Predicting Solar Energy Efficiency Levels of Buildings in Lagos

BY: TEAM NWAPA

AI Saturdays

October, 2025

PROJECT OBJECTIVES

- Classify energy installations into efficiency categories: **Low, Medium, High**
- Predict energy efficiency using building and installation features
- Support **data-driven decisions** for energy planning
- Deploy a **practical ML model** for real-world application

ABOUT DATASET

Rows: 200,000+

Columns: Estimated_building_height, Capacity_density, Energy_potential_per_year, etc.

Target: Energy efficiency category (Low / Medium / High)

Area_utilization_ratio	Energy_density	Capacity_density	System_efficiency	Energy_category
0.593005	253.625508	0.193966	0.149267	Medium
0.738660	254.914540	0.194002	0.149998	High
0.739046	254.916774	0.194010	0.149993	High

Data Cleaning & Preprocessing

Removed invalid entries: Eliminated zeros or negative values from critical features

Handled infinities: Replaced inf / -inf with NaN

Handled missing values: Dropped rows with nulls

Index reset: Reindexed dataset after cleaning

Exploratory Data Analysis (EDA) Performed:

- Checked feature distributions
- Computed summary statistics (mean, median, std)
- Correlation heatmap
- Identified strongly and weakly correlated features
- Detected extreme or outlier values
- Visualized target class balance

Raw Data → Cleaned Data → Features Ready for Modeling

Baseline Model

Logistic Regression:
A simple linear classifier used to estimate the probability of class membership; serves as a baseline to compare performance against more complex models.

Overall accuracy: 72%

Class-wise performance:

- **High:** Precision 0.85 | Recall 0.86 | F1-score 0.85
- **Low:** Precision 0.73 | Recall 0.70 | F1-score 0.72
- **Medium:** Precision 0.59 | Recall 0.61 | F1-score 0.60

Provides a **reference point** for improvement with more complex models

Model Comparison – Performance Improvement

Compared multiple classifiers to capture **linear relationships** and improve baseline performance:

- Logistic Regression (baseline)
  Decision Tree
- Gradient Boosting
- Random Forest

Model	Mean CV Accuracy	Test Accuracy
Random Forest	0.789245	0.789419
Gradient Boosting	0.787195	0.788477
Logistic Regression	0.722436	0.724007
Decision Tree	0.715599	0.715928

Random Forest achieved the highest test accuracy and became the selected model for tuning

Selected Model & Hyperparameter Tuning

Chosen Model: Random Forest

- Selected for **best performance** in model comparison (highest accuracy)
- Robust to overfitting, captures non-linear relationships effectively

Hyperparameter Tuning:

- **Performed using RandomizedSearchCV on a sample of the dataset**
- **Best parameters found:**
  - **n_estimators = 200**
  - **min_samples_split = 5**
  - **min_samples_leaf = 2**
  - **max_features = 'sqrt'**
  - **max_depth = None**

Final Model & Key Takeaways

Final Model: Random Forest trained on full dataset

Test Accuracy: 78.89%

Key Insights:

- **Captures non-linear relationships effectively**
- **Handles strong and weak feature correlations better than regression or simpler models**
- **Baseline Logistic Regression struggled with Medium efficiency category; Random Forest improved predictions across all classes**

Outcome: Model is ready for deployment and real-world application

“Random Forest outperformed all other models, demonstrating robust classification performance across energy efficiency categories.”

Real-World Impact

Project Significance & Real-Life Applications:

Provides a **data-driven method** to classify energy installations by efficiency

Supports **policy-making and energy planning** for sustainable infrastructure

Helps identify **high-potential areas** for energy optimization

Enables **resource allocation** based on predicted efficiency categories

Can be integrated into **smart building systems** or energy monitoring platforms

Local Impact:

- Helps Lagos and Nigeria optimize energy infrastructure by identifying high-efficiency installations.
- Supports sustainable urban planning and informed decision-making for national energy policies.

“This project bridges data science and energy management, offering actionable insights for sustainable energy deployment.”

References & Acknowledgments

References:

- Scikit-learn, Pandas, NumPy, Joblib documentation
- Python official docs and ML resources

Acknowledgments:

- AI Saturdays Lagos Cohort 4
- Mentors, peers, and dataset providers

Thank You!