class revealed distinctive patterns:

1. **High efficiency countries** showed diverse economic development levels but generally maintained lower energy consumption per unit of GDP
2. **Medium efficiency countries** clustered in the middle ranges of both GDP and energy consumption
3. **Low efficiency countries** showed a more scattered pattern with several outliers having high energy consumption relative to economic output