Implementation of Naïve Bayes Classifier-based Machine Learning to Predict and Classify New Students at Matana University

Simon Prananta Barus

Matana University, ARA Center CBD Barat Kav. 1st ,Tangerang 15810, Indonesia

Email: simon.barus@matanauniversity.ac.id

Abstract. Data is growing fast, triggered by the wider use of information technology (IT) by the public or organizations. In line with data growth, data processing is also developing. One of them is data mining. Data mining is very necessary in business to obtain the information that needed to make business strategy such as marketing strategy. At Matana University, the Marketing Department needs to utilize data mining to optimize the achievement of its targets, particularly to predict and classify the prospective student data. Current data mining processing machines already have the high performances with large storage capacities. To improve the performance of data mining processing machines, machine learning is applied. The machine becomes intelligent and able to learn from the provided data. The application of machine learning to predict and classify new students is based on supervised learning by applying the Naïve Bayes Classifier (NBC) algorithm. The data used is the data of prospective students who have registered at Matana University. Machine learning is built, using the Python programming language. The result of the application has an accuracy of 0.73 (73%) and very helpful to the head of marketing in making marketing strategies. In the future, several developments can be done such as using other algorithms, accessed by smartphone, using dashboards in visualization, or adding data attributes such as parental income, religion, hobby, future goal and so on.

1. Introduction

Advances in information technology (IT) and its utilization by the community and organizations results in data being easily obtained and growing fast (becoming large). Data is the raw material to produce information. Before it becomes useful information, data needs to be processed. Currently to process it, we can utilize data mining, where data will be extracted to obtain more valuable and useful information [1]. There are two main tasks of data mining [2]. There are predictive (consisting of classification, prediction and time-series analysis) and descriptive (consisting of association, clustering and summarization). The computer now has high performance and large storage capacity so that the data mining process can run smoothly.

This research is leaning to predictive by using the Naïve Bayes Classifier (NBC) algorithm. NBC algorithm is an algorithm based on Bayes theory (invented by Thomas Bayes). The theory is as follows [3]

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

Legend:
P (A | B): the posterior probability
P (B | A): the likelihood
P (A): prior probability
P (B): predictor prior probability

This study uses the NBC algorithm because it is supervised algorithm, the data used is not large and reliable [4]. NBC is also used for machine learning.

Machine learning is a subset of Artificial Intelligence (AI) [5]. Machine learning has three types of supervised learning, unsupervised learning and reinforcement learning [6]. Supervised learning has labeled data and the data is predetermined. Applications that implement machine learning have the intelligence and ability to learn from the data provided. In this study using supervised learning machine that consisting of regression and classification. The NBC algorithm includes an algorithm for supervised learning. Many researches have been done on the use of machine learning based on the Naïve Bayes Classifier, but it is rarely used in marketing at the University, just similar [7][8].

Marketing is an important thing in a business. Currently, marketing enters the 4.0 era [9] where there is a shift from traditional to digital. In the future there will be a very significant change in marketing due to the role of artificial intelligence [10]. Matana University is a private tertiary institution located in Gading Serpong, Tangerang, Banten province, established in 2014. Matana University has a marketing department that has a target to get new students. At present the marketing department has not utilized data mining or machine learning, specifically in predicting and classifying prospective student data, to achieve its targets.

Making an application using the Python programming language is because it is easy to learn, having many libraries to support data analysis and growing rapidly [11]. Some libraries that used for machine learning are Numpy, Panda, Sklearn and etc. [12].

2. Methodology

The research method used are:

2.1. Literature study

The Literature study is carried out to obtain basic concepts and find out the extent of related research previously carried out. The literature studied is data mining, the Naïve Bayes Classifier algorithm, machine learning, marketing and Python programming.

2.2. Data collection

The data collected is student data at Matana University. This data was obtained from the marketing department in Excel form. Student data attributes used as in table 1.

Table 1. Student data attributes.

Name	Sex	Origin School	Address	Major	Study Program

2.3. Data analysis

For selecting data, the student data used is the student data for the 2018 and 2019 classes. Smoothing of data, is done by trimming data from different formats in each excel, deleting data where duplication of data occurs or incomplete attributes. Correcting some data that has a typo. Data transformation is carried out so that it can be processed using the Naïve Bayes Classifier algorithm. The results of the transformation attribute will be processed into as in table 2.

Table 2. The results of the transformation attribute.

Frequency	JaBoDeTaBek	Major	Study Program	Admission Status

The frequency indicates the number of schools (SMA / K) the students become students at Matana, table 3:

Table 3. The frequency.

Frequency	Remarks
3	More than twice
2	Once
1	None

JaBoDeTaBek (Jakarta Bogor Depok Tangerang Bekasi) shows the location of the prospective student's high school, table 4.

Table 4. JaBoDeTaBek.

JaBoDeTaBek	Remarks
2	Yes, the location of the school in the areas of Jakarta, Bogor, Depok, Tangerang and Bekasi
1	No, the location of the school is outside JaBoDeTaBek

The admission status shows three potential levels of the prospective students, namely very potential, potential and less potential.

The major consists of two, namely IPA or IPS, table 5.

Table 5. Major.

Major	Remarks
1	IPA, related to exact science
2	IPS, related social science

From the results of the transformation the following dataset is obtained, table 6.

Table 6. The results of the transformation.

Frequency	JaBoDeTaBek	Major	Study Program	Admission Status
2	2	1	1	Very Potential
2	2	1	4	Very Potential
2	2	1	9	Very Potential
2	2	1	10	Very Potential
2	2	1	7	Very Potential
2	2	1	6	Very Potential
2	2	1	3	Very Potential
2	2	2	9	Very Potential
2	2	2	7	Potential
2	2	2	6	Very Potential
2	2	2	2	Very Potential
2	2	2	3	Very Potential
2	2	2	5	Very Potential
1	2	1	1	Potential
1	2	1	4	Potential
1	2	1	9	Potential
1	2	1	10	Potential
1	2	1	8	Potential

Frequency	JaBoDeTaBek	Major	Study Program	Admission Status
1	2	1	7	Potential
1	2	1	6	Potential
1	2	1	2	Potential
1	2	1	3	Potential
1	2	1	5	Potential
1	2	2	1	Potential
1	2	2	9	Potential
1	2	2	10	Less Potential
1	2	2	8	Less Potential
1	2	2	7	Potential
1	2	2	6	Potential
1	2	2	2	Potential
1	2	2	3	Potential
1	2	2	5	Potential
1	1	1	1	Less Potential
1	1	1	4	Potential
1	1	1	9	Potential
1	1	1	10	Potential
1	1	1	8	Less Potential
1	1	1	7	Potential
1	1	1	6	Less Potential
1	1	1	2	Less Potential
1	1	1	3	Potential
1	1	1	5	Less Potential
1	1	2	9	Less Potential
1	1	2	7	Less Potential
1	1	2	6	Potential
1	1	2	2	Potential
1	1	2	3	Potential
1	1	2	5	Potential

2.4. Software development

The program is made in the Python programming language, with version 3.7.3. The library used is Numpy, Pandas, and Sklearn (scikit-learn). The dataset for testing is taken using the random function in the Python library. The process flow for the program can be seen in Figure 1.

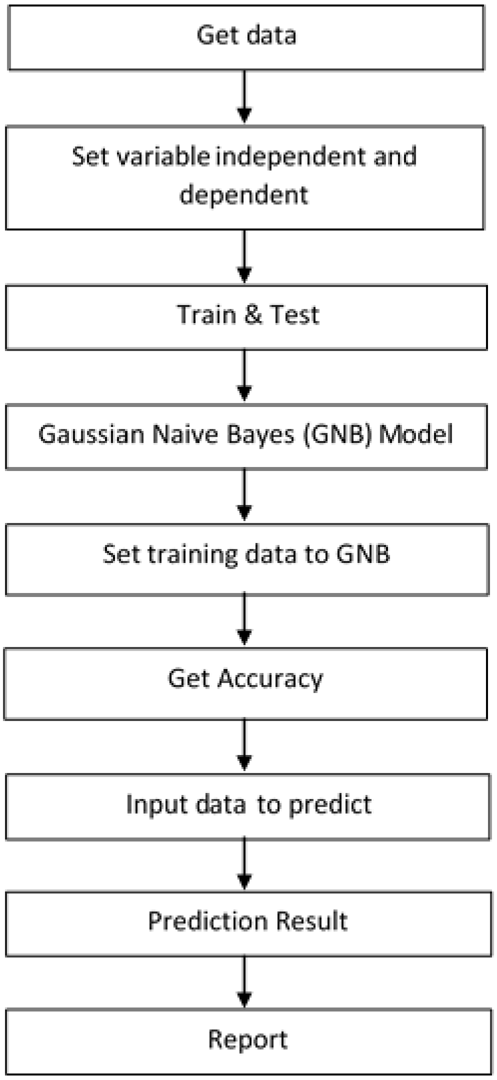

Figure 1. The process flow for the program.

3. Result and Discussion

Accuracy obtained from experiments based on the scale of training and testing, can be seen in the table 7. The accuracy obtained from the training and testing dataset is 0.73 or 73%. To obtain Accuracy, adequate training data and testing data are required. From the above experiments, the difference between the training and testing is 20 to 40.

Table 7. Accuracy based on the scale of training and testing.

No	Training : Testing	Accuracy
1	50 : 50	0.67
2	60 : 40	0.70
3	70 : 30	0.73
4	80 : 20	0.70
5	90 : 10	0.40

Based on an accuracy rate of 73%, a prediction experiment was conducted. The results of these experiments can be seen in the table 8. The results of training and testing with an accuracy of 0.73 (73%) can be seen in Figure 2.

Table 7. Prediction Experiment with accuracy rate of 73%.

No	Input Data	Prediction	Real
1	2, 1, 2, 1	Very Potential	Not Available
2	2, 2, 2, 1	Very Potential	Not Available
3	2, 1, 1, 2	Very Potential	Not Available
4	1, 1, 1, 9	Less Potential	Potential
5	1, 2, 2, 7	Potential	Potential
6	1, 2, 1, 5	Potential	Potential

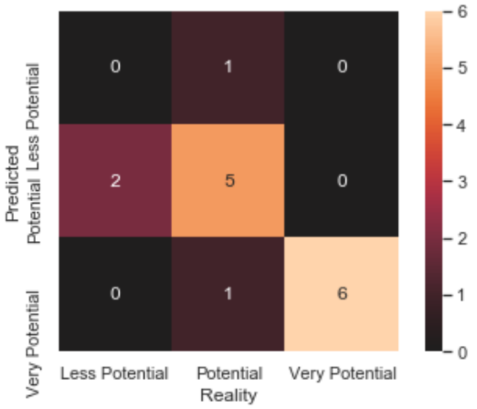

Figure 2. The Results of Training and Testing.

From the results of the six-time randomized trial of data, there is only one error in number 4. As for real data that has no application available, it can provide predictions for consideration in strategic decision making.

The results of the experiments show a close relationship to produce reliable information, i.e., data reliability and process reliability. These results can help the marketing department at Matana University in making decisions, especially to reach new students to achieve marketing targets. Information generated from this application can further enhance the effectiveness and efficiency of the marketing department in reaching new students. By knowing the status of these prospective students, the marketing department can immediately create the right strategy to reach them.

4. Conclusion

The application of machine learning based on the Naïve Bayes Classifier to predict and classify prospective new students at Matana University has been successfully carried out with an accuracy rate of 0.73 or 73%. There are two important things to produce accuracy, on the data side and on the process side. Reliable data and reliable processes will produce reliable information. One of the data or the problematic process results in information no longer reliable. Thus, this reliable information supports the marketing department's decision making so that the outreach of new prospective students is more effective and efficient.

Suggestions for further research are to improve accuracy by increasing the amount of data or modifying the model. It takes the development of an integrated marketing information system to reduce problem data, the use of artificial intelligence or machine learning can be more optimal and access data or information faster.

Acknowledgments

This research was supported by Matana University. I thank all those who have helped my research.

References

[1] Charu C. A 2015 Data Mining: The Textbook (Springer)
[2] Jiawei H, Micheline K, Jian P 2011 Data Mining: Concepts and Techniques 3rd (Morgan Kaufmann)
[3] Gordon S L, Michael J.A. B 2011 Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (Wiley Publishing, Inc.)
[4] Daniela X, Christopher J. H, Roger G. S 2009 Naïve Bayes vs. Decision Trees vs. Neural Networks in the Classification of Training Web Pages International Journal of Computer Science Issues (IJCSI), Vol. 4, No. 1, 2009 pp. 16-23
[5] Zsolt N 2018 Artificial Intelligence and Machine Learning Fundamentals (Packt Publishing Ltd)
[6] Peter G 2019 AI in Marketing, Sales and Service: How Marketers without a Data Science Degree can use AI, Big Data and Bots (Macmillan)
[7] Pulung H P, Zawiyah S, Ibrahim A 2017 Analisis Data Atribut Mahasiswa untuk Menentukan Strategi Promosi Kampus Menggunakan Metode Data Mining Seminar Nasional Teknik Elektro dan Informatika (SNTEI) pp. 81-86
[8] Ahmad H M 2019 Application of Naive Bayes Classifier Algorithm in Determining New Student Admission Promotion Strategies Journal of Information Systems and Informatics, Vol. 1, No. 1, March 2019 pp. 14-28
[9] Jim S 2017 Artificial Intelligence for Marketing, Practical Applications (John Wiley & Sons)
[10] Philip K, Hermawan K, and Iwan S 2017 Marketing 4.0, Moving from Traditional to Digital (John Wiley & Sons)
[11] Thomas D, Abhijit G, Dhruv G and Timna B 2019 How artificial intelligence will change the future of marketing Journal of the Academy of Marketing Science, (Springer, US) pp.1-19
[12] Jake V 2017 Python Data Science Handbook (O'Reilly Media, Inc.)