Indonesian Journal of Electrical Engineering and Computer Science
Vol. 33, No. 3, March 2024, pp. 1726~1735
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v33.i3.pp1726-1735

A hybrid data mining for predicting scholarship recipient students by combining K-means and C4.5 methods

Halifia Hendri[1], Harkamsyah Andrianof[1], Riska Robianto[1], Hasri Awal[1], Okta Andrica Putra[1], Romi Wijaya[1], Aggy Pramana Gusman[1], Muhammad Hafizh[2], Muhammad Pondrinal[3]
[1]Department of Computer System, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia
[2]Department of Informatics Engineering, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia
[3]Department of Accountancy, Faculty of Economic and Business, Universitas Putra Indonesia YPTK Padang, Padang, Indonesia

Article Info

Article history:

Received Oct 20, 2023
Revised Jan 2, 2024
Accepted Jan 11, 2024

Keywords:

Hybrid method
C.45 method
K-means mathod
Students
Scholarship recipients
University

ABSTRACT

This scholarly investigation delves into the strong desire for academic scholarships within the student body, especially prominent among socioeconomically disadvantaged individuals. The study aims to formulate a hybrid data mining paradigm by synergizing the K-means and C4.5 methodologies. K-means is applied for clusterization, while C4.5 facilitates prediction and decision tree instantiation. The research unfolds in sequential phases, commencing with data input and progressing through meticulous pre-processing, encompassing data selection, cleaning, and transformation. The novelty lies in successfully integrating the K-means and C4.5 methodologies, culminating in the hybrid data mining method. The dataset comprises 200 students seeking scholarships, revealing effective stratification into three clusters—cluster 0, cluster 1, and cluster 2—with 119, 48, and 33 students, respectively. The K-means method proves highly suitable, especially when combined with C4.5, for predicting scholarship recipients. A subset of 81 students from clusters 1 and 2 undergoes predictive modeling using C4.5, resulting in a commendable 85% accuracy, with 17 accurate forecasts and 3 minor inaccuracies. This research significantly enhances scholarship selection efficiency, particularly benefiting socioeconomically disadvantaged students.

Corresponding Author:

Halifia Hendri
Department of Computer System, Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang 25221, Padang, West Sumatera, Indonesia
Email: halifia_hendri@upiyptk.ac.id

1. INTRODUCTION

A student requires various resources to fortify their educational journey, encompassing essential requisites such as communication, internet access, transportation, sustenance, accommodation, literature, fees, and sundry ancillary expenditures [1]–[4]. Foremost among the diverse expenditures necessitated by educational pursuits are tuition fees [5]–[7], which students remit to their respective educational institutions. These fees assume prominence as the most pivotal among the varied financial obligations incurred during the pursuit of education. Significantly, the costliest echelon of education, namely tertiary education or college, commands the highest average expense in comparison to other strata of educational attainment [8]–[11]. Universities frequently confer designations upon students, a practice accompanied by the requisite payment of sundry fees to the educational institution of their enrollment. Scholarships, emanating from diverse sources including governmental, commercial, and public institutions, albeit valuable, are constrained in their availability [12]–[14].

The process of extracting, aggregating, or mining crucial information from an extensive dataset is commonly referred to as data mining [15]–[18]. This involves the application of statistical analysis methodologies in data mining procedures and mathematical techniques, predominantly leveraging advancements in artificial intelligence technologies [19]–[21]. Educational institutions employ a diverse array of methodologies to identify and select individuals eligible for scholarships. While some individuals undertake this task manually, others opt for computerized approaches. Numerous computer-based methods exist for predicting potential scholarship recipients. Within the realm of computerization, a multitude of algorithms find application, with two algorithms, namely the K-means algorithm [22], [23] and the C4.5 algorithm [24], [25], standing out prominently, developed by researchers. The amalgamation of these two algorithms yields a more precise predictive computation compared to the utilization of a single algorithm alone. The K-means Algorithm functions to cluster student data into three distinct clusters. Subsequently, the researcher employs the C4.5 technique, a data mining process utilized to unveil data-driven forecasts or predictions.

Upon the completion of this study, the objective is to identify methodologies for anticipating or predicting scholarship recipients, employing the K-means algorithm and C4.5 method, specifically tailored for students of computer science. Situated in Padang, Universitas Putra Indonesia (UPI) under the auspices of Computer College Foundation (YPTK) Padang stands as a distinguished and renowned private institution in Indonesia, particularly within the West-Sumatra Province. The Faculty of Computer Science (FILKOM) represents one of the faculties of computer science within the university. For this study, the data source comprises students enrolled in the FILKOM at UPI YPTK Padang in the year 2022. Scholarships, including the Student Study Assistance (BBM) scholarships and those available to class 1 and 2 winners, as well as the Bidik Misi scholarships, are accessible within the FILKOM Faculty at UPI YPTK Padang. These scholarships are distributed equitably among qualifying students, necessitating a meticulous selection process to identify eligible recipients. To facilitate this process, predictive measures utilizing data mining methodologies, such as the C4.5 algorithm, become imperative. By employing such techniques, prospective students can gain insights into the determining factors that influence their chances of securing a scholarship.

Previous research is done by Cordeiro de Amorim and Makarenkov [26]. In this research, they explore the relationship between the average number of iterations K-means takes to converge and the structure of a data set under study. They demonstrate that this number of iterations is related to the clustering quality and can be used to identify irrelevant features in a given data set and improve the results of existing feature selection algorithms. Additionally, the research shows that there is a strong relationship between the number of iterations and the number of clusters in a data set, which can be used to find the true number of clusters it contains. Overall, this research provides valuable insights into the use of k-means clustering for data analysis. The novelty of this research lies in the demonstration of a previously unknown relationship between the average number of iterations k-means takes to converge and the structure of a data set under study. This relationship has important implications for data analysis, including identifying irrelevant features present in a given data set and improving the results of existing feature selection algorithms. Additionally, the research shows that there is a strong relationship between the number of iterations and the number of clusters in a data set, which can be used to find the true number of clusters it contains. Overall, this research provides valuable insights into the use of k-means clustering for data analysis.

Previous research also done by Nasyuha *et al.* [23]. The objective of this research was to manage cosmetic products and find the right strategy to increase business in the field of sales and improve sales services. The researchers used data mining algorithms, specifically the K-means clustering algorithm, to analyze cosmetic product sales transactions and identify the best-selling products, products that are quite in demand, and products that are not selling well. The novelty lies in using clustering techniques to analyze sales transactions and identify products that are not selling well, thus preventing the accumulation of unsold products. This approach can help cosmetic companies improve their sales strategies and increase profits. Additionally, the research highlights the effectiveness of the K-means algorithm in solving grouping problems and encourages further research in different product grouping cases.

Previous research also done by Wang and Gao [27]. They discuss research conducted on the C4.5 algorithm improvement strategy based on MapReduce. The authors explore the use of MapReduce to improve the performance of the C4.5 algorithm and evaluate the effectiveness of this approach through experiments. The article provides details on the methodology, results, and conclusions of the research. The novelty of this research lies in the combination of the C4.5 decision tree algorithm with the MapReduce parallel model in Hadoop platform. This approach allows the C4.5 algorithm to be executed in parallel, which improves its efficiency and performance. The authors evaluate the effectiveness of this approach through experiments and provide insights into the potential applications of this research in real-world scenarios.

2. METHOD

2.1. Research framework

The research framework serves as the primary guide in executing the research. It delineates the sequential progression of the research from initiation to culmination. Figure 1 presented below illustrates the comprehensive research framework employed in this study.

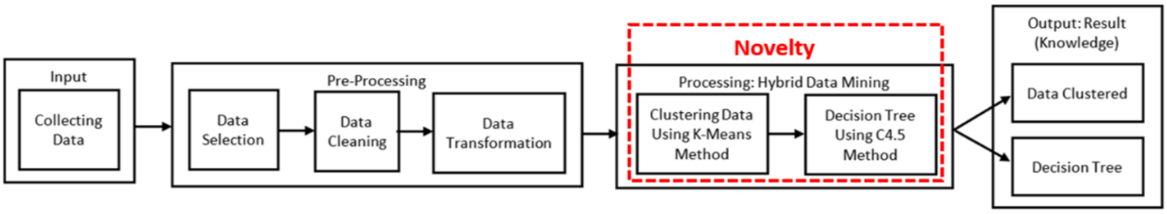

Figure 1. Research framework

2.2. Research framework details

A research framework refers to the structure or outline used in the context of research to provide direction, foundation, and boundaries for a research project. It serves as a guide for researchers in planning, designing, and conducting their research in a structured and systematic manner. This research consists of four steps: the first is input steps, followed by pre-processing steps, then processing steps, and finally, output steps, which yield the knowledge gained. The first step of this research is the Input step. Input steps generally refer to the process or sequence of actions involved in providing input to a system, program, or process. In the context of computing and programming, input refers to the data or information supplied to a program or system for processing. In this research, Input steps involve collecting data from scholarship applicants. We collect data from all scholarship applicants obtained from the Vice Dean III (WD III) of the Computer Science Faculty (FILKOM) at UPI YPTK, Padang. The WD III provided extensive data for this study. Given the volume of data received, we need to undertake pre-processing steps to ensure the best possible results.

The input data for this research has been obtained; however, not all the data required for the research is available, and the existing data is not yet free from noise. Additionally, there are instances where the data does not conform to the provided data format. Therefore, pre-processing the data is necessary before utilization. This pre-processing stage is divided into three types: data selection, followed by data cleaning, and finally, data transformation. Data selection involves the process of choosing and retrieving a subset of data from a larger dataset for further analysis. The goal of data selection is to focus on relevant information and reduce the volume of data to be processed, thereby making the analysis more efficient and effective. In this research, data selection entails choosing only the crucial data needed for processing. The selected data fields include student name, parents' income (PI), cumulative index (GPA), and parents' status (PS).

After obtaining the selected data, the next pre-processing step is data cleaning. Data cleaning refers to the process of identifying and correcting errors or inconsistencies in datasets. It is a crucial step in the data pre-processing phase of data mining. The goal of data cleaning is to improve the quality of the data, ensuring that it is accurate, reliable, and suitable for analysis. In this research, after conducting data cleaning, we eliminated 20 rows of data, resulting in a cleaned dataset of 200 rows. This dataset represents a total of 200 data entries, equivalent to 200 students. Following the data cleaning process, the next pre-processing step is data transformation. Data transformation refers to the process of converting and modifying data into a suitable format for analysis. This step is a part of the broader data pre-processing phase, where the goal is to prepare the raw data for effective use by data mining algorithms. Data transformation involves various operations to enhance the quality of the data and improve the performance of mining algorithms. In this research, after completing the data transformation, we converted non-numeric data into numeric format. Specifically, the data in the parents' status (PS) field was transformed from "Have" and "Haven't" into 1 or 0, respectively.

After obtaining the data transformation, the next step is processing data using a hybrid data mining method. In this research, the hybrid data mining method combines clustering techniques, namely the K-means method and C4.5 method [28]–[31]. This step involves organizing the data into clusters using the K-means method. The algorithm for the K-means method is as follows [26], [32], [33]:

1) Regulate desired amount of data clusters that will be created, which is referred to as amount 'k.'
2) Randomly assign each mean (centroid value) pre-defined class.
3) Identify the cluster center closest for each data point using value of centroid. To calculate it, employ following algorithm:

$$d_{Euclidean}(x,y) = \sqrt{\sum(x_i - y_i)^2} \quad (1)$$

where: $d_{Euclidean}(x,y)$ = each data row has a distance value and a centroid value, $y_i = y_1, y_2, y_3, \ldots.etc$, $x_i = x_1, x_2, x_3, \ldots.etc$,

4) Regulate nearest for each row of data, group (cluster) by comparing the closest distance value, acquired in the preceding phase, as well as changing center value of group using the algorithm:

$$Cluter\ Center = \sum \frac{a_i}{n} \quad (2)$$

where: $Cluster\ Center$ = The value of cluster center, a_i = Each cluster's value, n = total clusters number.

5) Repeat steps 3–5 until no data is transferred from one group to another for each row of data.

After obtaining the results of clustering the data, the next step is to implement the C4.5 method. The algorithm for the C4.5 method [27], [34], [35] is as follows:

1) Identify the characteristics of the data to be employed as root nodes or predictors in decision tree, and tally the occurrences each data row's YES and NO values.
2) Identify branch stemming from given each value, the root subsequent to establishing the decision tree's root. This is achieved by computing the Gain value. Gain is calculated as follows:

$$Gais(S,A)Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (3)$$

where: $Entropy(S_i)$ = Entropy Value of each attribute, $Gain(S,A)$ = The attribute's total gain value, n = number of *cluster*, $Entropy(S)$ = Entropy Value Total.

Entropi algorithm is:

$$Entropy(S) = \sum_{i=1}^{n} - pi * log_2\, pi \quad (4)$$

Information: pi = proportion of S_i for S, $Entropy(S)$ = Entropy Value Total,

3) Cases should be separated on each existing branch.
4) Repeat steps 2–3 for until each branch in every branch case achieves same class.

The novelty inherent in this research lies in the successful integration of two distinct data mining methodologies—namely, the K-means method and the C4.5 method—culminating in what is denoted as the hybrid data mining method. Unlike prior studies that predominantly utilized a singular method [26], [27] or amalgamated it with others [23], the innovation here resides in the fusion of the K-means and C4.5 methods. Remarkably, this marks a departure from conventional approaches, where such a combination was not hitherto explored by researchers. The confluence of the K-means and C4.5 methods proves particularly apt in the context of predicting scholarship recipients within a university setting. The next step is then output steps or result of this research that resulting the knowledge. Two kind knowledge that has been get in this research. They are data clustered and decision tree.

3. RESULTS AND DISCUSSION

Based on the methods of this research, the first step is inputting data by collecting data. The data that we collected is obtained from Vice Dean III of the Computer Science Faculty at UPI YPTK Padang. The collected input data is shown in Table 1.

Table 1. The input data

No	BP Number	Student name	Place of birth	Date of birth	Parents income (PO) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)
1	17..027	Wahyudi Nasti	Padang	January 10, 2004	2,000,000	3.33	Have
2	18..007	Rizki Saputra	Bukittinggi	June 16, 2004	800,000	2.86	Haven't
3	18..020	Reza Oktivani	Padang	March 4, 2004	1,000,000	3.87	Haven't
….	…	…			…	…	…
220	17..444	Admel Brina	Solok	July 20, 2004	2,000,000	3.25	Have

Drawing upon the data presented in Table 1 above, it is observed that the initial dataset comprises 220 students who registered as scholarship recipients, featuring 8 data columns or fields. All of this data has been meticulously collected and entered into the established system. Notably, not all collected data is employed in this research. The acquired data undergoes a comprehensive pre-processing procedure, with the initial step being data selection. The resulting dataset post data selection in this research is delineated in Table 2.

Table 2. The data selection

No	Student name	Parents income (PO) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)
1	Wahyudi Nasti	2,000,000	3.33	Have
2	Rizki Saputra	800,000	2.86	Haven't
3	Reza Oktivani	1,000,000	3.87	Haven't
….	…	…	…	…
220	Admel Brina	2,000,000	3.25	Have

Derived from the data presented in Table 2 above, it is discernible that data selection was executed on three data columns—specifically, BP Number, Place of Birth, and Date of Birth data. This action was undertaken as these three columns were deemed unnecessary for the analysis in this research, resulting in a dataset with four remaining columns. Notably, this research does not incorporate the entirety of the data selected due to its unclean nature, which still contains noise. Consequently, the subsequent step in the preprocessing phase is data cleaning. The resultant dataset post data cleaning in this research is elucidated in Table 3.

Table 3. The data cleaned

No	Student name	Parents income (PI) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)
1	Wahyudi Nasti	2,000,000	3.33	Have
2	Rizki Saputra	800,000	2.86	Haven't
3	Reza Oktivani	1,000,000	3.87	Haven't
….	…	…	…	…
200	Fauzan Satria	1,200,000	3.35	Have

Drawing insights from Table 3, it is evident that 20 rows of data have been omitted. This action is undertaken as part of the pre-processing data cleaning, which involves the removal of duplicate, inconsistent, and redundant data. The resultant dataset post data cleaning in this research is explicated in Table 4.

Table 4. The data transformation

No	Student name	Parents income (PI) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)
1	Wahyudi Nasti	2,000,000	3.33	1
2	Rizki Saputra	800,000	2.86	0
3	Reza Oktivani	1,000,000	3.87	0
….	…	…	…	…
200	Fauzan Satria	1,200,000	3.35	1

Derived from the data presented in Table 4 above, it is observable that the data, previously not formatted in numerical representation, has now been converted into a numerical format. The transformed data pertains to the parental status (PS) variable. In this transformation, if a student has both parents or one of them, it is denoted by the status 1; conversely, if both parents are absent, the status is represented as 0. Subsequent to the completion of the pre-processing procedure, the next step in the hybrid data mining process involves initiating clusterization through the utilization of the K-means method. The number of clusters in this research is set to 3, with a fixed value of centroids as seen in Table 5. The outcomes of the data clusterization are depicted in Figure 2 and expounded upon in Table 5.

Derived from Figure 2, the outcomes of data clusterization are unequivocally discernible. The blue dots denote the data situated within Cluster 0, while the red dots meticulously signify the data within Cluster 2, and the light green dots denote the data situated within Cluster 1. The black dots correspond precisely to the centroid values of the data within their respective clusters. Concretely, Cluster 0 is comprised of 119 data points, represented by blue dots, whereas Cluster 2 encompasses 48 data points, symbolized by red dots, and Cluster 1 comprises 33 data points, represented by light green dots.

Based on Table 5 above, in this research, we use two types of data clusters to predict students who will receive scholarships, namely data included in Cluster 1 and those included in Cluster 2. This is done because the data in Clusters 1 and 2 are students with parental income below Rp3,300,000 every month. The number of data with members in Cluster 1 and Cluster 2 is 81 students. To implement the C4.5 method, we need to add one more column to the right of the last column, namely "Have Received a Scholarship," with the answer being either "Yes" or "No." In Table 6, we can review the data with the additional column.

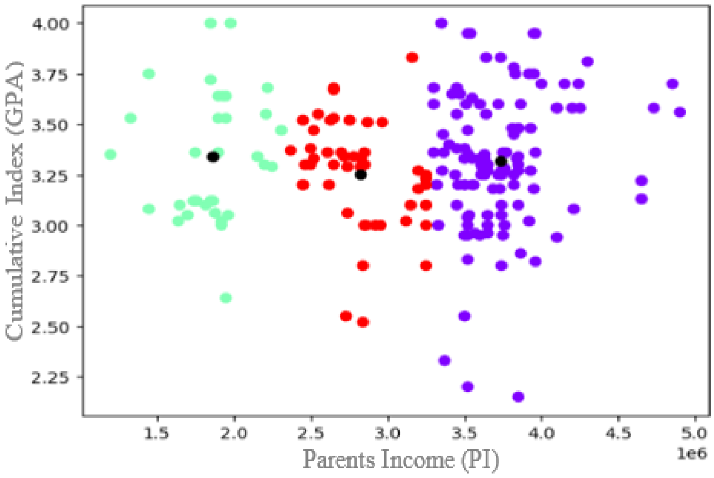

Figure 2. Result of data clusterization

Table 5. The centroid value and member of cluster data

No	Student name	Parents income (PI) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)	Cluster
1	Wahyudi Nasti	2,000,000	3.33	1	0
2	Rizki Saputra	800,000	2.86	0	2
3	Reza Oktivani	1,000,000	3.87	0	1
….	…	…	…	…	…
200	Fauzan Satria	1,200,000	3.35	1	0
	Centroid 1	4,900,000	4.00	1	
	Centroid 2	2,850,000	3.34	1	
	Centroid 3	1,200,000	2.55	0	

Table 6. The Table use to C4.5 method

No	Student name	Parents income (PI) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)	Cluster	Have received a scholarship
1	Rizki Saputra	800,000	2.86	0	2	Yes
2	Reza Oktivani	1,000,000	3.87	0	1	Yes
3	Annisa Meiza	2,900,000	3.55	1	2	No
….	…	…	…	…	…	…
81	Syahrul Furqan	2,500,000	3.60	1	1	Yes

After adding one column of data, the research continues with the C4.5 method to predict which students will receive a scholarship and create a decision tree. The data with members of Cluster 1 and Cluster 2, totaling 81 students, is divided into two types of data: training data and testing data. We determined training data as 61 data, covering rows 1 to 61, and testing data as 20 data, covering rows 62 to 81. Table 7 shows the training data and the testing data. The columns PI, GPA, and PS serve as the features columns, and the column HRS serves as the label column. Based on the data in Table 7, we then processed the data using the C4.5 method. Figure 3 shows the decision tree and prediction “Yes” and “No” students will get scholarship orwill not get scholarship from testing data.

Table 7. The trainning data and testing data

No	Student name	Parents income (PI) (Rp/Mounth)	Cumulative index (GPA)	Parents status (PS)	Have received a scholarship (HRS)
		The trainning data			
1	Rizki Saputra	800,000	2.86	0	Yes
2	Reza Oktivani	1,000,000	3.87	0	Yes
3	Annisa Meiza	2,900,000	3.55	1	No
….	…	…	…	…	…
61	Aditio Donera	2,200,000	3.50	1	No
62	Engli Saputra	2,900,000	2.95	1	Yes
63	Dhea Yolanda	3,150,000	3.20	1	Yes
64	Rahmat Ikhsan	2,850,000	2.85	0	No
….		…	…	…	…
81	Syahrul Furqan	2,500,000	3.60	1	Yes

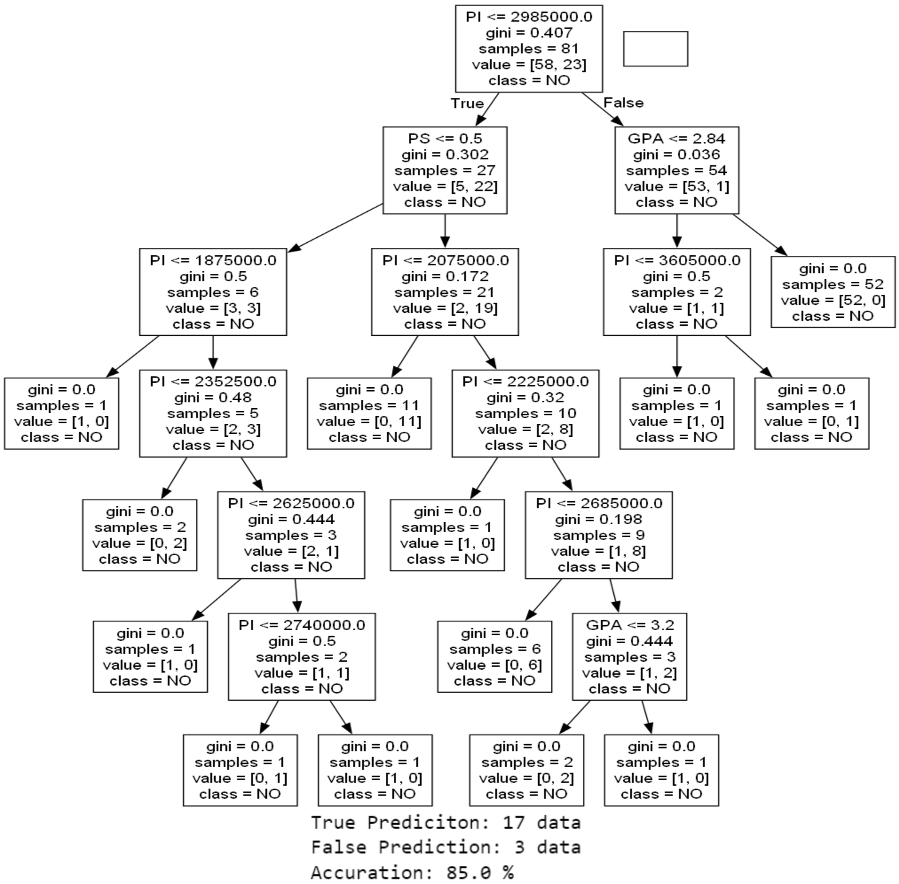

Figure 3. Decision tree

4. CONCLUSION

The research yields pivotal conclusions. Firstly, data designated for clustering must adhere to the numerical array format for successful processing. The C4.5 method requires bifurcation into training and testing datasets, each comprising features and label columns. The dataset encompasses 200 students competing for scholarships, revealing nuanced stratification into three clusters—cluster 0, cluster 1, and cluster 2—with 119, 48, and 33 members, respectively. Further analysis involves subjecting 81 instances from clusters 1 and 2 to C4.5 for predictive modeling. The dataset is divided into a training set of 61 instances and a testing set of 20 instances. The outcomes highlight the model's efficacy, with 17 accurate forecasts and a marginal discrepancy of 3 inaccuracies, achieving an 85% accuracy. The novelty lies in integrating the K-means and C4.5 methods, forming the hybrid data mining method. This research is crucial for a university in predicting scholarship recipients, benefiting stakeholders such as the vice dean III. The challenge lies in determining the most appropriate criteria as indicators for predicting scholarship recipients.

REFERENCES

[1] M. Raji, J. Duggan, B. DeCotes, J. Huang, and B. V. Zanden, “Modeling and visualizing student flow,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 510–523, Jul. 2021, doi: 10.1109/TBDATA.2018.2840986.

[2] F. Ye and A. G. Bors, “Lifelong teacher-student network learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6280–6296, Oct. 2022, doi: 10.1109/TPAMI.2021.3092677.

[3] F. Ye and A. G. Bors, “Dynamic self-supervised teacher-student network learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2022, doi: 10.1109/TPAMI.2022.3220928.

[4] S. Shen *et al.*, “Monitoring student progress for learning process-consistent knowledge tracing,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–15, 2022, doi: 10.1109/TKDE.2022.3221985.

[5] T. C. Nguyen, A. Hafeez-Baig, R. Gururajan, and N. C. Nguyen, “The hidden reasons of the Vietnamese parents for paying private tuition fees for public school teachers,” *Social Sciences & Humanities Open*, vol. 3, no. 1, 2021, doi: 10.1016/j.ssaho.2021.100105.

[6] Z. J. A. Belmonte *et al.*, “How important is the tuition fee during the COVID-19 pandemic in a developing country? Evaluation of filipinos’ preferences on public university attributes using conjoint analysis,” *Heliyon*, vol. 8, no. 11, Nov. 2022, doi: 10.1016/j.heliyon.2022.e11205.

[7] F.-C. Cheng *et al.*, “Tuition and fees for medical education and dental education in Taiwan from 1993 to 2021,” *Journal of Dental Sciences*, vol. 17, no. 3, pp. 1106–1114, Jul. 2022, doi: 10.1016/j.jds.2022.04.026.

[8] O. Hernon, E. McSharry, I. MacLaren, and P. J. Carr, “The use of educational technology in teaching and assessing clinical psychomotor skills in nursing and midwifery education: A state-of-the-art literature review,” *Journal of Professional Nursing*, vol. 45, pp. 35–50, Mar. 2023, doi: 10.1016/j.profnurs.2023.01.005.

[9] V. Marone and B. D. Heinsfeld, “‘Everyone pursuing their dreams’: Google’s and Microsoft’s discourse on educational technology,” *Computers and Education Open*, vol. 4, Dec. 2023, doi: 10.1016/j.caeo.2023.100138.

[10] B. Huntington, J. Goulding, and N. J. Pitchford, “Expert perspectives on how educational technology may support autonomous learning for remote out-of-school children in low-income contexts,” *International Journal of Educational Research Open*, vol. 5, Dec. 2023, doi: 10.1016/j.ijedro.2023.100263.

[11] P. Rodway and A. Schepman, “The impact of adopting AI educational technologies on projected course satisfaction in university students,” *Computers and Education: Artificial Intelligence*, vol. 5, 2023, doi: 10.1016/j.caeai.2023.100150.

[12] P. T. Sibiya and P. Ngulube, “Perceptions of employers in South Africa on library and information science graduates’ skills, knowledge and competencies on digital scholarship,” *Heliyon*, vol. 9, no. 2, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13531.

[13] B. Poirier, D. Haag, G. Soares, and L. Jamieson, “Whose values, what bias, which subjectivity?: The need for reflexivity and positionality in epidemiological health equity scholarship,” *Australian and New Zealand Journal of Public Health*, vol. 47, no. 5, Oct. 2023, doi: 10.1016/j.anzjph.2023.100079.

[14] A. N. Almassri, “Critical realist autoethnography in international scholarships impact research: An illustrative proposal,” *International Journal of Educational Research*, vol. 122, 2023, doi: 10.1016/j.ijer.2023.102254.

[15] M. R., A. W. Banu, and D. Mavaluru, “An efficient feature selection algorithm for health care data analysis,” *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 3, pp. 877–885, Jun. 2020, doi: 10.11591/eei.v9i3.1744.

[16] A. S. Ahmed and H. A. Salah, “A comparative study of classification techniques in data mining algorithms used for medical diagnosis based on DSS,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 5, pp. 2964–2977, Oct. 2023, doi: 10.11591/eei.v12i5.4804.

[17] K. Badapanda, D. P. Mishra, and S. R. Salkuti, “Agriculture data visualization and analysis using data mining techniques: application of unsupervised machine learning,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 1, pp. 98–108, Feb. 2022, doi: 10.12928/telkomnika.v20i1.18938.

[18] M. A. Febriantono, S. H. Pramono, R. Rahmadwati, and G. Naghdy, “Classification of multiclass imbalanced data using cost-sensitive decision tree C5.0,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 9, no. 1, pp. 65–72, Mar. 2020, doi: 10.11591/ijai.v9.i1.pp65-72.

[19] A. P. Gusman and H. Hendri, “Expert system to diagnose child development growth disorders with forward chaining method,” *Journal of Physics: Conference Series*, vol. 1339, no. 1, Dec. 2019, doi: 10.1088/1742-6596/1339/1/012045.

[20] G. W. Nurcahyo, A. P. Gusman, and H. Hendri, “Literature study on online learning as an impact of COVID 19 pandemic in education,” in *2021 International Conference on Computer Science and Engineering (IC2SE)*, Nov. 2021, pp. 1–5. doi: 10.1109/IC2SE52832.2021.9792065.

[21] H. Hendri, S. Enggari, Mardison, M. R. Putra, and L. N. Rani, “Automatic system to fish feeder and water turbidity detector using Arduino mega,” *Journal of Physics: Conference Series*, vol. 1339, no. 1, Dec. 2019, doi: 10.1088/1742-6596/1339/1/012013.

[22] D. S. Maylawati, T. Priatna, H. Sugilar, and M. A. Ramdhani, “Data science for digital culture improvement in higher education using K-means clustering and text analytics,” *International Journal of Electrical and Computer Engineering*, vol. 10, no. 5, pp. 4569–4580, 2020, doi: 10.11591/IJECE.V10I5.PP4569-4580.

[23] A. H. Nasyuha, Z. Zulham, and I. Rusydi, “Implementation of K-means algorithm in data analysis,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 2, pp. 307–313, Apr. 2022, doi: 10.12928/telkomnika.v20i2.21986.

[24] F. Aziz and A. Lawi, “Increasing electrical grid stability classification performance using ensemble bagging of C4.5 and classification and regression trees,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 2955–2962, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2955-2962.

[25] S. A. Kokatnoor and B. Krishnan, “Root cause analysis of COVID-19 cases by enhanced text mining process,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 2, pp. 1807–1817, Apr. 2022, doi: 10.11591/ijece.v12i2.pp1807-1817.

[26] R. Cordeiro de Amorim and V. Makarenkov, “On k-means iterations and Gaussian clusters,” *Neurocomputing*, vol. 553, Oct. 2023, doi: 10.1016/j.neucom.2023.126547.

[27] H.-B. Wang and Y.-J. Gao, “Research on C4.5 algorithm improvement strategy based on MapReduce,” *Procedia Computer Science*, vol. 183, pp. 160–165, 2021, doi: 10.1016/j.procs.2021.02.045.

[28] E. L. Cahapin, B. A. Malabag, C. S. Santiago Jr., J. L. Reyes, G. S. Legaspi, and K. L. Adrales, “Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms,” *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3647–3656, Dec. 2023, doi: 10.11591/eei.v12i6.4849.

[29] N. M. Mahfuz, M. Yusoff, and Z. Idrus, “Clustering heterogeneous categorical data using enhanced mini batch K-means with entropy distance measure,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 1, pp. 1048–1059, Feb. 2023, doi: 10.11591/ijece.v13i1.pp1048-1059.

[30] H. Hendri, H. Awal, and Mardison, “Solar-cell implementation for supporting tourist facilities and tourism promotion media,” *Journal of Physics: Conference Series*, vol. 1783, no. 1, Feb. 2021, doi: 10.1088/1742-6596/1783/1/012058.
[31] H. Hendri, Masriadi, and Mardison, “A novel algorithm for monitoring field data collection officers of Indonesia’s central statistics agency (BPS) using web-based digital technology,” *International Journal on Advanced Science, Engineering and Information Technology*, vol. 13, no. 3, pp. 1154–1162, Jun. 2023, doi: 10.18517/ijaseit.13.3.18302.
[32] Q. Huang, S. Chen, and Y. Li, “Selection of seismic noise recording by K-means,” *Case Studies in Construction Materials*, vol. 19, Dec. 2023, doi: 10.1016/j.cscm.2023.e02363.
[33] Y. Yu, M. Liu, D. Chen, Y. Huo, and W. Lu, “Dynamic grouping control of electric vehicles based on improved k-means algorithm for wind power fluctuations suppression,” *Global Energy Interconnection*, vol. 6, no. 5, pp. 542–553, Oct. 2023, doi: 10.1016/j.gloei.2023.10.003.
[34] X. Chen, “Design and research of MOOC teaching system based on TG-C4.5 algorithm,” *Systems and Soft Computing*, vol. 5, Dec. 2023, doi: 10.1016/j.sasc.2023.200064.
[35] A. Joshuva, R. S. Kumar, S. Sivakumar, G. Deenadayalan, and R. Vishnuvardhan, “An insight on VMD for diagnosing wind turbine blade faults using C4.5 as feature selection and discriminating through multilayer perceptron,” *Alexandria Engineering Journal*, vol. 59, no. 5, pp. 3863–3879, Oct. 2020, doi: 10.1016/j.aej.2020.06.041.

BIOGRAPHIES OF AUTHORS

Halifia Hendri is a dedicated lecturer in the Computer System Departement in Faculty of Computer Science at Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He earned his Bachelor's Degree from Universitas Negeri Padang (UNP) in the Electronics Engineering Education program in Faculty of Technics. Then he pursued a Master's Degree at Universitas Putra Indonesia YPTK Padang, specializing in Computer Science. Currently, He is engaged in doctoral studies at Universitas Putra Indonesia YPTK Padang, focusing on Information Technology within the Computer Science Faculty. His research endeavors traverse diverse domains, with particular expertise in image processing, data mining, and pattern recognition. He welcomes communication and collaboration, and he can be reached via email at: halifia_hendri@upiyptk.ac.id.

Harkamsyah Andrianof is a dedicated lecturer in the Computer System Departement within the Faculty of Computer Science at Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He earned his Bachelor's Degree from Universitas Putra Indonesia YPTK Padang in the Information System program under the Faculty of Computer Science. Subsequently, he pursued a Master's Degree at Universitas Putra Indonesia YPTK Padang, specializing in Computer Science. His academic pursuits reflect a commitment to continuous learning and scholarly excellence. His research endeavors traverse diverse domains, with particular expertise in expert system, data mining, and data science. He welcomes communication and collaboration. He can be reached via email at: harkamsyah.andrianof@upiyptk.ac.id.

Riska Robianto is a dedicated lecturer in the Computer Systems Departement in Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree and Master Degree from Universitas Putra Indonesia YPTK Padang in the Computer Systems Departement in Faculty of Computer Science. Currently, he is working on several mobile-based programming languages and is involved in several web-based projects. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cross a variety of domains, with particular expertise in mobile programming (Android programming), animation and multimedia. He welcomes communication and collaboration. He can be contacted via email at: riskarobianto@upiyptk.ac.id.

Hasri Awal is a dedicated lecturer in the Computer Systems Study Program at the Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Putra Indonesia University YPTK Padang in the Computer Systems program under the Faculty of Computer Science. Next, he continued his Masters of Computer Science at Universitas Putra Indonesia YPTK Padang with a specialization in Computer Science. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the spectrum, with particular expertise in expert systems, renewable energy and robotics. He welcomes communication and collaboration. He can be contacted via email at: hasriawal@upiyptk.ac.id.

Okta Andrica Putra is a dedicated lecturer in the Computer Systems Study Program at the Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Universitas Putra Indonesia YPTK Padang in the Computer Systems program under the Faculty of Computer Science. Next, he continued his Masters of Computer Science at Universitas Putra Indonesia YPTK Padang with a specialization in Computer Science. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the computer architecture, computer organization, and digital system. He welcomes communication and collaboration. He can be contacted via email at: okta.andrica@upiyptk.ac.id.

Romi Wijaya is a dedicated lecturer in the Computer Systems Study Program at the Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Putra Indonesia University YPTK Padang in the Computer Systems program under the Faculty of Computer Science. Next, he continued his Masters of Computer Science at Universitas Putra Indonesia YPTK Padang with a specialization in Computer Science. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the computer networks, websites, system performance analysis, algorithms, and programming. He welcomes communication and collaboration. He can be contacted via email at: wijayaromi@upiyptk.ac.id.

Aggy Pramana Gusman is a dedicated lecturer in the Computer System Study Program at the Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Universitas Putra Indonesia YPTK Padang in the information System program under the Faculty Computer Science. Next, he continued his Masters of Computer Science at Universitas Putra Indonesia YPTK Padang with a specialization in information System. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the introduction of information system, Introduction to Business Organizations, and databse. He welcomes communication and collaboration. He can be contacted via email at: apgusman@gmail.com.

Muhammad Hafizh is a dedicated lecturer in the Informatics Engineering Study Program at the Faculty of Computer Science, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Universitas Putra Indonesia YPTK Padang in the Informatics Engineering program under the Faculty of Computer Science. Next, he continued his Masters of Computer Science at Universitas Putra Indonesia YPTK Padang with a specialization in Computer Science. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the computer architecture, computer organization, and digital system. He welcomes communication and collaboration. He can be contacted via email at: muhammad_hafizh@upiyptk.ac.id.

Muhammad Pondrinal is a dedicated lecturer in the Accountancy Study Program at the Faculty of Economic and Business, Universitas Putra Indonesia YPTK Padang, Sumatera Barat, Indonesia. He obtained a Bachelor's degree from Universitas Putra Indonesia YPTK Padang in the accountancy program under the Faculty of Economic and Business. Next, he continued his Masters of Management at Universitas Putra Indonesia YPTK Padang with a specialization in Accountancy. His academic pursuits reflect a commitment to continuous learning and scientific excellence. His research efforts cut across the introduction of accounting, financial accounting, and advanced financial accounting. He welcomes communication and collaboration. He can be contacted via email at: m.pondrinal@gmail.com.