The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm

To cite this article: D Kurniadi et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 434 012039

View the article online for updates and enhancements.
The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm

D Kurniadi1,2*, E Abdurachman2, H L H S Warnars2 and W Suparta2

1 Department of Informatics, Sekolah Tinggi Teknologi Garut, Jalan Mayor Syamsu 1, Garut 44151, Indonesia
2 Department of Computer Science, Graduate Program Doctor of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia

*dede.kurniadi@sttgarut.ac.id

Abstract. This article aims to implement the algorithm model of k-Nearest Neighbor (k-NN) in analyzing, predicting, and classifying students who have potentials to get scholarships in universities. The k-NN algorithm works by making a prediction based on the closest data points between the old data history as training data and the new data as testing data. The data collected totals 1018 students with 24 scholarship receiver candidate students are used as the dataset for the test purposes. The attributes used in the prediction process are a semester, parents’ income, number of family dependents, and Cumulative Grade Point Average. The distance calculation of the value from testing attribute to each training attribute uses Euclidean Distance equation, while the test of the model accuracy value is calculated using Confusion Matrix. The results of the simulation of the prediction model show that the determining factor of training data from both the number and the variation of different values can improve the performance of the k-NN algorithm with the best accuracy rate of 95.83 percent in predicting students who have the greatest chance of getting the scholarship.

1. Introduction

The success of education accomplishment in universities is not only influenced by factors related to the academic achievements of the students, but also another aspect such as economic aspect. In Indonesia, incapability of paying tuition is one of the most significant obstacles for university students to finish their study, especially in private universities, which are not run by the government. Therefore, there needs to be scholarship provided by such bodies as the government, companies, or foundations, to help those students.

Sekolah Tinggi Teknologi Garut, Garut School of Technology (later to be referred as STTG), is one of the private campuses whose active students, in the academic year of 2016/2017, are 1018 [1]. Of that number of those students, some of them come from a low economic and social status yet with a tremendous academic performance which makes them in need of a scholarship to help finish their study. In this relation, STTG annually distributes scholarship coming from different institutions such as the government, private companies, and foundations.

However, due to limited allocation of the scholarship, there is a gap between the number of scholarship available and the number of students in need. To cope with it, STTG sets a rigorous selection so that the scholarship goes to the right hand. In this case, the implementation of ICT can help make
decisions objectively, fast, and properly [2],[3]. The more significant the number of students proposing scholarship is, the more various the attributes of data of the students will get. Thus, the potentials of data digging are arising.

To help make decisions on the scholarship-receiving students, a supporting system namely Expert System can be implemented. It is a system using human knowledge recorded in the computer to solve problems usually taken care of by humans [4],[5],[6],[7]. There are a variety of techniques in Expert System; one of which is by doing classification [8],[9],[10]. In this technique, the most frequently-used calculation is k-Nearest Neighbor (k-NN) algorithm. The algorithm is a method using supervised algorithm [11],[12].

Data used in this study are data of STTG students who have ever been proposed to or those who have received the scholarship. The data themselves are called data training. Meanwhile, attributes used in the prediction process are a semester, parents’ income, number of people in the family, and GPA index. In the previous study, the k-NN algorithm was used to predict student academic achievement in universities; the results show that the accuracy of the k-NN was 82% [13].

The purpose of the study is to model k-NN algorithm to analyze, predict, and group students with potentials of receiving a scholarship in STTG. In addition, the algorithm is also used to predict the level of accuracy as one of the most precise models in predicting something [14], which, in this context, is related to the students with potentials to receive the scholarship. It is expected that in the future, this is a permanent system in selecting students to receive the scholarship so that related divisions in the university such as academic division and the division of student affairs can select eligible students.

2 Literature review

2.1. Scholarship

Simply put, the scholarship is a financial aid presented to someone (a student) aiming at helping them finish their study. Scholarship can be given by the government, private companies, or foundations [15]. In STTG, the types of scholarship distributed to the students are Beasiswa Peningkatan Prestasi Akademik (Academic Improvement Scholarship, BPPA), Provincial Aid Scholarship, and Bidikmisi Scholarship.

2.2. k-Nearest neighbor (k-NN)

K-Nearest Neighbor is a method using the supervised algorithm. It is an algorithm using classification towards objects based on the nearest data (to the objects) [12]. Following is the formula of distance search using Euclidian formula [16].

\[
d_i = \sqrt{\sum_{i=1}^{p} (x_{2i} - x_{1i})^2}
\]

(1)

Where: d: distance; p: data dimension; i: data variable; X1: sample data; X2: testing data.

In general, the processes of the k-NN algorithm are as follows.

a. Preparing sample data in the form of an array.
b. Preparing testing data in the form of an array.
c. Calculating the distance between attributive values of testing to each training using Euclidean Distance.
d. Sorting the distance results based on the lowest values and the predetermined number of neighbors.
e. Obtaining the prediction results based on the calculation of the highest number.
f. Calculating the accuracy based on the prediction.
2.3. Testing accuracy
To measure the accuracy model, this paper uses Confusion Matrix tool. It is commonly used to evaluate the classification model to predict correct and incorrect objects. In other words, it usually contains information on actual values and prediction on classification. What follows is the calculation formula of accuracy rate.

\[
\text{Accuracy Value} = \frac{\text{Number of True Values}}{\text{Total Data Amount}} \times 100\%
\]  

(2)

3. Methodology

3.1. Conceptual research framework
The conceptual framework of research can be seen in Figure 1.

![Conceptual research framework](image)

**Figure 1.** Conceptual research framework

3.2. Data collection
The primary data collected in this study is data of BPPA scholarship in STTG in 2016. There are several criteria for students to be able to be eligible to receive the scholarship in STTG. Based on the primary data, the data contains full student name, sex, semester, parent monthly income, number of dependants, GPA, and other information related to the selection of scholarship recipients in the campus.

4. Result and discussion
To simplify data modeling, the scholarship chosen explicitly in this study is BPPA. The scholarship is also chosen since it is the most wanted scholarship in the campus every year. Based on the rules, one of the requirements for the scholarship recipients candidates, those who propose should have the minimum GPA of 3.00 and should be at least in semester 6.

4.1. Manual data processing
Manual data processing is necessary to clean up the data and decrease noise effect in the process of calculation and removing the unused attributes. For the following illustrations are presented examples of manual data processing. The sample data used comprises 24 students out of a total of 1018 students described in table 1.

| Student Name | Semester | GPA    | Parent Income | Dependents | Grantee |
|--------------|----------|--------|---------------|------------|---------|
| Student 1    | 6        | 3.57   | 2,933,300     | 7          | Yes     |
| Student 2    | 6        | 3.51   | 2,849,063     | 2          | Yes     |
| Student 3    | 6        | 3.33   | 1,800,000     | 5          | Yes     |
| Student 4    | 6        | 3.35   | 500,000       | 4          | Yes     |
Table 1. Cont.

| Student | Semester | GPA  | Parent Income | Dependents | Grantee |
|---------|----------|------|---------------|------------|---------|
| Student 5 | 4        | 3.73 | 3,091,200     | 3          | Yes     |
| Student 6 | 4        | 3.7  | 1,000,000     | 5          | Yes     |
| Student 7 | 4        | 3.43 | 1,500,000     | 4          | Yes     |
| Student 8 | 4        | 3.34 | 1,500,000     | 3          | Yes     |
| Student 9 | 2        | 3.83 | 2,682,400     | 1          | Yes     |
| Student 10 | 2      | 3.85 | 2,500,000     | 4          | Yes     |
| Student 11 | 2       | 3.97 | 3,908,000     | 4          | Yes     |
| Student 12 | 2       | 3.83 | 2,000,000     | 3          | Yes     |
| Student 13 | 4       | 3.09 | 3,250,000     | 4          | No      |
| Student 14 | 6       | 3.21 | 2,250,000     | 1          | No      |
| Student 15 | 2       | 3.11 | 3,900,000     | 3          | No      |
| Student 16 | 6       | 3.01 | 3,800,000     | 2          | No      |
| Student 17 | 2       | 3.15 | 2,500,000     | 2          | No      |
| Student 18 | 4       | 3.31 | 2,750,000     | 1          | No      |
| Student 19 | 6       | 3.41 | 3,900,000     | 3          | No      |
| Student 20 | 2       | 3.23 | 2,300,000     | 2          | No      |
| Student 21 | 2       | 3.45 | 3,250,000     | 1          | No      |
| Student 22 | 4       | 3.17 | 3,200,000     | 2          | No      |
| Student 23 | 2       | 3.29 | 2,150,000     | 2          | No      |
| Student 24 | 4       | 3.03 | 2,900,000     | 3          | No      |

To decrease the distance, data normalization is done by converting the data using a minimum-maximum parameter on each criterion. The results of the data normalization are shown in table 2.

Table 2. Normalized student sample data.

| Student Name | Semester | GPA | Parent Income | Dependents | Grantee |
|--------------|----------|-----|---------------|------------|---------|
| Student 1    | 1.00     | 0.58| 0.71          | 1.00       | Yes     |
| Student 2    | 1.00     | 0.52| 0.69          | 0.17       | Yes     |
| Student 3    | 1.00     | 0.33| 0.38          | 0.67       | Yes     |
| Student 4    | 1.00     | 0.35| 0.00          | 0.50       | Yes     |
| Student 5    | 0.50     | 0.75| 0.76          | 0.33       | Yes     |
| Student 6    | 0.50     | 0.72| 0.15          | 0.67       | Yes     |
| Student 7    | 0.50     | 0.44| 0.29          | 0.50       | Yes     |
| Student 8    | 0.50     | 0.34| 0.29          | 0.33       | Yes     |
| Student 9    | 0.00     | 0.85| 0.64          | 0.00       | Yes     |
| Student 10   | 0.00     | 0.88| 0.59          | 0.50       | Yes     |
| Student 11   | 0.00     | 1.00| 1.00          | 0.50       | Yes     |
| Student 12   | 0.00     | 0.85| 0.44          | 0.33       | Yes     |
| Student 13   | 0.50     | 0.08| 0.81          | 0.50       | No      |
| Student 14   | 1.00     | 0.21| 0.51          | 0.00       | No      |
| Student 15   | 0.00     | 0.10| 1.00          | 0.33       | No      |
| Student 16   | 1.00     | 0.00| 0.97          | 0.17       | No      |
| Student 17   | 0.00     | 0.15| 0.59          | 0.17       | No      |
| Student 18   | 0.50     | 0.31| 0.66          | 0.00       | No      |
| Student 19   | 1.00     | 0.42| 1.00          | 0.33       | No      |
| Student 20   | 0.00     | 0.23| 0.53          | 0.17       | No      |
| Student 21   | 0.00     | 0.46| 0.81          | 0.00       | No      |
| Student 22   | 0.50     | 0.17| 0.79          | 0.17       | No      |
| Student 23   | 0.00     | 0.29| 0.48          | 0.17       | No      |
| Student 24   | 0.50     | 0.02| 0.70          | 0.33       | No      |
4.2. *k*-Nearest neighbor calculations

To proceed with k-NN calculation process, the following analogy can be used.

| Student Name | Conversion Results | Student Name | Conversion Results |
|--------------|-------------------|--------------|-------------------|
| NoName       | Semester: 4       | NoName       | Semester: 0.50    |
|              | GPA: 3.45         |              | Grade Point: 0.46 |
|              | Parent Income: 1,850,000 | | Income Parent: 0.40 |
|              | Dependents: 3     |              | Dependents: 0.33  |

The data are calculated using k-NN to decide the recipients of the scholarship finally. The value of k used is 5. Euclidian formula 1 is also used to find the distance of student name with “NoName” towards the sample data. From the results of calculation of Euclidean Distance for each data, then done sorting data based on the values of the distance of the smallest to largest, then taken to a number of values of K, i.e., 5 top data starting from the smallest distance value. Then, the obtained results as shown in Table 3.

Table 3. Top data sorting distance (5 smallest to largest).

| Student Name | Grantee | Distance |
|--------------|---------|----------|
| Student 8    | Yes     | 0.15     |
| Student 7    | Yes     | 0.20     |
| Student 18   | No      | 0.45     |
| Student 5    | Yes     | 0.47     |
| Student 6    | Yes     | 0.49     |

Based on the order from the smallest to the largest distance, 5 data of the K value are obtained. Those are 5 decisions of scholarship recipients that consist of “Yes” with the number 4, and “No” with the number 1. Thus, the student name with "NoName" actually has a “Yes” in the acceptance decision of the scholarship.

4.3. Testing accuracy

Since k-NN does not use the set parameters for the decision making, there needs to be the calculation of accuracy level from the sample data. It has been known that the K value in the testing is 5. Thus, the system validity testing is done to the sample data by calculating each data on each row on Table 1. The results can be seen in Table 4.

Table 4. Test sample data result.

| Student Name | Sample Target | k-NN Output | Conclusion |
|--------------|---------------|-------------|------------|
| Student 1    | Yes           | Yes         | Accurate   |
| Student 2    | Yes           | No          | Not Accurate |
| Student 3    | Yes           | Yes         | Accurate   |
| Student 4    | Yes           | Yes         | Accurate   |
| Student 5    | Yes           | Yes         | Accurate   |
| Student 6    | Yes           | Yes         | Accurate   |
| Student 7    | Yes           | Yes         | Accurate   |
| Student 8    | Yes           | Yes         | Accurate   |
| Student 9    | Yes           | Yes         | Accurate   |
| Student 10   | Yes           | Yes         | Accurate   |
| Student 11   | Yes           | Yes         | Accurate   |
| Student 12   | Yes           | Yes         | Accurate   |
| Student 13   | No            | No          | Accurate   |
| Student 14   | No            | No          | Accurate   |
| Student 15   | No            | No          | Accurate   |
Based on table 4, see that there is one piece of test data are “Student2” classified into output k-NN is wrong or not by the classes of the target sample. So, the accuracy of the algorithm can be calculated by using the formula 2 confusion matrix as follows.

\[
\text{Accuracy Value} = \frac{\text{Number of True Values}}{\text{Total Data Amount}} \times 100\%
\]

\[
\text{Accuracy Value} = \frac{23}{24} \times 100\% = 95.83\%.
\]

The accuracy values indicated that formula proves that k-NN algorithm is accurate in predicting the recipients of the scholarship.

5. Conclusions
In this study, it has been proven k-NN algorithm will function well if the data are also well. The characters of the data regarding the number of input of the variation of the values can improve the accuracy level and performance of the k-NN algorithm. In conclusion, k-NN is a right method to determine the scholarship receiver candidates, evidenced by the value of 95.83 percent achieve accuracy based on the results of testing that has been done.

References
[1] D Kurniadi, Sasmoko H L H S Warnars, and F L Gaol 2017 Software size measurement of student information terminal with use case point in IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom) pp. 164–169.
[2] H Aulawi, M A Ramdhani, C Slamet, H Ainissyifa, and W Darmalaksana 2017 Functional Need Analysis of Knowledge Portal Design in Higher Education Institution Int. Soft Comput. 12(2) pp. 132–141
[3] L Kamelia, M A Ramdhani, A Faroqi, and V Rifadiapiyana 2018 Implementation of Automation System for Humidity Monitoring and Irrigation System in IOP Conference Series: Materials Science and Engineering 288(1)
[4] D Maylawati, M A Ramdhani, W B Zulfikar, I Taufik, and W Darmalaksana 2017 Expert system for predicting the early pregnancy with disorders using artificial neural network in 2017 5th International Conference on Cyber and IT Service Management CITSM
[5] A Rahman, C Slamet, W Darmalaksana, Y A Gerhana, and M A Ramdhani 2018 Expert System for Deciding a Solution of Mechanical Failure in a Car using Case-based Reasoning in IOP Conference Series: Materials Science and Engineering 288(1) p. 12011
[6] D S Maylawati, W Darmalaksana, and M A Ramdhani 2018 Systematic Design of Expert System Using Unified Modelling Language in IOP Conference Series: Materials Science and Engineering 288(1) p. 12047
[7] A Taofik, N Ismail, Y A Gerhana, K Komarujaman, and M A Ramdhani 2018 Design of Smart System to Detect Ripeness of Tomato and Chili with New Approach in Data Acquisition in IOP Conference Series: Materials Science and Engineering 288(1) p. 12018
[8] W B Zulfikar, Jumadi, P K Prasetyo, and M A Ramdhani 2018 Implementation of Mamdani
Fuzzy Method in Employee Promotion System in IOP Conference Series: Materials Science and Engineering 288(1) p. 12147

[9] C Slamet, A Rahman, M A Ramdhani, and W Dharmalaksana 2016 Clustering the verses of the holy qur’an using K-means algorithm Asian J. Inf. Technol. 15(24) pp. 5159–5162

[10] C Slamet, R Andrian, D S Maylawati, Suhendar, W Darmalaksana, and M A Ramdhani Web Scraping and Naïve Bayes Classification for Job Search Engine in IOP Conference Series: Materials Science and Engineering 288(1) p. 12038

[11] Al-Shehri H, Al-Qarni A, Al-Saati L, Batoaq A, Badukhen H, Alrashed S, and Olatunji S O 2017 Student performance prediction using Support Vector Machine and K-Nearest Neighbor. In Electrical and Computer Engineering (CCECE) 2017 IEEE 30th Canadian Conference on pp. 1-4 IEEE

[12] J Han, J Pei, and M Kamber 2011 Data mining: concepts and techniques Elsevier

[13] M Mustakim and G Oktaviani 2016 Algoritma K-Nearest Neighbor Classification Sebagai Sistem Prediksi Predikat Prestasi Mahasiswa J. Sains dan Teknol. Ind. 13(2) pp. 195–202

[14] Y A Gerhana, W B Zulfikar, A H Ramdani, and M A Ramdhani 2018 Implementation of Nearest Neighbor using HSV to Identify Skin Disease in IOP Conference Series: Materials Science and Engineering 288(1) p. 12153

[15] Wikipedia 2018 Scholarship-Wikipedia Encyclopedia Article [online]: http://en.wikipedia.org/wiki/Scholarship. 2018.

[16] B K Baradwaj and S Pal 2012 Mining educational data to analyze students’ performance arXiv Prepr. arXiv/201.3417