Project Report: Electricity Cost Prediction

1. EDA Insights

The dataset contains 10,000 records and 9 columns, including:
- Numeric features such as site area, water consumption, recycling rate, air quality index, and more.
- A categorical feature: structure type (Residential, Mixed-use, Commercial).

Key insights:
- Electricity cost ranges from ₹500 to ₹6446, with a mean around ₹2800.
- Correlation analysis shows site area, resident count, and water consumption are positively correlated with electricity cost.
- Distribution of structure types and their influence on cost showed variation: e.g., Mixed-use types tend to have higher consumption.

2. Feature Preprocessing Details

Features were preprocessed using a ColumnTransformer:

- *Numerical features* (site area, water consumption, recycling rate, utilisation rate, air quality index, issue resolution time, resident count) were:
  - Imputed using the median (if missing values exist)
  - Scaled using StandardScaler

- *Categorical feature* (structure type) was:
  - Imputed with the most frequent value
  - One-hot encoded using OneHotEncoder with handle_unknown='ignore'

All preprocessing was encapsulated within a `Pipeline`, which ensures transformation consistency during both training and inference.

3. Model Selection Reasoning

RandomForestRegressor was chosen for the following reasons:
- It handles non-linear relationships well
- It is robust to outliers and doesn't require feature scaling explicitly (although we scaled for other benefits)
- It provides strong performance on tabular data

GridSearchCV was used to tune the following hyperparameters:
- n_estimators: 100, 150, 200
- max_depth: None, 5, 10, 20

The best model was selected using 5-fold cross-validation based on negative MSE.

4. Model Performance Metrics

The final model's performance on the test set:

- *Root Mean Squared Error (RMSE)*: ~221.0
- *R^2 Score*: 0.96

This indicates that the model explains 96% of the variance in electricity cost and predicts with an average error of ₹221, which is less than 8% of the average cost.

5. API Endpoints

The application was deployed using Flask with a simple web interface and form.

`GET /`
- Renders the input form (`index.html`)

`POST /predict`
- Accepts form input via POST request
- Input fields:
  - site_area (int)
  - water_consumption (float)
  - recycling_rate (int)
  - utilisation_rate (int)
  - air_quality_index (int)
  - issue_resolution_time (int)
  - resident_count (int)
  - structure_type (str: Residential | Mixed-use | Commercial)
- Returns: Predicted electricity cost formatted to 2 decimal places