TIERS Information Technology Journal
Vol.3, No.2, December 2022, pp. 69~75
ISSN: 2723-4533 / E-ISSN: 2723-4541
DOI: 10.38043/tiers.v3i2.3865

Prediction of Education Level in Population Data Using Naïve Bayes Algorithm

Pungkas Subarkah[1], Siti Alvi Solikhatin[2], Irma Darmayanti[3], Ali Nur Ikhsan[4], Debby Ummul Hidayah[5], Rayinda Maya Anjani[6]
subarkah18.pungkas@gmail.com[1], alvi.sholikhatin@gmail.com[2], irmada@amikompurwokerto.ac.id[3], alinurikhsan@amikompurwokerto.ac.id[4], debbyummul@amikompurwokerto.ac.id[5], rayinda40@gmail.com[6]
[1,3,4] Department of Informatics, Universitas Amikom Purwokerto, Jawa Tengah
[2]Department of Digital Business, Universitas Amikom Purwokerto, Jawa Tengah
[5,6]Department of Information System, Universitas Amikom Purwokerto, Jawa Tengah

ABSTRACT

Education is the key to improving human resources. The Ministry of National Education is implementing a curriculum that requires students to study it (MoNE) and as part of this program, all Indonesian citizens are required to attend three years of primary education, which includes SD, MI/Equivalent, three years for SMP/Equivalent, and three years for high school/equivalent level. This study aims to determine whether or not a government program that requires Indonesian citizens to study for 12 years is required, therefore it is necessary to test predictions of data on the level of education in Blitar Regency. This study conducted a prediction test by implementing the Naive Bayes algorithm on education-level data in Blitar Regency as of 2020 which was taken from the satudata.go.id website. In the data processing, there are values of precision, recall, f-measure, Weighted Avg, and also Confusion Matrix. The accuracy results of the Naive Bayes Algorithm on education level data in Blitar district show that the district has implemented government policies regarding the 12-year compulsory education program, this is based on the results of data processing which shows an accuracy value of 98.4848% and category good classification.

Keywords: Prediction; Education; Naive Bayes; Algorithm

Article Info

Accepted : 19-12-2022
Revised : 10-11-2022
Published Online : 25-12-2022

Correspondence Author:

Pungkas Subarkah
Department of Informatics,
Universitas Amikom Purwokerto,
Jl. Letjend Pol. Soemarto No.127, Watumas, Purwanegara, Kec. Purwokerto Utara, Kabupaten Banyumas, Jawa Tengah 53127
Email: subarkah18.pungkas@gmail.com

1. INTRODUCTION

Education is a deliberate and organized effort to establish a learning environment and learning process in which students actively develop their potential to have self-control, intelligence, noble character, and skills needed by themselves, society, and the state [1]. The right to education is guaranteed for all Indonesian citizens by Article 31 Paragraph 1 of the 1945 Constitution of the Republic of Indonesia. According to this article, the government must monitor the progress of Indonesian education to minimize the loss of every citizen's right to an education[2]. This is important in evaluating the level of education in School aims at all schools and institutions education. To find out internal factors and external influences on learning outcomes It is important for such students to benefit the learning process can be maximized.

Concerning Law Number 20 of 2003, Article 3 of the National Education System with an Educational Function, "National Education develops valuable skills, personality, and national civilization, devoted to the Almighty, noble personality, healthy, knowledgeable, creative, independent, democratic, and being a responsible citizen. Education has a national goal that is stated in Law no. 20 of 2003 which states that education will be pursued by humans as they are (actualization) by considering various possibilities that are what they are (potentiality), and will be directed towards the realization of the human being that should be or the human being aspired to. In other words, the implications of education are to develop and realize various potentials in humans in the context of the dimensions of morality, sociality, diversity, individuality, and culture as a whole and with integrity[3].

The function of education in Indonesia is to develop valuable skills, character, and national civilization in the context of the nation's intellectual life. From this function, it can be seen that Indonesian national education prioritizes the transformation of attitudes, character, and philosophical values of the Indonesian state. It aims to strengthen national sentiment and compete in international competition. In addition, education also has very complex benefits, including education can provide information and understanding, deepen knowledge, improve careers, form scientific thinking patterns, prevent duping, teaching social functions in society, optimize one's talents, shaping character, and educating individual.

Compulsory education is one of the educational initiatives that have been introduced by the governments of each country. Each state government has a different policy regarding compulsory education. In Indonesia, compulsory education is a program that ensures everyone has received a basic understanding of the responsibilities of central and local government. The Ministry of National Education is implementing a curriculum that requires students to study it (Depdiknas). As part of this program, all Indonesian citizens are required to attend school for a total of nine years at the basic education level, which includes grades 1 to 9 of Elementary School (SD), Madrasah Ibtidaiyah (MI), or other equivalent schools. Each region is obliged to ensure that every resident has completed high school education or its equivalent in order to implement the 12-year compulsory education program[4].

Due to the growing awareness of the value of education in developing a country, many countries have made 12 years of school mandatory. This activity will be carried out by implementing the Naive Bayes Algorithm to facilitate processing and guessing predictions[5]. The method used is the Naive Bayes Classifier method which is one of the classification techniques in data mining. Where the analysis will be carried out to obtain information about population data at the education level[6]. The Naive Bayes algorithm is a probability-based prediction approach based on the application of the principles of a strong premise or independence. Naive Bayes is used for Machine Learning. The Naive Bayes algorithm takes only one time to scan the data, and is used to manage missing attribute values and continuous data. This study aims to determine whether or not a government program that requires Indonesian citizens to study 12 years by conducting data mining in order to obtain information from the data collected[7].

As Referring to previous research, Sherlin, Saniati and Ade conducted a study at SMK Teluk Betung Taman to estimate a web-based model to predict student enrollment potential at SMK Teluk Betung Taman. The purpose of this research is that new students will register at SMK Teluk Betung Taman and need assistance in developing a prediction system and applying naive bayes in determining accuracy. The researcher uses the Naive Bayes Algorithm in determining the accuracy results. From the results obtained, the accuracy is 86%[8].Similar to the first reference, there are other studies that we use as a second reference. The research conducted by this study discusses the diagnosis of heart disease by considering the previous data and information. Researchers implemented naive bayes to predict risk factors for heart disease. The data used have attributes of age, BP, cholesterol, gender, blood sugar, etc. From the results of the data that has been implemented in naive bayes, the accuracy obtained is 89.77%[9]. Research that discusses the analysis of government policy sentiment towards the Corona case. As a starting point for evaluating the efficacy of TF-IDF and N-gram feature extraction using the Naive Bayes approach, this study aims to determine whether public opinion on the New Normal policy is favorable or not. The results obtained from TF-IDF and N-gram using the Naive Bayes approach is an accuracy value of 81%[10].

Other research discusses the implementation of data mining using the Naive Bayes Algorithm to predict birth rates. In this study, data mining was used to solve the problems faced by the village office in predicting the birth rate. The recommended method is the Naive Bayes classifier. The purpose of using the Naive Bayes Algorithm is to see the predictive pattern for each attribute in the dataset and to test the training data with test data and see whether the data modeling is appropriate or not. The results of this survey are intended to facilitate the Lalang Village Office in managing population data and assist in the data input process, data search, and population reporting[11]. It can be seen from the three previous studies, the application of the Naive Bayes Algorithm is considered very suitable for carrying out probability prediction analysis activities for the case studies used[12], so in this study we also implement the Naive Bayes Algorithm to predict the level of education in Blitar Regency.

.

Based on the background of the problem and previous research that reviews the use of the Naive Bayes algorithm in education. For renewal in this study is to predict the education population level using the Naive Bayes algorithm and using secondary data in our research data processing.

2. RESEARCH METHOD

The method proposed by researchers in research on the Prediction of Education Level in Population Data Using Naïve Bayes Algorithm several stages can be seen in Figure 1.

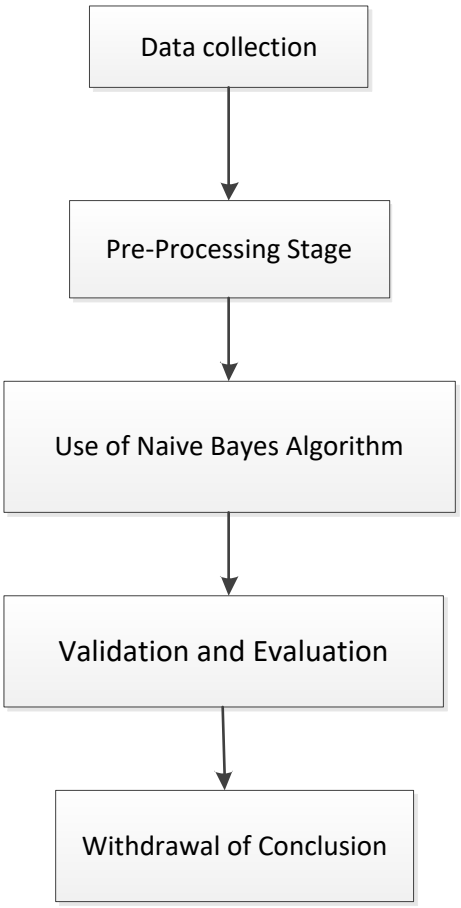

Figure 1. Research Stages

2.1. Data Collection

Data collection is required for Naïve Bayes testing. Accurate data must be collected and then divided into groups based on some rules. Classification is used to facilitate data selection[13]. After the raw data is obtained, the next step is to classify the data. education level in East Java with case studies only in Blitar Regency per year 2020. Then the data was only taken per sub-district with 132 records and 6 attributes (5 attributes and 1 target attribute)). The following is an explanation of the education level attribute, which can be seen in Table 1

Table 1. Attribut Dataset

Attribute Name	Information
Name of District	Bakung, Binangun, Doko, Gandusari, Garum, Kademangan, Kanigoro, Kesamben, Nglegok, Panggungrejo, Ponggok, Sanankulon, Selopuro, Selorejo, Srengat, Sutojayan, Talun, Udanawu, Wates, Wonodadi, Wonotirto.
Year	2020
Age	<6 - <13, <12, 12-15, <14, 15-18, <17

Education Level	Graduated from Elementary School/Equivalent, Did not graduate SD/Equivalent, Graduated from high school/equivalent, Did not graduate from high school/equivalent, Graduated from high school/equivalent, Did not graduate from high school/equivalent
Total amount	Number of people per district
Participate in the 12-Year Compulsory Education Government Program	Yes, No

2.2. Pre-Processing Stage

The pre-processing stage is a data mining process that is first carried out to get good quality data to be processed before the classification process, one of which is data transformation. Data transformation is an essential part of the health dataset in the pre-processing stage[14].

2.3. Use of Naive Bayes Algorithm

Stages of using the classification method. One of the tasks of data mining is classification. Classification represents the most widely used data mining technique[15]. At this stage the data is classified. Classification is the process of determining whether a data belongs to one of a number of predefined categories. To develop models from raw data, classification uses models to categorize new data. Classifications can be described in terms of a specific purpose that corresponds to each set of one of various class labels (features). The classification system is expected to classify all data accurately. The classification system should be evaluated to determine its effectiveness. In general, performance evaluation is a common practice. A confusion matrix is used to classify information[16].

2.4. Validation and Evalution

For decision making, data validation verifies whether a certain set of values has been selected from among the pre-determined sets as acceptable or not. By identifying data quality, data validation ensures that the data in the final data recording is of high quality. If the data used is relevant to education da ta and the government's 12-year compulsory education program, then the quality of the da ta also proves to be very good[17]. The validation and evaluation stages are carried out by measuring the accuracy of the results achieved by the confusion matrix and K-fold cross-validation technique models

2.5. Withdrawal Conclusion

This stage concludes the results obtained from the study using the Prediction of Education Level in Population Data Using Naïve Bayes Algorithm, which provide accurate results for classifying heart failure based on the precision, recall, and F-Measure values of each algorithm[18]. with the classification level, as follows:

a. Excellent classification = 0.90 – 1.00
b. Good classification = 0.80 – 0.90
c. Fair classification = 0.70 – 0.80
d. Poor classification = 0.60 – 0.70
e. Failure = 0.50 – 0.60

3. RESULTS AND DISCUSSION

3.1. Pre-Processing Stage

The pre-processing stage is through data transformation, which aims to facilitate the calculation of the value between the class attributes at the classification stage. The following results from the change of non-numeric data types into numeric data can be seen in table 2.

.

Table 2. Number of Datasets Based

Classification Type	**Number of dataset records**
YES	66
NO	66
Amount	**132**

Then at this pre-processing stage identification and adjustment of attributes are carried out as well as selection of population level educational datasets so that the data obtained is data that is truly ready to be used in the next stage. The results of the attribute-adjusted population-level education dataset for the Weka application are shown in the table 3.

Table 3. Data Pre-Processing

Original Data	Pre-processing Result Data	Information
Wates	Wates	Name of District
2020	2020	Year
18	18	Age
S1	S1	Education Level
225	225	Total amount
Yes	Yes	Participate in the 12-Year Compulsory Education Government Program

3.2. Use of Naive Bayes Algorithm

Before doing manual calculations. The education level dataset processed using WEKA will produce a Confusion Matrix based on the 10-fold cross validation evaluation method, and the dataset will be divided into 10 subsets (9 subsets as training sets and 1 subsets as testing set) with a total of 10 iterations. The classifier used for testing the dataset is Nave Bayes (Naive Bayes algorithm) which will produce visualization of the overall data and target classes (class YES and class NO) that can be seen in Figure 2.

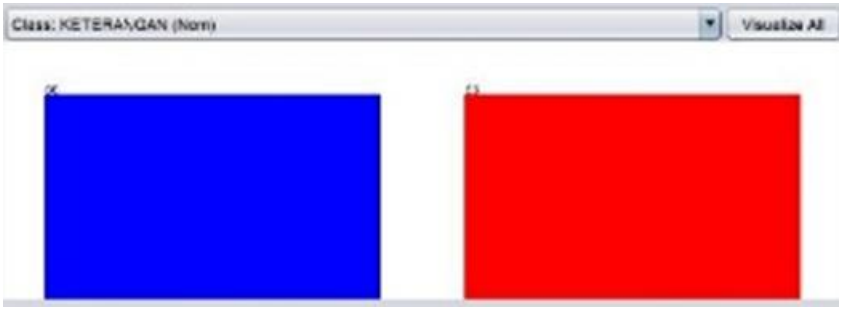

Figure 2. Overall data visualization

Processing of education level datasets that apply Naive Bayes classification is also carried out without the use of software. Calculations are done manually to determine the accuracy of the results. In implementing the dataset on the Weka software, the results of the Confusion Matrix are obtained. The confusion matrix is used as a performance measurement for machine learning classification problems. The output is in the form of two classes that have a table with 4 different combinations of predicted values and actual values. The following is Table 4. Confusion Matrix.

Table 4. Confusion Matrix

Classifies as	***Confusion Matrix***	
	a	**b**
Yes	66	0
No	2	64

In the Naive Bayes Algorithm, if manual calculations are carried out, precision, recall, and F-Measure will produce an accuracy value of 98.4848 %. The manual calculation process can be seen in table 5. Yes class confusion

Table 5. Yes Class Confusion

66 (True Positive)	0 (False Negative)
2 (False Positive)	64 (True Negative)

From table 4 and table 5, the precision, recall, and F-Measure values and accuracy values are obtained based on the Confusion Matrix, as follows:

Table 6. Accuracy Value Based On Confusion Matrix Using Naive Bayes Algorithm

Class	*Precision*	*Recall*	*F-Measure*
Yes	0,971	1	0,985
No	1	0,970	0,985
Weighted Avg	0,985	0,985	0,985

3.3. Validation and Evaluation

The evaluation and validation stages are used to measure the accuracy of the classification algorithm presented in table 7. In table 7., the table generated by the confusion matrix from dataset testing using the Naive Bayes algorithm with the 10-fold cross validation method.

Table 7. Confusion matrix Education Level in Population

	Class Yes	*Class No*
Class Yes	66	2
Class No	0	64
132	66	66

Table 7, it can be described as follows that the number of data generated by the rule that is in the Yes category is the same educational level as the testing data which is also 66. Furthermore, the number of data generated by the rule that is not at the educational level with testing data which is Yes Education Level is 66. Then the amount of data generated by rules that are at the educational level and testing data that is not at the educational level is 64. Finally, the number of data generated by rules that are not at the same educational level as the testing data that are also not at the educational level is 66.

After the results described above, the discussion of this study is that by using a dataset on the level of the educational population predicted using the Naive Bayes algorithm, an accuracy value of 98.4848% is obtained. With the acquisition of an accuracy value that is classified as very good, in line with several studies that the acquisition of accuracy results using the Naive Bayes algorithm is classified as good [19][20][21] and Naive Bayes is an algorithm or method that is most effective and efficient for design machine learning and data mining[22].

3.4. Withdrawal Conclusion

Based on the results of calculations that have been carried out using the Naive Bayes algorithm on the education level dataset, an accuracy value of 98.4848% is obtained, with details of a precision value of 0.985%, a recall value of 0.985% and an F-Measure value of 0.985%.

CONCLUSION

Based on the research that has been done regarding the identification of the educational level population using the Naive Bayes algorithm, it can be concluded that the confusion matrix values obtained an accuracy of 98.4848%, with details of a precision value of 0.985%, a recall value of 0.985% and an F-Measure value of 0.985%. From the results of the accuracy value categorized as very good classification. through this research for further research include: 1) Adding a classification algorithm that aims for comparison or comparison so as to get a better accuracy value in the education level population, and 2) Adding a feature selection process in identifying population level education.

ACKNOWLEDGEMENT

The author expresses his gratitude and sincere appreciation to Amikom University in Purwokerto for his unconditional support for this research.

REFERENCES

[1] E. Risdianto, "Analisis Pendidikan Indonesia di Era Revolusi Industri 4.0," *Res. Gate*, no. April, pp. 0–16, 2019.

[2] I. A. Nafrin and H. Hudaidah, "Perkembangan Pendidikan Indonesia di Masa Pandemi Covid-19,"

.

Edukatif J. Ilmu Pendidik., vol. 3, no. 2, pp. 456–462, 2021.
[3] I. W. C. Sujana, "Fungsi Dan Tujuan Pendidikan Indonesia," *Adi Widya J. Pendidik. Dasar*, vol. 4, no. 1, p. 29, 2019.
[4] I. A. Sugardha, "Upaya Ke Arah Wajib Belajar 12 Tahun Di Kabupaten Majalengka; Pendekatan Kebijakan," *J. Adm. Pendidik.*, vol. 25, no. 2, pp. 252–263, 2018.
[5] S. Hidayatul, A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," vol. 2, no. 9, pp. 2546–2554, 2018.
[6] H. Annur, "Klasifikasi Masyarakat Miskin Menggunakan Metode Naive Bayes," *Ilk. J. Ilm.*, vol. 10, no. 2, pp. 160–165, 2018.
[7] P. Subarkah, A. N. Ikhsan, and A. Setyanto, "The effect of the number of attributes on the selection of study program using classification and regression trees algorithms," in *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 2018.
[8] S. Eka *et al.*, "Penerapan Model Naive Bayes Untuk Memprediksi Potensi," vol. 1, no. 1, pp. 82–87, 2021.
[9] I. Loelianto, H. Angriani, B. Mappakasunggu, and K. Makassar, "IMPLEMENTASI TEORI NAÏVE BAYES DALAM KLASIFIKASI CALON," vol. 3, no. 2, pp. 110–117, 2020.
[10] A. R. Isnain, N. S. Marga, and D. Alita, "Sentiment Analysis Of Government Policy On Corona Case Using Naive Bayes Algorithm," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 15, no. 1, p. 55, 2021.
[11] M. Idris, "Implementasi Data Mining Dengan Algoritma Naive Bayes Untuk Memprediksi Angka Kelahiran," *J. Pelita Inform.*, vol. 7, no. 3, pp. 421–428, 2019.
[12] M. S. Nawaz, B. Shoaib, and M. A. Ashraf, "Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization," *Heliyon*, vol. 7, no. 5, p. e06948, May 2021.
[13] A. Khan *et al.*, "PackerRobo: Model-based robot vision self supervised learning in CART," *Alexandria Eng. J.*, vol. 61, no. 12, pp. 12549–12566, 2022.
[14] M. Han, J., & Kamber, *Data Mining Concepts, Model and Techniques 2nd Edition*. San Fransisco: Elsevier, 2006.
[15] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques 3rd Edition*. San Fransisco: Morgan Kauffman, 2012.
[16] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020.
[17] A. Nurfauzan and W. Maharani, "Klasifikasi Emosi Pada Pengguna Twitter Menggunakan Metode Klasifikasi Decision Tree," 2021.
[18] F. Gorunescu, *Data mining Concepts, Models and Techniques*. Verlen Berlin: Springer, 2011.
[19] I. Parlina *et al.*, "Naive Bayes Algorithm Analysis to Determine the Percentage Level of visitors the Most Dominant Zoo Visit by Age Category," *J. Phys. Conf. Ser.*, vol. 1255, no. 1, 2019.
[20] W. Chen, S. Zhang, R. Li, and H. Shahabi, "Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling," *Sci. Total Environ.*, vol. 644, pp. 1006–1018, 2018.
[21] P. Subarkah, W. Risma, and R. Aditya, "Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification," vol. 14, no. 2, pp. 120–125, 2022.
[22] G. Bermejo-Martín *et al.*, "Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy," *IEEE Access*, vol. 9, no. 1, p. 121, 2021.