Scientax
Jurnal Kajian Ilmiah Perpajakan Indonesia

djp

VOLUME 6 NO. 2 | APRIL 2025 Journal Page: ejurnal.pajak.go.id ISSN 2686-5718

Data Mining to Detect Fraud Patterns in a Taxpayer's Financial Statement

Achmad Ginanjar[a], Agung Septia Wibawa[b]
[a]The University of Queensland, Brisbane, Australia. Email: aginoffice@gmail.com
[b]Nation Universitas Gajah Mada, Yogyakarta, Indonesia. Email: agung.septia@gmail.com

* Corresponding author: aginoffice@gmail.com

ABSTRACT

The application of machine learning in the analysis of financial statements is a relatively underexplored area compared to mainstream data mining fields, such as natural language processing (NLP) and image analysis, yet it holds significant potential. This study investigates the use of advanced linear regression techniques to identify patterns in taxpayers' financial statements, employing a conceptual approach that combines both vertical and horizontal financial statement analysis methods. Using financial statement data reported to the Indonesian tax administration and historical taxation audit records, this study determines the presence of identifiable patterns. This study applies linear regression to financial statement account values to measure changes over the years and uses yearly account values to create unique data points representing each entity. A clustering method is then employed to group entities with similar patterns. The findings indicate that the proposed method can effectively analyse how entities report their financial statements over time and cluster them based on the likelihood of committing fraud, as inferred from historical audit records. These patterns are validated by instances of underpayment or overpayment of corporate income taxes identified during tax audits. By examining the clustering results, the study reveals that certain clusters accurately align with labelled patterns, correctly identifying 2 of 3 labels. The comparison between unsupervised clustering and labelled criteria demonstrates a significant fitness probability.

Keywords: accounting, machine learning, clustering, horizontal analysis, vertical analysis

1. INTRODUCTION

Machine learning (ML) has introduced new approaches to understanding and solving problems in many field. Supported by advancement in technology, mathematics, and statistics, researchers now able to develop advanced machine-learning models that have never been thought before. Multidimensional visualization, for example, with the new formula like TSNE, can now be performed with a consumer processing unit like a laptop (Maaten & Hinton, 2008). The advancement of technology has also created a new way to create, store, and access data. The ease of data creation has brought the world into the era of big data. In the big data era, data are too large to be processed. Hence, more machine-learning approaches are needed to help humans in decision-making based on abundant data.

Finance and accounting are areas that have a bright future for the implementation of machine learning applications. Big accounting companies like EY, Pricewaterhouse Coopers (PwC), Deloitte, and KPMG, have developed their own methods to approach ML in pursuit of a better

DOI: 10.52869/st.v6i2.571
Received: August 16, 2022; Revised: March 7, 2025; Accepted: March 14, 2025; Published: April 30, 2025

How to Cite:
Ginanjar, A., & Wibowo, A. S. (2025). Data mining to detect fraud patterns in a taxpayer's financial statement. *Scientax: Jurnal Kajian Ilmiah Perpajakan Indonesia*, *6*(2), 135–150. https://doi.org/10.52869/st.v6i2.571

understanding of their data. Many of the available methods are used to understand data as a single data source, such as fraud detection in financial reports or determining a group of businesses based on a financial report. In addition, ML approaches can be used to determine the practice of tax fraud. Vasco et al. (2018) conducted a study on taxpayer segmentation for personal income tax fraud using data mining techniques with internal tax administration data. This study shows that by examining the characteristics of potential fraud taxpayers through the probability distribution of variables, the profile of potential fraud taxpayers can be clustered into groups that show the degree of probability of committing tax fraud.

Adopting these approaches, this paper employs machine learning techniques and analyses the financial statements reported to tax administration agencies to examine patterns of taxpayer accounting behaviors. The Model will focus on the degree of change in each company's accounts from its historical financial statement entity and then cluster them based on reporting patterns. Furthermore, clusters generated from the processes are then validated with tax audit reports that indicate whether taxpayers from each cluster are fraud or not fraud based on tax assessments issued. Based on these finding, this research is expected to be able to investigate the potential to build a model of data mining techniques combined with a financial statement analysis approach to be used to detect corporate income tax fraud from financial statement patterns.

2. LITERATURE REVIEW

Prior studies have shown that the use of data mining techniques can help with data analysis in accounting and taxation. Experiments were conducted by Cai et al. (2016) and Becirovic et al. (2020), who observed the usage of machine learning theory in accounting scopes. Both studies used a similar method in terms of accessing the accounting report values, and the experiments accessed the account value as a single value to solve their problem. The result was the cluster of each report that also represented the entity itself. This finding provides an opportunity for researchers to explore more relevant implementations of machine learning methods for accounting research. Furthermore, the use of machine learning tends to be useful to help researchers gain insights from untouched accounting data, such as using it to detect the probability of a company committing fraud. Many studies have focused on how this method can provide an efficient and effective approach to capture the pattern of fraudulent companies' financial statements. One example of the use of machine learning to detect financial fraud was proposed by Alvarez et al. (2017) in their study of the money laundering investigation of more than 600 Spanish companies. They combined the use of Benford's law as a detection tool for anomalies in books and accounting records and applied four machine learning models of pattern recognition (logistic regression, neural network, decision trees, and random forests) to detect the presence of financial fraud especially in money laundering. The study found that the random forest (Biau & Scornet, 2016) algorithms with SMOTE transformation obtained reasonably confident results, with 96.15% true negative results and 94,98% true positive results.

The use of the machine learning approach has also inspire researchers in the taxation field to conduct similar studies to examine taxation data. A study conducted by Dias et al. (2016) provides an analytical framework that combines dimensionality reduction and data mining techniques to obtain sample segmentation according to potential fraud possibility in personal income tax. Furthermore, Dias et al. (2016) classified individual taxpayers based on their tax evasion risk using cluster analysis methodology to organize observations into homogeneous groups, which would enable them to identify risky firms in a more effective manner.

However, we could not find any studies that studied the changes in financial account entries as a whole. In general, most studies employ classical accounting techniques, such as ratio analysis or vertical and horizontal to capture partial patterns in financial statements. Therefore, to fill this gap, this paper aims to investigate the possibility of studying overtime account changes as a whole with the help of a machine learning methodology. In addition, this study aims to detect specific patterns in changes in financial statements with the possibility of tax fraud. For example, we want to confirm whether this approach can provide evidence of a specific pattern in which fraudulent taxpayers behave in their tax assessment. This study, however, would like to focus more on evaluating value changes in financial statement accounts from historical values. The main idea of this study is to detect the behavior or patterns formed from an entity's financial statements over the years. This approach adopts horizontal analysis, which compares the values of financial statements throughout the years and detects anomalies that may rise, thereby indicating abnormal patterns or behavior in the financial statements. At the same time, this study also adopts vertical analysis by using all accounts as features. This study is a mix of machine learning approaches, as we employed both supervised and unsupervised methods. Linear regression involves supervised learning. In contrast, the proof-of-concept stage, which clusters the data, involves unsupervised learning. The study then matches the cluster produced with real-world labels produced by a complex process business from the Tax Office. The label is a permanent legal force; therefore, its trueness is valid. The experiment, unlike others, focuses on how posting behavior changes over time. This is achieved by exploiting the βi value of the regression formula 1.

$$y = \beta o + \beta ix + e, \{i \geq 1\} \quad (1)$$

The Bi above formula expresses how the change of x affect the value of y. This study using 10 years of data from each account and calculates the linear regression formula.

3. RESEARCH METHODOLOGY

3.1 Analysis Methods

3.1.1 Horizontal and Vertical Analysis

Auditors and fraud examiners can use financial analysis techniques to find anomalies and examine indications of impropriety in financial information. Unexpected lapses in values are likely to indicate wrongdoing but may also indicate illegal or fraudulent actions (Albrecht et al., 2006). The process of investigating financial statement fraud should begin by determining the areas of operation in identifying potential fraud schemes, noting the red flags associated with the identified schemes, establishing effective audit measures to look for indicators, and conducting further investigations to validate the detection or suspicion of red flags (Coopers, 2004). Financial analysis can also be employed to examine whether there are tax frauds, as taxpayers have to attach financial statements to their tax reports. Horizontal and vertical analyses are sample forms of classical financial analysis (Albrecht et al., 2006). These financial analyses are proficient techniques that can be used for detecting fraud as well as spotlighting areas of concern (Edelman, 2009). Horizontal Analysis basically analyses the changes in the value of an account in financial statements from one period to another. In contrast, Vertical Analysis analyses the proportion of an account's value from a base account (i.e. sales or total assets) during a reporting period. In some cases, there are frauds that the standard use of ratio analysis in audit techniques can miss, but they can be effectively detected using horizontal and vertical analysis (Albrecht et al., 2006).

However, performing horizontal and vertical analysis in large sets of data requires enormous effort and may not be an efficient tool if there are many entities as well as large number of accounts that need to be analysed. Therefore, for this study, we use regression slopes to proxy changes in the financial report accounts, as performed in the horizontal analysis. Furthermore, we use account value and period as features in our clustering method so that the proportion of account value can be captured, as is typically done in vertical analysis.

3.1.2 Linear Regression Analysis

Regression is a mathematical model in which the result is a number (float or integer) that can take any value (continuous). Linear means that the ideal problem-solving result follows a straight line, either decreasing or increasing in a consistent value. The linear regression is described as follows:

$$y = \beta o + \beta ix + e, \{i \geq 1\} \quad (1)$$

Where:

- y = is the model result or predicted value. In this experiment, this value is the value of an account for a specific year x.
- x = is the feature value or the known variable. In this experiment, this value is the value of the year of y
- β_0= is the intercept value. This represents the value of y when x is equal to 0
- β_i= is the dimensional parameter value. It can also be understandable how large the changes in y are for each change in x. Also known as slope.
- e = error noise. Typically present in the form of a normal distribution random value.

3.2 Data Processing

This study uses data for 10 years of accounting report data. The study year was 2011-2020. For each entity being observed, there might be several accounts consisting of 10 years of data. All 10 years of data must exist regardless of the account name, which means that an entity has sets of its own unique account names. In this study, this condition will be called as the total year and data availability limitations. The data structure is illustrated in Table 1.

In Table 1, entity ABC has two account names that are used for calculation. The account names are 'cash' and 'Short-term Investment'. Both account names include 10 years of data from the studied years. Similarly, entity XYZ has 'Trade Receivable from Related Parties' consisting of data from 2011 to 2020.

3.2.1 Horizontal and Vertical Analysis

The dataset used in this study initially contained over 900 account names representing a total of 11,102 entity-years. However, due to the variability and completeness of each entity's annual data, not all account names were used for the analysis. Below are some samples of the account names used. The original data cannot be shared due to Indonesia's legal restrictions on maintaining the confidentiality of taxpayer data.

Table 2 shows the total of 617 account names used in the experiment. This study did not evaluate the details or type of account names. For each entity, all 617 account names within 10 years of the

Table 1
Attributes used in the analysis

Taxpayer ID	Account Name	Year	Value (IDR)
ABC	cash	2011	200000
. . .	. . .	. . .	. . .
ABC	cash	2020	700000
ABC	Short-term Invesment	2011	1200000
. . .	. . .	. . .	. . .
ABC	Short-term Invesment	2020	5700000
XYZ	Trade Receivable from Related Parties	2011	4000000000
. . .	. . .	. . .	. . .

observed data create a unique dimension for the respected entity.

3.2.2 Applying Regression

Unlike traditional horizontal and vertical analyses, this study uses a different strategy to analyse the data. In the horizontal analysis, the differences in account values for some periods are calculated, while in the vertical analysis, the percentages of an account relative to the overall account values are used (Albrecht et al.,2006). In contrast, this study evaluates both vertical and horizontal values using its algorithm. The vertical values, which are an account's yearly values, are the base values for the regression calculation. A linear regression formula is used to extract 10 years of 'degree of changes' from each account. Generally, the 'degree of change' is also known as the slope.

Table 2
Account name in total

ID	Account Name
1	Cash and cash equivalents
2	Short-term Investment
3	Trade Receivable from Third Parties
4	Trade Receivable from Related Parties
5	Other Receivable from Third Parties
6	Other Receivable from Related Parties
7	Allowance for Uncollectible Accounts
8	Inventory
9	Prepaid Expense
10	Purchase Down Payment
11	Other Current Assets
12	Long-term Receivable
13	Land and Building
14	Other Fixed Assets
15	Subtract: Accumulated Depreciation
16	Investment in Associated Companies
17	Other Long-term Investment
…	…
…	…
…	…
617	Intangible Assets

Other researchers have demonstrated how to process a single record as a feature for the clustering method. However, this study focuses on the 'degree of change' in each account as a feature. This degree of change is collected by using the B_i value from the above linear regression formula. The B_i value represent the relation between x and y. This B_i can be seen as how change x directly affects y. Therefore, this value can also be interpreted as the 'degree of changes' value.

3.2.3 Dimensional Reduction

This section discusses the dimensional reduction process for the easiness of clustering and visualization. As proposed by Bellman (2013) about the course of dimensionality, when processing high-dimensional data, various phenomena might occur that do not occur in low-dimensional data analysis. To reduce the possibility of this problem, the TSNE algorithm (Maaten & Hinton, 2008) and PCA algorithm are applied to the data.

In general, the TSNE algorithm results can be seen as an unsupervised clustering model. It works by pulling together data points that have similar mathematical features. This study applies TSNE with n = 2. This means that the result of the model has two dimensional features. On the other hand, (PCA) works by determining the most important value. The principle of PCA is simple: reduce dimensionality while preserving statistical information as much as possible (Jollife & Cadima, 2016). PCA works by extracting a set of values that represent the distribution of the original dataset. The set of values is usually marked as $\{PCA_n\}$ where $n \geq 1$. For example, PCA_1 is considered more significant to the dataset compared to PCA_2, and so on. This study applies PCA_1 and $PCA_{\neq 1}$represents (x, y) a two-dimensional area.

4. RESULTS AND DISCUSSIONS

4.1 Analysis Methods

The result from applying linear regression to the data is data that consists of 'id object', id account, slope βi and intercept $\beta 0$. The 'id object' represents a unique entity. The 'id account' is the representation of the account's name in an accounting journal, such as 'cash', 'debt', etc.

The 'slope' is the slope value, previously called 'degree of change', from the linear regression of the correspondence account. The 'intercept' is the intercept value from linear regression of the correspondence account. All data that is shown in this paper have been masked to represent only visual concepts. The table mined from linear regression is shown in Table 3.

The data consists of 40.000 rows from 11102 entities. The unique number of account names is 617. As a common practice, an account name might have variations and typos. However, account names are not treated with any treatment, such as corrections to manual or state-of-the-art NLP algorithms.

After the previous step, a table was built. The table has an entity as an index and an account name as a column name. The values of the table are the B1 values referring to the index and column names. This process was performed by pivoting the data, see Table 4.

4.2 Visualized Data

4.2.1 TSNE

The result of applying TSNE algorithm to the pivoted data with n = 2 can be seen in Table 5 The two-dimensional data are meant for easiness of data visualization. In this experiment, the TSNE result has "Feat 1" and "Feat 2" as features. "Feat 1" is used as x point and "Feat 2" is used as y point. The result then plotted for visualization. At this level, the visualization does not explain a lot. However, several conclusions can be drawn by examining the plot. From the TSNE Visualization in Figure 1. First, several individual clusters can be

Table 3

Regression Results

id_object	id account	β_1	β_0
1	0	-3,15E+16	6,35E+19
1	1	-6,81E+15	1,38E+19
1	2	1,44E+17	-2,90E+20
1	3	-3,15E+16	6,35E+19
…		…	…
400000	1	1,75E+15	-3,53E+18
400000	2	-2,80E+15	5,67E+18
400000	3	2,10E+15	-4,20E+18
400000	4	2,41E+15	-4,82E+18
400000	5	1,24E+15	-2,52E+18

Table 4
Pivoted Data

	1	2	3	4	5	6	7	8	9	10	…	875
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
0.000000e+00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0
-3.266133e+09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	…	0.0

Table 5
TSNE Result

Index	Feat 1	Feat 2
0	2.292.285	6.388.725.586
1	358.360.596	663.375.610
2	290.112.335	424.451.019
3	34.406.998	952.894.653
4	34.406.998	952.894.653
5	34.406.998	952.894.653
6	28.423.534	650.764.404
7	34.406.998	952.894.653
8	34.406.998	952.894.653
9	17.357.542	693.593.445

clearly identified. TSNE works by pulling together similar point. The individual clusters spotted on the figure can be seen as similar points. In other words, there are patterns in the data points.

4.2.2 PCA

The principal component analysis (PCA) algorithm works quite differently. PCA attempts to substitute data with similar variance to the original data. The PCA results of this study are presented below. This study attempts to evaluate the value of PCA1, PCA2, PCA3, and PCA4. As shown in Figure 2, no pattern was clearly observed in the data.

Figure 1

TNSE Visualisation

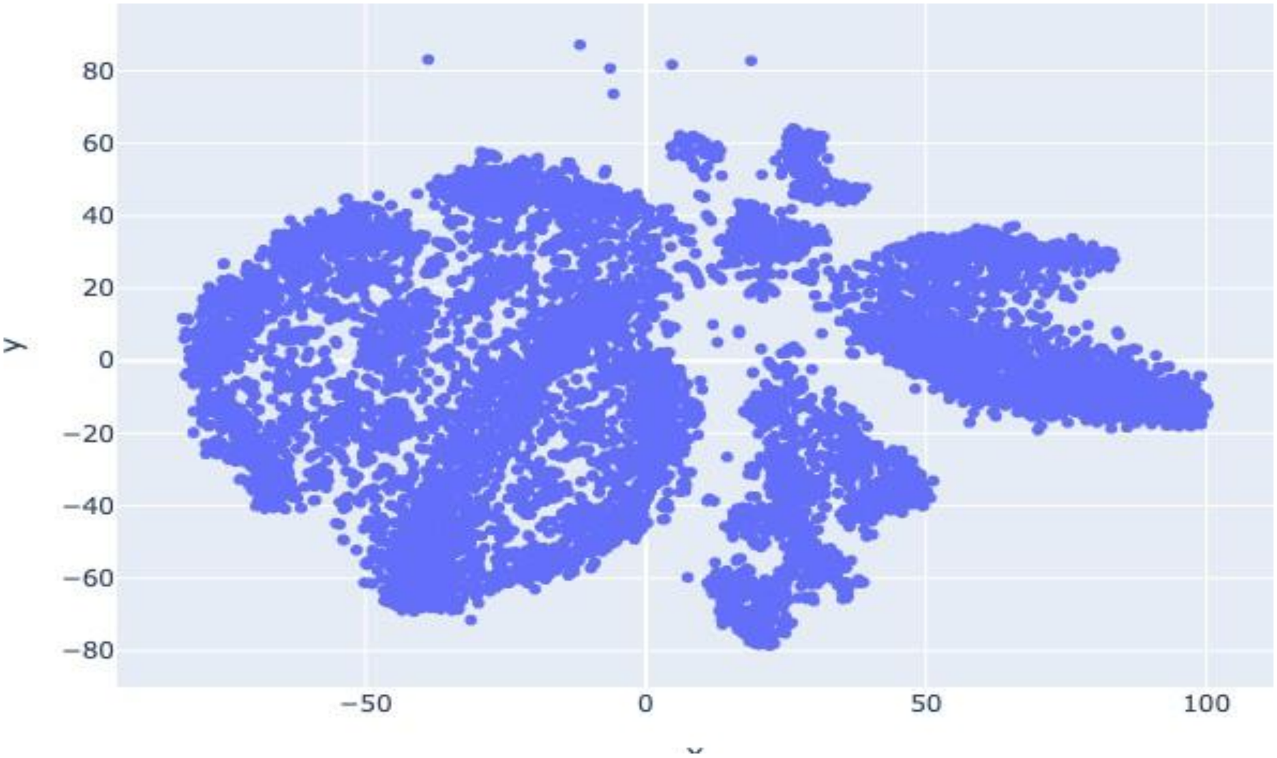


Figure 2

PCA Visualisation

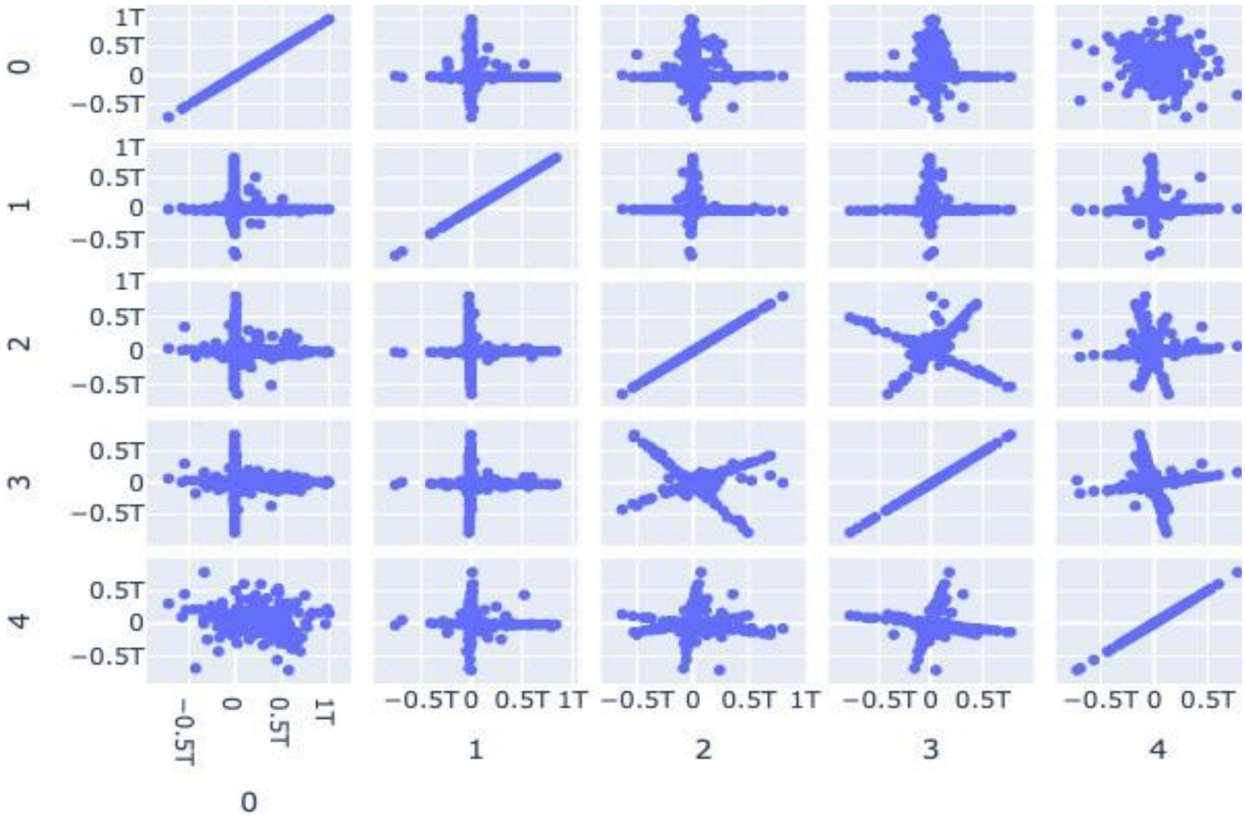
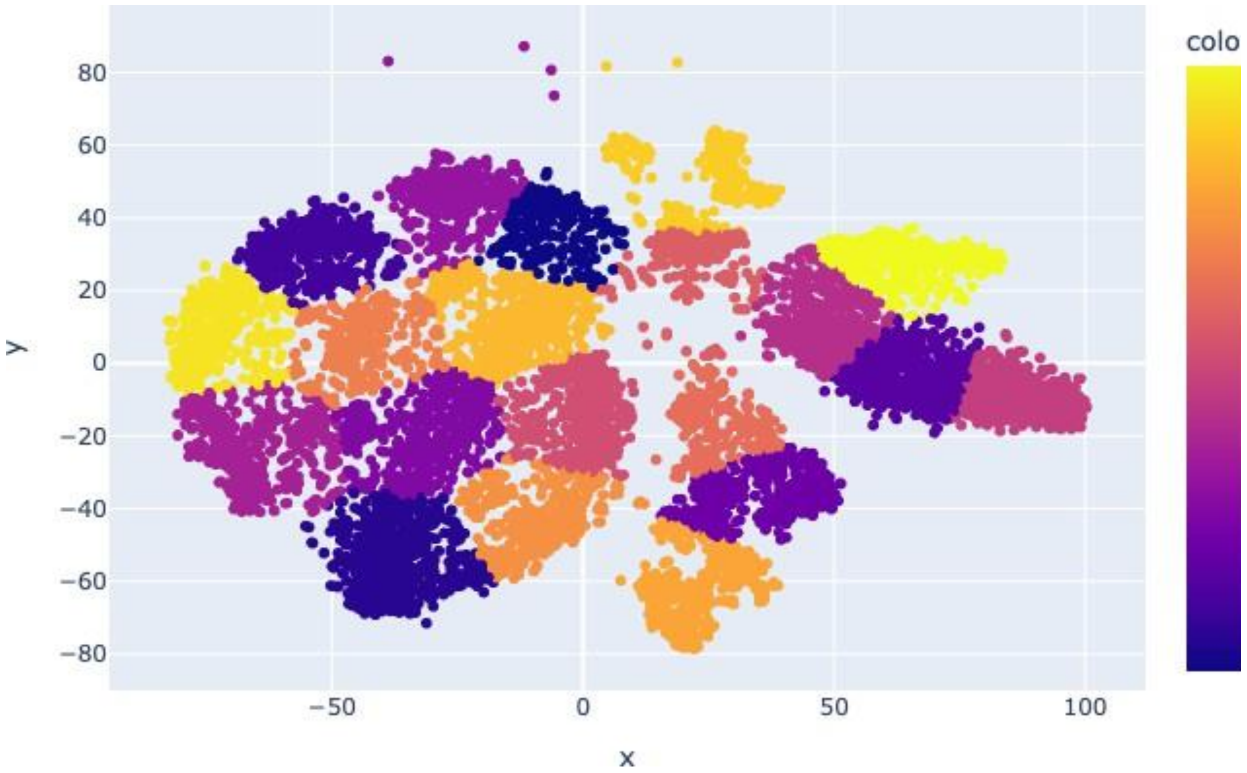


Figure 3

K- Means Plot on TSNE



4.3 K-Means Clustering

The TSNE data were then clustered using the K-Means algorithm. For this experiment, we used n=20. This also means that the experiment attempt to predict 20 clusters in the dataset. At this point, the use of K-Means was solely used to cluster data to obtain 20 regions. The result will be used to compare the region from the K-Means cluster with real label data. Figure 3 shows the visualization of k-means.

4.4 Proof of Concept with Labels

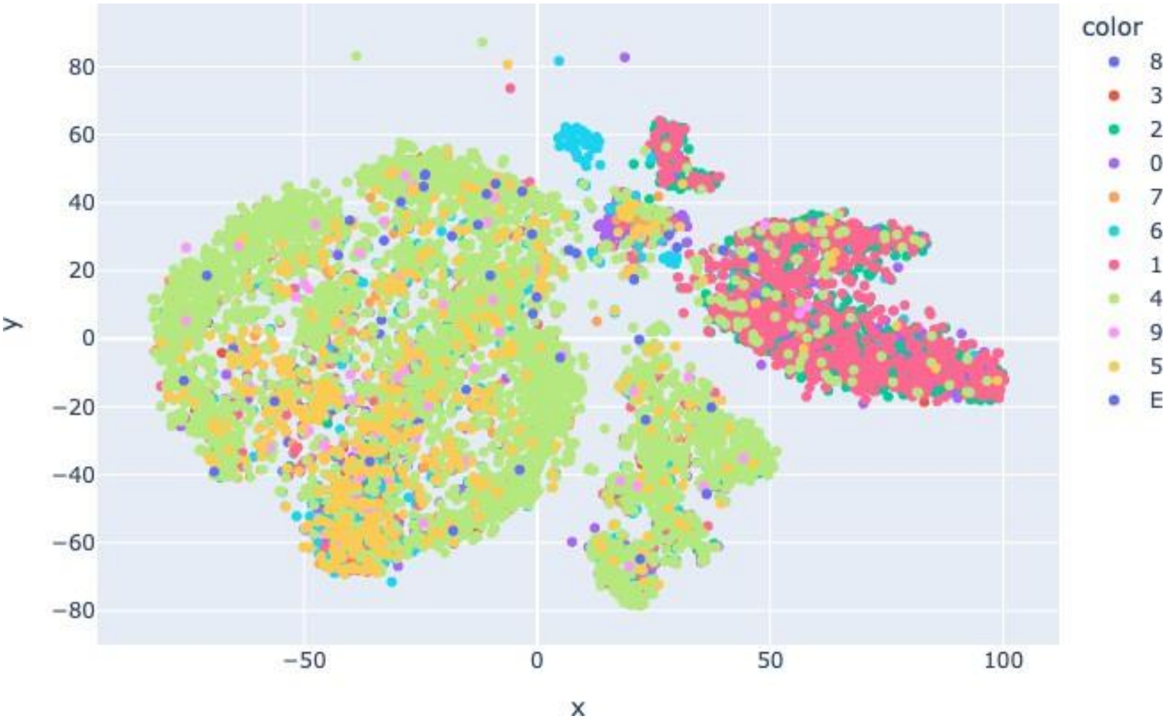
4.4.1 Business Categories

Business categories identified by 'Klasifikasi Lapangan Usaha' (KLU) are the first categories to be tested. The label comes from 11 general categories of taxpayer business. The 11 categories are labels assigned to the respective taxpayers during registration and are included in the experiment data. The results of the plotted data are shown in Figure 4. Figure 4 shows that some clusters have some dominant colours while others are not. Some business categories can be well separated by the TSNE cluster. In other words, it can be said that for this type of business area, there are patterns in how accountants log their reports over the studied years. In contrast, for some business categories, the algorithm could not detect any similarities in patterns. This means that the algorithm is not fit for the data, or the data labelling process itself has a human error.

Further evaluation of the experimental result confirmed the above conclusion. The confusion matrix table in Table 6 between the K-Means cluster and the Business category on the TSNE cluster shows that category 4 (Construction, Wholesale and Retail Trade; Car and Motor- cycle Repair and Maintenance) dominates almost every cluster. Category 2 (Processing industry) dominates clusters 0 and 2, with respective scores of 0.67 and 0.5. the numbers are not significant because they are close to 0.5 or random probability..

Figure 4

Plot of Business Area on TSNE



4.4.2 Fraud Group

- Labelled Criteria 1 in this section, label criteria 1 are applied. The label criteria 1 is about the occurrence of a taxpayer applying for an overpayment on their past tax payment. When a taxpayer applies for overpayment for any value, in this study, the taxpayer is assigned 1 or True or yellow. Value 0 or False or blue colour is given for taxpayer data that does not apply any overpayment.

  Figure 5 shows that there are clear patterns of the labelled case. Most clusters are completely covered with 0 or blue colour. This means that most taxpayers who were observed did not apply for overpayment. The blue colour is concentrated in the middle. For the cluster in the middle is completely blue. Yellow or true lines can be seen in the outer area of the TSNE cluster. Several clusters at the bottom are dominated by yellow. The observation of the confusion matrix in Table 8 between K-Means and Fraud Criteria Label for the TSNE cluster confirms the above result. 10 out of 20 of the K-Means clusters were dominated by the Non-fraud label, with a probability above 0.7. Clusters 3, 6, 14, and 19 have probabilities greater than 0.7 for fraud labels. The remaining 7 of K-Means clusters have probabilities below 0.7, which are not considered significant.
- Labelled Criteria 2 The label criteria 2 is about the occurrence of a taxpayer being assigned as an underpayment of their past tax payment. When a taxpayer is assigned an underpayment for any value, in this study, the taxpayer is labelled with 1 or True or yellow. Value 0 or False or blue colour is given for taxpayer data that did not show any signs of underpayment. Figure 6 shows a notable pattern between labelled criteria 1 and 2. The figure shows a fairly balanced composition of yellow and blue. The central TSNE cluster primarily exhibits a value of 0, which corresponds to blue. Some clusters on the right and bottom parts of the figure are yellow, indicating they are labelled as true. In the larger cluster, areas closer to the centre are blue, suggesting they are labelled as false. Conversely, as one moves further away from the centre, the likelihood of encountering yellow is increased. The confusion matrix in Table 9 between K-Means cluster versus labelled criteria 1 agrees with the previous figure. 10 K-Means clusters out of 20 are labelled with

True with probability more than 0.7 or 70%. The remainder of the cluster has no significant probability because the probability for both labels is near 0.5 or 50% of chance.

Figure 5

Fraud Criteria 1 plot on TSNE

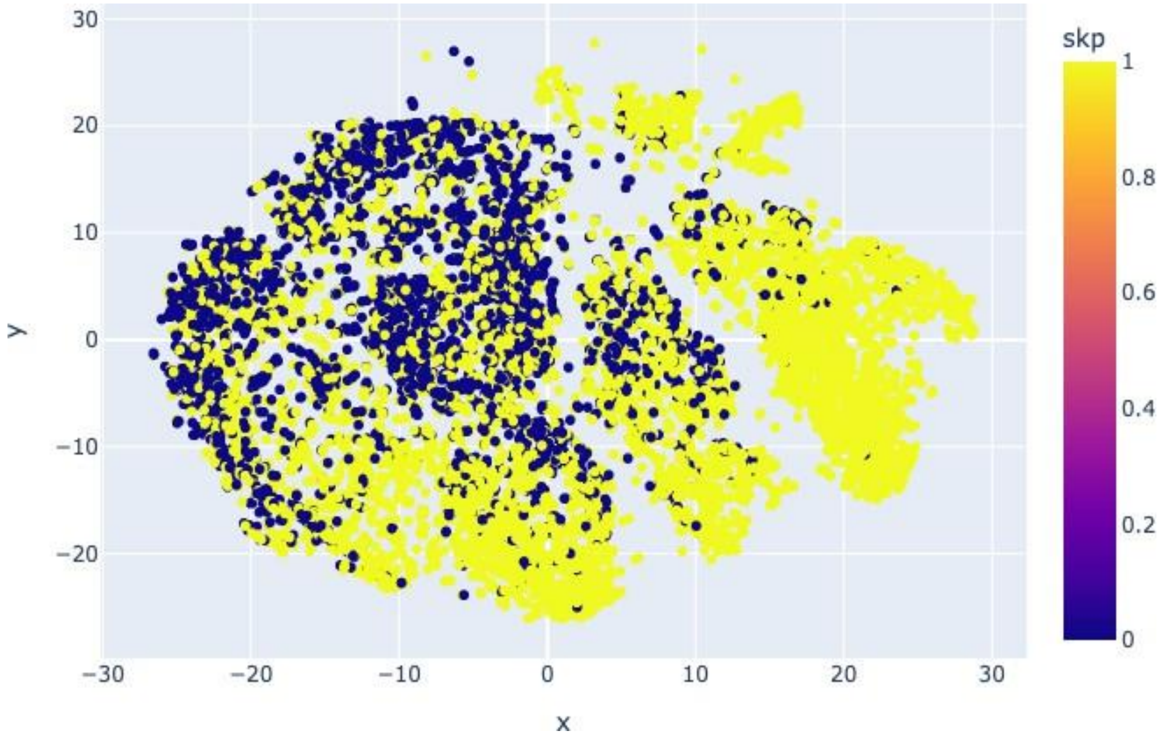


Figure 6

Fraud Criteria 2 plot on TSNE

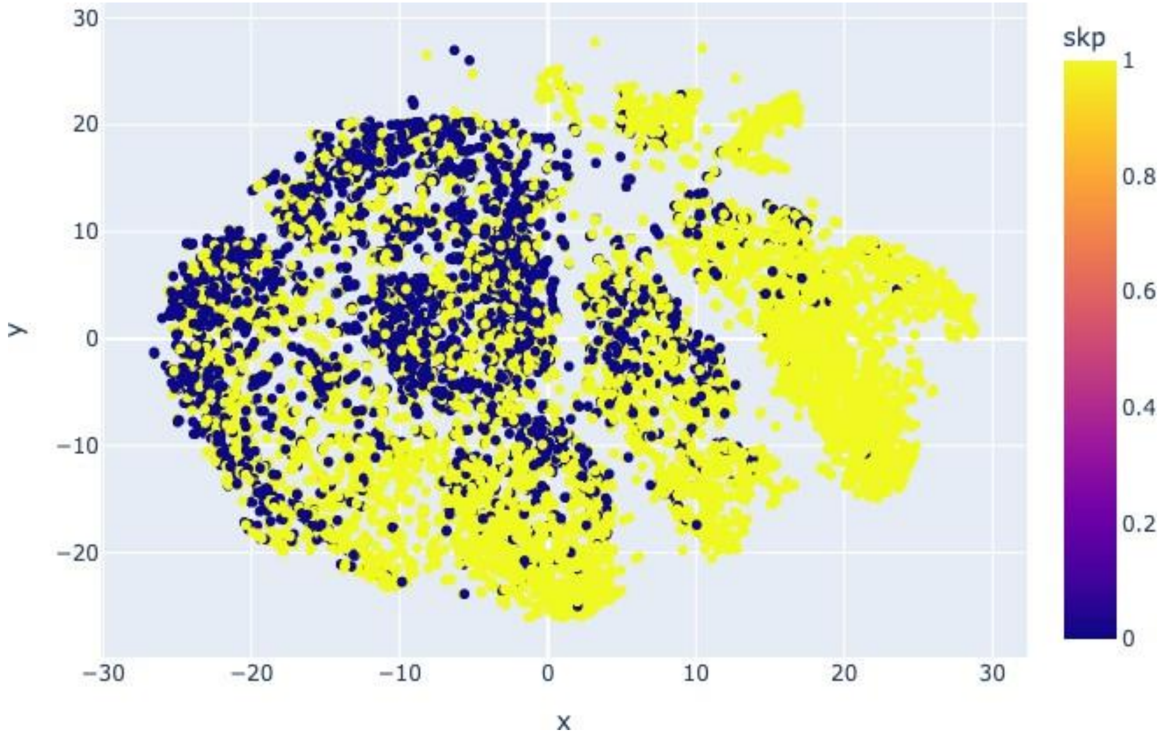


Table 6

K-Means vs Business Group

BC	0	1	2	3	4	5	6	7	8	9	E
Cluster											
0	0.01	0.23	0.67	0.04	0.01	0.00	0.03	-	-	-	0.00
1	0.08	0.03	0.03	0.03	0.45	0.13	0.12	0.07	0.05	0.02	-
2	0.04	0.36	0.50	0.06	0.03	0.00	0.01	0.00	0.00	-	-
3	0.07	0.05	0.02	0.01	0.56	0.10	0.06	0.10	0.01	0.01	0.00
4	0.01	0.01	0.01	0.01	0.82	0.05	0.03	0.06	0.01	0.00	-
5	0.13	0.03	0.03	0.05	0.25	0.17	0.19	0.08	0.06	0.02	-
6	0.04	0.07	0.04	0.01	0.72	0.03	0.05	0.03	0.00	0.01	-
7	0.01	0.04	0.01	0.01	0.79	0.05	0.04	0.03	0.02	0.01	0.01
8	0.01	0.02	0.01	0.00	0.78	0.04	0.02	0.10	0.02	0.00	0.00
9	0.03	0.10	0.05	0.01	0.56	0.06	0.09	0.03	0.03	0.01	0.02
10	0.04	0.43	0.39	0.06	0.07	0.00	0.01	0.00	0.00	0.00	0.00
11	0.01	0.03	0.01	-	0.83	0.02	0.03	0.07	0.01	0.01	0.00
12	0.03	0.06	0.03	0.01	0.77	0.02	0.03	0.03	0.01	0.00	0.00
13	0.04	0.38	0.47	0.05	0.04	0.01	0.02	0.01	-	-	-
14	0.02	0.03	0.01	0.00	0.72	0.05	0.08	0.06	0.02	0.01	0.00
15	0.04	0.04	0.05	0.01	0.76	0.02	0.04	0.02	0.00	0.01	0.00
16	0.03	0.31	0.57	0.05	0.02	0.01	0.00	0.00	-	-	-
17	0.03	0.03	0.02	0.01	0.69	0.07	0.03	0.10	0.02	0.00	-
18	0.03	0.04	0.02	0.00	0.67	0.06	0.08	0.07	0.01	0.01	-
19	0.31	0.03	0.07	0.04	0.16	0.03	0.27	0.07	0.01	0.01	0.01

Table 7

Business Categories

Digit	Category
0	Mining Agriculture, Forestry, and Fisheries
1	Processing industry
2	Processing industry
3	Processing industry Water Supply, Waste Management and Recycling, Disposal and Cleaning of Waste and Garbage Procurement of Electricity, Gas, Steam/Hot Water and Cold Air
4	Construction Wholesale And Retail Trade; Car and Motorcycle Repair and Maintenance Transportation And Warehousing
5	Information and Communication Provision of Accommodation and Provision of Food and Drink Transportation And Warehousing
6	Information and Communication Financial Services and Insurance Professional, Scientific, and Technical Services Real Estate
7	Rental Services, Employment, Travel Agencies, and Other Business Support Professional, Scientific, and Technical Services
8	Government Administration and Mandatory Social Security Health Services and Social Activities Education Services Rental Services, Employment, Travel Agencies, and Other Business Support
9	Individual Services Serving Households; Activities that produce goods and services by households that are used alone to meet needs Culture, Entertainment, and Recreation Activities of International Agencies and Other Extra International Agencies Other Service Activities

Table 8

K-Means vs Fraud Criteria 1

Criteria 1	Fraud Criteria	
K-means Cluster	Not Fraud	Fraud
0	0.382550	0.617450
1	0.834739	0.165261
2	0.736527	0.263473
3	0.233390	0.766610
4	0.852126	0.147874
5	0.819149	0.180851
6	0.146771	0.853229
7	0.337950	0.662050
8	0.918330	0.081670
9	0.920181	0.079819
10	0.516579	0.483421
11	0.937402	0.062598
12	0.616123	0.383877
13	0.912621	0.087379
14	0.073298	0.926702
15	0.559271	0.440729
16	0.730882	0.269118
17	0.598485	0.401515
18	0.870073	0.129927
19	0.025455	0.974545

Table 9

K-Means vs Fraud Criteria 2

Criteria 2	Fraud Criteria	
K-means Cluster	Not Fraud	Fraud
0	0.001736	0.998264
1	0.556180	0.443820
2	0.137821	0.862179
3	0.514979	0.485021
4	0.036125	0.963875
5	0.642741	0.357259
6	0.397408	0.602592
7	0.556593	0.443407
8	0.054662	0.945338
9	0.038270	0.961730
10	0.613953	0.386047
11	0.630691	0.369309
12	0.258232	0.741768
13	0.114413	0.885587
14	0.668975	0.331025
15	0.017713	0.982287
16	0.220848	0.779152
17	0.202797	0.797203
18	0.569069	0.430931
19	0.436077	0.563923

5. CONCLUSION

Our study focuses on identifying patterns through linear regression applied to annual financial statements submitted for tax reporting purposes. In addition, it utilizes t-SNE (t-distributed Stochastic Neighbour Embedding) to visualize the clustering of the data clearly. Our research demonstrates how to analyse the slope of yearly accounting reports to uncover the patterns that exist. This study provides evidence that linear regression can be effectively employed to identify how accountants log their bookkeeping and present their financial statements. The findings suggest that the detected patterns may indicate specific behaviours related to certain labels or purposes, particularly in identifying fraudulent taxpayers in corporate income tax reporting. Initial findings from preliminary research indicate the potential of using linear regression to detect patterns in yearly accounting reports.

Nevertheless, this study has limitations that should be addressed in future research. First, it does not account for the magnitude of fraud, which could be indicated by the amount of underpayment or overpayment. Consideration of the magnitude may provide additional insights into the patterns detected across different levels of fraud. Second, this study is based on the assumption that data can be represented as linear functions. Due to the size of the dataset, the linearity of the data was not tested explicitly; future research may explore other algorithms, including nonlinear ones. Lastly, while t-SNE was used as the clustering method in this study, it was not originally designed for this purpose. However, the nature of the algorithm allows for such intrepetations. Other unsupervised clustering methods may be better suited for this data, although they may require more effort to visualize and understand. Future studies could build on the findings of this research by, incorporating better data preprocessing and exploring different (nonlinear) algortithms to conduct similar experiments.

REFERENCES

Albrecht, W. S., Albrecht, C. O., Albrecht, C. C., & Zimbelman, M. F. (2006). *Fraud examination* (p. 696). Thomson South-Western.

Álvarez-Jareño, J. A., Badal-Valero, E., & Pavía, J. M. (2017). *Using machine learning for financial fraud detection in the accounts of companies investigated for money laundering* (No. 2017/07). University of Valencia.

Becirovic, S., Zunic, E., & Donko, D. (2020, March). A case study of cluster-based and histogram-based multivariate anomaly detection approach in general ledgers. In *2020 19th International Symposium Infoteh-Jahorina (INFOTEH)* (pp. 1–6). IEEE. https://doi.org/10.1109/INFOTEH48170.2020.9066342

Bellman, R. (2013). *Dynamic programming*. Dover Publications.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Cai, F., Le-Khac, N. A., & Kechadi, T. (2016). Clustering approaches for financial data analysis: A survey. *arXiv preprint arXiv:1609.08520*. https://doi.org/10.48550/arXiv.1609.08520

Coopers, P. H. (2004). *The emerging role of internal audit in mitigating fraud and reputation risk*. PricewaterhouseCoopers.

Dias, A., Pinto, C., Batista, J., & Neves, E. (2016). Signaling tax evasion, financial ratios and cluster analysis. *BIS Quarterly Review*, *51*, 1–34.

Edelman, B. (2009). Deterring online advertising fraud through optimal payment in arrears. In R. Dingledine & P. Golle (Eds.), *Financial cryptography and data security: 13th international conference, FC 2009, Accra Beach, Barbados, February 23–26, 2009. Revised selected papers* (Vol. 5628, pp. 17–31). Springer. https://doi.org/10.1007/978-3-642-03549-4_2

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vasco, C. G., Rodríguez, M. J. D., de Lucas Santos, S., & de Madrid, U. A. (2018). Characterization and detection of potential fraud taxpayers in personal income tax using data mining techniques. *Journal of Applied Economic Sciences*, *13*(2), 559–569.