# Comparison of Naive Bayes Classifier and C4.5 in Predicting Student Study Period

Wildan Budiawan Zulfikar[1], Indra Falah[2], Yana Aditia Gerhana[3], Mohamad Irfan[4]
{wildan.b@uinsgd.ac.id[1], indrafalah@gmail.com[2], yanagerhana@uinsgd.ac.id[3],
irfan.bahaf@uinsgd.ac.id[4] }

Department of Informatics, UIN Sunan Gunung Djati Bandung, Indonesia[1,3,4]

**Abstract.** A department of an Islamic State University in Indonesia has an average graduation presentation on time of 13.5%. This is very worrying, it has an impact on various things. Therefore, the department will find it difficult to obtain optimal assessment values. The purpose of this study was to analyze the data of active students to gain knowledge of what caused and what factors influenced the student's study period. This study proposes a classification model that can be used to predict student study periods using the Naive Bayes Classifier and C4.5 algorithms. In the determinant analysis phase, this work discovers that there are several attributes of active students who influence on the student's study period such as gender, GPA, college entry scheme, tahfidz, etc. In the testing phase, this work conducted that Naive Bayes Classifier has a better accuracy rate than C4.5.

**Keywords:** Naive bayes classifier, C4.5, classification, prediction, student study period.

## 1 Introduction

Students are one of the assets owned by each university. Generally, a university has a variety of study programs. Every year, universities or study programs graduate students. In not much different times, universities recruit new students. The number of students who enroll and be accepted in a study program shows that the study programs owned by the university have popularity and produce graduates who are qualified in the view of the stakeholders. The increasing number of students is also directly proportional to the university's income.

In a case, a study program has a phenomenon where there is an imbalance between the number of students who graduate and students accepted. This is because there are several students who cannot complete the lecture process on time. While the university routinely conducts the acceptance process every year. Of course, this makes the number of active students in the study program very much. This will be a serious problem and has a negative impact, especially for study programs that have student advantages. The study program will find it difficult to achieve the ideal ratio of lecturers and students[1]. Therefore, the study program will obtain a value that is not optimal in the study program accreditation process.[2]

This study proposes a classification model that can be used to predict student study periods using the Naive Bayes Classifier and C4.5 algorithms. Based on previous research, both of them showed quite a high accuracy [3][4][5][6]. Previous research, machine learning that used several algorithms implemented to classify student performance [7][8][9]. The other previous research, namely the comparison test of classification algorithms, namely C4.5, Naive Bayes Classifier,

KNN and SVM on skin disease predictions showed that the algorithm tested had values of accuracy, precision, and recall above 94%. Whereas in the present study the comparative test uses the Naive Bayes Classifier and C4.5. C4.5 is a decision tree classification algorithm that is widely used because it has the main advantages of other algorithms [10][11]. The advantages of C4.5 algorithm can produce decision trees that are easily interpreted, have an acceptable level of accuracy, are efficient in handling attributes of the discrete type and can handle discrete and numeric types of the attribute [10].

Another study, Naive Bayes Classifier was used for the dress-up method where the results stated that naive bayes classifier had an accuracy of 84%. The next study is the classification of student graduation using the Naive Bayes algorithm. The results of the study stated that the accuracy of the Naive Bayes algorithm was quite high at 82% [12]. This algorithm working properly in several cases include image data, text mining, medical case, academic, sport, etc [13][14][15][16][17]. But based on the researchers the results are due to the lack of data complexity. Suggestions for further research are comparing with the C4.5 algorithm or other classification methods so that they can find out the advantages of each [18][19][20].

This work compares the Naive Bayes Classifier algorithm with C4.5 using student data at an Islamic State University. The criteria for students who are declared to graduate on time are students who are studying for 4 years referring to legislation. The results of this study are expected to be a source of information that can show student data which is predicted not to graduate on time. Furthermore, the university can conduct several treatments to prevent this.

## 2 Methodology

This work uses training data taken from one of the study programs at state Islamic universities in Indonesia. Technically, this work uses data originating from the system to be built in the form of student data which contains several attributes including student entry points, cumulative grade point grades, majors from high school, and information about tahfidz. The research data was obtained from graduation data and master data of study program students at a state Islamic university in Indonesia. Graduation data was taken from students of 2011 and 2012, which numbered 109.

### 2.1 Data Understanding

Student master data has many attributes. However, in this case, selected attributes that influence on the student's study period include: The *admission type* is used to examine the relationship between the student's study period and the path taken by students. *High school major* also used to determine the relationship between the study period of students with majors from school at the high school level. *Gender* was used to determine the ratio between influential male and female sexes in the study period. *Activities* are used as a comparison between students who are active in organizations in departments and faculties and who are not active in organizations.

Also, this work also uses student graduation data such as *GPA* and *tahfidz exam*. *GPA* is used to determine the relationship between the level of academic achievement of the student's study period. The graduation of the *tahfidz exam* is used to find out the relationship between the student's study period and the student's exact time or late taking the *tahfidz exam*. The student indicator is on time, that is, maximally taking the exam one year after the *tahfidz* certificate is

issued. Each attribute needs to be analyzed to make it easier when the data is processed. Table 1 describes the predicate of student graduation in general. This data is divided into 3 groups, namely *excellent*, *very good*, and *good*.

**Table 1.** Predikat Kelulusan

| No | GPU Range | Predicate |
|---|---|---|
| 1 | 3,51 – 4,00 | Excellent |
| 2 | 2,76 – 3,50 | Very Good |
| 3 | 2,00 – 2,75 | Good |

The next attribute is *tahfidz test*. Directly, this attribute can be categorized into two parts, right and late. A student is said to have completed the *tahfidz test* in a timely manner if it meets the conditions if it is completed in less than or equal to one year since the ratification of the *tahfidz decree*. A student is said to be late if he cannot meet the conditions described in the previous statement.

**Table 2** Tahfidz Test

| No | Description | Predicate |
|---|---|---|
| 1 | Students complete the Tahfidz test less or equal to 1 year after the decree is passed | On Time |
| 2 | Students complete the Tahfidz test more than 1 year after the decree is passed | Late |

The last attribute is extra activities. Based on the results of the analysis of the data, this attribute divides students into three categories namely *high*¸ *medium*, and *low*. *High* shows that students who have a high level of activity can be proven by a large number of organizations attended by the students concerned. This work is summarized as a student affiliated with two or more organizations. *Medium* shows that the student has a busy life that is not too high. Finally, *low* level for students who have a low level of activity.

**Table 3.** Extra Activities

| No | Description | Activity Level |
|---|---|---|
| 1 | Students are affiliated with 2 or more organizations | High |
| 2 | Students are affiliated with 1 organizations | Medium |
| 3 | Students are not affiliated with any organization | Low |

## 2.2 Data Preparation

In the data preparation phase, training data is stored in a table prepared specifically for the calculation process. The table includes attributes, total data, the amount of data that has been classified based on the specified target, in this case, that is graduating on time or not, and the

entropy and gain columns. The next stage is the application of the C4.5 algorithm to calculate the Entropy and Gain values for each attribute to be used as Tree shapes. A tree is a form of classification rules that will be applied to the testing process.

The testing process is a step taken to enter test data or predictive data. The attributes used in this process must be in accordance with the attributes in the training process. Each attribute is compared to the rules that have been formed in the calculation of previous training data. Furthermore, the data will be classified based on the target you want to know, that is the data of students with this attribute condition can be passed on time or late in detail described in Fig. 1.

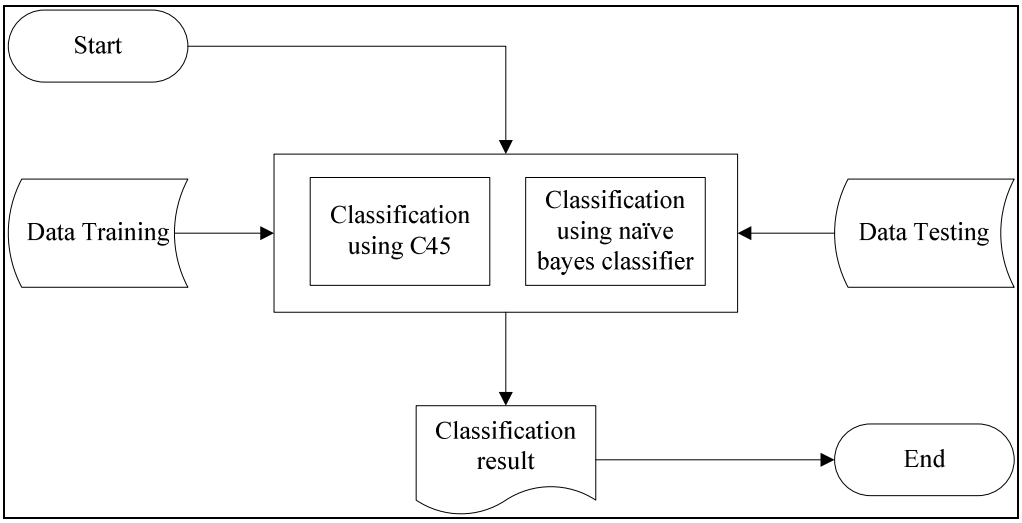


**Fig. 1.** General Classification Model

## 2.3 Modelling

### 2.3.1 Naive Bayes Classifier Modelling

In general, the procedure for calculating the Naive Bayes Classifier algorithm is divided into the following steps:

1. Calculates the number of classes from the graduation statement column based on the classification formed.
   a. C1 (class label="*on time"*) = the amount of "*on time*" on data training
      = 31/109 = 0.2844
   b. C2 (class label="*late*") = the amount of "*late*" on data training
      = 78/109 = 0.7155

**Table 4.** Data Testing

| Gender | GPU | Admission Type | Tahfidz | High School Major | Activities |
|---|---|---|---|---|---|
| M | Very Good | SBMPTN | On Time | IPA | Medium |

2. Second, this work calculates the same number of cases on each attribute of the graduation statement class based on testing data.
   a. P (gender = "*M*" | class label = "*on time*")
      = 14/31
      = 0.4516

   b. P (gender = "*M*" | class label = "*late*")
      = 57/78
      = 0.7307

   c. P (GPU = "*very good* " | class label = "*on time*")
      = 14/31
      = 0.4516

   d. P (GPU = "*very good*" | class label = "*late*")
      = 73/78
      = 0.9358

   e. P (admission type = "SBMPTN" | class label = "*on time*")
      = 12/31
      = 0.387

   f. P (admission type = "SBMPTN" | class label= "*late*"
      = 7/78
      = 0.089

   g. P (tahfidz = "*on time*" | class label = "*on time*")
      = 31/31
      = 1

   h. P (tahfidz = "*on time*" | class label = "*late*")
      = 15/78
      = 0.1923

   i. P (high school major = "IPA" | class label = "*on time*")
      = 27/31
      = 0.8709

   j. P (high school major= "IPA" | class label = "*late*")
      = 61/78
      = 0.782
   k. P (activities = "*medium*" | class label = "*on time*")
      = 5/31
      = 0.1612

   l. P (activities = "*medium*" | class label = "*late*")
      = 8/78

= 0.1025

3. The Accumulate of these values:

   a. All values with class label = "*on time*"
      P (X | class label = "*on time*")
      = 0.4516 x 0.4516 x 0.387 x 1 x 0.8709 x 0.1612
      = 0.0110

   b. All values with class label = "*late*"
      P (X | class label = "*late*")
      = 0.7307 x 0.9358 x 0.089 x 0.1923 x 0.782 x 0.1025
      = 0.0009

   c. Prior probability with class label = "*on time*"
      P (Ci) | class label = "*on time*" x P (X | class label ="*on time*")
      = 0.0110 x 0.2844
      = 0.0031

   d. Prior probability with class label = "*late*"
      P (Ci) | class label = "*late*" x P (X | class label ="*late*")
      = 0.0009 x 0.7155
      = 0.00067

   e. Calculate each atribute value of data testing toward data training
      P(X) = P (gender = "M") x P(GPU = "*very good*") x P(Admission type = "SBMPTN")
      x P(high school major = "IPA") x P(activities = "*medium*")
      = (71/109) x (63/109) x (19/109) x (46/109) x (88/109) x (13/109)
      =0.0036

### 2.3.1 C4.5 Modelling

The second algorithm is C4.5, table 5 describes 10 of the 109 training data used to determine the study period of students based on three determinants, namely GPU, *admission type*, and *student activities*. GPU is divided into three categories, namely *excellent*, *very good*, and *good*. While the entry path is divided into four categories, namely SBMPTN, SNMPTN, PPA, and *other*. Furthermore, student activities are divided into 3, namely *high*, *medium* and *low*.

**Table 5.** Sample of Data Training

| GPU | Admission Type | Activities | Label |
|---|---|---|---|
| Excellent | SBMPTN | High | On time |
| Excellent | Other | Medium | On time |
| Excellent | SBMPTN | High | On time |
| Very good | Other | Low | Late |

| | | | |
|---|---|---|---|
| Very good | Other | Low | Late |
| Excellent | Other | Low | Late |
| Very good | Other | Low | Late |
| Very good | Other | Medium | Late |
| Very good | SNMPTN | Low | Late |
| Excellent | PPA | Low | Late |

The first step is to calculate entropy using (1):

$$Entropy(S) = \sum_{i-1}^{n} - pi * log_2 pi \quad (1)$$

The results of entropy calculation are presented in table 6 where for the category *on time* with frequency 3 obtains a value of 0.521089678. For the label *late* has a frequency of 7 students and the entropy value is 0.360201221.

**Tabel 6.** Entropy Result

| Class Label | Frequency | *Pi * log₂ pi* |
|---|---|---|
| **On Time** | 3 | 0.521089678 |
| **Late** | 7 | 0.360201221 |
| | Entropy(S) | **0.881290899** |

The next step is to calculate the root value. Table 7 describes the gain calculations for the GPU, *admission type*, and *activities*. In Table 7 it can be seen that the organization has the highest gain with a value of 0.681291.

**Table 7** Node 1

| *Node* | | | Freq | On Time | Late | Entropy | Gain |
|---|---|---|---|---|---|---|---|
| **1** | GPU | Excellent | 5 | 3 | 0 | 0.970951 | **0.395816** |
| | | Very Good | 5 | 0 | 5 | 0 | |
| | | Good | 0 | 0 | 0 | 0 | |
| | Admission Type | SBMPTN | 2 | 2 | 0 | 0 | **0.491277** |
| | | Other | 6 | 1 | 5 | 0.650022 | |
| | | SNMPTN | 1 | 0 | 1 | 0 | |
| | | PPA | 1 | 0 | 1 | 0 | |
| | Activities | High | 2 | 2 | 0 | 0 | **0.681291** |
| | | Medium | 2 | 1 | 1 | 1 | |
| | | Low | 6 | 0 | 6 | 0 | |

Based on the provisions of the C4.5 algorithm, each attribute that has the highest gain then that attribute becomes root or gain. Based on the results of calculations, the *activities* have the highest gain value. Therefore, the *activities* become the root value. The *high* activities are the right leaf pass because the amount of data and the amount of data that passes right is the same.

While the *low activities* is a late leaf pass because there is a lot of data and the number of *low* activities data. Then, we can draw the node one decision tree as follows:

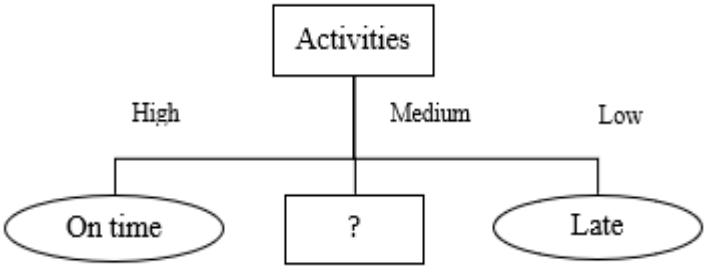


**Fig 2.** The Result of Node 1

The *medium* activity has a value, it must be recalculated to determine the branch node of the *medium* activity. Table 8 describes the calculations for branches for medium activity.

**Table 8** Node 2

| ***Node*** | | | **Freq** | **On Time** | **Late** | **Entropy** | **Gain** |
|---|---|---|---|---|---|---|---|
| **2** | GPU | Excellent | 1 | 1 | 0 | 0 | **0.881291** |
| | | Very Good | 1 | 0 | 1 | 0 | |
| | | Good | 0 | 0 | 0 | 0 | |
| | Admission Type | SBMPTN | 0 | 0 | 0 | 0 | **0.681291** |
| | | Other | 2 | 1 | 1 | 1 | |
| | | SNMPTN | 0 | 0 | 0 | 0 | |
| | | PPA | 0 | 0 | 0 | 0 | |

From the table above, there is a gain for each attribute. Where the gain results from each of these attributes include:

1. GPU = 0.881291
2. Admission type = 0.681291

Based on the table above, GPU is the attribute that has the highest gain. Therefore, GPUs are nodes of activities. *Excellent* GPU class is a leaf passed for *on time* while the *very good* GPU is a leaf passed for *late*.

In addition, it can be seen for the *excellent* and *very good* GPU attribute values that have the amount of data. However, for the *good* GPU attribute value does not have data which means that in the training data there is no data that shows *medium* activities and *good* GPU. So that the *good* IPK branch is not in the decision tree.

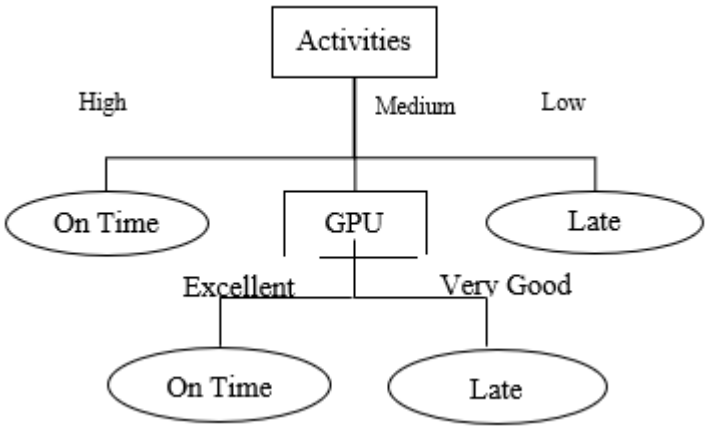


**Fig 3.** The Result of Node 2

The picture above is the final result of the tree formation process. The rules of the tree above are as follows:

1. If activities = high then class = on time.
2. If activiteis = low then class = late.
3. If activities = medium dan GPU = excellent then class = on time
4. If activities = medium and dan GPU = very good then class = late.

## 3 Results and Discussions

These models testing another 100 student data. The results of testing the algorithm can be seen in the table. The following is the formula for calculating accuracy obtained by comparing the amount of data that corresponds to the one that is not appropriate:

$$Accuration = \frac{\text{Correct Result}}{Data\ Training}\ x\ 100\% \tag{2}$$

(2) Implemented on the calculation of results on the system then compared with the data contained in the testing data. The test results produce accuracy values in the form of percentages as explained in (3) and (4).

$$NBC\ accuration = \frac{88}{100}\ x\ 100\% = 88.00\% \tag{3}$$

$$KNN\ accuration = \frac{87}{100}\ x\ 100\% = 87.00 \tag{4}$$

Based on the calculation of the accuracy that has been obtained through testing 109 data, it is obtained the results of 88% accuracy for the Naive Bayes Classifier algorithm, and 87% for the C4.5 algorithm. In addition, this study implements the proposed model in the form of a web-based application as described in fig. 4.

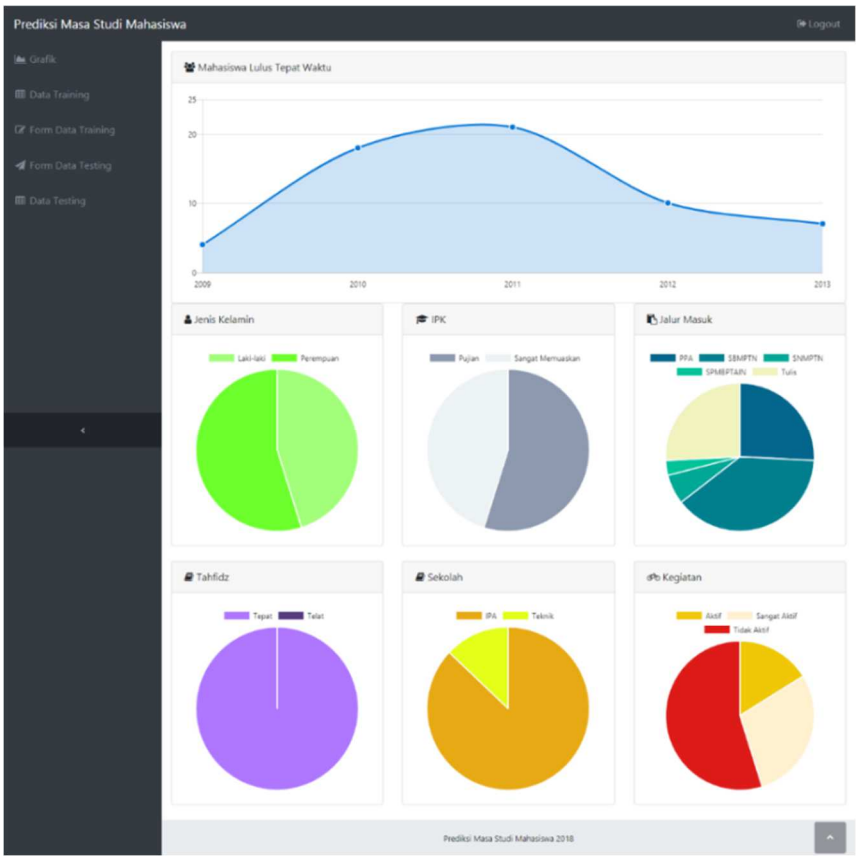


**Fig 4.** Web-based Statistic of Data

## 4 Conclusions

Based on the results of the experiment, the Naive Bayes Classifier algorithm has a better level of accuracy compared to the C4.5 algorithm, which is 88% and 87%. However, when viewed from the speed, the C4.5 algorithm has a better speed than the Naive Bayes Classifier algorithm. The results of the two classification models are largely determined by training data. In this work, the two algorithms have almost the same performance in classifying student study periods. Further work, it is necessary to add and review more detailed attributes that are allegedly influencing the student's study period such as personal and family economic conditions, analyzing relationships between students and lecturers, personality, psychology, and student activities on social media.

## References


[1] A. Satapathy and J. L. L. M, "An Intelligent Framework Prototype for Monitoring Students in Virtual Classroom," Indones. J. Electr. Eng. Comput. Sci., vol. 12, no. 3, p. 1151, Dec. 2018.

[2] A. M. Abdullahi, M. Makhtar, and S. Safie, "The patterns of accessing learning management system among students," Indones. J. Electr. Eng. Comput. Sci., vol. 13, no. 1, p. 15, Jan. 2019.

[3] W. N. L. W. H. Ibeni, M. Z. M. Salikon, A. Mustapha, S. A. Daud, and M. N. M. Salleh, "Comparative analysis on bayesian classification for breast cancer problem," Bull. Electr. Eng. Informatics, vol. 8, no. 4, Dec. 2019.

[4] S. A. Diwan Alalwan, "Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 1, p. 88, Apr. 2019.

[5] A. Adeleke, N. A. Samsudin, Z. A. Othman, and S. K. A. Khalid, "A two-step feature selection method for quranic text classification," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 2, pp. 728–734, Nov. 2019.

[6] B. M. Susanto, "Naïve Bayes Decision Tree Hybrid Approach for Intrusion Detection System," Bull. Electr. Eng. Informatics, vol. 2, no. 3, pp. 225–232, Sep. 2013.

[7] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, Dec. 2019.

[8] F. Jauhari and A. A. Supianto, "Building student's performance decision tree classifier using boosting algorithm," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 3, p. 1298, Jun. 2019.

[9] N. H. M. Ariffin and S. N. H. Askol, "Academician perceptions towards online student evaluation," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 2, pp. 952–958, Nov. 2019.

[10] M. M. Saad, N. Jamil, and R. Hamzah, "Evaluation of Support Vector Machine and Decision Tree for Emotion Recognition of Malay Folklores," Bull. Electr. Eng. Informatics, vol. 7, no. 3, pp. 479–486, Sep. 2018.

[11] Z. Saringat, A. Mustapha, R. R. Saedudin, and N. A. Samsudin, "Comparative Analysis of Classification Algorithms for Chronic Kidney Disease Diagnosis," Bull. Electr. Eng. Informatics, vol. 8, no. 4, Dec. 2019.

[12] Y. S. Nugroho, "DATA MINING MENGGUNAKAN ALGORITMA NAÏVE BAYES UNTUK KLASIFIKASI KELULUSAN MAHASISWA UNIVERSITAS," pp. 1–11, 2009.

[13] N. M. Samsudin, C. F. binti Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, "Youtube spam detection framework using naïve bayes and logistic regression," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 3, p. 1508, Jun. 2019.

[14] A. D. Poernomo and S. Suharjito, "Indonesian online travel agent sentiment analysis using machine learning methods," Indones. J. Electr. Eng. Comput. Sci., vol. 14, no. 1, p. 113, Apr. 2019.

[15] O. Somantri and D. Apriliani, "Opinion Mining on Culinary Food Customer Satisfaction Using Naïve Bayes Based-on Hybrid Feature Selection," Indones. J. Electr. Eng. Comput. Sci., vol. 15, no. 1, p. 468, Jul. 2019.

[16] F. H. K. Zaman, J. Johari, and A. I. M. Yassin, "Learning face similarities for face verification using hybrid convolutional neural networks," Indones. J. Electr. Eng. Comput. Sci., vol. 16, no. 3, Dec. 2019.

[17] H. A. Nugroho, I. M. D. Maysanjaya, N. A. Setiawan, E. E. H. Murhandarwati, and W. K. . Oktoeberza, "Feature analysis for stage identification of Plasmodium vivax based on digital microscopic image," Indones. J. Electr. Eng. Comput. Sci., vol. 13, no. 2, p. 721, Feb. 2019.

[18] M. Ghavamzadeh, S. Mannor, J. Pineau, and A. Tamar, "Convex Optimization: Algorithms and Complexity," Found. Trends® Mach. Learn., vol. 8, no. 5–6, pp. 359–483, Nov. 2015.

[19] E. Angelino, M. J. Johnson, and R. P. Adams, "Patterns of Scalable Bayesian Inference," Found. Trends® Mach. Learn., vol. 9, no. 2–3, pp. 119–247, Nov. 2016.

[20] P. H. S. Torr, "Bayesian Model Estimation and Selection for Epipolar Geometry and Generic Manifold Fitting," Int. J. Comput. Vis., vol. 50, no. 1, pp. 35–61, 2002.