Power outage prediction by using logistic regression and decision tree

Power outage prediction by using logistic regression and decision tree

Alia Yasmin Nor Saidi[1], Nor Azuana Ramli[2], Noryanti Muhammad[3], Lilik Jamilatul Awalin[4]

[1] Electrical Engineering Section, Universiti Kuala Lumpur British Malaysian Institute, 53100 Gombak, Selangor, Malaysia.
[2,3] Centre for Mathematical Sciences, College of Computing & Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang
[4] Sekolah Teknologi Maju dan Multidisiplin, Universitas Airlangga, Surabaya, Indonesia

[2]Email: azuana@ump.edu.my

Abstract. The occurrence of the power outage caused inconvenience to the customers including the energy suppliers. There are various factors that can trigger the power outage such as lightning, weather or animal. In this paper, the power outage prediction has been performed by using the datasets provided which are lightning data and tripping report. The machine learning method was carried out to predict the power outage occurrence by using the Classification Learner App in MATLAB. Before performing the machine learning method, the data went through the data pre-processing to ensure the data is clean and the significant feature for prediction can be selected to run in the Classification Learner App. The results of this research have shown that Fine Tree is the most suitable model to be used for the prediction of power outage. The results have been compared by using the Area Under Curve (AUC) in Receiving Operating Characteristic (ROC). Logistic Regression and Coarse Tree shows the lowest value of AUC compared to other model and Fine Tree has the highest value of AUC.

Keywords. Machine learning; power outage; prediction; MATLAB

1. Introduction

Power system is a complex and huge interconnected system that delivers electricity in a safe and good condition [1]. Most of the infrastructural system functionality depends on the reliability of the power grid [2].The occurrence of interruption in a system whether in a transmission line or distribution line defined as a power outage can cause bad effect to customers. These occurrence of power outages come from various factors such as lightning, animals or weather (for example extreme wind). The extreme weather event is one of the biggest factors that causes the power outage. A lot of countries in this world are suffering from this type of event. If power outage occurred, it will not only be affecting one area, but it could also affect a huge area where the time needed to restore can takes up a few days. This will give a huge impact as in this present day, power system is known as a backbone to all operations such as safety, security, health and welfare, and economy in a country [2].

Hence, to solve this problem, research have been done and various methods have been proposed to be used in predicting the power outage. One of the most used method is machine learning such as support

vector machine, artificial neural network, decision tree, logistic regression, and others. The findings from these types of research usually involving which method are the best or most accurate to be applied in predicting power outage. However, the accuracy of machine learning model is differ based on the "no free lunch" theorem. Machine learning model not just differ for every problem but for every type of dataset too. Therefore, this study was conducted to obtain a suitable model that suit our case study which is Malaysia.

Machine learning method alone is not enough to obtain the good results. Data handling is one of the important elements too especially when the research involving high volume of data. One of the previous studies proposed logistic regression and the results obtained showed that the complexity of the decision can be handled efficiently but it still need a lot more data to achieve stability [3]. In this research, datasets obtained have high volume which means big data analytics need to be used to process the huge data to come out with a good prediction of power outage. Although, the big data analytics part has already been discussed in [4] as this paper is the continuation of the study and the focus is more to build the predictive model.

The aim of this paper is to obtain the best method to predict the power outage. By using the data of lightning, weather and tripping report, machine learning method is proposed in predicting the power outage in Malaysia. The rest of paper is organized as follows, where the next section will be discussed on history of power outage in Malaysia. Then, proposed methodology will be briefly explained before discussion on all the results obtained. Lastly, conclusion on the research will be made and a few recommendations will be listed for future study that can be done to improve this research.

2. History of power outages in Malaysia

Power outage take place where there is any interruption in the electrical system and all consumers that stays within the area of the power outage. The history of major power outages has been listed in The Star newspaper dated 14 January 2005. On 29 June 1985, the East Coast of Malaysia had a blackout in 11 states that was caused by a trip at the transmission line. While in 31st of July in 1992, the West Coast of Malaysia had a power failure and tripped causes by the lightning that includes 15 power stations and time needed to restore the power almost 10 hours. Next, in 1996, a huge power outage that involves a whole area of Peninsular Malaysia caused by the cascading effect. All these incidents showed the importance of predicting the power outage before it happens.

Based on the research [5], the power outage could happen if there is a vegetation or tree encroachment in the specific area. The main reason Malaysia power outage caused by vegetation is that Malaysia is a tropical country where the tree is growing fast under the transmission line and it cannot be avoided. The research in [6] showed that 70% of power outage in Malaysia caused by lightning. Due to this causes, Malaysia has a high number of lightning incident that lead to death and severe damage to the property. The type of lightning that always occur is intra-cloud lightning and the lightning that affect the overhead distribution is cloud-to-ground lightning.

3. Proposed methodology

This research started with the data collection and data pre-processing. The data were obtained from the energy supplier company and the datasets consists of lightning data, weather data, and tripping report. For the lightning data, all the dataset were in txt file and one file consists of 2 days data. There are 365 txt files collected just for lightning data and all of these data were combined through command prompt. For the weather data and tripping report, both datasets were in Excel files. Although the datasets for weather data consists of 4 years data, it is still not consider as a big data since the datasets can fit in Excel file. Only lightning dataset has a high-volume data. Hence, the big data analytics was applied to pre-process the lightning data. The details of data pre-processing were already explained in the previous study [4].

In this study, our research methodology started with data analysis. The analysis including analysis of lightning criteria, weather data, and the correlation with the tripping report. The correlation analysis is different with the previous study in [4] as in this study, the correlation is to see the relationship

between lightning and power outage. Then, this research continued with building the predictive model by using a supervised machine learning methodology. The predictive model applied in this study were Logistic Regression (LR) and Decision Tree (DT). LR was proposed in this study since it can predict the categorical of dependent variable which the binary variable that contain any data as yes and no or 1 and 0. The LR model predict the equation of P(Y=1) as the function of *X*. The DT was also proposed due to the same reason as this method provides the decisions or possible event outcomes by using the "if then, else" construction. In other words, the problems were categorized until it determined the last category. Before the data was used to train the model, data partitioning was done to separate the data into a training group and a testing group. This process is important to prevent overfitting problems. Finally, the performance evaluation was done to find the best model for each predictive model. The models were compared, and the most accurate predictive model was chosen. Figure 1 below shows all the processes involved in this study.

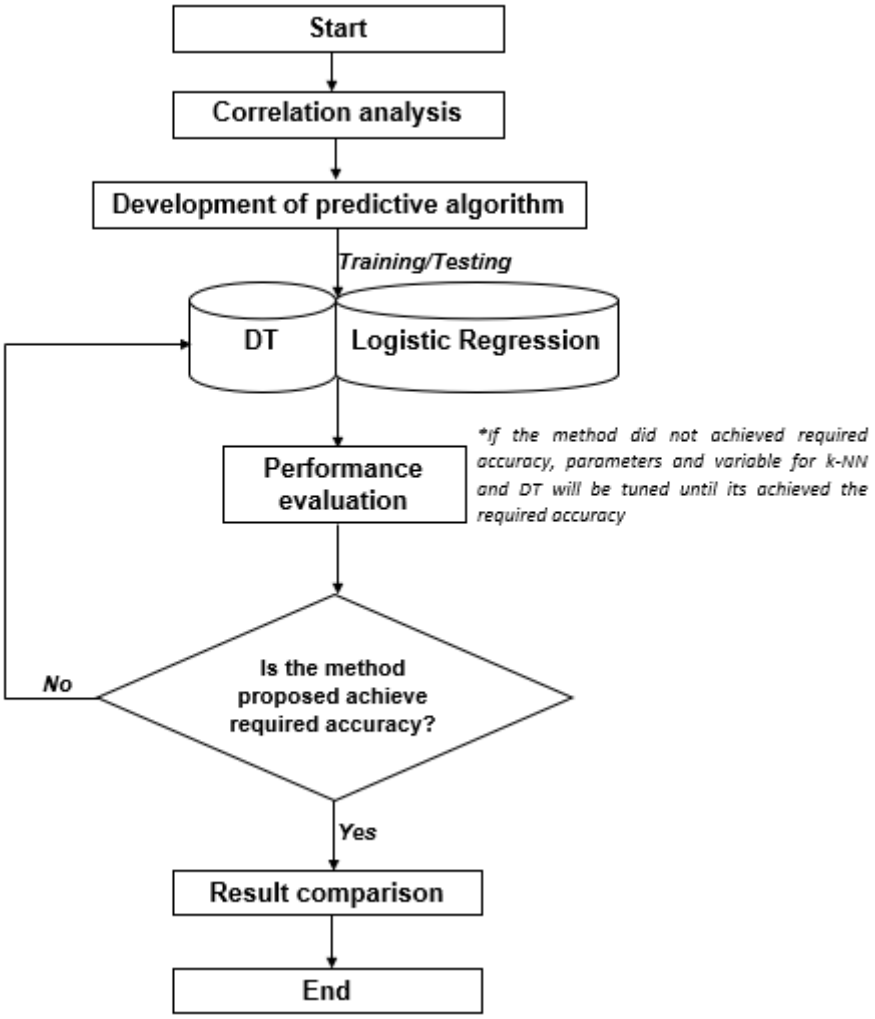

Figure 1. Flowchart of the power outage prediction by using machine learning.

3.1. Correlation analysis between lightning and power outage

The most frequent reason of the power outage occurrences is lightning. If there are any power outage happen in the electrical system utility, the lightning will be most likely to be analysed first before other causes. Thus, the analysis of both lightning and power outage can be utilized to predict the lightning state and the outage that may happen in the electrical system utility. Pearson's Coefficient Correlation was used in this study where the formula is as below:

$$r = \frac{n(\sum ab) - (\sum a)(\sum b)}{\sqrt{[\sum a^2 - (\sum a)^2][n\sum b^2 - (\sum b)^2]}} \quad (1)$$

This formula was used to see the two-variable relationship where the correlation coefficient value, r must be in range between -1 to +1. The value that is closer to +1 means it has the strong positive correlation and the value closer to -1 indicates the strong negative correlation.

3.2. Machine learning

Machine Learning (ML) is a program or system that can learn and improve itself by analysed any data given whether it is big or small data. Mainly, the techniques used by ML are classification, regression and clustering. There are three types of ML which are supervised learning, unsupervised learning, and reinforcement learning. In these three types of learning, there are a lot of algorithms that has been created or devised to help solve the problem given. In this research, the supervised learning was chosen, and the proposed classifiers were logistic regression and decision tree.

3.2.1. Logistic regression. The logistic regression model is a univariate, but it can be a multivariate technique sometimes. This method is used to model the conditional probability Pr (Y=1|X=x) as a function of x when there is a binary output variable, Y and any unknown parameters in the functions are to be estimated by maximum likelihood. In this study, the power outages were made as dependent variable where the outcome is Y=0 when there is no outage and Y=1 when there is an outage. Logistic regression finds the relationship between the independent variables and a function of the probability of occurrence and the linear probability model of the occurrence of power outage can be defined as:

$$\mathrm{P}_i = \Pr(Y = 1 \mid X_i) = \beta_1 + \beta_2 X_1 + ... + \beta_n X_i \quad (2)$$

where X_i is the indicators and $\beta_1 + \beta_2 X_1 + \cdots + \beta_n X_i$ is our familiar equation for the regression line. Anderson [7] described this model as an exact description in a wide variety of situations including the first situation when the class-conditional densities are multivariate normal with equal covariance matrices; second situation when multivariate discrete distributions following a loglinear model with equal interaction terms between groups; and last situation when both continuous and categorical variables describe each sample when the previous two situations are combined.

The probability of the occurrence of the power outage also can be written as:

$$\mathrm{P}_i = \frac{e^{\beta_1+\beta_2 X_1+....+\beta_n X_i}}{1+e^{\beta_1+\beta_2 X_1+....+\beta_n X_i}} = \frac{1}{1+e^{-(\beta_1+\beta_2 X_1+....+\beta_n X_i)}} \quad (3)$$

This equation (3) is also known as the cumulative logistic distribution function. An estimation of the problem needs to be created since P_i is nonlinear not only in X but also in β. This means that OLS procedure cannot be used to estimate the parameters. Hence, the probability of there is no power outage:

$$1 - \mathrm{P}_i = \frac{1}{1+e^{\beta_1+\beta_2 X_1+....+\beta_n X_i}} \quad (4)$$

The equation (4) above can be written as,

$$\frac{\mathrm{P}_i}{1-\mathrm{P}_i} = \frac{1+e^{\beta_1+\beta_2 X_1+....+\beta_n X_i}}{1+e^{-(\beta_1+\beta_2 X_1+....+\beta_n X_i)}} \quad (5)$$

$\mathrm{P}_i/(1 - \mathrm{P}_i)$ is the odds in favour of occurrence of a power outage. The odds describe the ratio of the number of occurrences to the number of non-occurrences. Taking the natural log of equation (5) above,

$$\mathrm{L}_i = \log\left[\frac{\mathrm{P}_i}{1-\mathrm{P}_i}\right] = \beta_1 + \beta_2 X_1 + ... + \beta_n X_i, i = 1, 2, ..., n \tag{6}$$

That is, the log-odds or logit transformations is not only linear in *X* but also linear in the parameters. L is called the logit.

3.2.2. Decision tree. Tree-based methods which also known as a decision tree is a multistage decision process and the decision is made in the binary form at each stage. It has been called as a tree-based method since this method has nodes and branches and their nodes are designed as an internal or a terminal node. The difference between internal and terminal node is an internal node can be separated into two children in contrast with a terminal node as it has no children at all. Additionally, a terminal node has a class label associated with it. Tree-based methods are conceptually simple but are known as powerful methods. The most popular tree-based methods are classification and regression tree (CART), multivariate adaptive regression spline (MARS), iterative dichotomizer (ID3) and C4.5. These four methods have been used widely to solve a variety of problems in different fields of study.

In this study, concerned are more regarding classification tree than regression tree since the target outcome is to take the binary values and involving two class problems. In classification tree, a feature vector is presented to the tree in order to use it. The decision will move to the left child when the value of a feature vector is less than certain number and for the opposite, the decision will move to the right child. This process will keep on going until it reaches one of the terminal nodes and its class label is the one that is assigned to the pattern. There are few heuristic methods on how to construct decision tree classifiers. Generally, decision tree classifiers are constructed top to down, but it begins at the root node since its root is at the top and its leaves are at the bottom. The construction usually involves with three steps only which are splitting, determining terminal nodes and finally, assigning class labels to the terminal nodes. Splitting is a step that is needed for users to decide which variables or probably combination of them should be used at a node in order to divide the samples into subgroups, and then also to decide what the threshold on that variable should be.

A classification tree can be constructed by using a labelled dataset, $L = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ where $\boldsymbol{x}_i$ are the lightning data and y_i the corresponding class labels. Let $N(t)$ denote the number of samples of L for which $\boldsymbol{x}_i \in u(t)$ and $N_j(t)$ be the number of samples for which $\boldsymbol{x}_i \in u(t)$ and $y_i = \omega_j(\sum_j N_j(t) = N(t))$, then

$$p(t) = \frac{N(t)}{n} \tag{7}$$

and an estimate of $p(x \in u(t))$ is based on L;

$$p(\omega_j \mid t) = \frac{N_j(t)}{N(t)} \tag{8}$$

and an estimate of $p(y = \omega_j \mid x \in u(t))$ based on L;

$$p_L = \frac{p(t_L)}{p(t)}\, p_R = \frac{p(t_R)}{p(t)} \tag{9}$$

where $t_L = l(t)$, $t_R = r(t)$, as estimates of $p(\boldsymbol{x} = u(t_L) \mid x \in u(t))$ and $p(\boldsymbol{x} = u(t_R) \mid x \in u(t))$ based on L respectively. Label to each node, t can be assigned according to the proportions of samples from each class in $u(t)$ while label ω can be assigned to node t if

$$p(\omega_j \mid t) = \max_i p(\omega_j \mid t) \tag{10}$$

3.2.3. Machine learning with MATLAB. By utilizing the MathWorks MATLAB resources, the predictive analytics was used for data analysis and to develop the power outage predictive model. This study used Classification Learner App (CLA) that is provided in the MATLAB. Based on MATLAB, CLA helps in develop a real-world machine learning application and it helps to achieve the accurate model. It also includes the process of selecting the algorithms, optimize the model parameters, and avoid the overfitting of the model. Figure 2 shows the Classification Learner App in MATLAB.

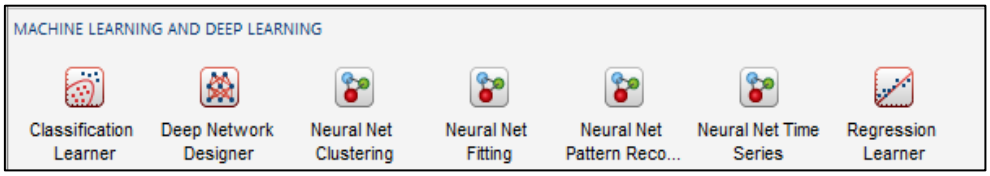

Figure 2. Classification Learner App in MATLAB.

3.3. Training and validating data

The whole prediction dataset of power outage has been partitioned into 70% training and 30% testing. For data validation, 10-fold cross validation technique was used to prove the effectiveness of the prediction technique and it is widely used for machine learning.

3.4. Performance evaluation

Confusion Matrix (CM) was used to evaluate the performance of accuracy and precision of the data prediction. CM formed a table that shows a summary of predicted and actual result on the prediction system by using classifications. The value of predictions data was summarized and classified by each class. In this research, the predicted classes can be positive and negative occurrences and the entries for the confusion matrix have the following meaning. Table 1 shows the CM used for this research in evaluating the performance of prediction data.

Table 1. The confusion matrix to evaluate the performance of prediction data.

		Predicted	
		Negative	Positive
Actual		A	B
		C	D

where A represents the value of current prediction for negative instances, B represents the value of outage prediction for positive instances, C represents the value of outage prediction for negative instances, and D represents the value of current prediction for positive instances.

The accuracy of the data prediction using CM is the quantity of the total number of correct predictions the same can be and is calculated using:

$$Accuracy = \frac{A + D}{A + B + C + D} \tag{11}$$

Besides accuracy, the result for both machine learning methods was compared by using the Receiver Operating Characteristic (ROC) curve. ROC curve shows the classes that has been distinguished by the model. Table 2 below shows the Area Under Curve (AUC) or c-statistic of the ROC curve that indicates which model is better in classifying the data.

Table 2. Interpretation of the Area Under the Curve (AUC) [9]

AUC = 0.5	**No discrimination, e.g., randomly flip a coin**
0.6 ≥ AUC > 0.5	Poor discrimination
0.7 ≥ AUC > 0.6	Acceptable discrimination
0.8 ≥ AUC > 0.7	Excellent discrimination
AUC > 0.9	Outstanding discrimination

4. Results and discussion

All the results that have been obtained in this research are presented in this section.

4.1. Data correlation

For data correlation analysis, the correlation between polarity amplitude and power outage has been investigated. The reason dataset in November was taken for analysis because the power outage data in November was higher and completed without any missing values compared to the other months. The results of correlation were analyzed by using Pearson's correlation coefficient, *r*. To see the strength relationship between both data, the value of coefficient needed to be between -1 to 1. There were two columns selected which are polarity amplitude and power outage. By using Microsoft Excel, the results obtained for *r* value was -0.00275 which indicates a weak negative correlation. This shows that one variable is inversely proportional to the other. In this research, the polarity amplitude did not affect the power outage data. The main reason both columns did not have any correlation in is because the power outage data were not enough to support the analysis. Based on study done in [8], the lightning data that was used to do the correlation with the polarity amplitude was from September 2009 to September 2010 and it was a one-year data. This concluded that the results from this study could be improved if the datasets obtained provide more data without any missing values.

4.2. Prediction analysis

The lightning data and power outage data that has been combined previously were used for prediction by using the Logistic Regression (LR) and Decision Tree (DT) method. The combination of both datasets was uploaded into the MATLAB. In this research, both LR and DT method used a classification technique to predict the power outage for May, October, and November.

4.2.1. Prediction using LR method. The model of LR obtained 100% accuracy due to the limited data of power outage. The training process might not learn the general trend of the data instead it learns the detail of the dataset given. Nevertheless, the result can still be compared through the Receiver Operating Characteristic (ROC) curve. Figure 3 below shows the Receiving Operating Characteristic for Logistic Regression Model.

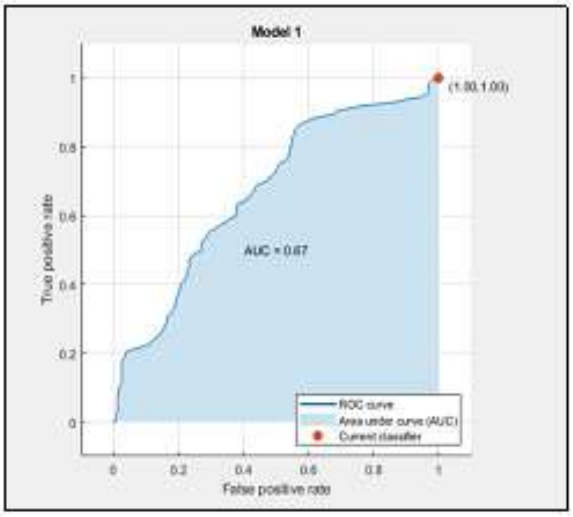

Figure 3. Receiving Operating Characteristic (ROC) curve for Logistic Regression Model.

The AUC for Logistic Regression model falls in poor discrimination. This is because the point (1,1) in the curve showed that even though all the power outage data was classified correctly, the LR model also incorrectly classified all the false data. Constructed on the Confusion Matrix, the true class of 1 and predicted class 0 has the value of 299. In this box, the value of 299 is considered as the occurrence of power outage which is true positive. However, the true class of 0 and the predicted class of 0 has the value of 948813 and this box determined that the false data of power outage has been predicted to be power outage occurrences which is false positive. The point (1,1) that is plotted on the curve is based on True Positive Rate (TPR) and False Positive Rate (FPR). The formula for both TPR and FPR is shown below:

$$\text{True Positive Rate (TPR)} = \frac{299}{299+0} = 1 \tag{12}$$

$$\text{False Positive Rate (FPR)} = \frac{948813}{948813+0} = 1 \tag{13}$$

4.2.2. Prediction using DT method. For Decision Tree (DT) three types of DT model has been trained and the accuracy for all these three models are 100% due to the limited data of power outage. These models were trained with the same data of LR model.

(a) Fine Tree

The AUC for Fine Tree model shown in figure 4 is 0.79 and it falls in a category of acceptable discrimination. In this model, the proportion of correctly classified data of power outage occurrence (true positive) is greater than the proportion of the data that were incorrectly classified as no power outage occurrence (false positive).

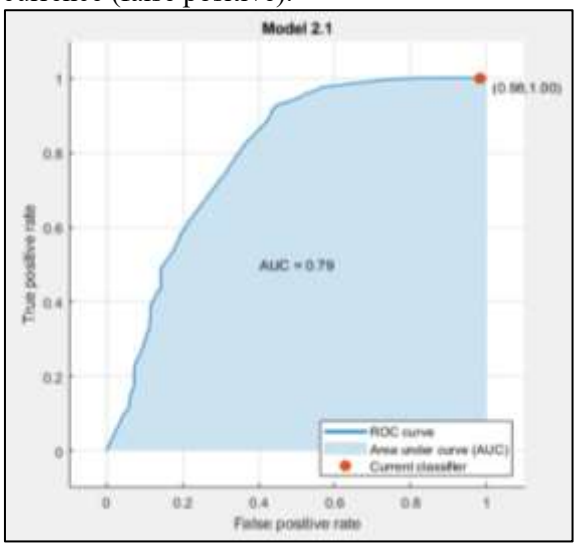

Figure 4. Receiving Operating Characteristic (ROC) curve for Fine Tree

(b) Medium Tree

As for the Medium Tree model (figure 5), the AUC falls into acceptable discrimination with the value 0.73. Even though the point that is plotted in the ROC curve same as the Logistic Regression model, but the Confusion Matrix has a slightly differences in terms of value in each box. In terms of AUC, the value is lower than the Fine Tree model and higher than the Logistic Regression model. Although the proportion of the correct classified data (true positive) is better than the proportion of incorrectly classified data (false positive), the result still cannot outrun the Fine Tree model.

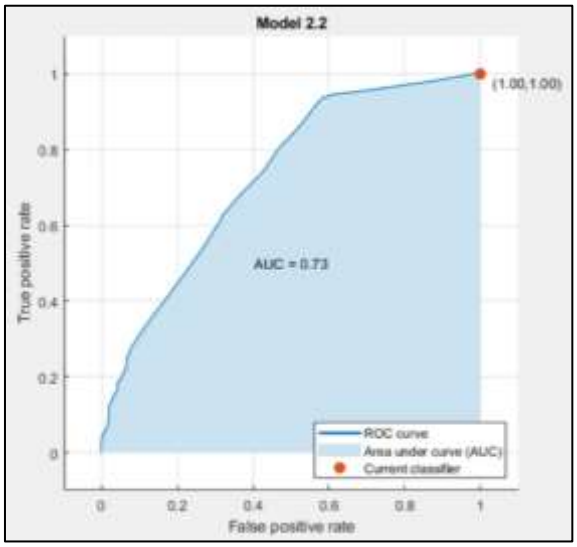

Figure 5. Receiving Operating Characteristic (ROC) curve for Medium Tree

(c) Coarse Tree

Figure 6 shows the results for Coarse Tree model. The result is the same with the Logistic Regression model and there is no difference in terms of value stated in the Confusion Matrix box including the point plotted in the ROC curve. The AUC value for the ROC curve is 0.67 which also shows no difference to the LR model. Both models have the lowest AUC value compared to Fine Tree model and Medium Tree model.

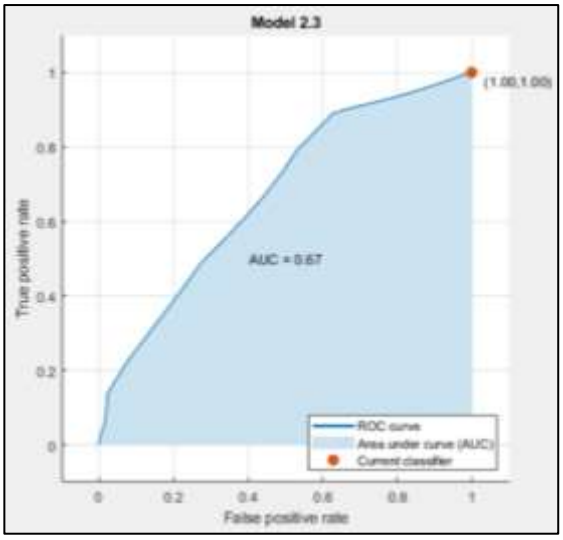

Figure 6. Receiving Operating Characteristic (ROC) curve for Coarse Tree

5. Conclusion

In this research, the analysis of the lightning events has been carried out by only utilized three months datasets which are May, October, and November in 2019. All three months data has gone through data pre-processing method before prediction was made by using Machine Learning method. From this process, a few features in the lightning data that is unnecessary have been take out and some features were sorted following their sequence. The correlation between polarity amplitude and lightning data has been done to know the relationship between both variables. Only one month data was taken for correlation analysis since there were not enough data for power outage in other months datasets. Then, the prediction for the power outage has been done by using Logistic Regression and Decision Tree. Both methods provided 100% accuracy but the ROC curve and AUC for each model showed different results. Among all the model, the best model was the Fine Tree model because the AUC value is the

highest among other which is 0.79. To sum up, the results of correlation and prediction could be improved if there were enough datasets provided for this study. For future study, it is hoped that more data can be obtained to get the better results. Since more data collected will result in high data volume, proper big data analytics with machine learning should be utilized such as Hadoop with Spark, Microsoft Azure Machine Learning Studio or Machine Learning on AWS.

Acknowledgments
This Research was funded by a grant from Ministry of Higher Education of Malaysia (FRGS Grant RDU 1901190) (FRGS/1/2019/STG06/UMP/02/11).

References

[1] Zhang Y, Ilic M D and Tonguz O K 2011 *Proc. IEEE* **99**.
[2] Mensah A F and Dueñas-osorio L 2014 *Int. Conf. Probabilistic Methods Appl. to Power Syst.*(UK: Durham/IEEE) p 1.
[3] Eskandarpour R and Khodaei A 2017 *IEEE Trans. Power Syst.* **32**.
[4] Ali R, Ramli N A and Awalin L J 2020 *Int. Conf. on Data Analytics for Business and Industry: Way Towards a Sustainable Economy*. (Online/IEEE)
[5] Asuhaimi A, Zin M, Member S and Karim S P A 2007 *Tenaga Nasional Berhad Malaysia* **22** (Malaysia) p 2047.
[6] Kadir M Z A A, Misbah N R, Gomes C, Jasni J, Ahmad W F W and Hassan M K 2012 *Int. Conf. Light. Prot.(Vienna, Austria)* p 1
[7] Anderson J A 1982 *Handbook of Statistics* vol 2, ed P R Krishnaiah and L N Kanal (Amsterdam: Elsevier) p 169.
[8] Hassan M K, Rahman R A and Kadir Z 2011 *Journal of Theoretical and Applied Information Technology* **34**(2) p 202-214.
[9] Yang S and Berdine G 2017 *The Southwest Respiratory and Critical Care Chronicles* **5**(19) p 34–36. https://doi.org/10.12746/swrccc.v5i19.391