Implementation of Naive Bayes Algorithm in Analyzing Acceptance of Poor Student Assistance

Amri Muliawan Nur1,*, Muhammad Farid Wazdi2, Bambang Harianto3, Muhammad Farid Zaini4 “Universitas Hamzanwadi Lombok Indonesia”
*Email: muliamriga@gmail.com

Abstract. The cost of education that is not insignificant is felt by the community, especially among the lower-middle-class people who have low-income levels. MA Hizbul Wathan NW Semaya Sikur is one of the schools where some students still have a lower-middle economic level. To overcome the existing problems the government provides assistance to poor students (BSM) in schools, but the problems felt in schools are still difficult in analyzing students who are entitled to receive BSM, therefore it is necessary to build a data processing system using the principles of data mining with the aim of helping the school in analyzing students who are entitled to receive assistance. The Naive Bayes method was chosen because it can predict future opportunities based on experience. The test results obtained were 95.27% of 169 student data with an AUC value of 1,000. This value is categorized as an excellent classification. Therefore the Naive Bayes method can be used as a reference in analyzing poor student assistance data very well.

Keywords: School, Poor Student Assistance, Naive Bayes.

1. Introduction
Education is a compulsory basis for development, for that, every Indonesian citizen is obliged and entitled to receive education, especially 12-year compulsory education programmed by the government[1]. To take education, a significant amount of money is needed, this is very much felt by the middle to lower-income groups who have low-income levels. To address this problem, the government provides assistance to students in the form of poor student assistance (BSM) in schools, starting from elementary to elementary with an equal high school. MA Hizbul Wathan NW Semaya Sikur is one of the schools located in the district of East Lombok, where some students are classified as middle to lower economic level, in determining who has the right to receive assistance from poor students, the school still subjectively makes decisions, making it difficult the school in analyzing students who are entitled to receive assistance from poor students (BSM). From these problems an appropriate data processing system needs to be made so that it can help the school in analyzing poor student aid recipients, so it is hoped that students who really need BSM can be on target. The author processes data using data mining principles. Data mining is widely used in various fields including retail, marketing, manufacturing, finance, communication, medical services, and others. Data mining itself is a technique used to process, extract (find patterns) and mine data from data whose information has not been known before[2]–[4]. In processing BSM data using the principle of data mining, the writer uses the Naive Bayes algorithm. Naive Bayes algorithm itself can process large amounts of data sets at high speed, Also, Naive Bayes algorithm can predict future opportunities based on past experience[5]–[8]. For this algorithm is chosen by hoping that the data processed can produce information that can be used as a reference to improve school policy, in providing assistance to poor students in the future.
2. Method

The method used in this study can be seen in the stages below:

![Research Model](image.png)

**Figure 1. Research Model**

2.1. Data collection

The method used in this study can be seen in the stages below. The method of data collection is done by using the literature study. A literature study is conducted to obtain information as reference material as well as authentic evidence. Research begins by collecting datasets. The dataset used in the study was sourced from data obtained from MA Hizbul Wathan NW Semaya Sikur, East Lombok Regency in the form of data on BSM recipient students in 2018 totaling 169 datasets.

2.2. Initial data processing

The initial data processing dataset that will be used is processed first. The dataset used is selected based on the required attributes with attribute selection. Furthermore, the dataset will be processed using the proposed method.

2.3. Proposed model

The model/method that will be proposed in this study is to use the Naive Bayes method. Naive Bayes is the classification method most widely used in the knowledge of probability and statistics and works based on the Bayes theorem\[9\]–\[11\]. The following is an overview of data processing using the Naive Bayes model using data mining principles.

![Proposed method](image.png)

**Figure 2. Proposed method**

The model starts from inputting the dataset, then determining the value of Validation using the X-Validation principle by using 10 fold-Validation. From the fold-validation value, the data partition is divided into two parts, namely training data and testing data, then the formation of an algorithm model using Naive Bayes. The two partitions are linked using the apply model command to test the model for
optimal values. Furthermore, the performance results of the Naive Bayes algorithm will be evaluated, the results of the model can be in the form of accuracy, precision, recall, ROC and AUC.

2.4. Experimentation and testing
At this stage, the test is carried out with the principles of data mining. Data mining can be defined as the process of finding knowledge from large data sets. To find knowledge in this process is done by several techniques and the results of the process are used to assist in decision making[12], [13]. To process data based on data mining principles, RapidMiner tools are used which are tools that provide an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics that are open source[14], [15]. RapidMiner has good data mining capabilities, it contains add-ons that make it possible to use different algorithms in this tool and others with operators that help develop modeling processes that apply to data analysis[16].

2.5. Evaluation and validation
In this last stage, evaluation and validation of the results of the application of the research model is carried out to determine the level of data accuracy of the proposed model, with the evaluation results in the form of accuracy, Confusion Matrix, ROC and AUC.

3. Results and Discussion
This test was conducted to determine the performance of the Naive Bayes method in classifying the classes that have been determined. Validation functions to divide the amount of training data on the data being tested. In the trial, carried out 10 times, one of the tests uses k-Fold Validation 10, which is dividing 10 parts into 169 data of students tested. The 10 sections consist of 9 parts training data and 1 part testing data. Following is an image of the data processing work area using RapidMiner to get accuracy results.

![RapidMiner work area with K-Fold Validation 10](image)

From the tests carried out 10 times using the Naive Bayes Algorithm, the best accuracy is 95.27%, this value can be classified as excellent classification. The following results of accuracy that have been tested can be seen in the image below.

![Accuracy results with K-Fold Validation 10](image)
Based on training data totaling 169 records, a confusion matrix is obtained from the Naive Bayes method, namely: The number of true negatives (TN) is 126 records classified as Ineligible and as many as 8 false-negative records are classified as Ineligible classes but in reality is a feasible class. Next 35 records for positive true (TP) are classified as Eligible, and 0 records for (FP) are classified as Eligible classes but apparently not Eligible and can be calculated to find the value of accuracy, precision, and recall. The results of the three can be seen in the equation below.

\[
\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \times 100\% \\
= \frac{35 + 126}{35 + 126 + 0 + 8} \times 100\% \\
= 92.75\%
\]

With the K-Fold Validation 10 test, there are 169 data to be processed and obtained an accuracy of 95.27%.

\[
\text{Precisi} = \frac{tp}{tp + fp} \times 100\% \\
= \frac{35}{0 + 35} \times 100\% \\
= \frac{35}{35} \times 100\% = 100\%
\]

From the above equation, we get a false-positive result (FP) which is the result of being feasible but apparently not feasible is 100.00%, this is because 35 students are selected from 35 students predicted.

\[
\text{Recall} = \frac{tp}{fn + tp} \\
= \frac{35}{8 + 35} \\
= \frac{35}{43} = 81.40\%
\]

From the above equation, we get a true positive result (TP), which is the result of feasible is 81.40%, this is because 35 students are selected from the 43 students predicted.

In addition to accuracy, precision and recall there are also results of calculations using the ROC (receiver operating characteristic) and AUC (area under the curve) curves. The calculation results are visualized with the ROC curve. A comparison of the two classes can be seen in the ROC curve image for the Naive Bayes algorithm. The ROC curve expresses the confusion matrix. Horizontal lines are false-positive and vertical lines are true positive. The area under the AUC curve is a prediction ordered by a score from highest to lowest. The optimistic version of the AUC divides positive results before dividing negative results. The optimistic AUC (area under the curve) results are 1,000 with the excellent classification value category.
Figure 5. The area under the curve (AUC) is optimistic using K-Fold Validation 10

Furthermore, the AUC is normal (neutral), where the normal version of the AUC is calculated by taking the average AUC optimistic and AUC pessimistic. Normal AUC can be seen in the image below:

Figure 6. The normal area under the curve (AUC) using K-Fold Validation 10

Horizontal lines are false-positive and vertical lines are true positive. AUC normally performs two or more of the same results, but the equation is not well defined. The normal AUC (area under the curve) results are 1,000 with an excellent classification value. Next measurement for pessimistic AUC values.

Figure 7. The area under the curve (AUC) pessimistic using K-Fold Validation 10

Horizontal lines are false-positive and vertical lines are true positive. It generates a normal AUC (Area Under Curve) value of 1,000 with an excellent classification value category. The pessimistic version of the AUC divides negative results before dividing positive results.

4. Conclusion

Based on the results of processing using the Naive Bayes method and conducting experiments using K-Fold Validation 10, it can be concluded that the processing of student data in MA Hizbul Wathan NW Semaya can be known as the results obtained by 95.27% of 169 student data with an AUC value of
This value is categorized as an excellent classification. From the results of data processing above, the method used can analyze the reception of poor student assistance very well.

5. References

[1] L. Hakim, “Pemerataan akses pendidikan bagi rakyat sesuai dengan amanat Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional,” EduTech J. Ilmu Pendidik. dan Ilmu Sos., vol. 2, no. 1, pp. 53–64, 2016.

[2] P. Kavipriya, “A Review on Predicting Students’ Academic Performance Earlier, Using Data Mining Techniques,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 6, no. 12, pp. 101–105, 2016.

[3] C. Anuradha and T. Velmurugan, “Feature Selection Techniques To Analyse Student Academic Performance Using Naïve Bayes Classifier,” 3rd Int. Conf. Small Mediu. Bus., no. February, pp. 345–350, 2016.

[4] M. Kumar, A. J. Singh, and D. Handa, “Literature Survey on Educational Dropout Prediction,” Int. J. Educ. Manag. Eng., vol. 7, no. 2, pp. 8–19, 2017, doi: 10.5815/ijeme.2017.02.02.

[5] H. Almarabeh, “Analysis of Students’ Performance by Using Different Data Mining Classifiers,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 8, pp. 9–15, 2017, doi: 10.5815/ijmecs.2017.08.02.

[6] H. Sharma and S. Kumar, “International Journal of Computer Science and Mobile Computing PREDICTION OF MU STUDENT’S PERFORMANCE USING DATA MINING TECHNIQUE,” Int. J. Comput. Sci. Mob. Comput., vol. 6, no. 5, pp. 405–417, 2017.

[7] I. B. A. Peling, I. N. Arnawan, I. P. A. Arthawan, and I. G. N. Janardana, “Implementation of Data Mining To Predict Period of Students Study Using Naïve Bayes Algorithm,” Int. J. Emerg. Technol., vol. 2, no. 1, p. 53, 2017, doi: 10.24843/ijjet.2017.v02.i01.p11.

[8] E. Sugiharti, S. Firmansyah, and F. R. Devi, “Predictive evaluation of performance of computer science students of unnes using data mining based on naïve bayes classifier (NBC) algorithm,” J. Theor. Appl. Inf. Technol., vol. 95, no. 4, pp. 902–911, 2017.

[9] S. Mukherjee and P. K. Bala, “Sarcasm detection in microblogs using Naïve Bayes and fuzzy clustering,” Technol. Soc., vol. 48, pp. 19–27, 2017, doi: 10.1016/j.techsoc.2016.10.003.

[10] H. Zhang, Z. X. Cao, M. Li, Y. Z. Li, and C. Peng, “Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals,” Food Chem. Toxicol., vol. 97, pp. 141–149, 2016, doi: 10.1016/j.fct.2016.09.005.

[11] P. Tsangaratos and I. Ilia, “Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size,” Catena, vol. 145, pp. 164–179, 2016, doi: 10.1016/j.catena.2016.06.004.

[12] C. Oliveira et al., “ScienceDirect ScienceDirect The 3rd International Workshop on Healthcare Interoperability and Pervasive Intelligent Systems Benchmarking Business Analytics Techniques in Big Data Benchmarking Business Analytics Techniques in Big Data,” Procedia Comput. Sci., vol. 160, pp. 690–695, 2019, doi: 10.1016/j.procs.2019.11.026.

[13] J. Silva, J. Silva, L. Hernandez, T. Romero, O. Bonerge, and P. Lezama, “ScienceDirect ScienceDirect Big Data Aplication for Selecting Theses Topics Big Data Aplication for Selecting Theses Topics,” Procedia Comput. Sci., vol. 160, pp. 538–542, 2019, doi: 10.1016/j.procs.2019.11.051.

[14] H. A. Cuesta, D. L. Coffman, C. Branas, and H. M. Murphy, “Using decision trees to understand the influence of individual- and neighborhood-level factors on urban diabetes and asthma,” Heal. Place, vol. 58, no. April, p. 102119, 2019, doi: 10.1016/j.healthplace.2019.04.009.

[15] A. Naik and L. Samant, “Correlation Review of Classification Algorithm Using Data Mining Tool: WEKA, Rapidminer, Tanagra, Orange and Knime,” Procedia Comput. Sci., vol. 85, no. Cms, pp. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

[16] J. Rodriguez et al., “ScienceDirect ScienceDirect Determining Student Interactions in a Virtual Learning Determining Student Interactions in a Virtual Learning Environment Using Data Mining Environment Using Data Mining,” Procedia Comput. Sci., vol. 155, no. 2018, pp. 587–592, 2019, doi: 10.1016/j.procs.2019.08.082.