NAIVE BAYES METHOD FOR PROCESS FOR SELECTION OF PPA SCHOLARSHIP TO STUDENTS:
CASE STUDY OF INFORMATICS AND COMPUTER MANAGEMENT ACADEMY OF PAKARTI LUHUR
TANGERANG

BADRUS ZAMAN SAPTO MARDI PUTRO[1], Krisna Adiyarta M[2], Sri Wahyuningsih[3], Pepi Permatasari[4]

Master of Computer Science-Postgraduate Program, Budi Luhur University
Jalan Raya Ciledug, Petukangan Utara, Kebayoran Lama, South Jakarta 12260
Tel. (021) 5853753, Fax. (021) 5869225
E-mail : badrus.zaman@pakartiluhur.ac.id[1], krisna.adiyarta@gmail.com[2],
sri.wahyuningsih@pakartiluhur.ac.id[3] , pepi.permatasari@pakartiluhur.ac.id[4]

ABSTRACT

The selection process for PPA Scholarship students currently at Pakarti Luhur AMIK is not subjective, namely through meeting results so that the accuracy of the results is reduced. This decision is adjusted by several requirements criteria so that the selection of PPA scholarship recipients in accordance with the requirements contained in the Pakarti Luhur scholarship can be chosen, which is based on the Academic Achievement Scholarship (PPA) guidelines, circulated by the Ministry of Research, Technology and Higher Education.
This study is in the form of a case study with research and development methods whose subject of research is the Information and Computer Management Academy of Pakarti Luhur. This model is designed for the selection process of students who have the opportunity to win PPA scholarships, with the Naive Bayes method used. The results of the model with the Naive Bayes Method in the process of selecting desktop-based PPA Scholarship winners display features of data class input menu, input data requirements, student input to be selected, data conversion process to training data, validation process, and data testing.

The results of the model design can provide objective results and ease in helping to carry out the selection process. The test results also show that the use of the Naive Bayes method can be used as an alternative in making decisions in the selection of PPA Scholarship students who have an accuracy of 93.4783%.

1. INTRODUCTION

All citizens who are at school age, from the primary level, to the level of lectures must receive education, so that they become intelligent people, as quoted from the Indonesian Constitution. There are still many obstacles that parents have to send their children to school, due to unaffordable education costs.

With such conditions, the Government seeks the allocation of funds so that economically disadvantaged students cannot be helped by the cost of education, and for students who have high achievements in curricular or extracurricular activities will be given scholarships by the Government. The government will channel it through the Directorate General of Higher Education.

With the guidance of the Director General of Higher Education with the aim that students can propose themselves as prospective students who win scholarships or help with tuition fees easily, students who achieve it can be lighter to exercise their rights and obligations, and so that the program can be accomplished with a target, amount and the right time.

So far, what has been done by AMIK Pakarti Luhur for the selection process for scholarship admissions is still done subjectively, namely from the results of the meeting so that the results of the selection process are not accurate based on the specified criteria. Therefore, it is necessary to have a system for selecting scholarship receipts, whether students who will receive the scholarship are accepted or rejected.

Naïve Bayes Classifier, one of the classification methods that may be used by BAAK AMIK Pakarti Luhur to select scholarship receipts.

2. THOUGHT BASIS
2.1. Scholarship

In order to fulfill the rights of outstanding students, the government must provide a scholarship.

Assistance or mitigation of education costs, loan funds that must be repaid after graduation and or obtaining employment, and without interest, have been set forth in Law No. 12 of 2012 concerning Higher Education, in Article 76 Paragraph (2) [1] is clearly stated.

2.2. Academic Achievement Improvement Scholarship (PPA)

Explained in 2017, in the Academic Achievement Improvement Scholarship (PPA) Guidelines [2] by the Directorate General of Learning and Student Affairs Ministry of Research, Technology and Higher Education. Then, the term Academic Achievement Improvement Scholarship (PPA) and Student Learning Aid (BBM), in 2012 was adjusted to be Academic Achievement Improvement Scholarship (PPA-Scholarship) and Academic Achievement Improvement Education Assistance (BPP-PPA). With the publication of this guideline, the selection process, distribution or scholarship is expected to run optimally, and the tuition fees for students can be alleviated and the education attended by smooth students, as well as their educational achievements can be improved and the education undertaken by students can also be completed properly and appropriately the time.

2.3. SDLC (System Development Life Cycle)

An information system can be said to be good if the implementation is managed properly. In developing information systems, SDLC (System Development Life Cycle) is a widely used method. This is a reference for the development and management of an information system well. Understanding SDLC is an ongoing process of planning, analysis, design and implementation.
Which in each process is carried out gradual improvement (Dennis et all, 2012) [3], we can see in Figure II-1.

Figure II-1 SDLC Stages (Dennis et all, 2012).

2.4. Unified Modelling Language (UML)

UML (Unified Modelling Language) Is an object-based general vocabulary and engineering diagram that is effective for modeling each system development project from the analysis stage to the design and implementation stage (Dennis et all, 2012) we can see in Figure II-2.
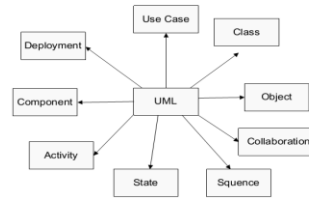

Figure II-2 UML Diagram (Dennis et all, 2012)

2.5. Black Box Testing

Black box testing is testing that is done by observing the results of execution through test data and checking the functional software.

Figure II - 5: Black Box Testing (Nidhra and Dondeti, 2012) [4]

Black box testing is also called functional testing, a functional testing technique that designs case tests based on the specifications of an information (Nidhra & Dondeti, 2012).

2.6. Naïve Bayes

Naïve Bayes is a probabilistic method of simple classification based on Bayes's Theorem published by Thomas Bayes, where classification is carried out through an efficient training data. Naïve Bayes has the assumption that giving input to a value from the attribute in the class does not depend on the value of the other attributes.

Bayes theorem is a theorem put forward by British scientist Thomas Bayes, which predicts an opportunity in the future based on previous experience so that it is known as the Bayes Theorem (Diasrina Dahri, Fahrul Agus, Dyna Marisa Khairina, 2016) [5]. The Bayes Theory Equation is:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \qquad (1)$$

Where :

X : Data on a class that is unknown.
C : a specific class based on the X Data Hypothesis.
P (C|X): Probability of hypothesis C based on condition X (posterior probability).
P(C):Probability of hypothesis C (prior probability)
P(X|C):Probability X is based on the conditions of the hypothesis. C
P(X) : Probability of X.

To explain the Naive Bayes theorem, please note that the classification process requires a number of instructions to determine what class is suitable for the sample analyzed [9].

Therefore, the bayes theorem in equation (1) is adjusted to equation (2):

$$P(C|X1 \dots Xn) = \frac{P(C)P(X1\dots XnvC)}{P(X1\dots Xn)} \qquad (2)$$

Where variable C represents the temporary class variable X1 ... Xn represents the characteristics of the instructions needed to classify or criteria. Then the formula explains that the probability of entering a certain characteristic sample in class C (Posterior) is the opportunity for the emergence of class C (before the entry of the sample, often called prior), multiplied by the probability of the appearance of the characteristics of the sample in class C (also called likelihood), divided with the probability of the emergence of sample characteristics globally (also called evidence).

Therefore, the formula above can also be written simply in equation (3):

$$Posterior = \frac{Priorxlikelihood}{evidence} \qquad (3)$$

Evidence value is always fixed for each class in one sample. The value of the posterior is chosen to be compared with the other class's posterior values to determine to what class a sample will be classified. Further elaboration of the Bayes formula is done by describing (C | X1, ..., Xn) using the multiplication rules as follows:

$$P(C|X1,\dots,Xn) = P(C)P(X1,\dots,Xn|C)$$
$$P(C)P(X1|C)P(X2,\dots,Xn|C,X1)$$
$$P(C)P(X1|C)P(X2|C,F1)P(X3,\dots,Xn|C,X1,X2)$$
$$P(C)P(X1|C)P(X2|C,X1)P(X3,\dots,Xn|C,X1,X2,X3)$$
$$P(C)P(X1|C)\dots P(Xn|C,X1,X3,\dots,Xn-1)$$

It can be seen that the results of the elaboration cause more and more complex factors that affect the probability value, which is almost impossible to analyze one by one. As a result, the calculation becomes difficult to do. This is where a very high (naïf) independence assumption is used, that each criterion (X1, X2, ..., Xn) is independent (independent) of each other. With this assumption, equation (4) applies:

$$P(Pi|Xj) = \frac{P(Xi \cap Xj)}{P(Xj)} = \frac{P(Xi)P(Xj)}{P(Xj)}$$
$$= P(Xi)$$

For i / = j, so

$$P(Xi|C,Xj) = P(Xi \vee C) \qquad (4)$$

From equation (4) it can be concluded that the naïf independence assumption makes the opportunity condition simple, so that it is possible to do a calculation.

Furthermore, the translation of P (C | X1, ..., Xn) can be simplified into equation (5):

$$P(X2|C)P(X3|C)\dots$$
$$P(C|X1,\dots,Xn) = P(X1|C)$$
$$= \prod_{i=1}^{n} P(Xi|C) \qquad (5)$$

Information :

$$\prod_{i=1}^{\tilde{n}} P(Xi|C) = \text{Perkalian rating antar atribut}$$

Equation (5) is a model of the Naïve Bayes theorem which will then be used in the classification process. For classification with continuous or numeric data, the Gauss Density formula is used:

$$P(Xi = xi|C = cj)\frac{1}{\sqrt{2\pi\sigma ij}}e^{\frac{-(fi-\mu ij)^2}{2\sigma 2ij}} \qquad (6)$$

Infomation :

P : Opportunity

Xi : Attribute to-i

Xi : The i attribute value

C : Class sought
Cj : Sub class C sought
μ : Mean, states the average of all attributes
s : Standard Deviation, states variants of all attributes p: 3.141592654 e: 2.718281828

Naïve Bayes is a method that uses a probability approach to produce classifications. This method combines probability terms with probability categories to determine the probability of a successful category. From the explanation of the naïve Bayes algorithm, it can be concluded that the steps of Naïve Bayes work are:

a. Determine the training data and test data that you want to classify.
b. Calculate P (Ci) which is the prior probability for each sub class C that will be generated using the equation:

$$P(Ci) = \frac{Si}{s} \qquad (7)$$

c. Where Si is the number of training data from the Ci category, and s is the total number of training data.
d. Calculating P (Xi | Ci) which is the posterior probability of Xi with the condition C using equation (5).

If xi is a numeric data, then to calculate P (Xi | Ci) uses the gaussian distribution contained in equation (6). Maximizing P (Xi | Ci). P (Ci) to get class C that wants to be classified by multiplying P (X | Ci) and P (Ci) for all possible classifications:

$$P(Ci) \prod_{i=1}^{n} P(Xi|Ci) \qquad \dots(8)$$

In other words, the results assigned to the Ci class are those that have the maximum $P(X_i \mid C_i) P(C_i)$.

3. RESEARCH DESIGN

Descriptive analysis is the method used in this study, which aims to get a more in-depth and complete picture of the object to be studied by using observations directly in the field, conducting interviews with related parties in the research, documentation and study literature. Naive Bayes is a method used for designing or modeling.

3.1. Methods of Selecting Populations and Samples

Population is a generalization area that consists of objects / subjects that have certain qualities and characteristics set by researchers to be studied and then drawn conclusions (Sugiyono, 2013) [6]. The sample is the part or number and characteristics possessed by the population. The researcher will take a sample of the population due to limited funds, energy and time, and if the population is large, and researchers are not likely to learn everything in the population. The conclusion of what has been learned from the sample will be applied to the population. For that sample taken from the population must be truly representative (Sugiyono, 2013). Student data available in Tangerang Pakarti Luhur AMIK were sampled in this study. There were 46 student data obtained by using purposive sampling technique.

Methods of Collecting Data

In research, data collection techniques are the most strategic step, because the main purpose of the research is to obtain data from Sugiyono, 2013).

Data and information obtained in this study are derived from the data and information obtained and in accordance with the data needed and collected using the following method: Primary Data Collection This research was conducted to obtain primary data by observation and interview. The reasons underlying the use of this method in the research process are as follows, observation is an analytical technique carried out by conducting observations in a systematic manner and systematically recording, interviews are meetings of two people to exchange information and ideas through question and answer, so that meaning can be constructed in a particular topic, and the document is a record of past events. Documents can be in the form of writing, pictures, or monumental works from a person. Documents in the form of writing such as diaries, life histories, stories, biographies, regulations, policies. Documents in the form of pictures such as photos, live images, sketches and others. Documents in the form of works such as artwork, which can be in the form of pictures, sculptures, films and others. Document study is complementary to the use of observation and interview methods in qualitative research. Secondary Data Collection Secondary data is obtained by observing data, reading studying and quoting from literature books, journals, internet and other sources that are closely related to research.

3.2. Instrumentation

The research instruments used by the author in order to collect the variables being studied are: To obtain valid data in this study, it is done by conducting interviews and collecting data in accordance with the research needs of the competent parties. Documents in the form of a selection report of PPA scholarship recipients on Pakarti Luhur AMIK until 2017.

The software used to analyze is Microsoft Windows 7 Ultimate 32 bit, the method in this study is the Naive Bayes method for the formation of training data, models, and testing using the data obtained. The hardware used to run the software is a laptop with Intel CoreDuo processor specifications, 2 GB RAM.

3.3. Data Analysis and Testing Techniques

3.3.1. Data Analysis Techniques

The analysis technique will be carried out using the Naive Bayes method. The method will be used by the author to process the selection data of PPA scholarship recipient students from students who have been obtained, in order to produce accurate predictions in the analysis of the selection of PPA scholarship recipient students with the requirements variables that have been determined by AMIK Pakarti Luhur which can be seen in table III- 1.

Table III-1 Research Variables Selection Requirements for PPA Scholarship Recipients.

NO	Observed Variables	Requirements
1	Indeks Prestasi Kumulatif (IPK)	>=3,50
2	Semester	>=2
3	Prestasi Mahasiswa	Aktif/-

(Announcement of Requirements for Scholarship Recipients, 2017) [7].

The data obtained will be divided into two sets, namely as training data (as well as test data) and evaluation data.

3.3.2. Classification Proccess on Prototype

The process to be designed in the system prototype includes:

a. Data input

Data input is done to enter data that will be predicted into the prototype that will be designed then the data is processed into a file. The format of the input process data is in the form of .dat.

b. Preprocessing

The contents of the input data as a whole are then converted into training data by the prototype to adjust the data in processing.

c. Conversion Proccess

Data can take the form of data based on the required attributes in accordance with the requirements. Prediction process After the data is clean, predictions will be made on the test data using the method with the best accuracy values that have been through the comparative stages in the analysis process in the study. The prediction results in the form of predictive accuracy in determining the student candidates for PPA scholarship recipients.

3.4. Teknik Pengujian Data

In this study the authors conducted data testing using academic reporting data and for testing the authors used the Naive Bayes method. Research Steps In creating a Model Determination System for PPA Scholarship Recipients Students the whole process must go through several stages. The steps at the research stage can be seen in Figure III-1.

Figure III-1 Steps to Research Research Schedule

Table III - 4 Research Schedules

4. DISCUSSION RESEARCH RESULTS

Analysis of Problem Analysis Problems in this study are outlined in the making of a system model. The system model that will be created has the following limitations:

1) Model design is made using the Naive Bayes method.

Table IV-1 variable Selection of Student Candidates for PPA Scholarship Recipients

NO	Observed Variables	Requirements
1	Indeks Prestasi Kumulatif (IPK)	>=3,50
2	Semester	>=2
3	Prestasi Mahasiswa	Aktif/Tidak

Tabel IV-2 class Seleksi Mahasiswa Calon Penerima Beasiswa PPA

NO	Class
1	Accept
2	Reject

2) Converting data obtained into training data and testing data, where the attributes of the data obtained will be selected according to the needs of the model used for testing. Data Analysis Samples are a part of the number and characteristics possessed by the population according to (Sugiono, 2013).

4.1. Sample Data

Table IV-3 Sample data of Student Candidates for PPA Scholarship Recipients

No	NAMA_MHS	Kelamin	SMT	IPK	PENGHASILAN	PRESTASI	Status
1	Fierda Partiana	Perempuan	3	3.85	2500000	Aktif	Terima
2	Dewi Putri Agustina	Perempuan	5	3.83	2442000	Aktif	Terima
3	Narang Bandriana	Laki-Laki	5	3.81	1500000	Aktif	Terima
4	Iga Pamella Sari	Perempuan	5	3.75	2400000	Aktif	Tolak
5	Sigit	Laki-Laki	3	3.73	1500000	Aktif	Tolak
6	Jdchatun	Perempuan	3	3.57	1500000	Aktif	Tolak
7	Eko Sumira	Laki-Laki	5	3.55	4000000	Aktif	Tolak
8	Dini Oktorika	Perempuan	5	3.56	2455200	Aktif	Tolak
9	Siti Masitoh	Perempuan	3	3.52	2500000	Aktif	Tolak
10	Retno Wati Fitriani	Perempuan	5	3.35	2378000	Aktif	Tolak
11	Solikah	Perempuan	3	3.79	2200000	Aktif	Terima
12	Winda Oktavianti	Perempuan	3	3.71	1500000	Aktif	Terima
13	M. Umar Ismawan	Laki-Laki	3	3.86	2000000	Aktif	Terima
14	Nur Cahyasari	Perempuan	3	3.62	1500000	Aktif	Terima
15	Yasir Fahmi	Laki-Laki	3	3.63	2000000	Aktif	Terima
16	Neni Solfiana	Perempuan	3	3.51	3000000	Aktif	Terima
17	Mimi Rofiah	Perempuan	3	3.47	500000	Aktif	Terima
18	Sarip Hidayatullah	Laki-Laki	3	3.39	2674500	Aktif	Terima
19	Esti Siti Daerah	Perempuan	3	3.4	2500000	Aktif	Terima
20	Egar Triansyah	Laki-Laki	3	3.44	2250000	Aktif	Terima
21	Desi Natalia	Perempuan	3	3.35	750000	Aktif	Terima
22	Yunar Indriani	Perempuan	3	3.34	2500000	Aktif	Terima
23	Herdi Firdaus	Laki-Laki	3	3.19	1300000	Aktif	Terima
24	Mohammad Sekhudin	Laki-Laki	3	3.36	900000	Aktif	Terima
25	M. Arya Alfatihah	Laki-Laki	3	3.49	2000000	Aktif	Terima
26	Lukman Hakim	Laki-Laki	3	3.23	3500000	Aktif	Terima
27	Kris Pramono SMA	Laki-Laki	3	3.53	4500000	Aktif	Terima
28	Raza Alfanti Larasati	Perempuan	3	3.5	2000000	Aktif	Terima
29	Desi Nurfazriyah	Perempuan	3	3.22	1800000	Aktif	Terima
30	Neli Hartati	Perempuan	3	3.34	2100000	Aktif	Terima
31	Delfiyanti	Perempuan	3	3.34	2300000	Aktif	Terima
32	Rani	Perempuan	3	3.58	4000000	Aktif	Terima

Sample data is data obtained from the interview and documentation process, then converted into training data and data testing to be tested. Of the 46 data obtained 43 data were identified and 3 data that could not be identified, because it has a value of "-".

4.2. Data Training and Data Testing
4.2.1. Table IV-4 Data Training and Data Testing

Training data is data that is data obtained from verified data samples according to the data that will be used in the data testing process. Criteria and Probability The probability value of each criterion is obtained from the training data in table IV-Table IV-5.

First sample Data Testing :

No.	NAMA MHS	SMT	IPK	PRESTASI	Status
1.	Fierda Partiana	3	3.85	Aktif	?

The probability value of each criterion can be seen in the test with Weka tools in figure IV-1.

```
--- Classifier model (full training set) ---

Naive Bayes Classifier

                       Class
Attribute        Terima   Tolak
                 (0.77)   (0.23)
===============================
SMT
  mean              4       4
  std. dev.     0.3333  0.3333
  weight sum       36      10
  precision         2       2

IPK
  mean          3.5299  3.5612
  std. dev.     0.1878  0.1256
  weight sum       36      10
  precision     0.0203  0.0203

PRESTASI
  Aktif           37.0     8.0
  -                1.0     4.0
  [total]         38.0    12.0

Time taken to build model: 0 seconds
```

Figure IV-1 Probability Value of Each Criteria

From the probability value above, 46 data of student data will be tested and completed with the following solutions: Before determining the probability of each criterion, the separation between likelihood status of Accept and likelihood status is done first.
Likelihood Status

- Acceot Status :
  Accept Probability = (SMT>=2/Status_Terima)x(IPK>=3,50/Status_Terima)x(Prestasi Aktif/Status_Terima) = (46/36)x(27/36)x(43/36) = 1,144676
- Reject Status :
  Probabilitas Tolak = (SMT>=2/Status Tolak)x(IPK>=3,50/Status_Tolak)x(Prestasi Aktif/Status_Tolak) = (46/10)x(27/10)x(43/10) = 0 Probability Percentage Status: Accept = 1,4467593/(1,4467593+0) x 100% = 100%
  Reject = 0/(1,4467593+0) x 100% = 0%

From the results of probability calculations and percentages can be seen in table IV-6.

Table IV-6 Probability Percentage Status Status Results

Status	Probability	Presentage
Accept	1,14467593	100
Reject	0	0

From the above results obtained the results of the amount of data received as many as 36 data from 46 data, SMT data> = 2 as many as 46 of 46 data, GPA> = 3.50 as many as 27 data from 46 data, and Achievement = "Active" as much as 43 data from 46 data. Then the results of the above calculations, obtained: Accept = 1.14467593 = 100% Denied = 0 = 0% By comparing the results received with rejected, it can be concluded that the data gets the results of "Accept" status, which is equal to 1.14467593 or with a percentage value of 100%. Design and Implementation After the selection system model for prospective recipients of PPA scholarships is obtained, the next step is to design and implement a Student Selection System for PPA Scholarship Recipients.
The steps are as follows:

A. Use Case Design

Use Case prototype application of the Naive Bayes method in the selection process of prospective students who receive the Pakarti Luhur PPA AMIK scholarship can be seen in Figure IV-3.

Figure IV-3 Use the Case prototype of the application of the Naive Bayes method in the selection process for prospective students who receive the Pakarti Luhur PPA AMIK scholarship.

5. CONCLUSION

This research was conducted to determine the selection process for prospective PPA scholarship recipients at the Tangerang Pakarti Luhur AMIK. Selection is taken from input variables, consisting of: Semester, GPA, Achievement, Status. Test the accuracy of the system using Cross Validation. This research is very important and useful because it can be used to accelerate the selection of prospective candidates for PPA scholarships that were previously carried out subjectively from the results of the meeting.

Based on research on the selection of prospective candidates for PPA scholarships, some conclusions can be drawn as follows: Based on data from prospective students who obtained PPA scholarships obtained, the process of applying the Naive Bayes method in obtaining information from the results of classifications of prospective PPA scholarship students. The Naive Bayes method utilizes training data to generate the probability of each criterion for different classes, so that the probability values of these criteria can be optimized to predict students who are candidates for PPA scholarships based on the classification process carried out by the Naive Bayes method itself.

Based on students' academic data which is used as training data, the Naive Bayes method successfully classifies 43 student data from 46 data tested. So that the Naive Bayes method is successful in predicting students who are candidates for PPA scholarships with an accuracy percentage of 93.4783%.

BIBLIOGRAPHY

Undang-Undang Negara Republik Indonesia Nomor 12 Tahun 2012, Pasal 76 Ayat (2), Tentang Pendidikan Tinggi Direktorat Jendral Pembelajaran dan Kemahasiswaan, Kementrian Riset, Riset, dan Pendidikan Tinggi, "Pedoman Beasiswa Peningkatan Prestasi Akademik (PPA)". 2017

Dennis et al. 2012. Systems Analysis & Design With UML Version 2.0; An Object-Oriented Approach 4th Edition. John Wiley & Sons, Inc.Edition, McGraw-Hill Book, Co.

Nidhra, Srinivas dan Jagruthi Dondeti. 2012. Black box and White box Testing Techniques– A Literature Review. Internasional Journal of Embedded System and Applications (IJESA) Vol.2, No.2.

Diasrina Dahri, Fahrul Agus, Dyna Marisa Khairina. 2016, METODE NAIVE BAYES UNTUK PENENTUAN PENERIMA BEASISWA BIDIKMISI UNIVERSITAS MULAWARMAN, Samarinda:Jurnal Informatika Mulawarman Vol. 11 No. 2. Sugiyono. 2013, Metode Penelitian Kuantitatif, Kualitatif dan R&D. Bandung : Afabeta Pengumuman Persyaratan Mahasiswa Calon Penerima Beasiswa, 2017.