Comparison Of K-N Earest Neighbor And Naïve Bayes Algorithms For *Prediction Of* Aptikom Membership Activity Extension In 2023

Fathia Alisha Fauzia[1)] **, Kannisa Adjani** [2)] **, Christina Juliane** [3)]
[1,2,3)]STMIK LIKMI, Bandung, Indonesia
[1)]fathiaalisha@uniga.ac.id, [2)]kannisaadjani68@gmail.com, [3)]christina.juliane@likmi.ac.id

Submitted : Dec 30, 2022 | **Accepted** : Dec 30, 2022 | **Published** : Apr 1, 2023

Abstract: So far APTIKOM as the Informatics and Computer Higher Education Association has provided many opportunities for registered members to participate in discussions on the development of science among fellow association members, access to various professional experts, as well as technical and non-technical guidelines in the field of education. With the various opportunities above, it is hoped that all members will support the activities of each member who has joined or has just joined so that a good association can be created. This study aims to find out about the problems that occur in APTIKOM, namely members who have registered as members but rarely renew their membership which results in data accumulation in APTIKOM. This research method uses the k-nn and naïve Bayes algorithms by using data sets from 2012 to 2022. The dataset used is APTIKOM member data and has 5 attributes namely name, gender, last education, institution and validation secret. To calculate the research test using a rapid miner. The purpose of this study is to predict whether in the following year there will be a membership renewal process for all APTIKOM members who have been recorded from 2012 to 2022. Furthermore, the results of this study have a different level of accuracy. Where for k-nn the resulting accuracy is 94.00% and for the result of naïve Bayes is 91.35%.

Keywords: *APTIKOM , K-NN, Naïve Bayes, Algorithms, Predictions.*

INTRODUCTION

Information technology profession in Indonesia cannot be separated from the important role of the information technology associations that have existed in Indonesia to this day. Association itself in KBBI is explained as an association of people who have the same interest [1]. In the world of computer science higher education in particular, campuses are required to apply information technology in various management and lecture processes in order to produce graduates who are able to compete with the world of work [2], so that universities often need new information and knowledge from professors, consultants and IT experts. Universities sometimes experience difficulties in choosing an IT curriculum or attending scientific meetings in the framework of the required information technology development. Than Information and Computer Higher Education Association (APTIKOM) emerged.

APTIKOM is one of the associations that houses universities throughout Indonesia that have informatics and computer study programs. A PTIKOM has an important role for universities to always get information and technical assistance for competence development in the IT field [3]. A PTIKOM has a commitment that actually efforts to educate the nation's life as mandated in the Preamble to the 1945 Constitution, are the duties and responsibilities of the family, community and government in a

*name of corresponding author

national education system. So that there are many applicants from every tertiary institution in Indonesia who register at APTIKOM for the needs of the individual or the college itself .

However, with so many registrants from every tertiary institution in Indonesia who have become members of APTIKOM itself, they often do not re-register to renew their membership, which causes many members to be inactive and do not renew every year. One of the efforts made by the APTIKOM management is to increase the activity of its members by imposing sanctions if they do not renew each year by being considered resigned and providing *benefits* to active members each time they renew their membership at APTIKOM. With that, we will try to compare the two algorithms to find out which algorithm is the most suitable for calculations in this study. Several researchers have conducted research related to naïve Bayes on data with a large number of attributes. Hasan proposed optimizing attribute selection on naive bayes using *forward selection* to improve the accuracy of naïve bayes to predict the smoothness of credit payments [4]. The results showed that the accuracy value of naïve Bayes based on *forward selection* reached an accuracy value of 71.97%. Meanwhile, Rinawati proposed optimizing attribute weights for naïve Bayes using *particle swarm optimization* to predict accurate ratings. The results show a higher accuracy value of 75.90% and an AUC value of 0.773. This resulted in an increase in accuracy of 3.5% and an increase in AUC of 0.008 [5].

From this research we need an appropriate and effective method in determining active members in APTIKOM by using data mining techniques, namely the K-NN and Naive Bayes methods to find out which algorithm is more accurate in predicting the success rate of the two methods because these methods can produce good accuracy in predictions [6].

So the purpose of this study is to find out whether there are active and inactive members in renewing each year. By predicting whether in 2023 each of these members will extend their membership in APTIKOM using the K-NN algorithm and the Naive Bayes algorithm as a comparison of the two methods to determine the accuracy of each method, tests are carried out to reduce errors and ensure the resulting output is in accordance with which are desired. One way is to apply the *Confusion matrix* as a classification model [7]. *The confusion matrix* is used to obtain *precision* , *recall* and *accuracy values. Confusion matrix* values are usually shown in units of percent (%)[8]. The test was carried out based on the criteria determined by the researcher. The results of the application of the K-NN and Naive Bayes algorithms are expected to become a reference value with selection results that match the specified criteria.

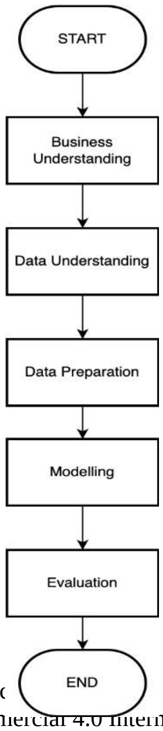
LITERATURE REVIEW

Research Stages

In this study using data mining. Data mining itself is used to process data or big data to generate useful information for associations or organizations.

Stages of Data Mining Process

In [9] states that the data mining process described as in Figure 1.



*name of corresponding author

Figure 1. Research Stage

a. Business/Organization Understanding

In this stage, understanding what the researcher will design is needed to answer questions and solve problems. And in this study the researcher made predictions about the membership data that will be extended in 2023 so that it can be seen the level of activity of members in the APTIKOM association,

b. Data Understanding

After understanding what will be designed, the next step is to understand the data that will be used. This stage is needed to check inaccurate data. Researchers checked APTIKOM member data by grouping according to the variables to be examined,

c. Data Preparation

Furthermore, this stage uses a combination of two or more data sets together, reducing the data sets only for the variables that are of interest in the research to be carried out. After the member data has been grouped, the next researcher prepares the data set to be applied to the K-NN and Naïve Bayes algorithm methodologies,

d. modelling

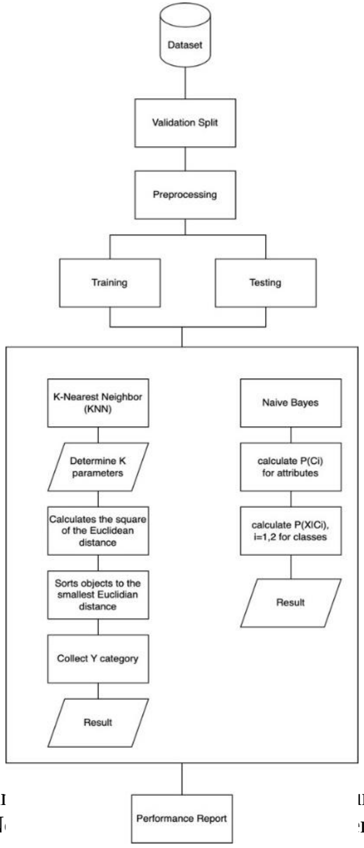
Modeling is done to find, identify and display any patterns or messages in research. So the researchers used 2 models to see differences in the level of accuracy, K-Nearest Neighbor modeling and Naive Bayes Algorithm used for this research.

e. Evaluation

The evaluation phase is carried out using several techniques to determine the usefulness of the model used previously. There is a confusion matrix value to determine the performance of the algorithm and ROC Curve rules are also used, called the Area Under the ROC Curve (AUC) to assess or measure the performance of data mining models. Algorithm performance is said to be good if the curve is close to point 0.1 and is said to be bad if the resulting curve is close to the transverse line from point 0.0 or the baseline line. The AUC value that is closer to 1 means that the model prediction is getting better [10].

METHOD

The research method used in this study is shown in Figure 2.


Dataset
Validation Split
Preprocessing
Training
Testing
K-Nearest Neighbor (KNN)
Determine K parameters
Calculates the square of the Euclidean distance
Sorts objects to the smallest Euclidian distance
Collect Y category
Result
Naive Bayes
calculate P(Ci) for attributes
calculate P(X|Ci), i=1,2 for classes
Result
Performance Report

*name of corresponding author

Figure 2. Research Method

K-Nearest Neighbor (KNN)

In the journal [11] K-NN is carried out by searching for cases by calculating the closeness between old cases and new cases based on weight matching. KNN is an example of a learning-based algorithm, where the training data set (*training*) is stored, so that the classification for new *records* that are not classified is obtained by comparing the *records* that are most similar to the *training set* .

KNN steps as follows [12]:

1. Determining the K parameter (number of closest neighbours), the K parameter in testing is determined based on the optimum K value during *training* ,
2. Calculating the square of the *Euclidean distance* of each object against the given sample data,
3. Sort these objects into groups that have the smallest *Euclidian distance,*
4. Gathering category Y (KNN),
5. By using the majority category, you will get results.

In general, to define the distance between two objects *x* and *y, the Euclidean* distance formula is used in the following equation [13]:

$$D(x,y) = \sqrt{}(x_1 - x_2)^2 + (y_1 - y_2)^2 \qquad (1)$$

Information:
X : training *data*
Y: data testing
I : *the Ith record* (row) *of* the table
n : amount of *training data*

Naive Bayes

Naive Bayes is a statistical classification, this clarification is done to predict the probability of class membership such as the probability of a tuple belonging to a particular class. Naive Bayes assumes that the effect of certain class attribute values is independent of the values of other attributes [14].

The Naive Bayes formula is as follows [15]:

$$P(Y|X)=P(Y)\ II\ P(X|Y) \qquad (2)$$

Explanation:
P(X|Y) : probability data with vector X in class Y
P(Y) : initial probability of class Y and P(XI
Y) is class Y independent probability of all features in vector X

RESULTS

Data Implementation

The data obtained from APTIKOM from 2012 to 2022 has 5 attributes. The attributes is member names, titles , institutional names, gender and secretarial validations. This data is still in the form of raw data, as in table 1 and represents the 7000 members who are in APTIKOM and only 1 member is taken each year which will be processed

at a later stage. The raw data is then done *with data preparation* to be used as a calculation variable, namely the name of the institution, gender and validation of secretions as stated in Table 1.

Table 1. Raw Data Samples

Name	Gender	Title	Institution Name	Validation secret
Member 1	Man	S3	ITG	2012
Member 2	Man	S2	STMIK AMIK	2013
Member 3	Woman	S3	Univ. Telkom	2014
Member 4	Woman	S2	Univ. Riau	2015
Member 5	Woman	S2	Univ. Riau	2016
Member 6	Man	S2	STMIK ASIA	2017
Member 7	Man	S2	UPI	2018
Member 8	Woman	S3	UNIBBA	2019
Member 9	Man	S3	UNJ	2020
…	…	…	…	…
Member 5086	Woman	S3	Univ. Subang	2022

Based on table 1, there are several attributes that are not used in the prediction process so that attribute selection is carried out or commonly called feature selection. The results of the attribute selection are described in Table 2.

Table 2. Sample *Data Preparation*

Institution Name	**Gender**	**Validation secret**
ITG	Woman	2012
STMIK AMIK	Man	2013
Univ. Telkom	Man	2014
Univ. Riau	Man	2015
Univ. Riau	Woman	2016
STMIK ASIA	Woman	2017
UPI	Woman	2018
UNIBBA	Man	2019
UNJ	Man	2020
UIN BDG	Woman	2021
Univ. Subang	Man	2022

Preprocessing

The preprocessing stage is an important stage in the prediction process. This stage is the process of preparing data that was still in the form of raw data into data that is ready for use or prediction. In this study, a cleaning process was carried out, namely the process of removing duplicate data because there were data on duplicate attributes.

Data Mining Process

From the data of members who have re-registered every year starting from 2012 to 2022 it can be used as the main capital to find out the pattern of member activity in APTIKOM. The data taken is the data that performs the initial registration and re-registration. The data received from APTIKOM is still in the form of raw data so there is a need for an understanding of the data. Not all attributes from APTIKOM membership are used, only using 2 attributes from the raw data, as shown in table 1.

Furthermore, the results of the model testing carried out are comparing which algorithm is more accurate and increases accuracy by using K-NN and *Naïve Bayes* in the Rapid Miner *framework* with the following model designs in Figure 3.

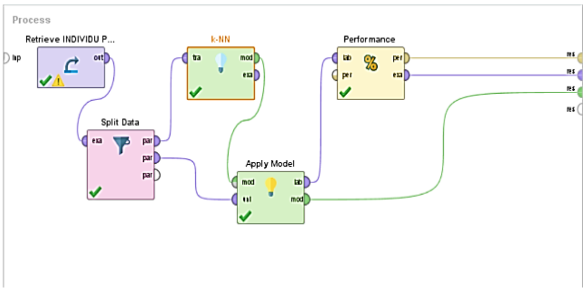


Figure 3. KNN *Validation* Testing Scheme

Figure 3 is a testing model of the K-NN algorithm using Rapid Miner, starting with entering data then processing *setroll* which will determine the label. After that, *filter* incomplete data or attributes that have *missing values* to optimize data processing, then proceed to the next stage, selecting the K-NN calculation model.

Figure 3. xplained that the process in the K-NN algorithm uses clean data that has been *preprocessed* . Then the results are processed back into a validation process consisting of *training* data and *testing data* . Furthermore, the results that have been calculated will produce an accuracy value.

Besides using the K-NN model, the Naïve Bayes model is also implemented. The following is the Naïve Bayes ALgorithm Testing Scheme in Figure 4.

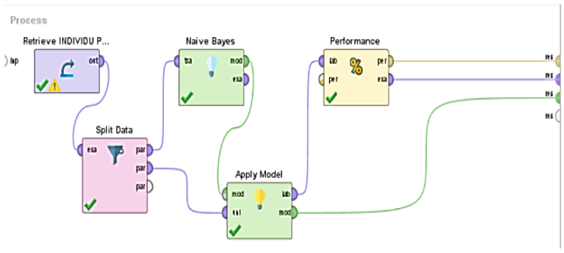


Figure 4. *Naïve Bayes Validation* Testing Scheme

Figure 4. is a testing model of the *Naïve Bayes algorithm* using Rapid Miner, starting with entering data and then processing *setroll* which will determine the label. After that, *filter* incomplete data or attributes that have *missing values* to optimize data processing, then proceed to the next stage, selecting the *Naïve Bayes calculation model* .

Figure 4. explained that the process in the *Naïve Bayes algorithm* uses clean data that has been *preprocessed* . Then the results are processed back into a validation process consisting of *training* data and *testing data* . Furthermore, the results that have been calculated will produce an accuracy value.

DISCUSSIONS

After going through the process of modeling or implementing the KNN algorithm, the resulting accuracy for the K-NN algorithm is as follows in Table 3.

Table 3.
KNN accuracy

*name of corresponding author

Accuracy: 94.00%	True NO	true YES	Class precision
Pred NO	888	34	96.31%
Pred YES	27	68	71.58%
Class recall	97.05%	66.67%	

Based on Table 3. explained that the accuracy in this study was 94.00% with the following details:

a. Prediction results Yes and it turns out Yes as many as 34,
b. Prediction results Yes and it turns out No as many as 888,
c. Prediction results No and it turns out yes as many as 68,
d. Prediction results No and it turns out not as many as 27.

This study also made comparisons with other methods, namely the Naïve Bayes algorithm. The prediction process of this algorithm is certainly different from the K-NN algorithm. The following is the result of the accuracy of the Naïve Bayes method as shown in Table 4.

Table 4.
***Naïve bayes* accuracy**

Accuracy: 91.35%	True NO	true YES	Class precision
Pred NO	864	37	96.89%
Pred YES	51	65	56.03%
Class recall	94.43%	68.73%	

Based on Table 4. explained that the accuracy in this study was 91.35% with the following details:

a. Prediction results YES and it turns out Yes as many as 37,
b. Prediction results Yes and it turns out No as many as 864,
c. The prediction results were no and it turned out to be Yes as many as 65,
d. The prediction results are not and it turns out that there are not as many as 51.

CONCLUSION

From the results of the research that has been done, the researcher can draw several conclusions and comparisons of the 2 algorithms used in the data set. The implementation of the data is done by cleaning the data so that you can see the variables that will be implemented in Rapid Miner. For the results of the K-NN algorithm, the accuracy value in the calculation is 94.00%. Whereas the Nive Bayes algorithm produces an accuracy value of 91.35%. From these conditions it can be concluded that the K-NN algorithm has better performance compared to the Naïve Bayes algorithm in the prediction numbers of members who will carry out the APTIKOM membership renewal process in 2023.

REFERENCE

U. Rahardja, N. Lutfiani, and M. S. Alpansuri, “Pemanfaatan Google Formulir Sebagai Sistem Pendaftaran Anggota Pada Website Asosiasi,” vol. 2, no. 4, 2018.

H. Budiman, “PERAN TEKNOLOGI INFORMASI DAN KOMUNIKASI DALAM PENDIDIKAN,” *Al-Tadzkiyyah: Jurnal Pendidikan Islam*, vol. 8, no. E-ISSN: 2528-2476, pp. 75–92, 2017.

A. R. Dayat and L. Angriani, “PEMANFAATAN MODEL-VIEW-CONTROLLER (MVC) DALAM RANCANG BANGUN SISTEM INFORMASI RAKORNAS APTIKOM 2017,” 2017.

M. Hasan, “Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis Forward Selection,” *ILKOM Jurnal Ilmiah*, vol. 9, no. 3, pp. 317–324, 2017, doi: 10.33096/ilkom.v9i3.163.317-324.

R. Rinawati, “Penentuan Penilaian Kredit Menggunakan Metode Naive Bayes Berbasis Particle Swarm Optimization,” *J-SAKTI (Jurnal Sains Komputer dan Informatika)*, vol. 1, no. 1, p. 48, 2017, doi: 10.30645/j-sakti.v1i1.28.

*name of corresponding author

D. Gustian, I. Suciati, S. Saepudin, P. Studi Sistem Informasi, U. Nusa Putra Sukabumi, and I. Jl Raya Cibolang Kaler No, “SISTEM PAKAR DENGAN ALGORITMA NAIVE BAYES UNTUK PREDIKSI HASIL PRODUKSI AYAM BROILER PLASMA (STUDI KASUS : PT.SEKAWAN SINAR SURYA).”

W. Jember, R. K. Midiarso, R. Umilasari, S. Pd, and M. Si, “PEMANFAATAN ALGORITMA NAIVE BAYES UNTUK KLASIFIKASI STATUS ALUMNI SMK BUSTANUL ULUM AL-GHAZALI,” 2022.

W. I. Rahayu, C. Prianto, and E. A. Novi, “PERBANDINGAN ALGORITMA K-MEANS DANNAÏVE BAYES UNTUK MEMPREDIKSI PRIORITASPEMBAYARAN TAGIHAN RUMAH SAKITBERDASARKAN TINGKAT KEPENTINGAN PADAPT. PERTAMINA (PERSERO,” *Jurnal Teknik Informatika*, vol. 1, no. 2, pp. 22–29, 2021.

M. Translated, “BAGIAN SATU : DASAR-DASAR DATA MINING,” pp. 1–11.

S. Narulita, A. Tigor Oktaga, and I. Susanti, “Media Aplikom Pengujian Model Prediksi Menggunakan Metode Data Mining Classification Decision Tree Untuk Penentuan Peminatan Peserta Didik,” *Media Aplikom*, vol. 13, no. 2, pp. 89–95, 2021, doi: 10.33488/1.ma.2.1.305.

A. Yandi Saputra and Y. Primadasa, “Penerapan Teknik Klasifikasi Untuk Prediksi Kelulusan Mahasiswa Menggunakan Algoritma K-Nearest Neighbour Implementation of Classification Method to Predict Student Graduation Using K-Nearest Neighbor Algorithm,” *Techno.Com*, vol. 17, no. 4, p. 9, 2018.

S. Wahyuningsih, D. R. Utari, U. B. Luhur, D. Tree, and K. Validation, “Perbandingan Metode K-Nearest Neighbor , Naïve Bayes dan Decision Tree untuk Prediksi Kelayakan Pemberian Kredit,” pp. 8–9, 2018.

R. Palupi, D. A. Yulianna, and S. S. Winarsih, “Analisa Perbandingan Rumus Haversine Dan Rumus Euclidean Berbasis Sistem Informasi Geografis Menggunakan Metode Independent Sample t-Test,” *JITU : Journal Informatic Technology And Communication*, vol. 5, no. 1, pp. 40–47, Jul. 2021, doi: 10.36596/jitu.v5i1.494.

N. R. Indraswari, P. S. Informatika, U. M. Surakarta, Y. I. Kurniawan, P. S. Informatika, and U. M. Surakarta, “Aplikasi prediksi usia kelahiran dengan metode naive bayes,” vol. 9, no. 1, pp. 129–138, 2018.

I. Kurniawan and A. Susanto, “Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019,” pp. 1–10, 2019, doi: 10.30864/eksplora.v9i1.237.

*name of corresponding author