SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Performance Comparison of Data Mining Algorithm to Predict Approval of Credit Card

Ipin Sugiyarto
STMIK Nusa Mandiri
Jakarta, Indonesia
ipin.sugiyarto@gmail.com

Bibit Sudarsono
Universitas Bina Sarana Informatika
Sukabumi, Indonesia
bibit.bbs@bsi.ac.id

Umi Faddillah
Universitas Bina Sarana Informatika
Jakarta, Indonesia
umi.umf@bsi.ac.id

Abstract**—** *Credit analysis needs to identify and assess the factors that can affect customers in returning credit. Accurate measurement and good management ability in dealing with credit risk is an effort to save the economic operations unit and be beneficial for a stable and healthy financial system. Data mining prediction techniques are used to determine credit risk. Using the Cross-Industry Standard Process for Data Mining (CRISP-DM) method which consists of several stages, namely Business Understanding (dataset), Data Processing (Feature Selection Principle Component Analysis & Dimension Reduce), Algorithm Models (Neural Network + Particle Swarm Optimize, Support Vector Machine, Logistic Regression), Evaluation (Validation and Accuracy). This study has tested the model using a neural network using the Principle Component Analysis (PCA) selection feature and optimized with the Particle Swarm Optimize (PSO) algorithm to predict credit card approval. Several experiments were conducted to see the best results. The results of this study prove that the use of a single Neural Network method produces an accuracy of 80.33%. whereas the use of PCA + Neural Network + PSO hybrid method has been proven to increase accuracy to 82.67%. Likewise, the AUC NN value of 0.706 increased to 0.749 when the Neural Network was optimized using PSO and used feature selection.*

Keywords— *Credit Analysis, Risk, Neural Network, PCA Feature Selection, Particle Swarm Optimize.*

I. INTRODUCTION

Failure to identify credit risk causes loss of income and extends the risk of bad credit to be a threat to profitability. Errors in credit analysis lead to credit risks, such as loss of customers, uncertainty about loan repayments, and even the inability of customers to repay loans according to (Zurada & Kunene, 2011).

Data mining classification techniques can be used to determine credit risk (Lee, Chiu, Lu, & Chen, 2002). Data mining is a computational process that expresses patterns of data collected using methods such as artificial intelligence, machine learning, statistics and others (Lee et al., 2002). The method used in data mining was investigated in two groups as predictive and descriptive. In the predictive method, the model is created using a dataset whose results are known. For example in a bank, the nature of customers who repay their loans can be revealed and models can be made using previous data sets on client funding. After that the model can be used on new ones. Customers to determine the possibility of paying their credit back. In the descriptive method, a relationship can be searched between two data sets, for example the shopping habits of two different cultures can be investigated for similarity.

The selection of data mining classification algorithms to determine credit risk that occurs in lending transactions is based on several previous studies. Data mining classification techniques in determining credit quality improvement and credit risk reduction using Logistic Regression, Discriminant Analysis, K-Nearest Neighboor, TAN Technique, Naives Bayes, Decission Tree (C.45),

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Associative Classification, Neural Network and Support Vector Machine.

From several studies conducted that artificial neural networks can do data processing well to predict the default of the credit card client and show the highest accuracy (Pasha, Fatima, Dogar, & Shahzad, 2017).

Artificial neural networks approach in terms of the average level of correct classification and can estimate the cost of incorrect classification (Akkoç, 2012). Prediction of artificial neural networks will be even better if it is optimized by several methods such as the hybird model to get help from discriminant analyzes in providing statistical support, reducing the number of input variables and providing a better initial solution (Lee et al., 2002). The study indeed supports the hypothesis that the two-stage hybrid credit scoring approach used in previous studies will have better credit rating accuracy and better convergence characteristics for the neural network model designed. In a comparative analysis of previous studies using cross validation, confusion matrix, ROC curve and T-Test for several data mining classification algorithms it can be concluded that the Linear Regression algorithm has the highest accuracy (Menarianti, 2015).

Based on the results of previous studies, it can be concluded that the Neural Network, Naives Bayes and Logistic Regression methods have a good level of accuracy so that for future studies, we will try to compare the Logistic Regression algorithm, Support Vector Machine and Neural Network based on feature selection PCA and optimize PSO parameters for know the results of the best level of accuracy in improving credit quality and reducing credit risk.

II. LITERATURE REVIEW

2.1. Artificial Neural Network

Artificial Neural Networks have a working principle similar to the workings of nerve cells, which have a part that functions to process information. This section will be connected to similar parts through a network that has functions as a series of inputs (outputs) and outputs (outputs). The part that has input and output functions similar to the Neurit and Dendrite functions found in human nerve cells. So it can be said that ANN is a distributed information processing structure and works in parallel, consisting of processing elements that are connected with connections.

Each processor element has a single fan connection to the desired number of guaranteed connections. The output of the processing element can be the result of data processing which can be a mathematical equation, function or method. The process that occurs in each processor element must be carried out locally, that is, the output depends only on the input value when the data is obtained locally, that is, the output depends only on the input value at the time obtained through the connection and the value stored in local memory. ANN is a processor that is distributed in parallel and has the ability to store knowledge gained from experience and remains available for use.

Source: (Soner Akkoc, 2012)

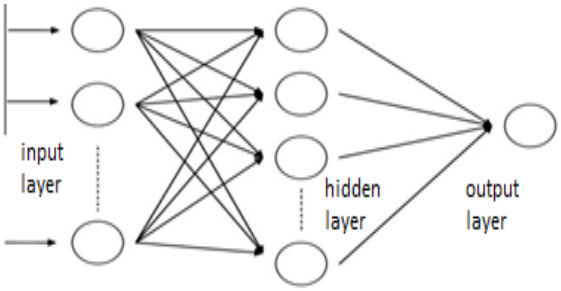


Figure 1. Artificial Neural Network Model.

Neural network is a category of soft computing that adopts the workings of the human brain that is able to provide stimulation, stimulation, process, and provide output.

Output obtained from neural networks in the form of variations of stimulation and processes that occur in the human brain. The ability of humans to process the information obtained is a result of the complexity of processes in the human brain. Neural networks send information to each other and produce information that can be represented as a target or destination. Neural network algorithms can be applied for classification, pattern recognition, function estimation, control, and datamining to look for separate data forms or commonly referred to as knowledge discovery.

Characteristics of artificial neural networks can be seen from the pattern of relationships between neurons, the method of determining the weight of each connection and its activation function. In general, artificial neural network models consist of, the input layer, the hidden layer, the Output Layer and the activation function.

2.2 Support Vector Machine

Support Vector Machine (SVM) is a machine learning method that works based on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyper plane separating two classes in

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

the input space. The best hyper plane is an airplane that is located halfway between two sets of objects from two classes. The best hyperplane separator between the two classes can be found by measuring the hyperplane's margin and finding its maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called a support vector.

2.3. Logistic Regression

Regression models are a method commonly used to predict the relationship of goods between more than two groups. The constraint in this case the model is the nature of the target that must be of binary value. Analogous with multiple regression, which is strong with the techniques it offers. The logistic regression model also provides ANOVA for unbroken answers. The independent function variables y1, y2, y3 ... yn with binary responses in nature, are part of an exponential family with "log (∏1 / (1-∏1), ... log (∏n / (1-∏n) "as canonical parameters of scientific conditions the correlation between canonical parameters and vector vector descriptive x is expressed as the log equation (∏i (1-∏i) = xβi.

The linear membership between the descriptor and the probability logarithm creates a non-linear membership between the probability of y equals 1 and the explanatory variable in the vector. ∏iexp (xβi / (1 + exp (xβi)) To deal with classification problems, logistic regression is the right algorithm, however, the calculated results can be displayed as probabilities

2.4. Principle Components Analysis

Personal Components Analysis (PCA) or principal component analysis is a technique used to simplify data, by changing data linearly to form new coordinates with maximum variance. Principal component analysis can be used to reduce data dimensions without significantly reducing data characteristics.

The main purpose of principal component analysis is to reduce the dimensions of interrelated variables and the number of variables sufficient so that it is easier to interpret the data. The method used is to determine the main components by transferring orthogonal variations or forming a linear combination of Y = A 'X. From this process, several components of total data diversity will be selected.

2.5. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is an iteration algorithm with random search. Each individual solution is a particle with no volume and no mass in the search space. In each iteration, the particle updates itself according to two pieces of extreme information. The first part is the best optimal p-individual value found by individual particles and the other is the global optimal value found by the population. In this way, the particle population can approach individual particles that have good adaptation values and finally find the optimal solution.

Algortima PSO can be explained as follows: for example in the search space for D-dimensional targets, N particles forming the particle population, the position and velocity of particle I are expressed as: Xj = (Xj1, Xj2 ... XjD) and Vj = (Vj1, Vj2. ..., VJD). Optimal individual position of individual particles, namely: Pj = (Pj1, Pj2, ..., PjD) (pbest). The global optimal position of all particles is Pg = (Pg1, Pg2, ..., PgD) (gbest). Each particle updates its speed and position iteratively according to the following formula:

$$V_{jd}^{t+1} = \omega V_{jd}^{t} + C_1 r_1 + (P_{jd}^{t} - X_{jd}^{t}) + C_2 r_2 (P_{gd}^{t} - X_{jd}^{t})$$

$$X_{jd}^{t+1} = X_{jd}^{t} + V_{jd}^{t+1}$$

The following is a description of the formula; i = (1,2, ... N), d = (1,2, ... D); ω is the weight of inertia; C1 and C2 are constant correlations; r1 and r2 are random numbers that are evenly distributed between 0 and 1; Vjdt is the current particle velocity i; Vjdt + 1 updated particle speed i; Xjdt is the current position of particle i; Xjdt + 1 is the latest position of particle i.

III. Proposed Method

In this study several stages of the study were carried out, as follows:

Data collection

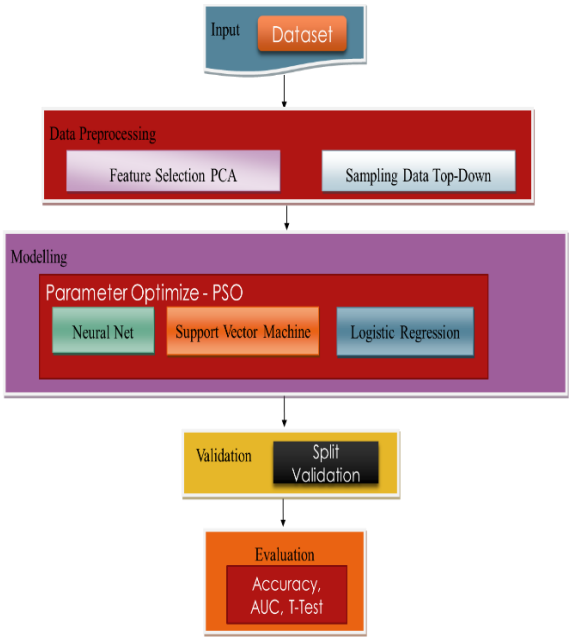
The data used in this study were obtained from the UCI dataset repository about credit cards in Taiwan banks consisting of 23 predictor attributes, namely, bales, sex, education, marriage, age, pay_0-pay_6, bill_amt1-bill_amt6 and pay_amt1-pay_amt6 and 1 label with numbers 1 and 0.

Initial Data Processing

In the initial data processing, there are 3 stages: (1) top-down data sampling, (2) data transformation and (3) feature selection with PCA. For top-down data sampling, the dataset is taken from 3,000 of the 30,000 rows of data from the top row of the dataset totaling 1,500 rows and the lowest row is 1,500 rows. Data is taken in a 50:50 balance based on data that has been labeled 1 and 0.

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The next step is to transform the data by converting numerical data into nominal form. The data that has been transformed is already in the form of simplified data with a range of the lowest 0 and the top 1. Feature selectin PCA in this stage is done by selecting the data based on the PCA algorithm (Principle Component Alaysis) namely selecting the most influential data and eliminating the data included in the data criteria take effect. The PCA process itself reduces the number of predictor attributes so that a new predictor attribute is formed from the results of the PCA algorithm process with the number of previous 23 attributes being 15 new predictor attributes.



Sources: (Sugiyarto & Gata, 2018)

Figure 2. Research Method

This stage is a research method design consisting of 5 parts namely, Input dataset, data preprocessing with PCA feature selection and top-down data sampling, neural net algorithm modeling, support vector machines and logistic regression, split validation and evaluation in the form of accuracy values.

Model Making

The data mining method or algorithm used in this study uses the CRISP-DM method approach, namely the neural network algorithm, support vector machine and logistic regression then compared using the PCA feature selection and optimize PSO parameters. Split Validation for data validation testing, evaluation in the form of accuracy and AUC values, deployment in the form of presentation of data processing in the form of user applications to calculate credit card customer approval calculations with outputs in the form of credit status approved and not approved. At the deployment stage the application development uses the C # program as a compiler program based on the neural network algorithm as the basis for calculations to predict customer credit card approvals. The modeling of this research is presented in Figure 3 below:

Sources: (Sugiyarto & Gata, 2018)

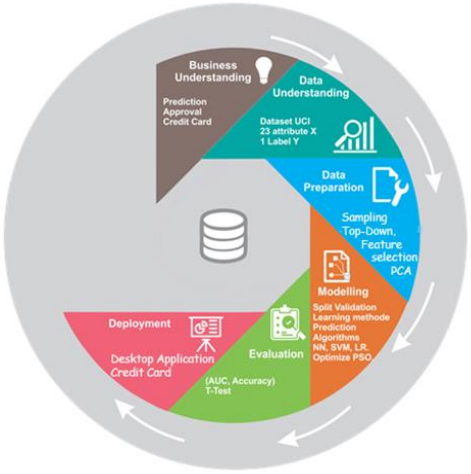
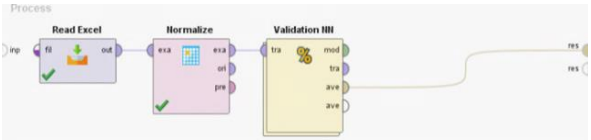


Figure 3. Proposed CRISP-DM Modified Method

This research will use several comparisons of data mining methods with the CRISP-DM model approach on neural network algorithms, support vector machines and logistic regression to conduct data experiments. The tool used is rapid miner. Data processing steps using rapid miner as shown below:

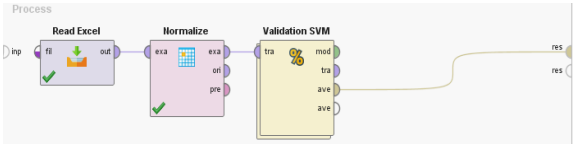


Sources: (Sugiyarto & Gata, 2018)

Figure 4. Process Model Neural Network Method.

This model contains read excel operator as source data, normalize operator for data normalization and operator validation for processing results with Neural Net method.

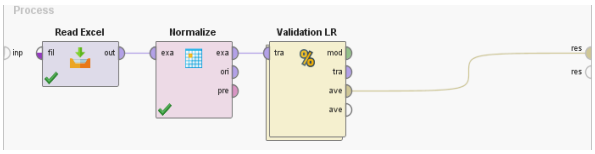
SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X



Sources: (Sugiyarto & Gata, 2018)

Figure 5. Process Model Support Vector Method.

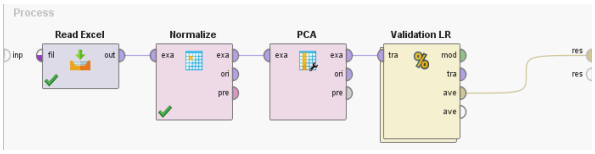
This model contains read excel operator as source data, normalize operator for data normalization and operator validation for processing results with SVM method.



Sources: (Sugiyarto & Gata, 2018)

Figure 6. Process Model Logistic Regression Method.

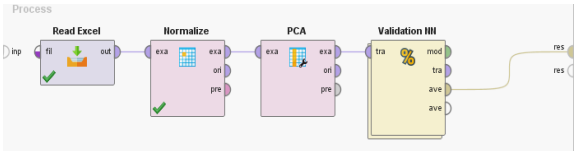
This model contains read excel operator as source data, normalize operator for data normalization and operator validation for processing results with Logistic Regression method.



Sources: (Sugiyarto & Gata, 2018)

Figure 7. Process Model Logistic Regression + PCA Method.

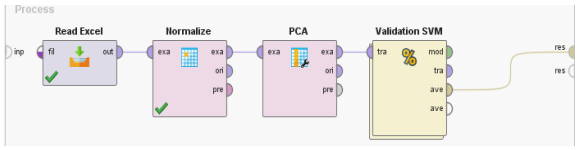
This model includes read excel operators as source data, normalize operators for data normalization, PCA operators for attribute selection and validation operators for processing results with logistic regression methods.



Sources: (Sugiyarto & Gata, 2018)

Figure 8. Process Model Method of NN + PCA.

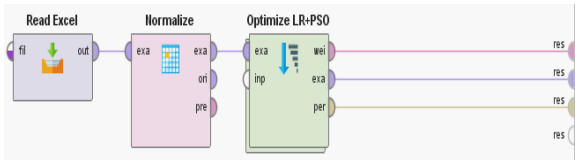
This model includes read excel operators as source data, normalize operators for data normalization, PCA operators for attribute selection and validation operators for processing results with neural network methods.



Sources: (Sugiyarto & Gata, 2018)

Figure 9. Process Model SVM + PCA Method.

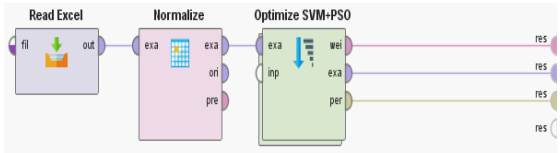
This model includes read excel operators as source data, normalize operators for data normalization, PCA operators for attribute selection and validation operators for processing results with support vector machine methods.



Sources: (Sugiyarto & Gata, 2018)

Figure 10. Process Model Logistic Regression + PSO Method.

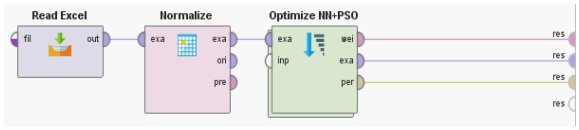
This model includes read excel operators as source data, normalized operators for data normalization, and operators for processing logistic regression methods that are optimized with PSO.



Sources: (Sugiyarto & Gata, 2018)

Figure 11. Process Model SVM + PSO method.

This model includes read excel operators as data sources, normalize operators for data normalization, and operators for the process of supporting vector machine methods that are optimized with PSO.



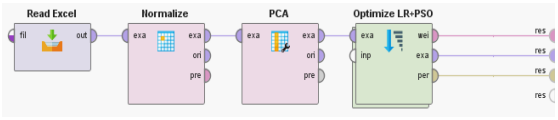
Sources: (Sugiyarto & Gata, 2018)

Figure 12. Process Model NN + PSO Method.

This model includes read excel operators as data sources, normalize operators for data normalization,

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

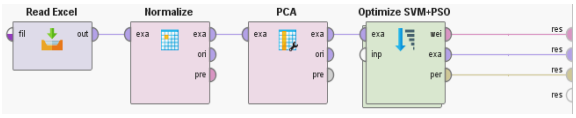
and operators for the process of neural network optimization results with PSO.



Sources: (Sugiyarto & Gata, 2018)

Figure 13. Process Model LR + PCA + PSO Method

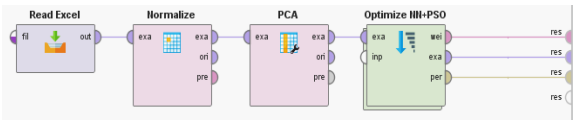
This model contains read excel operators as source data, normalize operators for data normalization, PCA operators for attribute selection and operators for the process of logistic regression method results that are optimized with PSO.



Sources: (Sugiyarto & Gata, 2018)

Figure 14. Process Model SVM + PCA + PSO method.

This model contains read excel operators as source data, normalize operators for data normalization, PCA operators for attribute selection and operators for the process of supporting vector machine methods that are optimized with PSO.



Sources: (Sugiyarto & Gata, 2018)

Figure 15. Process Model Method of NN + PCA + PSO.

This model includes read excel operators as data sources, normalize operators for data normalization, PCA operators for attribute selection and operators for the process of neural network optimization results with PSO.

IV. Result and Discussion

This stage is the result of processing in the validation process to find out the predictive comparative accuracy of the 3 methods used, namely Neural Network Algorithms, Support Vector Machines and Logistic Regression and the addition of PCA feature selection methods and Optimize parameters with PSO. The following are the results of the presentation of data that has been processed in the form of accuracy values and accuracy comparison charts in related research in the previous year and current research.

Table 1. Comparative Results of Data Mining Algorithms in Credit Card Approval Research.

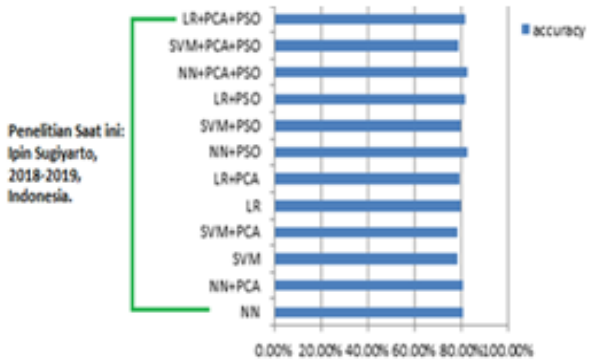
Research	Algorithm	Accuracy
Ipin Sugiyarto, Indonesia, (2018)	Neural Network	0.803
	Support Vector Machine	0.782
	Logistic Regression	0.794
	Neural Network + Principle Component Analysis	0.806
	Support Vector Machine + Principle Component Analysis	0.782
	Logistic Regression + Principle Component Analysis	0.793
	Neural Network + Particle Swarm Optimization	0.824
	Support Vector Machine +Particle Swarm Optimization	0.794
	Logistic Regression + Particle Swarm Optimization	0.813
	Neural Network + Principle Component Analysis + Particle Swarm Optimization	0.826
	Support Vector Machine + Principle Component Analysis + Particle Swarm Optimization	0.787
	Logistic Regression + Principle	0.816

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

	Component Analysis + Particle Swarm Optimization

Sources: (Sugiyarto & Gata, 2018)

Based on table 1, the prediction results are obtained with the best algorithm based on Neural Net + PCA + PSO with an accuracy of 82.60%.

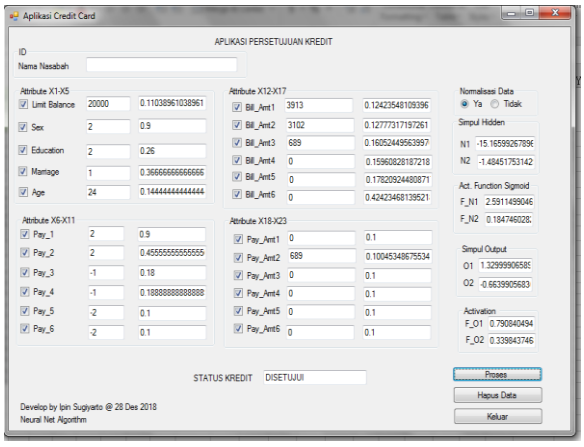


Sources: (Sugiyarto & Gata, 2018)

Figure 16. Comparison of Research Accuracy Charts on Approval Credit Cards.

Deployment

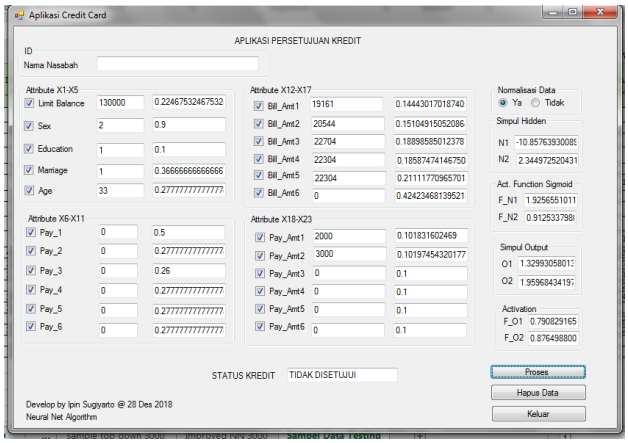
This stage is testing the application for credit card approval calculations based on neural network algorithms. There are 23 predictive attribute inputs in the text box and one label for creditor status. Process button as calculation execution, delete data button as clear data and exit button to exit the application.



Sources: (Sugiyarto & Gata, 2018)

Figure 17. Application Results of the Approved Credit Calculation Process.

This stage is testing the application model for credit card approval calculation with 23 input box attributes and credit card approval label results with the approved status based on the algorithm used, namely the neural network model with an output activation value of 0.790840494 and an activation output 2 of 0.339843746.



Sources: (Sugiyarto & Gata, 2018)

Figure 18. Application Results of the Credit Calculation Process Not Approved.

This stage is testing the application model of credit card approval calculation with 23 input box attributes and credit card approval label results with an unapproved status based on the algorithm used, namely the neural network model with an activation value of output 1 of 0.790829165 and activation of output 2 of 0.876498800.

V. CONCLUSION AND SUGGESTION

Based on the research conducted by the author, testing the model using a neural network using PCA feature selection and optimized with the Particle Swarm Optimize (PSO) algorithm to predict credit card approval. Several experiments were conducted to see the best results. The results of this study prove the use of the Neural Network method produces the best accuracy with an accuracy value of 80.33%, while the use of feature selection and parameter optimize with the Neural Network + PCA + PSO method has been proven to increase accuracy to 82.67%.

The model that has been formed can then be developed and can be implemented into an application. So that it can help and make it easy for stakeholders to make a decision to predict credit approval for sacrificed customers. Suggestions for research this is that parameter optimization on the Neural Network uses the Particle Swarm Optimize can improve the accuracy and feature selection PCA

Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X

is used for reduce the dimensions of the data so that the processing is more efficient than that data many and varied. But there are several factors that can be tried for research Furthermore, to find methods that both have the best results, that is to develop a credit approval prediction decision decision information system managerial information systems in banking and conducting further research has an influence on the predicted results of creditor customers.

VI. REFERENCES

Akkoç, S. (2012). An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data. *European Journal of Operational Research*, *222*(1), 168–178. https://doi.org/10.1016/j.ejor.2012.04.009

Budiharto, Widodo. (2016). Machine Learning & Computational Intelligence. ANDI, Yogyakarta.

Inspiring. (2018). Understanding Debit and Credit in Financial Statements. Retrieved from https://www.inspiring.id/peng understanding-debit-and-credit/

I. C. Yeh & C. Hui Lien. (2009). The comparison of data mining technique for predictive accuracy of probability of default of credit card client. Expert Syst. Appl., Vol.36, No.2 Part 1, pp. 2473-2480.

Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

Menarianti, I. (2015). Data mining classification in determining lending for cooperative customers. Journal of Scientific Science, 1 (1), 1–10.

Pasha, M., Fatima, M., Dogar, A. M., & Shahzad, F. (2017). Performance Comparison of Data Mining Algorithms for the Predictive Accuracy of Credit Card Defaulters. *International Journal of Computer Science and Network Security*, *17*(3), 178–183.

Pohan. B. Achmad & Sensuse I. Dana. (2014). Optimasi artificial neural network menggunakan genetic algorithm untuk prediksi uji coba marshal pada campuran aspal beton. Journal Ilmiah Prodi. Magister Ilmu Komputer. STMIK Nusa Mandiri.

Ronald L., Iman & W. J. Conover. (2012). A measure of top-down correlation. Technimetrics, Vol.29, No.3

S. F. Putra, R. Pradina & I. Hafidz. (2016). Feature selection pada dataset faktor kesiapan bencana pada provinsi di Indonesia menggunakan metode PCA (Princial Component Analysis). J. Tek. Its, vol.5, No.2, pp. 5-9.

Sugiyarto, I & Gata, W. (2018). thesis book

S. Umair. (2014, 1 Nov). A comparative study of data mining process models (KDD, CRIPS-DM and SEMMA). IJISR, Vol.12, No.1, pp. 217-222.

T. S. Lee, C. C. Chiu, C. J. Lu & I. F. Chen. (2002). Credit scoring using the hybird neural discriminant tehcnique. Expert Syst. Appl. Vol.23, No.3, pp. 245-254.

Vankatesh, A. & Jacob, G. S. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers.

Yuxi. Gao. (2018). An improved hybird group intelligent algorithm based on artificial bee colony and particle swarm optimatizion. International Conf. On Virtual Reality and Intelligent System.

Y. B. Wah & I. R. Ibrahim. Using data mining predictive models to classify credit card applicant. pp. 394-398.

Zurada, C. & Kunene, K. (2011). Comparison of the Performance of Computational Intelligence Methods for Loan Granting Decisions. Proceeding of the 44th Hawaii International Conference on System Science.

SinkrOn
JURNAL & PENELITIAN TEKNIK INFORMATIKA
Journal Publications & Informatics Engineering Research
Volume 4, Number 1, October 2019
DOI : https://doi.org/10.33395/sinkron.v4i1.10181
e-ISSN : 2541-2019
p-ISSN : 2541-044X