**Project Title: Predicting Carbon Emissions from Flight Data**

**1. Project Overview**

The objective of this project is to build a predictive machine learning model that estimates **carbon emissions** of flights based on a range of aircraft, flight, and environmental features. The dataset used is synthetically generated but represents real-world flight parameters, which makes it suitable for developing a proof-of-concept model and analysis pipeline.

**2. Dataset Description**

The dataset realistic\_synthetic\_flight\_data\_single\_file.csv contains  **million rows** and **50 columns**, each representing distinct measurable or categorical features related to:

* **Flight operations** (e.g., Flight\_Duration, Distance, Taxi\_Time)
* **Aircraft specifications** (e.g., Aircraft\_Weight, Engine\_Hours, Fuel\_Consumption)
* **Environmental conditions** (e.g., Altitude, Humidity\_Level, Outside\_Temperature)
* **Performance metrics** (e.g., Speed, Thrust\_Level, Fuel\_Efficiency)
* **Emission indicators** (e.g., CO2\_Emission, SO2\_Emission)
* **Maintenance and operational states** (e.g., Maintenance\_Flag, Sensor\_Error\_Code)

The **target variable** is:

* **Carbon\_Emissions** — the amount of carbon emitted during a flight (in tons)

**3. Project Flow**

1. **Data Ingestion**
2. **Data Exploration & Profiling**
3. **Exploratory Data Analysis (EDA)**
4. **Feature Engineering**
5. **Model Preparation**
6. **Model Evaluation**
7. **Conclusion & Recommendations**

**4. Data Ingestion**

* Load the dataset into a PySpark or Pandas environment.
* Check for schema correctness, missing data, duplicate records, and data types.

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

**5.1. General Data Profiling**

* Total rows, columns
* Data types per column
* Memory usage and loading time
* Basic statistics (mean, median, min, max, std) using .describe()

**5.2. Target Variable Exploration: Carbon\_Emissions**

* Distribution plot (histogram / KDE)
* Outlier detection (boxplot)
* Skewness and kurtosis
* Check if data is normally distributed or needs transformation (e.g., log)

**5.3. Missing Value Analysis**

* Count and percentage of missing values per column
* Visualization using heatmaps or missingno plots
* Strategy to handle missing values: imputation vs. deletion

**5.4. Correlation Analysis**

* Compute Pearson correlation matrix
* Visualize heatmap for top correlated features with Carbon\_Emissions
* Detect multicollinearity (VIF or pairwise correlations)

**5.5. Univariate Analysis**

* Distributions of key features like Flight\_Duration, Fuel\_Consumption, Speed, etc.
* Use histograms, KDE plots, and boxplots
* Log transformation for skewed distributions

**5.6. Bivariate Analysis**

* Scatter plots of each feature vs. Carbon\_Emissions
* Trendlines to observe linear/non-linear relationships
* Categorical columns: bar plots showing average emissions per category (if any)

**5.7. Multivariate Exploration**

* 3D scatter plots (e.g., Fuel\_Consumption vs. Flight\_Duration vs. Carbon\_Emissions)
* Feature combinations that might jointly impact emissions
* PCA or t-SNE for pattern detection

**5.8. Outlier Detection**

* Identify extreme values in continuous features
* Use Z-score or IQR methods
* Impact of outlier removal on Carbon\_Emissions

**6. Feature Engineering**

* **Transformations**: Log scaling, normalization, or standardization
* **Interaction Terms**: Combine Speed \* Aircraft\_Weight or Altitude / Distance
* **Derived Features**:
  + Fuel per km = Fuel\_Consumption / Distance
  + Emissions per km = Carbon\_Emissions / Distance
  + Efficiency Score = Fuel\_Efficiency / Thrust\_Level
* **Handling multicollinearity**: Drop or combine highly correlated features

**7. Model Preparation**

**7.1. Train-Test Split**

* 80–20 or 70–30 split
* Stratify if using categories (e.g., aircraft type in a real-world scenario)

**7.2. Model Candidates**

* **Linear Regression**
* **Random Forest Regressor**
* **Gradient Boosted Trees (e.g., XGBoost)**
* **Support Vector Regressor**
* **Neural Networks (if using deep learning frameworks)**

**7.3. Baseline Model**

* Mean Predictor or Linear Regression as baseline

**7.4. Model Evaluation Metrics**

* **R² Score**
* **Mean Absolute Error (MAE)**
* **Root Mean Squared Error (RMSE)**
* **Residual Plots**

**8. Model Tuning and Optimization**

* Use **Grid Search** or **Random Search** for hyperparameter tuning
* **Cross-validation** (k-fold or time-based if temporal data)
* Feature importance plots from tree-based models

**9. Conclusion & Recommendations**

* Highlight the **most important features** influencing carbon emissions
* Provide **recommendations** to reduce emissions:
  + Optimizing fuel consumption
  + Adjusting cruise speed or altitude
  + Monitoring engine conditions
* Evaluate whether the model is production-ready or requires more robust real-world data