Check for updates

Journal of Climate Change Research 2023, Vol. 14, No. 5, pp. 569~586
DOI: https://doi.org/10.15531/KSCCR.2023.14.5.569

Understanding the characteristics of residential electricity consumption in Korea: Applying machine learning techniques to household-level energy survey data

Moon, Jongwoo

Research Fellow, Climate and Air Quality Research Group, Korea Environment Institute, Sejong, Korea

ABSTRACT

Demand-side approaches become an important pillar for energy analysis, and their roles for achieving climate targets have been increasingly emphasized globally. Particularly, Korea is one of the countries experiencing a rapid transition of demographic and household structures, and accordingly, the current and future energy demand could be significantly affected. As per the importance of the understanding the energy demand characteristics, this study contributes to understanding the electricity consumption of households by analyzing how the various household characteristics can be used to understand the household's electricity consumption with household-level survey and machine learning techniques. This study utilizes the Household Energy Standing Survey published in 2022 and selects key household, housing, and appliance ownership and usage characteristics from the entire dataset. Afterward, the study applies Support Vector Machine, Random Forest, and Decision Tree classifiers to classify the household's monthly electricity consumption. The results suggest that the Random Forest classifier provides slightly better performances in general compared to the other models. Moreover, the feature importance suggests that the housing characteristics, such as the size of housing, and appliance usage information, and some household characteristics, such as the number of household members and household income, are relatively important features for classification. Although the study finds some evidence of the importance of household and behavioral information in understanding the household's electricity consumption, the study also identifies the limitation of the survey dataset in extracting the behavioral information.

Key words: Residential Electricity Consumption, Machine Learning, Household Characteristics, Electricity Demand, Household Energy Survey

1. Introduction

The understanding of the energy demand-side is increasingly important in achieving Net Zero, and especially, it would be important for Korea to achieve the NDC target and eventually reach Net Zero. It becomes more and more difficult to reduce greenhouse gas emissions in industry sectors, largely composed of energy- and carbon-intensive sectors, and the power sector with limited potential for renewable energies. To overcome the challenges and reach the national mitigation targets, inducing changes in demand-side to reduce future energy demands would be particularly important. There have been a large number of studies on the supply-side, such as renewable energy and technologies, but very limited numbers of studies have been conducted on demand-side of electricity. In this sense, forecasting electricity demand has been a very difficult challenge for Korea. Under 8th Basic Plan on Electricity Demand and Supply, the forecasted level of the target peak electricity demand was only 89.1GW in 2020, but the actual level exceeded 90 GW (Ministry of Trade, Industry and Energy, 2020), and

†**Corresponding author** : jwmoon@kei.re.kr, (30147, #1028, 370 Sicheong-daero, Sejong, Korea. Tel: +82-44-415-7603) **ORCID** Moon, Jongwoo 0000-0003-3147-3102

Received: July 19, 2023 / Revised: August 9, 2023 1st, September 8, 2023 2nd / Accepted: October 20, 2023

the electricity demand has been continuously under-forecasted.

Moreover, the demographic and household characteristics of Korea have been rapidly changing. According to Statistics Korea, the record low fertility rates in recent years and longer life expectancies lead to the accelerated transformation of the demographic and household characteristics of Korea. The share of the working-age population continues to decrease, but the share of people aged 65 or older gradually increases. This trend is expected to continue, and the share of people aged 65 or older would exceed 45% of the population in 2070 (Statistics Korea, 2022b). The aging population is a global trend, but the pace in Korea is exceptionally fast. Moreover, the composition of households continues to change, and the size of a household becomes smaller and smaller, and the share of single-person household continues to increase. The share of single-person households reached 31.2% in 2020, and it would reach 39.6% of the entire households (Statistics Korea, 2022a). These changes, in particular, of the shrinking size of a household and the increasing share of elderly people, are very likely to affect the behavior of people and households, and eventually, all of these would cause a significant impact on energy consumption and carbon emissions through various channels. Thus, understanding the impacts of household characteristics on electricity consumption and demand would become a key issue in establishing various policies.

This study applies machine learning techniques, particularly Support Vector Machine and Decision Tree, to analyze the residential electricity consumption of Korea by using household-level survey data. The paper reviews the existing studies analyzing the socio-economic drivers of electricity consumption and machine learning applications in energy studies. Chapter 3 covers the data and methodology used in this study, and Chapters 4 and 5 provide the results and the implications of the study.

There have been studies to understand the socio-economic drivers of electricity demand globally. Yu et al. (2018) used the framework of the Extended Snapshot tool, a bottom-up engineering model, and analyzed the impacts of changes in demographic structure on energy consumption and carbon emissions of households in China. This study considered various scenarios with different demographic components of the households in Sichuan Province. The study included the trend of shrinking household size and aging population in the region, and it suggested that these trends would increase the energy demand, such as cooling and heating and consumption. Loi and Ng (2018) analyzed the socio-economic drivers of residential electricity consumption in Singapore between 2005 and 2014. This study applied the one-way fixed effects and the fully modified Least Squares estimators, and it tried to understand how income and prices affect residential electricity consumption in Singapore. It also identified that household size is a significant component determining residential electricity consumption, but income and electricity prices could be less important. In Korea, Keum et al. (2018) used panel data of 16 municipalities over 1996-2013 and used a dynamic panel FD GMM (First-Differenced Generalized Method of Moments) to analyze the determinants of the electricity demand function. This study showed the impact of the positive income elasticity is much larger than that of the negative price elasticity, and the aging population could reduce residential electricity consumption. Also, CDD (cooling degree days) would positively affect residential electricity consumption in Korea. Shin (2018) applied the fixed effects threshold panel regression to analyze the relationship between residential electricity consumption and the aging population by using the panel data of 16 municipalities over the period 2003-2015. This study found the effect of the price elasticity (negative) is larger than that of the income elasticity (positive), and the aging population trend would lead to smaller income electricity but larger price elasticity. Noh and Lee (2013) suggested the aging population could increase electricity consumption by applying a panel model using data of 15 municipalities over the period 2001-2010. Most of the studies applied econometric methodologies, particularly panel models, to analyze the relationship between the household or demographic characteristics and the

residential electricity consumption.

The application of machine learning techniques in energy research is becoming popular, and already various areas, such as energy price or supply/demand forecasting, are adopting these techniques. Compared to the traditional econometric methods, these machine learning techniques require less strict assumptions, such as the parameter distributions, and allows the use of various types and size of data. Fan et al. (2020) used a hybrid model combining machine learning models, such as Support Vector Regression and Particle Swarm Optimization, and econometric methods, such as AR-GARCH, to forecast the electricity consumption of New South Wales in Australia. By using both types of models, the study proposed a forecasting model for electricity consumption with enhanced efficiency and efficacy. Dahl et al. (2018) developed a day-ahead heat load forecasting model by using various data, including weather and calendar data in Denmark. These weather and calendar data are included to understand consumer behaviors, and the Multilayer Perceptron and Support Vector Regression models are applied to the analysis. It showed the use of calendar and weather data can improve the power of forecasting significantly. Amasyali and El-Gohary (2021) adopted four machine-learning algorithms, such as CART, EBT, ANN, and DNN, and predicted the building energy consumption with consideration of occupant-behaviors. This analysis simulated over 5,000 model setting using EnergyPlus, set machine learning algorithms and predicted hourly cooling energy consumption by using the simulation results. This study found a significant impact of occupant behavior in energy cooling demand and showed that the neural network models, such as ANN and DNN, provide better prediction but with higher computational costs, compared to CART and EBT. There has been an increasing number of studies utilizing machine learning techniques in energy demand analysis, but the lacking interpretability has been a critical constraint of using machine learning techniques, compared to the econometric models; thus, much of research is focusing on forecasting or predictions of variables. Burnett and Kiesling (2022) applied machine learning algorithms to analyze the household's energy demand by utilizing the residential energy consumption survey in the US. This study collected the residential energy consumption survey information over 2001-2015 and applied various models, including random forest, k-nearest neighbors, penalized regression, and gradient boosting method. The study evaluated the out-of-sample predictions among the trained models and found the random forest method indicates the best relative prediction of the demand. While the regression models can provide the estimates of parameters, random forest model suggests the relative importance of features, and the study found the amount of space as the most important feature. The next important features were natural gas prices, number of bedrooms, type of housing, and electricity prices.

The existing literature mainly focuses on the prediction of the consumption of energy sources, such as electricity and natural gas, based on real-time data, as well as the development of models for better forecasting electricity consumption. However, there are a limited number of studies available utilizing the national-level household survey to understand the household's energy consumption. Moreover, the existing studies in Korea, such as (Lee et al., 2019, 2022), try to understand the household's characteristics by using statistical information of the survey or use the survey information to build characteristics of a representative household. This study can contribute to applying machine learning techniques, which have more flexibility in handling a large number of variables with different types, and to understanding the key features for determining the household's electricity consumption.

2. Material and Methods

2.1. Household Energy Standing Survey

Household Energy Standing Survey (hereafter "HESS") is prepared by Korea Energy Economics Institute (KEEI, 2022), and this yearly survey collects various information, including household characteristics, appliance ownership, and energy consumption, including electricity. The

previous surveys included 2,520 households, but HESS 2019 expanded the survey subjects to 8,010 households in 17 municipalities. The final survey result provided information of 6,597 surveyed households. This survey provides a wide range on information of surveyed households, including some information of usage patterns of the home appliances, household characteristics, such as income, number of household members, and age of household head, and housing features. The regional composition of the survey is quite similar to the actual regional composition of households in 2019, but there are limited numbers of single-person households in the survey. In 2020, the single-person household accounted for approximately 31% of the entire household, but only about 17% of the surveyed households were single-person households. The distribution of the monthly income of surveyed households is largely located between 200 million KRW and 600 million KRW. The shares of surveyed households with monthly income between 200 million KRW and 400 million KRW and between 400 million KRW and 600 million KRW are 33.4% and 30.7%, respectively. The survey provides information on appliance ownership and usage and consumption of various energy types, but this study focuses on the electricity consumption of households.

Table 1. Descriptive statistics (example)

Variable Name	Description	Average	Standard Deviation	Min	Max
elec_all	Electricity consumption (annual)	2.322	0.010	1	4
g_r_s10_101	housing type	2.204	0.011	1	3
m2_r_s10_106_10	Size of housing (m2)	72.457	0.295	13	244
s10_107	number of rooms	2.704	0.008	1	6
g_r6_s10_201_300	Main heating fuels	4.009	0.012	1	7
r_s10_201_4000	Supplement heating devices	2.917	0.017	1	4
s10_203_10	in-door temperature during Winter (at home)	24.135	0.034	13	40
s10_203_20	in-door temperature during Winter (when leave)	20.372	0.044	10	40
s10_801	number of household members	2.680	0.015	1	9
single	Single household dummy	0.169	0.005	0	1
old	Aged household dummy	0.363	0.006	0	1
s10_806_adj	annual income (pre-tax)_adjusted	1760.497	17.929	75	40000
g_s10_204_20	cooling temperature (air conditioner)	1.388	0.014	0	5
tv_all	monthly TV electricity consumption	5.795	0.083	0	83.16
wash_all	monthly washing machine electricity consumption	2.660	0.049	0	72.612
air_all	summer-time air conditioner electricity consumption	174.283	3.176	0	5304
fan_all	summer-time fan electricity consumption	13.926	0.186	0	202.5
ref_all	monthly refrigerator electricity consumption	19.284	0.172	0	152
dish_all	monthly dish washer electricity consumption	0.114	0.019	0	52.75368
comp_all	monthly computer electricity consumption	0.969	0.035	0	73.0716
cook_all	monthly cooker electricity consumption	28.395	0.456	0	442.08
clean_all	monthly cleaning device electricity consumption	4.543	0.212	0	688.632
airpuri_all	monthly air purifier electricity consumption	0.936	0.037	0	56
home_set	number of digital settop, blue-ray and other video/audio devices	0.235	0.004	0	4
coffee_water	number of coffee and water purifiers	0.173	0.003	0	2
elec_wave	number of electronic waves and ovens	0.367	0.004	0	10

(Source: KEEI, 2022)

2.2. Methodology

In this study, Support Vector Machine, Random Forest, and Decision Tree classifiers are used to classify the household's electricity consumption in Korea. Support Vector Machine is a popular machine learning method in classification-type research questions. This is a supervised machine learning method, and it classifies the dataset by finding the optimal hyperplane separating data points of different classes. The hyperplane separates the data points of different classes, and the hyperplane with maximum margin, which is the distance between the hyperplane and the closest data point, is considered as the optimal one. The following equation is the objective of the Support Vector Machine classifier that maximizes the margin by minimizing $\|w\|^2$ with a penalty of misclassified data points (Winters-Hilt and Merat, 2007)

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \; subject\ to\ \xi_i \geq 0, y_i(w^T\phi(x_i)+b) \geq 1-\xi_i$$

The study uses 'Scikit-learn' library to conduct the Support Vector Machine, and the kernel function is allowed to use. The kernel function transforms data to a higher-dimensional space and supports the classification of the dataset, which is difficult to be linearly separated. The study uses 'polynomial', 'radial basis function', 'sigmoid', and 'linear' kernels provided by the library.

This method considers the distance between the hysperplane and data points, so it is necessary to standardize the dataset (Hsu et al., 2016; Menon, 2009; Murty and Raghava, 2016; Park, 2019; Winters-Hilt and Merat, 2007). Particularly, this dataset contains various categorical and numerical variables, so this study applies 'StandardScaler'. This standardization approach uses a z-score for standardization, and 'OneHotEncoding' from the scikit-learn library is used. A feature of this is to create a binary column for each category and assigns a value of one to the feature for categorical variables.

Next, the Random Forest classifier is an ensemble method establishing multiple decision trees with a subset of the dataset. Each tree makes its own prediction, and it aggregates the predictions, and selects the class with the highest votes. It splits the trees based on the Gini impurity, which is a probability of incorrect classification of randomly chosen data. The strengths of this method are

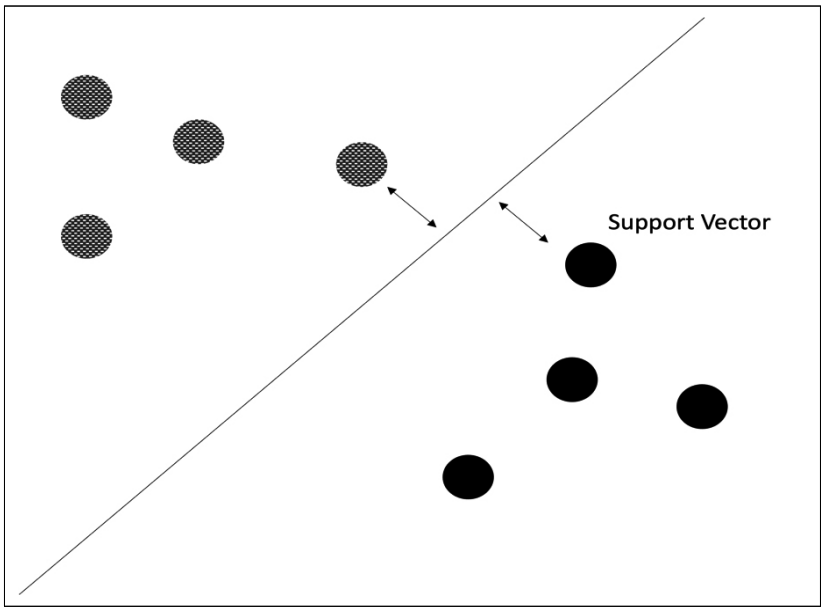

Fig. 1. Graphical illustration of support vector machine
(Source: Winters-Hilt and Merat, 2007)

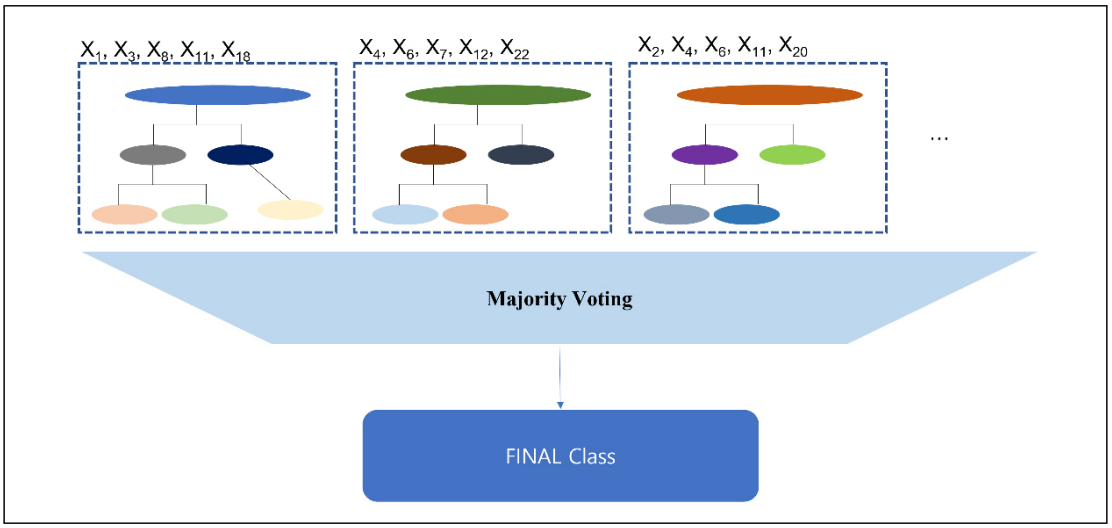

Fig. 2. Graphical illustration of random forest (source: Moon, 2022)

limiting overfitting issues, handling large datasets with various features, and information of importance for feature selections, which help the understanding of the classification process.

The last method used in this study is decision tree classification. It is a supervised machine learning method, and it splits datasets based on the decision rule. The highest node is called the root node, where the split starts. It moves to the decision nodes with branches, and the leaf nodes are the final output of decisions (Géron, 2019; Park, 2019).

These methodologies are typically used for classification problems. While Support Vector Machine is considered as black box, Random Forest and Decision Tree suggest feature importance, which indicates the relative importance of a feature to other features. The comparison of three classification methodologies indicates their classification abilities based on various types of

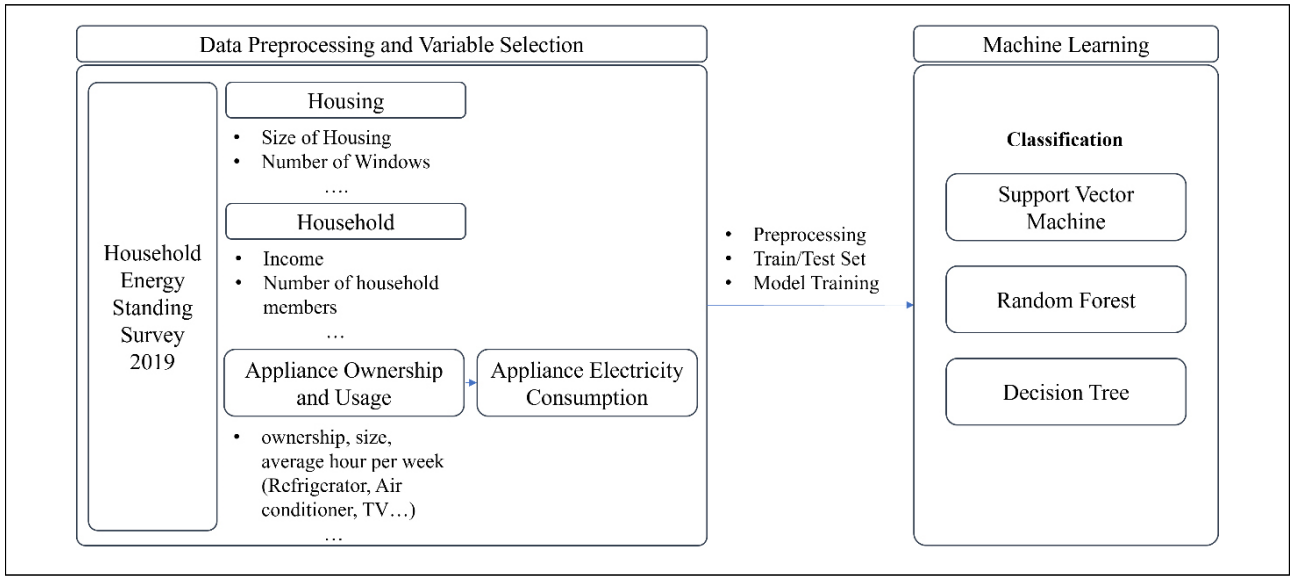

Fig. 3. Preprocessing and application of machine learning methodologies

variables, and Random Forest and Decision Tree classification methodologies provide the relative importance of features in classification and suggest some interpretability of the classification results.

The dataset includes various categorical and numerical variables. While it provides numerical household electricity consumption (kwh), the information on many variables is provided as categorical or dummy variables. Instead of directly using the information, the study set a classification research question, which focuses on how each variable affects the classification of the dependent variable. Thus, the study converted the dependent variable, the electricity consumption of households, from a numerical to a categorical variable. The study excluded some surveyed samples with missing key variables, such as floor level, and the final dataset for the analysis was 6,557 surveyed households.

Moreover, this study utilizes the household survey, and the survey contains various household and housing characteristics, which are fixed during the year. It is difficult to reflect the impacts of electricity prices and the seasonal differences. To indirectly consider those factors, the study considers the bracket of the progressive pricing structure in Korea in determining the classes of the dependent variable. Moreover, the residential electricity demand reaches a peak during summer (KEPCO, 2019), and the Korea Electric Power Corporation sets different ranges for the progressive pricing structure during summer, so the study considers summer separately with a different set of independent variables.

This study constructs a classification problem to understand the residential electricity consumption. A large portion of variables in survey information and used in this study is categorical variable, so there are limitations to use those variables to train the models and predict or forecast the amount of household electricity consumption. Instead, the study establishes a classification problem of a household's electricity consumption with consideration of the progressive electricity pricing, which affect the electricity consuming behaviors of each household. The study converted the annual electricity consumption of each household into monthly average electricity consumption. Considering a number of categorical variables in the dataset, the study divided the household's monthly average electricity consumption into four groups. To divide the groups, the study considered the progressive electricity pricing in Korea. However, a large number of households consume the level 2 pricing between 200 kWh and 400 kWh per month, so the study sets the range of electricity consumption as 100 kWh, instead of 200 kWh suggested by the progressive electricity pricing. Thus, the study sets four groups of monthly electricity consumption as following: (Group 1) below 200 kWh per month; (Group 2) 200 kWh ~ 300 kWh per month; (Group 3) 300 kWh ~ 400 kWh per month; and (Group 4) above 400 kWh per month. In addition, this study conducted a similar process to the household's monthly electricity consumption during summer time (June − August). During summer, the electricity demand for cooling increases significantly, and KEPCO tends to ease the progressive billing during summer to reduce the economic burden of households. Thus, the study set four groups of monthly electricity consumption during summer time as follows: (Group 1) below 250 kWh per month; (Group 2) 250 kWh ~ 350 kWh per month; (Group 3) 350 kWh ~ 450 kWh per month; and (Group 4) above 450 kWh per month.

Also, the study selected key independent variables from various household and housing characteristics. The following table indicates the variables selected as the independent variables in this study. Based on the ownership and usage information of key appliances provided by the survey, the study calculates the monthly average consumption of electricity by the key appliances. The survey information provides information, such as type and size, with different units (e.g., Wh per use, Wh per kg, and W for washing machine), so the study calculates the average electricity consumption of each appliance by using appliance's technical specification and usage information for securing the comparability. If there is only ownership information without usage, the study considers the number of appliances owned by the household. Also, the study created dummy variables of "single" and "old" to understand whether single-person

households or households with members aged 65 or older affect the classification.

The different sets of variables are considered in understanding the household's electricity consumption in two different time setting in this study. There are distinctive differences of using appliances in different seasons, including the use of heating and cooling devices. During the summer, the use of heating appliances would be quite limited, and the city gas and oil are mainly used for heating, so the study excludes heating-related variables in summer. The study sets two different specifications: one considers all characteristics, including ownership and usage of appliances, and the other specification considers variables closely related to the summer season and excluded characteristics, such as heating-related information.

Table 2. Selected variables for classification

variable	explanation	All	Summer
single	single = 1; other = 0	Y	Y
old	old = 1; other = 0	Y	Y
s10_city	Province Code	Y	Y
g_r_s10_101	housing type (1= detached; 2= multi-family; 3= apartment)	Y	Y
s10_102_21	floor level	Y	Y
s10_104	Housing direction (South/South East/South West = 1; other = 0)	Y	Y
g_m_r_s10_105	Construction Year (1= before 1970, 2= 1970s; 3= 1980s; 4= 1990s; 5= 2000s: 6= after)	Y	Y
m2_r_s10_106_10	Size of housing (m2)	Y	Y
s10_107	number of rooms	Y	Y
s10_110	number of windows	Y	Y
s10_201_2000	Installed heating devices (1= nothing; 2= regional/district; 3= gas/electricity; 4= renewables; 5= others)	Y	Y
g_r6_s10_201_300	Main heating fuels (1= briquet, kerosene; 2= city gas; 3= district heating; 4= electricity, 5= others)	Y	Y
r_s10_201_4000	Supplement heating devices (1= nothing; 2= regional/district; 3= kerosene, city gas,etc. 4= electricity; 5= renewable; 6= others)	Y	Y
r_2_s10_gassup_04	no supply of city gas = 1; other = 0	Y	Y
s10_203_10	in-door temperature during Winter (at home)	Y	
s10_203_20	in-door temperature during Winter (when leave)	Y	
s10_204_000	number of cooling devices	Y	Y
g_s10_204_20	cooling temperature (air conditioner): (0: no; 1: below 20; 2: 20-22; 3: 22-24; 4: 24-26; 5: 26 or above)	Y	Y
s10_205_100	number of cooking devices	Y	Y
r4_s10_205_200	main cooking fuel (1: electricity; city gas; 3: other; 4: none)	Y	Y
r_s10_206_100	installation of renewable (no = 1; yes = 2)	Y	Y
s10_207_70	decide to install renewables by themselves (yes = 1; no = 2) [awareness]	Y	Y
r3_s10_207_40	Size of solar (W)	Y	Y
r_s10_208_41	Size of solar heat (W)	Y	Y
s10_208_70	use of solar heat	Y	Y

Table 2. Selected variables for classification (continued)

variable	explanation	All	Summer
r_s10_209_40	size of geothermal (kw)	Y	Y
s10_209_60	use of geothermal	Y	Y
s10_601	Awareness of appliance energy efficiency labelling	Y	Y
s10_602	Whether checking previous year or month's electricity bill	Y	Y
s10_603	Satisfactory level of cooling	Y	Y
s10_604	Satisfactory level of heating	Y	
s10_607	Registration of energy saving program	Y	Y
s10_803_1	household head's sex: Male = 1; female = 0	Y	Y
g_m2_r_s10_803_3	age of household head: 1: 29 or below; 2 = 30s; 3 = 40s; 4 = 50s; 5= 60s or above)	Y	Y
s10_801	number of household members	Y	Y
s10_803_2	household head's education; 1 = middle-school or less; 2 = high-school graduate; 3 = university graduate; 4 = graduate school	Y	Y
s10_803_4	occupations: 1= full-time; 2 = temporary; 3 = self-employment; 4= others	Y	Y
s10_806_adj	annual income (pre-tax)_adjusted (no missing)	Y	Y
tv_all	monthly TV electricity consumption	Y	Y
wash_all	monthly appliance electricity consumption	Y	Y
air_all	summer-time air conditioner electricity consumption	Y	Y
fan_all	summer-time fan electricity consumption	Y	Y
ref_all	monthly appliance electricity consumption	Y	Y
dish_all	monthly appliance electricity consumption	Y	Y
cook_all	monthly appliance electricity consumption	Y	Y
clean_all	monthly appliance electricity consumption	Y	Y
airpuri_all	monthly appliance electricity consumption	Y	Y
home_set	number of digital settop, blue-ray and other video/audio devices	Y	Y
coffee_water	number of coffee and water purifiers	Y	Y
elec_wave	number of electronic waves and ovens	Y	Y
elec_cook	number of electronic cooking devices	Y	Y
airfri	number of air friers	Y	Y
elec_pot	number of electronic pots	Y	Y
foodwaste	number of food waste disposal machine	Y	Y
humid	number of humidifier and dehumidifier	Y	Y
styler	number of cloth styler	Y	Y
printer	number of printer and fax	Y	Y
app11_31	number of electric pad	Y	
app11_32	number of electric hotwater pad	Y	
app11_33	number of electric stove	Y	
app11_34	number of electric heater	Y	

3. Results

3.1. Results: Annual

The study applies three machine learning methodologies, which are Support Vector Machine, Random Forest, and Decision Tree, and conducts the classification of the household's electricity consumption. Fig. 4 suggests the classification result of the Support Vector Machine. RandomizedSearchCV is used to find the optimal hyperparameters of the Support Vector Machine. Instead of checking all combinations of parameter values, this method randomly selects the parameter values from the specified distributions. While the performance of this method could be less well, it could reduce the computation burden and require a shorter time. The types and range of the hyperparameters in the search are suggested in the following table.

The selected hyperparameters from RandomizedSearchCV are {'C': 100, 'gamma': 0.001, 'kernel': 'rbf (radial basis function''}, and the confusion matrix is as follows. As shown in the performance metrics, the weighted averaged f1-score is 0.54. By each class, the model predicts relatively well (65%) in class 2, and the classification performance of class 3 and class 4 are somewhat similar, but it provides relatively bad performance in classifying class 1 (34%). The result shows that the model predicts a large number of test samples as class 2 (200 kwh ~ 300 kwh per month), where a large number of observations are located. The classification result of the Support Vector Machine indicates the surveyed households' characteristics and the ownership and usage of appliances are not very different to precisely classify the household's monthly electricity consumption, particularly of class 1 with monthly consumption below 200 kwh per month.

Next, the study applies the Random Forest classifier to classify the same dataset and uses RandomizedSearchCV to find the optimal hyperparameters. The selected hyperparameters are {'n_estimators': 474, 'min_samples_

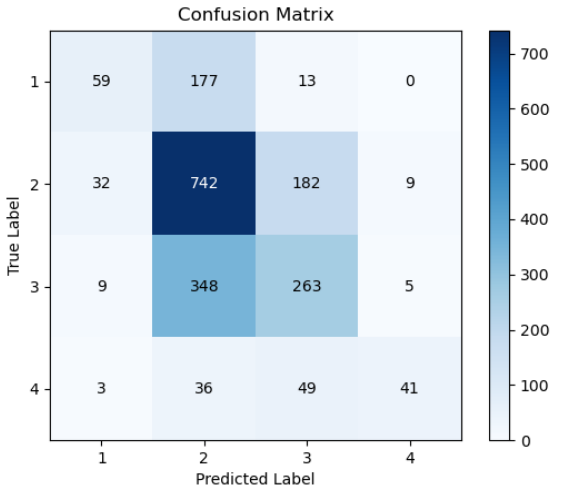

Fig. 4. Confusion matrix (support vector machine, annual)

Table 3. Hyperparameters used for RandomizedSearchCV (support vector machine)

Type	Values					
Kernel	Polynomial	RBF	Sigmoid	Linear		
C	0.001	0.01	0.1	1	10	100
Gamma	0.001	0.01	0.1	1	10	100

Table 4. Performance metrics (support vector machine, annual)

Type	Precision	Recall	F1-Score	Support
1	0.57	0.24	0.34	249
2	0.57	0.77	0.65	965
3	0.52	0.42	0.46	625
4	0.75	0.32	0.45	129
Accuracy			0.56	1,968
Macro Average	0.60	0.44	0.47	1,968
Weighted Average	0.57	0.56	**0.54**	1,968

split': 2, 'min_samples_leaf':5, 'max_features': 'sqrt', 'max_depth': 44, 'bootstrap': 'false', 'random_state':0}, and the confusion matrix and the performance metrics are as follows. The result is quite similar to the case of Support Vector Machine classification. A distinctive feature of Random Forest classification is the feature importance, which computes the level of the contribution of each feature in the model for decreasing the impurity. In this model, the highest feature is 'size of housing', and the usage information of appliances, such as tv, air conditioner, refrigerator, and cooker, are considered as the next important features. Some housing characteristics, such as main heating fuels, city, and cooling temperature, as well as household characteristics, such as household income and number of household members also showed relatively high feature importance.

Lastly, the decision tree classifier is applied, and the range of hyperparameters are {'max_depth': list(range(3,10)), 'min_samples_split': [2,3,4]}. The selected hyperparameter is {'max_depth':5}, and the classification result and the feature importance are shown as follows. The performance metrics are relatively worse than the previous two models. Interestingly, this classification suggests 'single' as the highest feature importance, and 'size of housing' and 'main heating fuels' as the next important features. This result seems reasonable as the dependent variable is the classification result of a household's monthly electricity consumption, and the single-member households would be likely to be classified as lower classes with less electricity consumption.

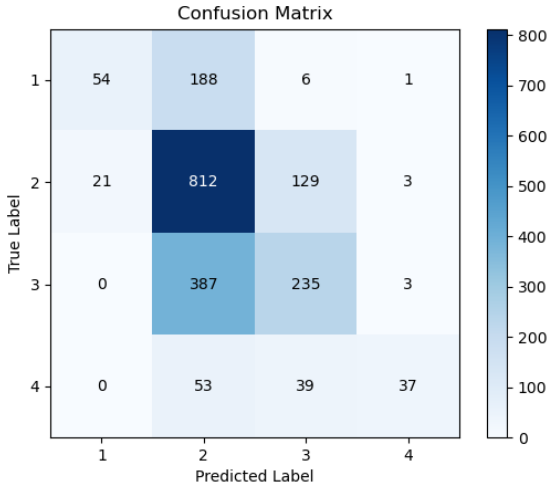

Fig. 5. Confusion matrix (random forest, annual)

Table 5. Hyperparameters used for RandomizedSearchCV (random forest)

Type	Values
n_estimators	Int(x) for x in np.linspace(start = 10, stop = 500, num = 20)
max_features	'None', 'sqrt', 'log2'
max_depth	Int(x) for x in np.linspace (1, 50, num = 50)
min_samples_split	Int(x) for x in np.linspace(start=2, stop = 10, num = 5)
min_samples_leaf	Int(x) for x in np.linspace(start=1, stop = 10, num = 5)
bootstrap	True, False
random_state	0

Table 6. Performance metrics (random forest, annual)

Type	Precision	Recall	F1-Score	Support
1	0.72	0.22	0.33	249
2	0.56	0.84	0.68	965
3	0.57	0.38	0.45	625
4	0.84	0.29	0.43	129
Accuracy			0.58	1968
Macro Average	0.67	0.43	0.47	1968
Weighted Average	0.61	0.58	**0.55**	1968

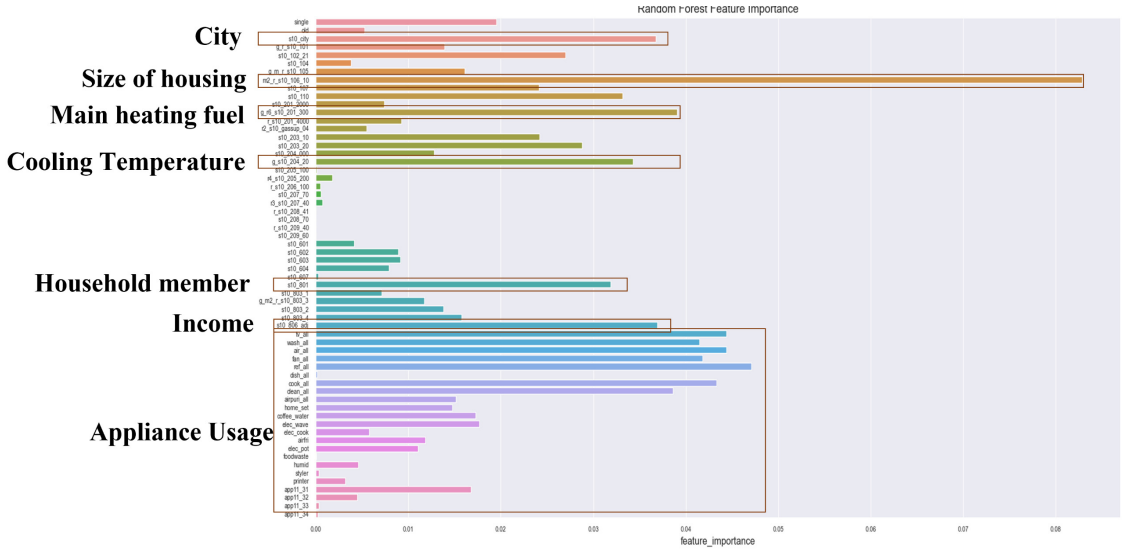

Fig. 6. Feature importance (random forest, annual)

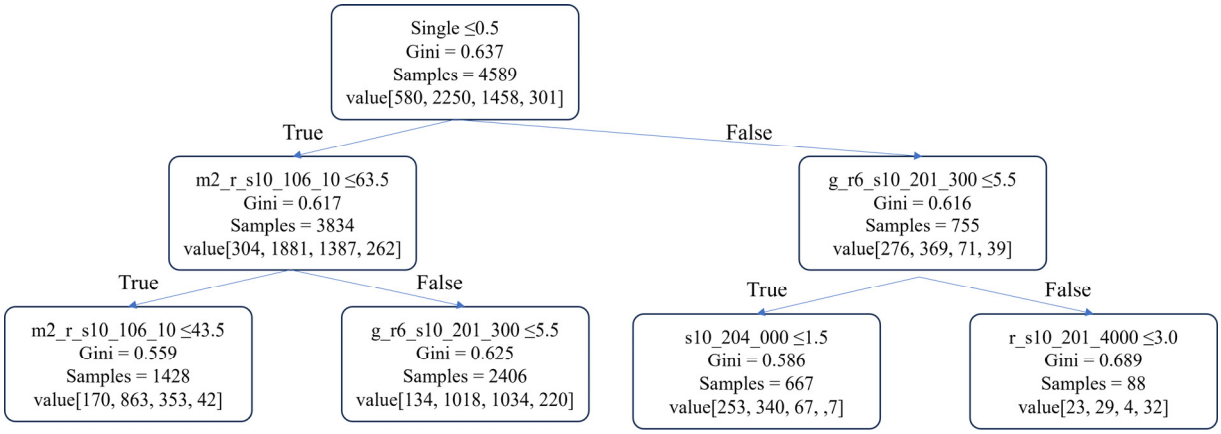

Fig. 7. A part of decision tree (annual)

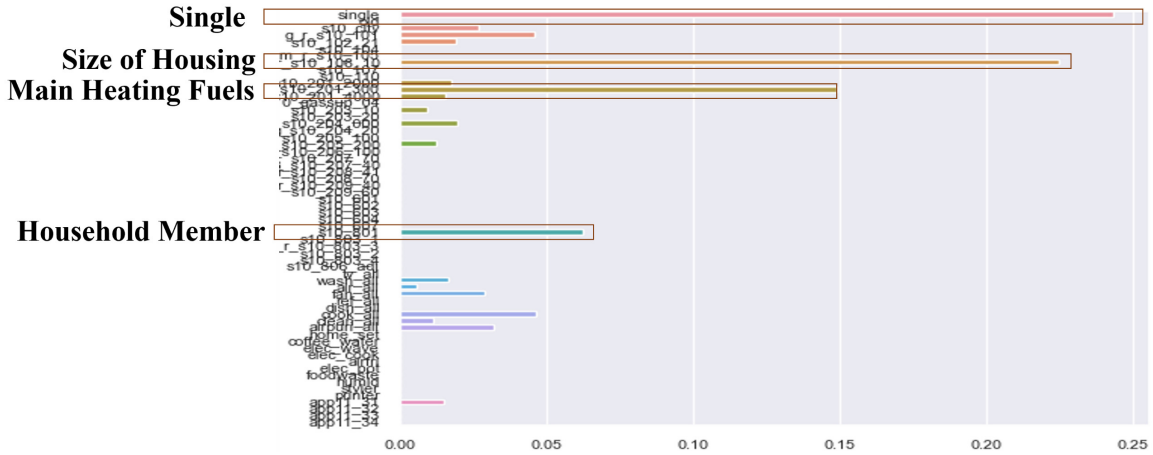

Fig. 8. Feature importance (decision tree)

Table 7. Performance metrics (decision tree)

Type	Precision	Recall	F1-Score	Support
1	0.47	0.18	0.26	249
2	0.56	0.67	0.61	965
3	0.47	0.50	0.48	625
4	0.74	0.31	0.44	129
Accuracy			0.53	1,968
Macro Average	0.56	0.41	0.45	1,968
Weighted Average	0.53	0.53	**0.51**	1,968

Table 8. Performance metrics (support vector machine, summer)

Type	Precision	Recall	F1-Score	Support
1	0.52	0.33	0.41	417
2	0.47	0.71	0.57	841
3	0.44	0.32	0.37	533
4	0.67	0.21	0.32	177
Accuracy			0.48	1,968
Macro Average	0.52	0.39	0.42	1,968
Weighted Average	0.49	0.48	**0.46**	1,968

The results indicate that the random forest model shows slightly better performance compared to the other two methodologies. The feature importance of the random forest and decision tree models identify the housing characteristics, such as size of housing, the electricity consumption of appliances, and the number of household members as the important features in common. However, those models show different results in household's characteristics. While random forest model finds 'single' as a relatively less important feature and 'income' as a relatively important feature, decision tree model identifies 'single' as the most important feature and does not find 'income' as an important feature.

3.2. Results: Summer

In addition to the monthly average electricity consumption, the study classifies the household's monthly electricity consumption in summer (June − August). For Support Vector Machine, the same methods and the range of hyperparameters are considered. The selected hyperparameters are {'C': 10, 'gamma': 0.1, 'kernel': 'rbf (Radial Basis Function)'}. The confusion matrix and the performance metrics are as follows. Similar to the previous case, a large number of test samples are classified as class 2, but the f1-score of classifying class 1 becomes slightly higher than the previous case. Partly, this can be explained that the change in the range of

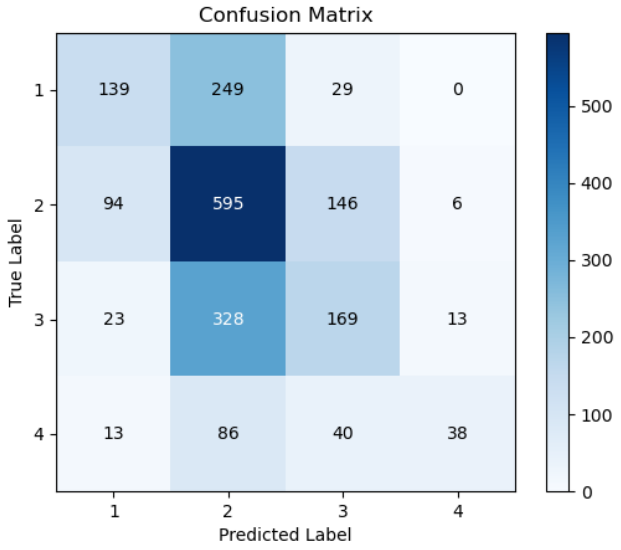

Fig. 9. Confusion matrix (support vector machine, summer)

monthly electricity consumption leads to more samples being classified as class 1 and enhance the classification performance. However, the general weighted average f1-score is 0.43, which is relatively lower than that of monthly electricity consumption for the entire year.

Next, The Random Forest classifier is applied to classify the household's monthly electricity consumption during summer time. The selected hyperparameters are {'random_state': 0, 'n_estimators': 87, 'min_samples_split': 4, 'min_samples_leaf': 3, 'max_features': 'log2', 'max_depth': 50, 'bootstrap': False}. The confusion matrix and the performance metrics are as follows. The performance metric of the weighted average of the F1-score is slightly better than those of Support Vector Machine, but this result indicates relatively low

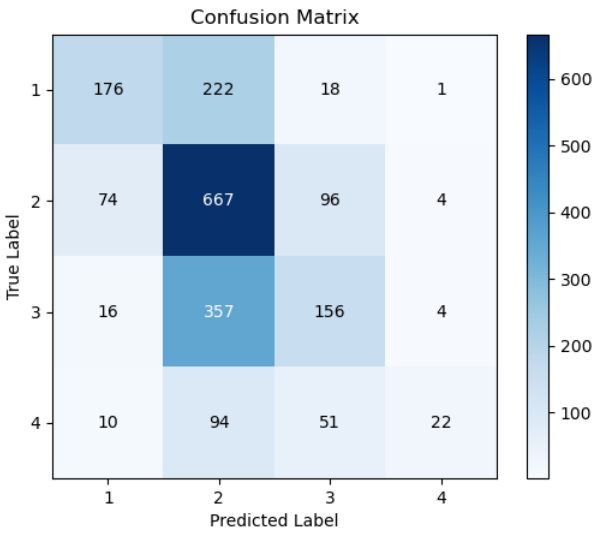

Fig. 10. Confusion matrix (random forest, summer)

Table 9. Performance metrics (random forest, summer)

Type	Precision	Recall	F1-Score	Support
1	0.64	0.42	0.51	417
2	0.50	0.79	0.61	841
3	0.49	0.29	0.37	533
4	0.71	0.12	0.21	177
Accuracy			0.52	1,968
Macro Average	0.58	0.41	0.42	1,968
Weighted Average	0.54	0.52	**0.49**	1,968

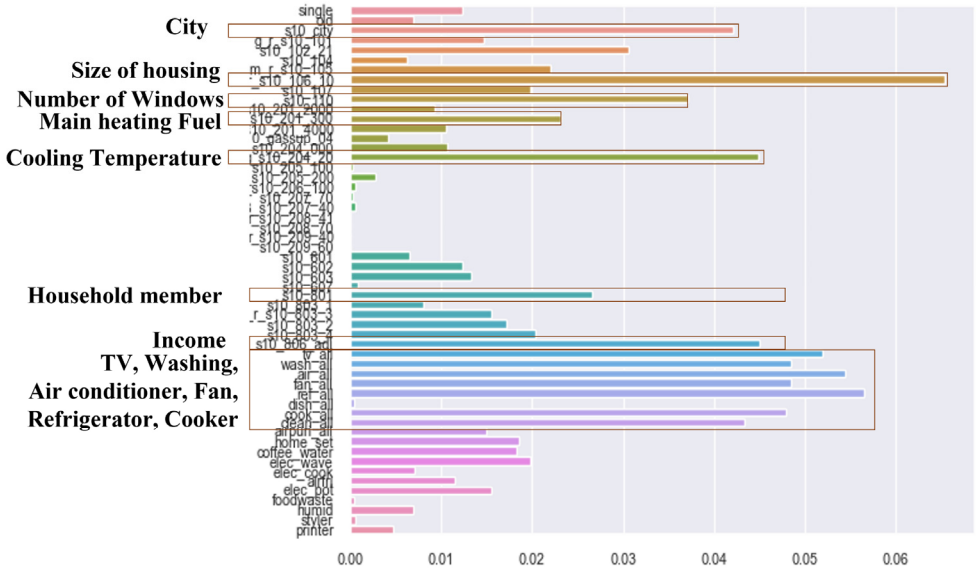

Fig. 11. Feature importance (random forest, summer)

performances in classifying large electricity-consuming households with relatively small samples. Similar to the annual case, the feature importance shows 'size of housing' as the highest feature, and the usage information of appliances are considered as the next important features. Also, the cooling temperature, number of windows, and province, also show some importance in classification.

For the decision tree classifier, the selected hyperparameter is {'max_depth':7, 'min_samples_split':4}, and the classification result and the feature importance are shown as follows. The performance metrics indicate similar results to the previous two models. Similar to the annual case, it finds 'single' feature as the highest importance and recognizes 'size of housing' as the next important feature. Also, 'cooling temperature' and 'number of household members' also show some importance in classification.

Similar to the previous case, random forest model indicates slightly better performance compared to the other models. While the feature importance of random forest model shows a similar result to the previous case, it finds 'number of windows' as relatively more important but 'main heating fuel' as relatively less important. Also, decision tree classifier finds 'cooling temperature', 'the electricity consumption of air conditioner', and 'income'

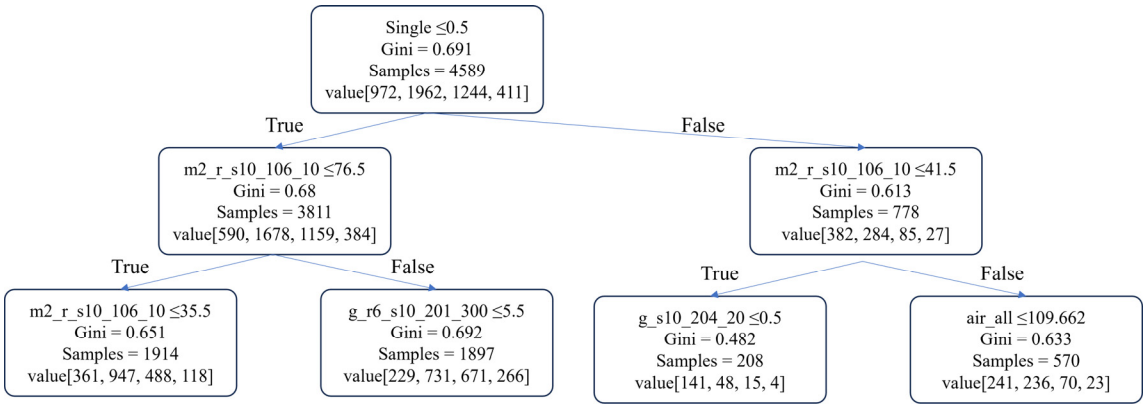

Fig. 12. A part of decision tree (summer)

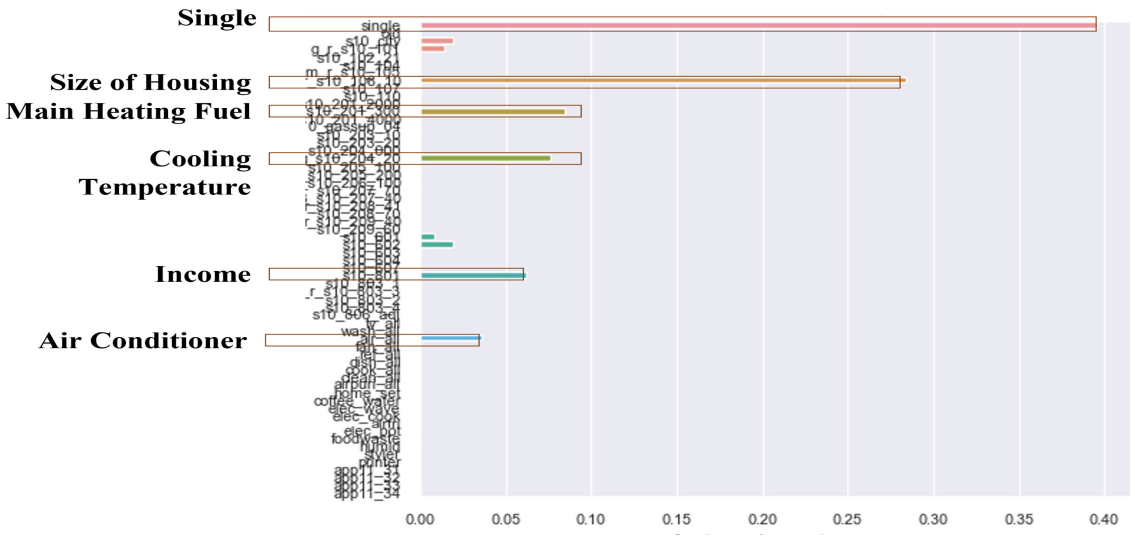

Fig. 13. Feature importance (decision tree, summer)

Table 10. Performance metrics (decision tree, summer)

Type	Precision	Recall	F1-Score	Support
1	0.51	0.36	0.42	417
2	0.49	0.69	0.57	841
3	0.41	0.32	0.36	533
4	0.52	0.19	0.27	177
Accuracy			0.48	1,968
Macro Average	0.48	0.39	0.41	1,968
Weighted Average	0.47	0.48	0.46	1,968

as important features. Compared to annual electricity consumption, features, such as 'cooling temperature' and 'air conditioners', which are more relevant to summer electricity consumption, show a relatively higher importance.

4. Conclusions and Discussions

This study identifies important social, economic, and behavioral characteristics of a household's electricity consumption and applies machine learning techniques, such as SVC, RF, and DT, to classify the household's electricity consumption. From Household Energy Standing Survey, the study brings various types of information, such as electricity consumption, housing and household characteristics, and appliance ownership and usage, and classifies the average monthly electricity consumption of households. First, it starts the classification at the annual level, and the study additionally analyzes it in summer when the cooling demand becomes significant, and electricity consumption reaches a peak. The study finds that the random forest model generally provides better performance metrics compared to the other two methods in classification, as similar to literature (Burnett and Kiesling, 2022).

The feature importance of the random forest and decision tree models allows some interpretation of the results and provides information on the relative importance of independent variables in classification, although those models do not provide the parameter coefficient. The results indicate that housing characteristics are important in understanding the electricity consumption of a household, and the study finds some evidence of the importance of understanding the ownership and usage information of appliances. The future trends of changes in the number of household members per household, and accordingly, the size of housing per household or per capita, which is an indicator for quality of residential environments (Statistics Korea, 2023), would be an important indicator for understanding the residential electricity consumption. The size of housing can be related to the size of key appliances, such as refrigerators, air conditioners, and TVs. In some cases, household characteristics show some importance, but this study could not find their distinctive importance in classifying the household's electricity consumption. A possible explanation for this would be that the housing and appliance usage information can be closely associated with household characteristics, and many households have similar characteristics, such as structure, size, and so on. Also, the importance of cooling temperature is identified as a relatively important feature for classification in summer. The setting of cooling temperature is related to each household's cooling behavior, and this implies the policies for changing behaviors, such as incentives or campaigns for raising cooling temperature, could affect the residential electricity consumption. However, at the same time, this finding implies that an increasing trend in average temperature in summer and the number of hot days, intensified by climate change, would affect the residential electricity consumption. Moreover, the study finds some evidence that information on appliance ownership and usage is an important indicator for understanding residential electricity consumption. This

suggests the need to enhance efforts on collecting data and information on household behaviors. While the survey questions in HESS identify general ownership and usage of appliances, they have limitations on the detailed behaviors behind households' electricity consumption, which would require different types of data collection, such as real-time. However, this study finds limited evidence of the feature 'single household' in the classification of the households' electricity consumption.

The three supervised machine learning methods provide somewhat limited performances in classifying the households' electricity consumption in both cases (annual and summer). Each household shares similar characteristics, such as a number of household members, and a large share of households lives in apartments, so these may lower the performance of the classification of the households' electricity consumption. Particularly, the classification performance of 'class 1' (the lowest electricity consuming group) is relatively lower than the others, and the model predicts those households, which actually consume less than 200 kWh per month, are located in 'class 2' in the annual case. It shows some evidence that the survey information could not distinguish the typical household and other types of households effectively. Also, the dataset includes mixed types of variables with numerical and categorical variables, which make the classification more difficult. Moreover, the models predict many of the households in the test set as the typical group, which contains the largest number of observations, and it implies that many households share similar characteristics, and the survey information reveals limited information about the behavior of households.

In a further study, the different combinations of housing characteristics, household characteristics, and appliance ownership and usage patterns will be considered to understand how each type of characteristic can contribute to understanding the household's electricity consumption. Moreover, the study can be further developed to analyze the characteristics of all four seasons to understand the Korean household's electricity consumption in a better way. The analysis can be further developed to find the implications by comparing it with the analysis based on traditional econometric methods. Moreover, the survey includes the weight information that each surveyed household represents how many of the households in Korea, and the summation of the weights constructs the entire households in Korea. In a further study, sampling methods, such as bootstrapping, could be additionally applied to the surveyed information, and this could generate results, which better represent the households in Korea.

Acknowledgements

The study is based on the Ph.D. dissertation, titled "An Analysis on Sustainable Electricity Supply and Demand in Korea with Application of Machine Learning Techniques."

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A5B5A17088675). This work was also supported by Korea Environment Industry & Technology Institute (KEITI) through "Climate Change R&D Project for New Climate Regime.", funded by Korea Ministry of Environment (MOE) (2022003560010). This study was also based on the findings of the research project "Analysis of Low-energy demand scenarios with Global CGE model" (RR2023-01), funded by the Korea Environment Institute (KEI).

References

Amasyali K, El-Gohary N. 2021. Machine learning for occupant-behavior-sensitive cooling energy consumption prediction in office buildings. Renew Sustain Energy Rev 142: 110714. doi: 10.1016/j.rser.2021.110714

Burnett JW, Kiesling LL. 2022. How do machines predict energy use? Comparing machine learning approaches for modeling household energy demand in the United States. Energy Research & Social Science 91, 102715.

Dahl M, Brun A, Kirsebom OS, Andresen GB. 2018. Improving short-term heat load forecasts with calendar and holiday data. Energies 11(7): 1678. doi: 10.3390/en11071678

Fan GF, Wei X, Li YT, Hong WC. 2020. Forecasting electricity consumption using a novel hybrid model. Sustain Cities Soc 61: 102320. doi: 10.1016/j.scs.2020.102320

Géron A. 2019. Hands-on machine learning with Scikit-Learn, Keras & TensorFlow: Concepts, tools, and techniques to build intelligent systems. Sebastopol: O'Reilly.

Hsu CW, Chang CC, Lin CJ. 2003. A practical guide to support vector classification. Taipei, Taiwan: National Taiwan University.

KEEI (Korea Energy Economics Institute). 2022. 10th household energy standing survey (year 2019); [accessed 2023 Mar 8]. https://www.kesis.net/sub/sub_0001_04.jsp

KEPCO (Korea Electric Power Corporation). 2019. Statistics of electric power in Korea. Naju, Korea: Author.

Keum Y, Kwon OS, Goo YW. 2018. Estimation of residential electricity demand in Korea: By using dynamic panel GMM model. Korean Energy Econ Rev 17(1): 37-65 (in Korean with English abstract). doi: 10.22794/keer.2018.17.1.002

Lee NH, Kim HJ, Seo DH. 2019. Analysis of residential energy use features with respect to location, housing type, gross area and construction year from household energy standing survey. J Korean Inst Archit Sust Environ Build Syst 13(6): 545-558 (in Korean with English abstract). doi: 10.22696/jkiaebs.20190047

Lee NH, Kim HJ, Seo DH. 2022. Estimation of end-use consumption and building energy efficiency rating criteria of prototypical residential buildings based on household energy standing survey. J Korean Inst Archit Sust Environ Build Syst 16(1): 80-93 (in Korean with English abstract). doi: 10.22696/jkiaebs.20220008

Loi TSA, Ng JL. 2018. Analysing households' responsiveness towards socio-economic determinants of residential electricity consumption in Singapore. Energy Policy 112: 415-426. doi: 10.1016/j.enpol.2017.09.052

Menon AK. 2009. Large-scale support vector machines: Algorithms and theory; [accessed 2023 Jun 29]. https://cseweb.ucsd.edu/~akmenon/ResearchExam.pdf

Ministry of Trade, Industry and Energy. 2020. Announcement of 9th basic plan on electricity demand and supply (2020~2034). Sejong, Korea: Author.

Moon J. 2022. An analysis on sustainable electricity supply and demand in Korea with application of machine learning techniques [doctoral dissertation]. Yonsei University.

Murty MN, Raghava R. 2016. Support vector machines and perceptrons: Learning, optimization, classification, and application to social networks. Cham: Springer. doi: 10.1007/978-3-319-41063-0

Noh SC, Lee HY. 2013. An analysis of the factors affecting the energy consumption of the household in Korea. J Korea Plann Assoc 48(2): 295-312 (in Korean with English abstract).

Park HS, translator. 2019. Introduction to machine learning with Python. Seoul: Hanbit Media.

Shin D. 2018. An analysis on the relationship between aging and residential electricity consumption in Korea. Korean Energy Econ Rev 7(1): 95-129 (in Korean with English abstract). doi: 10.22794/keer.2018.17.1.004

Statistics Korea. 2022a. Household prospects: 2020~2050; [accessed 2023 Apr 24]. https://kostat.go.kr/board.es?mid=a10301020600&bid=207&act=views&list_no=418919&tag=&nPage=1&ref_bid=

Statistics Korea. 2022b. World and Korea population prospects based on 2021 population prospects; [accessed 2023 Apr 24]. https://kostat.go.kr/board.es?mid=a10301020600&bid=207&act=view&list_no=420361

Statistics Korea. 2023. Housing size per capita; [accessed 2023 Sep 14]. https://www.index.go.kr/unify/idx-info.do?idxCd=4257

Winters-Hilt S, Merat S. 2007. SVM clustering. Proceedings of the Fourth Annual MCBIOS Conference. Computational Frontiers in Biomedicine; 2007 Feb 1~Feb 3; New Orleans, LA, USA: MidSouth Computational Biology and Bioinformatics Society. p. S18. doi: 10.1186/1471-2105-8-S7-S18

Yu B, Wei YM, Kei G, Matsuoka Y. 2018. Future scenarios for energy consumption and carbon emissions due to demographic transitions in Chinese households. Nat Energy 3(2): 109-118. doi: 10.1038/s41560-017-0053-4