Journal of Engineering and Applied Technology
Vol. 2, No. 2, August 2021, pp. 86-91

E-ISSN: 2716-2265
P-ISSN: 2716-2257

K-Nearest Neighbor (K-NN) algorithm with Euclidean and Manhattan in classification of student graduation

Nur Hidayati[1], Arief Hermawan [2,*]

[1] Magister Program of Information Techology, Technology University of Yogyakarta, Jl. Siliwangi (Ringroad Utara), Jombor, Sleman, D. I Yogyakarta, 55285, Indonesia
[2] Universitas Teknologi Yogyakarta, Jl. Siliwangi (Ringroad Utara), Jombor, Sleman, D. I Yogyakarta, 55285, Indonesia
E-mail: ariefdb@uty.ac.id *
* Corresponding Author

ABSTRACT

K-Nearest Neighbor (K-NN) algorithm is a classification algorithm that has been proven to solve various classification problems. Two approaches that can be used in this algorithm are K-NN with Euclidean and K-NN with Manhattan. The research aimed to apply the K-NN algorithm with Euclidean and K-NN with Manhattan to classify the accuracy of graduation. Student graduation was determined by the variables of gender, major, number of first-semester credits, number of second-semester credits, number of third-semester credits, grade point on the first semester, grade point on the second semester, grade point on the third semester, and age. These variables determined the accuracy of student graduation, timely or untimely. The implementation of the K-NN algorithm was carried out using Rapidminer software. The results were obtained after testing 380 training data and 163 testing data. The best accuracy system was achieved at K=7 with a value of 85.28%. The two algorithmic approaches did not affect the accuracy of the results. Furthermore, the addition of the value of K did not completely affect the accuracy.

This is an open-access article under the CC–BY-SA license.

ARTICLE INFO

Article history
Received:
31 July 2021
Revised:
16 August 2021
Accepted:
18 August 2021

Keywords
Euclidean
Manhattan
Classification
Graduation

1. Introduction

Student graduation rate is one of the indicators of the success of higher education. To achieve a proper graduation rate, universities must plan the learning process so that students can timely graduate [1]. The development of information technology can be used by universities to process data rapidly and accurately [2]. One of the benefits of using information technology is its use to predict student graduation [3]–[6]. Prediction of student graduation can be carried out by using student data in the first year. Prediction of student graduation can be further applied to assist universities in evaluating and improving the learning system that universities can produce qualified and timely graduates [7].

Prediction of student graduation can be conducted by classifying student graduation. One of the algorithms that can be utilized to classify is K-Nearest Neighbor (K-NN) using Euclidean and Manhattan Distance. To solve the problem of predicting student study time, the Euclidean Distance method can predict study time with an accuracy of 85.71% at K=10 [8]. In addition, to predict graduation based on Tryout scores, the Manhattan Distance method is proven to perform

with an accuracy of 97.30%. The accuracy is obtained when K=3 [9]. For the problem of predicting student graduation time, KNN with Euclidean Distance and Manhattan Distance can predict with an accuracy of 82.26% [10]. The KNN method with Euclidean can make predictions with an accuracy of 83% at K=10 [11]. Moreover, to predict the qualification of the National Examination, the Euclidean Distance method can perform with the accuracy of 88.42% with K=7 [4]. It is in line with the result of a study in SMA Negeri 12 Tangerang, the Euclidean Distance method can perform with an accuracy of 89.126% with K = 7 in predicting the qualification of the National Examination.

Based on the findings, it can be inferred that the accuracy value is highly dependent on the K value and the problem being solved. The research aims to build a student graduation prediction system using Euclidean and Manhattan. The variables used as input are gender, major, number of credits for semester one, number of credits for semester two, number of credits for semester three, grade point on semester one, grade point on semester two, grade point on semester three, age, and graduation status (timely/untimely). This research is expected to be able to provide benefits for universities to formulate policies to secure students can graduate properly [12]. The highest level of accuracy of 98.5 percent was attained when k = 3 according to the results of prediction testing on 60 data for students in 2015-2016. The accuracy of the K-Nearest Neighbor algorithm calculation is also improved when more samples and training data are used. [13]. 240 student scores were used to test the algorithm's performance. These 240 students have graduated, and the cluster is labeled based on their graduation dates. There are 7 clusters with a silhouette value of 0.2416 as an outcome. The range of student graduation times is used to designate each cluster. The variance in each cluster is attributable to the presence of students with similar scores in the majority but varying graduation times. Other factors influencing the range of graduation times in each cluster include academic leave or extending the thesis completion period. The average prediction accuracy of 99.58 [14] is obtained by k-folding 240 data into 5 subsets. Predicting student graduation based on 667 tests completed by the author of the training data. In the first test, with a value of k = 1, records had the maximum accuracy of 88.16 percent. [15].

2. Method

2.1. Data Set

This study used students' data from the Informatics Department at the University of Technology Yogyakarta in the academic year of 2014 and 2015. The attributes and data types are presented in Table 1.

Table 1. Data set

Number	Column	Type of Data
1	Gender	Integer
2	Origin of Schools/ Major	Integer
3	Credits 1	Real
4	Credits 2	Real
5	Credits 3	Real
6	Grade Point on Semester 1	Real
7	Grade Point on Semester 2	Real
8	Grade Point on Semester 3	Real
9	Age	Integer
10	Graduation status	Timely/untimely

The data used was 543 students, consisting of 444 male students and 99 female students. The number of students who timely graduated was 83 students and 460 students whom untimely graduated. 380 students were used as training data, while 163 students were used as testing data.

2.2. Research Procedure

To achieve the research objectives, the following steps were carried out:

(1) The data were divided into 2 groups

The first group was used for training, i.e. 380 data were used as training data, namely 57 students who timely graduated and 323 students who untimely graduated. The second group was used for student testing. 163 students were used as testing data; 31 students who timely graduated and 137 students who untimely graduated. The first group was used as training data for developing the model, and the second group was used for testing data.

(2) Developing a model using group A and testing the model using group B.

Rapidminer software was utilized to develop and test the model. There were 2 types of models used, namely K-NN with Euclidean and K-NN with Manhattan. K-NN with Euclidean was designed liberary by setting up the K-NN Numerical Measure menu as shown in Fig.1 and Fig. 2. Model testing that has been formed is carried out by performing the function parameter division of 70% for training data and 30% for testing data, as presented in Fig.3.

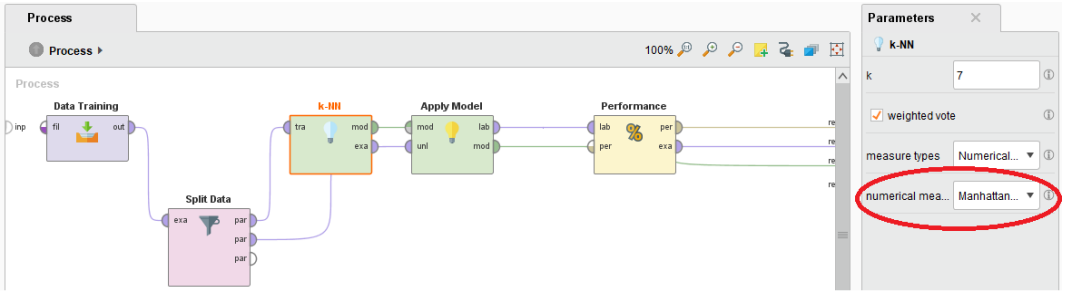

Fig. 1. Manhattan

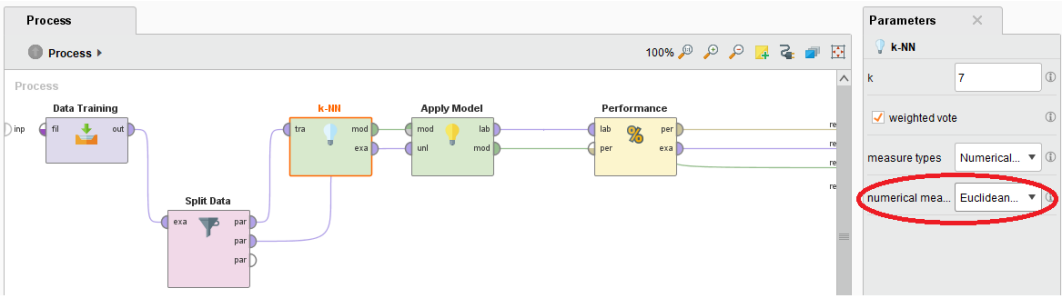

Fig. 2. Euclidean

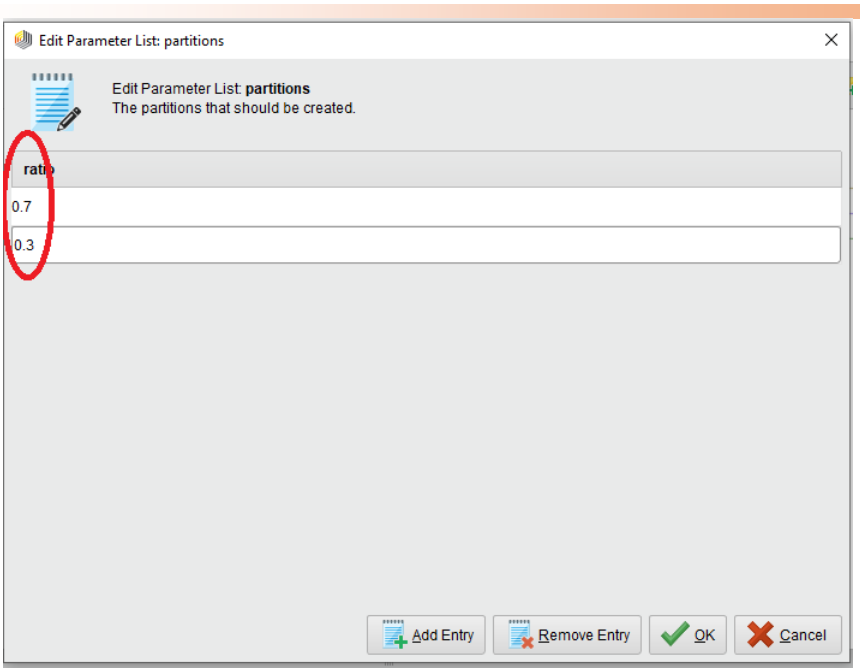

Fig. 3. Parameter list

The model development was carried out for the values of K=1, K=3, K=7, K=9, K=11. More details can be seen in Fig. 4.

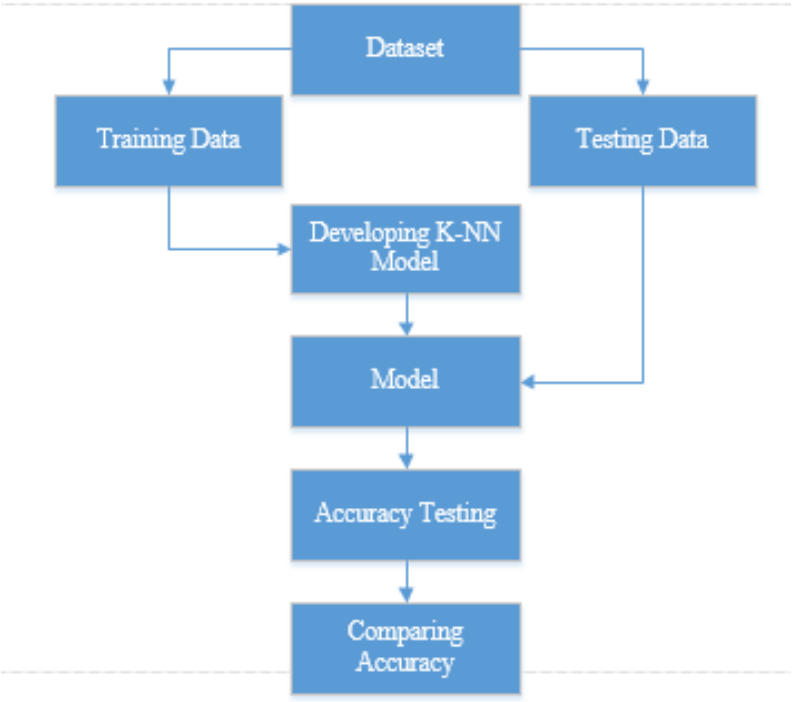

Fig. 4. Research procedure

3. Results and Discussion

After carrying out all steps on the research procedure, the results were obtained as shown in Table 2.

Table 2 . Testing result

Method	Accuracy					
	K=1	***K=3***	***K=5***	***K=7***	***K=9***	***K=11***
Euclidean	77.91%	82.82%	84.66%	85.28%	84.66%	84.66%
Manhattan	76.07%	82.21%	83.44%	85.28%	85.28%	84.66%

Based on Table 2, it can be inferred that the use of Euclidean and Manhattan Distance methods for classifying student graduation obtained the highest accuracy of 85.28% at K=7. These results indicated that the use of the K-NN algorithm with Euclidean and Manhattan Distance did not affect the classification accuracy. Moreover, from these results, it can be figured out that the distribution of student data did not affect the distance from the group separated by the Euclidean and Manhattan algorithms. In addition, it can be seen in table 3 that the addition of the value of k was not entirely beneficial for increasing accuracy, K=7 is the maximum value.

4. Conclusion

In this research, the use of Euclidean and Manhattan methods did not affect prediction accuracy. The highest prediction accuracy for the two methods is at K=7, which is 85.28%. The difference in distance calculation from the Euclidean and Manhattan Distance methods did not affect the results of the data classification. Furthermore, it was also found that the addition of the K value was not fully beneficial in affecting the accuracy value, the value of K=7 was the maximum value.

References

[1] M. Masrizal and A. Hadiansa, "Prediksi jumlah lulusan mahasiswa STMIK Dumai menggunakan jaringan syaraf tiruan," *Informatika*, vol. 9, no. 2, p. 9, 2019, doi: 10.36723/juri.v9i2.98.

[2] I. Vhallah, S. Sumijan, and J. Santony, "Pengelompokan mahasiswa potensial drop out menggunakan metode Clustering K-Means," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 2, no. 2, pp. 572–577, 2018, doi: 10.29207/resti.v2i2.308.

[3] T. Asril and S. M. Isa, "Prediction of students study period using K-Nearest Neighbor algorithm," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 6, pp. 2585–2593, Jun. 2020, doi: 10.30534/IJETER/2020/60862020.

[4] S. Mulyati, S. M. Husein, and R. Ramdhan, "Rancang bangun aplikasi data mining prediksi kelulusan ujian nasional menggunakan Algoritma (Knn) K-Nearest Neighbor dengan metode Euclidean Distance pada SMPN 2 Pagedangan," *JIKA (Jurnal Informatika)*, vol. 4, no. 1, p. 65, 2020, doi: 10.31000/jika.v4i1.2288.

[5] M. B. Musthafa, N. Ngatmari, C. Rahmad, R. A. Asmara, and F. Rahutomo, "Evaluation of university accreditation prediction system," *IOP Conference Series: Materials Science and Engineering*, vol. 732, no. 1, p. 012041, Jan. 2020, doi: 10.1088/1757-899X/732/1/012041.

[6] A. P. Salim, K. A. Laksitowening, and I. Asror, "Time series prediction on college graduation using KNN algorithm," *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*, Jun. 2020, doi: 10.1109/ICOICT49345.2020.9166238.

[7] B. A. Arifiyani and R. S. Samosir, "Sistem simulasi prediksi profil kelulusan mahasiswa dengan Decison Tree Bekti," *Jurnal Sains dan Teknologi Kalbi Scientia*, vol. 5, no. 2, pp. 115–123, 2018.

[8] H. E. Wahanani, M. H. Prami Swari, and F. A. Akbar, "Case based reasoning prediksi

waktu studi mahasiswa menggunakan metode Euclidean Distance dan normalisasi Min-Max," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 7, no. 6, p. 1279, 2020, doi: 10.25126/jtiik.2020763880.

[9] M. Erlangga Dwi Kurniawan, "Implementasi algoritma K-Nearest Neighbor dengan metode klasifikasi dan pengukuran jarak Manhattan Distance untuk prediksi kelulusan UN berdasarkan hasil nilai tryout berbasis Java Desktop pada SMA Harapan Jaya 2," *Skanika*, vol. 1, no. 1, pp. 76–81, 2018.

[10] F. E. Prabowo and A. Kodar, "Analisis prediksi masa studi mahasiswa menggunakan Algoritma Naïve Bayes," *Jurnal Ilmu Teknik dan Komputer*, vol. 3, no. 2, p. 147, 2019, doi: 10.22441/jitkom.2020.v3.i2.008.

[11] D. Z. Abidin, S. Nurmaini, and R. F. Malik, "Penerapan metode K-Nearest Neighbor dalam memprediksi masa studi mahasiswa (Studi kasus : mahasiswa STIKOM Dinamika Bangsa)," in *Prosiding Annual Research Seminar*, 2017, vol. 3, no. 1, pp. 133–138.

[12] P. Y. Santoso and D. Kusumaningsih, "Algoritma K-nearest Neighbor dengan menggunakan metode Euclidean Distance untuk memprediksi kelulusan ujian nasional berbasis desktop SMA Negeri 12 Tangerang," *Skanika 2018*, vol. 1, no. 1, pp. 123–129, 2018.

[13] R. Muliono, J. H. Lubis, and N. Khairina, "Analysis K-Nearest Neighbor Algorithm for improving prediction student graduation time," *SinkrOn*, vol. 4, no. 2, p. 42, 2020, doi: 10.33395/sinkron.v4i2.10480.

[14] L. Cahaya, L. Hiryanto, and T. Handhayani, "Student graduation time prediction using intelligent K-Medoids Algorithm," in *Proceeding of 2017 3rd International Conference on Science in Information Technology: Theory and Application of IT for Education, Industry and Society in Big Data Era, ICSITech 2017*, 2017, vol. 2018-Janua, pp. 263–266, doi: 10.1109/ICSITech.2017.8257122.

[15] M. Imron and S. A. Kusumah, "Application of data mining classification method for student graduation prediction using K-Nearest Neighbor (K-NN) Algorithm," *IJIIS: International Journal of Informatics and Information Systems*, vol. 1, no. 1, pp. 1–8, 2018, doi: 10.47738/ijiis.v1i1.17.