Non-Compliance Level of Motor Vehicle Taxpayer Classification

2023 14th International Conference on Information & Communication Technology and System (ICTS) | 979-8-3503-1216-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTS58770.2023.10330868

Ratih Nur Esti Anggraini
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ratih_nea@if.its.ac.id

Arianto Nugroho
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ariantonugroho83@yahoo.com

Raditia Wahyuwidayat
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
raditia.w@gmail.com

Riyanarto Sarno
Informatics Department
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— Motor vehicle taxes play a significant role in generating revenue for local governments, making it imperative to prioritize the collection of these taxes. In order to achieve the predetermined targets for motor vehicle tax revenue, it becomes necessary to closely monitor taxpayers who fail to comply with their tax payment obligations. This study focuses on classifying taxpayers into three distinct categories using the Objects and Subjects Data Collection Letter (SPOS), Tax Calculation Note (NPP), and Tax Bill Note (NTP) as classification criteria. To compare the performance of different Machine Learning algorithms, the classification process is conducted. The evaluation of algorithm performance is carried out using the 10 Folds Cross Validation method with a data split of 60:40, and metrics such as Accuracy, Precision, Recall, and F1-score are employed to assess algorithm effectiveness.**

Keywords*— *machine learning, classification, taxpayer non-compliance, motor vehicle tax

I. INTRODUCTION

Taxes play a crucial role in generating funds for the government at both the national and local levels [1] [2]. These funds are utilized for the development of various government sectors. Motor vehicle tax, categorized as a local government tax, falls under the purview of Law No. 28/2009 on Regional Taxes and Levies. According to this law, motor vehicle tax is imposed on individuals or entities who possess or have control over motorized vehicles [2]..

Motor vehicle taxpayer compliance is where the taxpayer is obedient or aware in fulfilling his tax obligations by making payments for the tax year from motorized vehicles by applicable tax provisions. Taxpayer compliance is important because if the taxpayer does not comply, the taxpayer will tend to take tax avoidance actions [1, 3], thus impacting motor vehicle tax revenues that are achieved not by the set targets. Therefore, it is necessary to supervise non-compliant taxpayers, namely taxpayers who are late in making payments for the tax year from motorized vehicles, by billing for late payments of motor vehicle taxes through the issuance of tax papers.

Motor Vehicle Tax is one of the local revenue sectors managed by the Regional Government. According to Law Number 28 of 2009 of Indonesia, Motor Vehicle Tax is a tax on ownership and/or control of motor vehicles [4]. The increasing number of motorized vehicles is in line with the increasing number of taxpayers. Likewise, the number of non-compliant taxpayers tends to increase. The high level of non-compliance with taxpayers will affect the amount of local revenue obtained. This will indirectly affect the process of implementing development in East Java Province.

The purpose of this study was to determine the taxpayer's non-compliance quality, which was carried out by classifying the level of non-compliance of taxpayers in the East Java Province. By knowing the quality of taxpayer non-compliance, the East Java Provincial Revenue Agency can determine a billing strategy so that it can streamline the efforts to collect the local revenue.

In this study, we use several Machine Learning algorithms: Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Logistic Regression, and Multi Layer Perceptron (MLP). The rest of the paper is organized as follows. In section 2 we provide the related work related to the implementation of the classification method. Section 3 introduces our research methodology. Section 4 describes the experimental results, and Section 5 is the conclusion of this paper.

II. RELATED WORKS

Several studies have been carried out to detect tax evasion and classify taxpayer compliance. A study [5] detected tax evasion in food and textile companies using a hybrid intelligence system that combines several data mining classification algorithms. Another research [6] predicted the risk of non-compliance of a group of small businesses to tax obligations using several data mining classification algorithms. Another study [7] conducted a study to estimate the tax avoided by potential tax evaders.

In previous studies, the level of taxpayer compliance has been determined for income tax data using several machine learning classification algorithms [8]. This study aims to classify taxpayer compliance into four distinct groups. These groups include corporate taxpayers who comply both formally and materially, corporate taxpayers who comply formally but are not required to comply materially, corporate taxpayers who comply materially but are not required to comply formally, and corporate taxpayers who do not comply both formally and materially.

In 2018, a study on the introduction of corporate tax evasion behaviour was conducted using the Levenberg-Marquardt (LM) Neural Network [10]. From this research, it can be seen that the LM Neural Network can classify well and can effectively determine whether a company commits tax evasion.

Research on the classification of taxpayer compliance has also been carried out on Land and Building Taxpayer data using the Naive Bayes Algorithm [11]. With data of 1467 Land and Building taxpayers, the algorithm is able to produce an accuracy of 99.33%.

In 2018 research was also carried out in the field of tax avoiding predictions [12]. To conduct this research, two datasets were used. Initially, a dataset comprising firm characteristics data of 1032 firms from Compustat for the fiscal year 2011 and their corresponding tax avoidance data for 2012 was collected. This dataset was utilized to ascertain the companies that are likely to exhibit low-tax behavior in the subsequent year based on the data from the current year.

In a separate investigation, the identification of corporate tax arrears, particularly in the Nanhai region of Foshan, China, was conducted using an ensemble learning model [9]. The study employed six base classifiers, namely Multi Layer Perceptron, K-Nearest Neighbor, Random Forest, Extremely Randomized Trees, Gradient Tree Boosting, and XGBoost.

III. Research Method

In our research, the classification of the level of non-compliance of motor vehicle taxpayers is based on tax papers, namely the tax object and subject data collection letter (SPOS), tax calculation notes (NPP), and tax collection note (NTP). Non-compliant taxpayers of SPOS come from taxpayers who make payments of motor vehicle taxes exceeding the due date plus 15 days to 29 days. Non-compliant taxpayers of NPP come from taxpayers who make payments of motor vehicle taxes exceeding the due date plus 30 days to 59 days. Non-compliant taxpayers of NTP come from taxpayers who make payments more than 59 days after the due date of motor vehicle taxes.

The level of non-compliance of motor vehicle taxpayers based on tax papers is shown in Fig 1. The research methodology can be seen in Fig 2.

A. Dataset

The dataset used in this study is secondary data obtained from East Surabaya Samsat Office. The data taken is the Rungkut subdistrict payment data of Motor Vehicle Tax for the tax period in 2020 with the amount of data as much as 14374 data. There are 17 attributes used in modelling as in Table 2.

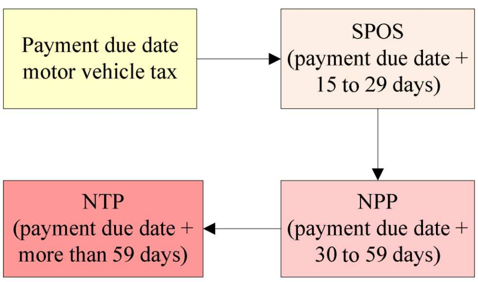

Fig 1. Tax papers

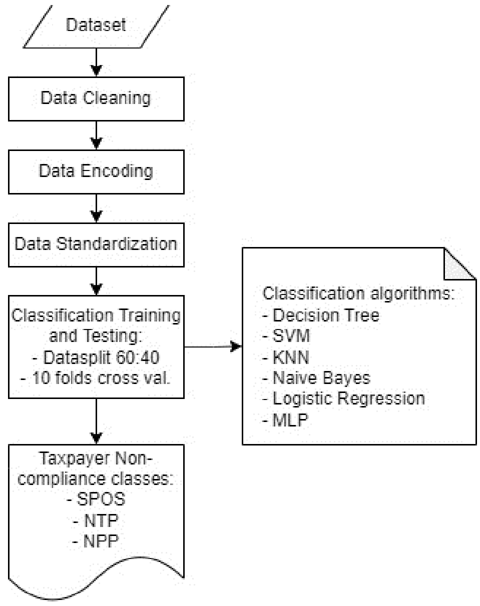

Fig 2. Methodology

Table 1. Data Attributes

No	Attribute	Description
1	KDDESA	Urban Village Code
2	RW	Hamlet Code
3	RT	Neighborhood Code
4	GOLKEND	Vehicle Class Code
5	KDASAL	Vehicle Country of origin code
6	KDMERK	Vehicle Brand Code
7	KDTYPE	Vehicle Type Code
8	KDJENISPOL	Vehicle Kind Code
9	KDPLAT	Vehicle Number Plate Code
10	THBUAT	Vehicle Year of Manufacture
11	JMLTP	Frequency of late-paying taxpayers
12	JMLPKBTP	Amount of late-paying tax
13	JMLHARITP	The Average of Number Days taxpayer late-paying
14	JMLP	Frequency of taxpayers not late paying
15	JMLPKBP	Nominal amount, note late paying tax
16	JMLHARIP	The Average Number of Days Taxpayers are Not Late Paying
17	SURAT	Class

B. Data Cleaning

It is necessary to ensure that there is no missing value in the data used before modelling. Missing data refers to the absence of a value assigned to an observation [13]. In a given dataset, missing data occurs when an input or generated value is not recorded for a particular item. This pervasive issue is encountered in various real-world datasets and cuts across different fields of research [14].

Missing value problems can be solved by deleting or imputing empty data. From the results of the evaluation on the dataset, it was found that one data record used contained a missing value. Based on these conditions, it was decided to delete records with missing values because the amount is not relatively significant.

C. Data Encoding

The encoding of categorical data is a crucial task as a significant portion of real-life data consists of categorical string values. However, most machine learning models operate solely with integer or other compatible value types. Hence, encoding categorical data becomes necessary to transform the categorical values into a numeric format. By performing data encoding, the categorical values on the KDJENISPOL attribute are converted from their original categorical form to a numerical representation. This allows the data to be appropriately processed by models, thereby enhancing the accuracy and effectiveness of predictions.

D. Data Standarization

Standardization is a method of scaling in which values are adjusted to have a mean of zero and a standard deviation of one. By applying standardization, the attribute's mean is shifted to zero, resulting in a distribution with a unit standard deviation..

Data standardization is the process of bringing data into a uniform format that allows analysts and others to research, analyze, and utilize the data. In this study, it is necessary to standardize the data because the range of data in the dataset is too far. This condition can affect the performance of the algorithm.

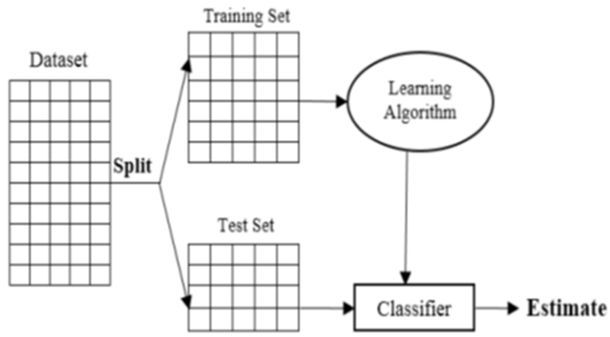

Fig 3. Data Split Method

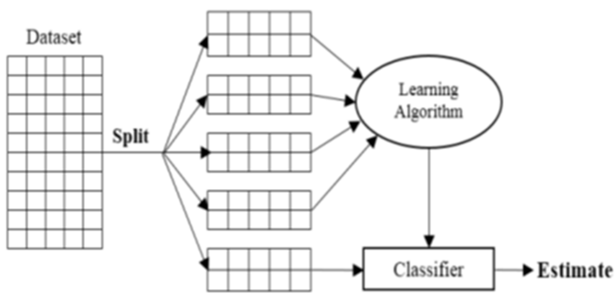

Fig 4. Cross validation method

E. Classification Training and Testing

Before being used in the classification process, the standardized dataset will be divided into Training and Testing Data. In this study, the 60% Split Percentage method was used. With this method 60% of the data from the dataset will be used as Training Data and the remaining 40% as Testing Data.

K-Fold Cross Validation method was used in this research. In this method the dataset is divided into k parts and k-1 parts will be used as testing data. The method of Data Split and Cross Validation is shown in Fig 3 and Fig 4, consecutively.

The performance of several machine learning algorithms will be evaluated.

1. The Decision Tree is a classification technique that transforms data/tables into a branching structure to facilitate the decision-making process [15, 16]. The branching of the decision tree is determined using the entropy function.
2. Support Vector Machine (SVM) is a classification method to find the best hyperplane that functions as a separator between classes [17]. The hyperplane can be found by measuring the margins and finding the maximum point. SVM uses kernel functions to map input space to feature space.
3. K-Nearest Neighbor (KNN) is a classification method that works by comparing testing data and training data. The KNN method's accuracy level is strongly influenced by the number of its closest neighbors, namely the optimal k value [18].
4. Naïve Bayes is a classification method based on simple probability and statistical methods to predict future opportunities based on past experiences. It is designed to be used with the assumption that between one class and another class are not interdependent (independent) [19].
5. Logistic regression is a statistical method that determines the relationship of the dependent and independent variables with the model [20]. The important feature of logistic regression is the categorical dependent variable. This is what distinguishes logistic regression from conventional regression.
6. The Multi Layer Perceptron is a classification method that utilizes an artificial neural network algorithm inspired by the functioning of neural networks in living organisms. It is regarded as a dependable algorithm due to its ability to perform directed learning processes [21].

To determine the performance of each algorithm, the values of Accuracy, Precision, Recall and F-Score are employed using Equation 1, 2, 3, and 4 consecutively.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Score = \frac{precision \times recall}{precision + recall} \times 2 \tag{4}$$

IV. Experiment and Results

This section describes the results of the experimental design. The output of this classification will get the level of non-compliance of the Taxpayer in accordance with the Taxation Letter that will be received, namely SPOS, NTP, and NPP. To determine the performance of each algorithm, the values of Accuracy, Precision, Recall and F-Score are used. The experiment results are shown in Table 2.

Table 2. Evaluation Results

Algorithm	Accuracy	Precision	Recall	F1-score
DT	**99.6**	**99.7**	**99.7**	**99.7**
SVM	94.4	95.0	95.0	95.0
KNN	70.8	61.7	56.7	57.3
NB	87.0	91.5	87.0	87.9
LG	97.9	98.0	97.9	97.9
MLP	98.8	98.8	98.8	98.8

Based on the experiment, we can conclude that Decision Tree gives the best performance in terms of Accuracy, Precision, Recall, and F1-score. In addition, an evaluation on the time taken to do the classification was also evaluated. It is described in Table 3.

Table 3. Time taken to build the model

Algorithm	*Time taken (seconds)*
DT	0.002
SVM	0.003
KNN	0.001
NB	0.01
LG	0.49
MLP	15.00

V. CONCLUSIONS

Taxpayer classification is very necessary in collecting Motor Vehicle Tax to increase local revenue in East Java Province. Therefore, a reliable and robust classification model is needed to be able to manage local revenue collection.

In this study, taxpayers were classified using the Naive Bayes algorithm, MLP and Logistic Regression. The experiment was carried out using the 60% Percentage Split and 10 Folds Cross Validation method. Based on the experiment, Decision Tree showed the best result. On the other hand, Multi Layer Perceptron took very long time to train the model even though the performance result is good.

Further research, the dataset can be developed by adding other taxpayer variables and using different classification algorithms to add to the diversity of comparisons of the classification results of motor vehicle taxpayer non-compliance. It could also use a different case study but still within the scope of the motor vehicle tax.

REFERENCES

[1] R. A. V. Y. Sari and N. Susanti, “Faktor-Faktor Yang Mempengaruhi Kepatuhan Wajib Pajak dalam Membayar Pajak Kendaraan Bermotor (Pkb) Di Unit Pelayanan Pendapatan Provinsi (UPPP) Kabupaten Seluma,” EKOMBIS Rev. J. Ilm. Ekon. dan Bisnis, vol. 2, no. 1, pp. 63–78, 2015, doi: 10.37676/ekombis.v2i1.5

[2] F. A. A. Ahmad, A. Febriansyah, M. R. Mulyana, H. N. Bagja, R. D. A. Andica, and I. F. Muttaqin, “Influence of Motor Vehicle Tax to the Local Revenue of West Java Province in Bandung City,” Rev. Int. Geogr. Educ. Online, vol. 11, no. 5, pp. 3207–3217, 2021, doi: 10.48047/rigeo.11.05.211.

[3] M. Jupri and R. Sarno, “Data Mining, Fuzzy AHP And Topsis For Optimizing Taxpayer Supervision,” Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 2, pp. 1–12, 2019, doi:10.11591/ijeecs.v14.i2.ppab-cd.

[4] Pajak Daerah dan Retribusi Daerah, UU No 28 (2009). https://djpk.kemenkeu.go.id/attach/post-law-of-the-republic-indonesia-number-2008-year-2009-concerning-local-taxes-and-charges/UU-559-753-uu282009verE_Edited.pdf

[5] 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, doi: 10.1016/j.accinf.2016.12.002

[6] G. Cao, A. Downes, S. Khan, W. Wong, and G. Xu, “Taxpayer Behavior Prediction in SMS Campaigns,” Proc. - 2018 5th Int. Conf. Behav. Econ. Socio-Cultural Comput. BESC 2018, pp. 19–23, 2018, doi: 10.1109/BESC.2018.8697317

[7] J. Mathews, P. Mehta, S. Kuchibhotla, D. Bisht, S. B. Chintapalli, and S. V. K.Visweswara Rao, “Regression Analysis towards Estimating Tax Evasion in Goods and Services Tax,” Proc. - 2018 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2018, pp. 758–761, 2019, doi: 10.1109/WI.2018.00011

[8] M. Jupri and R. Sarno, “Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP,” 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, vol. 2018-Janua, pp. 297–303, 2018, doi: 10.1109/ICOIACT.2018.8350710.

[9] A. Su, Z. He, J. Su, Y. Zhou, Y. Fan, and Y. Kong, “Detection of tax arrears based on ensemble leaering model,” Int. Conf. Wavelet Anal. Pattern Recognit., vol. 2018-July, pp. 270–274, 2018, doi: 10.1109/ICWAPR.2018.8521362.

[10] X. Xiangyu, Y. Youlin, and X. Qicheng, “Intelligent Identification of Corporate Tax Evasion Based on LM Neural Network,” Chinese Control Conf. CCC, vol. 2018-July, pp. 4507–4511, 2018, doi: 10.23919/ChiCC.2018.8483951.

[11] Q. Widayati, “Penerapan Data Mining Menggunakan Metode Teknik Classification Untuk Melihat Potensi Kepatuhan Wajib Pajak Bumi Dan Bangunan,” J. Ilm. Matrik, vol. 20, no. 2, pp. 157–168, 2019, doi: 10.33557/jurnalmatrik.v20i2.119.

[12] J. Lismont, E. Cardinaels., L. Bruynseels, S. De Groote, B. Baesens, W. Lemahieu, J. Vanthienen, “Predicting tax avoidance by means of social network analytics,” Decision Support Systems, vol. 108, pp. 13–24, 2018, doi: 10.1016/j.dss.2018.02.001.

[13] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider, “Missing data in surveys: Key concepts, approaches, and applications,” Res. Soc. Adm. Pharm., no. March, 2021, doi: 10.1016/j.sapharm.2021.03.009.

[14] R. Razavi-Far, B. Cheng, M. Saif, and M. Ahmadi, “Similarity-learning information-fusion schemes for missing data imputation,” Knowledge-Based Syst., vol. 187, p. 104805, 2020, doi: 10.1016/j.knosys.2019.06.013

[15] R. N. E. Anggraini, N. F. Ariyani, A. F. Septiyanto, R. Sarno, Z. D. Meilani and T. Soendoro, "A Decision Tree Knowledge-based System for Reviewing Research Ethics Protocol," 2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Solo, Indonesia, 2022, pp. 50-55, doi: 10.1109/COMNETSAT56033.2022.9994414.

[16] I. A. Sabilla, Z. A. Cahyaningtyas, R. Sarno, A. Al Fauzi, D. R. Wijaya, and R. Gunawan, “Classification of Human Gender from Sweat Odor using Electronic Nose with Machine Learning Methods,” Proc. - 2021 IEEE Asia Pacific Conf. Wirel. Mobile, APWiMob 2021, pp. 109–115, 2021, doi:10.1109/APWiMob51111.2021.9435205.

[17] Mihuandayani, E. Utami, and E. T. Luthfi, “Text mining based on tax comments as big data analysis using SVM and feature selection,” 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, vol. 2018-Janua, pp. 537–542, 2018, doi: 10.1109/ICOIACT.2018.8350743.

[18] S. Zhang et al., “Efficient kNN Classification With Different Numbers of Nearest Neighbors,” IEEE Trans. Neural Networks Learn. Syst., pp. 1–12, 2017, [Online]. Available: http://ieeexplore.ieee.org.

[19] Ainurrochman et al., "Ensemble Methods Classifier Comparison for Anomaly Based Intrusion Detection System on CIDDS-002 Dataset," 2021 13th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2021, pp. 62-67, doi: 10.1109/ICTS52701.2021.9608714.

[20] A. A. H. Ahmadini, “A novel technique for parameter estimation in intuitionistic fuzzy logistic regression model,” Ain Shams Engineering Journal, 13(1), p.101518.

[21] M. A. Hasan, R. Sarno, and S. I. Sabilla, “Optimizing machine learning parameters for classifying the sweetness of pineapple aroma using electronic nose,” Int. J. Intell. Eng. Syst., vol. 13, no. 5, pp. 122–132, 2020, doi: 10.22266/ijies2020.1031.12.