International Journal of Information System & Technology
Akreditasi No. 158/E/KPT/2021 | Vol. 7, No. 6, (2024), pp. 386-391

Improved Naive Bayes Algorithm with Particle Swarm Optimization to Predict Student Graduation

Embun Fajar Wati[1]*, Elvi Sunita Perangin-Angin*[2]*, Anggi Puspita Sari*[3]
[1,2]*Universitas Bina Sarana Informatika, Indonesia*
Email: embun.efw@bsi.ac.id[1]*, elvi.evt@bsi.ac.id*[2]*, anggi.apr@bsi.ac.id*[3]

Abstract

Timely graduation is very important for educational institutions such as universities, especially for students. Because it can prove that the University and students are able to undergo the learning process theoretically and practically. But many students do not pay attention to graduation, especially those who are already working or married. Therefore, analysis is needed to predict student graduation so that solutions can be found by the University. Data mining was chosen as a method to process data to get new information. The algorithm used in data mining is Naïve Bayes. The research stages include loading data into excel, cleaning empty data, selecting databases related to graduation and taking data from 300 students majoring in Informatics Engineering. The next stage is data transformation by categorizing student data, namely personal data attributes (gender, age, marital status, job status) and academic data (grade). Data testing, application of Naïve Bayes algorithm and accuracy testing were carried out with Rapis Miner software version 10.3.001. The results of data processing with Rapid Miner using the Naïve Bayes algorithm are shown with the Confusion Matrix and ROC Curve. The results of confusion matrix from data processing with Naïve Bayes in the form of accuracy, precision, and recall have the same result of 100%. The percentage of the Confusion Matrix indicates that the model created can classify correctly and accurately. The ROC curve depicted with AUC yields a value of 1, which means that the test showed excellent results.

Keywords: *graduation, students, data mining, naïve bayes, university*

1. Introduction

The rapid development in technology can be enjoyed by everyone [1] Advances in information technology have penetrated into all areas of life [2]. The level of higher education is one of the basic requirements in finding a job, where universities will prepare qualified undergraduate candidates and have skills in their fields [3]. Timely graduation from higher education becomes one of the indicators of the quality of a country's education system [4]. Each university has academic data and biodata for each student based on initial enrollment to graduation [5]. The data can be used to find new information.

The study of student graduation classification in a university aims to help the university understand student academic development and to find solutions in improving student graduation development in a timely manner [6]. Prediction of student graduation accuracy is designed to support study programs to guide students to graduate on time [7]. The resulting predictions become valuable input for universities to improve the quality of educational standards [8]. Students' on-time graduation has a huge impact on education [9]. The quality of students can be measured by the period of education in college. Timely graduates who have the title of quality students [10]. To get a good and timely graduation rate. Students are highly dependent on influences from on- and off-campus factors [11].

Based on the above, a system is needed that can predict student graduation so that evaluation can be carried out [12]. The purpose of this study is to classify students who graduate late, so that universities can provide policies such as accelerating graduation in the academic [13]. Several methods in data mining are used in predicting graduation on time in the form of classification methods [14]. Data mining is an extraction activity to

ISSN : 2580-7250

obtain important implicit information that was previously unknown from a data [15]. The model used uses the Naïve Bayes algorithm which serves as one of the models to perform classification [16]. The use of this method was chosen because it is easy to implement freely and the features owned by the data are not related to other features in the same data [17]. The naïve bayes method has advantages, including being able to predict the number of passes based on concrete data, so that the results obtained can be accounted for and used for further predictions [18]. Several attributes were used in the prediction in this study such as, personal data attributes (gender, age, marital status, job status) and academic data (grade). Then from the attributes and models used, researchers use a tool to manage data, namely Rapidminer to process datasets that have been prepared [19].

2. Research Methodology

Stages in research become very important in supporting the success of research, so as to get good results. Some of the stages used in this study are taken from a combination of previous research [9] [7] :

a) Data load is a process of entering and processing data into a program in the form of an Excel document file for processing.
b) Data cleaning is the first stage where noise data (odd values), data that is not used in the data mining process, data whose values are incomplete are cleaned from the student data loaded. For example, one student's name is not known to be scored, then the name will be deleted because it is considered noise data.
c) Database selection is a phase where determining which data parameters will be used in the mining stage, because valid data will be retrieved from the database. Not all attributes are taken in the mining process, the attributes that will be processed in the method are personal data attributes (gender, age, marital status, job status) and academic data (grade). In addition to these attributes will be removed.
d) Data transformation is a phase where data is converted, that is, data is categorized [20]. As in table 1, we can see the attributes of age (young: 19 - 24, old: 25 - 50) and grade (large: 3 - 4, small: 1 - 2.9).

Table 1. Data Transformation

Age	Young	19 – 24 years old
	Old	25 – 50 years old
Grade	Large	3 – 4 years old
	Small	1 – 2.9 years old

e) Data testing is a set of data with attributes to be tested to display prediction results. Testing data is processed using the application of algorithms to find out how accurate the performance of this study is. The study used 300 data taken from a database for data testing.
f) The application of the algorithm is the application process to the research system using the Naive Bayes algorithm with particle swarm optimization and validated 10 times with the 10-fold Cross Validation method. Naive Bayes is the best probability value search and is one of the classification type algorithms [10]. Particle swarm optimization utilizes information sharing as in birds, and their populations from evolutionary irregularity to order and can obtain optimal solutions [13].

   K-fold cross validation is one of the techniques used to sort data into train data and test data. This technique is widely applied by researchers because it is found to reduce the bias obtained in sampling [19]. K-fold cross validation applies to continuously divide data into train data and test data, so that each data will have the opportunity to become test data [21]. K is the large number of data sorting used for the division of trains and tests. K-fold cross validation was used in this study 10 times.

g) Accuracy testing, namely checking accuracy using the Confusion Matrix table. Confusion matrix is a method used to measure the performance of a classification algorithm as an array of four values representing the results of the classification [22]. The four values are True Positive (TP) values, namely students are expected to graduate on time (positive) and actually (correctly) graduate on time, True Negative (TN) values are students cannot graduate on time (negative) and correct. (True) did not graduate on time, False Positive (FP) grades are students who should have graduated on time (positive) but actually did not graduate on time (false) and False Negative (FN scores) are students who should not have graduated on time (Negative) but actually graduated on time.

3. Results and Discussion

The stages in the research are beginning with data load by taking data from the database of students majoring in informatics engineering. Next, the data cleaning stage is by cleaning incomplete or empty or incorrect data because it is considered missing value and noise. All data is complete, and this indicates that all data can be used. Database selection is to select data in the form of attributes related to graduation and not all student data. The student data taken was only 300 informatics engineering majors. Data transformation is carried out with the aim of facilitating classification with Rapid Miner. The transformed data can be seen in table 1. After the data is cleansing, selection and testing, the next stage is data testing which uses all data of 300 students. Data totaling 300 students, can be seen in table 2.

Table 2. Student Data

Gender	Age	Marital Status	Grade	Job Status
Male	Young	Married	Low	Employee
Female	Young	Single	Low	Employee
Male	Young	Married	Low	Unemployee
Male	Young	Single	Low	Employee
Male	Young	Married	Low	Unemployee
…	…	…	…	…
…	…	…	…	…
Male	Young	Married	Low	Unemployee
Female	Young	Single	Low	Employee
Male	Young	Married	Low	Unemployee

Application of Naïve Bayes algorithm with particle swarm optimization and validated 10 times with 10-fold Cross Validation method. Designed with Rapid Miner software that aims to classify student graduation data. The accuracy of the data will be shown by the confusion matrix table that will be displayed when running the rapid miner. The picture of the application of these algorithms can be seen in figure 1.

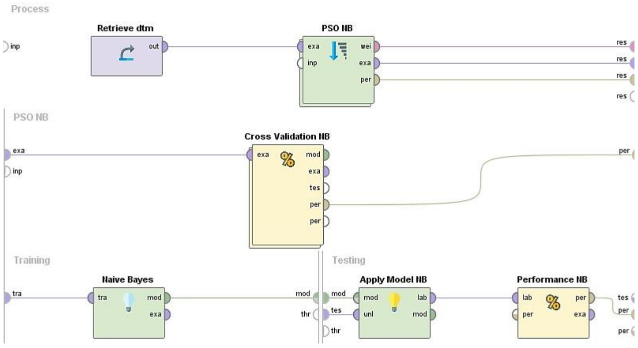
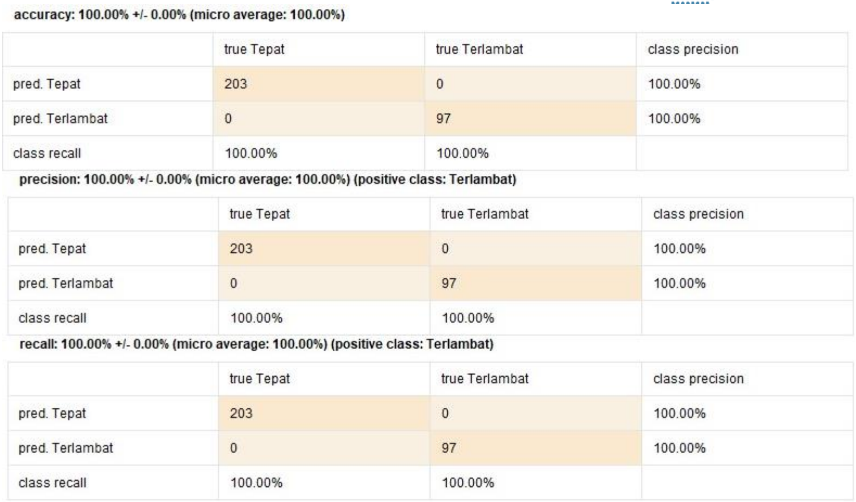


Figure 1. Application of the Naïve Bayes Algorithm

The result of applying the Naïve Bayes algorithm to Rapid Miner is the Confusion Matrix in the form of accuracy, precision, and recall can be seen in figure 2.



accuracy: 100.00% +/- 0.00% (micro average: 100.00%)

	true Tepat	true Terlambat	class precision
pred. Tepat	203	0	100.00%
pred. Terlambat	0	97	100.00%
class recall	100.00%	100.00%	

precision: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	203	0	100.00%
pred. Terlambat	0	97	100.00%
class recall	100.00%	100.00%	

recall: 100.00% +/- 0.00% (micro average: 100.00%) (positive class: Terlambat)

	true Tepat	true Terlambat	class precision
pred. Tepat	203	0	100.00%
pred. Terlambat	0	97	100.00%
class recall	100.00%	100.00%	

Figure 2. Confusion Matrix

In figure 2, there is a confusion matrix result from data processing with Naïve Bayes in the form of accuracy, precision, and recall which has the same result of 100%. The percentage of the Confusion Matrix indicates that the model created can classify correctly and accurately. The ROC curve, depicted with AUC in figure 3, yields a value of 1, which means that the test showed excellent results.

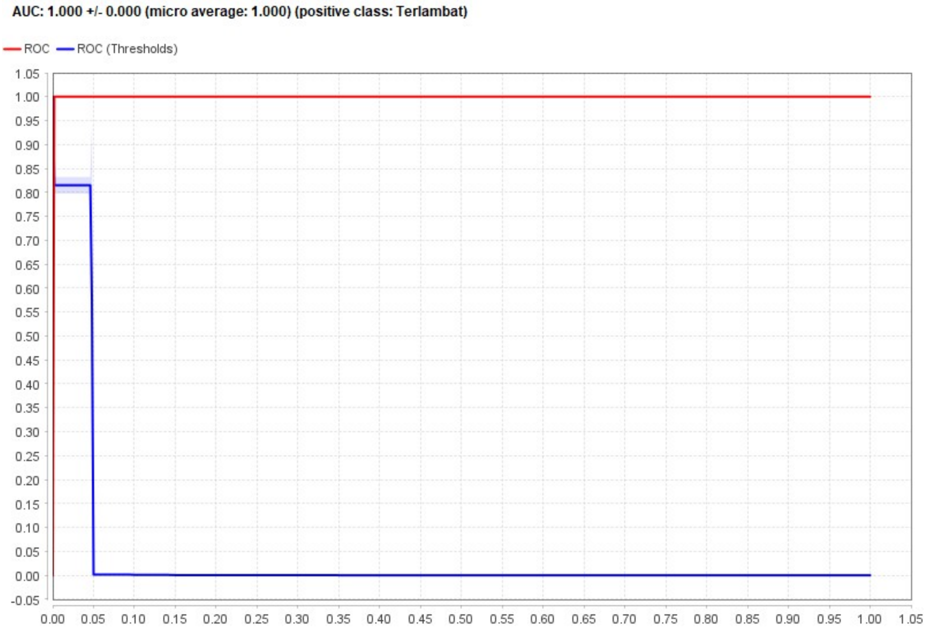


Figure 3. ROC Curve with AUC

4. Conclusion

The use of the Naïve Bayes algorithm in processing data on student graduation predictions using Rapid Miner software version 10.3.001 produces a correct and accurate classification. This can be shown by the value in the Confusion Matrix consisting of, accuracy, precision, and recall which has the same result of 100%. While the ROC curve depicted with AUC, produces a value of 1, which means that the test shows very good results.

In future research, a comparison of several classification algorithms will be used, so that the percentage of each algorithm can be known and how accurate the algorithm is in predicting student graduation. In addition, Naïve Bayes can also be used for data other than graduation. Data can be pulled from twitter. So it can be known how the results of the accuracy of Naïve Bayes on other data.

References

[1] Hartatik, “Optimasi Model Prediksi Kelulusan Mahasiswa Menggunakan Algoritma Naive Bayes,” *Ijai (Indonesian Journal Of Applied Informatics)*, Vol. 5, No. 1, Pp. 32–38, 2020.

[2] M. T. Sembiring And R. H. Tambunan, “Analysis Of Graduation Prediction On Time Based On Student Academic Performance Using The Naïve Bayes Algorithm With Data Mining Implementation (Case Study: Department Of Industrial Engineering Usu),” *Iop Conf Ser Mater Sci Eng*, Vol. 1122, No. 1, P. 012069, Mar. 2021, Doi: 10.1088/1757-899x/1122/1/012069.

[3] Moh. Zainuddin, “Perbandingan 4 Algoritma Berbasis Particle Swarm Optimization (Pso) Untuk Prediksi Kelulusan Tepat Waktu Mahasiswa,” *Jurnal Ilmiah Teknologi Informasi Asia*, Vol. 13, No. 1, Pp. 1–12, 2019.

[4] Indra Irawan, M Riski Qisthiano, Muhammad Syahril, And Pamuji M. Jakak, “Optimasi Prediksi Kelulusan Tepat Waktu: Studi Perbandingan Algoritma Random Forest Dan Algoritma K-Nn Berbasis Pso,” *Jurnal Pengembangan Sistem Informasi Dan Informatika*, Vol. 4, No. 4, Pp. 26–36, 2023.

[5] Robi Sepriansyah And Susan Dian Purnamasari, “Prediction Of Student Graduation Using Naïve Bayes,” *Budapest International Research And Critics Institute-Journal (Birci-Journal)*, Vol. 5, No. 3, Pp. 24255–24268, 2022.

[6] Evi Purnamasari, Dian Palupi Rini, And Sukemi, “Seleksi Fitur Menggunakan Algoritma Particle Swarm Optimization Pada Klasifikasi Kelulusan Mahasiswa Dengan Metode Naive Bayes,” *Jurnal Resti*, Vol. 5, No. 3, Pp. 469–475, 2021.

[7] Embun Fajar Wati, Elvi Sunita Perangin-Angin, And Anggi Puspita Sari, “Prediction Of Student Graduation Using The K-Nearest Neighbors Method,” *International Journal Of Information System & Technology*, Vol. 7, No. 3, Pp. 211–216, 2023.

[8] Shilpa Mehta, “Playing Smart With Numbers: Predicting Student Graduation Using The Magic Of Naive Bayes,” *International Transactions On Artificial Intelligence (Italic)*, Vol. 2, No. 1, Pp. 60–75, 2023.

[9] Imam Riadi, Rusydi Umar, And Rio Anggara, “Prediksi Kelulusan Tepat Waktu Berdasarkan Riwayat Akademik Menggunakan Metode Naïve Bayes,” *Rio Anggara*, Vol. 4, No. 1, Pp. 191–203, 2024.

[10] Embun Fajar Wati And Biktra Rudianto, “Penerapan Algoritma Knn, Naive Bayes Dan C4.5 Dalam Memprediksi Kelulusan Mahasiswa,” *Jurnal Format*, Vol. 11, No. 2, Pp. 168–175, 2022.

[11] Aulia Putri *Et Al.*, “Comparison Of K-Nn, Naive Bayes And Svm Algorithms For Final-Year Student Graduation Prediction,” *Malcom: Indonesian Journal Of Machine Learning And Computer Science*, Vol. 3, No. 1, Pp. 20–26, 2023.

[12] Ida Bagus Adisimakrisna Peling, I Nyoman Arnawan[, I Putu Arich Arthawan, And Ign Janardana, “Implementation Of Data Mining To Predict Period Of Students Study Using Naive Bayes Algorithm,” *International Journal Of Engineering And Emerging Technology*, Vol. 2, No. 1, Pp. 53–57, 2017.

[13] E. F. Wati, A. P. Sari, E. T. Alawiah, M. H. Siregar, And B. Rudianto, “Particle Swarm Optimization Comparison On Decision Tree And Naive Bayes For Pandemic Graduation Classification,” In *2nd International Conference On Advanced Information Scientific Development (Icaisd)*, 2021, Pp. 1–11.

[14] Sudriyanto, Rudi Rizaldi, And M. Ainun Rofiq Hariri, "Implementasi Particle Swarm Optimization (Pso) Untuk Optimisasi Algoritma Naive Bayes Dalam Memprediksi Mahasiswa Lulus Tepat Waktu," *Jurnal Kecerdasan Buatan, Komputasi Dan Teknologi Informasi*, Vol. 2, No. 1, Pp. 62–68, 2021.

[15] S. Suwitno And A. Wibowo, "On-Time Graduation Prediction System Using Data Mining Classification Method," In *Proceedings Of The Proceedings Of The 1st Workshop On Multidisciplinary And Its Applications Part 1, Wma-01 2018, 19-20 January 2018, Aceh, Indonesia*, Eai, 2019. Doi: 10.4108/Eai.20-1-2018.2281900.

[16] M. R. Qisthiano, T. B. Kurniawan, E. S. Negara, And M. Akbar, "Pengembangan Model Untuk Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Dengan Metode Naïve Bayes," *Jurnal Media Informatika Budidarma*, Vol. 5, No. 3, P. 987, Jul. 2021, Doi: 10.30865/Mib.V5i3.3030.

[17] E. F. Wati And B. Sudrajat, "Application Of Naive Bayes Method For Diagnosis Of Pregnancy Disease," *International Journal Of Information System & Technology*, Vol. 6, No. 1, Pp. 93–100, 2022.

[18] Nur Mahar Aji, Vihi Atina, And Nugroho Arif Sudibyo, "Pemodelan Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes Di Uniba," *Jurnal Manajemen Informatika Dan Sistem Informasi*, Vol. 5, No. 2, Pp. 148–158, 2023.

[19] M Riski Qisthiano, "Penerapan Model Klasifikasi Naïve Bayes Untuk Prediksi Mahasiswa Tepat Waktu," In *Seminar Nasional Riset Dan Inovasi Teknologi (Semnas Ristek)*, 2023, Pp. 164–168.

[20] S. Allan, "Migration And Transformation: A Sociomaterial Analysis Of Practitioners' Experiences With Online Exams," *Research In Learning Technology*, Vol. 28, No. 2279, Pp. 1–14, Jan. 2020, Doi: 10.25304/Rlt.V28.2279.

[21] Tengku Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *Klik: Kajian Ilmiah Informatika Dan Komputer*, Vol. 2, No. 5, Pp. 178–185, 2022.

[22] W. Yusuf, R. Witri, And C. Juliane, "Model Prediksi Penjualan Jenis Produk Tekstil Menggunakan Algoritma K-Nearest Neighbor (K-Nn)," *Ijcit (Indonesian Journal On Computer And Information Technology)*, Vol. 7, No. 1, Pp. 1–6, Jul. 2022, Doi: 10.31294/Ijcit.V7i1.11973.