International Journal of Computer and Information System (IJCIS)
Peer Reviewed - International Journal
Vol : Vol. 06, Issue 02, June 2025
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Performance Comparison K-Nearest Neighbors and Random Forest on Predicting The Performance New Polimedia Student Admissions

[1]Dwi Riyono, [2]Cholid Mawardi
[1,2]*Politeknik Negeri Media Kreatif, Jakarta, Indonesia*
Email : [1]dwirion@polimedia.ac.id, [2]cholid@polimedia.ac.id

Abstract *- New student admissions are at the forefront of the school's operational process. the success of each college's input stems from this. Polimedia always conducts new student admissions every year with various strategies used. Polimedia has 23 study programmes that can enable it to move in the creative industry that can be utilised by the community. in this study, a strategy using a prediction algorithm is used to be able to see the possible opportunities that occur if implemented in the coming year. with a dataset of 3738 data received by new students, an analysis will be carried out on prospective students who have re-registered or who have not re-registered. The classification model with 2 classes will be used. by conducting a data analysis process using exploratory data analysis (EDA) and also performing data cleansing so that the data modelling process runs well. The method used uses the main model of K-Nearest Neighbors by comparing with other machine learning models such as decision tree and random forest. It is expected that this research can produce high accuracy values 86.90% with powerful machine learning model comparisons. This research is also expected to be a reference for other studies that also conduct performance testing processes with machine learning models using various objects.*

Keyword*: Student, Performance, KNN, Machine Learning*

I. INTRODUCTION

Admission of new students is a priority for public universities including Politeknik Negeri Media Kreatif (Polimedia). New student admissions at polytechnics usually follow several stages and registration channels, depending on the rules of each institution In 2023 new student admissions, polimedia achieved an admission ratio over capacity of 72.2%. Meanwhile, in 2024, polimedia's new student admissions reached 87.7%, higher than the previous year. A prediction is used to estimate the results in a later period in the coming year and so on whether it is better or even less good than before. Artificial intelligence is needed to be able to classify data in the function of predicting a class in achieving the problems that occur [1]. The prediction method in machine learning can be used by determining the re-registration scheme or not who re-registered in the previous admission period. From these results, it will be given the benefit that the cause of the accuracy of the re-registration results can be used as a reference to be implemented in the admission of new students in the previous year.

Previous articles have discussed the Naïve Bayes and K-Nearest Neighbor Method Classification methods for Determining Poor Families from the author Riza Marsuciati, from the article concluded that the best method for classifying low-income families is the Naïve Bayes method [2]. Then research conducted by aisyah et al, in analyzing performance (accuracy, precision, recall and f-measure) on the image dataset of infected malaria and uninfected malaria [3]. Of course, from previous research, the KNN model or other machine learning models can be used to classify between two cold classes processed with several datasets that have been determined by class.

II. METHOD

In this study, researchers used two classification methods, namely Random Forest and k-Nearest Neighbor, where this study was looking for the classification method with the best performance to determine prospective new students who chose to re-register after being accepted and also chose not to re-register when they were accepted as prospective new students.

KNN (K-Nearest Neighbors) is one of the machine learning algorithms often used for classification and regression [4]. Although KNN is not a fundamental mathematical algorithm such as quadratic or integral equations, it does involve some mathematical concepts to calculate distances and determine nearest neighbors [5]. The main formula in KNN is to calculate the distance between two data points, which is generally done with the Euclidean Formula. Here is the Euclidean distance formula in KNN:

If there are two data points $A = (x1, y1)A$ dan $B = (x2, y2)$ in two-dimensional space, then the Euclidean distance d between them is:

$$d(A,B) = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \quad (1)$$

For higher dimensional spaces (e.g., 3D or beyond), the formula becomes:

$$d(A,B) = \sqrt{\begin{matrix}(x2 - x1)^2 + (y2 - y1)^2 + (z2 - z1)^2 2 + \\ \cdots + (n2 - n1)^2 2\end{matrix}} \quad (2)$$

Where n is the number of dimensions.

Meanwhile, random forest is a machine learning model that is included in the ensemble learning method[6]. This algorithm is used for both classification and regression problems. Random Forest works by building many decision trees during the training process, and produces a final decision by voting for classification or averaging for regression.

Each tree in Random Forest is trained using a random subset of the data obtained by bootstrapping[7]. This means the data is randomly selected with replacement. Mathematically, from the original dataset D of size N, we take random samples N times with replacement to get a new dataset D.

If $D = \{(x1, y1), (x2, y2), \ldots, (xN, yN)\}$, then D will be a subset of D, but may contain some duplication due to sampling with replacement.

If there are T trees in a Random Forest, the final prediction result for classification is the majority vote of all trees. Mathematically, for input x, the final prediction y is:

$$y = mode\{h1(x), h2(x), \ldots, hT(x)\} \quad (3)$$

Where $hi(x)h\_i$ is the prediction of the $-i$ tree, and mode is the class most frequently selected by the trees. In short, Random Forest is a combination of many decision trees that work together through bagging and random feature selection to improve model accuracy and prevent overfitting.

Data Preparation

The data to be processed comes from new student admission data when re-registering with 18 parameters, namely id, no_pendaftaran, nama_lengkap, jalur, major_id, skor, golongan, golongan_ditetapkan, golongan_ditetapkan_final, program_studi, role, status_survei, status_prodi, status_data_induk, status_biodata, daftar ulang, status_finalisasi, verificator_id.

a. Initial Process

Before the dataset is calculated, the data obtained needs to be processed first because not all data obtained is used in calculations in python [8]. There are two stages carried out in this preprocessing process, namely the data cleaning stage and the data selection stage [9].

Data Cleaning

In the dataset that has been obtained, there are 18 parameters, but in using the random forest classification model and also k-nearest neighbor, it is necessary to select parameters so that the calculation results of the two models are maximized. The data cleaning function to select from 18 parameters to 12 parameters, namely id no_pendaftaran, nama_lengkap, jalur, major_id, skor, golongan_ditetapkan_final, status_survei, status_prodi,status_data_induk, status_biodata dan daftar ulang. The data from this parameter reduction will be used to classify random forest and k-nearest neighbor (KNN)[10].

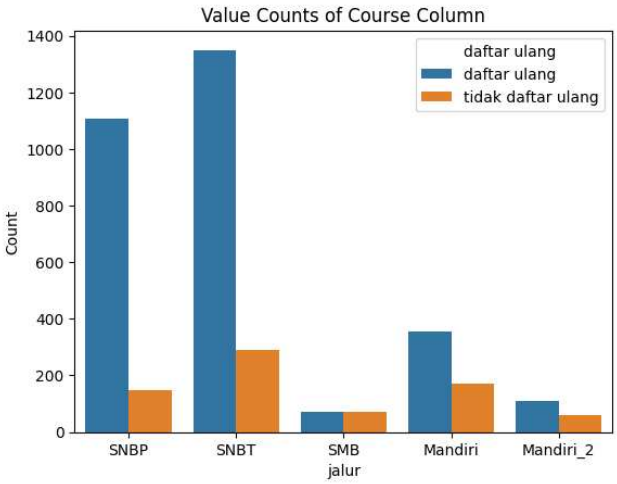

Figure 1. Re-Enrollment Classification Of Admission Pathways

Data Selection

From the dataset, there are several records that do not have values for each parameter as shown in Figure 2. To avoid errors or reduce the performance of the random forest and k-nearest neighbor classification models, it is necessary to do a dataset by selecting data by filling in data with empty parameter values with values already listed in the option [11]. Part of the data pre-processing is to analyze how numerical features relate to prospective students who re-enroll or not as well as other categories before proceeding to the modeling stage. The numerical features selected are 'skor', 'golongan', and 'jalur'. The variable numerical_features stores the names of these columns from the df_train dataset. These are the features that we want to analyze in more depth to see their distribution by status 'daftar ulang'.

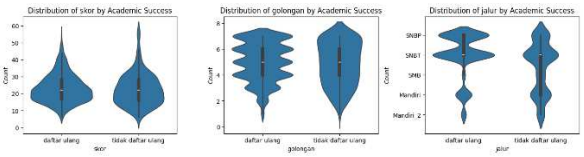

Figure 2. Violin Plot of The Distribution of New Student Candidate Variables

Evaluation Model

The research classification methods used are naïve bayes and k-nearest neighbor classification methods. Training data will use 20 percent of the data obtained, and the rest will be used as test data for 80 percent of the data obtained. Data processing uses python by adding validation in the form of performance to see the accuracy of the two classification methods. In performance, testing will use the confusion matrix method which consists of accuracy, precision, and recall[12]. Where in the performance matrix accuracy is used to test how accurate the classification model is.

$$Accuraccy = \frac{TP+FN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP+FN}{TP+TN+FP+FN}$$

$$Sensitivity = \frac{TP}{TP+FN}$$

$$F1-Score = \frac{TP+FN}{TP+TN+FP+FN}$$

International Journal of Computer and Information System (IJCIS)
Peer Reviewed - International Journal
Vol : Vol. 06, Issue 02, June 2025
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

III. RESULT AND DISCUSSION

Before testing using Python which will be used as test data and training data, preprocessing is first carried out, namely cleaning the data first from the original data which originally had 18 parameters which were reduced to 12 parameters. The parameters taken are adjusted to the needs of the random forest and k-nearest neighbor classification methods. Then data selection is carried out, namely replacing some empty values based on the choice of each parameter. The goal is to reduce errors and maximize performance.

Then prepare a dataset for training of 20% of the total data. And also the data used for testing is 80% of the total data. After the data is ready to use, then testing is carried out using the random forest classification method and also k- nearest neighbor. In testing, performance measurement is also added using the confusion matrix method which will produce accuracy, precision, and recall.

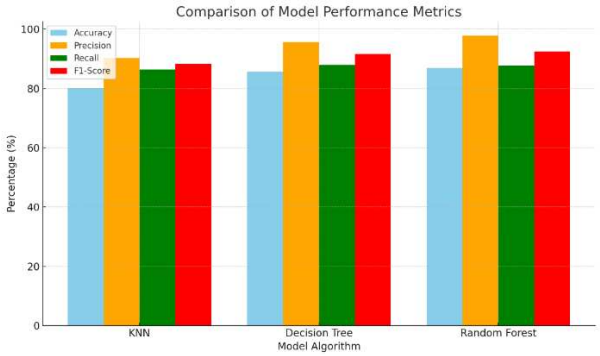

Figure 3. Comparison of Classification Testing Models

Figure 3 above shows the comparison results of the three different models. As an additional analysis, this research also adds another model as a comparison, namely decision tree.

Table 1. Comparison Results Of Model Testing With Other Machine Learning

Model Algorithm	Accuracy	Precision	Recall	F1-Score
KNN	80.21%	90.28%	86.36%	88.27%
Decision Tree	85.56%	95.62%	87.93%	91.61%
Random Forest	86.90%	97.73%	87.77%	92.48%

In table 1 above shows the results of the comparison of the three models, the accuracy value of random forest shows good results than the other two models with an accuracy of 86.90%. While the accuracy value of KNN only gets 80.21%. Decision tree gets a value of 85.56% only has a slight difference from random forest. This test uses the same parameters and data in terms of comparing with machine learning models. From various combinations of dataset ratios and experiments with various K values and other parameters, the highest accuracy value is obtained at a ratio of 80: 20 which is 86.90% but the recall value looks low because the accuracy of the classification is more inclined to precision than recall, meaning that the classification is less precise in estimating re-enrolled participants. The highest accuracy value in this study is found in the combination of 20:80 datasets and experiments with various K values which obtained an accuracy of 86.90%, where the precision shows a high value, meaning the balance of the classification process in predicting the prediction of new students who re-register.

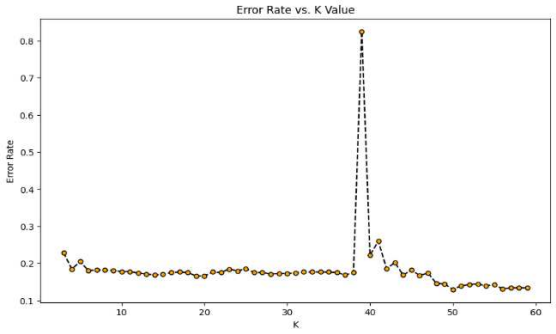

Figure 4. Error Rate At K-Value Knn

Figure 4 that you have shown illustrates the Error Rate against the K-Value of the K-Nearest Neighbors (KNN) algorithm. K in KNN is the number of nearest neighbors considered for classification or prediction. The value of K here ranges from 1 to about 60. At low values of K (about 5 to 10), the error rate seems to be quite low, with values around 0.18 to 0.2. As the K value approaches around K = 40, there is a very significant spike in the error rate which almost reaches 0.8. This indicates that at that K, the KNN model makes a lot of errors. After K = 40, the error rate again decreases drastically and stabilizes, with an error rate value of about 0.15 to 0.2 when K is in the range of 50-60. Based on this graph, the optimal K value selection is likely to be in the range of K = 10 to 30, depending on the balance between error rate and model complexity. Too large or too small a value of K results in poor model performance, as indicated by the drastic increase in error rate around K = 40.

IV. CONCLUSION

Based on the results of this study, the authors can conclude that in the range of K = 10 to 30, the error rate is quite low and stable, around 0.18 to 0.2, which indicates that the KNN model works quite well with a minimal error rate in that interval. In terms of overall model evaluation, the results using random forest show better accuracy than decision tree and KNN.

V. ACKNOWLEDGMENTS

The author would like to thank the Research and Community Service 2021 Funding from the Politeknik Negeri Media Kreatif, Ministry of Education, Culture, Research and Technology of Indonesia.

REFERENCES

[1] Mawardi C, Buono A, Priandana K, Herianto. Performance Analysis of ResNet50 and Inception-V3 Image Classification for Defect Detection in 3D Food Printing. *Int J Adv Sci Eng Inf Technol*. 2024;14(2):798-804. doi:10.18517/ijaseit.14.2.19863

[2] Marsuciati R, Gumelar G, Prietno R. Klasifikasi Metode Naïve Bayes dan K-Nearest Neighbor untuk Menentukan Keluarga Tidak Mampu. *Pros SISFOTEK*. 2021;5(1):246-249.

[3] Aisyah A, Anraeni S. Analisis Penerapan Metode K-Nearest Neighbor (K-NN) pada Dataset Citra Penyakit Malaria. *Indones J Data Sci*. 2022;3(1):17-29. doi:10.56705/ijodas.v3i1.22

[4] Chai BX, Eisenbart B, Nikzad M, et al. Process Prediction. Published online 2023.

[5] Zhang S, Li J. KNN Classification With One-Step Computation. *IEEE Trans Knowl Data Eng*. 2023;35(3):2711-2723. doi:10.1109/TKDE.2021.3119140

[6] Mienye ID, Sun Y. A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*. 2022;10(September):99129-99149. doi:10.1109/ACCESS.2022.3207287

[7] Han S, Kim H, Lee YS. Double random forest. *Mach Learn*. 2020;109(8):1569-1586. doi:10.1007/s10994-020-05889-1

[8] Performance A, Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, Moh S. The Effect of Preprocessing Techniques , Applied to Numeric. *Data*. 2021;6(11).

[9] Fitriyadi F, Muqorobin M. Prediction System for the Spread of Corona Virus in Central Java with K-Nearest Neighbor (KNN) Method. *Int J Comput Inf Syst*. 2021;2(3):80-85. doi:10.29040/ijcis.v2i3.41

[10] Koziarski M, Woźniak M, Krawczyk B. Combined Cleaning and Resampling algorithm for multi-class imbalanced data with label noise. *Knowledge-Based Syst*. 2020;204. doi:10.1016/j.knosys.2020.106223

[11] Cui L, Zhang Y, Zhang R, Liu QH. A Modified Efficient KNN Method for Antenna Optimization and Design. *IEEE Trans Antennas Propag*. 2020;68(10):6858-6866. doi:10.1109/TAP.2020.3001743

[12] Krstinić D, Braović M, Šerić L, Božić-Štulić D. Multi-label Classifier Performance Evaluation with Confusion Matrix. Published online 2020:01-14. doi:10.5121/csit.2020.100801