The Application of Determining Students' Graduation Status of STMIK Palangkaraya Using K-Nearest Neighbors Method

Lili Rusdiana[1] and Marfuah[2]
[1] STMIK Palangkaraya, Indonesia
[2] Universal University, Batam, 29456, Indonesia
E-mail: fasliiana7@gmail.com

Abstract: K-Nearest Neighbors method is one of methods used for classification which calculate a value to find out the closest in distance. It is used to group a set of data such as students' graduation status that are got from the amount of course credits taken by them, the grade point average (AVG), and the mini-thesis grade. The study is conducted to know the results of using K-Nearest Neighbors method on the application of determining students' graduation status, so it can be analyzed from the method used, the data, and the application constructed. The aim of this study is to find out the application results by using K-Nearest Neighbors concept to determine students' graduation status using the data of STMIK Palangkaraya students. The development of the software used Extreme Programming, since it was appropriate and precise for this study which was to quickly finish the project. The application was created using Microsoft Office Excel 2007 for the training data and Matlab 7 to implement the application. The result of K-Nearest Neighbors method on the application of determining students' graduation status was 92.5%. It could determine the predicate graduation of 94 data used from the initial data before the processing as many as 136 data which the maximal training data was 50data. The K-Nearest Neighbors method is one of methods used to group a set of data based on the closest value, so that using K-Nearest Neighbors method agreed with this study. The results of K-Nearest Neighbors method on the application of determining students' graduation status was 92.5% could determine the predicate graduation which is the maximal training data. The K-Nearest Neighbors method is one of methods used to group a set of data based on the closest value, so that using K-Nearest Neighbors method agreed with this study.

Keywords: *Extreme Programming, Graduation Status, K-Nearest Neighbor, STMIK Palangkaraya*

1. Introduction

The application of data mining to overcome the existing problems at this time is so widely used as for prediction. Data mining is used for data analysis of students who have bad record so that it can be found the solution for the problem solving on those students [1]. Data mining has a variety of methods, one of which is prediction such as a research conducted to predict the results of placement for college activities [2]. The research was conducted by comparing two methods namely Fuzzy Logic and K-Nearest Neighbor

(KNN). The result was that the KNN got superior with a value of 97.33% while the Fuzzy Logic of 92.67% to predict.

The so-called research of modified KNN method is conducted to improve KNN performance by modifying it that is to use robust neighbors in training data [3]. The development is aimed to get better results.

Based on the previous research using KNN method and the research in predicting and the research using students'data and its development, hence this research also related with data mining about prediction is that the classification by using data of university students, to classify the data into passed and not passed. The grouping of data obtained is based on amount of course credits has been taken, grade point average (GPA), and mini-thesis grade. The basis of such data grouping as one of requirements at Universities such as STMIK Palangkaraya as the one of universities that apply the regulation of the calculation of course credits amount, GPA, andmini-thesis grade as the requirements of students' graduation. The aim of this research is to know the results of application constructed using KNN concept in determining university students' graduation status (passed or not passed) based on amount of course credits, GPA, and mini-thesis grade of STMIK Palangkaraya students.

2. Literature Review

Classification is widely used in a group of data to look for an inter-data relationship. One of the methods used in the classification is KNN which classifies a tuple class by finding the closest value of the training data and after the closest data known, the tuple class will follow the closest data class that previously has been known through training.

KNN method is included into data mining such as a research ever conducted to analyze medical data [4].This study was conducted because of the development of patient data stored in a hospital. The data may be hidden though it may be helpful to analyze for more accurate diagnostic results. KNN algorithm was used to obtain more accurate diagnostic results. The results of research were an application that can overcome errors in diagnosing.

KNN algorithm is used to analyze distance measurement using a set of data. Classification is a process of analyzing input and building a model for a class. KNN is used to search for the unknown class tuple and as a method to find accuracy in distance measurement [5].

2.1 K-Nearest Neighbor Method

There are many methods can be used in classification, one of which is KNN method. KNN method requires test data and training data to calculate distance. Like other methods in general that have algorithms, KNN also has algorithms that do classification by calculating the proximity between the data. A value at a distance will be used to find the value of proximity or similarity between test data and training data. If there are results that have a difference then the number data will be selected.

The KNN algorithms and its formulation as follow [6]:

a. z = (x', y'), is test data with x' and y' unknown class label
   Where:
   x' : vector/ test data attribute
   y' : class label of unknown test data
b. counting distance d (x', y'), the distance between z-test data to each training data vector, stored into D
c. Select Dz D, where K is the nearest neighbor of z
d. y' = argmax (xi,yi) Dz I(v=yi)

To handle the largest amount of data, it is usually added the use of weights to calculate the class candidates that should be taken by the test data from the nearest KNN.

The classification requires test data and training data by finding the closest data from the training data. After finding it, the closest data will be followed. The number of values will be used as training data. If the majority of data is in a class then the value obtained will go into it based on the *range* of data provided.

2.2 *Extreme Programming*

Extreme programming is one of agile system development with an approach to develop software through some stages as follows [7]:

a. Planning
   Planning is a stage to start defining the needs, the resulting output, the services developed in the application and its function.
b. Design
   This stage is part of designing a simple application designing in accordance with the needs of the system or application.
c. Coding
   Coding stage is a stage in prepare the code that will be used in the application development so that it can be a problem solving.
d. Testing
   This is a stage to test the services or features and functionality contained in the built application.

2.3 Screen Design Worksheet

One of the tools to design the display is by using screen design worksheet. The interface design can use it in the form of paper sheet to facilitate the documentation. It will be very useful for programmers in building a program[8]. Screen design worksheet consists of 4 parts:

a. Worksheet number
   It based on the number of display made.
b. Display
   The design of the display that will appear on the screen.
c. Navigation
   Explanations of display designed.
d. Description
   Description of the display parts.

3. Analysis and Design

3.1 Needs Analysis

Needs analysis is a discussion of some needs and or requirements related to input data, processes and output results. The needs or requirements were obtained from the data used to students' graduation status and the academic division is that the data of STMIK students in information system programs of study. Based on the obtained data, the results of needs analysis were as follows:

a. Needs of input
   The built application required input data, they were students data such as students ID number, name, course credits collected, GPA, and grade of mini-thesis.
b. Needs of process
   The process used to process input data was clustering technique using KNN algorithm.
c. Needs of output
   Output was the analysis of the results of the use of KNN algorithm so that could be known the information pattern from input data.

3.2 Data Collection Technique

In the data collection, the researcher used some techniques in the research. It helped the researcher to gather the right data and could be processed. The techniques used were:

a. Observation
   Observation was to collect data by taking samples of students' data information system programs of study in STMIK Palangkaraya and made observation on the objects under the study to find out the current system in STMIK Palangkaraya especially information system programs of study of year 2011. From the observation, the researcher found that the graduation is based on 3 things namely amount of course credits, GPA, and grade of mini-thesis.

b. Library
   Library was to collect data by reading the literature or books related to students' graduation and KNN as well as visiting websites that fit the issues discussed in this study such as visiting the website to find out the graduation requirements at the college.

3.3 Data Analysis Technique

Data analysis technique used in this study was adopting the technique contained in the Agile model is that Extreme Programming as in Figure 1.

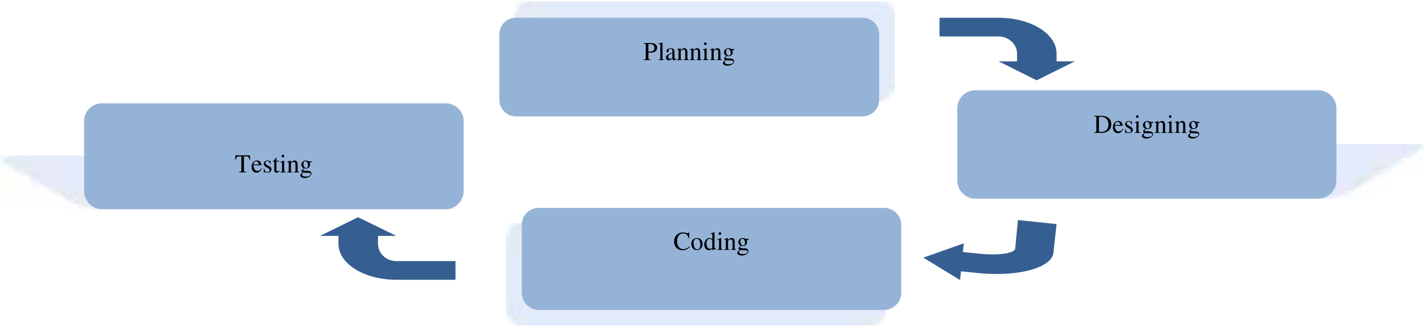


Figure1. Data Analysis Technique

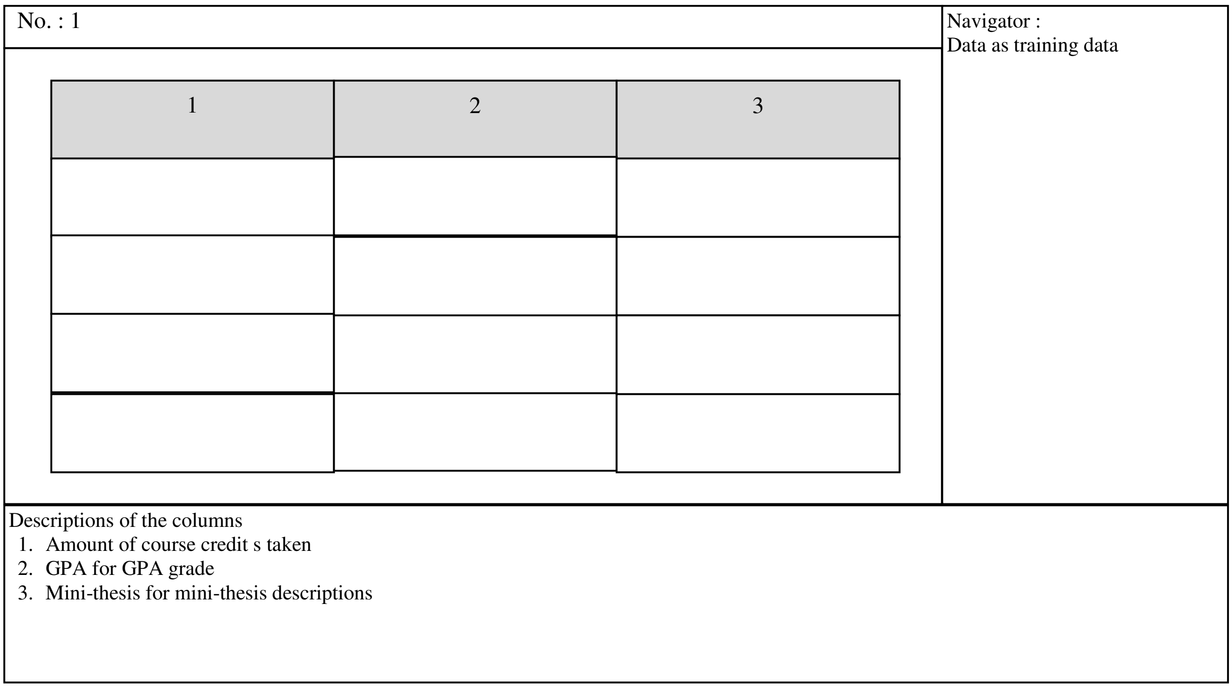
The steps of Extreme Programming used in this study as in Figure 1 as follows:

1) Planning
2) To acknowledge the scope of work as a plan to model the system, including searching for data and site information related to the needs of students' graduation data, features, and functions in accordance with system development is that students data such as name, GPA, amount of course credits, and information about grade of mini-thesis.
3) Design
4) This stage made a design of the work flow and a design that was needed in this research is that application development using Microsoft Office Excel software to design training data then it was processed using Matlab 7 software as well as the test data.
5) Coding
6) Coding stage used the codes contained in the Matlab 7 software to use training data and test data in which the test data previously were compiled using Microsoft Office Excel software.
7) Testing
8) Model testing used a simple application to know the results of model used so it got a conclusion about the services or features and functionality in the constructed application.

3.4 Design

The interface design used display worksheet for calculation view design and data analysis of the completed calculation using each method. The display worksheet in this study used Microsoft Office Word in which the worksheet itself consisted of worksheet number, display section, navigator, and description. The design facilitated preparing a simple program for the use of KNN method in general, software design considered the user needs, they are the interface display and the correctly calculation/computation process. In this study, software design was only a process and the results of calculation/computation using KNN method using students' data to know the students' graduation status. The initial design for the data used as training data was using Microsoft Office Excel 2007 software while application design and code used Matlab 7.

The interface design used display worksheet to design display of calculation and data analysis based on input and output data. The display worksheet applied Microsoft Office word consisting of number, display section, navigator, and description. The interface can make changes to the data before they entered by using KNN method such as initialization by allocating all data in a group randomly as shown in figure2 for the design and figure 3 the interface design in figure2, number 1 to 3 shows the input data and number 4 the output data.



No. : 1

Navigator :
Data as training data

1	2	3

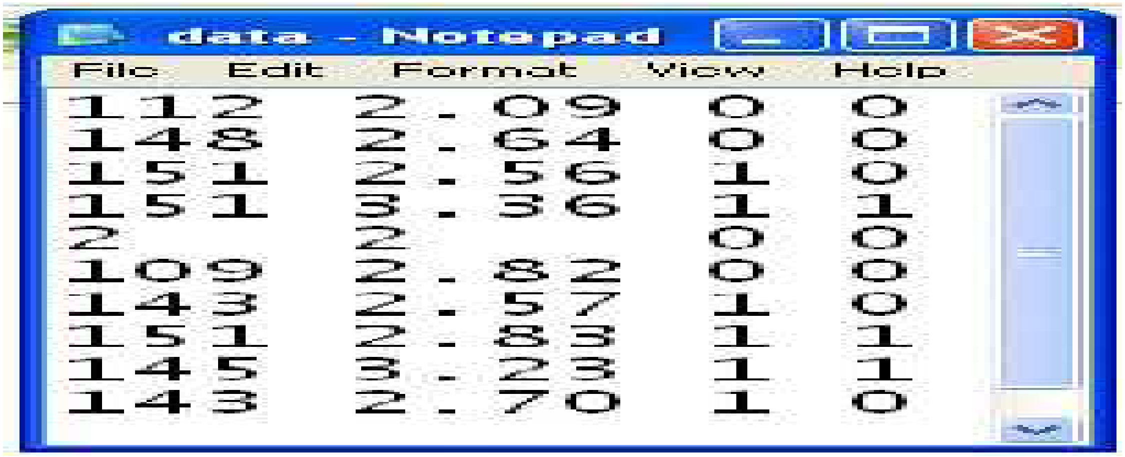
Descriptions of the columns
1. Amount of course credit s taken
2. GPA for GPA grade
3. Mini-thesis for mini-thesis descriptions

Figure2. Interface Design of Data initialization

4. Implementation and Discussion

4.1 Implementation

Interface was based on the input and output data required to implement the use of KNN method. Figure 3 shows the data from 10 training data. It would be used as the nearest search is that 0 and 1.



```
112 2.09 0 0
148 2.64 0 0
151 2.56 1 0
151 3.36 1 1
2   2    0 0
109 2.82 0 0
143 2.57 1 0
151 2.83 1 1
145 3.23 1 1
143 2.70 1 0
```

Figure 3. 10 training data

The interface was based on the design using display worksheet intended for algorithm usage and for data analysis of completed algorithm results. From the interface of Matlab itself, the researcher could make data input required before the data went to the concept of KNN method like doing initialization by randomly taking sample data of a group.

Figure 4 shows the interface of constructed application along with the samples of input and output of the application.

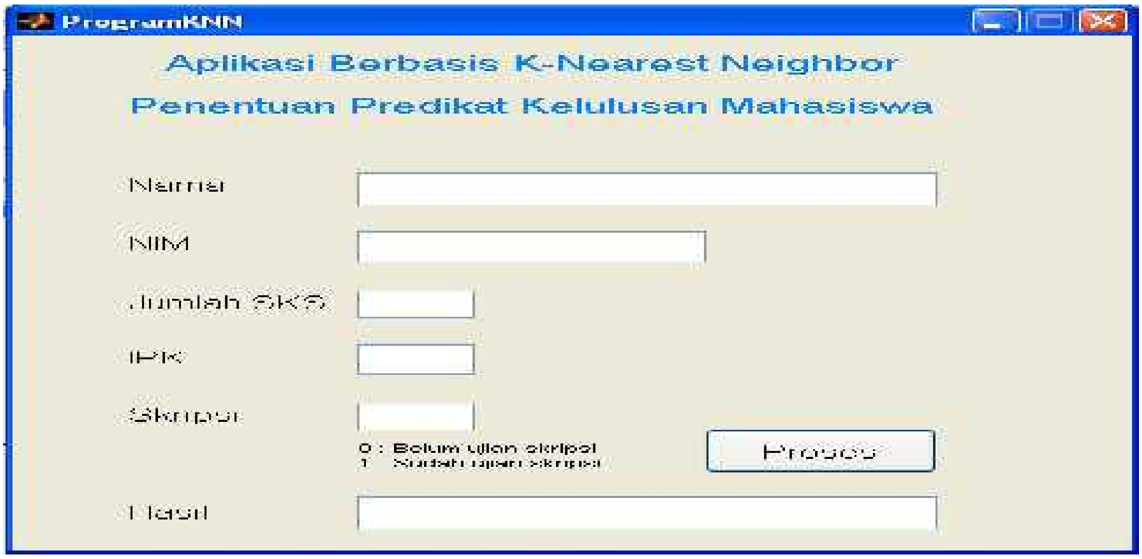


Figure 4. *Interface* Application View

Based on the process using clustering technique and KNN algorithms to determine students graduation through the input data was that amount of course credits taken, GPA achieved by the students, and mini-thesis grade then it was found the results for 10 training data and for each 50 training data given 2 phases of test that were testing for training data and all data. By the test of 10 training data, it was found that there was 10 suitable training data and none was. So, all data was recognized. While the test of all 94 data based on 10 training data was found that 88 data was suitable and 6 data was not. The test of training test based on 50 training data was 45 suitable data and 5 other was not. While the test of all data was that 94 data found that 87 data was suitable while 7 data was not. The comparison of training data testing and all data as shown in Table 1.

Table1. Comparison of data testing

Training data	10		50	
Test data	10	94	50	94
Suitable	10	88	45	87
Not suitable	0	6	5	7

Based on Table.1 it was found the percentage comparison as shown in Table 2.

Table2. Percentage Comparison of Data Testing

Training data	10		50	
Test Data	10	94	50	94
Suitable	100%	93,6%	90%	92,5%
Not suitable	0%	6,4%	10%	7,5%

At the analytical stage using KNN method, the data used as a test in this study was a combination of input and output in the form of a matrix of 3 columns. They were a column for amount of course credits, one for GPA achieved by students, and the other for mini-thesis grade description by giving 2 descriptions was that 0 (zero) and 1 (one). 0 (zero) for not passed and 1(one) for passed.

4.2 Discussion

On the clustering, the results were in the form of membership of each data on each cluster and the dominant data went into a certain cluster by finding the nearest value because the principle of KNN method is that finding the neighbors or nearby value. The use of a set of initial data as many as 136 data and after getting pre-processing the data became 94 used as data in the application of using KNN method by eliminating some data which could not be used such as double data.

Table 1 shows comparison of two tests using training data and it did not need training as in KNN method. The inconsistent value was smaller and the corresponding value was greater than the test stages performed on the more extensive data. The comparative calculation between standard testing and system testing as the calculation of accuracy using MAPE formula and the results as in table 2 showing comparison of training and testing in the form of percentage with the highest value of 100% system could recognize all test data and 0% system could recognize data from all data by using minimal training data. But through the increasing of training data the number of percentages of conformity decreased but not dramatically, the improvement remained in the test of 50 training data on 94 test data. From the results it was found that the more data trained the better suitability of data testing achieved, it could be seen from the improvement of system performance on determining students graduation status.

5. Conclusions and Suggestions

5.1 Conclusions

In this research can be concluded that the 136 initial data became 94 after experiencing pre-processing. It was served as data in the application of KNN concept. The data divides into training data and test data. The test results showed that training data could be recognized. But from the results of testing the all data, not all could be really recognizable. It indicates that the input data can be used for this research.

From the test results, it is obtained that the more data are trained and tested, the application will be better in tewting. Maximum results show 92.5% KNN-based application can recognize data from the maximum of training dara and test data used.

5.2 Suggestions

Development can be done on this research as in the following suggestions:

1) The application can be developed by adding the amount of data in which not only for one program of study, such as having others data on another program of study so the amount of data used will be even more to know the accuracy of application and method.
2) Other methods can be developed for the same case so that later it can be a new study on the comparison methods on the use of the same data to determine the quality of method used.
3) Other methods can be combined or developed in building the application in order to know comparison of results such as combining with substractive clustering method.
4) This application can be a recommendation for the parties who need such other educational institutions for research, development and use of the application itself as developed into an intelligent system or expert system.

6. References

[1] Pranav P., "*A Study of Student's Academic Performance Using Data Mining Techniques*". International Journal of Research In Computer Applications And Robotics. ISSN 2320-7345. Vol. 3, Issue 9, 2015, 59-63.

[2] Sheetal, B.M and Bakare, S., "*Prediction of Campus Placement Using Data Minng Algorithm-Fuzzy logic and K nearest neighbor*". International Journal of Advanced Research in Computer and Communication Engineering. ISSN 2278-1021. Vol. 5. Issue 6. 2016, 309-312.

[3] Parvin, H., Alizadeh, H., and B Minati, "*A modification on k-nearest neighbor classifier". Global Journal of Computer Science and Technology*. Vol.10, Issue 14, 2010, 37-41

[4] Khamis, H.S., Kriptu, W.C., and Stephen, K., "*Application of k-Nearst Neighbour Classification in Medical Data Mining*". International Journal of Information and Communication Technologi Research. Vol.4, No.4, 2014, 121-128.

[5] Mulak, P. and Nitin, T., "*Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset*". International Journal of Science and Research. Vol.4, Issue 7, 2015, 1201-1204.

[6] Prasetyo, E., *Data Mining Konsep dan Aplikasi* Menggunakan *Matlab*, Andi, Yogyakarta, 2012.

[7] Schach, R.S., Object-Oriented and Classical Software Engineering. Eighth Edition. McGraw-Hill. New York, 2011, 59.

[8] Santoso, I*., Interaksi Manusia dan Komputer*, Andi, Yogyakarta, 2010