

Improving the risk profile of Indonesian enterprise taxpayers using multilabel classification

Teguh Prasetyo, Budi Susetyo, Anang Kurnia

Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, Indonesia

Article Info	ABSTRACT
<p>Article history:</p> <p>Received Nov 28, 2023 Revised Feb 29, 2024 Accepted Mar 21, 2024</p> <p>Keywords:</p> <p>Deep learning Feature importance Machine learning Multilabel classification Tax revenue</p>	<p>Optimizing tax revenues is difficult in Indonesia due to obstacles such as tax evasion and tax avoidance. It is closely related to an organization's compliance with tax regulations, known as the taxpayers risk profile. However, this mechanism does not accurately detect tax avoidance and tax evasion risks. To overcome this limitation, we use a multilabel classification machine learning method in this study, which classifies a single observation into one or more labels at once. The approach involves problem transformation (binary relevance and label powerset), algorithm adaptation (multilabel k-nearest neighbor (ML-kNN) and multilabel-adaptive resonance associative map (ML-ARAM)), and ensemble (label space partitioning and random k-label sets with disjoint (RAkELd)). Based on the model performance comparisons, we discovered that the ML-ARAM method based on deep learning is the best, with an average F1-score of 95.5% and a hamming loss of 7.4%. We also examine the feature importance of the best model to reduce the dimensions of features so that we can identify the dominant factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion. The findings of this study improve the accuracy of tax avoidance risk detection and tax evasion risk profiles using machine learning methods, ensuring maximum tax revenues in Indonesia.</p> <p><i>This is an open access article under the CC BY-SA license.</i></p>



<p>Corresponding Author:</p> <p>Teguh Prasetyo Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University Bogor, Indonesia Email: teguhprasetyo@apps.ipb.ac.id</p>

1. INTRODUCTION

Taxation is Indonesia's most significant state revenue, which is critical for economic development. In practice, taxpayers engage in tax avoidance and tax evasion efforts to avoid or even fail to pay tax obligations, resulting in suboptimal tax revenues in Indonesia [1]. According to Sinaga *et al.* [2], tax avoidance occurs in Indonesia by reducing profit margins to avoid applicable tax rates or by exploiting legal loopholes to reduce tax liabilities. Meanwhile, tax evasion, according to Postea [3], is used by deception that not legal. The Indonesian government is attempting to reduce tax avoidance and evasion in several ways, including increasing tax compliance through the taxpayer risk profile. The more compliant an entrepreneur is, the greater the opportunity for tax revenue to be generated. The obstacle in the field is that the current risk profile mechanism fails to accurately detect tax avoidance and tax evasion risks, making it ineffective and inefficient to implement. This leads to suboptimal tax revenue in Indonesia.

Several previous studies used machine learning methods [4]–[6] to identify the risk of tax avoidance and tax evasion. Those previous studies examine the use of classification methods in detecting tax avoidance and tax evasion risks, as well as predict regional development project policies in several countries using binary-class classification (BCC) method. However, the risks of tax avoidance and tax evasion in Indonesia

are diverse, with more than two label categories. Furthermore, because a single taxpayer entity can engage in one or more types of tax avoidance or tax evasion simultaneously, BCC modeling cannot be used. To address this limitation, we present a modeling solution in this study that employs the multilabel classification (MLC) machine learning method, which can model two or more label classes simultaneously. This study not only improves the risk profile for detecting tax avoidance and tax evasion in Indonesia, but it also investigates the dominant factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion.

The classification method is used to model response variables as label class categories, which are widely used in various fields, including customer target marketing, disease diagnosis, document classification, and social network analysis [7]. One observed value is typically classified into one class or label on the response variable, called BCC or multi-class classification (MCC). The MLC method associates a single observation with multiple labels [8]. Several applications of the MLC method, among others, were conducted in [9]–[12], which were limited only to modeling and method comparison. This study has yet to examine data limitations such as class imbalance in the MLC model and how to address it. In field conditions, data indicating class imbalance are frequently found, as evidenced by this study's empirical findings. In addition to comparing methods and identifying risk factors for tax avoidance and tax evasion, this study discusses how to deal with class imbalance, which can help with the implementation of the MLC model in real-world cases.

We use empirical data from the Indonesian Ministry of Finance on historical tax avoidance and tax evasion data from 2018 to 2022. In this study, we compare six MLC methods to determine which model performs the best based on F1-score and hamming loss. Using the best model, we identify the risk element factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion based on the feature importance value. Next, we use these dominant factors to reduce risk elements, making them easier to interpret and decide. According to this explanation, using the MLC machine learning method in this study will provide a solution to improve the accuracy of tax avoidance and tax evasion risk detection, resulting in increased tax revenue for Indonesia. The added value is that this research will reveal the dominant factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion. It will be used to enhance Indonesia's tax risk assessment mechanism.

2. METHOD

2.1. Dataset

We use data from 61 risk elements and explanatory variables (features) in three main categories and five types of tax avoidance and tax evasion response variables carried out by taxpayer entities. The data type is a numeric score (0-100) with a total number of observations of 498 entities, as explained in Table 1. According to [13], the resampling-bootstrap method can provide accurate model performance results. The algorithm used is nonparametric bootstrap, which is the most straightforward and widely used method [14]. The weight of tax avoidance or tax evasion on the response variable labels in this study is considered equal, so there is no particular order regarding priority labels for the types of those labels.

Table 1. Explanatory variables (features) and target variables

Feature			Target Variables
1. Registration dimension (R): as many as 22 elements related to:	2. Operational dimension (O): as many as 10 elements related to:	3. Local dimension (L): as many as 29 elements related to:	1. Label 1: Fake, secondhand, or no tax stamp documents
a. The types of business entities	a. Current asset ratio	a. Factory security system	2. Label 2: Wrong personalization of tax stamp documents
b. Tax obligations	b. Production ratio	b. Sales transaction system	3. Label 3: Wrong purpose tax stamp documents
c. Accounting system	c. Audit history	c. Distributor list	4. Label 4: Administrative sanctions or fines
d. Production components			5. Label 5: Suspension of licensing or revocation of facilities

2.2. Research method

This study is divided into several stages: data exploration, MLC modeling, model evaluation, model performance comparison, and the identification of dominant tax avoidance and tax evasion factors based on feature importance value. The research framework is depicted in Figure 1. Throughout the research process, we used Python and R software. This decision was made based on the ease of integration with the MLC library and the need for understandable analysis output.

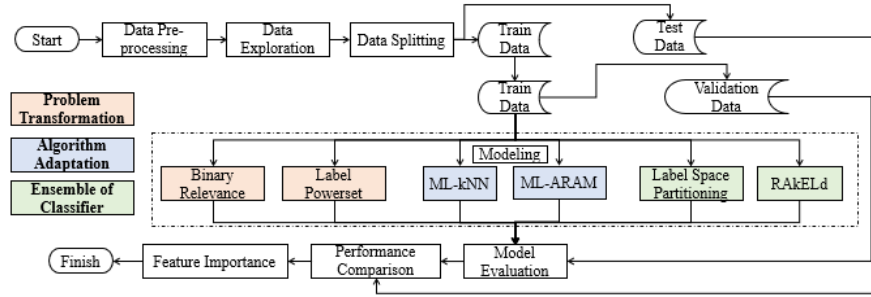


Figure 1. Research framework

2.3. Data exploration

We examined the data to get a broad picture because MLC data differs from BCC and MCC data: a single observation can be classified into multiple label classes [15]. We observed characteristics such as label distribution, label relationships, and label class imbalance [16]. The MLC method measures the label distribution using the cardinality value, which is the average number of active labels or pairs of labels for each observed value, and label density, which is the density value of each label in a lower dimension, described in (1).

$$Card(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i|; Dens(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \quad (1)$$

where D represents the number of observations or instances and L represents the number of labels.

Next, we look for class imbalances in the data. Our results show an imbalance in the number of observations between label classes. We use the imbalance ratio (IRLbI), MeanIR, and coefficient of variation values to determine the magnitude of the class imbalance, as describe in (2).

$$IRLbI(y) = \frac{\argmax_{(y' \in L)} (\sum_{i=1}^{|D|} h(y', Y_i))}{\sum_{i=1}^{|D|} h(y, Y_i)}, h(y, Y_i) = \begin{cases} 1 & y \in Y_i \\ 0 & y \notin Y_i \end{cases} \quad (2)$$

$$MeanIR = \frac{1}{|L|} \sum_{y \in L} (IRLbI(y))$$

Class imbalance in MLC is classified as imbalance within labels, imbalance between multiple labels, and imbalance of all labels [17]. It can be approached in three ways: using data, algorithms, and a combined approach. According to Kurniawati and Prabowo [18], both oversampling and undersampling methods have shortcomings, such as overfitting the model and losing information in observations. Based on this, in this MLC study, imbalances are addressed by stratifying the train and test data (multilabel stratified shuffle split). If the data is randomly split, the train and test data may have different proportions. This is because the MLC's labels are linked and have a relationship [19]. The stratification approach has been shown to improve cross-validation accuracy, bias values, and classification models [20].

MLC exhibits characteristics of one or more correlated label classes. The data we use contains several label sets. We measure this using the score of concurrence among imbalanced labels value (SCUMBLE) [21], explained in (3):

$$SCUMBLE_{ins}(i) = 1 - \frac{1}{IRLbI_{il}} (\prod_{l=1}^{|L|} IRLbI_{il})^{\frac{1}{|L|}} \quad (3)$$

$$SCUMBLE(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} SCUMBLE_{ins}(i)$$

2.4. Multilabel methods

The modeling stage uses the scikit-multilearn module in Python software [22]. The MLC method uses three approaches: problem transformation, method adaptation, and the ensemble of classifiers [23]. We use two methods to compare each approach, considering that each approach has its advantages and disadvantages. The first approach is the problem transformation, which works by converting the MLC model into a simpler model (BCC/MCC) before the modeling stage. In this approach, we use the binary relevance method (transform to binary classification problem) introduced by [24], which works by adapting the one-versus-all approach. The binary relevance method has several advantages, including a simple algorithm and ease of understanding [25]. We also applied the label powerset method (transform to multiclass classification problem) introduced in [26]. The working principle is to convert the MLC model into multiple MCC models. Each pair of observations

on label is treated as a separate observation. According to [27], if there is a strong correlation between labels, the label powerset method is recommended.

The second approach is algorithm adaptation, which works by adapting a classification model algorithm for direct application to the MLC case [23]. This approach relies on artificial neural networks, specifically multilabel-adaptive resonance associative map (ML-ARAM) based on deep learning method. According to [28], the working principle of ML-ARAM is to use adaptive resonance theory (ART) to speed up model classification. [29] explains that the ML-ARAM method employs fuzzy logic principles, with ART being a fuzzy model artificial neural network module with two hidden layers, F1 and F2. The learning process in the ARTa and ARTb layers is described as (4):

$$W_j^{(new)} = \beta(A \wedge W_j^{(old)}) + (1 - \beta)W_j^{(old)} \quad (4)$$

W is the weight of the neural network, A is the input (explanatory variable/features), and β is the learning rate (0,1). We also employ the multilabel k-nearest neighbor (ML-kNN) method proposed by [30], which is the most effective observation-based MLC method. The ML-kNN algorithm uses the k-nearest neighbor (kNN) algorithm in cluster analysis, which is a nonparametric method that uses the close distance between observations. There is a modification of the ML-kNN method that uses the Bayes probability principle, as explained in (5):

$$\vec{y}_t(l) = \operatorname{argmax}_{b \in \{0,1\}} \frac{P(H_b^l) \cdot P(E_{\vec{C}_{t(l)}}^L | H_b^l)}{P(E_{\vec{C}_{t(l)}}^L)} \quad (5)$$

Where t is a new dataset, H_b^l is an event whether t has a label or not, and $E_{\vec{C}_{t(l)}}^L$ is the number of events as many as K nearest neighbors of t with a label l .

The third approach is the ensemble of classifier, which has the working principle of modeling MLC using several basic problem transformation methods and algorithm/adaptation methods simultaneously, such as boosting or bagging. In this approach, we use the label space partitioning classifier method, with the working principle of dividing pairs of response variable observations into several small clusters and building a multiclass classification model using the label space clusterer method separately for each pair of observations using a hybrid partition approach [31]. Then, we also used the random label space partitioning with label powerset method (RAkELd), which works by randomizing several pairs of observations from labels and building a multiclass classification model using the label powerset method separately for each pair of observations [23]. The RAkELd method has an advantage over the label powerset method: because it can predict label pairs that have not yet to appear in the data. It facilitates the generalization of predictions when the data sample does not yet represent population characteristics [32].

2.5. Model evaluation

We compared the model performance using the cross-validation method by partitioning the data into train and test data in 10 partitions. Furthermore, cross-validation can be used to estimate the model average prediction error when the train and test data are randomly selected from the overall data [33]. Model performance evaluation yields the best model based on the F1-score and hamming loss. The MLC model performance is evaluated differently from the BCC or MCC models due to the different types and structures of response variables [34]. In addition to method selection, model performance is also influenced by data characteristics and label class separation patterns [35]. We use the F1-score, a label-based metric, considering the precision and recall values as (6) [36]:

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}; \text{ or: } F1 - Score = \frac{1}{N} \sum_{i=1}^N \frac{2|Y_i \cap \hat{Y}_i|}{|\hat{Y}_i| + |Y_i|} \quad (6)$$

N represents the number of observations in the data, Y_i is the true/actual label value, and \hat{Y}_i is the predicted value. The intersection of Y_i and \hat{Y}_i indicates the number of labels predicted correctly by the model. The higher the F1-score value, the better the model performance [37]. The F1-score, according to [38] which considers precision and recall as measures of model goodness, is robust to indications of imbalanced labels in the data. In addition to the F1-score, we use the hamming loss, an Instance-based metric, by evaluating the number of classification errors of the label/class response variable pairs from the resulting model, as formulated in (7):

$$Hamming Loss = \frac{1}{N} \sum_{i=1}^N \frac{|Y_i \neq \hat{Y}_i|}{M} \quad (7)$$

M represents the maximum number of labels on the response variable. The fewer mispredict labels, the lower the hamming loss, and the higher the model performance [39].

2.6. Model performance comparison

We compare the model performance using the analysis of variance method. The MLC method and data separation method is compared by the mean value of the MLC model performance (F1-score and hamming loss). The comparison ensures that the diversity between groups is heterogeneous using the analysis of variance method [40], as formulated in (8):

$$y_{ij} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ij} \quad (8)$$

τ_i denote the MLC methods and β_j are the data separation methods.

2.7. Feature importance

In the following stage, we identify risk element factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion using the feature importance value. [41] explained that machine learning models must be interpreted, even though some models have varying degrees of difficulty. Some methods, such as tree-based methods, are straightforward to understand. This study employs the permutation feature importance value, which describes the effect of each explanatory variable (features) on the prediction results. The advantage of this approach is that the interpretation is simple to understand [42]. According to [43], the permutation method is widely used to determine the feature importance of a machine learning model. Machine learning models that use high-dimensional data become more complex. Simple models are easier to interpret in decision-making, especially when considering good model performance. Dimensionality reduction is a method for converting high-dimensional data into lower-dimensional data, that employs two approaches: feature selection and feature extraction [44]. Next, we use these dominant factors to reduce risk elements, making them easier to interpret and decide.

3. RESULTS AND DISCUSSION

Our findings are organized into three: data exploration, model evaluation, and performance comparison, as well as feature importance and selection. This study aims to address the study objectives, which include dealing with class imbalance, comparing MLC methods, and identifying dominant factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion. We provide a discussion of each analysis result to support the study findings.

3.1. Data exploration

Among the five classes of data labels, label 4 dominates with a frequency of more than 200 observations, while label 1 has the fewest. Meanwhile, the label pairs with the highest frequency of observations are label 4 and 2, as well as the pair of labels 2 and 4. Labels with more pairs tend to have a lower frequency of observations. The empirical data contains 21 pairs of labels (single and paired labels). As the number of label pairs, the MLC modeling stage become more complicated. Figure 2 depicts the label distribution in empirical data. A data cardinality value of 1.04 indicates that one observation results in an average of one to two labels. A data density value of 0.21 indicates that on a scale of 1–10, 21% of labels have the greatest number of observations. The higher the cardinality and density values, the more label categories are generated from a single data observation, which complicates the MLC modeling stage.

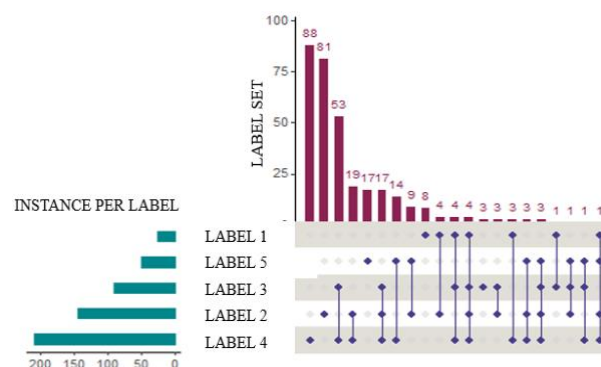


Figure 2. Frequency of each label and label pairs for empirical data

Figure 3 illustrates class imbalance comparing the number of observations between label classes. We can see visual signs of class imbalance on four of the five labels in the data. Category 0 (does not engage in tax avoidance or tax evasion) dominates all labels. Label 1 has the highest IRLbl value at 8.40. Meanwhile, the MeanIR data value is 3.49, meaning the average IRLbl across all labels is one versus 3.49.

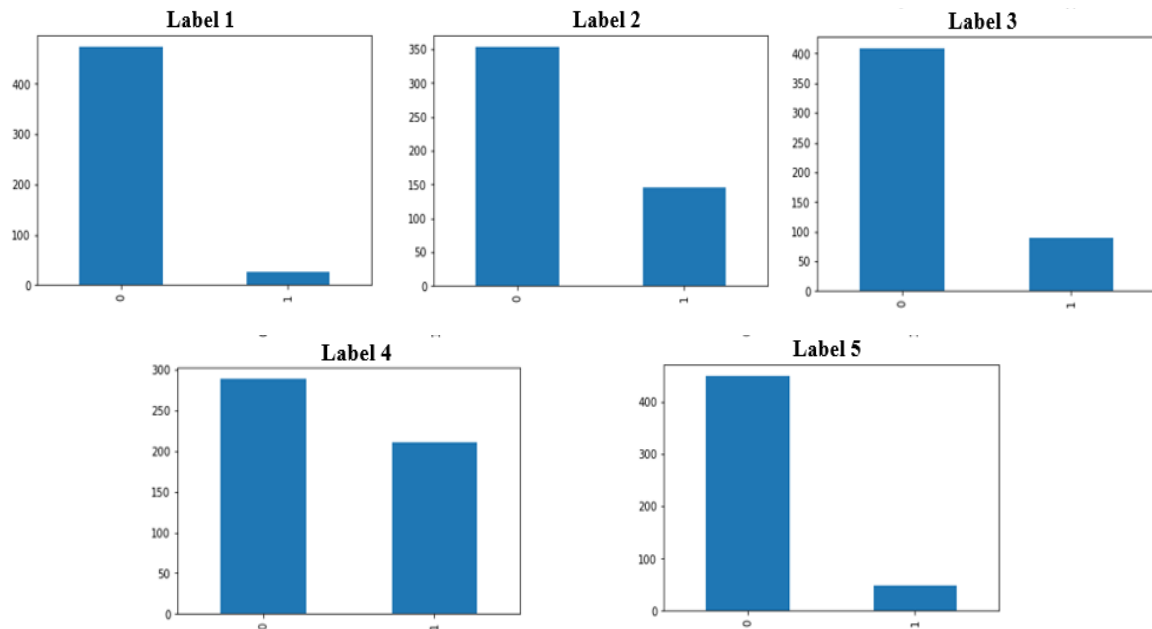


Figure 3. Comparison of the number of class observations between labels

We discovered that the label class with the highest relationship value was label 2, followed by label 4 with SCUMBLE(i) values of 1.82 and 1.35. We calculate the average value of the relationship between labels with the SCUMBLE(D) value, where the higher the SCUMBLE(D) value, the more complicated the MLC modeling. The SCUMBLE(D) value in the data is 0.03, which is not particularly high compared to other multilabel data types, such as images and text data. The higher the SCUMBLE value, the more observations produce the label simultaneously. Figure 4 depicts the relationship between labels as seen through the number of observations, with label 4 leading the label with the most observations.

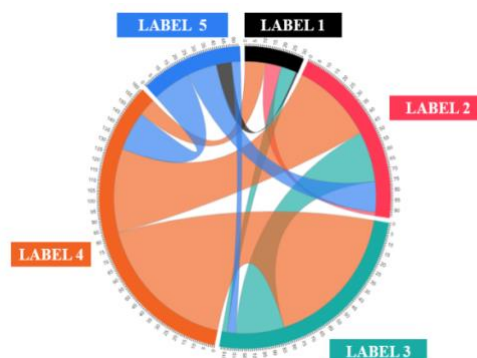


Figure 4. Relationships between labels

We address class imbalance in the data using a stratified data separation method (multilabel stratified shuffle split) to produce balanced class strata both train and test data. Furthermore, the label class strata generated in the train and test data are similar to those in the initial data, as shown in Figures 5(a) and 5(b)

respectively. The balance of these strata allows the MLC model to produce good results while being unaffected by class imbalance. According to [45], if the SCUMBLE(D) value is <0.1 , resampling (undersampling/oversampling) is unnecessary to address class imbalance.

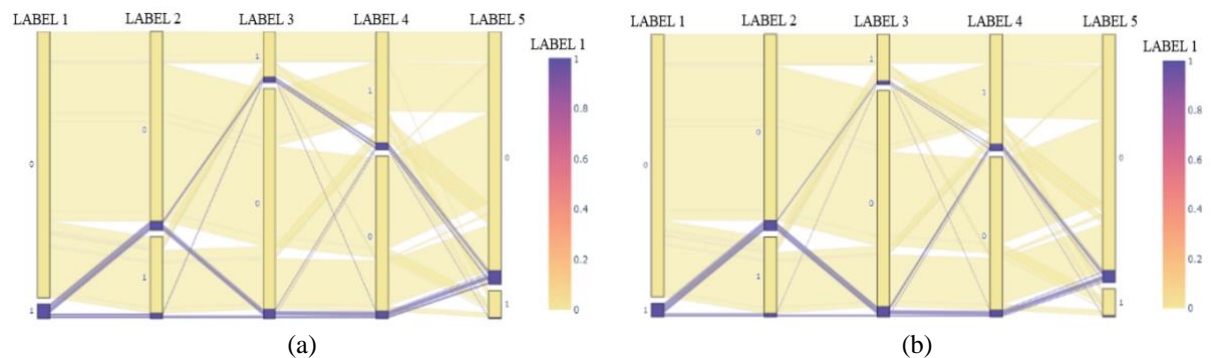


Figure 5. Class strata between labels on (a) initial data and (b) stratification results

3.2. Model performance evaluation

With a 100-repetition resampling scenario shown in Figure 6. The F1-score of model performance is shown in Figure 6(a) whereas the hamming is shown in Figure 6(b), we discovered that the ML-ARAM method has excellent model performance, as evidenced by a higher performance mean and less variability when compared to other methods using the F1-score and hamming loss. The F1-score obtained using ML-ARAM method is 0.9, indicating that the ML-ARAM method performs almost perfectly MLC classification performance. Meanwhile, the ML-ARAM method produces a minimal the hamming loss, less than 0.1. In addition to ML-ARAM method, the label powerset method works well with a stratified data separation. The stratified data separation method has been shown to improve model performance values in both case transformation and method adaptation approaches. However, the model performance value drops slightly in the ensemble of classifier approach. This is most likely due to the use of multiple methods at once in the ensemble of classifier approach, which can handle class imbalance without first stratifying the train and test data. With these findings, we propose that in future similar study, researchers use a stratified data separation method to reduce the impact of class imbalance in the MLC method on the results of [20].

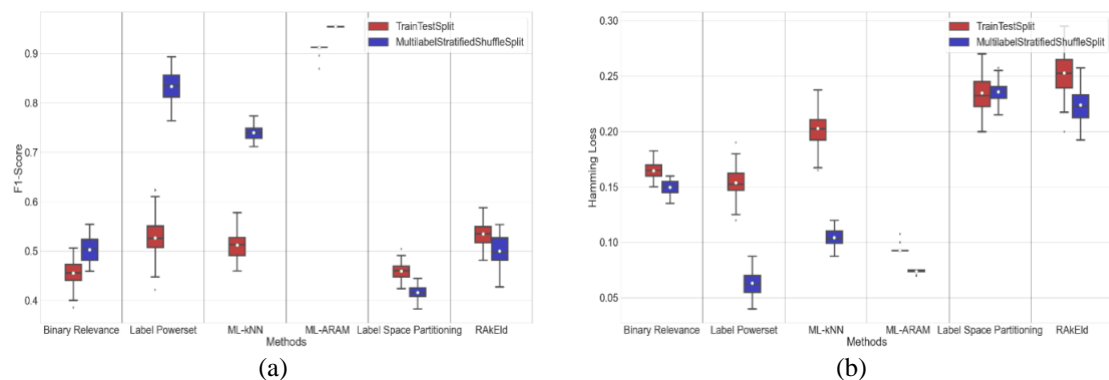


Figure 6. Comparison of model performance based on (a) F1-score \uparrow and (b) hamming loss \downarrow

3.3. Model performance comparison

Table 2 is analysis of variance shows the model performance comparison result. Based on the p-value ($\alpha=0.05$), we conclude that the factors of the MLC method, data separation method, and their interaction between the MLC method and the data separation method affect the mean value of model performance. Meanwhile, if we look at the interaction plot presented in Figure 7, it is concluded that the effect of data separation methods is specific to each MLC method. Figure 7(a) shows the interaction based on F1-score, meanwhile Figure 7(b) shows the interaction based on hamming loss. The stratified data separation method can improve model

performance in most MLC methods. When interacting with the data separation method, ML-ARAM combined with the stratified data separation method achieves the best performance (F1-score value of 0.955). The label powerset combined with the stratified data separation method yields the best model performance based on a hamming loss of 0.063. The ML-ARAM method, when combined with stratified or random data separation methods, still performs well in terms of hamming loss.

Table 2. Model performance analysis of variance

Source	F1-score ↑		Hamming loss ↓	
	F-statistics	P-value	F-statistics	P-value
MLC methods	12158.84	0.000*	6514.75	0.000*
Data separation methods	4589.42	0.000*	4213.44	0.000*
MLC methods	12158.84	0.000*	6514.75	0.000*

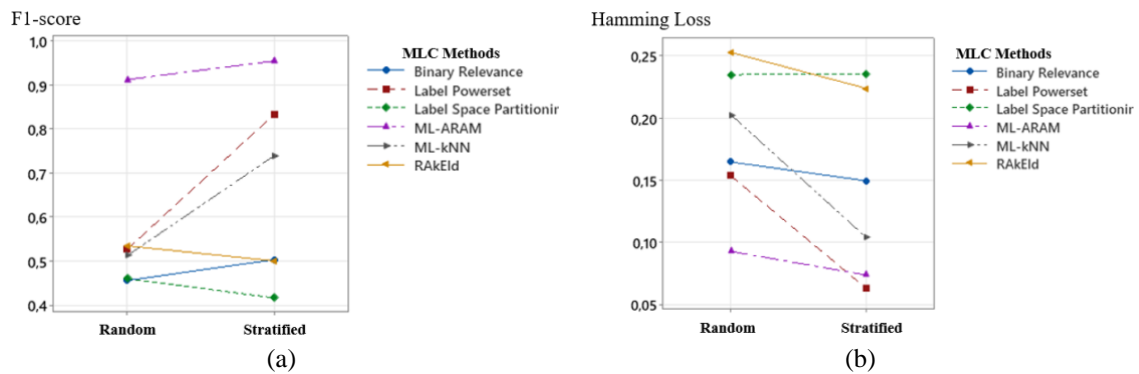


Figure 7. Interaction MLC and data separation methods based on (a) F1-score ↑ and (b) hamming loss ↓

Using a post hoc test on the MLC method factors that have been separated based on data separation methods presented in Table 3, we conclude that ML-ARAM has the best performance in the three conditions of combinations. ML-ARAM performs more than other methods based on the FI-Score value produced at more than 90 percent and the hamming loss at less than 10 percent. This result shows that this combination can improve the performance of the MLC model on data containing class imbalance.

Table 3. Post hoc test for interactions of MLC methods with data separation methods

Data separation	MLC methods	F1-score ↑	F1-score group	MLC methods	Hamming loss ↓	Hamming loss group
Random	ML-ARAM	0.912	A	ML-ARAM	0.093	A
	RakEld	0.535	B	Label powerset	0.154	B
	Label powerset	0.527	B	Binary relevance	0.165	C
	ML-kNN	0.512	C	ML-kNN	0.203	D
	Label space partitioning	0.459	D	Label space partitioning	0.235	E
Stratified	Binary relevance	0.456	D	RakEld	0.253	F
	ML-ARAM	0.955	A	Label powerset	0.063	A
	Label powerset	0.833	B	ML-ARAM	0.074	B
	ML-kNN	0.739	C	ML-kNN	0.104	C
	Binary relevance	0.503	D	Binary relevance	0.149	D
	RakEld	0.500	D	RakEld	0.224	E
	Label space partitioning	0.416	E	Label space partitioning	0.236	F

3.4. Feature importance

Using the ML-ARAM method combined with the stratification method as the best model in the previous stage, we calculated the importance values of 61 explanatory variables (features). We discovered that many explanatory variables have negative feature importance values. This means that the presence of these explanatory variables has no significant impact on the prediction accuracy of the MLC model; thus, it can complicate the model and increase the risk of overfitting. To simplify the MLC model of the risk profile, we

decided to perform the feature selection stage, considering the excellent prediction accuracy, specifically the F1-score, which remains at 0.87, and hamming loss, which is at 0.12. Table 4 summarizes 20 explanatory variables (features) for the feature selection results, most of which are local (L) category risk elements. The top three feature codes are company age (R2), value of administrative sanctions in the form of fines (O9), and procedures for importing and releasing finished goods (L8). The higher the feature importance value, the greater its influence in detecting tax avoidance and tax evasion. Our findings on these 20 explanatory variables (features) are the result of identifying risk element factors that encourage a taxpayer entity to engage in tax avoidance or tax evasion using feature importance values.

Table 4. Summary of feature selection results

No.	Feature importance	Code	No.	Feature importance	Code	No.	Feature importance	Code
1.	0.0722	R2	8.	0.0425	L13	15.	0.0362	L11
2.	0.0652	O9	9.	0.0414	R1	16.	0.0347	R10
3.	0.0474	L8	10.	0.0400	L23	17.	0.0344	L29
4.	0.0465	L28	11.	0.0395	O3	18.	0.0337	L17
5.	0.0453	R22	12.	0.0390	R8	19.	0.0332	O4
6.	0.0445	L9	13.	0.0382	L12	20.	0.0362	O6
7.	0.0441	R17	14.	0.0378	O1			

4. CONCLUSION

This study successfully identified the best MLC model for detecting tax avoidance and tax evasion risks in Indonesia. The best method is the ML-ARAM method based on deep learning, which has excellent model performance and outperforms to other methods (F1-score of 95.5% and hamming loss of 7.4%). The MLC model used in this study will improve the accuracy of tax avoidance and tax evasion detection in the risk profile mechanism. Furthermore, we discovered conditions where the data contained class imbalance, which the stratified data separation method (multilabel stratified shuffle split) effectively addressed. This is an input for future study, so researchers can think about using a stratified data separation method to reduce the impact of class imbalance in the MLC method. We also identified risk element factors that encourage taxpayer entities to engage in tax avoidance or tax evasion using feature importance values. These factors will be used to improve the risk profile mechanism and optimize tax revenues in Indonesia. We recognize that this study has limitations, specifically that the characteristics of taxpayer behavior will always change dynamically response to changes in tax regulations. To address this, we recommend that the analysis be repeated in the future to produce valid conclusions in improving the risk profile, resulting in a more relevant profile. We also recommend conducting similar study to compare methods for dealing with class imbalance in MLC data and comparing other feature importance methods, such as the shapley additive explanations (SHAP) method to learn more information about the local influence of each explanatory variable (features) on the prediction accuracy of the MLC model.

ACKNOWLEDGEMENTS

The author would like to thank those who contributed to this study, including the Department of Statistics, Faculty of Mathematics and Natural Sciences, IPB University for facilitating the study, the Indonesia Endowment Fund for Education Agency (LPDP) No. 0000564/FOR/M/ASN-2021 (202205110209116) for funding the study, and the Indonesian Ministry of Finance for providing the supervision and datasets.

REFERENCES




- [1] A. Hajawiyah, T. Suryarini, Kiswanto, and T. Tarmudji, "Analysis of a tax amnesty's effectiveness in Indonesia," *Journal of International Accounting, Auditing and Taxation*, vol. 44, 2021, doi: 10.1016/j.intaccudtax.2021.100415.
- [2] R. T. Sinaga, E. Evana, and F. Dharma, "Research of tax Avoidance in Indonesia: A bibliographic study," *Asian Journal of Economics, Business and Accounting*, vol. 24, no. 8, pp. 53–67, 2022, doi: 10.9734/ajeba/2022/v22i830587.
- [3] M. M. Postea, "Theoretical and methodological approaches on tax evasion," *DIEM: Dubrovnik International Economic Meeting*, vol. 6, no. 1, pp. 183–190, 2021, doi: 10.17818/diem/2021/1.19.
- [4] N. Alsadhan, "A multi-module machine learning approach to detect tax fraud," *Computer Systems Science and Engineering*, vol. 46, no. 1, pp. 241–253, 2023, doi: 10.32604/csse.2023.033375.
- [5] B. F. Murorunkwere, D. Haughton, J. Nzabanita, F. Kipkoge, and I. Kabano, "Predicting tax fraud using supervised machine learning approach," *African Journal of Science, Technology, Innovation and Development*, vol. 15, no. 6, pp. 731–742, 2023, doi: 10.1080/20421338.2023.2187930.
- [6] R. A. Rahman, S. Masrom, N. Omar, and M. Zakaria, "An application of machine learning on corporate tax avoidance detection model," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 4, pp. 721–725, 2020, doi: 10.11591/ijai.v9.i4.pp721-725.
- [7] C. C. Aggarwal, *Data classification: algorithms and applications*, Boca Raton, Florida: Chapman and Hall/CRC, 2014, doi:

- 10.1201/b17320.
- [8] E. K. Y. Yapp, X. Li, W. F. Lu, and P. S. Tan, "Comparison of base classifiers for multi-label learning," *Neurocomputing*, vol. 394, pp. 51–60, 2020, doi: 10.1016/j.neucom.2020.01.102.
 - [9] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, 2022, doi: 10.1016/j.is.2021.101718.
 - [10] E. Deniz, H. Erbay, and M. Coşar, "Multi-label classification of e-commerce customer reviews via machine learning," *Axioms*, vol. 11, no. 9, pp. 1–17, 2022, doi: 10.3390/axioms11090436.
 - [11] W. Liu, H. Wang, X. Shen, and I. W. Tsang, "The emerging trends of multi-label learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7955–7974, 2022, doi: 10.1109/TPAMI.2021.3119334.
 - [12] D. Afdhal, K. W. Ananta, and W. S. Hartono, "Adverse drug reactions prediction using multi-label linear discriminant analysis and multi-label learning," in *2020 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2020*, Oct. 2020, pp. 69–76, doi: 10.1109/ICACSIS51025.2020.9263166.
 - [13] L. Uwineza, H.-G. Kim, C. K. Kim, B. Kim, and J.-Y. Kim, "Accuracy assessment of typical meteorological year data for a photovoltaic system using a bootstrap method," *Journal of the Korean Solar Energy Society*, vol. 41, no. 4, pp. 115–129, 2021, doi: 10.7836/kse.2021.41.4.115.
 - [14] E. Zivot, *Introduction to computational finance and financial econometrics*, 2011. Accessed: Oct. 18, 2023. [Online]. Available: <https://bookdown.org/compfinezbook/introcompfinr/>
 - [15] F. Charte, "A comprehensive and didactic review on multilabel learning software tools," *IEEE Access*, vol. 8, pp. 50330–50354, 2020, doi: 10.1109/ACCESS.2020.2979787.
 - [16] F. Charte and F. D. Charte, "Working with multilabel datasets in r: the mldr package," *The R Journal*, no. 7, vol. 2, pp. 149–162, 2015.
 - [17] A. N. Tarekegn, M. Giacobini, and K. Michalak, "A review of methods for imbalanced multi-label classification," *Pattern Recognition*, vol. 118, 2021, doi: 10.1016/j.patcog.2021.107965.
 - [18] Y. E. Kurniawati and Y. D. Prabowo, "Model optimisation of class imbalanced learning using ensemble classifier on over-sampling data," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 276–283, 2022, doi: 10.11591/ijai.v11.i1.pp276-283.
 - [19] B. Liu, K. Blekas, and G. Tsoumakas, "Multi-label sampling based on local label imbalance," *Pattern Recognition*, vol. 122, 2022, doi: 10.1016/j.patcog.2021.108294.
 - [20] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data," *Machine Learning and Knowledge Discovery in Databases*, pp. 145–158, 2011, doi: 10.1007/978-3-642-23808-6_10.
 - [21] F. Charte, A. Rivera, M. J. D. Jesus, and F. Herrera, "Concurrence among imbalanced labels and its influence on multilabel resampling algorithms," in *Hybrid Artificial Intelligence Systems*, Verlag: Springer, 2014, pp. 110–121, doi: 10.1007/978-3-319-07617-1_10.
 - [22] P. Szymanski and T. Kajdanowicz, "Scikit-multilearn: A python library for multi-label classification," *Journal of Machine Learning Research*, vol. 20, no. 6, pp. 1–22, 2019.
 - [23] F. Herrera, F. Charte, A. J. Rivera, and M. J. D. Jesus, *Multilabel classification problem analysis, metrics and techniques*. Switzerland: Springer, 2016, doi: 10.1007/978-3-319-41111-8.
 - [24] S. Godbole and S. Sarawagi, "Discriminative methods for multi-labeled classification. in advances in knowledge discovery and data mining," *Advances in Knowledge Discovery and Data Mining*, pp. 22–30, 2004, doi: 10.1007/978-3-540-24775-3_5.
 - [25] F. Li, X. Ma, and Y. Wang, "A multi-label method of state partition and fault diagnosis based on binary relevance algorithm," *2020 IEEE 9th Data Driven Control and Learning Systems Conference (DDCLS)*, pp. 567–572, 2020, doi: 10.1109/DDCLS49620.2020.9275199.
 - [26] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004, doi: 10.1016/j.patcog.2004.03.009.
 - [27] A. S. Ardianto and S. Adi, "The best problem transformation method in multi-label classification text for thesis abstract," *2022 6th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 326–329, 2022, doi: 10.1109/ICITISEE57756.2022.10057824.
 - [28] F. Benites and E. Sapozhnikova, "HARAM: a hierarchical ARAM neural network for large-scale text classification," in *15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*, Jan. 2016, pp. 847–854, doi: 10.1109/ICDMW.2015.14.
 - [29] F. Benites and E. Sapozhnikova, "Improving scalability of ART neural networks," *Neurocomputing*, vol. 230, pp. 219–229, 2017, doi: 10.1016/j.neucom.2016.12.022.
 - [30] M. L. Zhang and Z. H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, Jul. 2007, doi: 10.1016/j.patcog.2006.12.019.
 - [31] E. C. Gatto, M. Ferrandin, and R. Cerri, "Multi-label classification with label clusters," *Research Square*, 2023, doi: 10.21203/rs.3.rs-3133411/v1.
 - [32] J. M. Moyano, E. L. Gibaja, K. J. Cios, and S. Ventura, "Review of ensembles of multi-label classifiers: Models, experimental study and prospects," *Information Fusion*, vol. 44, pp. 33–45, 2018, doi: 10.1016/j.inffus.2017.12.001.
 - [33] S. Bates, T. Hastie, and R. Tibshirani, "Cross-validation: what does it estimate and how well does it do it?," *Journal of the American Statistical Association*, pp. 1–43, 2023, doi: 10.1080/01621459.2023.2197686.
 - [34] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proceedings of the 34th International Conference on Machine Learning*, pp. 3780–3788, 2017.
 - [35] L. Xue, X. Zhang, W. Jiang, and K. Huo, "A classification performance evaluation measure considering data separability," *Artificial Neural Networks and Machine Learning – ICANN 2023*, pp. 1–13, 2023.
 - [36] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*, Cham: Springer, 2018, doi: 10.1007/978-3-319-78503-5.
 - [37] J. R. -Salazar, *Data science and analytics with Python*, Boca Raton, Florida: Chapman and Hall/CRC, 2017.
 - [38] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2018, doi: 10.1016/j.aci.2018.08.003.
 - [39] M. Rapp, E. L. Mencia, J. Fürnkranz, V. L. Nguyen, and E. Hüllermeier, "Learning gradient boosted multi-label classification rules," *Machine Learning and Knowledge Discovery in Databases*, pp. 124–140, 2021, doi: 10.1007/978-3-030-67664-3_8.
 - [40] H. S. A. Kutubi, "On randomized complete block design," *International Journal of Sciences: Basic and Applied Research*, vol. 53, no. 2, pp. 230–243, 2020.
 - [41] C. Molnar, *Interpretable machine learning a guide for making black box models explainable*, Victoria, Canada: Lean Publishing, 2022.
 - [42] D. A. Otchere, M. Aboagye, M. A. Abdalla, and T. B. Boakye, "Enhancing drilling fluid lost-circulation prediction using model




- agnostic and supervised machine learning,” SSRN, 2022, doi: 10.2139/ssrn.4085366.
- [43] Z. Zhou and G. Hooker, “Unbiased measurement of feature importance in tree-based methods,” *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 2, pp. 1–20, 2021, doi: 10.1145/3429445.
- [44] R. Zebari, A. Abdulazeez, D. Zeebaree, D. Zebari, and J. Saeed, “A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction,” *Journal of Applied Science and Technology Trends*, vol. 1, no. 2, pp. 56–70, 2020, doi: 10.38094/jastt1224.
- [45] F. Charte, A. J. Rivera, M. J. del Jesus, and F. Herrera, “Dealing with difficult minority labels in imbalanced multilabel data sets,” *Neurocomputing*, vol. 326–327, 2018, pp. 39–53, 2019, doi: 10.1016/j.neucom.2016.08.158.

BIOGRAPHIES OF AUTHORS






Teguh Prasetyo    is currently a Business & Data Analyst at the Ministry of Finance, Indonesia specializing in Machine Learning and Econometrics. He completed his Bachelor of Statistics in 2013 from Brawijaya University, Malang, Indonesia. He is currently pursuing Master of Science in Statistics & Data Science at IPB University, Bogor, Indonesia. Since 2014 he has been active in several activities at the Ministry of Finance, Indonesia as a core member of the Ministry of Finance – Data Analytics Community (MoFDAC) by providing several lectures and training related to Data Analysis and Machine Learning within the Ministry of Finance, Indonesia. In 2020, he was a finalist in the Ministry of Finance Data Analytics Competition with the topic “Analysis of predictions for violations in the use of excise tax stamp” in Indonesia. He can be contacted at email: teguhprasetyo@apps.ipb.ac.id or teguh.prasetyo@kemenkeu.go.id.



Budi Susetyo    is currently a lecturer at IPB University, Bogor, Indonesia. He completed his Bachelor of Statistics in 1985 and Master of Science in 1990 from IPB University. Then, he completed his Doctorate in biometrics in 1997 from Justus Liebig University, Giessen, Germany. He wrote various papers in the field of statistical modeling, multivariate analysis, and sampling. He was chairman of the Department of Statistics from 2003–2005. Apart from that, from 1998–2018 he was active as an education consultant at the Ministry of Education Indonesia, Culture, Research, and Technology in various projects funded by the government and donor agencies (World Bank, ADB, AUSAID, and the Dutch Government). In 2018–2023 he was a member of the National Accreditation Board for Schools and Madrasah (BAN-S/M). He is appointed secretary of the National Accreditation Board for Early Childhood Education, Basic Education, and Secondary Education for the 2023–2028 period. He can be contacted at email: budisu@apps.ipb.ac.id.



Anang Kurnia    is an Associate Professor of Statistics and Data Science and Vice Dean of Academic and Student Affairs at the School of Data Science, Mathematics, and Informatics at IPB University, Indonesia. He is also Chairman of the Indonesian Statistical Association. He is former head of the Department of Statistics at IPB University (2014–2023) and the former head of the Indonesian Statistics Higher Education Association (2014–2018). His main teaching and research interests include statistical machine learning, statistical inference, generalized linear mixed models, data science, and small area estimation. He has published several research articles in international journals in the area of statistics and data science. He can be contacted at email: anangk@apps.ipb.ac.id.