## **Predicting Household Energy Usage Using Historical Consumption Data**

### **📄 Abstract**

This project aims to predict the energy consumption of a household using historical data collected at one-minute intervals over a span of four years. Leveraging statistical and machine learning techniques, we built models to forecast energy usage based on various electrical features. The process involved thorough data cleaning, preprocessing, and modeling using Linear Regression and Decision Tree algorithms.

### **🎯 1. Introduction**

Energy usage prediction is crucial for efficient energy management and conservation. By accurately forecasting household energy demand, both consumers and utility providers can optimize usage, reduce costs, and ensure sustainability. This project focuses on analyzing detailed energy consumption data to build predictive models.

### **📁 2. Data Collection & Description**

* **Source**: The dataset includes measurements of electric power consumption in a single household recorded at one-minute intervals for nearly 4 years.
* **Features**:  
  + Date: Timestamp of the measurement
  + Global\_active\_power: Total active power consumed (kilowatts)
  + Global\_reactive\_power: Total reactive power consumed (kilowatts)
  + Voltage: Voltage (volts)
  + Global\_intensity: Current intensity (ampere)
  + Sub\_metering\_1: Energy consumed by kitchen appliances
  + Sub\_metering\_2: Energy consumed by laundry appliances
  + Sub\_metering\_3: Energy consumed by water heater and air conditioner

### **🧹 3. Data Cleaning & Preprocessing**

* **Null Values**: Removed all rows with missing values.
* **Outlier Treatment**: Used the IQR (Interquartile Range) method to detect and treat outliers.
* **Duplicate Handling**: Removed duplicate rows from the dataset.
* **Feature Scaling**: Applied StandardScaler to normalize the data for model compatibility.
* **Multicollinearity Check**: Used **Variance Inflation Factor (VIF)** to detect and eliminate multicollinear features.

### **🧠 4. Modeling Approach**

* **Train-Test Split**: The data was divided into training and testing sets.
* **Linear Regression**:  
  + Trained on the preprocessed dataset.
  + Applied cross-validation to assess model robustness.
  + Evaluation Metrics: R² Score, Mean Squared Error (MSE), Mean Absolute Error (MAE).
  + Visualized actual vs predicted values.
* **Decision Tree Regression**:  
  + Hyperparameter tuning performed using RandomizedSearchCV.
  + Evaluation done using the same metrics as above.
  + Plotted predictions vs actual values for interpretation.

### **📊 5. Evaluation Metrics**

| **Model** | **R² Score** | **MSE** | **MAE** |
| --- | --- | --- | --- |
| Linear Regression | (value) | (value) | (value) |
| Decision Tree | (value) | (value) | (value) |

*Note: Insert actual evaluation values after running the models.*

### **📈 6. Visualizations**

* Correlation heatmaps to assess relationships between features
* Time series plot of energy usage over time
* Actual vs Predicted line plots for both models

### **🔧 7. Tools and Technologies**

* **Python**
* **Pandas** for data manipulation
* **Matplotlib & Seaborn** for visualizations
* **Scikit-learn** for preprocessing and modeling

### **⚠️ 8. Challenges Faced**

* High granularity of data (1-minute intervals) required aggregation for modeling.
* Multicollinearity affected regression performance initially.
* Balancing bias-variance tradeoff for Decision Tree required careful tuning.