Scientax Jurnal Kajian Ilmiah Perpajakan Indonesia

VOLUME 5 NO. 1 | OCTOBER 2023 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Data Analytics for Identifying Fraudulent Tax Invoice Issuer in Value Added Tax System Fraudulent

Yusrifaizal Gumilar Winata[a], Fauziah Noor[b], Muhammad Futhra Bahar[c], Aris Budi Santoso[d], Eddy Sukarno[e]

[a] Directorate General of Taxes, Jakarta, Indonesia. Email: yusrifaizal.gumilar@kemenkeu.go.id
[b] Directorate General of Taxes, Jakarta, Indonesia. Email: noor.fauziah@pajak.go.id
[c] Directorate General of Taxes, Jakarta, Indonesia. Email: muhammad.futhra@pajak.go.id
[d] Directorate General of Taxes, Jakarta, Indonesia. Email: aris.budi@pajak.go.id
[e] Directorate General of Taxes, Jakarta, Indonesia. Email: eddy.sukarno@pajak.go.id

* Corresponding author: yusrifaizal.gumilar@kemenkeu.go.id

ABSTRACT

Taxes are the backbone of state revenue; however, tax revenues are not received optimally because of tax crimes. 39.8% of all tax crimes are derived from the issuance of Tax Invoices that are not based on actual business transactions. One case of Fraudulent Tax Invoices Crimes able to harm state revenues estimated at 244 billion Rupiah. This practice injures the state revenue arises from Value Added Tax (VAT). Moreover, it can take state money through VAT refunds. Dealing with this, DGT needs; more than ever; to perform audit more effectively and efficiently by utilizing data mining techniques. This study aims to build a model to detect issuer of tax invoices not based on actual transactions. We developed the model using a mixed methods approach, based of Cross Industries Standard Process for Data Mining. Data processing and machine learning conducted with Python programming language. The dataset to train and evaluate the model consist of 1.071 taxpayers that issue illegal tax invoices and 2.142 non-issuer taxpayers. The research provided a machine learning model that has Prediction Efficiency of 83.56%, reduction in Examination Effort of 69.31%, and Strike Rate of 90.77%. Then, we run this model with Streamlit. Using deployment data of 1.000 rows and probability threshold adjusted to 75%, it predicted eight issuer taxpayers.

Keywords: data mining, crisp dm, value added tax, illegal tax invoice

1. INTRODUCTION

Tax revenue serves as the backbone of state finances, but currently the Directorate General of Taxes (DGT) is facing significant challenges, as indicated by the still-low tax ratio. According on 2021 DGT Annual Report, tax ratio in 2021 was 9.11 percent, showing an improvement from 8.33 percent recorded in 2020. This increase, however, while a positive step, is still considered low when compared to the ideal ratio of 15 percent as suggested by Gaspar et al. (2016). Moreover, it is significantly lower than average tax ratio of OECD countries, that is 34.1 percent (OECD, 2022). Furthermore, international organizations such as

doi:
Received: August 17, 2022; Accepted: August 21, 2023; Published: October 31, 2023

Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia is Sinta 4 Journal (https://sinta.kemdikbud.go.id/journals/profile/9121)
How to Cite:
Winata, Y. G., Noor, F., Bahar, M. F., Santoso, A. B., & Sukarno, E. (2023). Application of data mining to taxpayers issuing fictitious tax invoice using classification techniques. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia, 5*(1), 58–73. https://doi.org/10.52869/st.v5i1.573

the United Nations Conference on Trade and Development (UNCTAD) and the High-Level Panel on International Financial Accountability, Transparency, and Integrity for Achieving the 2030 Agenda (FACTI Panel), along with others, underscore the persistent challenges posed by tax-related crimes. These issues significantly affect the economic development and stability of nations, particularly by depriving developing countries of their vital and limited revenue (Brun et al., 2022). Considering these challenges, addressing the prevalence of tax crimes becomes crucial. The problem of tax crimes represents an inherent risk stemming from government tax policies and requires comprehensive mitigation and management. In response to these challenges, the Directorate General of Taxation must enhance its law enforcement processes to enable early warning detection of tax crime indicators and to effectively act as a deterrent.

Looking at the 2021 DGT Annual Report, the most common tax crime handled by tax investigators is Issuance of Tax Invoices Not Based on Actual Transactions (NBAT) or Fraudulent Tax Invoice with a total of 39,80 percent as shown in the table 1 above. According to the recent news, a single case of Tax Invoice NBAT Crimes led to an estimated loss of 244 billion Rupiah in state revenue (Detik.com, 2023). This raises concerns about the potential impact, especially if the 41 cases reported in 2021 as shown in table above could result in similar losses and if the number of such cases continues to rise in the future.

Taxpayers employ intricate and dynamically changing methods to commit tax crimes. According to Turksen & Abukari (2020), a study conducted by PROTAX (2018) involved 13 case studies across 10 European Union Member States, revealing that taxpayers, especially transnational corporations, employ sophisticated strategies for engaging in tax crimes. These strategies encompass the use of shell companies (commonly known as 'missing trader'), trusts, VAT carousels, profit-shifting, and the utilization of free-ports. Facing this challenge, DGT must adapt quickly to prevent the loss of state revenue, primarily due to the fraudulent tax invoices commonly associated with missing trader cases and VAT carousels.

Tax Invoice NBAT, is essentially a misrepresentation of a tax invoice. It identified as 'fraudulent' because the identity of the issuer does not align with the actual transaction, or there is no substantial delivery of goods and/or services, despite formally meeting the provisions of the Value Added Tax (VAT) Law. Taxpayers resort to the issuance of such invoices due to the ease of obtaining VAT refunds, which can result in greater gains from tax evasion compared to the costs incurred (Yamin & Putrantri, 2009). Mismanagement of this practice can significantly reduce tax revenue from the VAT sector, posing a threat to state finances.

Table 1 Tax Crime Cases handled by Tax Investigators
Source: Annual Report of the Directorate General of Taxes for 2021

Descriptions	Total Cases
Tax invoices do not match actual transactions	41
File Tax Return inaccurately	30
Withheld tax but not paid	10
Do not file Tax Return	18
Money laundering	1
Not registering for TIN/Taxable Person for VAT Purposes Identification Number	2
TIN/Taxable Person for VAT Purposes Identification Number misuse	1
Total	103

With a staggering 66,351,573 taxpayers in Indonesia (Directorate General of Taxes, 2022), the task of supervision becomes a daunting challenge. In light of this immense scale, the Directorate General of Taxes (DGT) can harness the power of data mining to analyze and make informed decisions. Data Mining is the process of collecting important information which can be in the form of correlations, patterns, and trends from large data using artificial intelligence, statistical techniques, mathematics, machine learning, and so on (Larose, 2005).

Looking at previous research, Wu et al. (2012) have applied data mining on VAT reporting compliance in Taiwan. Their study highlighted the potential of data mining in enhancing tax evasion screening, thereby increasing efficiency in audit processes. This research contributes that the use of data mining can support screening activities for inappropriate VAT reports in a more scientific way compared to relying on audits based on manual methods and personal judgments alone. The results of the data mining are expected to be combined with the personal experience of the tax auditor to obtain more effective and efficient results. Naturally this will help DGT in auditing considering that the audit coverage ratio is still low at 0.86 percent.

To identify the most effective model for tax auditing purposes which generated through data mining activities, Gupta & Nagadevara (2007) assess model performance using Strike Rate (SR), Prediction Efficiency (PE), and Reduction in examination effort (EF). Said research emphasizes precision over recall as a top priority. Said research emphasizes precision over recall as a top priority. The primary objective is to enhance the accuracy of audits and simultaneously establish a strong deterrent effect on other taxpayers.

In theoretical context, this study is expected to contribute to data analytics research in taxation sector especially in the use of variables and algorithms used in this research. On the other hand, in the practical context, this study aims to demonstrate the potential of data analytics in taxation as a screening tool for identifying fraudulent tax invoice issuers and inspire tax authorities to integrate data analytics into their decision-making processes. The anticipated outcome from this research is an enhancement in the effectiveness and efficiency of tax audit activities, particularly within the Directorate General of Taxes, with a specific focus on the VAT sector.

THEORETICAL FRAMEWORK

1.1 Theory of Planned Behavior

Theory of Planned Behavior is explained by Ajzen (1991) as a person's intention to perform a certain behavior. A person's intention can be seen from attitudes toward behavior (attitude), subjective norms related to behavior (subjective norms), and control over perceived behavior (perceived behavioral control). The measure of effort expended to perform a particular behavior reflects the level of a person's intention. TPB also argues that a person will perform certain behaviors depending on the benefits and costs of an activity.

Simply put, a certain behavior can describe a person's intentions. In connection with this study, if tax invoice issuers had intention to issue fraudulent tax invoice, that intention should be reflected in their behavior and shows some patterns which will then be analyzed using data analytics. Recent study conducted in DKI Jakarta by Saputra (2019) shows that taxpayer behavior is induced by the intention to adhere to tax regulation which is reflected by attitudes, subjective norms and behavioral controls. The study also shows that TPB theory is still relevant to current situations and conditions in describing taxpayer behavior.

1.2 Theory of Tax Evasion

The difference between tax avoidance and tax evasion lies in the legality of the actions taken by the taxpayer. Sandmo (2005) explains tax avoidance as an act of legal by taking advantage of legal loopholes to reduce the tax that must be paid. Taxpayers do not need to worry if tax avoidance activities are detected by the tax authorities because their actions do not violate the law. Meanwhile, tax evasion is explained as reducing tax through violation of the law by not reporting income that should be taxed or carrying out other illegal activities that make him responsible for administrative actions from the tax authorities.

Research on tax evasion was conducted by Chen et. al. (2010). This study found that family companies tend to have a lower level of tax aggressiveness than other companies. Since family companies more concern about the sustainability of the company and avoid conflicts with the tax authorities so as not to tarnish the family name. From this study it can be inferred that company owners can affect the level of tax aggressiveness.

According to the 2021 DGT Annual Report, tax evasion cases that frequently occur are primarily related to Value Added Tax (VAT). These instances of tax evasion should not be disregarded, given that VAT significantly contributes to the total tax revenues. Recent research conducted by Vanhoeyveld et al. (2020) emphasizes the critical role of efficient fraud detection in mitigating government financial losses. Their findings suggest that robust fraud detection measures not only assist in recovering lost revenue but also act as a deterrent for taxpayers. Building upon this insight, this study aims to explore the potential of data analytics, specifically data mining, as a tool to enhance fraud detection in the context of VAT compliance.

1.3 Economic Deterrence Model

This model incorporates the concept of an economically rational taxpayer who will avoid taxes as long as the results of illegal tax evasion are greater than the penalties received when caught by the tax authorities (Hasseldine & Bebbington, 1991). According to this model, taxpayers are less likely to engage in illegal tax evasion when the potential penalties outweigh the potential profits. Therefore, it is imperative for tax authorities to ensure suitable penalties for tax crimes. Simultaneously, the effective detection and enforcement of penalties for illegal tax evasion become paramount priorities, as they can deter taxpayers from committing such crimes while upholding the law.

Tax authorities should focus on preventing tax evasion through tax audits and penalties for non-compliance (Carvalho & Pacheco, 2014). In accordance with this perspective, Wu et al. (2012) demonstrated the effectiveness of data mining techniques in auditing Monthly VAT returns in Taiwan, showcasing a more systematic approach to identifying non-compliant taxpayers compared to random audits or relying solely on auditor experience.

Supporting that matter, the impact of penalty rate on taxpayers' decision to comply is greater when the audit probability is higher, and vice versa, as noted by Alm & Malézieux (2020). However, Indonesia's audit ratio is currently low at only 0.86 percent. This indicates the need for more efficient and effective audit methods. The use of data analytics, specifically data mining, can enhance the selection of audit targets based on risk assessment. Drawing from the findings of previous research (Alm & Malézieux, 2020; Wu et al., 2012) and the context discussed, it is evident that the utilization of data analytics, particularly data mining, can significantly increase the audit probability and enhance the efficiency of the auditing process in order to help law enforcement.

1.4 Value Added Tax

Value Added Tax is a tax imposed on the delivery of taxable goods (Barang Kena Pajak - BKP) and/or taxable services (Jasa Kena Pajak - JKP) within the customs area (territory of the Republic of Indonesia) as regulated in VAT Law. Value Added

Tax is objective, not cumulative, and is an indirect tax with tax subjects consisting of VAT-Registered Businesses and non-VAT-Registered Businesses (Badan Kebijakan Fiskal, 2023).

Furthermore, in the context of Value Added Tax (VAT) collection, it's important to note that VAT payments are not directly remitted to the state treasury. Instead, VAT-registered businesses withhold these payments until they are transferred to the treasury. In its calculation, Value Added Tax has Input Tax (Pajak Masukan) which can be credited and Output Tax (Pajak Keluaran) which becomes tax debt. The amount of this Input Tax directly affects the amount of VAT that needs to be deposited to the state because it reduces the tax debt from the Output Tax. If the Input Tax has a larger amount than the Output Tax, then the taxpayer can get a refund.

1.5 Fraudulent Tax Invoice

Each Input Tax and Output Tax that is reported in the Taxpayer Periodic VAT tax return is recorded based on the Tax Invoices received and/or issued by the Taxpayer itself. VAT-related tax crimes often occur because tax invoices are not issued based on actual transactions. So, taxpayers receive higher input tax than it should and reduce the amount of VAT that must be deposited to the state. According to Regulation of the Director General of Taxes Number (No.) PER-19/PJ/2017, Fraudulent Tax Invoices are Tax Invoices issued not based on actual transactions and/or Tax Invoices issued by Entrepreneurs who have not been confirmed as Taxable Entrepreneurs. For this crime, the Directorate General of Taxes imposed a penalty based on General Provisions and Tax Procedures Law, namely imprisonment for a minimum of 2 years and a maximum of 6 years, as well as fines ranging from at least 2 times to a maximum of 6 times the amount of tax on the fraudulent tax invoice.

2. RESEARCH METHOD

This study aims to build a model for the detection of tax invoice issuers that are not based on actual transactions. This research was conducted using mixed methods, involving both qualitative and quantitative analysis. The qualitative method is used to determine which variables should be used based on regulations and relationship with fraud cases that have occurred. The quantitative method is used to identify patterns in numerical data that have been selected from quantitative analysis.

These two methods are used in the development of machine learning models with reference to the data mining process. The research flow will follow the Cross Industry Standard Process for Data Mining (CRISP-DM), as described by Brown (2014) as a step-by-step data mining process created by data miners for data miners. The data mining process is divided into six cycle stages, which are:

a) Business Understanding
This stage is carried out with a qualitative approach to obtain information about business problems in the research object which is the case study site.

b) Data Understanding
After information about business problems is learned, the next step is to collect data that is relevant to the research being conducted. Data acquisition was carried out by observing DGT's Data Warehouse, conducting profiling to find out the amount, size and type of data for each data element, and extracting structured data for use in this study.

c) Data Preparation
After the data is obtained, then the initial processing is carried out to form a labeled dataset. This stage includes the extraction of independent variables/features that are used as predictors and the activity of labeling the class that is the target of the prediction.

d) Modeling
This stage includes activities to train machine learning models with either a single algorithm or an ensemble algorithm. Examples of single

algorithms include K Nearest Neighbor, Decision Tree, and Logistic Regression, while examples of ensemble algorithms include Random Forest and AdaBoost. This activity will produce many machine learning models to compare their performance at the evaluation stage.

e) Evaluation

This stage includes activities to compare model performance and select the machine learning model that has the best performance. The performance measure of the model used is the F1-Score which is the harmonic mean of precision and recall. According to Gupta & Nagadevara (2007), the model can be further evaluated for tax audit purposes as follows:

- The model's ability to predict the Issuing Taxpayer is measured using the Prediction Efficiency (PE) with the formula PE = TP/ (TP+FN).
- Model's ability to reduce the effort required in an audit by reducing in Examination effort (EF) with the formula EF = 1-(FP+TN)/Total Cases.
- The model's ability to obtain Issuer Taxpayers if each taxpayer predicted to be an issuer is audited can be measured using the Strike Rate (SR) with the formula SR=TP/(TP+FP).

f) Deployment

After the model with the best performance measurement is obtained, then the model is applied to an application so that it can be used to detect invoice issuers that are not based on actual transactions.

3. RESULT DAN DISCUSSION

The research was conducted using Jupyter-Notebook which is supported by other libraries available in the Package Installer for Python (PIP) which is commonly used for data mining activities. The research was conducted by following the six stages of the CRISP-DM model with the following explanation.

3.1 Business Understanding

One of the threats to tax revenues is tax evasion in withholding taxes, such as Value Added Tax (VAT). The embezzlement scheme used is in the form of issuing and using tax invoices that are not based on actual (fraudulent) transactions. VAT-Registered businesses collect VAT from sales transactions without issuing a tax invoice, then issue a fraudulent tax invoice to other parties (with a smaller value than the actual transaction for which a tax invoice was not issued), the tax invoice is used as a tax credit by other parties, thereby reducing the amount of tax to be paid by the party. Puppet company networks were created to disguise this embezzlement scheme with a series of fraudulent transactions, so that this scenario is difficult to detect by tax examiners(Mehta et al., 2019).

3.2 Data Understanding

This study aims to build a detection model for fraudulent invoice issuers using taxpayer characteristic data as a predictor. Characteristics of taxpayers can be obtained from data on submission of tax returns (SPT Masa PPN), both annual tax returns and periodic VAT returns.

The Directorate General of Taxes has implemented a data warehouse. All data from various business processes have been collected in DGT's Data Warehouse. The required data is obtained by querying the data warehouse.

Predictive target data for the machine learning model to be built are obtained from law enforcement activities, namely preliminary evidence checks and/or year investigations with a time span of 2018 - 2020, as well as e-Faktur databases and Taxpayer Master Files with a timeframe of 2014 - 2020.

Data is received in the form of comma separated values (.csv) from different tables and databases so that it is prone to getting null values when merging is performed. The file received has the smallest size having a size of 59 Kilo Bytes containing approximately 1500 lines of data and the file with the largest size having a size of 600 Mega Bytes containing approximately 10,000,000 lines of data.

3.3 Data Preparation

The data is processed using Pandas, which is a Python library used for handling panel data. The data from these tables undergo merging processes to form a single data table with the following variables.

The merging process is carried out using the inner join method to ensure that only the rows with matching or common values in specified columns from both datasets are included in the resulting merged dataset. This method is chosen to maintain data integrity and relevance, as it ensures that only the data entries with corresponding information in both datasets are considered (Pawki, 2023).

However, there was a null value in the number of employees variable. To address this, linear regression was employed as the chosen method for imputation. This decision was based on its ability to predict the missing values accurately by leveraging factors such as the number of submissions in a year, the Input VAT value, and the Monthly Income Tax Return (Article 21) Reporting Compliance, which were identified as having the highest correlation with the 'number of employees' variable. Linear regression was selected because it offers a precise way to estimate missing values when no other suitable alternatives are available, ensuring that the imputed data aligns closely with the underlying patterns of the dataset (Subrahmanya, 2018). Using measures like the mean or median, in this case, would likely yield inaccurate results. Each variable in table 2 is explained as follows.

3.3.1. Attributes of the Taxpayer issuing the Fraudulent Invoice (Y)

This variable serves as the basis for distinguishing between positive and negative labels, enabling the algorithm to learn before making predictions. In order to fill this variable, manual labeling is conducted by referring to the results of preliminary evidence inspection activities. A taxpayer with a positive label is a taxpayer who issues a fraudulent tax invoice. On the other hand, a taxpayer who does not issue a fraudulent tax invoice is marked with a negative label. Taxpayers are classified as

Table 2 Variables used in Data Mining
Source: Processed by Authors

Code.	Variables	Type of Data
Y	Attributes of the Taxpayer issuing the Fraudulent Invoice	boolean
X1	Ratio of Underpayment to Output VAT	*Float*
X2	Age of VAT-Registered Business (categories divided into 0-3, 4-6,7-9, and >10)	*int*
X3	Number of output VAT sheets (average)	*int*
X4	Number of input VAT sheets (average)	*int*
X5	Number of Consignment in a year (average)	*Float*
X6	Number of employees in Monthly Income Tax Return (Article 21) (TLK)	*int*
X7	Monthly Income Tax Return (Article 21) Compliance (*average*)	*int*

issuers of fraudulent tax invoices if they issue Tax Invoices that are not based on actual transactions, as stipulated by PER-19/PJ/2017, as explained in the previous section. Conversely, taxpayers are labeled as non-issuers when they do not issue such fraudulent tax invoices.

The Positive label data consists of taxpayers who have been identified as issuers of fraudulent tax invoices. A total of 1,071 rows of data were selected from the initial dataset, which originally contained 1,551 rows of data. This data is sourced from taxpayer records that have undergone preliminary evidence and/or investigations within the SIGAKUM database for the period 2018 to 2020.

On the other hand, the negative label data initially consisted of 57,541 rows of data, from which 2,142 rows were randomly selected for training, and an additional 1,000 rows were reserved for deployment. Under sampling was deemed necessary due to significant class imbalance, as the positive label dataset contained only 1,071 rows. Balancing the dataset is crucial for model performance. While an ideal 1:1 ratio was considered, it was essential to account for the considerably larger number of negative label instances. Following guidance from Brownlee (2021), a 1:2 ratio was adopted for under sampling.

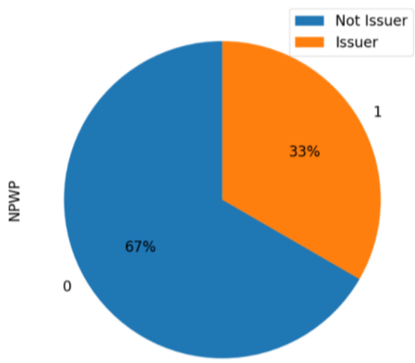


Figure 1 Distribution of Positive (issuer) and Negative (Not Issuer) Label Data
Source: Processed by Authors using Jupyter-Notebook

This variable then combines the 1,071 rows of positive label data obtained from the SIGAKUM database for 2018-2020 with randomly sampled negative label data from the e-Faktur and Taxpayer Master File databases, as previously described. The training data used has a distribution of labels which is illustrated by the Figure 1.

3.3.2. Ratio of Underpayment to Output VAT (X1)

The variable ratio of underpayment to output VAT is obtained from the amount of underpayment or overpayment coupled with the output VAT reported by the taxpayer in the Periodic VAT reports. The value of this variable is taken as the average value per year.

3.3.3. Age of VAT-Registered Business (X2)

The age of VAT-Registered business is taken from the year the taxpayer was confirmed or established himself as a VAT-Registered business in the Taxpayer Master File, reduced by the first year's data recorded in the SIGAKUM database for data with a positive label and the last year's data is in the e-Faktur database for data with negative labels.

Due to the wide range of the data, we simplified them by categorizing into four classes as shown in the Table 3.

Table 3 Age Division into Categories
Source: Processed by Authors

Categories	Age Range
0	0-3 years
1	4-6 years
2	7-9 years
3	>10 years

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import scipy as stats
from sklearn.metrics import accuracy_score, confusion_matrix
```

Figure 2 Libraries used for Data Mining
Source: Processed by Authors using Jupyter-Notebook

3.3.4. Number of output VAT sheets and Number of input VAT sheets (X3 & X4)

For the variable Number of Output VAT Sheets and Number of Input VAT Sheets, the average value of the taxpayer issuing a tax invoice and crediting the tax invoice were taken from e-Faktur database.

3.3.5. Number of Consignment in a year (X5)

The Variable Amount of Delivery in a year is the variable the value of the goods and/or services delivered by the taxpayer to another party, which in general is a sale and purchase transaction of taxable goods or taxable services. This variable is the basis for the imposition of output VAT which becomes a tax debt for taxpayers. Data for this variable is obtained based on the average value in the e-Faktur database.

3.3.6. Number of employees on Monthly Income Tax Return (Article 21) (X6)

The variable number of employees on Monthly Income Tax Return (Article 21) is obtained based on the last value or Take Last Known (TLK) in the Taxpayer's Master File based on the latest Monthly Income Tax Return (Article 21) reported by the taxpayer. Naturally, if the taxpayer has high value and routine submissions, then the taxpayer requires a large number of workers as well.

3.3.7. Monthly Income Tax Return (Article 21) Compliance (X7)

The Monthly Income Tax Return (Article 21) Compliance variable is obtained based on the average number of taxpayers reporting their Monthly Income Tax Return (Article 21) Reports per year. Variables have a minimum value of 0 and a maximum value of 12.

3.3.8. VAT reports compliance (X8)

VAT reports Compliance variable is obtained based on the average number of taxpayers reporting VAT reports each month per year as recorded in the e-Faktur database. The authors did not use the Annual Tax Returns variable because from the data received, every registered taxpayer complied with reporting the Annual SPT, so the writing team used the VAT reports Compliance variable to test monthly taxpayer compliance with VAT obligations.

3.3.9. Value of Input VAT in a year (X9)

The input VAT value variable in a year is obtained from the average number of Input VAT credited by the taxpayer during the year. This variable is suspected to have a relationship with the fraudulent tax invoices because input VAT will reduce the amount of tax debt that must be paid to the state.

3.4 Modeling

After the training data is prepared, the next step is modeling or variable fitting of the machine learning algorithm for data mining. At this stage, the authors used libraries as shown in the Figure 2.

The selection of machine learning algorithms in this study, including Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors (KNN), was influenced by two factors. Firstly, these algorithms align with the authors' computer equipment capabilities and

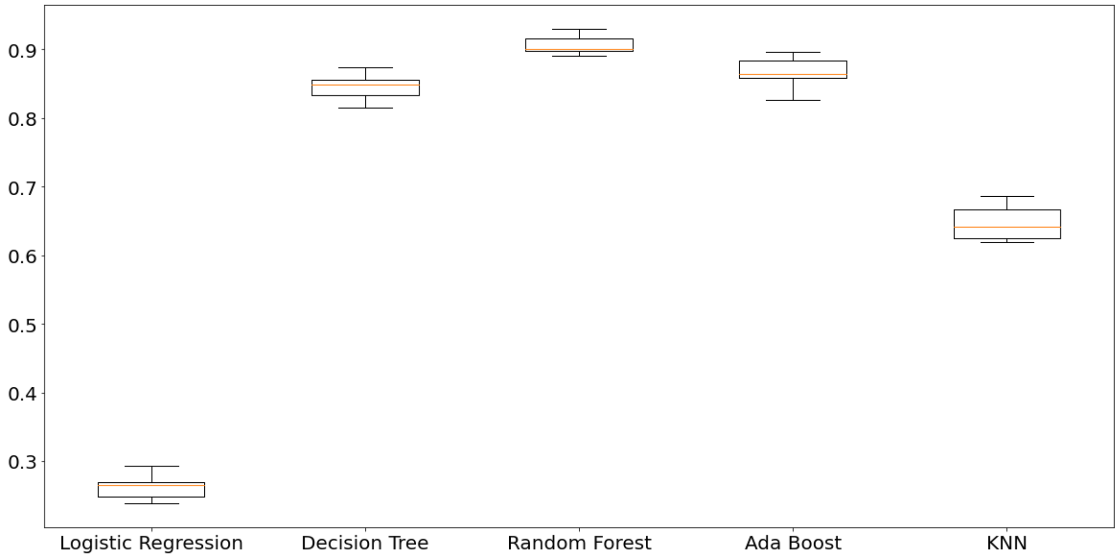


Figure 3 Each Models compared using Subplot graphic.
Source: processed by authors using Jupyter-Notebook

constraints, taking into account technology and device limitations. Secondly, these algorithms are well-known and widely used in data mining classification techniques, as highlighted by Gong (2022). Furthermore, the AdaBoost algorithm was incorporated into the analysis as an additional method to complement the previous ensemble methods and to evaluate its performance in comparison. This decision was motivated by the goal of reducing the risk of overfitting, as elucidated by Singhal (2020).

The five algorithms are measured using the F1-Score because even though precision is the main goal of the writing team, Gupta & Nagadevara (2007) explain that recall is also needed to prove the model can be used even if only to certain limits. Considering that precision and recall are needed, the F1-Score is the right measure in measuring a model's ability (Lanier, 2020).

In calculating the F1-Score, a K-Fold Cross-Validation approach with 10 folds was employed. This choice of K-Fold Cross-Validation is particularly suited for situations where data scarcity is a concern, as elaborated by Hastie et al. (2009) Which is relevant in the context of this research due to limited positive label data availability. It's worth noting that Hastie et al. (2009) recommends using either 5 or 10 folds in cross-validation. In this study, 10-fold cross-validation was chosen, aligning with the default value provided by the Python library utilized.

Hastie et al. (2009) also explain Cross Validation will divide the training data into 10 parts with 9 parts for training data and 1 part for data testing which is then tested 10 times to get more

Table 4 F1-Score comparison of each models
Source: processed by authors using Jupyter-Notebook

Algorithm models	*F1-Score (mean)*
Logistic Regression	26,17%
Decision Tree	84,66%
Random Forest	90,67%
AdaBoost	86,60%
kNN	64,78%

```python
from sklearn.model_selection import GridSearchCV, KFold, cross_validate, LeaveOneOut
def tuning( x , y , model , params , cv=10 ,verbose=10 ):
    clf = GridSearchCV(estimator= model,
                       param_grid= params,
                       scoring= 'f1_macro',
                       cv= cv,
                       verbose= verbose, return_train_score=True,
                       n_jobs= -1)

    clf.fit( x , y )
    return clf.best_params_ , clf.best_score_, clf.cv_results_
```

```python
params = {"n_estimators":[i for i in range(5,100,2)], "max_depth":[i for i in range(1,30,1)]}
best_params, best_score, cv_result = tuning(features, target, RandomForestClassifier(),
                                            params,cv=cv, verbose=2)
print(" Best Parameters:",best_params,"with score of:",best_score)
```

```
Fitting 10 folds for each of 1392 candidates, totalling 13920 fits
 Best Parameters: {'max_depth': 17, 'n_estimators': 59} with score of: 0.9090344740117207
```

Figure 5 Syntax used for parameter tuning.
Source: processed by authors using Jupyter-Notebook

objective test results. After fitting is done to each algorithm, the models are compared using the subplots depicted in the Figure 3.

Based on Figure 3, the Random Forest is the best model because it has the top position and has a fairly low standard deviation compared to the other four models. The mean value of the F1-Score for each model can be seen in the Table 4.

From the Random Forest algorithm model obtained a True Positive Rate of 84%, a True Negative rate of 96%, a Positive Precision Rate of 91% and a Negative Precision Rate of 92%. When visualized with the confusion matrix, jupyter-notebook generated the Figure 4.

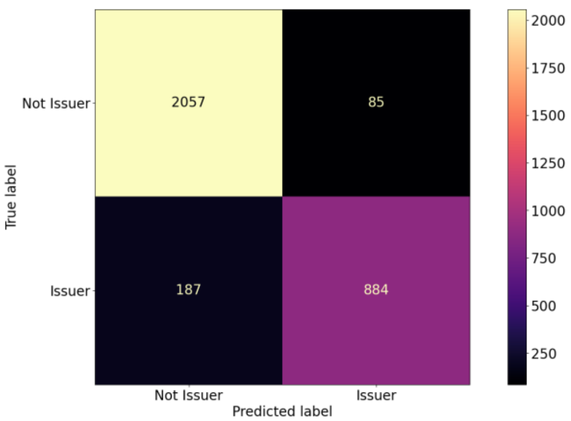


Figure 4 Confusion Matrix of Random Forest algorithm model
Source: processed by authors using Jupyter-Notebook

3.5 Model Evaluation

Out of the five previous algorithm models, the Random Forest algorithm model was chosen as the best algorithm model for the data you have. To optimize its performance, a hyperparameter tuning process was conducted using the GridSearchCV function. This approach, following the guidance provided by Badvelu (2021) on hyperparameter tuning, aimed to identify the best values for two key parameters: n-estimator, representing the number of trees in the model, and max_depth, indicating the size of an individual tree within the model. This parameter tuning process was executed within a Jupyter Notebook environment, as illustrated in Figure 5.

From the test we found out that the Random Forest algorithm model will work better, if it uses max_depth = 17 and n-estimators = 59 outperforming 1392 other candidates, with an F1-score of 90.90%. Drawing from the test results, we proceed with an in-depth analysis of the Random Forest model's performance after parameter tuning. Figure 6 displays the confusion matrix, offering a comprehensive view of the model's performance, especially in the context of tax audit purposes.

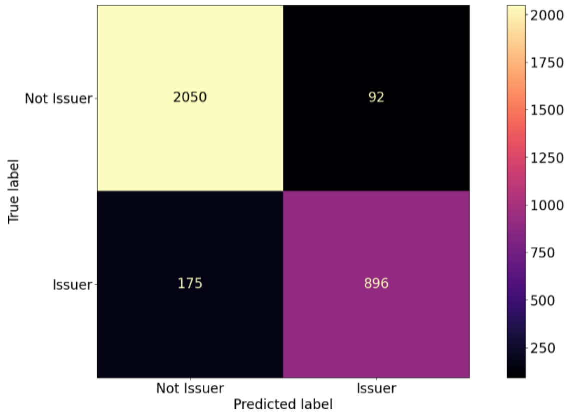


Figure 6 Confusion Matrix of Random Forest Model After Tuning
Source: processed by authors using Jupyter-Notebook

The Random Forest model after tuning has a True Positive Rate of 83.56%, a True Negative Rate of 96%, a Positive Precision Rate of 90.9%, and a Negative Precision Rate of 93%. Then for further evaluation with audit purposes, the model has the following capabilities.

a. PE of 895/895+176 = 83,56%;
b. EF of 1-(91+895)/3213 = 69,31%; and
c. SR of 895/(895+91) = 90,77%.

Out of the three evaluations based on research conducted by Gupta & Nagadevara (2007), the SR takes precedence among these three evaluations. The primary objective of the model is to optimize the allocation of audit resources effectively and efficiently. Consequently, when auditors investigate taxpayers suspected of being fraudulent invoice issuers, the model offers a high probability of yielding results that align with the predictions. The Prediction Efficiency (PE) metric supports the model's ability to outperform random audits in predicting fraudulent taxpayers. Significant EF signifies that employing the model can reduce audit effort. Finally, the model is saved using the pickle library in (.sav) format for usage in the next part.

3.6 Deployment

At the deployment stage, the saved model is implemented in a simple application using Streamlit. Streamlit is a python framework that is used to create web-based applications that are commonly used for data mining deployments. The

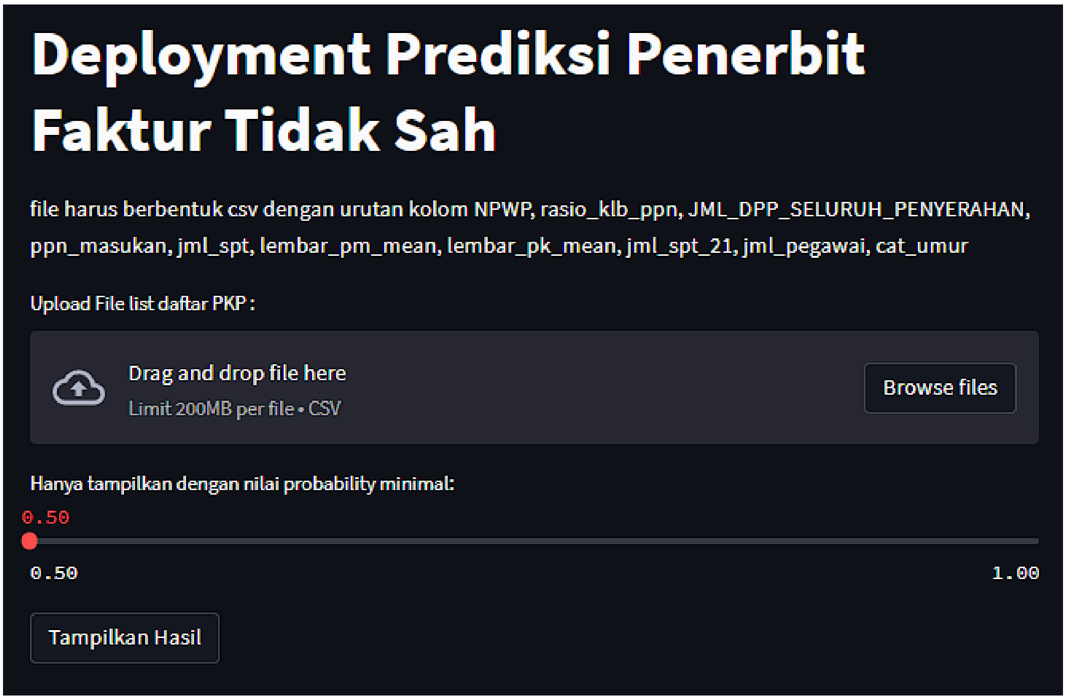
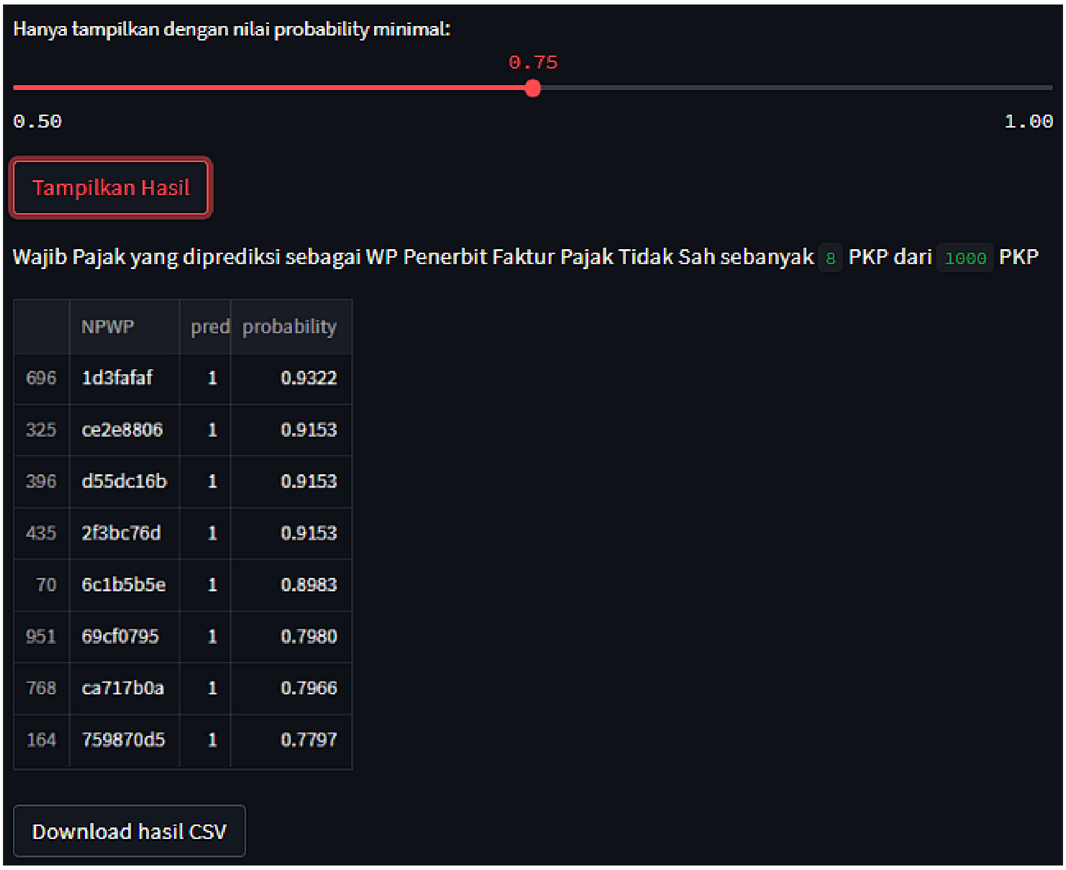


Figure 7 Interface of Application Deployment using Streamlit
Source: processed by authors using Jupyter-Notebook



	NPWP	pred	probability
696	1d3fafaf	1	0.9322
325	ce2e8806	1	0.9153
396	d55dc16b	1	0.9153
435	2f3bc76d	1	0.9153
70	6c1b5b5e	1	0.8983
951	69cf0795	1	0.7980
768	ca717b0a	1	0.7966
164	759870d5	1	0.7797

Figure 8 Display of taxpayers who are predicted as issuers by applications with a minimum probability value of 75%
Source: processed by authors using Jupyter-Notebook

interface of application that has been made has the following display in Figure 7.

This application allows an input file in the form of comma-separated values (.csv) with the order of the NPWP (Tax Identification Number) column, the ratio of under/overpayment of output VAT, consignment in a year, input VAT in a year, total VAT reports in a year, input tax invoice sheets in a year, output tax invoices sheets in a year, the number Monthly Income Tax Return (Article 21) per year, the number of employees, and the age category of VAT-registered Business. It also provides a slider to adjust the probability value provided by the Random Forest so that it can only display taxpayers who are predicted to be issuers with a certain level of probability.

Furthermore, the deployment data that has been prepared at the data preparation stage is uploaded via the Upload List menu of VAT-Registered Business lists and the slider is shifted to 0.75 for slider testing and simplification of the display of results.

As can be seen in Figure 8, the application will only display the NPWP of the taxpayer who is predicted to be the issuer and has a minimum probability of 75% according to the slider. The results show that there are 8 taxpayers who are suspected of being issuers with a probability exceeding 75% which are displayed sequentially based on the highest probability with the aim of the top taxpayers being prioritized for a tax audit.

4. CONCLUSION

Based on the analysis that has been described previously, we conclude that Random Forest is the most suitable algorithm model to use for the data owned compared to the Logistic Regression, Decision Tree, kNN, and AdaBoost algorithm models with F1-score of 90.61%. Subsequently, the parameters of the Random Forest algorithm were fine-tuned using GridSearchCV. The tuned Random Forest Algorithm, with adjusted max depth and n-estimators, achieved an improved F1-Score of 90.90%. This suggests that the parameter tuning may not had a significant impact on model performance, even though yielding better results compared to the untuned version.

The Random Forest algorithm model is evaluated to be able to correctly predict invalid invoice issuer taxpayers if each taxpayer is examined with a SR of 90.77%. Then the model can reduce the number of examinations compared to random examinations with a EF of 69.31%. The model can also classify fraudulent tax invoice issuers from all actual fraudulent tax invoice issuer taxpayers with a PE of 83.56%.

Furthermore, the model that has been created can be used in simple applications using Streamlit and can predict that 8 taxpayers are taxpayers who issue invoices with a probability that exceeds 75% of the 1000 taxpayer data in the deployment data. For deployment with different data, this can be done as long as the file uploaded is in the form of comma-separated values (.csv) and has a column for NPWP (Tax Identification Number) column, the ratio of under/overpayment of output VAT, consignment in a year, input VAT in a year, total VAT reports in a year, input tax invoice sheets in a year, output tax invoices sheets in a year, the number Monthly Income Tax Return (Article 21) per year, the number of employees, and the age category of VAT-registered business, sequentially.

5. IMPLICATIONS AND LIMITATIONS

From academics' standpoint, this study demonstrates the effectiveness of Data Analytics in detecting fraudulent tax activities. This can provide valuable insights for researchers in the field of machine learning, fraud detection, and taxation. Multiple variables used in this study should be able to inform future research on identifying additional variables that may improve the accuracy of fraud detection models.

In practitioners' side, this study provides a useful tool for tax authorities to detect fraudulent tax activities, reduce number of examinations, and improve the efficiency of tax audit processes. The findings can help tax practitioners to identify potential fraudulent tax activities among their clients and take appropriate measures to prevent or address them.

For regulators, the study provides evidence that machine learning algorithms can improve the effectiveness and efficiency of tax audit planning. Based on this finding, when implementing policies and procedures, we encourage the tax authorities to incorporate data analytic into their tax audit framework.

This study is limited by several factors. First, the audit coverage ratio is low at 0.86%, which limits the availability of positive label data. Additionally, negative label data may be invalid as some taxpayers may not have been examined or proven to be fraudulent invoice issuers. Another limitation is the limited data available to the authors, as authors has limited access to data. Furthermore, the study is dependent to the quality of the data sources and the accuracy of the predictions made by the Random Forest algorithm. There is also a lack of transparency in the decision-making process of the algorithm, which may be difficult to interpret by stakeholders not familiar with machine learning.

Finally, the study focuses solely on machine learning algorithms for detecting fraudulent tax activities and does not consider other factors that may contribute to tax fraud, such as economic conditions or social norms. These limitations should be considered when interpreting the findings of this study.

ACKNOWLEDGEMENTS

Our team would like to thank our colleague Ajar Parama Adhi, for his assistance in obtaining the data needed for this study. In addition, we acknowledge:

a. Mr Ag. Sigit Adi Satmoko, Section Head of Data Science at Directorate of Tax Data and Information, Directorate General of Taxes; and
b. Mr. Leonard Yulianus, Subdivision Head of General Affair at Sanggau Small Taxpayer Office, Directorate General of Taxes;

for the guidance and input given to the authors during this research process.

REFERENCES

[1] Ajzen, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

[2] Alm, J., & Malézieux, A. (2020). 40 years of tax evasion games: a meta-analysis. *Experimental Economics 2020 24:3*, *24*(3), 699–750. https://doi.org/10.1007/S10683-020-09679-3

[3] Badan Kebijakan Fiskal. (2023). *Pajak Pertambahan Nilai*. https://fiskal.kemenkeu.go.id/fiskalpedia/2021/07/13/173618726358430-pajak-pertambahan-nilai-ppn

[4] Badvelu, J. (2021, October 31). *Hyperparameter tuning for Machine learning models*. Towardsdatascience.Com. https://towardsdatascience.com/hyperparameter-tuning-for-machine-learning-models-1b80d783b946

[5] Brown, M. S. (2014). *Data mining for dummies*. John Wiley & Sons, Inc.

[6] Brownlee, J. (2021). *Random Oversampling and Undersampling for Imbalanced Classification*. MachineLearningMastery.Com. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/

[7] Brun, J.-P., Cebreiro Gomez, A., Julien, R., Waruguru, J., Ndubai, J. W., Owens, J., Siddhesh, R., & Soto, Y. E. (2022). Taxing Crime: A Whole-of-Government Approach to Fighting Corruption, Money Laundering, and Tax Crimes. *Taxing Crime: A Whole-of-Government Approach to Fighting Corruption, Money Laundering, and Tax Crimes*. https://doi.org/10.1596/978-1-4648-1873-8

[8] Carvalho, D., & Pacheco, L. M. (2014). *Tax compliance, corruption and deterrence: An application of the slippery model.* http://repositorio.uportu.pt:8080/handle/11328/837

[9] Chen, S., Chen, X., Cheng, Q., & Shevlin, T. (2010). Are family firms more tax aggressive than non-family firms? *Journal of Financial Economics*, *95*(1), 41–61. https://doi.org/10.1016/J.JFINECO.2009.02.003

[9] Detik.com. (2023, February 6). *Bikin Rugi Negara Rp 244 M, Dua Tersangka Faktur Pajak Fiktif Ditangkap*. Detik.Com. https://finance.detik.com/berita-ekonomi-bisnis/d-6553430/bikin-rugi-negara-rp-244-m-dua-tersangka-faktur-pajak-fiktif-ditangkap

[10] Directorate General of Taxes. (2022). *Directorate General of Taxes collaborative effort for national economic recovery*. https://www.pajak.go.id/sites/default/files/2022-11/Laporan%20Tahunan%20DJP%202021%20-%20English.pdf

[11] Gaspar, V., Jaramillo, L., & Wingender, P. (2016). *Tax Capacity and Growth: Is there a Tipping Point? IMF Working Paper Fiscal Affairs Department Tax Capacity and Growth: Is there a Tipping Point?* (WP/16/234). https://www.imf.org/external/pubs/ft/wp/2016/wp16234.pdf

[12] Gong, D. (2022, February 23). *Top 6 Machine Learning Algorithms for Classification*. TowardsDataScience.Com. https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

[13] Law Number 6 of 1983 concerning General Provisions and Tax Procedures as amended several times, the latest by Law Number 7 of 2021, (2021).

[14] Law Number 8 of 1983 on Value Added Tax of Goods and Services and Sales Tax on Luxury Goods as already amended several times the latest by the decree of Law Number 21 of 2021, (2021).

[15] Gupta, M., & Nagadevara, V. (2007). Audit selection strategy for improving tax compliance: application of data mining techniques. *Foundations of Risk-Based Audits. Proceedings of the Eleventh International Conference on e-Governance*, 28–30.

[16] Hasseldine, D. J., & Bebbington, K. J. (1991). Blending economic deterrence and fiscal psychology models in the design of responses to tax evasion: The New Zealand experience. *Journal of Economic Psychology*, *12*(2), 299–324. https://doi.org/10.1016/0167-4870(91)90018-O

[17] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. https://doi.org/10.1007/978-0-387-84858-7

[18] Larose, D. T. (2005). An introduction to data mining. *Traduction et Adaptation de Thierry Vallaud*.

[19] Mehta, P., Mathews, J., Kumar, S., Suryamukhi, K., Sobhan Babu, C., & Kasi Visweswara Rao, S. V. (2019). Big data analytics for nabbing fraudulent transactions in taxation system. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11514 LNCS*, 95–109. https://doi.org/10.1007/978-3-030-23551-2_7/COVER

[20] OECD. (2022). *The impact of COVID-19 on OECD tax revenues*. https://www.oecd.org/tax/tax-policy/revenue-statistics-highlights-brochure.pdf

[21] Pawki. (2023, August). *Different Types of Joins in Pandas*. Geeksforgeeks.Org. https://www.geeksforgeeks.org/different-types-of-joins-in-pandas/

[21] Regulation of the Director General of Taxes Number (No.) PER-19/PJ/2017 concerning Treatment of Issuance and/or Use of Illegal Tax Invoices by Taxpayers, (2017).

[22] Sandmo, A. (2005). The theory of tax evasion: A retrospective view. *National Tax Journal*, *58*(4), 643–663. https://doi.org/10.17310/NTJ.2005.4.02

[23] Saputra, H. (2019). Analisa kepatuhan pajak dengan pendekatan teori perilaku terencana (Theory of Planned Behavior): Terhadap wajib pajak orang pribadi di provinsi dki jakarta). *Jurnal Muara Ilmu Ekonomi Dan Bisnis*, *3*(1), 47–58. https://doi.org/10.24912/JMIEB.V3I1.2320

[24] Singhal, G. (2020, June 25). *Ensemble Methods in Machine Learning: Bagging Versus Boosting | Pluralsight*. Pluralsight. https://www.pluralsight.com/guides/ensemble-methods:-bagging-versus-boosting

[25] Subrahmanya, S. (2018, June). *Missing Data Imputation using Regression*. Kaggle.Com. https://www.kaggle.com/code/shashankasubrahmanya/missing-data-imputation-using-regression

[26] Wu, R.-S., Ou, C. S., Lin, H., Chang, S.-I., & Yen, D. C. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, *39*(10), 8769–8777.

[27] Yamin, L., & Putrantri, T. M. (2009). Model Penyelewengan Pajak Menggunakan Faktur Pajak Fiktif. *Bisnis & Birokrasi, Jurnal Ilmu Administrasi Dan Organisasi*, *16*(1), 1–7.