Energy Consumption Prediction Project

# 1. Introduction

This document summarizes the complete workflow followed in building a machine learning model to predict equipment energy consumption. It includes data preprocessing, advanced feature engineering, model training and tuning, and evaluation following best practices and guidance similar to the reference GitHub repository.

# 2. Data Loading and Exploration

The dataset was loaded into a pandas DataFrame and basic EDA was performed to understand structure, missing values, and descriptive statistics.

# 3. Data Cleaning

- Dropped irrelevant columns like 'random\_variable'.  
- Removed or filled missing values.  
- Timestamp was parsed to extract temporal features.

# 4. Advanced Feature Engineering

We created features based on domain knowledge and data insights:  
- Extracted time-based features: hour, day\_of\_week, month, is\_weekend, is\_peak\_hour, is\_working\_hour.  
- Generated interactions and transformations: temp\_diff\_3\_5, rolling\_temp\_zone3, saturation\_zone3, etc.

# 5. Feature Scaling

StandardScaler was used to scale numeric features, improving model performance and convergence.

# 6. Train-Test Split

Data was split using an 80/20 train-test ratio with a fixed random state for reproducibility.

# 7. Model Training

Random Forest Regressor was trained with manually tuned hyperparameters initially:  
- n\_estimators=150  
- min\_samples\_split=2  
- min\_samples\_leaf=2  
- max\_features='log2'  
- bootstrap=False

# 8. Model Evaluation

Evaluation was performed using:  
- MAE (Mean Absolute Error)  
- R² Score  
- RMSE (Root Mean Squared Error)

# 9. Hyperparameter Tuning

GridSearchCV was used with 3-fold cross-validation on Random Forest. Best hyperparameters were found:  
- n\_estimators=200  
- max\_depth=None  
- min\_samples\_leaf=2  
- min\_samples\_split=5  
- bootstrap=False

# 10. Feature Importance

The importance of each feature was computed using Random Forest's built-in method and visualized using bar plots.

# 11. Final Model Performance

Best achieved metrics:  
- MAE: ~0.441  
- R² Score: ~0.544  
- Cross-validated MAE: ~0.634

# 12. Recommendations

- Continue exploring interaction features between temperature and humidity.  
- Try other ensemble models (XGBoost, LightGBM).  
- Use SHAP or LIME for interpretability.  
- Perform error analysis on low/high predictions.