CLASSIFICATION OF SCOUT SKILLS USING NAIVE BAYES ALGORITHM

[1]RIZKY YUDHA PRAMUDHIKA, [2]RIFKI FAHRIAL ZAINAL, [3]RANI PURBANINGTYAS

[123]Informatics Engineering, Bhayangkara Surabaya University
Jl. Ahmad Yani No 114 Surabaya
E-mail: [1]pramudhika13@gmail.com, [2]rifki@ubhara.ac.id, [3]rani@ubhara.ac.id

ABSTRACT

In the development of skills and skills of students Scout often finds the problem is the difficulty of developing the skills of learners caused by mistakes in determining the skills dominated by learners. How you can do it to solve that problem is by making a class determination or classification of areas of expertise controlled by the learner. This study aims to build a system to determine the inner class Scout skills area using Naïve Bayes algorithm for Scout Coach to determine the area of expertise of learners. So Scout Coach can develop the skills of learners in accordance with the areas of expertise that are owned optimally. Assessment criteria used are the values of General Knowledge, Scout Knowledge, Sign Language (Password, Morse, Semaphore), Node Bond, Pioneering, and Hasta Karya. All assessment criteria are used numerically. The resulting classification is the students who are included in the class Intelligence (Intelligence), Physical (Strength), or Creativity. The results of this study is a program that can perform calculations to determine skill class scout field. The Scouting Skill Classification Program using the Naïve Bayes Algorithm has been tested and its accuracy or accuracy is 100%.

Keywords: *Classification, Naïve Bayes, Java, Scouting*

1. INTRODUCTION

There are several skills taught in Scouting organizations. These skills can be applied in social life. The skills are tested in the Scouting competition. Skill members' development skills often experience difficulties. This is due to errors in determining the skills controlled by learners. Computer programs can be used to solve the problem. The problem formulation is how to design a system to classify Skill skills using Naïve Bayes algorithm. The data used comes from 5 junior high schools in Tumpang District and surrounding areas. Each data consists of general Knowledge, Scouting Knowledge, Sign Language (Password, Morse, Semaphore), Knot Ties, Pioneering, and Hasta Karya. The results of the classification are classes of Intelligence, Physical, or Creativity. The purpose of this study is to facilitate Scout Coach and Coach to determine which areas of expertise are mastered by students.

2. METHODOLOGY

2.1 System Analysis

This stage aims to explain the problems of the system, analyzing the needs of the system in the process of classification areas of Scouting skills.

2.1.1 Problem Analysis

Scouting activities have become part of the Indonesian government's program to improve the quality of education in Indonesia. One way to measure the ability of Scout members is to follow the race. Participants of the competition are Scout members from educational institutions. The Scouting Competition includes classes of expertise that can be classified into classes of intelligence, physics, and creativity. Using the data mining system, specifically the Naïve Bayes classification, a program was developed to classify areas of Scouting expertise.

2.1.2 Data Analysis

Data used in this research is data from 2015-2016 from 5 junior high schools in Tumpang District, Malang Regency, and surrounding areas. Each data consists of General Knowledge Scores, Scouting Knowledge, Sign Language (Password, Morse, Semaphore), Knot Ties, Pioneering, and Hasta Karya, as well as the original class of participants. There are 712 data available, consisting of 447 training data and 265 test data.

2.2 System Design
The designed system consists of 2 main parts, namely the master data process and the process of calculating the Naïve Bayes algorithm.

2.2.1 Master Data Process
The master data process contains data to be used as training data and test data. In the data master process there is a function to add data, change data, and delete data. The input data required by the system is participant data for training and participant data for testing.

2.2.2 Calculation Process of Naïve Bayes Algorithm
Naïve Bayes model development has several stages. The steps are described in the following figure.

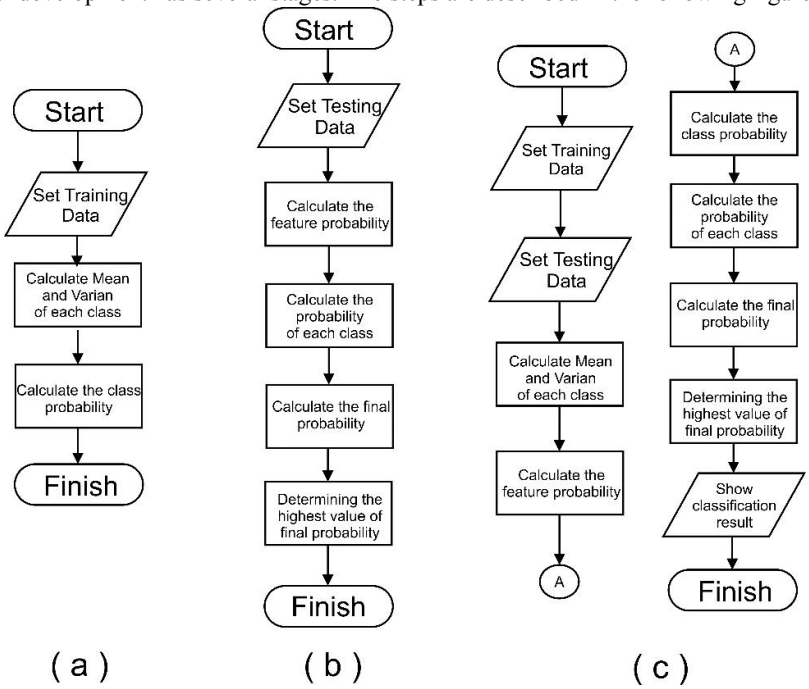

Figure 1. (a) Flowchart of Train Process, (b) Flowchart Test Process, (c) Flowchart Implementation Process

All features are numeric, so the feature probability calculation is not included in the training process. But the feature probability is calculated in the testing process by including the mean and variance of the training process. The entire process of training and testing is included in the implementation process. Once the highest end probability value is known, the system will display the classification results with the Naïve Bayes algorithm.

2.2.2.1 Example Calculation Classification Skills Scouting Field Using Naive Bayes
To perform the process of calculating Naïve Bayes classification, the training data and test data are required by algorithm. There are some data to be used as training data.

Table 1. Example of Train Data

NO	NAME	SCHOOL	PU	PK	SMS	SI	PIO	HK	CLASS
1	BAGAS IMAM ABDILLAH	SMP 1 TUMPANG	80	84	86	96	98	78	FISIK
2	RAKA IQBAL RAMADHAN	SMP 1 TUMPANG	88	86	88	76	74	98	KREATIFITAS
3	BEBBYTO AYUMI R.E.	SMP 1 TUMPANG	100	100	100	70	60	72	INTELIGENSI
4	RYAN HIDAYAT	SMP 2 TUMPANG	94	90	90	82	76	78	INTELIGENSI
5	KHOIRUL DWI ANAM	SMP 1 PAKIS	80	78	74	88	90	80	FISIK
6	EKO SETIAWAN	SMP DIPONEGORO TUMPANG	76	78	78	88	82	76	FISIK
7	RISKA FANIDHATUS	SMP 1 PONCOKUSUMO	70	78	64	94	88	78	FISIK
8	EKA NUR WAHYUDI	SMP DIPONEGORO TUMPANG	94	84	88	80	70	78	INTELIGENSI
9	FEBBY ANTI S.	SMP 1 PAKIS	80	78	72	84	76	94	KREATIFITAS
10	ROSALIA PUTRI ERLANDA	SMP 1 PAKIS	82	84	76	72	68	90	KREATIFITAS

From the sample train data, the mean and variance values are calculated. The mean of each feature is calculated based on belonging in the class where the data is. For example, there are 3 data in the class of intelligence. That is, only the values of those 3 data are calculated to find the mean of each feature for the class of intelligence. The mean of the training data can be seen in the following table.

Table 2. Mean Of Train Data

CRITERIA	INTELIGENSI	FISIK	CREATIVITY
PU	96.00	76.50	83.33
PK	91.33	79.50	82.67
SMS	92.67	75.50	78.67
SI	77.33	91.50	77.33
PIO	68.67	89.50	72.67
HK	76.00	78.00	94.00

The next step is to calculate the variant and final variant of the training data, following example calculation of variant and final variant for PU matlom Intelligence class:

$$s^2 \text{ PU} \mid \text{INTELIGENSI} = \frac{(100-96)^2+(94-96)^2+(94-96)^2}{3-1}$$

$$s^2 \text{ PU} \mid \text{INTELIGENSI} = \frac{24}{2} = 12$$

$$s \text{ PU} \mid \text{INTELIGENSI} = \sqrt{12} = 3.46$$

The value of the variant and the final variant can be seen in the following table.

Table 3. Variant Values From Examples of Train Data

CRITERIA	INTELIGENSI	FISIK	CREATIVITY
PU	12.00	22.33	17.33
PK	65.33	9.00	17.33
SMS	41.33	83.67	69.33
SI	41.33	17.00	37.33
PIO	65.33	43.67	17.33
HK	12.00	2.67	16.00

Table 4. The End Variance Of Example Train Data

CRITERIA	INTELIGENSI	FISIK	CREATIVITY
PU	3.46	4.73	4.16
PK	8.08	3.00	4.16
SMS	6.43	9.15	8.33
SI	6.43	4.12	6.11
PIO	8.08	6.61	4.16
HK	3.46	1.63	4.00

After the mean, varian, and final variants are obtained, the next step is to calculate the probability value of features for numerical features. For that required test data. Here is the data for testing.

Table 5. Examples of Test Data

NAME	SCHOOL	PU	PK	SMS	SI	PIO	HK	CLASS
KHARISMA PRAMISWARA	SMP 1 TUMPANG	86	88	86	80	72	74	

The Naive Bayes algorithm computes the data one by one to classify it by its class through several stages in it. So one data for testing is enough to test the calculation of Naïve Bayes. All features are numerical, here is a formula for calculating the probability of numerical features:

$$P(X_i = x_i \mid Y = y_i) = \frac{1}{\sqrt{2\pi}\,\sigma ij} \exp^{-\frac{(xi-\mu ij)^2}{2\sigma ij^2}}$$

Where :

$P(X_i = x_i \mid Y = y_j)$ = Probabilitas kondisional dari fitur numerik pada tiap kelas.

μ_{ij} = Mean dari sampel X_i ($\bar{x}$) dari semua data latih dalam kelas y_j.

σ_{ij} = Varian akhir dari sampel (s) dari data latih.

σ^2_{ij} = Varian dari sampel (s^2) dari data latih.

Here is the probability value of the numerical feature of the participant named Kharisma Pramiswara which is used as the test data.

Table 6. Probability of Numerical Features of Participants Kharisma Pramiswara

CLASS	CRITERIA					
	PU	PK	SMS	SI	PIO	HK
INTELLIGENCE	0.0018	0.0453	0.0363	0.057	0.0453	0.0975
PHYSICAL	0.0112	0.0024	0.0226	0.0019	0.0018	0.0122
CREATIVITY	0.0781	0.0422	0.0325	0.0594	0.0946	0.0000

After the probability of numerical features is obtained, then calculate the probability value of the class. The probability of a class is calculated by dividing the amount of data in each class by the sum of all training data. Here is the class probability value.

Tabel 7.Nilai Probabilitas Kelas dari Data Latih

CLASS	INTELIGENSI	FISIK	CREATIVITY
SUM	3	4	3
P	0.3	0.4	0.3

The next step is to calculate the probability of each class. This calculation requires the probability value of numerical features of the participants of Kharisma Pramiswara. This calculation is done by multiplying all probability values of numerical features in the same class. Here is the formula of calculating the probability value of each class.

$$P(\text{Class}) = P(PU \mid \text{Class}) \times P(PK \mid \text{Class}) \times P(SMS \mid \text{Class}) \times P(SI \mid \text{Class} \times P(PIO \mid \text{Class}) \times P(HK \mid \text{Class}) \quad (3)$$

Here are the calculations for the class of intelligence:

P (INTELLIGENCE) = 0.0018 × 0.0453 × 0.0363 × 0.057 × 0.0453 × 0.0975

= 0.0000000007394

Here is a calculation for the physical class:

P (PHYSICAL) = 0.0111 × 0.0024 × 0.0226 × 0.0019 × 0.0018 × 0.0122

= 0.00000000000002649

Here are the calculations for the creativity class:

P (CREATIVITY) = 0.0781 × 0.0422 × 0.0325 × 0.0594 × 0.0946 × 3.718e-07 = 0.0000000000002237

After the probability value of each class is found, then done the final probability calculation of the test data. To calculate the final probability value, the probability value of each class is obtained from the class probability data from the training data. Here is the calculation.

Calculation of the final probability value of the class of intelligence:

P (INTELIGENSI) = 0.0000000007394 × 0.3

= 0.0000000002218

Calculation of final physical probability value of class:

P (FISIK) = 0.00000000000002649 × 0.4

= 0.00000000000001059

Calculation of final grade probability value of creativity:

$$P(KREATIFITAS) = 0.0000000000002237 \times 0.3$$

$$= 0.00000000000006712$$

The final stage of Naïve Bayes algorithm calculation is to compare the value of each final probability that has been obtained before. The data will be classified based on the final probability value tertunggi. The highest probability end value of the test data is the class of intelligence. That is, participants of Kharisma Pramiswara including having a field of Scouting skills in the class of intelligence.

2.3 Data Flow Diagram (DFD)

Data flow diagram is a method of development of a structured data system. DFD describes all activities in the system clearly.

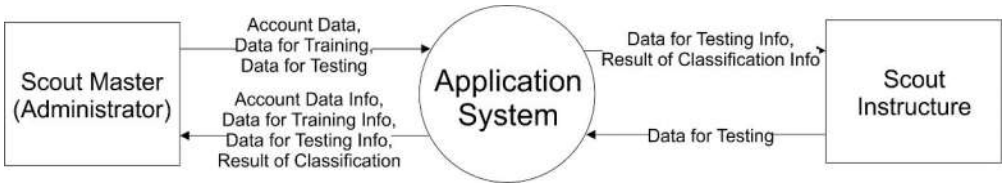

Figure 2. Data Flow Diagram

Based on Figure 2, the Scout Coach is the system administrator, and the Scout instructor is a system user. Scout coaches have full access to the system. Scout master can manage the training data and test data, but it can also access the classification calculation process. While Scout instructors can only access test data and classification results.

2.4 Entity Relation Diagram (ERD)

ERD is a major data modeling that helps organize data in projects into entities and determines relationships between entities. ERD is also a model to explain the relationship between data in the database based on objects that have relationships between entities. In this case, the ERD design explains the relationship between attributes used in Naïve Bayes algorithm calculations.

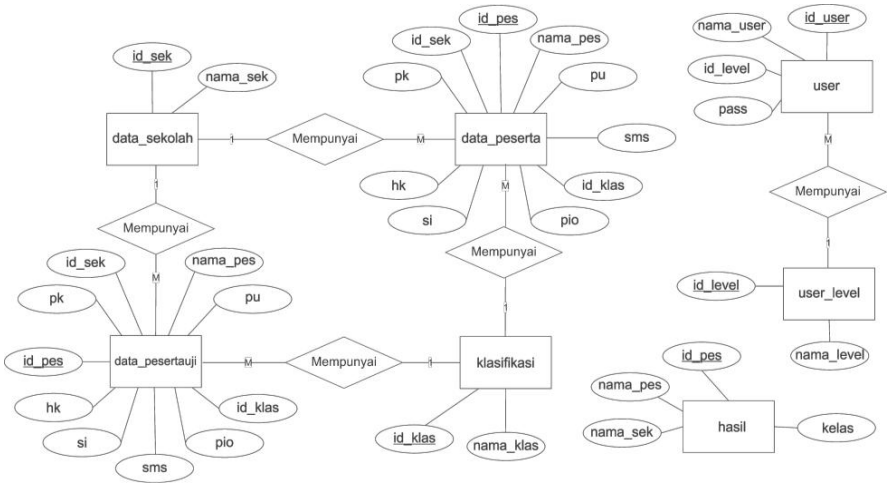

Figure 3. Entity Relation Diagram of the System

3. RESULTS

This research produces a program that can classify areas of expertise controlled by Scout members. The program is built based on Java programming language. The program interface is shown in figure 4.

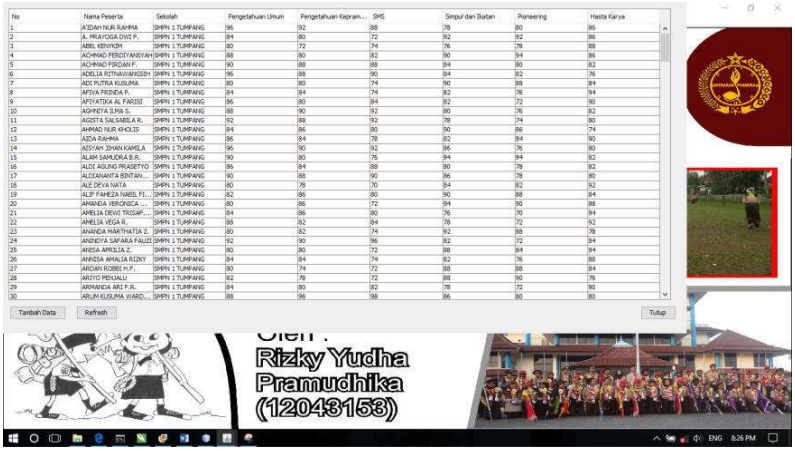

Figure 4. Master Data Interface of the Program

In the Master Data process, the administrator can manage the data used in the system. By clicking on the name of the data, the administrator can change the details of the data.

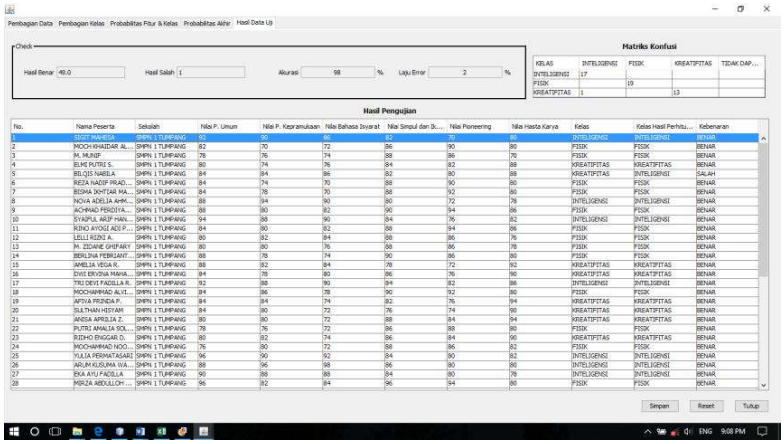

Figure 5. Results Classification in Naïve Bayes Calculation Process

To perform calculations, the first time the user must set the train data. Users simply type in the amount of train data, then click the "Data Set" button. Then click the "Data Set" button on the right side of the test data amount column. Then click "Next" button and follow the steps. In the last panel, the user can see the result of classification. Users can choose whether or not to save their classification results. Unsaved classification results, can not be seen in the classification report page.

3.1 System Testing

The testing phase is done to find the accuracy value of system calculation using Naïve Bayes algorithm. Because the calculation of the participant's son and daughter is done separately, the test of the system is performed twice for each calculation.

Table 8. Test Result Analysis

Testing	***Gender***	**Number of Train Data**	**Number of Test Data**	**Same Results**	**Different Results**	**Accuracy Value (%)**	**Rate Values Error (%)**
First	Men	150	125	125	0	100	0
	Women	150	140	140	0	100	0
Second	Men	207	125	125	0	100	0
	Women	240	140	140	0	100	0
Average						100	0

As can be seen in table 8, with some of the training data used, the system calculation accuracy has reached 100% for both genders. Similarly, testing using all available trainer data, test results show that there are no test data that are classified as different from the original class. This means that the program's accuracy reaches 100%. The average value of the calculation accuracy is 100%, and the error rate is 0%.

3.1.1 Analysis of Class Effects of Train Data on Classification Results

From system testing, it is known that its accuracy reaches 100%. Furthermore, analysis of the tests will be carried out by including incorrectly classified data into the training data. There are 60 incorrect data for each gender. The class of data has been changed into the wrong class. The following is the result of the test.

Table 9. Analysis of Class Effects of Train Data on Classification Results

Gender	**Number of Train Data**	**Number of Test Data**	**Same Results**	**Different Results**	**Accuracy Value (%)**	**Rate Values Error (%)**
Men	207	125	113	12	90.4	9.6
Women	240	140	134	6	95.7	4.3

As shown in table 9, there are some data that are classified differently from the original class. For testing on the male participants, there are 12 different results. The accuracy value decreased to 90.4%, and the error rate increased to 9.6%. While the test of female participants resulted in 6 different results. Accuracy decreased to 95.7% while the error rate increased to 4.3%. This suggests that the determination of the training data class has an effect on the final classification results using the Naïve Bayes algorithm.

4. CONCLUSION

The Naïve Bayes algorithm has a high degree of accuracy in the Scout classification skill classification, which achieves 100% accuracy when using all available trainer data. The available train data are 207 data for son and 240 data for daughter. Based on the results of the tests using accurate data for training, as well as incorporating inappropriate data for the training, it can be seen that variations in the training data class may affect the classification results using the Naïve Bayes algorithm. Data class determination when adding new data for training can affect the variation of class probabilities in the calculation, thus affecting the final classification results using the Naïve Bayes algorithm. Data for training is still taken randomly by the system, for further research it is advisable to develop a program with a function to select data for training. For further development, it is hoped that the program will be built based on web browser for wider access by users.

REFERENCES

[1] Aditya, Muhammad. (2013), Klasifikasi Calon Pendonor Darah dengan Metode Naïve Bayes Classifier, Brawijaya University, Malang.
[2] Agung Baktiar, Yoga. (2013), Implementasi Metode Naïve Bayes untuk Klasifikasi Kenaikan Grade Karyawan Pada Fuzzyfikasi Data Kinerja Karyawan(Studi Kasus PT PJB UP Brantas), Brawijaya University, Malang.
[3] Anugroho, Prasetyo. (2011), Klasifikasi Email Spam dengan Metode Naïve Bayes Classifier Menggunakan Java Programming, State Electronic Polytechnic of Surabaya-Institute of Technology Sepuluh November, Surabaya.
[4] Bob Sunardi, Andri. (2014), Boyman, Ragam Latih Pramuka, Ninth Print, Nuansa Muda, Bandung.
[5] Bustami. (2012), Penerapan Algoritma Naïve Bayes Untuk Mengklasifikasi Data Nasabah Asuransi, Malikussaleh University, Banda Aceh.
[6] Distiawan Trisedya, Bayu. (2009), Klasifikasi Dokumen Menggunakan Algoritma Naïve Bayes dengan Penambahan Parameter Probabilitas Parent Category, Indonesia University, Jakarta.
[7] Fahrurozi, Ahmad, (2014), Klasifikasi Jenis Kayu Dengan Menggunakan Naïve Bayes Classifier, Institute of Technology Sepuluh November, Surabaya.
[8] Hamzah, A. (2012), Klasifikasi Teks dengan Naïve Bayes Classifier (NBC) untuk Pengelompokan Teks Berita dan Abstract Akademis, Institute of Science and Technology AKPRIND, Yogyakarta.
[9] Natalius, Samuel. (2011), Metoda Naïve Bayes Classifier dan Penggunaannya pada Klasifikasi Dokumen, Institute of Technology Bandung, Bandung.
[10] Prasetyo, Eko. (2012), Data Mining Konsep dan Aplikasi Menggunakan Matlab, Ed.1, Andi, Yogyakarta.
[11] Ridwan, Mujib et al. (2013), Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naïve Bayes Classifier, Brawijaya University, Malang.
[12] Rozaq, Abdur. (2013), Klasifikasi Dokumen Teks Berbahasa Arab Menggunakan Algoritma Naïve Bayes, Institute of Technology Sepuluh November, Surabaya.
[13] Saleh,Alfa. (2015),Implementasi Metode Klasifikasi Naïve Bayes dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga, Potensi Utama University, Medan.
[14] Satria,Danil.(2009),Gerakan Pramuka Indonesia,https://id.wikipedia.org/wiki/Gerakan_Pramuka_Indonesia, accessed at March 24, 2016.
[15] Septian Nugroho, Yuda. (2014), Data Mining Menggunakan Algoritma Naïve Bayes untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro, Dian Nuswantoro University, Semarang.