 ISSN:2302-8149
e-ISSN:2540-9719

Integrating K-Means Clustering and K-Nearest Neighbor Classification for Effective Scholarship Recipient Selection

[1]Suandi Daulay, [2]Rizky Wandri*
[1]Information Systems, Faculty of Engineering and Pekanbaru College of Technology
[2]Informatics Engineering, Faculty of Engineering and Islamic University of Riau
[1]Jl. Dirgantara No.4, Sidomulyo Tim., Kec. Marpoyan Damai, Pekanbaru, Riau, Indonesia
[2]Jl. Kaharuddin Nst No.113, Simpang Tiga, Kec. Bukit Raya, Pekanbaru, Riau, Indonesia
*e-mail: rizkywandri@eng.uir.ac.id

(**received:** 17 November 2024, **revised:** 11 January 2025, **accepted:** 20 January 2025)

Abstract

This research is important because public interest in the KIP Kuliah Scholarship continues to increase. However, many educational institutions still use manual selection which is prone to bias and less effective in data management. Therefore, a method is required to make the selection process more efficient; the K-Means and K-Nearest Neighbor methods are two data processing methods that have been proven effective in various applications, including in the field of data processing. In this study, the K-Means and K-Nearest Neighbor methods are used to select scholarship recipients to increase efficiency in the process. Based on the processing carried out, there were 1257 participants who were then grouped into three clusters: Cluster 0 with 739 data points, Cluster 1 with 290 data points, and Cluster 2 with 228 data points. Testing using the K-Nearest Neighbor algorithm was carried out by evaluating the appropriate k values, specifically 27, 31, 35, 41, 45, and expanded to 185 to obtain the optimal value, namely K-155 and produced as many as 155 people who were deemed worthy and qualified according to the specified criteria. The combination of K-Means and K-NN algorithms resulted in an accuracy of 89.72% accomplished in 16 seconds. This combo can recognize data with excellent accuracy in a fast time while minimizing errors. The test results suggest that this technique is effective in selecting applicants based on the criteria and quotas established, thus it can be used as a guideline for future selection.

Keywords: Scholarship selection, smart indonesia program, KIP college scholarship, k-means method, k-nearest neighbor method.

1 Introduction

The Smart Indonesia Program (PIP) is assistance in the form of cash [1] and learning opportunities the government provides to students from underprivileged families [2], [3], [4]. The government is committed to placing access to higher education for the entire community as a development priority [5]. One of the PIP programs is the KIP College Scholarship, which replaces the Bidikmisi scholarship [6]. This scholarship aims to increase access to and opportunities to study at universities and prepare intelligent and competitive Indonesian people [7]. This scholarship program provides students with tuition fees and living expenses for up to 8 semesters [8] of study, which has increased public interest in it rapidly.

The development of technology in the current digital era [9] has become an important aspect for social and economic development [10], in supporting accessibility to education. One way to support accessibility is by providing scholarships to students who excel but are financially disadvantaged. However, in its implementation, the scholarship selection process is crucial. In the field of computer science, data processing methods are tools to assist in various data processing processes to obtain more precise and objective results. The increasing use of information technology creates a great opportunity to improve efficiency in various aspects of administration in higher education, one of which is in the selection process for scholarship recipients. With the presence of information technology, the process becomes more efficient in its completion. However, many educational institutions still use manual selection processes, which take a long time [11], [12] and are not effective in the data management process. Based on this, the Problem Formulation that will be used in this study is obtained: first, how to carry out an efficient selection process; second, how to process data in

 ISSN:2302-8149
e-ISSN:2540-9719

a technology-based selection process; and third, how efficient the technology-based data processing process would be if implemented.

Therefore, a data processing method is needed to complete the process efficiently. Researchers will use a combination of two methods to obtain accurate results. The K-Means method is a clustering algorithm [13], [14] that partitions data by performing an iterative process in forming data groups [15], [16] through a series of iterative partitions to reduce the average distance between each data and the corresponding cluster center [17]. The K-Nearest Neighbor method is one of the classification techniques [18], [19] included in supervised learning [20], which performs the proximity of the location (distance) of data to other data [21], [22]. The K-Nearest Neighbor algorithm can select or reduce data features at the pre-processing stage [23]. K-Nearest Neighbor classifies unknown data by finding the *k* nearest data points [24]. The optimal *k* value is contingent upon the dataset; generally, a higher *k* diminishes noise in classification but renders the borders between classifications progressively indistinct [25]. Data processing methods such as K-Means and K-Nearest Neighbor have begun to be applied in various contexts outside the field of information technology, including in the scholarship recipient selection process.

In recent years, using the K-Means and K-Nearest Neighbor methods can be one solution in data processing. However, research that focuses on integrating these two methods still needs to be improved by using more than one method and comparing specific methods. Combining the two methods is expected to provide results with a better level of accuracy than using one method. This study's novelty is developing a model that combines both methods to increase efficiency in the scholarship selection process, which can be used as a reference for future selection processes. The urgency of this research is to help the administrative process in the scholarship recipient selection process with a method that will calculate the level of efficiency in processing scholarship applicant data. These results will determine how accurately the method can help the administrative process in the scholarship recipient selection process.

2 Literature Review

Research conducted by [26] indicates a trend in employing classification algorithms to enhance efficiency and objectivity in the scholarship selection process. This research uses the K-Nearest Neighbors (KNN) algorithm to assess student eligibility based on historical data from 350 students for the 2022-2023 academic year. The KNN algorithm, employing Euclidean distance as a similarity metric, attains a prediction accuracy of up to 93%. This methodology is executed in Python on Google Colab, encompassing data normalisation, division into training data (75%) and test data (25%), and model assessment via a confusion matrix. This study highlights that KNN is a straightforward yet effective method, rendering it a suitable choice for enhancing data-driven decision-making in scholarship administration selection. This study has limitations, including a constrained data set and a singular emphasis on one algorithm without comparative evaluation against alternative methodologies. Furthermore, the utilisation of the 2022-2023 dataset implies that testing on more recent datasets has not occurred.

Prior research by [13] employed the k-means and C4.5 algorithms, utilising application data from 2022. This study adeptly integrates two methodologies, specifically clustering and classification, to achieve superior outcomes. The outcomes derived from manual testing and the utilisation of the RapidMiner program are coherent and valid in accordance with the results of the application testing. A total of 1289 individuals were classified into three clusters based on the conducted processing. The results indicated that cluster 0 comprised 327 individuals, all of whom were students with high scores. The resultant decision tree demonstrates the subsequent pattern: A participant with an Indonesia Smart Card who achieves a score over 70 points on the "Total Income" criterion qualifies for a scholarship. This study has a limitation, specifically the lack of a conducted test evaluation.

Previous research had flaws, such as not using the most recent data or using a combination of clustering and classification algorithms in an integrated manner. The current work aims to close this gap by analysing the 2023 data set and evaluating the effectiveness of combining clustering and classification approaches. Furthermore, this study emphasises the significance of validating the results using an evaluation technique to ensure that the predictions are correct. Thus, this study not only

 ISSN:2302-8149
e-ISSN:2540-9719

expands on prior findings, but it also offers a more complete and up-to-date strategy for selecting scholarship applicants.

3 Research Method

This study aims to process data in the scholarship recipient selection process using the K-Means and K-Nearest Neighbor methods. The results of this study are expected to be a reference to help the administrative process in higher education. In this study, researchers used applicant data in 2023. The data will be processed according to the Knowledge Discovery in Database (KDD) process as follows:

a. Data Cleansing
The initial part of the KDD process, data cleansing, is to find and correct flaws or inconsistencies within the dataset. This stage is essential, as data quality directly affects the results of following analysis. Efficient data cleansing guarantees that the data utilised for analysis is precise and dependable [27]. This analysis utilises scholarship application data until 2023.

b. Data Integration
Subsequent to data cleansing, data integration entails amalgamating data from many sources to form a cohesive dataset. This stage is essential in KDD since it facilitates a more thorough analysis by utilising varied datasets. Integrating diverse data sources can enhance the analytical context, especially in intricate situations like climate data research [28].

c. Data Selection
The procedure for selecting data deemed pertinent to the study in accordance with the regulations of the Ministry [29].

d. Data Transformation
The data transformation process is defined as the process of changing data into a form that is appropriate to the form required by the mining procedure [30].

e. Data Mining
The extraction of knowledge through the application of specific algorithms, methods, and techniques [31]. Data processing will be conducted in this study using the K-Means and K-Nearest Neighbor methods to identify prospective patterns that generate valuable data.

1. K-Means

Step 1. Get the data sampling ready
This phase is essential, as the quality and pertinence of the data directly affect the clustering outcomes. K-means clustering was utilised to classify courses according to student enrolment, highlighting the significance of meticulously collected data for efficient clustering [32].

Step 2. Ascertain how many clusters there are
Subsequently, ascertaining the quantity of clusters is a crucial step in the K-means procedure. Multiple techniques are available to determine the ideal number of clusters (k). Assessment of various indices for identifying the number of clusters, providing significant insights into the selection procedure [33].

Step 3. Find the centre point or centroid value
After establishing the number of clusters, the subsequent step is to identify the initial centroids or cluster centres. The choice of centroids profoundly influences convergence and the ultimate clustering outcomes [34].

Step 4. Determine each centroid distance
Subsequent to centroid initialisation, the algorithm computes the distance from each data point to the centroids. The predominant metric employed is the Euclidean distance, which measures the similarity between data points and centroids. This phase is essential as it determines the allocation of data points to clusters [35].

Step 5. Use the shortest distance to group data
Finally, data points are categorised according to their proximity to the centroids, resulting in the establishment of clusters. This process is iterative, since the

 ISSN:2302-8149
e-ISSN:2540-9719

algorithm adjusts the centroids according to the current cluster assignments until convergence is reached [36], [37].

2. K-Nearest Neighbor

K-Nearest Neighbour (K-NN) is a straightforward yet efficient classification technique employed in numerous machine learning and data processing applications. It functions on the premise that analogous objects are likely to be situated near one another in feature space. This is a comprehensive elucidation of the phases of the K-NN algorithm:

Step 1. Prepare the Data to be Classified

The initial phase entails the preparation of the dataset designated for classification. This entails choosing pertinent features and verifying that the data is sanitised and appropriately prepared for analysis. The quality of the data profoundly affects the efficacy of the K-NN algorithm, as it depends on the proximity of data points within the feature space [38], [39].

Step 2. Determine the Number of Nearest Neighbors (k)

The subsequent step is to specify the parameter (k), which denotes the quantity of nearest neighbours to be taken into account during classification. The selection of (k) is essential; a diminutive (k) can result in susceptibility to noise, whilst an excessive (k) may obscure the differences between classes [40], [41]. To determine the value of (k) in the K-Nearest Neighbours algorithm, a straightforward and commonly used method is as follows [42]:

$$k = \sqrt{n} \quad (1)$$

Description:

n = The total number of data in the dataset.

Step 3. Calculate Distance with Euclidean Distance

After the data is produced and (k) is established, the method computes the distance between the new data point and every point in the training dataset. The Euclidean distance is the predominant metric utilized for this purpose, although alternative metrics such as Manhattan distance may also be employed depending on the context [43], [44]. Utilise the Euclidean distance formula to compute the distance as outlined below [42]:

$$d_{ij} = \sqrt{(x_{1j} - x_{1j})^2 + (x_{2j} - x_{2j})^2 + \cdots + (x_{ki} - x_{kj})^2} \quad (2)$$

Description:

d_{ij} = Distance from data point i to cluster centroid j

x_{kj} = Data from i to k data attribute

x_{kj} = Data from j to k data attribute

Step 4. Sort the Distance Calculation Results

Subsequent to computing the distances, the following step is to arrange these distances in ascending order. This sorting enables the program to determine which data points are nearest to the new instance requiring classification [45].

Step 5. Take the Value of the Nearest Neighbor

The method finds the k nearest neighbours from the ordered list of distances. This decision is predicated on the minimal distance values, which denote the nearest data points to the new instance [46].

Step 6. Perform New Object Classification

Finally, the method categorises the new data point according to the predominant class among the (k) nearest neighbours. This majority voting process is a straightforward and efficient method for ascertaining the class label of a new instance, as it utilises the pooled expertise of the nearest neighbours [47].

f. Knowledge Presentation

The concluding phase of the KDD process is knowledge presentation, which entails exhibiting the outcomes of the data mining process in a comprehensible and practical format for users. This step is crucial as it dictates the efficacy of communicating insights derived from the data to

 ISSN:2302-8149
e-ISSN:2540-9719

stakeholders. Proficient knowledge dissemination is essential for facilitating informed decision-making grounded in analysis [48].

4 Results and Analysis

The objective of this investigation is to ascertain the likelihood of potential scholarship recipients by employing the K-Means and K-Nearest Neighbors algorithms. The findings of this investigation are anticipated to assist educators in obtaining predictions during the selection process. In this section, researchers employ data from scholarship applicants who will be awarded scholarships in 2023. Knowledge Discovery in Database (KDD) will be implemented to evaluate the data. The K-Means process and the K-Nearest Neighbors algorithm will be implemented in the subsequent stage after the data is prepared for data mining. Table 1 shows data that has gone through the KDD process up to the data transformation stage, as follows:

Table 1. Data set (transformation process)

Initials	DTKS	P3KE	KIP	…	Father's & Mother's Income	Home Ownership
P-1	100	75	0	…	100	100
P-2	100	0	0	…	80	20
…	…	…	…	…	…	…
P-1256	0	100	0	…	100	80
P-1257	100	100	0	…	100	100

To reduce data into groups or clusters of all data, the initial mining procedure implements the k-means algorithm.

Step 1. Determine the number of clusters in the dataset, which is three.

Step 2. Determine the center value (centroid) using the elbow method.

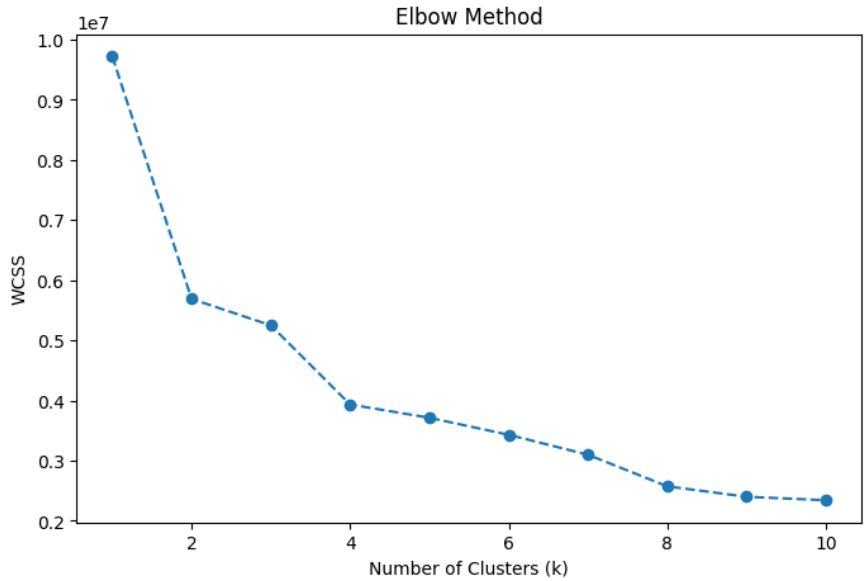

Figure 1. Results of the elbow method in python

Figure 1 of the graph indicates that the elbow occurs at k = 3. The ideal decision is three clusters, as indicated by the elbow graph, derived from selecting three random data points from the existing dataset. Table 2 presents the randomly assigned centroid values.

Table 2. Initial centroid value

Initials	DTKS	P3KE	KIP	…	Father's & Mother's Income	Home Ownership

 ISSN:2302-8149
e-ISSN:2540-9719

C1	100	50	0	…	100	100
C2	0	100	0	…	80	20
C3	0	62,5	0	…	100	100

Step 3. Determine the object's Euclidean distance from the centroid.

$$C(1) = \sqrt{\begin{matrix}(a_1-c_1)^2+(b_1-c_1)^2+(c_1-c_1)^2\\+(d_1-c_1)^2+(e_1-c_1)^2+(f_1-c_1)^2\\(g_1-c_1)^2+(h_1-c_1)^2+(i_1-c_1)^2+\\(j_1-c_1)^2+(k_1-c_1)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-100)^2+(75-50)^2+\\(0-0)^2+(0-0)^2+\\(50-50)^2+(33{,}3-33{,}3)^2+\\(100-50)^2+(100-33{,}3)^2+\\(20-10)^2+(100-100)^2+\\(100-100)^2\end{matrix}}$$

$$= 87{,}58$$

$$C(2) = \sqrt{\begin{matrix}(a_1-c_2)^2+(b_1-c_2)^2+(c_1-c_2)^2\\+(d_1-c_2)^2+(e_1-c_2)^2+(f_1-c_2)^2\\+(g_1-c_2)^2+(h_1-c_2)^2+(i_1-c_2)^2\\+(j_1-c_2)^2+(k_1-c_2)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-0)^2+(75-100)^2+\\(0-0)^2+(0-0)^2+\\(50-75)^2+(33{,}3-33{,}3)^2+\\(100-100)^2+(100-33{,}3)^2+\\(20-40)^2+(100-80)^2+\\(100-20)^2\end{matrix}}$$

$$= 151{,}31$$

$$C(3) = \sqrt{\begin{matrix}(a_1-c_3)^2+(b_1-c_3)^2+(c_1-c_3)^2\\+(d_1-c_3)^2+(e_1-c_3)^2+(f_1-c_3)^2\\+(g_1-c_3)^2+(h_1-c_3)^2+(i_1-c_3)^2\\+(j_1-c_3)^2+(k_1-c_3)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-0)^2+(75-62{,}5)^2+\\(0-0)^2+(0-0)^2+\\(50-100)^2+(33{,}3-33{,}3)^2+\\(100-100)^2+(100-33{,}3)^2+\\(20-40)^2+(100-100)^2+\\(100-100)^2\end{matrix}}$$

$$= 132{,}29$$

Step 4. Arrange items in order of their distance from the adjacent centroid.

The following are the results of the data calculation at the centroid center point for each existing cluster.

Table 3. First iteration results of centroid calculation for each cluster and the shortest distance

Initials	C1	C2	C3	Shortest Distance
P-1	87,58	151,31	132,29	C1
P-2	147,99	32,02	105,39	C2

 ISSN:2302-8149
e-ISSN:2540-9719

…	…	…	…	…
P-1256	129,03	70,89	58,79	C3
P-1257	80,78	129,61	109,69	C1

Table 3 displays the results of computing the centroid of each cluster and the shortest distance from the data set. The initial computation of the distance between the data and the cluster center point yields the findings shown in Table 4.

Table 4. First iteration cluster results

Cluster	Results
C1	736
C2	342
C3	179

Step 5. Repeat steps 3–4 until the centroid is at its most optimal

Determine the new center point by combining the data from each cluster member. Table 5 displays the updated center centroid values based on cluster results.

Table 5. New centroid values

Initials	DTKS	P3KE	KIP	…	Father's & Mother's Income	Home Ownership
C1	100	72,57	22,28	…	99,62	98,83
C2	0,58	80,26	2,05	…	72,05	30,41
C3	0,56	63,97	1,12	…	90,50	88,38

Then, in the second iteration, calculate the shortest distance of the data and the new centroid value with the data set value. And in the second iteration, compute the distance of the data to the cluster center point. The iteration procedure is terminated when the centroid value obtained from the previous iteration is either equal to the current value or is optimal (i.e., does not change). This technique concludes at the fifth iteration, yielding the cluster result values presented in Table 6 below:

Table 6. Fifth iteration cluster results

Cluster	Results
C1	739
C2	290
C3	228

The outcomes of data processing via the k-means algorithm align with the results derived from the Python programming language, as illustrated in Figure 2 below, which specifically depicts the existence of three clusters (1, 2, and 3), with the quantity of items corresponding to the test results.

```
Jumlah item di Cluster 1: 739
Jumlah item di Cluster 2: 290
Jumlah item di Cluster 3: 228
```

Figure 2. Results of the fifth iteration cluster in python

The subsequent phase will entail analyzing the first data mining outcomes with the K-Nearest Neighbors method to generate a decision tree from the processed data. In the subsequent phase, the initial data mining outcomes will be analyzed utilizing the K-Nearest Neighbors

 ISSN:2302-8149
e-ISSN:2540-9719

(KNN) algorithm to provide predictions depending on the configuration of the processed data. The subsequent dataset comprises the data findings from cluster 1, derived using the k-means algorithm, as presented in table 7 below:

Table 7. Dataset derived from K-means outcomes (cluster 1)

Initials	DTKS	P3KE	KIP	…	Father's & Mother's Income	Home Ownership
P-1	100	75	0	…	100	100
P-14	100	50	0	…	100	100
…	…	…	…	…	…	…
P-1255	100	100	100	…	100	100
P-1257	100	100	0	…	100	100

Step 1. Determine the parameter K (the number of closest neighbors).
To ascertain the value of k, a basic principle is to select k as the square root of the data quantity utilizing formula (1).

$$k = \sqrt{n}$$
$$k = \sqrt{739}$$
$$k = 27.18 = \mathbf{27}$$

To determine the appropriate k value, a model performance evaluation will be conducted using cross-validation based on the dataset. The cross-validation procedure will involve testing several k values (27, 31, 35, 41, 45, …, 185) to determine the optimal k performance. The value k = 155 derived from the conducted tests will be utilized as the current value of k, given that the university quota is about 150 recipients.

Step 2. Calculate the square of the Euclidean distance of each object to the given sample data using the equation.

$$C(1) = \sqrt{\begin{matrix}(a_1-c_1)^2+(b_1-c_1)^2+(c_1-c_1)^2\\+(d_1-c_1)^2+(e_1-c_1)^2+(f_1-c_1)^2\\(g_1-c_1)^2+(h_1-c_1)^2+(i_1-c_1)^2+\\(j_1-c_1)^2+(k_1-c_1)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-100)^2+(75-100)^2+\\(0-100)^2+(0-100)^2+\\(50-100)^2+(33{,}3-100)^2+\\(100-100)^2+(100-100)^2+\\(20-100)^2+(100-100)^2+\\(100-100)^2\end{matrix}}$$

$$= 184{,}31$$

$$C(2) = \sqrt{\begin{matrix}(a_1-c_1)^2+(b_1-c_1)^2+(c_1-c_1)^2\\+(d_1-c_1)^2+(e_1-c_1)^2+(f_1-c_1)^2\\(g_1-c_1)^2+(h_1-c_1)^2+(i_1-c_1)^2+\\(j_1-c_1)^2+(k_1-c_1)^2\end{matrix}}$$

$$= \sqrt{\begin{matrix}(100-100)^2+(50-100)^2+\\(100-100)^2+(100-100)^2+\\(100-100)^2+(100-100)^2+\\(100-100)^2+(100-100)^2+\\(40-100)^2+(100-100)^2+\\(100-100)^2\end{matrix}}$$

$$= 78{,}10$$

Step 3. Then sort the objects into groups that have the smallest Euclidean distance.

 ISSN:2302-8149
e-ISSN:2540-9719

Step 4. Determine the closest distance up to size k.
Step 5. Pair the appropriate classes.
Step 6. Find the number of classes from the closest neighbors and set that class as the class of data to be evaluated.

Table 8. Performance results of the K-NN model

Euclidean Distance	Euclidean Ranking	K = 155	Prediction Data Classification
78,10	1	Y	P-4
154,56	65	Y	P-16
162,91	125	Y	P-27
145,31	35	Y	P-40
143,33	30	Y	P-43
…	…	…	…
166,17	142	Y	P-1171
162,48	119	Y	P-1198
166,43	144	Y	P-1212
165,29	138	Y	P-1236
152,03	52	Y	P-1255

Table 8 displays the results of implementing the K-NN method with a K value of 155. The Euclidean Distance value indicates the distance between data points, with ranking defining order based on the closest to farthest distance. The choice of K = 155 produces consistent categorization results while meeting the university's quota. The outcomes of data processing utilizing the K-NN algorithm align with the test findings acquired through the Python programming language, as illustrated in Figure 3 below:

```
Euclidean Distance Classification Results (K=155):
+----+-----------+--------------------+------------+
|    | Inisial   | Distance           | Centroid   |
+====+===========+====================+============+
|  0 | P-0004    | 78.10249675906654  | C1         |
+----+-----------+--------------------+------------+
|  1 | P-0016    | 154.56030825826173 | C1         |
+----+-----------+--------------------+------------+
|  2 | P-0027    | 162.907608443832   | C1         |
+----+-----------+--------------------+------------+
|  3 | P-0040    | 145.3061901258473  | C1         |
+----+-----------+--------------------+------------+
|  4 | P-0043    | 143.33333333333334 | C1         |
+----+-----------+--------------------+------------+
|  5 | ...       | ...                | ...        |
+----+-----------+--------------------+------------+
|  6 | P-1171    | 166.17427264438044 | C1         |
+----+-----------+--------------------+------------+
|  7 | P-1198    | 162.48076809271922 | C1         |
+----+-----------+--------------------+------------+
|  8 | P-1212    | 166.41856533719093 | C1         |
+----+-----------+--------------------+------------+
|  9 | P-1236    | 165.28804823364842 | C1         |
+----+-----------+--------------------+------------+
| 10 | P-1255    | 152.03252575974946 | C1         |
+----+-----------+--------------------+------------+
```

Figure 3. Performance results of the K-NN model in python

Leave-One-Out Cross Validation (LOOCV) is the assessment technique employed because the K-NN algorithm uses all datasets as training data. When the full dataset is utilized as training data without any divisions, the LOOCV method is the most accurate way to evaluate. In this study, the LOOCV approach is a good choice because the amount of data to be processed is not too large. The accuracy findings of the LOOCV technique test are as follows:

 ISSN:2302-8149
e-ISSN:2540-9719

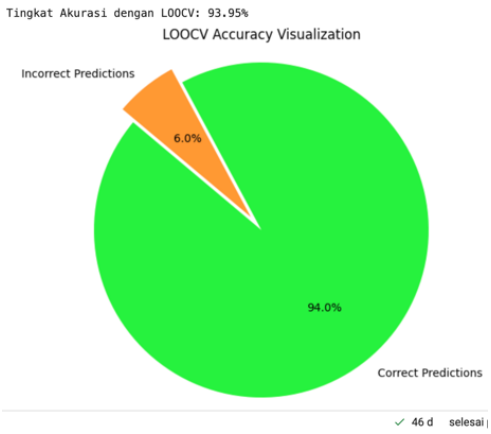

Figure 4. Evaluation results of K-Means and K-NN algorithms (1257 dataset)

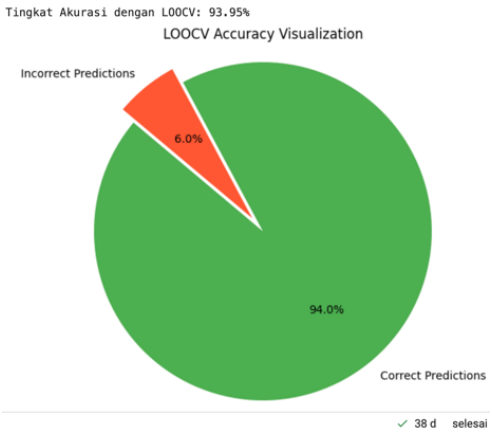

Figure 5. Evaluation results of the K-NN algorithm (1257 dataset)

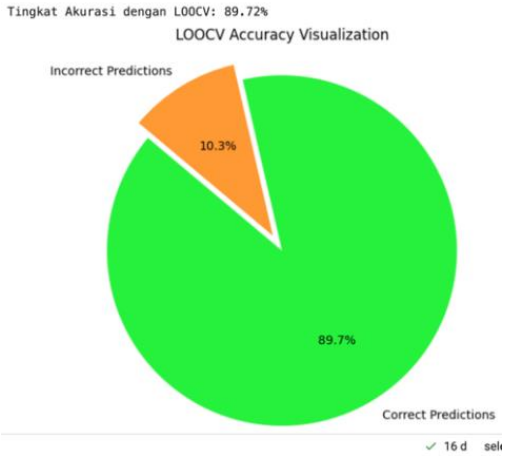

Figure 6. Evaluation results of K-Means and K-NN algorithm (739 Dataset)

Three distinct tests were performed: First, the K-means and K-NN algorithms were applied. Following the completion of the clustering findings, all 1257 data points were processed using the K-NN algorithm, yielding an accuracy of 93.95% and an assessment completion time of 46 seconds. Second. Utilizing a singular K-NN method, all 1257 datasets were processed, yielding an accuracy of 93.95% and an evaluation completion time of 38 seconds. Third, after completing the clustering findings with the K-means and K-NN algorithms, data from cluster 1, comprising 739 entries, were processed using the K-NN technique, yielding an accuracy of 89.72% and an evaluation time of 16 seconds. The conducted tests demonstrate that employing a combination of clustering algorithms (K-Means) and classification (K-NN) yields favorable evaluation results in data processing, achieving an accuracy rate of 89.72% within 16 seconds, as illustrated in Figure 6. This combined engineering approach improves data processing reliability in terms of time and accuracy, resulting in better values and new knowledge for future choices.

5 Conclusion

The research findings indicate that employing a combination of the K-Means and K-Nearest Neighbors (K-NN) algorithms resulted in a considerable enhancement in data processing performance compared to the usage of either technique in isolation. From the 1257 people who underwent processing, three groups were formed: Cluster 0 with 739 data, Cluster 1 with 290 data, and Cluster 2 with 228 data. According to the findings, cluster 0 included 739 high-point-scoring pupils. Examining the suitable k numbers 27, 31, 35, 41, 45, and expanding to 185 yielded the optimal value K-155 using the K-Nearest Neighbor approach. The K-155 figure was chosen because it corresponded to the university's quota, which was 155 people declared competent and eligible based on the specified

 ISSN:2302-8149
e-ISSN:2540-9719

standards. The trials showed that the combination of the K-Means and K-NN algorithms was effective in selecting candidates based on the criteria and quotas stated. The combination of the two methods yields a respectable accuracy of 89,72% accomplished in 16 seconds. This combo can recognize data with excellent accuracy in a fast time while minimizing errors. The combination of the K-Means and K-NN algorithms is a more effective and efficient way to enhance classification results than using them separately. This study demonstrates the importance of investigating hybrid techniques in order to achieve more effective results, and can thus be used as a guideline for future selection.

Acknowledgement

Thank you to all parties who have provided support and contributions so that this research can be completed properly. I would like to express my gratitude to the Ministry of Education, Culture, Research, and Technology for providing this research grant, the DRTM Beginner Lecturer Research Scheme (PDP). This grant has helped provide the resources needed to conduct research more deeply and comprehensively. Thank you, Pekanbaru College of Technology, for providing administrative support throughout the research process.

Reference

[1] Y. Saputra, D. Jaelani, and E. S. Nurpajriah, "Implementasi Algoritma Smart untuk Beasiswa Kip-K di Perguruan Tinggi (Studi Kasus: UIN Sunan Gunung Djati Bandung)," *Jurnal Sistem Informasi Dan Bisnis Cerdas*, vol. 17, no. 1, pp. 59–71, 2024.

[2] N. Indriyani, A. Fauzi, A. Bayu, and H. Yanto, "Pemodelan Prediksi Penerima Beasiswa KIP Kuliah menggunakan Metode *Weight Product*," 2024. [Online]. Available: http://jurnal.bsi.ac.id/index.php/imtechno

[3] M. Safii and Amanda, "Optimisasi Algoritma MOOSRA pada Seleksi Penerima Beasiswa KIP Kuliah," *Jurnal SAINTIKOM (Jurnal Sains Manajemen Informatika dan Komputer)* , vol. 22, no. 2, pp. 555–561, 2023, [Online]. Available: https://ojs.trigunadharma.ac.id/index.php/jis/index

[4] N. W. A. Ulandari, N. L. G. P. Suwirmayanti, and I. P. W. Putra, "Seleksi Penerima Beasiswa pada ITB STIKOM Bali dengan Metode *Weighted Aggregated Sum Product Assessment*," *Jurnal Teknik Informatika Unika ST. Thomas (JTIUST)*, vol. 08, no. 01, pp. 2657–1501, 2023.

[5] M. D. V. Elvira, I. Muda, and A. Suharyanto, "Implementasi Peraturan Menteri Pendidikan dan Kebudayaan Nomor 10 Tahun 2020 tentang Program Indonesia Pintar pada SMAN 4 Kisaran Kabupaten Asahan," *Strukturasi: Jurnal Ilmiah Magister Administrasi Publik*, vol. 4, no. 1, pp. 87–95, 2022, doi: 10.31289/strukturasi.v4i1.1187.

[6] Z. Saputra, D. Sartika, and M. H. Irfani, "Prediksi Calon Mahasiswa Penerima KIP pada Universitas Indo Global Mandiri menggunakan Algoritma *Decision Tree*," *RESOLUSI : Rekayasa Teknik Informatika dan Informasi*, vol. 4, no. 3, pp. 231–240, 2024, [Online]. Available: https://docs.python.org/3.13/tutorial/index.html

[7] P. Apriyani Br Rangkuti *et al.*, "Manajemen Pengelolaan Keuangan Mahasiswa Penerima Beasiswa KIP Kota Medan (Studi Kasus Mahasiswa di Kota Medan)," *Jurnal Akuntansi Keuangan dan Bisnis*, vol. 1, no. 2, pp. 38–43, 2023, [Online]. Available: https://jurnal.ittc.web.id/index.php/jakbs/index

[8] H. Kesuma and S. Hamidani, "Penerapan Data Mining menggunakan Algoritma *K- Means Clustering* dalam Pengelompokan Penerima Beasiswa KIP Kuliah," *Jurnal Ilmiah Binary STMIK Bina Nusantara Jaya Lubuklinggau*, vol. 5, no. 1, pp. 86–92, Apr. 2023, doi: 10.52303/jb.v5i1.102.

[9] N. Haryanti, M. Hasanah, and S. Utami, "Pengaruh Game Online terhadap Prestasi Belajar dan Motivasi Belajar Siswa MI Miftahul Huda Sedang Tulung Agung," *Bahasa dan Pendidikan*, vol. 2, no. 3, pp. 131–138, 2022.

[10] A. Hanafiah, H. O. Nasution, Y. Arta, and R. Wandri, "Perkembangan Portal Informasi berbasis Website Di SMK YKWI Pekanbaru," *Jurnal Pengabdian Masyarakat dan Penerapan Ilmu Pengetahuan*, vol. 5, no. 1, pp. 14–18, 2024.

[11] E. Indriati, N. ' Ainun, S. Azisa, E. I. Sihombing, Z. Sukma, and D. Mokodompit, "Implementasi Algoritma *K-Means Clustering* untuk Pengelompokan Status Penerima KIP

 ISSN:2302-8149
e-ISSN:2540-9719

Kuliah Mahasiswa Universitas Papua," *Jurnal Mahasiswa Teknik Informatika*, vol. 7, no. 6, pp. 3458–3463, 2023.

[12] E. Novianto, A. Hermawan, and D. Avianto, "Perbandingan Metode *K-Nearest Neighbor* dan *Support Vector Machine* untuk memprediksi Penerima Beasiswa Keringanan UKT," *Jurnal Media Informatika Budidarma*, vol. 8, no. 1, pp. 654–662, 2024, doi: 10.30865/mib.v8i1.6913.

[13] R. Wandri, Y. Arta, A. Hanafiah, and R. Oktaviani, "*Prediction of Student Scholarship Recipients using the K-Means Algorithm and C4*," *Indonesian Journal of Computer Science Attribution*, vol. 12, no. 1, pp. 74–88, 2023.

[14] N. S. Ngaeni and K. Kusrini, "Analisis Kombinasi Algoritma *K-Means Clustering* dan TOPSIS untuk menentukan Pendekatan Strategi Marketing berdasarkan *Background* Target *Audiens*," *Journal of Computer System and Informatics (JoSYC)*, vol. 5, no. 2, pp. 393–403, 2024, doi: 10.47065/josyc.v5i2.4948.

[15] I. Irawan, U. Rizki, P. M. Jakak, M. B. Prayogi, and M. Rahman, "Penerapan Metode *K-Means Clustering* dalam Pengembangan Strategi Promosi berbasis Data Penerimaan Mahasiswa Baru (Studi Kasus :Universitas Nurul Huda)," 2024.

[16] J. Faran and R. T. Aldisa, "Perbandingan Algoritma *K-Means* dan *K-Medoids* dalam Pengelompokan Kelas untuk Mahasiswa Baru Program Magister," *Journal of Information System Research*, vol. 5, no. 2, pp. 509–519, 2024, doi: 10.47065/josh.v5i2.4753.

[17] L. Awaliyah, N. Rahaningsih, and R. D. Dana, "Implementasi Algoritma K-Means dalam Analisis Cluster Korban Kekerasan di Provinsi Jawa Barat," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 1, pp. 188–195, 2024.

[18] N. L. Putri, B. Warsito, and B. Surarso, "Pengaruh Klasifikasi Sentimen pada Ulasan Produk Amazon berbasis Rekayasa Fitur dan <i>K-Nearest Negihbor </i>," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 65–74, Feb. 2024, doi: 10.25126/jtiik.20241117376.

[19] M. Fauzan and S. Kurnia Gusti, "Penerapan Seleksi Fitur untuk Klasifikasi Penerima Bantuan Sosial Pangkalan Sesai menggunakan Metode *K-Nearest Neighbor*," *Jurnal Sistem Komputer dan Informatika (JSON) Hal: 1−*, vol. 10, no. 1, 2023, doi: 10.30865/json.v5i1.6654.

[20] A. N. Ikhsan, P. Subarkah, and R. S. Alifian, "Komparasi Algoritme *K-NN, Naïve Bayes*, dan *Cart* untuk memprediksi Penerima Beasiswa," *JST (Jurnal Sains dan Teknologi)*, vol. 12, no. 2, Oct. 2023, doi: 10.23887/jstundiksha.v12i2.51745.

[21] F. Karepesina and L. Zahrotun, "Penerapan Data Mining untuk Penentuan Penerima Beasiswa dengan Metode *K-Nearest Neighbor (K-NN)*," *Techno*, vol. 24, no. 1, pp. 1–9, 2023.

[22] H. Saleh, "K-Nearest Neighbor berbasis Seleksi Atribut *Chi Square* untuk Klasifikasi Penerima Beasiswa Kurang Mampu," *Jurnal SIMETRIS*, vol. 14, no. 1, pp. 39–47, 2023.

[23] T. Xie *et al.*, "*Application of the Improved K-Nearest Neighbor-based Multi-Model Ensemble Method for Runoff Prediction*," *Water (Switzerland)*, vol. 16, no. 1, Jan. 2024, doi: 10.3390/w16010069.

[24] R. Aprilian, R. Habibi, and M. Y. H. Setyawan, *Algoritma KNN dalam memprediksi Cuaca untuk menentukan Tanaman yang Cocok sesuai Musim*. Kreatif, 2020.

[25] U. Hidayah, A. Sifaunajah, and M. Kom, *Cara Mudah memahami Algoritma K-Nearest Neighbor Studi Kasus Visual Basic 6.0*. Lembaga Penelitian dan Pengabdian kepada Masyarakat Universitas KH. A. Wahab …, 2019.

[26] U. O. R. Permatasari, W. J. Shudiq, and M. Jasri, "Prediksi Kelayakan Mahasiswa sebagai Penerima Beasiswa Bank Indonesia pada Tahap Seleksi Administrasi di Universitas Nurul Jadid menggunakan Algoritma *K Nearest Neighbor*," *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, vol. 06, no. 01, Apr. 2024, doi: 10.33650/jeecom.v4i2.

[27] A. O. R. Rodríguez, M. A. Riaño, P. A. G. García, and C. E. M. Marín, "*Application of Learning Analytics for Sequential Patterns Detection Associated with Moments of Distraction in Students in E-learning Platforms*," *Computer Applications in Engineering Education*, vol. 32, no. 1, 2023, doi: 10.1002/cae.22682.

[28] D. Munandar, B. N. Ruchjana, and A. S. Abdullah, *"Principal Component Analysis-Vector Autoregressive Integrated (Pca-Vari) Model using Data Mining Approach to Climate Data in the West Java Region*," *Barekeng Jurnal Ilmu Matematika Dan Terapan*, vol. 16, no. 1, pp. 099–112, 2022, doi: 10.30598/barekengvol16iss1pp099-112.

 ISSN:2302-8149
e-ISSN:2540-9719

[29] U. O. R. Permatasari, W. J. Shudiq, and M. Jasri, "Prediksi Kelayakan Mahasiswa sebagai Penerima Beasiswa Bank Indonesia pada Tahap Seleksi Administrasi di Universitas Nurul Jadid menggunakan Algoritma *K Nearest Neighbor*," *Journal homepage: Journal of Electrical Engineering and Computer (JEECOM)*, vol. 6, no. 1, pp. 252–260, 2024, doi: 10.33650/jeecom.v4i2.

[30] S. Daulay, W. Apriani, and Y. Perwira, *"Application of Data Mining for Prediction of Students Out of College Algorithm C4.5*," *Jurnal ICT : Information and Communication Technologies*, vol. 13, no. 1, pp. 2086–7867, 2022.

[31] E. Buulolo, *Data Mining untuk Perguruan Tinggi*. Deepublish, 2020.

[32] D. Anggraeni and R. Rizaldi, "*K-Means Clustering Calculation to Determine Mainstream Domination of Courses*," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 10, no. 1, pp. 193–198, 2023.

[33] A. Rykov, R. C. De Amorim, V. Makarenkov, and B. Mirkin, "*Inertia-based Indices to Determine the Number of Clusters in K-Means*: *An Experimental Evaluation*," *IEEE Access*, 2024.

[34] G. Gunadi, "Penerapan Algoritma *K-Means Clustering* untuk menganalisa Transaksi Penjualan Jasa Cetak pada Unit *Print on Demand (Pod)* Percetakan Gramedia," *Infotech Journal of Technology Information*, vol. 8, no. 2, pp. 117–126, 2022, doi: 10.37365/jti.v8i2.148.

[35] A. Al Masykur, S. K. Gusti, S. Sanjaya, F. Yanto, and F. Syafria, "Penerapan Metode *K-Means Clustering* untuk Pemetaan Pengelompokan Lahan Produksi Tandan Buah Segar," *Jurnal Informatika*, vol. 10, no. 1, 2023, doi: 10.31294/inf.v10i1.15621.

[36] G. Feng, M. Fan, and C. Yu, *"Analysis and Prediction if Students' Academic Performance based in Educational Data Mining," IEEE Access*, vol. 10, pp. 19558–19571, 2022, doi: 10.1109/access.2022.3151652.

[37] F. Marisa, A. R. Wardhani, W. Purnomowati, A. V. Vitianingsih, A. L. Maukar, and E. W. Puspitarini, "*Potential Customer Analysis using K-Means With Elbow Method," Jiko (Jurnal Informatika Dan Komputer)*, vol. 7, no. 2, p. 307, 2023, doi: 10.26798/jiko.v7i2.911.

[38] P. Violita, G. J. Yanris, and M. N. S. Hasibuan, *"Analysis of Visitor Satisfaction Levels using the K-Nearest Neighbor Method," SinkrOn*, vol. 8, no. 2, pp. 898–914, Apr. 2023, doi: 10.33395/sinkron.v8i2.12257.

[39] R. N. Angraeni, B. Priyatna, A. Hananto, and S. S. Hilabi*, "Application of the K-Nearest Neighbor Method to Predict Demand for Goods from Customers at PT Sinergi Prima Enjineering," Instal : Jurnal Komputer*, vol. 16, no. 02, pp. 99–109, Jun. 2024, doi: 10.54209/jurnalinstall.v16i02.200.

[40] A. S. Paramita, I. Maryati, and L. M. Tjahjono*, "Implementation of the K-Nearest Neighbor Algorithm for the Classification of Student Thesis Subjects*," *Journal of Applied Data Sciences*, vol. 3, no. 3, pp. 128–136, 2022.

[41] E. Gavagsaz, *"Efficient Parallel Processing of k-Nearest Neighbor Queries by using a Centroid-based and Hierarchical Clustering Algorithm*," *Artificial Intelligence Advances*, vol. 4, no. 1, pp. 26–41, May 2022, doi: 10.30564/aia.v4i1.4668.

[42] E. Buulolo, *Data Mining untuk Perguruan Tinggi*. Deepublish, 2020.

[43] D. S. F. Azzahrah and A. Alamsyah, *"Comparison of Probabilistic Neural Network (PNN) and k-Nearest Neighbor (k-NN) Algorithms for Diabetes Classification*," *Recursive Journal of Informatics*, vol. 1, no. 2, pp. 73–82, Sep. 2023, doi: 10.15294/rji.v1i2.66078.

[44] K. Tingkat *et al.*, *"Classification of The Severity 0f Traffic Accident Victims in the City of Samarinda uses the K-Nearest Neighbor and Naive Bayes Algorithms*," *Jurnal EKSPONENSIAL*, vol. 14, no. 2, 2023, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial99

[45] Riana, M. I. Mazdadi, I. Budiman, Muliadi, and R. Herteno, *"Implementation of Information Gain and Particle Swarm Optimization on Sentiment Analysis of Covid-19 Handling using K-NN*," *Jurnal Informatika dan Komputer) Accredited KEMENDIKBUD RISTEK*, vol. 6, no. 1, 2023, doi: 10.33387/jiko.v6i1.5260.

[46] A. Faturrahmi, Zamahsary Martha, Y. Kurniawati, and F. Fitri, *"Sentiment Analysis of Prabowo Subianto as 2024 Presidential Candidate on Twitter using K-Nearest Neighbor*

 ISSN:2302-8149
e-ISSN:2540-9719

*Algorith*m," *UNP Journal of Statistics and Data Science*, vol. 1, no. 5, pp. 385–391, Nov. 2023, doi: 10.24036/ujsds/vol1-iss5/101.

[47] L. Xiang, Y. Xu, J. Cui, Y. Liu, R. Wang, and G. Li, *"GM (1, 1)-based Weighted K-Nearest Neighbor Algorithm for Indoor Localization," Remote Sens (Basel)*, vol. 15, no. 15, p. 3706, 2023.

[48] C. M. Huerta, A. S. Atahua, and J. V. Guerrero, *"Data Mining: Application of Digital Marketing in Education*," *Advances in Mobile Learning Educational Research*, vol. 3, no. 1, pp. 621–629, 2023, doi: 10.25082/amler.2023.01.011.