DOI: 10.14421/ijid.2022.3804 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 11, No. 2, 2022, Pp. 232-241

Classifying High School Scholarship Recipients Using the K-Nearest Neighbor Algorithm

Ica Dwi Yulisa
Department of Information System
UIN Raden Fatah
Palembang, Indonesia
1830803060@radenfatah.ac.id

Gusmelia Testiana
Department of Information System
UIN Raden Fatah
Palembang, Indonesia
Gusmeliatestiana_uin@radenfatah.ac.id

Imamulhakim Syahid Putra
Department of Information System
UIN Raden Fatah
Palembang, Indonesia
imamulhakim@radenfatah.ac.id

Article History
Received November 30th, 2022
Revised January 13th, 2023
Accepted January 13th, 2023
Published February 2023

***Abstract*—YBM PLN UIWS2JB provides scholarships for high school students who cannot afford school tuition due to poverty or other economic conditions. Since the target is specific, the foundation must carefully select the recipients to ensure the scholarship is granted to those who deserve it. The predetermined criteria combined with the limited quota available become a difficulty in itself as a large number of applications are coming in. Data mining with a classification method using the K-Nearest Neighbor algorithm is believed to be one of the alternative solutions to solve this problem. This study aims to examine how this method could help in the selection process to determine who is eligible to receive the scholarship, and it also aims to evaluate the algorithm's performance with the optimal K value. The findings of this research showed that the classification method using K-Nearest Neighbor is the potential to be applied in a case such as this. The results found an accuracy of 91% in the selection process and are included in the Excellent Classification category. The optimal K value obtained is K = 5.**

Keywords--data mining; classification; predetermined criteria; selection; algorithm's performance

1 INTRODUCTION

Yayasan Baitul Maal PLN or commonly abbreviated as YBM PLN UIWS2JB is a foundation that regularly provides scholarships named Beasiswa Cahaya Pintar. It is part of their concern about education in Indonesia, especially in the city of Palembang. This scholarship aims to help students, especially high school students, who cannot afford school fees, due to economic issues. Therefore, it must be distributed carefully to those who deserve. it

The process to select high school scholarship recipients involves screening with a number of predetermined criteria to ensure it is granted to the right person. Besides, YBM PLN UIWS2JB set a limited quota of scholarships. The increasing number of scholarship applicants creates challenges and difficulties for the YBM PLN UIWS2JB to decide who is truly entitled to receive the scholarships. Moreover, the selection process has so far been quite ineffective and inefficient as it was still conducted manually. proper scholarship recipients. Therefore, the foundation must seek some solutions to this problem.

Processing data mining by using the right methods and algorithms can be an alternative to overcome this [1]. One method in data mining that is appropriate to use in this case is classification, namely the process of obtaining models, patterns, or formulas that can classify new data by studying old data sets. Classification using data mining is needed to take advantage of a set of old data that already exists and was previously only stored and then processed to find out the model or pattern contained in it so that it can become useful knowledge [2]. Therefore, this method is the potential to be used to help the YBM PLN UIWS2JB to select the high school scholarship recipients precisely, effectively, and efficiently.

There are many algorithms in the classification including K-Nearest Neighbor. The concept of this algorithm is to see from the closest distance training data with an object based on the value of K [3]. The KNN algorithm has advantages including strong consistency and resistance to training data error or outliers and is effective if the training data is large [4]. Several studies comparing several classification algorithms also show that the KNN algorithm has better accuracy performance than other classification algorithms. In a research that classified the type of cattle by comparing the algorithm K-Nearest Neighbor and Support Vector Machines, it was found that the final result obtained algorithm K-Nearest Neighbor with an accuracy of 100% while Support Vector Machine by 80% [5].

Similar research had also been conducted by Julaiha et all using K-Nearest Neighbor algorithms on the classification of prospective scholarship recipients at the Raja Ali Haji Maritime University [6]. The data used in this study were the data of Bidikmisi registrants in 2019. In the test modeling, the k value used was equal to 5 (k=5). The final result of the test showed that the accuracy obtained was 83.13%, the precision value was 82.35%, and the recall by 89.36% [6].

Furthermore, another research was by Jaman et all that applied the algorithm K-Nearest Neighbor for data classification of prospective Bidikmisi recipients in 2018 with a total data of 358 Bidikmisi registrant data with 5 variable attributes in it [7]. For sharing test data and training data, the researchers used the training set option, while for the best accuracy value, the results of processing data mining were evaluated using a confusion matrix with elements namely accuracy, precision, and recall. Based on the test results, the accuracy obtained was 97.2067%, the highest precision value was 0.972, and the highest recall value was 0.972 [7]. The calculation of the KNN algorithm in this study was carried out using WEKA Software 3.8.3

Subsequent research also applied K-Nearest Neighbor algorithms for the classification of prospective UNS Bidikmisi recipients [[8]. The data used by the researcher was approximately 2039 data from the registrants which contained 8 variables. The final result of the calculation process using the algorithm K-Nearest Neighbor showed an accuracy value of 84.4% and the value of kappa total of 0.63 [8]. The calculation of the KNN algorithm in this study was carried out using R-Studio Software.

Based on several previous studies and from what has been explained in the background, this research will classify scholarship recipients using an algorithm. K-Nearest Neighbor to be able to assist the process selection of high school scholarship recipients by YBM PLN UIWS2JB. In this study, the data that will be used as the object of research are taken from the high school scholarships by YBM PLN UIWS2JB. For the variable attributes, this research used approximately 15 influential variable attributes, including school origin, father and mother status, father's occupation, father's income, mother's occupation, mother's income, final semester report cards, home conditions, certificate of incapacity, letters of recommendation, academic achievements and non-academic achievements. As for the novelty and different contributions of this research to previous research, it is the first for algorithm calculations using the software RapidMiner. In addition, several K values will be used in this research. Each of these will be evaluated to determine the optimal K value, contrary to previous studies where the value of K used was directly determined.

2 METHOD

The method used in this study is the quantitative method with a descriptive approach. Quantitative research is a method of research used to examine a population or certain sample, data collection using research instruments, quantitative analysis or statistical data [16]. The purpose of quantitative research methods is to identify the problems defined as goals in research, make predictions to decide the problem, get data to predict, and interpret or analyze data to see if they support the prediction [17]

The quantitative method used will be elaborated. A descriptive approach is an attempt to describe or explain a symptom, event, or event that occurs [18]. Based on the definitions of quantitative and descriptive approaches as mentioned above, it can be concluded that descriptive research with a quantitative approach is conducted by looking for

information related to existing symptoms through data collection in numbers which will be followed by data interpretation.

The stages of this research are shown in Figure 1 as follows:

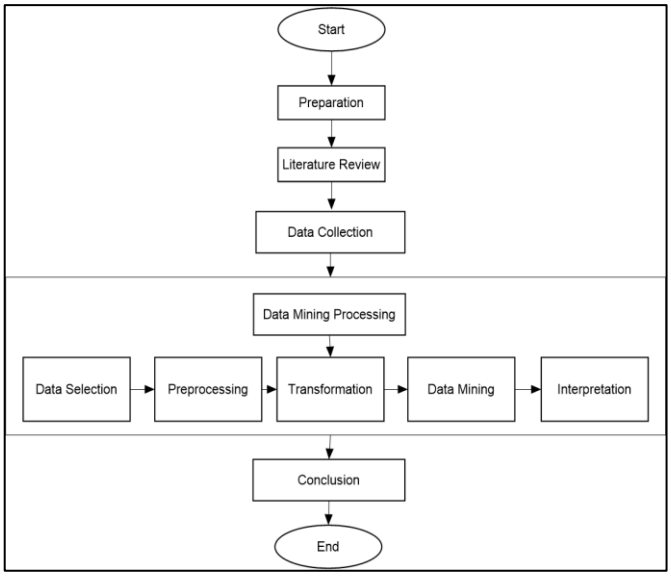

Figure 1. Stages of Research

Fig.1 shows the stages of the research, which are explained as follows:

1. The first stage is preparation, which is to prepare the research plan.
2. The second stage is a literature review, which is a study of the sources of books and others that are related to the problem under study.
3. The third stage is data collection, one of which is done by way of interviews.
4. The fourth stage is data mining processing, which is carried out with the data processing stage to form knowledge discovery in database.
5. The last stage is the conclusion, which is done by making conclusions regarding the results that have been obtained.

In stage 4, data mining includes an activity to analyze and explore large amount of data to find something of value that would be important and very useful knowledge and information [12]. The stages of data mining for knowledge discovery in a database (KDD) are shown in Figure 2 [13].

The following is an explanation from Fig. 2 regarding the stages of the KDD:

1. Selection is the process of selecting relevant data from the data set to be used.
2. Pre-processing is a process carried out with the aim of cleaning data that contains errors, duplicate, and data which are inconsistent.
3. Transformation is a process that is carried out by transforming data into a form suitable for data mining processes.

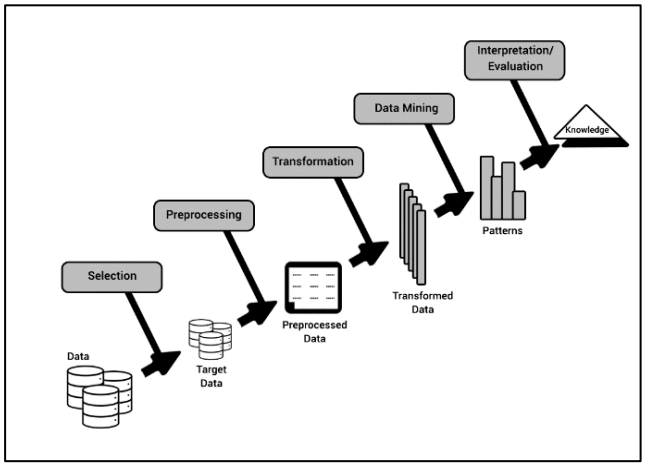

Figure 2. Stages of Data Mining for knowledge discovery in database

4. Data mining is the main stage. This stage is carried out to find patterns, relationships or interesting information contained in the selected data from the previous process by using a particular way or method. The data mining technique that will be used in this case is classification for algorithm using the K-Nearest Neighbor.
5. Interpretation/Evaluation is a process that includes observation pattern, information or results obtained.

Meanwhile the stages of the KNN algorithm are shown in Figure 3 as follows:

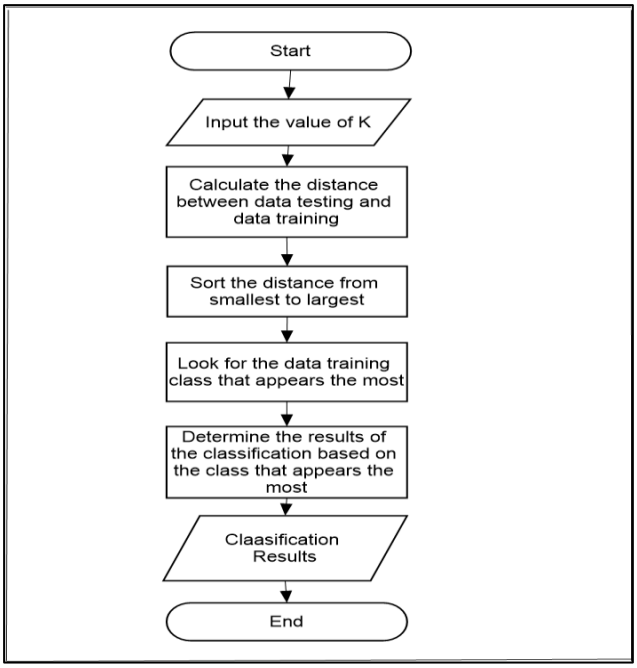

Figure 3. Stages of KNN

The following is an explanation from Fig. 3 regarding the stages of the KNN algorithm:

1. Determining the number of nearest neighbors symbolized by the value of the K parameter. There is no definite method for determining the right K value, so in this study, a test will be carried out on each K value parameter to find the optimal K value. The K value to be tested is odd K of 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, and 21.
2. Calculating distance test data to training data by using the formula Euclidean Distance and determining the value of the square of the distance between the object test data and the training data provided. Euclidean Distance is used because it has the best accuracy among other distance formulas [15]. The formula Euclidean Distance is shown in Formula 1 as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{K}(xi - yi)^2} \qquad (1)$$

Information:

d(x,y) : Euclidean Distance

K: Number of attributes

yi: Data Testing to i

xi: Data Training to i

i: 1, 2, 3, …, n

3. Sorting the calculation results in stage two in order from the lowest value to the highest value.
4. Collecting data on the nearest neighbor classification category based on the K value.
5. Classification of the nearest neighbor based on the majority of K values is the final result of the classification.

3 RESULT AND DISCUSSION

Results and discussion are carried out by following the data mining KDD stages.

3.1 Data Selection

The data used in this study is scholarship data for high school level obtained from YBM PLN UIWS2JB as the organizer of the scholarship with a total of 200 data consisting of 95 data of "pass" and 105 data of "no pass", and 23 variables or attributes in it, consisting of 22 determinant variables and 1 decision class variable. Table 1 shows the data.

Data Selection is a stage that aims to select data that are relevant to what is needed for the process of data mining. This is done by eliminating variables that are not used during the process of data mining conducted.

Table 1. Attributed List

Attributed Used	Removed Attributed
School Origin	Name
Father & Mother's Status	Gender
Father & Mother's Job	Father & Mother Education
Father & Mother's Income	Name of Father & Mother
The number of dependents	FC KK & FC KTP
Certificate of incapacity	
Home Condition	
Letter of recommendation	
Academic & non-academic achievements	
Decision status	

3.2 Data Preprocessing

At the preprocessing stage, what is done is a process of cleaning data. This process includes, among others, cleaning data that contain errors, duplicate data, and inconsistent data. In the dataset used in this study, there were no data records that had to be cleaned, so the dataset used remained at 200.

3.3 Data Transformation

At this stage, the transformation of the selected data will be carried out and preprocessed, so it becomes a form suitable for the process of data mining. The variables or attributes that need to be transformed are all variables except the status variable. This attribute change is based on the appropriate theory.

3.3.1 *School Origin:* The school origin is a determining attribute that must be transformed based on their place or area of schooling. Because the high school scholarship by YBM PLN Palembang is a scholarship intended for students in the city of Palembang, the school origin information is a determinant of the registrant's school of origin. The attribute transformation is shown in Table 2.

Table 2. School Origin Variable Transformation

Variable	Sub Variable	Variable Value (Lower is Better)
School Origin	1. Palembang City	1
	2. Outside Palembang	2

3.3.2 *Father and Mother's Status:* The status of the father and the mother are classified as still alive and dead. The attribute transformation is shown in Table 3.

Table 3. Variable Transformation of Father's Status and Mother's Status

Variable	Sub Variable	Variable Value (Lower is Better)
Father and Mother Status	1. Died	1
	2. Still Alive	2

3.3.3 *Father and Mother's Job:* According to Badan Pusat Statistik, the types of work in Indonesia are classified into 11 groups [19]. The attribute transformation is shown in Table 4.

Table 4. Variable Transformation of Father & Mother's Job

Variable	Sub Variable	Variable Value (Lower is Better)
Father and Mother's Job	1. Not working	1
	2. Worker Rough, Power Cleanliness, and Power Related to That (YBDI)	2
	3. Operator and assembler Machine	3
	4. Power Processing and Crafts (YBDI)	4
	5. Power Effort Agriculture and Farm	5
	6. Power Effort Service and Power Sales shop and Market	6
	7. Power System Effort	7
	8. Technician and Assistant Professionals	8
	9. Power Professional	9
	10. Office Institution Legislative, office height, and Manager	10
	11. Member Soldier National Indonesia (TNI) and Indonesian	11

3.3.4 *Monthly Income of Father & Mother:* In transforming the income variable of father/mother 3. 80-89 3 4. 90-100 4 theory is used based on the classified income level. According to the Central Statistics Agency, there are four income groups of the population, namely the low-income group with an average of IDR. 1,500,000 per month, medium income group with an average of IDR 1,500,000 – IDR 2,500,000, high income group with an average of IDR 2,500,000 - IDR 3,500,000, and very high class with an average of IDR 3,500,000 [20]. The attribute transformation is shown in the Table 5.

Table 5. Variable Transformation of Father & Mother's Income Per Month

Variable	Sub Variable	Variable Value (Lower is Better)
Father & Mother's Income Per Month	1. None	1
	2. ≤ IDR 1,500,000	2
	3. IDR 1,500,000 - IDR 2,500,000	3
	4. ≥IDR 2,500,000 - IDR 3,500,000	4
	5. ≥IDR 3,500,000	5

3.3.5 *Number of Dependents:* According to the BKKBN, family types are categorized into two types, namely nuclear family and extended family. Family size is measured based on the number of family members and grouped into small families with 4 members, medium families with 5-6 members, and large families with 7 members [21]. The attribute transformation is shown in Table 6.

Table 6. Variable Transformation Number of Dependents

Variable	Sub Variable	Variable Value (Higher is Better)
	1. ≤ 4 People	1
	2. 5-6 People	2
	3.≥ 7 people	3

3.3.6 Score

In transforming the attribute variable of the average school value is based on the existing guidelines on report cards and categories that have been commonly used. Generally, the categories used are very good (A) with a value range of 90-100, good (B) with value range of 80-89, sufficient (C) with a value range of 70-79, and the last one is less than 60-69. The attribute transformation is shown in Table 7.

Table 7. Transformation Score Variable

Variable	Sub Variable	Variable Value (Higher is Better)
Score	1. 60-69	1
	2. 70-79	2
	3. 80-89	3
	4. 90-100	4

3.3.7 *Certificate of Incapacity:* In transforming the certificate of incapacity classified as present and non-existent, the attribute transformation is shown in Table 8.

Table 8. Transformation Certificate of Incapacity Variable

Variable	Sub Variable	Variable Value (Lower is Better)
Certificate of Incapacity	1. There	1
	2. None	2

3.3.8 *Letter of Recommendation:* In transforming the letters of recommendation classified into being and not being, the attribute transformation is shown in Table 9.

Table 9. Variable Transformation of Letter of Recommendation

Variable	Sub Variable	Variable Value (Lower is Better)
Letter of Recommendation	1. There	1
	2. None	2

3.3.9 *Home Condition:* In transforming the state of the house, it is classified into supporting criteria and not supporting criteria. The attribute transformation is shown in the Table 10.

Table 10. Home Condition Variable Transformation

Variable	Sub Variable	Variable Value (Lower is Better)
Home Condition	1. Supporting Criteria	1
	2. Not Supporting Criteria	2

3.3.10 *Academic Achievement:* In transforming the academic achievement variable as is found in Table 14, the training data are divided into 6 categories, namely school level achievement, city level achievement, provincial level achievement, national level achievement, and international level achievement [13]. The attribute transformation is shown in Table 11.

Table 11. Variable Transformation of Academic Achievement

Variable	Sub Variable	Variable Value (Higher is Better)
Academic achievement	1. None	1
	2. Achievement at School Level	2
	3. Achievement at City Level	3
	4. Achievement at Province Level	4
	5. Achievement at National Level	5
	6. Achievement at International Level	6

3.3.11 *Non-Academic Achievements:* In transforming the academic achievement, the variable is divided into 5 categories, namely city level achievement, provincial level achievement, national level achievement, and international level achievement [13]. The attribute transformation is shown in Table 12.

Table 12. Variable Transformation of Non-Academic Achievement

Variable	Sub Variable	Variable Value (Higher is Better)
Non-Academic Achievement	1. None	1
	2. Achievement at School Level	2
	3. Achievement at City Level	3
	4. Achievement at Province Level	4
	5. Achievement at National Level	5

The results of the transformation is shown in Table 13.

Table 13. Transformation Results

No	School Origin	Father Status	Father's Job	Father's Income/Month	Mother Status	...	Home Condition	Academic Achievement	Non-Academic Achievement	Status
1	1	2	2	3	2	⋮	1	1	1	Pass
2	1	2	2	3	2	⋮	1	1	1	Pass
...	...	...	...	...	...	⋮	...	...	...	...
200	1	2	2	3	2	⋮	1	1	1	Not Pass

After transforming the dataset, at this stage, the dataset will also be divided into training data (Table 14) and testing data (Table 15). The distribution is carried out with a comparison ratio of 80% of training data and 20% of testing data. Therefore, the results of the training data are 160 data and the results for data testing are 40 data.

Table 14. Training Data

No	School Origin	Father Status	Father's Job	Father's Income/Month	Mother Status	...	Home Condition	Academic Achievement	Non-Academic Achievement	Status
1	1	2	2	3	2	⋮	1	1	1	Pass
2	1	2	2	3	2	⋮	1	1	1	Pass
...	...	...	...	...	...	⋮	...	...	...	...
160	1	2	2	2	2	⋮	2	1	2	Not Pass

Table 15. Testing Data

No	School Origin	Father Status	Father's Job	Father's Income/Month	Mother Status	...	Home Condition	Academic Achievement	Non-Academic Achievement	Status
1	1	2	2	3	2	⋮	1	1	1	Pass
2	1	1	1	1	2	⋮	1	2	2	Pass
...	...	...	...	...	...	⋮	...	...	...	...
40	1	2	2	3	2	⋮	1	1	1	Not Pass

3.4 Data Mining

3.4.1 *Calculation of K-Nearest Neighbor:* Based on the stages of data mining for K-Nearest Neighbor algorithms, the steps from K-Nearest Neighbor are as follows:

- Determination of K Value. In determining the optimal K value, there is no definite formula. In this study, an odd value of K will be used because for an even value of K with an even number of classifications, there will be a possibility of voting from both classifications getting the same vote, but for an odd value of K with an even number of classifications, it will be easier because it is guaranteed that both classes will not get the same sound [22]. This study used several odd-value of K values from the range 1-21, namely K=1. K=3, K=5, K=7, K=9, K=11, K=13, K-15, K=17, K=19, and K=21. Therefore, at the time of implementation RapidMiner testing and evaluation of the model that has been made using these K values were carried out to determine the optimal K value.

- Calculate distance test data to training data by using the formula given in Formula 1.

$$d(1,1) = \sqrt{(1+1)^2 + (2+2)^2 + \cdots + (1-1)^2} = 1.414213562$$

$$d(2,1) = \sqrt{(1+1)^2 + (2+2)^2 + \cdots + (1-1)^2} = 1$$

$$d(3,1) = \sqrt{(1+1)^2 + (2+2)^2 + \cdots + (1-1)^2} = 1.732050808$$

...

$$d(160,1) = \sqrt{(1+1)^2 + (2+2)^2 + \cdots + (2-1)^2} = 2$$

The distance between training data and test data was then calculated using similar method, so the calculation results obtained Euclidean distance among training data in order to test data 1. Table 16 shows the results.

Table 16. Distance Calculation Results

No	School Origin	Father Status	Father's occupation	…	Home Condition	Academic Achievement	Non-Academic Achievement	Status	Distance
1	1	2	2	⋮	1	1	1	Pass	1.4142136
2	1	2	2	⋮	1	1	1	Pass	1
…	…	…	…	⋮	…	…	…	…	…
160	1	2	2	⋮	2	1	2	Pass	2

- Sort the Euclidean Distance from smallest to largest. Table 17 shows the results.

Table 17. Distance Sorting Results

No	School Origin	Father Status	Father's occupation	…	Home Condition	Academic Achievement	Non-Academic Achievement	Status	Distance
54	1	2	2	⋮	1	1	1	Pass	0
2	1	2	2	⋮	1	1	1	Pass	1
…	…	…	…	⋮	…	…	…	…	…
99	2	2	5	⋮	2	2	1	Not Pass	8.7749644

- Determine the test data group based on the majority label of the K nearest neighbors (Table 18).

Table 18. Data Group with Example K=5

No	School Origin	Father Status	Father's occupation	…	Home Condition	Academic Achievement	Non-Academic Achievement	Status	Distance
54	1	2	2	⋮	1	1	1	Pass	0
2	1	2	2	⋮	1	1	1	Pass	1
11	1	2	2	⋮	1	1	1	Pass	1
13	1	2	2	⋮	1	1	2	Pass	1
15	1	2	2	⋮	1	1	1	Pass	1

- Determine the decision class. Table 19 shows the results.

Table 19. Overall Result Classification Decision

Data Testing	K=1	K=3	K=5	K=7	…	K=21
1	Pass	Pass	Pass	Pass	…	Pass
2	Pass	Pass	Pass	Pass	…	Pass
3	Pass	Pass	Pass	Pass	…	Pass
4	Not Pass	Not Pass	Not Pass	Not Pass	…	Not Pass
5	Not Pass	Not Pass	Not Pass	Not Pass	…	Not Pass
6	Pass	Pass	Pass	Pass	…	Pass
7	Pass	Pass	Pass	Pass	…	Pass
8	Pass	Pass	Pass	Pass	…	Pass
9	Pass	Pass	Pass	Pass	…	Pass
10	Not Pass	Not Pass	Not Pass	Not Pass	…	Not Pass
…	…	…	…	…	…	…
40	Pass	Not Pass	Not Pass	Pass	…	Pass

3.4.2 *Implementation of RapidMiner:* Implementation of RapidMiner in the classification of scholarship recipients can be done by adding an algorithm operator of K-Nearest Neighbor and Apply Model and then connecting all operator functions to display the results of the classification of scholarship recipients, whether or not they were predicted to pass or not pass. Figure 4 shows the operations. In Fig 4. for operators of K-Nearest Neighbor, we enter the K value to be used, so that the classification results were obtained, namely the decision to pass or not pass as given in Table 20.

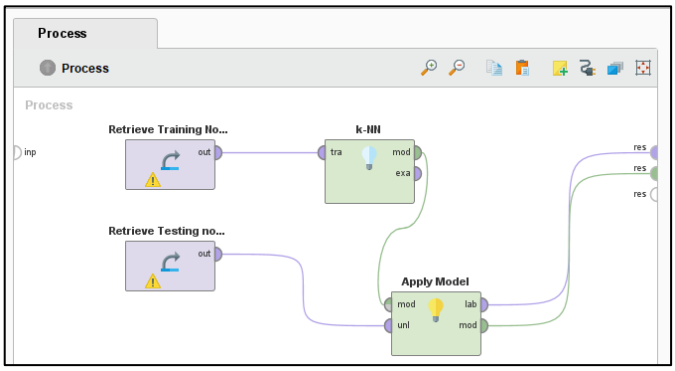

Figure 4. RapidMiner Classification Model

Table 20. RapidMiner Classification Results

No	Status	Prediction	Confidence (Pass)	Confidence(Not Pass)	School Origin	Father's Status	Father's Occupation	...	Non Academic Achievement
1	Pass	Pass	1	0	1	2	2	⋮	1
2	Pass	Pass	1	0	1	1	1	⋮	2
3	Pass	Pass	1	0	1	2	2	⋮	1
4	Not	Not	0	1	1	2	2	⋮	1
5	Not	Pass	0.202	0.798	1	2	9	⋮	1
...	...	...	...	...	...	...	...	⋮	...
40	Not	Pass	0.793	0.207	1	2	2	⋮	1

Table 20 displays the results obtained with the value of K = 5. From these results, there are similarities with the classification results obtained from manual calculations using Microsoft Excel, which were taken from the 40 data tested, 18 data were predicted to fail and 22 data were predicted to pass. Likewise, with the results of the classification using other K values, namely K = 1, K = 3, K = 7, K = 9, K = 11, K = 13, K = 15, K = 17, K = 19, and K = 21, the same results were obtained for the manual classification results. The following Table 21 shows the overall classification results using RapidMiner. Some of the K are not displayed to fit to the paper.

Table 21. RapidMiner Overall Classification Results Final

Data Testing	K=1	K=3	K=5	K=7	...	K=21
1	Pass	Pass	Pass	Pass	...	Pass
2	Pass	Pass	Pass	Pass	...	Pass
3	Pass	Pass	Pass	Pass	...	Pass
4	Not Pass	Not Pass	Not Pass	Not Pass	...	Not Pass
5	Not Pass	Not Pass	Not Pass	Not Pass	...	Not Pass
6	Pass	Pass	Pass	Pass	...	Pass
7	Pass	Pass	Pass	Pass	...	Pass
8	Pass	Pass	Pass	Pass	...	Pass
9	Pass	Pass	Pass	Pass	...	Pass
10	Not Pass	Not Pass	Not Pass	Not Pass	...	Not Pass
...	...	...	...	...	...	...
40	Pass	Not Pass	Not Pass	Pass	...	Pass

3.5 Evaluation

This stage is the final stage of KDD data mining. At this stage, the results obtained from the calculation process using the algorithm K-Nearest Neighbor were evaluated using the operator Cross Validation on RapidMiner (Figure 5 and 6).

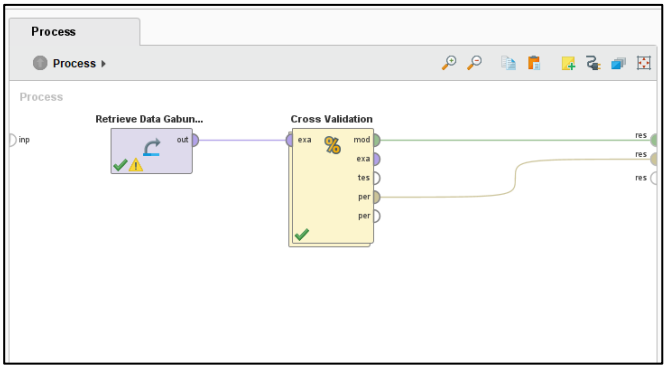

Figure 5. Cross Validation

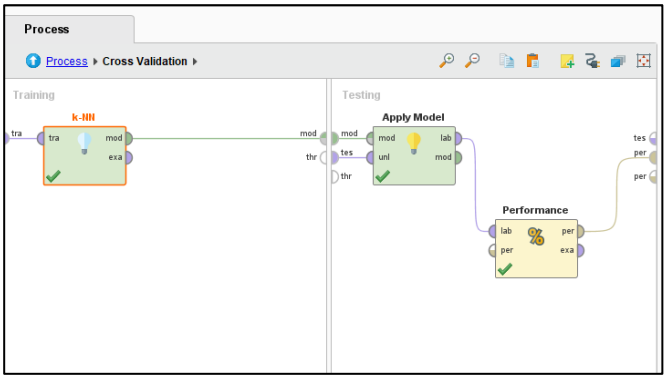

Figure 6. Operators in Cross Validation

After testing each classification model using the value of K = 1, K = 3, K = 5, K = 7, K = 9, K = 11, K = 13, K = 15, K = 17, K = 19, and K = 21, the evaluation results were obtained. They are shown in Table 22 as follows.

Table 22. Accuracy Results

K Value	Accuracy
K=1	88%
K=3	90,50%
K=5	91%
K=7	89,50%
K=9	89%
K=11	89%
K=13	88%
K=15	89%
K=17	89,50%
K=19	89,50%
K=21	90%

The evaluation results showed that classification could be done using the optimal K value of K = 5 because it had an accuracy rate of 91% and was the highest among other K values after testing the model. Based on the results of the evaluation, it was also found that the accuracy for the value of K after K = 5, namely K = 7, K = 9, etc., did not experience a percentage increase in accuracy but on the contrary, experienced a decrease in percentage below the value of K = 5. Xu and Kumar in their book entitled *Top 10 Algorithms in Data Mining* states that the value of K which is too large in the calculation of the algorithm K-Nearest Neighbor will create a boundary in the classification, so it becomes blurred [23]. Therefore, the use of K values greater than five is not a good idea in the case of this study.

Based on the results obtained related to the value of K optimal performance and accuracy of modeling K-Nearest Neighbor algorithms can be made into a flowchart to make it easier to understand and can be a knowledge and recommendation for YBM PLN UIWS2JB in determining scholarship recipients by using a model or K-Nearest Neighbor algorithm. The flowchart model can be seen in Fig 7.

As previously explained, KNN does not have a definite formula for determining the optimal K values, so it is necessary to experiment by using and evaluating several K values so that the optimal K value can be found and can be applied in a classification model. Previous studies did not use several K values to see which one was the most optimal but instead directly determined the 1 K value without first making comparisons of several K values. Therefore, the results of this research could enrich the study that applies K-Nearest Neighbor algorithm. Besides, this research can be a future reference on the use of RapidMiner Software. Previous studies in this type of research usually used WEKA and R-Studio.

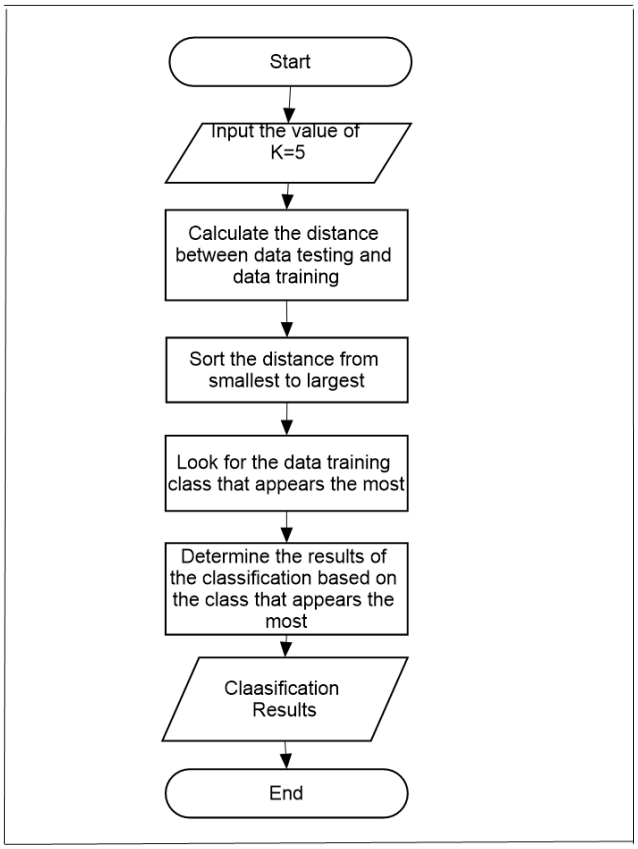

Figure 1. Flowchart Model

4 CONCLUSION

Based on the results of the research, it can be concluded that the classification of YBM PLN UIWS2JB scholarship recipients using the K-Nearest Neighbor algorithm can be done accurately by making a classification model using K = 5 against 200 scholarship datasets obtained with several determinant attribute variables in it. It is evident in the similarity between the results of manual classification and the results of classification using RapidMiner. From the 40 data tested, there were 18 data classified as “no pass” and 22 data classified as “pass”. And it is also apparent in the results of testing or evaluation of the K = 5 classification model. The accuracy results obtained with a percentage of 91% is considered in Excellent classification although the data used were not too large. Therefore, the classification model with K-Nearest Neighbor algorithm is recommended to be used in classifying scholarship recipients for high school level at YBM PLN UIWS2JB accurately and precisely.

The advantage of the results of the classification model using K-Nearest Neighbor algorithm is that it can classify easily by looking at the nearest neighbor (K values). However, there are also deficiencies in this algorithm. For examples, it does not have definite rules or formulas in determining the optimal K value, so the determination of the K value in this study must be carried out using several values which are then evaluated to obtain the optimal K value. Regardless of the limitations of this study, the results show that this model can be a recommendation for YBM PLN UIWS2JB in the future when making decision.

AUTHOR CONTRIBUTIONS

As the first author, Ica Dwi Yulisa mainly wrote this article and managed technical issues. She did the research with the supervision from the second author, Gusmelia Testiana, regarding the concept and contents of the research. And then the third author, Imamulhakim Syahid Putra also provided supervision regarding the theory and writing methods.

COMPETING INTEREST

Complying with the publication ethics of this journal, Ica Dwi Yulisa, Gusmelia Testiana an Imamulhakim Syahid Putra as the authors of this article declare that it is free from Conflict Of Interest (COI) or Competing Interest (CI).

REFERENCE

[1] M. Kholil, Kusrini, and Henderi, "Penerapan Metode K Nearest Neighbord Dalam Proses Seleksi Penerima Beasiswa," *Semin. Nas. Sist. Inf. Dan Teknol. Inf.*, Pp. 13–18, 2018.

[2] O. Kristanto, "Penerapan Algoritma Klasifikasi Data Mining Id3 Untuk Menentukan Penjurusan Siswa Sman 6," 2013.

[3] L. S. Putri, "Klasifikasi Minyak Goreng Berdasarkan Frekuensi Penggorengan Menggunakan Metode K-Nearest Neighbor Berbasis Raspberry Pi," Universitas Brawijaya Malang, 2018.

[4] S. A. P. Aji, H. Oktavianto, and Q. A'yun, "Klasifikasi Penerima Bantuan Dana Desa Menggunakan Metode Knn (K-Nearest Neighbor)," Pp. 1–11, 2020.

[5] S. Farah, A. Wijaya, K. Usman, and S. Saidah, "Analisis Perbandingan K- Nearest Neighbor Dan Support Vector Machine Pada Klasifikasi Jenis Sapi Dengan Metode Gray Level Coocurrence Matrix," Vol. 2, No. 2, Pp. 93–102, 2022.

[6] S. Julaiha, M. Bettiza, and D. A. Purnamasari, "Penerapan Algoritma K-Nearest Neighbor (Knn) Untuk Klasifikasi Calon Penerima Bidikmisi (Studi Kasus: Universitas Maritim Raja Ali Haji)," *Student Online J.*, Vol. 2, No. 1, Pp. 230–235, 2021.

[7] J. H. Jaman, and S. A. Fahlevi, "Klasifikasi Calon Mahasiswa Bidikmisi Dengan Algoritma K-Nearest Neighbor," *Pros. Annu. Res.*, Vol. 5, No. 1, Pp. 1–5, 2019.

[8] H. N. Zerlinda, I. Slamet, and E. Zukhronah, "Klasifikasi Calon Penerima Bidikmisi Dengan Menggunakan K-Nearest Neighbor," *Semin. Nas. Penelit. Pendidik. Mat. Umt*, Pp. 88–93, 2019.

[9] D. Sartika, and J. Jumadi, "Clustering Penilaian Kinerja Dosen Menggunakan Algoritma K-Means (Studi Kasus: Universitas Dehasen Bengkulu)," *Semin. Nas. Teknol. Komput. Sains*, Pp. 703–709, 2019, [Online]. Available: Https://Seminar-Id.Com/Semnas-Sainteks2019.Html.

[10] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Deepublish, 2020.

[11] S. Hendrian, "Algoritma Klasifikasi Data Mining Untuk Memprediksi," Vol. 11, No. 3, Pp. 266–274, 2018.

[12] F. Gorunescu, *Data Mining "Concepts, Models, And Techniques."* Berlin: Springer, 2011.

[13] S. R. Cholil, T. Handayani, R. Prathivi, and T. Ardianita, "Implementasi Algoritma Klasifikasi K-Nearest Neighbor," *Indones. J. Comput. Inf. Technol.*, Vol. 6, No. 2, Pp. 118–127, 2021.

[14] A. Yudhana, Sunardi, and A. J. S. Hartanta, "Algoritma K-Nn Dengan Euclidean Distance Untuk Prediksi Hasil Penggergajian Kayu Sengon," *Transm. J. Ilm. Tek. Elektro*, Vol. 22, No. 4, Pp. 123–129, 2020.

[15] M. Nishom, "Perbandingan Akurasi Euclidean Distance , Minkowski Distance , Dan Manhattan Distance Pada Algoritma K-Means Clustering Berbasis Chi-Square," Vol. 04, No. 01, Pp. 20–24, 2019, Doi: 10.30591/Jpit.V4i1.1253.

[16] Sugiyono, "Metode Penelitian Kuantitatif Kualitatif Dan R&D." 2013.

[17] I. Santoso, and H. Madistriyanto, "Metode Penelitian Kuantitatif." Indigo Media, 2021.

[18] S. Arikunto, "Prosedur Penelitian: Suatu Pendekatan Praktik." Rineka Cipta, 2014.

[19] B. P. Statistik, *Klasifikasi Baku Jenis Pekerjaan Indonesia*. Jakarta: Badan Pusat Statistik, 2002.

[20] L. S. Rakasiwi, and A. Kautsar, "Pengaruh Faktor Demografi Dan Sosial Ekonomi Terhadap Status Kesehatan Individu Di Indonesia," *Kaji. Ekon. Dan Keuang.*, Vol. 5, Pp. 146–157, 2021.

[21] Elmanora, I. Muflikhati, and Alfiasari, "Kesejahteraan Keluarga Petani Kayu Manis," *J. Ilmu Kel. Dan Konsum.*, Vol. 5, Pp. 58–66, 2012.

[22] I. K. Hasan, Resmawan, and J. Ibrahim, "Perbandingan K-Nearest Neighbor Dan Random Forest Dengan Seleksi Fitur Information Gain Untuk Klasifikasi Lama Studi Mahasiswa," Vol. 5, No. 1, Pp. 58–66, 2022.

[23] X. Wu *Et Al.*, "Top 10 Algorithms In Data Mining," Vol. 06, 2007.