**1. Approach**

This project aims to predict household energy consumption using a machine learning model. The goal is to understand patterns in historical energy data and develop a reliable model to forecast future energy usage.

**Steps followed:**

* Data cleaning and preprocessing
* Exploratory data analysis (EDA)
* Model selection and training
* Hyperparameter tuning
* Performance evaluation
* Visualization of results and interpretation

**Dataset:**

Contains minute-wise energy consumption metrics from a household

Key variables include:

* Global\_active\_power (target)
* Global\_reactive\_power
* Voltage
* Sub\_metering\_1, Sub\_metering\_2, Sub\_metering\_3
* Time-based features (hour, day, weekday)

**Preprocessing:**

* Converted Datetime to datetime object
* Handled missing values
* Extracted additional time features (hour, day, month)

**Findings:**

* Energy usage peaks during evening hours
* Weekends have higher average consumption
* There is a smaller rise in energy use between 7 AM and 9 AM in the morning.
* A bigger and more noticeable increase happens in the evening between 6 PM and 9 PM, with the highest usage around 8 PM.

**Model Selection and Evaluation**

Three models were evaluated on both training and testing datasets.

| **Model** | **Dataset** | **MAE** | **RMSE** | **R²** |
| --- | --- | --- | --- | --- |
| **Linear Regression** | Train | 0.3583 | 0.5308 | 0.7478 |
|  | Test | 0.3580 | 0.5303 | 0.7492 |
| **Random Forest Regressor** | Train | 0.2717 | 0.4345 | 0.8310 |
|  | Test | 0.2730 | 0.4365 | 0.8301 |
| **Gradient Boosting Regressor** | Train | 0.2469 | 0.4022 | 0.8552 |
|  | Test | 0.2492 | 0.4065 | 0.8526 |

**Model Performance:**

* The Gradient Boosting Regressor gave the best results. It had the lowest errors and could predict energy usage patterns better than the other models.
* The Random Forest model also worked well and gave reliable results.
* The Linear Regression model didn’t perform as well as the others. It had higher errors and wasn't as accurate in predicting the values.

**Performance Visualizations:**

* Actual vs Predicted Plot: Shows how closely predictions are to the real energy usage.
* Residual Plot: Detects the prediction errors. It suggests the model is reliable.
* Feature Importance: Shows which features had the biggest impact on the predictions.