e-ISSN: 2614-8404
p-ISSN:2776-3234

Analysis of the Use of Particle Swarm Optimization on Naïve Bayes for Classification of Credit Bank Applications

Yoga Religia[1]*), Gatot Tri Pranoto[2], I Made Suwancita [3]
[1]Informatics Engineering Study Program, Faculty of Engineering, Pelita Bangsa University
[2]Information Systems Study Program, Faculty of Creative Industries and Telematics, Trilogy University
[3]Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Budi Luhur
email: [1]yoga.religia@pelitabangsa.ac.id , [2]gatot.tp@gmail.com, [3]isuwancita@gmail.com

Abstract –The selection of prospective customers who apply for credit in the banking world is a very important thing to be considered by the marketing department in order to avoid non-performing loans. The website www.kaggle.com currently provides South German Credit data in the form of supervised learning data. The use of data mining techniques makes it possible to find hidden patterns contained in large data sets, one of which is using classification modeling. This study aims to compare the classification of South German Credit data using the Naïve Bayes algorithm and compare the classification of South German Credit data using the Naïve Bayes algorithm with particle swarm optimization (PSO). The test was carried out using a confusion matrix to determine the accuracy, precision and recall values of the research model. Based on the test, it is known that PSO is able to increase the accuracy and recall of Nave Bayes, but PSO has not been able to increase the precision value of Nave Bayes. The test results show that PSO optimization gives Naïve Bayes an increase in the value of accuracy by 0.46%, and gives Naïve Bayes an increase in recall value by 3.02%.

Keywords – Data Mining, Classification, Nave Bayes, PSO Optimization, bank credit acceptance.

I. INTRODUCTION

The banking marketing department needs to select prospective customers to find out which customers can be given credit financing by considering various factors. Credit financing is the provision of funds by the bank to the customer based on a loan agreement that requires the customer to repay the loan within a certain period of time.[1]. Therefore, the selection of prospective customers is needed so that a marketing bank is able to keep their customers from experiencing non-performing loans[2]. One way that can be used to reduce the possibility of non-performing loans is to utilize data mining techniques, so that it is possible to mine information from pre-existing credit application data sets.[3].

In general, data mining is divided into two categories, namely predictive and descriptive. Predictive methods can be done with a classification model. The use of the classification model can be done by changing the data record into a set of the same class[4]. Nowsitewww.kaggle.com has provided the South German Credit data set consisting of 21 attributes with 800 instances of credit application and there is no missing value, so that it can be used to build a creditworthiness classification model [5]. The label attribute contained in the South German Credit data is the "Credit" attribute with 600 instances with the description "accepted" and 200 instances with the description "rejected", thus making South German Credit data including imbalance data.

It takes a good algorithm to create an optimal classification model. One algorithm that has been widely used for classification modeling with good performance is the Naïve Bayes algorithm. Several previous studies have stated that the Naïve Bayes algorithm is able to provide better classification performance when compared to other classification algorithms[6] [7] [8]. The Naïve Bayes algorithm can also be used on imbalanced data[9] [10], so it is suitable for classifying South German Credit data. Currently, independent assumptions are rarely discussed in the Naïve Bayes classification. One way to try independent assumptions in the Naïve Bayes algorithm is by attribute weighting[11]. It is necessary to propose an attribute weighting method to reduce the independent assumption[12]. One of the weighting optimization methods that can be used is to use particle swarm optimization (PSO).[13].

PSO has significant advantages in handling non-linear fittings and multi-input parameters [14]. PSO does not have evolution operators such as crossover and mutation, so it is easy to implement and there are very few parameters to adjust[15]. Based on several previous studies, it was stated that the combination of PSO and Naive Bayes was able to provide better imbalance data classification performance results than using Naive Bayes alone.[16] [17], even PSO is able to increase the accuracy of Naive Bayes by more than 10% [18].

Based on previous research showing that both PSO is able to improve classification performance on Nave

e-ISSN: 2614-8404
p-ISSN:2776-3234

Bayes, it is necessary to do further testing regarding the use of PSO optimas on Naïve Bayes for South German Credit data classification. This study will compare the Naïve Bayes classification on South German Credit data with and without PSO optimization.

A. Bank Credit Financing

The word credit comes from the Italian word credere which means trust. The trust referred to here is the trust of the creditor that the debtor will return the loan and the interest in accordance with the agreement that has been agreed by both parties.[19]. The implementation of the granting of credit is carried out through several steps, namely credit application, credit application examination, credit analysis, credit approval, credit realization and the last is credit supervision.[20]. In general, most of the bank's wealth is in the form of credit which is a source of bank income which is commonly referred to as productive assets. Management must use the precautionary principle so that loans are in the current category. Often there are several customers whose interest and principal payments are not smooth which makes them fall into the category of non-performing loans (NPL). The higher the NPL indicates the greater the potential loss, so the bank must be able to reduce its lending[21].

B. Data Mining

Data mining has been around since the 1990s as an effective way to extract previously unknown patterns and information from a data set [22]. Data mining is one of the most important fields in research that aims to obtain information from data sets. Data mining is the process of extracting meaningful information and structures in complex data sets[23]. In its implementation, data mining can use various parameters to examine data including association, classification and clustering. Data mining involves key steps which include problem definition, data exploration, data preparation, modeling, and evaluating and deployment[24].

Data mining techniques are used to find relationships between data to perform classifications that predict the values of several variables (classification), or to divide known data into groups that have similar characteristics (clustering). Using data mining techniques it is possible to search, analyze, and sort through large data sets to discover new patterns, trends, and relationships contained within them.[25].

C. Classification with Naïve Bayes

The Naïve Bayes algorithm is a supervised learning algorithm based on the Bayes Theorem with the assumption of independence between predictors. That is, features in a class do not depend on other features[26]. Naïve Bayes is widely used to solve classification problems in real-world applications, this is because it is easy to build and interpret data, and has good performance.[12]. The Naive Bayes classifier can also be used for continuous and categorical variables. It is based on the Bayes formula which is the probability of event A given proof of B which can be seen in the following equation[27]:

$$P(A, B) = P(A)P(B) \quad (1)$$

Through equation (1) and using the concept of Bayes' theorem, the final equation of the Naïve Bayes algorithm is obtained as follows:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (2)$$

Based on equation (2), it is known that A represents the class and B represents the instance. A is the dependent event which means the predicted variable and B is the previous event which means the predictor attribute. The last step of the Naive Bayes algorithm is to find the maximum probability that will be used as a predictor class.

D. Particle Swarm Optimization(PSO)

Particle swarm optimization or commonly abbreviated as PSO is an optimization technique whose concept is based on the behavior of a swarm of insects, such as ants, termites, or bees. [28]. The PSO approach is like a collection of particles that simultaneously explore the problem search space with the aim of finding the optimal global configuration[29]. PSO has proven to be very effective in solving various engineering problems and solving them very quickly[30].

The basic assumption behind the PSO algorithm is that birds find food in groups and not individually. This gives rise to the assumption that information is shared in flocking. The herd initially has a population of random solutions. Each potential solution is called a particle (agent), is assigned a random velocity and is flown through the problem space[30]. All particles have a memory and each particle keeps track of the previous best position (Pbest) and the corresponding match value. Flock has another value called Gbest, which is the best value of all Pbests. By using this concept, PSO can provide a technique for solving attribute selection quickly.

II. RESEARCH METHODOLOGY

A. Data used

This study uses secondary data in the form of a South German Credit data set taken from the sitewww.kaggle.com [5]. The number of data instances contained in the South German Credit data is 800 instances consisting of 21 attributes and there is no missing value, so there is no need for pre-processing data. Based on the existing 21 attributes, there is 1 label attribute contained in the South German Credit data, namely the "Credit" attribute. On the label there are 600 instances with the description "good" and 200 instances with the description "bad", so that the South German Credit data includes imbalance data.

South German Credit Data chosen because it is free from missing values, so there is no need for preprocessing data to be used in making classification models. As for more clearly about 21 attributes and 1 label from South German Credit data, it can be seen in Table 1.

e-ISSN: 2614-8404
p-ISSN:2776-3234

Table1. Data AttributeSouth German Credit

Attribute	Information
status	status of the debtor's checking account with the bank (categorical)
duration	credit duration in months (quantitative)
credit history	history of compliance with previous or concurrent credit contracts (categorical)
purpose	purpose for which the credit is needed (categorical)
amount	credit amount in DM (quantitative; result of monotonic transformation; actual data and type of trans...
savings	debtor's savings (categorical)
employment duration	duration of debtor's employment with current employer (ordinal; discretized quantitative)
installment rate	credit installments as a percentage of debtor's disposable income (ordinal; discretized quantitative...
personal status sex	combined information on sex and marital status; categorical; sex cannot be recovered from the variab...
other debtors	Is there another debtor or a guarantor for the credit? (categorical)
present residence	length of time (in years) the debtor lives in the present residence (ordinal; discretized quantitative...
property	the debtor's most valuable property, ie the highest possible code is used. Code 2 is used, if code...
age	age in years (quantitative)
other installment plans	installment plans from providers other than the credit-giving bank (categorical)
housing	type of housing the debtor lives in (categorical)
number credits	number of credits including the current one the debtor has (or had) at this bank (ordinal, discretiz...
job	quality of debtor's job (ordinal)
people liable	number of persons who financially depend on the debtor (ie, are entitled to maintenance) (binary,d...
telephone	Is there a telephone landline registered on the debtor's name? (binary; remember that the data are f...
foreign workers	Is the debtor a foreign worker? (binary)
credit risk	Has the credit contract been complied with (good) or not (bad) ? (binary)

B. Research Model

The classification model built in this study was carried out using South German Credit data. The label used is the attribute "Credit Risk" with the values "Good" and "Bad". As many as 77% of the instances in the South German Credit data are instances with the class label "good", while the rest are instances with the label "bad". Tests in this study were carried out 2 times which will later be analyzed the results obtained. The first test was carried out using Naïve Bayes with PSO optimization, while the second test was carried out using Nave Bayes without PSO optimization.

Spit validation used as a process of validating research data which aims to divide South German Credit data into training data and testing data. Split validation is used to divide South German Credit data into training and testing data with a comparison of 90% and testing data of 10%. The training data will be used for classification modeling using the Naïve Bayes algorithm. The resulting model is then used as an apply model for use in data testing. After the classification has been carried out, the performance of the formed classification model is measured in the form of accuracy, precision, and recall values.

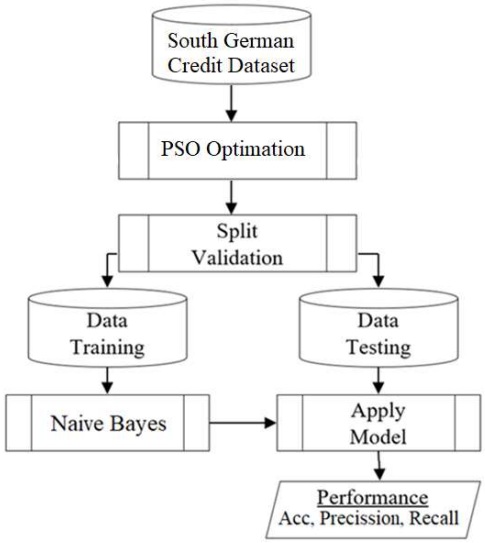

Figure 1. First Test: Nave Bayes Classification with PSO Optimization

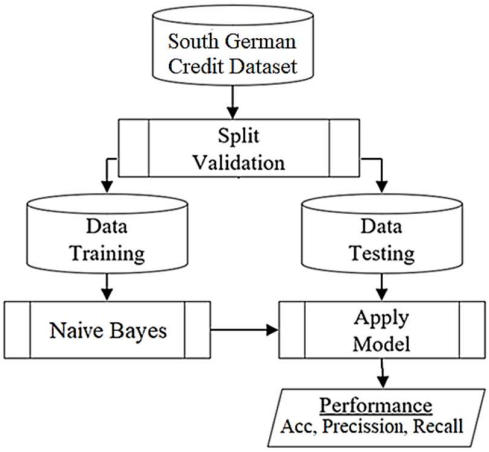

Figure 2. Second test: Nave Bayes Classification without PSO Optimization

Based on Figure 1 and Figure 2 it shows that the testing in this study was carried out 2 times, namely: (1) Classification of South German Credit data using Naïve Bayes with PSO optimization, (2) Classification of South German Credit data using Naïve Bayes without PSO optimization. The performance results of the two tests will be compared and then analyzed to obtain research findings.

III. RESULTS AND DISCUSSION

A. Testing Step

The testing tools in this study used RapidMiner version 5.0. The use of RapidMiner tools is done because RapidMiner can be used for rapid prototyping, and supports all steps of the data mining process[31]. The first step in making this research model is to call South German Credit data. The second step is to perform the multiply function to perform two tests at once, namely testing using PSO optimization and testing without using PSO optimization. The third step is to distribute the data into the split validation process. Validation process by dividing training data by 90% and testing data by 10%

e-ISSN: 2614-8404
p-ISSN:2776-3234

from South German Credit data. More clearly about the data calling and validation process in this study can be seen in Figure 3.

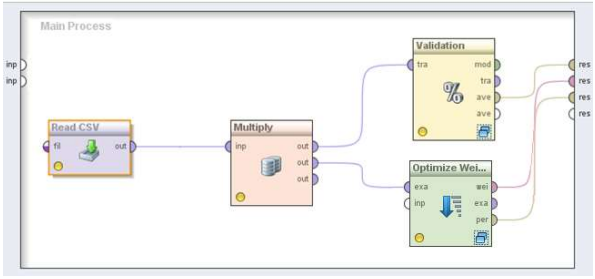

Figure 3. Data Calling and Validation Process

In each validation process in Figure 3, it contains a learning process using the Naïve Bayes algorithm which is then applied to the model to measure its accuracy, precision and recall performance. The learning process formed in this study can be seen in Figure 4.

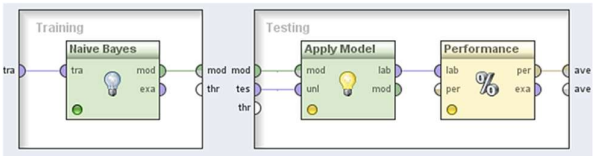

Figure 4. Learning Process and Apply Model

After the entire research model has been formed, the last step is to run the model that has been built in RapidMiner to see the results of its accuracy, precision and recall.

B. Test result

After 2 tests, the accuracy, precision, and recall values of the two models were obtained. More complete test results can be seen in Table 2.

Table2. Test result

No	Algorithm	Accuracy	Precision	Recall
1	Naive Bayes	85.43%	89.37%	85.80%
2	PSO + Naive Bayes	85.89%	88.23%	88.82%

Based on Table 2, it can be seen that PSO is able to increase the accuracy and recall of Nave Bayes, but PSO has not been able to increase the precision value of Nave Bayes. The test results show that with an accuracy of 85.89%, PSO optimization gives Naïve Bayes an increase in accuracy value of 0.46% and an increase in recall value of 3.02% for South German Credit data classification. However, it turns out that the use of PSO also reduces the precision value by 1.14% from the use of Naïve Bayes for South German Credit data classification. This is presumably because of the 20 attributes (non attribute labels) in the South German Credit data, it turns out that there are only 5 attributes that are weighted by PSO. This weighting result also explains why the increase in accuracy and recall provided by PSO is not too large.

In Table 3 it can be seen that there are only 5 attributes that are weighted by PSO. This shows that, based on PSO optimization, these 5 attributes are the most important to consider when classifying South German Credit data. The attributes are: credit history, savings, property, age, and job.

Table 3. PSO Weighting Results on DataSouth German Credit

Attribute	Weighting
status	0
duration	0
credit history	1
purpose	0
amount	0
savings	1
employment duration	0
installment rate	0
personal status sex	0
other debtors	0
present residence	0
property	1
age	1
other installment plans	0
housing	0
number credits	0
job	1
people liable	0
telephone	0
foreign workers	0

C. Results Discussion

Based on the results of the tests that have been carried out, it is known that the use of PSO optimization on Nave Bayes for the classification ofSouth German Credit data is able to improve the performance of Naïve Bayes in terms of accuracy and recall even though it is not too big. The small increase in performance given by PSO is thought to be because the attributes that are weighted by PSO are only 5 attributes out of 20 predictor attributes in South German Credit data. This makes the probability calculation process on Naïve Bayes become irrelevant. Even looking at the precision side, it turns out that the use of PSO optimization actually makes Naïve Bayes' precision performance decrease.

Although the optimization of PSO does not give maximum results, by using PSO it can be seen which attributes can be used as evaluation priorities to consider loan application approval. By looking at the attributes given the weighting by PSO, it can be used as a reference to consider these attributes as the main focus to avoid the risk of non-performing loans. The attributes that are given weighting by PSO are: credit history, savings, property, age, and job. This finding is expected to provide a practical contribution to the decision to provide credit by marketing parties in order to minimize the occurrence of non-performing loans.

IV. CONCLUSION

This research has tested the use of Naïve Bayes algorithm and PSO optimization to classify South German Credit data. Based on the tests that have been carried out, some results are shown as follows:

1. PSO optimization is able to improve the performance of Naïve Bayes in classifying South German Credit data in terms of accuracy and recall, although it does not have a big impact.
2. This study found that the use of PSO has not been able to improve the performance of Naïve Bayes in classifying South German Credit data in terms of precision, even the precision performance of Nave Bayes has decreased in value.

e-ISSN: 2614-8404
p-ISSN:2776-3234

3. By using PSO optimization, it can be seen that there are 5 attributes that need to be the main consideration in lending in the banking world, namely credit history, savings, property, age, and job.

This study has not been able to provide a good enough accuracy value for the classification of South German Credit data, so further research is needed to obtain a better classification model. Based on the findings of this study, it is suggested that further research can apply different optimization methods such as Bagging, Genetic Algorithm, Adaboost or other optimizations to significantly improve the performance of Naïve Bayes in classifying South German Credit data.

REFERENCES

[1] AT Rahmawati, M. Saifi and RR Hidayat, "Analysis of Credit Provision Decisions in Minimizing Non-Performing Loans," Journal of Business Administration, vol. 35, no. 1, pp. 179-186, 2016.

[2] S. Somadiyono and T. Tresya, "Criminal Responsibility for Marketing according to the Banking Law for Non-performing Financing at Bank Muamalat Indonesia, Tbk," Journal of Lex Specialis, vol. 21, pp. 22-38, 2015.

[3] S. Masripah, "Comparison of Data Mining Classification Algorithms for Evaluation of Credit Provisions," Bina Insani ICT Journal, vol. 3, no. 1, pp. 187-193, 2016.

[4] S. Umadevi and KSJ Marseline, "A Survey on Data Mining Classification Algorithms," at the International Conference on Signal Processing and Communication, Coimbatore, India, 2017.

[5] "Kaggle," kaggle.com, 2020. [Online]. Available: https://www.kaggle.com/c/south-german-credit-prediction/overview/data-overview. [Accessed 2 November 2020].

[6] IAA Amra and AYA Maghari, "Students Performance Prediction Using KNN and Naïve Bayesian," at the 8th International Conference on Information Technology (ICIT), Al-Zaytoonah University of Jordan, Jordan, 2017.

[7] F. Osisanwo, J. Akinsola, O. Awodele, JO Hinmikaiye, O. Olakanmi and J. Akinjobi, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pp. 128-138, 2017.

[8] EN Azizah, U. Pujianto, E. Nugraha and Darusalam, "Comparative Performance Between C4.5 and Naive Bayes Classifiers in Predicting Student Academic Performance in A Virtual Learning Environment," in the 4th International Conference on Education and Technology (ICET), Malang, Indonesia, 2018.

[9] K. Madasamy and M. Ramaswami, "Data Imbalance and Classifiers: Impact and Solutions from A Big Data Perspective," International Journal of Computational Intelligence Research, vol. 13, no. 9, pp. 2267-2281, 2017.

[10] EM Hassib, AI El-Desouky, E.-SM El-Kenawy and SM El-Ghamrawy, "An Imbalanced Big Data Mining Framework for Improving Optimization Algorithms Performance," Journal & Magazines, vol. 7, no. 1, pp. 170774-170795, 2019.

[11] S. Chen, GI Webb, L. Liu and X. Ma, "A Novel Selective Naïve Bayes Algorithm," Knowledge-Based Systems, vol. 192, pp. 1-15, 2020.

[12] L. Jiang, L. Zhang, L. Yu and D. Wang, "Class-Specific Attribute Weighted Naive Bayes," Pattern Recognition, vol. 88, no. 1, pp. 321-330, 2019.

[13] S. Ernawati, R. Wati, N. Nuris, LS Marita and ER Yulia, "Comparison of Naïve Bayes Algorithm with Genetic Algorithm and Particle Swarm Optimization as Feature Selection for Sentiment Analysis Review of Digital Learning Application," Journal of Physics: Conference Series , vol. 1641, pp. 1-7, 2020.

[14] X. Liu, Z. Liu, Z. Liang, S.-P. Zu, JAFO Correia and AMPD Jesus, "PSO-BP Neural Network-Based Strain Prediction of Wind Turbine Blades," Materials, vol. 12, no. 12, pp. 2-15, 2019.

[15] S. Srivastava, J. Gupta and M. Gupta, "PSO & Neural-Network Based Signature Recognition for Harmonic Source Identification," in IEEE Region 10 International Conference TENCON, Singapore, 2009.

[16] M. Misdram, E. Noersasongko, A. Syukur, Purwanto, M. Muljono, HA Santoso and DRIM Setiadi, "Analysis of Imputation Methods of Small and Unbalanced Datasets in Classifications using Naïve Bayes and Particle Swarm Optimization," in International Seminar on Application for Technology of Information and Communication (ISemantic), Semarang, Indonesia, 2020.

[17] I. Romli, T. Pardamean, S. Butsianto, TN Wiyatno and EB Mohamad, "Naive Bayes Algorithm Implementation Based on Particle Swarm Optimization in Analyzing the Defect Product," Journal of Physics: Conference Series, vol. 1845, no. 1, pp. 1-6, 2021.

[18] J. Li, L. Ding and B. Li, "A Novel Naive Bayes Classification Algorithm Based on Particle Swarm Optimization," The Open Automation and Control Systems Journal, vol. 6, no. 1, pp. 747-753, 2014.

[19] MS Hasibuan, Banking Fundamentals, Jakarta: PT Bumi Aksara, 2004.

[20] R. Widayati and M. Efriani, "Activities of Business Loans at PT. Batang Kapas Rural Bank," in OSF Preprints, Batang, Indonesia, 2019.

[21] B. Panuntun and Sutrisno, "Determining Factors in Banking Credit Disbursement Case Study in Conventional Banks in Indonesia," Dewantar Journal of Accounting & Finance Research, vol. 1, no. 2, pp. 57-66, 2018.

[22] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241-266, 2013.

[23] MS Başarslan and ID Argun, "Classification Of A Data Bank Set On Various Data Mining Platforms," in Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), Istanbul, Turkey, 2018.

[24] V. Krishnaiah, G. Narsimha and N. Chandra, "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques," International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39-45, 2013.

[25] K. Sumiran, "An Overview of Data Mining Techniques and Their Application in Industrial Engineering," Asian Journal of Applied Science and Technology, vol. 2, no. 2, pp. 947-953, 2018.

[26] K. Yadav and R. Thareja, "Comparing the Performance of Naive Bayes and Decision Tree Classification Using R," IJ Intelligent Systems and Applications, vol. 12, pp. 11-19, 2019.

[27] S. Sankaranarayanan and TP Perumal, "Analysis of Naïve Bayes Classification for Diabetes Mellitus," International Journal of Computer Sciences and Engineering, vol. 6, no. 12, pp. 520-524, 2018.

[28] M. Jaenal, A. Nugroho and I. Romli, "Analysis of Effectiveness Particle Swarm Optimization in Improving The Performance of Naïve Bayes Algorithm," in ICID Proceedings, Yogyakarta, Indonesia, 2018.

[29] B. Chopard and M. Tomassini, "Particle Swarm Optimization," in An Introduction to Metaheuristics for Optimization, Springer, Cham, Natural Computing Series, 2018, p. 97–102.

[30] YS Haruna, YA Yisah, GA Bakare, MS Haruna and SO Oodo, "Optimal Economic Load Dispatch of the Nigerian Thermal Power Stations Using Particle Swarm Optimization (PSO)," The International Journal Of Engineering And Science (IJES), vol. 6, no. 1, pp. 17-23, 2017.

[31] A. Jeyaraj, R. S and MR Raja, "A study of classification algorithms using Rapidminer," International Journal of Pure and Applied Mathematics, vol. 119, no. 12, pp. 15977-15988, 2018.