Sana'a University Journal of Applied Sciences and Technology https://journals.su.edu.ye/index.php/jast
Vol. 1 | No. 4 | Page 304 – 312 | 2023 |
ISSN: 2958-9568
Doi: 10.59628/jast.v1i4.528

A Comparative Study of the Performance of Machine Learning Models on a Tax Dataset of Yemen to Detect Levels of Tax Evasion

Abeer Abdullah Shujaaddeen[1, *], Fadl M.M. Ba-Alwi[1]
[1] Department of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen.
***Corresponding author:** *abeer41036@gmail.com*

ARTICLE INFO
Article history:
Received: July 20, 2023
Accepted: Oct 27, 2023
Published: Nov, 2023

KEYWORDS
1. ML techniques
2. K-Fold Validation
3. Dataset of Tax
4. Performance measures.

ABSTRACT

The performance of a classification model in machine learning is affected by many factors, such as the type of machine learning technology used. Accuracy varies from method to method. This paper presents a comparison between the performance of different models in terms of the machine learning technique used (e.g. KNN, NB, SVM, DT, RF, MLP). Based on the data provided by the Tax Authority of Yemen, which is related to the commercial and industrial profits tax, which consists of 760 attributes, after the preprocessing of data. The dataset partition technique used k-fold validation. The paper shows that the e Naïve Bayes (NB) classifier gave the highest result in accuracy and other measures. Then KNN, SVM, and RF gave the same results in accuracy 99.87%, but in SVM, KNN the results were also the same in the rest measures, while in RF models the rest measures were 97.91%,99.95%, and 98.91% in Recall, Precision and F-score in order. MLP gave 98.42 in accuracy with 66.62%, 64.21%, and 64.40 in the recall, Precision, and F-score, then DT gave 97.76% in accuracy with 57.006% ,99.24% and 72.41% in the recall, precision, and F-score.

CONTENTS

1. Introduction
2. Related Work
3. Methodology
4. Splitting data by K-Fold Validation technique
5. The algorithms that used for experiments
6. Model Evaluation
7. Experiments
8. Discussion the Results
9. Conclusion and Future work
10. References

1. **Introduction:**

Taxes are considered one of the most important revenues for developed and undeveloped countries alike, because of their importance in raising the level of the country. Taxes are an amount that the state imposes on companies and individuals.

However, many taxpayers try to evade tax by not paying their taxes in several ways, such as lying on the declaration form, hiding part of the data for tax fraud, and other ways and methods [1]. Therefore, many countries have implemented many procedures and regulations to reduce tax evasion. Recently, it has resorted to artificial intelligence techniques such as machine learning (ML) and deep learning (DL) such as neural networks, decision trees, random forests, clustering techniques such as K-Mean, and others to reduce tax evasion. In this paper, we will present comparisons between a group of

ML techniques and two types of splitting datasets by conducting some experiments on a dataset of taxes in Yemen after the preprocessing for the data, and make our notes.

2. Related Work

In [2] the researchers concentrated on the effectiveness of using a hybrid intelligent system. It combined (MLP) neural network, (SVM), and logistic regression (LR) classification models with harmony search (HS) optimization algorithm for detecting tax evasion for the Iranian National Tax Administration (INTA). The results showed from out-of-sample data that MLP neural network in combination with HS optimization algorithm outperforms other combinations with 90.07% and 82.45% accuracy, 85.48% and 84.85% sensitivity, and 90.34% and 82.26% specificity, respectively in the food and textile sectors. In addition. There was also a difference between the selected models and obtained accuracies.

A strategy was presented in [3] for evaluating and predicting corporate financial fraud forecasts. It was found the method presented performed well, and it showed a high improvement over its basic algorithms, SVM and ID3. They proposed a 6.66 percent improvement over the ID3 algorithm and an 8.27 percent improvement over SVM. Then they worked with the Bayesian network algorithm. They also investigated and it was found the proposed method performs better than the Bayesian algorithm and has higher accuracy. The Bayesian algorithm performed better than the SVM and ID3 algorithms. However, it was observed if the MSE error rate is investigated, the ID3 has an error rate lower than the Bayesian, because the MSE is dependent on TP, TN, FP, and FN.

In [1] The goal of the research was to contribute to the detection of tax fraud, concerning personal income tax returns (IRPF, in Spanish) that were filed in Spain. Through using Machine Learning advanced predictive tools. By applying (MLP) models.

The using of the neural networks enabled taxpayer segmentation. Also, calculation of the probability concerning an individual taxpayer's propensity for attempting to evade taxes. The results showed the selected model has an efficiency rate of 84.3%, implying an improvement over the other models that are utilized in tax fraud detection.

in [4] the researcher proposed a system based on ML techniques capable of classification whether a company is involved in fraud or not based on financial and tax data from various companies, four different classifiers (Random Forest (RF), k-Nearest Neighbors (KNN), Neural Network, and Support Vector Machine (SVM)), was trained and used to indicate fraud. The model achieved the best performance a macro averaged F1-score of 92.98% with the (RF). The work presented a system that relies on machine learning for detecting tax evasion in the state of Esp´ırito Santo (Brazil). The results showed that RF achieved the best performance with a macro-averaged F1 score of 92.98%. The KNN and SVM achieved statistically equivalent performance, and the lowest F1score was achieved with Neural Network.

In [5] it was explored the application of machine learning technique (ML) for predicting fraudulent financial fraud statements by using analyzing the text content of publicly available financial statements. Implementation of two textual analysis methods besides a third method that combined between the two methods showed a promising result for the application of ML technique for predicting fraudulent financial statements, through analysis of the content of textual. The combined method produced the best results for sensitivity, accuracy, and type II error with a ratio of 93%,79%, and 7% respectively.

The paper proposed a system based on machine learning able to classify whether a company is involved in fraud or not. Rely on tax and financial data from different companies, four classifiers – k-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), and a Neural Network (NN) were trained, then used for indicating fraud. The best model was RF where it achieved a macro averaged F1-score of 92.98% [6].

In [7] the results based on the accuracy indicated that the PNN was the best performing Naives bays and SVM gives good results with the NSL-KDD dataset (99,02%, 98,8%) for credit card fraud. Also, they (98.09%) followed by the Genetic algorithm (95%) gave lower accuracies in most cases.

Methodology

The full methodology of the proposed study to compare the performance of a set of machine learning models on the datasets provided by the Tax Authority of Yemen is shown in Figure 1. The proposed study was designed on supervised machine learning techniques.

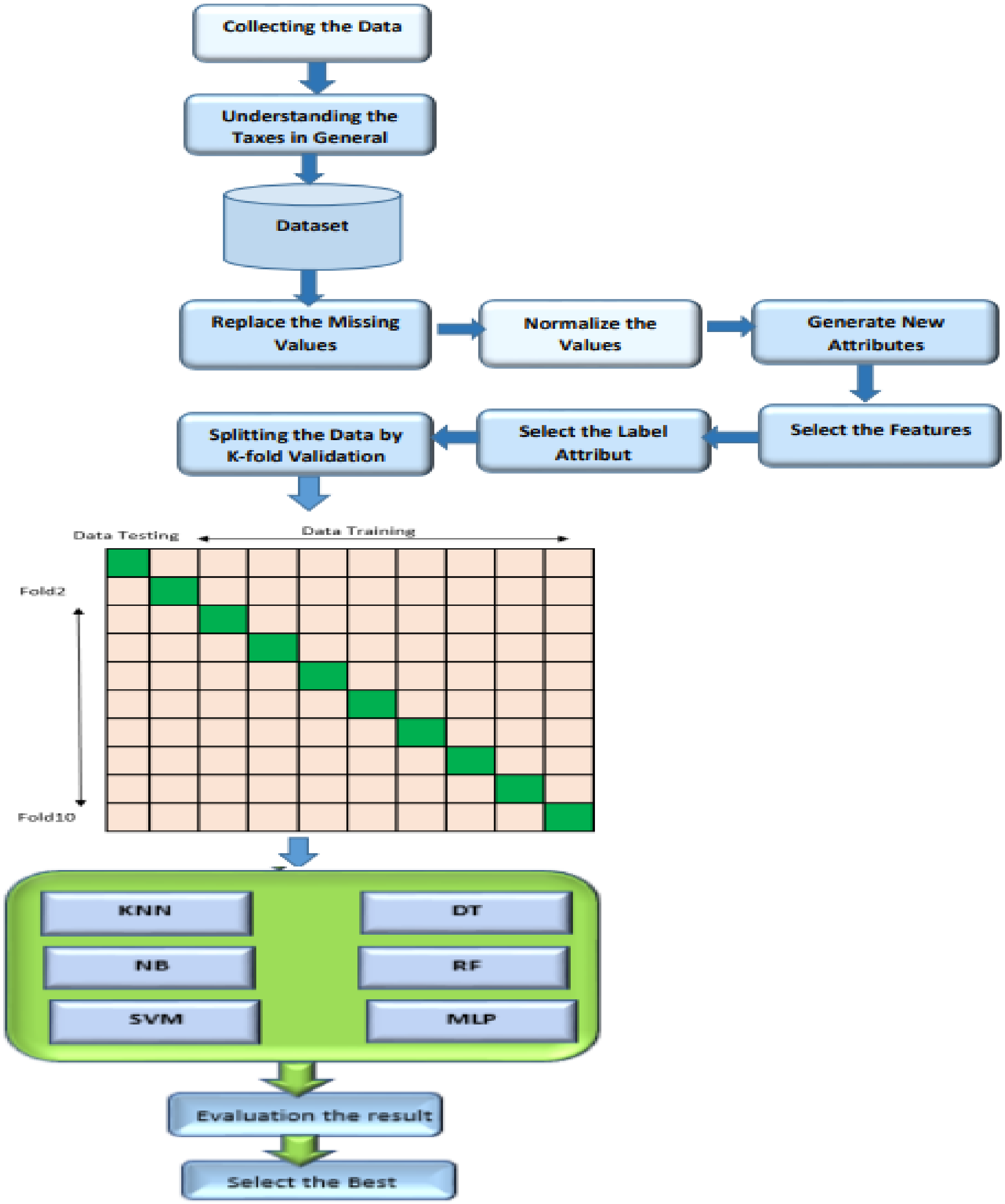

Figure. 1: Compression Mode

3.1-Data collection

Data collection is the most important part of the research. We collected data from the Tax Authority of Yemen.

We have taken one type of tax, which is the commercial and industrial profits tax because this type of tax is the backbone of the rest of the other types of taxes. The data were described and the meanings of the fields were clarified and the extent of their impact on other variables from the tax accounting side, are as shown in Table 1.

Table 1: The variables of Taxes

No	Name of variable	The description	Type
1	TIN	Tax Number	Integer/number
2	TN	Trade Name	Var Char
3	LE	legal entity	Var Char
4	TP	Tax period	Integer/number
5	T_type	Tax type	Var Char
6	BN	Business Number	Integer/number
7	Tax	Tax	Integer/number
8	Punder	Payable under account	Integer/number
9	Dtax	Due tax	Integer/number
10	Fine	Fines	Integer/number
11	Damount	Deserved amount	Integer/number
12	Tax_L div BN_L	Tax rate to turnover	Integer/number
13	Tax_C div BN_C	Tax rate to turnover	Integer/number
14	Ratio_C	Tax rate to turnover for the previous year	Integer/number

We maintained the necessary features and eliminated unnecessary features that did not help in the decision-making process from the training dataset. This step is very important in reducing the dimensions of the input, which reduces the execution time increases the prediction accuracy, and eliminates confusion from the data in the case of adding unnecessary features and variables. In our study, we had 14 columns and we reduced the dimensions to reach 6 columns as shown in the table2.

Table 2: Features Selection

No	Name of variable	The description	Type
1	TIN	Tax Number	Integer/number
2	BN	Business Number	Integer/number
3	TAX	Tax	Integer/number
4	Paid under	Payable under account	Integer/number
5	Fine	Legal Fine	Integer/number
6	Ratio_C	Tax rate to turnover for the previous year	Integer/number

3.2 Data Preprocessing:

It is a set of operations used to modify the raw data as follows.

Data cleaning is very important because of its impact on the accuracy of classification, and the decision-making process is compromised when the data set is missing or incorrect.

Data cleaning in our data we replaced the missing data with zero. We made the normalization method to let the values between (0,1), and then we generated new attributes (computed attributes)

for adoption as tax risk criteria as follows:

If the tax payable is less than or equal to zero, there is a risk.

If the payment under the account is greater than zero and the tax is negative or zero, there is a risk If the payment under the account is greater than zero and the turnover is zero, there is a risk.

-If the payment under the account is more than the tax due, there is a risk.

If the Business number *1% is less than the appropriate tax, there is a risk.

If there is a fine, that means the fine is greater than zero, then there is a risk.

We set a score for each criterion according to priority so that the total is 100%.

The restrictions were studied and reviewed with experts in the tax and accounting field, and comparisons were made to determine the extent of the taxpayer's commitment or not. As shown in Table 3.

Table 3: Generated New Attributes

No	Name of Attributes	Function Expressions
1	Tin	
2	Finecode_C	if([Fine_C]>0,20,0)
3	Taxcode_C	if([Taxdue_C]<0,20,0)
4	BN_C*1	BN_C /100
5	Paid and BN_C	if([Paid Under_C]>0 && [BN_C]<=0,15,0)
6	Paid and Tax_C	if([Paid Under_C]>[Taxdue_C],15,0)
7	Tax_C div BN_C	[Taxdue_C] /[BN_Cu]
8	BN_Cu	if([BN_C]== 0,1,[BN_C])
9	Taxcode_L	if([Taxdue_L]<0,20,0)
10	Finecode_L	if(Fine_L>0,20,0)
11	BN_La	if([BN_L]==0,1,[BN_L])
12	Tax_L div BN_L	[Taxdue_L] /[BN7]
13	Ratio_C	if([Tax_C div BN_C]<[Tax_L div BN_L],10,0)
14	BN and Tax_C	if([BN_C*1]<[Taxdue_C],20,0)
15	Sum_C	[BN and Tax_C]+[Paid and BN_C]+[taxcode_C]+[finecode_C]+[Paid and Tax_C]+[Ratio_C]
16	Class_C	if([sum_C]>=65,"F",if([sum_C]>=25,"PT",if([sum_C]>=10,"OT","E")))

Then we selected features (feature reduction) (to choose the appropriate features). for the second time to work within the model. As shown in Table 4.

Table 4: Select features for the second time

No	Attributes
1	BN and Tax_C
2	Paid and Bn18_C
3	Paid and Tax_C
4	Ratio_C
5	Class_C
6	Finecode_C
7	Sum_C
8	Taxcode_C

After cleaning the data, normalizing, generating new attributes, selecting features, and devising new features, we choose the attribute to be a label for the model.

The Actual data after processing

In our study we divided the behaviors of taxpayers into four categories bases on the level of evasion as the follow:

Complete evasion.
Partial evasion.
Simple evasion.
Tax committed

Each of these cases has a specific treatment. Complete evasion is examined by comprehensive examination at the taxpayer's headquarters, and partial evasion is examined by partial examination at the taxpayer's headquarters, and simple evasion is examined by a desk examination at the tax administration, and the obligated taxpayer is not examined or visited, but may be give him a set of tax concessions.

We made label for each case of treatment as follow: -

comprehensive examination= F.
partial examination=PT.
desk examination =OT
obligated taxes=E .

After the pre-processing, we get the next chart as shown in Fig2.

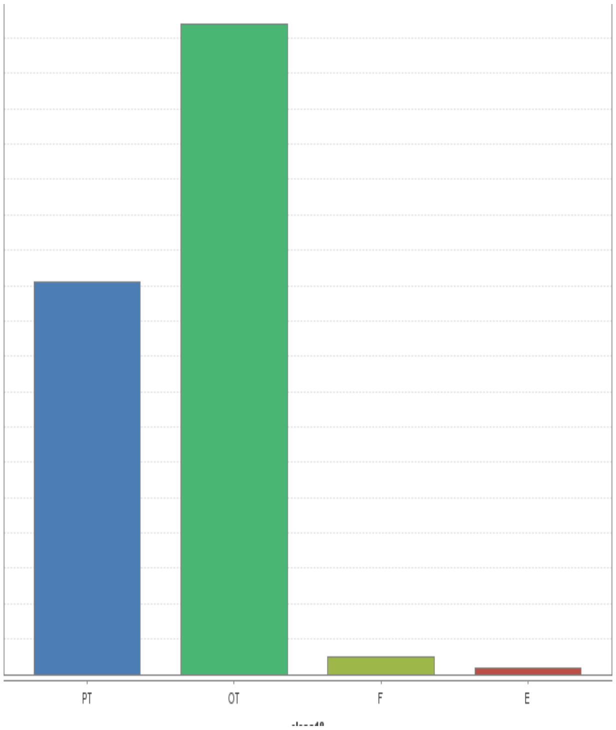

Figure. 2: The Actual data after processing

Then it was time for training and testing. The training data is used to train the model, and the dependent variable is known as Test Data. The test data is used to make the predictions from the model that is already trained on the training data

The data was trained using the K-Fold Validation technique.

4. Splitting data by K-Fold Validation technique

Cross-validation is a resampling process used for evaluating machine learning models, on a limited data sample.

The process has a single parameter called k, which refers to the number of groups a given data sample is to be split into. This procedure is often called k-fold cross-validation. When choosing a specific value for k, it can be used instead of k, in the reference to the model, like k=10 becoming 10-fold cross-validation.

Cross-validation is used in applied machine learning for estimating the skill of a machine learning model, on data that is unseen. That is, using a limited sample for estimating how the model is expected to perform in general. When used for making predictions on data, that is not used during the training of the model.

It is a popular method because it is simple to understand, and because it generally results in a less biased, or less optimistic estimate for the model skill than other

methods. Such as a simple train/test split.

The general process is as follows: -

1. Shuffling the dataset randomly.
2. Splitting the dataset into k groups.
3. For each unique group
   - Taking the group as a holdout or test data set.
   - Taking the remaining groups for a training data set.

Fitting a model on the training set, and evaluating it on the test set.

Retaining the evaluation score, and discarding the model.

4. Summarizing the skill of the model by using the sample of model evaluation scores. Importantly, each observation in

the data sample is assigned to an individual group and stays in that group for the duration of the process. That means, that each sample is given the same opportunity to be used in the hold-out set 1 time, and used for training the model k-1 times. This approach includes randomly the set of observations into k groups. Or folds, equal in size approximately. It is treated with the first fold as a validation set, and the method is to fit the remaining k – 1fold [7].

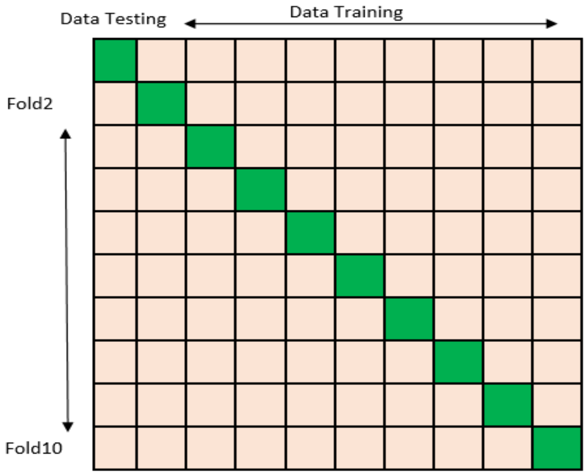

Figure. 3: K-Fold Validation

5. The algorithms that used for experiments

- KNN.
-NB.
-SVM.
- DT.
-Random Forest RF.
-Malty Layer Perceptron MLP.

Model Evaluation

This activity is responsible for describing the evaluation parameters, and results of the defined model. The model is evaluated by using an evaluation parameter, that compares the number of data points, that are properly and erroneously classified in the confusion matrix, which includes values of true positives (correct classifications) and false positives (incorrect classifications), to evaluate the various classification models.

6.1 Confusion Matrix

The confusion matrix is an N x N matrix, used to evaluate the performance of a classification model, as N is the number of target classes.
By visualizing the confusion matrix. An individual can determine the accuracy of the model, by observing the diagonal values to measure the number of accurate.
classifications. The confusion matrix is a square matrix, where the column represents the actual values, and the row depicts the predicted value for the model and vice versa.

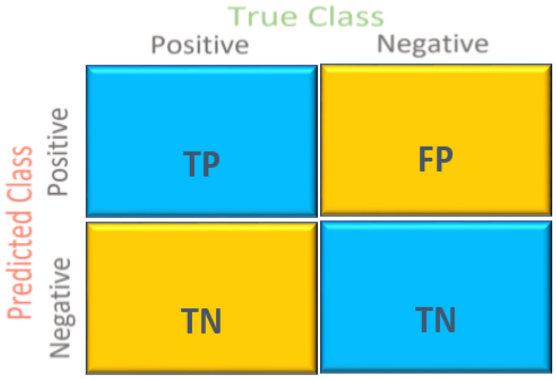

Figure. 4: Confusion Matrix

TP: True Positive: The actual value is positive, and the model predicts a positive value.

FP: False Positive: the prediction is positive, but it is false. (Known as Type 1 error).

FN: False Negative: the prediction is negative, and the result is also false. (Known as Type 2 error.

TN: True Negative: An actual value is negative, and the model also predicted a negative value [8][9].

A. Accuracy

Accuracy is a measure of the number of correct predictions in the model, that made for the complete test dataset. Accuracy is a good metric to measure the model performance, in unbalanced datasets the accuracy becomes a poor metric.

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}\ \textit{(1)}$$

B. Precision

Precision tells us, how many of the cases that are correctly predicted actually, and turned out to be positive. This will determine whether the model is reliable or not. Precision is a useful metric in cases where a False Positive is a higher concern than a False Negative.

$$\boldsymbol{Precision} = \frac{TP}{TP+FP}\ (2)$$

C. Recall

Recall tells us about how many actual positive cases, we were able to predict with our model. The recall is a useful metric, in cases where the False Negative is on False Positive.

$$\boldsymbol{Recall} = \frac{TP}{TP+FN}\ (3)$$

D. F-Score

F-measure (F1 score) is defined as the mean of precision and recall. It is a measure that combines accuracy and recall into a single performance measure. Averaging recall and accuracy yielded the F1-score. Precision and recall contribute equally to the F1 score [10].

$$\boldsymbol{Fscore} = \frac{2*(Precision*Recall)}{Precision+Recall}\ (4)$$

6. Experiments

A. Experiment 1

The First experiment was done by using a KNN classifier to predict taxpayers‟ compliance levels after completing the data processing. In the KNN model, we achieved 99. 87% of accuracy, 99.95% recall, 98.04% in Precision, 98.98% in F-Score.

B. Experiment 2

The Next experiment was done by using the Naive Bayes (NB)classifier to predict taxpayers‟ compliance levels after completing the data processing. In the BN model, we achieved 100% of accuracy, 100% recall, 100% in Precision, and 100% in F-Score.

C. Experiment 3

The Third experiment was done by using the Support Vector Machine (SVM) classifier to predict taxpayers‟ compliance levels after completing the data processing. In the SVM model, we achieved 99.87 accuracies, 99.95% recall, 98.04% in Precision 98.98% in F-score

D. Experiment 4

The Fourth experiment was done by using a Decision Trees (DT) classifier to predict taxpayers‟ compliance level after completing the data processing. In the DT model, we achieved 97.76% accuracy, 57.006% recall, 99.24% in Precision 72.41% in F-score.

E. Experiment 5

The Fifth experiment was done by using a Random Forest (RF) classifier to predict taxpayers‟ compliance level after completing the data processing. In the RF model, we achieved 99.87% of accuracy, 97.91% recall, 99.95% in Precision 98.91% in F-score.

F. Experiment 6

The Sixth experiment was done by using an MLP classifier (MLP) to predict taxpayers‟ compliance levels after completing the data processing. In the MLP model, we achieved 98.42% accuracy, 66.62% recall, 64.21% in Precision 64.40% in F-score.

7. Discussion the Results

We summarized the five evaluation metrics that evaluated ML models that used the tax dataset. Its accuracy, Precision, Recall, F-Score, and misclassification are as follows:

The e Naïve Bayes (NB) classifier gave the highest result in accuracy and other measures. Then KNN, SVM, and RF gave the same results in accuracy 99.87%, but in SVM, KNN the results were also the same in the rest measures as follows 99.95%,98.04%, and 98.98 % in Recall, Precision and F-score, while in RF models the rest measures were 97.91%,99.95% and 98.91% in Recall, Precision and F-score. MLP gave 98.42 in accuracy with 66.62 recall, 64.21 Precision, and 64.40 in F-score, then DT gave 97.76% in accuracy with 57.006% ,99.24%, and 72.41% in recall, precision, and F-score, as it showing in the table5.

Table 5: The Summaries of Experiments

Performance measures	ML Techniques					
	KNN	(NB)	(SVM)	D T	RF	(MLP)
Accuracy	99.87%	100%	99.87%	97.76%	99.87%	98.42%
Recall	99.95%	100%	99.95%	57.006%	97.91%	66.62%
Precession	98.04%	100%	98.04%	99.24%	99.95%	64.21%
F-Score	98.98%	100%	98.98%	72.41%	98.91%	64.40%

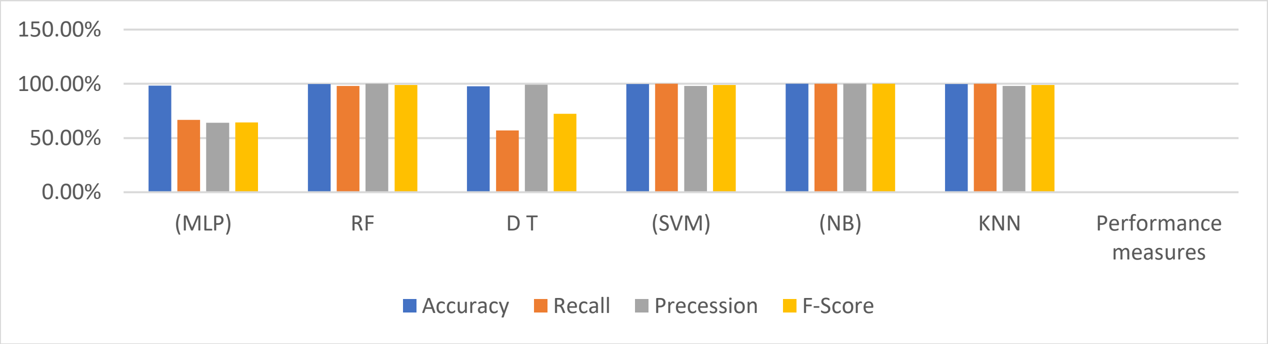

Figure. 5: Performance Measures of ML Modules

8. Conclusion and Future work

This paper presents a comparison of the performance of a set of machine learning models as follows (KNN, NB, SVM, DT, RF, MLP) on the dataset provided by the Tax Authority of Yemen. The results showed that the e Naïve Bayes (NB) classifier gave the highest result in accuracy and other measures. Then KNN, SVM, and RF gave the same results in accuracy 99.87%, but in SVM and KNN the results were also the same in the rest measures, while in the RF model, the rest measures were 97.91%,99.95%, and 98.91% in Recall, Precision, and F-score. MLP gave 98.42 in accuracy with 66.62 recall, 64.21 Precision and 64.40 in F-score, then DT gave 97.76% in accuracy with 57.006%,99.24%, and 72.41% in recall, precision, and F-score. We will develop a new technique to build a new model based on the dataset of the Tax Authority of Yemen to detect tax fraud and get the best results.

9. References

[1] C. P. López, M. J. D. Rodríguez, and S. de L. Santos, “Tax fraud detection through neural networks: An application using a sample of personal income taxpayers,” Futur. Internet, vol. 11, no. 4, 2019, doi: 10.3390/FI11040086.

[2] E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, “International Journal of Accounting Information Systems Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran,” Int. J. Account. Inf. Syst., vol. 25, pp. 1–17, 2017, doi: 10.1016/j.accinf.2016.12.002.

[3] A. Javadian, A. Ali, P. Aghajan, and M. Hosseini, “A Hybrid Model Based on Machine Learning and Genetic Algorithm for Detecting Fraud in Financial Statements,” J. Optim. Ind. Eng., vol. 14, no. 2, pp. 169–186, 2021, doi: 10.22094/JOIE.2020.1877455.1685.

[4] J. P. A. Andrade et al., “A MachineLearning-based System for FinancialFraud Detection,” pp 165–176, 2021, doi:10.5753/eniac.2021.18250

[5] C. Page, “Master’ s Thesis Predicting Fraudulent Financial Statement using Textual Analysis and Machine-Learning Techniques by DIMAS Lagusto September 2018 Master’ s Thesis Presented to Ritsumeikan Asia Pacific University In Partial Fulfillment of the Requirement,” Ritsumeikan Asia Pacific University, 2018.

[6] J. P. A. Andrade et al., “A Machine Learning-based System for Financial Fraud Detection,” pp 165–176, 2021, doi:0.5753/eniac.2021.18250.

[7] I. Sadgali, N. Sael, and F. Benabbou, “Performance of machine learning techniques in the detection of financial frauds,” Procedia Comput. Sci., vol. 148, no. Icds 2018, pp. 45–54, 2019, doi: 10.1016/j.procs.2019.01.007.

[8] Z. Karimi, “Confusion Matrix,” no. October, pp. 0–4, 2021.

[9] A. Tasnim, M. Saiduzzaman, M. A. Rahman, J. Akhter, and A. S. M. M. Rahaman, “Performance Evaluation of Multiple Classifiers for Predicting Fake News,” J. Comput. Commun., vol. 10, no. 09, pp. 1–21, 2022, doi: 10.4236/jcc.2022.109001.

[10] R. Analysis, H. Reliability, P. S. Assessment, S. V. Machine, and A. Kulkarni, “Confusion Matrix,” ScienceDirect, pp. 1–22, 2022.