Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

EVALUATING PREDICTIVE ANALYTICS MODEL PERFORMANCE ACCURACY FOR NETWORK SELECTION MECHANISM

M. I. A. Halim[1], W. Hashim[1,*], A. F. Ismail[2], S. H. Suliman[1], A. S. Yahya[1], R. M. A.Raj[1]

[1]Institute of Informatics & Computing in Energy; Universiti Tenaga Nasional

[2]Dept. of Elec. & Comm. Eng., Intl. Islamic University of Malaysia

Published online: 01 February 2018

ABSTRACT

Predictive analytics has been widely used and adopted in many fields. The idea of anticipating change rather than reacting to change has appealed to many system designers. In this paper, we evaluate the feasibility of applying predictive model into a network selection mechanism to choose most reliable network with higher speed for a communication device such as a modem. The predictive model will attempt to predict best network download speed at a given time of day based on the historical data that we have measured at specific location under studies. This paper also focuses on the accuracy of these predictive models on our sample data, which are measured by calculating its Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) values. Based on the results, a decision tree model outperforms linear regression and M5 models in terms of accuracy. The findings of these studies help us to improve our cognitive network selection algorithm in making decision for best network.

Keywords: Network Selection Mechanism, Predictive Models, Linear Regression, Decision Trees, M5

Author Correspondence, e-mail: Wahidah@uniten.edu.my

doi: http://dx.doi.org/10.4314/jfas.v10i2s.14

1. INTRODUCTION

Machine learning recently has been the frequent buzzword. With the advancement of modern technology thus increasing computing power and efficiency of end devices, machine learning is easily accessible by anyone. Furthermore, with the availability of open source software and libraries such as Tensorflow, Scikit-learn, Pytorch, Rapidminer, Weka and many more, one can easily develop a machine learning model from scratch with minimum practical knowledge. Fundamentally, machine learning is defined as the domain of computational intelligence which is concerned with the question of how to construct computer programs that automatically improve with experience, and what are the fundamental laws that govern all learning processes[1],[2]. In contrast with traditional programming in which each set of rule needs to be defined to reach an outcome, machine learning switches that approach [3]. From the input data that we give to a machine learning model, we demand a certain output to be reached and the machine learns the step to accurately reach that output. However, to get the output accurately, we must also consider the classifier used for the problem. A good rule of thumb is "a good input will produce a good output". This means that the output greatly depends on the input that we provide, thus if we want a good prediction, we also need to supply high quality data which can be translated as the data having meaningful features and significant instances and reduced missing values in the data. In tandem, no machine learning models are unconditionally better than the other as proved by Wolpert and Macready with the no free lunch theorem (NFL)[4], it depends on the type of data it is used on, whether it is continuous or discrete data or if we are solving a regression or classification problem.

2. MACHINE LEARNING MODELS

Generally machine learning models are developed to solve either a regression or classification problem. A regression problem means to predict a future value from a series of continuous data [5]. Whereas, a classification problem is to predict to which class of the data (usually discrete) it belongs to from the input data [6]. Beyond that, it is also a good practice to know whether you are doing a supervised learning approach or unsupervised approach as to cut out unnecessary assumptions in favor of Occam's Razor [[7]]. Reinforced learning approach is also another available alternative. A supervised learning indicates that each of the input data is labelled, which means the machine can distinguish data from different classes [8]. On the other hand,

unsupervised learning is the approach where the input data is unlabeled, the machine will have to learn by itself and label the data itself [9]. As a consequence, supervised learning is always precedent than the unsupervised learning approach. However, raw data can be messy and complex. In most cases, it is benign yet rare to have a perfectly labelled dataset to use. Thus, it is harder to use supervised learning in each case. Consequently, the data used for machine learning models are the utmost prominent. This is in lieu of traditional programming where we have to define each singular step to gain our results. Machine learning is overly reliant on data for the learning process [10]. Simply put, limited, incomplete and insignificant data using a good classifier can give worse results than having extended, thorough and significant data fed into a terrible classifier. This is due to the learning process of the machine learning models which are heavily dependent on the data itself.

In this paper, we study the relationship of various machine learning models with the accuracy of prediction using the same dataset. This paper is organized as follows. Section 2 formally defines predictive modelling and its uses test cases. Section 3 describes various predictive model employed and how it is derived theoretically. Section 4 outlines the methodology for the experiment. Section 5 presents the result. Finally, section 6 concludes with a brief yet concise summary and possible upcoming future works.

3. PREDICTIVE MODEL

3.1 Linear Regression (LR)

Linear Regression is a staple of machine learning algorithm. It is frequently used to forecast continuous data. Linear Regression works by mapping a function f(x) that returns a best fit, with the assumption of R being any real number as shown in equation (1). It also assumes a linear relationship between the independent and dependent variable [11]. The model can also be represented in such a way that the forecast depends on the line of best fit [12]. To reduce the error of the line of best fit, we can use the technique known as least squared estimation. This technique aims to reduce the sum of squared errors [13]. Also, to optimize the result and reduce overfitting of data, we can use regularization, thus further reduce the error of the equation. Amongst the technique that can be used is Least Absolute Shrinkage and Selection Operator (Lasso), Gradient Descent and Ridge Regression for regularization.

$$f(x) = \beta_i + \beta_1 x + \ldots + \beta_a x + \varepsilon \quad (1)$$

for i = 1, … ,a.

where ε is the error derived from the line of best fit

To find the correlation coefficient (r), we can calculate it using equation (2) whether the data is collinearly related or not [9]. If there is no value for the correlation coefficient, it suggests that the method of linear regression is not suitable for the dataset. Figure 1 shows the derivation steps for linear regression technique. It can be observed that the steps are rather simple to simulate and incorporate into an algorithm.

$$R = r = \Sigma [(x_i - \quad)(y_i - \quad)] / \Sigma [(x_i - x)^2] \quad (2)$$

where " " is the accumulated mean value of x and " " is the accumulated mean value of y

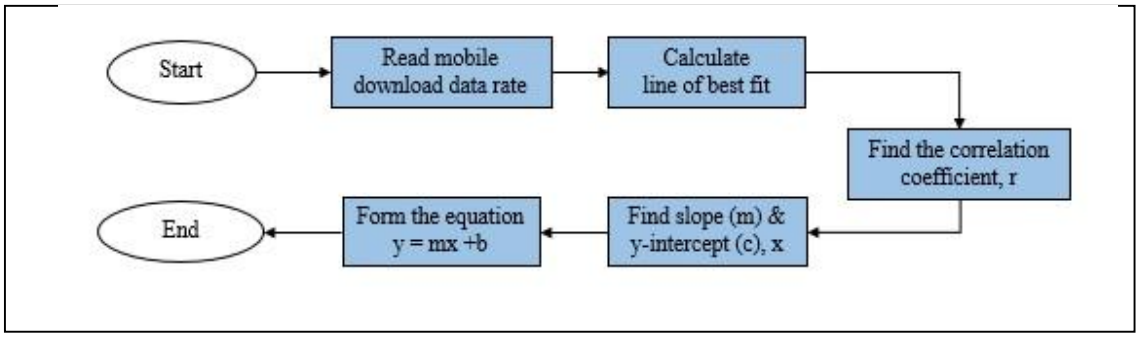

Fig.1. Linear regression derivation steps for network selection mechanism technique

3.2Decision Tree (DT)

Decision tree is basically a classifier in the form of a tree structure. It consist of either leaf nodes which can be used to label the value of the target attribute (class), or it can be a decision node which output on a single attribute-value, with one branch and sub- tree for each possible outcome of the test[14]. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance[15]. Decision tree induction is a typical inductive approach to learn knowledge on classification. Figure 2, depicts the simplified derivation of decision tree applied to our data.

3.3 M5

M5 was introduced back in 1992 by Quinlan [16]. The M5 model primarily utilizes a divide and conquers approach to manifest the relationship between the independent and dependent variables. M5 is suitable to be applied for both classification and regression problems and it also has the capability to process both qualitative and quantitative data [17]. Unlike model trees, this has made the M5 model more preferable because of this flexible capability of data that can be handled. M5 model is akin to piece-wise linear functions as it combines both the linear regression and regression tree concepts [18]. The regression tree approach basically splits the data into subsets, which is also denoted as leaves, thus their relations at the leaves are shown by average numeric values. The linear regression approach formulates a relationship between the independent and dependent variable by fitting a line of best fit for the two variables, thus formulating a linear regression equation at the decision nodes.

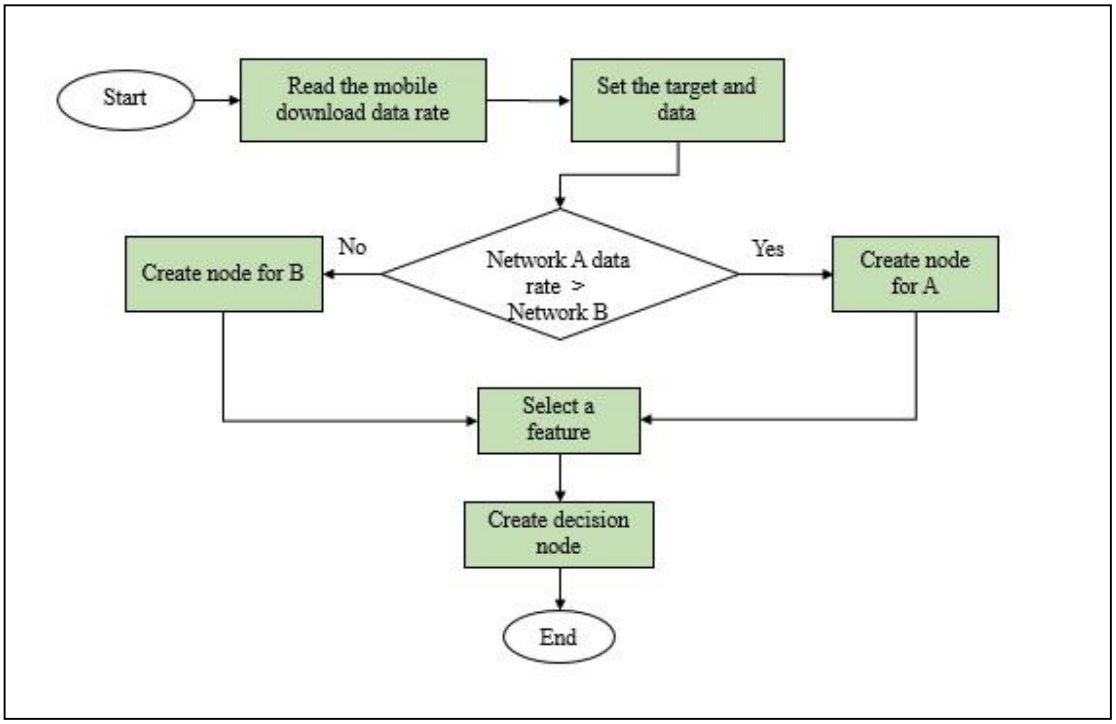

Fig.2. Proposed decision tree derivation steps for network selection mechanism technique

4. METHODOLOGY

Our empirical studies were conducted using various classifiers, namely Linear Regression, Decision Tree, M5P and M5Rules. The classifiers were trained on the identical dataset that we

have collected on our own. Our dataset consists of features such as time of day which is in the format HH:mm:ss, ping in milliseconds (ms), download speed & upload speed in Megabyte per second (Mbps). The network speed was recorded at an interval of every 30 minutes.

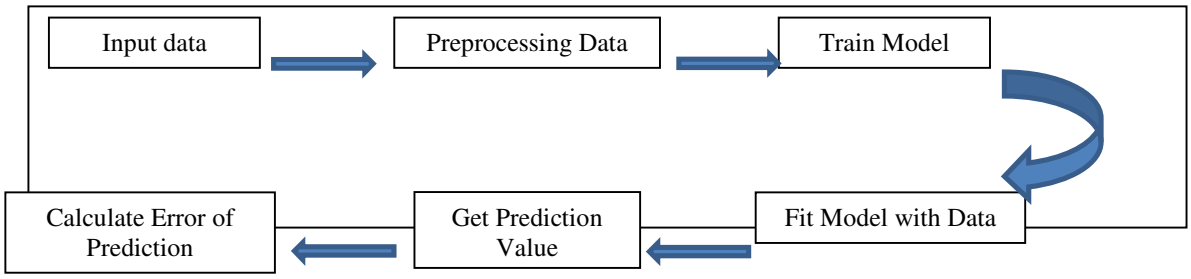

Fig.3. Proposed generalized predictive models

Figure 3 shows the mechanism for our generalized proposed predictive model that has been applied in these studies. The input data is preprocessed by choosing only rows of instances that does not have the feature download speed missing. Beyond that, we also ensured that we split the data into training and test set with a ratio of 70:30. The training set is used as the name implies, to train the model. The trained model will then try to predict the isolated new data point which is the test dataset. In other words, the trained model will try to fit a best line to the dataset, to produce the predicted value. After that, the predicted results need to be validated and error is to be calculated. The lower the error, the closer the predicted values to the actual values.

5. RESULTS & ANALYSIS

The network speed vs time of day was plotted on a graph. The predicted network speed is also displayed on the graph as a comparison. The predicted network speed was obtained by using various predictive models, amongst them are M5, Decision Tree and Linear Regression. Figure 4 shows the graph for the prediction of M5. For the following Figure 4,5 and 6, the vertical axis represents the download speed which was measured in Mbps, whilst the horizontal axis indicates time of day in the format of HH:mm:ss, which was measured at a defined interval of 15 minutes. However, it is apparent that the fluctuation of network speed and predicted network speed throughout the day is quite high and occurs frequently. Consequently, it is a complex process to

model the fluctuations accurately. To describe the fluctuations sufficiently, a Piecewise equation is probably needed since it is not a smooth line. Figure 4 shows the graph for the prediction of M5, whereas Figure 5 displays the graph for the prediction of Linear Regression. Next, Figure 6 indicates the graph for the prediction using Decision Tree.

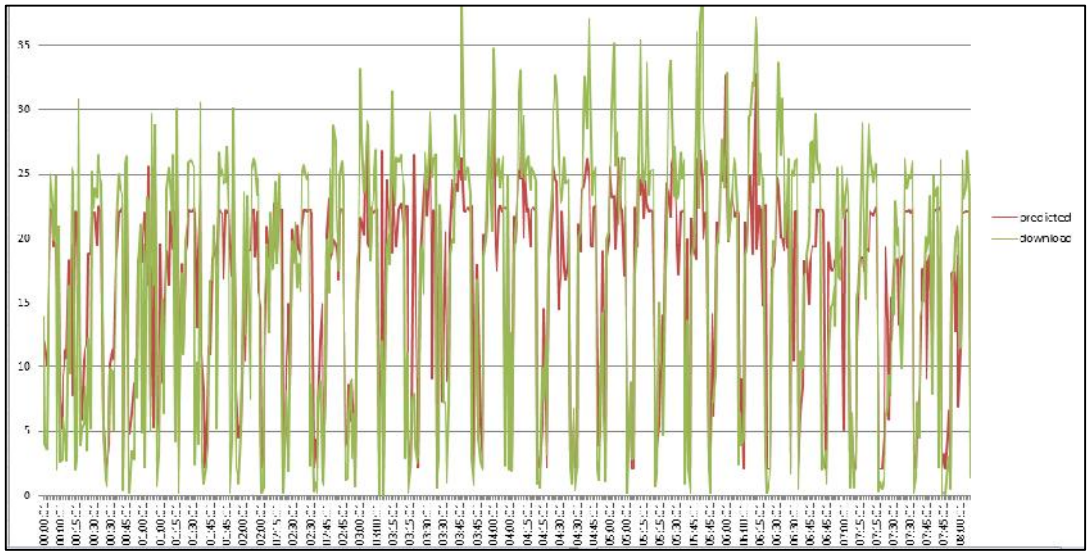

Fig.4. Raw data of M5 predictive results (red) as compared to the actual (green)

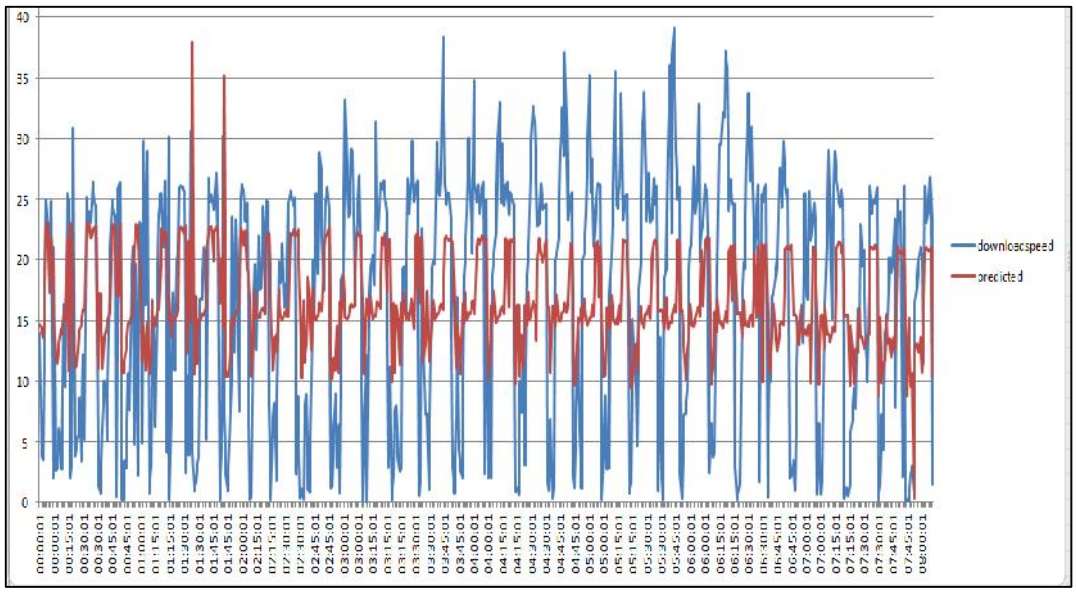

Fig.5. Raw data of linear regression predictive results (red) as compared to the actual (blue)

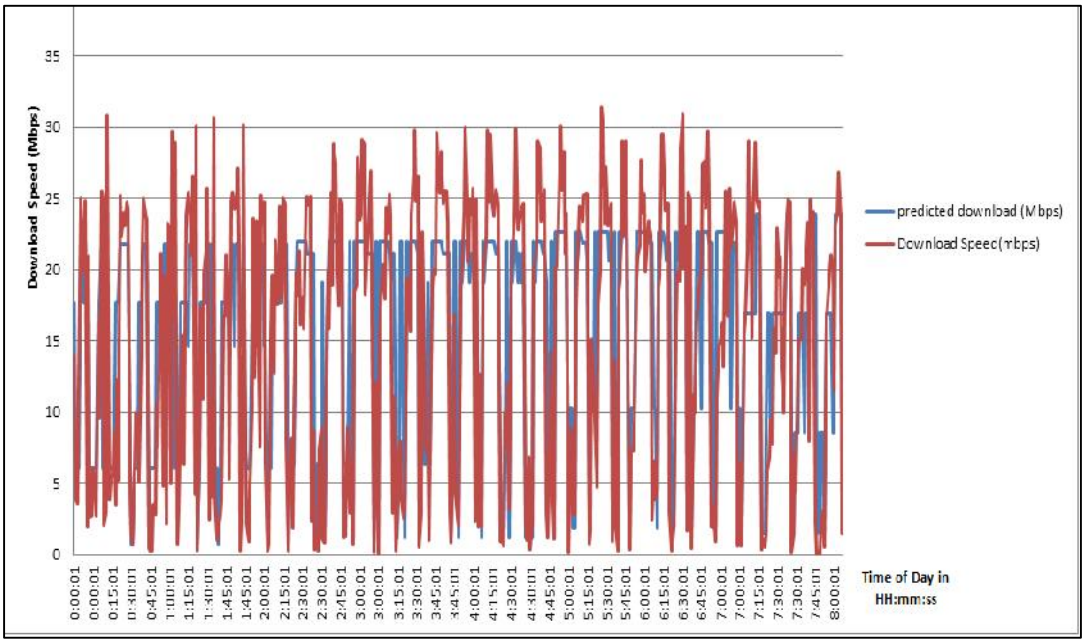

Fig.6. Raw data of decision tree predictive results (blue) as compared to the actual (red)

Table 1. Performance criteria of predictive models

Model	Correlation Coefficient	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)
Linear Regression	0.207	8.9359	10.012
M5	0.2791	8.6393	9.8269
Decision Tree	0.3866	8.0229	9.4379

Table 1 presents the performance criteria for each of the predictive models. Three prevalent criteria are the correlation coefficient, mean absolute error (MAE) and root mean squared error (RMSE). Correlation coefficient value indicates how much of the variance in your data is represented accurately by the predictive model. A low value is not inherently terrible if it truly is the best model available, but a higher value is generally more sought after. Mean absolute error is the average distance the models predictions are from the actual data points. Whereas, absolute in the title indicates that predictions below data points will not be represented as negative values. Root mean squared error is usually calculated by the difference of the predicted and the actual value of the data. It can be intuitively deduced that for this experiment, the decision tree model is the best since it has the lowest MAE and RMSE and also the highest correlation coefficient amongst the three models.

6. CONCLUSION

In this paper, we examined different predictive models to forecast the network download speed ahead of time. For this comparison, it was observed that decision tree has the highest correlation coefficient. This indicates that decision tree is relevant to the problem than other predictive models. However, there are still more work needed to be conducted in this area, since complex predictive models could introduce high processing power in the algorithm.

7. ACKNOWLEDGEMENT

We would like to state our outmost gratitude to the Ministry of Higher Education of Malaysia for sponsoring our project based on Fundamental Research Grant Scheme (FRGS) under grant number FRGS/1/2015/ICT02/UNITEN/02/1.

8. REFERENCES

[1] T. M. Mitchell, *Machine Learning*, vol. 4. 1997.

[2] T. M. Mitchell, “The Discipline of Machine Learning,” *Mach. Learn.*, vol. 17, no. July, pp. 1–7, 2006.

[3] P. Domingos, “A few useful things to know about machine learning,” *Commun. ACM*, vol. 55, no. 10, p. 78, 2012.

[4] D. H. Wolpert and W. G. Macready, “No free lunch theorems for optimization,” *IEEE*

Trans. Evol. Comput., vol. 1, no. 1, pp. 67–82, 1997.

[5] A. Bayati, V. Asghari, K. Nguyen, and M. Cheriet, "Gaussian Process Regression based Traffic Modeling and Prediction in High-Speed Networks," 2016.

[6] S. Ubik and P. Žejdl, "Evaluating application-layer classification using a machine learning technique over different high speed networks," *Proc. - 5th Int. Conf. Syst. Networks Commun. ICSNC 2010*, pp. 387–391, 2010.

[7] T. Lattimore and M. Hutter, "No free lunch versus Occam's Razor in supervised learning," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7070 LNAI, pp. 223–235.

[8] A. Ng, "1. Supervised learning," *Mach. Learn.*, pp. 1–30, 2012.

[9] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *Int. J. Adv. Res. Artif. Intell.*, vol. 2, no. 2, pp. 34–38, 2013.

[10] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015.

[11] K. L. Sainani, "Understanding linear regression," *PM R*, vol. 5, no. 12, pp. 1063–1068, 2013.

[12] N. Altman and M. Krzywinski, "Points of Significance: Simple linear regression," *Nat. Methods*, vol. 12, no. 11, pp. 999–1000, 2015.

[13] K. H. Zou, K. Tuncali, and S. G. Silverman, "Correlation and Simple Linear Regression," *Radiology*, vol. 227, no. 3, pp. 617–628, 2003.

[14] W. N. H. W. Mohamed, M. N. M. Salleh, and A. H. Omar, "A comparative study of Reduced Error Pruning method in decision tree algorithms," in *Proceedings - 2012 IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2012*, 2013, pp. 392–397.

[15] M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2227–2240, 2014.

[16] J. R. Quinlan, "Learning with continuous classes," *Mach. Learn.*, vol. 92, pp. 343–348, 1992.

[17] S. Samadianfard and A. A. Sadraddini, "M5 Model Tree and Gene Expression Programming Based Modeling of Sandy Soil Water Movement under Surface Drip," no. 3, pp.

178–190, 2014.

[18] C. Chinnarasri and P. Ditthakit, “Estimation of Pan Coefficient using M5 Model Tree,” *Am. J. Environ. Sci.*, vol. 8, no. 2, pp. 95–103, 2012.

How to cite this article:
Halim M I A, Hashim W, Ismail A F, Suliman S H, Yahya S A, Raj R M A. Evaluating predictive analytics model performance accuracy for network selection mechanism. J. Fundam. Appl. Sci., 2018, *10(2S), 162-172.*