1. Introduction

In modern intelligent infrastructure systems, **energy optimization** is critical for both economic and environmental sustainability. This project addresses the challenge of predicting **equipment energy consumption** using a rich dataset collected across **nine environmental zones**, capturing both temporal and meteorological information.

🔍 Objectives

- Accurately forecast equipment energy consumption using historical sensor data.
- Identify key drivers of energy usage via correlation and feature importance.
- Apply transformations to correct data skewness and improve model reliability.
- Compare multiple regression algorithms to identify the most performant model.

📌 Scope

- Data wrangling and cleaning
- Exploratory Data Analysis (EDA)
- Feature engineering & transformation
- Predictive modeling with hyperparameter tuning
- Model evaluation using RMSE and R^2

---

2. Data Preprocessing & Exploratory Data Analysis (EDA)

📂 Dataset Overview

- Format: CSV
- Records: Multiple daily readings across 9 zones

- Key features: `temperature`, `humidity`, `wind_speed`, `zone`, `hour`, `equipment_energy_consumption`

🧹 Preprocessing Steps

- **Missing values**: Checked using `data.isnull().sum()`, handled through imputation or exclusion.
- **Outliers**: Identified via distribution plots and boxplots, treated using transformations (e.g., log, square).
- **Duplicates**: Removed to prevent bias and overfitting.

📊 Visualizations & Findings

🔸 Humidity Distribution by Zone

- **Visualization**: `sns.distplot` and scatterplots
- **Insight**: All zones except **Zone 6** show humidity distributions centered between **35–45%**, aligning with a **normal distribution**. Zone 6 skews higher (~55%), contributing to **wider energy usage variance**.
- **Business Implication**: HVAC behavior in Zone 6 may require **special calibration** due to its anomalous humidity response.

🔸 Hourly & Monthly Trends

- Energy consumption **spikes from 8 AM to 9 PM**, aligning with peak operational hours.
- Monthly analysis shows **August** as the highest consumer, possibly due to seasonal temperature and humidity patterns.

🔸 Correlation Heatmap

- Implemented using `sns.heatmap()` on the top correlated features.
- **Strong positive correlations** observed between temperature, humidity difference, and energy consumption.

- Insightful for identifying **redundant features** and potential **multicollinearity**, guiding feature selection.

---

3. Feature Engineering & Transformation

🛠️ Feature Engineering Highlights

- **Humidity difference squared**: Captures non-linear relationship with energy usage.
- **Log transformation** of `wind_speed`: Reduces right skewness.
- **Box-Cox transformation** on `equipment_energy_consumption`: Normalizes the target distribution for better regression modeling.

❓ Why Transform?

Several features exhibited skewed distributions that can degrade model performance:

- Logarithmic and Box-Cox transformations ensure **linear relationships** and **normality assumptions**, crucial for models like Linear Regression and Tree-based regressors.
- **Outcome**: Reduced heteroscedasticity and improved generalization.

---

4. Model Development & Evaluation

🔍 Models Trained

- **Linear Regression**: Baseline model, limited by linearity assumptions.
- **Decision Tree Regressor**: Captures non-linear relationships, prone to overfitting.
- **Random Forest Regressor**: Ensemble model, more robust and accurate.
- **XGBoost Regressor**: Gradient boosting technique, efficient and powerful.

- **ExtraTrees Regressor**: Randomized tree ensemble, **best performing** in this use case.

🧪 Evaluation Metrics

- **Root Mean Squared Error (RMSE)**: Measures prediction accuracy.
- **R² Score**: Explains variance captured by the model.

Model	RMSE	R² Score
Linear Regression	0.542	0.431
Decision Tree	0.410	0.587
Random Forest	0.372	0.638
XGBoost	0.358	0.651
ExtraTrees	**0.341**	**0.668**

✅ Final Model: ExtraTrees Regressor

- Chosen for its high accuracy and ability to capture complex feature interactions.
- Saved using `pickle` for deployment and reuse.

---

5. Business Impact & Recommendations

📈 Positive Impacts

- **Data-driven energy optimization**: Helps facilities reduce wastage and anticipate demand surges.
- **Zone-specific insights**: Zone 6's abnormal humidity patterns can inform targeted system redesign or sensor calibration.

- **Model integration**: Real-time prediction models can be embedded in energy dashboards or building management systems (BMS).

⚠️ Challenges & Future Work

- **Zone 6 anomaly** suggests potential sensor drift or infrastructure inefficiency – needs physical verification.
- Expand feature set to include **external weather APIs**, occupancy data, or machinery load profiles.
- Consider time series modeling (e.g., LSTM, Prophet) for sequential forecasting.

---

6. Conclusion

This project demonstrates a full pipeline — from raw sensor data to actionable insights and a deployed machine learning model. The process showcases:

- Rigorous **EDA and feature engineering**
- Effective **use of transformations**
- Thorough **model comparison**
- Deliverable insights for **energy optimization strategies**

With an R^2 of **0.668**, the final model represents a significant step toward smarter energy systems. Continued enhancement, such as **real-time ingestion and time-series analysis**, can drive further efficiencies.