

Prediction of New Student Registration at Universitas Muhammadiyah Yogyakarta using the Naïve Bayes Classification Algorithm

Lusiana Ning Saputri*, Asroni, and Slamet Riyadi

Universitas Muhammadiyah Yogyakarta, Jln.Brawijaya, Tamantirto, Kasihan, Bantul, Yogyakarta 55183, Indonesia

**Corresponding author: lusiana.ning.2012@ft.umy.ac.id*

Abstract

All public and private universities perform new student admission each year, with the number of applicants reaching the thousands. Prospective new students who intend to continue and meet the higher education criteria must re-register according to the university's timetable after enrolling through their selected pathway and being pronounced passed. Due to the movement of prospective students to different departments or universities, the number of applications typically does not match the number of individuals who have re-registered. If the probability of a new student candidate departing can be discovered early, higher education management can take the necessary steps to retain the prospective student. Data mining and the Naïve Bayes algorithm were employed to analyze the data. The information was extracted from Universitas Muhammadiyah Yogyakarta's database of Information Technology freshmen applicants for 2016-2017. Microsoft Excel, RapidMiner, and SQL Server 2014 Management Studio were utilized.

Keywords: *Data Mining, Naïve Bayes, RapidMiner, Re-registration Prediction.*

1. Introduction

Universitas Muhammadiyah Yogyakarta (UMY) is located at South Ring Road, Kasihan, Bantul, the Special Region of Yogyakarta (DIY), with an A accreditation status under BAN-PT Decree No. 5237/SK/BAN-PT/Akred/PT/XII/2017. UMY conducts new student admission annually, with thousands of applicants registering. After enrolling through their selected pathway and being announced passed, students who decide to continue their studies and fulfill higher education standards must re-register at the time designated by the university. However, the number of applications frequently does not match the number of re-registered students because prospective students have transferred to other departments or universities.

The growth of information technology in systems and data processing is accelerating and becoming more precise. A server-based storage solution can effectively manage the thousands of new student data added annually. Using data mining, it is feasible to extract old data to generate new information for future needs. The number of registrants at a university can be utilized to predict the number of students who will re-register. In order for the university to anticipate future possibilities regarding the number of candidates. Suppose early detection of a potential new student's likely resignation is feasible. In that case, the higher education administration can take the required steps to retain the prospective student[1]. According to the Gartner Group, data mining is discovering significant links, patterns, and trends by analyzing massive data sets in a database using pattern recognition techniques such as statistics and mathematics[2]. Naïve Bayes is a straightforward probability classifier based on Bayes' theorem[3].

In the study of the Bayesian Classification Algorithm for Predicting New Student Registration at Stikes Hang Tuah Pekanbaru, it was found that from the results of registration data for new students who registered based on Pekanbaru City, Midwifery Department, Diploma 3 and Glombang 1, it can be seen that Posteriors who register are larger than those who do not register. So with these results it can be seen that the student will do the re-registration [4]. Previous research entitled Bayesian Classification Algorithm for Predicting New Student Registration at Stmik Widya Pratama, using the Naïve Bayes Algorithm in Data Mining Techniques, found that the number of registrants who registered was 658 people, and those who did not register were 255 people, out of a total of 913 people. who did the registration[5].

Naïve Bayes is a machine learning method that utilizes probability and statistical calculations proposed by the British scientist Thomas Bayes, namely predicting future probabilities based on previous experience. Bayes' theorem is an approximation to uncertainty as measured by probability[6].

The process of data mining is incoming data from various sources, this is integrated and placed in a data storage area or can be called a database. then taken and analyzed. After that it is passed to a data mining algorithm which produces output in the form of rules or some other type of 'pattern'. These are then interpreted to provide new discovery knowledge and potentially useful knowledge[7].

This study employs the Naïve Bayes classifier method to predict the data on re-registered students at UMY.

2. Method

2.1. Place and Time of Research

This study was performed in the Bureau of Information System (BSI) room of the AR Fachruddin B building at UMY from May 2019 to November 2020.

2.2. Tools and Materials

This study utilized additional tools to simplify research and data implementation.

A. Research Tools

The research tools consisted of hardware and software.

a. Hardware

1. Personal Computer
2. Intel i3-3217U 1.80 GHz with Nvidia GeForce 720M Cuda 2Gb
3. 4 Gb Random Acces Memory (RAM)

b. Software

1. Operating system windows 8.1 64 bit
2. RapidMiner
3. SQL Server 2014 Management Studio
4. Microsoft Office 2010

2.3. Research Flow

This study applied the Software Development Life Cycle (SDLC) waterfall model [8]. The flow is depicted in Figure 1. Waterfall model is very common model in software development. It needs results certainty in each steps so that going back to previous step is unnecessary except some condition applied it [9]. Thus, it suits well for this research work.

2.4. Literature Study

The initial phase of this research was a literature study. It was undertaken to gather the necessary information for the research by reviewing relevant literature. The use of Naïve Bayes Algorithm are popular in the research that involve academics data. In 2017, Makhtar et. al. [10] used the algorithm to analyze on student performance in one of Malaysian university. The work also applied 10 fold cross-validation and resulted 73.4 accuracy.

The second work that used Naïve Bayes Algorithm was performed by Maitra et. al. [11] to analyze the authenticity of student feedback for faculty. They presented a proactive and outcome-based faculty feedback analysis model that employs the Nave Bayes Classifier to

sift and categorize the feedback provided by the students into invalid or valid categories based on the relative effect of the aforementioned quality features on the feedback measure.

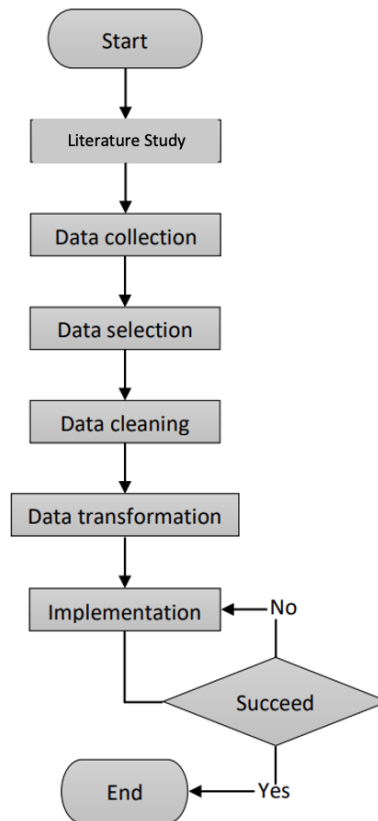


Figure 1 Research Flowchart

2.5. Data Collection

The second phase was data collection. It is crucial since it can influence the implementation process and the findings. The data were obtained from the BSI of UMY.

2.6. Data Selection

Thousands of general data required analysis and filtering to obtain more specific data before the existing database could be processed.

2.7. Data Cleaning

Data cleaning is the detection and correction of inaccurate and incomplete data. Data extracted from a database contains imperfect information, including missing, invalid, and duplicate information. This study possessed incorrect and unnecessary data attributes, necessitating data cleaning.

2.8. Data Transformation

During data transformation, the data were modified or combined into a format appropriate for data mining. This study employed Microsoft Excel to convert the data into CSV format. RapidMiner software could expedite the mining process through data transformation.

2.9. Implementation

Implementation is the primary procedure for discovering data. This study employed the Naïve Bayes algorithm and RapidMiner. RapidMiner was utilized to reprocess the data gathered from data mining to generate more precise analysis results. This study utilized training data containing Information Technology Department registrants in 2016, with the attributes of gender, province of origin, and registration pathway. Meanwhile, the data testing encompassed the same attributes of registrants in 2017. The prediction analysis applying RapidMiner and the Naïve Bayes algorithm yielded percentage-based findings.

3. Implementation and Discussion

3.1 Data Collection

This study utilized a data warehouse for UMY's students of the Information Technology Department Class of 2016. SQL Server 2014 Management was employed to access the data warehouse. Data warehouse access rights were acquired by making a copy of the BSI, along with a written approval letter from the BSI.

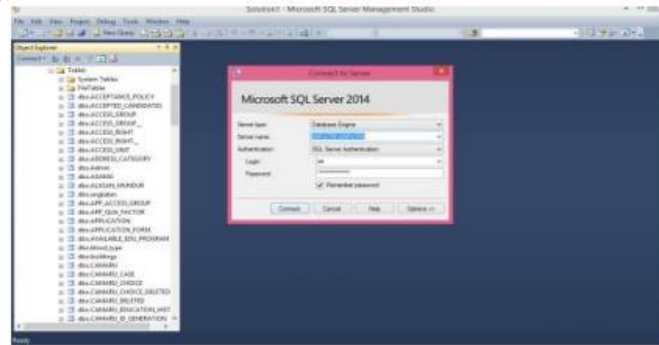


Figure 2 Entering the Database Server

In SQL Server 2014 Management, the database server must be connected as the initial step. A view to facilitating the retrieval of data was developed.

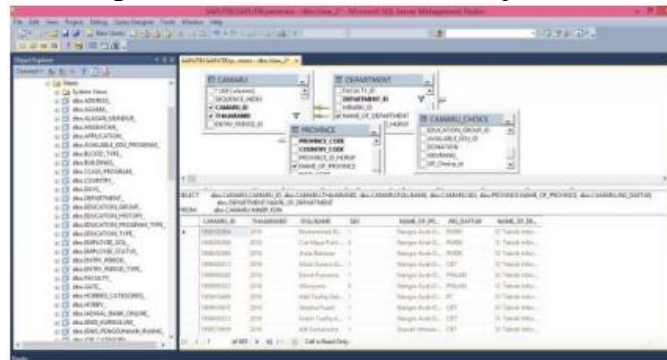


Figure 3 View Data from Training Data of the Information Technology Department Class of 2016

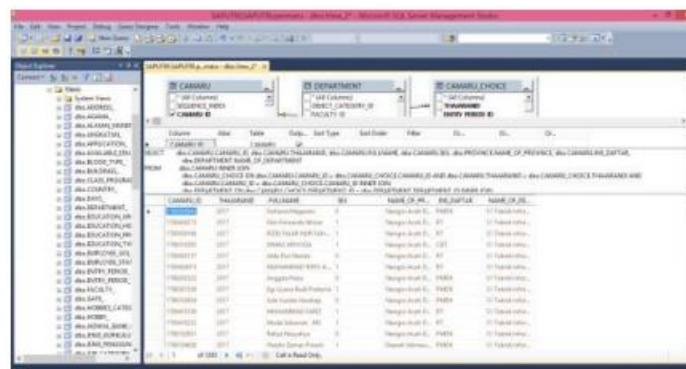


Figure 4 View Data from Testing Data of the Information Technology Department Class of 2017

The data comprised prospective students of the Information Technology Department Class of 2016 who had and had not re-registered, totaling 805 registrants. In addition, the 2017 Class of the Information Technology Department possessed 1,293 prospective new students. These were pure data from the database of UMY.

3.2 Data Selection

Data Selection is the process of selecting data for analysis from a database based on research requirements. The data possessed the attribute names sex, province name, and non-subsidized pathway, derived from data collected using View Data in SQL Server Management. Sex is the gender, the province name represents the province of origin, and the non-subsidized pathway refers to the registration pathway prospective new students took. To simplify research, the labels of these attributes were altered; specifically, sex became gender, province name became province, and non-subsidized pathway became registration pathway.

3.3 Data Cleaning

Collecting data necessitated a cleaning procedure to prevent data duplication before it could be mined for insights. After the data had been thoroughly cleaned, they were saved in a new Microsoft Excel CSV-formatted dataset. The data extracted from the Information Technology Department contained null data; thus, it must be eliminated because the data mining system employing the Naïve Bayes method could not analyze null data.

In training data, 246 records were discovered to include the same data. However, there were 347 records in testing data, resulting in the registrant registering several times and entering incorrect data.

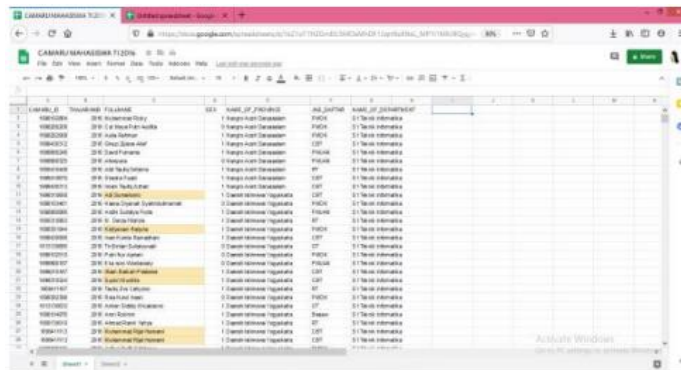


Figure 5 Conditional Formatting 1 in Training Data

Conditional formatting is a feature used to automatically modify cell or range formats in Microsoft Excel or Google Spreadsheets based on specified criteria. This study employed criteria based on the number of FullName column cells containing identical data.

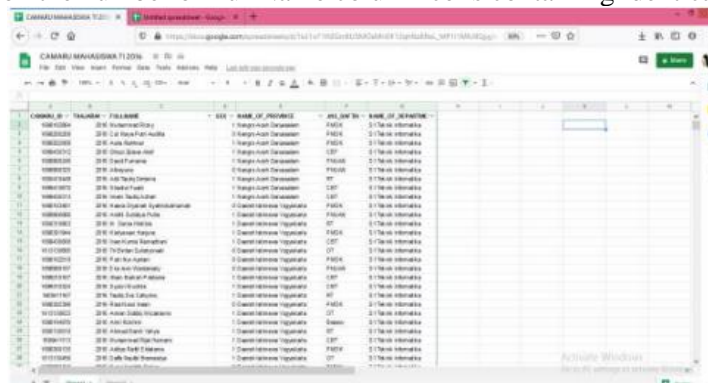


Figure 6 Conditional Formatting 2 in Training Data

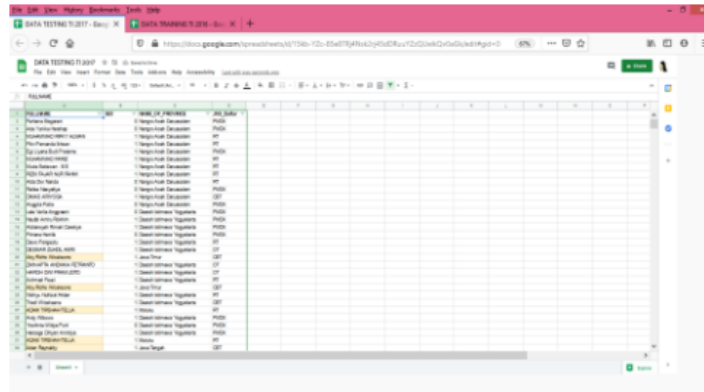


Figure 7 Conditional Formatting 1 di Testing Data

Conditional formatting is a feature to automatically change cell or range formats in Microsoft Excel or Google Spreadsheets under determined criteria. This study utilized criteria following the number of FullName column cells containing identical data.

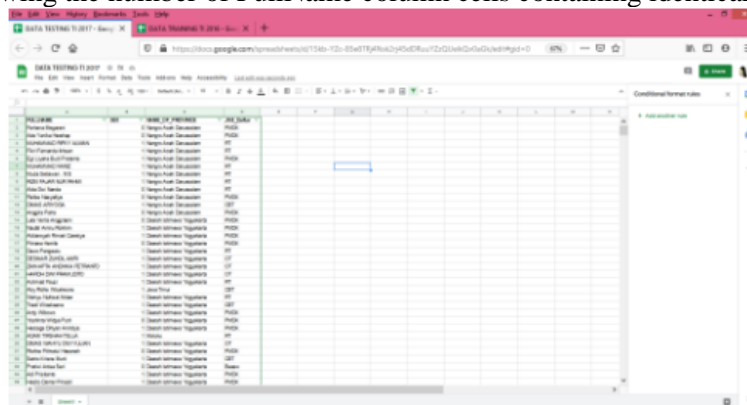


Figure 8 Conditional Formatting 2 di Testing Data

Sorting duplicate training data resulted in a reduction of 706 records and the production of 1,169 records of testing data.

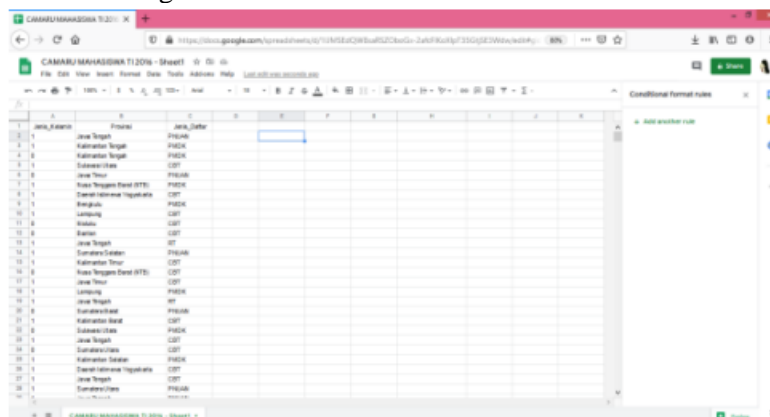


Figure 9 Attributes of Gender, Province, and Registration Pathway of Training Data

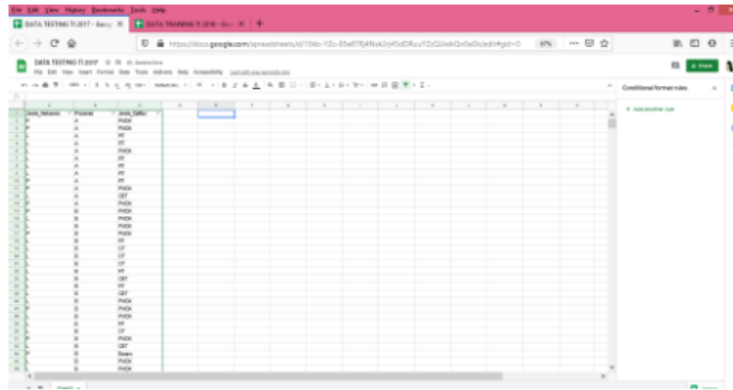


Figure 10 Attributes of Gender, Province, and Registration Pathway of Testing Data

Attributes used to be processed in the application were first initialized.

Table 1. Gender Initialization

Gender	Initialization
0 (Female)	F
1 (Male)	M

Table 2 Gender Initialization

Province of Origin	Initialization
Sumatera, Riau, Bangka Belitung	A
Jawa, Bali, Madura	B
Kalimantan	C
Sulawesi	D
NTB, NTT, Papua	E
Overseas	F

3.4 Data Transformation

This study processed the data through the SQL Server 2014 Management Studio database. Then, they were converted into a CSV file for data processing in the RapidMiner software. Data attribute names sex, province name, and non-subsidized pathway were changed to gender, province, and registration pathway.

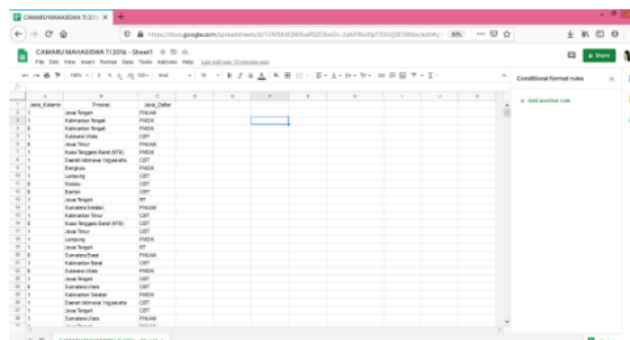


Figure 11 Data Transformation

3.5 Implementation

This study utilized probability (prediction) based on the Naïve Bayes algorithm. This method could better predict possibilities by comparing past experiences.

Using the RapidMiner software, training and testing data were divided and evaluated separately. The Information Technology Department yielded 706 training and 1,169 testing data records.

After selection, cleaning, and initialization, the following training data in CSV format were prepared for data mining.

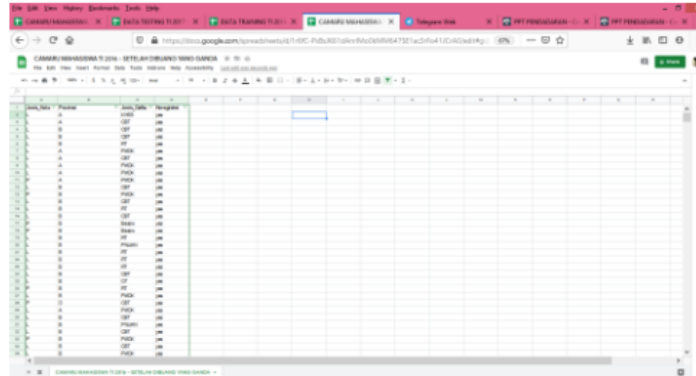


Figure 12 Training Data in CSV Format

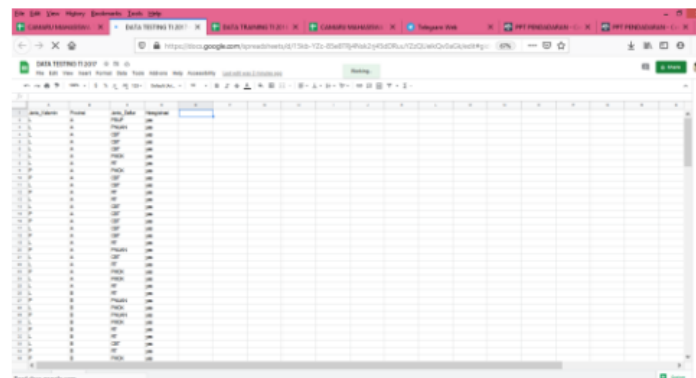


Figure 13 Testing Data in CSV Format

3.6 RapidMiner Software Testing

In this study, the re-registration attribute served as the label. The projections of prospective new students who would re-register based on Information Technology Department information were examined. The data must be in CSV format to be processed by the Rapidminer software.

After the data were converted to CSV format, importing was performed. It was accomplished by including the read CSV operator in the view process (Import - Data - Read CSV), followed by renaming training and testing data through drag-and-drop, as displayed in the following figures.

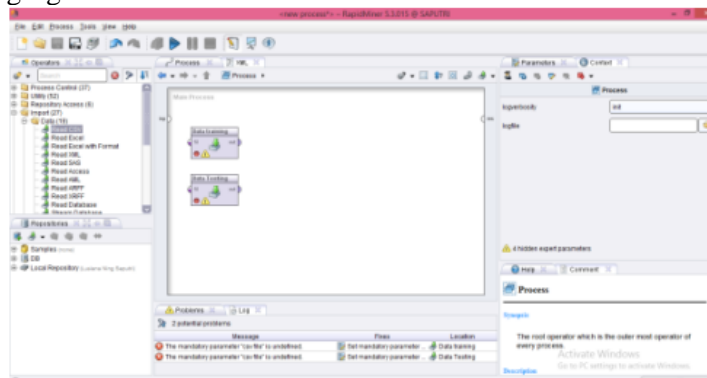


Figure 14 Drag and Drop and the Read CSV Operator

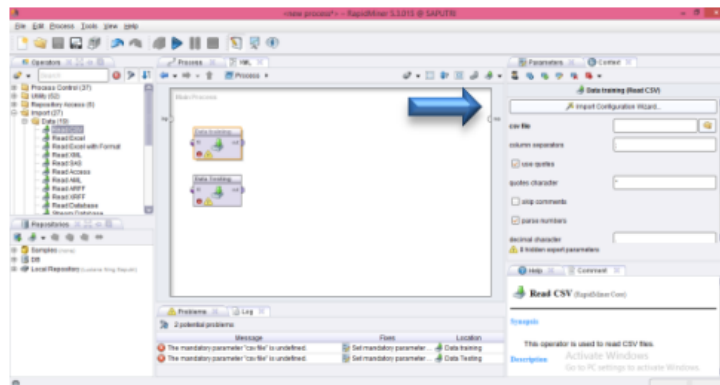


Figure 15 Importing Training File

The Import Configuration Wizard must be selected to bring up the Data Import Wizard form of the first step. In this step, the file location of the training data must be determined.

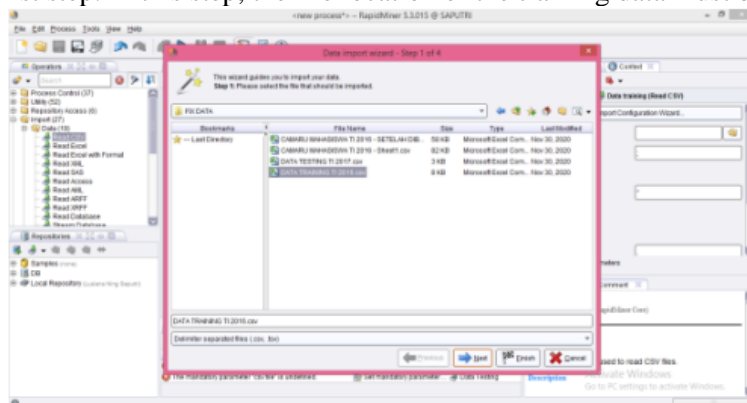


Figure 16 Data Import Form of Training Data

After selecting the training data, Next was clicked, and the Data Import Wizard form from the second step appeared.

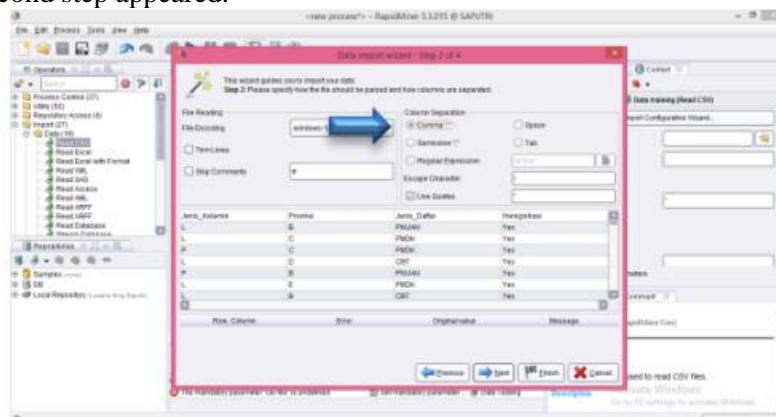


Figure 17 Data Import Wizard of Step 2 of Training Data

Then, the second step form appeared, as illustrated in Figure 3.16. Then, Column Separation Comma was clicked. The comma separated one attribute from another since the data would merge if the file were in CSV format. By clicking Next, the third step form for the Data Import Wizard emerged.

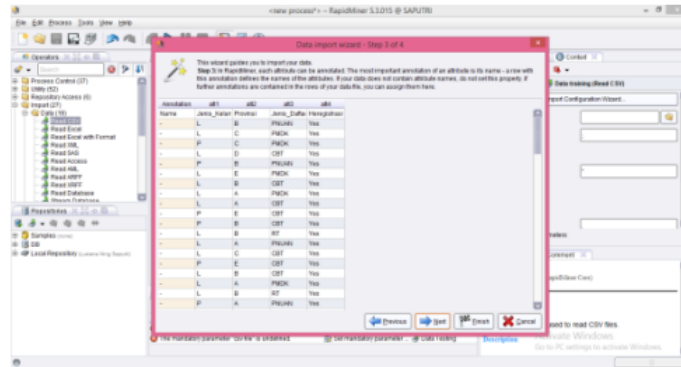


Figure 18 Data Import Wizard of Step 3 of Training Data

In the third step, the contents of the training data appeared. After clicking Next, the Data Import Wizard of the fourth step emerged. Then, the attribute column serving as the label was selected and modified. The re-registration attribute was designated as the label attribute. Afterward, Finish was clicked, and the process of importing training data was complete. Subsequently, testing data were inputted in the same manner as training data.

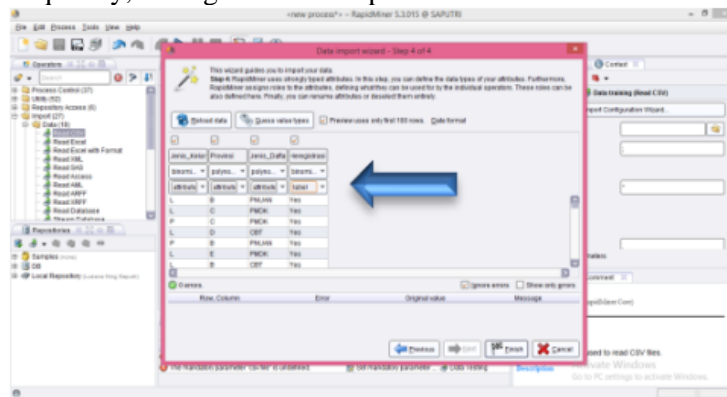


Figure 19 Data Import Wizard of Step 4 of Training Data

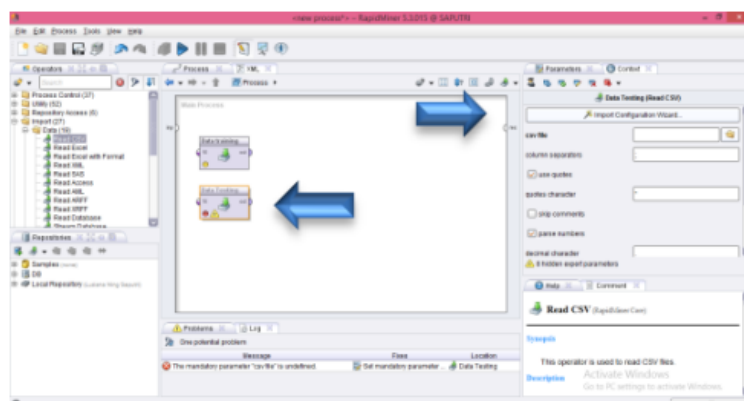


Figure 20 Importing Testing File

The Import Configuration Wizard was selected to bring up the Data Import Wizard form of the first step. During this step, the file location of the testing data for the research process was selected.

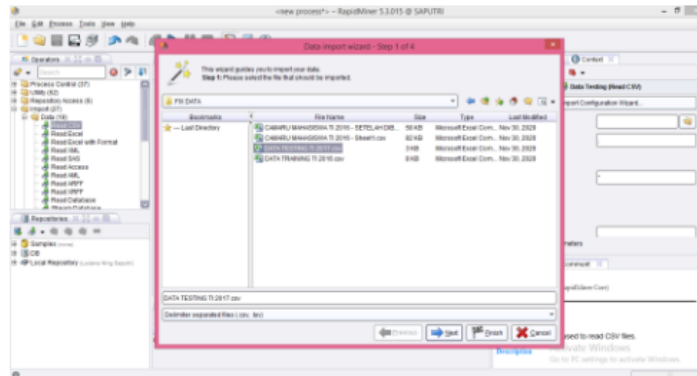


Figure 21 Import Data Form of Testing Data

After selecting data testing, Next was clicked, and the Data Import Wizard form of the second step appeared.

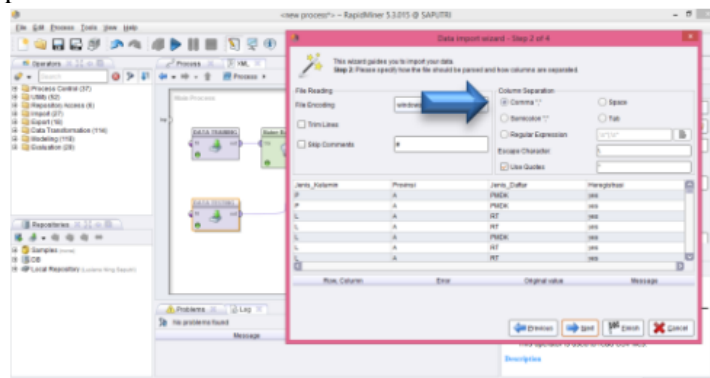


Figure 22 Data Import Wizard Form of Step 2 of Testing Data

Then, the second step form appeared, as depicted in Figure 3.21. Then, Column Separation Comma was selected. The comma separated one attribute from another since the data would merge if they were in CSV format. Subsequently, Next was clicked, and the Data Import Wizard form of the third step appeared.



Figure 23 Data Import Wizard Form of Step 3 of Testing Data

In the third step, the contents of the testing data appeared. After clicking Next, the Data Import Wizard of the fourth step emerged. The attribute column serving as the label was determined and modified. The re-registration was selected as the label attribute. Then, Finish was clicked, and the procedure of importing testing data was complete.

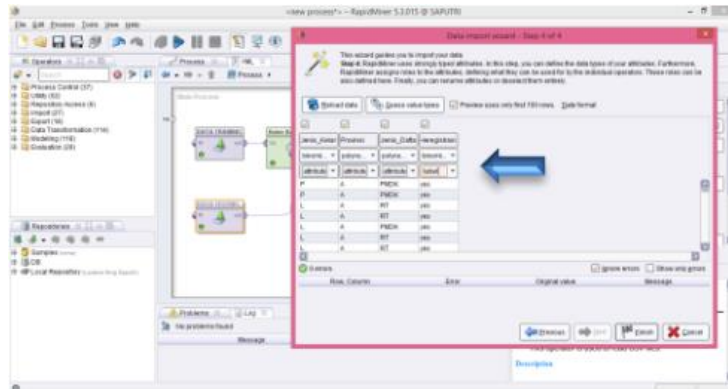


Figure 24 Data Import Wizard Form of Step 4 of Testing Data

The last step was to transfer the Naïve Bayes model into the operator, serving as the research method. Drag and drop was employed to navigate to the processes view. Afterward, Modeling – Classification and Regression – Bayesian Modeling – Naïve Bayes were clicked. As demonstrated in Figure 3.20, the read CSV of Training Data was linked to the Naïve Bayes operator. The Naïve Bayes algorithm was applied to process the data.

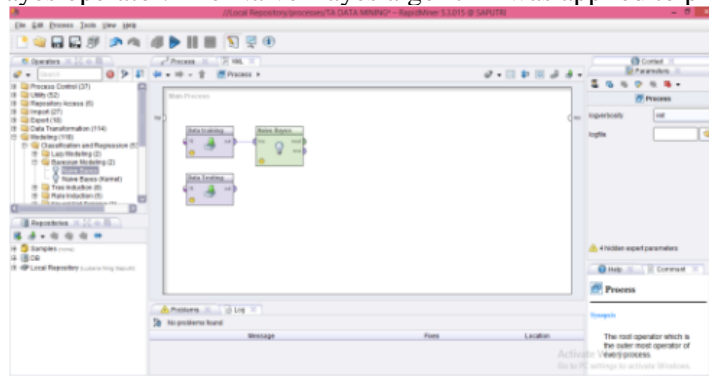


Figure 25 Read CSV File of Training Data Connected to the Naïve Bayes Operator

The subsequent step was to incorporate the Apply model operator into the process view, performed by selecting Application Model – Apply Model. Furthermore, the Naïve Bayes operator mod port was connected to the Apply model port, and the read CSV output port of testing data was to the UNL Apply model port.

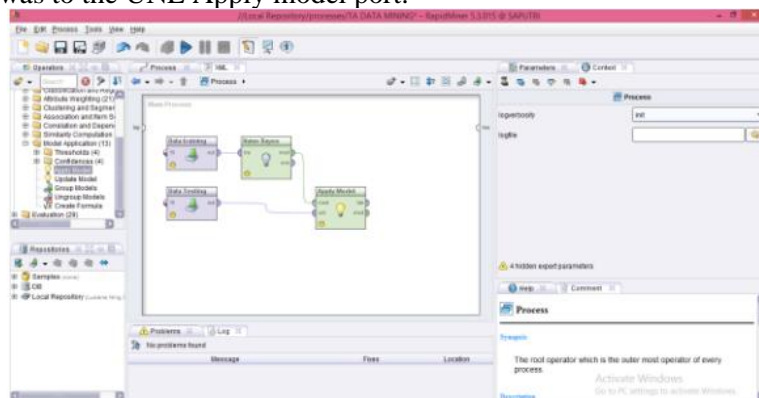


Figure 26 The Naïve Bayes Model and Read CSV of Testing Data Connected to the Apply Model

The Apply model served as a repository for the outcomes of data computations processed using the Naïve Bayes algorithm, subsequently confirmed by data testing. Therefore, the objective of predictive data testing was to calculate the training data results.

It was performed by selecting Evaluation - Performance Measuring - Performance as the next step to add a performance operator. This operator was applied to determine the amount of precision and inaccuracy associated with the computation results in the Apply model, as illustrated in Figure 3.26.

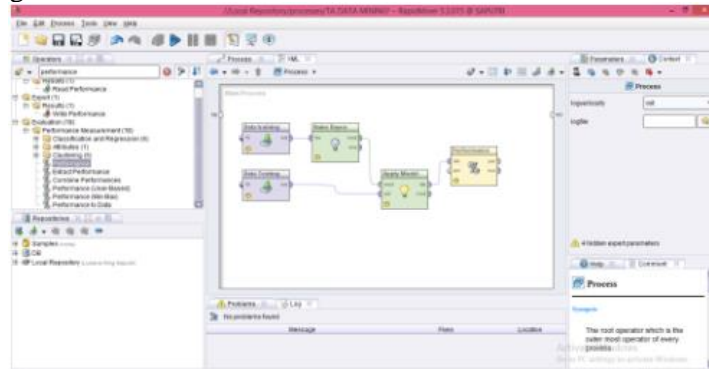


Figure 27 Operator Performance

The lab output of the Apply model was connected to the lab input for operator performance. After everything had been hooked, the per and exa ports on the operator performance were connected to the res port on the process view's right side.

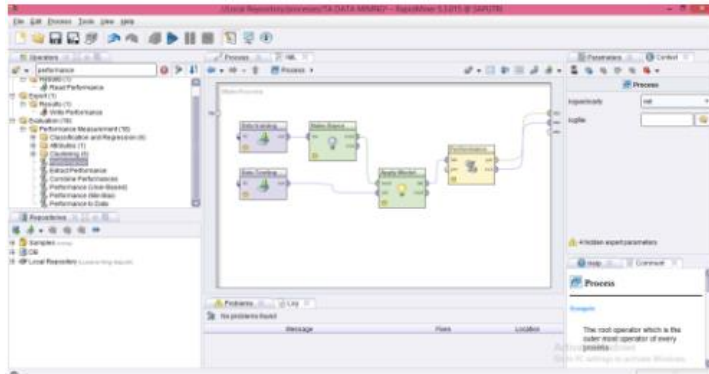


Figure 28 The Operator Performance was Connected to the Res Port in the Process View

After everything was connected, the run button on the toolbar was selected. The calculation results appeared in a few moments.



Figure 29 Running the RapidMiner

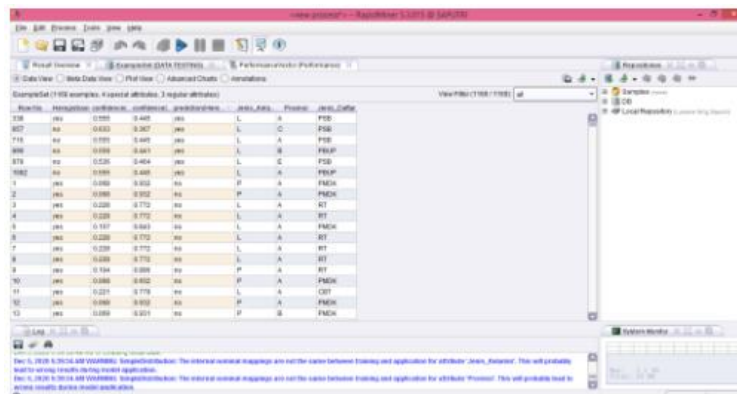


Figure 30 Calculation Results

RapidMiner software calculation with the prediction method from testing and training data resulted in a prediction column (Re-registration). This column provided statistics on prospective new students who re-registered (Yes) and did not (No). The performance vector was selected to determine the accuracy of the Naïve Bayes algorithm.

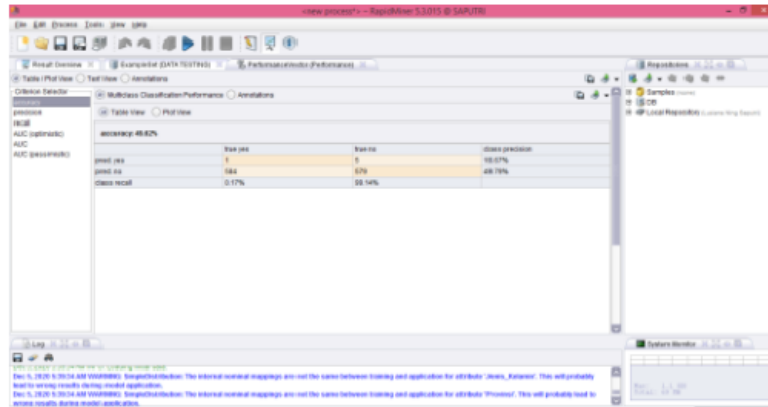


Figure 31 Accuracy

3.7 Naïve Bayes Algorithm

The following data were generated using the Naïve Bayes algorithm.

Table 3 Training Data of Prospective Students of Information Technology Class of 2016

No.	Gender	Province	Registration Pathway	Re-registration
1	M	B	PNUAN	Yes
2	M	C	PMDK	Yes
3	F	C	PMDK	Yes
4	M	D	CBT	Yes
5	F	B	PNUAN	Yes
...	M	B	CBT	Yes
...	M	A	PNUAN	No
...	M	E	CBT	No
706	M	B	RT	No

Table 4 Testing Data of Prospective Students of Information Technology Class of 2017

No.	Gender	Province	Registration Pathway	Re-registration
1	F	A	PMDK	Yes
2	F	A	PMDK	Yes
3	M	A	RT	Yes
4	M	A	RT	Yes
5	M	A	PMDK	Yes
...	M	A	PNUAN	No
...	M	A	CBT	No
...	M	A	RT	No
1169	M	A	CBT	No

This study utilized 706 records from Table 3.3 as training data and 1,169 from Table 3.4 as testing data, with 50% Yes data included in testing data. The following formula was

applied to calculate the probability value or prediction of prospective new students who would re-register.

$$P(H|X) = \frac{P(X|H) \cdot P(H)}{P(X)}$$

The first step was determining the number of Yes and No training data, generating the following results.

- Yes: 103
- Male: 81
- Female: 22
- Province B: 58
- Registration Pathway of PMDK: 17
- No: 603
- Male: 396
- Female: 207
- Province A: 155
- Registration Pathway of RT: 84
- Total: 706 records

After determining the number of prospective new students who re-registered and did not register, the probability of each attribute from testing data was calculated as follows.

- Gender = Male, Province = B, Registration Pathway = PMDK
- $F(\text{Re-registration} = \text{Yes} / \text{Number of Students of Training Data}) = 103 / 706 = 0.146$
- $F(\text{Gender} = \text{Male} = \text{Yes} / \text{Re-registration}) = 81 / 103 = 0.787$
- $F(\text{Province} = \text{B} = \text{Yes} / \text{Re-registration}) = 58 / 103 = 0.563$
- $F(\text{Registration Pathway} = \text{PMDK} = \text{Yes} / \text{Re-registration}) = 17 / 103 = 0.165$

Subsequently, the probability of No was calculated using the same testing data as the calculation of Yes.

- Gender = Male, Province = A, Registration Pathway = RT
- $P(\text{Re-registration} = \text{No} / \text{Number of Students of Training Data}) = 603 / 706 = 0.854$
- $P(\text{Gender} = \text{Male} = \text{No} / \text{Not Re-registered}) = 396 / 603 = 0.657$
- $P(\text{Province} = \text{A} = \text{No} / \text{Not Re-registered}) = 155 / 603 = 0.257$
- $P(\text{Registration Pathway} = \text{RT} = \text{No} / \text{Not Re-registered}) = 84 / 603 = 0.140$

4. Conclusion

The conclusion obtained from the research is that the Naïve Bayes Algorithm can be used to predict the enrollment of prospective new students at Yogyakarta Muhammadiyah University. The Naïve Bayes algorithm in predicting the registration of prospective new students has an accuracy of 49.62%, and the information obtained in this study is that prospective new students who are predicted to register at the time are students who are male, students from provinces outside Java, and students with a non-scholarship registration path.

References

- [1] Kusrini, & Emha T. Luthfi., 2009, Data Mining Algorithm, Andi Offset, Yogyakarta. (*in Indonesian*)
- [2] Larose D.T, Discovering Knowledge in Data. New Jersey : John Willey & Sons, Inc, 2005.
- [3] Prasetyo, E. (2012). Data Mining Concepts and Applications using Matlab. Yogyakarta: Andi. (*in Indonesian*)
- [4] HERI, S. (2017). Bayesiyan Classification Algorithm for Predicting New Student Registration at Stikes Hang Tuah Pekanbaru (Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau). (*in Indonesian*)

- [5] Sugianti, D. (2012). Bayesian Classification Algorithm for Predicting New Student Heregration at STMIK Widya Pratama. *Jurnal Ilmiah ICTech*, 10(2), 1-5. *(in Indonesian)*
- [6] Febrian, F. (2011). Comparative Analysis of Data Mining Classification Algorithms on Life Reinsurance Facultative Data Acceptance. STMIK Eresha, Jakarta, MasterThesis. *(in Indonesian)*
- [7] Bramer, M. (2007). Principles of data mining (Vol. 180). London: Springer.
- [8] Bassil, Y. (2012). A simulation model for the waterfall software development life cycle. arXiv preprint arXiv:1205.6904.
- [9] Ruparelia, N. B. (2010). Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3), 8-13.
- [10] Makhtar, M., Nawang, H., & WAN SHAMSUDDIN, S. N. (2017). ANALYSIS ON STUDENTS PERFORMANCE USING NAÏVE BAYES CLASSIFIER. *Journal of Theoretical & Applied Information Technology*, 95(16).
- [11] Maitra, S., Madan, S., Kandwal, R., & Mahajan, P. (2018). Mining authentic student feedback for faculty using Naïve Bayes classifier. *Procedia computer science*, 132, 1171-1183.