p-ISSN : 2443-2210
e-ISSN : 2443-2229

Deteksi Tindak Kecurangan Penjualan di Perusahaan Distribusi Menggunakan *Machine Learning*

http://dx.doi.org/10.28932/jutisi.v9i2.6932

Riwayat Artikel
Received: 12 Juli 2023 | *Final Revision*: 24 Juli 2023 | *Accepted*: 25 Juli 2023

Budi Wibowo Suhanjoyo✉ #1, Bernard Renaldy Suteja *2, Hapnes Toba #3

# Magister Ilmu Komputer Fakultas Teknologi Informasi, Universitas Kristen Maranatha
Jl. Prof. drg. Surya Sumantri, M.P.H. No.65, Bandung, 40164, Indonesia

[1]2179005@maranatha.ac.id
[2]bernard.rs@it.maranatha.edu
[3]hapnes.toba@it.maranatha.edu
✉Corresponding author: 2179005@maranatha.ac.id

Abstract — **Penjualan pada perusahaan distribusi menjadi salah satu tempat sering terjadinya tindak kecurangan. Tindak kecurangan tersebut terjadi dengan berbagai cara dan menimbulkan kerugian yang berdampak cukup besar bagi perusahaan. Tindak Kecurangan tersebut memiliki pola tertentu. Pola-pola yang terjadi pada praktek tindak kecurangan tersebut dipelajari oleh para ahli internal auditor perusahaan. Pengalaman para ahli tersebut diolah menjadi suatu sistem yang disebut dengan *Expert System*. Diperlukan suatu alat bantu yang berbasis teknologi agar tindak kecurangan bagian penjualan dapat terdeteksi sejak dini. Target penelitian ini adalah agar dapat memberikan manfaat bagi perusahaan dengan deteksi dini tindak kecurangan pada bagian penjualan. Pada saat penelitian ini dilakukan, belum ada ditemukan penelitian sejenis dengan objek yang sama. Pada penelitian ini akan dilakukan komparasi dari berbagai model algoritma *machine learning* dengan tujuan agar dapat diketahui apakah dengan menggunakan tekonologi *machine learning* dapat membantu mendeteksi tindak kecurangan dengan nilai akurasi tinggi. Metode algoritma yang digunakan adalah metode *supervised learning*. Model algoritma yang akan dikomparasi adalah *Decision Tree, K-Nearest neighbors, Random Forest, Support Vector Machine* dan Regresi Logistik. Diharapkan dengan menggunakan teknologi *machine learning* maka tindak kecurangan dapat dideteksi sejak dini, sehingga tingkat kerugian dan risiko penjualan dapat diminimalkan.**

Kata kunci— ***Fraud*; *Machine Learning*; *Supervised Learning*.**

Fraud Detection in Sales of Distribution Companies Using Machine Learning

Abstract — ***The sales department of a distribution company is one of the places where fraud often occurs. This fraud occurs in various ways and causes massive losses for the company. These frauds have certain patterns. The patterns that occur in these practices are studied by the company's internal auditor experts. The experience of these experts is processed into a system called the Expert System. The support of a technology-based tool is needed to detect sales fraud early. The purpose of this research is to be able to provide benefits for companies with early detection of fraud in the sales department. At the time this research was conducted, no similar research with the same object had been found. In this research, a comparison of various machine learning algorithm models will be carried out to***

know whether using machine learning technology can help detect fraud with a high accuracy value. The algorithm method used is the supervised learning method. The algorithm models to be compared are Decision Tree, K-Nearest Neighbors, Random Forest, Support Vector Machine, and Logistic Regression. It is expected that by using machine learning technology, fraud can be detected early so that the level of loss and risk of sales can be minimized.

Keywords— **Fraud; Machine Learning; Supervised Learning.**

I. INTRODUCTION

Companies engaged in distribution services generally have a very large number of transactions. Some of the characteristics of distribution companies are a very large number of daily transactions, a large number of sales, a very large number of customers or consumers, namely thousands for certain areas at the district level, and also the number of items of goods which are also of course very large. Based on the large number of transactions, good company management and management of fraud detection that may occur in the company are needed.

Based on the characteristics of a large amount of transaction data and others, fraud in the sales department is very likely to occur. The large number of transactions is precisely the gap for fraudsters to commit fraud. Likewise, the large number of customers or consumers is one of the benchmarks for the number of fraudulent acts in the sales department.

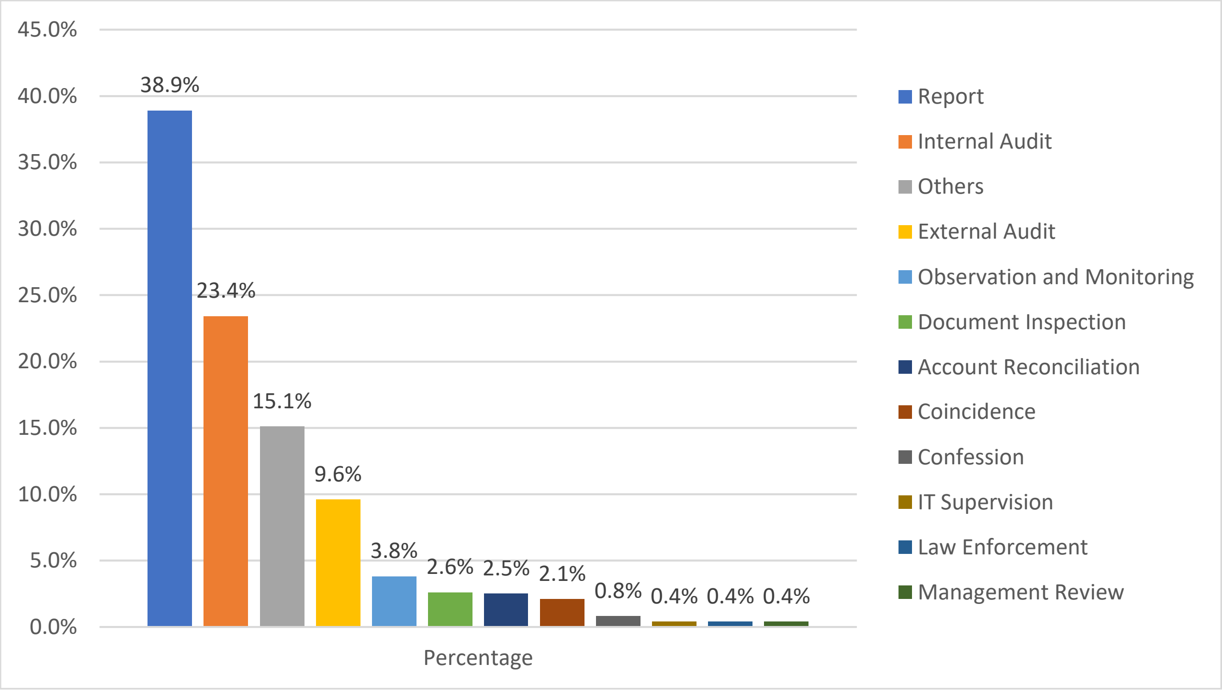


Figure 1. Fraud detection media by survey fraud Indonesia 2019

Figure 1 is derived from research conducted by the Association of Certified Fraud Examiners [1], the IT supervision media that contributed to the disclosure of fraud incidents was 0.4% or equivalent to 1 case out of a total of 250 cases. Based on the survey results, it can be concluded that the share of IT supervision is very minimal and should be further developed.

In the same survey results, IT supervision has a 100% ratio on the duration of fraud detection, which is less than 12 months. The speed of detection is because information technology tends to be able to detect in real-time. This rapid detection is one of the main factors in the importance of IT's role in disclosing sales fraud.

Based on the need for a system that is capable of conducting early detection of fraud, research was conducted to develop a tool for early detection of fraud using machine learning technology. At the time this research was conducted, no similar research with the same objects had not been found.

The sales fraud detection will use machine learning technology. This research will compare several methods. The utilization of these technologies and methods is expected to produce accurate early detection.

According to Albrecht [2] the types of fraud can be divided into 5 types. Fraud in this research is the type of fraud that causes losses to certain parties and is carried out intentionally by related parties. The types of fraud *are* divided into the following, see Table 1:

TABLE 1
TYPES OF FRAUD ACCORDING TO ALBRECHT [2]

Fraud Type	Victims	Performers	Explanation
Embezzlement by employees or employment fraud	Employer or company.	Worker or employee	Employees directly or indirectly steal from the employer or company
Fraudulent acts by management	Shareholders, and those who rely on financial reports	TOP Management	TOP Management provides false reports, usually financial reports.
Investment fraud	Investor	Individuals	Individuals defraud investors by putting money into fraudulent investments
Vendor Fraud	The organization that purchases goods or services from the vendor	Organizations that sell goods or services	Delivery of goods or performance of services is not made even though payment has been made
Customer Fraud	Organizations that sell goods or services	Customer	The customer deceives the seller by asking for something that is not his right,

According to [3], an opportunity is a situation where fraud can be possible. This opportunity can occur due to weak internal control, ineffective supervision by management, and abuse of authority. According to [4], machine learning is a branch of science that combines ideas from other branches of science, including artificial intelligence, statistics, mathematics, information technology, and others. Machine Learning is a part of artificial intelligence that has the ability to help find answers to various problems.

The algorithm of machine learning begins with the stage of collecting data, then the stage of observing the available data, then the stage of determining the components that will be used as predictors [5]. These components can be divided into several parts, namely:

- Attributes: A variable that serves as input in the prediction stage.
- Target: A desired outcome of a prediction process, usually known as a label, response, dependent variable, and outcome.

Research [6] also conducted research on fraud. The parameters used were Recency, Frequency, and Monetary. This research compares classification and regression methods.

Research [7] has the same division of label categories, namely red, yellow, and green, which will also be used in this study. The algorithm model used in this research is similar to the research to be carried out.

Research [8] uses the SVM method, but the data used is only 100 data, so the accuracy level cannot be obtained properly. Similar research on fraud detection was also conducted by [9] using the application of deep learning. By using SMOTE this research can improve accuracy and precision quite significantly.

Research [10] conducted research with retail consumer objects at a financing company, namely the object of customer receivables where there is a possibility that the customer may commit fraud by not making payments or instalment payments to the company. The study used several methods, namely KNN, Decision Tree, Random Forest, Logistic Regression, and SVM. The results of the study state that the Random Forest method has the highest test score and has an accuracy value of 0.77. The methods used are the same but the datasets and data objects are not the same.

Research [11] also used SMOTE but with the C4.5 Decision Tree algorithm. By using SMOTE, the research also managed to increase the accuracy value very significantly. Research [12] conducted a comparison using seven algorithms. The results obtained in this study are Boosted Trees has an accuracy value of 99.85%.

Based on some of the previous research above, different results were obtained for the superior algorithms in their respective studies. This can happen because the datasets used with one another are not necessarily suitable for one and the other algorithm.

This research will use the supervised learning method. The algorithm model that will be used in the comparison is Decision Tree, K-Nearest Neighbors, Random Forest, SVM, and Logistic Regression.

A Decision Tree is a machine learning algorithms model with a shape like a tree structure. The accuracy rate of the Decision Tree Method is higher if it has a large amount of data [13]. The Decision Tree algorithm was chosen based on the expert system that had been formed previously by the internal auditor. The Decision Tree method itself was chosen based on several considerations that this method is considered the most suitable for detecting fraud early on, in line with the expert system.

K- Nearest Neighbor according to [14] is a model of the simplest classification algorithm. Images are classified into labels by this method. Classification in this method is based on the closest distance to other objects.

According to research [15], Random Forest is an adaptation of the Decision Tree method, where each branch is developed by bootstrapping examples using basic *training* data. This Algorithm model is a collection of Decision Tree models which then operate so that the collection is functional.

Support Vector Machine method according to [16] has a high accuracy value in predicting classification. The advantage of SVM is that it can classify a pattern even though it has limited datasets. However, there are limitations to the SVM method, namely when the number of attributes has large number so that it can result in the burden of the device in performing computational calculations becoming heavier so that the results obtained become less accurate.

According to research [17] logistic regression is a data analysis technique that has the aim of knowing the relationship of variables that have categorical properties on the response variable.

This research is conducted so that sales fraud can be detected early on. Preventive action for losses in distribution companies can be done by early detection of sales fraud with the help of machine learning technology. A comparison of various algorithm models is carried out to determine the most accurate model for detecting sales fraud in a distribution company.

II. METHODS

The following is a flow image of the research stages. Starting from the beginning of the research until the final stages:

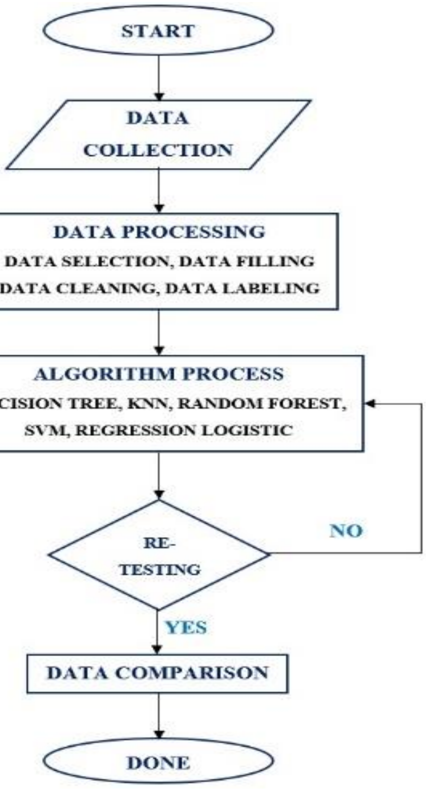


Figure 2. Research stages

Figure 2 explains the stages of the research starting with data collection and then performing the Data Processing stage. After that, the test and training stages are carried out using the specified algorithm model. The algorithm models used are Decision Tree, KNN, Random Forest, SVM, and Logistic Regression. Then a check will be made of the results of the algorithm process if it is deemed necessary to retest the process. Then the stages will repeat the algorithm process until the results obtained are deemed sufficient, then it will continue to the data comparison stage. After the comparison stage, all stages have been completed.

A. *Data Collection*

In this initial stage, data collection is carried out from the company's database. The data taken are two data tables: sales data and payment data. Data is taken by querying the company's database. Each table has hundreds of columns.

The dataset has 18 columns to which 1 label column is added. The dataset has 4427 rows of data. The dataset is taken from the sales data of a metal steel roof distribution company in 2022.

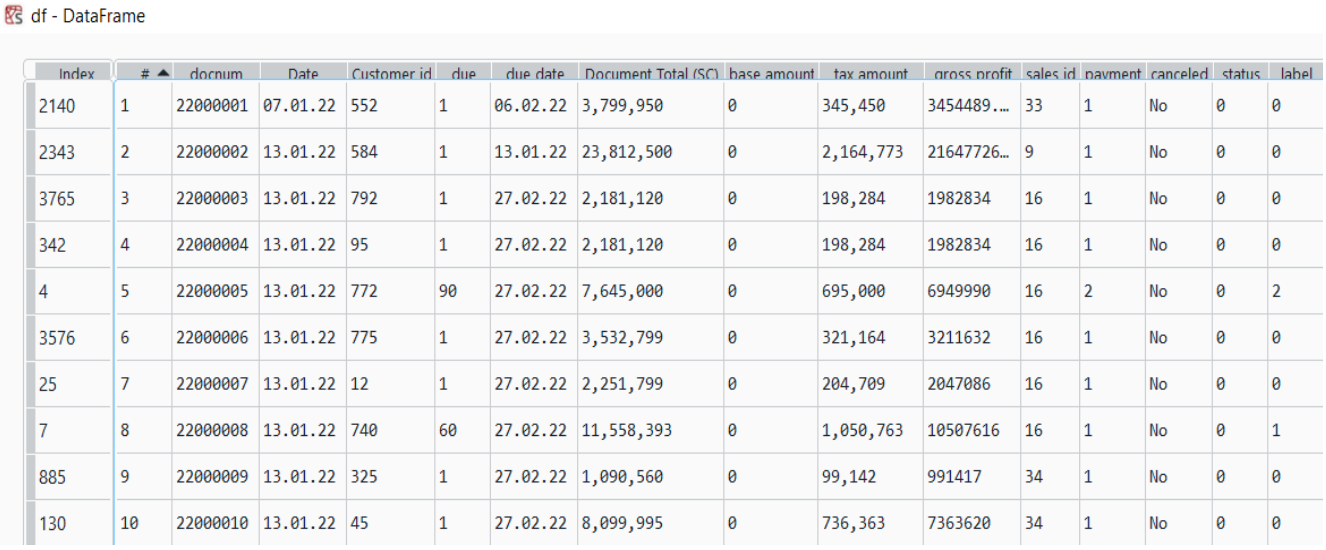
B. *Data Processing*

At this stage, data processing is carried out. This stage is carried out so that the data can be processed using machine learning technology. This stage consists of Data Selection, Data Filling, Data Cleaning, and Data Labeling.

1) *Data Selection*: This process is done by sorting data columns that will be useful in this research. Some of the columns include the payment duration column (Due), and the instalment amount (Payment). Other data selected are customer code, payment amount, and receivable status.

2) Data Filling: Process to complete missing data. The process is done manually. This process also ensures that all data has been filled in so that the data is ready to be processed.

3) *Data Cleaning:* This process is done manually in the dataset. It is done by deleting columns that are not used. Out of 18 columns in the dataset, 3 columns were used as the main parameters to detect fraud.



df - DataFrame

Index	# ▲	docnum	Date	Customer id	due	due date	Document Total (SC)	base amount	tax amount	gross profit	sales id	payment	canceled	status	label
2140	1	22000001	07.01.22	552	1	06.02.22	3,799,950	0	345,450	3454489.…	33	1	No	0	0
2343	2	22000002	13.01.22	584	1	13.01.22	23,812,500	0	2,164,773	21647726…	9	1	No	0	0
3765	3	22000003	13.01.22	792	1	27.02.22	2,181,120	0	198,284	1982834	16	1	No	0	0
342	4	22000004	13.01.22	95	1	27.02.22	2,181,120	0	198,284	1982834	16	1	No	0	0
4	5	22000005	13.01.22	772	90	27.02.22	7,645,000	0	695,000	6949990	16	2	No	0	2
3576	6	22000006	13.01.22	775	1	27.02.22	3,532,799	0	321,164	3211632	16	1	No	0	0
25	7	22000007	13.01.22	12	1	27.02.22	2,251,799	0	204,709	2047086	16	1	No	0	0
7	8	22000008	13.01.22	740	60	27.02.22	11,558,393	0	1,050,763	10507616	16	1	No	0	1
885	9	22000009	13.01.22	325	1	27.02.22	1,090,560	0	99,142	991417	34	1	No	0	0
130	10	22000010	13.01.22	45	1	27.02.22	8,099,995	0	736,363	7363620	34	1	No	0	0

Figure 3. Data research after data cleaning stage

Figure 3 contains research data after the data-cleaning process is carried out. From a total of 18 columns and 1 label column, 4 columns were removed. The removed column is a column related to the identity of the customer whose confidentiality is maintained.

4) *Data Labeling:* The stage of labeling the data is done manually. This labeling is also based on confirmation to the company based on the history of checking by the internal auditor. The label itself is divided into 3 categories:

- The Green category is a safe category, where the data entered in this category will not be checked by the internal auditor. However, the same data, which is included in this green category, can become another category in the future. If this is found, an additional learning process will be carried out on machine learning.
- The Yellow category is the alert category. In this category, further checking by the internal auditor is recommended. Data in this category can become a red or green category through a manual checking process and label updating process.
- The Red category is a category where the level of suspicion of fraud is very high. Data entered in this category must be quickly cross-checked with the actual conditions in the field.

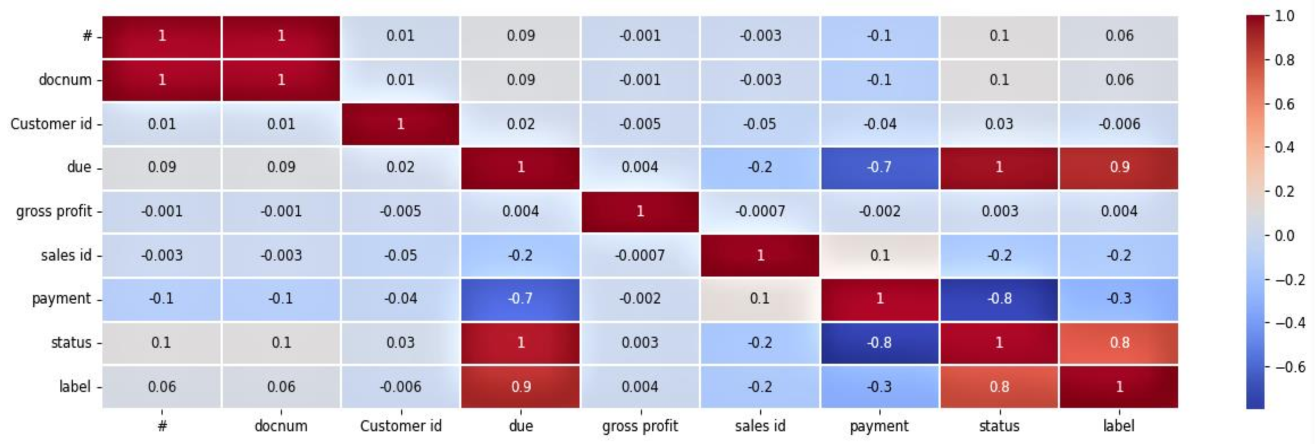


Figure 4. Correlation matrix

In the Correlation Matrix Figure 4, it is clear that the due and status columns correlate with a value of 1. Meanwhile, the Label column has a relationship with due worth 0.9 and status worth 0.8. Payment and status also have an inverse correlation of -0.8. While the due and payment columns correlate -0.7.

The correlation between the label column and the due column and the status column is quite highly correlated. This is because due and status are the basis for labeling, as well as the payment column. The important parameters are due, payment, and status which will be the main parameters in this research.

C. *Algorithm Process*

At this stage, the Train and Test process is carried out using the predetermined algorithm model. The Train and Test process is using 3 kinds of ratios. The ratio is 80:20, ratio 70:30, and ratio 60:40.

The machine learning method used in this research is supervised learning, namely classification and regression. The dataset is processed in such a way through data processing so that the dataset can be processed and produce the desired output. One of the processes carried out is labeling data categories. The labels given are red, yellow, and green labels.

The parameters used are payment data (due column), instalment data (payment column), and remaining receivables data (remaining column). Table 2 is how to label based on the parameters that have been determined:

TABLE 2
PARAMETER INSTALMENT PAYMENT AND DUE

	Due 1	**Due 15**	**Due 30**	**Due 45**	**Due 60**	**Due 90**
Payment 0						Yellow
Payment 1	Green	Green	Green	Yellow	Yellow	Yellow
Payment 2		Green	Green	Yellow	Red	Red
Payment 3		Yellow	Yellow	Red	Red	Red

At this stage, data processing is carried out. This stage is carried out so that the data can be processed using machine learning technology. This stage consists of Data Selection, Data Filling, Data Cleaning, and Data Labeling.

Table 2 is a label table based on the payment and due columns. The instalment payment column is data that contains the number of times the amount of payment for one receivable. The Due column is the number of days from the date of the receivable until it is paid.

Explanation of instalment payment column:

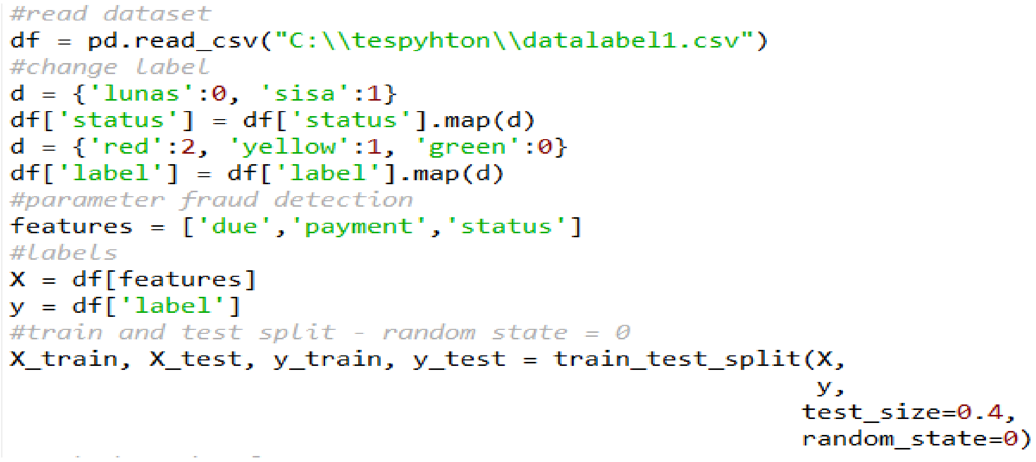
- Payment 0 (zero): unpaid receivables
- Payment 1 (one): receivables are paid once directly in full
- Payment 2 (two): receivables are paid twice until they are paid off.
- Payment 3 (three): receivables are paid three times until paid off

Explanation of the Due column:

- Due 1: receivables paid within 1 day
- Due 15: receivables payable within 15 days

- Due 30: receivables payable within 30 days
- Due 45: receivables payable within 45 days
- Due 60: receivables payable within 60 days
- Due 90: receivables still unpaid past 90 days

The stages of processing raw data into labeled data are done with the help of verification by a team of internal auditors. The data is taken from the company dataset from the sales year of 2022. From hundreds of columns, 18 columns are selected which can help in the process of early detection of fraud. The total row data is 4.427 data. The total green label is 4.210 data. The total yellow label is 121 data. The total red label is 19 data.



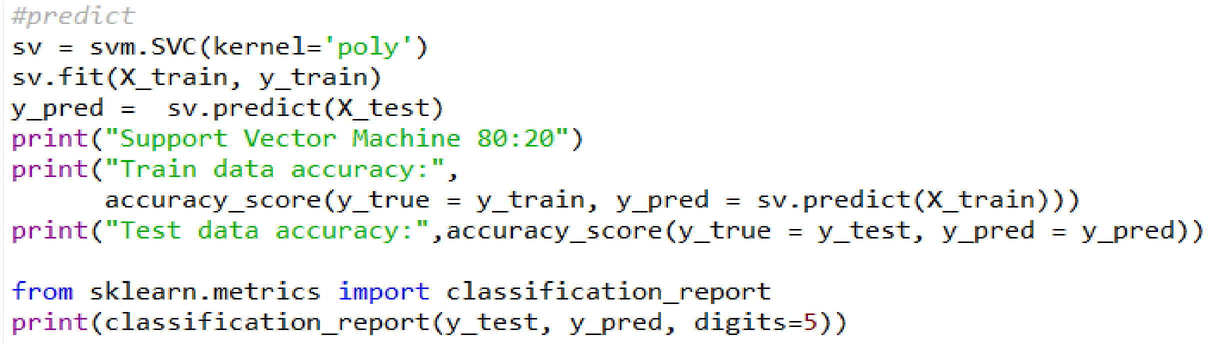
```
#read dataset
df = pd.read_csv("C:\\tespyhton\\datalabel1.csv")
#change label
d = {'lunas':0, 'sisa':1}
df['status'] = df['status'].map(d)
d = {'red':2, 'yellow':1, 'green':0}
df['label'] = df['label'].map(d)
#parameter fraud detection
features = ['due','payment','status']
#labels
X = df[features]
y = df['label']
#train and test split - random state = 0
X_train, X_test, y_train, y_test = train_test_split(X,
                                                      y,
                                                     test_size=0.4,
                                                     random_state=0)
```

Figure 5. Step algorithm model

In Figure 5, we can see the process of reading the dataset, then proceed with the process of changing labels to numbers. Then there is the process of selecting parameters used to detect fraud through features. Next is the process of defining x and y, which are features and label parameters. Then there is the process of split training and testing.

The Supervised learning model using pandas as an open-source library. Sklearn as a Python module integrating machine learning algorithms is used to help classify the algorithm model. All the supervised learning in this research is using sklearn to help with the algorithm model.

Model Decision Tree is predicted using Decision Tree Classifier from the sklearn tree. K-Nearest Neighbors scenario is using K Neighbors Classifier from sklearn neighbors. The n value with this model is using n-neighbors = 2. The Random Forest algorithm model scenario is using Random Forest Classifier from the sklearn ensemble. Logistic Regression is using Logistic Regression with solver=’lbfgs’ from the sklearn linear model.



```
#predict
sv = svm.SVC(kernel='poly')
sv.fit(X_train, y_train)
y_pred =  sv.predict(X_test)
print("Support Vector Machine 80:20")
print("Train data accuracy:",
      accuracy_score(y_true = y_train, y_pred = sv.predict(X_train)))
print("Test data accuracy:",accuracy_score(y_true = y_test, y_pred = y_pred))

from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred, digits=5))
```

Figure 6. SVM prediction

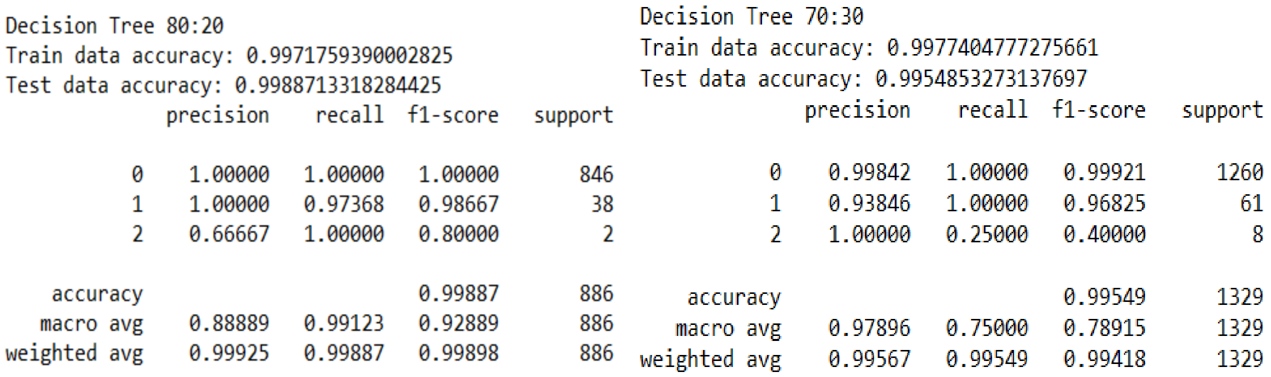
Figure 6 is Python programming for support vector machine prediction. Support vector machine is using SVM SVC with the poly kernel. SVC is Support Vector Classifier, the implementation is based on the library SVM.

D. *Data Comparison*

The data comparison stage is carried out to find out which algorithm model has the highest accuracy value. This stage is the final stage in this research. The data comparison stage in this research compares supervised learning methods, namely the Decision Tree, KNN, Random Forest, SVM, and Logistic Regression algorithm models.

III. RESULT AND DISCUSSION

A. *Result*

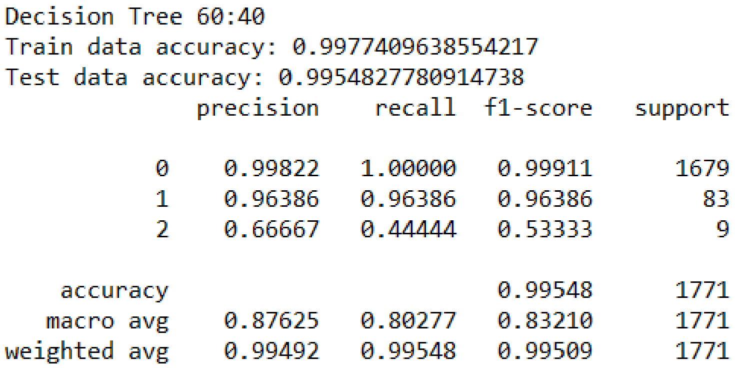


```
Decision Tree 80:20
Train data accuracy: 0.9971759390002825
Test data accuracy: 0.9988713318284425
              precision    recall  f1-score   support

           0    1.00000   1.00000   1.00000       846
           1    1.00000   0.97368   0.98667        38
           2    0.66667   1.00000   0.80000         2

    accuracy                        0.99887       886
   macro avg    0.88889   0.99123   0.92889       886
weighted avg    0.99925   0.99887   0.99898       886
```

```
Decision Tree 70:30
Train data accuracy: 0.9977404777275661
Test data accuracy: 0.9954853273137697
              precision    recall  f1-score   support

           0    0.99842   1.00000   0.99921      1260
           1    0.93846   1.00000   0.96825        61
           2    1.00000   0.25000   0.40000         8

    accuracy                        0.99549      1329
   macro avg    0.97896   0.75000   0.78915      1329
weighted avg    0.99567   0.99549   0.99418      1329
```

Figure 7. Decision tree trial and test accuracy score using ratio 80: 20 and ratio 70:30

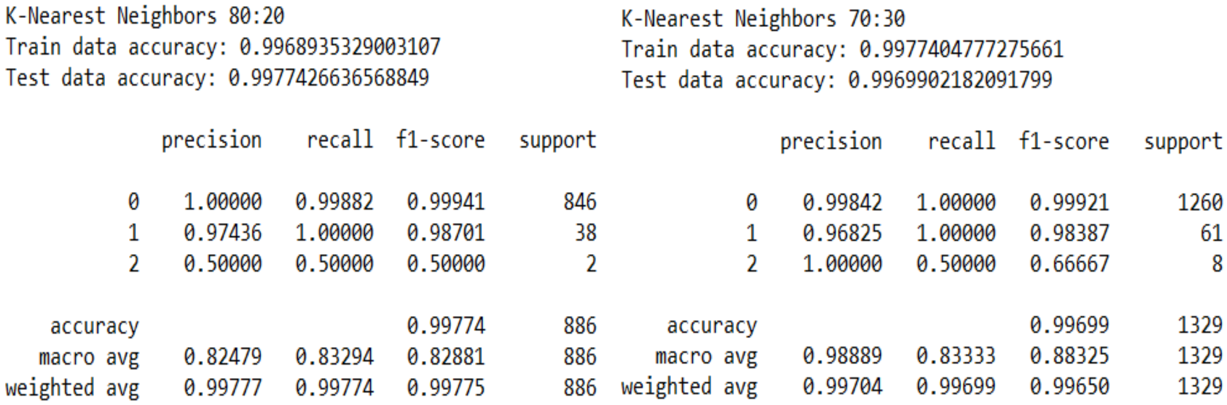
In Figure 7, a test was conducted using the Decision Tree algorithm model. This test uses a ratio of 80:20. The accuracy data test result is 0.9988. Model algorithm Decision Tree ratio 70:30 test data accuracy is 0.995485.



```
Decision Tree 60:40
Train data accuracy: 0.9977409638554217
Test data accuracy: 0.9954827780914738
              precision    recall  f1-score   support

           0    0.99822   1.00000   0.99911      1679
           1    0.96386   0.96386   0.96386        83
           2    0.66667   0.44444   0.53333         9

    accuracy                        0.99548      1771
   macro avg    0.87625   0.80277   0.83210      1771
weighted avg    0.99492   0.99548   0.99509      1771
```

Figure 8. Decision tree trial and test accuracy score using a ratio of 60:40

In Figure 8, a test was conducted using the Decision Tree algorithm model. This test uses a ratio of 60:40. The accuracy data test result is 0.995482. Decision Tree model algorithm test and trial using ratios 80:20, 70:30, and 60:40. The highest accuracy is 0.9988 using ratio 80:20.

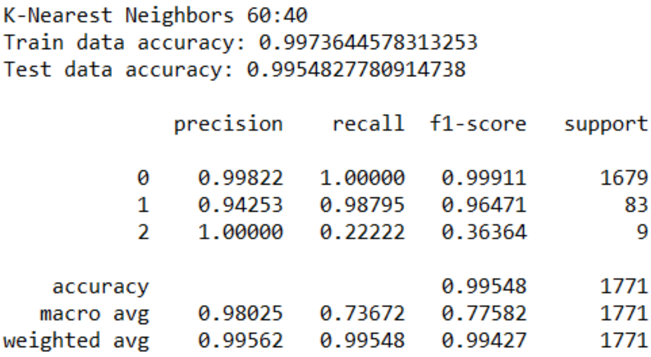


```
K-Nearest Neighbors 80:20
Train data accuracy: 0.9968935329003107
Test data accuracy: 0.9977426636568849

              precision    recall  f1-score   support

           0    1.00000   0.99882   0.99941       846
           1    0.97436   1.00000   0.98701        38
           2    0.50000   0.50000   0.50000         2

    accuracy                        0.99774       886
   macro avg    0.82479   0.83294   0.82881       886
weighted avg    0.99777   0.99774   0.99775       886
```

```
K-Nearest Neighbors 70:30
Train data accuracy: 0.9977404777275661
Test data accuracy: 0.9969902182091799

              precision    recall  f1-score   support

           0    0.99842   1.00000   0.99921      1260
           1    0.96825   1.00000   0.98387        61
           2    1.00000   0.50000   0.66667         8

    accuracy                        0.99699      1329
   macro avg    0.98889   0.83333   0.88325      1329
weighted avg    0.99704   0.99699   0.99650      1329
```

Figure 9. K-Nearest Neighbors trial and test accuracy score using ratio 80:20 and 70:30

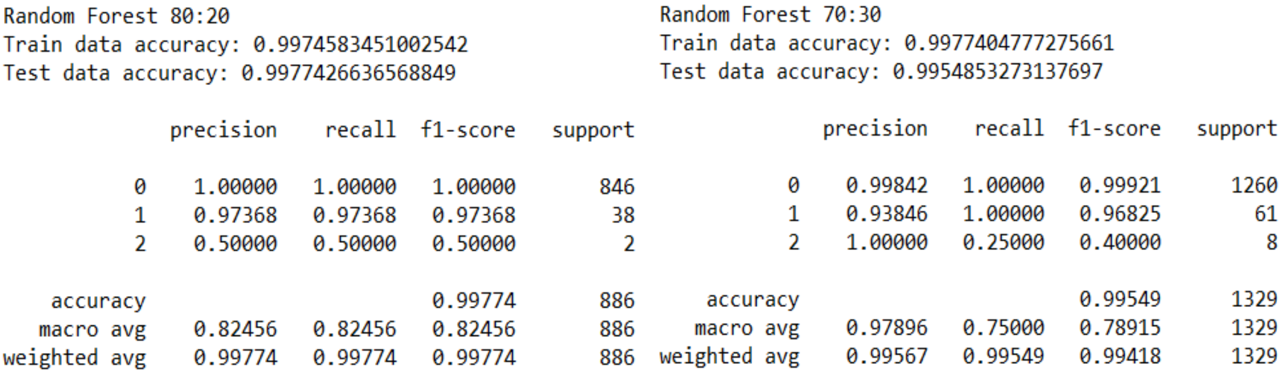
In Figure 9, a test was conducted using the K-Nearest Neighbors algorithm model. This test uses a ratio of 80:20. The accuracy data test result is 0.997742. Model algorithm K-Nearest Neighbors ratio 70:30 test data accuracy is 0.996990.



```
K-Nearest Neighbors 60:40
Train data accuracy: 0.9973644578313253
Test data accuracy: 0.9954827780914738

              precision    recall  f1-score   support

           0    0.99822   1.00000   0.99911      1679
           1    0.94253   0.98795   0.96471        83
           2    1.00000   0.22222   0.36364         9

    accuracy                        0.99548      1771
   macro avg    0.98025   0.73672   0.77582      1771
weighted avg    0.99562   0.99548   0.99427      1771
```

Figure 10. K-Nearest Neighbors test and trial accuracy using a ratio of 60:40

In Figure 10, Model K-Nearest Neighbors using n=2 and the highest accuracy score is 0.99548 by using a ratio of 60:40.

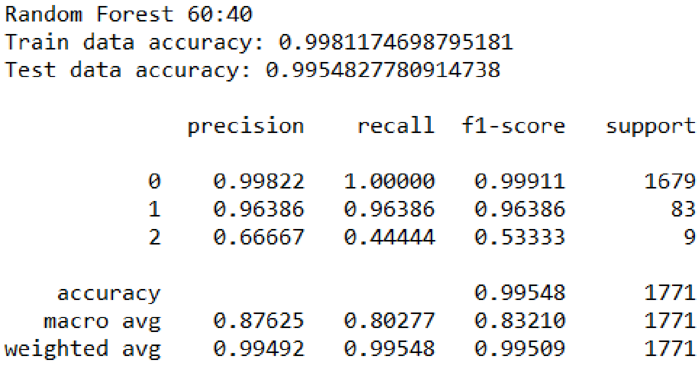


```
Random Forest 80:20
Train data accuracy: 0.9974583451002542
Test data accuracy: 0.9977426636568849

              precision    recall  f1-score   support

           0    1.00000   1.00000   1.00000       846
           1    0.97368   0.97368   0.97368        38
           2    0.50000   0.50000   0.50000         2

    accuracy                        0.99774       886
   macro avg    0.82456   0.82456   0.82456       886
weighted avg    0.99774   0.99774   0.99774       886
```

```
Random Forest 70:30
Train data accuracy: 0.9977404777275661
Test data accuracy: 0.9954853273137697

              precision    recall  f1-score   support

           0    0.99842   1.00000   0.99921      1260
           1    0.93846   1.00000   0.96825        61
           2    1.00000   0.25000   0.40000         8

    accuracy                        0.99549      1329
   macro avg    0.97896   0.75000   0.78915      1329
weighted avg    0.99567   0.99549   0.99418      1329
```

Figure 11. Random forest trial and test accuracy score using ratio 80:20 and ratio 70:30

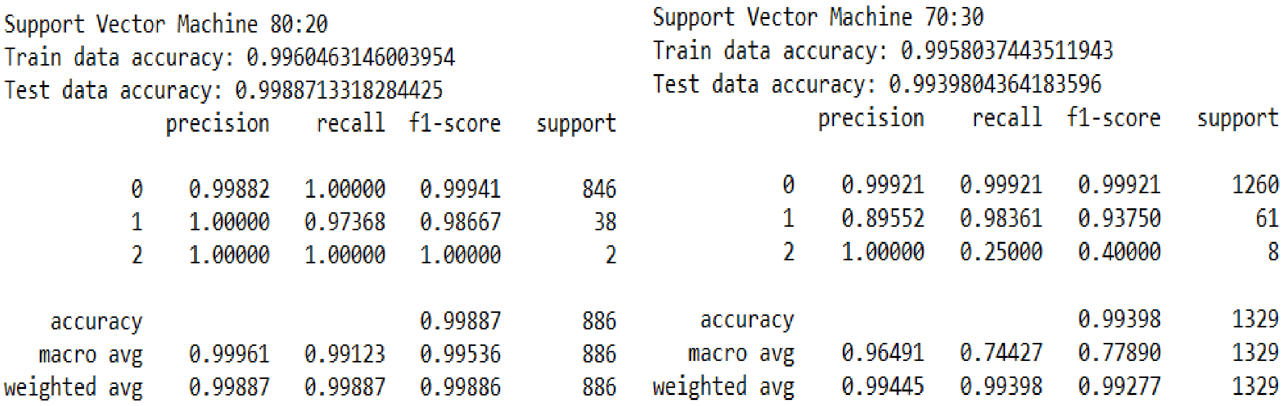
In Figure 11, a test was conducted using the Random Forest algorithm model. This test uses a ratio of 80:20. The accuracy data test result is 0.997742. Random Forest model algorithm test and trial using a ratio 70:30 accuracy score is 0.995458.



```
Random Forest 60:40
Train data accuracy: 0.9981174698795181
Test data accuracy: 0.9954827780914738

              precision    recall  f1-score   support

           0    0.99822   1.00000   0.99911      1679
           1    0.96386   0.96386   0.96386        83
           2    0.66667   0.44444   0.53333         9

    accuracy                        0.99548      1771
   macro avg    0.87625   0.80277   0.83210      1771
weighted avg    0.99492   0.99548   0.99509      1771
```

Figure 12. Random forest trial and test accuracy score using ratio 60:40

In Figure 12, a test was conducted using the Random Forest algorithm model. This test uses a ratio of 60:40. The accuracy data test result is 0.995482. Random Forest all ratio 80:20, 70:30, and 60:40. The highest accuracy is 0.997742 using ratio 80:20.

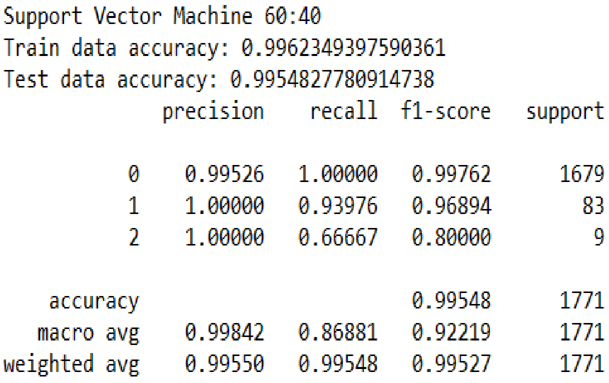


```
Support Vector Machine 80:20
Train data accuracy: 0.9960463146003954
Test data accuracy: 0.9988713318284425
              precision    recall  f1-score   support

           0    0.99882   1.00000   0.99941       846
           1    1.00000   0.97368   0.98667        38
           2    1.00000   1.00000   1.00000         2

    accuracy                        0.99887       886
   macro avg    0.99961   0.99123   0.99536       886
weighted avg    0.99887   0.99887   0.99886       886
```

```
Support Vector Machine 70:30
Train data accuracy: 0.9958037443511943
Test data accuracy: 0.9939804364183596
              precision    recall  f1-score   support

           0    0.99921   0.99921   0.99921      1260
           1    0.89552   0.98361   0.93750        61
           2    1.00000   0.25000   0.40000         8

    accuracy                        0.99398      1329
   macro avg    0.96491   0.74427   0.77890      1329
weighted avg    0.99445   0.99398   0.99277      1329
```

Figure 13. SVM trial and test accuracy score using ratio 80:20 and ratio 70:30

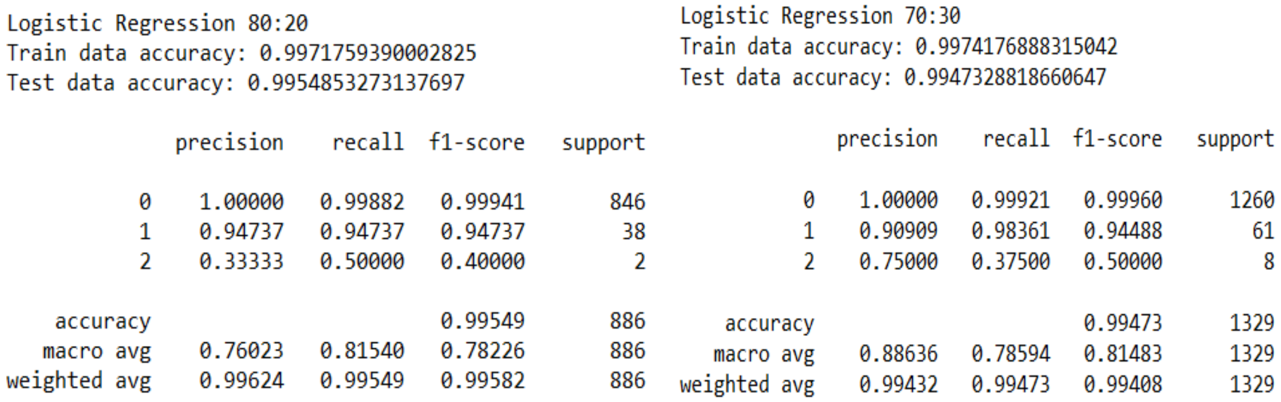
In Figure 13, a test was conducted using the Support Vector Machine algorithm model. This test uses a ratio of 80:20. The accuracy data test result is 0.998871. Random Forest model algorithm test and trial using a ratio 70:30 accuracy score is 0.99398.



```
Support Vector Machine 60:40
Train data accuracy: 0.9962349397590361
Test data accuracy: 0.9954827780914738
              precision    recall  f1-score   support

           0    0.99526   1.00000   0.99762      1679
           1    1.00000   0.93976   0.96894        83
           2    1.00000   0.66667   0.80000         9

    accuracy                        0.99548      1771
   macro avg    0.99842   0.86881   0.92219      1771
weighted avg    0.99550   0.99548   0.99527      1771
```

Figure 14. SVM trial and test accuracy score using a ratio of 60:40

In Figure 14, a test was conducted using the Support Vector Machine algorithm model. This test uses a ratio of 60:40. The accuracy data test result is 0.995482. Random Forest all ratio 80:20, 70:30, and 60:40. The highest accuracy is 0.998871 using ratio 80:20.

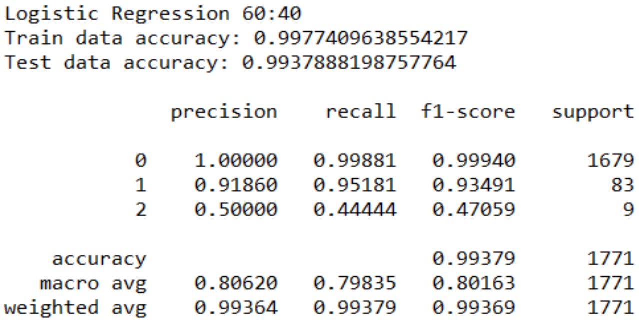


```
Logistic Regression 80:20
Train data accuracy: 0.9971759390002825
Test data accuracy: 0.9954853273137697

              precision    recall  f1-score   support

           0    1.00000   0.99882   0.99941       846
           1    0.94737   0.94737   0.94737        38
           2    0.33333   0.50000   0.40000         2

    accuracy                        0.99549       886
   macro avg    0.76023   0.81540   0.78226       886
weighted avg    0.99624   0.99549   0.99582       886
```

```
Logistic Regression 70:30
Train data accuracy: 0.9974176888315042
Test data accuracy: 0.9947328818660647

              precision    recall  f1-score   support

           0    1.00000   0.99921   0.99960      1260
           1    0.90909   0.98361   0.94488        61
           2    0.75000   0.37500   0.50000         8

    accuracy                        0.99473      1329
   macro avg    0.88636   0.78594   0.81483      1329
weighted avg    0.99432   0.99473   0.99408      1329
```

Figure 15. Logistic regression trial and test accuracy score using ratio 80:20 and ratio 70:30

In Figure 15, a test was conducted using the Logistic Regression algorithm model. This test uses a ratio of 80:20. The accuracy data test result is 0.995485. Logistic Regression model algorithm test and trial using a ratio 70:30 accuracy score is 0.994732.



```
Logistic Regression 60:40
Train data accuracy: 0.9977409638554217
Test data accuracy: 0.9937888198757764

              precision    recall  f1-score   support

           0    1.00000   0.99881   0.99940      1679
           1    0.91860   0.95181   0.93491        83
           2    0.50000   0.44444   0.47059         9

    accuracy                        0.99379      1771
   macro avg    0.80620   0.79835   0.80163      1771
weighted avg    0.99364   0.99379   0.99369      1771
```

Figure 16. Logistic regression trial and test accuracy score using ratio 60:40

In Figure 16, a test was conducted using the Logistic Regression algorithm model. This test uses a ratio of 60:40. The accuracy is 0.993788. Logistic Regression all ratios 80:20, 70:30, and 60:40. The highest accuracy is 0.995485 using ratio 80:20.

B. Comparison

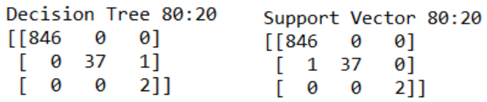
After going through various series of trial processes from various algorithm models, the results of the data are obtained to be compared. The data comparison becomes a reference to determine the algorithm model that has the highest accuracy and the most suitable algorithm model for early detection of fraud in the sales department, see Table 3.

TABLE 3
TEST DATA ACCURACY TABLE COMPARISON

	80:20	70:30	60:40
Decision Tree	0.998871	0.995485	0.995482
K-Nearest Neighbors	0.997742	0.996990	0.995482
Random Forest	0.997742	0.995485	0.995482
Support Vector Machine	0.998871	0.993980	0.995482
Logistic Regression	0.995485	0.994732	0.993788

C. *Discussion*

Based on the result of the data comparison in Table 3, It can be seen that the results only have a slight difference below 0,1%. This is because the labeling process has gone through an identification process by the internal auditor. The parameters used are also closely related to the labeling process.



```
Decision Tree 80:20        Support Vector 80:20
[[846   0   0]             [[846   0   0]
 [  0  37   1]              [  1  37   0]
 [  0   0   2]]             [  0   0   2]]
```

Figure 17. Confusion matrix decision tree and SVM

In Figure 17, The Decision Tree algorithm model with a ratio of 80:20 and the Support Vector Machine model with a poly kernel ratio of 80:20 have the highest accuracy value results. The Decision Tree model predicts 1 False Positive on the yellow label, this can be assumed to be better than the prediction results of the Support Vector Machine, which is 1 False Negative on the green label. Based on the assumption that the label is better to be categorized as yellow fraud so that it will be checked then to be labeled as green and not checked.

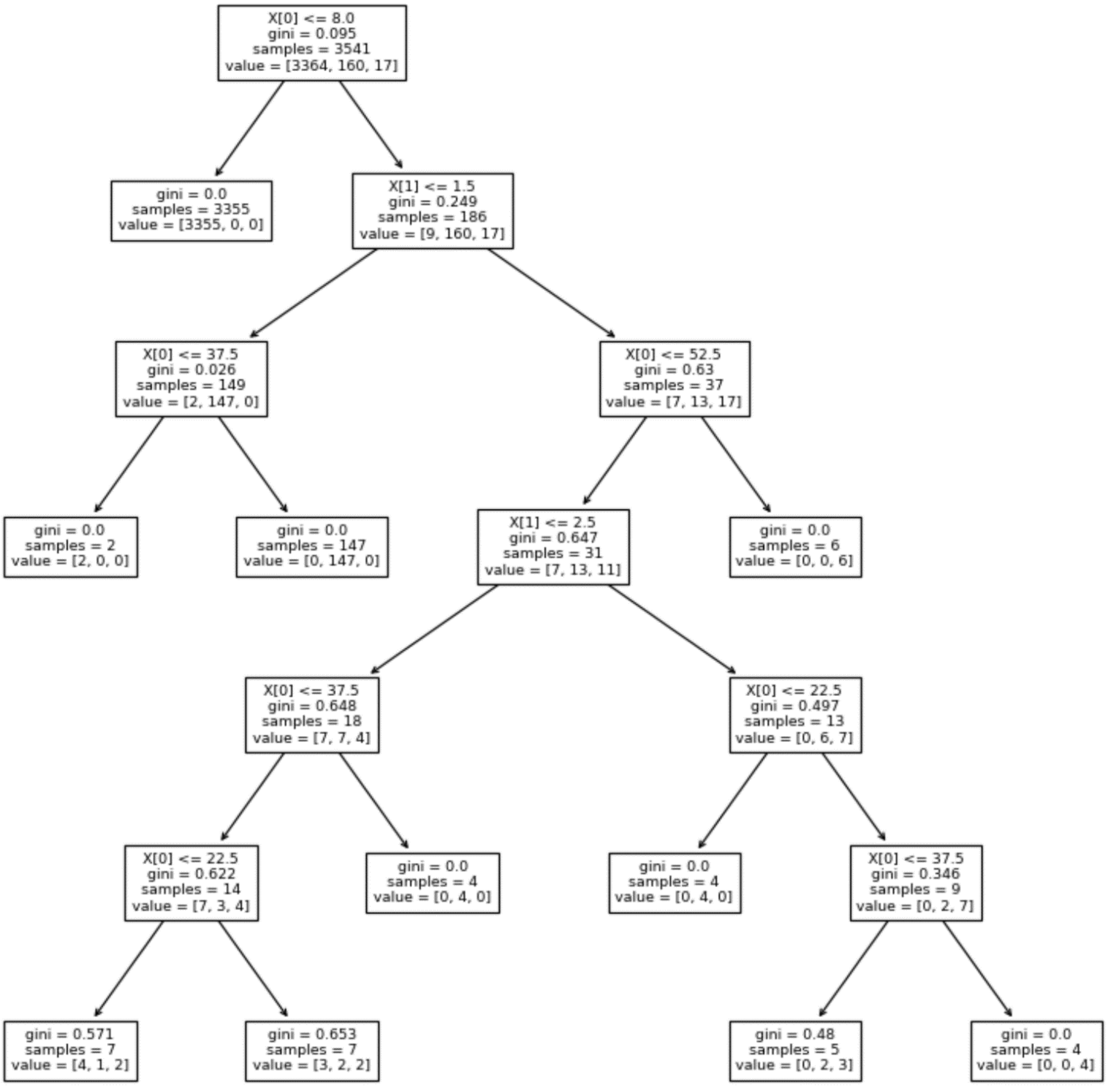


Figure 18. Model decision tree using ratio 80:20

Figure 18 shows the Decision Tree model using the 80:20 ratio with 7 levels. 80% of the training data, which is a total of 3541 data from data total of 4427 data. It reads that the green label is 3364 data then the yellow label is 160 data and 17 red labels data (first level). And then continued to the 2nd level from 3541 data divided into 3355 accurate green labels with a gini value of 0.0. And with the gini = 0,249, it reads 9 green labels, 160 yellow, and 17 red label data.

Continued in the 4th level branch can be seen fix data with gini = 0.0, 2 green labels, 147 yellows labels, and 6 red labels. In the 6th level branch with gini = 0.0, there is a total of 8 yellows labels coming from 2 separate 5th level branches. In the 7th level branch with a gini value of 0.0, there are 4 red labels.

IV. CONCLUSIONS

Machine learning technology can be used as a tool in the early detection of sales fraud in distributor companies. The time needed to detect fraud can be shortened by using machine learning technology, namely a pre-pared algorithm. Decision Tree is an algorithm with the highest accuracy value of 0.9988 in this research.

Further research can be developed by comparing other algorithms. The research will be continued by creating an application that is easy to use so that the operation of early detection of fraud can be done by the general user without involving the development team. This application will be used in the company. In addition, the application will be integrated via API for general use. Furthermore, the API will be developed so that the application can be used on many types of databases. Further research will also be carried out using a larger database so that results can be more varied.

The time required for machine learning to detect fraud is very fast when compared to user experts (internal auditors), provided that the dataset and programming code for the algorithm model have been prepared beforehand. This is proven by the speed required by machine learning to detect fraud in just seconds, compared to the detection time by user experts of more than one hour for 200-300 data.

ACKNOWLEDGMENTS

Gratitude is extended to Maranatha Christian University. Gratitude is also expressed to those who have helped so that this research can be completed properly. Praise be to the Almighty God.

REFERENCES

[1] ACFE, “Survei Fraud Indonesia 2019,” *Acfe Indonesia Chapter,* 2019.

[2] Albrecht.W.Steve and C. O. Albrecht, Fraud Examination, Western: Thomson South, 2002.

[3] Wahyuni and S. B. Gideon, “Fraud Traingle Sebagai Pendeteksi Kecurangan Laporan Keuangan,” *Jurnal Akuntansi,* vol. XXI, pp. 47-61, 2017.

[4] J. Andreanus and A. Kurniawan, “Sejarah, Teori Dasar dan Penerapan Reinforcement Learning: Sebuah Tinjauan Pustaka,” *Jurnal Telematika,* vol. 12, no. 2, pp. 113-118, 2017.

[5] M. Bowles, Machine Learning in Python: Essential Techniques for Predictive Analysis, Wiley, 2015.

[6] J. Badvelu, *Comparative Study of Different Machine Learning Models for Sales Prediction and Fraud Detection,* GitHub, 2020.

[7] J. Narabel and S. Budi, “Deteksi Dini Status Keanggotaan Industri Kebugaran Menggunakan Pendekatan Supervised Learning,” *JuTISI,* vol. 6, no. 2, 2020.

[8] Y. Yazid and A. Fiananta, “Mendeteksi Kecurangan Pada Transaksi Kartu Kredit Untuk Verifikasi Transaksi Menggunakan Metode SVM,” *IJAI (Indonesian Journal of Applied Informatics),* vol. 1, no. 2, pp. 61-66, 2017.

[9] F. Zamachsari and N. Puspitasari, “Penerapan Deep Learningdalam Deteksi Penipuan Transaksi Keuangan Secara Elektronik,” *J. RESTI (Rekayasa Sist. Teknol. Inf.),* vol. 5, no. 2, pp. 203-212, 2021.

[10] N. I. Mustika, B. Nenda and D. Ramadhan, “Machine Learning Algorithms in Fraud Detection: Case Study on Retail Consumer Financing Company,” *Asia Pacific Fraud Journal,* vol. 6, no. 2, 2021.

[11] L. D. Perwara, F. A. Bachtiar and I. Indriati, “Penerapan Algoritma Decision Tree C4.5 Untuk Deteksi Fraud Pada Kartu Kredit dengan Oversampling Synthetic Minority Technique (SMOTE),” *J-PTIIK,* vol. 4, no. 8, pp. 2664-2669, 2020.

[12] H. Sunata, “Komparasi Tujuh Algoritma Identifikasi Fraud ATM Pada PT. Bank Central Asia Tbk,” *Jurnal Teknik Informatika dan Sistem Informasi,* vol. 7, no. 3, pp. 441-450, 2020.

[13] C. N. Dengen, Kusrini and E. T. Luthfi, “Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu,” *Jurnal Ilmiah SISFOTENIKA,* vol. 10, no. 1, 2020.

[14] L. Farokhah, “Implementasi K-Nearest Neighbor untuk Klasifikasi Bunga Dengan Ekstraksi Fitur Warna RGB,” *JTIIK,* vol. 7, no. 6, 2020.

[15] E. Renata and M. Ayub, “Penerapan Metode Random Forestuntuk Analisis Risiko pada dataset Peer to peer lending,” *JuTISI,* vol. 6, no. 3, 2020.

[16] T. B. Sasongko, “Komparasi dan Analisis Kinerja Model Algoritma SVM dan PSO-SVM (Studi Kasus Klasifikasi Jalur Minat SMA),” *JuTISI,* vol. 2, no. 2, 2016.

[17] E. D. Anggara, A. Widjaja and B. R. Suteja, “Prediksi Kinerja sebagai Rekomendasi Kenaikan Golongan dengan Decision Tree dan Regresi Logistik,” *JuTISI,* vol. 8, no. 1, 2022.