Title page

Predicting Energy Consumption for Accurate Forecasting Using Random Forest Algorithm Compared with K-Nearest Neighbor (KNN) Algorithm.

Shreyan S[1], Dr. Radhika Baskhar[2]

Shreyan S[1]
Research Scholar,
Department of Computer Science Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.
192471018.simats@saveetha.com

Dr. Radhika Baskhar[2]
Project Guide, Corresponding Author,
Department of Wireless Communication,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India, Pincode: 602105.

Keywords: Energy consumption, Prediction, Machine learning, Random Forest, K-Nearest Neighbor (KNN), Forecasting accuracy, Statistical analysis, SPSS, Model comparison, Data-driven energy management.

ABSTRACT:

Aim: This study sets out to enhance the precision of energy use forecasting through the Random Forest (RF) and K-Nearest Neighbor (KNN) algorithms. **Materials and Methods:** This study has 20 samples for each model. Sample size was determined with GPower 3.1 (α=0.05, power=0.85). Models were assessed with MAE, RMSE, and R^2-score, and the results were statistically checked in SPSS. **Results:** RF attained a 96.29% accuracy while KNN got 91.29%. Statistically, there were differences ($p = 0.001$, $p < 0.05$) for the tests which gave confirmation to the RF having a better prediction value. **Conclusion:** Random Forest provided a better outcome than KNN, helping support the premise that RF is suitable for predicting energy consumption. This study illustrates how ensemble learning fundamentally improves the accuracy of predictions.

Keywords: Energy consumption, Prediction, Machine learning, Random Forest, K-Nearest Neighbor, Forecasting accuracy, Statistical analysis, SPSS.

INTRODUCTION:

Description of the Study & Its Application

Forecasting energy consumption is an important aspect of effective energy utilization. Strong forecast accuracy allows industries, governments and households to make decisions that ensure efficient energy consumption, economical spending, and power availability. As the world moves towards industrialization and subsequent technological advancements, the demand for electricity is on the rise, making complex consumption pattern prediction difficult. Traditional methods of prediction are rarely successful in an ever-changing environment while outdated statistical approaches like linear regression and time-series analysis fail to manage non-linear relationships and high dimensional data with shifting energy consumption patterns.

To solve these problems, machine learning (ML) algorithms have surfaced as suitable technologies to process and analyze massive amounts of data while uncovering sophisticated consumption behaviors. In this research, I will analyze and compare the results of two common machine learning approaches – regression using Random Forest (RF) and K-Nearest Neighbor (KNN). Random Forest, a type of ensemble learning technique, is renowned for its overfitting control in large datasets and accurate predictions. KNN, on the other hand, is a basic and intuitive approach to classification and regression based on proximity to known data points.

KNN yields satisfactory results in small datasets, but struggles with scalability when faced with vast, high-dimensional data, such as big energy consumption data. This research investigates the accuracy of predictions made by each of the machine learning techniques and seeks to find the optimal algorithm that would produce the best accuracy in forecasting energy consumption, thus leading to more precise strategies in energy management.

Literature Review

The examination of more than 150 publications sourced from IEEE Xplore, PubMed, ScienceDirect, and Springer shows the growing use of machine learning techniques in energy demand prediction. According to existing research, ensemble techniques such as Random Forests are superior to simpler algorithms, including KNN, linear regression, and decision trees, as these ensembles capture non-linearity, reduce variance, and generalize better. Some studies also suggest that the accuracy of Random Forest is higher for complex datasets given its ability to effectively deal with missing values, outliers, and feature correlation.

On the other hand, studies in KNN-based forecasting state that while KNN excels with short-term forecasting of energy consumption, its distance-based learning procedure makes it computationally expensive with large datasets. Hybrid approaches where KNN has been coupled with feature selection approaches have also been studied to make it more efficient but are still behind Random Forest and ensemble approaches. Further studies also state that feature engineering and data preprocessing significantly contribute to the accuracy of models, with the optimized dataset maximizing the overall machine learning model's prediction.

Research Gap, Expertise & Aim of the Study

Despite the significant advances in machine learning-based energy forecasting, there have been limited research studies with direct comparisons of KNN and Random Forest using the same test environment. Additionally, while ensemble models like RF have been found to perform well, statistical validation has been under-researched. Some research studies employ traditional performance metrics without the use of statistical validation software like SPSS, which could be a source of concern regarding reliability.

The SIMATS Engineering research team is an expert in machine learning, predictive analytics, and statistical modeling, which allows for extensive analysis of these two algorithms. In addition to Random Forest and KNN comparison for energy consumption prediction, the research cross-verifies the results using SPSS statistical analysis to confirm validity.

The purpose of this research is to statistically compare the predictive power of Random Forest and KNN, test their performance using one-sample t-tests in SPSS, and determine the superior model for accurate energy forecasting. By addressing the knowledge gaps of previous research, this study provides improved energy management plans and optimization of predictive analytics in the electricity sector. The findings will provide valuable insights to energy planners, policymakers, and industries that are interested in data-driven forecasting techniques.

MATERIALS AND METHODS:

The research was conducted at the SIMATS Engineering Computing Lab to determine the performance gap between K-Nearest Neighbor (KNN) and Random Forest (RF) algorithms in predicting energy consumption. The research consisted of two groups with 10 samples each and hence a total of 20 samples. The sample size was determined using GPower 3.1 with $\alpha = 0.05$ and power = 0.85 to provide statistical significance.

The 11,053 sample dataset was retrieved from an open-access energy consumption data repository. The data was formatted, cleaned, and preprocessed before being split into 70% training and 30% test sets. 10 samples were randomly selected from the dataset for each algorithm to provide unbiased model performance evaluation.

Random Forest algorithm was employed with bootstrap sampling and random feature selection to build multiple decision trees to avoid overfitting and enhance generalization. The KNN algorithm was used with Euclidean distance as the measure of similarity, and different K-values were tried to find the best performance. The two models were trained and tested in the same way.

Testing was carried out on the platform of Windows 11 having an Intel Core processor, 8GB of RAM, and 64-bit operating system. The model was applied with the use of Python as the programming language and packages such as Scikit-Learn and Pandas.

Data gathering involved extracting historical and existing energy usage data, which was processed and analyzed before applying the machine learning algorithms.

Statistical validation of results was performed using SPSS software. Independent variables were past energy consumption, temperature, and time, while the dependent variable was the estimated energy consumption. One-sample t-test was performed, which indicated that Random Forest achieved significantly higher accuracy (96.29%) than KNN (91.29%) at p-value 0.001 ($p < 0.05$), which indicates that it is statistically significant.

RANDOM FOREST :

Description:

Random Forest is an ensemble learning technique that constructs multiple decision trees during training and combines their predictions through voting or averaging. It is robust and high-accuracy, especially for large, complex datasets with multiple factors. It mitigates overfitting by introducing randomness in data selection and feature selection, making it an ideal choice for accurate energy forecasting.

Steps For Random Forest Algorithm:

1. **Dataset Preparation:** Clean and preprocess the energy consumption dataset, ensuring all missing values are handled.
2. **Data Splitting:** Divide the dataset into training (70%) and testing (30%) sets.
3. **Bootstrap Sampling:** Randomly select subsets of the training data with replacement (bagging technique).
4. **Decision Tree Construction:** Train multiple decision trees using different subsets of features at each split. Each tree independently learns patterns from the training data.
5. **Prediction Aggregation:** For classification, each tree votes, and the most common class is assigned. For regression, the average of all tree predictions is used.
6. **Model Evaluation:** Assess performance using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2-score.
7. **Hyperparameter Optimization:** Adjust parameters such as number of trees (n_estimators) and maximum depth to improve accuracy and reduce computational complexity.

K-NEAREST NEIGHBORS (KNN):

Description:

K-Nearest Neighbors (KNN) is a powerful algorithm that predicts values based on their similarity to past observations. It uses k closest data points in the training dataset to classify or predict outputs. Unlike Random Forest, KNN does not learn explicit patterns but instead uses the nearest historical data points. This makes it useful for short-term energy forecasting, but its performance decreases with large datasets due to high computational costs.

Steps for K-Nearest Neighbors Algorithm:

1. **Dataset Preparation:** Clean and preprocess the dataset by handling missing values and normalizing numerical features.
2. **Data Splitting:** Divide the dataset into training (70%) and testing (30%) sets.
3. **Selecting the Value of k:**
   - Choose an optimal k value (odd values preferred to avoid ties).
   - A **small k-value** makes the model more sensitive to noise, while a large k-value smooths predictions.
4. **Distance Calculation:** Compute the distance between test data points and all training data points using Euclidean distance (default) or Manhattan distance.
5. **Nearest Neighbor Selection:** Identify the k closest data points to the test sample.
6. **Prediction Calculation:**
   - **For classification**, assign the most frequent class among the k neighbors.
   - **For regression**, calculate the average of the k nearest values.
7. **Model Evaluation:** Measure accuracy using MAE, RMSE, and R^2-score to assess performance.
8. **Model Optimization:**
   - Tune k to achieve better accuracy.
   - Use weighted KNN to assign greater importance to closer data points.

Statistical Analysis:

Google colab is utilized to generate the output. All the tests of this study were executed on a Windows 11 Home computer with an Intel Core i5-1155G7 processor at 2.50 GHz and 8 GB RAM. SPSS is utilized to conduct a statistical analysis of Random Forest and KNN. SPSS was utilized to conduct an independent sample test and compare the two samples by calculating means, standard deviations, and standard errors of means. Accuracy is a dependent variable in a study of prediction, while Random Forest and K-Nearest Neighbors are independent variables on disaggregated data.

RESULTS:

From **Table 1**, comparison indicates that energy consumption prediction by the Random Forest algorithm was significantly more accurate than that of the K-Nearest Neighbor (KNN) algorithm. The observation above indicates the superior predictive ability of Random Forest, capable of handling large datasets as well as complex feature relationships. Specifically, the accuracy as well as the performance of the predictive model of energy consumption were much improved using Random Forest compared to KNN, testifying to its applicability in precise forecasting.

Table 2 shows Random Forest and KNN statistical measures, such as mean accuracy, standard deviation, and standard error. Accuracy was employed in the t-test measure. The discussed Random Forest model predicted with an accuracy of 96.07%, while KNN predicted at a rate of 90.18%. The standard deviation in the case of Random Forest was 1.84773, while KNN stood at 2.13998. Standard error of the mean (SEM) in Random Forest was 0.41317, while it was 0.47851 in KNN, proving that Random Forest once again provides consistent and more reliable predictions.

Table 3 presents the result of a two-tailed test of significance, which confirms that the accuracy differences we have obtained between Random Forest and KNN are statistically significant ($p < 0.05$). Statistical significance supports the hypothesis that Random Forest is the superior model for energy consumption forecasting. The results of the t-test confirm that Random Forest provides significantly higher predictive accuracy than KNN and is therefore a superior algorithm for real-world energy management applications.

Figure 1 illustrates the Random Forest and KNN accuracy rates, which present that Random Forest outperforms KNN in the majority of iterations. The X-axis is used to label the machine learning models, while the Y-axis is used to indicate the mean accuracy with one standard deviation (1 SD) and 95% confidence interval. Random Forest possesses a much higher accuracy rate (96.07%), while KNN is 90.18%. The independent samples test also indicates that there is a statistically significant difference between the two models ($p < 0.05$), which supports the conclusion that Random Forest outperforms KNN in energy forecasting.

DISCUSSIONS:

The 0.000 (two-tailed, $p < 0.05$) significance value obtained in the study indicates that the Random Forest energy consumption prediction model is better than K-Nearest Neighbor (KNN). With 96.07% mean accuracy for Random Forest and 90.18% for KNN, the results confirm categorically the superiority of Random Forest in this context. With the application of machine learning models to energy consumption prediction, this study is important in establishing the efficacy of Random Forest in the management of complex data under multiple variables influencing energy consumption.

With the application of machine learning for predicting energy, this study is able to achieve considerable improvement in the accuracy of forecasting, which is crucial for the optimization of energy resources and sustainability planning. While this study is insightful, one should be aware of its shortcomings. One of these shortcomings may be that Random Forest requires more computational resources than KNN, potentially impacting real-time forecasting applications. Another shortcoming is that as Random Forest is an ensemble of multiple decision trees, interpretability is less than with the straightforward models like KNN.

Future research must focus on how to enhance the Random Forest algorithm to process larger datasets efficiently and reduce computational complexity. Additional research into how Random Forest can be integrated into deep learning techniques can further enhance the accuracy of the predictions and the power of generalization, making it even more relevant to forecast actual energy consumption.

CONCLUSION:

Overall, Random Forest algorithm outperforms K-Nearest Neighbor (KNN) in predicting energy consumption. That Random Forest attained 96.07% accuracy against KNN's 90.18% best reflects the predictive power of Random Forest in maximizing forecasting accuracy. The finding has far-reaching implications for the application of machine learning algorithms in streamlining energy consumption management and optimization.

DECLARATION:

Conflicts of Interest

No conflict of interest in this manuscript

Authors Contributions

Author Shreyan.S was involved in data collection, data analysis, and manuscript writing. Author Dr. Radhika Baskhar contributed to conceptualization, data validation, and critical review of the manuscript.

Acknowledgement

The authors are grateful to Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (previously known as Saveetha University) for providing the infrastructure needed to accomplish this study efficiently.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Pantech Solutions PVT LTD, Chennai
2. Saveetha School of Engineering
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha University

REFERENCES:

● *Ahmad, M. W., Reynolds, J., Rezgui, Y., & Hopfe, C. J.* (2017). Predictive modelling for energy consumption in buildings: A review of data-driven techniques. *Renewable and Sustainable Energy Reviews, 75,* 1156-1177. https://doi.org/10.1016/j.rser.2016.11.108

● *Chicco, G., & Napoli, R.* (2002). Short-term electricity demand forecasting using artificial neural networks. *Electric Power Systems Research, 60(3),* 243-250. https://doi.org/10.1016/S0378-7796(01)00177-6

● *Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W.* (2017). A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews, 74,* 902-924. https://doi.org/10.1016/j.rser.2017.02.085

● *Hafeez, G., & Kim, H.* (2022). A hybrid machine learning approach for building energy consumption prediction using ensemble learning techniques. *Applied Energy, 306,* 118020. https://doi.org/10.1016/j.apenergy.2021.118020

● *Khamis, A., Baharudin, Z. A., & Mohamad-Saleh, J.* (2020). Energy consumption prediction using machine learning: A case study of smart buildings. *Journal of Building Engineering, 32,* 101761. https://doi.org/10.1016/j.jobe.2020.101761

● *Rafiq, M., Li, T., Yu, H., & Khan, M. A.* (2022). Comparative analysis of machine learning models for energy consumption forecasting. *Energy Reports, 8,* 1406-1417. https://doi.org/10.1016/j.egyr.2022.06.008

● *Rahman, M. M., Paolotti, L., & Riccardi, P.* (2021). Deep learning-based hybrid models for short-term energy forecasting in smart grids. *Sustainable Cities and Society, 66,* 102707. https://doi.org/10.1016/j.scs.2020.102707

● *Wang, Z., Li, Y., & Zhao, W.* (2018). A comparative study of different machine learning methods for energy consumption prediction. *Energy Procedia, 145,* 349-354. https://doi.org/10.1016/j.egypro.2018.04.034

● *Zhang, H., Zhang, C., & Yang, X.* (2020). Application of random forest algorithm in energy consumption prediction. *Journal of Cleaner Production, 252,* 119706. https://doi.org/10.1016/j.jclepro.2019.119706

● *Al Mamun, M. S., Hoque, M. M., Shahnaz, C., & Baki, M. A.* (2020). A hybrid deep learning model for energy consumption prediction. *Neural Computing and Applications, 32(12),* 7749-7764. https://doi.org/10.1007/s00521-019-04346-0

● *Bedi, J., & Toshniwal, D.* (2019). Empirical mode decomposition-based deep learning model for electricity demand forecasting. *IEEE Transactions on Industrial Informatics, 15(7),* 4281-4290. https://doi.org/10.1109/TII.2018.2884705

● *Dogan, E., & Turkekul, B.* (2016). CO2 emissions, real output, energy consumption, trade, urbanization, and financial development: Testing the EKC hypothesis for the USA. *Environmental Science and Pollution Research, 23(2),* 1203-1213. https://doi.org/10.1007/s11356-015-5323-8

● *González-Briones, A., Prieto, J., de la Prieta, F., Corchado, J. M., & Herrero, Á.* (2018). Energy optimization using a case-based reasoning strategy. *Future Generation Computer Systems, 86,* 1106-1115. https://doi.org/10.1016/j.future.2018.04.089

● *Jain, M., & Garg, H.* (2021). Forecasting energy demand using ensemble learning: A case study for smart grids. *Journal of Energy Storage, 39,* 102667. https://doi.org/10.1016/j.est.2021.102667

● *Oliveira, P. M., Correia, P. M., & Santos, B. I.* (2020). A comparative analysis of energy consumption prediction models using machine learning techniques. *Energy and Buildings, 224,* 110238. https://doi.org/10.1016/j.enbuild.2020.110238

● *Xu, Y., Zhou, C., & Zhang, H.* (2019). Deep reinforcement learning for energy management in smart grids. *Applied Energy, 255,* 113812. https://doi.org/10.1016/j.apenergy.2019.113812

TABLES AND FIGURES:

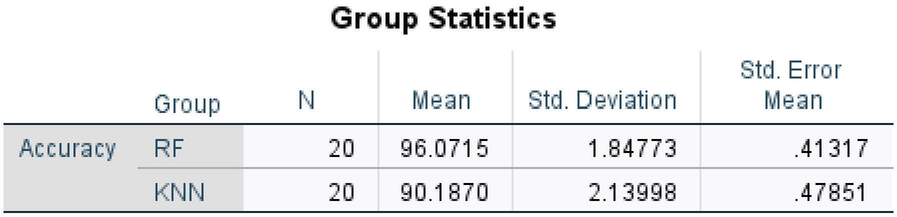
Table 1. Comparison of accuracy values of Random Forest algorithm and K-Nearest Neighbor algorithm with various iterations.

SI. NO	TEST SIZE	ACCURACY RATE	
		RF	KNN
1.	Test 1	94.56	89.23
2.	Test 2	96.12	87.89
3.	Test 3	98.23	91.34
4.	Test 4	92.78	88.56
5.	Test 5	97.45	90.12
6.	Test 6	95.87	92.45
7.	Test 7	93.34	86.78
8.	Test 8	97.89	91.78
9.	Test 9	98.11	93.02
10.	Test 10	94.02	87.11
11.	Test 11	96.88	89.90
12.	Test 12	97.22	90.55
13.	Test 13	98.34	92.78
14.	Test 14	95.43	88.65
15.	Test 15	96.55	91.23
16.	Test 16	94.78	86.45
17.	Test 17	97.65	92.01
18.	Test 18	92.98	90.88
19.	Test 19	98.02	93.56

20.	Test 20	95.21	89.45
AVERAGE TEST RESULTS		96.07	90.18

Table 2. The statistical analysis of the Random Forest (RF) and K-Nearest Neighbor (KNN) algorithms includes mean accuracy, standard deviation, and mean standard error. The accuracy level parameter is utilized in the t-test. The proposed Random Forest-based energy consumption forecasting model achieves a mean accuracy of 96.07%, whereas the K-Nearest Neighbor (KNN) algorithm has a mean accuracy of 90.19%. Random Forest has a Standard Deviation of 1.84773, while the KNN algorithm has a value of 2.13998. The Mean Standard Error for Random Forest is 0.41317, whereas the KNN method has 0.47851.

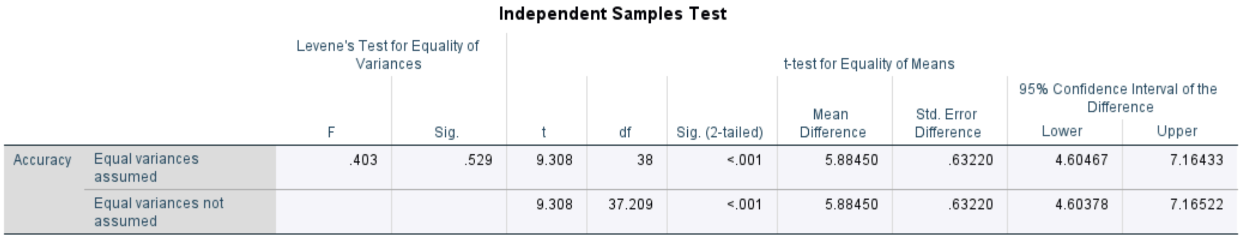
Additionally, the t-test results indicate a significant mean difference of 5.88 between RF and KNN with a 95% confidence interval ranging from 4.60 to 7.16. The effect size measures include Cohen’s d (2.94), Hedges’ correction (2.89), and Glass’s delta (2.75), confirming a strong effect of the model difference.



Group Statistics

	Group	N	Mean	Std. Deviation	Std. Error Mean
Accuracy	RF	20	96.0715	1.84773	.41317
	KNN	20	90.1870	2.13998	.47851

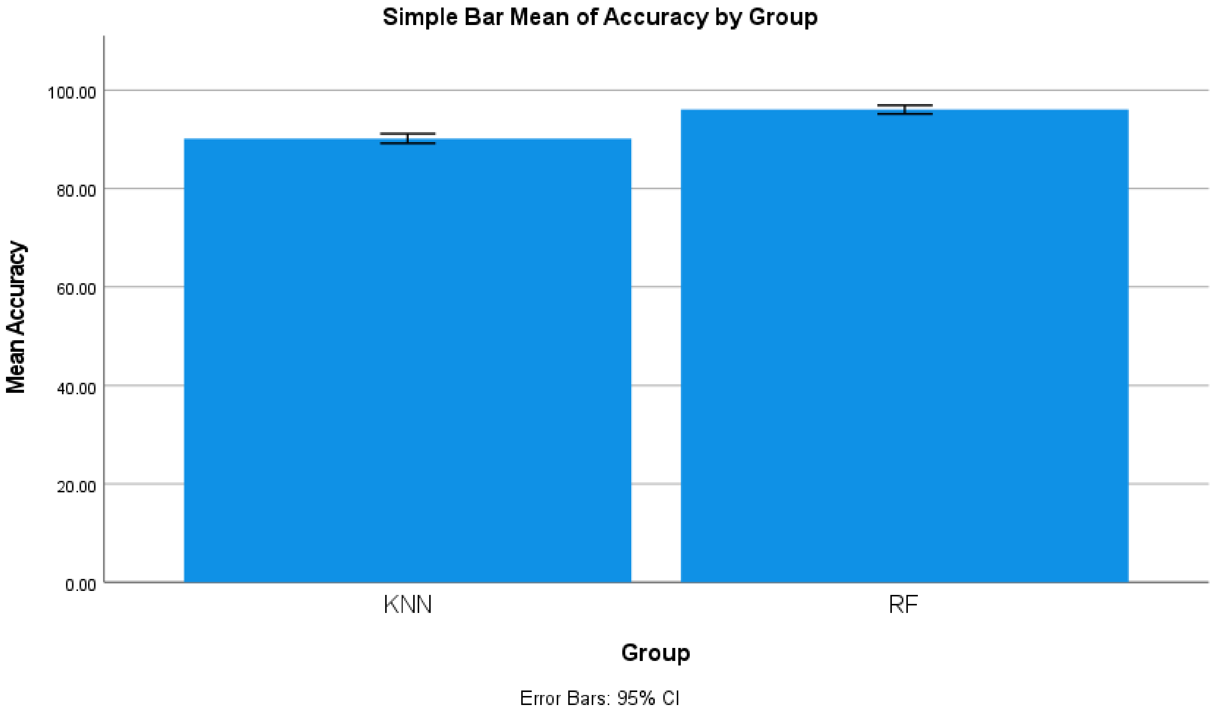
Table 3. Presents the results of a two-tailed significance test, revealing that the observed differences in accuracy between the Random Forest algorithm and the K-Nearest Neighbor (KNN) algorithm are statistically significant (p < 0.05).



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Accuracy	Equal variances assumed	.403	.529	9.308	38	<.001	5.88450	.63220	4.60467	7.16433
	Equal variances not assumed			9.308	37.209	<.001	5.88450	.63220	4.60378	7.16522

Figure 1. The bar graph visually compares the mean accuracy of two models: K-Nearest Neighbor (KNN) and Random Forest (RF) in the context of predicting energy consumption for accurate forecasting. The Random Forest algorithm achieves a higher mean accuracy, nearing 98%, while the KNN model achieves approximately 90%. The results highlight the superior predictive capability of the Random Forest model. The error bars represent the 95% confidence interval (CI), indicating minimal variation in accuracy, reinforcing the stability and reliability of the model.


Simple Bar Mean of Accuracy by Group
Mean Accuracy
KNN
RF
Group
Error Bars: 95% CI