Energy Consumption Prediction Report

1.Problem Statement

The objective of this project is to provide the methodology of building a robust machine learning pipeline that predicts equipment energy consumption and providing the insights that has been extracted from the EDA Process.

2.Approach

I have designed the completed end-to-end machine learning pipeline. The pipeline has following components:

1. Data ingestion pipeline
2. Data preprocessing pipeline
3. Feature Engineering pipeline
4. Model building pipeline
5. Model evaluation pipeline

Note: *First the Experiment was performed on jupyter notebook then converted into pipeline*

Approach for data preprocessing:

1. Timestamp Conversion
    - Converted the timestamp column from object to proper datetime format to enable time-based operations like sorting, lag features, and time window aggregation.
2. Fixing Numeric Columns
    - Several numeric columns (e.g., equipment_energy_consumption, lighting_energy, zone1_temperature, etc.) were incorrectly stored as objects due to non-numeric characters.
    - Removed unwanted characters using regex and converted them to numeric using pd.to_numeric() with coercion for invalid parsing.
3. Replacing Impossible Values

- Identified and replaced invalid values (e.g. negative humidity, negative energy) with NaN, as they are physically impossible in real-world environmental data.

4. Outlier Treatment
    - Handled outliers using three main strategies:
        - Winsorization: Capped extreme values at calculated IQR-based limits for features like equipment_energy_consumption, visibility_index, and atmospheric_pressure.
        - Domain-Based Capping: Applied temperature bounds (-5°C to 50°C) and humidity bounds (0% to 100%) for all zones.
5. Missing Value Treatment
    - Assessed missing value percentages (~4–5% per column).
    - Applied forward fill followed by backward fill for zoneX_temperature and zoneX_humidity using the time order, preserving temporal consistency.
    - For non-time-sensitive numeric features (e.g., lighting_energy, dew_point, etc.), imputed missing values with the median.
    - Applied time-aware forward-backward fill for outdoor_temperature and outdoor_humidity.
6. Column Cleanup
    - Removed unnecessary or auxiliary columns:
        - Unused random variables (random_variable1, random_variable2)
        - Since they do not have a strong correlation with target variable verified through heatmap and scatter plot

Approach for data engineering:

Indoor-Outdoor Temperature Differences

- For each of the 9 zones (zone1_temperature to zone9_temperature), calculate the difference between indoor and outdoor temperatures

Time-based Categorical Features

- Weekend Indicator (is_weekend): 1 if the timestamp is a weekend (Saturday/Sunday), else 0.
- Business Hours Indicator (is_business_hours): 1 if the hour is between 8 AM and 6 PM, else 0.
- Season Category:
    - Map months to seasons (winter, spring, summer, fall).
    - Apply one-hot encoding to create season-specific binary columns (season_winter, etc.).

Rolling Averages

- 24-hour Rolling Averages:
    - equipment_energy_24h_avg: 24-hour rolling mean of equipment_energy_consumption.
    - lighting_energy_24h_avg: 24-hour rolling mean of lighting_energy.

Zone Averages

- Average Zone Temperature: Mean of all 9 zone temperatures.
- Average Zone Humidity: Mean of all 9 zone humidity readings.

3. Data Insights

→ I have done two things first I have cleaned the data and then generate the insights from the cleaned data

→ I also did one thing after carefully observing the data I came to know that some instances have very high energy consumption so I perform separate analysis of those instances

Insights from cleaned Data

Mean	**88.91wh**
Median	**60.00wh**
Minimum	**10.00wh**
Maximum	**250.00wh (90% quantile)**

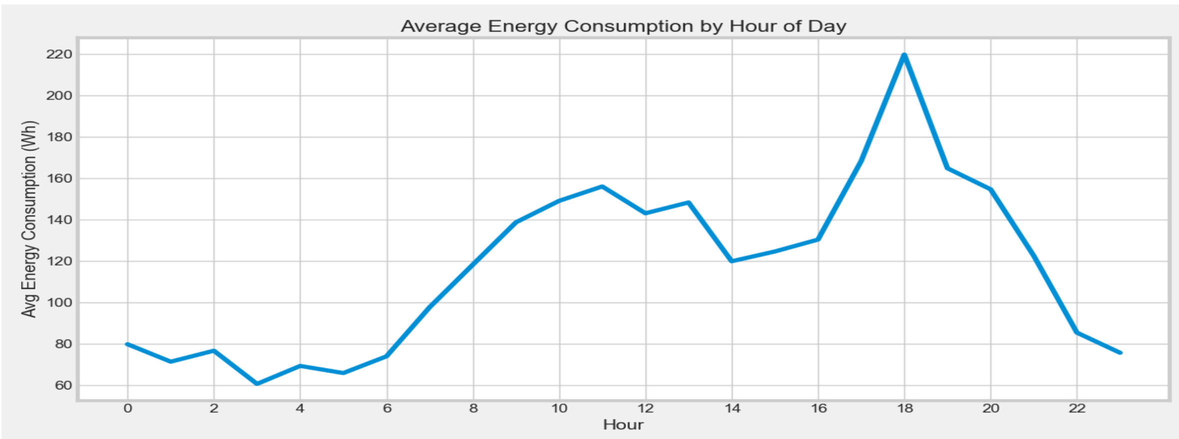
Peak consumption hour: **18:00**

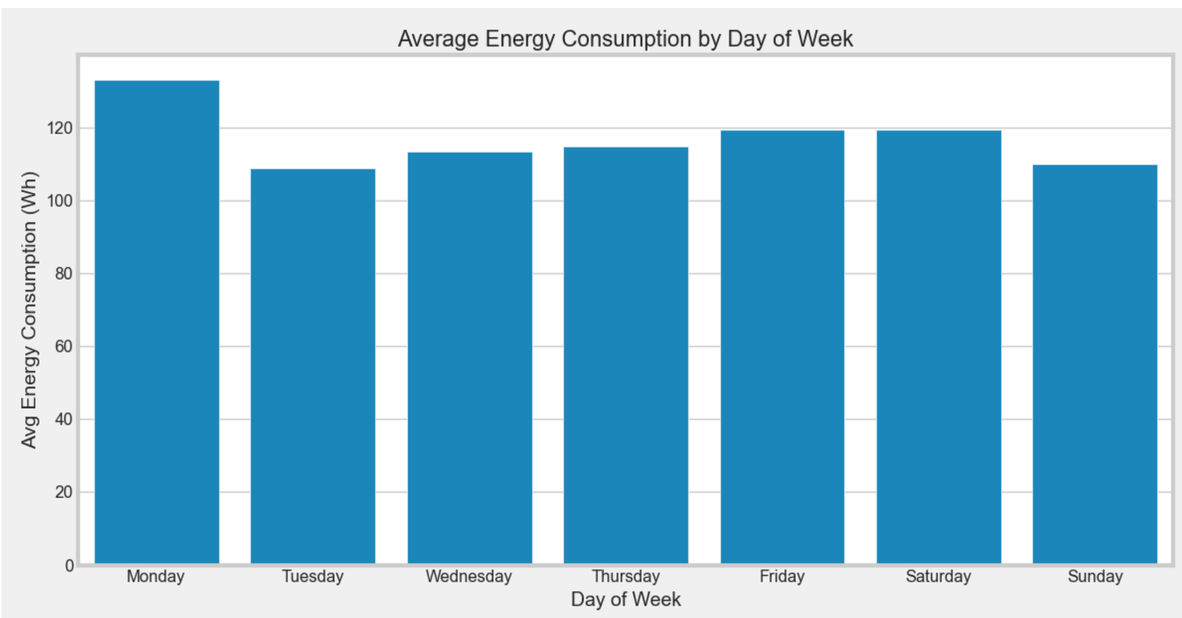
Minimum consumption hour: **3:00**

Peak consumption day: **Monday**

Minimum consumption day: **Tuesday**

Energy consumption trend across the day:





-> There is no correlation between the random variables and target so I removed them

Insights from high Energy data:

Mean	**114.78wh**
Median	**60.00wh**
Minimum	**10.00wh**
Maximum	**1139.00wh (90% quantile)**

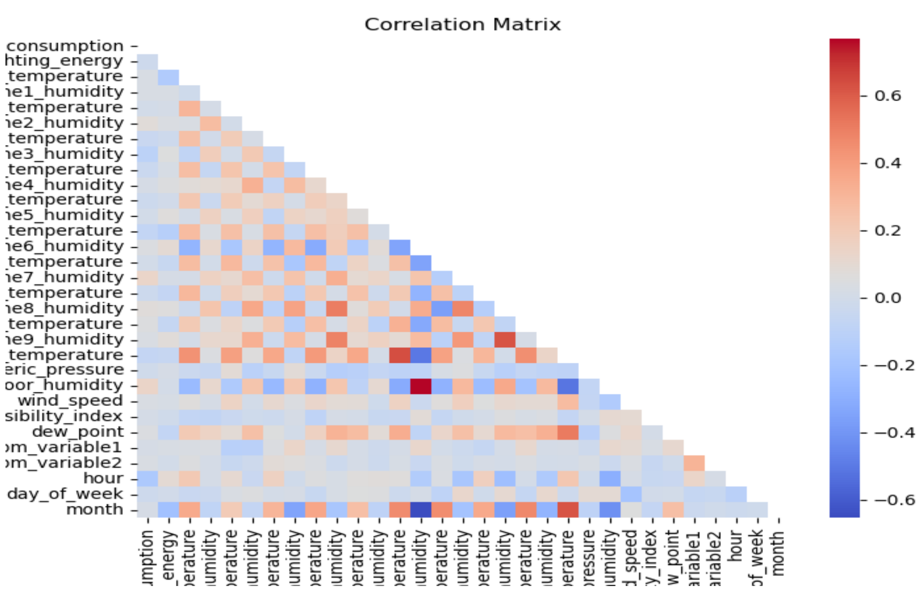
Peak consumption hour: **18:00**

Minimum consumption hour: **3:00**

Peak consumption month: **March**

Minimum consumption month: **September**

Top Correlated Features:



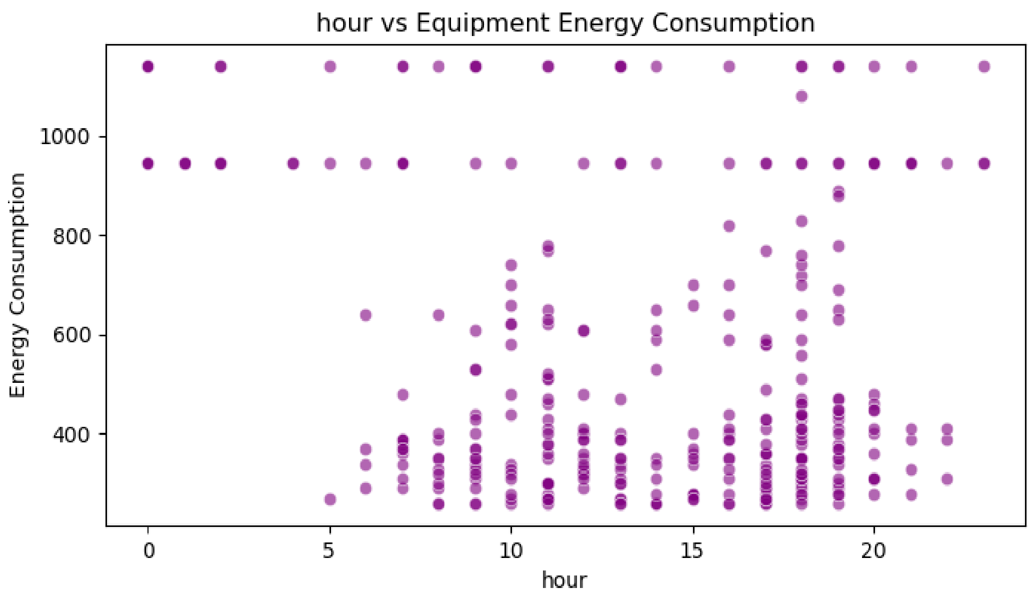
outdoor_humidity 0.139515

zone7_humidity 0.129377

zone2_humidity 0.075357

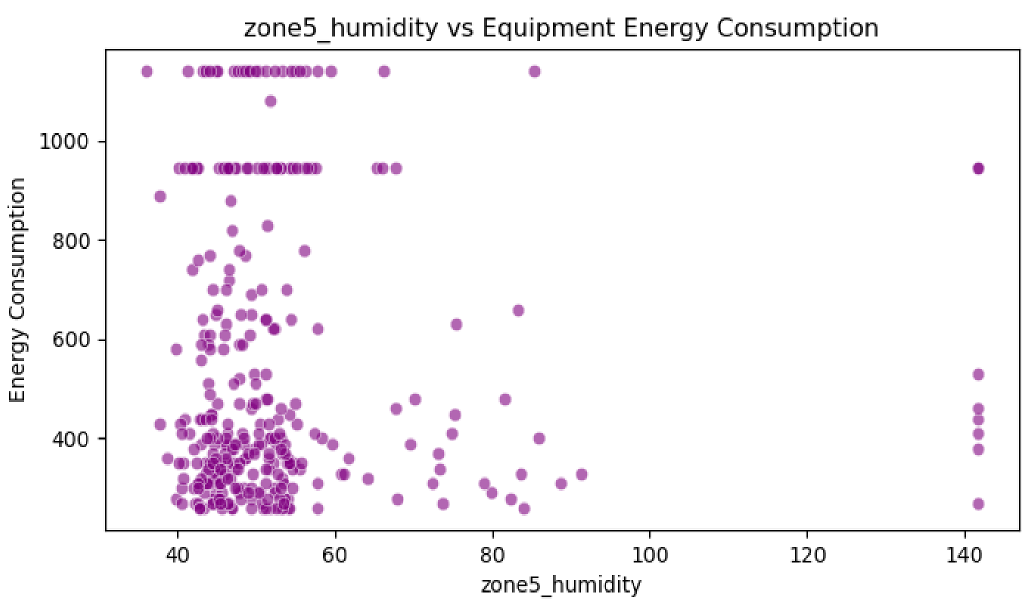
zone8_humidity 0.071302

Hour vs energy consumption:



-> **Energy consumption slightly increases with certain hours of the day**, likely during peak operational periods

Zone5 humidity vs energy consumption:



Most data points are concentrated in the **humidity range of 35–65%**, where equipment energy consumption ranges from **250 to 400**. This suggests **normal humidity levels are associated with moderate energy usage**.

High Energy Outliers:

A number of points exceed **1000 units** of energy consumption, particularly at **mid humidity (40–60%)**. these may be **outliers or heavy usage events**.

Extremely High Humidity (100%–140%):

Very few data points, but the energy consumption in this range is still **moderate** (300–500).

These values might be **sensor anomalies** or **unusual environmental conditions**

4. Model Development

I have performed the splitting of data for training and testing by considering the timestamp column so I have used the first 70% data for training and remaining for testing

I have performed Hyper parameter tunning using Optuna and since our dataset has a lot of outliers and data entry error(based on my knowledge) I have preferred to try with tree based algorithm which is confirmed by optuna. The best model is Decision tree with **95%** accuracy.

6. Recommendation

1. Validate Sensor Data for Accuracy

Anomalous values in humidity (e.g., negative or >100%) indicate possible sensor errors.

2. Address Equipment Inefficiency or Overuse

Several outliers show extremely high equipment consumption, even under normal conditions (in zone5_humidity and zone6_humidity plots).

Conduct preventive maintenance on equipment with abnormal consumption patterns.

3. Optimize Equipment Usage During Peak Hours

From the hour vs equipment_energy_consumption plot, energy consumption peaks during working hours (8 AM to 6 PM).

Schedule non-essential equipment during off-peak hours.

4. Audit and Maintain Lighting Systems

Positive trend observed between lighting_energy and equipment energy consumption.

Switch to **LED lighting** with **motion sensors or timers**.

Audit lighting schedules to prevent overlap with daylight hours.

Name:

Kanhaiya Kumar Jha