ISSN 2278-3091
Volume 9, No.1.4, 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse9491.42020.pdf
https://doi.org/10.30534/ijatcse/2020/9491.42020

A Web-Based Non-Intrusive Appliances Monitoring System Using Logistic Regression Technique

Ling Yee Tan, Ahmad Asrul Ibrahim*, Muhammad Ammirrul Atiqi Mohd Zainuri, Nor Azwan Mohamed Kamari
Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia, * ahmadasrul@ukm.edu.my

ABSTRACT

Energy supplies are expected to be expanded in order to meet the emerging demands in this modern era due to highly depends on electricity. Energy demand has been reported to increase by 2.3% in 2018 and this causes increasing in electricity price. Energy consumption in houses or factories is given more attention due to our concern on every single cent that we need to pay for the electricity bills. Consequently, the status of each appliance usage should be monitored for appropriate saving actions to be taken. However, this requires a tremendous amount of meter being clamped on the electrical appliances to monitor their status. This paper presents a non-intrusive monitoring approach using logistic regression technique to identify whether the appliances are switched on or off from the total consumption. The logistic regression estimates the parameters of a logistic model giving output in form of binary such as on or off. Individual logistic model is proposed for each appliance to give better estimation. Then, the developed logistic regression models are embedded in a web-based user interface so that user can monitor the status of appliances easily. As a result, consumers will be able to monitor their energy consumption and status of electrical appliances usage at anywhere and everywhere.

Key words: Energy consumption, Logistic regression, Non-intrusive load monitoring, Status of appliances; Web-based user interface.

1. INTRODUCTION

The use of electricity has undoubtedly played an important role in daily life especially during the modern era today. The global contributions from buildings including both residential and commercial, have steadily increased and reached figures between 20% and 40% of total energy consumption in developed countries [1]. As the demand for electricity usage increase tremendously day by day, energy efficiency in buildings becomes a prime objective for energy policy at regional, national and international level [1]. This includes encourage more renewable energy integrations [2] and actively engage with consumers in distribution systems [3]. At domestic level, consumers are encouraged to use electricity wisely where they are provided with enough information relates to their energy usage behavior. Research on this topic has attracted attention among researchers and local authorities within their efforts aimed at overall energy savings and sustainable use [4].

In the current situation, consumers are notified about their monthly electricity consumption only after receiving their utility bill. However, the status and consumption at any particular time cannot be identified. In our daily lives, there are times when we are unsure if electrical appliances at home are switched off while we are away. One of the solutions is that all appliances are equipped with monitoring devices or meters. Adequate monitoring scheme is important to give enough information but too much allocation on it causes redundancy and not cost-effective [5-6]. Besides that, the status of appliances whether they are switched on or off can be identified from the existing meter at the entrance which is known as Non-Intrusive Load Monitoring (NILM). This is where individual loads can be detected and separated from rapid sampling of power signal at a single point that serves a number of equipment, for example the electrical service entrance for an entire house or all of the central space-conditioning equipment in a commercial building [7]. The features of each appliance's consumption can be extracted from measurement at the entrance point using time series signal processing techniques such as S and TT transform [8] or Hilbert Huang transform [9]. However, these techniques are complicated and unable to learn from new features. Alternatively, machine learning is a viable solution to solve the problem and easy to be implemented.

Machine learning is a subset of Artificial Intelligence (AI). It is an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment. They are considered the working horse in the new era of the so-called big data [10]. In machine learning, it consists of two important process which are the training and testing process. Training data set is a subset used to train a model while testing dataset is a subset used to test

the trained model. Both set of data should be large enough in order to yield statistically meaningful results [11]. Training and testing set of data usually will be break into two portion which normally around 8:2 or 7:3 [12]. In machine learning, it breaks into supervised learning, semi-supervised learning, unsupervised learning and reinforcement learning [13] as depicted in Figure 1.

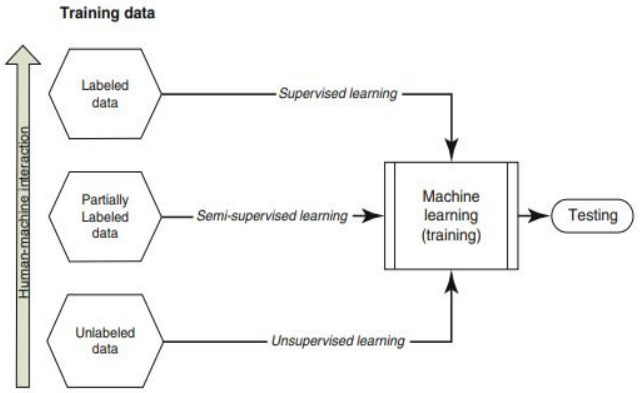

Figure 1: Categories of Machine Learning Algorithms According to Training Data Nature [10]

Among of the categories, supervised learning is very helpful in classification problems [14]. In the supervised learning, learning function maps an input to an output which based on an input-output example pair. The function must be provided with a set of training examples. In [15], a sentiment classification of online reviews on travel destinations is solved using supervised machine learning approaches based on the model of Naïve Bayes. They realized that travel-related information is important to travelers for planning process. Therefore, they obtained reviews towards seven popular travel destinations in US and Europe through different travel blogs and uses the concept of supervised machine learning to classify them into positive or negative reviews. According to [16], simple linear regression performed better than other algorithms such as Random Forest, Decision Table, SMOreg and LWL in classification of student's performance. This clearly indicates that regression technique is a good alternative to solve the classification problems. However, this work focuses to identify the status of appliances at all time. Hence, logistic regression which falls under the category of supervised learning has been chosen to deliver the task. It has been reported that the logistic regression is good in solving binary (0/1) problems [17]. Therefore, the performance of the logistic regression in NILM application to identify the status of appliances should be investigated.

2. LOGISTIC REGRESSION

As mentioned earlier, logistic regression can be used as non-intrusive technique in load monitoring system. It can be derived from the algorithm of linear regression which applies the linear equation as follows:

$$y = mx + c$$

If more inputs are needed, a multiple-linear regression is required and expressed as the following:

$$y = \sum_{i=1}^{N} m_i x_i + c \tag{2}$$

where, N is total number of inputs. Linear regression or multiple-linear regression is normally used for continuous domain problems. In logistic regression, the output from (2) is estimated into binary domain (0/1). The output estimation can be derived from the following equations [18]:

$$p = \frac{1}{1+e^{-y}} \tag{3}$$

$$Y = \begin{cases} 1 & \text{, if } p \geq 0.5 \\ 0 & \text{, otherwise} \end{cases} \tag{4}$$

where, Y is output in binary to indicate status either on (1) or off (0).

3. NON-INTRUSIVE MONITORING DEVELOPMENT

The logistic regression has been developed using Python 3 within Google Colaboratory environment. This is because dataset obtained requires a large size of memory which is difficult to be supported by local computer. In this case, the Google Colaboratory will store the dataset in cloud which allow the program to run faster and easy to access. Figure 2 shows a block diagram of the entire system development from data collection until produce a web-based user interface for monitoring the status of appliances.

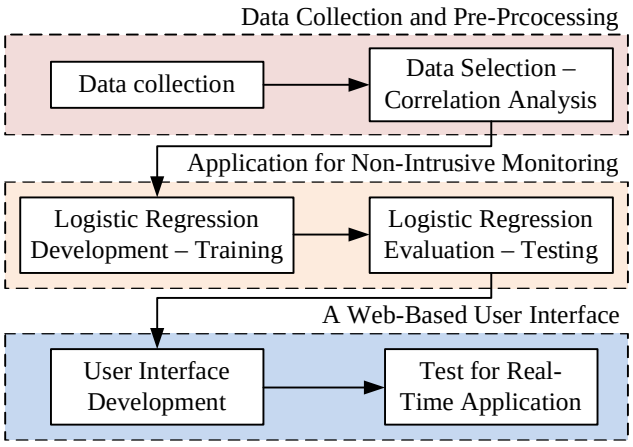

Figure 2: A Block Diagram of Web-Based Non-Intrusive Monitoring System Development

3.1 Data Collection and Pre-Processing

A dataset of the energy consumption for home appliances has been obtained as suggested in [19]. The dataset consists of date, time, global active power, global reactive power, voltage, global intensity and the power consumption for four years from December 2006 until November 2010. There are three measurements from sub-meters that can be used to represent home appliances usage. Reading from sub-meters 1, 2 and 3 represent microwave oven, washing machine and air conditioner, respectively. From observation, the global active power, global reactive power and global intensity are not corresponding to any of the three sub-meters. In other words, the data also consist of other appliances which are unknown. Hence, a net real power, power factor and current have been calculated based on the data provided so that the data corresponding solely to the three sub-meters.

A heatmap is produced to check the correlation between inputs and outputs so that suitable inputs can be selected for the training process of logistic regression as shown in Figure 3. In the heatmap, it shows a range of -1 to 1. If the scales show near to 0 that means, there is very low or no correlation between the parameters and if the scale is near to 1 it means 100% correlation between the parameters. Negative scales show that the relationship between parameters is inversely proportional. Real power and current will be used as input as they show close relationship with the sub-meters which can observed from power and current that the corresponding relationship scales are more than 0.5 towards the three sub-meters (as highlighted with green boxes). Although the scale for power factor and voltage portrays a very low scale correspondingly in the heatmap, they are still taken into consideration as they are one of the few contributions in energy consumption.

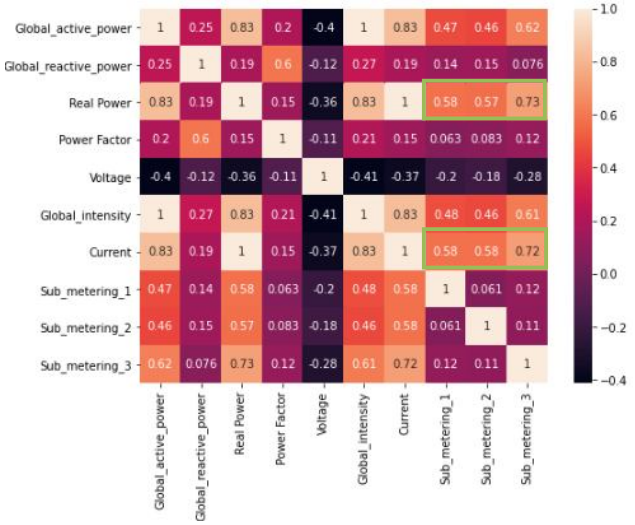

Figure 3: Heatmap of Correlation between Parameters

Apart from selection of suitable input parameters, appropriate range of the input data is also important factor to give more accurate result. From the collected data, the value of power factor is in the range of 0.9 to 1.0 whereas value for voltage is around 240. This will affect the performance of logistic regression if they are not treated well. Therefore, a function of *StandardScaler*() is used to normalized the input data. As a result, the range of inputs will be fixed in a certain range based on the following expression [20]:

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

where, μ and σ are mean and standard deviation, respectively. For the output, the value of electrical consumption for each sub-meter needs to be expressed in binary form to represent the status of appliances (on or off). This can be done by using the following expression:

$$A_i = \begin{cases} 1 & \text{, if } P_i > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{6}$$

where, A_i is status of appliance *i*. A_i is set to 1 to indicate that the appliance *i* is switched when its power consumption, *Pi* shows a reading (> 0) or otherwise 0 to indicate switch off.

3.2 Logistic Regression Application

In order to develop NILM approach based on logistic regression, 85% of the data are used for training and the rest to evaluate its performance as a test data. Based on equations (2)-(4), the logistic regression is designed for one output problem. Therefore, dataset from each sub-meter has been trained and tested separately using logistic regression algorithm. If they are trained and tested together in one sort instead of individually, it will diminish the performance of logistic regression model.

The distributions of zeros and ones in the expected output dataset are imbalance. It is very common to have imbalanced dataset when performing machine learning [21]. There are situations where some group having more data points which is known as majority class while some group having less data point which is referred as minority class. However, this scenario can be solved by tuning the hyperparameter of *class_weight* in logistic regression. If *class_weight* is not assigned, it will be set as one by default [20]. Therefore, the *class_weight* parameter has been adjusted in according to the dataset using expression in (7). The parameter settings are tabulated in Table 1. With an appropriate adjustment of *class_weight* parameter, the probability in making wrong decision could be reduced.

$$class\_weight = \frac{N_s}{N_c + N_b(A)} \tag{6}$$

where,

N_s = number of samples
N_c = number of classes
$N_b(A)$ = number of binary counts in output

Table 1: Logistic regression parameter settings

Sub-Meter	Class_Weight_0	Class_Weight_1
1	0.54	6.09
2	0.70	1.75
3	0.93	1.08

3.3 A Web-Based User Interface

A web-based user interface is then developed to make the output of NILM more accessible to the users. The user interface platform is developed within Flask environment using Python code. Flask is a micro web framework and it is installable from the Python Package Index (PPI). The trained logistic regression model in form of Python code is embedded into the user interface platform to provide the NILM function. Apart from Python, interaction between HTML and JavaScript is required to establish a web user interface and give more accessible to the users.

4. RESULTS AND DISCUSSION

The training of logistic regression has been carried out on a PC with processor of 2.2 GHz and 13 GB RAM. Performances of the training for individual sub-meters are tabulated in Table 2. Since the size of data is the same for all three sub-meters, they took almost similar computational time (35 seconds) to execute the logistic regression model. However, there are significant differences in terms of accuracy. Sub-meter 3 has shown the highest training accuracy at 89.19% followed by sub-meters 2 (86.4%) and 1 (83.99%). This clearly shows that usage behavior from sub-meter 1 is harder to capture as compared to sub-meters 2 and 3. Although the training accuracies are different, very small accuracy reductions can be observed in test performance where they only deviate around 0.01-0.04% as compared to train performance. This indicates that all three models are well trained.

Table 2: Accuracy of each sub-meter

Sub-Meter	Execution Time	Training Accuracy	Testing Accuracy
1	34.25s	83.99%	83.97%
2	34.57s	86.40%	86.39%
3	34.78s	89.19%	89.15%

Table 3 shows data obtained from confusion matrix to further consolidate the findings in Table 2. Confusion matrix is often used to describe the performance of classification model on the test data for which they are true [22]. The confusion matrix shows the way a logistic regression model is confused in making predictions and gives an insight of errors being made. Three important parameters in confusion matrix are used to evaluate the logistic regression model such as precision (*Pc*), recall (*Rc*) and F1-Score as used in [23-24].

According to Table 3, predictions that say appliance is off give high precision (above 0.87) for all sub-meters. However, predictions that say appliance is on for sub-meter 1 is relatively low as compared to other sub-meters. This is mainly due to appliance at sub-meter 1 is rarely used so that it becomes hard to correctly predict when the appliance is switched on. Apart from precision, recall is used to show the proportion of actual positives is correctly retrieved. Since the *class_weight* parameter is adjusted accordingly, the predictions from all sub-meters show a good proportion (above 0.82) in any conditions. There is situation where one has better precision and the other has better recall. Therefore, F1-Score is used to give relative performance of the prediction model. Since F1-Score will be always nearer to the smaller value of precision or recall, it portrays lowest F1-Score in prediction of appliance is switched on from sub-meter 1 as it has low precision. As a result, sub-meter 1 gives the lowest accuracy as discussed earlier.

Table 3: Confusion matrix for each sub-meter

Sub-meter 1			
Status	**Precision**	**Recall**	**F1-Score**
On	0.98	0.84	0.91
Off	0.32	0.82	0.46
Sub-meter 2			
Status	**Precision**	**Recall**	**F1-Score**
On	0.96	0.85	0.90
Off	0.70	0.91	0.79
Sub-meter 3			
Status	**Precision**	**Recall**	**F1-Score**
On	0.87	0.94	0.90
Off	0.93	0.83	0.88

A comparison of overall performance between the proposed individual model for each sub-meter and universal model (all sub-meters in one sort) is presented in Table 4. In order to make a fair comparison, solution given by the proposed individual model is only counted as correct when all three sub-meters are predicted correctly. This can be seen clearly that the proposed model performs better than universal model in which it shows the overall accuracy increased more than 20%. This is because each sub-meter has unique feature of appliance usage behavior that causes conflicts when they are modelled together. Therefore, it is more accurate when they are modelled separately as proven in this work.

Table 4: Overall accuracy of individual and universal logistic regression models

Model	Training (%)	Testing (%)
Individual	70.51	70.50
Universal	50.68	50.05

Figure 4 shows a web-based user interface from the smart meter measurement. Although there are "Start" and "Cancel" buttons displayed on the homepage, the buttons are used to replicate the real-time application where data starts to be fed into the platform and close the homepage when required. Once "start" button is pressed, a graph of energy consumption will be shown in a dynamic graph whereby it will update the total energy consumption. A section underneath the "Start" and "Cancel" buttons gives a live update of measurement taken from smart meter. In the next section below the graph, three pictures of appliances with a status box beside them to indicate their operating conditions either on or off. This is where the NILM based on logistic regression is applied. The status box will turn green showing that the appliance is on or turn red if it is off. In the figure, all status boxes are red because all appliances are currently off and total consumption shows 0 W. This interface is useful to alert users for their home appliances usage anywhere at any time.

5. CONCLUSION

This paper presents a non-intrusive load monitoring approach based on logistic regression to identify status of home appliances from measurement at the main entrance point. This helps to reduce the number of sub-meters to be clamped on appliances and thereby reduce installation cost and complexity of the entire monitoring system. To achieve this, usage behavior of each appliance is modelled separately. The results clearly show that the proposed combination of individual logistic regression models gives better overall performance as compared to a universal model. In addition, a web-based user interface that is integrated with the non-intrusive load monitoring based on logistic regression technique is developed to provide enough information to users at one's fingertips. This will help them to make decision and use electricity wisely.

ACKNOWLEDGEMENT

Publication for this work was sponsored by Universiti Kebangsaan Malaysia under grant project GUP-2018-024.

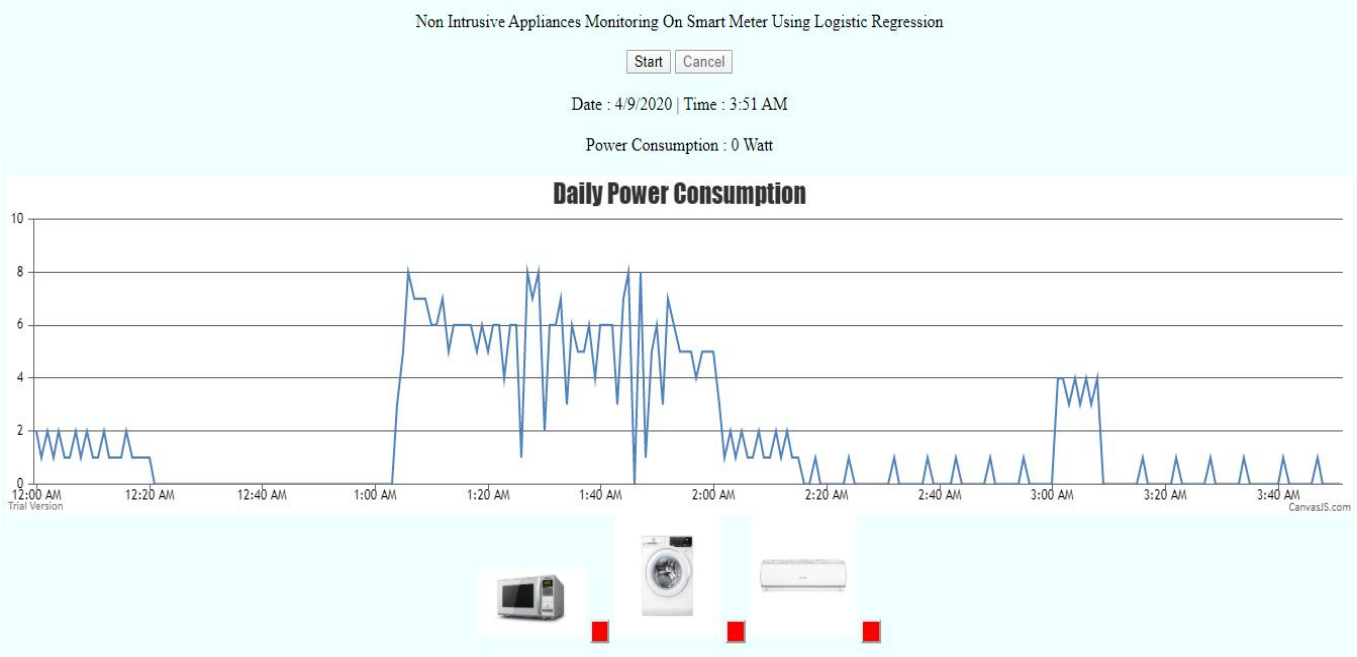

Figure 4: A web-based user interface platform

REFERENCES

1. R. Saidur. **Energy consumption, energy savings, and emission analysis in Malaysian office buildings**, *Energy Policy*, vol. 37, no. 10, pp. 4104–4113, Oct. 2009.
2. A. F. A. Kadir, A. Mohamed, H. Shareef, A. A. Ibrahim, T. Khatib, and W. Elmenreich. **An improved gravitational search algorithm for optimal placement and sizing of renewable distributed generation units in a distribution system for power quality enhancement**, *Journal of Renewable and Sustainable Energy*, vol. 6, no. 3, pp. 033112, May 2014.
3. A. A. Ibrahim, N. A. M. Kamari, and M. A. A. M. Zainuri. **Optimal scheduling of plug-in hybrid electric vehicles operation in distribution networks using gravitational search algorithm**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 1.6, pp. 219-224, 2019. https://doi.org/10.30534/ijatcse/2019/3381.62019
4. I. Vassileva, F. Wallin, and E. Dahlquist. **Understanding energy consumption behavior for future demand response strategy development**, *Energy*, vol. 46, no. 1, pp. 94-100, Oct. 2012.
5. A. A. Ibrahim, A. Mohamed, and H. Shareef. **Optimal placement of power quality monitors in distribution systems using the topological monitor reach area**, in *Proc. 2011 IEEE International Electric Machines & Drives Conference (IEMDC)*, Niagara Falls, ON, 2011, pp. 394-399.
6. A. A. Ibrahim, A. Mohamed, H. Shareef, and S. P. Ghoshal. **A new approach for optimal power quality monitor placement in power system considering system topology**, *Przeglad Elektrotechniczny*, vol. 88, no. 9a, pp. 272-276, Sept. 2012.
7. L. K. Norford, and S. B. Leeb. **Non-intrusive electrical load monitoring in commercial buildings based on steady-state and transient load-detection algorithms**, *Energy and Buildings*, vol. 24, no. 1, pp. 51-64, 1996.
8. H. Shareef, A. Mohamed, and A. A. Ibrahim. **Identification of voltage sag source location using S and TT transformed disturbance power**, *Journal of Central South University*, vol. 20, pp. 83–97, Jan. 2013.
9. W. L. Ai, H. Shareef, and A. A. Ibrahim. **A single monitor method for voltage sag source location using Hilbert Huang transform**, in *Proc. 2012 10th International Power & Energy Conference (IPEC)*, Ho Chi Minh City, 2012, pp. 374-379.
10. I. El Naqa and M. J. Murphy. **What Is Machine Learning?**, in *Machine Learning in Radiation Oncology*, I. El Naqa, R. Li and M. J. Murphy, Springer, Cham, 2015, pp. 3-11.
11. K. K. Dobbin, and R. M. Simon. **Optimally splitting cases for training and testing high dimensional classifiers**, *BMC Medical Genomics*, vol. 4, no. 31, pp. 1-8, Apr. 2011.
12. V. Gholami, K. W. Chau, F. Fadaee, J. Torkaman, and A. Ghaffari, A. **Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers**, *Journal of Hydrology*, vol. 529, no. 3, pp. 1060-1069, Oct. 2015. https://doi.org/10.1016/j.jhydrol.2015.09.028
13. H. Hormozi, E. Hormozi, and H. R. Nohooji. **The classification of the applicable machine learning methods in robot manipulators**, *International Journal of Machine Learning and Computing*, vol. 2, no. 5, pp. 560–563, Jan. 2012.
14. M. Iqbal, and Z. Yan. **Supervised machine learning approaches: A survey**, *International Journal on Soft Computing*, vol. 5, no. 3, pp. 946–952, Apr. 2015.
15. Q. Ye, Z. Zhang, and R. Law. **Sentiment classification of online reviews to travel destinations by supervised machine learning approaches**, *Expert Systems with Applications*, vol. 36, no. 3, Part 2, pp. 6527–6535, Apr. 2009.
16. A. Dhankhar, K. Solanki, A. Rathee, and Ashish. **Predicting student's performance by using classification methods**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, pp. 1532-1536, 2019. https://doi.org/10.30534/ijatcse/2019/75842019
17. A. F. Cabrera. **Logistic regression analysis in higher education: An applied perspective**, in *Higher Education: Handbook of Theory and Research*, J. C. Smart, New York: Agathon Press, 1994, pp. 225-256.
18. C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll. **An introduction to logistic regression analysis and reporting**, *Journal of Educational Research*, vol. 96, no. 1, pp. 3–14, 2002.
19. A. Parate, and S. Bhoite. **Individual household electric power consumption forecasting using machine learning algorithms**, *International Journal of Computer Applications Technology and Research*, vol. 8, no. 9, pp. 371-374, 2019. https://doi.org/10.7753/IJCATR0809.1007
20. G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller. **Scikit-learn: machine learning without learning the machinery**, *GetMobile: Mobile Computing and Communications*, vol. 19, no. 1, pp. 29-33, Jan. 2015.
21. D. Ramyachitra, and P. Manikandan. **Imbalanced dataset classification and solutions: A review**, *International Journal of Computing and Business Research*, vol. 5, no. 4, pp. 1-29, July 2014.
22. X. Deng, Q. Liu, Y. Deng, and S. Mahadevan. **An improved method to construct basic probability assignment based on the confusion matrix for classification problem**, *Information Sciences*, vol. 340–341, no. 1, pp. 250–261, May 2016. https://doi.org/10.1016/j.ins.2016.01.033
23. D. Bowes, T. Hall, and D. Gray. **Comparing the performance of fault prediction models which report multiple performance measures: Recomputing the**

confusion matrix, In *Proc. the 8th International Conference on Predictive Models in Software Engineering*, Lund, Sweden, 2012, pp. 109–118.

24. R. B. Jadhav, S. D. Joshi, U. G. Thorat, and A. S. Joshi. **A software defect learning and analysis utilizing regression method for quality software development**, *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8, no. 4, pp. 1275-1282, 2019. https://doi.org/10.30534/ijatcse/2019/38842019