iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84312

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Estimating the Cost of Electricity Using Classification Techniques with Machine Learning

Jyothika Nambari[1], Pravitha R Prasad[2]
[1]Student, Master of Computer Applications (MCA), AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Visakhapatnam, Andhra Pradesh, India
[2]Assistant Professor, Department of MCA, AQJ Centre for P.G. Studies (Affiliated to Andhra University, Visakhapatnam), Visakhapatnam, Andhra Pradesh, India

Abstract: Electricity has become one of the most important resources for residential, commercial, and industrial consumers. Due to increasing electricity demand and changing tariff rates, estimating future electricity costs has become a significant challenge. Traditional electricity billing systems provide consumers with electricity costs only after the completion of the billing cycle, making it difficult to monitor and control energy expenses in advance. This paper presents a machine learning-based classification system for estimating electricity cost categories using historical electricity consumption data. The proposed approach utilizes multiple input features, including appliance usage, monthly operating hours, electricity tariff rate, city, and electricity company, to classify electricity costs into Low, Medium, and High categories. Several supervised machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting, XGBoost, and LightGBM, were trained and evaluated. Experimental results showed that tree-based ensemble models achieved the highest performance on the prepared dataset. A Streamlit-based web application was also developed to provide a simple interface for predicting electricity cost categories. The proposed system can assist consumers in making informed decisions about electricity usage and supports intelligent energy management.
Keywords: Machine Learning, Electricity Cost Prediction, Classification, Random Forest, Decision Tree, XGBoost, LightGBM, Streamlit, Energy Management, Artificial Intelligence.

I. INTRODUCTION

Electricity is one of the fundamental resources supporting modern society. Residential households, industries, hospitals, educational institutions, and commercial organizations depend on electricity for their daily activities. As the number of electrical appliances continues to increase, electricity consumption has grown significantly, leading to higher electricity bills for consumers. Managing electricity costs has therefore become an important concern for both consumers and electricity providers.

Traditional electricity billing systems calculate electricity charges only after the billing period has ended. Consequently, consumers have limited opportunities to adjust their electricity usage before receiving the final bill. This lack of early information often results in excessive electricity consumption and increased financial burden.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have enabled the development of intelligent systems capable of learning from historical data and making accurate predictions. Machine learning algorithms have demonstrated excellent performance in various prediction tasks, including energy demand forecasting, electricity load prediction, customer behavior analysis, and smart grid management.

This research proposes a machine learning-based classification system for estimating electricity cost categories before bill generation. Instead of predicting the exact electricity bill amount, the proposed system classifies electricity costs into three categories:

- Low Cost
- Medium Cost
- High Cost

The system analyzes multiple influencing factors, including appliance usage, monthly operating hours, electricity tariff rates, city information, and electricity company details. Multiple machine learning algorithms are compared to determine the most suitable model for electricity cost classification.

The primary contributions of this work include:

- Development of an intelligent electricity cost classification model.
- Comparison of multiple supervised machine learning algorithms.
- Performance evaluation using standard classification metrics.
- Deployment of a Streamlit web application for user-friendly predictions.
- A framework that can be extended for future smart energy management systems.

The proposed system demonstrates the practical application of machine learning in energy management and provides a foundation for future research in intelligent electricity cost prediction.

II. LITERATURE REVIEW

The rapid growth of electricity consumption and the increasing availability of smart meter data have encouraged researchers to apply Artificial Intelligence (AI) and Machine Learning (ML) techniques for electricity demand forecasting, energy consumption prediction, and electricity cost estimation. Various machine learning algorithms have been proposed to improve prediction accuracy and support intelligent energy management systems. This section reviews important research studies related to electricity cost prediction and identifies the research gap addressed in this work.

Author et al. (2020) proposed a machine learning framework for household electricity consumption prediction using Decision Tree and Random Forest algorithms. The study showed that Random Forest provided better prediction accuracy than a single Decision Tree because of its ensemble learning approach. However, the work mainly focused on electricity consumption forecasting and did not classify electricity cost categories.

Author et al. (2021) developed a Support Vector Machine (SVM)-based model for predicting residential electricity demand. The proposed model produced satisfactory results for medium-sized datasets but required careful parameter tuning and longer training time. The study demonstrated that SVM is effective for nonlinear prediction problems, although its computational complexity increases for large datasets.

Author et al. (2021) applied Artificial Neural Networks (ANNs) to forecast electricity usage in residential buildings. The model successfully learned complex relationships among different input variables and achieved good prediction accuracy. However, the performance depended heavily on the availability of large amounts of training data, making implementation more challenging for smaller datasets.

Author et al. (2022) introduced Extreme Gradient Boosting (XGBoost) for electricity load forecasting. XGBoost outperformed traditional regression and classification techniques because of its efficient gradient boosting mechanism and built-in regularization. The research demonstrated that ensemble learning algorithms can significantly improve prediction performance while reducing overfitting.

Author et al. (2022) presented an Internet of Things (IoT)-based smart energy monitoring system integrated with machine learning. The system collected real-time electricity consumption data from smart meters and used machine learning algorithms to predict future energy usage. Although the approach supported intelligent monitoring, it required specialized hardware and continuous data collection.

Author et al. (2023) investigated the use of LightGBM for electricity demand prediction. The study reported faster training speed and lower memory consumption compared with other boosting algorithms while maintaining high prediction accuracy. The researchers concluded that LightGBM is suitable for handling large-scale datasets efficiently.

Author et al. (2023) compared several machine learning algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Naïve Bayes, for energy consumption prediction. Their results indicated that tree-based algorithms generally achieved higher accuracy because they effectively captured nonlinear relationships among input features.

Author et al. (2024) proposed a hybrid machine learning model combining feature engineering and ensemble learning techniques for electricity bill prediction. The proposed approach improved prediction accuracy compared with individual classifiers. However, the research mainly focused on estimating numerical electricity bills rather than classifying electricity costs into different categories.

Author et al. (2024) used smart meter data and deep learning models such as Long Short-Term Memory (LSTM) networks to forecast electricity demand. The model successfully captured temporal dependencies in electricity usage patterns but required high computational resources and longer training time compared with traditional machine learning methods.

Author et al. (2025) presented a comparative analysis of modern machine learning algorithms for smart grid applications. The study evaluated Random Forest, XGBoost, LightGBM, Gradient Boosting, and deep learning models. The results confirmed that ensemble learning techniques consistently produced high prediction accuracy and were suitable for practical energy management applications.

Overall, the literature indicates that machine learning techniques have significantly improved electricity consumption forecasting and demand prediction. Tree-based ensemble algorithms such as Random Forest, Gradient Boosting, XGBoost, and LightGBM generally outperform traditional classification methods in terms of accuracy and robustness. However, many existing studies focus on forecasting electricity consumption or estimating numerical electricity bills rather than classifying electricity cost categories into meaningful groups for end users.

A. Research Gap

Based on the literature survey, the following research gaps were identified:

- Most previous studies focus on electricity consumption forecasting rather than electricity cost classification.
- Many models estimate the exact electricity bill, which can be more sensitive to tariff fluctuations and data variations.
- Limited research compares a wide range of machine learning classification algorithms using the same dataset and evaluation metrics.
- Few studies provide a simple, user-friendly web application for predicting electricity cost categories.
- There is limited emphasis on integrating preprocessing, exploratory data analysis, model comparison, and deployment into a single end-to-end framework.

To address these gaps, the proposed work develops a machine learning-based electricity cost classification system that compares multiple supervised learning algorithms and deploys the best-performing model using a Streamlit web application.

III. PROPOSED METHODOLOGY

A. Overview

The proposed system estimates electricity cost categories using supervised machine learning classification techniques. Instead of predicting the exact electricity bill amount, the system classifies electricity costs into one of three categories: Low, Medium, or High. The overall workflow consists of data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model training, model evaluation, and deployment through a Streamlit web application.

The methodology is designed to provide an accurate and user-friendly framework for electricity cost estimation. Multiple machine learning algorithms are trained and compared to identify the best-performing classifier.

B. Dataset Description

The dataset used in this study contains 45,345 records and 12 original attributes related to electricity consumption.

Dataset Features

Feature	Description
Fan	Number of fans used
Refrigerator	Number of refrigerators
AirConditioner	Number of air conditioners
Television	Number of televisions
Monitor	Number of computer monitors
MotorPump	Number of motor pumps
Month	Month of electricity usage
City	Consumer location
Company	Electricity distribution company
MonthlyHours	Total appliance operating hours
TariffRate	Electricity tariff per unit
ElectricityBill	Monthly electricity bill

A new target variable named CostCategory was created based on the electricity bill values.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

The categories are:
- Low Cost
- Medium Cost
- High Cost

This transformation converts the prediction problem into a multiclass classification task.

C. Data Preprocessing
Data preprocessing is an essential step in building an accurate machine learning model. The following preprocessing operations were performed:
Step 1: Dataset Loading
The dataset was imported into Python using the Pandas library.
Step 2: Missing Value Analysis
The dataset was checked for missing values.
Result:
No missing values were found.
Step 3: Duplicate Checking
Duplicate records were identified.
Result:
No duplicate rows were found.
Step 4: Feature Encoding
Since City and Company are categorical variables, they were converted into numerical values using Label Encoding.
Step 5: Target Generation
The ElectricityBill values were divided into three cost categories:
- Low
- Medium
- High

using threshold-based categorization.

D. Exploratory Data Analysis (EDA)
Exploratory Data Analysis was performed to understand the characteristics of the dataset before model training.
The following visualizations were generated:
- Distribution of Electricity Cost Categories
- Histogram of Monthly Hours
- Tariff Rate Distribution
- Appliance Usage Distribution
- Correlation Heatmap
- Boxplots
- Pairwise Relationship Analysis
- Feature Distribution Graphs

EDA helped identify relationships among features and confirmed that the dataset was balanced across the three electricity cost categories.

E. Feature Selection
The following input variables were used during model training:
- Fan
- Refrigerator
- AirConditioner
- Television
- Monitor

- MotorPump
- Month
- City
- Company
- MonthlyHours
- TariffRate

The target variable was:
CostCategory
Feature selection reduces unnecessary information and helps improve model efficiency.

F. Machine Learning Algorithms
Nine supervised machine learning algorithms were implemented and compared.
1. Logistic Regression
Logistic Regression is a statistical classification algorithm used as a baseline model. It predicts class probabilities using a logistic function.
2. Decision Tree
Decision Tree creates hierarchical decision rules based on input features. It is easy to understand and provides fast predictions.
3. Random Forest
Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
4. K-Nearest Neighbors (KNN)
KNN classifies a sample by identifying its nearest neighboring data points.
5. Support Vector Machine (SVM)
SVM constructs an optimal decision boundary that separates different classes while maximizing the margin between them.
6. Naïve Bayes
Naïve Bayes is a probabilistic classifier based on Bayes' theorem and assumes feature independence.
7. Gradient Boosting
Gradient Boosting builds a sequence of weak learners, where each learner corrects the errors made by previous learners.
8. XGBoost
Extreme Gradient Boosting (XGBoost) is an optimized boosting algorithm known for high prediction accuracy and efficient computation.
9. LightGBM
LightGBM is a gradient boosting framework designed for faster training, lower memory usage, and high scalability on large datasets.

G. Model Training
The processed dataset was divided into:
- 80% Training Data
- 20% Testing Data

Each classifier was trained using the same training dataset to ensure a fair comparison.
The trained models were stored for later evaluation.

H. Model Evaluation
The models were evaluated using the following performance metrics:
- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

- Classification Report
- Cross Validation

The best-performing model was selected based on overall classification performance.

I. *Streamlit Deployment*

The selected machine learning model was integrated into a Streamlit web application.

The application allows users to:

1) Enter appliance usage details.
2) Provide tariff and monthly usage information.
3) Click the **Predict** button.
4) Instantly view the predicted electricity cost category.

This deployment demonstrates the practical applicability of the proposed model for everyday users.

J. *Methodology Workflow*

The complete workflow of the proposed system is shown below:

Electricity Dataset
▼
Data Preprocessing
▼
Exploratory Data Analysis
▼
Feature Engineering
▼
Machine Learning Model Training
▼
Performance Evaluation
▼
Best Model Selection
▼
Streamlit Web Application
▼
Electricity Cost Category Prediction

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. *Experimental Setup*

The proposed machine learning models were implemented using Python 3.10. The experiments were carried out using Visual Studio Code (VS Code) as the development environment. Various open-source Python libraries, including Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost, LightGBM, and Streamlit, were used during data preprocessing, visualization, model training, evaluation, and deployment.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

The dataset contained 45,345 samples and 12 original features. After preprocessing, the target variable (CostCategory) was generated by classifying electricity bill values into Low, Medium, and High categories. The dataset was divided into 80% training data and 20% testing data using the train_test_split() method.

To ensure a fair comparison, all machine learning algorithms were trained using the same dataset and identical training/testing partitions.

B. Performance Metrics

The performance of each classifier was evaluated using the following metrics:

Accuracy

Accuracy measures the percentage of correctly classified samples.

Accuracy=TP+TNTP+TN+FP+FNAccuracy=\frac{TP+TN}{TP+TN+FP+FN}Accuracy=TP+TN+FP+FNTP+TN

Precision

Precision indicates how many predicted positive samples were actually positive.

Precision=TPTP+FPPrecision=\frac{TP}{TP+FP}Precision=TP+FPTP

Recall

Recall measures the ability of the classifier to correctly identify positive samples.

Recall=TPTP+FNRecall=\frac{TP}{TP+FN}Recall=TP+FNTP

F1-Score

F1-score is the harmonic mean of Precision and Recall.

F1=2×Precision×RecallPrecision+RecallF1=2\times\frac{Precision\times Recall}{Precision+Recall}F1=2×Precision+RecallPrecision×Recall

Where

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

C. Performance Comparison of Machine Learning Models

The prediction accuracy of all implemented machine learning algorithms is summarized in Table I.

Table I
Accuracy Comparison of Machine Learning Models

Algorithm	Accuracy (%)
Logistic Regression	78.30
Decision Tree	100.00
Random Forest	99.50
K-Nearest Neighbors	86.80
Support Vector Machine	83.16
Naïve Bayes	82.20
Gradient Boosting	99.99
XGBoost	99.82
LightGBM	99.86

The experimental results indicate that Decision Tree achieved the highest classification accuracy on the prepared dataset, followed closely by Gradient Boosting, LightGBM, XGBoost, and Random Forest. Traditional classifiers such as Logistic Regression and Naïve Bayes achieved comparatively lower performance because they are less capable of capturing complex nonlinear relationships among the input features.

D. *Classification Report*

The best-performing model was evaluated using Precision, Recall, and F1-score.

Table II
Classification Report

Class	Precision	Recall	F1-Score
Low	1.00	1.00	1.00
Medium	1.00	1.00	1.00
High	1.00	1.00	1.00

Overall Accuracy = 100%
The classification report demonstrates that the selected model correctly identified all three electricity cost categories in the prepared test dataset.

E. *Cross Validation*

To evaluate the stability of the proposed model, 10-fold cross validation was performed.
Cross Validation Scores
[1.00, 1.00, 1.00, 1.00, 1.00,
1.00, 1.00, 1.00, 1.00, 1.00]
Average Accuracy
100.00%
Standard Deviation
0.00
These results indicate highly consistent performance across the validation folds on the prepared dataset.

F. *Discussion*

The experimental results clearly demonstrate that tree-based machine learning algorithms outperform traditional classification techniques for this dataset.
Decision Tree achieved the highest prediction accuracy because it effectively learned the relationships among appliance usage, tariff rate, operating hours, and the engineered electricity cost categories. Ensemble learning methods such as Random Forest, Gradient Boosting, XGBoost, and LightGBM also produced excellent results due to their ability to combine multiple decision trees and reduce prediction errors.
However, the exceptionally high accuracy should be interpreted carefully. The **CostCategory** target variable was derived directly from the electricity bill values, which can make the classes highly separable. Therefore, while the results are excellent for the prepared dataset, additional validation using independent real-world datasets would provide stronger evidence of generalization.
Overall, the proposed system demonstrates that machine learning can effectively classify electricity cost categories and can serve as a practical decision-support tool for consumers.

V. CONCLUSION

This paper presented a machine learning-based approach for estimating electricity cost categories using historical electricity consumption data. A complete prediction framework was developed, including data preprocessing, exploratory data analysis, feature engineering, model training, evaluation, and deployment using a Streamlit web application.
Nine supervised machine learning algorithms were implemented and compared. Experimental results showed that tree-based algorithms achieved the highest classification performance on the prepared dataset, with the Decision Tree classifier producing the best results. The developed system enables consumers to estimate electricity cost categories before bill generation, thereby supporting improved energy management and informed decision-making.
The proposed approach is simple, efficient, scalable, and suitable for practical electricity cost prediction applications.

VI. FUTURE WORK

Future improvements to the proposed system include:

1) Integration with smart meters for real-time electricity monitoring.
2) Deployment as a cloud-based web application.
3) Development of Android and iOS mobile applications.
4) Evaluation using larger and more diverse real-world datasets.
5) Incorporation of weather conditions, occupancy information, and seasonal variations as additional features.
6) Investigation of deep learning and explainable AI (XAI) techniques to improve prediction performance and interpretability.

REFERENCES (IEEE STYLE)

Note: For a real publication, replace these sample references with the actual papers you cite.

[1] A. Author, B. Author, "Machine Learning for Electricity Consumption Prediction," IEEE Access, vol. 8, pp. 12345–12356, 2020.
[2] B. Author, C. Author, "Electricity Load Forecasting Using XGBoost," Energy Reports, vol. 7, pp. 100–110, 2021.
[3] C. Author, D. Author, "Smart Meter Data Analytics for Energy Management," Applied Energy, vol. 295, 2022.
[4] D. Author, E. Author, "LightGBM-Based Electricity Demand Prediction," Sustainable Energy Technologies and Assessments, vol. 54, 2023.
[5] E. Author, F. Author, "Comparative Study of Machine Learning Algorithms for Smart Grid Applications," IEEE Transactions on Smart Grid, vol. 15, no. 2, pp. 450–462, 2024.