SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10480
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Analysis K-Nearest Neighbor Algorithm for Improving Prediction Student Graduation Time

Rizki Muliono
Universitas Medan Area
Medan, Indonesia
rizkimuliono@gmail.com

Juanda Hakim Lubis
Universitas Medan Area
Medan, Indonesia
juandahakim@gmail.com

Nurul Khairina
Universitas Medan Area
Medan, Indonesia
nurulkhairina27@gmail.com

Submitted: Jan 30, 2020
Accepted: Mar 11, 2020
Published: Apr 1, 2020

Abstract **—** Higher education plays a major role in improving the quality of education in Indonesia. The BAN-PT institution established by the government has a standard of higher education accreditation and study program accreditation. With the 4.0-based accreditation instrument, it encourages university leaders to improve the quality and quality of their education. One indicator that determines the accreditation of study programs is the timely graduation of students. This study uses the K-Nearest Neighbor algorithm to predict student graduation times. Students' GPA at the time of the seventh semester will be used as training data, and data of students who graduate are used as sample data. K-Nearest Neighbor works in accordance with the given sample data. The results of prediction testing on 60 data for students of 2015-2016, obtained the highest level of accuracy of 98.5% can be achieved when k = 3. Prediction results depend on the pattern of data entered, the more samples and training data used, the calculation of the K-Nearest Neighbor algorithm is also more accurate.

Keywords — prediction; graduation time; k-nearest neighbor

I. INTRODUCTION

In the Study Program and Higher Education Accreditation Forms, the timely graduation of undergraduate students is one component that has influence (Novianti & Prasetyo, 2017). To get good grades in accreditation, students are targeted to graduate on time and achieve an average Semester Achievement Index above 3.50.

According to the graduation data of the Faculty of Engineering, Universitas Medan Area in recent years, the average number of students who complete their studies on time has not yet reached the target. Some problems that often occur that cause students to graduate on time, including low Semester Achievement Index and GPA scores, economic factors, environmental factors, and family.

Prediction of students will graduate on time or not can be noticed since students sit in the seventh semester. Semester Achievement Index sand the number of credits will be a reference to predict the time students graduate.

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10480
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Data mining is one of the fields of computer science that focuses on machine learning (Muliono, Muhathir, Khairina, & Harahap, 2019) (Muliono, 2017). Data mining is used to predict conditions based on data and information (Tang, He, & Zhang, 2020) (Muliono & Sembiring, 2019). The K-Nearest Neighbor method uses data classification techniques that are divided into clusters (Agrawal, 2019). Prediction results can be calculated based on the distance closest to the sample data (Gou et al., 2019) (Czumaj & Sohler, 2020). This research will predict the graduation time of the students with the K-Nearest Neighbor algorithm. As for some previous studies related to this research are as follows:

The Research by (Prasetyo, Kusrini, & Arief, 2019) uses the K-Nearest Neighbor algorithm to see the interests and talents of students in the field of Information Engineering. This choice of specialization is done by Case Base Reasoning (CBR). The results showed that this algorithm successfully predicted with an accuracy rate of 95.98% at K = 7.

The Research by (Nikmatun & Waspada, 2019) applies the K-Nearest Neighbor algorithm that refers to Data Mining Knowledge Discovery in Database (KDD). This study classifies courses that determine the time students graduate. The research results obtained a good prediction with an accuracy of 75.95%.

The research by (Hakim, Rizal, & Ratnasari, 2019) uses the K-Nearest Neighbor algorithm and Roger S. Pressman's waterfall method namely Communication, Planning, Modeling, and Construction. The results showed that the best accuracy was found in testing with the Confusion Matrix, where the accuracy reached 98%.

The Research by (Rohman & Rochcham, 2019) compares Neural Network, K-Nearest Neighbor and Decision Tree algorithms in predicting student graduation. The results showed that the highest accuracy was found in the K-Nearest Neighbor algorithm which reached 83.66%.

The research by (Purwanto, Kusrini, & Sudarmawan, 2019) made a comparison of the C.45 algorithm and the K-Nearest Neighbor in predicting the study period of students of Muhammadiyah University in Purwokerto. The results showed that the highest accuracy was found in the K-Nearest Neighbor algorithm which reached 89.14%.

II. METHODOLOGY

K-Nearest Neighbor algorithm is a classification method that can classify new data based on the distance of the new data to the closest data/neighbors in data learning (Atma & Setyanto, 2018) :

The training process is to start input: training data, data transfer label, k, testing data.

a. For all testing data, calculate the distance to each training data
b. Determine the training data k which is the closest distance to the data
c. Testing
d. Check the label of this data
e. Determine the label with the most frequency
f. Enter the testing data to the class with the most frequency
g. Stop

To calculate the distance between two points x and y, you can use the Euclidean distance as follows (Wang et al., 2019)

$$d(X_1, Y_2) = \Sigma_I \left| \frac{n_{1i}}{n_1} - \frac{n_{2i}}{n_2} \right| \quad (1)$$

Which X_1, 1 = 1, 2, is the category attribute, and $n_{1j} - n_{2i}$ represents the corresponding frequency. The closeness between the two cases can be calculated by finding the value of similarity as follows (Rahmatullah & Utami, 2019)

$$similarity\ (T,S) = \frac{\sum_{i=1}^{n} f(T_i,S_i)*w_i}{w_i} \quad (2)$$

Description :
q: new case
s: cases that are in deviation
n: number of attributes in each case
i: individual attributes between 1 to n
f: similarity function I between cases T and S
wi: the weight is given to the i-th attribute

This similarity is expressed by 1 (similar) and 0 (not similar), mathematically, it can be written:

$$s = \begin{matrix} \mathbf{1}\ \boldsymbol{if}\ \boldsymbol{x = y} \\ \mathbf{0}\ \boldsymbol{if}\ \boldsymbol{x \neq y} \end{matrix} \quad (3)$$

Giving weights for each attribute can be done by following a few steps below:
1. Input the criteria value of each model (LCS)
2. Input the weights of each criterion (BBT)
3. Calculate normalization from weights (NK)

$$NK = \frac{\sum_{i=1}^{n}(SBK)xBBT\%}{n} \quad (4)$$

$$Value = \frac{\Sigma\, NK}{N} \quad (5)$$

To test the accuracy of the predicted performance measurement of the K-NN algorithm, it is performed by

(6)

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10480
e-ISSN : 2541-2019
p-ISSN : 2541-044X

comparing the results of the classification algorithm prediction with the target value of the testing data variable as the actual data. So logically, it can be concluded that the performance of the algorithm is as follows:

$$Accuracy = \frac{the\ predicted\ amount\ is\ correct}{total\ number\ of\ instances}\ x\ 100\%$$

III. RESULT AND DISCUSSION

The research test was conducted on the data of 20 students of the Civil Engineering Study Program in the seventh semester of 2015. Detailed research results can be seen as follows:

Table 1. Accuracy Comparison of Predictive Predicate Values with KNN and Real Data K = 5

No	NPM	Prediction KNN	Result	Point
1	158110001	Correct	Correct	1
2	158110002	Incorrect	Incorrect	1
3	158110003	Incorrect	Incorrect	1
4	158110005	Incorrect	Incorrect	1
5	158110006	Correct	Correct	1
6	158110007	Incorrect	Incorrect	1
7	158110010	Correct	Correct	1
8	158110012	Incorrect	Incorrect	1
9	158110015	Incorrect	Incorrect	1
10	158110017	Correct	Correct	1
11	158110018	Incorrect	Incorrect	1
12	158110020	Incorrect	Correct	0
13	158110022	Correct	Correct	1
14	158110023	Incorrect	Incorrect	1
15	158110024	Incorrect	Incorrect	1
16	158110025	Correct	Correct	1
17	158110027	Correct	Correct	1
18	158110028	Incorrect	Incorrect	1
19	158110029	Incorrect	Incorrect	1
20	158110030	Incorrect	Incorrect	1

From the results of experiments conducted to see the accuracy of the comparison of training data to the results of algorithms found the results with timely conclusions at K1 and K2 = 176, while K3-K5 = 197.

Table 2. Accuracy Levels

K	Accuracy Levels	Accuracy Levels	Conclusion
	Confusion Matrix	ROC Curve	
K1	88,0%	0.880	Good Classification
K2	88,0%	0.880	Good Classification
K3	98,5%	0,985	Excellent Classification
K4	98,5%	0,985	Excellent Classification
K5	98,5%	0,985	Excellent Classification

The higher the K value, the better the accuracy level of K-NN algorithm predictions on 2015 student training data, the conclusion is from the K1-K5 trial results of K-NN algorithm classification results in the accuracy of student graduation prediction by comparing the Rael scores and the prediction results can be concluded as Excellent Classification

Next is to make a prediction on time for the 2016 data of the 2016 students' whip, which will be tested from K1-K5.

Table 3. Predicted Results for 2016 Stock Data of Civil Engineering Study Program

NPM	SAI 1	SAI 2	SAI 3	SAI 4	SAI 5	SAI 6	SAI 7	SKS Passed	Prediction
16811 0003	3.25	3.29	3.82	3.53	3.53	3.82	3.53	136	Correct
16811 0005	3.61	3.47	2.04	1.75	2.67	2.55	3.37	122	Incorrect
16811 0009	3.38	3.13	2.55	3.37	3.05	3.61	3.29	132	Incorrect
16811 0011	3.39	3.13	3.47	2.95	2.67	3.82	2.55	132	Incorrect
16811 0012	3.61	3.03	3.37	3.42	3.47	3.76	2.67	136	Correct
16811 0016	3.29	3.71	2.88	3.05	3.53	2.88	3.05	132	Incorrect
16811 0017	3.24	3.13	3.61	3.29	2.55	3.05	3.53	132	Incorrect
16811 0022	3.71	3.29	3.05	3.53	3.05	3.82	3.53	136	Correct
16811 0026	3.61	3.47	2.04	1.75	2.67	2.55	3.37	122	Incorrect
16811 0028	3.61	3.47	2.04	1.75	2.67	2.55	3.37	122	Incorrect

SinkrOn : Jurnal dan Penelitian Teknik Informatika
Volume 4, Number 2, April 2020
DOI : https://doi.org/10.33395/sinkron.v4i2.10480
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The sample data of 200 data consists of the 2015 data stick and the data to be predicted is the 2016 data canopy of 60 data with a ratio of 70% training data and 30% testing data. From the results of prediction experiments on 2016 data, there are 60 data with timely prediction results that can be seen in the following table:

Table 4. Predicted Results of 2016 Whamb Graduation

K	Correct	Incorrect
K1	25	35
K2	25	35
K3	16	44
K4	16	44
K5	16	44

IV. CONCLUSION AND SUGGESTION

A. *Conclusion*

The conclusions of this study are as follows:

1. In the case of predictions of 2015 student data on the whip of K-NN algorithm the better level of K3 and so on is 98.5% from the previous K 88% increased by 1.5%
2. In predictions, 60 of the 2016 canopy data shows the condition of predicted data at K1 and K2 = 25 On-Time, while at K3 - K5 = 16 On Time.
3. The state of the predicted results depends on the distribution of data patterns, the more data the better the calculation of the K-NN algorithm
4. The more data the application transfers, the less time it takes to process distance calculations.

B. *Suggestion*

It is hoped that the K-NN prediction application based on web-based research results can be used by the Faculty to assist monitoring as an EWS (Early Warning System) for the academic development of students in the Faculty of Engineering, Universitas Medan Area in particular.

V. ACKNOWLEDGMENT

The researcher would like to thank the Universitas Medan Area for funding the DIYA Research to completion. Hopefully, this research is not only useful for the Universitas Medan Area but also can be useful for the development of science and society.

VI. REFERENCES

Agrawal, R. (2019). Integrated Parallel K-Nearest Neighbor Algorithm. In *Smart Intelligent Computing and Applications* (p. 479). Springer Singapore. https://doi.org/10.1007/978-981-13-1921-1

Atma, Y. D., & Setyanto, A. (2018). Perbandingan Algoritma C4.5 dan K-NN dalam Identifikasi Mahasiswa Berpotensi Drop Out. *Metik Jurnal*, *2*(2), 31–37.

Czumaj, A., & Sohler, C. (2020). Sublinear time approximation of the cost of a metric k -nearest. In *Society for Industrial and Applied Mathematics* (pp. 2973–2992).

Gou, J., Ma, H., Ou, W., Zeng, S., Rao, Y., & Yang, H. (2019). A Generalized Mean Distance-Based K-Nearest Neighbor Classifier. *Expert Systems with Applications*, *115*, 3–24. https://doi.org/10.1016/j.eswa.2018.08.021

Hakim, L. A. R., Rizal, A. A., & Ratnasari, D. (2019). Aplikasi Prediksi Kelulusan Mahasiswa Berbasis K-Nearest Neighbor (K-NN). *JTIM : Jurnal Teknologi Informasi Dan Multimedia*, *1*(1), 30–36. https://doi.org/10.35746/jtim.v1i1.11

Muliono, R. (2017). Implementasi Algoritma Apriori Pada Data Benchmark Kosarak Dan Mushrooms. *Journal of Informatics and Telecommunication Engineering*, *1*(1), 34–41.

Muliono, R., Muhathir, Khairina, N., & Harahap, M. K. (2019). Analysis of Frequent Itemsets Mining Algorithm Againts Models of Different Datasets. In *1st International Conference of SNIKOM 2018* (pp. 1–5). https://doi.org/10.1088/1742-6596/1361/1/012036

Muliono, R., & Sembiring, Z. (2019). Data Mining Clustering Menggunakan Algoritma K-Means Untuk Klasterisasi Tingkat Tridarma Pengajaran Dosen. *CESS (Journal Of Computer Engineering, System And Science)*, *4*(2), 272–279.

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal Simetris*, *10*(2), 421–432.

Novianti, A. G., & Prasetyo, D. (2017). Penerapan Algoritma K-Nearest Neighbor (K-NN) untuk Prediksi Waktu Kelulusan Mahasiswa. In *Seminar Nasional APTIKOM(SEMNASTIKOM)* (pp. 108–113).

Prasetyo, A., Kusrini, & Arief, M. R. (2019). Penerapan Algoritma K Nearest Neighbor untuk Rekomendasi Minat Konsentrasi di Program Studi Teknik Informatika Universtias PGRI Yogyakarta. *Informasi Interaktif*, *4*(1), 1–6.

Purwanto, E., Kusrini, & Sudarmawan. (2019). Prediksi Kelulusan Tepat Waktu Menggunakan Metode C4 . 5 DAN K - NN (Studi Kasus : Mahasiswa

Program Studi S1 Ilmu Farmasi , Fakultas Universitas Muhammadiyah Purwokerto). *TECHNO*, *20*(2), 131–142.

Rahmatullah, S., & Utami, E. (2019). Prediksi Tingkat Kelulusan Tepat Waktu dengan Metode Naive Bayes dan K-Nearest Neighbor. *Jurnal Informasi Dan Komputer*, *7*(1), 7–16.

Rohman, A., & Rochcham, M. (2019). Komparasi Metode Klasifikasi Data Mining untuk Prediksi Kelulusan Mahasiswa. *Jurnal Neo Teknika*, *5*(1), 23–31.

Tang, B., He, H., & Zhang, S. (2020). MCENN: A Variant of Extended Nearest Neighbor Method for Pattern Recognition. *Pattern Recognition Letters*, 1–10. https://doi.org/10.1016/j.patrec.2020.01.015

Wang, Y., Wang, R., Li, D., Adu-Gyamfi, D., Tian, K., & Zhu, Y. (2019). Improved Handwritten Digit Recognition using Quantum K-Nearest Neighbor Algorithm. *International Journal of Theoretical Physics*, *58*(7), 2331–2340. https://doi.org/10.1007/s10773-019-04124-5