1. Approach

This project aimed to develop a machine learning model to predict equipment energy consumption in a manufacturing facility and provide actionable insights to reduce energy usage. The following approach was adopted to achieve this objective:

1.1 Data Loading and Exploration

The first step involved loading the provided sensor data into a pandas DataFrame. The dataset contained several columns, including `equipment_energy_consumption`, environmental parameters such as temperature and humidity, as well as random variables. A detailed exploration was conducted:

- **Descriptive Statistics**: Summary statistics were computed for all numerical columns (mean, median, standard deviation, etc.) to understand the distribution and central tendency of the data.

- **Data Types**: The data types for each column were reviewed to ensure they were appropriately assigned (e.g., float for continuous variables, object for categorical).

- **Missing Values**: Missing values were identified and analyzed to determine the best approach for imputation or removal.

- **Correlation Analysis**: The relationship between `equipment_energy_consumption` and other variables, such as temperature, humidity, and the random variables, was assessed using Pearson's correlation coefficient.

1.2 Data Cleaning

The data cleaning process involved:

- **Handling Missing Values**: Missing values in the `equipment_energy_consumption` column were filled using the median of the observed values to maintain consistency and avoid bias.

- **Outlier Detection**: Outliers in the dataset were detected using the Interquartile Range (IQR) method. Any data points falling outside 1.5 times the IQR from the quartiles were removed.

- **Duplicate Data**: Duplicate rows were identified and removed to avoid skewing the model's performance.

1.3 Data Preparation and Feature Engineering

Feature engineering played a critical role in enhancing the model's predictive power:

- **Time-Based Features**: Extracted time-based features from the timestamp, including the hour of day, day of the week, and month of the year. These features allowed the model to account for any temporal patterns in energy consumption.

- **Interaction Terms**: Interaction terms between environmental factors (e.g., zone temperatures) and equipment energy consumption were created to capture complex relationships.

- **Random Variables**: The role of random variables was evaluated by calculating their correlations with the target variable (`equipment_energy_consumption`). Variables showing little to no correlation were excluded from the final model.

1.4 Data Splitting

The dataset was split into training, validation, and test sets. A **temporal split** approach was adopted to ensure the data used for validation and testing represented future periods relative to the training data. This method simulates a real-world scenario where future energy consumption is predicted based on past data.

1.5 Model Training and Optimization

Several regression models were trained to predict equipment energy consumption, including:

- **RandomForestRegressor**: A robust ensemble method, good at handling non-linear relationships and feature importance.

- **GradientBoostingRegressor**: An ensemble method that iteratively refines its predictions to minimize errors.

- **LinearRegression**: A simple baseline model for comparison.
  Hyperparameter optimization was carried out using **GridSearchCV** on the validation set. Hyperparameters such as the number of estimators, learning rate, and maximum depth were tuned for the Random Forest and Gradient Boosting models.

1.6 Model Evaluation and Feature Importance

The model performance was evaluated using the following metrics:

- **RMSE (Root Mean Squared Error)**: Measures the average magnitude of the error between predicted and actual values.

- **MAE (Mean Absolute Error)**: Gives an average of the absolute errors between predictions and actual values.

- **R-Squared (R^2)**: Measures the proportion of variance in the target variable explained by the model.

Feature importance was calculated for the **RandomForestRegressor** to identify the most influential features. Key variables influencing energy consumption included temperature zones, time-based features, and interaction terms.

1.7 Data Visualization

To better understand the relationships in the data and the model's predictions:

- **Model Predictions vs. Actuals**: Scatter plots and line graphs were used to visualize the alignment of predicted energy consumption values with actual values.

- **Feature Relationships**: Heatmaps and correlation plots were created to show how different features, such as zone temperatures and time-based factors, interact with energy consumption.

- **Model Performance Metrics**: Graphs displaying RMSE, MAE, and R^2 were generated to show the performance comparison across the models.

---

2. Key Insights from the Data

The data analysis yielded the following valuable insights:

2.1 Weak Correlation with Random Variables

- Both `random_variable1` and `random_variable2` exhibited negligible linear relationships with `equipment_energy_consumption`. Their correlation coefficients were close to zero, indicating no significant predictive value. This justified their exclusion from the model.

2.2 Interaction Terms as Key Predictors

- **Zone Temperatures**: Interaction terms between `zone1_temperature` and `zone3_temperature` with energy consumption showed strong predictive power. These factors played a crucial role in determining energy usage, particularly in areas with fluctuating temperatures.

2.3 Temporal Patterns

- Energy consumption exhibited patterns that aligned with the time of day, day of the week, and seasonality (month of year). For instance, higher energy consumption was observed during specific hours of the day when machinery usage peaked, which suggested that scheduling optimization could play a significant role in reducing energy usage.

---

3. Model Performance Evaluation

The Random Forest model, after hyperparameter tuning, performed exceptionally well on both validation and test sets:

- **Validation Set**:
    - **RMSE**: 0.85
    - **MAE**: 0.14
    - **R-Squared**: 0.9995
- **Test Set**:
    - **RMSE**: 2.03
    - **MAE**: 0.65
    - **R-Squared**: 0.997

The results indicate that the model is highly accurate in predicting energy consumption, with minimal error on both the validation and test datasets.

---

4. Recommendations for Reducing Equipment Energy Consumption

The following recommendations are made to optimize energy usage in the manufacturing facility:

4.1 Optimize Zone 1 and Zone 3 Temperatures

- Implement dynamic temperature control mechanisms based on equipment activity levels, especially in zones where equipment usage is high. Consider using automated cooling/heating systems that adjust temperatures during peak periods to optimize energy usage.

4.2 Improve Insulation

- Focus on enhancing insulation in Zone 1 and Zone 3 to minimize heat loss or gain, particularly during extreme temperature fluctuations. This would reduce the energy required to maintain stable temperatures in these zones.

4.3 Optimize Equipment Scheduling

- Schedule energy-intensive operations during off-peak hours, when electricity rates are lower, and when temperature requirements are less demanding. This would allow the facility to reduce its energy costs while still maintaining operational efficiency.

4.4 Investigate Non-linear Relationships

- Further investigation into non-linear relationships or interactions involving environmental and random variables could uncover additional optimization opportunities. For instance, exploring how humidity or equipment-specific factors might interact with energy consumption could lead to more targeted energy-saving measures.

4.5 Implement Energy Monitoring Systems

- Set up real-time energy monitoring and predictive maintenance systems that can alert operators when energy usage exceeds expected thresholds. This would help identify inefficiencies promptly and allow for timely intervention.