Taxpayer Compliance Classification Using C4.5, SVM, KNN, Naive Bayes and MLP

M. Jupri
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
m.jupri.17092@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
riyanarto@if.its.ac.id

***Abstract*— Tax revenue has a very important role to fund the State's finances. In order for the optimal tax revenue, the tax authorities must perform tax supervision to the taxpayers optimally. By using the self-assessment taxation system that is taxpayers calculation, pay and report their own tax obligations added with the data of other parties will create a very large data. Therefore, the tax authorities are required to immediately know the taxpayer non-compliance for further audit. This research uses the classification algorithm C4.5, SVM (Support Vector Machine), KNN (K-Nearest Neighbor), Naive Bayes and MLP (Multilayer Perceptron) to classify the level of taxpayer compliance with four goals that are corporate taxpayers comply formally and materially required, corporate taxpayers comply formally required, corporate taxpayers comply materially required and corporate taxpayers not comply formally and materially required. The classification results of each algorithm are compared and the best algorithm chosen based on criteria F-Score, Accuracy and Time taken to build the model by using fuzzy TOPSIS method. The final result shows that C4.5 algorithm is the best algorithm to classify taxpayer compliance level compared to other algorithms.**

Keywords—*classification rule learning, taxpayer compliance, data mining, fuzzy TOPSIS.*

I. INTRODUCTION

Tax is a source income of a country that has a very big role compared to other sources income. For corporate, income tax as a deduction of net income earned by the corporate so that there is a tendency for corporates to do income management and tax management [1]. The self-assessment tax system will build enormous data which is a challenge quickly for tax authorities to detect non-compliant taxpayers for further research. To know the non-compliant taxpayer must be made the right method, accurate and the selection of many variables that determine the goal.

The understanding of business process and the obligation of taxpayers will assist in choosing the right variable to determine a success in this research. Types of tax obligation to corporate taxpayers in Indonesia consist of Income Tax Article 21 / Article 26, Income Tax Article 23, Income Tax Article 26, Final Income Tax, Income Tax Article 25, Income Tax Article 29 and Value Added Tax (VAT). Each type of tax has a tax reporting form, reporting due date and payment due date. The tax reporting form is called the Tax Return consist of Periodic Tax Return and Annual Tax Return. The Periodic Tax return is a tax obligation which is conducted every month and the annual tax return is a tax obligation which is conducted every year. In the consolidation of law in the Republic Of Indonesia Number 6 of 1983 concerning General provisions and tax procedures as lastly amended by the Law Number 28 of 2007 mentioned that the Taxpayer can repair the tax return has been reported as long as it hasn't been audited. It makes tax administration data will always grow and become very large.

The algorithm classification C4.5, SVM, KNN, Naive Bayes and MLP are used in this research to determine the level of corporate taxpayer compliance which distinguished based on four goals are corporate taxpayers comply formally and required materially, corporate taxpayers comply formally required, corporate taxpayers comply materially required and corporate taxpayers not comply formally and materially required. The C4.5 algorithm is the best algorithm to classify the taxpayer compliance level because it has the highest preference value based on criteria F Score, Accuracy, and Time taken to build the model compared to other algorithms.

II. PREVIOUS RESEARCH

Detecting Tax avoidance is an interesting research because the tax has the nature of coercion and the tax benefits for taxpayers can't be obtained directly. Therefore, the tendency of taxpayers does not report their earnings correctly is greater.

The previous research has detected tax frauds by the tax type Value Added Tax (VAT) by using data mining association algorithm [2]. There is also research to detect tax fraud of corporate taxpayers by using hybrid intelligence systems for corporate taxpayers who have business in food and textile in Iran [3]. Both types of research are relevant to be applied in Indonesia if the business processes, data structures, and tax regulations have similarities. The research based on the financial statements for a particular type in

business is irrelevant if doing for different types of business because every company has different business processes, different costs, different profits and difference financial statements standards of each country.

This research offers a method to classify corporate taxpayers in Indonesia based on four levels of compliance. The variables in this research can be used for all types of corporate taxpayers which cover all types of taxes. To select the best algorithm in this research is by using fuzzy TOPSIS method based on criteria F Score, Accuracy and Time taken to build the model.

III. PROPOSED METHOD

The main idea in this research is how to create dataset consisting of variables that affect the level of taxpayer compliance from the enormous tax data which will always grow continuously and to examine it use algorithms classification C4.5, SVM, KNN, Naive Bayes, and MLP. Each algorithm was compared to find out which algorithm has the highest preference value based on the criteria F Score, Accuracy, and Time taken to build the model (fig 1). This research object is the corporate taxpayer because there are some variables that can not be applied to individual taxpayers as well as vice versa.

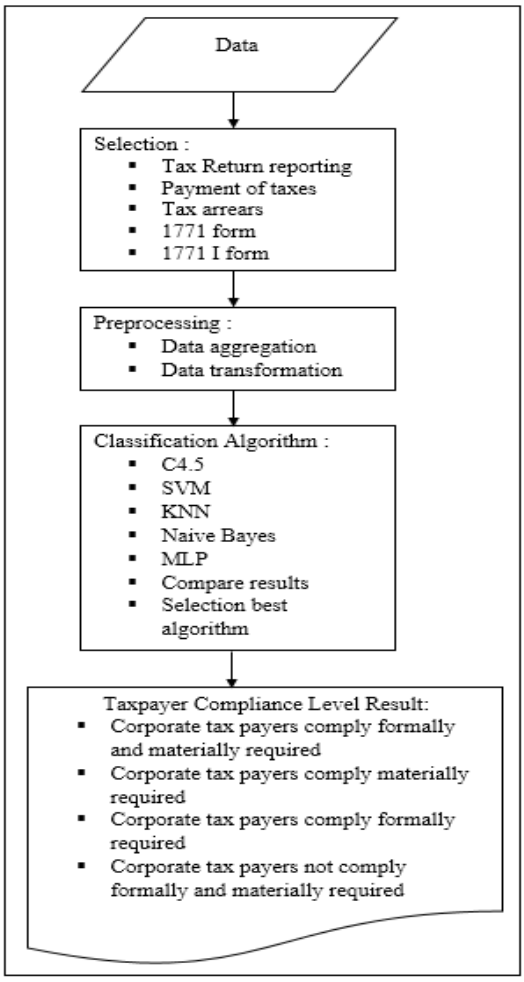

Fig. 1. Research Overview Diagram

A. Data Selection

Data mining is designed to find out the right variable to determine the purpose of the decision to be taken [4]. The choice of the right variable greatly determines the accuracy of the data mining to be performed. The data used in this research comes from the recapitulation of taxpayers tax return and taxpayer payments in certain regions of the fiscal year 2014 and 2015. The tables that become objects of this research are the report table of tax return, tax payment table, tax arrears table, form 1771 table, and form 1771 Attachment I table.

B. Preprocessing

B.1. Data Aggregation

Aggregation is an attempt made to make a summary of a data [5]. Periodic Tax Return data, Annual Tax Return data, and tax payment data are aggregated to obtain information amount late reports in a year, non-reporting amount in a year and the amount of late payment in a year for each taxpayer, and type of tax. The result of the aggregation process is the variable that will determine corporate taxpayers comply formally required.

B.2. Data Transformation

Existing data should be consolidated into a form suitable for the purpose of data mining [5]. Selected data are aggregated within a year to perform data equalization techniques. Equalization technique is the process of comparing Tax Return data of taxpayer both Tax Return data with Tax Return data and Tax Return data with tax payment data. The purpose of this technique is to detect an indication of fraud taxpayer in performing taxation obligations. The result data of this technique is transformed and made into a variable to determine corporate taxpayers that comply material required.

The level of corporate taxpayer compliance in this research consists of formal compliance and material compliance. Corporate taxpayers comply formal required if do not report late more than three times in a year and not late to pay more than three times. Corporate tax payers comply material required are taxpayers who have no tax arrears in a year, taxpayers who have obligations under Government Regulation number 46 has paid the income tax based on to Government Regulation number 46, there is no difference of payment between income tax article 25 in the Corporate Annual Tax Return paired with the payment of the income tax installment article 25, there is no difference between the underpayment tax in the Corporate Annual Tax Return with the payment of income tax article 29, for a Taxpayer confirmed as a Taxable Entrepreneur there is no difference between the underpayment of VAT tax returns with VAT payments in a year, there is no difference between the Underpayment Tax on the Income Tax Article 23 paired with the payment of Income Tax Article 23 in a year, there is no difference between the underpayment tax in income tax Article 21 paired with the payment of income tax Article 21

in a year, there is no difference Income Tax Installment Article 25 in the Annual Tax Return is paired with the income tax Article 25 on the April tax period of the next tax year, there is no difference turnover in the Annual Tax Return paired with the VAT tax base in a year and turnover not less than the amount withholding tax slip in a year.

Table 1. Research dataset

No	Variable	Description	Type
1	JTL	Amount of late reports in a year	Numeric
2	JTdkL	Non-reporting amount in a year	Numeric
3	JTP	Amount of late payment in a year	Numeric
4	Arrears	There are arrears, no arrears	Nominal
5	Equalization1	Difference turnover in the Annual Tax Return with the VAT tax base	Numeric
6	Equalization2	Difference turnover where turnover less than the amount withholding tax slip	Numeric
7	PP46	0 = Not Government Regulation number 46, 1 = must Government Regulation number 46 and not paid	Nominal
8	Equalization3	Difference income tax article 25 in the Corporate Annual Tax Return with payment	Numeric
9	Equalization4	Difference underpayment tax in the Corporate Annual Tax Return with the payment of income tax article 29	Numeric
10	Equalization5	Difference Income Tax Installment Article 25 in the Annual Tax Return with the income tax Article 25 on the April tax period of the next tax year	Numeric
11	Equalization6	Difference between the underpayment of VAT tax returns with VAT payments	Numeric
12	Equalization7	Difference income tax article 23 with payments	Numeric
13	Equalization8	Difference income tax article 21 with payments	Numeric
14	Equalization9	Difference income tax article 4 paragraph 2 with payments	Numeric
15	Tax Compliance	Goal in research	Nominal

C. Classification Process

Data classification is a process to search knowledge from the data of a model to predict the desired label/goal [5].

C.1. C4.5

C4.5 is an algorithm classification developed by J. Ross Quinlan that is a substitute for the ID3 (Iterative Dichotomiser) algorithm which it has developed. This algorithm builds a classification model of a top-down decision tree that evaluates all attributes using a statistical measure called information gain [5]. The first C4.5 algorithm process has selected the attribute as the root. The attribute that has the highest gain value is chosen as the root (1).

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \left(\frac{|S_v|}{|S|} Entropy(S) \right) \quad (1)$$

S : The Set of Cases
A : Atribut
n : Number of partition attributes A
|Sv| : Number of cases on i partition
|S| : Number of cases in S

To calculate the gain value, the first thing finds the value of entropy (2). Entropy is diversity. The more diversity of data, the greater the value of entropy.

$$Entropy \quad (S) = \sum_{i=1}^{n} - p_i \log_2 p_i \quad (2)$$

S : The Set of Cases
n : Number of partitions S
p *i*: The proportion of Sv against S

After selecting the attribute as the root, the next step is to create a branch for each attribute value. Then divide the case in the branch and repeat the process for each branch until all the cases on the branch have the same class.

C.2. Support Vector Machine

Support Vector Machine (SVM) is a classification method for linear and nonlinear data by using nonlinear data mapping to transform training data to a higher dimension [5]. This method will find hyperplane by maximizing margin or distance between classes. The best hyperplane is that is located between two sets of objects of two classes. If + b = +1 is supporting hyperplane of class +1 (+ b = +1) and + b = -1 is supporting hyperplane of class -1 (+ b = -1), the margin between two classes can be calculated by finding the distance between two supporting hyperplane of both classes. Specifically, the margin is calculated in the following way (3)

$$\left(\frac{w}{\|w\|} (x_1 - x_2) \right) = \frac{2}{\|w\|} \quad (3)$$

C.3. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification method that classifies data testing based on the distance function between data testing to the nearest training data (Neighbor) which has the highest number [6]. This algorithm will compare data testing with similar training data. if the data is not known then the data will be given the class of training data closest to the pattern space. This method is also called lazy learning method.

C.4. Naive Bayes

Naive Bayes (NB) is a classification method that uses Bayes theory (4) that is based on probability and statistical knowledge [7]. This method was discovered by Thomas

Bayes in the 18th century. Decision-making on the Bayes theorem relates to inference probabilities that serve to collect knowledge about previous events by predicting events through the rule base [8]. The Naïve Bayes classification has independent input variables that assume the presence of an articular feature of a class is unrelated to the presence of other features [9].

$$P(h_j|x) = \frac{p(x|h_j)\,P(h_j)}{p(x)} \quad (4)$$

$P(h_j|x)$ = States the probability arises h_j if known x.
$p(x|h_j)$ = The likelihood function of hj to x
$P(h_j)$ = Prior probability
$p(x)$ = Evidence

C.5. Multilayer Perceptron

Multilayer Perceptron (MLP) is an artificial neural network model that can be used for data classification [10]. Artificial neural network terminology is how neurons in the human brain to function and interact in parallel for recognition, reasoning, and recovery of damage [11]. The learning process of this algorithm finds the most appropriate synaptic weight to classify patterns in the data set of training. Synaptic is a link that connects neurons with other neurons in the network.

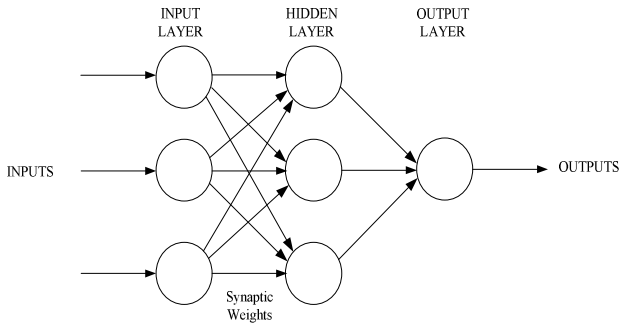

Fig. 2. Architecture of artificial neural network

Figure 2 above shows that the ANN structure consisting of the input layer, the hidden layer and the output layer where each layer contains some neurons that have some weights associated with it for further processing.

C.6. Compare Results

The classification results of C4.5, SVM, KNN, Naive Bayes and MLP algorithms are paired to determine the validity of data with Confusion Matrix, while the measurement used is precision, recall, and accuracy. Time taken to build the model for each algorithm will also be evaluated.

Confusion matrix helps to provide classification performance information against the goal of how the results of classification are correct and how wrong the classification [12]. Correctly classify Instances are shown by elements True Positive (TP) and True Negative (TN) while incorrectly classify Instances are shown by elements False Positive (FP) and False Negative (FN) fig 3 [13]. The results from classification algorithm are compared with the results from trusted external assessment classifier known as True Positive and False Positive [14].

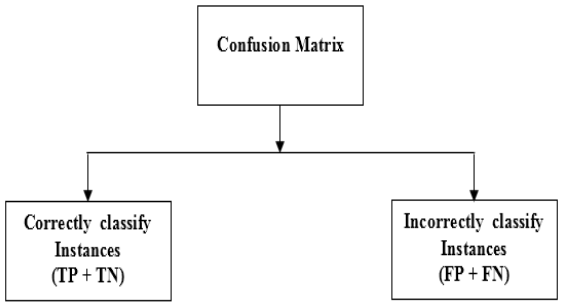

Fig. 3. Confusion Matrix

Precision is the level of accuracy between the information requested by the user and the answers provided by the system (5). The Recall is the success rate of the system in rediscovering an information (6). To avoid measurements that cause the wrong deviation, a combination of Precision and recall is used (7). Accuracy is the degree of proximity between the predicted value and the actual value (8).

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (5)$$

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (6)$$

$$\text{F Score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (7)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (8)$$

Separate Training and Test Sets is a testing method by dividing into two parts that is training set and test set (Figure 4) [12].

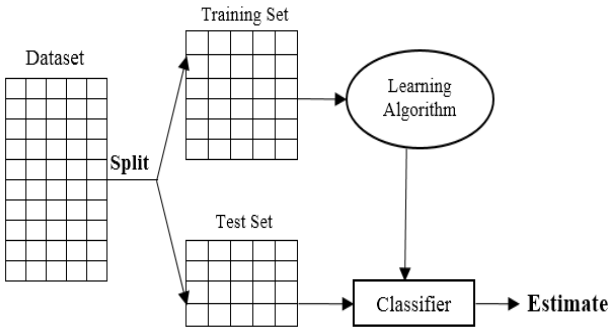

Fig. 4. *Separate Training and Test Sets*

k-fold Cross-validation is a method of testing by dividing training data as much as k section, k-1 part is used as data for

training and the rest is used as test data and has the characteristic that k is small such as 5 or 10 (fig 5) [12].

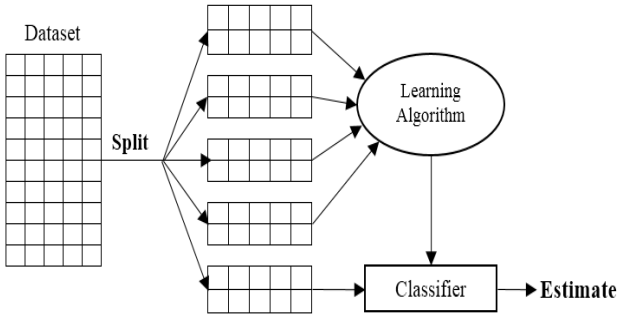

Fig. 5. *k*-fold Cross-validation

C.7. Selection best algorithm

Accuracy results generated by classification algorithms are not the main factors to measure the best performance of a classification algorithm even though most people assume that way [12]. Therefore, in this research, the value of F Score, Accuracy, and Time taken to build the model is weighted to determine the ranking by using Fuzzy TOPSIS method [15]. TOPSIS method is multi-criteria decision analysis method proposed in 1981 by Hwang and Yoon to find the optimal alternative with a Positive Ideal Solution (PIS) and to find the furthest optimal alternative with a Negative Ideal Solution (NIS) [16]. The steps of Fuzzy TOPSIS method are as follows [15] [17][18] :

1. Normalization of decision matrix (9).

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}, i = 1,2,3...,m$$
$$j = 1,2,3...,n \quad (9)$$

r_{ij} is the normal i-th number on the alternative.

2. The weighted normalization matrix (10).

$$v_{ij} = w_j \cdot r_{ij}, i = 1,2,3,\ldots m$$
$$j = 1,2,3,\ldots n \quad (10)$$

w_j is the weight value for each criterion.

3. Determining ideal solutions with negative ideal and positive ideal (11).

$$A^- = \{v_1^-, v_2^-, v_3^- ... v_n^-\}$$
$$A^+ = \{v_1^+, v_2^+, v_3^+ ... v_n^+\} \quad (11)$$

A- is used as a negative ideal while A+ is used for a positive ideal.

4. Calculate the negative ideal distance and positive ideal in every alternative with the following formula (12):

$$D_i^- = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^-)^2} \quad D_i^+ = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^+)^2}$$
$$i = 1,2,3...,m \quad (12)$$

D_i^- is used to calculate the distance of each alternative negative ideal while D_i^+ is used to calculate the distance of each alternative positive ideal.

5. Calculate the preference value (13).

$$C_i = \frac{D_i^-}{(D_i^+ + D_i^-)}$$
$$i = 1,2,3...,m \quad (13)$$

6. Rank each alternative based on the greatest preference value.

IV. RESULT AND DISCUSSION

The dataset that has been created is tested using the weka tools version 3.8.1. The dataset containing 2.998 data was tested with 60% Percentage Split mode and Cross-validation 10-folds with C4.5, SVM, KNN, Naive Bayes and MLP algorithms.

Table 2. Confusion Matrix on Percentage Split 60%

	C4.5	SVM	KNN	NB	MLP
Correctly Classified	1181	1055	1121	625	1068
Incorrectly Classified	18	144	78	574	131

Table 3. Comparison on Percentage Split 60%

	C4.5 (%)	SVM (%)	KNN (%)	NB (%)	MLP (%)
Precision	98.50	89.70	93.70	74.20	90.30
Recall	98.50	88.00	93.50	52.10	89.10
F Score	98.50	88.84	93.60	61.22	89.70
Accuracy	98.50	87.99	93.49	52.13	89.07

Table 2 dan Table 3 shows that by using 60% percentage split mode which means 60% is training data and 40% test data, to know taxpayer compliance level with four goals it is known that C4.5 algorithm has highest correctly classified, precision, recall, and accuracy compared with other classification algorithms. The Naive Bayes algorithm has the lowest correctly classified, precision, recall, and accuracy compared to other algorithms.

Table 4. Confusion Matrix on Cross Validation 10-folds

	C4.5	SVM	KNN	NB	MLP
Correctly Classified	2966	2654	2814	1493	2676
Incorrectly Classified	32	344	184	1505	322

Table 5. Comparison on Cross Validation 10-folds

	C4.5 (%)	SVM (%)	KNN (%)	NB (%)	MLP (%)
Precision	99.00	90.00	94.00	75.00	90.00
Recall	99.00	89.00	94.00	50.00	89.00
F Score	99.00	89.50	94.00	60.00	89.50
Accuracy	98.93	88.53	93.86	49.80	89.26

Table 4 and Table 5 shows that by testing the Cross-validation 10-folds mode, the C4.5 algorithm has the highest correctly classified, precision, recall, and accuracy compared with other classification algorithms. Naive Bayes algorithm has the lowest correctly classified, precision, recall, and accuracy compared with other classification algorithms.

Table 6. Time taken to build model (seconds)

	C4.5	SVM	KNN	NB	MLP
Percentage Split 60%	0.13	1.36	0.02	0.08	11.95
Cross Validation 10-folds	0.06	0.87	0	0.03	9.97

Table 6 shows time required to create a model for each algorithm. K-Nearest Neighbor algorithm has the fastest time compared to other algorithms while Multi Layer Perceptron algorithm takes the longest time to create a model.

Table 7. Weight of Criteria

Criteria	C1	C2	C3
Weight	0.5	0.4	0.2

The results algorithm comparison is calculated to view the best ranking by assigning weights for Accuracy (C1), F Score (C2) and Time taken to build model (C3) criteria as shown in table 7. By using the equations (9), (10), (11), (12) and (13) obtained the preference ratings of each algorithm as shown in tables 8 and 9 below:

Table 8. Ranking alternative on Percentage Split 60%

Ranking	Alternative	Preference
1	C4.5	0.993
2	KNN	0.935
3	SVM	0.837
4	NB	0.580
5	MLP	0.359

Table 9. Ranking alternative on Cross Validation 10-folds

Ranking	Alternative	Preference
1	C4.5	0.995
2	KNN	0.935
3	SVM	0.852
4	NB	0.569
5	MLP	0.371

Good taxpayer supervision is the more variables used to determine the level of taxpayer compliance with high precision, recall and accuracy results and can be used for all taxpayers types. Therefore, next research can add variables to know the level of taxpayer compliance and these variables can be used for all taxpayers types not only corporate taxpayers.

V. CONCLUSION

Based on the comparison of classification algorithms C4.5, Support Vector Machine, K-Nearest Neighbor, Naive Bayes and Multi Layer Perceptron, to know taxpayer compliance level can be concluded that:

1. Both Percentage Split 60% and Cross Validation 10-folds test get the same result that is C4.5 is the best classification algorithm based on the criteria F Score, Accuracy and Time taken to build model
2. Naive Bayes algorithm and Support Vector Machine algorithm are better than the Multilayer Perceptron algorithm based on criteria F score, Accuracy and Time taken to build model although the Multilayer Perceptron algorithm has a higher f score and accuracy.

REFERENCES

[1] M. S. Mulyadi and Y. Anwar, "Corporate Governance, Earnings Management and Tax Management," *Procedia - Soc. Behav. Sci.*, vol. 177, pp. 363–366, 2015.

[2] R. S. Wu, C. S. Ou, H. Y. Lin, S. I. Chang, and D. C. Yen, "Using data mining technique to enhance tax evasion detection performance," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8769–8777, 2012.

[3] E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran," *Int. J. Account. Inf. Syst.*, vol. 25, pp. 1–17, 2017.

[4] I. H. Witten, E. Frank, and M. a. Hall, *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*, vol. 54, no. 2. 2011.

[5] M. K. A. J. P. Jiawei Han, "Data Mining: Concepts and Techniques, Third Edition - Books24x7," *Morgan Kaufmann Publ.*, p. 745, 2012.

[6] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," in *Proceedings of 2015 International Conference on Data and Software Engineering, ICODSE 2015*, 2016, pp. 170–174.

[7] H. Zhang, Z. X. Cao, M. Li, Y. Z. Li, and C. Peng, "Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals," *Food Chem. Toxicol.*, vol. 97, pp. 141–149, 2016.

[8] M. Mayilvaganan and D. Kalpanadevi, "Comparison of classification techniques for predicting the performance of students academic environment," *Commun. Netw. Technol. (ICCNT), 2014 Int. Conf. Comput. Intell. Comput. Res.*, pp. 113–118, 2014.

[9] P. Suryachandra, "Comparison of Machine Learning Algorithms," *3rd Int. Conf. Sci. Technol. - Comput. Comp.*, vol. 8, no. 5, pp. 2241–2247, 2017.

[10] A. M. Mubarek and E. Adali, "Multilayer perceptron neural network technique for fraud detection," *2017 Int. Conf. Comput. Sci. Eng.*, pp. 383–387, 2017.

[11] G. Singh and M. Sachan, “Multi-layer perceptron (MLP) neural network technique for offline handwritten Gurmukhi character recognition,” in *2014 IEEE International Conference on Computational Intelligence and Computing Research, IEEE ICCIC 2014*, 2015.

[12] M. Bramer, *Principles of Data Mining*. 2007.

[13] D. L. Gupta, A. K. Malviya, and S. Singh, “Performance Analysis of Classification Tree Learning Algorithms,” *Int. J.*, vol. 55, no. 6, pp. 39–44, 2012.

[14] R. A. E.-D. Ahmeda, M. E. Shehaba, S. Morsya, and N. Mekawiea, “Performance study of classification algorithms for consumer online shopping attitudes and behavior using data mining,” *Proc. - 2015 5th Int. Conf. Commun. Syst. Netw. Technol. CSNT 2015*, pp. 1344–1349, 2015.

[15] U. Yudatama and R. Sarno, “Priority Determination for Higher Education Strategic Planning Using Balanced Scorecard, FAHP and TOPSIS (Case study: XYZ University),” in *IOP Conference Series: Materials Science and Engineering*, 2016, vol. 105, no. 1.

[16] D. Walczak and A. Rutkowska, “Project rankings for participatory budget based on the fuzzy TOPSIS method,” *Eur. J. Oper. Res.*, vol. 260, no. 2, pp. 706–714, 2017.

[17] O. Sohaib and M. Naderpour, “Decision making on adoption of cloud computing in e-commerce using fuzzy TOPSIS,” in *IEEE International Conference on Fuzzy Systems*, 2017.

[18] M. Dağdeviren, S. Yavuz, and N. Kilinç, “Weapon selection using the AHP and TOPSIS methods under fuzzy environment,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8143–8151, 2009.