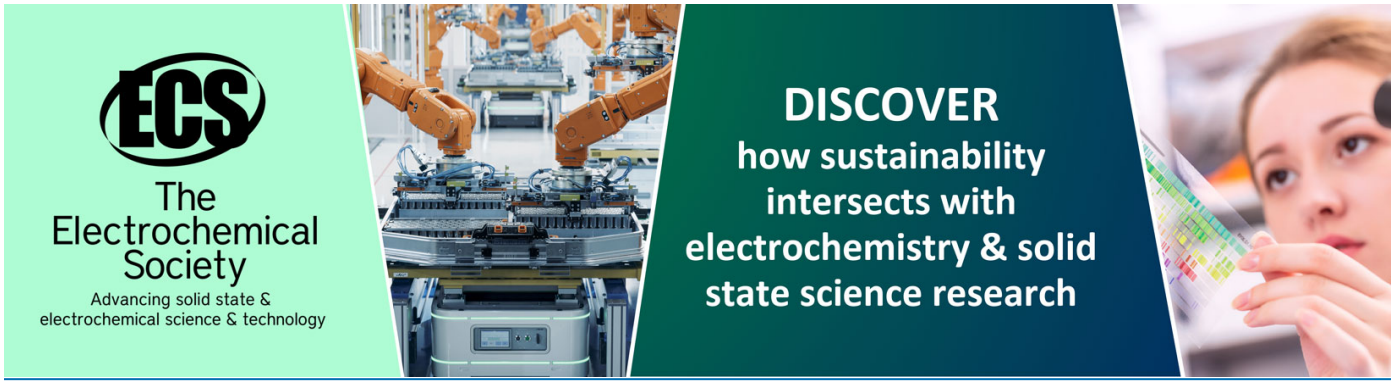
Prediction of Vocational Students Behaviour using The k-Nearest Neighbor Algorithm



Prediction of Vocational Students Behaviour using The k-Nearest Neighbor Algorithm

Fadliansyah Nasution[1], Elviawaty Muiza Zamzami[2]

[1]Master Programme in Informatics, Universitas Sumatera Utara
[2]Faculty of Computer Science and Information Technology, Universitas Sumatera Utara

*E-mail: fadliansyah.nasution@gmail.com

Abstract. This article discusses the implementation of the k-NN algorithm in predicting student behavior. The school is a management unit that has data that correlate with students. All student data is stored in an academic information system that can be processed to predict student behavior. One of the data assessing student behavior is in the database of counseling guidance. Some data that will be processed include attendance, lateness notes of problems, teacher responses, tuition payments, broken home. The sample being tested was the data of 100 vocational students from various classes and various majors and divided into two categories. From this experiment, it can be seen that the most accurate K value is the value of K = 1, 3 and 4. The accuracy of the testing data generated is 94.9%.

1. Introduction

K-Nearest Neighbor is one classification tool using all training samples in classifications that cause high level of computational complexity. where the nearest neighbor is calculated based on the k value in determining how many closest neighbors should be considered to decide the class of the sample data point, correcting the K-Nearest Neighbor based on the weight of the k value. Training is given weight according to distance from the sample data points, but computational complexity and memory remain a major concern [1],[5].

School is a place to form the character and personality of a human being. Starting from kindergarten to college. Vocational high schools are part of the education system in Indonesia which is the target of current government achievements. Vocational School consists of several majors that have specialization from students who come from a variety of behaviors that can be applied to various disciplines in accordance with the needs of competencies and majors.

Nowadays, the development of telecommunications, media and informatics is currently receiving a positive reception in the community [6]. Some schools have begun to build academic information systems to assist schools in managing student data, administrative data and other data. One of the data that is part of this information system is student counseling data. Student counseling is a field that deals with student behavior

at school dealing with problem students. Therefore we need a method that is able to predict student behavior to be able to find out earlier the tendency of students to behave.

2. Literature Review

2.1. 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) [7]. Following is the formula of distance search using Euclidian formula [8], [9]

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

Where: d: distance; p: data dimension; i: data variable; X_{1i}: sample data; X_{2i}: 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.

2.2. Accuracy Testing

The process of nearest neighbor performance is by adding a location which has the nearest distance to the location visited last time [10]. Therefore, 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 predictions on classification. What follows is calculation formula of accuracy rate.

$$\text{Accuracy Value} = \frac{\text{Number of True Values}}{\text{Total Data Amount}} \text{x } 100\% \quad (2)$$

3. Methodology

3.1 Conceptual research framework

The conceptual framework of research can be seen in Figure 1

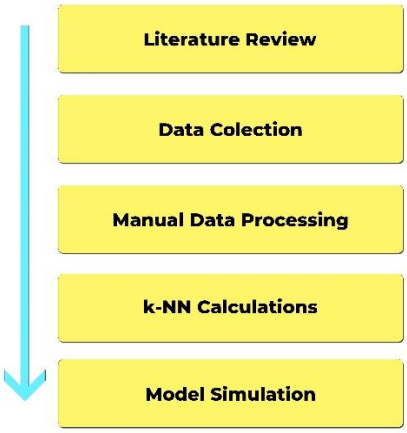


Figure 1. Conceptual framework research

3.2. Data Collection

The data is obtained from one of the private vocational schools in Medan, it is SMKS IT Marinah Al Hidayah. The data source is taken from the academic information system (sia.smkmarinah.sch.id). In this data there are some features is needed to be processed by the method specified.

We also obtain data manually from the assessment of teachers and high school students, which is entirely a student counseling database. Data includes attendance, lateness, violations, responses from teachers, classmates, parents' income, student's score in the report book. This information is information that might be related to student behavior.

4. Result and discussion

4.1 Manual data processing

Behavior of students will be tested on 100 students data from grades 10 to 12. In previous, all students had been categorize as a problematic and non-problematic according to data from the counseling guidance teacher. Data is gathered through assessment and school administration data which taken is the most recent data from the beginning of the teaching by supporting previous data that came from before students entered the class. The sample is 100 students as respondents consisting of various classes and majors. Table 1 shows student data that will be tested.

Table 1. Student sample data

No.	Student Name	Class	absensi	Keterlambatan	Catatan Masalah	Respon Guru	Pembayaran SPP	Broken Home	Kategori
1	Febri	12	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Aman	Ya	Tidak Bermasalah
2	Mushir	12	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Tidak Aman	Ya	Tidak Bermasalah
3	Irham	12	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Tidak Aman	Tidak	Tidak Bermasalah
4	Firza	12	Sedikit	Tidak Ada	Tidak Ada	Baik	Aman	Tidak	Tidak Bermasalah
5	Aliya	12	Tidak Ada	Sedikit	Tidak Ada	Baik	Aman	Tidak	Tidak Bermasalah
6	Asruli	12	Banyak	Banyak	Ada	Kurang	Tidak Aman	Tidak	Bermasalah
7	Farhan	12	Banyak	Banyak	Ada	Kurang	Aman	Tidak	Bermasalah
8	Ibnu	12	Banyak	Banyak	Ada	Kurang	Aman	Ya	Bermasalah
9	Elsa	12	Banyak	Sedikit	Tidak Ada	Baik	Aman	Tidak	Bermasalah
10	Khaviz	12	Sedikit	Banyak	Tidak Ada	Baik	Aman	Tidak	Bermasalah
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
98	Azizah	10	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Aman	Tidak	Tidak Bermasalah
99	Putri Ayu	10	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Aman	Tidak	Tidak Bermasalah
100	Putri Fadila	10	Tidak Ada	Tidak Ada	Tidak Ada	Baik	Aman	Tidak	Tidak Bermasalah

Feature transformation is needed to processing data which run in the program. For the transformation of features can be seen in Table 2.

Table 2. Feature Transformation

	absensi	Keterlambatan	Catatan Masalah	Respon Guru 1	Pembayaran SPP	Broken Home	Kategori
0	0	0	0	1	1	1	0
1	0	0	0	1	0	1	0
2	0	0	0	1	0	0	0
3	1	0	0	1	1	0	0
4	0	1	0	1	1	0	0

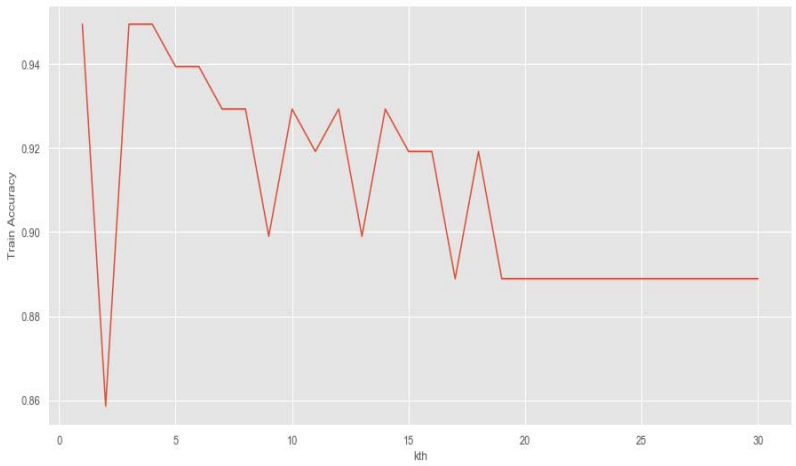
4.2 Testing accuracy

Since k-NN does not use the features specified in decision making, there is important to calculate level of accuracy of data sample. Therefore the value of testing the system validity is carried out on the sample data by counting each data in each row in Table 1. The results can be seen from Table 3 below

Table 3. Test sample data result

	absensi	Keterlambatan	Catatan Masalah	Respon Guru	Pembayaran SPP	Broken Home	Sample Target	KNN Output	Conclusion
41	1	1	0	1	0	0	0	1	not accurate
30	2	2	1	0	0	0	1	1	accurate
96	2	2	1	0	0	0	1	1	accurate
25	2	2	1	0	0	0	1	1	accurate
11	1	1	1	0	0	0	1	1	accurate
23	2	1	1	0	0	0	1	1	accurate
16	1	1	1	0	1	0	1	1	accurate
34	2	1	1	0	0	0	1	1	accurate
19	1	1	1	0	0	0	1	1	accurate
57	1	1	1	0	0	0	1	1	accurate
33	2	1	1	0	0	0	1	1	accurate

In Table 3, we can see that there is inaccurate data in the test. In sequence number 41 the results of the target sample that has a value of "0" are different from the KNN output that has a value of "1". This value affects the accuracy of the results obtained. From the results of the experimental data the K score is obtaining optimal accuracy results. The K scores are 1, 3 and 4 with the percentage accuracy being 94.9%. Information on the accuracy value of K can be explained in figure 2. Graph K training data and the results of its accuracy



	kth	Train Accuracy		kth	Train Accuracy
0	1	0.949495	**15**	16	0.919192
1	2	0.858586	**16**	17	0.888889
2	3	0.949495	**17**	18	0.919192
3	4	0.949495	**18**	19	0.888889
4	5	0.939394	**19**	20	0.888889
5	6	0.939394	**20**	21	0.888889
6	7	0.929293	**21**	22	0.888889
7	8	0.929293	**22**	23	0.888889
8	9	0.89899	**23**	24	0.888889
9	10	0.929293	**24**	25	0.888889
10	11	0.919192	**25**	26	0.888889
11	12	0.929293	**26**	27	0.888889
12	13	0.89899	**27**	28	0.888889
13	14	0.929293	**28**	29	0.888889
14	15	0.919192	**29**	30	0.888889

Figure 2. Graph K training data and graph accuracy results

5. Conclusion

In this experiment, it can be concluded that K-Nearest Neighbor is able to predict student behavior from school data related to student behavior. At a smaller K value, the accuracy rate is better, even reaching 93%, and with a larger k value the accuracy is reduced. In this experiment, it becomes a problem when the data

is taken from student data, so it needs more data for testing and its level of accuracy. The test feature is also a support in its level of accuracy. The more features so it will be better, but the amount of data tested must also be large, otherwise there will be overfitting. So for schools with many features students must be large enough for the accuracy of the test data.

References

[1] M E Saputra, H Mawengkang, and E B Nababan 2019 Gini Index With Local Mean Based For Determining K Value In K-Nearest Neighbor Classification *J. Phys.: Conf. Ser* **1235** p. 012006
[2] M E Saputra, H Mawengkang, and E B Nababan 2018 Determination value k in k-nearest nieghbor with local mean euclidean and weight gini index *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 420, p. 012098,
[3] N Bhatia and C Author 2016 Survey of Nearest Neighbor Techniques vol. 8, no. 2, p. 4
[4] A A Nababan, O S Sitompul, and Tulus 2018 Attribute Weighting Based K-Nearest Neighbor Using Gain Ratio," *J. Phys.: Conf. Ser.* **1007** p. 012007
[5] P WiraBuana, S Jannet D.R.M., and I Ketut Gede Darma Putra 2012 Combination of K-Nearest Neighbor and K-Means based on Term Re-weighting for Classify Indonesian News *IJCA*, 50 11, pp. 37–42
[6] A Setiadi, M A Adnandi, S T Wibowo, D Pratama, and A B Rizky 2015 Analisa Sistem Pakar Untuk Identifikasi Kepribadian Siswa Menggunakan Algoritma Fuzzy Pada Siswa Slta p. 6
[7] Jiawei H, Micheline K, Jian P 2011 Data Mining Concepts and Techniques The-Morgan-Kaufmann-Series in Data Management Systems 3rd Ed-Morgan-Kaufmann
[8] D Kurniadi, E Abdurachman, H. L. H. S. Warnars, and W.Suparta 2018 The prediction of scholarship recipients in higher education using k-Nearest neighbor algorithm IOP *Conf. Ser.: Mater. Sci. Eng.*, **434**, p. 012039
[9] B Kumar and S Pal 2011 Mining Educational Data to Analyze Students Performance *Int. of Advanced Computer Science and Application*s **2** 6
[10] M I Mulyadewi, Handriyono, and F N D Nadia 2019 Analysis of 3 Kg LPG determination route decision using nearest neighbor algorithm and local search method *Int. J. Sci. Technol. Res*. **8** 6, pp. 159–162