JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 7, No. 2, December 2024, pp. 138~144
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: 10.36378/jtos.v7i2.3866

Naive Bayes Algorithm Classification for Predicting Graduation Rate

Pradani Ayu Widya Purnama[1], Nurmaliana Pohan[2]
[1,2]Department of Informatics Engineering, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

Article Info

Article history:

Received 11 13, 2024
Revised 11 15, 2024
Accepted 11 28, 2024

Keywords :

Data Mining
Naïve Bayes Algorithm
Prediction
Graduation

ABSTRACT

Classification refers to the process of identifying a model or function that clarifies or differentiates concepts or categories of data, with the goal of predicting the class of an object. Naïve Bayes is a machine learning technique that employs probability computations. In this case study, various algorithms are used for modeling classification, and the naïve bayes algorithm is applied to examine the graduation rate. By utilizing this method, accuracy is assessed, which allows for an analysis based on criteria such as School Major, First Choice of College, Second Choice of College, Average Graduation Value, and Graduation Information. The outcome of the computation utilizing the Naïve Bayes Algorithm (Information Systems | Option 1) > (Information Engineering | Option 2) is 53.32% > 0%, which allows us to infer that the First Option of Information Systems and the Second Option of Informatics Engineering yield an Average Score of 75.00, resulting in a Graduation Information status of PASS, thus, Information Pass (Option 1-Information Systems).

Corresponding Author :

Pradani Ayu Widya Purnama
Department of Informatics Engineering
Universitas Putra Indonesia YPTK Padang
Padang, Indonesia
Email : pradaniwid@gmail.com

1. Introduction

Naive Bayes is a classification method that utilizes probability and statistics to predict future outcomes based on past experiences, under the assumption of strong independence (naive). Strong independence means that the attributes are assumed to be neither interdependent nor related to one another. In this study, researchers aim to calculate the accuracy level of the Naive Bayes method. This method can be applied to data with either categorical or continuous attributes. For continuous attributes, Naive Bayes assumes that the data follows a specific distribution and calculates the parameters of the distribution using training data. Commonly, the Gaussian distribution is employed to compute the conditional probabilities of continuous attributes within a class. The parameters of the Gaussian distribution include the mean and the standard deviation. [9].

This study aims to assist universities in assessing students accepted into their first or second choice of program. The calculation results using the Naïve Bayes Algorithm show that (Information Systems | First Choice) > (Information Engineering | Second Choice), with values of 53.32% > 0%. Therefore, it can be concluded that the First Choice is Information Systems, the Second Choice is Information Engineering, the Average Value is 75.00, the Graduation Category is PASS, and the Status is Pass (First Choice - Information Systems)..

1.1 Knowledge Extraction

Data mining is a series of processes to find added value in the form of information that has not been known manually from a database. The information obtained is produced by extracting and identifying important patterns or searching from data in the database [2]. Data mining is generally used to find knowledge contained in large databases, so it is often called Knowledge Discovery in Databases (KDD). Classification is a process to evaluate data objects so that they can be included in certain categories from various existing classes. In classification there are two main tasks that are carried out, namely building a model as a prototype that is stored as memory and using this model to perform recognition/classification/prediction on other data objects so that it can be known which class the data object is in the model that has been stored. [3].

There are 7 stages of data mining, with the following explanations [5]:

1. Data Cleaning: The process of eliminating inconsistent data or irrelevant data
2. Data Integration: Merging or combining data from several sources.
3. Data Selection: Selecting data from a set of operational data before obtaining knowledge discovery information in the database.
4. Data Transformation: Data is changed or combined into a format suitable for processing in data mining.
5. Data Mining: The main process when the method is applied to find knowledge or information from data.
6. Pattern Evaluation: To identify interesting patterns into the knowledge based found.
7. Knowledge presentation: Knowledge about the method used to obtain knowledge obtained by users.

1.2 Naïve Bayes

Naïve Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Naïve Bayes' competitive performance in the classification process even though it applies the assumption of attribute independence (no relationship between attributes). This assumption of attribute dependence in real data is rare, but even if the assumption of attribute dependence is violated, the performance of naïve Bayes classification remains high, which has been proven in various empirical studies. [8]. The Naive Bayes algorithm is one of the methods in classification techniques. Naive Bayes is a classification method that uses probability and statistics, proposed by British scientist Thomas Bayes, which projects future possibilities based on past experience, so it is known as Bayes' Theorem. The theorem is combined with Naive, which assumes that the conditions between attributes are mutually independent. Naive Bayes classification assumes that the presence or absence of certain features in a class is unrelated to the features of other classes. [1].

Uses of Naïve Bayes [10]:

1. Classifying text documents such as news texts or academic texts
2. As a machine learning method that uses probability
3. To make medical diagnoses automatically
4. Detecting or filtering spam

Advantages of Naïve Bayes [10]:

1. Can be used for quantitative and qualitative data
2. Does not require a large amount of data
3. No need to do a lot of training data
4. If there is a missing value, it can be ignored in the calculation
5. The calculation is fast and efficient
6. Easy to understand
7. Easy to make
8. Document classification can be personalized according to each person's needs
9. If used in a programming language, the code is simple
10. Can be used for classifying niner or multiclass problems

Naive Bayes is a classification with probability and statistical methods based on Bayes' theorem, namely predicting future opportunities based on previous experience where it is assumed that the conditions between attributes are mutually independent. The equation of Bayes' Theorem is [4] :

$$P(C|F) = \frac{P(C) * P(F|C)}{P(F)}$$

where,

F : Data with unknown class
C : Hypothesis of data is a specific class
P(C | F) : Probability of hypothesis C given F (posterior probability)
P(C) : Probability of hypothesis C (prior probability)
P(F | C) : Probability of hypothesis F given
C : Probability of hypothesis F

Class determination is done by comparing the probability value of a sample being in one class with the probability value of a sample being in another class. Class determination is done by comparing the probability value of a sample being in one class with the probability value of a sample being in another class. [7].

2. Research Method

This study applies a classification method based on the Naive Bayes algorithm to predict student graduation rates by considering several attributes. The design of this study is quantitative with an experimental approach, where the classification model is designed and tested using certain data. This approach allows researchers to evaluate the accuracy of the Naive Bayes algorithm in classification, so it must be arranged in the form of a research framework. This research framework can be seen in Figure 1 below:

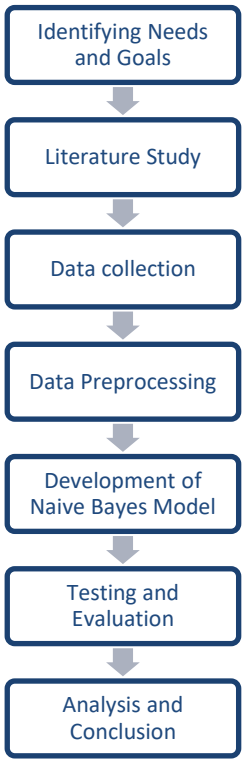

Figure1. Research Framework

This research framework serves as a clear guide regarding the flow of research implementation, from the preparation stage to the final analysis, with the following explanation :

a. Problem Identification, namely how to predict student graduation rates by considering several attributes through the Naive Bayes algorithm. The determination of this problem is based on the need to provide accurate prediction information to help universities plan student admissions. After finding the problems and needs, the next step is to formulate the intention of how Naïve Bayes can be applied to predict

graduation rates.b. Studi Literatur, dalam melaksanakan penelitian ini penting untuk mengumpulkan referensi yang berkaitan dengan Naive Bayes dan aplikasinya dalam klasifikasi serta prediksi. Literatur diperoleh dari berbagai sumber seperti buku, jurnal akademik, dan prosiding. Penelitian ini membantu peneliti mengerti konsep dasar algoritma dan metode penerapannya dalam studi klasifikasi.

c. Data Collection, at the data collection stage, it is carried out on datasets that have attributes such as school majors, first and second study program choices, average grades, and graduation status. This dataset was collected using purposive sampling and will be further processed at the preprocessing stage..

d. Data Pre-processing, the collected data is processed to overcome inconsistencies, restructure attribute formats, and prepare the data to suit the Naive Bayes algorithm. This process involves data cleaning, data integration, and data transformation. This pre-processing is crucial to improve the quality of the data before it is applied in the prediction model..

e. Development of the Naive Bayes Model, after the data is collected and processed, an analysis is carried out using a descriptive statistical approach to explain the characteristics of the data. Probabilistic analysis is carried out using the Naive Bayes algorithm, through the calculation of posterior values based on Bayes' theorem. The final results are compared between estimates and real values to assess the accuracy of predictions in classifying graduation data. In this development, the Naive Bayes algorithm is implemented using the Python programming language and the scikit-learn library. This algorithm works by calculating the likelihood of each attribute to predict the target class. This model will produce possibilities based on the characteristics of each sample.

f. Testing and Evaluation, after the model is developed, testing is carried out using test data. The model that is created is tested using different test data. The trial is carried out by comparing the results of the model prediction with the actual passing grade. Model accuracy is calculated to assess how effective the Naive Bayes algorithm is in predicting graduation using the provided attribute data. Evaluation of model accuracy is carried out to evaluate how effective the Naive Bayes algorithm is in predicting the graduation rate by comparing the prediction results with the actual data. Assessment metrics such as accuracy are used in this phase. To assess the effectiveness of the model, the accuracy metric is used, which shows the percentage of correct predictions from the total predictions made. Evaluation of accuracy is carried out by comparing the probability values of passing in the first and second choices of the study program.

g. Analysis and Conclusion, after the model is tested, the resulting data is analyzed to understand the graduation pattern based on certain attributes. The resulting conclusions indicate the effectiveness of the Naive Bayes model in predicting graduation. This analysis is based on the calculation of probability and statistics from the model..

3. Result and Discussion

In the application of the Naive Bayes Algorithm to predict the graduation rate, there are 15 training and testing data, which can be seen in the following table 1 :

Table 1. Training and Testing Data

No	School Majors	First Choice College	Second Choice College	Average Passing Grade	Graduation Statement	Graduation Statement
1	IPA	Information Systems	Information Technology	80.00	PASSED	Choice 1
2	IPA	Information Technology	Information Systems	78.50	PASSED	Choice 2
3	IPA	Information Systems	Information Technology	92.25	PASSED	Choice 2
4	IPA	Information Systems	Management	83.00	PASSED	Choice 1
5	IPA	Information Technology	Information Systems	75.00	PASSED	Choice 1
6	IPS	Information Technology	Information Systems	65.20	PASSED	Choice 2
7	IPS	Management	Akuntansi	80.30	PASSED	Choice 1

8	IPA	Information Systems	Management	65.55	PASSED	Choice 1
9	TKJ	Information Systems	Information Technology	75.40	PASSED	Choice 2
10	IPS	Management	Accountancy	80.00	PASSED	Choice 1
11	TKJ	Accountancy	Management	90.00	PASSED	Choice 1
12	TKJ	Information Systems	Information Technology	85.25	PASSED	Choice 1
13	IPA	Information Technology	Information Systems	66.40	LULUS	Pilihan 1
14	IPS	Accountancy	Management	75.00	LULUS	Pilihan 1
15	IPA	Management	Accountancy	65.00	LULUS	Pilihan 2
16	**IPA**	**Information Systems**	**Information Technology**	**75.00**	**LULUS**	**?**

Calculating Choice 1 and Choice Majors 2 (P(vj))

To calculate the probability of theft using the formula:

$$P(Vj) = \frac{N}{Jumlah} \quad (1)$$

$$P(\text{Pilihan 1}) = \frac{10}{15} = 0.6666$$

$$P(\text{Pilihan 2}) = \frac{5}{15} = 0.3333$$

Calculating the probability of Science Major, First Choice Information Systems, Second Choice Informatics Engineering, Average Score 75.00, Graduation Remarks PASS (CHOICE 1)

P (Information System | Choice 1) = 4/10 = 0.4
P (Average Score 75.00 | Choice 1) = 2/10 = 0.2
P (Graduation PASS | Choice 1) = 10/10 = 1

Calculating the probability of Science Major, First Choice Information Systems, Second Choice Informatics Engineering, Average Score 75.00, Graduation Remarks PASS (CHOICE 2)

P (Information System | Choice 2) = 1/5 = 0.2
P (Average Score 75.00 | Choice 2) = 0/5 = 0
P (Graduation PASS | Choice 2) = 3/5 = 0.6

Determining the percentage of Information Technology and Information Systems Science
P (Information Systems | Choice 1) = 0.6666*0.4*0.2*1 = 0.05332 = (53.32%)
P (Information Technology | Choice 2) = 0.3333*0.2*0*0.6 = 0

Therefore, based on the calculation results above with the results (Information Systems | Option 1) > (Information Engineering | Option 2) which is 53.32% > 0%, it can be concluded that the First Choice is Information Systems, the Second Choice is Informatics Engineering, Average Score 75.00, Graduation Description PASS so that the Information Passed (Option 1-Information Systems).

Therefore, based on the calculation results obtained, where the probability for **Information Systems (Option 1)** is greater than for **Information Engineering (Option 2)**—specifically, **53.32% > 0%**—it can be concluded that:

- The **First Choice** is **Information Systems**.
- The **Second Choice** is **Informatics Engineering**.

The results indicate that the likelihood of acceptance into the Information Systems program as the first choice is significantly higher compared to the Informatics Engineering program as the second choice. This conclusion is supported by the calculated probabilities, where Information Systems has a strong chance of

success with a value of **53.32%**, while Informatics Engineering does not show any probability of acceptance in this scenario (0%).

Additionally, with an **Average Score of 75.00** and a **Graduation Description of PASS**, it confirms that the candidate meets the necessary criteria for passing. Therefore, the best decision is to prioritize **Option 1: Information Systems** as the primary recommendation. This conclusion highlights the suitability of Information Systems as the ideal choice based on the provided data and analysis.

4. Conclusion

Based on the results of the analysis that has been carried out, it can be concluded that for the case with the Science Department, the First Choice of Information Systems College, the Second Choice of Informatics Engineering College, the Average Passing Score of 75.00, Information PASS in the search process using the Naïve Bayes Algorithm obtained the results (Information Systems | Choice 1) > (Informatics Engineering | Choice 2) which is 53.32% > 0%, then it can be concluded that the First Choice is Information Systems, the Second Choice is Informatics Engineering, with an Average Score of 75.00, and Information PASS, so Information Pass (Choice 1-Information Systems).

Acknowledgement

The author would like to express his gratitude and thanks to all parties who have provided support and assistance during the implementation of this research. Special thanks are given to Universitas Putra Indonesia YPTK Padang for providing the opportunity and facilities to carry out this research. I would also like to thank my friends and family for their continued support and encouragement, as well as to the readers who took the time to understand the results of this research. Hopefully the results of this research can provide benefits for the advancement of science, especially in the field of data classification with the Naive Bayes algorithm.

References

[1] Astuti, et al. (2018). Algoritma Naive Bayes Dengan Fitur Seleksi Untuk Mengetahui Hubungan Variabel Nilai Dan Latar Belakang Pendidikan. Jurnal Simetris.
[2] Budi Santosa. (2007), Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis/Studi, 1st ed. Yogyakarta: Graha Ilmu.
[3] Kursini & Emha, T. L. (2009). Algoritma Data Mining.Yogyakarta : Penerbit Andi.
[4] Susana, H. et al. (2022). Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. JURSISTEKNI (Jurnal Sistem Informasi dan Teknologi Informasi), Vol 4, No.1, January 2022: Hal 1– 8 ISSN.
[5] Association Rule. Jurnal Riset Komputer (JURIKOM).
[6] Mulyanto, Agus, 2009, Sistem Informasi Konsep dan Aplikasi, Cetakan I, Pustaka Pelajar, Yogyakarta
[7] Permadi, V. A. (2020). Analisis Sentimen Menggunakan Algritma Naïve Bayes Terhadap Review Restoran Di Singapura. Buana Informatika, Vol.11, No.2:141–151.
[8] Putro, H. F., Vulandari, R. T., & Saptomo, W. L. Y. (2020). Penerapan Metode Naive Bayes Untuk Klasifikasi Pelanggan. Jurnal Teknologi Informasi Dan Komunikasi (Tikomsin), Vol. 8, No.2:19–24.
[9] Supriyatna, Mustika. (2018). Komparasi Algoritma Naive Bayes Dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil. Jurnal Sains Komputer & Informatika (J-SAKTI)
[10] Watratan, A., Puspita, A. B., Moeis, D., Informasi, S., & Profesional Makassar, S. (2020). Implementasi Algoritma Naive Bayes Untuk Memprediksi Tingkat Penyebaran Covid-19 Di Indonesia. In Journal of Applied Computer Science and Technology (Jacost), Vol. 1, No 1:7–14.
[11] Uysal, A. K., & Gunal, S. (2014). The Impact of Preprocessing on Text Classification. Information Processing & Management, 50(1), 104-112.
[12] Wu, X., et al. (2008). Top 10 Algorithms in Data Mining. Knowledge and Information Systems, 14(1), 1-37.
[13] Tan, P., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley.
[14] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.
[15] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.
[16] Mitra, S., Pal, S. K., & Mitra, P. (2002). Data Mining in Soft Computing Framework: A Survey. IEEE Transactions on Neural Networks, 13(1), 3-14.

[17] Bayes, T., & Price, R. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London, 53, 370-418.
[18] Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.
[19] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.
[20] Domingos, P., & Pazzani, M. (1997). On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Machine Learning, 29(2-3), 103-130.
[21] Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 13(1), 21-27.
[22] Cahyani, L., & Wicaksono, M. (2019). Implementasi Algoritma Naive Bayes Untuk Memprediksi Penyakit Diabetes. Prosiding Seminar Nasional Informatika.
[23] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
[24] Russell, S. J., & Norvig, P. (2016). Artificial Intelligence: A Modern Approach. Pearson.
[25] Patel, H., & Prajapati, P. (2018). Study and Analysis of Decision Tree Based Classification Algorithms. International Journal of Computer Sciences and Engineering, 6(10), 74-78.