[1]Dr. Mohamed Adel Al-Shaher

Classification Techniques to Predict the Behavior of Electrical Energy Consumption

JES Journal of Electrical Systems

Abstract: In this paper, we studied the feasibility of developing a complete intelligent system for predicting the behavior of electrical energy consumption as electricity is gaining room as energy source. This work is presenting an overview to organize and visualize all the data using different representations according to the specific goal of classifying the consumption of electricity usage with respect to its behavior. It has been interesting to provide a method that allowed the visual representation for each algorithm used for classification of electrical data and satisfies all the purposes sought. This facilitates the comparison among electricity usage, permits to access easily to the individual electric consumption main characteristics and detect possible irregularities or problems. Electricity is gaining room as energy source, its share will keep increasing constantly in the following decades. In this close future, smart grids and smart meters deployment will benefit both the utility and the consumer. This work has classified the electricity usage with respect to the consumption according to the similarities of their electrical load profiles, using the proportion of energy usage per hour as a common framework using the classification (five algorithms in particular) and clustering theory. The objective behind this segmentation or classification is to be able to provide personalized recommendations to each group in order to reduce their energy consumption and the associated costs, fostering energy efficiency measures and improving the electricity usage engagement. The desired classification is obtained by an iterative process, based on computational classification calculation (using Weka for analysis and results generation) and finalized by a post-clustering and classification analysis applying visualization and statistical techniques to detect the outliers and reallocate them to a more appropriate classes. Five different classification techniques (Decision Tree, Support Vector Machine, Naïve Bayes, Random Forest and Hybrid), were tested and compared, giving similar outputs. The solution from the Hybrid is the one that better adapts to the classification sought, which is used as the base of the post-classification stage to obtain the final classification.

Keywords: prediction, electricity consumption, smart meter, classification, clustering, intelligent system, data mining, electric usage

1. INTRODUCTION

Energy efficiency has been gaining importance in the energy world's priorities, it is named as the "invisible fuel" stating that the best choice is not to waste energy. In regions like Europe with highly external energy supply dependences, an optimization at all stages in the energy chain is a must from both environmental and economic point of view. Rules and obligations set by the Energy Efficiency Directive are in that direction. The target of the European Union is to reduce up to 20 % the energy consumption, achieving by 2020 an energy consumption lower than 1.474 of primary energy or less than 1.078 of final energy, but setting an objective adequate to each country characteristics [1-2].

This directive differentiates the energy efficiency in energy supply and energy efficiency in energy use. The measures adopted by the European Union can be summarized:

- Energy distributors and retailers must reduce annually 1.5% their energy sales
- Annual energy efficient renovation of at least 3% of buildings owned or occupied by governments
- Incentive the buildings renovation, i.e. adding insulation, double glaze windows, high efficient boilers; to

[1]Computer Science - Information Technology, Computer Department, College of Computer Science and Mathematics, University of Thi-Qar , Nassiriyah; Iraq

Email: alshaher_comp82@sci.utq.edu.iq ; alshaher2016@gmail.com ; alshaher2006@yahoo.com

improve their energy performance

- Mandatory energy performance certificates when renting or selling buildings
- Set of minimum labels or standards for a range of products, such as boilers, domestic appliances, lighting.
- Periodical energy audits for large companies and incentives for Small and Medium enterprises to undergo energy audits
- Protecting consumer rights to receive comprehensive information, access to real-time and historically energy consumption and billing data to a better consumption management
- Deployment of 200 million of electricity smart meter (72% of the total) by 2020

The energy efficiency in energy supply is principally focused on the highly-efficiency cogeneration plants (electricity + heat production), implementation of district heating and cooling and integration of renewables. Together with the energy savings attained due to suppliers obligations and the application of the white certificates [1]. Optimizations in addition to the Energy Efficiency Directive, there are three other directives devoted on building's thermal and electrical demand, as buildings represent around 40% of the total energy demand in Europe these directives are:

Energy performance of buildings directive: include building energy certificates when selling and renting buildings; finance renovation of building elements to a low energy needs, and all new buildings must be nearly-zero energy buildings.

Energy labelling directive: intending to help consumers to have more information in order to choose energy efficient products (air conditioners, television, washing machines, lights…)

Eco-design directive: directed on product manufacturers, regulating the requirements of manufacturers to establish the minimum energy efficiency standards.

In contrast, with the new Distributed production is not only delivering a service also including other features that improve the whole process. The production is also close to the consumption points, thanks to the small and medium plants built mostly by renewable energy sources (PV, wind…) or new technologies like fuel cells [3], the possibility to store the energy to use when and where is necessary, thanks to a network that becomes bidirectional, prioritizes the demand by using more efficiently the resources, taking advantage and the big data to shave consumption peaks and engage the customer, who will have an active role.

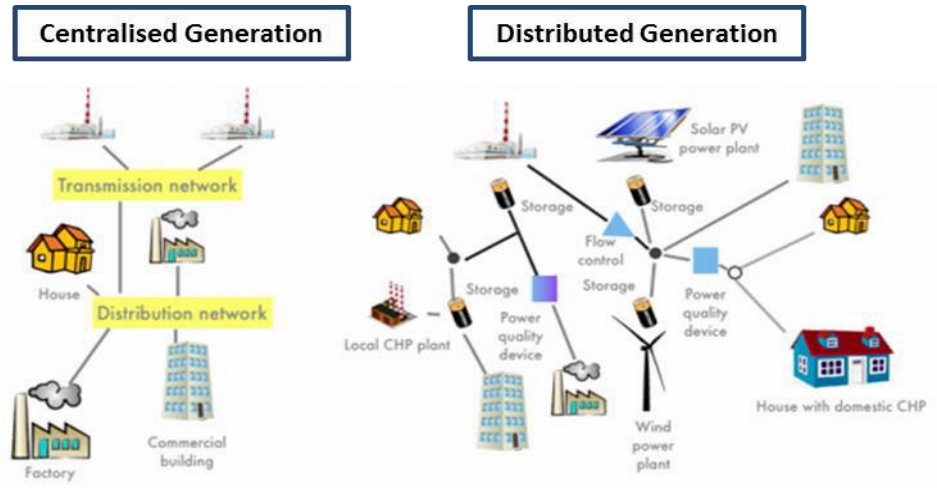


Figure 1: Mapping FEC chain operations to particular functional units [3].

2. BACKGROUND

Electricity is called to be the fastest-growing component on the energy supply portfolio; the increase on electricity consumption is expected to achieve the 40% of the total energy market by 2020 [12] and reaching an important share by 2050. This growth will be essentially led by the emerging economies and developing countries, as is shown in Figure 2, while a modest increment in other developed economies will occur.

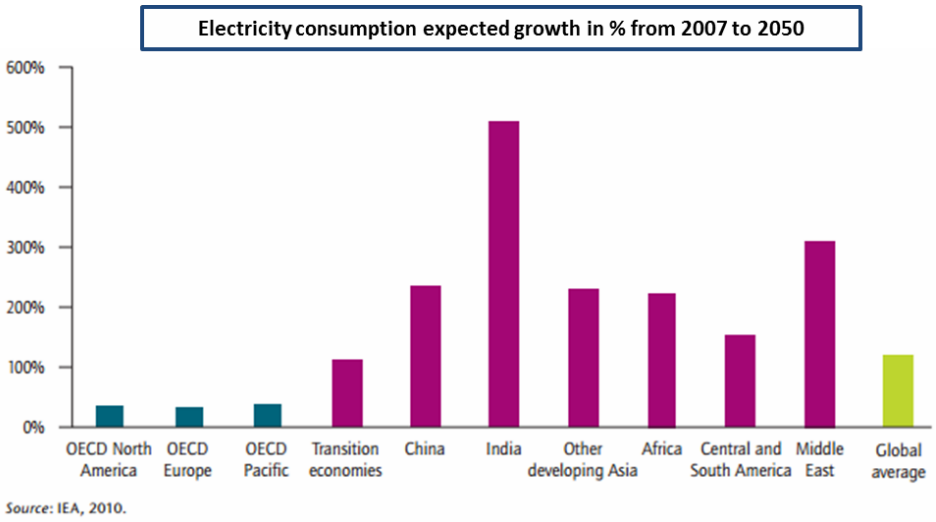


Figure 2 : Electricity consumption expected % increase by regions. [12]

This large growth on electricity usage is likely to arise in the:

- Residential heating and cooling due to the use of heat pumps. Either individual at home heat pumps or larger district heating and cooling heat pumps.
- The electrification of the transportation and mobility, due to vast deployment of Electric Vehicles and the Plug-in Hybrid Electric vehicles.
- Variable electricity generation, by using PV, wind, hydro, biomass, combined heat and power (CHP) technologies, decentralizing the production.

This new conjuncture will tackle the climate change to a reduction, while covering the same needs moving from rich-carbon sources to low-carbon ones. And maintain the economic and social development by using more sustainable systems; however the electrical grid will have to be ready and able to host these new features.

Polar For the present work, a special attention is given to the smart meters and to the Advanced Metering Infrastructure AMI, as from them the consumption data generated [3]. This infrastructure is still under development but many regulators and governments put the focus especially to the smart meter deployment. It is the case of the European Union [4], inside the energy strategy for 2020 program. Electrical markets were heavily regulated in some countries, fact that difficult the implementation of new technologies and quick changes to the conventional market scheme, due to its rigidity. This has arisen the implementation of private smart meters mainly to the industry, commercial and office building; as allows an energy management with future economic benefits. However in the residential sector the energy savings are not compensating the investment on private sub-metering equipment, and so relying on the "official" smart meter implementation to start managing their energy consumption. They have the role to smooth the operation of the power system, many of these tasks are responsibility of the regulated utility such as the maintenance and construction, meter reading or security management; and some of them can be outsourced to other domains.

However, [3] suggests a further investigation on the households' features and householders' characteristics is needed in order to be able to determine the possible causes and the origin of the load patterns and whether exists

or not any correlation with the households' features and the load patterns; with so it would be possible to provide more accurate and detailed energy-savings recommendations. For example, the load profile it won't reflect the same whether the kitchen is electric or gas; also whether the hot water boiler is electrical or gas; due to the fact that it they are electric the peaks are likely to be visible. The base consumption refers to the minimum consumption that the house constantly has, this is related to the night hours and the hours where there is no activity at home, for instance when the households are working [3].

Also, nowadays the time-of-use tariffs allows to set different prices each hour, the utilities in order to shave the peaks set higher prices to the peak time, usually the evening. So, again a behavioral change is needed in order to reduce the peaks, the simultaneity coefficient of devices used at the same time should be reduced. For instance, use the washing machine at night, out of peak hours that have been rather classified with known machine learning algorithms in previous study [5].

3. METHODOLOGY

Performing any data analysis of data quality is important and adjusting this data to the correct format is the key to starting the desired analysis. Therefore, it is important to know in advance the purpose of the analysis by having one objective or another process of access to data usage can change completely. Good data optimization will also report better and more accurate results. The first phase of the preliminary analysis, includes (a) Collection of consumer hourly data and features of household and homeowners, understanding how data is presented and reconstructed for ease of study. Once you are familiar with the database and have identified the problem; (b) Data purification process is performed to remove bad data (eggs, freezers, duplicate features). To close this section, a graphical image is required such as (c) to examine data and visualize in order to delve deeper into the data used in the research, and to help monitor its effectiveness.

The second phase is devoted to integration analysis to determine the desired consumer segregation that is the core of the work; divided into pre-collection, collection and assembly categories.

The pre-segment (d) section consists of conducting a study of the same textbooks, selecting the most appropriate input data format to allow accurate comparisons between users' power load profiles, and processing input data with that concept to determine the hourly power consumption per user.

In the (e) classification category; a review of existing consolidation strategies used for the separation of electricity consumption data is underway. Hierarchical, Naïve Bayes and Support vector machine techniques are selected and used individually in the multiplication process, based on computational computer calculations (using weka software) to obtain the total number of clusters with the correct number of members in each of them. Finally, visual and statistical comparisons of the results of each integration method were made to determine which solution was most appropriate.

The final stage, (f) post-separation; dedicated to obtaining the final and consumer segregation of the consumer through the redistribution of users. This is done by hand-operated interventions by the analyst using visual and statistical techniques to be able to identify vendors and re-assign them to the right group.

At the third stage, the (g) household and householders characteristics are analyzed. Selecting those features that could add value to the analysis according to related previous papers, in this case histograms are used to find the most common characteristics inside each group of consumers.

The aim is to organize and visualize all the data using different representations according to the specific goal of the visualization. It is interesting to provide a tool that allows the visual representation for each individual customer and satisfies all the purposes sought. This facilitates the comparison among consumers, permits to access easily to the individual electric consumption main characteristics that, for instance, facilitate a quick audit on its consumption and detect possible irregularities or problems.

Among these purposes are:

- Differentiate the weekdays and weekends load profile, are compare them to the load profile output accounting all days not differentiating weekdays and weekends.

- Differentiate the load profile for each day of the week (Monday-Sunday), to know if among the weekdays the load profiles are similar or not (Friday may differ from Monday-Thursday)
- Representation in absolute values, percentage of consumption per hour, accumulated in a period of time.
- Study if there is a consumption difference among months and seasons.
- Find the characteristic load profile per each month of the year
- Be able to specify exact dates and plot the consumption at that period
- Using different kinds of visualization graphs shapes (bars, points, lines, boxplots…) and
- color to facilitate the information visualization
- Facilitate the comparisons between users by using interactive graphs.

Hence, it is observed that diverse techniques can be used to achieve the electrical load pattern grouping, each technique has its own particular approach to reach the same final goal. In that sense, for the sake of the current analysis the considered methods are:

- Decision Tree, Random Forest
- Naïve Bayes classification
- Support vector machine (SVM)

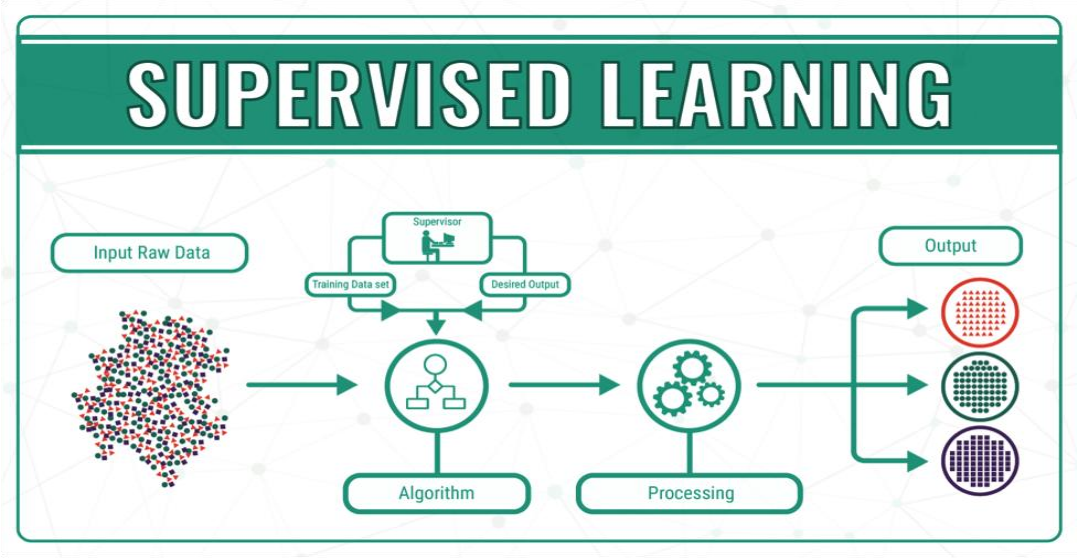


Figure 2: Supervised learning approach in which 5 algorithms were used in our case of study [5].

The data regarding the electricity consumption is the main one used in the later analysis; it is numerical data as its records the measurements of electrical consumption for every household participating into the project.

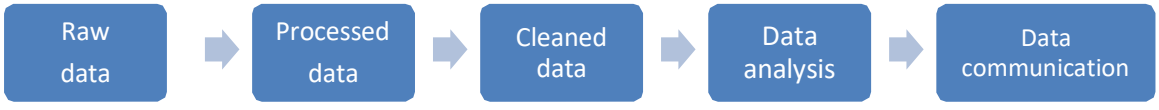


Figure 3: States of the data when performing a data analysis before classification.

When exploring the data, three aspects should be taken into account:

- First, regards the dataset format as in most of the cases it needs to be reshaped into the best format to have the desired graphical output.

- Second, the dates' treatment is a challenge when exploring the data, as time series data needs further and different approaches. Especially, if division between weekdays and weekends, or monthly distinction is needed.
- Third, focus on the specificity of the electricity use data, studying which is the best path to extract the information or conclusions sought.

3.1. Clustering

The similarity characteristics of the data are not known in advance. The dataset is divided into clusters or groups. Objects inside the same group have similar properties to each other; and differ from instances of other groups. It is an unsupervised learning, due to the absence of a training dataset able to provide prior knowledge. The objective is to label or group the observations, according to their similarity.

3.2. Classification

Exists a training dataset that was used to previously to subset this data in groups. New data is classified based on the training dataset. Algorithms find the group to which each new object belongs to due to its common characteristics. It is a supervised learning task due to the existence of a training statistical dataset that has labelled or grouped the data previously. The objective is each new data object into an existing group.

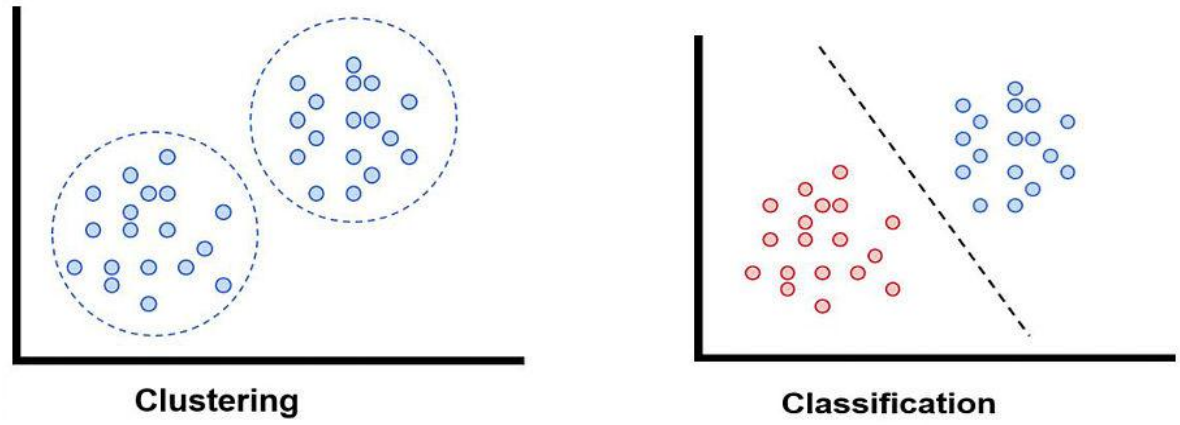


Figure 4: Clustering and Classification grouping in data mining while clustering on the left and classification on the right [5].

3.3 Classification Algorithm

Our main objective consists on predicting electricity consumption and this can be divided into two parts: 1) have a classification model that detects, with high accuracy, the occupancy of any household by using solely the electricity consumption data; 2) have a prediction model that predicts occupancy based on the detected occupancy data for an intelligent system. In the next sections we explain the theory behind each classification algorithm chosen.

3.3.1. Support Vector Machines

Vector support devices (SVM) are an efficient mechanical-guided algorithm that is widely used to detect patterns and differentiation problems such as facial recognition, genetic extraction and speaker identification. It can be used to perform line and non-linear separation by kernel methods and works well even with small training data samples (compared to neural networks) [4]. SVM also has excellent ability to handle high-volume data [6]. In this work, three types of SVM were used: linear kernel, radial (or Gaussian) kernel and polynomial kernel.

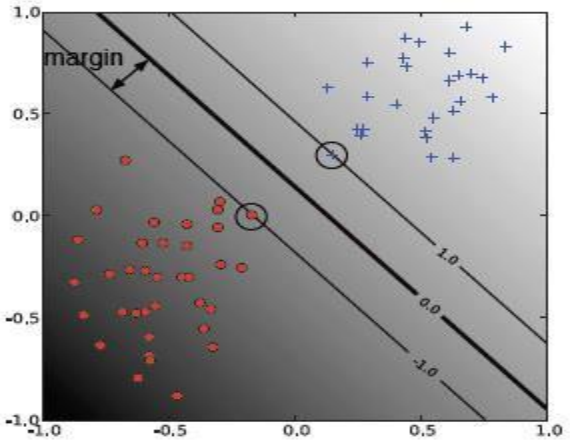


Figure 5: Example a two-class linear SVM classifier [6].

3.3.2 Random forest

A random forest is a machine-learning mechanism used to isolate and retreat activities. This method of learning together includes many trees of unrelated decision-making decisions obtained by combining the results in each decision tree. The combination of multiple separators reduces the model excessively, thus increasing the accuracy of the sections. The standard procedure used for compiling algorithms is called Bootstrap aggregating or Bagging and is shown in Figure 6. This method is used to replicate the data and to produce different training sets for each separator by entering data.

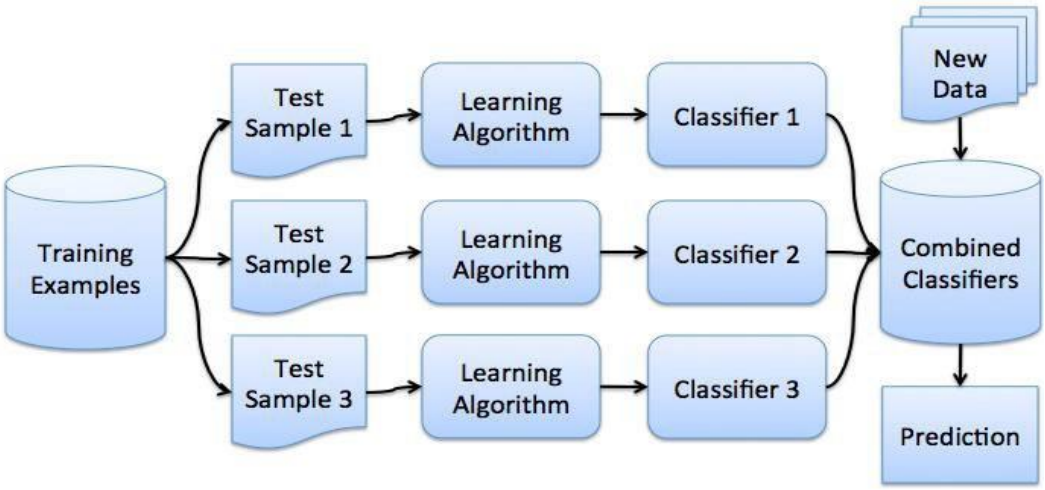


Figure 6: Phases of ensemble learning approaches to solve classification problems [7].

3.3.3. Decision Tree

The decisive tree makes the division of the data division and is made up of three types of nodes: root node (at1), internal node (at2, at3 and at4) and leaf nodes (yes with no results), as shown in Figure 7. The roots and the inner parts represent the test in the element / feature and the branch represents the result of that test. The tree is split in half until no further algorithm tests are performed to make the split. The tree ends with a leaf node (or terminal) that contains the splitting effect.

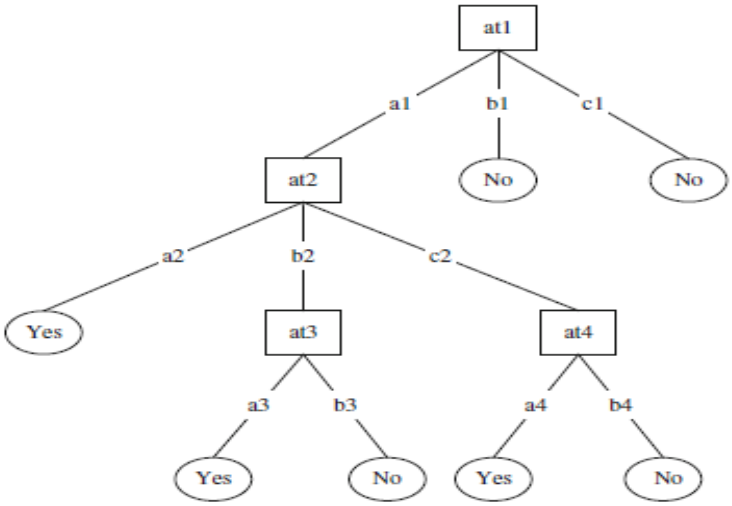


Figure 7: Decision tree algorithm structure [7].

3.3.4. Naïve Bayes

We selected the naïve bayes classification algorithm to make possible residential classes, tried later on the set. The probability that the house involved in a given day of the week and the length of time is registered by dividing the number of times involved in the total number of times available for each period and within a set period of the plan. For example, if our subdivision set has weeks of details, and if in these weeks the house was always given a share, then the chances of being at this time during the separation. The classification of naïve buses was then considered by setting the situation at all times of the week from classes.

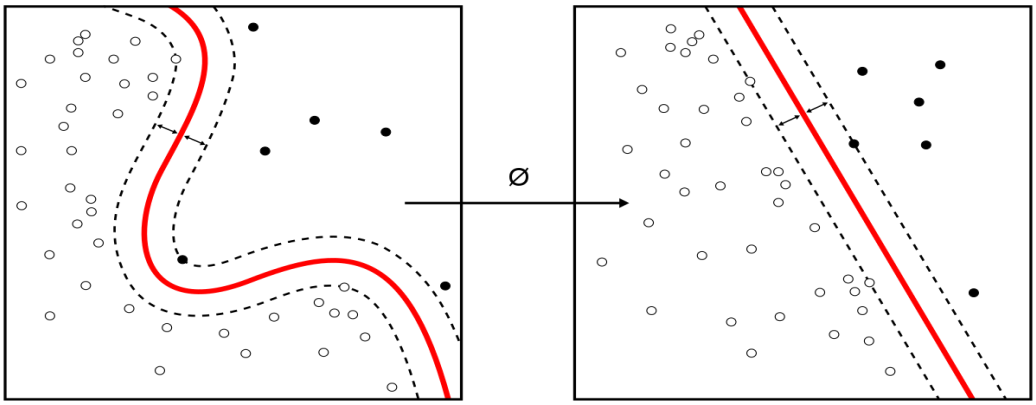


Figure 8: Naïve bayes algorithm on left with respect to support vector machine on right side for classification structure [8].

3.3.5. Hybrid Model

We use exactness as the fundamental assessment criteria for our order and forecast models using the combination of random forest and decision tree, which shows the level of right grouping or expectations. Different measurements were likewise utilized as a correlative measure, as depicted further in this in part. We assess the likelihood of foreseeing the inhabitance of a family unit dependent on its power utilization via preparing a model and testing it in a similar family. In any case, the fundamental business challenge is to utilize a conventional model that predicts with high exactness the inhabitance of any family unit. To this end, we prepared a model in a solitary family unit and tried in numerous families.

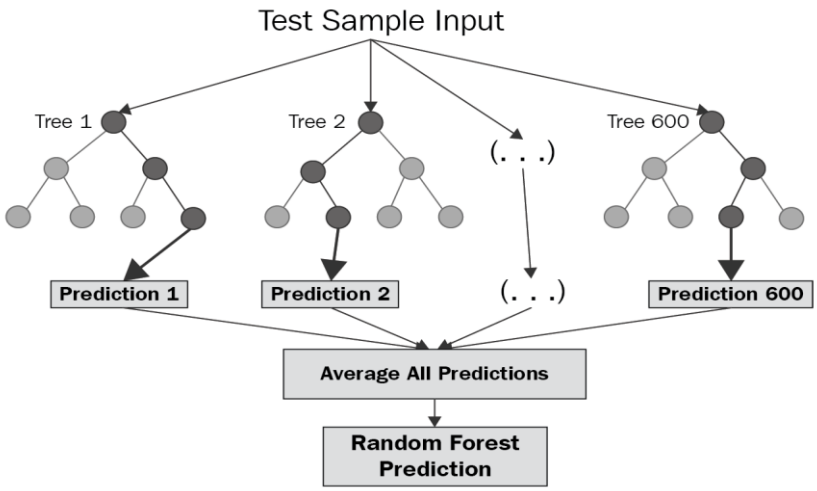
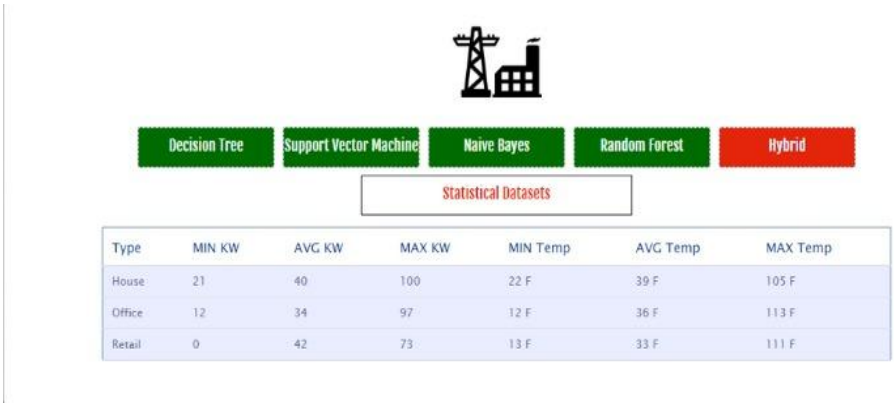


Figure 9 : Hybrid model algorithm present as combination of random forest and decision tree for classification structure .

4. RESULTS

In this work, the geometrical and statistical center of each class is computed first, and the distance between the two classes equals the distance between the two centroids for different types of classification algorithms.

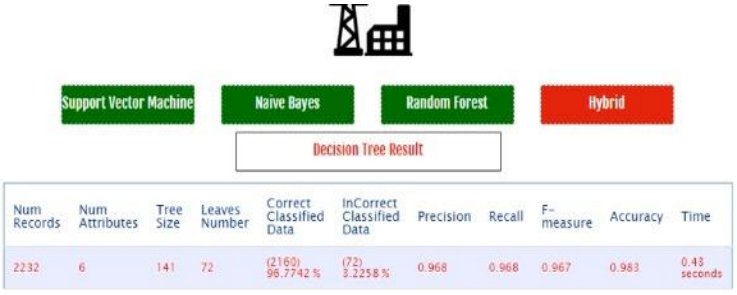


Type	MIN KW	AVG KW	MAX KW	MIN Temp	AVG Temp	MAX Temp
House	21	40	100	22 F	39 F	105 F
Office	12	34	97	12 F	36 F	113 F
Retail	0	42	73	13 F	33 F	111 F

Figure 10 : The overall classification results based on approaches used on statistical dataset.

4.3.1. Decision Tree Prediction

The output results used to carry out the classification analysis vary depending on the specific aim of the load profile's energy consumption segmentation plus classification using decision tree classification, and it is up to the analyst to decide which are the most convenient input data units for output results; the use of absolute values in result including number of records, precision, recall, F-measure, accuracy and time for the use dimensionless data with time factor for generating and execution is absolute 0.43 seconds with recorded precision of 0.968.



Num Records	Num Attributes	Tree Size	Leaves Number	Correct Classified Data	InCorrect Classified Data	Precision	Recall	F-measure	Accuracy	Time
2232	6	141	72	(2160) 96.7742 %	(72) 3.2258 %	0.968	0.968	0.967	0.983	0.43 seconds

Figure 11: The decision tree classification results based on approaches used.

4.3.2. Random Forest Prediction

The output results used to carry out the classification analysis vary depending on the specific aim of the load profile's energy consumption segmentation plus classification using random forest classification, and it is up to the analyst to decide which are the most convenient input data units for output results; the use of absolute values in result including number of records, precision, recall, F-measure, accuracy and time for the use dimensionless data with time factor for generating and execution is absolute 1.24 seconds with recorded precision of 0.949.

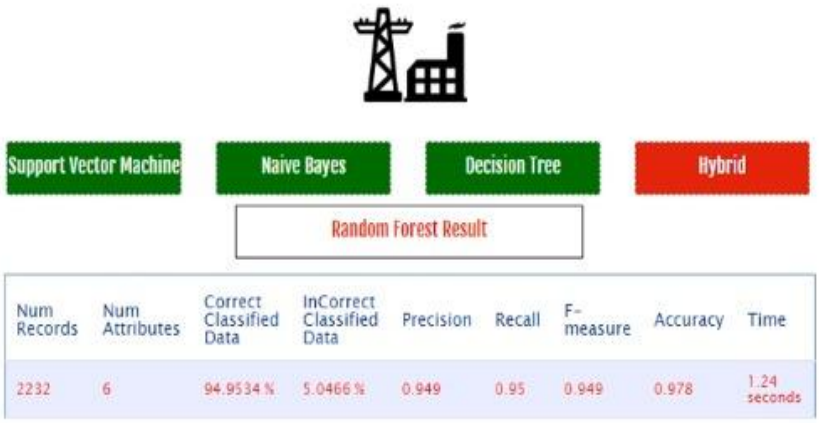


Figure 12: The random forest classification results based on approaches used.

4.3.3. Support Vector Machine Prediction

The output results used to carry out the classification analysis vary depending on the specific aim of the load profile's energy consumption segmentation plus classification using support vector machine classification, and it is up to the analyst to decide which are the most convenient input data units for output results; the use of absolute values in result including number of records, precision, recall, F-measure, accuracy and time for the use dimensionless data with time factor for generating and execution is absolute 0.99 seconds with recorded precision of 0.906.

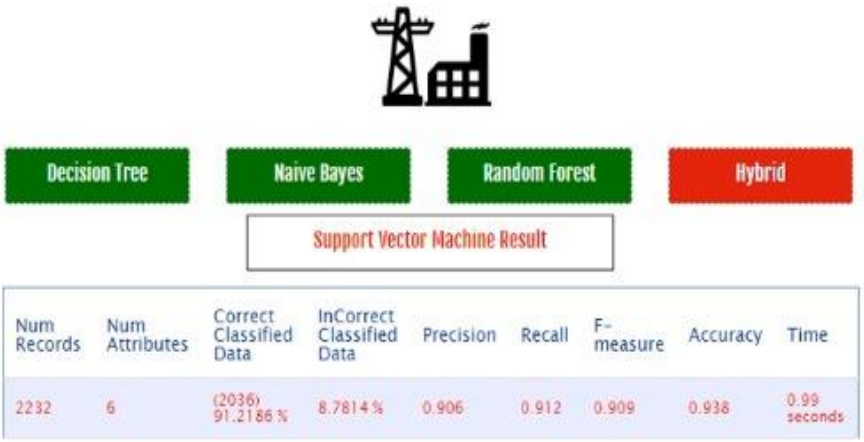


Figure 13: The support vector machine classification results based on approaches used.

4.3.4. Naïve Bayes Prediction

The output results used to carry out the classification analysis vary depending on the specific aim of the load profile's energy consumption segmentation plus classification using naïve bayes classification, and it is up to the analyst to decide which are the most convenient input data units for output results; the use of absolute values in result including number of records, precision, recall, F-measure, accuracy and time for the use

dimensionless data with time factor for generating and execution is absolute 0.24 seconds with recorded precision of 0.801.

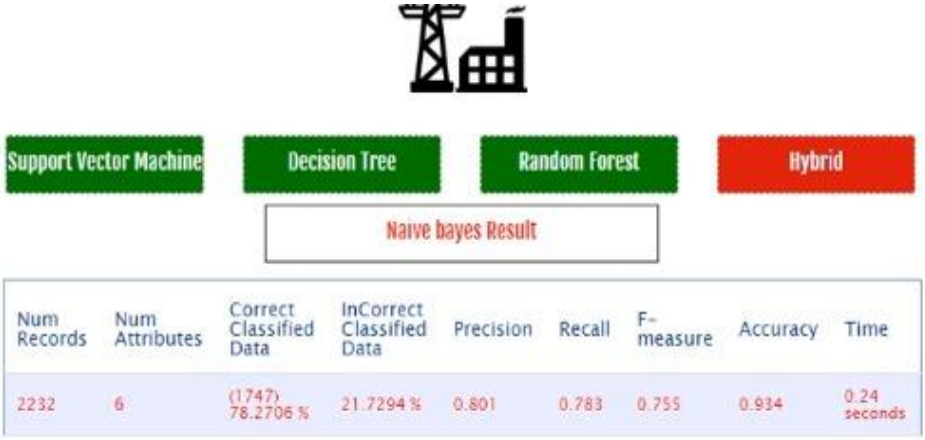


Figure 14: The naïve bayes classification results based on approaches used.

4.3.5. Hybrid Model Prediction

The output results used to carry out the classification analysis vary depending on the specific aim of the load profile's energy consumption segmentation plus classification using hybrid classification which knowingly involves both decision tree plus random forest, and it is up to the analyst to decide which are the most convenient input data units for output results; the use of absolute values in result including number of records, precision, recall, F-measure, accuracy and time for the use dimensionless data with time factor for generating and execution is absolute 2.23 seconds with recorded precision of 0.969.

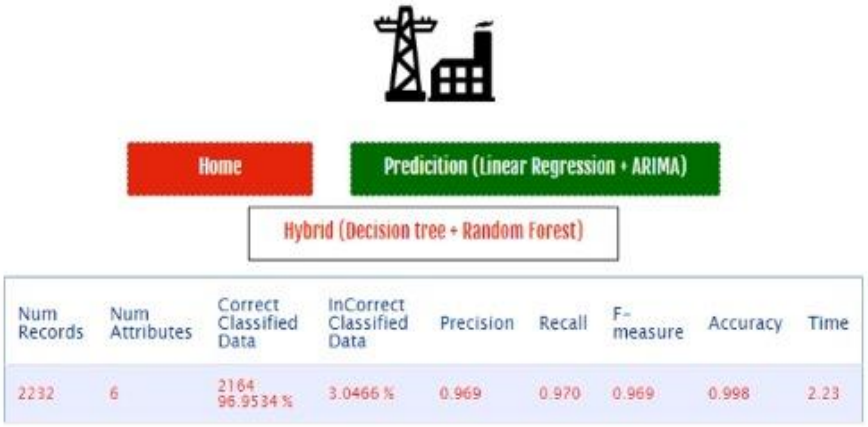


Figure 15: The hybrid model classification results based on approaches used.

4.3.6. Prediction Result

Calculating the AVG KW for the mean square error of each load profile respect to its class mean load profile, will allow the checking the class assignation and determining the classes outliers, which are the furthest load profile's to the mean and they might be reallocated in another prediction class between range of months with respect to type.

Type	Prediction Class	Range of Month	Range of Hours	AVG Temp	AVG KW	MSE (Mean Square Error)
House Retail Office	1 (Low)	3-6 and 9-11	4-9 and 18-21	60 F or less than 10F	less than 10per day	0.02
House Retail Office	2 (med)	7 and 12	9-12 and 12-1	40 F	less than 30per day	0.04
House Retail Office	3 (high)	8,1 and 2	4-9 and 18-21	less than 35 F or more than 100f	more than 40per day	0.03

Figure 16: The prediction of result for classifying the consumption of electricity in terms of mean square error.

5. DISCUSSION

In discussion section, after analyzing these features, they could detect the group of households with large energy savings potentials, mainly due to the high number of electrical appliances. And concluding that the characteristics that were more relevant to the define a consumer with high savings potential are: type of occupants employment, number of adults/children, type of space heating, type of domestic hot water, heating, total number of home appliances and dwelling construction year. However, for the current study case it is not available such amount of detailed data as in [9-10]; the features' data is much more limited in terms of samples and properties. The data referred to the household and householders is described in this work, this data is not fully complete and also needs to be further treated in order to eliminate bad data and duplicated features. Once refined, the data is able to be analyzed, in this case a graphical analysis using histograms is considered to represent the results and be the base to extract conclusions.

Table 1: compare the algorithms results

Algorithms	**Accuracy**
Decision tree	0.983
random forest	0.978
Support Vector Machine (SVM)	0.938
Naïve Bayes	0.934
Hybrid Model	0.998

6. CONCLUSION

The objective of this work is to study the feasibility of developing an intelligent system for predicting the behavior of electrical consumption through weka software on general-purpose-processor. The study is an example of a public administration pioneer initiative to engage the consumers and foster the electrical energy efficiency and consumption among them, aiming to provide energy knowledge, understanding and guidance to the user in order to reduce its consumption. Nonetheless, it was seen that providing the technical consumption data in kWh to the householders has limited influence and effect , so is necessary to go beyond and translate the technical information into call-to-action measures, guiding the user to smart energy choices (like a 5 data mining

approaches and algorithms), have a better response from the householder. Also the need of sub-metering devices to obtain the consumption data limits the scalability, as the cost associated is significant. To be able to reproduce this kind of project in a large-scale the access to the data from the classification algorithms are necessary. So, the combination of the smart meter deployment and the big data analytics are called to play an important on the energy sector. The data mining techniques generate large amounts of raw data that need to be managed and, once analyzed can be converted into useful information that benefits both the utility and the consumer, as aim to improve the customer engagement and the quality of the service. Creating new business opportunities, mainly related to data science and data analysis in response to the market needs.

REFERENCES

[1] Ardakanian, O. et al., 2014. Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference on CEUR-WS.org.

[2] Armaroli, N. & Balzani, V., 2011. Towards an electricity-powered world. Energy Environmental Science, pp. 4, 3193-3222.

[3] Beckel, C., Sadamori, L. & Santini, S., 2012. Towards automatic classification of private households using electricity consumption data. Embedded Sensing Systems for Energy-Efficiency in Buildings: Proceedings of the Fourth ACM Workshop, (BuildSys '12), pp. pp.169-176.

[4] BloomEnergy, 2015. Fuel Cell: Distributed Generation. [Online] Available at: http://www.bloomenergy.com/fuel-cell/distributed-generation/

[5] A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines," in Methods in molecular biology, 2010, pp. 223-239

[6] Chicco , G. & Ilie, I., 2009. Support vector clustering of electrical load pattern data. IEEE Trans. Power Syst, 24(3), pp. 1619-28.

[7] E. Goel and E. Abhilasha, "Random Forest: A Review," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 1, pp. 251-257, 2007.

[8] Chicco , G. et al., 2005. Emergent electricity customer classification. IEE Proc Gener Transm Distrib, 152(2), pp. 164-72.

[9] Chicco , G. et al., 2004. Load pattern-based classification of electricity customers. IEEE Trans. Power Syst, 19(2), pp. 1232-9.

[10] J. Kelly and W. Knottenbelt, "Neural NILM: Deep Neural Networks," in Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments, 2015.

[11] K. C. ARMEL, A. GUPTA, G. SHRIMALI and A. ALBERT, "Is disaggregation the holy grail of energy efficiency? The case of electricity," Energy Policy, vol. 52, p. 213–234, 2013.

[12] S.Raschka,"MachineLearningFAQ,"[Online].Available:https://sebastianraschka.com/aq/docs/evaluate-a-model.html. [Accessed 04 04 2017].