OPEN ACCESS
Vol. 7, No. 2, pp.148-162, May 2025

ISSN : 2622-8106 (ONLINE)
DOI: 10.20895/inista.v7i2.1630

Journal of Informatics, Information System, Software Engineering and Applications
(INISTA)

Data Mining Analysis of K-means Algorithm and Decision Tree for Early Detection of Students at Risk of Dropping Out

Imam Akbar[*1], Ita Sarmita Samad[2], Rahmat[3,] Sri Rosmiana[4]

[1]Sains and Technology Faculty, Muhammadiyah University of Enrekang, Indonesia
Jalan Jenderal Sudirman No.17 Enrekang, Sulawesi Selatan, 91711, Indonesia
[2]Language and Literature Faculty, State University of Makassar, Indonesia
Jalan Jenderal Sudirman No.17 Enrekang, South Sulawesi, 91711, Indonesia
[3&4]Teacher Training and Education, Muhammadiyah University of Enrekang, Indonesia
Jalan Jenderal Sudirman No.17 Enrekang, South Sulawesi, 91711, Indonesia

[1] imamakbar071093@gmail.com
[2]itaneverendita@gmail.com
[3]rahmatbastra01@gmail.com
[4]srirosmiana@gmail.com

Received on 25-10-2024, revised on 28-04-2025, accepted on 23-05-2025

Abstract

Dropout occurs in higher education, where students are unable to complete their studies within a specified timeframe. It has become a significant concern in education due to its substantial impact on individuals, institutions, and society. This study aims to develop a model for predicting the early potential for students' dropout using the K-Means Algorithm and decision trees. The research method consists of a Dataset, Data Preprocessing, K-means implementation, labeling student data, and Decision Tree implementation. This study resulted in 4 clusters. The students in Cluster 1 have an excellent average GPA, a substantial number of credits, and are very active. The students in Cluster 2 have a lower average GPA and are less active than in Cluster 1. The students in Cluster 3 show a relatively good average GPA, which is lower than in Clusters 1 and 2. The number of active students indicates that students in this cluster are much less active or at risk of D.O. than those in clusters 1 and 2. Cluster 4 indicates that the average GPA of students is very low, often close to zero, and they are generally inactive in academic activities. Thus, they are significantly at risk of D.O. at Universitas Muhammadiyah Enrekang. This research provides significant results, both in terms of accuracy and data interpretation. The resulting insights enable universities to make more strategic and targeted decisions, thereby reducing the risk of university dropout rates, increasing resource efficiency, and supporting the overall educational success of students. The accuracy of the resulting model is 98.52% which indicates that the model has excellent performance in classifying students at risk of D.O.

Keywords: *Data Mining, Decision Tree, Dropout, K-Means*

Corresponding Author:
*Imam Akbar
Sains and Technology Faculty, Muhammadiyah University of Enrekang, Indonesia
Jalan Jenderal Sudirman No.17 Enrekang, South Sulawesi, 91711, Indonesia
Email: imamakbar071093@gmail.com

I. Introduction

Higher education plays an important role in creating qualified graduates or human resources (HR) to support a country's social, economic, and technological development. However, a significant challenge faced by many countries, both developed and developing, is the high rate of students who do not complete their studies, known as dropouts (D.O). This phenomenon not only reflects problems with individuals but also impacts the efficiency and productivity of higher education as a whole. In developing countries, dropout rates are higher, reaching as high as 50%-70%, due to more complex challenges, such as limited access to education, poverty, and students' inability to balance learning needs with the pressures of daily life. According to the UNESCO report, student dropout has major implications for Sustainable Development Goal (SDG) 4, which is to ensure inclusive and equitable quality education and promote lifelong learning opportunities. In Indonesia, the issue of dropout in higher education is a serious concern, as it contradicts the government's vision to increase the Gross Enrollment Rate (GER) of higher education. Based on data from the Ministry of Education, Culture, Research, and Technology (MoECristek), the dropout rate of students in Indonesia ranges from 10% to 15% per year, despite the support programs that have been implemented.

Dropouts are a significant issue in education, as they can lead to financial losses, lower graduation rates, and a declining school reputation. [1]. DO has various definitions from experts [2][3][1]. However, in this case, dropping out is defined as a situation experienced by students who stop studying before they complete their studies. [4]. Based on preliminary data obtained through interviews with the Higher Education Database operator at Universitas Muhammadiyah Enrekang (UNIMEN), it was found that the dropout rate in 2021-2024 reached 20%. This fact is quite high and certainly affects the reputation of higher education. It is revealed that the factors causing students to drop out include exceeding the study period, inactivity, failing compulsory courses, not meeting the minimum number of credits, and difficulties in completing the final project. Thus, it is necessary to implement early prevention measures for student dropouts, thereby increasing the student graduation rate and positively impacting the quality of education.

This research is urgent to carry out, considering that UNIMEN is the only university in Enrekang Regency. The citizens' interest in continuing their studies at UNIMEN has increased since the campus transitioned from high school status to university status, with two faculties offering nine study programs. The community is greatly helped by the presence of a campus in Enrekang City because it reduces the cost of education compared to having to travel out of town to continue education. Students no longer need to pay boarding fees, lower living costs, and can still help parents at home. For this reason, if the dropout rate of UNIMEN students increases, the reputation will decrease and have an impact on decreasing public interest in continuing studies at UNIMEN[5]. Moreover, the dropout rate will also certainly affect the accreditation assessment of this institution [6]. Therefore, it is essential to conduct this research as a means to prevent student dropouts [7].

The problem formulations addressed in this research focus on several key questions. First, it examines how the K-means algorithm can be utilized to group students based on similar characteristics for early identification of those at risk of dropping out. Second, it examines the application of the Decision Tree algorithm in developing a predictive model that can distinguish between students who are likely to drop out and those who are not. Third, it considers how the resulting model can be interpreted to support effective decision-making at Muhammadiyah University of Enrekang. Lastly, the research investigates whether the combined analysis of K-means and Decision Tree algorithms can offer deeper insights into the effectiveness of early detection methods and contribute to reducing the dropout rate at the university.The approach used to solve the problem above involves using data mining with the K-means algorithm analysis model and a decision tree for early detection of students at risk of dropping out at UNIMEN. The purpose of this research is to develop a prediction model that can accurately and effectively distinguish students who have the potential to drop out from those who do not. By utilizing this model, educational institutions can deliver timely and targeted interventions to at-risk students, including academic guidance, emotional support, and specialized programs [6]. Thus, the ultimate goal is to increase student retention and reduce dropout rates, which in turn will improve the quality of education and the institution's reputation.

Research related to Big Data has grown rapidly in the last 10 years. Data Mining has become one of the most popular research studies among researchers. Data Mining is a process that combines several disciplines, such as computer science, mathematics, and statistics, among others, aiming to produce models for problem-solving [8]. Data Mining can be done using the K-means algorithm. [9], K-medoids [10], naïve bayes[11], decision tree [12], Support Vector Machine [13], and other algorithms. Specifically for this study, the research team used decision trees and K-Means algorithms. Data Mining using the K-Means algorithm or Decision tree has also been done several times, showing the effectiveness of the algorithm approach in producing problem-solving models. [14] Research using K-Means and Decision Tree in a hybrid way to predict student academic achievement. [15] used the Decision Tree algorithm to predict the DO student indicator model. [16] used the decision tree algorithm to model the relationship between learning styles and student academic achievement. [17] This research employs a hybrid approach, combining decision trees and k-means, to compare the results of both methods in linking learning styles to student achievement. The novelty of this research lies in the use of data mining with two different types of algorithms, namely decision trees and K-Means, to produce a model of potential student dropouts at UNIMEN. Previous research tends to use a single type of algorithm to predict potential student dropouts.

As for other studies that use the two types of algorithms, linking them to student academic achievement or student learning styles. So no one has connected it with the prediction of student dropouts.

II. RESEARCH METHOD

The research flow of this research is given in Figure 1.

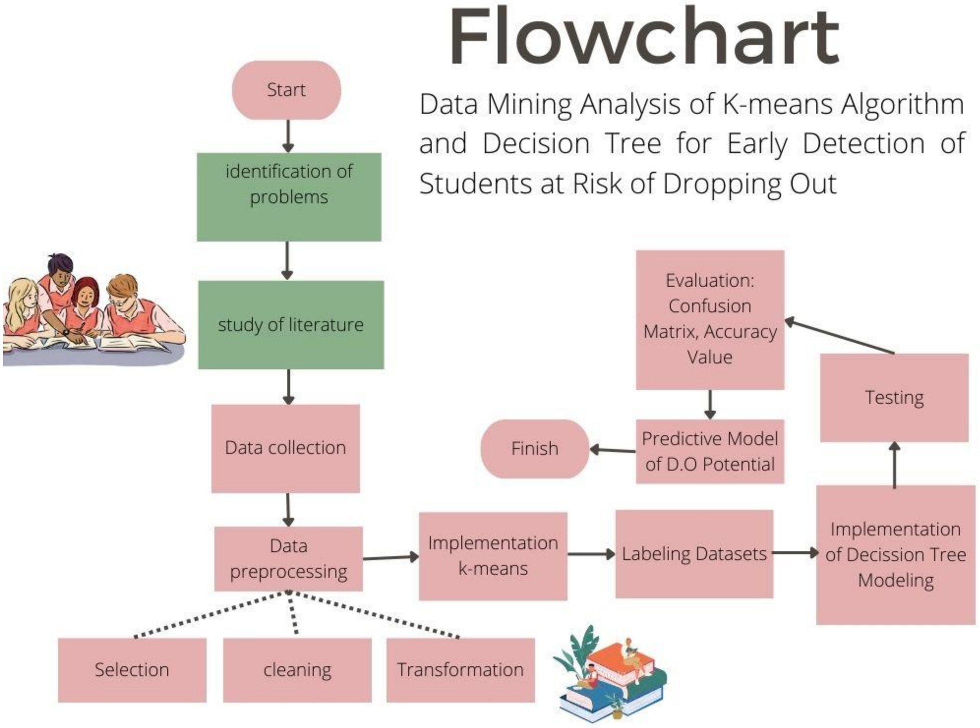

Flowchart
Data Mining Analysis of K-means Algorithm and Decision Tree for Early Detection of Students at Risk of Dropping Out
Start
identification of problems
study of literature
Data collection
Data preprocessing
Selection
cleaning
Transformation
Implementation k-means
Labeling Datasets
Implementation of Decission Tree Modeling
Testing
Evaluation: Confusion Matrix, Accuracy Value
Predictive Model of D.O Potential
Finish

Fig. 1. Flowchart Research Method

A. Data Collecting

At this stage, the data collected is in the form of primary and secondary data obtained from the Academic Information System (SIAKAD) and the UNIMEN Higher Education Database (PDDIKTI) with several criteria, namely:

a. DO because IPS < 1.75 for semester 1, 2 and 3.
b. DO because IPS < 2.00 for semester 4 and up.
c. Drop out because SKS > 8 or have E [15]
d. dropped out due to low activity.

B. Preprocessing Data

Preprocessing data [17], [18] is a process that is carried out to prepare data before carrying out the clustering or classification process, including data selection, data cleaning, and data transformation. Data preprocessing is crucial to ensure that the data is free from noise or inconsistencies that can affect the analysis results. It is relevant to identify patterns and insights related to dropout risk, and optimal for producing models with a high level of accuracy, clear interpretation, and good generalization to new data. The following is the data preprocessing stage:

a. *Data selection* is a process of selecting data to adjust the attributes in accordance with the needs, aiming to ensure the relevance and quality of the data used in the analysis process. Data selection is done because not all the data in the database is used. [19]. Variable identification is crucial in selecting relevant data. Based on the literature review and interviews with education experts, variables considered to influence dropout risk were identified. Some of the key variables include:

1. GPA (Grade Point Average): Measures the academic performance of students.
2. Credits Attempted: Looks at a student's level of study progress based on the number of credits completed.
3. Engagement Level: Measured by class attendance, participation in academic activities, and campus organizations.

These data variables can be collected from various sources, such as the campus academic information system (SIAKAD) and PDDIKTI.

b. *Data cleaning is* a process that involves removing duplicate data, verifying consistency, and correcting data errors. Data Cleaning or Missing Values Handling is done by paying attention to missing data (for example, GPA or number of credits is not recorded) and then using imputation techniques such as Mean/Median Imputation for numerical data such as GPA and SKS and Mode Imputation for categorical data such as activeness status If the amount of missing data is too large (>20%), the rows or columns of data are deleted to maintain the quality of the dataset.

c. *Data transformation* is a process that includes data normalization, data encoding, data imputation, and data scaling. At this stage, the best data is retrieved by reducing or modifying the standard data type, making it ready for presentation to data mining techniques. This selection aims to identify relevant variables used in the study. Data transformation includes data normalization, which processes numeric variables such as GPA and SKS using the Min-Max Scaling method. This method ensures that all values are in the 0-1 range, allowing models like K-Means to work more optimally.

C. Implementation of K-Means

The researchers used the k-means algorithm on the PDDIKTI and SIAKAD data, which had been processed to determine the optimal number of clusters for the k-means algorithm. This algorithm will group student status into clusters based on the similarity of their features. The K-Means algorithm itself is an unsupervised learning algorithm, and the clustering system then proceeds to partition. This algorithm performs very well on numerical data, making it particularly suitable for the data in this study. The K-Means algorithm is also considered capable of reducing the distance between data points and their respective clusters, which is better than the k-medoids algorithm.[20]. In addition, the K-Means Algorithm has a relatively small time and space complexity, is easy to implement, and can effectively separate the features in the space. [21]. Next, the k-means result data is used as labels for the next stage.

Specify the number of clusters to be formed.

One of the challenges in the K-Means algorithm is determining the optimal number of clusters (K) to use. Choosing the right K is critical to ensure the clustering results are relevant and provide meaningful insights. Here are three common methods used to determine the K value:

1. Elbow Method

   The elbow method works by calculating the Sum of Squared Errors (SSE) or inertia for various values of K (e.g., K = 1 to 10) Plot the SSE value against the number of K. then look for the “elbow point” or the point where the decrease in SSE starts to slow down significantly. The elbow method is easy to implement and provides an intuitive visual representation, making it suitable for datasets with well-defined clusters. However, elbow points are not always obvious, especially when the data has a weak or overlapping cluster structure.

2. Silhouette Method

   The silhouette method works by calculating the Silhouette Score value for each data point, which reflects how similar the data is to its cluster compared to other clusters. The average silhouette score for each K is calculated based on the highest average silhouette score among all Ks. Silhouette provides a quantitative measure of cluster quality, making it more accurate than relying solely on visualization and suitable for data that has distinctly different clusters.

3. Gap Stats Method

   Works by comparing the total within-cluster variation (SSE) of the actual data with that of randomly generated data (the baseline). Gap stats ensure that the clusters found are more significant than the randomly generated patterns and are particularly effective on high-noise datasets. Evaluation of clustering models is important to ensure that the clustering results are not only mathematically sound but also contextually relevant to the problem being analyzed. The combination of multiple evaluation methods helps provide more reliable results and supports better data-driven decision-making.

a. The initial K centroid was generated randomly using the Equation (1).

$$v = \frac{1}{N}\sum_{1=1}^{n} Xi \tag{1}$$

b. The distance of each object to each centroid is calculated from each cluster. Then, the distance between the object and the centroid is calculated using the Euclidean distance, as Equation (2).

$$d(x, y) \ = \sum_{i=1}^{n} (x_i - y_i)^2 \tag{2}$$

c. Allocate each object to the nearest centroid.
d. Perform iteration using the equation to determine the new centroid position [22].
e. If the centroid position is not the same, then the researcher repeats Step 3.

Labeling data

The process of labeling student datasets is an important stage in the development of data mining models, especially in this research, as it combines two different methods. This process involves assigning a label to each data in the dataset that indicates the category to which the data corresponds. The labeling process for student data is based on the results of k-means clustering.

Implementation of Decission Tree

Decision Tree is a flowchart-like structure where each internal node (not the outermost node) is a test of the attribute variable, each branch is the result of the test, while the outermost node, the leaf, is the label.[11]. The main advantage of this algorithm is that it can produce a decision tree that is easier to interpret, has an acceptable accuracy level, and is efficient in handling both discrete and numeric attributes, making its application very suitable for this study [23]. In formulating in Equation (3) for the decision tree algorithm, [24]:

$$Entropy(S) = \sum_{i=1}^{k} S_i \log_2 S_i \tag{3}$$

Explanation :
Entropy(S) : A measure of uncertainty or impurity in the dataset **S**.
k : The total number of classes in the dataset.
S$_i$: The proportion (probability) of data belonging to class **i** in the dataset **S**.
log$_2$ S$_i$: The base-2 logarithm of the proportion **S$_i$**

D. Testing

Testing is an evaluation model to understand how well a prediction model performs, especially in classification cases such as dropout risk detection. The following is an explanation of the F1-Score, ROC Curve, and Confusion Matrix, and how they provide a comprehensive picture of model performance. F1-Score is a metric that combines Precision and Recall into a single value using harmonic averaging. The F1-Score is particularly useful when the dataset has class imbalance (for example, the number of dropout students is significantly less than the number of students who remain active). The downside of this test is that it does not provide information on true negatives, making it unsuitable if the goal is to understand the model's overall performance. The ROC (Receiver Operating Characteristic) Curve is a graph that illustrates the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at various model probability thresholds. The downside of this test is that it is not ideal for highly imbalanced datasets, as the FPR can be very low in the majority class, making the curve appear “too good.

A confusion matrix is a table that summarizes the model's prediction results in a quantitative form, comparing the true label with the model's predicted value. This matrix consists of four elements:

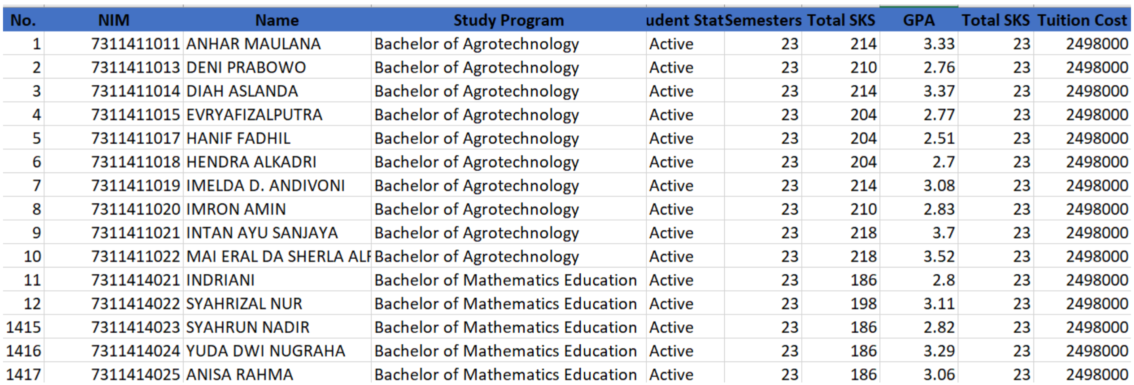
- True Positive (TP): The number of correctly predicted positive cases.
- True Negative (TN): The number of negative cases that were correctly predicted.
- False Positive (FP): The number of negative cases that were incorrectly predicted as positive (also called Type I Error).

- False Negative (FN): The number of positive cases that were incorrectly predicted as negative (also called Type II Error).

The advantages of this matrix provide a detailed analysis of model errors and allow the calculation of other metrics, such as Precision and Recall. Testing is performed using a Confusion matrix consisting of four classifications: Fast, Right, Late, and DO. To measure the performance of the resulting predictions, testing is carried out to evaluate the Precision, Recall, Accuracy, and Error rate values. [25].

III. RESULTS AND DISCUSSION

The result was analyzed descriptively before preprocessing the data, which is an initial step in the K-Means clustering process. This process involves cleaning unnecessary data to produce clean data. [26], Handling missing values and data normalization. Preprocessing of students' data includes data from the Faculty of Education and Teacher Training (FKIP) and the Faculty of Science and Technology (SAINSTEK) of Universitas Muhammadiyah Enrekang, specifically for the 2021 class. Figure 2 shows the results of data collection taken from the Higher Education Database (PDDIKTI)



No.	NIM	Name	Study Program	udent Stat	Semesters	Total SKS	GPA	Total SKS	Tuition Cost
1	7311411011	ANHAR MAULANA	Bachelor of Agrotechnology	Active	23	214	3.33	23	2498000
2	7311411013	DENI PRABOWO	Bachelor of Agrotechnology	Active	23	210	2.76	23	2498000
3	7311411014	DIAH ASLANDA	Bachelor of Agrotechnology	Active	23	214	3.37	23	2498000
4	7311411015	EVRYAFIZALPUTRA	Bachelor of Agrotechnology	Active	23	204	2.77	23	2498000
5	7311411017	HANIF FADHIL	Bachelor of Agrotechnology	Active	23	204	2.51	23	2498000
6	7311411018	HENDRA ALKADRI	Bachelor of Agrotechnology	Active	23	204	2.7	23	2498000
7	7311411019	IMELDA D. ANDIVONI	Bachelor of Agrotechnology	Active	23	214	3.08	23	2498000
8	7311411020	IMRON AMIN	Bachelor of Agrotechnology	Active	23	210	2.83	23	2498000
9	7311411021	INTAN AYU SANJAYA	Bachelor of Agrotechnology	Active	23	218	3.7	23	2498000
10	7311411022	MAI ERAL DA SHERLA ALF	Bachelor of Agrotechnology	Active	23	218	3.52	23	2498000
11	7311414021	INDRIANI	Bachelor of Mathematics Education	Active	23	186	2.8	23	2498000
12	7311414022	SYAHRIZAL NUR	Bachelor of Mathematics Education	Active	23	198	3.11	23	2498000
1415	7311414023	SYAHRUN NADIR	Bachelor of Mathematics Education	Active	23	186	2.82	23	2498000
1416	7311414024	YUDA DWI NUGRAHA	Bachelor of Mathematics Education	Active	23	186	3.29	23	2498000
1417	7311414025	ANISA RAHMA	Bachelor of Mathematics Education	Active	23	186	3.06	23	2498000

Fig. 2. Higher Education Database (PDDIKTI)

The data contains a total of 1417 rows and 10 columns. From the number of columns several variables was used such as NIM, IPS, GPA, semester credits, overall credits and tuition fees which are used to analyze and early detect students who are at risk of Drop Out (D.O) and several variables were deleted such as No, Name, College Status and Study Program [27]. Figure 3 shows the average IPS, GPA, and semester credits of the student data preprocessing results.

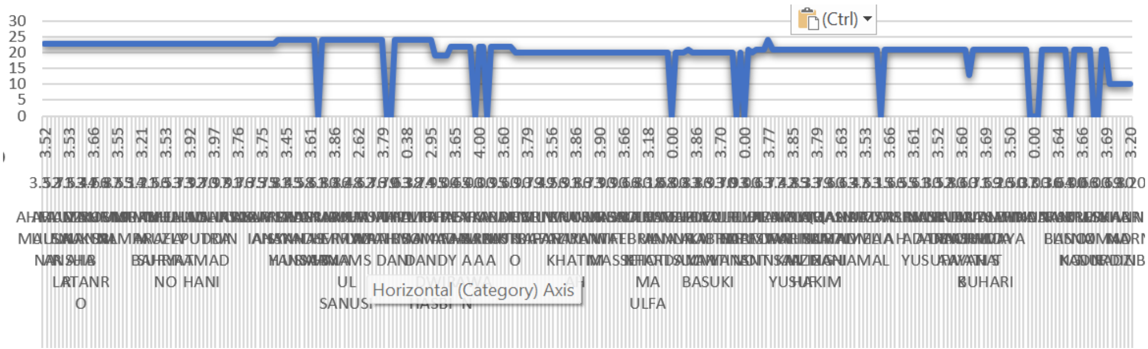
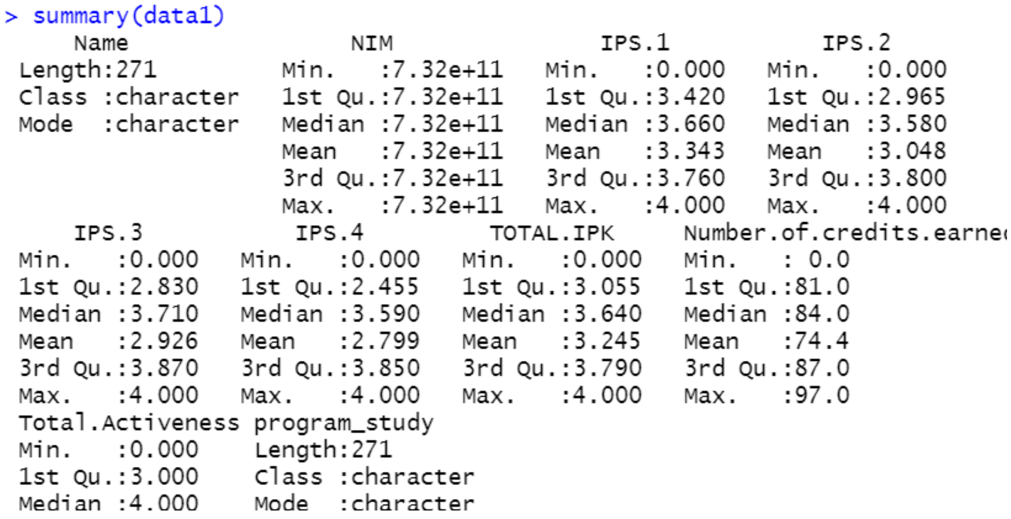


Fig. 3. The Average of IPS, IPK, and SKS Odd Semester 2021

Figure 3 shows the average number of IPS 1, IPS 2, IPS 3, IPS 4, GPA and SKS that have been taken by students of the 2021 Class of Muhammadiyah University of Enrekang with the number of semesters passed for 4 semesters, this variable is used to measure the extent of the student success process in carrying out the learning process as well as detecting students who have the potential to drop out (D.O) in the lecture process.



```
> summary(data1)
     Name                 NIM                 IPS.1           IPS.2      
 Length:271         Min.   :7.32e+11   Min.   :0.000   Min.   :0.000  
 Class :character   1st Qu.:7.32e+11   1st Qu.:3.420   1st Qu.:2.965  
 Mode  :character   Median :7.32e+11   Median :3.660   Median :3.580  
                    Mean   :7.32e+11   Mean   :3.343   Mean   :3.048  
                    3rd Qu.:7.32e+11   3rd Qu.:3.760   3rd Qu.:3.800  
                    Max.   :7.32e+11   Max.   :4.000   Max.   :4.000  
     IPS.3           IPS.4          TOTAL.IPK     Number.of.credits.earne
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   : 0.0         
 1st Qu.:2.830   1st Qu.:2.455   1st Qu.:3.055   1st Qu.:81.0         
 Median :3.710   Median :3.590   Median :3.640   Median :84.0         
 Mean   :2.926   Mean   :2.799   Mean   :3.245   Mean   :74.4         
 3rd Qu.:3.870   3rd Qu.:3.850   3rd Qu.:3.790   3rd Qu.:87.0         
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :97.0         
 Total.Activeness program_study     
 Min.   :0.000    Length:271        
 1st Qu.:3.000    Class :character  
 Median :4.000    Mode  :character  
```

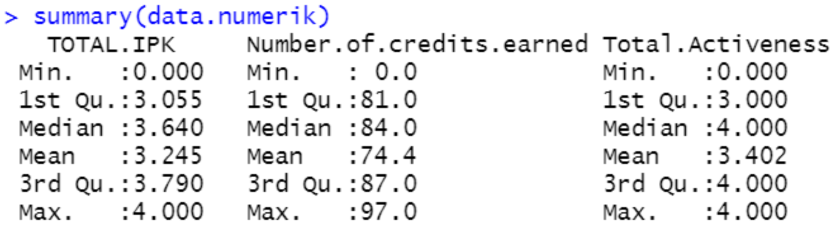
Fig. 4. Statistical Summary in R Studio

The output in Figure 4 is a *summary()* function in R Studio that displays summary statistics for datasets with multiple output structures. One of them is the character type of the name variable, which shows that there are 271 data points or observations. Then, IPS.1, IPS.2, IPS.3, IPS.4, total GPA, number of credits, and activity are of a numeric type. For a detailed explanation of the summary statistics, refer to the Table I.

TABLE I. THE STATISTICAL SUMMARY

No	Name	Description
1	Min	The smallest value in the data, for example, for IPS.1, is 0.000.
2	1st Qu	First quartile: Indicates the value that 25% of the data falls below, e.g., for IPS.1, it is 3.420.
3	Median	The median value, which is the value where 50% of the data is below it and the other 50% is above it
4	Mean	Average score. For IPS.1, the average was 3.343.
5	3rd Qu	Third quartile: Indicates the value that 75% of the data falls under. For IPS.1, the third quartile is 3,760.
6	Max	The largest value in data. For IPS.1, the maximum value is 4.00

The output in the figure above and the table provide concise information about the data distribution of some important variables. With this data, we can understand the distribution of students' social studies scores and GPAs, the number of credits they take, and their level of student activity. This data can be used to conduct further analysis, such as identifying students with high performance or potential, or examining the distribution of the number of credits taken by students across several semesters. The next step is data scaling, which is used to standardize or scale several variables, including Total GPA, Number of SKS obtained, and Total Activity, so that they are in a similar range. This helps maintain the proportion between variables when used for further analysis, as shown in Figure 5.



```
> summary(data.numerik)
   TOTAL.IPK     Number.of.credits.earned Total.Activeness
 Min.   :0.000   Min.   : 0.0             Min.   :0.000   
 1st Qu.:3.055   1st Qu.:81.0             1st Qu.:3.000   
 Median :3.640   Median :84.0             Median :4.000   
 Mean   :3.245   Mean   :74.4             Mean   :3.402   
 3rd Qu.:3.790   3rd Qu.:87.0             3rd Qu.:4.000   
 Max.   :4.000   Max.   :97.0             Max.   :4.000   
```

Fig. 5. Data Scaling

Standardization serves to change the data in such a way that the data distribution has a mean of 0 and a standard deviation of 1 [28]. This is important because it allows for the use of several machine learning algorithms, which helps avoid bias that may arise due to differences in scale between variables. Next, find the optimal K value using the silhouette method. Here are the optimal K results obtained using the silhouette method.

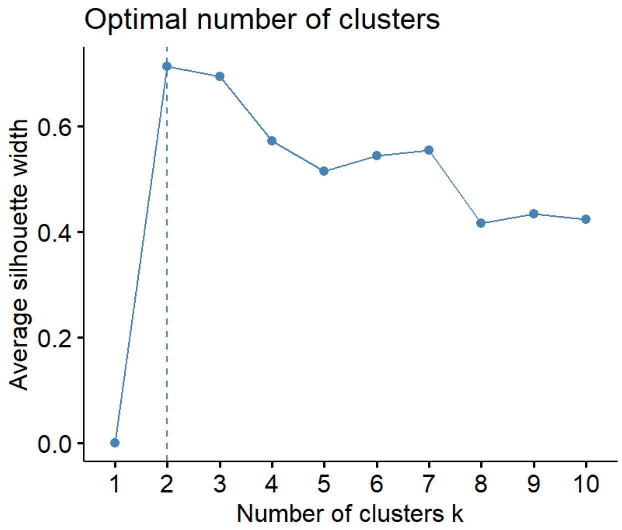


Fig. 6. Silhouette Method

Figure 6 shows the analysis result of the Silhouette Analysis method to determine the optimal number of clusters in a clustering algorithm. The X-axis indicates the number of clusters tested, ranging from 1 to 10. Each point on this axis represents the analysis result for a given number of clusters (k). The Y-axis shows the average silhouette width value for each number of clusters tested. Silhouette measures how well an observation is clustered and estimates the average distance between clusters [29]. Thus, silhouette width is a measure of how well the objects in a cluster are grouped. The highest point of the graph occurs at k = 2, with an average silhouette width of approximately 0.7. This suggests that two clusters are the optimal number of clusters for this dataset.

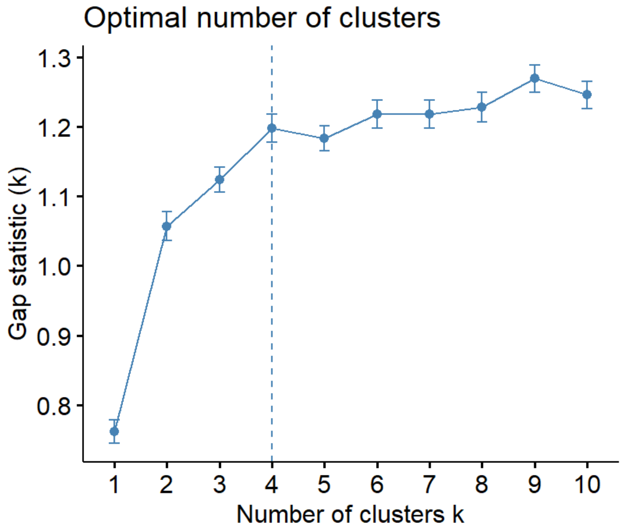


Fig. 7. Gap Stats

In Figure 7, the X-axis shows the number of clusters tested, ranging from 1 to 10. Each point on this axis represents the result for each number of clusters (k) tested using the Gap Statistic. The Y-axis represents the value of the Gap Statistic, which measures the quality of the clusters formed compared to random data. The Gap Statistic calculates the difference between the log of the total within-cluster dispersion of the actual data and the log of the total within-cluster dispersion of the random data. This technique is based on the change in within-cluster dispersion as the number of clusters in the data increases. The optimal number of clusters is selected based on the Gap Statistic value that first reaches a maximum or is significantly greater than previous values. The GAP Statistic is said to be the optimum cluster if the GAP Statistic value increases the most.[30]. In this figure, the point at k = 4 has a high Gap Statistic and is quite stable.

The graph also displays error bars at each point, which show the variability in the Gap Statistic calculation. If two points have overlapping error bars, it means that the difference between the Gap Statistic values may not be significant. In this graph, the error bars begin to overlap after k = 4, indicating that increasing the number of clusters beyond four does not yield a significant improvement in clustering quality. This graph demonstrates that the optimal number of clusters for the tested data is k = 4, as the Gap Statistic value at k = 4 is the highest and significantly larger than those of the other clusters. Although the Gap Statistic value remains high for k = 4, the improvement in cluster quality after k = 4 is not statistically significant.

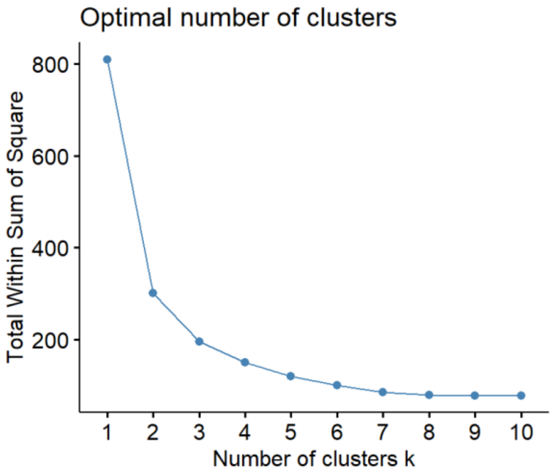


Fig. 8. Elbow

In Figure 8, the Y-axis shows the total variation within the cluster. It measures the distance of each point in a cluster from its centroid. The smaller the WSS value, the better the cluster minimizes internal variation. This graph shows a sharp decrease in the WSS value as the number of clusters increases. However, at some point, this decrease starts to slow down and forms an "elbow," where increasing the number of clusters no longer significantly reduces the WSS. k = 1: At this point, the WSS is very high (around 800), as all the data is grouped in one large cluster, so the internal variation is very large. K = 2 to 4: There is a significant decrease in WSS as the data is divided into more clusters, so the internal variation within each cluster is reduced. k = 4 to 5: At this point, the decrease in WSS starts to slow down, indicating that increasing the number of clusters beyond this no longer provides a significant reduction in internal variation. The "elbow" point on this graph is usually used to determine the optimal number of clusters. From this graph, the elbow point appears to be at k = 4. This means that using 4 clusters provides a good balance between low internal variation (WSS) and an efficient number of clusters. If the number of clusters is increased beyond 4, the larger the number of clusters k, the smaller the WSS value will be [31]. WSS values will still decrease, but at a slower rate, meaning that adding further clusters does not provide significant benefits in reducing within-cluster variation.

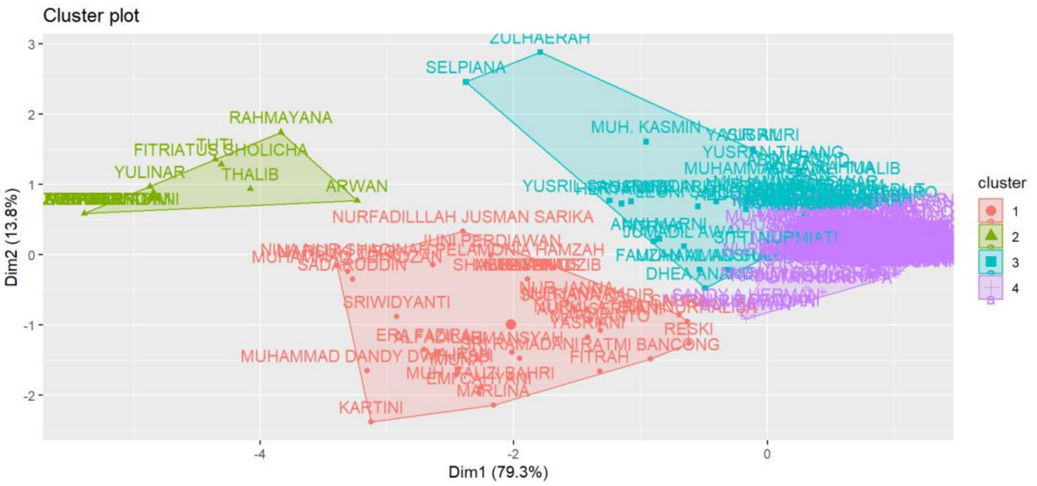
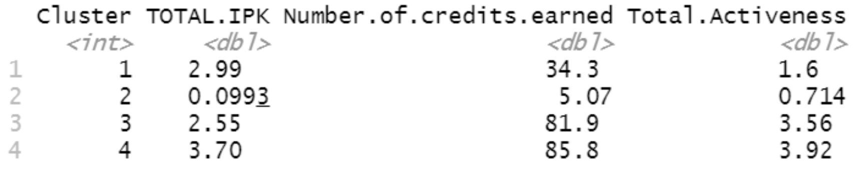


Fig. 9. Visualization of 4 Clusters

Figure 9 displays the results of a Cluster Plot using clustering visualization techniques, which was created using K-Means Clustering with mapping to two main dimensions (Dim1 and Dim2). Dim1 (79.3%) represents the horizontal axis, indicating the first principal component, which explains 79.3% of the variation in the data. This means that most of the information or variation in the dataset is represented by Dim1. Dim2 (13.5%) represents the vertical axis, indicating the second principal component, which explains 13.5% of the data variation. So, cumulatively, these two dimensions explain about 92.8% of the variation in the dataset. In Figure 9, four clusters are identified, each represented by a different color. Cluster 1 (Red color), cluster 2 (Green color), cluster 3 (Blue color), and cluster 4 (Purple color). The average grouping of K = 4, as indicated by the cluster, is visible in Figure 10.



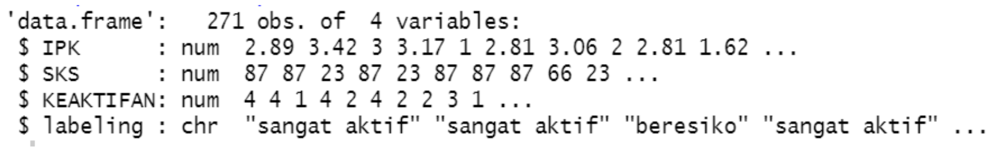
```
  Cluster TOTAL.IPK Number.of.credits.earned Total.Activeness
    <int>     <dbl>                    <dbl>            <dbl>
1       1    2.99                       34.3            1.6
2       2    0.0993                      5.07           0.714
3       3    2.55                       81.9            3.56
4       4    3.70                       85.8            3.92
```

Fig. 10. K = 4 cluster averages

Figure 10 displays the results of the clustering process, which shows the average value (centroid) of several variables for each cluster formed. Each row in the table represents a cluster and displays the average value for a particular variable within each cluster. The cluster 1 total GPA is 3.64, indicating students in this cluster have a very good average GPA. The number of credits earned is 85.6, indicating that students in this cluster also take a considerable number of credits, close to the maximum for most study programs. The number of activeness is 3.91, indicating that students in this cluster are very active, both in academic and extracurricular activities.

Cluster 2's total GPA is 2.24, indicating that students in this cluster have a lower average GPA than those in Cluster 1. The number of credits obtained is 81.0, indicating that they also took a relatively large number of credits, but fewer than those in cluster 1. The total activeness score is 3.4, indicating that these students are still active in activities, but slightly lower than those in cluster 1. Cluster 3 has a total GPA of 2.99, indicating a good average GPA. The number of credits earned is 34.3, indicating that students in this cluster take a significantly smaller number of credits compared to clusters 1 and 2. The number of activities is 1.6, indicating that students in this cluster are substantially less active and at a higher risk of D.O. compared to clusters 1 and 2. The cluster 4 total GPA is 0.0993, indicating that the average GPA of students in this cluster is very low, even approaching zero, which suggests they may be facing serious academic challenges. The number of credits earned is 5.07, indicating that students in this cluster take only a few credits, which may reflect their inability to take more courses due to academic problems. Total engagement is 0.714, indicating that students in this cluster are also very inactive in academic activities and are highly at risk of D.O. from Muhammadiyah University of Enrekang.

Figure 11 is the process of labeling student data used for classification using a Decision Tree. Dataset labeling, also known as annotation, is the process of providing information about a dataset [32]. In performing the classification process, the class must be determined first, as illustrated in Table II. The Next step is looking at the contents of the data to be classified using *R Studio*.



```
'data.frame':   271 obs. of  4 variables:
 $ IPK      : num  2.89 3.42 3 3.17 1 2.81 3.06 2 2.81 1.62 ...
 $ SKS      : num  87 87 23 87 23 87 87 87 66 23 ...
 $ KEAKTIFAN: num  4 4 1 4 2 4 2 2 3 1 ...
 $ labeling : chr  "sangat aktif" "sangat aktif" "beresiko" "sangat aktif" ...
```

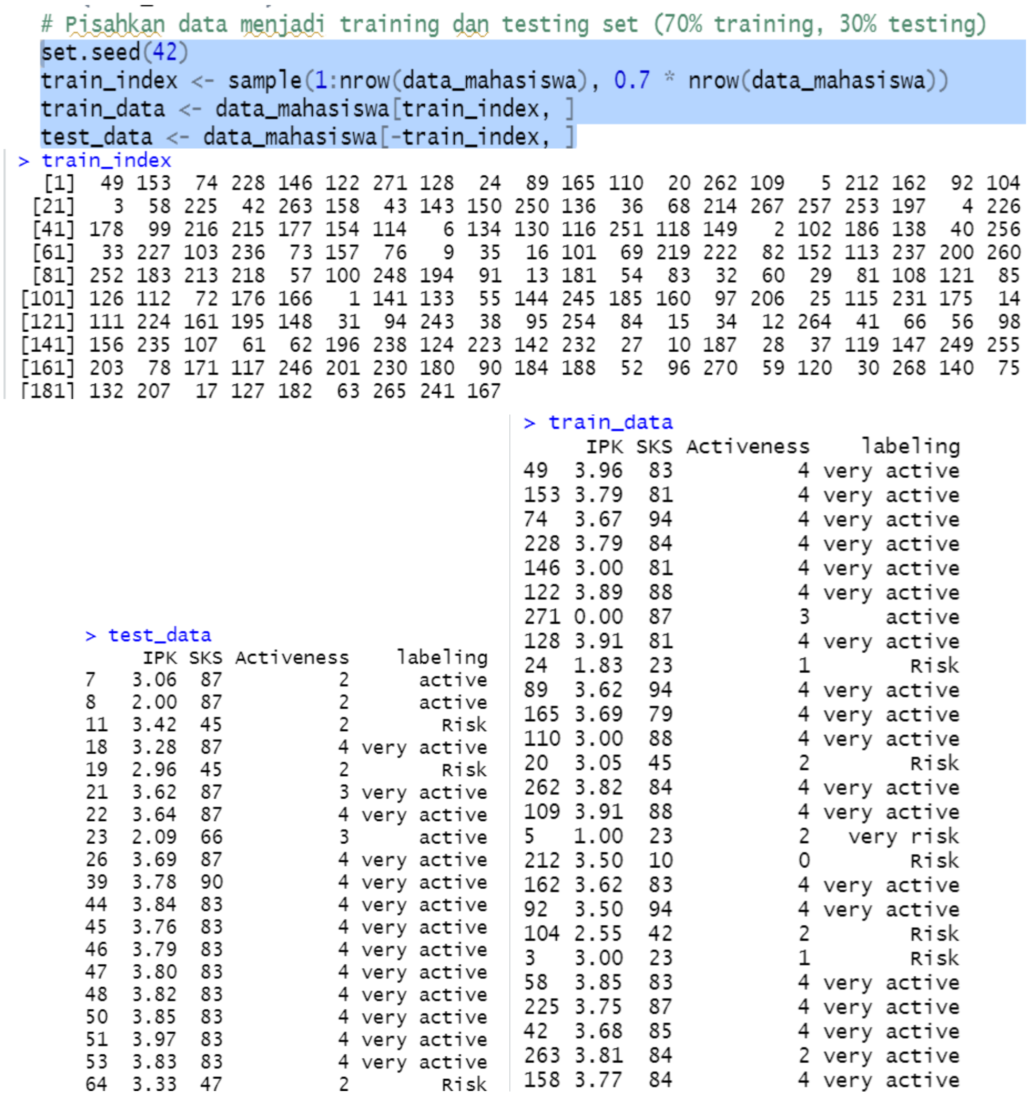
Fig. 11. Data Structure

TABLE II. LABELING DATA PROCESS

Cluster	Average of GPS	Average of Credits	Average of Activeness	Labelling
k=1	3.64	85	3.91	Very Active
k=2	2.24	81	3.4	Active
k=3	2.99	34	1.6	At Risk
k=4	0.0993	5.07	0.7	Very At Risk

Based on this data, the student data was classified with attributes such as GPA, SKS, and Activeness, using numeric data type and class labeling in character data type. To process the data using a decision tree, the class labeling data type must first be converted into a factor in accordance with the workings of the

Decision tree algorithm, which works by dividing the data based on category values that facilitate model interpretation.[33], This makes the model easier for humans to interpret.



```
# Pisahkan data menjadi training dan testing set (70% training, 30% testing)
set.seed(42)
train_index <- sample(1:nrow(data_mahasiswa), 0.7 * nrow(data_mahasiswa))
train_data <- data_mahasiswa[train_index, ]
test_data <- data_mahasiswa[-train_index, ]
```

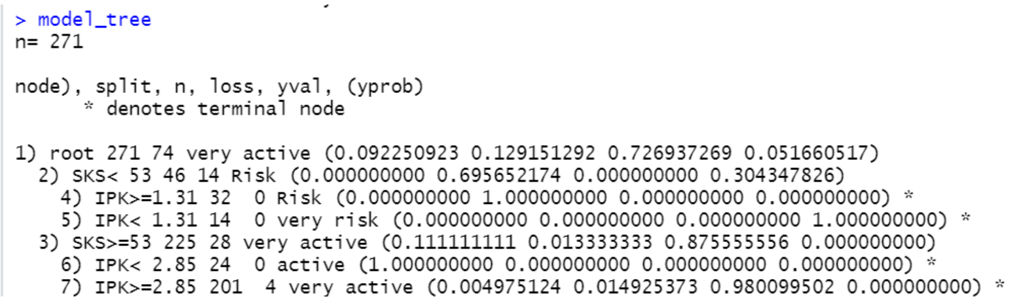
```
> train_index
  [1]  49 153  74 228 146 122 271 128  24  89 165 110  20 262 109   5 212 162  92 104
 [21]   3  58 225  42 263 158  43 143 150 250 136  36  68 214 267 257 253 197   4 226
 [41] 178  99 216 215 177 154 114   6 134 130 116 251 118 149   2 102 186 138  40 256
 [61]  33 227 103 236  73 157  76   9  35  16 101  69 219 222  82 152 113 237 200 260
 [81] 252 183 213 218  57 100 248 194  91  13 181  54  83  32  60  29  81 108 121  85
[101] 126 112  72 176 166   1 141 133  55 144 245 185 160  97 206  25 115 231 175  14
[121] 111 224 161 195 148  31  94 243  38  95 254  84  15  34  12 264  41  66  56  98
[141] 156 235 107  61  62 196 238 124 223 142 232  27  10 187  28  37 119 147 249 255
[161] 203  78 171 117 246 201 230 180  90 184 188  52  96 270  59 120  30 268 140  75
[181] 132 207  17 127 182  63 265 241 167
```

```
> train_data
     IPK SKS Activeness    labeling
49  3.96  83          4 very active
153 3.79  81          4 very active
74  3.67  94          4 very active
228 3.79  84          4 very active
146 3.00  81          4 very active
122 3.89  88          4 very active
271 0.00  87          3      active
128 3.91  81          4 very active
24  1.83  23          1        Risk
89  3.62  94          4 very active
165 3.69  79          4 very active
110 3.00  88          4 very active
20  3.05  45          2        Risk
262 3.82  84          4 very active
109 3.91  88          4 very active
5   1.00  23          2   very risk
212 3.50  10          0        Risk
162 3.62  83          4 very active
92  3.50  94          4 very active
104 2.55  42          2        Risk
3   3.00  23          1        Risk
58  3.85  83          4 very active
225 3.75  87          4 very active
42  3.68  85          4 very active
263 3.81  84          2 very active
158 3.77  84          4 very active
```

```
> test_data
    IPK SKS Activeness    labeling
7  3.06  87          2      active
8  2.00  87          2      active
11 3.42  45          2        Risk
18 3.28  87          4 very active
19 2.96  45          2        Risk
21 3.62  87          3 very active
22 3.64  87          4 very active
23 2.09  66          3      active
26 3.69  87          4 very active
39 3.78  90          4 very active
44 3.84  83          4 very active
45 3.76  83          4 very active
46 3.79  83          4 very active
47 3.80  83          4 very active
48 3.82  83          4 very active
50 3.85  83          4 very active
51 3.97  83          4 very active
53 3.83  83          4 very active
64 3.33  47          2        Risk
```

Fig. 12. Testing and Training Data

Figure 12 illustrates the division of the training data and testing data. Training data is used to create a subset of the student dataset. The training data comprises 70% of the data used to train the model (training set). The remaining 30% of the data (using an index not included in the training index) was used for testing and validation, allowing the model to be trained in order to understand the patterns and relationships in the data [34]. This is done by retrieving all rows from the student data that are not in the training index. Testing data is the data used to test the performance of the model after it has been trained (test set).



```
> model_tree
n= 271

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 271 74 very active (0.092250923 0.129151292 0.726937269 0.051660517)
  2) SKS< 53 46 14 Risk (0.000000000 0.695652174 0.000000000 0.304347826)
    4) IPK>=1.31 32  0 Risk (0.000000000 1.000000000 0.000000000 0.000000000) *
    5) IPK< 1.31 14  0 very risk (0.000000000 0.000000000 0.000000000 1.000000000) *
  3) SKS>=53 225 28 very active (0.111111111 0.013333333 0.875555556 0.000000000)
    6) IPK< 2.85 24  0 active (1.000000000 0.000000000 0.000000000 0.000000000) *
    7) IPK>=2.85 201  4 very active (0.004975124 0.014925373 0.980099502 0.000000000) *
```

Fig. 13. Tree Model

Figure 13 illustrates the output of a decision tree model known as Model Tree. Decision Tree Structure: Root Node (node 1). This is the first node, or root, of the tree. There are 271 data points (n=271) in this node, of which 74 have "loss" (inaccuracy), and most predicted labels are "very active". (0.0925, 0.1291, 0.7269, 0.0516), This is the probability distribution of each category. This means that for this node, the probability of the label "very active" is 0.7269 or 72.69%, which means it is the dominant prediction.

The tree branches (Nodes 2 and 3) then split the data based on certain features. Node 2, the first splitting condition is SKS < 53. There are 46 data points under this node, and most of them have the label "at risk". Here, the probability of the "at risk" class is 0.6956 or about 69.56%. Node 3, the opposite condition, SKS> 53, has 225 data points, and most students in this node are labeled "very active" with a probability of 87.56%. A sub-branch of Node 2, after dividing the data by SKS, the model performs further breakdowns. Node 4, for students with SKS < 53 and GPA <= 1.31, all of them (32 data) are labeled "at risk" with 100% probability (1.0000). This is the terminal node (leaf node), meaning there is no further division. Node 5: For students with a GPA > 1.31, there are 14 data points, all labeled "very at risk" (100%). This is also a terminal node.

The sub-branch of Node 3 on the branch where SKS >= 53 has two further breakdown conditions. Node 6: For students with a GPA <= 2.85, there are 24 data points that all have the label "active" with 100% probability. This is the terminal node. Node 7: For students with a GPA > 2.85, there are 201 data points, and the majority (98%) are labeled "very active". This is also a terminal node.

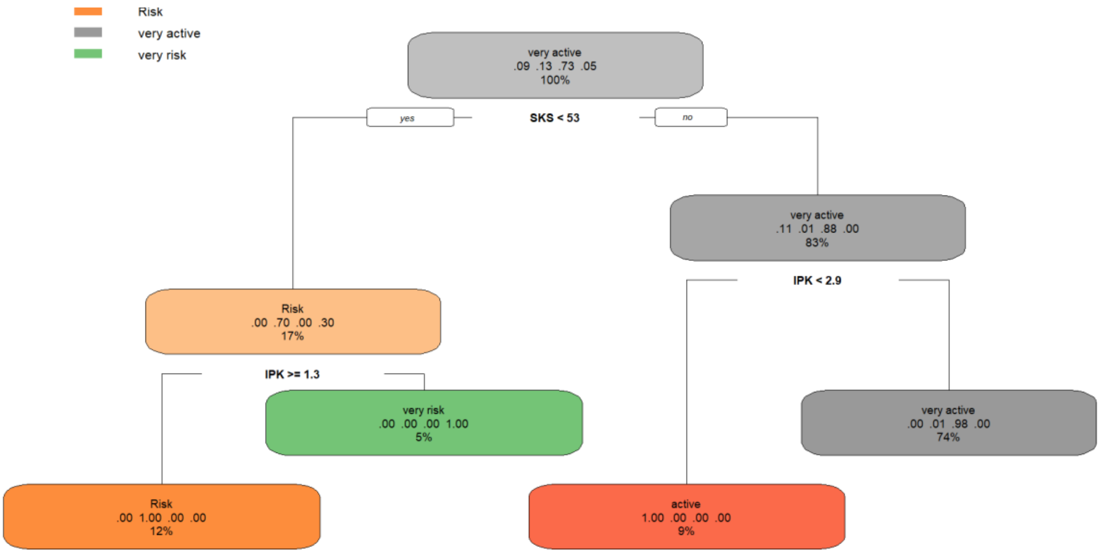
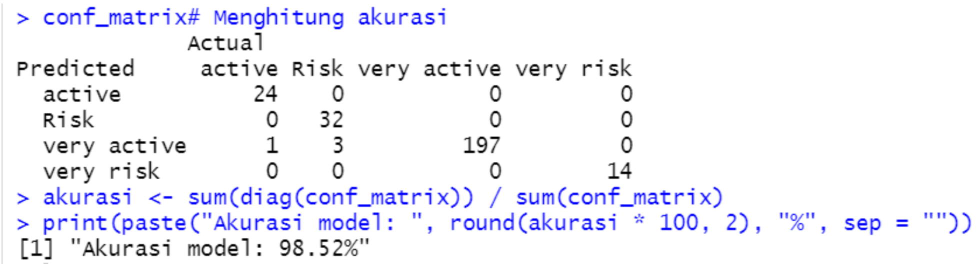


Fig. 14. Visualization of Decision Tree

Figure 14 illustrates the visualization of the decision tree, which is used to classify student data based on two primary features: SKS and GPA, with four output categories: active, at risk, very active, and very at risk. The Root Node SKS < 53 is the first separation condition. Initially, all student data (100%) are in one group, and most of them (73%) are classified as very active (in gray). This node splits the data into two branches, based on whether the student's SKS is less than 53 or not. Left Branch (SKS < 53) If SKS < 53, the data is further split based on GPA. Orange-colored nodes indicate that if GPA >= 1.3, the data is classified as at-risk with 100% probability. This accounts for approximately 17% of the data. Green colored nodes: If GPA < 1.3, the data is classified as very risky with 100% probability, accounting for 5% of the total data.

Right Branch (SKS >= 53) If SKS >= 53, there are two separation conditions based on GPA: Nodes are colored red: If GPA < 2.9, all data (9% of the total) is classified as active. Gray nodes: If the GPA is greater than or equal to 2.9, the data is classified as very active with 100% probability, which accounts for 74% of the total data. The color of the node represents the category of the gray prediction result: Very active (green), Orange (risk), Green (very risky), and Red (active). Each node displays the probability or distribution of classification. For example, the root node shows that, of all the data, 9% of students are classified as "active", 13% as "at risk", 73% as "very active", and 5% as "very at risk". On the left branch, after splitting by credits < 53, 100% of students with GPA >= 1.3 are classified as "at risk".



```
> conf_matrix# Menghitung akurasi
             Actual
Predicted     active Risk very active very risk
  active          24    0           0         0
  Risk             0   32           0         0
  very active      1    3         197         0
  very risk        0    0           0        14
> akurasi <- sum(diag(conf_matrix)) / sum(conf_matrix)
> print(paste("Akurasi model: ", round(akurasi * 100, 2), "%", sep = ""))
[1] "Akurasi model: 98.52%"
```

Fig. 15. Confusion Matrix

Figure 15 shows the confusion matrix, which displays the model's predicted results compared to the actual values. It serves to evaluate the performance of the classification model by calculating the correct and incorrect predictions for each class. The interpretation of this matrix is:

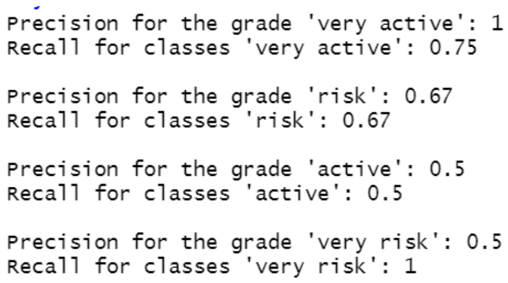
1. The model predicted 24 students as “active” , and all of them were active.

2. The model predicted 32 students as “at risk”, and all of them were correct.

3. The model predicted 197 students as “very active”, and these 197 students were very active.

4. The model predicted 14 students as “very at risk”, and all of them were indeed very at risk.

The confusion matrix in Figure 16 shows the performance of the classification model based on the predicted and actual values. The model's accuracy is 98.52%, indicating that it performs very well.



```
Precision for the grade 'very active': 1
Recall for classes 'very active': 0.75

Precision for the grade 'risk': 0.67
Recall for classes 'risk': 0.67

Precision for the grade 'active': 0.5
Recall for classes 'active': 0.5

Precision for the grade 'very risk': 0.5
Recall for classes 'very risk': 1
```

Fig. 16. Precision and Recall

The “very active” class performed best with perfect precision (1), but still had some problems with recall (0.75), indicating that some “very active” data was not detected. The “at risk” class performed quite well with the same precision and recall (0.67), indicating that the model has some errors but can still detect most of the data. The “active” class performed less well in terms of precision and recall at 0.5, indicating that the model was not effective in detecting data in this class. The “very risky” class has perfect recall (1), but with lower precision (0.5), indicating that all data that should be in this class is detected, but half of the predictions are incorrect.

IV. CONCLUSION

Based on the results of the research above, it can be concluded that during the preprocessing of student data for the Class of 2021, the initial dataset consists of 1417 rows and 10 columns, containing several variables. Still, after processing, 279 rows and 7 columns remain as variables that we will cluster using *k-means*. Based on the search for the optimal k value using the silhouette method, k = 2, the gap stats method, k = 4, and the elbow method, k = 4. From the results of determining the value of k, an effective method is employed to determine the optimal value of k, namely the elbow and gap statistics, with k = 4. From the results of visualization and determination of the average value of clustering using *k-means,* four cluster plots are identified. Cluster 1 has a high GPA (3.64), a high number of credits (85.6), and high activity (3.91). This is a cluster with excellent academic performance and a high level of activity. Cluster 2 has a lower GPA (2.24) and a higher number of credits (81.0), but remains quite active with a score of 3.4. Cluster 3 has a medium GPA (2.99), but a much lower number of credits (34.3), and is only (1.6) active. This cluster is categorized as potential D.O. students. Cluster 4 is the lowest performing cluster, where members have a very low GPA (0.0993), a very small number of credits (5.07), and low activeness (0.714), meaning students are very potential D.O. Classification using decision tree produces the right criteria in analyzing students who have the potential to drop out. The criteria include GPA, which is calculated from the accumulation of social studies over 4 semesters, the number of semester credits, and the number of activeness, which can affect decisions on the status of active, non-active, and potentially Dropout students. From the classification results, the decision tree model predicts 24 students as active, 32 students as "at risk", 197 students as "very active", and 14 students as "very at risk".

ACKNOWLEDGMENT

This research was supported by a research grant from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia through the 2024 Novice Lecturer Research Scheme, funded by BIMA. The authors would like to express their sincere gratitude for this financial support, which made the completion of this study possible.

REFERENCES

[1] Márquez-Vera C, Cano A, Romero C, Noaman AYM, Mousa Fardoun H, Ventura S. Early dropout prediction using data mining: A case study with high school students. Expert Syst. 2016;33(1):107–24.

[2] Casanova JR, Cervero A, Núñez JC, Almeida LS, Bernardo A. Factors that determine the persistence and dropout of university students. Psicothema. 2018;30(4):408–14.

[3] Agrusti F, Bonavolontà G, Mezzini M. University dropout prediction through educational data mining techniques: A systematic review. J E-Learning Knowl Soc. 2019;15(3):161–82.

[4] Gitto L, Minervini LF, Monaco L. University dropouts in Italy: Are supply side characteristics part of the problem? Econ Anal Policy. 2016;49(February):108–16.

[5] Panggabean DD, Motlan, Harahap MH, Irfandi, Sirait AP. Development of Accelerating Strategy on Improvement of Study Program Accreditation in Accordance With 9 Criterias of Ban-Pt in the State University of Medan. Adv Soc Sci Res J. 2020;7(1):483–91.

[6] Utari M, Warsito B, Kusumaningrum R. Implementation of Data Mining for Dropout Prediction using Random Forest Method. 2020 8th Int Conf Inf Commun Technol ICoICT 2020. 2020;

[7] Singh HP, Alhamad IA. A Data Mining Approach to Predict Key Factors Impacting University Students Dropout in a Least Developed Economy. Arch Bus Res. 2022;10(12):48–59.

[8] Roiger RJ. Data Mining: A Tutorial-Based Primer, Second Edition. Data Mining: A Tutorial-Based Primer, Second Edition. 2017. 1–487 p.

[9] Darwis M, Hasibuan LH, Firmansyah M, Ahady N. Implementation of K-Means C lustering A lgorithm in M apping the G roups of G raduated or D ropped-out S tudents in the Management Department of the National University. 04(01):1–9.

[10] Guntara M, Suprawoto T. Drop Out Student Clusterization Using the k-Medoids Algorithm. Jl Raya Janti Karang Jambe. 2022;(2):61–6.

[11] Harwati, Virdyanawaty RI, Mansur A. Drop out Estimation Students based on the Study Period: Comparisonbetween Naïve Bayes and Support Vector Machines Algorithm Methods. IOP Conf Ser Mater Sci Eng. 2016;105(1).

[12] Pérez B, Castellanos C, Correal D. Predicting student dropout rates using data mining techniques: A case study. Commun Comput Inf Sci. 2018;833:111–25.

[13] Dewi Purba S, Harahap L, Panggabean JFR. Prediction Of Students Drop Out With Support Vector Machine Algorithm. J Mantik. 2021;6(1):582–6.

[14] Ogwoka TM, Cheruiyot W, Okeyo G. A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms. Int J Comput Appl Technol Res. 2015;4(9):693–7.

[15] Sivakumar S, Venkataraman S, Selvaraj R. Predictive modeling of student dropout indicators in educational data mining using improved decision tree. Indian J Sci Technol. 2016;9(4):1–5.

[16] Safitri SN, Haryono Setiadi, Suryani E. Educational Data Mining Using Cluster Analysis Methods and Decision Trees based on Log Mining. J RESTI (Rekayasa Sist dan Teknol Informasi). 2022;6(3):448–56.

[17] Akbar I, Hazriani H, Arda AL, Samad IS. Analysis of Student Behavior Based on the History of Learning Activities in the Learning Management System Using the Pearson Correlation Method. Edumaspul J Pendidik. 2024;8(1):464–70.

[18] Iddrus I, Sari DW. Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja. J Adv Res Inform. 2023;1(02):1–7.

[19] Ramadhani A, Fazarany Noor R, Vernanda D, Herdiawan T. Klasifikasi Mahasiswa Berpotensi

Drop Out Menggunakan Algoritma C4.5 di Politeknik Negeri Subang. 18(1).

[20] Sugianto CA, Rahayu AH, Gusman A. Algoritma K-Means Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Cigugur Tengah.

[21] Selvi C, Sembiring D, Hanum L, Parsaoran Tamba S. PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS UNTUK MENENTUKAN JUDUL SKRIPSI DAN JURNAL PENELITIAN (STUDI KASUS FTIK UNPRI). J Sist Inf dan Ilmu Komput Prima). 2022;5(2).

[22] Abdul Majid MB, Cani YM, Enri U. Penerapan Algoritma K-Means dan Decision Tree Dalam Analisis Prestasi Siswa Sekolah Menengah Kejuruan. J Sist Komput dan Inform. 2022 Dec 31;4(2):355.

[23] Rifa'i H, Ryan Hamonangan, Dian Ade Kurnia, Kaslani, Mulyawan. Implementasi Algoritma Decision Tree Dalam Klasifikasi Kompetensi Siswa. KOPERTIP J Ilm Manaj Inform dan Komput. 2022;6(1):15–20.

[24] Keputusan Dirjen Penguatan Riset dan Pengembangan Ristek Dikti S, Nurkholis A, Susanto T. Terakreditasi SINTA Peringkat 2 Algoritme Spatial Decision Tree untuk Evaluasi Kesesuaian Lahan Padi Sawah Irigasi. Masa Berlaku Mulai. 2017;1(3):978–87.

[25] Orpa EPK, Ripanti EF, Tursina T. Model Prediksi Awal Masa Studi Mahasiswa Menggunakan Algoritma Decision Tree C4.5. J Sist dan Teknol Inf. 2019;7(4):272.

[26] Lailatul Ramadhania H, Zakaria L, Nusyirwan dan. Aplikasi Metode Sillhouette Coefficient, Metode Elbow dan Metode Gap Staticstic dalam Menentukan K Optimal pada Analisis K-Medoids. Vol. 04, Jurnal Siger Matematika. 2023.

[27] Nurani S, Syahra Y, Calam A. Penerapan Data Mining Dalam Clustering Pencapaian Target Penjualan Menggunakan Algoritma K-Means. J Sist Inf Triguna Dharma (JURSI TGD). 2023;2(3):355.

[28] Abdullah A, Utami PY. 1816-4463-1-Pb. :540–54.

[29] Utari DT. Analisis Karakteristik Wilayah Transmisi Covid-19 dengan Menggunakan Metode K-Means Clustering. J Media Tek dan Sist Ind. 2021;5(1):25.

[30] Sulistiyawan E, Hapsery A, Arifahanum LJA. PERBANDINGAN METODE OPTIMASI UNTUK PENGELOMPOKAN PROVINSI BERDASARKAN SEKTOR PERIKANAN DI INDONESIA (Studi Kasus Dinas Kelautan dan Perikanan Indonesia). J Gaussian. 2021;10(1):76–84.

[31] Hartanti NT. Metode Elbow dan K-Means Guna Mengukur Kesiapan Siswa SMK Dalam Ujian Nasional. J Nas Teknol dan Sist Inf. 2020;6(2):82–9.

[32] Khairunnas K, Yuniarno EM, Zaini A. Pembuatan Modul Deteksi Objek Manusia Menggunakan Metode YOLO untuk Mobile Robot. J Tek ITS. 2021;10(1).

[33] Raka Sujono M, Bahtiar A, Irawan B. Analisis Model Machine Learning Untuk Jenis Aspal Di Jawa Barat Menggunakan Algoritma Decision Tree Dan Random Forest. JATI (Jurnal Mhs Tek Inform. 2024;7(6):3886–91.

[34] Raharja AR, Jayadi, Pramudianto A, Muchsam Y. Penerapan Algoritma Decision Tree dalam Klasifikasi Data "Framingham" Untuk Menunjukkan Risiko Seseorang Terkena Penyakit Jantung dalam 10 Tahun Mendatang. Technol J. 2024;1(1).