G- Tech: Jurnal Teknologi Terapan
Vol. 10, No. 3, July 2026, p. 1216-1225
E-ISSN: 2623-064 X | P-ISSN: 2580-8737

Implementation K-Means Algorithm and Decision Tree for Prediction Graduation in Vocational School Student Diploma Data

Yohanes Brekmans M Darkel[1]✉, Hilarius Alfan[2], Gerfasius Take Piran[3], Henderikus Basilius Nong Muda[4]

[1, 2, 3, 4] Teknik Informatika, Faculty of Enginnering, Universitas Nusa Nipa, Indonesia

✉ **Corresponding Author :** Yohanes Brekmans M Darkel (e-mail: yohanes.darkel@nusanipa.ac.id)

Article Information

Article History

Received : May 31, 2026
Revised : June 17, 2026
Published : July 10, 2026

Keywords **:**
K-Means;
Decision Tree;
Vocational High School Student Diploma;
Data Mining;
RapidMiner.

ABSTRACT

This study aims to analyze vocational high school (SMK) students' diploma data using two data mining approaches: the K-Means algorithm for clustering and the Decision Tree for classification and graduation prediction. The application of these two methods provides insights into the clustering structure of students based on their diploma scores and produces a more accurate graduation prediction model. Data processing was conducted using RapidMiner through preprocessing, K-Means clustering, and classification modeling with Decision Tree. The results indicate that K-Means successfully forms student clusters based on the similarity of their scores, while Decision Tree generates classification rules that can be used to predict graduation categories. This study contributes to academic evaluation and enhances data-driven decision-making in the school environment.

INTRODUCTION

Education is a process of strengthening self use prepare self For Can compete according to the current times. Education has role important use form source Power quality human beings Because development progress something nation determined by the source Power the humans in it. By general education is factor main in all field life. So that education becomes need the most respected society important. Learning outcomes is very thing noticed by others, especially by oneself yourself and your parents. For reach results maximum learning That Alone sometimes own constraint in the learning process. Factors that influence the learning process consists of from two factor that is existing factors in self students (internal) and factors consisting of from outside students (external). Internal factors are factors originating from from the person himself and is biological whereas , the factor external is factors originating from from outside like condition environment , family , environment school and environment society that can influence self somebody For Study.(Maulana et al., 2026) In the world of education performance is something very important aspect important Where success in understand or control something material That is as supplies for his future. Achievement is A results obtained from the learning process that has been done.(Abdul Majid et al., 2022)

Development technology information has push institutions education for managing and utilizing academic data in a way more effective. Schools, in particular level School Intermediate Vocational (SMK), produces various data types such as diploma grades, grades eye lessons, and graduation status. However, some the size of the data only saved as archives without analyzed

more carry on for reveal pattern academics that can support the evaluation and decision-making process decision. Through application of data mining, schools can utilize the data For dig information and knowledge new useful in increase quality education.(Anisa et al., 2025)

Student diploma data contain information important about achievements academics that can analyzed for see pattern performance as well as potential graduation. Evaluation graduation ceremony manually often takes time time, less accurate, and is influenced subjectivity. Therefore that , it is needed approach analytical data -based for map achievements student as well as predict graduation in a way more objective and measurable.(Dwi Himatul Khoiriyah & Dr. Rita Ambarwati Sukmono, 2024)

K-Means algorithm is algorithm clustering that groups data based on point center nearest cluster (centroid) with data. Purpose of K-Means is grouping data with maximize data similarity in One cluster and minimize data similarity between cluster. Size similarity used in cluster is function distance. So that maximization data similarity obtained based on distance shortest between data and centroid point.(Akbar et al., 2025) (Khoiriyah & Ambarwati, 2024) The K-Means clustering method is one of the many algorithms used in grouping data based on level similarity value. In the field education , methods This can implemented For grouping student based on diploma grades so that school can see distribution category performance academic students , such as group achieve height and group achieve.(Limia Budiarti et al., 2025) (Kaloka et al., 2026). Decision Tree is something method classification that uses structure tree Where each node represents attributes and their branches presenting mark from attribute whereas the leaves used For presenting class.Node top from this decision tree called root, stated that method This is popular method for used Because results from the model formed easy For understood.(Falakhi, 2023) Named tree decision Because established rules similar with tree.Tree formed from the sorting process recursive binary in group data, so that response variable value at each group.Concept from tree decision is transform data into decision trees and rules decision.Benefits main from use tree decision is ability simplify the retrieval process decision so that maker decision can interpret solution For problem.(Reza Pahlevi Kurniawan & Ferdiansyah, 2023) (Mochammad Luthfan Nur Rafif Falah & Novia Amilatus Solekha, 2026)

A number of study previously discuss grouping performance or prediction graduation , but Not yet Lots use accuracy second method in a way simultaneously in analyzing vocational school student diploma data.(Arfan; & Paraga;, 2024) (Ridwan, 2025) Therefore that , research This implement K-Means algorithm for grouping mark academic students and the Decision Tree algorithm for predict graduation based on attribute diploma grades. Combination second method This give more analysis comprehensive in understand pattern academic students and support evaluation graduation data-based. With Thus, the results study expected can become reference for school in do evaluation achievements learning and taking decision academic in a way more objective and data-driven.

RESEARCH METHODS

This research is a quantitative study with a data mining approach that utilizes the K-Means algorithm for the clustering process and Decision Tree for the classification process and student graduation prediction. Both methods are used because they are able to provide pattern analysis and predictions that are easy to interpret in an educational context.(Meriana Milla et al., 2025)

1. Data and Data Sources

The data used in this study is vocational school student diploma data, which includes Core subject values, Productive subject scores, Average value, Gender, Graduation statement (Pass/Fail). The data is secondary and obtained from the school's official archives.

2. Research Flow.

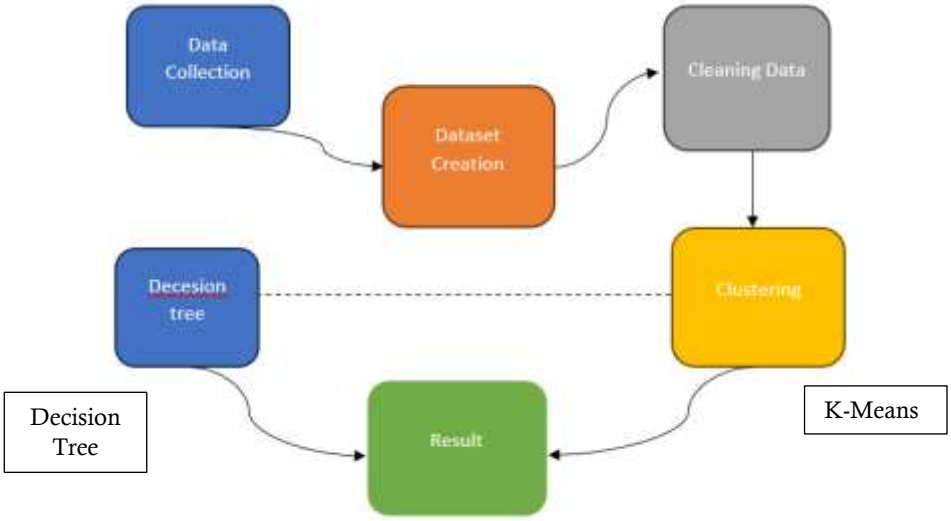

Figure 1. Research Flow

Data collection

Stage beginning study is collect data on SMK student diplomas from various department. The data taken includes, Currency value lesson general (Indonesian, Mathematics , English),b Currency value lesson productive vocational , Test scores school , and Information graduation. The data in this study comes from from diploma documents member groups that are voluntary made into as sample research. Data is summarized and compiled back into Excel format so can processed more continue to the preprocessing stage , grouping using K-Means, as well as prediction using Decision Tree.

Dataset Creation

Nama	Jenis kelamin	Jurusan	Agama	PPKN	Bahasa Indonesia	Matematika	Sejarah	Bahasa Inggris
Maria Yosefina Jari	P	Otomatisasi & Tata Kelola Perkantoran	89	96	94	90	86	80
Paulino Junifar Tokan	L	Teknik Komputer & Jaringan	85	80	85	85	80	90
Christopher Joseph Dwi Nugraha	L	Konstruksi Jalan, Irigasi & Jembatan	81	82	82	79	75	78
Lorensius Lelang Lelaona	L	Teknik Komputer & Jaringan	74	78	88	75	80	75
Sisilia Junarti	P	Teknik Komputer & Jaringan	84,17	77,14	78,94	75,66	76	78,49
Katarina Mone	P	Rekayasa Perangkat Lunak	74	65	74	73	89	76

Seni Budaya	PJOK	Muatan Lokal	Simulasi Digital	IPA	Dasar Program Keahlian	Kompetensi Keahlian	Rata-rata	Status	Keterangan
92	97,5	88	92,5	90	95	92	92,27	Lulus	Sangat Baik
77	85	88	82	80,5	80	84	82,69	Lulus	Sangat Baik
93	79	88	80	78	78	81	79,69	Lulus	Baik
80	80	88	75	76,35	77,63	77,43	78,39	Lulus	Baik
78,37	86,6	88	77,57	72,97	75,1	78,22	77,53	Lulus	Baik
66,8	78	88	70	70	89	73	73,55	Tidak Lulus	cukup

Figure 2. Student Diploma Data Set Table Vocational school students

Data that has been collected Then arranged become a research dataset. This stage includes make data table contains all values each students , make Graduation label column , combines data from a number of major become one dataset, ensuring a uniform data format. This dataset will be used for clustering and classification processes.(Falakhi, 2023)(Limia Budiarti et al., 2025) (Khoiriyah & Ambarwati, 2024)

Data Cleaning

To get accurate results , the process of deleting empty data (missing values), matching the number or decimal format , normalizing value to be ready used K-Means method , eliminating outliers if there is. Pre-processing very important because K-Means is sensitive to scale mark.

Clustering

Variables used in clustering is all over attribute mark eye subjects , such as , Religion , PPKN, Indonesian, Mathematics , Digital Simulation , Physical Education , Products Creative

Entrepreneurship , Average of All variables In clustering, labels are not used. This stage aims to group student based on similarity diploma grades. What is done in RapidMiner determines number of clusters (k), running K-Means, The results of this clustering help see pattern academic student without see major.(Aprilia & Aribowo, 2022)

Decision Tree

At the stage classification, used variable:

attributes: all mark eye lesson

Label (target): Description Graduation (Pass/Fail)

At this stage it is done prediction graduation based on diploma value. Steps in RapidMiner, determine attribute mark as input, set column Graduation as target label, run Decision Tree algorithm (C4.5), get rule graduation in the form of tree decisions , analyzing the most influential nodes to graduation. Decision Tree gives description factor mark what is most decisive students pass or not.

Results

In this section , the researcher explain characteristics each student cluster K-Means results, pattern graduation based on Decision Tree, value eye most influential lesson to graduation , implications results for school.

RESULTS AND DISCUSSION

Result

A. Decision Tree

Decision Trees work by creating tree-like classification rules that can be used to predict graduation categories. Modeling Objectives, The purpose of the Decision Tree stage is to predict student graduation. based on diploma grades and to obtain graduation rules in the form of a decision tree. The algorithm used in this research is Decision Tree (C4.5).

Data Process and Input In Decision Tree modeling in RapidMiner, the diploma grade columns (general subject grades, vocational productive, and school exam grades) are set as input (attributes). The “Graduation” column is set as the target label. Decision Tree uses data that has gone through the stages of data collection, dataset creation, and data cleaning.

B. Benefits of Results

Decision Tree provides an overview of what value factors are most determining students pass or not. By analyzing the nodes that most influence graduation, schools can gain information for academic evaluation and data-driven decision-making. The rules generated by the Decision Tree (e.g., if Math Score > X and Productive Score > Y, then Pass) can be used for early intervention or providing more targeted academic guidance for students at high risk of failing.

Figure 3. Rapidminer tool display design

This process flow aims to predict student graduation (Status) , based on diploma grades using the Decision Tree algorithm. The modeling process begins by loading student diploma data using Excel's Read operator. Preprocessing is then performed, starting with the Replace Missing Values operator to address missing values by filling them in (e.g., using the average). The next crucial step is the Set Role operator, which sets the Status column (Graduation) as the label (target variable) to be predicted by the model, ensuring all other value columns are input attributes. If non-numerical data is present, the Nominal to Numerical operator is used to convert it to a numeric format ready for modeling. The data is then divided using the Split Data operator into training data (e.g., 70%) and test data (e.g., 30%) to objectively test the model's performance. Finally, the Decision Tree operator is run to build a model that generates classification rules (decision trees) based on the input values, which are then evaluated using the Performance operator to determine the model's accuracy.

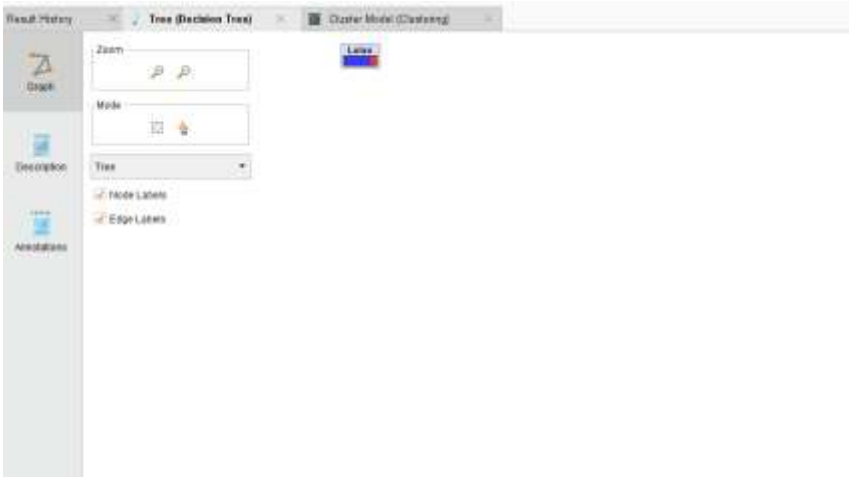

Figure 4. Display of the graph results of the decision tree method in rapidminer.

The results display a graphical representation of the decision tree. While the detailed structure isn't fully visible, the topmost node (the root) displays the label "Pass." This indicates that the final outcome or prediction displayed is centered around the student's passing category.

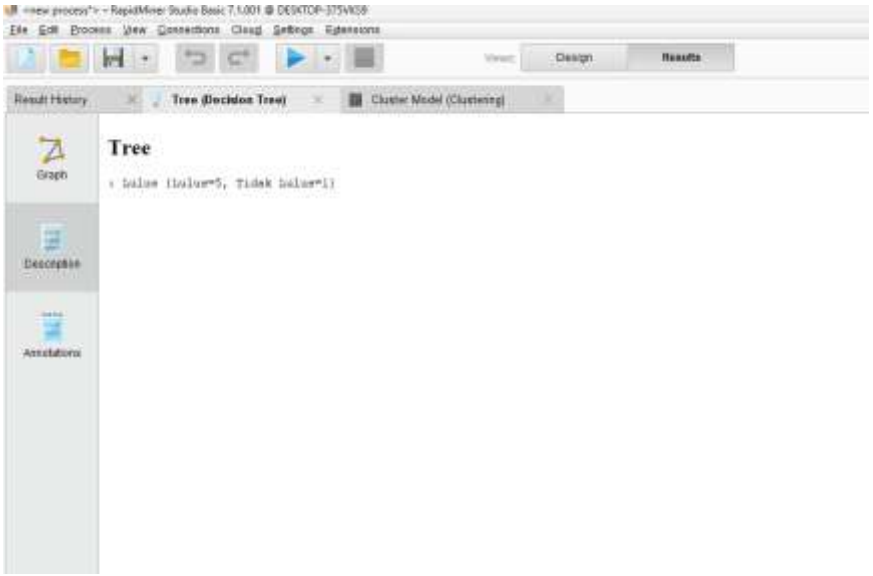

Figure 5. Display of the results of the Description Decision tree

The Decision Tree description section provides a statistical summary of the classification results. The visible output is:

Pass {Pass=5, Fail=1}
This means that from a total of 6 processed sample data (5 + 1), the Decision Tree model classifies the results as follows:
Pass = 5: There are 5 student data that are predicted or classified as "Pass".
Failed = 1: There is 1 student data that is predicted or classified as "Failed". This figure shows how the diploma data has been grouped and classified by the Decision Tree model based on the rules formed from the input values.

C. Clustering

The K-Means Clustering process in RapidMiner begins with the Read Excel operator to load student diploma data, followed by the Replace Missing Values preprocessing step to address missing values and maintain data quality. Next, the Select Attributes operator is used to select only diploma grade attributes (general subject scores, vocational productive scores, and school exam scores) as input for clustering. Then, the Nominal to Numerical operator is run to convert non-numerical data to numeric values.

Figure 6. Clustering display in rapidminer

Figure 6. is an important step because the K-Means algorithm must process data in numeric format. Once the data is ready, the Clustering operator (K-Means) is run, where the researcher determines the desired number of clusters (k), and the algorithm works iteratively to group each student into the cluster whose center (centroid) is closest to their grades, thus producing groups of students with similar grade characteristics.

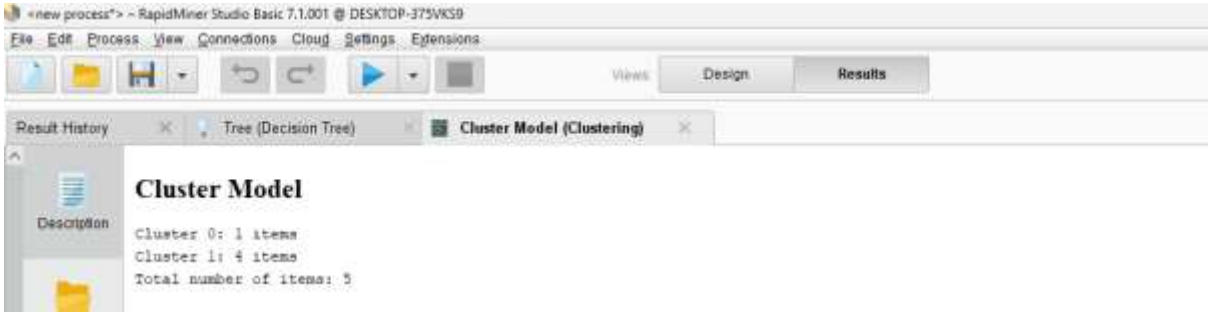

Figure 7. Cluster model results on rapiminer.

Figure 7. is The results of the Cluster Model description show the details of the data that was successfully grouped: These results show that from a total of 5 sample data processed for clustering, the K-Means algorithm successfully divided them into two groups. Cluster 0 is a group consisting of 1 student, and Cluster 1 is a group consisting of 4 students.

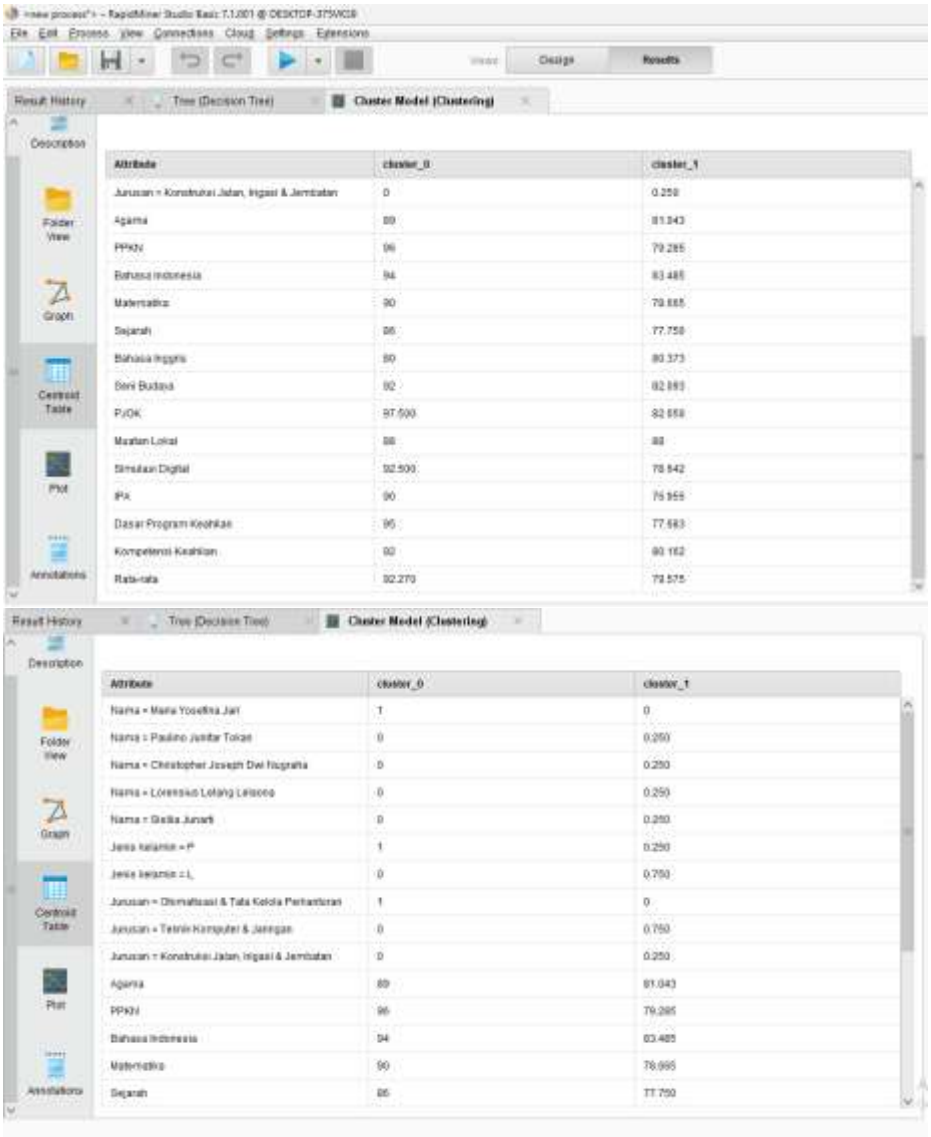

Figure 8. Centroid table clustering results in rapidminer.

Figure 8 displays the Centroid Table, showing the average value (centroid) of each attribute (subject score) within each cluster. This is key to interpreting the meaning of each cluster.

Interpretation of Results:

Cluster 0 (High Value Group):

Cluster 0 centroid shows consistently higher mean scores across almost all subjects (example: Total Mean 92.270).

This group represents students with characteristics of very good or above average diploma grades.

Cluster 1 (Medium/Intermediate Value Group):

The centroid of Cluster 1 shows a consistently lower mean value than Cluster 0 (e.g., Total Mean 79.575).

This group represents students with characteristics of diploma grades that are around or slightly above the standard passing limit.

Implications of Results

The results of this grouping help schools see students' academic patterns without being tied to majors. Through this mapping, schools can:

Academic Evaluation: Identifying which groups of students require further intervention and which groups can serve as role models.

Data-Driven Decision Making: Using these results as a basis for developing targeted guidance programs or curriculum.

Discussion

Comparison of K-Means and Decision Tree Results These two algorithms complement each other in academic data analysis, K-Means helps in understanding the data structure by grouping students who have similar grades, providing a picture of the academic profile purely based on grade data without being influenced by graduation labels. Decision Trees help in creating transparent predictive models, providing explicit rules about which value criteria will result in a particular passing prediction. In short, K-Means answers "Who are similar students?" based on grades, while Decision Tree answers "When is a student predicted to Pass or Fail?" based on grade rules.

Table 1. Comparison of K-Means and Decision Tree Results

Aspect Comparison	K-Means Algorithm (Clustering)	Decision Tree Algorithm (Classification)
main purpose	Clustering. Grouping student based on similarity characteristics their diploma grades	Classification and Prediction. Predicting category graduation students (Pass/ Fail).
Type Learning	*Unsupervised Learning* (No using target labels).	*Supervised Learning* (Using the target label, namely " Graduation ").
Main Output	Cluster Model (Data group) and Centroid Table (Mean value profile) each group).	Decision *Trees* and Rules Classification
Results Obtained	Forming group student based on profile value. Example : Cluster 0 (High Value) vs. Cluster 1 (Medium Value).	Produce rules that can used For prediction graduation. Example : If the Mathematics score > X, then Pass.
Benefit for School	Give description about structure grouping student in a way natural. Useful For segmentation and evaluation general academic without see results end.	Give description about factor mark what is most decisive students graduate or useless For intervention early and taking decision predictive.
Sample Data Results	Cluster 0 has 1 item and an average value of 92.270.	Prediction output shows 5 students passed and 1 student failed Failed from total 6 samples.

CONCLUSION

Based on the results of research on the application of the K-Means and Decision Tree algorithms to vocational high school student diploma data, it can be concluded that these two data mining methods are able to provide important information in the academic evaluation process and support data-based decision-making. First, the K-Means algorithm successfully grouped students into two clusters based on the similarity of diploma scores. Cluster 0 is the group with a higher average score, thus depicting students with good achievements, while Cluster 1 shows a lower average score and represents students with average academic achievement. The results of this grouping can be used by schools to identify students' academic profiles, design learning strategies, and determine more targeted academic interventions.

Second, the Decision Tree algorithm is capable of generating a classification model that predicts student graduation based on subject grade attributes. The model generates easy-to-understand graduation rules and can be used to assess the grade factors that most influence student graduation status. The classification results predict that most students will be in the Pass category, with some students in the Fail category, based on the characteristics of the input grades tested.

Overall, the combination of Clustering and Decision Tree methods proved effective for analyzing vocational high school students' academic data. The use of these two algorithms provided a comprehensive overview, both in terms of grade profile grouping and graduation prediction. The results of this study are expected to serve as a basis for schools to improve the quality of academic evaluations, provide more appropriate academic guidance, and support data-driven decision-making.

REFERENCES

Abdul Majid, M. B., Cani, Y. M., & Enri, U. (2022). Penerapan Algoritma K-Means dan Decision Tree Dalam Analisis Prestasi Siswa Sekolah Menengah Kejuruan. *Jurnal Sistem Komputer Dan Informatika (JSON)*, *4*(2), 355. https://doi.org/10.30865/json.v4i2.5299

Akbar, I., Samad, I. S., Rahmat, R., & Rosmiana, S. (2025). Data Mining Analysis of K-means Algorithm and Decision Tree for Early Detection of Students at Risk of Dropping Out. *Journal of Informatics Information System Software Engineering and Applications (INISTA)*, *7*(2), 148–162. https://doi.org/10.20895/inista.v7i2.1630

Anisa, A., Turmudi Zy, A., & Hadikristanto, W. (2025). Klasifikasi Kualitas Air dengan K-Means dan Decision Tree. *Jurnal Pustaka AI (Pusat Akses Kajian Teknologi Artificial Intelligence)*, *5*(2), 166–176. https://doi.org/10.55382/jurnalpustakaai.v5i2.1114

Aprilia, A. S., & Aribowo, B. (2022). Simulasi Rancangan Pemetaan Sekolah dengan Metode Algoritma Machine Learning Menggunakan Software RapidMiner. *JURNAL Al-AZHAR INDONESIA SERI SAINS DAN TEKNOLOGI*, *7*(1), 20. https://doi.org/10.36722/sst.v7i1.877

Arfan;, U., & Paraga;, N. (2024). The Comparison of K-Means , Naïve Bayes and Decision Tree Algorithm in Predicting Fuel Oil Sales Perbandingan Algoritma K-Means , Naïve Bayes dan Decision Tree dalam Memprediksi Penjualan Bahan Bakar Minyak. *Institut Riset Dan Publikasi Indonesia (IRPI) MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(October), 1379–1389.

Dwi Himatul Khoiriyah, & Dr. Rita Ambarwati Sukmono, S. E.. M. M. (2024). *Analisis Segmentasi Dinamis Untuk Layanan Ekspedisi: Mengintegrasikan K-Means dan Decision Tree*.

Falakhi, A. (2023). Pengolahan Data Pelanggan Dengan Tenik Clustering K-Means Di Aplikasi Weka. *Journal Computer Science and Information Systems : J-Cosys*, *3*(2), 54–60. https://doi.org/10.53514/jco.v3i2.394

Kaloka, T. P., Marta, A., & Syofrin, N. D. (2026). *K-Means and Decision Tree Models for Analysis Students Perception of Digital Marketing*. *6*(April), 453–461.

Khoiriyah, D. H., & Ambarwati, R. (2024). Dynamic Segmentation Analysis for Expedition Services: Integrating K-Means and Decision Tree. *Journal of Information Systems and Informatics*, *6*(1), 363–377. https://doi.org/10.51519/journalisi.v6i1.666

Limia Budiarti, R., Mulyati, S., & Zaidan, A. (2025). PENERAPAN DATA MINING UNTUK ANALISIS KESEHATAN BALITA MENGGUNAKAN METODE *K-MEANS* DAN *DECISION TREE* (STUDI KASUS PUSKESMAS PAAL LIMA). *FORTECH (Journal of Information Technology)*, *9*(2), 89–103. https://doi.org/10.53564/1kfq0k87

Maulana, H. A., Fitriyah, N., Baradei, D. El, Ardiyanto, F., & Arifin, M. (2026). *Analisis Data Pendidikan Menggunakan Metode Data Mining dengan Algoritma Decision Tree dan K-Means Clustering*. *6*(1).

Meriana Milla, Vinsensius Aprila Kore Dima, & Agustina Purnami Setiawi. (2025). Penerapan

Algoritma K-Means untuk Mengidentifikasi Minat Belajar Siswa di Sekolah Dasar Negeri Puu Naga. *Modem : Jurnal Informatika Dan Sains Teknologi.*, *3*(4 SE-Articles), 1–12. https://journal.aptii.or.id/index.php/Modem/article/view/630

Mochammad Luthfan Nur Rafif Falah, & Novia Amilatus Solekha. (2026). Determinan Pendidikan di Indonesia Menggunakan Metode K-Means Clustering. *Jurnal Harmoni Pendidikan*, *2*(1), 57–69. https://doi.org/10.64845/jhp.v2i1.161

Reza Pahlevi Kurniawan, & Ferdiansyah. (2023). Penerapan Algoritme K-Means Clustering Untuk Mengelompokkan Siswa Berdasarkan Nilai Akademik Di Smp Negeri 207 Ssn. *Prosiding Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI)*, *2*(2), 530–538.

Ridwan. (2025). Implementasi Data Mining Menggunakan Algoritma K-Means Clustering dan Random Forest untuk Prediksi Status Kesehatan Karyawan. *Jurnal Advance Research Informatika*, *4*(1), 1–10. https://www.ejournalwiraraja.com/index.php/JARS