# IMPLEMENTATION OF K-NEAREST NEIGHBOUR (KNN) ALGORITHM TO PREDICT STUDENT'S PERFORMANCE

**Slamet Wiyono**
Informatics Engineering Department
Politeknik Harapan Bersama
Email: slamet2wiyono@gmail.com

**Taufiq Abidin**
Informatics Engineering Department
Politeknik Harapan Bersama
Email: N3k4ther.otr@gmail.com

**ABSTRACT**

Salah satu unsur untuk menjadi penilaian akreditasi adalah ketepatan waktu lulusan siswa. Adanya siswa yang tidak aktif tentu akan mempengaruhi ketepatan waktu kelulusan. Prediksi kinerja siswa diperlukan untuk mencegah siswa yang tidak aktif. Algoritma KNN digunakan untuk memprediksi kinerja siswa dengan menggunakan metode klasifikasi. Penelitian ini untuk mengoptimalkan algoritma KNN untuk memprediksi kinerja siswa dengan metode klasifikasi. Penelitian yang telah dilakukan dengan menggunakan data Jurusan Teknik Informatika Politeknik Harapan Bersama menyimpulkan bahwa nilai K terbaik adalah 3, 6, dan 9 untuk mendapatkan prediksi terbaik. Hasil ini diperoleh dengan mencoba nilai K, 3 hingga 60. Nilai prediksi kemudian dibandingkan, hasil yang salah diprediksi dimana persentase terkecil adalah yang terbaik.

**Kata kunci:** KNN; optimasi; kinerja siswa.

***ABSTRACT***

*One of the elements to be an accreditation assessment is the timeliness of graduating students. The existence of non-active students will certainly affect the timeliness of graduation. Prediction of student performance is needed to prevent non-active students. KNN algorithm was used to predict student performance by using classification method. This research is to optimize KNN algorithm to predict student performance by classification method. The research had been done by using data of department Informatics Engineering Politeknik Harapan Bersama conclude that the best value K are 3, 6, and 9 to get the best predict. This result is obtained by trying the value of K is 3 to 60. The predicted value is then compared, the incorrectly predicted result of which the smallest percentage is the best.*

***Keywords:*** *KNN; optimization; student performance.*

## 1. INTRODUCTION

Each department will try to improve the quality of education and accreditation of the department. One of the elements to be an accreditation assessment is the timeliness of graduating students [1]. The more students who graduate on time the better the value of accreditation. The existence of non-active students will certainly affect the timeliness of graduation. The more non-active students will be more and more students who pass not on time. Thus, the more number of non-active students hence can affect the value of accreditation of study program.

Prediction of student performance is needed to prevent non-active students. Research on predictions of student performance several times had been conducted. Among the research that had been conducted are predictions of student activity using the KNN algorithm [2]. Similar studies had also been conducted, namely; research to predict students' graduation using KNN algorithm [3], In addition to using the KNN algorithm, the application of Fuzzy Inference System (FIS) had also been used to predict student activity [4], Random Forest algorithm to predict length of student study [5], Decision Tree C4.5 algorithm to predict potentially non-active students [6] and to predict the study period of students [7]. In addition to these studies, other similar studies had also been conducted; research about academic performance using decision tree techniques [8], predict student's performace using data mining technique [9], and estimating student's performance using Weka Environment [10].

KNN algorithm is perhaps one of the simplest machine learning algorithms, it is still used widely [11]. The KNN algorithm is highly dependent on the number of Kernels to get predicted results. Several studies have been conducted to optimize the number of Kernel KNN algorithms. Research that had been done; learning K on the KNN algorithm to make predictions [12], optimization techniques modified K Nearest Neighbor classification using

Genetic algorithm [13], and optimization of K parameters in K-Nearest Neighbor algorithm for classification of diabetes disease mellitus [14]. This research is to optimize KNN algorithm to predict student performance by classification method. Classification methods are widely applied in many sciences such as health sciences [15], science education [16][17], building science [18], and others. The focus of research that has been done is to optimize the kernel on KNN algorithm to predict (classification method) student performance.

## 2. METHODS

KNN algorithm was used to predict student performance by using classification method. Nearest Neighbor classifiers are defined by their characteristic of classifying unlabeled examples by assigning them the class of similar labeled examples. Despite the simplicity of this idea, nearest neighbor methods are extremely powerful [11]. Steps of the research process shown in Figure 1.

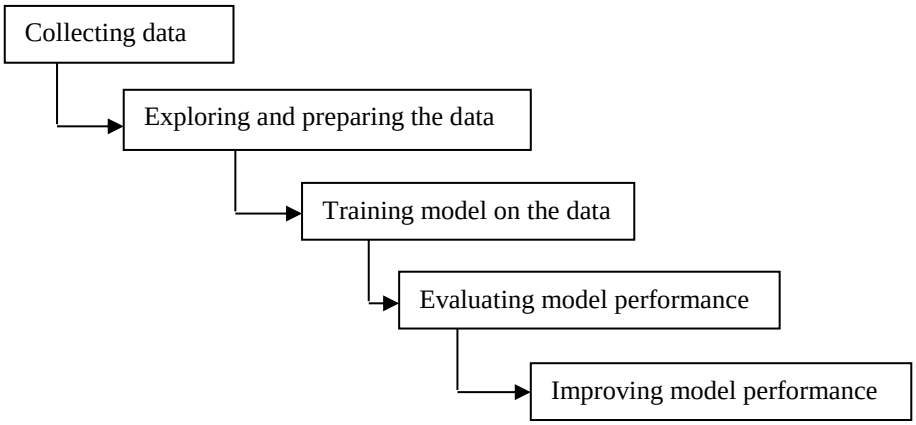


**Figure 1. Steps Of The Research Process**

### *2.1 Collecting Data*

We utilize the student academic data of Department Informatics Engineering Politeknik Harapan Bersama. The data used was 1530 rows with 7 attributes numeric. These 7 attributes are; grade point, grade point average, hometown, type of school, major at school, parent's job, and student performace. The student performance is coded as "A" to indicate active or "N" to indicate non-active.

### *2.2 Exploring and Preparing Data*

Data exploration and preparation was done to see the dataset to be used. If there is data that is not appropriate, then the data will be corrected. Data exploration and preparation was done using the str command in R Studio. Checking results show the dataset has been structured with 1,530 lines and 7 attributes as expected. The first few lines of the checking output are shown in Figure 2.

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       1530 obs. of  7 variables:
 $ GP            : num  3.2 3.9 3.6 3.55 3.9 4 2.25 3.9 3.65 3.7 ...
 $ GPA           : num  3.2 3.9 3.6 3.55 3.9 4 2.25 3.9 3.65 3.7 ...
 $ Hometown      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ Type of school: num  2 1 1 1 2 1 2 2 1 1 ...
 $ major         : num  1 3 3 2 1 3 1 1 3 3 ...
 $ parent's job  : num  1 4 2 4 3 2 2 2 3 4 ...
 $ AKTIF         : num  1 1 1 1 1 1 1 1 1 1 ...
```

**Figure 2. Structure Dataset**

Then I transformed by normalizing numerical data to equate the data. Normalizing numerical data using equation (1), and output are shown in Figure 3.

```
'data.frame':   1530 obs. of  7 variables:
 $ GP            : num  0.8 0.975 0.9 0.887 0.975 ...
 $ GPA           : num  0.8 0.975 0.9 0.887 0.975 ...
 $ Hometown      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Type.of.school: num  1 0 0 0 1 0 1 1 0 0 ...
 $ major         : num  0 1 1 0.5 0 1 0 0 1 1 ...
 $ parent.s.job  : num  0 0.75 0.25 0.75 0.5 0.25 0.25 0.25 0.5 0.75 ...
 $ AKTIF         : num  1 1 1 1 1 1 1 1 1 1 ...
```

**Figure 3. Structure Dataset After Normalizing**

After the data becomes the same, the next is preparing the data by creating training and test dataset. I had used about 70% of the data for training and about 30% for tests.

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)} \quad (1)$$

### *2.3 Training Model on The Data*

This step, the ready dataset is used to classify. For the KNN algorithm, at this stage it is not for model formation; the training process only involves storing input data in a structured format. Model training had been done using Kernel = 39. The number of Kernels used is 39 because one common practice is to begin with Kernel equal to the square root of the number of training examples. This stage has produced one model.

### *2.4 Evaluating Model Performance*

The next step of the process is to evaluate how well the had predicted. To do this, I used CrosTable function in the gmodels packages of R Studio. After loading the packet, I have created a cross-tabulation showing the agreement between two output vectors: label and prediction. The cross-tabulation is shown in Figure 4. Figure 4 shows the number of false negative is 0 and false positive is 9, so there 9% classified is incorrect.

| | prediction | | |
|---|---|---|---|
| test_label | 0 | 1 | Row Total |
| 0 | 113 | 9 | 122 |
| | 0.926 | 0.074 | 0.244 |
| | 1.000 | 0.023 | |
| | 0.226 | 0.018 | |
| 1 | 0 | 378 | 378 |
| | 0.000 | 1.000 | 0.756 |
| | 0.000 | 0.977 | |
| | 0.000 | 0.756 | |
| Column Total | 113 | 387 | 500 |
| | 0.226 | 0.774 | |

**Figure 4. Cross-Tabulation Output Label And Prediction**

### *2.5 Improving Model Performance*

In these step, attempted improving the model by trying several different values for kernel (K). By trying out different values of K, it is hoped that the best model will be obtained. The same 500 labels are classified using different K values. Then the numbers of the false negative and false positive are displayed each iteration.

## 3. RESULT AND DISCUSSION

The result of the research is the percentage of incorrect prediction student performance. Table 1 shows the results of research with attributes; K value, false negative, false positive, and predict incorrectly. 20 different K values are used to compare the percentage of incorrectly predicted. the best value of K to make predictions is 3, 6, and 9 with the percentage of incorrectly predicted 0%. Figure 5 shows that the greater the value of K the greater the incorrectly predicted, and the smaller the value of K the smaller the incorrectly predicted. Although the percentage of incorrectly predicted is getting greater, but the percentage is not to so different, just about 0,0%.

Figure 6 shows Relation k values with false negative and false positive. Whatever the value of K then the false negative value is 0, it indicates that the accuracy for predicting the inactive student (0) is very high. The greater the value of K the higher the false positive and otherwise. It shows that the greater the value of K the accuracy of the prediction decreases and otherwise.

**Table 1. Predicted result using different K value**

| ***K value*** | 3 | 6 | 9 | 12 | 15 | 18 | 21 | 24 | 27 | 30 | 33 | 36 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***False negative*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***False positive*** | 0 | 0 | 0 | 1 | 2 | 3 | 3 | 5 | 6 | 7 | 7 | 7 |
| ***Incorrectly predicted (%)*** | 0 | 0 | 0 | 0.002 | 0.004 | 0.006 | 0.006 | 0.01 | 0.012 | 0.014 | 0.014 | 0.014 |

| ***K value*** | 39 | 42 | 45 | 48 | 51 | 54 | 57 | 60 |
|---|---|---|---|---|---|---|---|---|
| ***False negative*** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***False positive*** | 9 | 9 | 10 | 10 | 12 | 12 | 13 | 13 |
| ***Incorrectly predicted (%)*** | 0.018 | 0.018 | 0.02 | 0.02 | 0.024 | 0.024 | 0.026 | 0.026 |

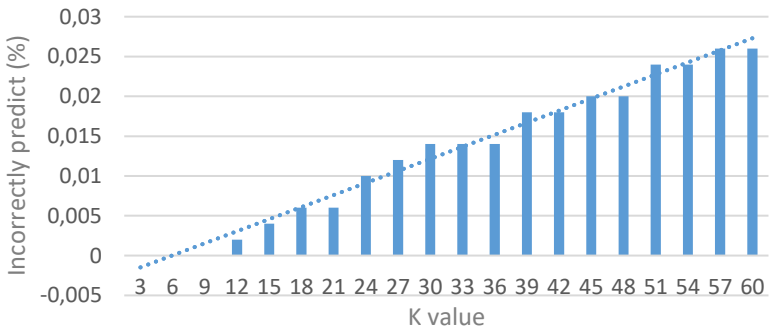


**Figure 5. Trendline Incorrectly Predicted Based Of K Value**

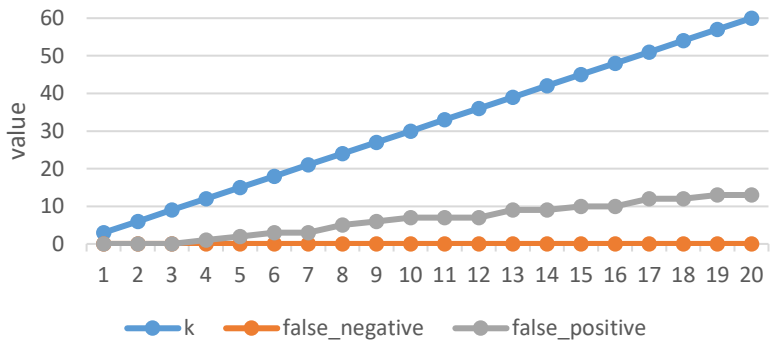


**Figure 6. Relation k Values With False Negative And False Positive**

## 4. CONCLUSIONS

The research had been done by using data of department Informatics Engineering Politeknik Harapan Bersama conclude that the best value K are 3, 6, and 9 to get the best predict. This result is obtained by trying the value of K is 3 to 60. The predicted value is then compared, the incorrectly predicted result of which the smallest percentage is the best.

## ACKNOWLEDGEMENTS

This research supported by Ministry of Research, Technology and Higher Education (RISTEKDIKTI) Indonesia

## REFERENCES

[1] BAN-PT, *Buku I Naskah Akademik Akreditasi Institusi Perguruan Tinggi*. Jakarta: BAN-PT, 2011.

[2] M. S. Mustafa and I. W. Simpen, "Perancangan Aplikasi Prediksi Kelulusan Tepat Waktu Bagi Mahasiswa Baru Dengan Teknik Data Mining (Studi Kasus: Data Akademik Mahasiswa STMIK Dipanegara Makassar)," *Creat. Inf. Technol. J. (CITEC Journal)*, vol. 1, no. 4, 2014.

[3] A. Rohman, "Model Algoritma K-Nearest Neighbor (K-NN) Untuk Prediksi Kelulusan Mahasiswa," *Neo Tek.*, vol. 1, no. 1, 2015.

[4] A. Wanto, "Analisis Penerapan Fuzzy Inference System (FIS) Dengan Metode Mamdani Pada Sistem Prediksi Mahasiswa Non Aktif (Studi Kasus : AMIK Tunas Bangsa Pematangsiantar)," *Semin. Nas. Inov. Dan Teknol. Inf. 3*, vol. 3, pp. 393–400, 2016.

[5] I. M. B. Adnyana, "Prediksi Lama Studi Mahasiswa Dengan Metode Random Forest (Studi Kasus : STIKOM Bali)," *CSRID J.*, vol. 8, no. 3, pp. 201–208, 2015.

[6] D. Untari, "Data Mining Untuk Menganalisa Prediksi Mahasiswa Berpotensi Non-Aktif Menggunakan," Universitas Dian Nuswantoro, 2014.

[7] D. A. S. Arga, U. Lestari, and E. Sutanta, "Implementasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritme C4.5 (Studi Kasus: Jurusan Teknik Informatika, Institut Sains & Teknologi Akprind Yogyakarta)," *J. Risiko*, vol. 5, no. 2, pp. 1935–1943, 2017.

[8] M. N. Quadri and N. V. Kalyankar, "Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques," *Glob. J. Comput. Sci. Technol.*, vol. 10, no. 2, pp. 2–5, 2010.

[9] A. U. Khasanah and Harwati, "A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 215, p. 012036, Jun. 2017.

[10] G. S. Gowri, R. Thulasiram, and M. A. Baburao, "Educational Data Mining Application for Estimating Students Performance in Weka Environment," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 263, no. 3, 2017.

[11] B. Lantz, *Machine Learning with R*, 2nd ed. Birmingham-Mumbai: Packt Publishing Ltd., 2015.

[12] S. Zhang, X. Li, M. Zong, X. Zhu, and D. Cheng, "Learning *k* for kNN Classification," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 3, pp. 1–19, 2017.

[13] S. Mutrofin, A. Izzah, A. Kurniawardhani, and M. Masrur, "Optimasi Teknik Klasifikasi Modified K Nearest Neighbor Menggunakan Algoritma Genetika," *J. Gamma*, no. September, pp. 130–134, 2014.

[14] Indrayanti, D. Sugianti, and M. A. Al Karomi, "Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit Diabetes Mellitus," in *SNATIF*, 2017, pp. 823–829.

[15] N. Jagdish, D. Kumar, and A. Rajput, "An Empirical Comparison by Data Mining Classification Techniques for Diabetes Data Set," *Int. J. Comput. Appl.*, vol. 131, pp. 6–11, 2015.

[16] A. H. M. Ragab, A. Y. Noaman, A. S. Al-Ghamdi, and A. I. Madbouly, "A Comparative Analysis of Classification Algorithms for Student College Enrollment Apporal Using Data Mining," in *Proceedings of the 2014 Workshop on Interaction Design in Educational Environments*, 2014, p. 106.

[17] O. Somantri and S. Wiyono, "Model Data Mining Untuk Klasifikasi Tingkat Penguasaan Materi Bahan Ajar," in *Conference: Seminar Nasional Teknologi Informasi (SNTI)*, 2017.

[18] A. Ashari, I. Paryudi, and A. Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree

and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 11, pp. 33–39, 2013.