International Journal of Engineering & Technology, 7 (4.44) (2018) 228-232

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

The Development of Classification System of Student Final Assignment Using Naive Bayes Classifier Case Study: State Community Academy of Bojonegoro

Pramana Yoga Saputra[1]*, Yoppy Yunhasnawa[2], Windy Fatmila[3], Faisal Rahutomo[4], Rosa Andrie Asmara[5], Dimas Wahyu Wibowo[6], Erfan Rohadi[7], Indrazno Siradjuddin[8], Awan Setiawan[9], Ahmad Hafidh Ayatullah[10], Arie Rahmad Syulistyo[11], Dewi Cahyandari[12]

[1,2,3,4,5,6,7,8,9,10,11]*State Polytechnic of Malang*
[12]*University of Brawijaya*
**Corresponding author E-mail: pramana.yoga@polinema.ac.id*

Abstract

In determining interest, students are faced with the choice of specialization in determining the final field of interest. Specialization in the Information Management Study Program of State Community Academy of Bojonegoro is divided into five specializations. The choice of specialization groups is an important part. This is because the accuracy in choosing specialization groups is part of the initial plan of students to determine the final assignment project. Thus, the field of specialization taken will be in accordance with the interests and abilities of the students and will have an impact on the process. In this work, we propose a system that can provide information about the classification of student final assignments. We use Naive Bayes Classifier (NBC) algorithm to do the classification. In this work, we used datasets, that obtained from the State Community Academy of Bojonegoro Informatics Management Study Program. Based on the accuracy testing of the classification results, the system gives higher result, than test manual calculation of 83.33%.

Keywords: *Data Mining, classification, Naive Bayes Classifier, Machine Learning, final assignment.*

1. Introduction

Every student who will complete his studies at the level of Higher Education is required to compile a scientific work in the form of a report called final assignment. The final assignment is one of the requirements for student graduation. Such as State Community Academy of Bojonegoro Diploma 2 Study Program students, are required to prepare and write a final assignment report, at the end oh their study period.

The purpose of the final assignment is to make a final review and assess the students, whether they can solve problems in a certain field of interest, by using their experience, that they get in the academy. It is expected that students are able to solve problems systematically and logically, critically and creatively, based on accurate data or information supported by appropriate analysis. And then write it in scientific writing.

In the final assignment, students are required to choose a specialization of the field of interest. Information Management Study Program of State Community Academy of Bojonegoro has five specializations. The selection of specialization is very important. Because it is the initial phase for the students to work their final assignment. By choosing the specialization, they can determine the topic for their final assignment. If the topic fit with their interest and ability, it will bring a good impact on the working process of the final assignment.

Based on the curriculum in the Informatics Management Study Program, specialization of students started at 4th semester, with specialization guidelines based on information and the history of three previous semesters grades. The grades are taken from supporting subjects that are in accordance with specialization. Then, it will be processed into new information in the form of datasets. From these datasets, students can be classified into specialization categories that are suitable with their grade.

We propose a system to classify the student final assignment, based on the specialization. In this work, we use data mining techniques. Data mining is a series of activities to find interesting patterns of large amounts of data. The data can be stored in a database, data warehouse or another information storage [1]. In addition, the use of data mining techniques can discover unknown relationships in data, and present the data clearly, and easily be understood by the user. So, the data relationship can be the basis of decision making [2].

The data mining technique for classification that will be used in this work is the Naive Bayes Classifier (NBC), which is a simple probabilistic classifier that applies the Bayes Theorem. Naive Bayes Classifier is a classification algorithm that is effective (getting the right results) and efficient (the reasoning process is done by using existing inputs in a relatively fast way). The Naive Bayes Classifier algorithm aims to classify data in certain classes. The performance of classifiers is measured by the value of predictive accuracy [3]. It is expected that this work can provide the results of the classification of specialization automatically and have high accuracy.

2. Study of Literature

2.1. Data Mining

Data mining is the process of extracting information from a data set that uses algorithms and techniques used in the fields of statistics, machine learning, and database management systems [4]. In general, data mining is mining or discovering new information by looking for certain structures or rules from a very large amount of data [5]. Data mining is often also referred to as knowledge discovery in database (KDD). KDD is an activity that includes collecting, using data, historically to find order, patterns or relationships in large data sets [6].

Data mining is an activity of finding interesting patterns from large amounts of data, data can be stored in a database, data warehouse, or other information storage. Data mining is related to other fields of science, such as database systems, data warehousing, statistics, machine learning, information retrieval, and high-level computing. In addition, data mining is supported by other sciences such as neural networks, pattern recognition, spatial data analysis, database image, signal processing [7].

2.2. Classification

Classification is a process of grouping data based on certain characteristics into predetermined classes. Classification is also a search process for a set of models that distinguish data classes, to be used to predict the class of an object whose class is unknown [8]. In achieving these objectives, the classification process forms a model that is able to distinguish data into different classes based on certain rules or functions. The model itself can be an "if-then" rule, in the form of a decision tree, or a mathematical formula.

The classification technique works by grouping data based on training data and classification attribute values. The grouping rules will be used to classify new data into existing groups. In the classification process there are usually two processes that must be done, namely [10]:

- Training Process
  In this process, labels or attributes of training set data or sample data was known. The training set data is used to build the model.
- Testing Process
  This purpose of this process is to determine the accuracy of the models that have been made in the training process. So, the system can determine the label of the tested data

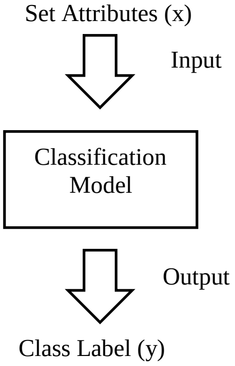


Fig. 1: Block classification model diagram

In Figure 1 the input data for classification is a collection of records or sample data. Each record is known as an instance, which is determined by a tuple (x, y), where x is a set of attributes and y is a certain attribute that states as a class label (also known as a category or target attribute) [9].

2.3. Request For Final Assignment

Student's individual interest can be seen from their tendency to be captivated or attracted to an experience and want to preserve that experience [10]. Specialization is a decision made by the student to choose a subject group according to their interest, talent, and ability as long as they study in academy. The selection of specialization is done on the basis of the need to work on the research report at the final assignment.

The final assignment is scientific papers worked by students. In the final project, students conduct a research to provide solutions to a problem. In conducting the research, students are guided by a supervisor.

The final assignment is one of the requirements for student graduation in completing the Diploma Program. The provisions regarding the final assignment are regulated by each faculty or department, following the standards of the College. The final assignment for diploma students is in the form of a final project. Specialization of the final assignment is the decision of students in choosing fields of interest in the preparation of the final assignment.

2.4. Naive Bayes Classifier

The Naive Bayes Classifier algorithm, also known as Bayesian Classification, is one of the classification techniques. Naive Bayes is a classification using probability and statistical methods presented by British scientist Thomas Bayes. Naive Bayes algorithm is based on probabilistic calculations with the assumption that each feature used is mutually independent.

Naive Bayes is the most popular text classification method used. This algorithm has advantages in terms of learning speed and its tolerance to the missing values of features. To handle numerical data, this algorithm uses probability density function, meaning that the data is considered to follow a normal distribution, then calculate the average value and its standard deviation [11].

To represent a class, there are characteristics of the instructions needed to do the classification. This instruction useful for explaining the probability of entering certain sample characteristics into the posterior class. Opportunities for the emergence of a class (before the entry of these samples), often called priors, multiplied by the chance of the emergence of global sample characteristics, which also called evidence. The evidence value is always fixed for each class in one sample. The posterior value is compared with the other posterior values of the class, to determine the sample class [12]. Classification of Naive Bayes is assumed that there are certain characteristics of a class that has nothing to do with the characteristics of other classes. The general equation of the Bayes theorem can be seen in the following equation 1:

$$P(H|X) = \frac{P(X|H)\ .\ P(H)}{P(X)} \tag{1}$$

Information:

X : data with unknown classes

H : the data hypothesis is a specific class

$P(H|X)$: the data hypothesis is a specific class

$P(H)$: the probability of hypothesis H (prior probability)

$P(X|H)$: the probability of hypothesis H (prior probability)

$P(X)$: probability X

The above equation is a model of the Naive Bayes theorem which will be used in the classification process. For classification with continuous data, the Gauss Density formula is used as shown in equation 2 below:

$$P(Xi = xi\,|Y = yj) = \frac{1}{\sqrt{2\pi\sigma ij}}\, e^{-\frac{(xi-\mu ij)^2}{2\sigma^2 ij}} \tag{2}$$

Informations:

P : probabilistic

Xi : attribute to i

Xi : attribute value to i
Y : class sought
Yj : Y class sub searched
μ : mean, expresses the average of all attribute
σ : standard deviation, declares a variant of all attributes

2.5. State Community Academy of Bojonegoro

State Community Academy of Bojonegoro is a PDD (Study Program Outside Domicile) of State Polytechnic of Malang in Bojonegoro. The legal basis for the program is:

- Law No. 12 of 2012,
- Minister of Education and Culture Decree No. 161 of 2012,
- Minister of Education and Culture No. 48 of 2013

Since 2012, Bojonegoro has been trusted by the Ministry of Education and Culture of the Republic of Indonesia as the organizer of the Study Program Outside the Domicile of State Polytechnic of Malang. The level of education held is the Diploma 2 Program, which aims to prepare a skilled middle workforce to meet the needs of workforces in Bojonegoro and its surroundings in the field of technology. State Community Academy of Bojonegoro Study Program includes Information Management, Accounting, and Automotive Engineering [13].

3. Methodology

For this work, there are three phase to do. First step is collecting data. And then we continue to train and test the system using data sets, that we get from collected data. And then we testing the system accuracy. The methodology diagram depicted in figure 2.

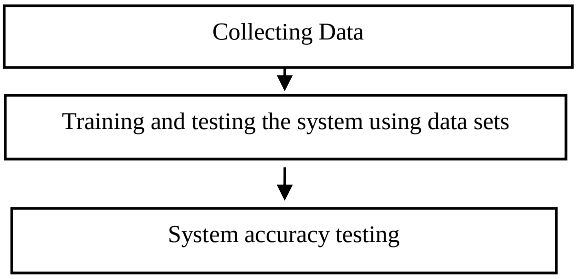


Fig. 2: Methodology diagram

3.1. Method of collecting data

The research begin with data collection. Data collection methods are used to collect data and information needed in the process of this work. We collecting the data by requesting the required data to the head of the Informatics Management Study Program State Community Academy of Bojonegoro.

From the data collection phase, data which obtained used as research objects, namely in the form of student data for the 2015-2016 period and student data for the 2016-2017 period, 2015/2016 curriculum, student's grade data, and actual data (data on prospective graduate in 2015).

3.2. Data Processing Methods

This phase is a core of this work. We use the Naive Bayes algorithm to process the data. In calculating the Naïve Bayes Classifier algorithm, several calculation steps are needed. The following are the stages in calculating the Naïve Bayes Classifier algorithm:

1. Learning by reading the training data and the form of classification on each training data. The example of training data can be seen on figure 3. Figure 3 is training data with numerical features. There are 17 numerical features which come from 17 supporting subjects [14]. The supporting subjects include basic multimedia, introduction to information technology, algorithms and programming, databases, operating systems, computer architecture organizations, human and computer interactions, visual programming, basic internet and web design, information systems, basic computer networks, management systems database, software engineering, web programming, applied multimedia, mobile programming, computer network management.
2. In this work, we using continuous or numerical data, then the data is processed by the Gauss Density formula, in the following way:
   a. Look for the mean value of each parameter (each feature) which is numeric data. The equation used to calculate the calculated average value (mean) can be seen in equation 3 and equation 4 as follows:

   $$\mu = \frac{\sum_{i=1}^{n} xi}{n} \tag{3}$$

   or

   $$\mu = \frac{x1 + x2 + x3 + \cdots + xn}{n} \tag{4}$$

   Information :
   xi : sample value to *-i*
   n *:* number of sample

   b. Look for the standard deviation of each parameter (each feature) which is numeric data. And the equation to calculate the standard deviation value can be seen in equation 5 as follows:

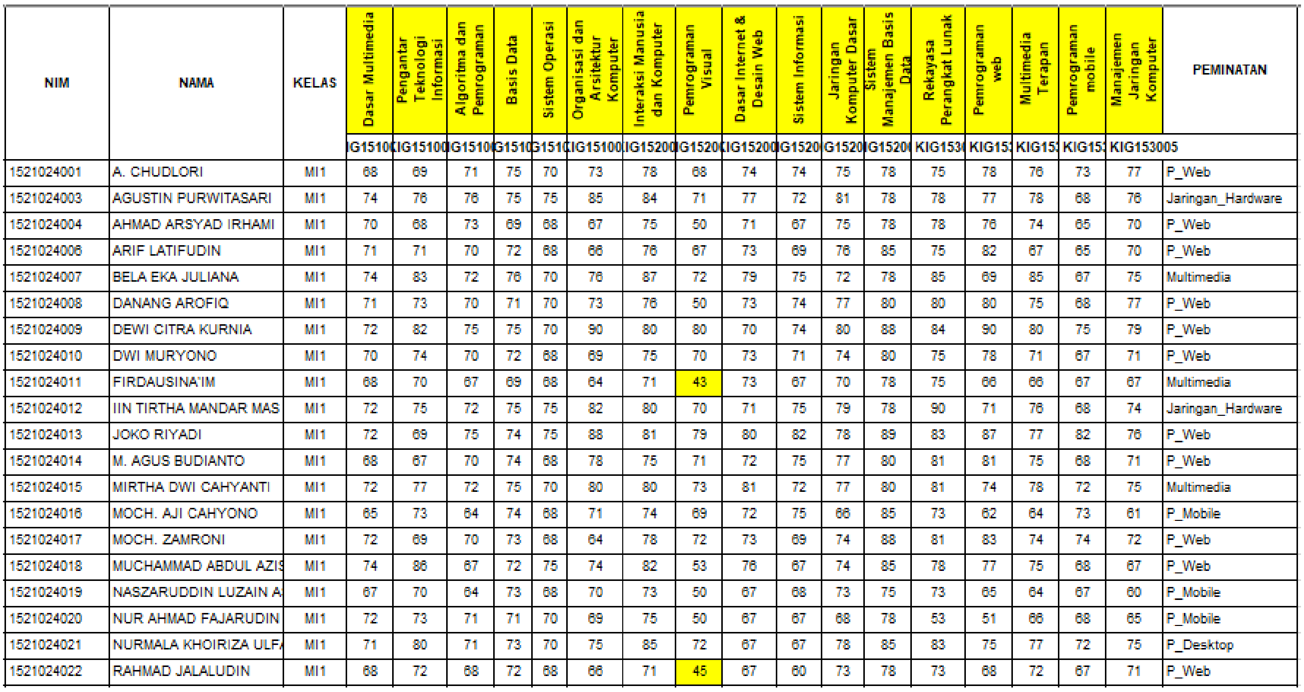
   $$\sigma = \sqrt{\frac{\sum_{1=1}^{n}(xi - \mu)^2}{n - 1}} \tag{5}$$

   Information :
   σ : standard deviation
   xi : value *x* to *-i*
   μ : average count (*mean*)
   n : number of sample

   c. Look for the probabilistic value of each parameter (each feature) by calculating the amount of the corresponding data from the same category divided by the amount of data in that category. And the equation for calculating probabilistic values (opportunities) can be seen in equation 2.
3. Get the value in the table mean, standard deviation and probability.
4. Calculate the total number of probabilistic values of all parameters or all features.
5. Get the results of the classification of student final assignments.

3.3. Testing Methods

The testing phase is the stage after the process of making the system is finished, the tests carried out to test the system are carried out with 2 testing steps, namely functional testing, and system accuracy testing. The test aims to ensure that the system built has been running in accordance with the Naive Bayes Classifier algorithm used.



NIM	NAMA	KELAS	Dasar Multimedia	Pengantar Teknologi Informasi	Algoritma dan Pemrograman	Basis Data	Sistem Operasi	Organisasi dan Arsitektur Komputer	Interaksi Manusia dan Komputer	Pemrograman Visual	Dasar Internet & Desain Web	Sistem Informasi	Jaringan Komputer Dasar	Sistem Manajemen Basis Data	Rekayasa Perangkat Lunak	Pemrograman web	Multimedia Terapan	Pemrograman mobile	Manajemen Jaringan Komputer	PEMINATAN
			IG1510(	IG15100	IG1510(	G1510	G151(	IG15100	IG15200	IG1520(	IG15200	IG1520	G1520	IG1520(	KIG153(	KIG15:	KIG15:	KIG15:	KIG153005	
1521024001	A. CHUDLORI	MI1	68	69	71	75	70	73	78	68	74	74	75	78	75	78	76	73	77	P_Web
1521024003	AGUSTIN PURWITASARI	MI1	74	76	76	75	75	85	84	71	77	72	81	78	78	77	78	68	76	Jaringan_Hardware
1521024004	AHMAD ARSYAD IRHAMI	MI1	70	68	73	69	68	67	75	50	71	67	75	78	78	76	74	65	70	P_Web
1521024006	ARIF LATIFUDIN	MI1	71	71	70	72	68	66	76	67	73	69	76	85	75	82	67	65	70	P_Web
1521024007	BELA EKA JULIANA	MI1	74	83	72	76	70	76	87	72	79	75	72	78	85	69	85	67	75	Multimedia
1521024008	DANANG AROFIQ	MI1	71	73	70	71	70	73	76	50	73	74	77	80	80	80	75	68	77	P_Web
1521024009	DEWI CITRA KURNIA	MI1	72	82	75	75	70	90	80	80	70	74	80	88	84	90	80	75	79	P_Web
1521024010	DWI MURYONO	MI1	70	74	70	72	68	69	75	70	73	71	74	80	75	78	71	67	71	P_Web
1521024011	FIRDAUSINA'IM	MI1	68	70	67	69	68	64	71	43	73	67	70	78	75	66	66	67	67	Multimedia
1521024012	IIN TIRTHA MANDAR MAS	MI1	72	75	72	75	75	82	80	70	71	75	79	78	90	71	76	68	74	Jaringan_Hardware
1521024013	JOKO RIYADI	MI1	72	69	75	74	75	88	81	79	80	82	78	89	83	87	77	82	76	P_Web
1521024014	M. AGUS BUDIANTO	MI1	68	67	70	74	68	78	75	71	72	75	77	80	81	81	75	68	71	P_Web
1521024015	MIRTHA DWI CAHYANTI	MI1	72	77	72	75	70	80	80	73	81	72	77	80	81	74	78	72	75	Multimedia
1521024016	MOCH. AJI CAHYONO	MI1	65	73	64	74	68	71	74	69	72	75	66	85	73	62	64	73	61	P_Mobile
1521024017	MOCH. ZAMRONI	MI1	72	69	70	73	68	64	78	72	73	69	74	88	81	83	74	74	72	P_Web
1521024018	MUCHAMMAD ABDUL AZIS	MI1	74	86	67	72	75	74	82	53	76	67	74	85	78	77	75	68	67	P_Web
1521024019	NASZARUDDIN LUZAIN A	MI1	67	70	64	73	68	70	73	50	67	68	73	75	73	65	64	67	60	P_Mobile
1521024020	NUR AHMAD FAJARUDIN	MI1	72	73	71	71	70	69	75	50	67	67	68	78	53	51	66	68	65	P_Mobile
1521024021	NURMALA KHOIRIZA ULFA	MI1	71	80	71	73	70	75	85	72	67	67	78	85	83	75	77	72	75	P_Desktop
1521024022	RAHMAD JALALUDIN	MI1	68	72	68	72	68	66	71	45	67	60	73	78	73	68	72	67	71	P_Web

Fig. 3: Example of raining data used, contain student data, grade, and specializations categories.

4. System Analysis and Planning

4.1. System Analysis

The developed system will be able to classify specialization in Web Programming, Desktop Programming, Mobile Programming, Multimedia, and Network / Hardware. The system is a unified process that can generate classification results according to the inputted data. The detail of the process, shown in figure 4.

The system has four sub-processes that compile so that it can produce a complete process. Sub-processes contained in the system are as follows:

1. Initial Phase
   This is the phase to process the dataset before it can be used for training and testing. The dataset obtained is student data in 2015 as many as 160 data and data in 2016 as many as 150 data. We preprocessing the data by deleting data when there is a blank data. This should be done, because if there is a blank value then the data cannot be processed. The final assignment requirement is the student must have completed the semester 1 to semester 3.
   After preprocessing, from the 2015 we get 100 data and 2016 get 150 data to be used as training data. Then, we give label or
   classify manually to the data. In manual labeling or classification, it is done by making a recommendation by calculating the value based on a cognate subject for each specialization category. Specialization categories are divided into Web Programming, Desktop Programming, Mobile Programming, Multimedia, and Network / Hardware. Labeling is done by the head of the Informatics Management Study Program. The labeling is intended to provide a manual classification of specializations that have been obtained. Giving these labels will be useful for the training and testing process. 60 student data from 2015 not labeled, because it will be used for the next learning process.
2. Training Phase
   We use Naive Bayes Classifier algorithm for classification, in the training phase. The grade of supporting subjects, which have been calculated and labeled in the initial stages, is used as input data for training. There are 250 data that used for the training, from 310 data of datasets.
3. Testing Phase
   The testing phase is a process to generate a classification of specializations based on the classifier model that has been generated from the training process. The grade of supporting subjects, which not labeled, is used as input data for testing. Later, the data will be labeled by the system, which applying Naive Bayes Classifier method. The data used for the testing process are 60 data, from the total number of datasets are 310 data. The result of testing data is a classification of specializations in Web Programming, Desktop Programming, Mobile Programming, Multimedia, and Network / Hardware.
4. Accuracy Testing Phase
   We test the level of accuracy by comparing the classification label given by the system with the actual data. The number of similarities between the classification label given by the system and the actual data is the value of accuracy obtained

5. Implementation and Testing

We do the implementation by creating code for the classification process, we also create the interface for the user. Afterward, we conduct two forms of testing, namely system functionality testing, and system accuracy testing. We use black box testing for functionality testing, to observe the results of execution through test data and inspect the function of the system that has been completed. Functionality testing scenarios include testing menus and features in the classification system of student final assignments.

After functionality testing, we conduct system accuracy testing for this work. The purpose of the test is to observe the results of the classification carried out by the system using the Naive Bayes Classifier algorithm using formula in equation 6 [15].

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Number\ of\ prediction} \times 100\% \quad (6)$$

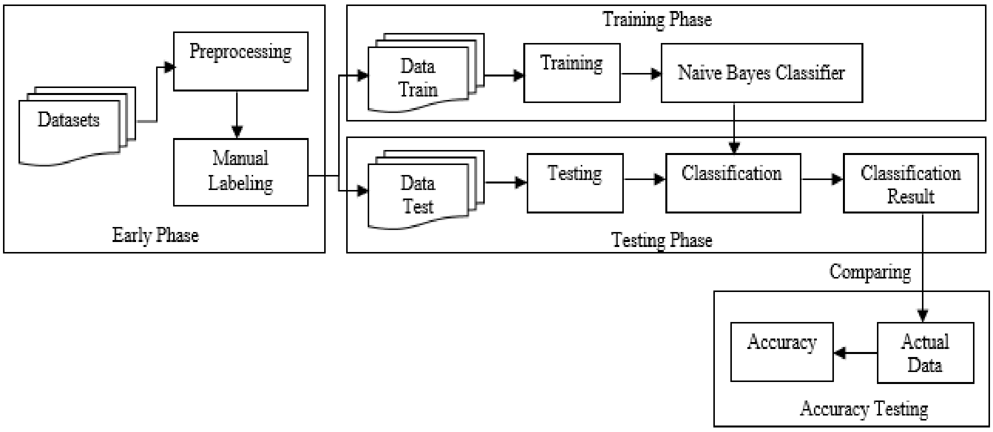


Fig. 4: Training data used, contain student data, grade, and specializations categories

We use 310 datasets, where the training data is 250 data and the testing data is 60 data. For the training data, 100 data taken from 2015 student data, and 150 from 2016. The testing data, 60 data taken from 2015 student data. The results of the testing, there are 50 data that are appropriate or data with the correct results and 10 data are not appropriate or the results are wrong. Using the equation 6, we calculate the accuracy as follows.

$$Accuracy = 50\ /60 \times 100\% = 83,33\%$$

Measurement results were carried out using 250 training data and 60 testing data resulting in an accuracy rate of 83.33%.

6. Discussion and Conclusion

In this research, the system has succeeded in giving a label. Labeling is based on patterns from training data which have been processed with machine learning using the Naive Bayes Classifier algorithm. In practice, there are anomalies. The following are our analysis related to these anomalies:

a. If the value data is different from the training data, it will affect the label.
b. If the value data has the same pattern as the training data, the data will remain the same, even though the value is reduced or added, it does not affect the label
c. If the value data does not have a pattern on the training data, it affects the label.
d. The inputted data for all supporting course values are the same value, for example 50. It will result in P_Mobile specialization as the default because P_Mobile Specialization is an initial specialization that is processed or counted first. So, for example, specialization of class X which is first processed/calculated for the first time will become the default interest.
e. Value data that is entered if it does not meet a class that is cognate with specialization then will affect the label.

Based on the research and testing that have been done we can take conclusions as follows:

- The Naive Bayes Classifier algorithm can be used for the classification of final assignments for students of the State Community Academy of Bojonegoro Informatics Management Study Program.
- The results of the classification system using the Naive Bayes Classifier algorithm are quite good, which produces an accuracy value of 83.33%.

Acknowledgement

We would like to thank to the State Polytechnic of Malang and the Research Center Office (P2M) who have supported this research project.

References

[1] Meilani BD & Susanti N (2014), “Aplikasi Data Mining Untuk Menghasilkan Pola Kelulusan Siswa Dengan Metode Naive Bayes”. Jurnal LINK , Vol.21, No.2, (2014), pp:1-6.
[2] McLeod JrR & Schell GP (2007), "Management Information System. 10 th ed" Pearson Education, Inc.
[3] Zhang H, Jiang L, & Su J (2005), “Augmenting Naive Bayes for Ranking”, Proceedings of the 22nd International Conference on Machine Learning, (2005), pp.1025-1032.
[4] Ranjan J (2007), “Application of Data Mining Technique in Pharmaceutical Industry”, Journal of Theoritical and Applied Information Technology, Vol 3, (2007), pp: 61 – 67.
[5] Davies & Beynon P. (2004), “Database Systems Third Edition”, Palgrave Macmillan, New York.
[6] Santoso B (2007), “Data Mining: Teknik Pemanfaatan Data Untuk Keperluan Bisnis”, Yogyakarta: Graha Ilmu.
[7] Han J & Kamber M. (2006), “Data Mining : Concept and Techniques Second Edition”, Morgan Kaufmann Publishers.
[8] Mulyanto A. (2009), “Sistem Informasi Konsep & Aplikasi”, Yogyakarta: Pustaka Pelajar.
[9] Nugroho A & Subanar (2013). "Klasifikasi Naive Bayes untuk Prediksi Kelahiran pada Data Ibu Hamil", Berkala MIPA , Vol.23, No.3, (2013), pp:297-308.
[10] Luthfiansyah DH (2016), “SPK Pemilihan Jurusan Berdasarkan Kuisoner Minat Bakat Menggunakan Metode Naive Bayes”, Seminar Informatika Aplikatif Polinema, (2016).
[11] Musthofa M (2016), “Pengembangan Sistem Pendukung Keputusan Penjurusan Bagi Siswa Baru Menggunakan Metode Naive Bayes”, Seminar Informatika Aplikatif Polinema, (2016).
[12] Saraswati NW (2013), “Naïve Bayes Classifier Dan Support Vector Machines Untuk Sentiment Analysis”, Seminar Nasional Sistem Informasi Indonesia, (2013).
[13] PDD Polinema Rintisan AKN Bojonegoro, “Politeknik Negeri Malang Program Studi Diluar Domisili Di Kabupaten Bojonegoro (Rintisan Akademi Komunitas Negeri Bojonegoro)”, Dasar Hukum, Program Studi, Unit Kegiatan Mahasiswa, Kerjasama Internasional, url: http://www.aknbojonegoro.ac.id/ [accessed: January 25th, 2018].
[14] Tim Penyusun (2015), "Pedoman Akademik AKN Bojonegoro, Bojonegoro".
[15] Muktamar BA, Setiawan NA & Adji TB (2015), "Analisis Perbandingan Tingkat Akurasi Algoritma Naïve Bayes Classifier Dengan Correlated-Naïve Bayes Classifier", Seminar Nasional Teknologi Informasi dan Multimedia, (2015), pp: 49-54.