JURNAL RISET INFORMATIKA **P-ISSN: 2656-1743 |E-ISSN: 2656-1735**
Vol. 5, No. 4. September 2023 DOI: https://doi.org/10.34288/jri.v5i4.242
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

Performance Improvement of K-Nearest Neighbor Algorithm in KIP Scholarship Recipient Selection

Manzilur Rahman Romadhon[-1*], M. Faisal[-2], M. Imamudin[-3]

Magister informatics
UIN Maulana Malik Ibrahim
Malang, Indonesia
19841010@student.uin-malang.ac.id[1], mfaisal@ti.uin-malang.ac.id[2], imamudin@ti.uin-malang.ac.id[3]
(*)Corresponding author

Abstract

Law 12 of 2012 mandates that the government increase access to higher education for high achievers and underprivileged people. One of the efforts to realize this is by providing KIP Lectures. To ensure that beneficiaries are indeed eligible for KIP scholarships, it is necessary to classify scholarship recipients with data mining classification techniques correctly. The classification technique chosen is k-Nearest Neighbor (K-NN). K-NN is a classification method that relies heavily on the k parameter in carrying out classification. K-NN was applied to the KIP Scholarship applicant dataset at UIN Malang in 2022. The test scenario in this research is to compare the k-odd and k-even parameters to find the most optimal k value in K-NN. The highest accuracy value obtained by k-odd is 0.71 or 71% when k=9, and the highest for k-even is 0.67 or 67% when k=10. Using optimal k parameters is proven to improve k-NN performance. The K-NN algorithm with k-odd parameters, namely k=9, is the best method for classifying KIP scholarship recipients in this research. The results of this research can be considered in determining KIP scholarship recipients worthy of using K-NN.

Keywords: k-NN; Parameter of k; KIP

Abstrak

Undang-undang nomor 12 tahun 2012 memberikan mandat kepada pemerintah untuk meningkatkan akses pendidikan tinggi bagi Masyarakat berprestasi dan kurang mampu. Salah satu usaha untuk merealisasikannya adalah dengan pemberian KIP Kuliah. Untuk memastikan bahwa penerima manfaat memang layak mendapatkan beasiswa KIP, maka perlu dilakukan langkah yang tepat untuk melakukan klasifikasi penerima beasiswa dengan Teknik klasifikasi data mining. Teknik klasifikasi yang dipilih adalah K-Nearest Neighbor (K-NN). K-NN merupakan metode klasifikasi yang sangat bergantung pada parameter k dalam melakukan klasifikasi. K-NN diterapkan pada Dataset pendaftar Beasiswa KIP di UIN Malang tahun 2022. Skenario pengujian dalam penelitian ini ialah melakukan perbandingan parameter k-ganjil dengan k-genap untuk mencari nilai k yang paling optimal dalam K-NN. nilai akurasi tertinggi yang diperoleh k-ganjil adalah 0,71 atau 71% ketika k=9 dan rata-rata nilai akurasi tertinggi untuk k-genap adalah 0,67 atau 67% ketika k=10. Penggunaan parameter k yang optimal terbukti dapat meningkatkan kinerja k-NN. algoritma K-NN dengan parameter k=9 merupakan metode yang paling baik dalam penelitian ini dalam melakukan klasifikasi penerima beasiswa KIP. Hasil penelitian ini dapat dijadikan pertimbangan dalam menentukan penerima beasiswa KIP yang layak menggunakan K-NN.

Kata kunci: K-NN; parameter k; KIP

INTRODUCTION

Law No. 12/2012 on Higher Education has mandated the government to realize affordability and equitable equity in gaining access to quality higher education that is relevant to the interests of society for progress, independence, and prosperity. The government must increase access and learning opportunities and prepare intelligent and competitive Indonesians (Wibowo et al., 2022). One of the government's efforts to improve community learning access is providing Indonesia Smart Card (KIP) scholarships.

The selection process for receiving KIP scholarships has an essential role in efforts to support equitable and inclusive access to education. The Indonesia Smart Card (KIP) Tuition Scholarship is one of the scholarship channels

offered by the government to increase access to higher education for people who are outstanding and economically disadvantaged (Agama, 2020). With the increasing quota of KIP Tuition recipients, it is necessary to take accurate steps in determining the eligibility of KIP Tuition recipients. In this context, improving the performance of the K-Nearest Neighbors (KNN) algorithm in selecting KIP scholarship acceptance is an essential aspect of ensuring that students entitled to receive benefits from this program can be identified more precisely.

The k-NN algorithm was selected by searching for similar previous references to support this research. The first reference was testing the likelihood of a new patient developing diabetes. The study used k-NN to analyze a pile of historical data. The result is an accuracy value of 68.30% (Syukri Mustafa & Wayan Simpen, 2019). The following research was conducted to identify the type of flower by applying k-NN. The result with the value of k-7 obtained the most fantastic accuracy of 71% (Salsabila et al., 2021). Inna Alvi Nikmatun and Indra Waspada conducted research using data on the history of students who graduated to classify student study periods with the k-NN algorithm. The classification results obtained an accuracy value of 75.95% (Nikmatun & Waspada, 2019). The k-NN algorithm was also applied to predict stroke disease using 80 training and 20 test data. The selected attributes include gender, age, hypertension, heart history, and average glucose levels. The accuracy result obtained using the k=9 value is 95% (Maskuri et al., 2022).

Research on optimizing k parameters in k-NN for priority classification of village development aid was carried out by Saiful Ulya, M Arief Soeleman and Fikri Budiman. The results have proven that k-nn parameter optimization based on genetic algorithms can increase accuracy values compared to other methods. (Ulya et al., 2021). Several other studies have also performed k-optimization for medical datasets, village aid priority classification, and hotel occupancy rate prediction (Akbar & Kusumodestoni, 2020; Fuadah et al., 2019; Prasetio, 2020; Ulya et al., 2021). In the research to be carried out, the classification process is carried out by varying the k parameter and the size of the testing data to find the best classification results based on the accuracy value. The resulting output is a recommendation for the eligibility of KIP recipient candidates with the best k value in k-NN. Through the experiments conducted, it is hoped that this research will contribute to improving the performance of the k-NN algorithm in the context of KIP scholarship selection. The results of this research can guide relevant agencies to optimize the method, and scholarships are given to deserving candidates.

RESEARCH METHODS

This research uses several stages, starting from determining the dataset, preprocessing the dataset, creating a classification model, which includes varying the parameter k and the test data, and evaluating the model, as shown in Figure 1.

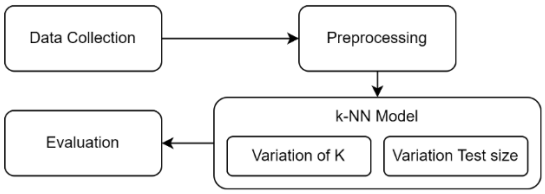

Figure 1. Research Flow Design

a. Data collection

Data collection is the stage of collecting data for the classification process using the k-NN model. The dataset used in the research is the dataset of KIP Lecture scholarship applicants at UIN Maulana Malik Ibrahim Malang in 2021. The dataset is obtained from the kip.uin-malang.ac.id website and combined with a dataset from the student affairs admin. The dataset contains 1044 records, seven feature attributes, and one target attribute are taken. Table 1 shows the identity of the dataset consisting of ID, father's salary, job, family members, average report value, academic achievement, electricity payment account, and recommendation.

Table 1. Identity of Scholarship Applicants KIP Lecture UIN Malang 2021

No	Attribute	Description	Type
1	Id	recipient's identity	Numeric
2	father's_salary	father's income	Categorical
3	job	father's occupation	Categorical
4	Family_members	parent's responsibility	Categorical
5	Average_report	average report card score	Categorical
6	Aca_achievement	academic achievement	Categorical
7	Electricity payment	electricity bill fees	Categorical
8	Recommendation	recommendation	Categorical

b. Preprocessing

Preprocessing is a series of techniques performed on a data set so that the data is not damaged, contains noise, and overcomes missing

values and other errors so the data is ready for further processing (Naufal et al., 2023). Preprocessing carried out in the research is the selection of columns that will be used as features and targets, the process of handling incorrect or inappropriate data so that the dataset that has been preprocessed is ready for classification using the k-NN algorithm.

Table 2 shows one of the preprocessing processes. The process is to replace the incorrect average_report data using the edit domain widget in the orange application. The domain editing stage is carried out after selecting the columns, the order of which can be seen in Table 2. This process is carried out so that the K-NN accuracy results are high.

Table 2 Process for handling inappropriate data

Edit	
Name	: average report
Type	: categorical
Values	: 07-Aug → 7 – 8 (merged) 08-Jul → 7 – 8 (merged)

c. k-Nearest Neighbor (k-NN) Algorithm

K-nearest neighbor (k-NN) is one of the algorithms for solving classification problems. The principle of the k-NN operation is to find the shortest distance between the data to be tested and the nearest neighbor in the training data (Nikmatun & Waspada, 2019). The K-nearest neighbor (K-NN) algorithm is one of the simplest algorithms for solving classification problems and often produces competitive and significant results (Prasetio, 2020). To calculate the distance, you can use the Euclidean distance. The Euclidean distance formula is specified in equation (1) (Id, 2021).

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad \text{.....(1)}$$

d. Evaluation

Evaluation is conducted to determine the performance of the k-NN algorithm on the KIP scholarship recipient dataset. The implementation of the k-NN algorithm used is a confusion matrix so that the accuracy, precision, and recall values can be known. Accuracy is the closeness between the predicted and actual values (Endang Etriyanti, 2021). Precision is defined as the ratio of selected relevant features to all selected features (Cahyanti et al., 2020). Precision can be understood as the correspondence between a request for information and the response to that request. Recall is the ratio of relevant items selected to the total number of relevant items available. The formulas for accuracy, precision, and recall are shown in equations 2, 3, and 4.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad \text{.....(2)}$$

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(3)}$$

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(4)}$$

Confusion matrix is a table containing the number of rows of test data predicted correctly and incorrectly by the classification model used. The confusion matrix table is needed to select the best performance of the classification model (Normawati & Prayogi, 2021). This method is often used for binary classification or multiclass classification. Confusion matrix is very suitable for research that measures the accuracy of model-based classification results that have been carried out. The confusion matrix table can be seen in Table 3 (Pratiwi et al., 2021).

Table 3. Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

RESULTS AND DISCUSSION

a. Testing

Figure 2 shows the simulation of k-NN model testing to find the most optimal k-NN model in classifying the eligibility of KIP Lecture scholarship recipients. Testing is done by varying the parameter k combined with various test data. The tool used for simulation is Orange 3.35. Orange Data Mining is an open-source data mining or analysis software through visual programming (Pradnyana, 2020).

The simulation flow starts from the CSV file import widget. The dataset is sent to the select columns device to determine the attributes used as features and target class attributes. The results of the unique columns process are sent to the edit domain widget, which functions to perform preprocessing. Preprocessing is done by handling missing data. Then, the data is sent to the data sampler widget to split into training and testing data. The model will be built using k-NN and sent to the test and score widget.

The test and score widget tests ten times according to the variation of the parameter k sent. Testing is done for odd k parameters and even k parameters. Training and testing data are obtained

from the split data process in the data sampler widget. The result is the accuracy value of each parameter k. After going through the test and score widget, evaluation is done in the confusion matrix widget.

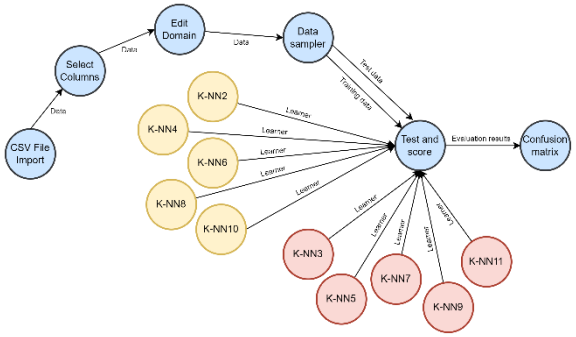

Figure 2 Simulation of k-NN Testing with k-odd and k-even

b. Analysis

The test results in Table 4 show the accuracy of k-NN for odd k parameters. The highest accuracy value is 0.71 or 71%, the precision value is 0.68 or 68%, and the recall value is 0.71 or 71%. This value is obtained when parameter k=9. The lowest accuracy, precision, and recall values are 0.68 or 68%, 0.63 or 63%, and 0.68 or 68%. The average accuracy, precision, and recall for the k-odd parameters are 0.70, 0.66, and 0.70. The trend shows that accuracy tends to be high and stable along with the large parameter k used. These results will be compared with the accuracy values for testing the even number k parameters.

Table 4. Accuracy, precision, and recall of k-NN Testing for k-odd

K	Accuracy	Precision	Recall
3	0,70	0,68	0,70
5	0,70	0,67	0,70
7	0,68	0,63	0,68
9	0,71	0,68	0,71
11	0,70	0,66	0,70
Average	**0,70**	**0,66**	**0,70**

The test results in Table 5 show the accuracy of k-NN for even k-parameters. The highest accuracy value is 0.67 or 67%, the precision value is 0.63 or 63%, and the recall value is 0.67 or 67%. This value is obtained when the parameter k=10 or k is highest. The lowest accuracy, precision, and recall values are 0.51 or 51%, 0.61 or 61%, and 0.51 or 51%. This value is obtained when the parameter k=2 or k is the lowest. The average accuracy, precision, and recall for the k-even parameters are 0.61, 0.63, and 0.61. The trend shows that the accuracy gain tends to be low but increases with the size of the k parameter. These results indicate that K-NN with k-odd parameters is superior to K-NN using k-even parameters.

Table 5. Accuracy, precision, and recall of k-NN Testing for k-even

k	Accuracy	Precision	Recall
2	0,51	0,61	0,51
4	0,58	0,63	0,58
6	0,64	0,65	0,64
8	0,63	0,62	0,63
10	0,67	0,63	0,67
Average	**0,61**	**0,63**	**0,61**

The confusion matrix widget in orange contains a confusion matrix table resulting from implementing the k-NN algorithm with various scenarios. Figure 3 shows the confusion matrix table produced by the k-NN algorithm using k=9 and dividing the training data into testing data 85 – 15. The amount of testing data used is 156 datasets. This number consists of 11 true positive data and 100 true negative data. Then, there were 35 false harmful data and ten classified as false positive. The accuracy, precision, and recall values of the k-NN algorithm can be calculated based on acquiring the confusion matrix.

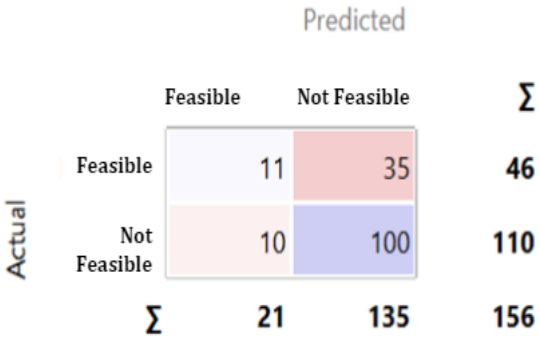

Figure 3 Confusion matrix k=9 15% test data

Figure 4 displays a graph comparing the accuracy values obtained by the k-NN model with k-odd and k-even parameters. The blue line shows the diagram for odd k-NN, and the orange line shows the graph of k-NN accuracy results for even k. the line for k-odd is always higher than the line for k-even. Even though the trend of k-even continues to increase with higher k parameters, it still cannot outperform the accuracy results of K-NN with odd k-parameters. This proves that k-NN with odd k parameters is still better than k-NN with even k parameters.

P-ISSN: 2656-1743 |E-ISSN: 2656-1735
DOI: https://doi.org/10.34288/jri.v5i4.242
Accredited rank 4 (SINTA 4), excerpts from the decision of the DITJEN DIKTIRISTEK No. 230/E/KPT/2023

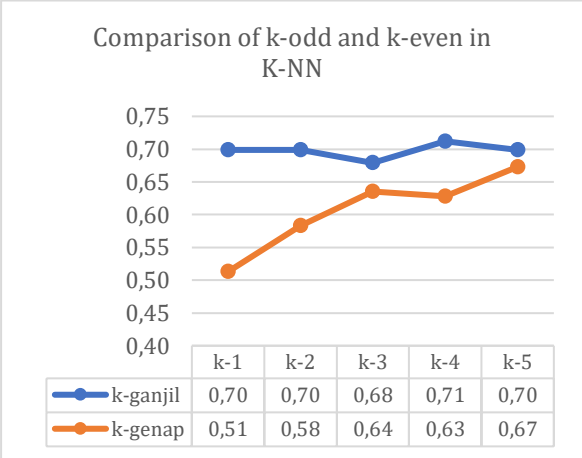

Figure 4 k-odd and k-even k-NN comparison

CONCLUSIONS

Based on the tests and analysis results, it can be concluded that the k-NN model with k-odd parameter values is superior to K-NN with k-even in classifying KIP Kuliah scholarship recipients. Several factors influencing accuracy, precision, and recall values are dataset quality, k parameters, data samples, training and testing data distribution, and attribute selection. Suggestions for further research are to consider several things to improve the results obtained.

REFERENCES

Agama, M. (2020). *KMA No. 361 Tentang Pedoman KIP Kuliah*. Kementerian Agama.

Akbar, A. S., & Kusumodestoni, R. H. (2020). Optimasi nilai k dan parameter lag algoritme k-nearest neighbor pada prediksi tingkat hunian hotel Optimization. *Jurnal Teknologi Dan Sistem Komputer*, *8*(3), 246–254. https://www.jtsiskom.undip.ac.id/index.php/jtsiskom/article/view/13648

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Endang Etriyanti. (2021). Perbandingan Tingkat Akurasi Metode Knn Dan Decision Tree Dalam Memprediksi Lama Studi Mahasiswa. *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, *3*(1), 6–14. https://doi.org/10.52303/jb.v3i1.40

Fuadah, Y. N., Magdalena, R., Palondongan, S., & Kumalasari, N. (2019). Optimasi K-Nearest Neighbor Untuk Sistem Klasifikasi Kondisi Katarak. *TEKTRIKA - Jurnal Penelitian Dan Pengembangan Telekomunikasi, Kendali, Komputer, Elektrik, Dan Elektronika*, *4*(1), 16–25. https://doi.org/10.25124/tektrika.v4i1.1832

Id, I. D. (2021). Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python. In *UNRI Press*. UNRI Press.

Maskuri, M. N., Harliana, H., Sukerti, K., & Bhakti, R. M. H. (2022). Penerapan Algoritma K-Nearest Neighbor (KNN) untuk Prediksi Penyakit Stroke. *Jurnal Ilmiah Intech: Information Technology Journal of UMUS*, *4*(01), 130–140. http://jurnal.umus.ac.id/index.php/intech/article/view/751

Naufal, M. F., Subrata, -, Susanto, A. F., Kansil, C. N., & Huda, S. (2023). Penerapan Machine Learning untuk Prediksi Potensi Hilangnya Nasabah Bank. *Techno.Com*, *22*(1), 1–11. https://doi.org/10.33633/tc.v22i1.7302

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. https://jurnal.umk.ac.id/index.php/simet/article/view/2882

Normawati, D., & Prayogi, S. A. (2021). Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter. *Jurnal Sains Komputer & Informatika (J-SAKTI*, *5*(2), 697–711. http://ejurnal.tunasbangsa.ac.id/index.php/jsakti/article/view/369

Pradnyana, G. A. (2020). *Data Mining Tanpa Skill Coding Menggunakan Orange*. Universitas Pendidikan Ganesha. https://www.youtube.com/watch?v=WOyfHJvq8_I&ab_channel=UniversitasPendidikanGanesha

Prasetio, R. T. (2020). Seleksi Fitur Dan Optimasi Parameter K-Nn Berbasis Algoritma Genetika Pada Dataset Medis. *Jurnal Responsif : Riset Sains Dan Informatika*, *2*(2), 213–221. https://doi.org/10.51977/jti.v2i2.319

Pratiwi, B. P., Handayani, A. S., & Sarjana, S. (2021). Pengukuran Kinerja Sistem Kualitas Udara Dengan Teknologi Wsn Menggunakan Confusion Matrix. *Jurnal Informatika Upgris*, *6*(2), 66–75. https://doi.org/10.26877/jiu.v6i2.6552

Salsabila, A., Yunita, R., & Rozikin, C. (2021). Identifikasi Citra Jenis Bunga menggunakan Algoritma KNN dengan Ekstrasi Warna HSV dan Tekstur GLCM. *Technomedia Journal*,

6(1), 124–137. https://doi.org/10.33050/tmj.v6i1.1667

Syukri Mustafa, M., & Wayan Simpen, I. (2019). Implementasi Algoritma K-Nearest Neighbor (KNN) Untuk Memprediksi Pasien Terkena Penyakit Diabetes Pada Puskesmas Manyampa Kabupaten Bulukumba. *SISITI: Seminar Ilmiah Sistem Informasi Dan Teknologi Informasi*, *8*(1), 1–10. https://ejurnal.dipanegara.ac.id/index.php/sisiti/article/view/1 -10

Ulya, S., Soeleman, M. A., & Budiman, F. (2021). Optimasi Parameter K Pada Algoritma K-NN Untuk Klasifikasi Prioritas Bantuan Pembangunan Desa. *Techno.Com*, *20*(1), 83–96. https://doi.org/10.33633/tc.v20i1.4215

Wibowo, C., Sukarno, S., Nursanti, Y. B., & Dewadi, F. M. (2022). Kebutuhan Perguruan Tinggi di Wonogiri sebagai Bagian dari Pengembangan Sumber Daya Manusia. *Visioner*, *4*(1 JUNI), 20–27. http://ojs.mputantular.ac.id/index.php/vis/article/view/660%0Ahttp://ojs.mputantular.ac.id/index.php/vis/article/download/660/501