Data Science and Engineering (DSE) Record, Volume 5, issue 1.

ANALYSIS OF FACTORS INFLUENCING ELECTRICITY CONSUMPTION OF UNIVERSITY'S DORMITORIES DURING POST-COVID-19 RECOVERY

Sukanya Sawanoi[1] and Pree Thiengburanathum [2]

[1] Master's Degree Program in Data Science, Chiang Mai University, Chiang Mai, Thailand
[2] College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand
sukanya_saw@cmu.ac.th

Abstract. In recent times, there has been a significant increase in global and Thai electricity consumption. This surge has led people to seek ways to save on electricity costs, such as installing solar panels. However, it appears to be a solution addressing the symptom rather than the root cause, as people continue to consume electricity at similar levels. Understanding the factors influencing electricity usage is crucial for tackling the root cause, as it enables the reduction of activities or behaviors leading to excessive energy consumption. This research aims to investigate the factors influencing electricity consumption in university dormitories, specifically focusing on the electricity bills. Data was collected through a total of 243 surveys, rigorously verified and prepared for analysis. The survey yielded a total of 35 factors, which were then analyzed to identify their relationships with electricity consumption. The 16 factors were found to correlate with electricity usage based on Spearman's correlation, while 19 factors were identified through MI. To simplify the data and reduce complexity, EFA was employed, resulting in only 7 common factors from both Spearman's correlation and MI analyses. Each dataset was utilized to build predictive models for electricity consumption using five algorithms: SVM, MLP, KNN, DT, and LR. The baseline model, performing best in terms of learning efficiency with the dataset analyzed for correlation with electricity consumption using MI's 19 factors and SVM, achieved a testing accuracy score of 0.5762. To enhance the processing efficiency of the baseline model, parameter tunings were made for the SVM, with C set to 1.5, gamma set to 4.699, and using the "rbf" kernel. Post-training and evaluation, the adjusted model exhibited a testing accuracy score of 0.7353, indicating that parameter tuning positively affected the predictive performance of the model for real-world scenarios. From the information gathered, it can be concluded that factors influencing electricity consumption include the number of notebooks operating on the Windows operating system, the duration of computer usage for both learning and gaming, activities such as ironing or using a hair dryer combined with turning on the air conditioner for heat dissipation, knowledge about electricity usage (e.g., choosing electrical appliances labeled with the number 5), and finally, attitudes towards electricity usage.

Keywords: Building Energy Consumption, Occupant Behavior, Influencing Factors.

1 Introduction

Nowadays, electricity consumption has increased steadily all over the world. Due to the business growth in some developing countries including Thailand (World Population Review. 2022), the electricity consumption has been rising significantly. According to figure 1, the graph showed an upward trend of the electricity consumption in Thailand between 2017 and 2021. Compared to the electricity consumption in 2017, the rate surged to 190,464 GWh in 2021 (by approximately 2.81%). The Ministry of Energy also mentioned that the percentage of Thailand's electricity consumption in 2022 increased by 1.3% (DEDE. 2022).

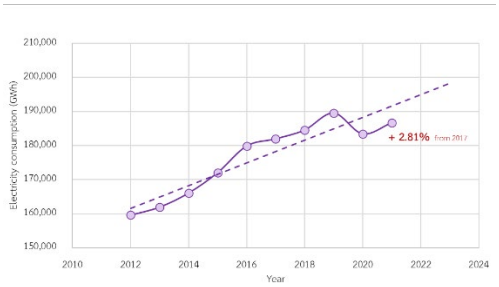


Figure1 The Electricity Consumption in Thailand from 2017 to 2021**.**

During the pandemic, people spending more time at home due to COVID-19 measures led to a peak in residential sector energy consumption in 2021 (around 54,290 GWh). Meanwhile, business sector electricity consumption gradually increased until 2019 and then consistently dropped to just over 41,000 GWh in 2021 (EPPO. 2022).

The university, viewed as a business entity, actively supports environmental and energy conservation policies. One crucial policy involves reducing energy consumption and promoting renewable energy on campus. The university's diverse buildings, with varying numbers, patterns, characteristics, and occupants like students, lecturers, officers, experience multiple functions in rooms like classrooms, labs, meeting areas, libraries, dorms, sports centers, and canteens. Operating in various capacities, the university faces significant monthly expenses, averaging millions of baht, for electricity bills (Srimode, P. 2010).

Universities often initiate electricity-saving campaigns, like Chiangmai University's "CMU Smart City – Clean Energy" (CMU'SDGs) in 2020. This campaign utilizes advanced technology to enhance people's lives sustainably and has been recognized as one of seven showcases by the Entity Conservation and Promotion Fund Office (ECPFO). Currently, over 40% of solar panels are installed on Chiang Mai University's campus, reducing dormitory electricity bills by 30%. While effective in lowering costs, the campaign does not directly address the main issue of minimizing energy consumption. Hence, careful consideration of factors influencing campus electricity consumption is imperative.

This study aims to identify the factors influencing electricity consumption in university's dormitories and develop a model that relevant factors influencing electricity consumption. In this study, we will collect data related to building energy through questionnaires and will distribute them on-campus dormitories during the post-COVID-19 recovery. In order to identify factors influencing electricity consumption, Spearman's rank correlation and Mutual Information will be utilized to analyze the relationship between factors and electricity usage in university dormitories. Exploratory Factor Analysis will be employed to examine the relationship between factors and reduce their number before utilizing the data to build a predictive model for electricity consumption. To establish a baseline classification model, five algorithms including SVM, KNN, MLP, DT, and LR will be employed. The baseline model with the best learning performance will be further developed by finding the most suitable parameter values. Finally, the performance of the optimized models will be compared to determine the most appropriate model. Besides, it is beneficial to understand which factors causing the higher demand of electricity consumption in each dormitory and be able to minimize the amount of energy use in the dormitories. These are expected to be the policy design for energy conservation of Chiang Mai university and the paradigm for the dormitory expansion in the future.

2 Literature Review

2.1 Analysis of factors influencing electricity consumption

Khumsayyai, A. (2007), studied electricity consumption behaviors among Chiang Mai University dormitory students via 378 questionnaires. The survey assessed attitudes toward energy use, energy literacy, and daily electricity consumption behaviors. Descriptive analysis revealed students' intermediate energy literacy levels. Most students somewhat agreed with electricity conservation attitudes, ranking their engagement in energy-saving behavior as above average.

Poolsawat, K. (2020) explored electricity characteristics in the residential sector and potential conservation in Thailand through descriptive analysis. The study focused on household and dwelling characteristics, along with energy consumption in electrical appliances. Enhancing Thailand's electricity conservation potential involved improving appliance performance. Short-term recommendations included prioritizing LED light bulbs due to their low investment cost, benefiting many low-income individuals. Long-term strategies involved upgrading air conditioners and refrigerators.

Laicane, I., Blumberga, D., Blumberga, A., & Rosa, M. (2015) conducted a study on household electricity savings, analyzing data from a survey and smart meter of a four-member family over one year (April 2013 to March 2014). The study focused on electricity appliance usage and daily timing. Energy efficiency evaluation indicated a potential 13% reduction in electricity consumption by 2020 compared to the current scenario. Forecasts suggested a decreasing trend in electricity consumption, particularly for electric water heaters in summer. Considering technological improvements, solar collectors were identified as a beneficial solution for energy savings.

2.2 Classification model for predicting electricity consumption

Thaiyanan, N. (2019), explores factors impacting electric energy use and forecasts electricity consumption in various campus buildings. Data were collected from six building categories (lecture, administration, multi-purpose, research-laboratory, dormitories, and parking) in 61 Kasetsart University and 81 Chulalongkorn University buildings. Factors like building characteristics, energy use, site and climate, building and system, user, and occupancy were analyzed. Feature selection identified significant factors, and a predictive model employed machine learning algorithms (Random Forest, Multiple Regression Analysis, and Neural Network). Results highlight building areas and levels as crucial factors influencing electricity consumption, with different buildings showing varied influences. Random Forest proved the most effective algorithm for predicting electricity consumption.

Shen, M., Sun, H., & Lu, Y. (2017) studied predicting household electricity consumption through multiple behavior interventions. Data, divided into three groups, covered demographics (family size, gender, income, education), energy-related behaviors (e.g., appliances), and big five personality traits. Support Vector Regression (SVR) was employed for electricity consumption prediction.Data and Methodology

3 Data and Methodology

3.1 Data

The data used in this study was collected through questionnaires distributed to students residing in university dormitories, specifically buildings 1 and 2 for male dorms and buildings 11 and 12 for female dorms, totaling four buildings. A total of 243 samples were collected with a 95% confidence level and a margin of error not exceeding 5%. The questionnaire was designed based on literature review and factors of interest, covering six sections: occupant characteristics, room information, occupant behaviors, knowledge of electricity consumption, attitude towards electricity consumption, and satisfaction with CMU's Wi-Fi network service.

3.2 Research framework

The overall design of the research framework illustrates the experimental process that begins with data collection, exploratory data analysis, data preprocessing, feature selection, and dimensionality reduction, aiming for prediction and evaluation.

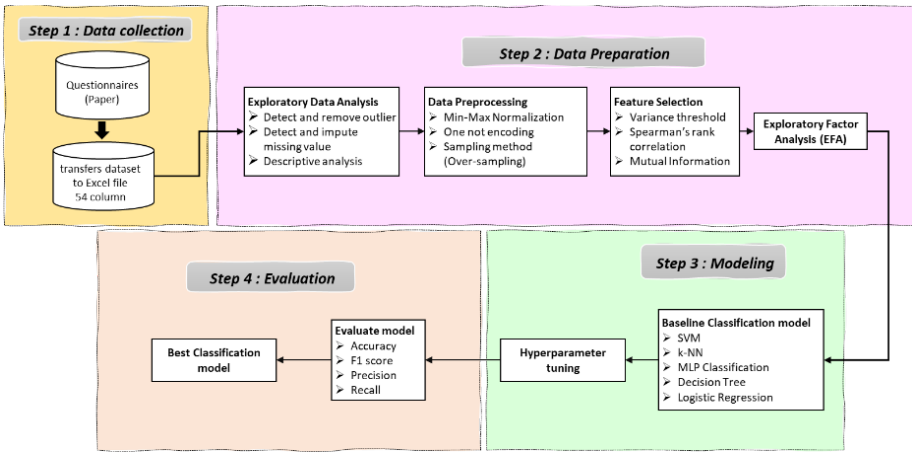


Figure 2 The research framework overall design

The research framework in Figure 2 comprises four steps. Step one involves conducting interviews with 243 samples using a questionnaire containing closed-ended and Likert scale questions. Following this, the paper data is digitized using Excel, resulting in a file with 54 columns and 243 samples.

The data preparation step is crucial before constructing the prediction model. Survey data often has incomplete information, necessitating a check for missing values in the Excel file. Samples with excessive missing values are removed, while acceptable ones are imputed with mean or mode values. Selected data is transformed for analysis, aggregating information for shared spaces. Descriptive Analytics explore preliminary data characteristics. Subsequently, data preprocessing involves normalization, encoding, and over-sampling. Feature selection follows, utilizing techniques like Variance threshold, Spearman's rank correlation, and Mutual Information. For high feature numbers, dimensionality reduction via Exploratory Factor Analysis (EFA) groups correlated features into factors.

In step three, the prepared dataset is split into the training dataset for model training and the testing dataset to assess prediction accuracy. Five algorithms, namely Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP), Decision Tree (DT), and Logistic Regression (LR), are used to create the baseline classification model. The 10-fold cross-validation technique is applied to prevent overfitting. The most efficient baseline model is chosen, and its parameters are adjusted for optimal category classification accuracy.

In the last step, models are evaluated and compared based on accuracy from each dataset. The model with the highest predictive performance is chosen, along with the data used, becoming the most influential factors impacting electricity consumption.

This research uses Python for programming which Pandas library is used for manipulating the data and data imputation. Sklearn library is used for building classification model. Factor analyzer library is used for EFA.

3.3 *Data Collection*

Upon completion of the questionnaire collection, the researcher transferred the data to Microsoft Excel, dedicating one column for each question. Certain questions were

expanded into multiple columns to accommodate multiple responses. For example, the first section of the questionnaire, focusing on general information about the first room occupant, utilized columns labeled "C7-1" and "C7-2" to capture multiple answers. Abbreviations for each questionnaire section, along with item numbers, were used as column labels. The entire questionnaire, comprising six sections with 45 questions, was entered into Microsoft Excel, resulting in a dataset of 243 samples and 54 variables.

3.4 *Data Preparation*

In the Data Preparation phase, an initial exploration of the collected data was performed to assess its accuracy, revealing a 34.16% missing value rate. Samples with missing values exceeding 30% were removed, and the remaining values were imputed using mean or mode. Outliers in variables related to electricity consumption and air conditioner temperature settings were identified and handled accordingly, resulting in a final dataset of 227 samples.

To facilitate machine learning analysis, transformations were applied to variables like the number of computers in the room (com_no) which is the aggregation of the number of computers for residents living in pairs in one room, the sum of scores for answering questions related to electricity usage knowledge (SumKnowledge) and range of electricity cost because the questionnaire, the specified range of electricity costs (R5new) does not align well with reality, resulting in a significant concentration of data in one class. Min-Max normalization was employed, and binary values were one-hot encoded. Random sampling addressed imbalances in the target variable, with the SMOTE technique generating synthetic samples.

The selection of variables is crucial, and methods like Variance Threshold, Spearman's correlation, and Mutual Information were used for this purpose. Additionally, factor analysis was employed to group related variables, identifying components explaining correlations between different variables.

These data preparation steps ensure the dataset's readiness for machine learning, enhancing its accuracy and relevance for subsequent analyses.

3.5 *Modeling and Evaluation*

In this step, we will begin by creating a baseline classification model to serve as a foundation or comparison for developing more complex models later on. This involves employing machine learning algorithms, specifically SVM, KNN, MLP, DT, and LR. Subsequently, the baseline model will be refined for enhanced performance through hyperparameter tuning.

Evaluation is crucial in ML as it assesses model performance and facilitates model comparison to select the best one. In this study, classification evaluation was con-ducted, utilizing the confusion matrix to assess prediction results from the generated ML model. The concept involves determining the proportions of what was predicted against what actually occurred. The confusion matrix provides various metrics such as Accuracy, Precision, Recall, and F1 score to evaluate prediction performance. Another technique employed for model assessment is K-Fold Cross Validation, which involves dividing data into parts and testing the model K times, using a different part as the test set in each iteration while using the remaining parts for training, enhancing the model's robustness and reliability.

4 Result

4.1 *Descriptive Analysis*

For explaining the characteristics of sample groups and gaining a basic understanding of the data, descriptive statistics were employed in the data analysis. In examining the room samples, it was found that the majority of occupied rooms accommodated two people per room, accounting for 81.9% of the sample, while only 18.1% had single occupants. It is plausible that rooms with two occupants may have a higher probability of consuming more electricity than single-occupancy rooms. However, considering only the number of residents in a room might not be sufficient; other variables should be taken into account, such as the number of computers used in each room. The analysis revealed that 56.83% of the rooms had two computers, 27.31% had one computer, and 6.17% of the rooms had no computers at all. Additionally, the floor level in the residence building was considered. Laicane, L. (2015), studied the electricity consumption in homes and found that the residential floor level could impact electricity consumption. Statistical analysis indicated that most residents lived on the third and fifth floors, constituting 26.87% and 22.30% of the sample, respectively. Therefore, it can be speculated that the residential floor level might influence electricity consumption, especially on the fifth floor.

For variables related to electricity usage behavior in the dormitory, such as the number of hours using air conditioning, it was found that the majority, 43.17%, tended to use air conditioning for no more than 2 hours. Additionally, there was a behavior of opening windows before turning on the air conditioning to release room heat, observed in 40.09% of cases. It is possible that this behavior may not significantly impact the electricity consumption in the dormitory. Furthermore, this behavior might align with the assessed interest in conserving natural resources and the environment, as indicated by survey responses, with 49.34% agreeing and 33.04% strongly agreeing with conservation values. Lastly, regarding the average knowledge score related to electricity usage, it was found to be 2.20 out of a maximum of 4 points. This indicates a moderate level of knowledge among respondents about electricity usage, consistent with the assessment of respondents' knowledge levels, where the majority (68.28%) perceived their knowledge about electricity usage as moderate.

In addition, there was an examination of the data distribution characteristics to enhance the appropriateness of the analysis technique selection. The evaluation included checking for a Normal Distribution, starting with the hypotheses H0: the data follows a Normal Distribution, and H1: the data does not follow a Normal Distribution. Subsequently, the K-S test statistic was computed based on the difference between the actual data distribution function and the expected Normal Distribution function. In the analysis, if the p-value obtained is less than the significance level of 0.05, H0 is rejected. The results indicated that for all variables analyzed, the p-values were less than 0.05. Therefore, it can be concluded that the data does not follow a Normal Distribution. When visualizing the data, as shown in Figure 3, it is evident that the distribution does not closely resemble a bell curve, exemplified by the behavior data of the residents.

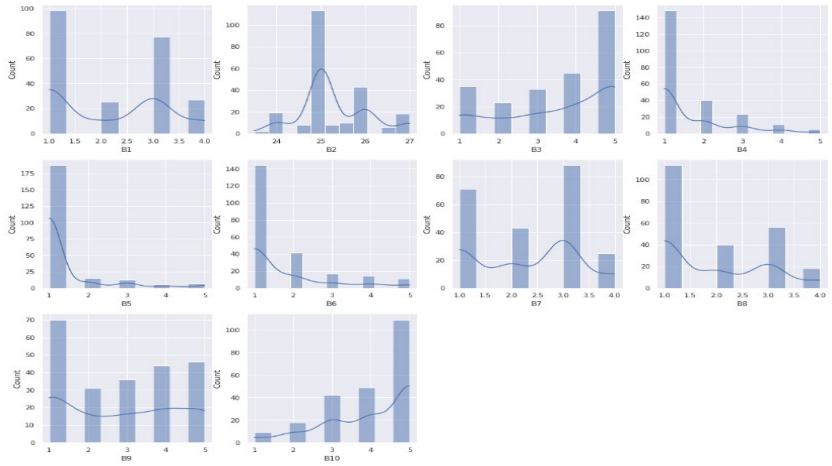


Figure 3 The distribution of the behavioral data of the student.

4.2 *Feature Selection*

Feature selection is crucial for enhancing the efficiency of the model and reducing the computational and memory resources required. Initially, the process involves removing variables where every sample has the same value or considering variables with zero variance. In this case, no variables are eliminated as each one has a variance greater than zero. This helps in improving the model's performance and optimizing resource usage in terms of computation and memory.

From Table 1, which displays the Spearman correlation coefficients between factor variables and electricity cost, it is observed that there are 16 variables significantly correlated with electricity cost at a significance level of 0.05. Starting with variables showing a positive correlation with electricity cost at a moderate level, “B1” is identified. Following this, variables positively correlated with electricity cost at a lower level include “C1_2”, “window”, “com_no”, “notebook”, “B8”, “A4”, and “R2_5”. Conversely, variables negatively correlated with electricity cost, indicating factors that do not contribute to an increase in electricity cost, include “C1_1”, “R2_1”, “B5”, “R3_0”, “R2_4”, “R2_2”, “K1”, and “B3”.

For assessing the relationship between various factors and electricity cost using MI, which can measure different types of relationships beyond linear correlations, a minimum MI value of 0 indicates no relationship between the paired variables. For this study, variables with MI values greater than 0.1 will be considered, as observed through the MI values graphically represented in Figure 4. The graph reveals a group of variables that cluster together or exhibit slight differences above the 0.01 threshold. These 19 variables include “A5”, “A4”, “B1”, “B2”, “B9”, “A1”, “A2”, “B3”, “SumKnowledge”, “B7”, “window”, “B10”, “B6”, “A3”, “K1”, “com_no”, “B8”, “notebook”, and “B4”.

Table 1 Correlation coefficient obtained by analyzing factors related to electricity cost by Spearman's correlation analysis.

Feature	Spearman's correlation	P-value	Feature	Spearman's correlation	P-value	Feature	Spearman's correlation	P-value
B1	**0.33**	**0**	R2_3	0.04	0.45	B10	-0.06	0.31
C1_2	**0.19**	**0**	macos	0.04	0.47	linux	-0.1	0.08
window	**0.19**	**0**	SumKnowledge	0.02	0.75	A1	-0.1	0.08
com_no	**0.19**	**0**	B9	0.02	0.79	**B3**	**-0.11**	**0.05**
notebook	**0.16**	**0**	desktop	0.01	0.8	**K1**	**-0.11**	**0.04**
B8	**0.14**	**0.01**	R1_1	-0.01	0.89	**R2_2**	**-0.14**	**0.01**
A4	**0.12**	**0.03**	A5	-0.01	0.85	**R2_4**	**-0.17**	**0**
R2_5	**0.11**	**0.05**	A3	-0.03	0.54	**R3_0**	**-0.17**	**0**
R3_1	0.09	0.1	A2	-0.04	0.52	**B5**	**-0.17**	**0**
B7	0.08	0.13	B2	-0.04	0.45	**R2_1**	**-0.23**	**0**
B4	0.08	0.14	R1_0	-0.04	0.42	**C1_1**	**-0.26**	**0**
B6	0.07	0.21	andriod	-0.05	0.39			

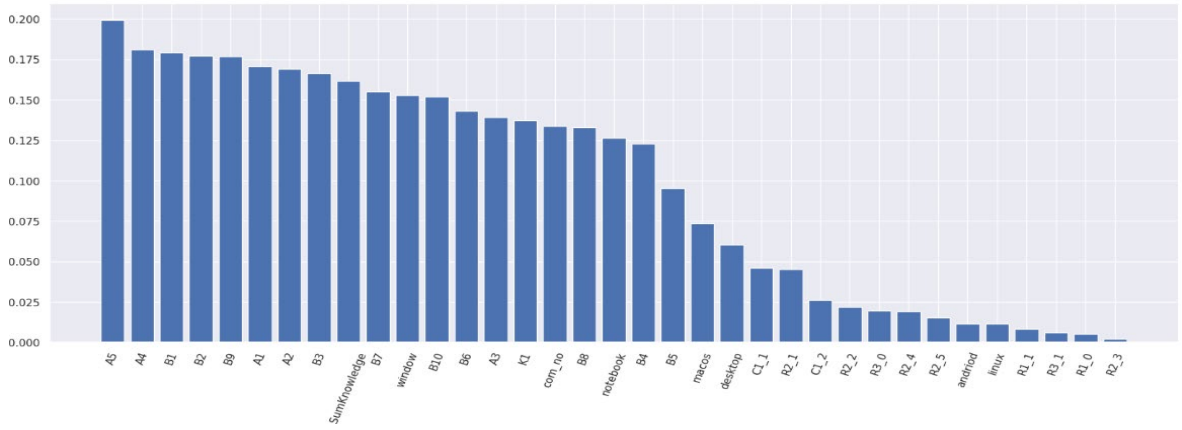


Figure 4 The Mutual Information score between factor and power

4.3 *Exploratory Factor Analysis*

The objective of conducting EFA is to identify factors with high interrelationships and reduce the number of variables before constructing a ML model for improved efficiency. However, before conducting EFA, it is essential to test the relationships between variables by conducting hypothesis testing using Bartlett's test. Additionally, the suitability of the data should be assessed using the KMO measure. For the data that will undergo EFA, there are three datasets that consist of the dataset includes data obtained after the data preparation process, comprising a total of 35 factors (feature_all), the dataset consists of data obtained after analyzing Spearman correlation relationships, en-

compassing a total of 16 factors (feature_spearman) and the dataset includes data obtained after analyzing MI correlation relationships, involving a total of 19 factors (feature_mi). These datasets serve as the foundation for the EFA process, contributing to the identification of relevant factors and the subsequent development of an ML model with enhanced efficiency.

From Table 2, the results of the Bartlett's test for relationship testing are presented. The hypothesis was set, and the outcomes were measured using Chi-Square and p-value. It was found that the 'feature_all' dataset cannot be computed for these values. Therefore, this dataset will not be used for EFA. However, for the 'feature_spearman' and 'feature_mi' datasets, it was observed that the hypothesis of variables having a relationship is accepted. This is evident from the Chi-Square values of 1829.05 and 1651.46, respectively, and the p-value of 0.00 for both datasets. This indicates a significant relationship between variables in both datasets, making them suitable for further component analysis. Regarding the consideration of the KMO value, as shown in Table 2, the KMO value for the 'feature_all' dataset is 0.4998, which is less than 0.5. Hence, this dataset is not suitable for component analysis. However, for the 'feature_spearman' and 'feature_mi' datasets, the KMO values are greater than 0.5, indicating their suitability for component analysis. Therefore, the datasets to be analyzed will consist of 'feature_spearman' and 'feature_mi'.

Table 2 Bartlett's Test of Sphericity and KMO of 3 datasets

Dataset	Number of factors	KMO	Bartlett's Sphericity test	
			Chi-Square	P-value
feature_all	35	0.4998	nan	nan
feature_spearman	16	0.6408	1829.05	0.00
feature_mi	19	0.6399	1651.46	0.00

When the data is deemed suitable for component analysis, communalities are considered to assess the multiple correlation of variables with the components. It was found that in the 'feature_spearman' dataset, the variable with the highest communality is 'C1_2' with a value of 0.9950, while the lowest is 'B5' with a value of 0.2911. This indicates that the variables in the 'feature_spearman' dataset can be clearly classified into specific components. Similarly, for the 'feature_mi' dataset, the variable 'com_no' has the highest communality at 0.9950, and the lowest is 'B3' with a value of 0.3724. Therefore, it can be concluded that the variables can be distinctly grouped into specific components in both datasets.

Next is the extraction of factors from the selected datasets. After the factors have been extracted, the Total Variance Explained is calculated, explaining the percentage of overall similarity or relationship in the data that the extracted factors can account for. This step is crucial in deciding how many factors to consider for explaining the data.

For factor extraction of the 'feature_spearman' dataset, as shown in Table 3 under Initial Eigenvalues, it is found that only 7 factors should be retained, as the first 7 factors have eigenvalues greater than 1. The most significant factor is Factor 1, which can explain the highest variance at 21.43%. When the factor loadings are rotated using Varimax rotation, each component becomes distinct, with an effort to maximize or minimize the weights of variables within each factor. The Rotation Sums of Squared Loadings in Table 3 after Varimax rotation show that the eigenvalues and % of Variance for each factor remain the same as before rotation.

Similarly, for the 'feature_mi' dataset, as presented in Table 4, it is also suggested to retain only 7 factors, consistent with the previous dataset. The Initial Eigenvalues indicate that the first 7 factors have eigenvalues greater than 1, with the most influential factor being Factor 1, explaining the highest variance at 16.51%. After Varimax rotation, the Rotation Sums of Squared Loadings in Table 4 demonstrate that the eigenvalues and % of Variance for each factor remain unchanged, similar to the pre-rotation values.

Table 3 Total variance explained of the *"feature_spearman"* dataset.

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cum.%	Total	% of Variance	Cum.%	Total	% of variance	Cum.%
1	3.43	21.43	21.4	3.43	21.43	21.4	3.43	21.43	21.4
2	1.57	9.81	31.2	1.57	9.81	31.2	1.57	9.81	31.2
3	1.39	8.69	39.9	1.39	8.69	39.9	1.39	8.69	39.9
4	1.38	8.61	48.5	1.38	8.61	48.5	1.38	8.61	48.5
5	1.27	7.93	56.5	1.27	7.93	56.5	1.27	7.93	56.5
6	1.13	7.07	63.5	1.13	7.07	63.5	1.13	7.07	63.5
7	1.02	6.41	69.9	1.02	6.41	69.9	1.02	6.41	69.9
8	0.94	5.87	75.8						
9	0.88	5.52	81.3						
10	0.77	4.82	86.2						
11	0.70	4.38	90.5						
12	0.64	4.02	94.6						
13	0.41	2.58	97.1						
14	0.28	1.76	98.9						
15	0.10	0.65	99.5						
16	0.07	0.45	100.0						

Table 4 Total variance explained of the *"feature_mi"* dataset.

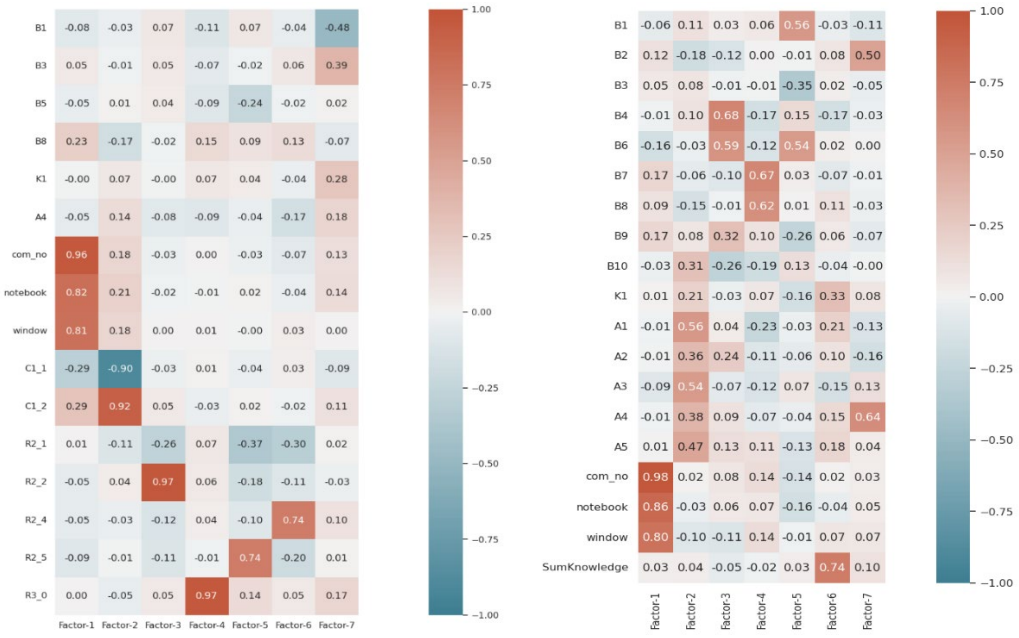
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cum.%	Total	% of Variance	Cum.%	Total	% of variance	Cum.%
1	3.14	16.51	16.5	3.14	16.51	16.5	3.14	16.51	16.5
2	2.23	11.73	28.2	2.23	11.73	28.2	2.23	11.73	28.2
3	1.76	9.26	37.5	1.76	9.26	37.5	1.76	9.26	37.5
4	1.37	7.22	44.7	1.37	7.22	44.7	1.37	7.22	44.7
5	1.36	7.14	51.9	1.36	7.14	51.9	1.36	7.14	51.9
6	1.26	6.63	58.5	1.26	6.63	58.5	1.26	6.63	58.5
7	1.05	5.52	64.0	1.05	5.52	64.0	1.05	5.52	64.0
8	0.88	4.61	68.6						
9	0.87	4.59	73.2						
10	0.82	4.32	77.5						
11	0.74	3.91	81.4						
12	0.71	3.71	85.1						
13	0.60	3.16	88.3						
14	0.56	2.94	91.2						
15	0.52	2.75	94.0						
16	0.42	2.23	96.2						
17	0.38	1.97	98.2						
18	0.25	1.30	99.5						
19	0.10	0.50	100.0						

Moving on, the next step involves organizing variables into each factor based on the factor loadings obtained after Varimax rotation, as depicted in Figure 5. In this assessment, if the factor loading of a variable within any particular factor is high (approaching +1 or -1), and the factor loadings of other factors are low (approaching 0), the variable is assigned to the factor with the highest factor loading. Variables with factor loadings less than 0.3 are not considered in any factor. Therefore, considering the factor loadings of variables in the 'feature_spearman' dataset, the variables are allocated to each factor as follows:

Factor 1: "com_no," "notebook," "window"
Factor 2: "C1_1" and "C1_2"
Factor 3: "R2_2"
Factor 4: "R3_0"
Factor 5: "R2_1" and "R2_5"
Factor 6: "R2_4"
Factor 7: "B1" and "B3"

Similarly, for the 'feature_mi' dataset, variables are assigned to factors based on the same criterion, resulting in the following grouping:

Factor 1: "com_no," "notebook," "window"
Factor 2: "A5," "A3," "A2," "A1," "B10"
Factor 3: "B4," "B6," "B9"
Factor 4: "B7" and "B8"
Factor 5: "B1" and "B3"
Factor 6: "SumKnowledge"
Factor 7: "B2" and "A4"



"feature_spearman" dataset "feature_mi" dataset

Figure 5 Factor loading heatmap

4.4 *Modeling*

The goal of this section is to identify factors influencing electricity usage in dormitory rooms. This involves comparing datasets obtained through feature selection using Spearman's and MI, along with the dataset after EFA. Another objective is to develop a model for electricity usage factors, reducing redundancy. This includes comparing baseline classification models (SVM, KNN, MLP, DT, LR) and selecting the most suitable algorithm. Hyperparameters will be adjusted to enhance the chosen baseline model, ultimately providing factors influencing electricity usage and a model for predicting costs in various price ranges.

For the process of building a model from training to testing, it starts with dividing the data into two parts: the training dataset (80%) and the testing dataset (20%). In the training dataset, the model is taught using 10-fold cross-validation to enhance its performance. This involves splitting the training dataset into 10 parts and testing the model 10 times. Each time, 9 sets are used for training, and the remaining set is used for testing. This ensures that every set of data is utilized in both training and testing processes. The testing dataset is then used to evaluate how well the model can predict unseen data. This dataset is not used in the model training process. To measure the performance of the classification model, the confusion matrix, a performance measurement table, is employed. This research will utilize values obtained from the confusion matrix to calculate single-value metrics that are commonly preferred over directly interpreting the confusion matrix, such as Accuracy, Precision, Recall, and F1-Score.

For the results of model creation using all five datasets and five algorithms, it was observed that the highest accuracy in predicting unseen data was 0.5762. This was achieved by the model generated from the 'feature_mi' dataset, utilizing the SVM algorithm, as shown in Table 6. This indicates that this model has better predictive capabilities compared to other models. Although examining the accuracy from the training scores presented in Table 5 shows that the highest accuracy is achieved by the model from the 'feature_all' dataset, using the KNN algorithm, with a score of 0.6474. However, the significance lies in testing the model with the testing dataset or evaluating the accuracy of the testing score. This is crucial as it reflects the model's performance in scenarios it has not encountered before and its ability to predict new data accurately. Therefore, the researcher selected the model created from the 'feature_mi' dataset using the SVM algorithm as the baseline classification model to identify factors influencing electricity usage.

Table 5 The accuracy of the training set for the baseline classification model.

Dataset	Number of factors	Training score				
		SVM	KNN	MLP	DT	LR
feature_all	35	0.6419	0.6474	0.6467	0.6407	0.6326
feature_spearman	16	0.6009	0.5989	0.5983	0.6064	0.5994
feature_mi	19	0.5929	0.5817	0.5843	0.5856	0.5707
feature_spearman_latent	7	0.5858	0.5613	0.5671	0.5617	0.5584
feature_mi_latent	7	0.5858	0.5748	0.5736	0.5710	0.5576

Table 6 The accuracy of the testing set for the baseline classification model.

Dataset	Number of factors	Testing Score				
		SVM	KNN	MLP	DT	LR
feature_all	35	0.5476	0.5167	0.5333	0.5327	0.5367
feature_spearman	16	0.4881	0.4655	0.4873	0.5161	0.5162
feature_mi	19	0.5762	0.5524	0.5413	0.5464	0.5443
feature_spearman_latent	7	0.3976	0.4060	0.4278	0.4131	0.4124
feature_mi_latent	7	0.5000	0.4929	0.5008	0.4607	0.4638

Selecting the best baseline classification model involves utilizing the dataset analyzed for its correlation with electricity consumption, applying MI, and employing the SVM algorithm. Researchers then refined the model's performance through parameter tuning. The results, displayed in Table 7 under the 'Optimal Model' section, show that the testing score accuracy of the optimal model is 0.7353, surpassing the baseline model's testing score accuracy. Therefore, adjusting the parameters significantly improves the performance of the model.

Table 7 The evaluation of the optimal classification model.

	Baseline model		Optimal model	
	Training score	Testing Score	Training score	Testing Score
Accuracy	0.5929	0.5762	0.7792	0.7353
Recall	0.5943	0.5611	0.7792	0.7353
Precision	0.5873	0.5561	0.8079	0.7531
F1-Score	0.5767	0.5220	0.7700	0.7335

4.5 *Conclusion*

For building the classification model using SVM, KNN, MLP, DT, and LR with the five datasets, the results from creating the baseline model revealed that the SVM baseline model generated from the “feature_mi” dataset exhibited the highest testing score accuracy at 0.5762, as shown in Table 6. Not only does the optimal baseline model provide accurate predictions, but it also identifies all 19 variables in the “feature_mi” dataset that significantly impact electricity consumption. These variables include “A5”, “A4”, “B1”, “B2”, “B9”, “A1”, “A2”, “B3”, “SumKnowledge”, “B7”, “window”, “B10”, “B6”, “A3”, “K1”, “com_no”, “B8”, “notebook”, and “B4”. Following this, the researchers enhanced the baseline model's performance by adjusting SVM parameters, as detailed in Table 7. The testing score accuracy increased from the baseline, indicating that the model has been refined and improved for real-world scenarios with complex and diverse datasets.

5 Conclusion and Future Work

5.1 Objectives revisited

1). To identify the factors affecting electricity consumption from electricity cost.

The primary objective of this research is to identify factors that impact electricity consumption levels based on the electricity costs. The study findings revealed that the number of computers used by student residents has a significant influence on electricity consumption since computers require electrical energy to operate. An increase in the number of computers in the dormitories leads to a substantial rise in energy consumption and heat generation from computers, which can elevate room temperatures. This, in turn, may result in increased energy usage by air conditioning systems as they work harder to maintain a comfortable environment. Additionally, the duration of computer usage each day, whether for studying, working, gaming, or other recreational activities, directly affects electricity consumption. The more extensive the computer usage, the higher the electricity usage. Factors related to air conditioning behavior, such as extended periods of air conditioning usage throughout the day or in conjunction with energy-releasing activities like ironing or using hairdryers, as well as the daily temperature settings for air conditioning, also contribute to increased electricity consumption. The more significant the adjustments made to lower the temperature, the more electricity the air conditioning unit requires to cool the room. Knowledge about electricity usage plays a crucial role in estimating electricity consumption since it enables residents to implement energy-saving practices. For instance, setting the air conditioner at 26-27 degrees Celsius while using a fan can help save electricity since fans distribute cool air throughout the room and consume less electricity compared to air conditioners. Lack of awareness about energy-saving practices, despite having knowledge about electrical appliances, may result in a failure to adopt energy-saving behaviors. Therefore, attitudes and awareness regarding electricity usage and energy conservation practices also play a significant role in electricity consumption levels.

2) To analyze the relationship among factors influencing electricity costs.

According to the EFA of the dataset that underwent analysis of the relationship with electricity consumption using MI, variables can be grouped into distinct factors due to their high correlations. The first clear group is associated with the number of notebooks using the Windows operating system, represented by the variables “com_no”, “notebook” and “window” Following that, there is a group of variables related to attitudes towards electricity consumption, including “A1”, “A2”, “A3”, “A4”, “A5” derived from questions about overall attitudes. Additionally, the variable “B9” related to using sleep mode during temporary computer usage, is part of this group. Sleep mode is designed to conserve energy, so regular use implies an awareness of energy-saving practices. The next crucial variable group concerns behaviors that release heat energy into the room while simultaneously using air conditioning for heat dissipation. This includes the variable “B4” representing the behavior of ironing clothes alongside air conditioning usage, and “B6” representing the behavior of blow-drying hair concurrently with air conditioning use. Subsequently, there is a group of variables related to the duration

of computer usage, whether for work or gaming. It is evident that variables "B7" and "B8" are interrelated. Lastly, there is a group of variables associated with temperature settings. Variables "B2" linked to the temperature settings of air conditioning, and "A4" related to attitudes towards setting the temperature of air conditioning to 27-28 degrees Celsius while using a fan to save electrical energy, demonstrate a connection.

3) To develop a model that relevant factors influencing electricity consumption and reduces the redundancy of these factors.

For the predictive modeling process of electricity cost, it begins with creating a baseline model using all five datasets and training with five machine learning algorithms: SVM, KNN, MLP, DT, and LR. The model generated from the dataset analyzed for its relationship with electricity cost using MI and trained with the SVM algorithm demonstrated superior performance compared to other models, achieving a testing score accuracy of 0.5762. Subsequently, the model underwent improvement through parameter tuning. By adjusting parameters, the accuracy of the testing score increased to 0.7353. The specific parameter values used for this enhancement were C equal to 1.5, gamma equal to 4.699, and employing the "rbf" kernel. This adjustment, known as hyperparameter tuning, resulted in the model exhibiting enhanced predictive capabilities for real-world scenarios.

5.2 Discussion

This research found that the number of floors in residential accommodations and gender has no impact on electricity usage. This finding may seem contradictory to the study by Yang, Y. (2021), which suggests that the number of residential floors and gender are factors affecting energy usage. This is because Yang, Y. (2021) studied seasonal factors throughout the year, revealing that the number of residential floors and gender significantly influence energy usage in both winter and summer seasons. However, according to Shen, M. (2017), it was found that energy usage behaviors related to air conditioning are significant factors influencing electricity usage. Additionally, the researchers found that temperature settings and duration of air conditioning usage have a greater impact on electricity usage. Furthermore, behaviors leading to heat generation, such as ironing clothes and using hair dryers, were found to affect electricity usage, which is consistent with Shen, M. (2017) study.

5.3 Limitation

The data collection process from the questionnaire continues to employ inadequate sampling methods, resulting in imbalanced data where the number of instances for each class is not approximately equal. Additionally, the questionnaire design is insufficient for capturing accurate information from residential units with two occupants. Another limitation of this research lies in the incompleteness of responses in the questionnaire, possibly stemming from misunderstandings or confusion regarding the wording of the questions.

5.4 Future of work

For future research regarding identifying factors influencing electricity usage in university dormitories, it is recommended to conduct surveys or interviews with residents before designing questionnaires to understand current behaviors or activities of residents that have changed. For example, the use of deodorizers, air purifiers to remove PM2.5, water filters, or even keeping pets in dormitories, which may require lights to be turned on or heat to be ventilated while residents are outside. This will lead to adding more questions about the number of electrical appliances used beyond this. There are also other interesting variables for studying factors affecting electricity usage, such as behaviors related to water heaters, elevator usage, and residents' income, among others. Finally, for model development, other baseline models besides SVM should undergo hyperparameter tuning, as other algorithms have numerous parameters to adjust.

References

1. World Population Review. (2022). Electricity Consumption by Country 2022. Retrieved May 16, 2022, from: http://worldpopulationreview.com/country-rankings/electricity-consumption-by-country
2. Department of Alternative Energy Development and Efficiency (DEDE). (2022). Energy Situation of Thailand. Retrieved May 16, 2022, from: https://www.dede.go.th/ewt_news.php?nid=47349
3. Energy Policy and Planning office (EPPO). (2022). Energy Statistics. Retrieved May 16, 2022, from: http://www.eppo.go.th/index.php/en/en-energystatistics/electricity-statistic
4. Srimode, P., Promwattanapakdee, T. (2010). Study for Specific Energy Consumption in Lecture Building.
5. CMU"s Sustainable Development Goals (SDGs). (2020). CMU Smart Campus. Retrieved May 16, 2022, from: https://sdgs.cmu.ac.th/en/ArticleDetail/b8517747-72ea-40ef-a023-59ba0c10778d
6. Hu, S., Yan, D., Guo, S., Cui, Y., & Dong, B. (2017). A survey on energy consumption and energy usage behavior of households and residential building in urban China. Energy and Buildings, 148, 366-378.
7. Singto, P., Nabnean, A. (2012). Electric power using behaviors of student residing in Silpakorn University Petchaburi IT campus dormitories.
8. Thaiyanan, N. (2019). Factor analysis and prediction of energy consumption in university buildings from different category.
9. Rinaldi, A., Schweiker, M., & Iannone, F. (2018). On uses of energy in buildings: Extracting influencing factors of occupant behaviour by means of a questionnaire survey. Energy and Buildings, 168, 298-308.
10. De Silva, M. N. K., & Sandanayake, Y. G. (2012). Building energy consumption factors: a literature review and future research agenda.
11. Khumsayyai, A. (2007). Electric power using behaviors of students residing in Chiang Mai University Dormitories.
12. Poolsawat, K., Tachajapong, W., Prasitwattanaseree, S., & Wongsapai, W. (2020). Electricity consumption characteristics in Thailand residential sector and its saving potential. Energy Reports, 6, 337-343.

13. De Silva, M. N. K., & Sandanayake, Y. G. (2012). Building energy consumption factors: a literature review and future research agenda.
14. Sirichotpundit, P., Poboon, C., Bhanthumnavin, D., & Phoochinda, W. (2013). Factors affecting energy consumption of households in Bangkok metropolitan area. Environment and Natural Resources Journal, 11(1), 31-40.
15. Hu, S., Yan, D., An, J., Guo, S., & Qian, M. (2019). Investigation and analysis of Chinese residential building occupancy with large-scale questionnaire surveys. Energy and Buildings, 193, 289-304.
16. Yang, Y., Yuan, J., Xiao, Z., Yi, H., Zhang, C., Gang, W., & Hu, H. (2021). Energy consumption characteristics and adaptive electricity pricing strategies for college dormitories based on historical monitored data. Energy and Buildings, 245, 111041.
17. Xie, X., Siau, K., & Nah, F. F. H. (2020). COVID-19 pandemic–online education in the new normal and the next normal. Journal of information technology case and application research, 22(3), 175-187.
18. Blognone. (2022). Canalys and IDC look at 2022 PC and peripheral markets still growing high for another year. Retrieved March 20, 2023, from: https://www.blognone.com/node/126685.
19. Statista. (2023). Market share held by the leading computer (desktop/tablet/console) operating systems worldwide from January 2012 to January 2023. Retrieved March 20, 2023, from: https://www.statista.com/statistics/268237/global-market-share-held-by-operating-systems-since-2009/
20. Statista. (2022). Online Food Delivery - Thailand. Retrieved March 20, 2023, from: https://www.statista.com/outlook/dmo/online-food-delivery/thailand#revenue
21. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
22. Shen, M., Sun, H., & Lu, Y. (2017). Household electricity consumption prediction under multiple behavioural intervention strategies using support vector regression. Energy Procedia, 142, 2734-2739.