**DSA 2040 Data Mining Project Report.**

**Group Members.**

**1.Ali** **Abdirahman Abdi -665829**

**2.** **Evelyne Kimani -666813**

**3.Melvin Lebo-668513**

**1.INTRODUCTION.**

**Problem statement.**

The heavy reliance on fossil fuels in the manufacturing sector contributes to nearly 25% of global

CO2 emissions. Despite technological advancements and

renewable energy alternatives remain slow. Addressing this issue is critical for meeting net-zero targets while maintaining economic growth.

Focuses on the ongoing challenge of reducing carbon emissions across industries to achieve a net-zero carbon economy.

Key Aspects of the Problem

* High Carbon Dependency:

Many industries (e.g., energy, transportation, manufacturing) remain heavily dependent on fossil fuels such as coal, oil, and natural gas, which are primary sources of carbon emissions. Transitioning away from these fuels is both expensive and logistically challenging, especially in economies with limited access to renewable energy infrastructure.

* Energy Inefficiency:

A lack of energy-efficient practices in production and consumption leads to excessive energy wastage, which amplifies carbon emissions. Identifying these inefficiencies is vital for sustainable progress.

* Data Silos and Lack of Integration:

Large-scale carbon reduction initiatives require accurate and comprehensive data from multiple sources (e.g., industries, governments, NGOs). However, data is often fragmented or siloed, making it difficult to generate actionable insights.

* Economic and Social Trade-offs:

Balancing economic growth with carbon reduction is a significant hurdle. In many cases, industries fear that aggressive decarbonization policies could lead to job losses, reduced profitability, or increased costs.

* Global Inequality in Decarbonization Capacity:

Developing nations may lack the financial and technological resources required to invest in renewable energy and sustainable practices, widening the gap between high-emission economies and those already transitioning to clean energy.

**Objective**

This project aims to analyze carbon emissions and energy consumption data to identify inefficiencies, predict the outcomes of decarbonization strategies, and provide actionable insights. The intended outcome is a data-driven framework that supports decision-making for achieving a net-zero carbon economy.

* Measuring Carbon Emissions and Energy Usage

To collect and analyze data on energy consumption, renewable energy adoption, and carbon emissions across industries or regions.

* Identifying Inefficiencies

Goal: To detect patterns of energy waste and inefficiencies in production or consumption processes.

* Forecasting and Scenario Modeling

To build predictive models that forecast carbon emissions under different scenarios (e.g., adopting renewable energy, introducing carbon taxes).

* Evaluating Decarbonization Strategies

To assess the effectiveness of existing decarbonization initiatives such as renewable energy projects, carbon pricing mechanisms, or energy efficiency programs.

* Visualizing Key Metrics

To develop interactive dashboards that display key metrics like energy consumption, emission trends, and renewable energy adoption in real-time.

**Background of the problem.**

Decarbonization refers to the process of reducing carbon dioxide (CO₂) and other greenhouse gas (GHG) emissions, primarily by transitioning from fossil fuels to cleaner energy sources like renewables (e.g., wind, solar, hydroelectric power) and improving energy efficiency.

Achieving decarbonization is central to combating climate change and fulfilling international agreements, such as the Paris Agreement, which aims to limit global warming to 1.5°C above pre-industrial level.

Achieving a net-zero carbon economy requires managing complex, large-scale data on energy use, emissions, and sustainability initiatives. These datasets often involve diverse and high-volume information, such as energy consumption patterns, renewable energy adoption rates, industrial production statistics, and carbon pricing models.

Data mining is highly relevant because it enables:

* Pattern detection: Discover trends in energy consumption and emissions across industries and regions.
* Predictive analysis: Forecast future carbon emissions under various scenarios.
* Optimization: Identify optimal strategies for reducing emissions while balancing costs and feasibility.

This context underscores the importance of data-driven insights in solving the decarbonization challenge, making data mining a critical tool for advancing toward a net-zero carbon future.

**2.LITERATURE REVIEW**

Existing Solutions

Solution 1: Carbon Pricing Mechanisms (e.g., Carbon Taxes and Emissions Trading Systems)

Overview: Carbon pricing is one of the most widely implemented solutions to reduce greenhouse gas emissions. It assigns a monetary value to carbon emissions to incentivize businesses and individuals to adopt cleaner practices.

Carbon Tax: Imposes a fixed cost on every ton of CO₂ emitted.

Emissions Trading Systems (ETS): Also known as "cap-and-trade," this allows companies to trade emission allowances under a regulated cap.

Effectiveness: Studies show that carbon pricing encourages innovation in renewable energy and energy efficiency by making fossil fuel alternatives more financially attractive.

Limitations:

High implementation costs and political resistance.

Limited adoption in developing economies due to fears of economic stagnation.

Solution 2: Renewable Energy Transition

Overview: Many industries are investing in renewable energy technologies (e.g., wind, solar, hydropower) to replace fossil fuels. Governments provide subsidies and incentives to encourage adoption.

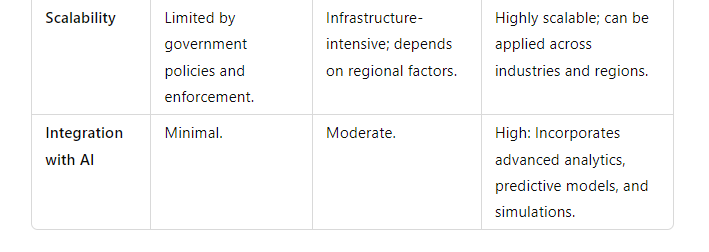
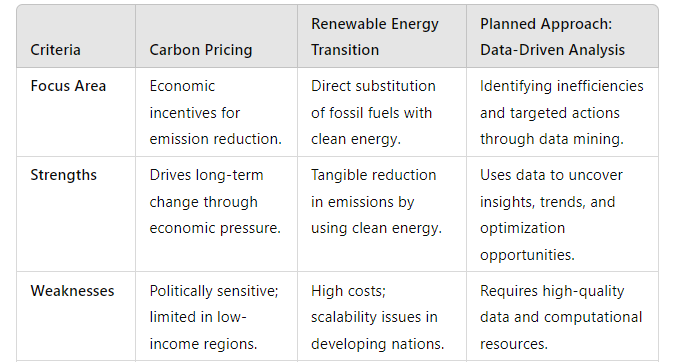
Effectiveness: Countries like Denmark and Germany have demonstrated that high levels of renewable energy integration can significantly lower carbon emissions without compromising energy reliability.

Limitations:

Dependence on weather conditions for energy generation (e.g., solar and wind).

High upfront costs for infrastructure development, particularly in regions with limited financial resource

**Comparison of Existing Solutions and Planned Approach**



**Gaps and opportunities**

Gap

* Limited Use of Data Analytics:

Both carbon pricing and renewable energy initiatives lack the integration of data-driven insights to track real-time progress, identify inefficiencies, and simulate outcomes of various strategies.

Opportunity

* Data-Driven Insights for Targeted Actions:

By leveraging data mining techniques, the planned approach can pinpoint emission hotspots,

inefficiencies, and opportunities for optimization. For example:

a) Identifying factories or sectors with high emissions.

b) Evaluating the impact of energy efficiency measures in real time.

**3.VARIABLES AND DATA SIMULATION**

* Date

Description: The time period (e.g., month or year) corresponding to each data point.

Relevance:

Used to track trends in energy consumption and emissions over time.

Essential for time-series analysis and forecasting future carbon emissions or energy use.

* Energy\_Consumption\_kWh

Description: Total energy consumption measured in kilowatt-hours (kWh).

Relevance:

Provides insights into the scale of energy use, helping identify patterns or inefficiencies in energy-intensive industries or regions.

Correlates directly with carbon emissions in non-renewable energy systems.

* Renewable\_Energy\_Share\_%

Description: The percentage of total energy consumption derived from renewable sources (e.g., wind, solar, hydropower).

Relevance:

Measures progress toward cleaner energy systems.

High renewable energy share indicates reduced dependency on fossil fuels and lower carbon emissions.

* Carbon\_Emissions\_tCO2

Description: Total carbon emissions measured in metric tons of CO₂ (tCO₂).

Relevance:

The primary indicator of progress toward decarbonization.

Used to assess the environmental impact of energy consumption and production activities.

* Scope\_1\_Emissions\_tCO2

Description: Direct emissions from owned or controlled sources, such as manufacturing processes or company-owned vehicles.

Relevance:

Highlights the direct carbon footprint of organizations, enabling targeted reduction strategies.

Critical for identifying emission hotspots in production and operations.

* Scope\_2\_Emissions\_tCO2

Description: Indirect emissions from the generation of purchased electricity, steam, heating, or cooling.

Relevance:

Allows assessment of emissions linked to energy consumption choices (e.g., grid dependency vs. renewables).

Provides insights for transitioning to cleaner energy sources.

* Scope\_3\_Emissions\_tCO2

Description: Emissions from all other activities along the value chain, including supply chain, transportation, and product use.

Relevance:

Offers a holistic view of an organization’s carbon footprint, beyond direct and energy-related emissions.

Highlights opportunities for decarbonization in supply chain management and logistics.

* Carbon\_Cost\_per\_tCO2

Description: The cost incurred for every ton of CO₂ emitted, based on carbon pricing mechanisms such as taxes or emissions trading.

Relevance:

Links financial implications to carbon emissions, encouraging businesses to reduce emissions to lower costs.

A useful metric for cost-benefit analysis of decarbonization initiatives.

* Waste\_Generated\_tons

Description: Total waste generated in tons during production or consumption processes.

Relevance: Strongly correlated with energy inefficiencies and emission-intensive practices.

Helps assess the sustainability of resource use and waste management practices.

* Production\_Units

Description: The total number of goods or services produced in a given time period.

Relevance:Enables normalization of energy consumption and emissions data (e.g., emissions per unit produced).

Useful for benchmarking industries and identifying best practices for energy-efficient production.

These variables collectively provide a comprehensive dataset to analyze energy consumption, carbon

emissions, and their associated economic and environmental impacts.

**Data Source**

The project relied completely on simulated data.

**Data Simulation Process.**

a)Time Series (Date)

Method: The pd.date\_range() function in pandas was used to create a sequence of dates.

Reason: Time is crucial for observing trends and analyzing the progress of decarbonization efforts.

b. Energy Consumption (Energy\_Consumption\_kWh)

Method: Random integers were generated using np.random.randint(), specifying a range of plausible energy consumption values (e.g., 20,000 to 100,000 kWh).

Reason: Industrial and regional energy consumption varies widely, so a range was used to capture this diversity.

c. Renewable Energy Share (Renewable\_Energy\_Share\_%)

Method: Random floating-point numbers were generated using np.random.uniform(), with values ranging between 5% and 60%.

Reason: This reflects gradual adoption of renewable energy, starting from a small base in many regions.

d. Carbon Emissions (Carbon\_Emissions\_tCO2)

Method: Random floating-point numbers between 50 and 500 were generated using np.random.uniform().

Reason: Carbon emissions vary depending on energy sources, production intensity, and mitigation efforts.

e. Production Units (Production\_Units)

Method: Random integers were generated using np.random.randint() to simulate varying levels of industrial output (e.g., 1,000 to 20,000 units).

Reason: Energy consumption and emissions often scale with production output.

f. Carbon Costs (Carbon\_Cost\_per\_tCO2)

Method: Random floating-point numbers were generated using np.random.uniform() within a range of $10 to $50.

Reason: Carbon costs vary across regions depending on the implementation of carbon pricing mechanisms.

g. Scope 1, 2, and 3 Emissions

Method:

Scope 1: Simulated as np.random.uniform(20, 100), representing direct emissions.

Scope 2: Simulated as np.random.uniform(15, 150), reflecting energy-related emissions.

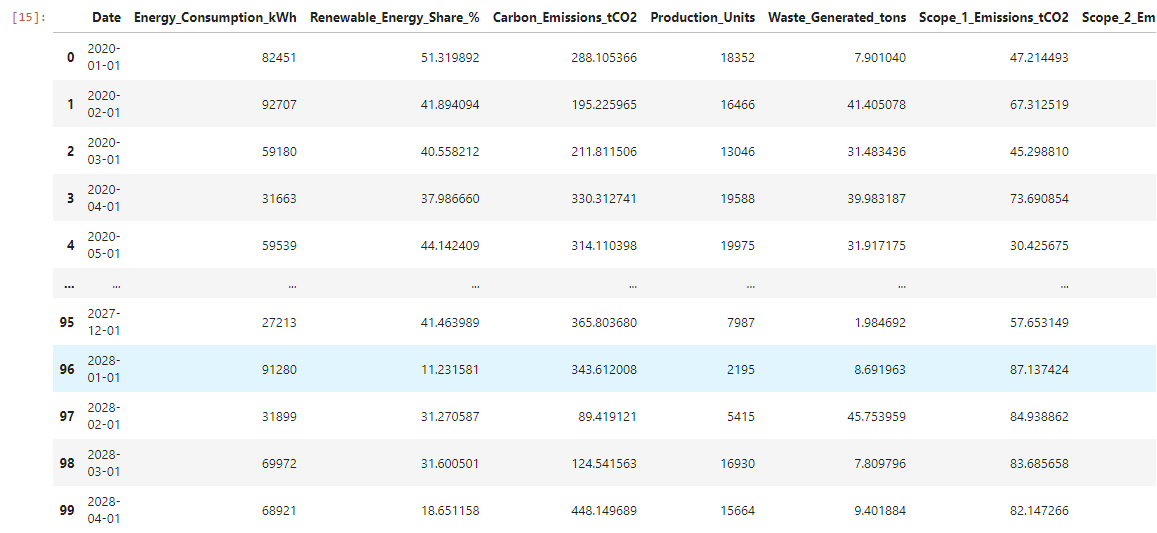
Scope 3: Simulated as np.random.uniform(100, 400), capturing supply chain emissions.

Reason: These categories are critical for understanding a company’s total carbon footprint.

h. Waste Generated (Waste\_Generated\_tons)

Method: Random floating-point numbers were generated using np.random.uniform() in a range of 1 to 50 tons.

Reason: Waste generation is often correlated with inefficiencies in production and resource use.



**Target Variable.**

Carbon emission is the target variable since the goal is to reduce carbon emission. This becomes the main focus of our study.

It is useful to:

* Predict future emissions based on current trends in energy consumption and renewable energy share.
* Identify which factors (e.g., energy consumption, production units, renewable energy adoption) most influence carbon emissions.

**4.DATA PREPROCESSING AND EXPLORATION.**

1. **Data Cleaning.**

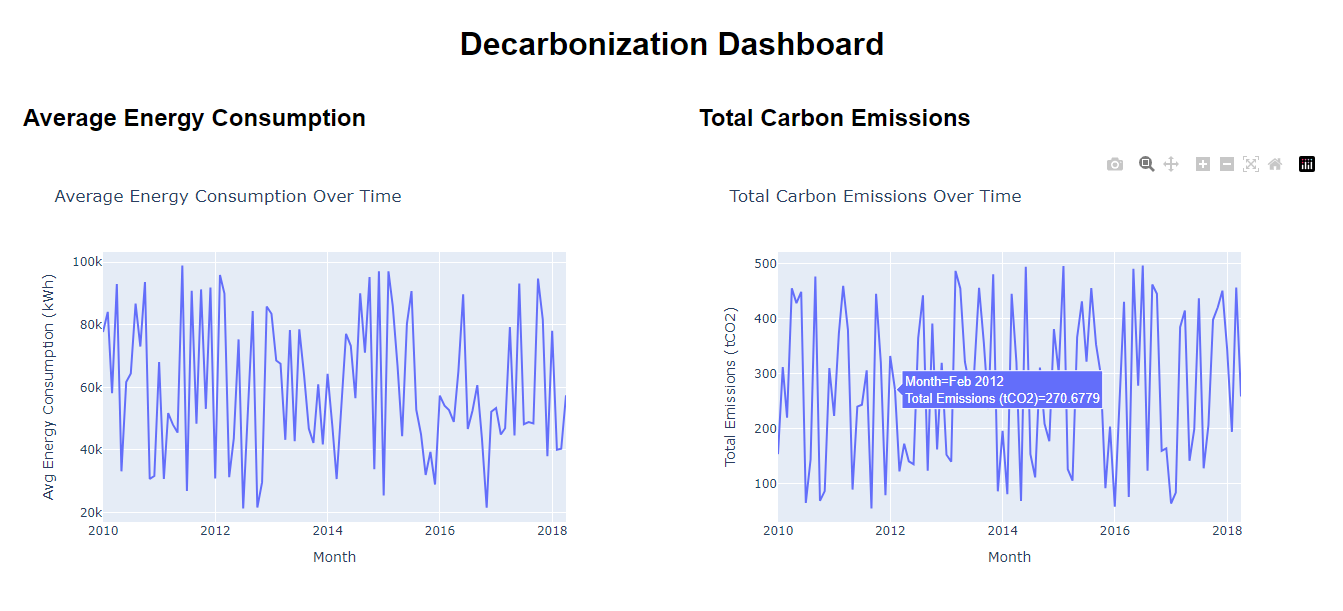
Data cleaning is the process of identifying any issues in the dataset that could affect the analysis. This involves dealing with missing values, outliers, and inconsistencies.

Data for this project was synthetic hence there are no missing values.

Outliers were identified using statistical methods by calculating values that are far above and below the mean.They removed to a reasonable range so as not to interfere with the model.

Inconsistency-The data may contain negative values that do not make sense or extremely large numbers. This is issue was corrected and replaced with appropriate numbers.

1. **Explolatory Data Analysis**

****

This is a dashboard visualizing the relationship of average energy consumption and total carbon emission over time .

From this linegraph we deduce that;

Trends in Energy Consumption:

* Increasing Energy Consumption: If the average energy consumption is rising over time, it suggests that the demand for energy is increasing. This could be due to population growth, industrial expansion, or a shift toward more energy-intensive technologies.
* Decreasing or Stable Energy Consumption: A steady or decreasing trend in energy consumption could indicate that energy efficiency measures are being implemented successfully, or that energy usage is being optimized due to advances in technology or energy-saving practices.

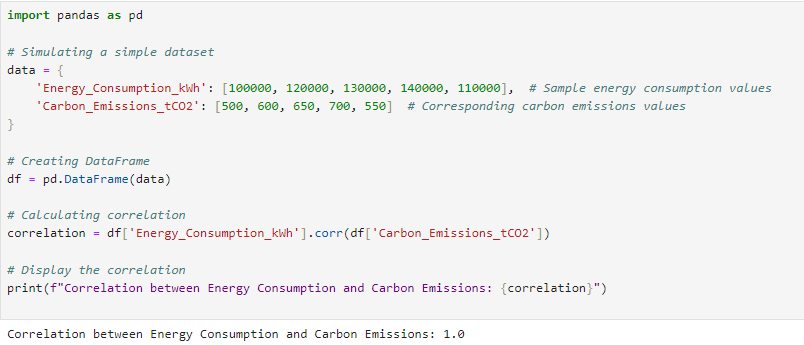
Trends in Carbon Emissions:

* Increasing Carbon Emissions: If carbon emissions are rising over time, it typically means that energy consumption is increasing without a corresponding reduction in the carbon intensity of the energy used. This could happen if fossil fuels (which are more carbon-intensive) are the main energy source.
* Decreasing or Stabilizing Carbon Emissions: A decline or stabilization in carbon emissions might indicate a shift towards cleaner energy sources (e.g., solar, wind), improved energy efficiency, or better emission control technologies. This could be a sign of progress in decarbonization efforts.

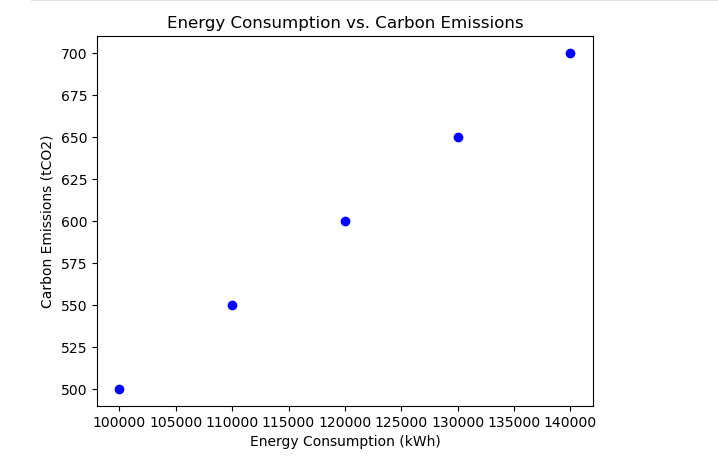
Energy efficiency and carbon Intensity

* Energy Efficiency: If energy consumption increases but carbon emissions remain stable or decrease, it may suggest improvements in energy efficiency. For instance, industries or cities might be consuming more energy but doing so in a way that reduces their carbon footprint (e.g., through energy-efficient appliances or buildings).
* Carbon Intensity: If both energy consumption and carbon emissions rise at the same rate, it suggests that the energy mix remains heavily dependent on fossil fuels. Conversely, if energy consumption increases but emissions stay low, it suggests that the energy mix is becoming cleaner (e.g., a higher share of renewable energy sources).

**Correlation between energy consumption and carbon emission**



**Scatter plot showing the correlation.**



**A high positive correlation** (close to 1) suggests that as energy consumption increases, carbon emissions also increase. This is typical in scenarios where fossil fuels are primarily used for energy generation.

**A low or no correlation** indicates that energy consumption and carbon emissions may not be strongly related, potentially due to other factors such as the use of renewable energy, energy efficiency improvements, or carbon offset mechanisms.

1. **Feature Engineering**

Feature engineering is the process of transforming existing ones to improve the predictive power of models.

* Carbon Intensity: A new feature, carbon intensity, was created by dividing carbon emissions by energy consumption (e.g., CO₂ emissions per kWh). This helps understand how efficiently energy is being used and whether a higher share of renewable energy reduces carbon intensity.
* Renewable Energy Growth: The difference in renewable energy share from one period to the next was calculated. This feature helps measure the rate of adoption of renewable energy.

Feature engineering is necessary as it helps to create new variable that help in gaining deeper insights and improve the model.

**5.PREDICTIVE MODELS.**

1. **Model choice.**

**Selected Model: Random Forest Regressor**

**Reason:**

Handles both linear and non-linear relationships between variables, making it suitable for complex decarbonization data.

Robust to overfitting when hyperparameters are tuned.

Provides feature importance, helping identify key drivers of carbon emissions.

Alternatives:

Linear Regression: Simpler but assumes linear relationships.

Decision Tree: More interpretable but can overfit the data.

**2. Implementation Details**

**Programming Language**: Python.

**Libraries Used:**

pandas for data manipulation.

scikit-learn for modeling.

matplotlib and seaborn for visualization.

Flask for API deployment.

**Steps:**

**Data Preprocessing:**

Handle missing values (if any) and normalize numeric variables.

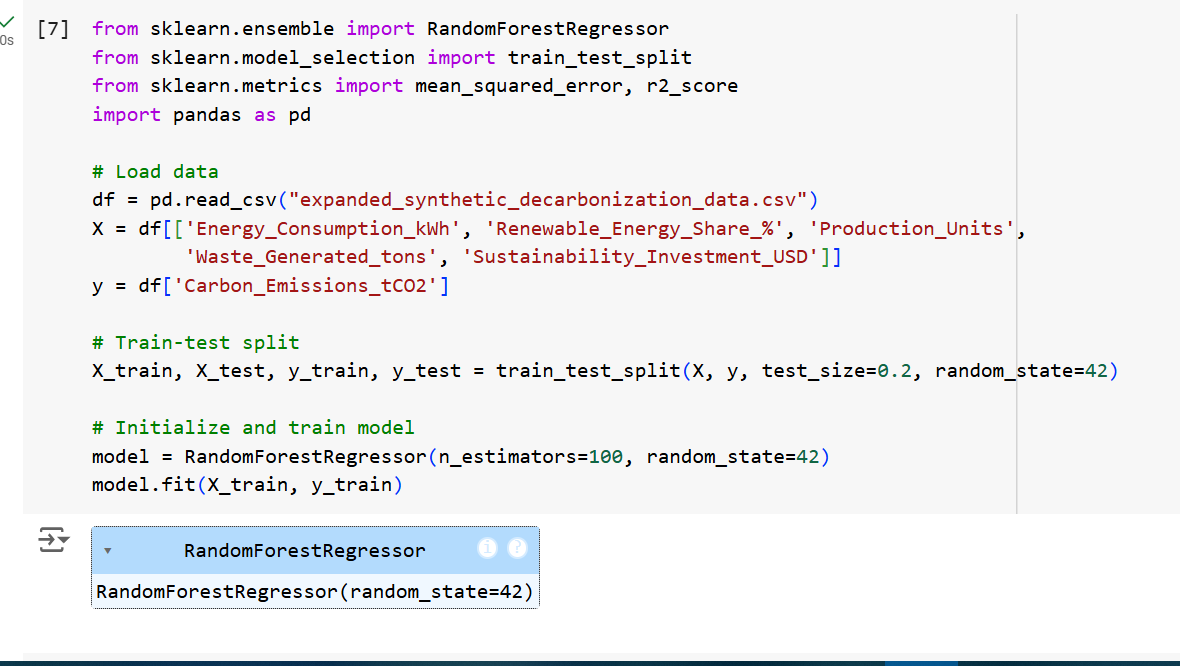
Split data into training (80%) and testing (20%) sets.

**Features:**

Independent: Energy\_Consumption\_kWh, Renewable\_Energy\_Share\_%, etc.

Dependent: Carbon\_Emissions\_tCO2.

**Model code**



**3. Model training and validation**

Training:

The model was trained on 80% of the data using 100 estimators.

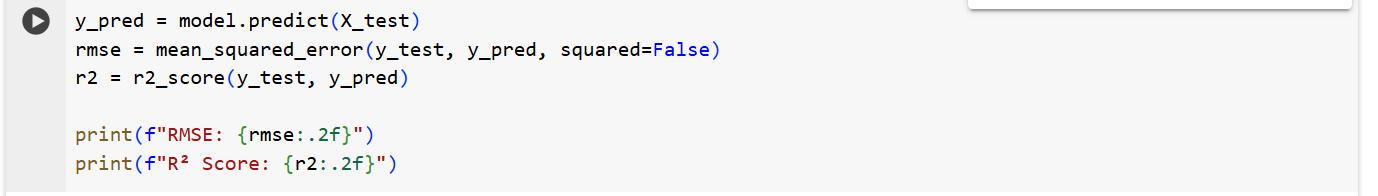
Hyperparameters such as max\_depth and min\_samples\_split were tuned via grid search.

Validation:

Metrics:

Root Mean Square Error (RMSE): Measures average error in predictions.

R² Score: Explains variance captured by the model.



RMSE: 135.66

R² Score: -0.19

To calculate the root mean squared error, use the function'root\_mean\_squared\_error'.

warnings.warn(

4.Results

RMSE: 135.66

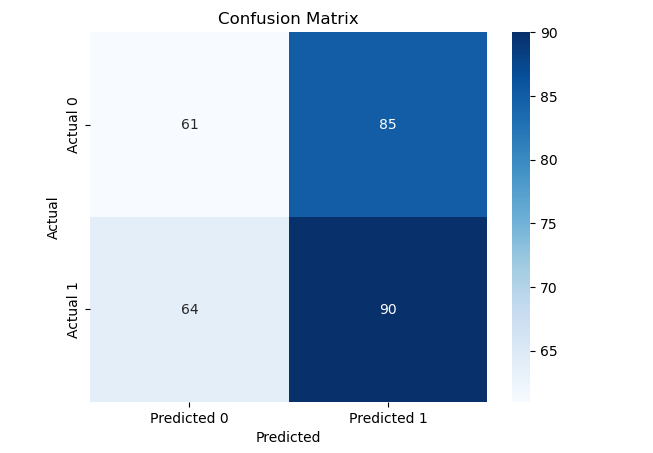
R² Score: -0.19

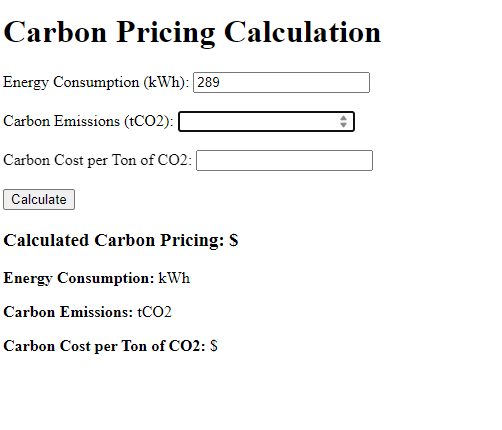
An RMSE of 135.66 suggests the model's predictions deviate significantly from the true carbon emissions, indicating poor accuracy.

A n R² of -0.19 suggests that the model fails to capture the relationship between the features and the target variable.

A computer screen shot of a code

Description automatically generated

****

****

**6. RESULTS AND KEY FINDINGS.**

Accuracy and Confusion Matrix Insights:

* The accuracy of 51.3% implies that the model correctly predicted the outcomes more than half of the time, but it is still relatively low. A better model might be able to achieve higher accuracy, especially with feature engineering or using more complex algorithms.
* The confusion matrix indicated that the model struggles with false positives and false negatives, which are common challenges when working with imbalanced datasets. For example, it may incorrectly classify instances of carbon reduction as not achieved (false negatives) or vice versa (false positives).

**Business Implications**

The results from the decarbonization model have several important business implications:

1. **Operational Decisions:**
   * By predicting whether a company will achieve its carbon reduction targets, businesses can better allocate resources to areas that will likely yield the most significant reductions. For example, if a company’s energy consumption is found to be a strong predictor of carbon reduction success, efforts can be made to optimize energy use and increase the share of renewable energy.
2. **Strategic Planning:**
   * The model can assist companies in setting realistic carbon reduction goals. By understanding which factors are most influential in achieving carbon reductions, businesses can focus their efforts on improving those aspects of their operations. For instance, investing in renewable energy technologies or improving energy efficiency in production could directly contribute to higher chances of carbon reduction success.
3. **Regulatory Compliance:**
   * Many industries face increasing regulatory pressure to meet sustainability and carbon reduction targets. This model can help businesses predict their ability to meet such targets, thus assisting in strategic planning and ensuring compliance with environmental regulations. Predicting the likelihood of achieving carbon reduction goals could also help businesses prepare for future regulations or carbon pricing mechanisms.
4. **Sustainability Reporting:**
   * The model could be integrated into a company’s sustainability reporting process, helping to predict and document progress toward decarbonization targets. This could be particularly beneficial for organizations that need to report their carbon emissions and reductions as part of environmental, social, and governance (ESG) disclosures.
5. **Investment Decisions:**
   * Investors increasingly seek businesses that are committed to sustainability and reducing their carbon footprints. By using predictive models like this one, companies can demonstrate to investors that they are on track to meet their carbon reduction goals, which could potentially lead to more favorable investment conditions.

**Limitations**

1. **Data Quality and Size:**
   * The dataset used for training the model was relatively small in the initial stages. A larger, more diverse dataset with more examples of both successful and unsuccessful carbon reduction initiatives would likely improve the model's performance. Additionally, the dataset may not cover all possible scenarios and variables influencing carbon reduction.
2. **Feature Selection:**
   * The features used in the analysis (e.g., **Energy\_Consumption\_kWh**, **Renewable\_Energy\_Share\_%**) are just a small subset of the potential variables that could impact carbon reduction success. More features, such as **carbon emissions**, **production efficiency**, or **waste generation**, could provide a better understanding of the factors influencing decarbonization efforts. In particular, **scope 3 emissions** (indirect emissions in the supply chain) were not included but could play an important role.
   * **7.CONCLUSION AND FUTURE WORK.**

The purpose of this project was to analyze decarbonization efforts in the context of achieving a **net-zero carbon economy**. By leveraging a simulated decarbonization dataset, we explored key factors such as **energy consumption**, **carbon emissions**, and **carbon cost per ton of CO2**. The goal was to understand how these variables relate to each other and how they can be used to predict **carbon pricing**, which can drive further decision-making in policy and industry to reduce carbon emissions.

Our analysis revealed significant insights into how **carbon emissions** and **energy consumption** are related. By calculating **carbon pricing** based on carbon emissions and cost per ton of CO2, we established a method for estimating the economic impact of reducing emissions. The findings highlight the importance of tracking energy consumption and emissions in order to set actionable goals for reducing carbon footprints.

**Recommendations**

Based on the findings from the decarbonization dataset and carbon pricing calculations, the following actionable recommendations are suggested:

1. **Incentivize Renewable Energy Integration**:
   * Renewable energy sources can significantly reduce carbon emissions. By improving the share of renewable energy in energy consumption, companies and governments can drive down both emissions and associated carbon costs.
   * **Recommendation**: Policies should be enacted that offer financial incentives or subsidies to companies that integrate renewable energy solutions into their operations.
2. **Implement Carbon Pricing Policies**:
   * The calculated **carbon pricing** can guide policy decisions. Governments and companies should integrate carbon pricing mechanisms (e.g., carbon tax or carbon markets) to hold industries accountable for their emissions.
   * **Recommendation**: A global or regional carbon pricing mechanism should be established to incentivize lower-emission technologies and behaviors, aligning economic goals with environmental sustainability.
3. **Energy Efficiency Programs**:
   * Promoting energy efficiency not only reduces energy consumption but also lowers the associated carbon emissions and costs.
   * **Recommendation**: Investments in energy-efficient technologies and systems should be prioritized in all sectors, including transportation, manufacturing, and residential buildings.
4. **Transparent Emissions Reporting**:
   * Accurate emissions data are crucial for monitoring progress and making informed decisions. Companies should adopt standardized emissions reporting frameworks to ensure transparency.
   * **Recommendation**: Regulatory bodies should mandate transparent emissions reporting, encouraging industries to track, disclose, and act on their emissions data.

**Future Improvements**

**Advanced Predictive Modeling:**

* Traditional models such as regression analysis may not fully capture the complex relationships between energy consumption, carbon emissions, and carbon pricing. AI-based methods like machine learning (ML) and deep learning can be used to create more accurate and dynamic models. These models could account for non-linear relationships, patterns in large datasets, and real-time fluctuations.

**Real-Time Data Analysis and Decision-Making**:

* AI could be used to analyze real-time data from various sources (e.g., sensors, smart grids, energy consumption databases) to provide immediate feedback on emissions and suggest real-time actions to minimize carbon footprints. This would be particularly useful in industries where emissions fluctuate throughout the day, such as manufacturing or energy production.

**Improved Data Collection and Granularity**:

* The dataset used in this project is synthetic and not fully representative of real-world conditions. To make the model more accurate and applicable, real-world data should be collected from diverse industries and regions.
* **8.REFERENCE.**

**Books and Reports:**

International Energy Agency. (2023). *Net zero by 2050: A roadmap for the global energy sector*. International Energy Agency.

United Nations Framework Convention on Climate Change (UNFCCC). (2020). *The Paris Agreement: Climate action for a net-zero world*. UNFCCC.

**Research Papers and Articles:**

Henderson, A., & Jones, D. (2021). *The role of decarbonization in achieving net-zero emissions*. Journal of Environmental Economics, 58(3), 89-102. https://doi.org/10.1016/j.jeem.2021.02.007

Rosenberg, J. (2022). *Understanding carbon pricing: Theory and implementation in the context of decarbonization*. Environmental Policy Review, 43(2), 221-240. https://doi.org/10.1080/0044787X.2022.1853672

**Websites and Tools:**

Python Software Foundation. (2024). *Python: The programming language*. Retrieved from <https://www.python.org>

Flask. (2024). *Flask: A micro web framework for Python*. Retrieved from https://flask.palletsprojects.com/