Journal of Theoretical and Applied Information Technology
30th April 2022. Vol.100. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BORUTA FEATURE SELECTION AND PARTICLE SWARM OPTIMIZATION FOR FRAUD DETECTION ON PAWN TRANSACTION

[1]MASDRA WAHYU KUNCORO, [2]SUHARJITO

[1,2] Computer Science Department, Master of Computer Science, Bina Nusantara University, Jakarta, 11480, Indonesia

E-mail: [1]masdra.kuncoro@binus.ac.id, [2]suharjito@binus.edu

ABSTRACT

The pawn company is one of the largest financial service companies in Indonesia, pawning is become an alternative for Indonesia people to obtain credit other than Bank. The number of pawn transaction raises the potential for fraud action. The use of PSO has shown promising result for improving classification accuracy. This will be a problem if the dataset used has many attributes. Previous research on fraud detection is usually carried out on credit card transaction, there has not been any research on fraud detection in pawn transaction. This research proposes based on data mining model to combining Boruta Feature selection and Particle Swarm Optimization (PSO). For classification using Gradient Boosted Tree (GBT) and C5.0 to measure the level of accuracy. The research is comparing several classifications and to know the highest accuracy of some classification. Pawn transaction data has been taken from pawn company in Indonesia. There are 216 transactions in 2019 until 2020. Among them, 26 transactions detected as fraud and 191 are no fraud. The attributes used is 24, among other is name of the customer, address, type of work, age, address, loan destination, identity number, collateral category, estimated value of collateral, loan money, credit time, type of product, type of transaction, class of collateral, maximum loan money, weight of collateral, and others. The results indicate that the combination of C5.0 optimizing by PSO and Boruta feature selection gives the highest classification accuracy of 96.82% and the GBT optimizing by PSO and Boruta feature selection reaching accuracy 93.57%.

Keywords: *Fraud Detection, Boruta, Particle Swarm Optimization, Gradient Boosted Tree, C5.0*

1. INTRODUCTION

Pawn companies in Indonesia have been around since 1901. Pawn transaction is an alternative way other than Bank to get credit for Indonesian people. In Indonesia there are a lot of pawn companies, one of the biggest is PT Pegadaian (Persero). Currently there are approximately 13 million pawn customers at the company in 2020, these customers are mostly in rural areas and not able to Bank. With the large number of customers and transaction, the risk of fraud is very large and the company had difficulty detecting fraud quickly.

Fraud can be defined as wrongful or criminal deception intended to personal gain, fraud can be doing by people from inside or outside the organization [1]. The fraudulent action is not always same as a crime. Fraudulent action that is not a crime categorized as operational risk. For the present, fraud is defined as any behavior by which one person gains or intend to gain a shared advantage over another [2]. Fraud can be declared to be a criminal act or an act to gain a dishonest advantage and violates constitutional provisions [3]. Fraud usually founded in corporate or government organization. Principally, fraudulent action in a company can be carried by employees or leaders, where the result of the fraudulent action is losses to the company [4]. The company's losses because fraudulent action can eventually lead to bankruptcy [5].

Techniques used in fraud detection can be divide into two : Supervised techniques where the past of fraud is known; and unsupervised techniques where there are no prior sets in which the state of the transaction are known to be fraud [6]. The most common classifier used to detect fraudulently is decision tree, neural network, support vector machine (SVM), logistic regression, k-means clustering and nearest neighbor algorithms.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Previous research has proven that fraud can be detected by applying data mining and machine learning methods. Fraud detection can be detected by applying a classification model with the SVM algorithm on credit card transactions [7]. The data used in the study with 100 training data, the classification process using SVM and define the attribute account number as label, month and transaction nominal as values. The case resolution in this method is made by looking for anomalies/outliers using hyperplane. The weakness of this research is that only uses relatively little test data and does not measure the accuracy of SVM performance in detecting fraud.

The clustering method is applying to detect fraud in credit card transactions [8]. Data is randomly generated using Microsoft SQL Server Management Studio because there are no original data on credit card transactions, the resulting data is used to detect fraud with the k-means clustering algorithm which is implemented using .NET programming language in Visual Studio 2012. In this study, the data are generated into several features, including transaction ID, transaction number, country transaction, transaction date, and credit card number. In his research, it was explained that the application of k-means clustering in detecting fraud by grouping the level of fraud into four categories, namely low level, high-level, high-risk and high-risk fraud represented by colors (orange, yellow, green, and purple). Fraud on credit card transactions cannot be detected 100%, but you can see transactions with their respective risk levels.

The completion of fraud detection cases using decision tree classifiers [9]. The study used data on 202 companies listed of the Chinese stock exchange with a total of 35 features. This study uses 18 important features after pre-processing using t-statistics. The types of decision Tre algorithm used in this study include Random Forest, Naïve Bayesian Tree, C4.5, RIPPER, CART and Tree Net. The accuracy result of C4.5 algorithm is 58% for the 18 features test and 57% for the overall feature test.

The C5.0 algorithm can be optimize with the feature section method and reduce error pruning (REP) [10]. From the research results, the comparison error value between the algorithm C5.0 which is optimized using the feature section and reduce error pruning (REP) compared to C5.0 without optimization is 0.9% for C5.0 with optimization and 6% for C5.0 without optimization, the overall value of the error rate is generated on the comparison testing dataset.

The classification of social assistance receipts by combining the K-NN and Gradient Boosted Tree methods, using a dataset from the Central Statistics Agency (BPS) in 2019 [11]. The research carried out was to predict the correct level of acceptance of social assistance in the correct category. Poor families and not poor families. From the results of the research, the results of the accuracy rate using the K-NN algorithm were 89.04%, using the Gradient Boosted Tree algorithm was 93.15%, the results of the accuracy test for the combination of the K-NN and Gradient Boosted Tree methods were 98.17%.

Integrated the Gradient Boosted Tree algorithm with SMOTE and bagging to examine student graduation rates [12]. The data used in the research were obtained from the Directorate of Higher Education for the 2018/2019 academic year. In this study, the classification of GBT with SMOTE and bagging is able to solve the problem of class imbalance (class imbalance) and reduction of errors in the classification model (misclassification) of the unbalanced dataset, so that it can improve the performance of a GBT model. The results of this study indicate that the best value is in the 90:10 split with an accuracy value of 80.57% and an AUC value of 0.858 and in the 90:10 split with an average accuracy value of 79.44% and an AUC value of 0.852. In this study, it was concluded that the application of SMOTE and bagging was proven to be able to provide solutions to the handling of class imbalance problems and to improve the performance of the GBT classification model.

Compared the Random Forest algorithm and the Support Vector Machine (SVM) to predict breast cancer [13]. In this study, the Boruta algorithm is used to select features. After selecting the features using the Boruta algorithm, the attribute classification is labeled as important and unimportant. The test results show the classification using Random Forest produces an accuracy of 90.90%, while the SVM classification produces an accuracy of 95.45%.

Predicted leukemia cancer patients using the AdaBoost classification, Regression Tree, Artificial Neural Network, Random Forest, Linear Discriminant Analysis, and Naïve Bayes and performed feature selection with 5 (five) methods, namely T -test, WCSRC test, RF, Boruta, and LASSO [14]. The data used were 72 patient data consisting of 7129 genes in which 25 patients had

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

leukemia cancer. From this research, the highest accuracy is 99.95% using a combination of LASSO and Naïve Bayes.

This study proposes to optimizing Gradient Boosted Tree (GBT) and C5.0 classification method using Particle Swarm Optimization (PSO) and Boruta feature selection. The main contributions of this study are as follows:

- We combining the Particle Swarm Optimization and Boruta Feature Selection to improve the classification accuracy of the GBT and C5.0 methods.
- We use the pawn transaction data to find the best method to fraud detection on pawn transaction.

2. RESEARCH METHOD

Data collection is by requesting pawn transaction data at one of the pawn companies in Indonesia. The data contains a lot of information related to pawn transactions. Such as the name of the customer, address, type of work, age, address, loan destination, identity number, collateral category, estimated value of collateral, loan money, credit time, type of product, type of transaction, class of collateral, maximum loan money, weight of collateral, and others all of the attribute will be rename as V1,V2,V3,V4,V5…V24. In determining which transactions have occurred fraud, researchers are assisted by internal company parties to determine which transactions have occurred fraud from the data, and then provide additional labels in the form of information "Fraud" and "No Fraud".

In this study, researchers only used pawn transaction data from one branch of pawn company. The data used in this study was 216 pawn transaction, including 26 transactions that occurred fraud, data was taken only from pawn product from 2019 to 2020. The master data obtained are credit data, customer data, non-cash transaction data, and data on fraud. After getting the master data, the next step is to merge the data into one file. From the combined master data, feature selection is carried out using the Boruta algorithm. The implementation of the Boruta algorithm is carried out using the R programming language using the RStudio application. From the results of the feature selection, a dataset is obtained which will be used for the classification test. The dataset is then divided into two parts, namely training data and testing data.

The framework proposed in this study is shown in Fig. 1. This study proposes the classification process is carried out using the Gradient Boosted Tree (GBT) and C5.0 algorithm, which is optimized using the Particle Swarm Optimization (PSO) algorithm and Boruta feature selection. From the classification results, the level of accuracy is measured and compared between the Gradient Boosted Tree (GBT) and C5.0 algorithm so that the best accuracy results are obtained.

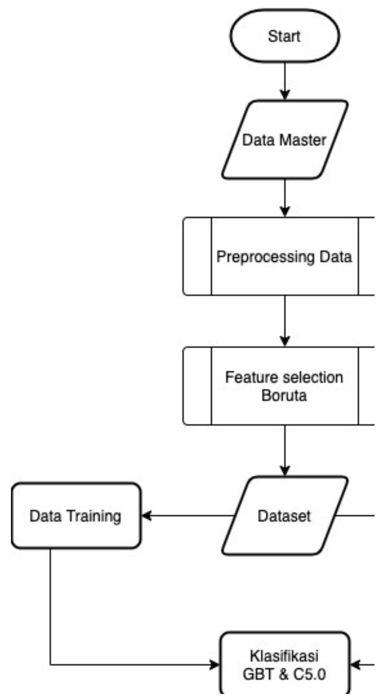

Figure 1 propose method

2.1 Feature Selection Techniques

Feature selection is one of the most important steps in machine learning [15]. When incorporating features into a model, the aim is to feed the model with the relevant features for predicting class. Including irrelevant features creates unnecessary noise issues in the data, which results in lower model accuracy. Generally, we use statistical feature selection methods such as ANOVA or Chi-squared test, evaluating the relationship between each predictor variable and the target variable [16].

Boruta feature selection is built around the random forest classification algorithm. Random forest is a classification method which is performed by voting of multiple unbiased decision trees built from samples of the training set [17]. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

importance of the feature is obtained from the loss of accuracy of classification.

2.2 Gradient Boosted Tree

Boosting is one of the predictive models with the main idea of combining a simple model (weak learner) so that it becomes one strong model iteratively. At each iteration, this algorithm seeks to obtain a weak learner who is able to predict better than that obtained in the previous iteration. The basis of this algorithm was first put forward [19] which was then applied to a predictive model [20] called the AdaBoost algorithm or Adaptive Boosting.

Furthermore Friedman [21] propose a new boosting model by embedding a statistical modeling framework on AdaBoost. This new model was then called the Gradient Boosting Machine). In 2016, GBM underwent a development in the application of its algorithm so that the GBM model could be implemented faster and produce more accurate predictions. This development was proposed by Tianqi Chen, who later gave the model the name XGBoost (Extreme Gradient Boosting).

2.3 C5.0

The C5.0 algorithm is a refinement of the ID3 algorithm and C4.5. In the process of forming a decision tree the highest gain information value will be selected as root for the next node. This algorithm begins with all data are used as the root of the decision tree while the attributes are selected will be the divider for that sample. [22]

$$Info(D) = -\sum_{i=1}^{m} pi\, log_2(pi) \quad (1)$$

Where Info (D) is the information needed to classify the class label of a tuple in D. pi is a non-zero probability with a random tuple in D. The log function uses base 2, because the information is encoded in bits. Info (D) is also known as entropy.

$$Info_A(D) = \sum_{J=1}^{Y} \frac{|Dj|}{D} x Info(Dj) \quad (2)$$

To get the information gain value on attribute A.

$$\boxed{\text{Gain (A)} = \text{Info(D)} - \text{Info (Dj)}} \quad (3)$$

Gain (A) states how many branches will be obtained on A. Attribute A with the highest information gain. The information, Gain (A), is selected as the attribute at node N.

2.4 Particle Swarm Optimization

The Particle Swarm Optimization (PSO) method was introduced by Doctors Kennedy and Elbert in 1995 based on research conducted on the behavior of birds and fish and is a global heuristic optimization method [23]. PSO is a population-based iterative algorithm. The population consists of many particles, which are initialized with a random solution population and used to solve optimization problems [24]. Each particle represents a candidate solution and moves towards the optimal position by changing its position according to the speed of the particle flying through the search space at a speed dynamically adjusted for historical behavior. Therefore, particles have a tendency to fly to better and better search areas during the search process [24].

$$V_i(t) = V_i(t-1) + c_1 r_1 [X_{pbesti} - X_i(t)] + c_2 r_2 [X_{gbest} - X_i(t)] \quad (4)$$

3. RESULTS AND ANALYSIS

The feature selection process is carried out using the Boruta algorithm in the RStudio application. The number of iterations carried out is as much as 124 times with a processing time of 5.20 seconds. From the results, it was found that 18 attributes had the important category and 6 other attributes that were unimportant, as can be seen in the picture Fig.2. Attribute with decision "Confirm" will be used to test the classification model and attribute with decision "Rejected" will be ignored. From the feature selection process, it can be seen that attribute with normHits value above 0.65 will be accept.

> attStats(boruta)

	meanImp	medianImp	minImp	maxImp	normHits	decision
V3	12.9736132	12.9555567	11.3813321	14.2241351	1.00000000	Confirmed
V6	2.7832584	2.8841915	0.5248168	5.4585129	0.66129032	Confirmed
V7	1.3692409	1.6083367	-0.5726824	3.1680530	0.08870968	Rejected
V9	3.1025433	3.1790850	0.8937121	4.4485706	0.79838710	Confirmed
V10	8.5412516	8.5199221	6.9931323	9.7417022	1.00000000	Confirmed
V11	2.6643051	2.6870129	0.7133897	4.3852687	0.65322581	Confirmed
V13	14.3252740	14.3264433	12.2689731	15.7309361	1.00000000	Confirmed
V14	14.2873449	14.3408084	12.4669909	15.8302164	1.00000000	Confirmed
V18	3.7276028	3.7389614	0.9854949	6.3816995	0.90322581	Confirmed
V19	4.0259405	4.0647686	1.2463569	6.0167387	0.90322581	Confirmed
V21	10.0433066	10.0091053	8.3927062	12.3147665	1.00000000	Confirmed
V22	8.0239895	7.9919459	6.6348407	9.4967665	1.00000000	Confirmed
V24	12.0619601	12.0412603	10.0435521	13.8070657	1.00000000	Confirmed
V1	14.7197750	14.7372129	12.6222817	16.9743111	1.00000000	Confirmed
V2	5.5192548	5.6098835	3.1181976	8.0466447	1.00000000	Confirmed
V4	10.0779970	10.0860613	8.0264089	11.5753359	1.00000000	Confirmed
V5	0.5078190	0.8091464	-1.2998586	2.0459176	0.00000000	Rejected
V8	0.7194974	1.0010015	-1.4157719	1.9131904	0.01612903	Rejected
V12	0.7120842	0.8211071	-1.2778927	1.8651497	0.01612903	Rejected
V15	3.8887051	3.7723920	2.4066372	5.4535282	0.95967742	Confirmed
V16	2.6512587	2.6566528	-0.4032245	4.5603694	0.65322581	Confirmed
V17	-0.7140349	-0.6413972	-2.3142918	0.8786189	0.00000000	Rejected
V20	0.3185206	0.3531374	-1.6690911	1.4170209	0.00000000	Rejected
V23	5.1250526	5.1612429	3.7991871	6.2797755	1.00000000	Confirmed

Figure 2 boruta feature selection result

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

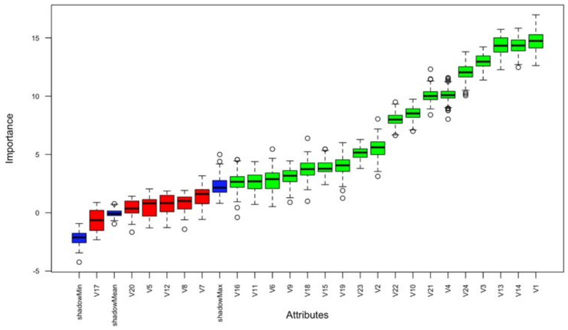

Figure 3 boruta feature selection stats

Fig. 3 shows the classification of attributes based on the value of the importance of these attributes. The green pin bar shows that the attribute has an important category, for the red pin bar it shows that the attribute has an insignificant category, while the blue pin bar shows the min shadow, mean shadow, and max shadow. From the results of the feature selection, it is known the value and ranking of attributes based on their importance, starting from V1, V14, V13, V3, V24, V4, V21, V10, V22, V2, V23, V19, V15, V18, V9, V6, V11, V16, V7, V8, V12, V5, V20, and V17. In the classification and optimization process only attributes that are in the important category will be used as a dataset, these attributes are V1, V14, V13, V3, V24, V4, V21, V10, V22, V2, V23, V19, V15, V18, V9, V6, V11, and V16 while the attributes that fall into the insignificant category will be ignored, those attributes are V7, V8, V12, V5, V20, and V17.

From the results of data testing, all data is tested to obtain accuracy, precision, recall and the number of fraud and no fraud values for each data. In the following table the highest value of each process is taken, this shows the best accuracy that can and has been achieved by each test data. The results of classification testing will be obtained, accuracy, precision, and recall values.

Table 1 result of GBT+PSO+Boruta model

Accuracy: 93.57% +/- 3.21% (micro average: 93.55%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	13	1	92.86%
Pred No Fraud	13	190	93.60%
Recall	50.00%	99.48%	

Table 2 Result of C5.0+PSO+Boruta model

Accuracy: 96.82% +/- 4.31% (micro average: 96.77%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	22	3	88.00%
Pred No Fraud	4	188	97.92%
Recall	84.62%	98.43%	

Table 3 Result of GBT+PSO model

Accuracy: 93.12% +/- 4.38% (micro average: 93.09%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	13	2	86.67%
Pred No Fraud	13	189	93.56%
Recall	50.00%	98.95%	

Table 4 Result of C5.0+PSO model

Accuracy: 96.32% +/- 3.60% (micro average: 96.31%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	19	1	95.00%
Pred No Fraud	7	190	96.45%
Recall	73.08%	99.48%	

Table 5 Result of GBT Model

Accuracy: 79.91% +/- 24.84% (micro average: 80.18%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	4	21	16.00%
Pred No Fraud	22	170	88.54%
Recall	15.38%	89.01%	

Table 6 Result of C5.0 Model

Accuracy: 91.69% +/- 4.77% (micro average: 91.71%)

	True Fraud	**True No Fraud**	**Class Precision**
Pred Fraud	10	2	83.33%
Pred No Fraud	16	189	92.20%
Recall	38.46%	2	

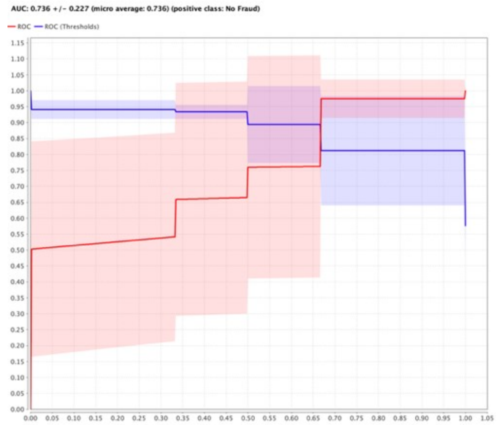

Figure 4 AUC curve results of GBT + PSO + Boruta model

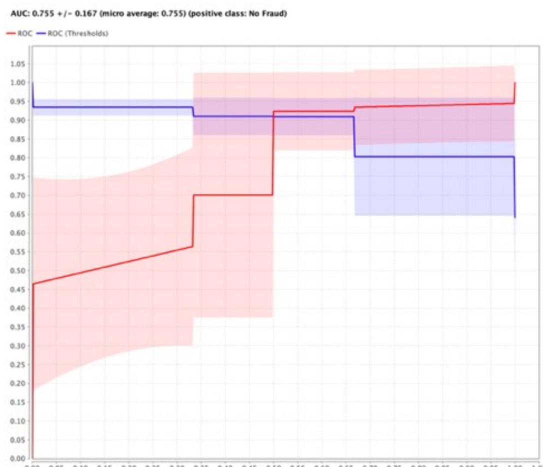

Figure 5 AUC curve results GBT + PSO model

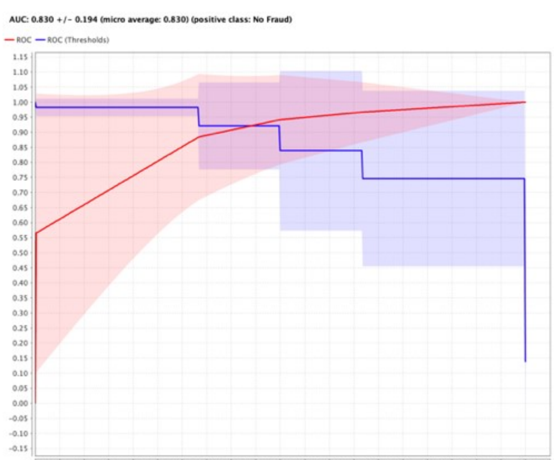

Figure 6 AUC curve results of C5.0 + PSO + Boruta model

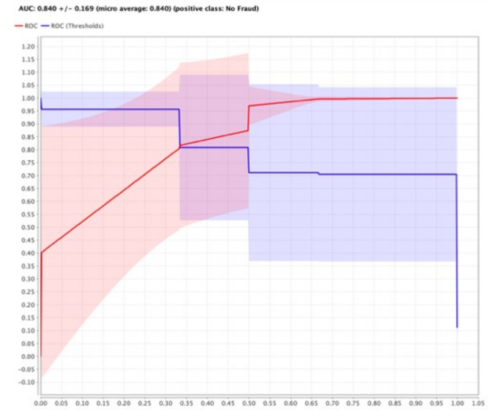

Figure 7 AUC curve results of C5.0 + PSO model

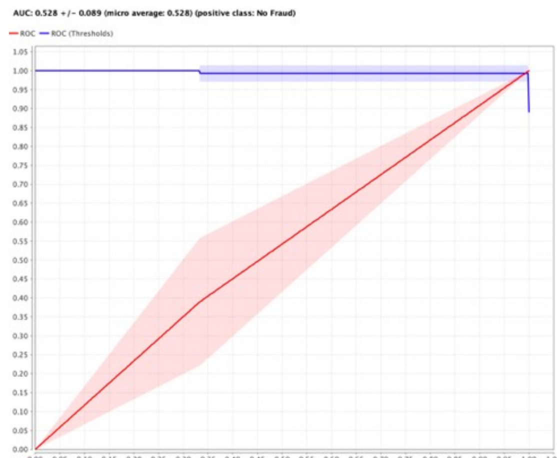

Figure 8 AUC curve results of GBT model

Figure 9 AUC curve results of C5.0 model

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Summary results from the classification model test can be seen in the table 7 :

Table 7 Classification Model Result

Model	Accuracy	Error Rate
GBT+PSO+Boruta	93.57%	**+/- 3.21%**
C5.0+PSO+Boruta	**96.82%**	+/- 4.31%
GBT+PSO	93.12%	+/- 4.38%
C5.0+PSO	96.32%	+/- 3.60%
GBT	79.91%	+/- 24.84%
C5.0	91.69%	+/- 4.77%

The classification model has the highest accuracy value is the C5.0 algorithm model optimized by PSO and Boruta with an accuracy value of 96.82%. Meanwhile, the model has the lowest accuracy is the Gradient Boosted Tree (GBT) without optimization or feature selection with an accuracy value of 79.91%. The best error rate value is the GBT + PSO + Boruta model with an error rate of +/- 3.21%. Meanwhile, the largest error rate is owned by the GBT model without optimization and feature selection with an error rate of +/- 24.84%.

Feature selection with Boruta and optimization with PSO is proven to increase the accuracy of the GBT classification by 13.66% from 79.91% to 93.57% and reduce the error rate by 21.63% from 24.84% to 3.21%. In addition, it can also increase the recall value, which initially had a recall value of 15.35% to properly detect fraud become 50%, there was an increase of 34.65% and also increase to detect whether it was not a fraud from 89.01% to 99.48% there was an increase of 10.47%. The value of precision in carrying out detection, there was also an improvement in the value of precision in predicting true fraud, which increased by 76.86% from 16% to 92.86%, for correct prediction of not fraud, there was an increase of 5.06% by 88.54% to 93.60%. From all the test of the GBT classification model that have been carried out, it can be concluded that the implementation of feature selection with Boruta and optimization with PSO can improve the performance of the classification model.

Performing feature selection with Boruta and optimization with PSO, it is proven that it can increase the accuracy rate of C5.0 classification by 5.50% from 91.32% to 96.82% and reduce the error rate by 0.45% from 4.77% to 4.32%. In addition, it can also increase the recall value, which initially had a recall value of 49.97%, to properly detect fraud to 84.62%, there was an increase of 34.65% and to detect whether it was not fraud from 98.95% to 98.43% there was a decrease of 0.52%. Meanwhile, for the value of precision in carrying out detection, there was also an improvement in the value of precision in predicting true fraud, which increased by 4.67% from 83.33% to 88%, for correct prediction of not fraud, there was an increase of 5.06% by 92.20% to 97.92%. From all the trials of the C5.0 classification model that have been carried out, it can be concluded that the implementation of feature selection using Boruta and optimization using PSO as a whole can improve the performance of the classification model.

4. CONCLUSION

Fraud in pawn transactions can be detected same as transactions on credit cards. Detecting fraud on pawn transactions will greatly assist to reducing company losses. The optimization method with Particle Swarm Optimization (PSO) and feature selection with Boruta can be implemented and is proven to increase the accuracy of the various classifications. The highest accuracy value is obtained through the C5.0 + PSO + Boruta model with an accuracy of 96.82% and the highest error rate value is obtained through the GBT + PSO + Boruta model with an error rate of +/- 3.21%.

REFRENCES:

[1] D. H. Murdock and D. H. Murdock, "Association of Certified Fraud Examiners (ACFE)," *Audit. Essentials*, pp. 7–10, 2018, doi: 10.1201/9781315178141-3.

[2] M. Jans, N. Lybaert, and K. Vanhoof, "Internal fraud risk reduction: Results of a data mining case study," *Int. J. Account. Inf. Syst.*, vol. 11, no. 1, pp. 17–41, 2010, doi: 10.1016/j.accinf.2009.12.004.

[3] N. Sagala and H. Tampubolon, "Komparasi Kinerja Algoritma Data Mining pada Dataset Konsumsi Alkohol Siswa," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 4, no. 2, p. 98, 2018, doi: 10.23917/khif.v4i2.7061.

[4] G. M. Zack and ACFE, "Using Data Analytics to Detect Fraud," pp. 1–16, 2015.

[5] M. (2018) Maulidiastuti, T., & Yusuf, "Analisis peran akuntansi forensik, data mining, continuous auditing, terhadap pendeteksian fraud serta dampaknya pada pencegahan fraud," vol. 3, no. 2, pp. 104–121, 2018.

[6] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Stat. Sci.*, vol. 17, no. 3, pp. 235–255, 2002, doi: 10.1214/ss/1042727940.

[7] Yazid and A. Fiananta, "Mendeteksi

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

kecurangan pada transaksi kartu kredit untuk verifikasi transaksi menggunakan metode svm," *Indones. J. Appl. Informatics*, vol. 1, no. 2, pp. 61–66, 2017.

[8] V. Vaishali, "Fraud Detection in Credit Card by Clustering Approach," *Int. J. Comput. Appl.*, vol. 98, no. 3, pp. 29–32, 2014, doi: 10.5120/17164-7225.

[9] H. Sharma and V. Ravi, "Detection of Financial Statement Fraud Using Decision Tree Classifiers," *Work. Pap.*, 2013.

[10] R. Pandya and J. Pandya, "C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *Int. J. Comput. Appl.*, vol. 117, no. 16, pp. 18–21, 2015, doi: 10.5120/20639-3318.

[11] E. Firasari *et al.*, "Kombinasi K-Nn Dan Gradient Boosted Trees Untuk Klasifikasi Combinations of K-Nn and Gradient Boosted Trees for," vol. 7, no. 6, pp. 1231–1236, 2020, doi: 10.25126/jtiik.202073087.

[12] A. Bisri and R. Rachmatika, "Integrasi Gradient Boosted Trees dengan SMOTE dan Bagging untuk Deteksi Kelulusan Mahasiswa," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 309, 2019, doi: 10.22146/jnteti.v8i4.529.

[13] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *Telkomnika (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, pp. 815–821, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14785.

[14] M. Alamgir Sarder, M. Maniruzzaman, and B. Ahammed, "Feature Selection and Classification of Leukemia Cancer Using Machine Learning Techniques," *Mach. Learn. Res.*, vol. 5, no. 2, p. 18, 2020, doi: 10.11648/j.mlr.20200502.11.

[15] B. Venkatesh and J. Anuradha, "A review of Feature Selection and its methods," *Cybern. Inf. Technol.*, vol. 19, no. 1, pp. 3–26, 2019, doi: 10.2478/CAIT-2019-0001.

[16] Y. L. Wu, C. Y. Tang, M. K. Hor, and P. F. Wu, "Feature selection using genetic algorithm and cluster validation," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2727–2732, 2011, doi: 10.1016/j.eswa.2010.08.062.

[17] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010, doi: 10.18637/jss.v036.i11.

[18] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta - A system for feature selection," *Fundam. Informaticae*, vol. 101, no. 4, pp. 271–285, 2010, doi: 10.3233/FI-2010-288.

[19] R. E. Schapire, "The Strength of Weak Learnability," *Mach. Learn.*, vol. 5, no. 2, pp. 197–227, 1990, doi: 10.1023/A:1022648800760.

[20] R. Barmaki, "Multimodal assessment of teaching behavior in immersive rehearsal environment - TeachLivE^{TM}," *ICMI 2015 - Proc. 2015 ACM Int. Conf. Multimodal Interact.*, vol. 139, pp. 651–655, 2015, doi: 10.1145/2818346.2823306.

[21] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, no. 4, pp. 367–378, 2002, doi: 10.1016/S0167-9473(01)00065-2.

[22] J. Han, M. Kamber, and J. Pei, "Data Mining Techniques, Third Edition," p. 847, 2011.

[23] A. Gopal, M. M. Sultani, and J. C. Bansal, "On Stability Analysis of Particle Swarm Optimization Algorithm," *Arab. J. Sci. Eng.*, vol. 45, no. 4, pp. 2385–2394, 2020, doi: 10.1007/s13369-019-03991-8.

[24] C. Grosan, A. Abraham, and M. Chis, "Swarm intelligence in data mining," *Stud. Comput. Intell.*, vol. 34, no. 2006, pp. 1–20, 2006, doi: 10.1007/978-3-540-34956-3_1.

[25] P. . Maya Gopal and R. Bhargavi, "Feature Selection For Yield Prediction Using Boruta Algorithm," *Int. J. Pure Appl. Math.*, vol. 118, no. 22, pp. 139–144, 2018.