**U.S. Carbon Dioxide Emissions Analysis**

**1. Problem Definition and Objectives**

**Problem Statement:**

Identifying high-risk U.S. states for carbon emissions and the factors contributing to these emissions.

**Objectives:**

1. Analyze historical emissions trends from 1970 onwards.
2. Identify key contributors by state, sector, and fuel type.
3. Predict future emissions using machine learning techniques.

**Significance:**

Understanding carbon dioxide emissions patterns helps policymakers target high-risk areas and sectors, transition to cleaner energy sources, and evaluate the effectiveness of emission-reduction policies.

**2. Questions Posed**

1. Which states have the highest total carbon dioxide emissions?
2. What are the trends in carbon dioxide emissions over time?
3. Which sector contributes the most to emissions in the top five high-emission states?
4. How do emissions from different fuel types (coal, petroleum, natural gas) vary across states?
5. Is there a correlation between a state’s total emissions and its industrial emissions?
6. Can we predict future emissions levels for high-risk states based on historical trends?
7. What factors most strongly predict a state’s total emissions?
8. Can we cluster states based on similar emissions patterns and sector contributions?

**3. Data Source and Description**

**Dataset:**

**U.S. Carbon Dioxide Emissions by State, Sector, and Fuel Type**

**Key Features:**

* **year:** Year of emissions record (1970 to 2021).
* **state-name:** Name of the U.S. state.
* **sector-name:** Sector contributing to emissions (e.g., industrial, residential).
* **fuel-name:** Fuel type used (e.g., coal, petroleum, natural gas, or all fuels combined).
* **value:** Carbon dioxide emissions in million metric tons.

**Dataset Relevance:**

This dataset provides comprehensive data to analyze emissions trends, identify patterns, and predict future outcomes, offering insights into the energy consumption and environmental impact of different regions and sectors.

**4. Data Wrangling and Cleaning**

**Steps Performed:**

1. **Outlier Removal:**
   * Identified outliers using the interquartile range (IQR) method and removed rows with extreme emission values.
2. **Data Type Conversion:**
   * Converted the year column to datetime format for time-series analysis.
3. **Encoding Categorical Variables:**
   * Applied one-hot encoding to columns like state-name, sector-name, and fuel-name for machine learning compatibility.
4. **Consistency Check:**
   * Verified the absence of missing values and ensured uniform data formatting.

**5. Exploratory Data Analysis (EDA)**

**Key Visualizations and Insights:**

1. **Total Emissions Over Time:**
   * Emissions trends show distinct patterns with spikes or reductions corresponding to economic activities or policy changes.
2. **Top 5 States by Emissions:**
   * States like Texas, California, and Pennsylvania consistently rank as the highest contributors to carbon dioxide emissions.
3. **Sector Contributions in Top State:**
   * In high-emission states, the industrial and electric power sectors dominate emissions.
4. **Fuel Type Contributions:**
   * Coal and petroleum are primary sources of emissions, with natural gas showing a moderate contribution.
5. **Anomalies and Trends:**
   * Significant reductions in emissions were observed in certain states and sectors, likely due to shifts to renewable energy or economic changes.

**6. Predictive Analysis**

**Model Used:**

**Random Forest Regressor**

**Features and Target:**

* **Features:** State, sector, fuel type, and year.
* **Target:** Carbon dioxide emissions (value column).

**Evaluation Metrics:**

1. **Root Mean Squared Error (RMSE):** Measures average prediction error.
2. **Mean Absolute Error (MAE):** Evaluates the average magnitude of errors.
3. **R² (R-Squared):** Explains the variance captured by the model.

**Performance:**

* The model demonstrated strong predictive capabilities with low RMSE and MAE values and a high R² score, indicating reliability for forecasting future emissions.

**7. Key Insights**

1. **High-Emission States:**
   * Certain states, such as Texas and California, are consistently major contributors to emissions.
2. **Sectoral Trends:**
   * Industrial and electric power sectors are critical drivers of emissions, necessitating targeted interventions.
3. **Fuel Dependency:**
   * Coal and petroleum remain the dominant sources, highlighting the importance of transitioning to cleaner energy.
4. **Future Trends:**
   * Predictive analysis offers actionable insights to guide environmental policy and clean energy initiatives.

**8. Recommendations**

1. Focus on high-emission states for policy-driven interventions.
2. Prioritize the industrial and electric power sectors for emission reduction.
3. Transition from coal and petroleum to cleaner energy sources.
4. Leverage predictive insights to proactively manage and mitigate emissions.

**9. Conclusion**

This project comprehensively analyzed U.S. carbon dioxide emissions, identifying critical trends and contributors. The insights gained are instrumental in guiding targeted environmental policies and promoting sustainable energy practices. Predictive modeling further enhances the ability to forecast emissions and plan for a cleaner, sustainable future.