Student Discipline Information System Nahdlatul Ulama University, Sidoarjo Using The Knn Method

Arief Saputro[1], Awang Andhyka[2], Syahri Mu'min[3], Rizky Aditya Nugroho[4], and Catur Lega Wibisono[5], Achmad Alif Ubaidillah[6]
Information Systems Department
Universitas Nahdlatul Ulama Sidoarjo
Sidoarjo 61218, Indonesia
ariefsyaputro.si@unusida.ac.id[1], alifubed3@gmail.com[2], awang85.si@unusida.ac.id[3], syahri.si@unusida.ac.id[4], rizkyaditya.si@unusida.ac.id[5], catur.si@unusida.ac.id[6]

Abstract

Discipline to the rules is very important for students in universities. The K-Nearest Neighbor or KNN algorithm has the advantage of classifying a new case based on the size of the similarity and the calculation process. This study aims to develop a student discipline information system at a university using the K-Nearest Neighbor algorithm. This research uses the literature study method and KNN algorithm in the case study of the Universitas Nahdlatul Ulama Sidoarjo, Indonesia. The research developed an automated analysis process system to simplify and speed up student discipline analysis. This information system uses three assessment variables: attendance, GPA, and student organization activity. The results showed that the lowest proximity value was used as a reference for new student discipline assessments based on data held by old students. The study showed 79.96% at a value of k = 5 and a value of k = 7. This indicates that the K-Nearest Neighbor method was an appropriate method used to solve classification problems.

Keywords
Information Systems, K-Nearest Neighbor, Classification, Student, Discipline

1. Introduction

This era of globalization has both positive and negative impacts on the world of education. One of the negative impacts of the current of globalization is the erosion of the nation's moral values due to the influence of foreign cultures which are sometimes not in accordance with the culture of the Indonesian nation. Nations that want smart, faithful, and pious citizens, need to pay attention to student education because the success of education is determined, among other things, by the education provided to students (Iswanti, et al, 2012). Nahdlatul Ulama University Sidoarjo (UNUSIDA), is an institution that functions as a vehicle for fostering and empowering students in contributing to increasing the competence and character of students with noble character. InstitutionThis requires the existence of rules and norms that become a reference for coaches, managers and residents so that they can help create a conducive atmosphere for student learning. Today's students also have a lot of undisciplined behavior due to coercion and restraint. Discipline can be seen from attitudes, behavior, and actions that are in accordance with organizational regulations, whether written or not. Discipline is the parameters in determining success, because we can see that successful people are disciplined people (Nitisemito, 2008).

According to Amri (2016), discipline comes from the word "disciple" which is someone who learns to voluntarily follow a leader. According to Hurlock in stating that discipline is the way society teaches children to behave morally approved by the group.

Meanwhile, according to Hani, (2008) states that discipline is a very absolute thing in human life, because a human being without strong discipline will damage the joints of life, which will endanger himself and other humans and even the environment.

In addition, the development of technology that is increasingly rapidly now has an impact on all aspects of human life. The obvious impact is that technology facilitates all human activities. It is undeniable that now everything that used to be manual has now turned automatic (Indriani et al., 2017).

In this era of globalization, of course, we are required to keep pace with the development of technology and information. Technology as a tool to support the management of resources and information as a guide to the management carried out. In these developments, of course, it is necessary to integrate technology and information so as to get maximum results.

Based on this background, this research will build a system that can determine whether or not students are disciplined by taking 3 data variables, namely attendance, GPA and organizational activity. There are several methods for classification such as Naive Bayes, K-Nearest Neighbor, Decision Tree, Support Vector Machine, or K-Means Clustering. The classification method used in this study is K-Nearest Neighbor. The K-Nearest Neighbor method uses the concept of minimizing variations between data in a cluster and maximizing variations with data in other clusters. This method was chosen because it is easy to implement and run in the time required to carry out this learning is relatively fast and easy to adapt.

2. Research Method

According to Han et al. (2012: 421-425), the K Nearest Neighbor method was first described in the early 1950s. As the use of available computing power increases, so that K Nearest Neighbor has developed to be used in the pattern recognition field. K Nearest Neighbor classification is based on analogies such as comparing test data with training data that are similar to it. Training data is given as many as n attributes. Each data is a point in n-dimensional space. All training data is stored in the n-dimensional pattern space. When the data is unknown, K Nearest Neighbor will search the pattern space for the k training data closest to the unknown data. K Nearest Neighbor classification, the data will be included in the most common class (many similarities) among the k closest neighbors. K Nearest Neighbor when given "unknown" data, the data will look for the space pattern of the k training data closest to the "unknown" data. The k training data is the k nearest neighbor of the “unknown” data. “Unknown” data is assigned to the class with the most similarity among k nearest neighbors. When k = 1, the "unknown" data is assigned to the class from the training data that is closest to the pattern. The algorithm for classifying K Nearest Neighbor is as follows. then the "unknown" data is determined by the class from the training data that is closest to the pattern. The algorithm for classifying K Nearest Neighbor is as follows. then the "unknown" data is determined by the class from the training data that is closest to the pattern. The algorithm for classifying K Nearest Neighbor is as follows.

a. Determine the parameter k, which is the number of nearest neighbors.
b. Calculate the distance between the new data and all the training data.
c. Sorts the distance between the new data and the training data.
d. Determine the nearest neighbor based on the kth minimum distance.
e. Check the class of the nearest neighbor.
f. Using a simple majority of the nearest neighbor class to find out the predicted value of the test data (new).

The accuracy of the K Nearest Neighbor grouping is believed to be very good at a certain distance. Extensive research was carried out as a development of the K Nearest Neighbor method which focused on a new way of distance determination by automated learning (Keller, 1985).

The K-Nearest Neighbor (KNN) algorithm is a method for classifying new objects based on (K) their closest neighbors (Gorunescu, 2011). K Nearest Neighbor includes a supervised learning algorithm, which results from a new query instance, classified according to the majority of the categories in K Nearest Neighbor. The class that appears the most will be the class resulting from the classification (Gorunescu, 2011).

In the K Nearest Neighbor algorithm, there are five ways, to find the nearest neighbor (Prasetyo, 2017), namely:

a. Euclidean distance
b. Manhattan distance
c. Cosine Distance
d. Correlation Distance
e. Hamming distance

In this research, the writer only uses Euclidean distance, so the formula for calculating distance with Euclidean is as follows:

$$\sqrt{\sum_{i=1}^{K}(Xi - Yi)^2}$$

Figure 1.Euclidean formula

The Xi value is the value in the Training data, while the Yi value is the value in the Testing data. The value of K is an attribute dimension.

Steps to calculate the K Nearest Neighbor algorithm:

a. Determine the value of k.
b. Calculates the square of the Euclidean distance (query instance) of each object to the given training data.
c. Then sort the objects into groups that have the smallest Euclidean distance.
d. Collecting class Y labels (Nearest Neighbor classification).
e. By using the category of Nearest Neighbor which is the majority, it can predict the value of the calculated query instance.

The influential value of k depends on the amount of data. In a related study, a high value of k will reduce noise in the classification process, but by setting a boundary for each classification, the results will be blurry. The optimal value of k can be obtained by optimizing parameters, namely using cross-validation. In a special case where a classification is predicted based on training data that has proximity (in other words, k = 1) is called the Nearest Neighbor algorithm. The success and accuracy of the K Nearest Neighbor algorithm is influenced by irrelevant features or if the weights of these features do not match their relevance to the classification.

K-Nearest Neighbor (KNN) is one of the top 10 most popular data mining methods (Wu & Kumar, 2009). The pure K Nearest Neighbor method is included in the lazy learner classification because it delays the training process (or even does not do the training at all) until there is test data that you want to know the class label for, then the new method will run the algorithm. The K Nearest Neighbor algorithm performs a classification based on the similarity of a data with other data (Tan et all, 2004). The advantages of K-Nearest Neighbor:

a. Easy to understand and implement. The K Nearest Neighbor algorithm is easy to understand and easy to implement. To classify an instance of x using K Nearest Neighbors, we simply define a function to calculate the distance between instances, calculate the distance of x to all other instances based on that function, and determine the class x as the class that occurs the most in the k closest instances.
b. More effective in large training data.
c. Can produce more accurate data.

3. Result And Discussion

a. Data Integration (data integration)

The data integration process is data obtained from different sources that can be combined into one dataset. In this study, data were obtained from different sources, namely student data from the Faculty of Computer Science, Information Systems Study Program, Nahdlatul Ulama University Sidoarjo combined into one dataset.

b. Data Selection (data selection)

The data selection process is the selection of data or attributes to be used and removing unused data from the dataset. At this stage the authors perform data selection manually using Microsoft Excel.

c. Data Transformation (change data)

Process data *transformation*is the process of changing the value of the data and converting it into the form of data to be mined. In this study, the author carried out the process of transforming student data manually using Microsoft Excel.

d. Data mining (data mining)

In this study the classification method uses *K-Nearest Neighbor* for information gathering. The correlation between prediction and classification is a classification prediction model, which is to distinguish data into classes. While the prediction model will look for new values in the future. Both prediction models use historical data. The K-Nearest Neighbor algorithm uses neighboring classification as a predictive value and a single data test.

e. Interpretation/Evaluation (pattern evaluation)

The pattern of information generated from the data mining process needs to be displayed in a form that is easily understood by interested parties. In this study the results of visualization in the form of accuracy results.

Activity The Data Training diagram, namely the interaction between the admin and the data training use case, is described in the Activity diagram in the image below.

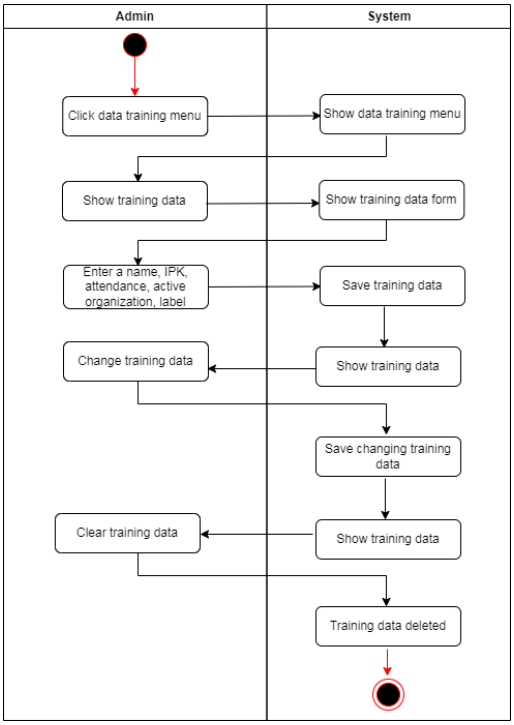

Figure 2. Activity diagram data Training

Activity The Data Testing diagram, namely the interaction between the admin and the data testing use case, is described in the Activity diagram in the image below

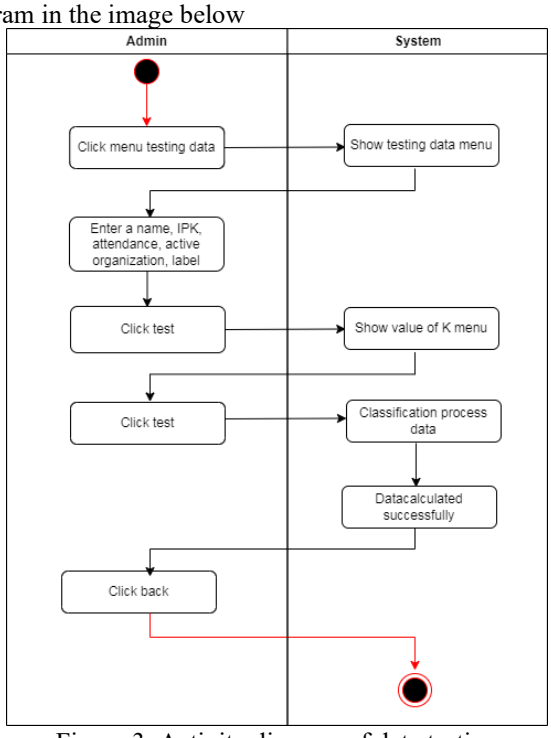

Figure 3. Activity diagram of data testing

The data used are 35 data. Of the 35 data divided into training data and testing data. 32 for training data and 3 for testing data.table for the calculation of data normalization that I took as an example of the calculation as many as 10. The formula for the normalization itself is:

$$Normalisasi = \frac{data_x - data_{min}}{data_{max} - data_{min}}$$

Figure 4. Normalization Formula

Where data x is the first row of a training data attribute, then data min is the smallest value of all data on attribute x also applies to data max which is looking for the largest value of all data attributes x.

$$d_{Euclidean}(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Figure 5. Euclidean Formula

Calculate the distance of each testing data one by one to the training data using the Euclidean Distance formula. In the K Nearest Neighbor algorithm, it uses neighboring groupings with predictive values from the new testing data. Determining the similarity used is the Euclidean Distance method. The following is the application to calculate the Euclidean Distance value in the K-Nearest Neighbor algorithm.

Table 1. Euclidean Distance

1 test data	2 test data	2 test data
1.68948	0.497012	1.001418
1.609048	0.517136	0.858796
1.584362	0.616309	0.714286
1.470784	0.500242	0.727273
1.278003	0.65943	0.480398
1.093744	0.588678	0.59439
0.765491	0.676412	1.349727
0.714286	0.932427	1.584362
1.091925	1.024194	0.326622
0.450129	0.943097	1.349263

Determining the value of k (nearest neighbour) is determining or selecting the smallest distance value that is ranked by a number of k values. For example, if the value of k is 3 then you have to find the smallest distance value of two values. The more values produced, the farther away the value of k (nearest neighbour) is.

Table 2. K . Value

1 test data	2 test data	2 test data
3	3	3
-	discipline	-
-	discipline	-
-	-	-
-	discipline	-
-	-	discipline
-	-	discipline
discipline	-	-
discipline	-	-
-	-	discipline
discipline	-	-

after testing using 3 datasets, the highest accuracy is 79.96%. The results of this accuracy are obtained from each experiment with the value of k (nearest neighbor) starting from 1, 3, 5, 7, 9. And the value of k that produces the highest accuracy is k = 5 and k = 7. The results of the accuracy and the number of nearest neighbors can be seen in the table below:

Table 3. Results % of the best K values

K . value	Right	Wrong	%
1	2	1	66.6
3	2	1	66.6
5	3	0	100
7	3	0	100
9	2	1	66.6

4. Conclusion

In research on the application of the K-Nearest Neighbor algorithm for predicting student discipline at Nahdlatul Ulama' Sidoarjo University, the following conclusions can be drawn:

a. The K-Nearest Neighbor algorithm has been successfully implemented.
b. The results of the accuracy of the K-Nearest Neighbor Algorithm for predicting student discipline at Nahdlatul Ulama University Sidoarjo were obtained at 79.96%.
c. The value of k (nearest neighbor) that produces the highest accuracy is k=5 and k=7.

References

Ayistiya,S(2014).NearestNeighbor Algorithm. http://ayistiya13.blogspot.co.id/2014/11/algoritma-nearest-neighbor.html.

Firman, A., Wowor, H., & Najoan, X. (2016). Web Application Based Library Information System. Journal of Business Information Systems, 1(2), 66–77. https://doi.org/10.21456/vol1iss2pp69-72.

Gorunescu, F. 2011. Data Mining Concepts, Models and Techniques, New York: Springer_Verlag.

Han, J., M. Kamber, & J. Pei. 2012. Data Mining Concepts and Techniques (3th ed.). Waltham: Morgan Kaufmann Publishers.

Hani (2008) Discipline Development Strategy at Madrasa Ibidaiyah Negri Olak-Alen Selorejo Blitar. http://thesis(Faculty of UIN Maliki Malang).

Indriani, M., Sholihin, M., & Aprilliani, NF (2017). Application for Data Filling Alumni of the Lamongan Islamic University Engineering Study Program Based on Android. Jouticla, 1(2), 1–6. https://doi.org/10.30736/jti.v1i2.48.

Iswanti.S, et al, (2012). Teacher's attitude and personality formation http://journal of research iswanti et al.pdf.

Jamal, M. (2013). Guidebook for Internalization of Character Education in Schools, Jogjakarta: Diva Press. http://journal of character education in schools.pdf.

Keller, JM, MR Gray, & J.A. Givens. 1985. A Fuzzy K-Nearest Neighbor Algorithm. IEEE Transactions on Systems, Man and Cybernetics, 15(4): 580-585.

Krisandi, N., Helmi, Prihandono, B. 2013. The K-Nearest Neighbor Algorithm in data classification of oil palm production at PT. Minamas, Parindu sub-district. Math Scientific Bulletin. Stats and Applied (Bimaster).

Mittal, Kavita & P. Mahajan. 2017. Performances Analysis of K-Nearest Neighbor and K-Means Clustering To Predict The Diagnostic Accuracy. Proceedings of International Conference on: Information, Communication and Computing Technology (ICICCT 2017), 26-36.

Muryanto. (2008). Creating Personal Children Easy to Get along with Semarang Cv.ghyas son. http://Journal of Easy Sociable Children's Personal.

Nitisemito. (2008) Contribution of Achievement Motivation and Work Discipline to

Biographies

Arief Saputro is a faculty member of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia.

Awang Andhyka is a faculty member of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia.

Syahri Mu'min is a faculty member of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia

Rizky Aditya Nugroho is a faculty member of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia.

Catur Lega Wibisono is a faculty member of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia.

Achmad Alif Ubaidillah is a student of the Information Systems Department at the Universitas Nahdlatul Ulama Sidoarjo, Sidoarjo, Indonesia.

.