Tax Avoidance Detection Based on Machine Learning of Malaysian Government-Linked Companies

Rahayu Abdul Rahman, Suraya Masrom, Normah Omar

Abstract— Machine learning has been widely used in solving the problem of prediction and classification. It is also beneficial in the problem of tax avoidance detection. This study presents the utilization of machine learning classification approach for detecting tax avoidance of Malaysian government-linked companies (GLCs). There were nine machine learnings algorithms have been used on the real dataset collected from datastream and companies annual reports. The performance of these algorithms have been observed based on different training approaches and different features selection. The findings have revealed that the accuracy of results from each machine learnings were divergent according to the training approaches and features selection.

Index Terms: Government-linked companies, machine learning, prediction, real dataset, tax avoidance.

I. INTRODUCTION

Corporate tax is a vital source of government revenues especially for developing countries. It is a compulsory levy by the government on companies' earnings to assist government's social and political objectives. Indeed, in Malaysia corporate tax is one of the major contributors to the government revenue. In 2018 direct taxes collected by the Malaysian Inland Revenue Board was RM137.035 billion, where the corporate tax was the biggest contributor at RM70.036 billion or 51.11 percent of the total collection.

However, such taxes represent a significant cost to the companies and a reduction in cash flows available to the shareholders [1]. Given that managers are more likely to engage in tax avoidance and tax evasion activities in order to reduce their tax burden. Tax avoidance is a legal activity consists of various tax planning strategies that are used to minimize tax liability. Tax evasion, on the other hand, is the illegal practice in which companies avoid paying actual tax liability. Although tax avoidance is a legal strategy but it will results in revenue losses to the country [2].

Due to its seriousness and negative impacts to the government and country, many tax evasion detection methods had been developed and implemented to help the regulators such as Malaysian Inland Revenue Boards and auditors to detect or to predict tax fraud. However, very limited studies on the tax avoidance prediction model. Therefore, research that focus specifically on tax avoidance prediction is needed.

In the era of Industrial 4.0, many urgent issues in industries can be effectively solved with big data techniques, including machine learning. Research shows that machine learning has been very useful in many problems of prediction and classification. Nevertheless, the use of machine learnings on tax fraud and tax avoidance prediction is either limited or not adequately documented in literature. Thus, this study attempts to explore the utilization of machine learning classification approach for detecting tax avoidance of Malaysian government-linked companies.

This study makes multifaceted contributions. First, the study expands on the existing body of knowledge by providing evidence on tax avoidance prediction model using machine learnings. Second, this study reveals that selection of parameters with different significant level have slightly affect the performances of machine learning algorithms tested on the real text avoidance dataset. This work is an extension of prior study by [3] on tax avoidance prediction.

In general, most selected algorithms can produce better or similar level accuracy results with better correlation of variables. However, there was no indicator can be concluded that using cross validation split training can affect the algorithms performance.

The remainder of the paper is organized as follows. Section two discusses on government-linked companies in Malaysia and reviews relevant prior research on tax avoidance detection and machine learnings. Section three elaborates the research methods. Section four presents and discusses the findings. The final section provides the summary and conclusions.

II. BACKGROUND OF STUDY

A. Government-Linked Companies

Government linked companies (GLCs) are a strong feature of the Malaysian corporate sector. GLCs are companies controlled by government via its government institutional investors namely government-linked investment companies (GLICs). GLICs consist of Employee Provident Fund (KWSP), Khazanah Nasional Berhad (KNB), Kumpulan Wang Amanah Pencen (KWP), Lembaga Tabung Angkatan Tentera (LTAT), Lembaga Tabung Haji (LTH), Menteri Kewangan Diperbadankan (MOF) and Permodalan Nasional

Revised Version Manuscript Received on September 16, 2019.

Rahayu Abdul Rahman, Faculty of Accountancy, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Malaysia.

Suraya Masrom, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Perak Branch, Tapah Campus, Malaysia.

Normah Omar, Accounting Research Institute, Universiti Teknologi MARA, Selangor, Malaysia.

Retrieval Number: B10830982S1119/2019©BEIESP
DOI: 10.35940/ijrte.B1083.0982S1119

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Berhad (PNB). As Malaysian government owns majority stake of GLCs, the government have a power to direct a firm to contribute to achieving social and political goals by maximizing tax revenues.

Prior studies argue that government shareholdings affect corporate tax avoidance strategies as managers of such firms are appointed and evaluated by the government owners. Thus, they are committed to assist the controlling shareholder (government) to increasing government revenues by maximizing tax payment in order to achieve social and political goals. In addition, some scholars stress that managers with a reputation for paying more taxes can enhance their political capital and increase their chances for promotion in government [4]. This in turn will reduce the degree of tax avoidance in such firms. Besides, state-owned firms also incur political costs such as excessive taxation due to the government's intervention in such firms [5]. Motivated by these arguments, this paper extends previous literature by being the first to build a tax avoidance prediction model which incorporates government ownerships of firms for developing country.

B. Tax avoidance prediction

In [3] is the first and only study on tax avoidance prediction. They developed a tax avoidance prediction model using network characteristics of firms as an input factor. This study applies three machine learning techniques, logistic regression, decision trees and random forests; to create five models using either firm characteristics, network characteristics or different combinations of both. The findings show that network features significantly improve the predictive ability of tax avoidance model. In particular, firms that have board members with no-connections to low-tax firms are less likely to be classified as low-tax.

C. Machine learning classification

Machine learning addresses the issue of how to use computer machine to assist human in planning and making decision with the rapid growing of complex problem of prediction and classification. It is one of today's most rapidly growing technical fields, lying at the intersection of artificial intelligence and statistics and at the core of data science. Currently, the progress in machine learning has been driven both by the introduction of variation types of machine learning algorithms and tools together with the ongoing explosion in the availability of online data and low-cost computation. The adoption of data-intensive machine-learning methods can be found throughout different domains of science [6], [7], technology [8], [9] and business [10].

Ability of machine learning to think like a human is an underlying concept that need the performance measurement before the real implementation. Researchers highlighted that features selection is one of the important factors. Feature selection is the process of identifying and removing as much irrelevant and redundant information as possible. This reduces the dimensionality of the data and may allow learning algorithms to operate faster and more effectively. In some cases, accuracy on future classification can be improved; in others, the result is a more compact, easily interpreted representation of the target concept. Researcher in [11] introduced different approach of features selection and one way is through the correlation relationships.

Evaluating machine learning algorithm enquires dataset to be grouped as training dataset and testing dataset according to separation ratio, which normally configured as 70:30, 80:20. Breaking the dataset into training and testing dataset is called as split training approach. The split training approach can be also classified as normal or cross validation. In cross validation, each model is trained against different combinations of training subsets and validated against the remaining parts of the training subsets. In a certain cases of dataset pattern, using different training approaches seemed to give some performance effect to the machine learning model [12].

III. METHODOLOGY

A. Dataset

The dataset is a collection of tax avoidance firms and non-tax avoidance firms among GLCs for the period 2010 to 2016. This study uses Effective Tax Rates (ETR) to measure tax avoidance, which measured as the ratio of the total tax expenses to the total income before tax. Consistent with prior study [2], GLCs will be classified as tax avoidance firms when the ETR lower than corporate statutory tax rates. Observations with a negative value for tax avoidance are coded as 1, represent as tax avoidance firms. Observation with a positive value for tax avoidance are coded as 0, represent non-tax avoidance firms.

Table 1 shows the set of features to develop the prediction model. This study uses 24 attributes or factors variables as the independent variables for predicting tax avoidance. As this paper use machine learning prediction, these variables are called as features. The features can be classified into four categories; government shareholdings (LTAT, KWSP, KNB, KWP, LTH and PNB), firms specific characteristics (SIZE, LEVERAGE, GROWTH, PROFIT), governance (BODind, BODsize, AUDsize, AUDind, MusCh, MaleCEO, Duality) and sector (Finance, IndustrilProd, Cons, Const, Plant, IPC, TradSER).

Table 1: Features in the dataset

Features	Measurement
LTAT	Total percentage of shareholding by LTAT
KWSP	Total percentage of shareholding by EPF
KNB	Total percentage of shareholding by KNB
KWP	Total percentage of shareholding by KWAP
LTH	Total percentage of shareholding by LTH
PNB	Total percentage of shareholding by PNB
SIZE	Natural log of total assets of firm *f* in year *y*
LEVERAGE	Total liabilities to total assets of firm *f* in year
PROFIT	Earnings (EBIT) to total assets firm *f* in year
GROWTH	Market to book ratio of firm *f* in year *y*
MusCh	1 if firm has Muslim Chairman and 0 otherwise
MaleCEO	1 if firm has Male CEO and 0 otherwise
Duality	1 if chairman and CEO position hold by same person and 0 otherwise
BODind	The proportion of independent directors on the board
BODsize	The number of directors on the board
AUDsize	The number of directors on the audit committee board
AUDind	The proportion of independent directors on the audit committee
Industrial Prod	1 if firm from industrial product sector and 0 otherwise
Finance	1 if firm from financial services sector and 0 otherwise
Cons	1 if firm from consumer product sector and 0 otherwise
IPC	1 if firm from infrastructure project companies sector and 0 otherwise
TradSER	1 if firm from trading/services sector and 0 otherwise
Plant	1 if firm from plantation sector and 0 otherwise
Const	1 if firm from construction sector and 0 otherwise

B. Features selection

Machine learning features selection is an important step of machine learning prediction. In this paper, features selection is divided into two groups. First group or *FeatureSetOne* involves all the independent parameters in the training dataset. Fig. 1 to 7 present the scatter plots with the value of Pearson correlation (*r*) of all independent variables to the dependent variable ETR. Then, only some variables were used for the second group of features selection named as *FeatureSetTwo.*

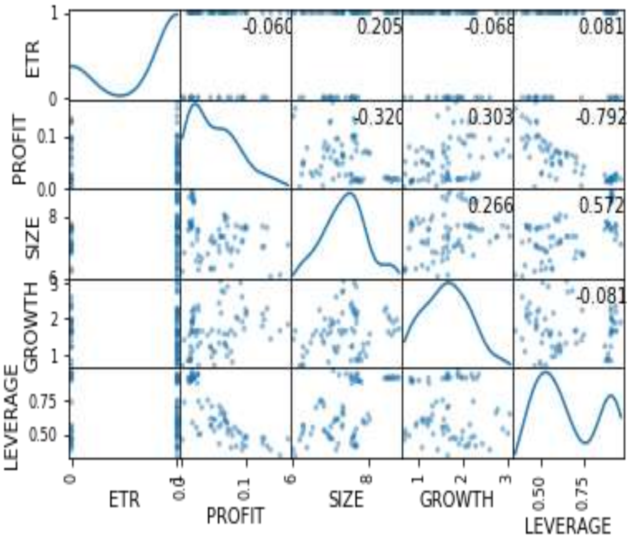

Fig. 1: Scatter Plot with Pearson correlation between ETR and the four independent variables of firms' specific characteristics

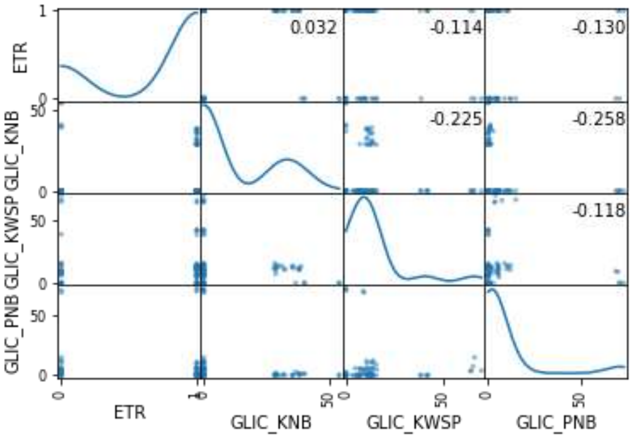

Fig. 2: Scatter Plot with Pearson correlation between ETR and the three independent variables of government shareholdings

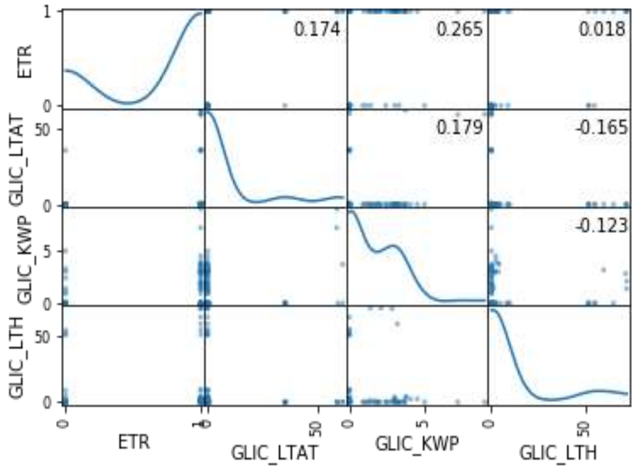

Fig. 3: Scatter Plot with Pearson correlation between ETR and the three independent variables of government shareholdings

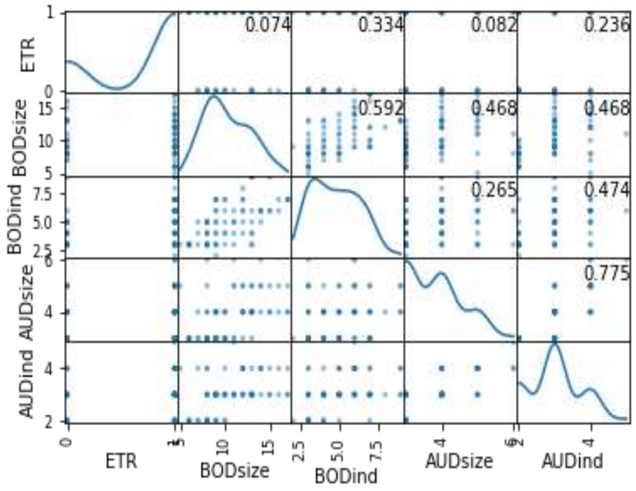

Fig. 4: Scatter Plot with Pearson correlation between ETR and the four independent variables of firms' governance

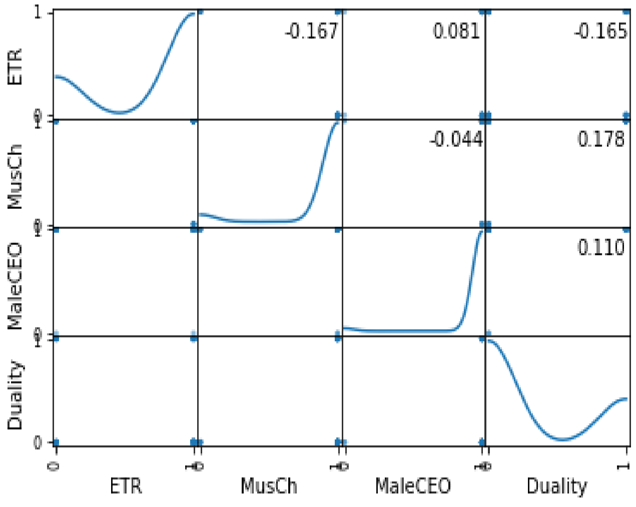

Fig. 5: Scatter Plot with Pearson correlation between ETR and the three independent variables of firms' governance

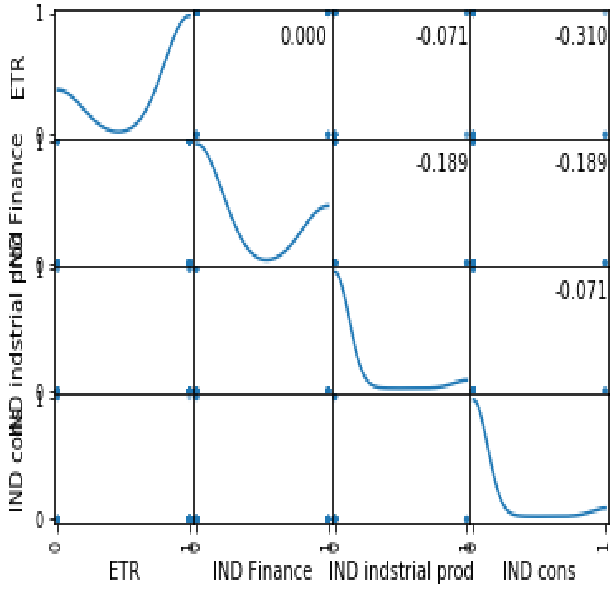

Fig. 6: Scatter Plot with Pearson correlation between ETR and the three independent variables of firms' sector

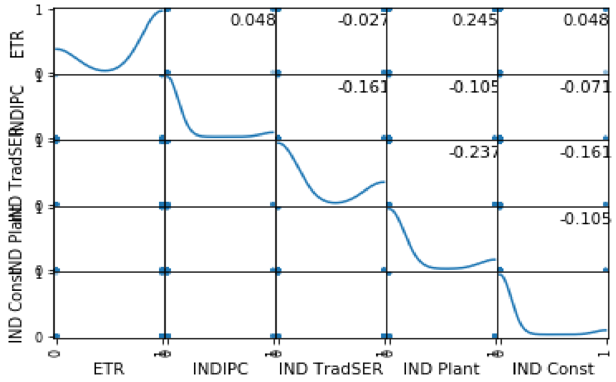

Fig.7: Scatter Plot with Pearson correlation between ETR and the four independent variables firms' sector

As all variables have very weak relationships with the ETR, *FeatureSetTwo* only used the variables that has correlation values between 0.1 to 0.35. The correlation matrix of these variables is presented in the following Python heat map at Fig. 8.

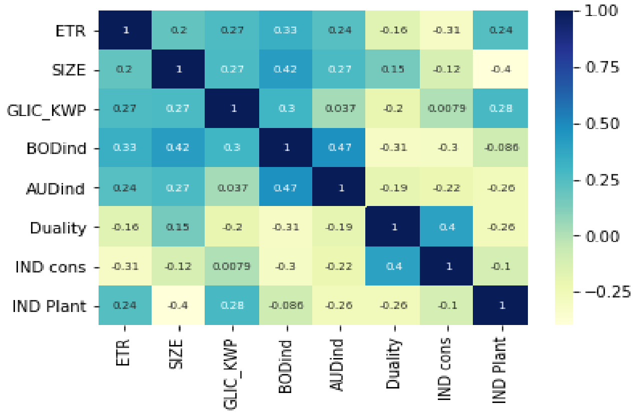

Fig. 8: Correlation matrix of ETR with the seven independent variables from *FeatureSetTwo*

C. Training approach

In this study, split and split with cross validation approaches have been used with ratio 80:20 of training and validation respectively. Cross validation is basically an extension of the basic split methodology. The advantage of it though is that it randomly divides the dataset multiple times, and in each time, it trains the tests the model on a slightly different dataset. In this study, *ShuffleSplit* cross validation technique has been used with 5 split numbers and 50 random state.

D. Performance measurements

In this study, the tax avoidance detection problem is categorized as classification problem, which uses measurements from the confusion matrix below:

N=15	Predicted Negative (No)	Predicted Positive (Yes)
Actual Negative (No)	True Negative (TN)	False Positive (FP)
Actual Positive (Yes)	False Negative (FN)	True Positive (TP)

Fig. 9: Confusion matrix

Based on confusion matrix, two types of accuracy and errors can be measured. TP and TN are related to accuracy while FP and FN are more relevant to classification error. TP is a number that detects fraud from the truly fraud occurrence while TN is a number that detects no fraud from the truly no fraud. True means that the classifier model can detects the positive/negative occurrence from the actual positive/negative occurrence. On the other hand, FP is a condition that the model presume there is a fraud while there is exactly no fraud. FN is when there is a fraud and not be detected by the model. Based on this confusion matrix, three important measurements can be used namely accuracy, precision and recall, where

$$Accuracy = TP / (TP + TN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

$$Recall = TP / (TP + FN) \quad (3)$$

Accuracy defined as the number of true negative or positive fraud occurrence can be detected by the model while precision is the number positive occurrence can be detected true by the model. The number that representing how often the fraud condition occurs in the sample tested dataset is referred as recall.

E. Machine learning algorithms

Table 2 lists the classification machine learning algorithms used in this study.

Table 2: List of algorithms

Algorithm Name	Python Algorithm Call With the Parameters
Logistic regression	logreg = LogisticRegression()
Support Vector Machine (SVM)	svc = SVC(C=2,kernel='poly')
k-Nearest Neighbor(k-NN)	knn = KNeighborsClassifier (n_neighbors = 8)
Gaussian NB	gaussian = GaussianNB ()
Perceptron	perceptron = Perceptron()
Linear Support Vector Machine (Linear SVM)	linear_svc = LinearSVC()
Stochastic Gradient Descent	sgd = SGDClassifier()
Decision Tree	dt = DecisionTreeClassifier(max_depth=2)
Random Forest	random_forest= RandomForestClassifier(estimator’s=100)

The accuracy score of these algorithms were compared and the most outperform algorithm was selected for getting the details information of confusion matrix.

F. Hardware and software for the implementation

The experiments for running the data analysis and testing the machine learning algorithms have been implemented with Python programming and *Scikit-learn* machine learning toolbox. The computer is a Lenovo notebook Intel i7 7th Generation processor, 16 GB RAM.

IV. RESULTS AND DISCUSSION

Table 3 shows the results of accuracy score from all algorithms used in this study.

Table 3: Result of accuracy for all tested algorithms with FeaturesSetOne

Algorithm	Split	Cross Validation Split (Average of 5 Splits)
Logistic regression	0.80	0.72
Support Vector Machines	0.83	0.83
k-NN	0.67	0.72
Gaussian NB	0.60	0.55
Perceptron	0.71	0.65
Linear SVC	0.69	0.78
Stochastic Gradient Descent	0.44	076
Decision Tree	0.67	0.83
Random Forest	0.73	0.83

All algorithms except Support Vector Machine, Gaussian NB, and Perceptron have gained higher score of accuracy with cross validation training compared to basic split training. Drastic increment is in the Stochastic Gradient Descent. The best results provided by Random Forest with cross validation split training approach. Presenting the confusion matrix for the Random Forest with cross validation split is not possible as the approach score is identified based on average score for each validation splits. Therefore, the confusion matrix of SVM is presented in the following Fig. 10. SVM able to produce remarkable accuracy result with the split training approach.

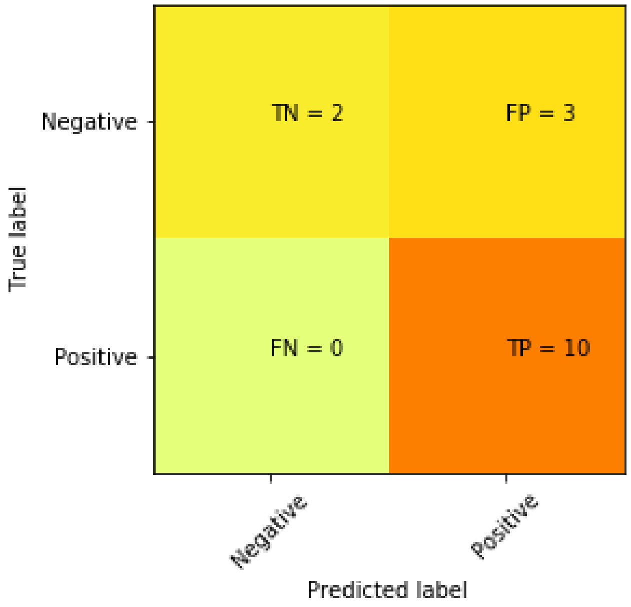

Fig. 10: Confusion matrix of Support Vector Machine

With SVM, 13 positive frauds and 2 negative frauds have been detected (see predicted label). The tested dataset created from the random split has 10 positive frauds and 5 negative frauds (see True label). Therefore, the SVM has 83% accuracy (9 truly detects fraud, 1 truly detects negative fraud from the total 15 tested records). Additionally, it has 0.77% precision (10 is true out of the 13 predicted as frauds) and 100% recall. Recall representing how often the fraud occurs from the tested dataset.

The following Table 4 is accuracy results of all algorithms when tested on the seven independent variables with correlation values in range of 0.1 to 0.35.

Table 4: Result of accuracy for all tested algorithms with FeaturesSetTwo

Algorithm	Split	Cross Validation Split (Average of 5 Splits)
Logistic regression	0.80	0.78
Support Vector Machines	0.83	0.8
k-NN	0.71	0.80
Gaussian NB	0.47	0.39
Perceptron	0.72	0.78
Linear SVC	0.80	0.84
Stochastic Gradient Descent	0.72	0.76
Decision Tree	0.53	0.67
Random Forest	0.87	0.84

Majority of algorithms except Gaussian NB and Decision Tree have some increment of accuracy with the FeatureSetTwo. The worst effect occurred on Perceptron algorithm. Although the 24 features have a very low correlation with the ETR, these variables however contribute to some level of efficiency to the machine learning algorithms. Even with the ranked lowest correlation features, there are still contain considerable information and are somewhat relevant. The Random Forest, outperform the rest of algorithm on the FeatureSetTwo and the following Fig. 11 is the confusion matrix.

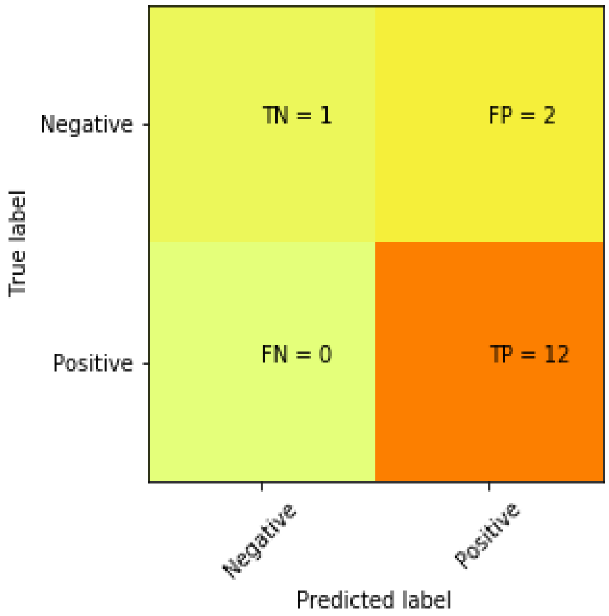

Fig. 11: Confusion matrix of Random Forest

The Random Forest algorithm detects 14 positive frauds and 1 negative fraud. The selected tested dataset from the random split training approach has 12 positive frauds and 3 negative frauds. Therefore, the Random Forest algorithm has 87% accuracy (12 truly detects fraud, 1 truly detects negative fraud from the total 15 tested records). Additionally, it has 86 percent precision (12 out of 14 expected frauds are true) and 100 percent recall.

V. CONCLUSION

This paper presents the review and finding of using machine learning algorithms for real data of Malaysian Government-Linked companies. The researchers demonstrated a real data analysis and presented the performances different machine learning models with different training approaches and features selection. Selection of features has been done based on the correlation values between dependent variable and the independent variables. In general, most selected algorithms can produce better or similar level accuracy results with better correlation of variables. However, there was no indicator can be concluded that using cross validation split training can affect the algorithms performance. However, the finding is limited to the tested dataset and therefore requires furthers investigations for different types of problems.

VI. ACKNOWLEDGMENT

This research was supported by Fundamental Research Grant Scheme, Ministry of Education, Malaysia grant number FRGS/1/2018/SS01/UITM/02/32.

REFERENCES

1. S. Chen, X. Chen, Q. Cheng, and T. Shevlin, "Are family firms more tax aggressive than non-family firms?," J. Financ. Econ., 95(1), 2010, pp. 41–61.
2. E. A. Abdul Wahab, A. M. Ariff, M. Madah Marzuki, and Z. Mohd Sanusi, "Political connections, corporate governance, and tax aggressiveness in Malaysia," Asian Rev. Account., 25(3), 2017, pp. 424–451.
3. J. Lismont, E. Cardinaels, L. Bruynseels, S. D. Groote, B. Baesens, W. Lemahieu, and J. Vanthienen, "Predicting tax avoidance by means of social network analytics," Decis. Support Syst., 108, 2018, pp. 13–24.
4. K. H. Chan, P. L. L. Mo, and A. Y. Zhou, "Government ownership, corporate governance and tax aggressiveness: evidence from China," Account. Financ., 53(4), 2013, pp. 1029–1051.
5. M. Zhang, M. Lijun, B. Zhang, and Z. Yi, "Pyramidal structure, political intervention and firms' tax burden: Evidence from China's local SOEs," J. Corp. Financ., 36, 2016, pp. 15–25.
6. R. Balakrishna and R. Anandan, "Early diagnosis of chronic and acute pancreatitis using modern soft computing techniques," International Journal of Recent Technology and Engineering, 7(4S), 2018, pp. 65–69.
7. S. Kavipriya and T. Deepa, "Dual Edge Classifier Based Support Vector Machine (DESVM) classifier for clinical dataset," International Journal of Recent Technology and Engineering, 7(6), 2019, pp. 331–338.
8. A. K. Rai, S. Agarwal, M. Khaliq, and A. Kumar, "Quantitative analysis of development environment risk for agile software through machine learning," International Journal of Recent Technology and Engineering, 7(6), 2019, pp. 83–89.
9. C. Rajinikanth and S. A. Lincon, "A semi supervised based Hyper Spectral Image (HSI) classification using machine learning approach," International Journal of Recent Technology and Engineering, 7(5S2), 2019, pp. 13–16.
10. J. Dean, Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. New Jersey: John Wiley Sons, 2014.
11. M. A. Hall, Correlation-based feature selection for machine learning. Phd thesis, Hamilton: University of Waikato, 1999.
12. C. Schaffer, "Technical note: Selecting a classification method by cross-validation," Mach. Learn., 13(1), 1993, pp. 135–143.