No. 317 | Received: 2020-05-11 | Revised: 2020-06-05 | Accepted: 2020-06-08

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390
http://tip.ppj.unp.ac.id

Jurnal Teknologi Informasi dan Pendidikan
Volume 13, No. 2, September 2020
P. 1 – 8 | https://doi.org/10.24036/tip.v13i2

ANALYSIS OF STUDENT TUITION FEE PAY DELAY PREDICTION USING NAIVE BAYES ALGORITHM WITH PARTICLE SWARM OPTIMATION OPTIMAZATION (CASE STUDY : POLITEKNIK TEDC BANDUNG)

Dini Rohmayani
Program Studi Teknik Informatika, Politeknik TEDC Bandung
Jl. Pesantren KM 2 Cibabat Cimahi Utara
Corresponding Author: dinirohmayani@poltektedc.ac.id : 081313217486

ABSTRACT

One source of funds that plays a very important role in education or teaching and learning activities is the Donation of Education Development or tuition fee, tuition fee is also one of the requirements for continuing operational activities at tertiary institutions and cannot be denied the importance of student tuition fee payments, especially at private universities, where all costs for the benefit of the campus are more heavily charged or incurred by the campus itself. The problem faced by TEDC Polytechnic of Bandung is based on historical data from the financial department there are still many students who are late in making tuition payments, for that the authors make an analysis in predicting late payment of fees by using the Naive Bayes algorithm which is compared with Particle Swarm Optimization (PSO) with the purpose of determining the classification pattern is right or late and to find out what indicators most influence the prediction of late payment of tuition fee. In testing the author uses 115 datasets from the results of a questionnaire filled out by students and these testing these divided into three models, the first is testing using the entire dataset and all attributes of the questionnaire results, the second test is done by using the entire dataset based on Particle Swarm Optimation (PSO), as well as the third test carried out using a dataset using the attributes that most influence to the prediction of late tuition payment. Off the 13 indicators, there are only 8 indicators that are most influential in predicting late payment of tuition fee, namely father's income, number of dependents of parents, pocket money / month, financial services, academic services, study programs, payment methods for tuition payment, and mother's occupation. The testing result of those three classification models using the highest Naive Bayes accuracy algorithm are testing using the Naive Bayes algorithm based on Particle Swarm Optimization (PSO), with an accuracy of 73.94%, precision 78.50%, 69% recall, and AUC 0.771 , even though the execution time is 3 seconds longer.

Keywords: *Tuition fee, Naive Bayes Algorithm, Particle Swarm Optimation*

INTRODUCTION

One source of funds that plays a very important role in the world of education or teaching and learning activities is the *Donation of Education Development or tuition fee*. Tuition fee is also one of the requiremen.s for continuing operational activities at Higher Education (HE) and cannot be denied the importance of tuition fee payments made by the students, especially in private universities, where all the costs for the benefit of the campus more or more charged or incurred by the campus itself. TEDC Polytechnic of Bandung as a private campus that provides facilities to conduct lecture activities has a policy of making tuition payments.

The policy is that students must make tuition fee payments at the beginning of the semester, for new students to register and previously registered students (active students) must register with the condition that they have paid 50% of the total tution fee / semester, at the time the Middle Exams will be held students must pay the tuition fee of 75%, and at the time of the Final Semester Examination the student must have paid the tuition fee of 100%. The policy is already good but based on historical data from the financial department there are still students who are late in making tuition payments and it happens every semester, of course this is a problem for the campus or students, the impact for students themselves will affect to the course scores obtained, Based on these problems, the authors make an

analysis to be able to predict the reliability of tuition fee payments by using the Naive Bayes algorithm based on Particle Swarm Optimization (PSO) which aims to determine the exact classification pattern or late and to find out which indicators are most influential on the prediction of delays tuition fee payment by distributing questionnaires to students and giving a number of questions that may be an indicator of late payment of tuiton fee, with several variables including (1) Study Programs, (2) Father's occupation, (3) Mother's occupation, (4) Father s Income/Month, (5) Mother s Income/Month, (6) Number of Dependents of Parents, (7) Expenditures of Parents/Months, (8) Allowance/Month, (9) Boarding/Living with Parents, (10) Parental Residence Status (Rent/Owned), (11) Financial Services, (12) Academic Services, (13)tuition fee Payment Method Cash/ Installment).

To analyze the prediction of late payment of tuition fee, one of the techniques that can be used is the classification technique using data mining. Data mining is the process of getting useful information from a large database that can help in decision making. The term data mining can also be called knowledge discovery.[1]

As for some previous studies that serve as a reference in this study include:

1. Ariyati, et al. In a study entitled "Implementation of Particle Swarm Optimization for Data Mining Optimization in the Evaluation of Lecturer Assistant Performance" the research aims to find alternative complex solutions in the evaluation of teaching assistants where the parameters obtained from the UCI Machine Repository. The results of testing using the Particle Swarm Optimization method can improve accuracy by 75.56% from the previous value of 51.75% and increase the kappa value of 0.632 from the previous kappa value of 0.276. This research can be a motivation for researchers so that the optimization of Particle Swarm Optimization can increase the level of accuracy of the performance of the Naive Bayes algorithm in analyzing the prediction of late payment of tuition fee.[2]
2. In a study conducted by Nugroho, et al who aimed to classify the complaint text of more than one table at the same time with Naive Bayes Classification which was optimized by using Particle Swarm Optimaton (PSO). The results of testing that NBC optimization using PSO achieves an accuracy of 84.44% which is better than KNN and NBC without PSO optimization.[3]
3. In a study conducted by Muqorobin, et al Naive Bayes method with Feature Selection Information Gain can be used as a method to classify late school tuition payment predictions with an accuracy value of 90%, for that Naive Bayes makes an appropriate method in this study.[4]
4. In research on diagnostic decision-making systems with the integration of k-means grouping and Naive Bayes algorithm with different initial selection can improve accuracy in diagnosing patients with heart disease, for that researchers are interested that Naive Bayes can be used as a method in the classification for prediction analysis of delay tuition fee payment.[5]
5. In Xindong Wu and Vipin Kumar's book "The Top Ten Algorithms in Data Mining" discusses the top 10 best algorithms, including Decision Tree C4.5, K-Means, SVM, Apriori, EM, Page Rank, AdaBoost, K- Nearest Neighbors, Naive Bayes, and CART.[6] For this reason, researchers use Naive Bayes as a method to analyze the prediction of late payment of tuition fee because Naive Bayes is included in the 10 best algorithms.

From a number of previous studies, researcher have been interested in making research on the prediction analysis of late payment of tuition fee by optimizing Particle Swarm Optimation (PSO)

Data Mining

Data mining is the process of finding knowledge found in a database or knowledge discovery in a database or usually abbreviated as KDD. This knowledge can be in the form of patterns or relationships between valid data. Data mining is also the process of discovering new patterns from very large data sets, which include sliced methods of artificial intelligence, machine learning, statistics, and database systems. Data mining can also extract (extract the essence) of knowledge from a collection of data to obtain structures that can be understood by humans and include data bases and data management, data processing, consideration of models and inferences, measures of interest, consideration of complexity, post-processing of structures found , visualization, and online updating.[7]

Classification Technique

Classification is a work that has data object value that can be entered into a certain class from the number of available classes. In the classification there are two main jobs performed, namely :[1]

1. Modeling the model as a prototype to be stored as memory
2. the user of the model is to make an introduction / classification / prediction of another data object so that it is known which class of data object in the model it has stored.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390

Naive Bayes Algorithm

Naive Bayes is a simple probability classification algorithm that can calculate a set of probabilities by adding up the frequency and combination of values from a given dataset. Bayes is a simple probability-based prediction technique based on the Bayes theorem (Bayes' rules) which assumes strong (naive) independence. The model used at Naive Bayes is an "independent feature model". also a classification using a statistical method that has been proposed by a British scientist named Thomas Bayes, which predicts future opportunities based on the previous period.

The advantage of the Naive Bayes method is that the Naive Bayes method only requires a small amount of training data to determine the parameters required in the classification process.[8]

The method in Bayes theorem's probability or conditional probability can be seen in equation 1.

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad \text{.....(1)}$$

Where X is proof, H is a hypothesis, P (H \ X) is the probability that the hypothesis H is true for proof X or P (H \ X) is the posterior probability of H on condition X, P (X \ H) is the probability that proof X true for hypothesis H or the posterior probability of X on conditions H, P (H) is the probability of prior hypothesis H, and P (X) is the probability of prior proof X.[7]

Rapidminer

Rapidminer is an application that can be downloaded for free (opensource) and can be used for data mining processes. One method of data mining is to use linear regression. Linear regression is a statistical method used to estimate or estimate based on existing data. Rapidminer Learning includes Extraction, Transformation, Loading (Extrac Tranfor Loading) and Preprocessing, visualization, modeling, and evaluation.[9]

Cross Validation

Cross Validation is a statistical method that can be used to evaluate and compare learning algorithms by dividing data into two segments, namely one is used to learn or train the model and another is used to validate the model. Cross Validation is also a technique in assessing and validating a model that is built based on a specific dataset.[10]

Particle Swarm Optimation (PSO)

Particle Swarm Optimization (PSO) was first introduced by Kennedy and Elbert in 1995 which is a population-based optimization technique that is inspired by the social behavior of a group of birds that are gathering (Swarm). This social behavior consists of individual actions and the influence of other individuals in a group. To find a more optimal solution the particles in the PSO move towards a better search after passing through the search process.[11]

In the PSO algorithm, the velocity vector is updated for each particle, then the sum of the velocity vector is added to the particle position. The process of updating the velocity is affected by the two solutions: adjusting the best position of the particle and adjusting to the best particle of the entire set (global best). At each iteration, each solution presented by the position of the particle is evaluated by inserting the solution into the fitness function. The procedure of the PSO algorithm is as follows :

1. Initialize populations of particles at random positions and speeds
2. Evaluate the fitness value for each particle
3. Comparison and renewal of the best particles and particles of the whole set for each particle based on the fitness function.

METODE

This research is a type of case study research, because in this research does not make tools to be analyzed, but in this study only observes phenomena associated with the observed object, the nature of this study is causal, that is, the researcher gets information from the financial section about the number students who are late in making tuition payments at the Politeknik TEDC Bandung, and the approach used in this study is a qualitative approach that aims to increase understanding of something.

This research was conducted based on variables that might be a factor in the late payment of tuition fee, in this study three tests would be conducted, the first to test with all datasets using Naive Bayes, the second to test with all datasets using Naive Bayes based on particle swarm optimization (PSO), and the third test using the attributes that most influences the late payment of tuition fee, so the final result will be a comparison of the results of the accuracy of the three tests and will get the most influential indicator on the prediction of late payment of tuition fee at Politeknik TEDC Bandung.

Jurnal Teknologi Informasi dan Pendidikan
Volume 13, No. 2, September 2020
http://tip.ppj.unp.ac.id

In Figure 1, we will explain the research flow that is used to predict late payment of tution fee by using the Naive Bayes method based on Particle Swarm Optimatin (PSO) in the form of a flow chart.

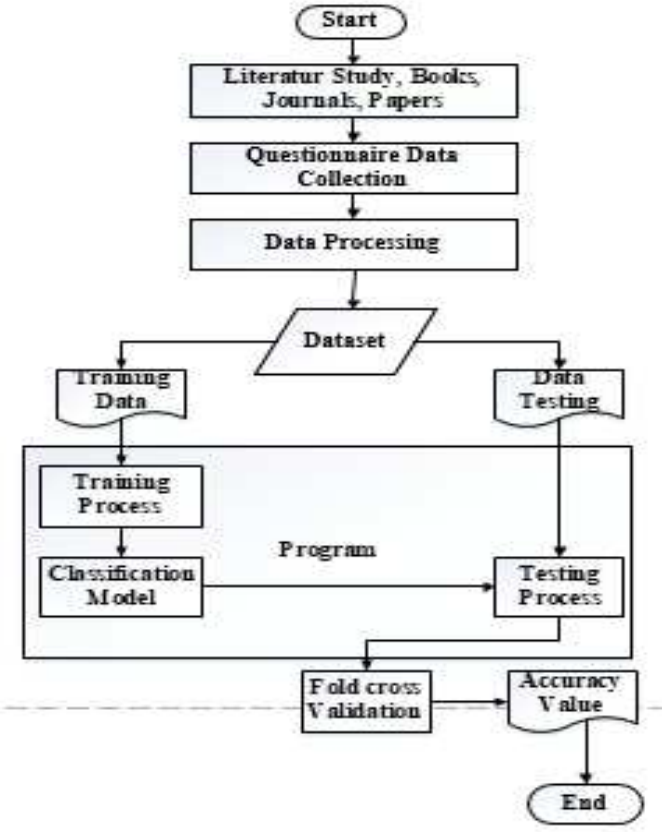

Figure 1. Research Flow

An explanation of Figure 1 regarding the research flow is as follows:

1. Literature Study, which is a literature review in finding references related to the method used by researchers, which is sourced from several books or relevant scientific journals.
2. Questionnaire data collection, which is collecting data based on data requirements that will be used in research. The data collected is the result of a questionnaire distributed to students.
3. Data Preprocessing, cleaning / cleansing data that is not needed, data integrity or merging data, data selection, and data transformation.
4. The dataset that will be tested in the Rapidminer tool and the Naive Bayes algorithm is divided into two: training data and testing data.
5. Classification models, in this study the test is divided into three models, the first test using the entire dataset obtained from the questionnaire consisting of 115 data, the second using a dataset with PSO optimization, and the third testing using the attributes that most influence on predictions late payment of tuition fees.
6. Fold Cross Validation, all classification models are tested using Fold Cross Validation so that it will produce an accuracy value on each model.

RESULT AND DISCUSSION

In this study the collection of data and information obtained from the questionnaire results that have been filled out by the Politeknik TEDC Bandung students on the link bit.ly/kuesionerprediksi-spp.

The dataset that was filled out by the students amounted to 115 datasets, before testing, the researchers conducted data preprocessing, namely cleaning / cleaning unneeded data, merging data, and selecting and transforming data in order to obtain a dataset that can be easily processed and fast and can produce conclusions and appropriate information.

Survey data in the form of questionnaires distributed to Politeknik TEDC Bandung students consisted of several factors and aspects of assessment, before testing the data, researcher determined several variables as listed in Table 1.

Table.1 Rating Attributes and Components

NO	Atribut	Komponen Penilaian
1	Study Progam	1. Akuntansi 2. Komputerisasi Akuntansi 3. Mekanik Industri dan Desain 4. Mesin Otomotif 5. Rekam Medik dan Informasi Kesehatan 6.Teknik Elektronika 7. Teknik Informatika 8. Teknik Komputer 9. Teknik Mesin 10.Teknik Otomasi Industri
2	Father s Occupation	1. Civil Servants 2. Private Employee 3. Fisherman/Farmers 4. Doesn t Work 5. TNI/POLRI 6. Entrepreneurs 7. Other..
3	Mother s Occupation	1. Civil Servants 2. Private Employee 3. Fisherman/Farmers 4. Doesn t Work 5. TNI/POLRI 6. Entrepreneurs 7. Other..
4	Father s Income/Month	1. More than Rp.5.000.000,- 2. 0 Rp. 2.000.000,- 3. Rp.2.000.000,- - 3.000.000,-

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390

		4. Rp.3.000.000,- - Rp.4.000.000,- 5. Rp.4.000.000,- - Rp.5.000.000,-
5	Mother s Income/Month	1. More than Rp.5.000.000,- 2. 0 Rp. 2.000.000, 3. Rp.2.000.000,- - 3.000.000, 4. Rp.3.000.000,- - Rp.4.000.000, 5. Rp.4.000.000,- - Rp.5.000.000,-
6	Number of Dependents of Parents	1. 1 2. 2 3. 3 4. 4 5. More than 4
7	Expenditures of Parents/Months	1. More than Rp.1.000.000,- 2. 0 Rp. 2.000.000,- 3. Rp.2.000.000,- - 3.000.000,- 4. Rp.3.000.000,- - Rp.4.000.000,- 5. Rp.4.000.000,- - Rp.5.000.000,-
8	Allowance/Month	1. More than Rp.1.000.000,- 2. 0 Rp.300.000,- 3. Rp.300.000,- - Rp.500.000,- 4. Rp.500.000,- - Rp.700.000,- 5. Rp.700.000,- - Rp.1.000.000,-
9	Residence	1. Living with Parents 2. Boarding 3. Other...
10	Parental Residence Status	1. Owned 2. Rent 3. Other...
11	Financial Services	1. Baik 2. Not good
12	Academic Services	1. Good 2. Not good
13	tuition fee Payment Method	1. Cash 2. Installment

Model Development

Model development is done by testing three models, the first is testing with the entire dataset, the second is by optimizing Particle Swarm Optimation, and the third is testing by using the attributes that most influence the prediction of late payment of tuition fee.Tools used are Rapidminer Study Free 8.2.001.

The dataset that will be imported into Rapidminer is a dataset in the form of an excel file, then proceed with determining each indicator for each attribute. For attributes that contain two classifications are binominal and polynominal, for status attributes (late and exact status) are changed to label because they are attributes that will be determined.

The results of setting and truncating the dataset in importing into the Rapidminer can be seen in Figure 2.

Figure 2. Import Dataset Process into Rapidminer

After the dataset is imported into Rapidminer it produces 56 statistics which are late data and 59 data that are correct in making tuition fee payments from 115 datasets. Detailed statistics can be seen in Figure 3.

Figure 3. Detailed Statistics Data on Rapidminer

Development, Testing and Performance Accuracy Results Using the Naive Bayes Algorithm

The test starts with a retrive dataset stored with the name prediction of tuition fee delay and then connected to cross validation, from cross validation there are four validations connected to the result, namely the validation model, validation example set, validation test, and validation performance I, validation performance functions to display Accuracy , Precision, Recall, and AUC. Testing using the Naive Bayes algorithm can be seen in Figure 4.

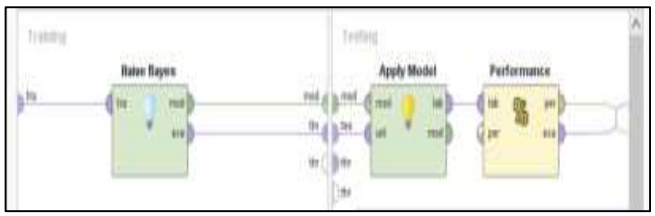

Figure 4. Testing Process with Naive Bayes Algorithm

Performance vector test results using Naive Bayes algorithm are 65.50% accuracy, Precision

Jurnal Teknologi Informasi dan Pendidikan
Volume 13, No. 2, September 2020
http://tip.ppj.unp.ac.id

66.63%, Recall 63.84%, and AUC 0.732. Performance vector and performance vector description can be seen in Figure 5 and Figure 6.

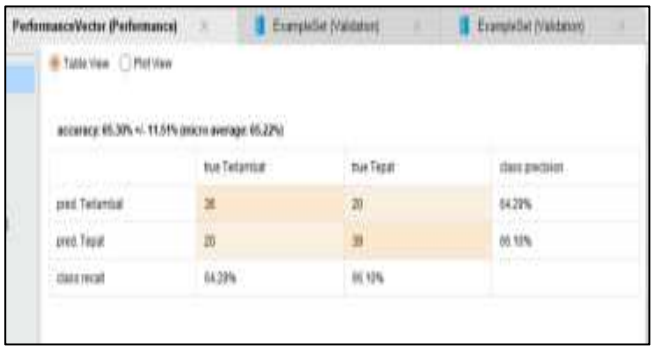

Figure 5. Vector Performance Results Using the Naive Bayes Algorithm

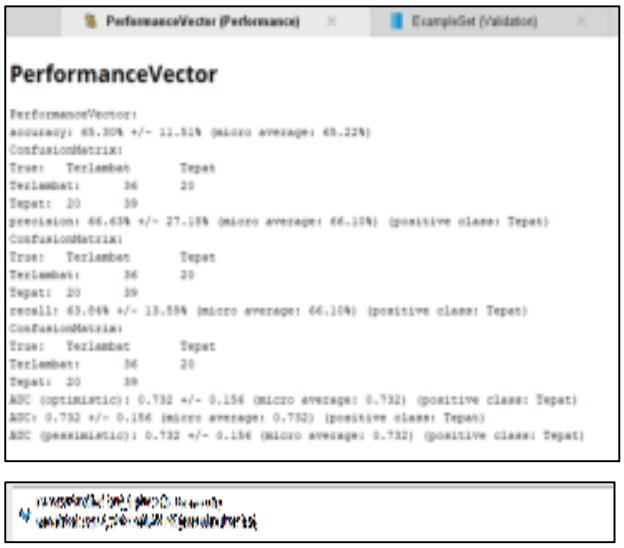

Figure 6. Description of Vector Performance

Test and Performance Results Using the Naive Bayes Algorithm Based on Particle Swarm Optimization (PSO)

The testing process using the PSO-based Naive Bayes algorithm can be seen in Figure 7.

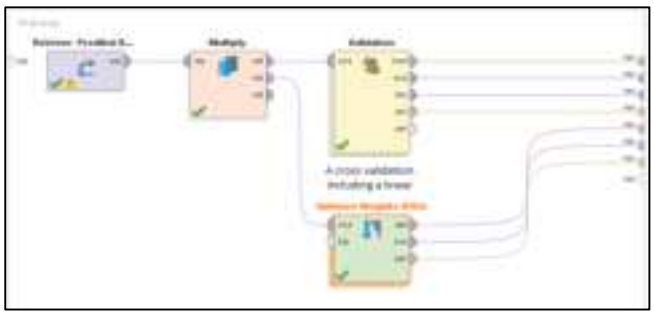

Figure 7. PSO-Based Naive Bayes Algorithm Testing Process

Vector performance testing results can be seen in Figure 8.

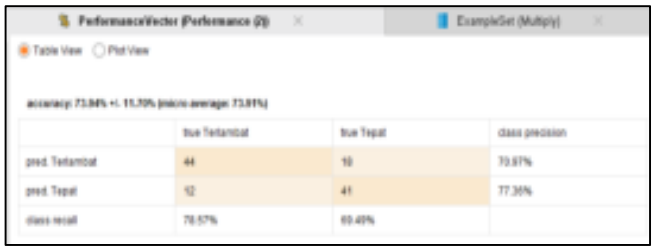

Figure 8. Results of Vector Prformance Using Naive Bayes Algorithm Based on PSO

From the test results using Naive Bayes algorithm based on PSO produced 73.94% accuracy, 78.50% precision, 69.00% recall, and AUC 0.771

Testing Naive Bayes algorithm based on PSO can bring up the most influential attribute in predicting late payment of tuition fee, the results can be seen in Figure 9.

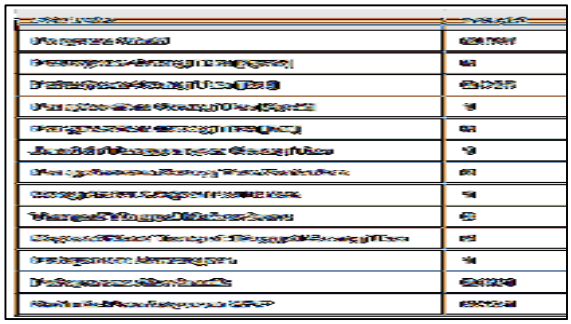

Figure 9. The most influential attribute values

Of the 13 attributes in predicting late payment of tuition fee, there are 8 most influential attributes namely Father's Income, Number of Dependents of Parents, Allowance / Month, Financial Services, Academic Services, Study Programs, tuition fee Payment Methods, and Mother's Work.

Testing Using the Most Influential Attributes in Predicting Late Payment of tuition fee Using the Naive Bayes Algorithm

Performance Vector of the Naive Bayes algorithm test results with the most influential attributes in predicting late payment of tuition fee is accuracy 71.14%, Precision 73.62%, recall 69.67%, and AUC 0.740, can be seen in Figure 10.

Figure 10. Performance Vector Test Results Using the Most Influential Attributes

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390

Analysis and Comparative Performance Classification Model

From the results of tests and evaluations that have been carried out on the Naive Bayes algorithm, the Particle Swarm Optimization (PSO) based algorithm model, and the Naive Bayes algorithm using the most influential attributes, the performance comparison results can be seen in Table 2.

Table 2. Comparison Results of the Naive Bayes Algorithm Performance Classification Model

Performance Parameters	Naive Bayes algorithm	Naive Bayes Algorithm Based on PSO	The Naive Bayes algorithm uses the most influential attributes
Accuracy	65,30%	73,94%	71,14%
Pecision	66,63%	78,50%	73,62%
Recall	63,84%	69%	69,67%
AUC	0,372	0,771	0,740
Execution Time	0S	3S	0S

CONCLUSION

Based on the results of research and discussion that has been done, it can be concluded that the application of Particle Swarm Optimization (PSO) with the Naive Bayes algorithm for prediction of late payment of tuition fee payments at TEDC Polytechnic of Bandung obtained better accuracy results which is 78.50%, whereas if without using PSO 65 , 30% and by using Naive Bayes and the most influential attribute is 73.62%. So in this study concluded that the optimization of the Naive Bayes method based on Particle Swarm Optimization can help in predicting late payment of the TEDC Polytechnic of Bandung with better accuracy, and from the use of the PSO method can also find out the most influential attributes in the prediction of late payment of SPP namely Father's Income, Number of Dependents of Parents, Allowance / Month, Financial Services, Academic Services, Study Programs, tuition fee Payment Methods, and Mother's Work.

SUGGESTION

Based on research that has been done as for suggestions or ideas that can be proposed for research related to the prediction of late payment of tuition fee, including:

1. Testing the prediction of late tuition can be done by using the most influential attributes, namely Father's Income, Number of Dependents of Parents, Allowance/ Month, Financial Services, Academic Services, Study Programs, tuition fee Payment Methods, and Mother's Work added other attributes such as name student, gender, and student achievement index.
2. Can use other algorithms such as SVM, KNN, which is optimized by the Particle Swarm Optimation method.

Perform comparisons using other methods such as the K-Nearest Noeighbor algorithm, SVM and others.

1. We recommend that the dataset used can be added with a greater number.

REFERENCES

[1] F. A. Hermawati, *Data Mining*. Yogyakarta: CV. Andi Offset, 2013.

[2] I. Ariyati, Ridwansyah, and Suhardjono, Implementasi particle swarm optimization untuk optimalisasi data mining dalam evaluasi kinerja asisten dosen, *Inform. dan Komput.* , vol. 3, no. 2, pp. 70 75, 2018.

[3] K. S. Nugroho, F. Marisa, I. Istiadi, and F. Marisa, Optimasi naive Bayes classifier untuk klasifikasi teks pada e-government menggunakan particle swarm optimization, *J. Teknol. dan Sist. Komput.*, vol. 8, no. November 2019, pp. 21 26, 2020.

[4] Muqorobin, Kusrini, and E. T. Luthfi, Optimasi Metode Naive Bayes Dengan Feature Selection Information Gain Untuk Prediksi Keterlambatan Pembayaran Sumbangan Pembinaan Pendidikan Sekolah, *J. Ilm. Sinus*, vol. 17, no. 01, pp. 1 14, 2019.

[5] A. Altayeva, S. Zharas, and Y. I. Cho, Medical Decision Making Diagnosis System Integrating k-means and Naïve Bayes algorithms Department of Computer Engineering , Gachon University , Seongnam , Seoul, in *International Conference on Control, Automation and Systems*, 2016, no. Iccas, pp. 1087 1092.

[6] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining, pp. v 201, 2009.

[7] Suyanto, *Data Mining Untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2017.

[8] A. Saleh, Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, *Citec J.*, vol. 2, no. 3, pp. 207 217, 2015.

[9] I. A.Muis and M. Affandes, Penerapan Metode Support Vector Machine (SVM) Menggunakan Kernel Radial Basis Function (RBF) Pada Klasifikasi Tweet, *J. Sains, Teknol. dan Ind.*, vol. 12, no. 2, pp. 189 197, 2015.

[10] Nursalim, Suprapedi, and H.himawan, Klasifikasi Bidang Kerja Lulusan Menggunakan Algoritma K-Nearest Neighbor, *Teknol. Inf.*, vol. 10, no. April, pp. 31 43, 2014.

[11] F. Y. Bisilisin, Y. Herdiyeni, and B. I. B. P. Silalahi, Optimasi K-Means Clustering Menggunakan Particle Swarm Optimization pada Sistem Identifikasi Tumbuhan Obat Berbasis Citra K-Means Clustering Optimization Using Particle Swarm Optimization on Image Based Medicinal Plant Identification System, *Ilmu Komput. Agri-Informatika*, vol. 3, no. 2002, pp. 38 47, 2014.

P.ISSN: 2086 – 4981
E.ISSN: 2620 – 6390