INDONESIAN TREASURY REVIEW
JURNAL PERBENDAHARAAN, KEUANGAN NEGARA DAN KEBIJAKAN PUBLIK

MACHINE LEARNING ANALYTICS FOR PREDICTING TAX REVENUE POTENTIAL

Raden David Febriminanto*
School of Business and Management, Institut Teknologi Bandung, Bandung
david_febriminanto@sbm-itb.ac.id

Meditya Wasesa
School of Business and Management, Institut Teknologi Bandung, Bandung
meditya.wasesa@sbm-itb.ac.id

*Corresponding author: david_febriminanto@sbm-itb.ac.id

ABSTRACT

In line with rapid business process digitalization in the Directorate General of Taxes, the size of the data stored in the institution has grown exponentially. However, there is a problem with generating value out of the valuable data assets. Correspondingly, this research provides machine-learning-based predictive analytics as a solution to the question of how to use taxpayers' trigger data as a decision support system to discover and realize unexplored tax potential. More specifically, this research presents predictive analytics models that can accurately predict which potential taxpayers are likely to pay their due. We developed three machine learning models: logistic regression, random forest, and decision tree. We analyzed 5,562 tax revenue potential data samples with eight predictors: trigger data nominal value, distance to tax office, type of taxpayer, media of tax report, type of tax, report status, registered year of taxpayer, and area coverage. Our study shows that the random forest model provided the best prediction performance. The resultant weight of each attribute indicated that the status of the tax report was the top tier of variable importance in predicting tax revenue potential. The analytics can help tax officers determine potential taxpayers with the highest likelihood to pay their due. Given the size of the data records, this approach can provide tax administrators with a powerful tool to increase work efficiency, combat tax evasion, and provide better customer service.

Keywords: prediction, tax, big data, data mining, decision tree, random forest, logistic regression

JEL Classification:
H21

HOW TO CITE:
Febriminanto, R. D. & Wasesa, M. (2022). Machine learning analytics for predicting tax revenue potential. *Indonesian Treasury Review: Jurnal Perbendaharaan, Keuangan Negara dan Kebijakan Publik*, *7*(3), 193-205.

INTRODUCTION

The Directorate General of Taxes (DGT) of the Republic of Indonesia had 49.82 million registered taxpayers by 2021. Figure 1 shows that the number of registered taxpayers in Indonesia has expanded multiple times in the last 20 years. In 2002, the number of taxpayers was 2.59 million, then in 2021, the number had increased almost twenty-fold to 49.82 million (Hariani, 2021)

Corresponding to the need for efficient management of the ever-growing taxpayers, KEP-178/PJ/2004 on the Policy Blueprint of the Directorate General of Taxes for 2001 to 2010 has mandated to modernize the national taxation administration system. In support of this decision, Account Representative (AR) was formed in 2006 to serve the following main tasks: (1) supervision of taxpayer compliance (i.e., material or formal), (2) consultation and services, (3) potential exploration and intensification, and (4) data and information collection and processing.

As a result, the policy creates extensive digitalization in various business processes and generates a large amount of data. The constantly growing number of taxpayers leads to rising data records on tax transactions. One of the most important datasets is the internal trigger data of small tax offices (Kantor Pelayanan Pajak Pratama). The internal trigger data refer to the financial transactions of taxpayers whose taxation obligations have not been carried out or followed up, meaning that they contain tax revenue potential.

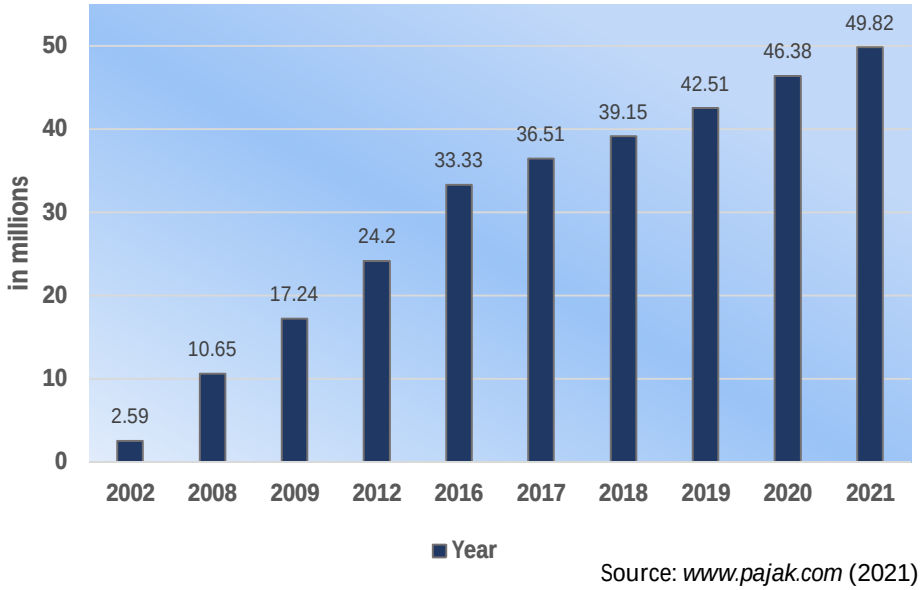
The head tax office delivers internal trigger data to an account representative in the small tax office in raw data format. According to the tax law, these trigger data are gathered from the tax reports that taxpayers must submit periodically according to their type. The account representative must validate and confirm the data, i.e., whether the tax obligations have been carried out and whether the tax obligations are correct or not. This also includes identifying the data containing tax revenue potential. the data also may potentially refer to discovering fraud cases. Suppose the tax revenue potential and fraud can be identified, the tax officer can take corresponding actions in the forms of notification, consultation, or any other action that can lead to tax revenue potential realization.

This research proposes a solution to the question of how to use these trigger data to realize unexplored tax potential by developing a machine-learning-based predictive analytics artefact. It is, in addition, very beneficial to focus on potential

APPLICATIONS FOR PRACTICE

- The tax officers should use the rapid and massive growth of potential tax data to get tax revenue with limited human capital and fund.
- The results of this study can determine the predictors of tax revenue potential that can help tax officers find out about tax revenue potential.
- This research presents accurate predictive analytics to predict which potential taxpayers are likely to pay their due.
- This research also references the classification method that yields the best result to predict tax revenue potential.

Figure 1 Number of Taxpayers Development



Source: *www.pajak.com* (2021)

taxpayers highly likely to pay their due. Also, we can reduce the time and cost inefficiencies in targetting random taxpayer. To do proper targeting, AR must be able to identify indicators of the trigger data that contain high potential tax revenue realization precisely. In response, this research presents an accurate predictive analytics model that can accurately predict which potential taxpayers are likely to pay their due. Given the size of the data records, this approach can provide tax administrators with a powerful tool to increase work efficiency, combat tax evasion, and provide better customer service (Strømme, 2018). The study on the use of machine learning to predict outcomes in tax law by Alarie et al. (2016) concluded that the usage of machine learning techniques can answer the provide predictions with greater accuracy rather than spending hours researching precedents, or relying on imperfect human memory.

LITERATURE REVIEW

The issue of tax revenue prediction/ determinants has been well studied throughout the years. Table 1 summarizes previous research works on the tax revenue topic and lists the details of each paper in terms of objective, method, dependent and independent variables, and type of data source (i.e., macroeconomic/ microeconomic data).

Several authors have conducted studies in the tax area, both forecasting or predicting and finding

Table 1 Summary of previous research on tax revenue indicators

Article	Variables		Indicator	Objective	Method
	Dependent	Independent			
Andrejovská & Puliková (2018)	Tax revenue	Statutory corporation, GDP, ETR, LoE, IN, PD, FDI	Macroeconomy	To quantify the impact of selected macroeconomic indicators on the total amount of tax revenues	Regression analysis
Petutschnig (2017)	Future orientation	Personal income taxes rates, capital gains taxes, taxes rates on dividend, corporate income taxes, VAT	Macroeconomy	To find out if various aspects of a country's tax system have a positive or negative influence on individuals' attitudes toward the future	Fixed effects panel regression
Lismont et al. (2018)	Tax avoidance	EBITDA, R&D, advertising, SG&A, capex, sales, leverage, cash, FOR, NOL, size, intangible, PP&E	Microeconomy	To create tax avoidance prediction models using three popular machine learning techniques, namely logistic regression, decision trees, and random forest.	Logistic regression, random forest, decision tress.
Brender & Israel (2014)	Tax revenue	GDP, change in GDP, real wage employee, import of consumption goods, sales of new dwellings, sales of shares by parties, credit rates	Macroeconomy	To examine an alternative model for predicting government tax revenues in Israel	The Engel-Granger method
Cezar & Lozano (2020)	Tax crime	Annual tax value as a specialized company, annual tax value as a small business	Microeconomy	This study applied machine learning and algorithms with the goal of predicting tax crimes.	Naïve bayes, decision tree, random forest, logistic regression
Hassan et al. (2021)	Tax revenue	Government stability, law and order, internal and external conflicts	Macroeconomy	To investigate the relationship between governance and tax revenue collection	Autoregressive distributive lag (ARDL)
Ogneru (2019)	Tax revenue	GVA, direct income, gross operating surplus, compensation of employees, mixed income, final consumption	Macroeconomy	To examine the nature of the gross value added (GVA) and to identify the best predictors of tax revenue for Romania	Time series regression
Javid & Arif (2012)	Revenue potential	GDP per capita, agriculture value addition to GDP, trade/GDP, debt/GDP, population growth, inflation, control of corruption, bureaucracy, law and order	Macroeconomy	To analyze revenue potential and revenue effort in developing Asian countries	Panel regression analysis
Sapiei et al. (2014)	Tax compliance	Corporate characteristic, tax compliance costs, tax attitudinal aspects	Microeconomy	To gain insight into the influence of some possible causes that affect tax compliance	Multiple regression analysis
Tarfa, et al. (2020)	Tax revenue generation	Illegal practice of taxpayers, objective of tax audit, tax audit techniques, provision of training for taxpayers	Microeconomy	To assess the effects of tax audit on revenue generation in a case of the Ethiopian Ministry of Revenue	Regression

Source: author's data

determinants of it. Lismont et al. (2018) developed a tax avoidance prediction model, and the results showed that random forest performed better than logistic regression and decision tree. Petutschnig (2017), using fixed panel regression, expanded the existing tax literature by providing evidence that tax could influence fundamental personal values, such as individuals' attitudes toward their future.

Javid & Arif (2012) reported that per capita GDP, share of agriculture in GDP, and foreign debt were strong determinants of tax revenue. Brender & Israel (2014) examined alternative models for the prediction of government tax revenues. Using the Engle-Granger method, their study pointed out that the main macroeconomic variables GDP, change in GDP, real wage employee, import of consumption goods, sales of new dwellings, sales of shares by parties, and credit rates affected government tax revenue. Sapiei et al. (2014) focused on the determinants of taxpayer compliance behavior with respect to corporate income tax reporting requirements utilizing a regression analysis, and they found that business size was a significant determinant of tax non-compliance behavior.

Andrejovská & Puliková (2018) used a regression analysis and pointed out the positive relationship between macroeconomic determinants (statutory corporation, GDP, ETR, LoE, IN, PD, and FDI) and tax revenues from corporate tax. At the same time, it pointed out some decisive factors, including employment rate, GDP, and foreign direct investment.

Lismont et al. (2018) developed a tax avoidance prediction model, and the results showed that random forest performed better than logistic regression and decision tree. Another author, also using econometric modeling regression, noticed a direct relationship between tax revenue and gross value added (GVA) and found that consumption was a very good predictor of tax revenue (Ogneru, 2019). From the perspective of tax audit practice, Tarfa et al. (2020) analyzed the effect of tax audit on revenue generation using panel regression, and the results showed that illegal practice of taxpayers was a significant and negative factor that affected tax revenue generation and that the objective of tax audit, audit technique, and training for taxpayers were significant and positive factors. Hassan et al. (2021) examined the relationship between governance and tax revenue collection, discovered external conflicts to remarkably have the largest effect on tax revenue.

Most of the existing quantitative research studies on tax revenue used statistical approaches such as regression for the purpose of explaining the impacts of independent variables on the dependent variable. To the best of our knowledge, Cezar & Lozano (2020) demonstrated to apply Naïve Bayes classifier, decision tree, random forest, and logistic regression machine learning techniques to predicting service tax crimes using fiscal data.

Therefore, this study contributes to the literature of machine learning models in the context of tax revenue potential prediction, with a focus on microeconomic variables as determinants of revenue performance. It does so by foregrounding three machine learning models (logistic regression, random forest, and decision tree) along with their abilities to predict tax revenue potential. Rather than using macroeconomic data, this research used microeconomic variables from a trigger dataset such as tax ID, trigger data nominal value, distance to tax office, type of taxpayer, media of tax report, type of tax, report status, registered year of taxpayer, and area coverage to predict tax revenue potential.

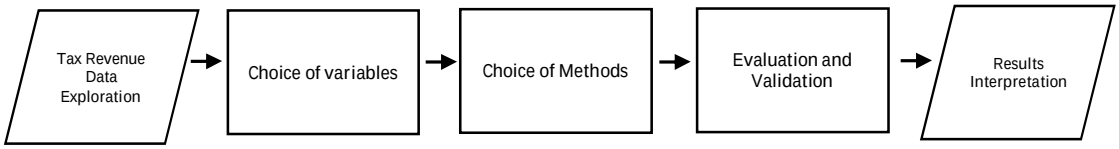
RESEARCH METHOD

Figure 2 depicts the research design of this study, which adopted the standard predictive analytics model development framework by Shmueli & Koppius (2011). The research consisted of five stages, namely exploring the data, selecting relevant predictor variables, determining the potential prediction model, evaluating, validating, and selecting the best model, and finally reporting the research results.

Tax Revenue Data Exploration

In this study, we utilized a sample of taxation data from small tax offices (microeconomic data level). A small tax office (kantor pelayanan pajak pratama) is a regional tax office that provides numerous tax services such as counseling, supervision, and law enforcement for taxpayers

Figure 2 Research Design



Source: Shmueli & Koppius (2011)

Table 2 Dataset Overview

No	Variables	Data Type	Definition	Data Code	n	%
Predictor Variables						
1	Trigger Data Nominal Value	Numerical	Nominal value of the tax potential trigger data of each taxpayer	-	5,562	100%
2	Distance to Tax Office	Categorical	Distance of the taxpayer's residence (district) to the tax office (in kilometer)	< 20 km > 20 km	3,237 2,325	58.20% 41.80%
3	Type of Taxpayer	Categorical	Type of taxpayer based on subjective and objective norms	Individual Corporate	4,881 681	87.76% 12.24%
4	Type of Tax	Categorical	Type of tax according to the object taxed	Income Tax Value-added Tax	5,388 174	96.87% 3.13%
5	Media of Tax Report	Categorical	Channel of annual tax report used	E-filing (online) Manual (offline)	4,709 853	84.66% 15.34%
6	Report Status	Categorical	Status of payment of annual tax report	Null Overpayment Underpayment	5,388 211 13	95.97% 3.79% 0.23%
7	Registered Year of Taxpayer	Categorical	Year of taxpayer registered	Long registered (before 2018) New registered (after 2018)	4,463 1,099	80.24% 19.76%
8	Area Coverage	Categorical	Total area of taxpayer residence (district) in square kilometers	< 25 km^2 > 25 km^2	4,514 1,048	81.16% 18.84%
Predicted Variable						
9	Tax Revenue Potential	Categorical	Target (dependent) variable, whether the data resulting in tax revenue potential or not	Tax Revenue Potential No Tax Revenue Potential	128 5,434	2.30% 97.70%

Source: author's data

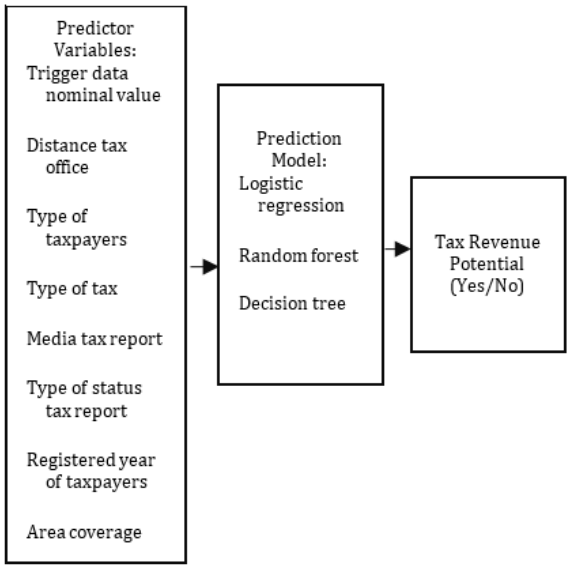
(WP) in the fields of income tax (PPh) and value-added tax (PPN). It is also responsible to collect tax data and ensure their validity in its region. The study used a sample of trigger tax data from 2018 to 2020, consisting of 5,562 rows and 10 columns, from taxpayers who resided in 26 districts. The data size in this study will eventually grow to big data as the period and the regional coverage of the data is expanding.

Table 2 portrays the overview of the dataset. To gain an understanding of the dataset, data exploration was conducted mainly to explore the values within each independent variable. The results showed that there was no difference in each independent variable, in which case most variables had 5,562 rows of data, indicating no missing values that required addressing. The predictand (i.e., tax revenue potential) also needed to be examined as to whether the numbers in two categories (potential and no potential) were balanced or not. Table 2 shows the counts for both potential and no potential categories. The absolute count of the predict and indicated a class imbalance in that the tax revenue potential count was only 128/5,434 or 2.30%. This issue is considered to be normal as this type of data (e.g., transaction, credit score) will not always result in balanced amounts (Brown et al., 2017).

Choice of Variables

Using this dataset, Figure 3 shows the framework of the prediction models. There were

Figure 3 Research Framework to predict tax revenue potentialpotential



numerical and categorical attributes or data features in this study: "Trigger Data Nominal Value", "Distance to Tax Office", "Type of Taxpayer", Type of Tax", "Media of Tax Report", "Report Status", "Registered Year of Taxpayer", "Area Coverage", and "Tax Revenue Potential".

We narrate the description of the prediction models' predictors as follows. First, trigger data nominal value is the nominal amount of tax potential that has not been followed up yet by

Account Representative to result in tax revenue potential. Distance to tax office is the distance between the taxpayer's location and the tax office in which it is registered. Type of taxpayer is the type of the taxpayer i.e., corporate or individual. Media of tax report is the channel through which the taxpayer submits annual tax return, either via E-filing through an online channel or over the counter, directly to the tax office or offline channel. Report status is the status of annual tax return of each registered taxpayer, which can be either one of these three: null, suggesting that the tax payment already meets the rule and rate; underpayment, suggesting that there is still an amount of the tax due to be paid; and overpayment, suggesting that the amount of tax paid exceeds the rule and rate.

Registered year of taxpayer is the first date on which the taxpayer received their Tax ID (NPWP) and on which obligations and rights of taxation first emerged. Lastly, area coverage refers to the district where the taxpayer lives. Tax revenue potential as predictand is categorized as YES if the trigger data have been followed up by Account Representative and generating actual realization of tax payment.

These features would be used to classify tax revenue potential based on the logistic regression, random forest, and decision tree models. Before developing the classification model, the taxpayer dataset was split into training and testing data. The training dataset was used for training the classification model, while the testing dataset was used to test the prediction performance. In this study, the dataset was split into standard composition of 70% training and 30% testing (Vrigazova, 2021). Data sources for all variables are presented along with the variable descriptive statistics (Table 3). Trigger data nominal value has the highest value of 81,164,951,203. The distance tax office has the code "1" for the taxpayer who lives at less than 20km from the tax office and code "2" for who lives more than 20km from tax office. Then for type of taxpayers, "1" is for the corporate type of taxpayer and "2" for the individual taxpayer. The type of tax code "1" for income tax and "2" for value-added tax. For media of tax report, code "1" is manual report and "2" is online or via e-filing. For the report status code "1" is for the status of null, "2" for underpayment, and "3" for overpayment. For the registered year of taxpayer, code "1" for who has registered before 2018 and "2" for who registered after 2018. And the last area coverage, code "1" for area width more than 25km^2 and "2" for area width less than 25km^2. And for the predictand variables, code "1" for there is no tax potential and code "2" for there is tax potential.

Choice of Methods

The focus of this study was to develop prediction models with binary classification that can give accurate predictions on whether the trigger data of tax revenue potential will be resulting in tax revenue or not. The classification models learned in three algorithms, namely logistic regression, random forest, and decision tree.

Logistic Regression

Logistic regression was used to predict the class (or category) of individuals based on one or multiple predictor variables (x). It was used to model a binary outcome, i.e., a variable, which could have only two possible values: yes/no or 0/1.

Logistic regression belongs to a family named Generalized Linear Model (GLM), developed for extending the linear regression model to other situations. The predictor variables in logistic

Table 3 Descriptive Statistics

	Mean	Std. dev	Min	Max	Data source
Predictors variables					
Trigger Data Nominal Value	210,801, 921	-	100	81,164,951,203	Pratama Tax office
Distance to Tax Office	1.42	0.49	1	2	Pratama Tax office
Type of Taxpayer	1.87	0.33	1	2	Pratama Tax office
Type of Tax	1.03	0.20	1	2	Pratama Tax office
Media of Tax Report	1.84	0.40	1	2	Pratama Tax office
Report Status	1.04	0.20	1	3	Pratama Tax office
Registered Year of Taxpayer	1.20	0.40	1	2	Pratama Tax office
Area Coverage	1.81	0.40	1	2	The Central Bureau of Statistics
Predicted variable					
Tax Revenue Potential	1.02	0.20	1	2	Pratama Tax office

Source: author's data

regression can be categorical or numerical. The predicted variable of logistic regression is binary or dichotomous. Logistic regression may have several weaknesses. It can often compete with other machine learning methods, such as neural networks, support vector machine, random forest, and gradient boosting (Nusinovici et al., 2020). The logistic regression function can be written as follows (Peng et al., 2002):

$$\pi = Probability(Y = outcome\ of\ interest \mid X = x, a\ specific\ value\ of\ X) = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n}} \quad (1)$$

where:

- π is the probability of the outcome of interest
- e is 2.71828 (the base of the system of natural logarithms)
- α is Y intercept
- β_n is the regression coefficients
- x_n is set of predictor variables ("Distance to Tax Office", "Type of Taxpayer", "Type of Tax", "Media of Tax Report", "Report Status", "Registered Year of Taxpayer", and "Area Coverage")
- Y is the class of tax revenue potential.

Random Forest

A random forest is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently, with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them (Breiman, 2001).

A random forest uses the law of large numbers, so it does not overfit and can be good for prediction (Breiman, 2001). Furthermore, it can be utilized for any dataset since it does not need an assumption of distribution. The formalization of the random forest classifier is stated as follows (Izquierdo-Verdiguier & Zurita-Milla, 2020):

$$\widehat{Y}_l = mode_{n=1\ldots N_{trees}} \widehat{Y}_n \quad (2)$$

where:

- $\widehat{Y}_l$ is the score of random forest
- N_{trees} is the total number of trees used in the random forest
- $\widehat{Y}_n$ is the score of a single tree
- $mode$ is the class that most often occurs

Decision Tree

Table 4 Confusion Matrix

		Actual class	
		Positive	Negative
Predicted	Positive	tp	fp
	Negative	fn	tn

Source: author's data

The C4.5 Algorithm is an algorithm used to form a decision tree (Mohankumar, 2016). It considers all the possible tests that can split the data and selects a test that gives the best information gain. This algorithm allows pruning of the resulting decision trees. Thus, it increases the error rates on the training data but, importantly, decreases the error rates on the unseen testing data. It can also deal with numeric attributes, missing values, and noisy data. Decision tree can divide large datasets into smaller records by applying a series of decision rules. The C4.5 algorithm formula is divided into two equations. The first equation is used to find the value of gain:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy\ (Si) \quad (3)$$

where:

- S is set of cases
- A is attributes
- n is the partition number of A
- $|Si|$ is the case number in i^{th} partition
- $|S|$ is the case number

Meanwhile, entropy value can is given below:

$$Entropy\ (S) = \sum_{i=1}^{n} - pi * \log_2 pi \quad (4)$$

where:

- S is set of cases
- n is the partition number of S
- pi is the proportion of S_i to S

Evaluation

Evaluation is done by calculating accuracy in a confusion matrix. As shown in Table 4, the confusion matrix classifies problems in two classes. Therefore, there are four possible different results that are forecasted.

Really positive and really negative is positive example of the class which is wrongly classified as negative. In the context of research entrance to confusion matrix have the following meaning (Kohavi & Provost, 1998)

- tp is the proportion of positive cases that are correctly identified
- fp is the proportion of negative cases that are incorrectly classified as positives

- fn is the proportion of positive cases that are incorrectly classified as negatives
- tn is defined as the proportion of negative cases that are classified correctly

The following standard terms are defined in a matrix in two classes: accuracy, precision, recall, and F score. Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy may be determined using the equation below:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + tn} \quad (5)$$

Then, precision or positive predictive value shows the fraction of predictive positive cases that are accurate and is calculated using the equation below:

$$Precision = \frac{tp}{tp + fp} \quad (6)$$

Recall is the proportion of positive cases that are properly identified. It can be calculated using the equation below:

$$Recall = \frac{tp}{tp + fn} \quad (7)$$

Finally, F Score is used to seek a balance between precision and recall when there is an uneven class distribution (large number of actual negatives). It is also used to evaluate a model: the higher the F score, the better. A score of 0 is the worst possible, and 1 is the best. F score can be calculated using the equation below:

$$F\ Score = 2\,x\,\frac{precision * recall}{precision + recall} \quad (8)$$

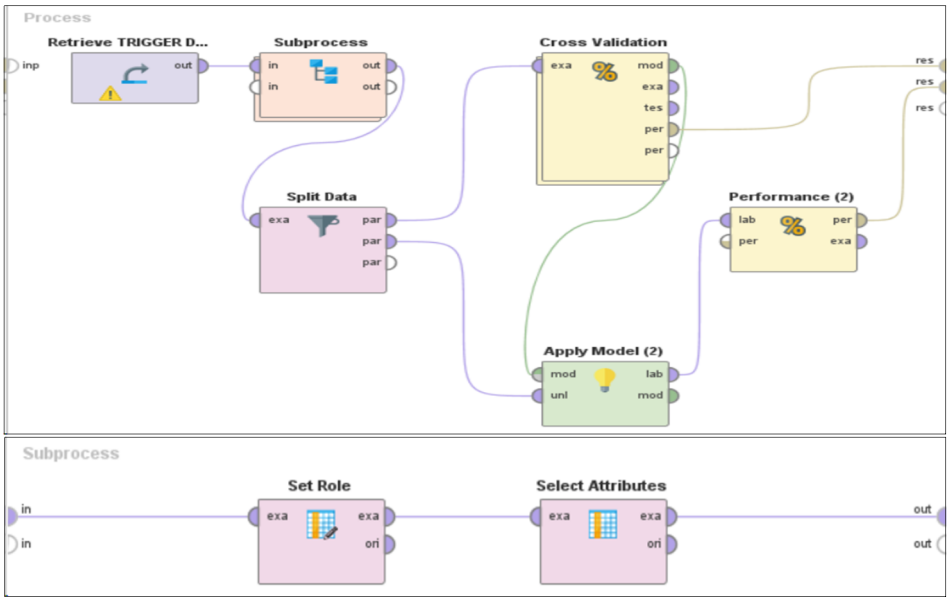
Implementation of the models using RapidMiner

We implemented the proposed classification models using free RapidMiner data science and machine learning platform. RapidMiner Studio software utilizes predictive data analysis and descriptive data analysis methods to provide every user with information and knowledge in the hope that they can make decisions based on data. This implementation consisted of four main steps, namely data retrieval, feature selection, data split, and model construction.

First, this study started from retrieving data. Data input blocks were modelled by the retrieve operator, as shown in Figure 4. The retrieve operator loaded a Rapid Miner object into the data flow process. In this case, the said data were the" Trigger Data" from small tax offices. The dataset was imported in csv (comma-separated values) format. When importing the data, we verified the data type as either binominal, polynomial, real, or integer.

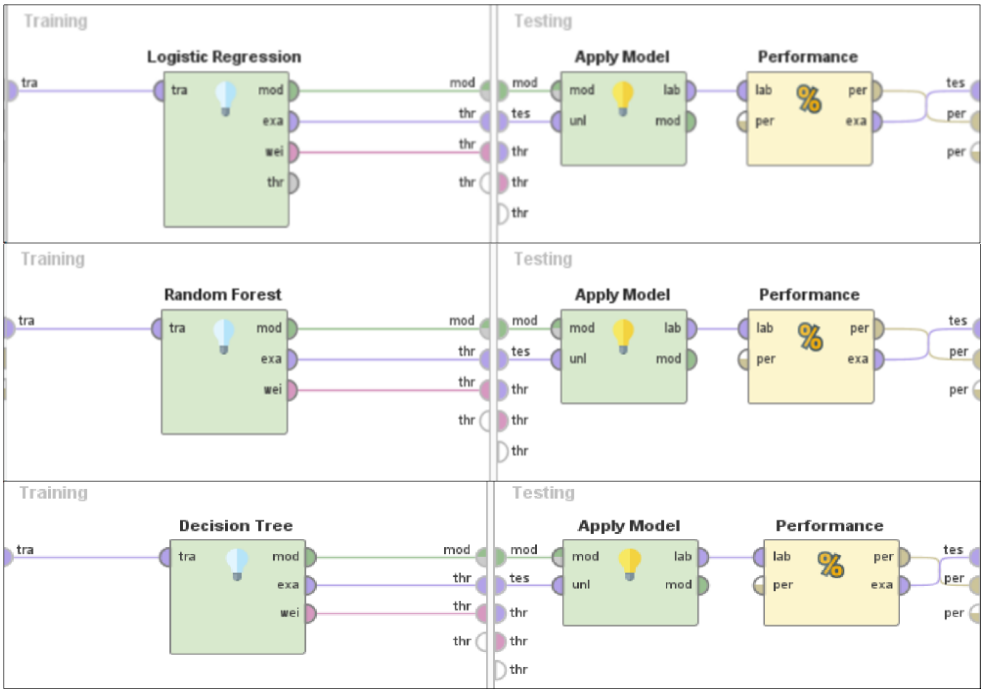
Second, we used the sub-process operator to assign the predictor and predicted the role of each data attribute in the trigger tax dataset (see Figure 4). The data size was then reduced by removing unnecessary attributes. Next, we used the select role operator to change the role of one or more

Figure 4 Retrieval and Splitting of Data



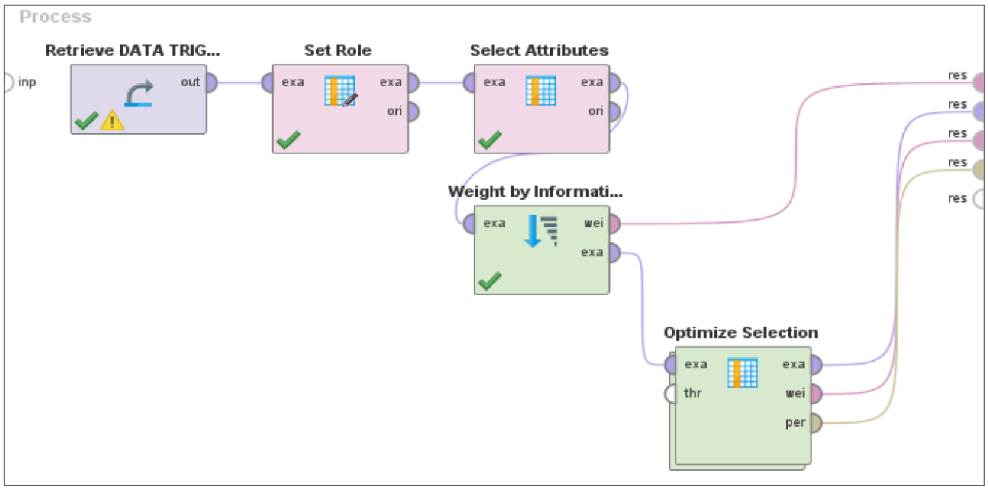
Source: author's data

Figure 5 The Construction of Logistic Regression, Random Forest, Decision Tree Models



Source: author's data

Figure 6 Optimize Selection Operator



Source: author's data

attributes. In this study, we selected tax revenue potential as the label since it would be the predicted variable in the models.

Third, we used the split data operator to split the data by stratified sampling into training and testing datasets in 70% and 30% partitions, respectively (see Figure 4). Lastly, in the modeling stage, we selected the intended algorithms, namely logistic regression, random forest, and decision tree (see Figure 5). The models were validated using the cross-validation operator to assess the prediction performance of the models.

Please note that we conducted feature selection (see figure 6) to select the most relevant attributes of the given dataset for a specific prediction model. In this feature selection step, we used the optimize selection and weight by information gain operators. The optimize selection operator selected subset or attributes from the original attributes or dataset. For examples, we selected several most correlated attributes from among original attributes or dataset with the highest correlation with the label attribute or predictand using certain algorithms (logistic regression, random forest, and decision tree). The optimize selection operator selected the most relevant attributes of the given dataset by trying all possible combinations of attribute selection.

The weight by information gain operator, meanwhile, calculated the relevance of the attributes based on information gain and assigned

weights to them accordingly. We present the weight by information gain from the best model to show the weight of each attribute in predicting tax revenue potential. By knowing the weight of each attribute, we could propose several input actions related to the effort to increase tax revenue potential. Figure 6 shows the optimize selection and weight by information gain operators on RapidMiner that was used to perform feature selection.

RESULTS AND DISCUSSION

We present the results of the prediction in the forms of confusion matrices, which classify the model's prediction outcome into four different states of prediction, namely true positive, true negative, false positive, and false negative.

Comparing the false positive and false negative, the model mostly incorrectly predicted 'tax revenue potential'. This happened due to the imbalance in the dataset, which had a higher proportion of 'no tax revenue potential'. Moreover, it can be inferred from the classification report of logistic regression (Table 5) that there was lower precision potential (row No) for 'tax revenue potential' and higher recall (tp/(tp+fn)) for 'no tax revenue potential', in which case the model predicted most 'no tax revenue potential' correctly.

Table 5 Confusion Matrix of Logistic Regression

	Actual: NO Tax Potential	Actual: YES Tax Potential
Predicted: NO Tax Potential	1622	29
Predicted: YES Tax Potential	8	9
Accuracy: 97.78%		
Classification Report		

	Precision	Recall	F score	Support
No	98.24%	99.51%	98.24%	1,651
Yes	52.94%	23.68%	32.72%	17
Macro Avg	75.59%	61.60%	65.80%	1,668

Source: Output Rapidminer Studio

The results of logistic regression presented in Table 5 where all independent variables were used show a 97.78% accuracy level, a 75.59% precision level, a 61.60% recall level, and an F score of 65.80%. Table 5 shows the confusion matrix and classification report for logistic regression.

Next is the analysis results of decision tree (see Table 6). The accuracy obtained with the decision tree model was 97.54%, which was slightly lower than logistic regression. The precision was 73.98%, the recall was 55.14%, and the F score for the decision tree model was 58.12%. Overall, the confusion matrix of the decision tree model performed lower than the logistic regression model. Table 6 shows the confusion matrix and classification report for decision tree.

Table 6 Confusion Matrix of Decision Tree

	Actual: NO Tax Potential	Actual: YES Tax Potential
Predicted: NO Tax Potential	1626	34
Predicted: YES Tax Potential	4	4
Accuracy: 97.54%		
Classification Report		

	Precision	Recall	F score	Support
No	97.95%	99.75%	98.84%	1,660
Yes	50.00%	10.53%	17.40%	8
Macro Avg	73.98%	55.14%	58.12%	1,668

Source: Output Rapidminer Studio

Table 7 depicts the prediction results of the random forest model. The accuracy obtained with random forest was 98.14%, which was the highest of all the three models. The precision, recall, and F score also showed the best results of all the three models. This means that random forest performed best in terms of accuracy, precision, recall, and F Score of all the three learning models. The high accuracy value proved that the data used were good, and the high F score indicated that the model was also good. Accuracy itself is the percentage of true positive and true negative in the overall prediction. F score measures a model's accuracy on a dataset by combining the precision and recall of the model. It is otherwise defined as the harmonic mean of the model's precision and recall. The confusion matrix of random forest can be seen in Table 7.

Table 7 Confusion Matrix of Random Forest

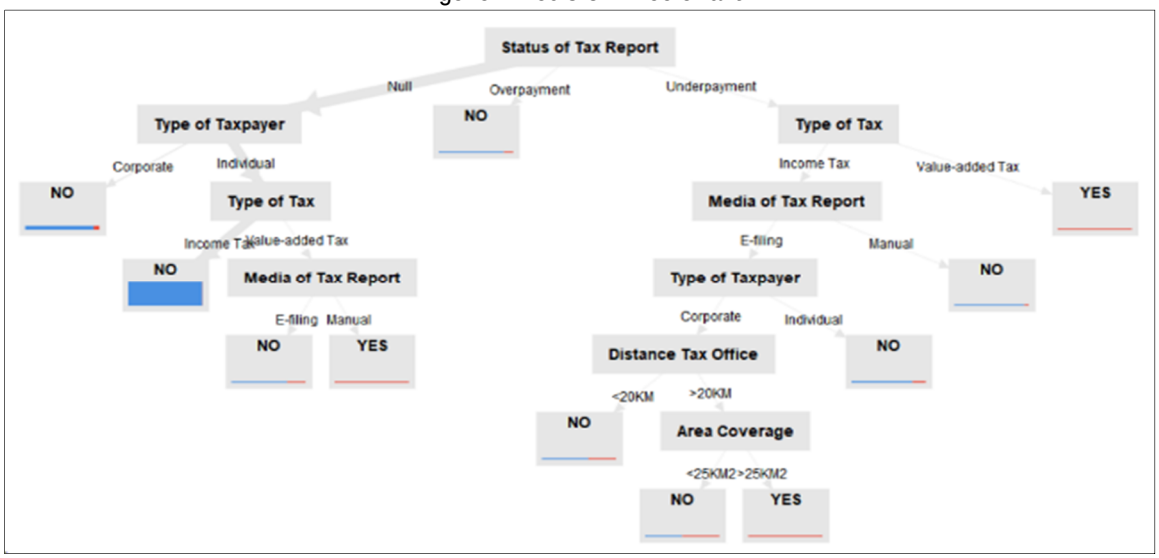
	Actual: NO Tax Potential	Actual: YES Tax Potential
Predicted: NO Tax Potential	1625	26
Predicted: YES Tax Potential	5	12
Accuracy: 98.14%		
Classification Report		

	Precision	Recall	F score	Support
No	98.43%	99.69%	99.06%	1,651
Yes	70.59%	31.58%	43.64%	17
Macro Avg	84.51%	65.64%	71.35%	1,668

Source: Output Rapidminer Studio

Table 8 shows the final comparison results of accuracy, precision, recall, and F score between the logistic regression, random forest, and decision tree classification models. The random forest model showed the best performance of the three models. It provided the best accuracy, precision, recall, and F score of 98.14%, 84.51%, 65.64%, and 71,35%, respectively. These findings showed that random forest, which employed an ensemble classifier, outperformed other singular classifiers (i.e., logistic regression and decision tree). Random

Figure 7 Decision Tree Chart



Source: Output Rapidminer Studio

forest demonstrated several advantages, for instance, robustness against overfitting and dealing with high-dimensional problems (Izquierdo-Verdiguier & Zurita-Milla, 2020). Furthermore, as a non-parametric method, random forest did not require distributional assumption for the training dataset. Nonetheless, the complexity and time taken in constructing the random forest model increased as the numbers of trees and training samples increased. Therefore, using a cross-validation procedure, we concluded that random forest performed best. Next, Figure 7 visualizes the decision tree chart of the random forest model. From the chart, we can view the level of significance of each predictor variable. We can see that report status was significant at predicting tax revenue potential. Furthermore, Table 9 shows the results of feature selection using the optimize selection and weight by information gain operators and using the best model, which was random forest model. According to the weights of the attributes, report status was the top tier of variable importance in predicting tax revenue potential. This means that, in gaining tax revenue potential, Account Representative can classify trigger tax potential data based on the status of tax first. The following factors were trigger data nominal value and type of taxpayer, meaning that after finding the list of taxpayers based on the status of annual tax report, we can sort the data based on the nominal value of trigger data and then classify them by type of taxpayer. Of all attributes, media of tax report was the least influential factor.

Table 8 Final Comparison between Logistic Regression, Random Forest, and Decision Tree

Aspect	Logistic Regression	Random Forest	Decision Tree
Accuracy	97.78%	98.14%	97.54%
Precision	98.24%	98.43%	97.95%
Recall	96.51%	99.69%	99.82%
F Score	65.80%	71.35%	58.12%

Source: Output Rapidminer Studio

Table 9 Weights of Attributes Using The Random Forest Model

Attributes	Weight
Media of Tax Report	0.00011
Area Coverage	0.00012
Distance to Tax Office	0.00209
Type of Tax	0.00291
Registered Year of Taxpayer	0.00326
Type of Taxpayer	0.02352
Trigger Data Nominal Value	0.02948
Report Status	0.02981

Source: Output Rapidminer Studio

CONCLUSION

A machine-learning based predictive analytics that operated on a small tax office has proven an applicable solution for AR to increase their performance and to the question of how to use tax trigger data to realize unexplored tax potential. It is also a fact that random forest model has better performance than the logistic regression and the decision tree models. The operation on the small tax office shows random forest has 98.14% accuracy while others have 97.78% and 97.54% respectively. It shows that the random forest model provided the best prediction performance than the logistic regression and decision tree. Given the size of the data records, this approach can provide tax administrators with a powerful tool to increase their work efficiency, combat tax evasion, and provide better customer service (Strømme, 2018). We analyzed a total of 5,562 tax revenue potential data with 8 predictors, namely trigger data nominal value, distance to tax office,

type of taxpayer, media of tax report, type of tax, report status, registered year of taxpayer, and area coverage.

From a practical point of view, the proposed machine learning model can support tax officers in predicting their tax revenue potential. They can improve their way to find potential sources of tax revenue and do tax extensification in gaining new sources of tax revenue. The results of this study could determine the predictors of tax revenue potential that could help tax officers in finding out about tax revenue potential and give a reference to the classification method that yields the best result to predict tax revenue potential. To the best of our knowledge, this study is one of the earliest studies to predict tax revenue potential using machine learning approaches.

This study has several limitations that open up opportunities for further research. First, this study only focused on small tax office data; thus, one may attempt to develop predictive models for wider regional contexts, including several cities and more small tax offices. Second, this paper focused on eight constructs as the predictors for the machine learning models. Further research may attempt to use more advanced data pre-processing and machine learning approaches.

ACKNOWLEDGEMENTS

The first author sends his gratitude for the research permission given by the anonymous tax office administration (letter number: S-387/RISET/PJ.09/2022). The first author received a master's degree scholarship from the Indonesia Endowment Fund for Education LPDP (scholarship no: S-387/RISET/PJ.09/2022)

REFERENCES

Alarie, B., Niblett, A., & Yoon, A. H. (2016). *Using Machine Learning to Predict Outcomes in Tax Law*, , no. 1–23. University of Toronto, Faculty of Law.

Andrejovská, A., & Puliková, V. (2018). Tax revenues in the context of economic determinants. *Montenegrin Journal of Economics, 14*(1), 133–141. https://doi.org/10.14254/1800-5845/2018.14-1.10.

Breiman, L., (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:101093340432.

Brender A., & Navon, G. (2010). *Predicting Government Tax Revenues and Analyzing* Forecast Uncertainty. *Israel Economic Review, Bank of Israel,7*(2), 81-111.

Brown, I., Brown, I., & Mues, C. (2017). *An experimental comparison of classification algorithms for imbalanced credit scoring data sets expert systems with applications. Expert Systems With Applications, 39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033.

Cezar, A., & Lozano, G. (2020). *Tax crime prediction with machine learning : A case study in the municipality of são paulo. Proceedings of the 22nd International Conference on Enterprise Information Systems, 1*: ICEIS, 452–459. https://doi.org/10.5220/0009564704520459.

Hariani A. (2021). *WP terdaftar naik 20 kali lipat di 20 tahun terakhir*. Retrieved December 13, 2021, from https://www.pajak.com/pajak/wp-terdaftar-naik-20-kali-lipat-di-20-tahun-terakhir/.

Hassan, M. S., Mahmood, H., Tahir, M. N., Yousef Alkhateeb, T. T., & Wajid, A. (2021). Governance: A source to increase tax revenue in Pakistan. *Complexity*, 6663536. https://doi.org/10.1155/2021/6663536.

Izquierdo-Verdiguier, E., & Zurita-Milla, R. (2020). *An evaluation of guided regularized random forest for classification and regression tasks in remote sensing.* I*nternational Journal of Applied Earth Observation and Geoinformation, 88* (October 2019), 102051. https://doi.org/10.1016/j.jag.2020.102051.

Javid, A. Y., & Arif, U. (2012). *Analysis of revenue potential and revenue effort in developing asian countries. Winter*, 365–380.

Kohavi, R. & Provost, F. (1998). Glossary of terms. in glossary of terms. machine learning—Special issue on applications of machine learning and the knowledge discovery process. *Machine Learning*, 30, 271–274. https://doi.org/10.1177/14034948135151 31.

Lismont, J., Cardinaels, E., Bruynseels, L., De Groote, S., Baesens, B., Lemahieu, W., & Vanthienen, J. (2018). *Predicting tax avoidance by means of social network analytics. Decision Support Systems*, 108, 13–24. https://doi.org/10.1016/j.dss.2018.02.001.

Mohankumar M., Amuthakkani S. & Jeyamala G. (2016). *Comparative analysis of decision tree algorithms for the prediction of eligibility of a man for availing bank loan. International*

Journal of Advanced Research in Biology Engineering Science and Technology, *2*(15), 360–366.

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., & Cheng, C. Y. (2020). *Logistic regression was as good as machine learning for predicting major chronic diseases*. *Journal of Clinical Epidemiology*, *122*, 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002.

Ogneru, V. (2019). Analysis of the relationship between tax revenue and gross value added in the Romanian economy. *Financial Studies*, *23* (2(84)), 37–55. http://hdl.handle.net/10419/231676.

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). *An introduction to logistic regression analysis and reporting*. *Journal of Educational Research*, *96*(1), 3–14. https://doi.org/10.1080/00220670209598786.

Petutschnig, M. (2017). *Future orientation and taxes: Evidence from big data*. *Journal of International Accounting, Auditing and Taxation*, *29*, 14–31. https://doi.org/10.1016/j.intaccaudtax.2017.03.003.

Sapiei, N. S., Kasipillai, J., & Eze, U. C. (2014). *Determinants of tax compliance behaviour of corporate taxpayers in malaysia*. *EJournal of Tax Research*, *12*(2), 383–409. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84920280986&partnerID=40&md5=08271e486a55d4b29dc37779388fe01c.

Shmueli, G., & Koppius, O. R. (2011). *Predictive analytics in information systems research*. *MIS Quarterly: Management Information Systems*, *35*(3), 553–572. https://doi.org/10.2307/23042796.

Strømme, Ø. (2018). *Increased compliance and efficiency with machine learning*, (Issue June, 2018), 50-52, Budapest, General Assembly of IOTA, WWW.IOTA-TAX.ORG

Tarfa, G.E. ; Tarekegn, G & Yosef, B. (2020). E*ffects of tax audit on revenue generation*. *Journal of International Trade, Logistics and Law*, *6*, 65–74.

Vrigazova, B. (2021). *The proportion for splitting data into training and test set for the bootstrap in classification problems*. *Business Systems Research*, *12*(1), 228–242. https://doi.org/10.2478/bsrj-2021-0015.