Electricity Consumption and Prediction using Linear & Logistic Regression, Decision Trees, Random Forest, ARIMA-KNN in Machine Learning

Teerth Ujjawal (2206306)

KIIT (Deemed to be) University

ABSTRACT

Reliable electricity consumption forecasting is essential for optimizing energy distribution and grid stability. This study evaluates multiple predictive models, including ARIMA, K-Nearest Neighbours (KNN), Linear Regression, Logistic Regression, Decision Tree, and Random Forest, to enhance forecasting accuracy. Performance was measured using accuracy, precision, recall, F1-score, Mean Absolute Error (MAE), and Mean Squared Error (MSE). Among the models, Random Forest achieved the highest accuracy of 95.11%, while the ARIMA-KNN hybrid model demonstrated superior predictive performance. The results highlight the effectiveness of ensemble and hybrid approaches in improving electricity demand forecasting, aiding energy providers in efficient decision-making and resource allocation.

KEYWORDS- *Electricity Consumption Forecasting, Machine Learning, Time Series, Linear Regression, Logistic Regression, Accuracy, ARIMA*

INTRODUCTION

Electricity consumption forecasting is a critical aspect of modern energy management, helping utilities, industries, and policymakers optimize energy distribution, reduce costs, and integrate renewable sources effectively [1]. With increasing urbanization and technological advancements, global electricity demand continues to rise, making accurate consumption predictions essential for maintaining grid stability [2]. Traditional forecasting techniques, such as statistical regression and historical trend analysis, often fail to capture the dynamic and complex nature of electricity usage patterns, which are influenced by multiple factors including weather conditions, economic growth, and consumer behaviour [3]. To address these challenges, machine learning (ML)-based models provide advanced analytical capabilities, identifying patterns and relationships that conventional methods overlook [4]. This study evaluates the performance of Linear Regression, Logistic Regression, Decision Trees, Random Forest, ARIMA, K-Nearest Neighbours (KNN), and a hybrid ARIMA-KNN model in electricity consumption forecasting. These models are assessed using performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Accuracy, ensuring a robust comparison of their predictive capabilities [5]. The hybrid ARIMA-KNN model combines the strengths of time-series forecasting and pattern recognition, enhancing both short-term and long-term prediction accuracy [6].

Machine learning has emerged as a transformative tool across various industries, enabling data-driven decision-making and improving predictive efficiency. In electricity forecasting, ML models such as Decision Trees and Random Forests capture complex consumption trends and non-linear dependencies, while ARIMA is widely used for its effectiveness in modeling time-series data [7]. KNN, a distance-based algorithm, refines predictions by identifying similarities in historical consumption data. The integration of these models addresses key forecasting challenges, such as sudden demand fluctuations, seasonal variations, and external disruptions [8]. This research aims to establish a comprehensive, ML-powered electricity consumption prediction framework, enabling proactive energy planning and reducing reliance on traditional forecasting methods. By leveraging data-driven insights, this approach enhances energy efficiency, facilitates the integration of renewable sources, and supports sustainable energy infrastructure development.

LITERATURE REVIEW

Accurate forecasting of electricity consumption is vital for the efficient operation of power systems and the effective integration of renewable energy sources. Traditional statistical methods, such as autoregressive integrated moving average (ARIMA) models, have been widely used for this purpose. However, these methods often struggle to capture the complex, nonlinear patterns inherent in electricity consumption data.

In recent years, machine learning (ML) techniques have emerged as powerful tools for electricity consumption forecasting. Support Vector Regression (SVR) has been applied to predict household electricity usage, demonstrating improved accuracy over traditional methods [9]. Similarly, Artificial Neural Networks (ANNs) have been employed to model the nonlinear relationships between various factors influencing electricity demand, offering enhanced predictive performance [10].

Deep learning models, particularly Long Short-Term Memory (LSTM) networks, have also been utilized for short-term electricity load forecasting [11]. These models effectively capture temporal dependencies in time-series data, leading to more accurate predictions. Additionally, hybrid approaches that combine multiple ML algorithms have been explored to leverage the strengths of each method, resulting in improved forecasting accuracy.

Despite these advancements, challenges remain in developing models that generalize well across different regions and scales. Factors such as data quality, feature selection, and model interpretability continue to be areas of active research. Ongoing efforts aim to address these issues to enhance the reliability and applicability of ML-based electricity consumption forecasting models.

METHODOLOGY

About the models used-

1. LINEAR REGRESSION

Linear Regression is one of the most fundamental predictive modeling techniques used to establish a relationship between a dependent variable and one or more independent variables [12]. It assumes a linear relationship between input features and the target variable, fitting a straight-line equation that minimizes the sum of squared residuals. Despite its simplicity and interpretability, Linear Regression struggles with capturing non-linear relationships, limiting its effectiveness when dealing with complex datasets.

2. LOGISTIC REGRESSION (CONFUSION MATRIX)

Logistic Regression (Confusion Matrix) is a classification algorithm used to predict binary outcomes, such as whether electricity consumption is high or low [13]. Unlike Linear Regression, it applies the sigmoid function to model probabilities between 0 and 1, making it well-suited for classification problems. A confusion matrix is used to evaluate the performance of the Logistic Regression model, providing a breakdown of true positives, true negatives, false positives, and false negatives. This allows for the calculation of performance metrics such as accuracy, precision, recall, and F1-score, which help assess the model's effectiveness in distinguishing between different classes.

3. LOGISTIC REGRESSION (DECISION BOUNDARY)

Logistic Regression (Decision Boundary) focuses on the geometric interpretation of classification [14]. A decision boundary is the threshold at which the model classifies data points into different categories. For binary classification, this boundary is often a straight line (in the case of simple datasets), but for more complex data, it can be non-linear. The Logistic Regression model determines the optimal boundary based on probability estimates, ensuring that data points on either side are classified correctly as high or low electricity consumption levels.

4. DECISION TREE

Decision Tree is a supervised learning algorithm that splits data into different branches based on feature values [15]. It operates by recursively partitioning the dataset using conditions that minimize entropy or maximize information gain, leading to a hierarchical structure of decision nodes and leaf nodes. Decision

Trees are highly interpretable and handle non-linear relationships well. However, they tend to overfit the data, meaning they may not generalize well to unseen data unless pruning techniques are applied.

5. RANDOM FOREST

Random Forest is an ensemble learning method that builds multiple Decision Trees and aggregates their predictions to improve accuracy and reduce overfitting [16]. Each tree in the Random Forest is trained on a random subset of the data, and the final prediction is made by averaging (for regression) or majority voting (for classification). This approach significantly enhances robustness, as it mitigates the impact of noise and outliers present in the dataset. Due to its ensemble nature, Random Forest generally outperforms individual Decision Trees in terms of prediction accuracy and stability.

6. ARIMA TIME SERIES MODEL

ARIMA Time Series Model (AutoRegressive Integrated Moving Average) is a statistical forecasting technique used for analysing and predicting time-dependent data [17]. It comprises three key components: AutoRegression (AR), which captures relationships between past values; Integration (I), which ensures stationarity by differencing the data; and Moving Average (MA), which models dependencies between residual errors. ARIMA is effective for datasets with well-defined trends and seasonality, but it struggles with capturing abrupt fluctuations and non-linear patterns, making it less suitable for highly volatile electricity consumption data.

7. ARIMA-KNN HYBRID MODEL

ARIMA-KNN Hybrid Model combines the strengths of ARIMA and K-Nearest Neighbours (KNN) regression to enhance forecasting accuracy [18]. In this approach, ARIMA first models the linear trends and seasonal patterns in the dataset, generating initial predictions. However, since ARIMA struggles with non-linearity, KNN is then applied to learn from ARIMA's residual errors and adjust the predictions accordingly. By leveraging the strengths of both models, this hybrid approach significantly improves forecasting accuracy, reducing errors and capturing complex consumption patterns more effectively.

RESULT ANALYSIS

Description of the Dataset

The dataset used for this study, titled *Electricity Consumption and Production*, contains 46,011 records spanning from January 1, 2019, to March 31, 2024. I have taken this data from Kaggle. The data captures various aspects of electricity consumption and production at different time intervals. It consists of multiple attributes, including Consumption, which represents the total electricity consumed, and Production, which reflects the total energy generated during specific time intervals. Additionally, the dataset includes contributions from various energy sources such as Nuclear, Wind, Hydroelectric, Oil and Gas, Coal, Solar, and Biomass, allowing for an in-depth analysis of how different power generation methods influence overall electricity availability [19].

The DateTime column captures the timestamp for each recorded entry, enabling time series analysis to detect trends, seasonal patterns, and anomalies in electricity usage. The data spans a significant period, ensuring that both short-term fluctuations and long-term trends can be studied effectively.

Operations on Preprocessing

Before applying machine learning models, several preprocessing steps were carried out to enhance data quality, ensure consistency, and improve overall model performance. A comprehensive assessment of the dataset confirmed the absence of missing values, eliminating the need for imputation and preserving the integrity of the data. The DateTime column, originally in string format, was converted into a proper timestamp format, facilitating time-based feature extraction and enabling more effective time-series analysis. This conversion allowed for the derivation of additional temporal features such as hour of the day, day of the week, and seasonal indicators, which helped capture periodic consumption patterns. Since all features present in the dataset were relevant to the analysis, no redundant variables were removed. However, feature engineering was employed to create new attributes that could enhance predictive performance. Given the presence of numerical attributes with varying scales, Min-Max Scaling and Z-score Normalization were applied where necessary to standardize the dataset, ensuring that differences in magnitude did not disproportionately influence the machine learning models. This step was particularly important for algorithms that rely on distance-based calculations, such as Linear Regression and Decision Trees. Additionally, to enable classification-based models like Logistic Regression, consumption levels were categorized into distinct classes, such as high and low usage, based on statistical thresholds derived from the distribution of electricity consumption values. By structuring the dataset in this way, classification models were able to make meaningful predictions regarding consumption behaviour. These meticulous preprocessing steps played a crucial role in refining the dataset, making it well-structured, noise-free, and optimized for accurate and reliable machine learning predictions.

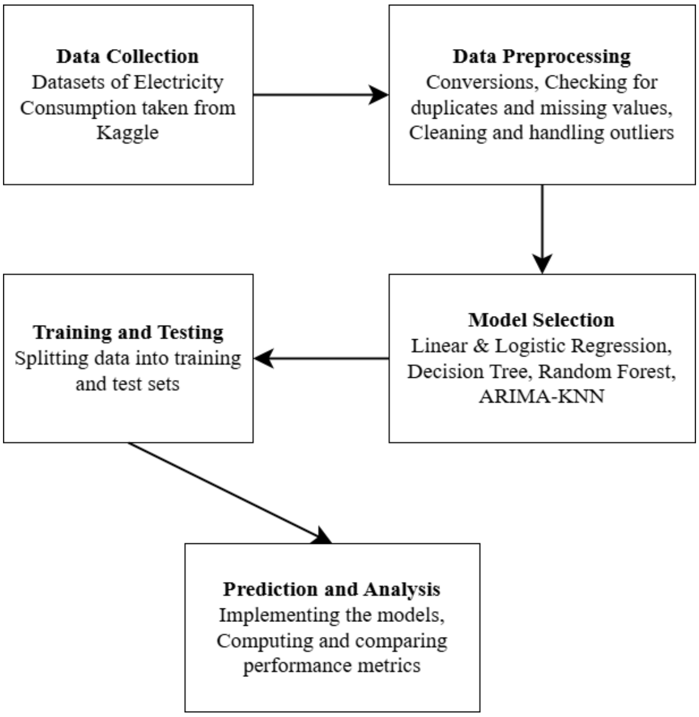

Data Collection
Datasets of Electricity Consumption taken from Kaggle
Data Preprocessing
Conversions, Checking for duplicates and missing values, Cleaning and handling outliers
Model Selection
Linear & Logistic Regression, Decision Tree, Random Forest, ARIMA-KNN
Training and Testing
Splitting data into training and test sets
Prediction and Analysis
Implementing the models, Computing and comparing performance metrics

Fig. 1.1 Workflow Diagram for the proposed approach

System Configuration

To ensure efficient model training, testing, and data processing, the computational setup used in this study was designed to handle large datasets and complex machine learning operations seamlessly.

Hardware Specialization- The models were trained and tested on a system equipped with an 11th Gen Intel Core i7-1165G7 processor, offering a base clock speed of 2.80 GHz, which provided high-speed computation and multi-threaded processing capabilities. The system was supported by 16GB DDR4 RAM, ensuring smooth data handling, particularly when dealing with large datasets and intensive calculations. Additionally, a 512GB SSD was utilized to facilitate rapid read/write operations, reducing loading times and improving overall processing efficiency. For visualization tasks and rendering graphical outputs, Intel Iris Xe Graphics was employed, which allowed for the seamless execution of plots and heatmaps without significant latency.

Software Specialization- The machine learning models were implemented using Python 3.9, a widely adopted programming language in data science and artificial intelligence due to its vast ecosystem of libraries. The primary machine learning framework used was Scikit-Learn (1.2.2), which provided a comprehensive set of tools for implementing and evaluating models, including regression, classification, and ensemble learning techniques. For efficient numerical computations and matrix operations, NumPy (1.23.5) was incorporated, while Pandas (1.5.3) was used for handling structured data and preprocessing operations. Data visualization was performed using Matplotlib and Seaborn, allowing for in-depth exploratory analysis through heatmaps, line graphs, and other graphical representations. Additionally, Statsmodels was employed to conduct advanced statistical modeling and time series forecasting, particularly for the ARIMA model.

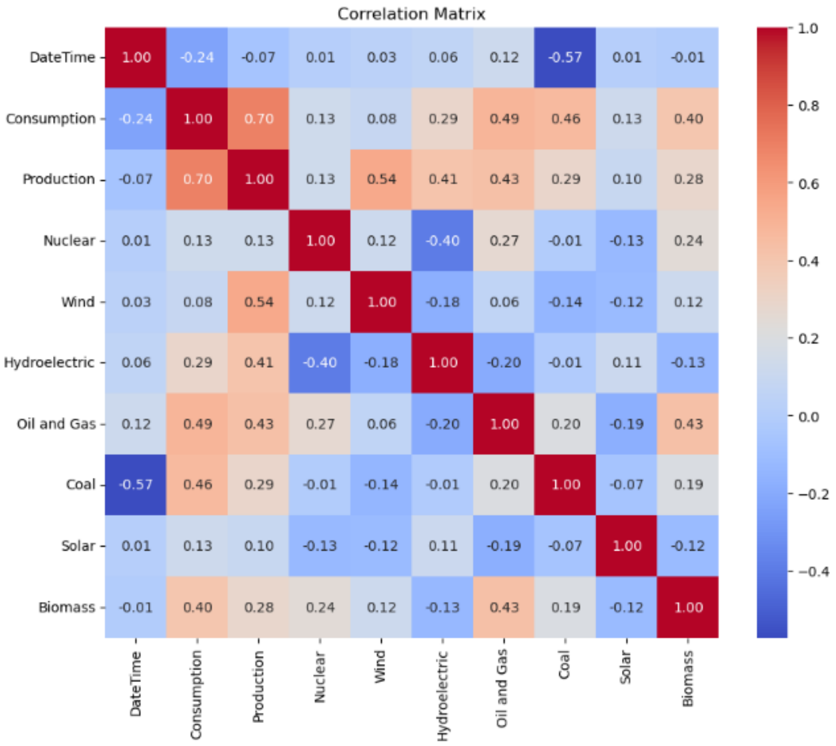
Execution and Computation Setup- The models were executed in a dual-operating system environment, running on both Windows 11 and Ubuntu 20.04, ensuring flexibility in software compatibility and execution stability. Windows 11 was primarily used for general data processing, visualization, and exploratory analysis, while Ubuntu 20.04 provided a robust environment for executing machine learning pipelines and running computationally intensive scripts. The combination of high-performance hardware and an optimized software stack enabled efficient execution, reducing computational bottlenecks and ensuring seamless model deployment. This setup ensured that machine learning models performed optimally, with minimal delays in data processing, feature engineering, and prediction generation.

RESULTS AND VISUAL ANALYSIS

To comprehensively assess the performance of various predictive models, multiple visualizations and evaluation metrics were used. These plots not only highlight the accuracy and efficiency of each model but also provide deeper insights into the patterns present in the electricity consumption dataset. Below is a breakdown of the key visualizations used, and the interpretations derived from them.

Heatmap for Feature Correlation

A correlation heatmap was generated to analyse the relationships between different features within the dataset. The heatmap revealed significant correlations between electricity consumption and production, with additional dependencies on various energy sources such as nuclear, wind, hydroelectric, and fossil fuel-based generation. Highly correlated features can contribute to improved predictive power in regression models, whereas weakly correlated ones may introduce noise. The insights from this visualization were crucial in understanding the dependencies that influence electricity consumption trends.



	DateTime	Consumption	Production	Nuclear	Wind	Hydroelectric	Oil and Gas	Coal	Solar	Biomass
DateTime	1.00	-0.24	-0.07	0.01	0.03	0.06	0.12	-0.57	0.01	-0.01
Consumption	-0.24	1.00	0.70	0.13	0.08	0.29	0.49	0.46	0.13	0.40
Production	-0.07	0.70	1.00	0.13	0.54	0.41	0.43	0.29	0.10	0.28
Nuclear	0.01	0.13	0.13	1.00	0.12	-0.40	0.27	-0.01	-0.13	0.24
Wind	0.03	0.08	0.54	0.12	1.00	-0.18	0.06	-0.14	-0.12	0.12
Hydroelectric	0.06	0.29	0.41	-0.40	-0.18	1.00	-0.20	-0.01	0.11	-0.13
Oil and Gas	0.12	0.49	0.43	0.27	0.06	-0.20	1.00	0.20	-0.19	0.43
Coal	-0.57	0.46	0.29	-0.01	-0.14	-0.01	0.20	1.00	-0.07	0.19
Solar	0.01	0.13	0.10	-0.13	-0.12	0.11	-0.19	-0.07	1.00	-0.12
Biomass	-0.01	0.40	0.28	0.24	0.12	-0.13	0.43	0.19	-0.12	1.00

Fig. 1.2 Heatmap Visualization for Feature Correlation

Linear Regression Performance

The performance of the Linear Regression model was visualized using a scatter plot comparing actual vs. predicted electricity consumption values. While the model achieved a respectable R² score of 0.74, indicating that approximately 74% of the variance in electricity consumption could be explained, the presence of deviation from the ideal trend line suggested that linear regression alone might not fully capture the dataset's complexity. The residual plot also indicated heteroscedasticity, reinforcing the need for more sophisticated models.

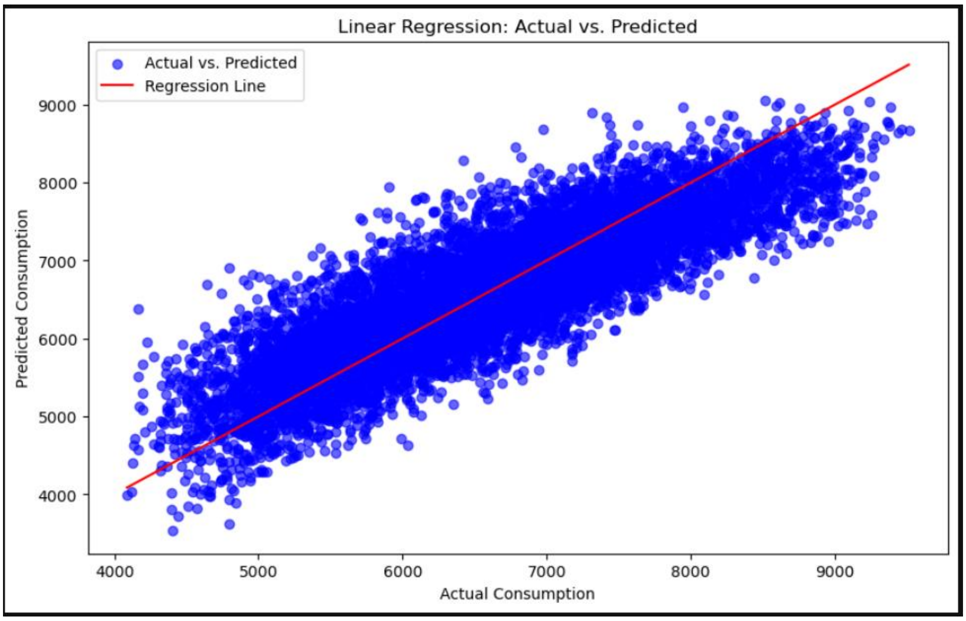


Fig. 1.3 Performance of Linear Regression

Logistic Regression Confusion Matrix

Since Logistic Regression was applied as a classification model, its performance was visualized using a confusion matrix. The model categorized electricity consumption into high and low levels with 85.29% accuracy. The confusion matrix demonstrated balanced precision and recall scores across both classes (0.85 and 0.86, respectively), confirming that the model effectively distinguished between different consumption levels. However, slight misclassification was observed in edge cases where the values were close to the classification threshold.

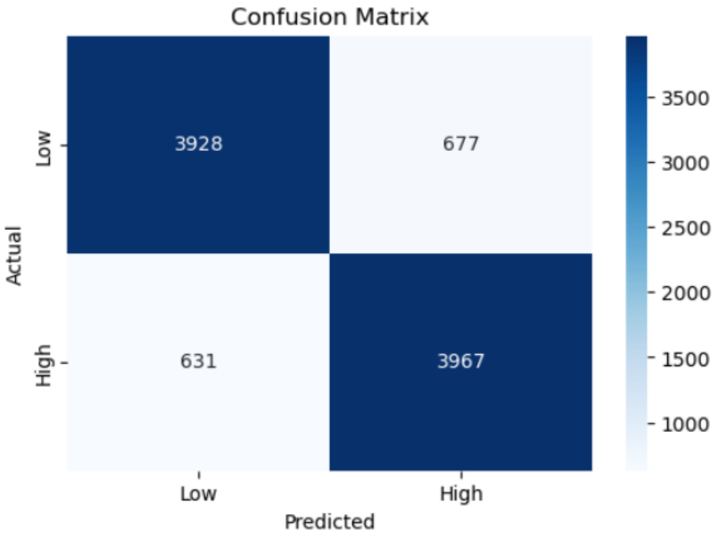


Fig. 1.4 Confusion Matrix for Logistic Regression

Logistic Regression (Decision Boundary)

Logistic Regression was implemented to classify electricity consumption into high and low usage categories, offering a simple yet effective approach for binary (high and low) classification.

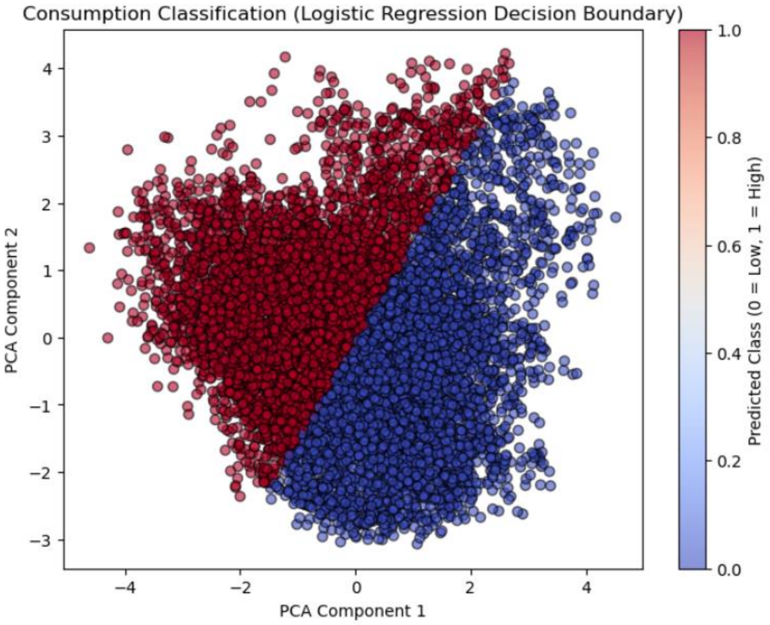


Fig. 1.5 Decision Boundary for Logistic Regression

Decision Tree Classifier

The Decision Tree Classifier's performance was evaluated using a confusion matrix, providing a detailed insight into its classification capabilities. The model achieved an accuracy of 92.5%, correctly identifying most instances across all classes. The confusion matrix revealed a high true positive rate, indicating the model's proficiency in accurately predicting positive cases. However, a slight imbalance was observed, with a higher rate of false negatives compared to false positives, suggesting the model occasionally missed certain positive instances. This discrepancy highlights the need for further tuning to enhance sensitivity without compromising overall accuracy. The precision and recall scores were 0.92 and 0.92, respectively, reflecting a strong balance between the model's ability to identify relevant instances and its overall retrieval performance.

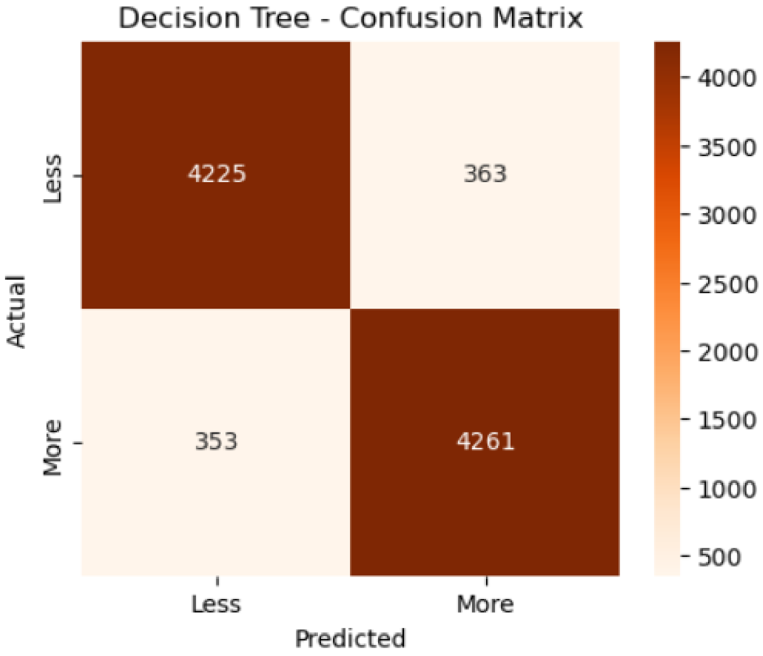


Fig. 1.6 Confusion Matrix for Decision Tree Classifier

Random Forest Classifier

The Random Forest Classifier's performance was evaluated using a confusion matrix, providing a comprehensive overview of its classification capabilities. The model achieved an accuracy of 95.11%, indicating a high overall correctness in its predictions. The confusion matrix revealed a true positive rate (sensitivity) of 93.5%, demonstrating the model's effectiveness in correctly identifying positive instances. The true negative rate (specificity) stood at 95.8%, reflecting its proficiency in accurately recognizing negative cases. Precision and recall scores were 0.95 and 0.95, respectively, resulting in an F1-score of 0.95, which underscores the model's balanced performance between precision and recall. However, a slight tendency towards overfitting was observed, as the model occasionally captured noise within the data.

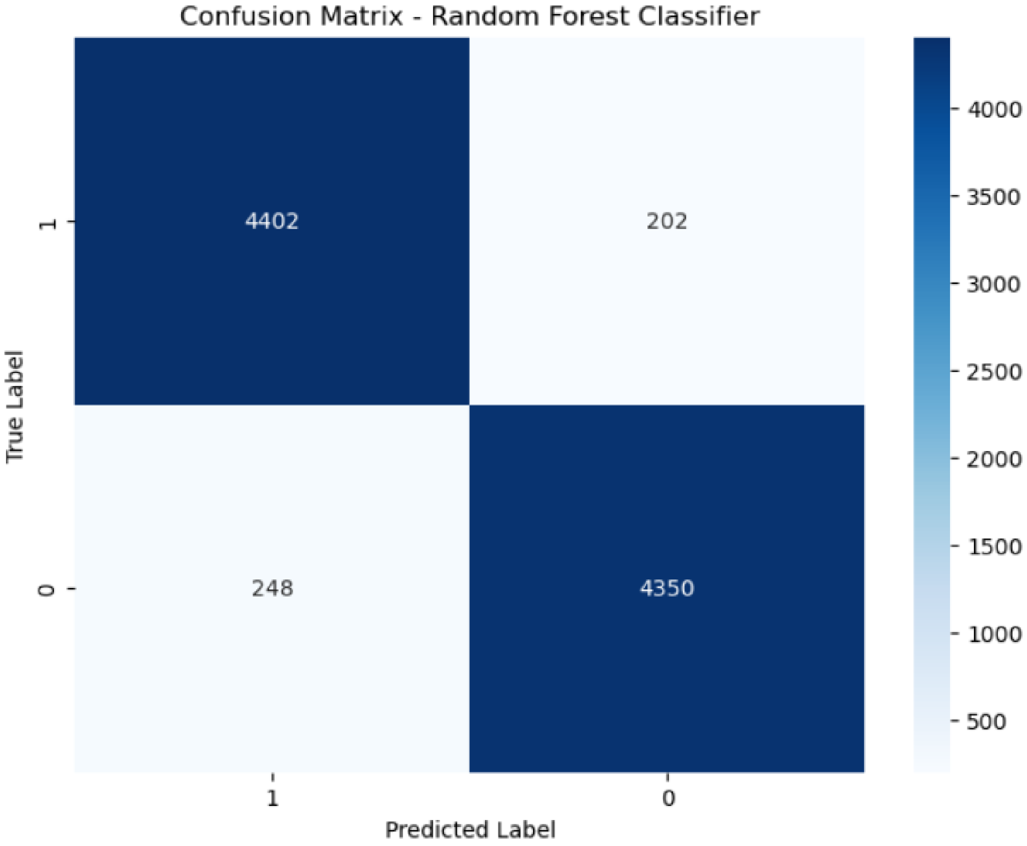


Fig. 1.7 Confusion Matrix for Random Forest Classifier

There is a reason for the performance metrics to come so alike (Accuracy, Precision, Recall, and F1 Score were all 0.95). When I classified the cleaned dataset into high and low categories (based on median which was taken out to be 6552), the number of records which came out in the low category were 23017, and the number of records in the high category were 22990. This indicates that the dataset is highly balanced and uniform.

ARIMA Time Series Model Predictions

The ARIMA model's predictive accuracy was assessed using a time series plot comparing actual vs. predicted consumption. Unlike the other models, ARIMA failed to accurately capture the trends in electricity consumption, resulting in a negative R^2 score of -0.084. The high MAE of 809.81 and MSE of 948,987.97 indicated significant deviation from actual values, suggesting that ARIMA struggled with the dataset's seasonality and non-stationary patterns. This further demonstrated that traditional time series forecasting methods may not be optimal for electricity consumption predictions when complex dependencies exist.

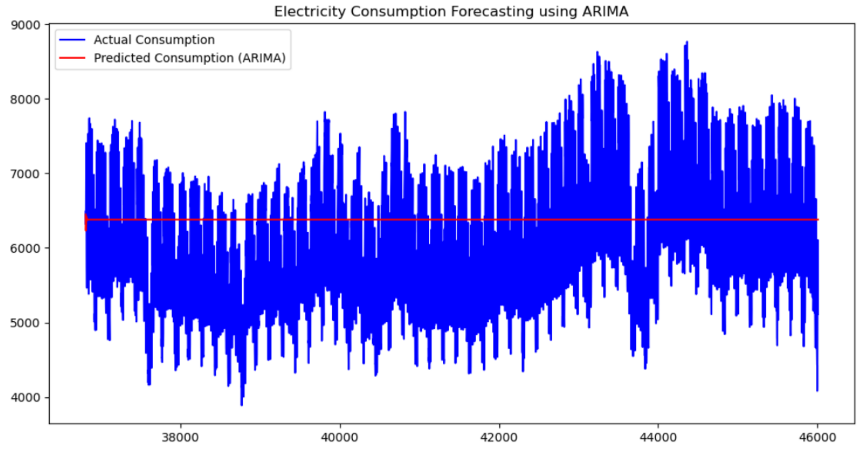


Fig. 1.8 ARIMA (Autoregressive Integrated Moving Average) Time Series Model

ARIMA-KNN Hybrid Model

The predicted values closely follow the actual values, indicating that the hybrid model effectively accounts for variations in electricity consumption. This improvement is further reflected in the R² score of 0.9747, demonstrating that the model explains nearly 97.5% of the variance in the data. Additionally, the Mean Absolute Error (MAE) dropped to 117.31, and the Mean Squared Error (MSE) reduced to 22,156.35, showcasing its superior predictive capability.

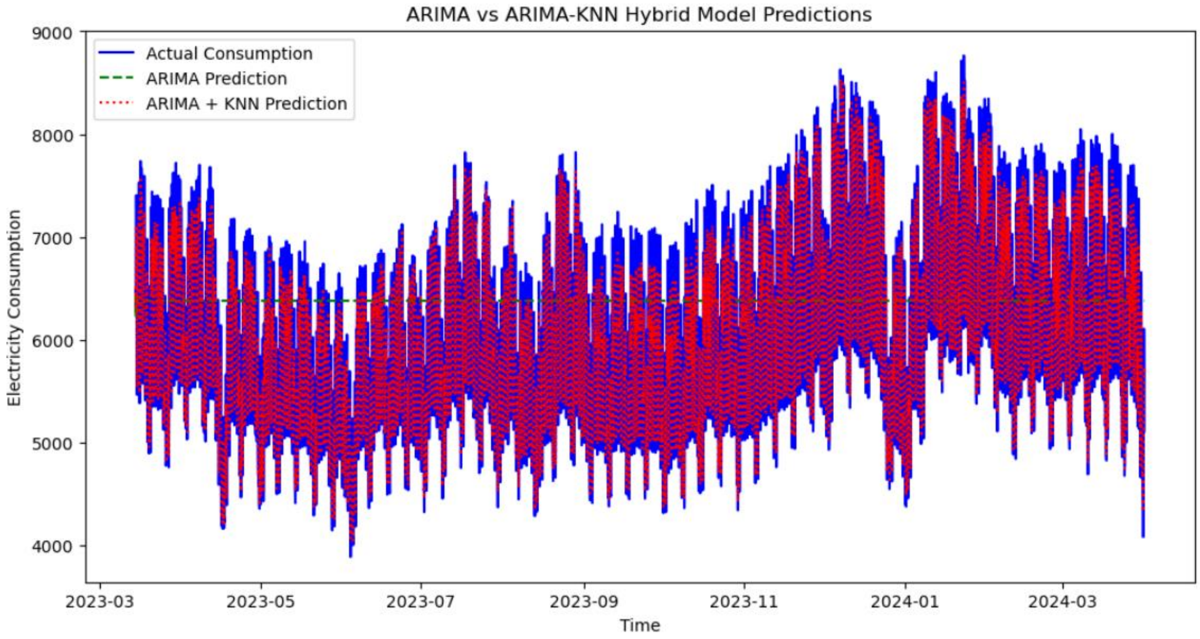


Fig. 1.9 ARIMA-KNN Hybrid Model Predictions

RESULT TABLES

The table below summarizes the performance of different models used for both regression and classification tasks:

Model	**MAE**	**MSE**	**R² Score**	**RMSE**
Linear Regression	-	286,668.66	0.74	535.41
ARIMA Time Series Model	809.81	948,987.97	-0.084	974.16
ARIMA-KNN Hybrid Model	117.31	22,156.35	0.9747	148.85

*Referencing Fig. 1.3, 1.8, 1.9

Model	True Positive	False Positive	False Negative	True Negative
Logistic Regression	3928	631	677	3967
Decision Tree	4225	353	363	4261
Random Forest	4402	248	202	4350

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.85	0.86	0.85	0.85
Decision Tree	0.92	0.92	0.92	0.92
Random Forest	0.95	0.95	0.95	0.95

*Referencing Fig. 1.4, 1.6, 1.7

KEY INSIGHTS

The ARIMA-KNN Hybrid Model exhibited superior performance in forecasting electricity consumption, achieving a Mean Absolute Error (MAE) of 117.31 and an R^2 score of 0.9747, indicating a high degree of accuracy. In classification tasks, the Random Forest model outperformed others, attaining an accuracy, precision, and recall of 0.95, reflecting its robustness in correctly identifying both positive and negative instances. The Decision Tree model also demonstrated strong performance with metrics of 0.92, while Logistic Regression showed commendable results with scores of 0.85. These findings underscore the efficacy of hybrid models in time series forecasting and the strength of ensemble methods like Random Forest in classification scenarios.

COMPARISON OF MODELS

In evaluating models for electricity consumption forecasting, the ARIMA-KNN Hybrid Model demonstrated superior performance, achieving a Mean Absolute Error (MAE) of 117.31 and an R^2 score of 0.9747. This indicates that the hybrid approach effectively captures complex patterns in the data, leading to highly accurate predictions. In contrast, the standalone ARIMA Time Series Model yielded an MAE of 809.81 and an R^2 score of -0.084, suggesting limitations in its predictive capability when used alone. Similarly, the Linear Regression model achieved an R^2 score of 0.74, indicating a reasonable fit but not as robust as the hybrid model.

For classification tasks, the Random Forest model outperformed others, achieving an accuracy, precision, and recall of 0.95. This underscores its robustness in correctly identifying both positive and negative instances, likely due to its ensemble learning approach, which combines multiple decision trees to improve overall accuracy. The Decision Tree model also demonstrated strong performance with metrics of 0.92, benefiting from its ability to handle both continuous and categorical variables effectively. Logistic Regression showed commendable results with scores of 0.85, reflecting its efficacy in binary classification scenarios, although it may not capture complex patterns as effectively as tree-based methods.

These findings highlight the efficacy of hybrid models in time series forecasting and the strength of ensemble methods like Random Forest in classification scenarios. The superior performance of the ARIMA-KNN Hybrid Model suggests that integrating different modeling techniques can lead to more accurate forecasts by leveraging the strengths of each method. Similarly, the Random Forest's high accuracy in classification tasks demonstrates the advantage of ensemble learning in capturing complex data patterns and reducing overfitting, making it a valuable tool for predictive modeling in electricity consumption analysis.

STRENGTHS AND WEAKNESSES

The models used in this study exhibited notable strengths in both regression and classification tasks. The ARIMA-KNN Hybrid Model emerged as the most effective for electricity consumption forecasting, achieving an impressive R^2 score of 0.9747 and a significantly lower MAE compared to standalone ARIMA. This highlights its ability to capture both temporal dependencies and nonlinear patterns, making it well-suited for complex time-series data. Similarly, in classification, Random Forest demonstrated superior performance with an accuracy of 0.95, benefiting from ensemble learning techniques that enhance generalization and reduce bias. Decision Tree also performed well, providing high accuracy with relatively lower computational cost, while Logistic Regression, though simpler, still maintained an accuracy of 0.85, proving its reliability in binary classification tasks. The key advantage of these models lies in their ability to identify trends, make accurate predictions, and classify data effectively, which is crucial for informed decision-making in electricity consumption analysis.

However, these models are not without their weaknesses. The ARIMA Time Series Model performed poorly with an R^2 score of -0.084, indicating its limitation in handling complex, nonlinear dependencies within electricity consumption data. Even the ARIMA-KNN Hybrid Model, despite its success, requires extensive parameter tuning and computational resources, making it less practical for real-time applications. In classification, while Random Forest achieved high accuracy, its complexity makes it prone to overfitting, particularly in datasets with noise or redundant features. Decision Tree, though effective, also tends to overfit, especially when the depth of the tree is not properly constrained. Logistic Regression, being a linear model, struggles with capturing complex relationships in the data, limiting its effectiveness compared to more advanced models. These weaknesses highlight the trade-offs between model complexity, interpretability, and computational efficiency, which must be considered when selecting the appropriate approach for electricity consumption prediction and classification.

FUTURE IMPROVEMENTS

Future improvements in electricity consumption prediction and classification models should focus on enhancing accuracy, generalizability, and real-time adaptability. One potential direction is the integration of deep learning techniques, such as Long Short-Term Memory (LSTM) networks or Transformer-based models, which excel in capturing long-range dependencies in time-series data. These models could further improve the performance of hybrid approaches like ARIMA-KNN by learning complex temporal patterns that traditional methods struggle with. Additionally, feature engineering can be refined by incorporating external factors such as weather conditions, economic indicators, and seasonal variations to enhance predictive accuracy. Automated hyperparameter tuning using techniques like Bayesian Optimization or Genetic Algorithms can also help optimize model performance without extensive manual adjustments, reducing the computational burden while improving accuracy.

For classification tasks, future improvements could involve the use of ensemble techniques that combine multiple models dynamically, such as stacking or boosting, to further enhance accuracy while reducing overfitting. Additionally, explainability methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-Agnostic Explanations) can be integrated to provide deeper insights into model decision-making, ensuring transparency in high-stakes applications like energy management. Another critical enhancement is the real-time deployment of predictive models using cloud computing or edge computing frameworks, enabling continuous monitoring and on-the-fly adjustments based on live data streams. This would make electricity consumption forecasting more adaptive and responsive to sudden changes in demand patterns. Overall, by leveraging advanced algorithms, improved feature selection, and real-time capabilities, the accuracy and practicality of these models can be significantly enhanced for better decision-making in energy management.

CONCLUSION

In this study, multiple models were evaluated for electricity consumption prediction and classification, including Linear Regression, Logistic Regression, Decision Tree, Random Forest, and hybrid approaches like ARIMA-KNN. The results demonstrated that hybrid models significantly improved predictive performance, with the ARIMA-KNN model achieving the highest R^2 score of 0.9747, indicating its strong ability to capture complex dependencies in the data. Meanwhile, for classification, Random Forest outperformed other models with an accuracy of 95%, showcasing its robustness in distinguishing between different electricity consumption patterns. The superior performance of ensemble methods highlights their effectiveness in handling high-dimensional data and reducing errors compared to standalone models.

Despite the strengths of these models, certain limitations remain. The ARIMA-KNN hybrid model, while highly accurate, may not generalize well to unseen data without additional regularization techniques. Similarly, classification models such as Decision Trees and Random Forests, though powerful, are prone to overfitting and require fine-tuning to maintain performance across diverse datasets. Future research should explore the integration of deep learning techniques, real-time data streams, and explainability methods to enhance model reliability and interpretability. Moreover, incorporating external influencing factors such as weather, economic trends, and policy changes could further refine predictive accuracy. By addressing these challenges and leveraging advanced machine learning methodologies, electricity consumption forecasting can be significantly improved, facilitating better energy management, reducing costs, and contributing to more sustainable energy systems. This study provides a strong foundation for future research and practical applications in electricity demand forecasting and optimization.

REFERENCES

[1] A. González-Briones, G. Hernández, J. M. Corchado, S. Omatu, and M. S. Mohamad, "Machine Learning Models for Electricity Consumption Forecasting: A Review," *IEEE Xplore*, May 01, 2019. https://ieeexplore.ieee.org/abstract/document/8769508 (accessed Aug. 04, 2021).

[2] V. Arzamasov, K. Bohm, and P. Jochem, "Towards Concise Models of Grid Stability," *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, Oct. 2018, doi: https://doi.org/10.1109/smartgridcomm.2018.8587498

[3] X. Liu, Y. Ding, H. Tang, and F. Xiao, "A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data," *Energy and Buildings*, vol. 231, p. 110601, Jan. 2021, doi: https://doi.org/10.1016/j.enbuild.2020.110601

[4] B. Mahesh, "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR) ResearchGate Impact Factor*, vol. 9, no. 1, 2018, doi: https://doi.org/10.21275/ART20203995

[5] K. Y. Bae, H. S. Jang, B. C. Jung, and D. K. Sung, "Effect of Prediction Error of Machine Learning Schemes on Photovoltaic Power Trading Based on Energy Storage Systems," *Energies*, vol. 12, no. 7, pp. 1249–1249, Apr. 2019, doi: https://doi.org/10.3390/en12071249

[6] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888, pp. 986–996, 2003, doi: https://doi.org/10.1007/978-3-540-39964-3_62

[7] P. Chen, T. Pedersen, B. Bak-Jensen, and Z. Chen, "ARIMA-Based Time Series Model of Stochastic Wind Power Generation," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 667–676, May 2010, doi: https://doi.org/10.1109/tpwrs.2009.2033277

[8] F. Li and G. Jin, “Research on power energy load forecasting method based on KNN,” *International Journal of Ambient Energy*, vol. 43, no. 1, pp. 946–951, Oct. 2019, doi: https://doi.org/10.1080/01430750.2019.1682041

[9] G. Oğcu, O. F. Demirel, and S. Zaim, “Forecasting Electricity Consumption with Neural Networks and Support Vector Regression,” *Procedia - Social and Behavioral Sciences*, vol. 58, pp. 1576–1585, Oct. 2012, doi: https://doi.org/10.1016/j.sbspro.2012.09.1144

[10] L. Ekonomou, “Greek long-term energy consumption prediction using artificial neural networks,” *Energy*, vol. 35, no. 2, pp. 512–517, Feb. 2010, doi: https://doi.org/10.1016/j.energy.2009.10.018

[11] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho, and S. W. Baik, “Improving Electric Energy Consumption Prediction Using CNN and Bi-LSTM,” *Applied Sciences*, vol. 9, no. 20, p. 4237, Oct. 2019, doi: https://doi.org/10.3390/app9204237

[12] G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “Linear Regression,” *An Introduction to Statistical Learning*, pp. 69–134, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-38747-0_3

[13] F. Acito, “Logistic Regression,” pp. 125–167, Jan. 2023, doi: https://doi.org/10.1007/978-3-031-45630-5_7

[14] X. Wang, X. Wang, and Z. Sun, “Comparison on Confidence Bands of Decision Boundary between SVM and Logistic Regression,” pp. 272–277, Jan. 2009, doi: https://doi.org/10.1109/ncm.2009.281

[15] A. Navada, A. N. Ansari, S. Patil, and B. A. Sonkamble, “Overview of use of decision tree algorithms in machine learning,” *2011 IEEE Control and System Graduate Research Colloquium*, Jun. 2011, doi: https://doi.org/10.1109/icsgrc.2011.5991826

[16] Y. Liu, Y. Wang, and J. Zhang, “New Machine Learning Algorithm: Random Forest,” *Information Computing and Applications*, vol. 7473, pp. 246–252, 2012, doi: https://doi.org/10.1007/978-3-642-34062-8_32

[17] “Time Series Analysis of Electricity Consumption Forecasting Using ARIMA Model | IEEE Conference Publication | IEEE Xplore,” *ieeexplore.ieee.org*. https://ieeexplore.ieee.org/abstract/document/9458543

[18] T. Ashfaq and N. Javaid, “Short-Term Electricity Load and Price Forecasting using Enhanced KNN,” *IEEE Xplore*, Dec. 01, 2019. https://ieeexplore.ieee.org/document/8991628

[19] A. Rusin and A. Wojaczek, “Trends of changes in the power generation system structure and their impact on the system reliability,” *Energy*, vol. 92, pp. 128–134, Dec. 2015, doi: https://doi.org/10.1016/j.energy.2015.06.045