# The K-Nearest Neighbor Algorithm using Forward Selection and Backward Elimination in Predicting the Student's Satisfaction Level of University Ichsan Gorontalo toward Online Lectures during the COVID-19 Pandemic

**Andi Bode [a,1]; Zulfrianto Y Lamasigi [a,2]; Ivo Colanus Rally Drajana [a,3*]**

[a] Universitas Ichsan Gorontalo, Jl. Achmad Nadjamuddin No.17 Kota Selatan, 96138. Gorontalo
[1] andibode22@gmail.com; [2] zulfrianto.dns@unisan.ac.id; [3]ivocolanusrally@gmail.com
* Corresponding author

## Abstract

Academic services are actions taken by state and private universities to provide convenience for student's academic activities. During the current covid-19 pandemic, every university remains active in academic activities. This study aimed to apply the K-Nearest Neighbor algorithm in predicting the level of student satisfaction with online lectures at University Ichsan Gorontalo. Our main aim was to obtain quantitative information to measure student satisfaction with online lectures during the pandemic, which should be taken into account when making decisions. K-Nearest Neighbor is a non-parametric Algorithm that can be used for classification and regression, but K-Nearest Neighbor are better if feature selection is applied in selecting features that are not relevant to the model. Feature Selection used in this research is Forward Selection and Backward Elimination. Seeing the results of experiments that have been carried out with the application of the K-nearest Neighbor algorithm and the selection feature, the results of the forecasting can be used for consideration or policy in decision making. The highest level of accuracy in the K-Nearest Neighbor algorithm model used Forward Selection with an accuracy rate of 98.00%. Thus, the experimental results showed that feature selection, namely forward selection, was a better model in the relevant selection variables compared to backward elimination.

**Keywords:** K-Nearest Neighbor; Feature Selection; Prediction; Online Lecture

## Introduction

Higher education is one of the educational institutions that produce qualified human resources (HR) [1] Academic services are an effort made by every university to provide scientifically easy services to students. Students, lecturers, and infrastructure have a major impact on the quality of academic services in universities. The three factors influence each other in the implementation of higher education activities [2]. Covid-19 is currently a trending topic for news coverage in Indonesia, because almost all regions have been affected by this outbreak [3]. The problem that existed during the COVID-19 pandemic was that all academic activities, especially teaching and learning, were carried out at home with online system to maintain productivity. To find out and evaluate the results of online teaching and learning activities that have been implemented by the university, since not all students responded to online teaching and learning activities, a method that is able to make predictions is needed. Data mining is a method for finding knowledge in high-quality piles of data. Data extraction is divided into several groups according to their respective tasks, in particular: description, evaluation, forecast, classification, grouping [4]. The K-Nearest Neighbor algorithm is one of the approaches used in classification, the main concept of the K-Nearest Neighbor algorithm is to find the closest distance between the k objects closest to the test data. K-Nearest Neighbor compares the test data with the training data by looking for the sample data template that is closest to the test data. K-Nearest Neighbor has advantages over high-noise training data sets that are effective against large amounts of training data. However, K-Nearest Neighbor has drawbacks in determining the value of k in selecting the best features. Feature selection is used to select elements from the original attribute, eliminating unrelated properties [5]. Forward Selection a step-by-step rule for multiplying controlled variables one by one into a specific alpha-based input. Alpha is an input value that examine the probability to add a predictive device to the model that is not currently in the model. Backward Elimination method that has the

function to optimize the performance of the model with the opposite choice. Features or variables that are significant or significant to the model are excluded from the model [6].

## Method

### A. Data Mining

Data mining is the process of using statistics, mathematics, artificial intelligence, and machine learning techniques to extract or identify useful information from relevant databases for relevant knowledge [7]. The term data mining is a discipline whose main goal is to find, examine or extract knowledge from data or information available to us. Data mining is also known as Database Lookup (KDD). KDD is an activity that collects certain historical data and finds patterns, or relationships in large data sets [8].

### B. Predction

The rapid development of science and technology expands understanding of various aspects of the environment and causes many predictable events, and predicts events more precisely which make it possible to provide better input in future decision making [9]. Prediction is a process of systematically forecasting future events from information from the past [10].

### C. K-Nearest Neighbor

K-Neraest Neighbor algorithm is a classification method [11]. The K-Nearest Neighbor algorithm is a method of guided algorithm which is divided into two types, namely controlled and uncontrolled learning. Controlled learning algorithm aims to obtain new patterns, while uncontrolled learning algorithms aim to maintain data patterns. The accuracy of the evaluation algorithm is determined by the presence or absence of irrelevant data or when the attribute weights match. K-Nearest Neighbor algorithm is one of the classification analysis methods, however, it has also been used for forecasting in recent decades. In the training data, it was used to find the shortest distance between the data and the nearest neighbor. This room is divided into several groups based on the classification of training data. At this point, a point is denoted by class c if class c is the most common classification found around point k. The K-Nearest operating system training data is designed in a multidimensional space where the data characteristics of each dimension are presented [12]. The best k value in this algorithm depends on the data, a good k value can be selected based on the parameter [13].

The K-Nearest Neighbor Algorithm equation model is as follows:

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \qquad (1)$$

Information:
x = Sample Data
y = Test Data
D = Distance

$$fknn(x)^2 = \frac{k}{1} \sum_{i \in N_{k(x')}} y_1 \qquad (2)$$

X'            = Estimate or Estimate
K             = Number of Nearest Neighbors
Nk(x')       = Nearest Neighbor
Yi            = Nearest Neighbor Output

### D. Forward Selection

Forward Selection Inverting the independent variable X based on the ratio to find a perfect equation. Advanced selection begins at each stage by selecting empty attributes, such as adding unused attributes [14]. Forward selection begins with no variables then adds them one by one [15].

### E. Backward Elimination

Backward Elimination is a technique used to optimize the performance of backward models with the preferred operating system. Variables are selected by further selection, checking all variables and then eliminating variables that are considered insignificant. Variables are processed one by one. If these variables are believed to have a insignificant effect on the model, then these variables are excluded from the model [16]. Backward Elimination measures the impact of the elimination of a set of features [17].

## Results and Discussion

The source of the data set in this study was taken directly at the University Ichsan Gorontalo. The data collected using questionnaire for assessing student satisfaction with online lectures which were distributed directly to students

via google form. This study aimed to determine the performance of the K-Nearest Neighbor algorithm using a selection feature that aimed to obtain the best model in predicting student satisfaction with online lectures.

### A. Parameter K-Nearest Neighbor

At this stage to determine the distance or number of nearest neighbors, training was conducted by dividing the training data and testing data and then testing using the K-Nearest Neighbor method.

**Table 1.** Results Accuracy K-NN Algorithm

| Of Validation | K | Accuracy |
|:---:|:---:|:---:|
| 10 | 1 | 68,63 |
| 10 | 3 | 68,66 |
| 10 | 5 | 72,06 |
| 10 | 7 | 72,06 |
| 10 | 9 | 69,15 |
| 10 | 11 | 68,01 |
| 10 | 13 | 68,56 |

**Table 1** shows the results of student satisfaction with online lectures by determining the model with the k-nearest neighbor algorithm using cross-validation 10 and k-values of 1,3,5,7,9 and 13 based on the highest accuracy. From the experimental results, the best model based on the accuracy value was found at k 5 and 7 with a number of validation 10 with a result of 72.06. The result of this highest accuracy was chosen to be the best model, which can be used as a measure that determined the degree of similarity between the measurement results and the given value actually measured. The higher the level of accuracy, the more accurate a value was with the true value.
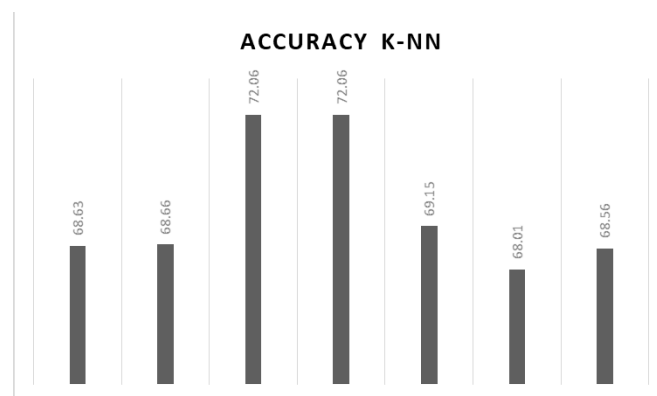


**Figure 1.** Comparison Graph Accuracy Value K-NN Algorithm

Based on **Figure 1**, it can be seen that the highest accuracy value is K 5 7 with an accuracy value of 72.06.

### B. Parameter K-NN Using Forward Selection

At this stage, experiments were carried out using the selection feature, namely the K-Nearest Neighbor algorithm using Forward Selection.

**Table 2.** Accuracy Results K-NN Algorithm Using Forward Selection

| Of Validation | K | Accuracy |
|:---:|:---:|:---:|
| 10 | 1 | 74.44 |
| 10 | 3 | 73.92 |
| 10 | 5 | 75.59 |
| 10 | 7 | 75.07 |
| 10 | 9 | 74.44 |
| 10 | 11 | 73.86 |
| 10 | 13 | 74.93 |

**Table 2** shows the experimental results of student's satisfaction from online lectures by determining the model using the k-nearest neighbor algorithm using cross-validation with a progress value of 13 k. The best model from the experimental results was made based on the accuracy. The highest accuracy was validation 10 k 5 with an accuracy of 75.59. The result of this highest accuracy was chosen to be the best model, which can

be used as a measure that determined the degree of similarity between the measurement results and the given value. The higher the level of accuracy, the more accurate a value is with the true value.
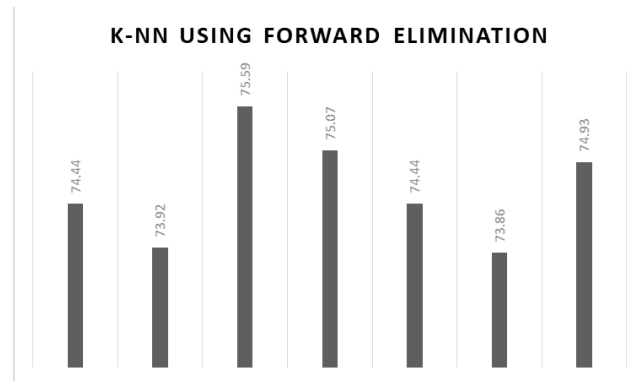


**Figure 2.** Comparison Graph K-NN Accuracy Values Using Forward Selection

Based on **Figure 2**, it can be seen that the highest accuracy value is in K5 with an accuracy value of 75.59

### C. *Parameter K-NN Using Backward Elimination*

At this stage an experiment was conducted with a selection function, namely the K-Nearest Neighbor algorithm with backward elimination feature selection.

**Table 3.** Results K-NN Accuracy Using Backward Elimination

| Of Validation | K | Accuracy |
|---|---|---|
| 10 | 1 | 72.09 |
| 10 | 3 | 71.01 |
| 10 | 5 | 73.95 |
| 10 | 7 | 73.89 |
| 10 | 9 | 72.78 |
| 10 | 11 | 72.09 |
| 10 | 13 | 71.54 |

**Table 3** shows the results of student satisfaction with their online lectures by determining the model using the k-nearest neighbor algorithm and the backward elimination selection feature. The validation 10 k 1,3,5,7,9,11 13  was observed based on the highest accuracy. The best model from the test results was made based on the highest accuracy value, namely the number of validation 10 k 5 with the accuracy value was 73.95. The result of this highest accuracy value was chosen to be the best model. It can be used as a measure that determined the degree of similarity between the measurement results and the given value measured. The higher the level of accuracy, the more accurate a value was toward the true value.
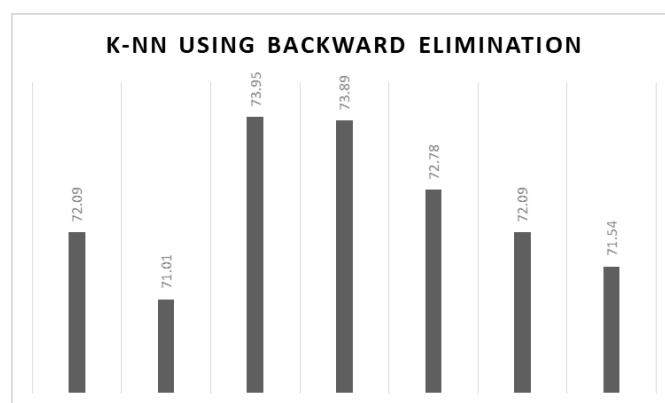


**Figure 3.** Comparison Graph K-NN Accuracy Values Using Backward Elimination

Based on **Figure 3**, the highest accuracy value was the value of k 5 with an accuracy of 73.95.

## D.  Evaluation

The selected model was based on the highest accuracy value obtained. Based on the experimental results, it was possible to determine the model that could be used to predict the student's satisfaction from their online lectures. The model with the highest accuracy can be seen in the table below:

**Table 4.** Model Comparison

| Model | Validation | K | Accuracy |
|---|---|---|---|
| K-NN | 10 | 5 | 72.06 |
| K-NN, Forward Selection | 10 | 5 | 75.59 |
| K-NN, Backward Elimination | 10 | 5 | 73.95 |

**Table 4** shows the results of the comparison of the K-Nearest Neighbor. The K-Nearest Neighbor algorithms using the forward selection and backward elimination features. The results showed that the best model was produced by the K-Nearest Neighbor algorithm with forward selection because its accuracy was 75.59.
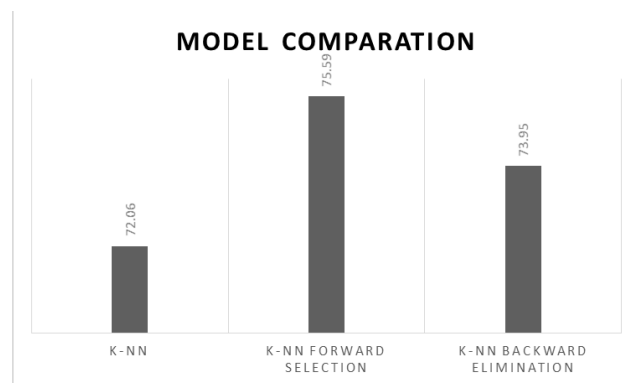


**Figure 4.** Comparison Graph K-NN and K-NN Models using Feature Selection

It can be seen in **Figure 4** the difference in the model. From the graph, it can be seen that the best model for student's satisfaction data towards online lectures was the K-Nearest Neighbor model using forward selection with an accuracy value of 75.59 at k 5 with a number of validation 10.

## E.  Implementation

At this stage, the application of the best model from the trial experiment was carried out with a sample data set of 5 records to predict the level of student satisfaction with online lectures, the results of predicting the level of student satisfaction with online lectures can be seen from the table as follows:

**Table 5.** Prediction Results

| Actual Data | Prediction Result |
|---|---|
| Very satisfied | Very satisfied |
| Satisfied | Satisfied |
| Very satisfied | Very satisfied |
| Less satisfied | Satisfied |
| Very satisfied | Very satisfied |

In accordance with **Table 5** above which is the result of predicting the level of student satisfaction with online lectures using the K-Nearest Neighbor algorithm using forward selection with a percentage accuracy rate of 98.00%. Thus the prediction results can be declared accurate and implemented at University Ichsan Gorontalo.

## Conclusion

Seeing the results of the experiments that have been carried out, the prediction of the student's satisfaction level with online lectures with the application of the K-Nearest Neighbor (K-NN) algorithm and the selection feature has been successfully carried out. Then the results of the forecast can be used for consideration or policy in decision making. The highest level of accuracy in the K-NN algorithm model used Forward Selection with an accuracy rate of 98.00%. Thus, the experimental results showed that feature selection, namely forward selection, was a superior model

in selecting the relevant variables, compared to backward elimination. The K-NN and forward selection algorithms have obtained better performance than the K-NN and backward elimination algorithms.

It is recommended for future research to use other selection features to improve the performance of the K-Nearest Neighbor (K-NN) algorithm with the aim of a higher level of accuracy.

## References

[1] Isa, Indra Griha Tofik, And Febie Elfaladonna. "Penilaian kinerja akurasi metode Klasifikasi dalam dataset penerimaan mahasiswa baru Universitas Xyz." *jepin (jurnal edukasi dan penelitian informatika)* 8.2: 292-298. 2022.

[2] Gustientiedina, g., siddik, m. And deselinta, Y.,. "Penerapan Naïve Bayes untuk memprediksi tingkat kepuasan mahasiswa terhadap pelayanan akademis". *Jurnal infomedia: teknik informatika, multimedia & jaringan*, *4*(2), pp.89-93. 2019.

[3] Rahmaddeni, Rahmaddeni, Et al. "Comparison of Support Vector Machine and Xgbsvm in analyzing public opinion on COVID-19 Vaccination" *ilkom jurnal ilmiah* 14.1. 2022.

[4] Julianto, Indri Tri, Et al. "Comparison of Data Mining algorithm for Forecasting Bitcoin Crypto Currency Trends." *jurnal teknik informatika (jutif)* 3.2. 2022: 245-248.

[5] Bode, A.,. "K-Nearest Neighbor dengan Feature Selection menggunakan Backward Elimination untuk prediksi harga komoditi kopi arabika". *Ilkom jurnal ilmiah*, *9*(2), pp.188-195. 2017.

[6] Drajana, I.C.R.,. "Metode Support Vector Machine dan Forward Selection prediksi pembayaran pembelian bahan baku kopra". *Ilkom jurnal ilmiah*, *9*(2), pp.116-123. 2017.

[7] Lasulika, Mohamad Efendi, And Andi Bode. "Komparasi algoritma data mining menggunakan Forward Selection pada prediksi harga jagung." *jurnal tecnoscienza* 5.2. 2021: 157-172

[8] Fitriani, Yetti, Sarjon Defit, And Gunadi Widi Nurcahyo. "Prediksi hasil belajar siswa secara daring pada masa pandemi COVID-19 menggunakan metode C4. 5." *jurnal sistim informasi dan teknologi*. 2021: 120-127.

[9] Boy, Ahmad Fitri. "Implementasi Data Mining dalam memprediksi harga crude palm oil (cpo) pasar domestik menggunakan algoritma Regresi Linier Berganda (Studi Kasus Dinas Perkebunan Provinsi Sumatera Utara)." *journal of science and social research* 3.2. 2020: 78-85.

[10] Larassati, Devina, Ati Zaidiah, And Sarika Afrizal. "sistem prediksi penyakit jantung koroner menggunakan metode Naive Bayes." *jipi (jurnal ilmiah penelitian dan pembelajaran informatika)* 7.2. 2022: 533-546.

[11] Erdiansyah, Umri, Ahmadi Irmansyah Lubis, And Kamil Erwansyah. "Komparasi metode K-Nearest Neighbor dan Random Forest dalam prediksi akurasi klasifikasi pengobatan penyakit kutil." *jurnal media informatika budidarma* 6.1. 2022: 208-214.

[12] Nabila, Silvana Puspa, Nurissaidah Ulinnuha, And Ahmad Yusuf. "Model Prediksi kelulusan tepat waktu dengan metode Fuzzy C-Means dan K-Nearest Neighbors menggunakan data registrasi mahasiswa." *network engineering research operation* 6.1. 2021: 38-46.

[13] Jumarlis, Mila, Mirfan Mirfan, And Abdul Rachman Manga. "Classification of Coffee Bean Defects using Gray-Level Co-Occurrence Matrix and K-Nearest Neighbor." *ilkom jurnal ilmiah* 14.1. 2022.

[14] Hasan, M.,. "Prediksi tingkat kelancaran pembayaran kredit bank menggunakan algoritma Naive Bayes berbasis Forward Selection". *Ilkom jurnal ilmiah*, *9*(3), pp.317-324. 2017.

[15] Sanjaya, Rangga, And Fitriyani Fitriyani. "Prediksi bedah toraks menggunakan seleksi fitur Forward Selection dan K-Nearest Neighbor." *jepin (jurnal edukasi dan penelitian informatika)* 5.3. 2019: 316-320

[16] Widhiantoyo, Abimanyu, Betha Nurina Sari, And Dadang Yusuf. "Penerapan algoritma Naïve Bayes dengan Backward Elimination untuk prediksi waktu tunggu alumni mendapatkan pekerjaan." *jiko (jurnal informatika dan komputer)* 4.3. 2021: 145-151.

[17] Harafani, Hani, and H. Aji Al-Kautsar. "Meningkatkan kinerja K-Nn untuk Klasifikasi Kanker Payudara dengan Forward Selection." *jurnal pendidikan teknologi dan kejuruan* 18.1. 2021: 99-110.