Application Development of Student's Graduation Classification Model based on The First 2 Years Performance using K-Nearest Neighbor

P Prasetyawan[1], M Faridz Abadi[2]

1. Faculty of Engineering and Computer Science, University of Teknokrat Indonesia, Bandar Lampung, Indonesia. purwono.prasetyawan@teknokrat.ac.id
2. Faculty of Engineering and Computer Science, University of Teknokrat Indonesia, Bandar Lampung, Indonesia. faridjmiftah@gmail.com

Abstract— A College keeps a lot of data such as, academic data, administration, student biodata and others. The existing student data has not been fully utilized. In the student education system is an important asset for an educational institution and for that it is necessary to note the graduation rate of students on time. Differences in the ability of students to complete the study on time required the monitoring and evaluation, so that it can find new information or knowledge to make decisions. The purpose of this study, to know the relationship between IP variables Semester 1, IP Semester 2, IP Semester 3, IP Semester 4, Gender, Student Status on Student Study Duration using k-nearest neighbor algorithm. The result of this research in the classification of students' graduation using the knn algorithm based on student status, gender, ip semester 1 - ip semester 4 with k-fold cross validation in can mean value of K1 accuracy 88%, K3 accuracy 88.67%, K5 accuracy of 93.78%, K7 86% accuracy, K9 accuracy 86.22%, K11 accuracy 92.44%, K13 accuracy 89.55%, K15 accuracy 93.78%, K17 accuracy 99.78%, and K19 accuracy 100 %. Of the 500 training data in the status of 188 students, 312 students, the status of students work longer in completing the lecture and in the gender of 290 men, 210 women, then women longer in finishing college. Finding the optimal k value using k-fold cross validation. The result of accuracy using k-fold cross validation is K19 with 100% accuracy. Keywords— classification; duration study; k-nearest neighbor.

1. Introduction

Advances in information technology has been growing rapidly in all areas of life. Lots of data generated by sophisticated information technology, ranging from industry, economics, science and technology and various other fields of life. The application of information technology in the world of education can also produce abundant data about students and the learning process that is produced. In a study related to data, it takes a method or technique that can help in the process of implementation [1]. Data mining is a process for finding relationships and patterns to draw conclusions from existing data warehouses to be analyzed and explored so as to be useful in decision making. Data mining also utilizes the experience or even mistakes in the past to improve the quality of the model and the results of its analysis, one of them with the learning ability of data mining techniques that is classification. Classification is a learning task that maps a new object into one of the class or category labels on a predefined old object [2].

Universities are required to provide quality education for students to produce knowledgeable, competent, creative, and competitive human resources. In the student education system is an important asset for an educational institution and for that it is necessary to note the graduation rate of students just in time. The ups and downs of the students' ability to complete a timely study is one of the elements of university accreditation assessment [3]. Until now, the existing student data has not been fully utilized so it needs to be processed to find new information or knowledge to make decisions. One of the research using data mining is Classification using one method of data mining algorithm that is k-Nearest Neighbor (KNN). The KNN algorithm works based on the shortest distance from the new object to the old object by determining the value of k. The value of k is a parameter to determine the closest distance between the new object against the old object [4]. In this study, the authors used the variables of student status, gender, IP semester 1, IP semester 2, IP semester 3, and IP semester 4 to

This research sponsored by: Direktur Riset dan Pengabdian Masyarakat (DRPM), Kemenristekdikti

determine the effect on the duration of student study. This study uses k-Fold Cross Validation to know the accuracy of the classification results.

2. Theoretical basis

2.1 Datamining

Data mining is a series of processes to explore the added value of a data set of knowledge that has been unknown manually [5]. The tasks in data mining are generally divided into two main categories: Predictive and Descriptive. The purpose of a predictive task is to predict the value of a particular attribute based on the value of the other attributes. Predictable attributes are commonly known as targets or non-free variables, while the attributes used to make predictions are known as explanatory or independent variables. The purpose of the descriptive task is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize key relationships in the data. The task of descriptive data mining is an investigation and often requires postprocessing techniques for validation and explanation of results [6].

2.2 K-Nearest Neigbor

K-Nearest Neighbor (KNN) algorithm is a method to classify objects based on learning data closest to the object. KNN includes a supervised learning algorithm in which the results of a new instance query are classified by the majority of the categories in KNN. The most emerging class will be the classified result class. The purpose of this algorithm is to classify new objects based on attributes and training samples.

The K-Nearest Neighbor algorithm uses neighbor classification as the value of the new query instance. The algorithm is simple, works based on the shortest distance from the query instance to the training sample to determine its neighbors [7]. The closest distance calculation method using Euclidean Distance.

The Euclidean Distance method is presented as follows:

$$\mathrm{di} = \sqrt{\sum_{i=1}^{p} (x_{1i} - x_{2i})^2} \quad (1)$$

Information:

x1 = Data sample
x2 = Test data or data testing
i = Variable data
d = Distance
p = Dimensions of data.

KNN algorithm steps:

a) Specifies the parameter k (the number of nearest neighbors).
b) Calculates the euclidean distance of each object against the given sample data.
c) Sort those objects into groups that have the smallest euclidean distances.
d) Collect category Y (nearest neighbor classification).
e) By using the majority category, it can be classification results.

2.3 K-Fold Cross Validation

K-Fold Cross Validation which is one of the methods used to determine the average success of a system by looping by scrambling the input attribute so that the system is tested for some random input attributes. With K = 5 or 10 can be used to estimate the error rate, because the training data on each fold is quite different from the original training data. Overall, 5 or 10-fold cross validation are both recommended and mutually agreed upon. Calculating the value of its accuracy can be done using equation [8]:

$$Akurasi = \frac{jumlah\ prediksi\ benar}{jumlah\ data\ uji} x\ 100\% \quad (2)$$

3. Research Method

The method used for calculation uses the k-nearest neighbor algorithm, as follows:

1). The data used amounted to 500 data

2). Conducting Pre-Processing data by using data transformation method that is manipulating raw data to produce single input. The goal is to be more efficient in the process of data mining and so that the resulting pattern is more easily understood. Attributes of pre-processing data are student status and gender. The status of student work is changed to 1 and student status changed to 0. Male gender changed to 1 and gender perempun changed to 0. Then done normalization with the formula:

$$newdata = \frac{(data - min) * (newmax - newmin)}{(max - \min) + newmin} \quad (3)$$

3). Calculation of K-Nearest Neighbor for example of raw data. which can be seen in the Table 1.

Table 1 Example of raw data

NO	Status Mahasiswa	Jenis kelamin	IPS 1	IPS 2	IPS 3	IPS 4	KELULUSAN
1	Bekerja	Perempuan	2,76	2,8	3,2	3,17	TERLAMBAT
2	Mahasiswa	Perempuan	3	3,3	3,14	3,14	TEPAT
3	Bekerja	Perempuan	3,5	3,3	3,7	3,29	TEPAT
4	Mahasiswa	Perempuan	3,17	3,41	3,61	3,36	TEPAT
5	Bekerja	Perempuan	2,9	2,89	3,3	2,85	TEPAT
6	Bekerja	Laki-Laki	2,95	2,82	3,09	3,1	TEPAT
7	Mahasiswa	Perempuan	2,76	3,14	2,6	2,95	TERLAMBAT
8	Bekerja	Perempuan	2,62	2,89	2,32	2,5	TERLAMBAT
9	Bekerja	Perempuan	3,6	3,54	3,52	3,39	TEPAT
10	Bekerja	Perempuan	2,71	2,55	1,77	2,11	TERLAMBAT
11	Bekerja	Perempuan	3,14	3,46	3,4	3,43	TEPAT
12	Bekerja	Perempuan	2,67	2,3	1,57	1,44	TERLAMBAT
13	Bekerja	Perempuan	2,57	2,82	2,2	2,45	TERLAMBAT
14	Bekerja	Perempuan	2,71	3	2,65	2,27	TERLAMBAT
15	Mahasiswa	Perempuan	3,24	3,38	3,44	3,3	TEPAT
16	Mahasiswa	Perempuan	2,7	2,8	2,9	2,91	?

Here are the steps of the K-Nearest Neighbor algorithm:

a. Determine the parameter K, for example K = 5

b. Calculates the euclidean distance of each object against the given sample data. the calculation results can be seen on Table 2. Column name at Table 2, means calculation of distance between d1 (data 1) with d16 (data 16 predicted)

Table 2. Calculation distance results

No	Nama	Disntace
1	d1,d16	0,9
2	d2,d16	0,67
3	d3,d16	1,52
4	d4,d16	1,14
5	d5,d16	0,92
6	d6,d16	1,19
7	d7,d16	0,46
8	d8,d16	0,72
9	d9,d16	1,62
10	d10,d16	1,62
11	d11,d16	1,34
12	d12,d16	2,2
13	d13,d16	1,17
14	d14,d16	1,07
15	d15,d16	1,04

c. Sort those objects into groups that have the smallest euclidean distances.

Table 3. Ascending sorting results

No	Nama	Disntace	Rangking
7	d2,d16	0,46	1
2	d5,d16	0,67	2
8	d11,d16	0,72	3
1	d1,d16	0,9	4
5	d6,d16	0,92	5
15	d12,d16	1,04	6
14	d10,d16	1,07	7
4	d13,d16	1,14	8
13	d9,d16	1,17	9
6	d7,d16	1,19	10
11	d1,d16	1,34	11
3	d14,d16	1,52	12
9	d8,d16	1,62	13
10	d3,d16	1,62	14
12	d4,d16	2,2	15

d. Determine k classification (k best based on rank)

Table 4. Fifth best based on Smallest Distance

No	Nama	Disntace	Rangking
7	d2,d16	0,46	1
2	d5,d16	0,67	2
8	d11,d16	0,72	3
1	d1,d16	0,9	4
5	d6,d16	0,92	5

e. By using the majority category, it can be classification results.

Table 5. Classification results

No	Nama	Disntace	Rangking	Kelulusan
7	d2,d16	0,46	1	TERLAMBAT
2	d5,d16	0,67	2	TEPAT
8	d11,d16	0,72	3	TERLAMBAT
1	d1,d16	0,9	4	TERLAMBAT
5	d6,d16	0,92	5	TEPAT

From the Figure 5 can be concluded the number of late as much as 3 and not late 2, so that the data in the test included in the category “TERLAMBAT” (LATE).

4. Implementation and Result

Applications built using the JAVA language and MySQL database. there are 500 data used for KNN calculations in predicting student graduation. testing using K-fold validation with K = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19.

4.1 Implementation of Application

This application has some functionality including: login system, data input and data classification

4.1.1 Login system

This login system uses the username and password of the user, if the user does not have permissions then he can not use this application.

Figure 1. Login system

4.1.2 Input Data

On this functionality, new data can be entered into the system. student data taken are: work status, gender, ip semester 1 to 4 and graduation. in addition to adding data on this form can manipulate data, both change and delete data. This form can been seen at figure 2.

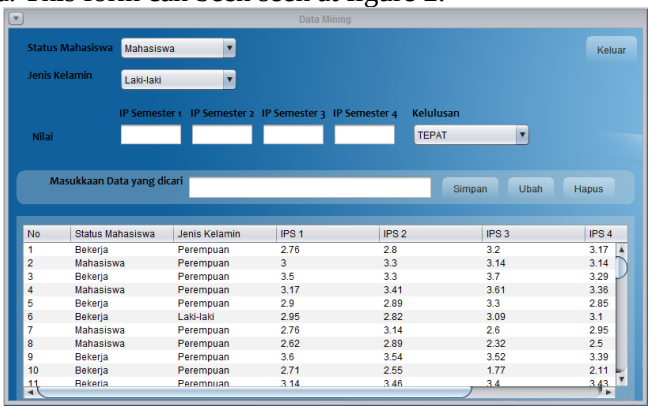

Figure 2. Form Manipulate Data

4.1.3 Classification using KNN

The timely or late graduation prediction function using the KNN algorithm can be seen in the figure 3.

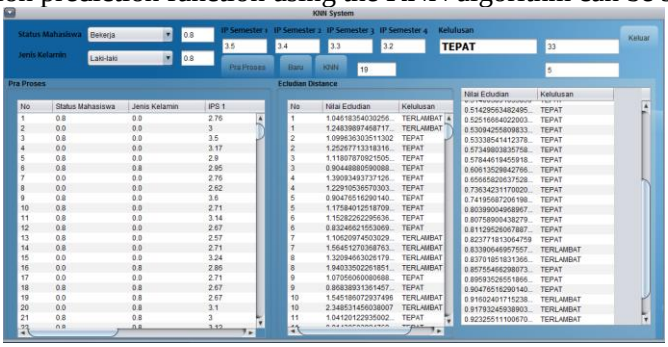

Figure 3. Form Manipulate Data

4.2 Testing using K-Fold Validation

The k-closeest validation algorithm is performed by k-fold cross validation to determine the average success of a system by looping by scrambling the input attribute so that the system is tested for some random input attribute. In cross validation we must specify the number of partitions or folds, the usual and famous standard used to obtain the best error estimation is 10 times the partition or tenfold cross validation. Training data amounting to 500 divided into 10 equal parts that is 50 pieces of data each piece of data testing. Each data testing is classified using the k-nearest neighbor algorithm by entering

the value of each variable. The k-nearest neighbor classification results are compared with real data and the correct number of classifications is calculated. The high degree of accuracy is chosen to be the optimal k value.

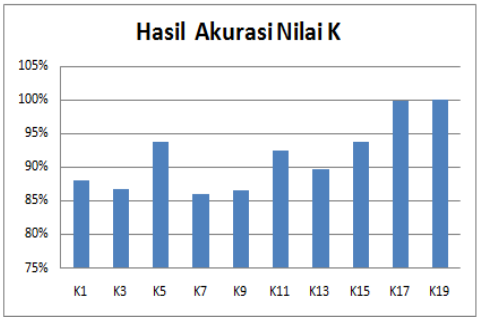

Figure 4. Result K-Value of K-Fold Validation

From 10 experiments conducted got the highest accuracy value that is k19 100% accuracy and that as k-optimal value in this research using 500 data. the results of this experiment can be seen in the figure4.

5. Conclusion

Based on the research result of Classification of Old Study Student Using K-Nearest Neighbor Algorithm, hence can be drawn conclusion as follows:

1) In the classification of students' graduation using the KNN algorithm based on student status, gender, ip semester 1 - ip semester 4 with k-fold cross validation in can the average value of K1 accuracy 88%, K3 accuracy 88.67%, K5 accuracy 93 , 78%, K7 86% accuracy, K9 accuracy 86.22%, K11 accuracy 92.44%, K13 accuracy 89.55%, K15 accuracy 93.78%, K17 accuracy 99.78%, and K19 accuracy 100%.
2) The optimal K value for the 500 accuracy results using the k-Nearest Neighbor algorithm is K19 with 100% accuracy.
3) From 500 training data in the status of student work 188, student 312, hence student status work longer in finish college and in can gender 290 male, woman 210, hence woman longer in finish college.

6. Acknowledgment

We would like to thank DRPM Kemenristekdikti for financial support for this research

References

[1] Kamagi, D.H. dan Hansun, S., 2014, *Implementasi Data Mining Dengan Algoritma C4.5 Untuk Memprediksi Tingkat Kelulusan Mahasiswa,* Tanggerang.

[2] Banjarsari, Mutiara A. 2015. *Pencarian k-Optimal pada Algoritma KNN untuk prediksi Kelulusan Tepat Waktu Mahasiswa Berdasarkan IP Sampai dengan Semester 4*. FMIPA Unlam : Banjarbaru.

[3] Buku VI *Matriks Penilaian Instrumen Akreditasi Program Studi Sarjana*, Badan Akreditasi Nasional Perguruan Tinggi, 2011.

[4] Ginting, S.L.,.dkk.,2014, *Teknik data mining untuk memprediksi masa studi mahasiswa menggunakan algoritma K-Nearest Neighborhood*, **Volume 3**, No.2.

[5] Pramudiono, l., 2007. *Pengantar Data Mining : Menambang Permata Pengetahuan di gunung data*.https://www.academia.edu/10378211/Kuliah_Pengantar_Data_Mining_Menambang_Permata_Pengetahuan_di_Gunung_Data, Diakses pada tanggal 15 September 2016

[6] Gorunescu, Florin. 2011. *Data Mining: Concepts, Models and Techniques. Computational Intelligence and Complexity.* (Verlag Berlin Heidelberg-Springer), **Volume 12**.pp. 1-43.

[7] Rizal, Azwar. 2013. *Perbandingan Performa antara Imputasi Metode Konvensional dan Imputasi dengan Algoritma Mutual Nearest Neighbor*. Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember. Surabaya.

[8] Pandie, Emerensye S. Y. 2012. *Implemetasi Algoritma Data Mining K-Nearest Neighbor (KNN) Dalam Pengambilan Keputusan Pengajuan Kredit*. Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana : Kupang