INDONESIAN JOURNAL OF MULTIDISCIPLINARY SCIENCE

P-ISSN: 2808-5957
E-ISSN: 2808-6724

Application of K-Means and C4.5 Algorithms for dropout risk prediction in vocational high schools

Syahrul Anwar*,
Politeknik Siber Cerdika Internasional, Indonesia
*Email for Correspondence: syahrul@polteksci.ac.id

ABSTRACT

The issue of dropout remains a serious challenge at the vocational high school (SMK) level, including at SMK Islam Nusantara. Over the past five years, the school has experienced a dropout rate of 20% of all enrolled students. This study aims to identify students at high risk of dropping out by applying the K-Means and C4.5 algorithms. The K-Means algorithm is used to cluster students based on factors such as academic performance, socioeconomic conditions, and the distance from home to school. Subsequently, the C4.5 algorithm is used to predict dropout risk based on the clustering results. The data used in this study were obtained directly from the Dapodik Data Center of SMK Islam Nusantara and include student data from the 2019 to 2024 academic years. The results indicate that socioeconomic factors, the distance from home to school, and academic performance significantly influence dropout risk. Students from low-income families and with poor academic performance were found to be at the highest risk of dropping out. This study makes a significant contribution to SMK Islam Nusantara by developing an early warning system that can help identify students at risk of dropping out. With more targeted interventions, such as academic counseling and socioeconomic support, it is expected that the dropout rate can be significantly reduced. Additionally, this predictive model can be applied to other vocational schools with similar conditions to improve the overall quality of vocational education.

Keywords:

C.45
dropout
data mining
K-Means
risk prediction

This is an open access article under the CC BY-SA license.

INTRODUCTION

The problem of dropout or dropping out of school is still a serious challenge in the global education system, especially at the Vocational High School (SMK) level (Wuni et al., 2020). The high dropout rate can have a negative impact on the quality of available human resources, as well as have a direct impact on the economy (Kim & Kim, 2018). This phenomenon is further exacerbated by various factors, such as socioeconomic conditions, lack of family support, and less effective interaction in the school environment (Kim & Kim, 2018; Liang et al., 2016). In Indonesia, SMK Islam Nusantara, located in Indramayu Regency, is experiencing a similar problem. In the last five years, around 20% of all registered students have experienced dropouts. This phenomenon reflects the conditions faced by many other vocational schools in Indonesia and even in the world (Yi et al., 2015).

Previous research has proven that dropout is influenced by various factors, both physical, social, economic, and academic (Kassim & Sukor Shaari, 2014; Marhayati et al., 2023; Markos et al., 2019; Mustakim et al., 2024; Tanjung & Ramadhani, 2023). Several studies highlight that the economic condition of families and the quality of social interaction in schools play a significant role in the increased risk of students dropping out of school. On the other hand, the use of data mining and machine learning techniques has been proven effective in predicting students who are at high risk of dropout (Muhamad et al., 2019). For example, the C4.5 decision tree algorithm has been used in several studies to predict dropout risk and show that geographical factors and social conditions have a significant influence (Daulay et al., 2022). In addition, several other studies have also

developed ensemble-based prediction models that aim to reduce dropout rates in educational institutions (Amalia et al., 2021; Yi et al., 2015).

This research is specifically focused on vocational schools because of the unique characteristics possessed by these vocational schools. Vocational schools have a vocational education orientation that prepares students to immediately enter the world of work after graduation. This makes the challenges faced by vocational school students more complex compared to students in Madrasah Aliyah (MA) or Senior High School (SMA), which generally focus more on preparing to continue their education to higher education. Factors such as readiness to enter the world of work, socio-economic conditions, and family background play a significant role in determining the risk of dropout in vocational schools (Liang et al., 2016). In addition, vocational school students generally come from more diverse socio-economic backgrounds, with many of them coming from families with lower middle economic levels, so economic factors have a significant influence on students' decisions to continue or leave education.

Identifying students at risk of dropout early on can allow for quicker and more targeted interventions, such as the provision of additional guidance, psychological support, or a specially designed mentoring program. In this context, this study focuses on the application of K-Means and C4.5 algorithms to group students based on their risk profile and predict the risk of dropout in vocational schools (Mduma et al., 2019). By using the K-Means algorithm for clustering and the C4.5 algorithm for prediction, it is hoped that schools can implement an effective early warning system in identifying students who are at high risk of dropping out of school (Guntara & Suprawoto, 2022). The results of this study are not only expected to contribute to the development of early warning systems, but also help educational institutions in designing more effective intervention strategies to reduce dropout rates.

METHOD

This study uses a quantitative approach with data mining and machine learning techniques to predict the risk of student dropout at SMK Islam Nusantara. The design of this study consists of several stages which include data collection, data processing, application of the K-Means algorithm for clustering, and C4.5 algorithm for prediction, as well as evaluation of results, as depicted in the figure below. This study uses a quantitative approach with the student population of SMK Islam Nusantara located in Indramayu Regency. The population targeted by the study included all students enrolled in the 2019 to 2023 school year. To ensure representative results, this study uses a random sampling technique, with samples taken based on the proportion of students who drop out and who remain in school.

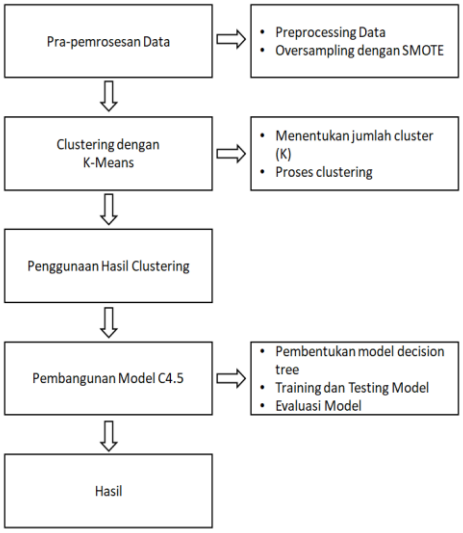

Figure 1. Data Processing Flow

The data used in this study was obtained directly from the Dapodik Data Center of SMK Islam Nusantara. This data includes information related to students' academic achievement (report card scores), socio-economic background (such as family income and parental employment), as well as student health data.

This research is specifically focused on the scope of SMK Islam Nusantara with the aim of identifying students who are at high risk of dropout, so that appropriate intervention can be carried out by the school.

Before analysis, the data will go through several stages of processing: Data Preprocessing: Carrying out data cleaning, handling missing values, and normalizing data if needed. This includes handling lost value, outliers, and data transformation (Tanjung & Ramadhani, 2023).

Oversampling with SMOTE: Since dropout data is usually unbalanced (fewer students dropout than students who remain in school), the SMOTE (Synthetic Minority Over-sampling Technique) technique will be applied to address class imbalances (Kassim & Sukor Shaari, 2014).

The K-Means algorithm will be used to group students based on their profile similarity (Rizky et al., 2021). The clustering process is carried out with the following steps:

1) Determining the number of clusters (K): Using the Elbow Method method to determine the optimal number of clusters to be used.
2) Clustering process: Once the number of clusters is determined, the K-Means algorithm will group students based on the variables that have been collected (academic achievement, socio-economic, health, and social integration).

The results of K-Means will provide an overview of groups of students based on their characteristics (Marhayati et al., 2023), which will then be used for dropout risk prediction.

After the grouping, the next step is to use the C4.5 algorithm to predict the risk of student dropout (Muhamad et al., 2019). The prediction process includes: (1) Decision tree model formation: The C4.5 algorithm will create a decision tree based on input variables and the results of K-Means clustering; (2) Training and Testing Model: The data will be divided into training data (80%) and testing data (20%) to train and test the accuracy of the prediction model; (3) Model Evaluation: The accuracy of the model will be evaluated using evaluation metrics such as Accuracy, Precision, Recall, and F1-Score. In addition, the Confusion Matrix will be used to understand the prediction distribution of the model.

RESULTS AND DISCUSSION

K-Means Algorithm Process

In K-Means clustering to identify clusters that may be at risk of dropout based on several important factors. The steps we will take are:

Selecting relevant features

To identify groups of students at risk of dropout, we used the K-Means clustering technique. This process begins by selecting the features that are considered relevant in influencing the risk of dropout, namely the parent's income, distance from home to school, and the student's academic grades. The selection of this feature was made based on the hypothesis that economic factors, school accessibility, and academic performance have a significant influence on a student's likelihood of dropping out of school.

Determining the number of clusters (K)

The next important step is to determine the optimal number of clusters (K). This number of clusters describes how many different groups there are among students in terms of dropout risk. To find the optimal number of clusters, we use the "Elbow Method" method. This method involves testing multiple K values and plotting the total within-cluster sum of squares (WCSS) for each K value (Santosa et al., 2021). The optimal number of clusters is usually indicated by the point at which the WCSS drop begins to slow down significantly, resembling an elbow shape on a graph—hence the so-called "Elbow Method" (Marhayati et al., 2023).

The Elbow Method graph shown shows the change in the Within Cluster Sum of Squares (WCSS) as the number of clusters (K) increases. The "elbow" point on the graph signifies the optimal number of clusters for this analysis. Based on the Elbow graph visible, the "elbow" point or significant change in WCSS appears to be around 3 or 4 clusters. This indicates that K=3 or K=4 is a good choice for the optimal number of clusters.

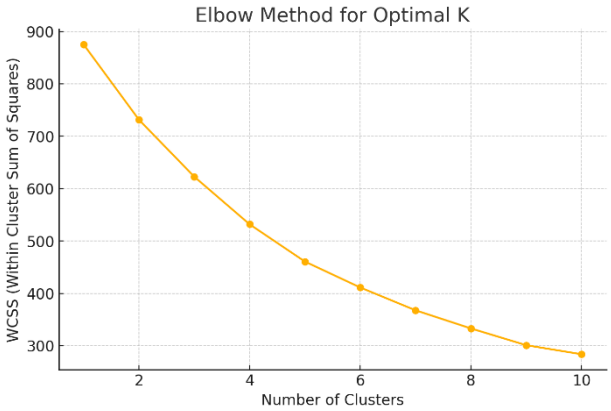

Figure 2. The Number of Clusters (K)

Clustering

As a first step, we tested the use of three clusters (K=3) to group student data based on relevant variables, namely parental income, distance from home to school, and academic grades.

Attribute	cluster_0	cluster_1	cluster_2
Berat Badan	50.337	45	53.750
Tinggi Badan	155.054	145	155.625
Jml. SaudaraKandung	2.301	3	2.625
Jarak Rumahke Sekolah (KM)	8.512	13	8
NILAI S1	85.295	77	85.750
NILAI S2	84.518	75	84.875
NILAI S3	85.151	78	87.875
Penghasilan Fix	1022590.361	10000000	4000000

Figure 3. Results of 3 Clusters

The clustering results showed three groups of students with the following average characteristics: Cluster 0 consisted of students with a family income of around 1.02 million, an average distance from home to school of 7.88 km, and academic scores ranging from 88.7 for S1, 85.4 for S2, and 79.7 for S3. Cluster 1 also has a family income of around 1.02 million, but with a longer distance from home to school, which is an average of 8.94 km, as well as an academic score of 82.99 for S1, 83.9 for S2, and 88.8 for S3. Meanwhile, Cluster 2 shows a much higher family income, around 4.67 million, with an average distance from home to school of 8.55 km, and an academic score of 84.7 for S1, 83.7 for S2, and 86.7 for S3. These findings show significant differences in economic factors, accessibility, and academic achievement among the groups of students identified through this clustering method.

Next, we test the use of four clusters (K=4) to find out if four clusters can be more optimal in their grouping.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Berat Badan	49.465	45	53.750	56.045
Tinggi Badan	154.812	145	155.625	156.636
Jml. SaudaraKandung	2.278	3	2.625	2.455
Jarak Rumahke Sekolah (KM)	8.340	13	8	9.636
NILAI S1	84.889	77	85.750	87.955
NILAI S2	84.722	75	84.875	83.182
NILAI S3	85.021	78	87.875	86
Penghasilan Fix	873263.889	10000000	4000000	2000000

Figure 4. Results of 4 Clusters

The results of clustering with four clusters showed different characteristics among the student groups. Cluster 0 consists of students with an income of around 1.08 million, an average distance from home to school of 12.15 km (the farthest among other clusters), and an academic score of 83.4 for S1, 81.5 for S2, and 86.6 for S3. Cluster 1 shows students with an income of around 1.00 million, the distance from home to the nearest school with an average of 4.53 km, as well as an academic score of 80.44 for S1, 83.9 for S2, and 87.3 for S3. Meanwhile, Cluster 2 includes students with the lowest income, which is around 983 thousand, the average distance from home to school is 7.87 km, and the highest academic score for S1 (90.65) and S2 (87.88), but the S3 score is lower than other clusters, which is 82.08. Cluster 3 consists of students with much higher incomes, around 4.66 million, an average distance from home to school of 8.55 km, and an academic score of 84.7 for S1, 83.7 for S2, and 86.7 for S3. Based on these characteristics, Cluster 2 can be identified as a group at higher risk of dropout due to its lowest average income and lower S3 value compared to other clusters.

A comparison between the clustering results with 3 clusters and 4 clusters showed that when using 3 clusters, the group of students could be differentiated based on income, distance to school, and academic grades, with one of the clusters (Cluster 0) showing a combination of factors at a higher risk of dropout. However, with 4 clusters, the analysis became more granular, revealing an additional group that showed a higher risk of dropout (Cluster 2) due to the lowest earnings and declining S3 values. The 4-cluster approach provides sharper detail in identifying groups of students who are potentially at risk of dropout, allowing for more targeted interventions.

Predicting Dropout Risk with the C4.5 Algorithm

The next step is to use the clustering results as a feature to predict students who have the potential to drop out using the C4.5 (Decision Tree) algorithm.

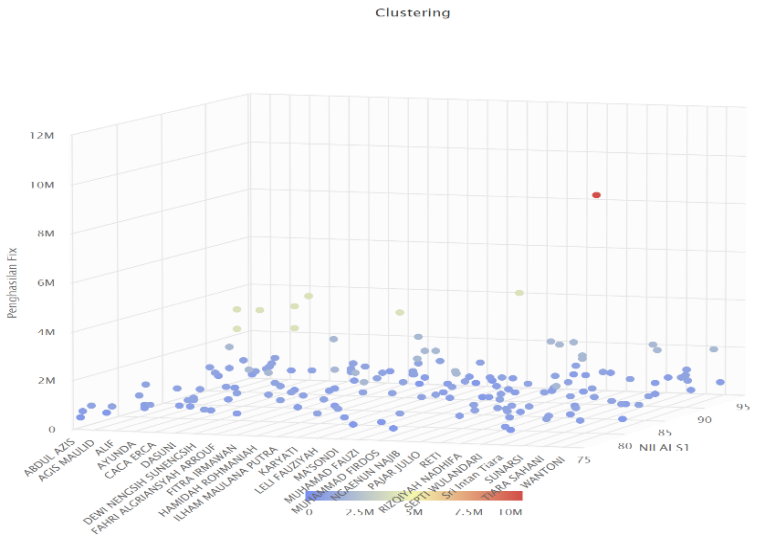

Figure 5. Dropout Risk Prediction

Flag students who are at risk of dropout

In machine learning, it is important to determine the features and targets to be predicted (Daulay et al., 2022). At this stage, we select some features from the dataset that we consider relevant to predict whether a student has the potential to drop out or not.

1) Feature: The input variable we use to train the model. The features selected from the dataset are:
   - Fix Income: The income of the student's parents.
   - Distance from Home to School (KM): The distance from the student's home to the school.
   - S1, S2, S3 GRADES: Academic grades of students from several subjects.
2) Target (Label): The output variable that we predict, i.e. whether the student has the potential to drop out or not. This is obtained from the results of the previous clustering, where:
   - 1 = Dropout (potential)
   - 0 = No Dropout (no potential).

These features and targets are separated to train the model.

Build and train the C4.5 model

After determining the feature of the criteria, we divide the dataset into two parts (Haji, 2024):

1) Training Data: The part of the data used to train the model. The model will learn patterns from this data.
2) Test Data: The part of the data that is not used in training. The model will be tested using this data to see how well the model performs on data that has never been seen before.

In this case, we use 70% of the training data and 30% of the test data.

Building the C4.5 Model (Decision Tree) and visualization

In the process of testing the imported data, the researcher uses training data and data testing to determine a decision tree model by using a comparison of 70% for training data and 30% for test data.

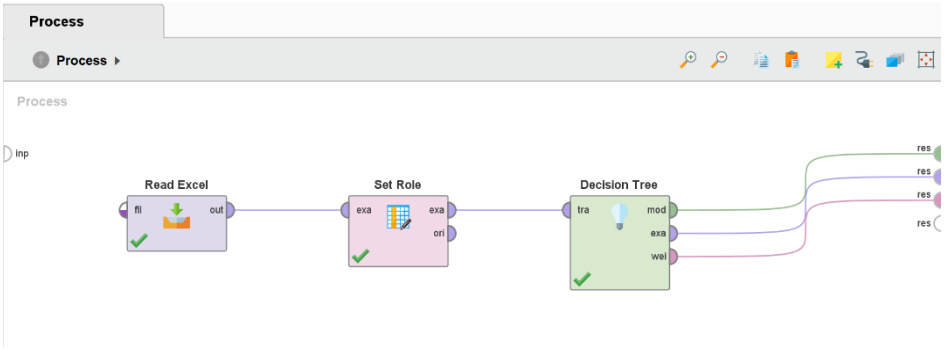

Figure 6. Data Processing

The decision tree visualization below shows how the model makes decisions to predict students who are likely to drop out or not, based on features such as income, distance from home to school, and academic grades.

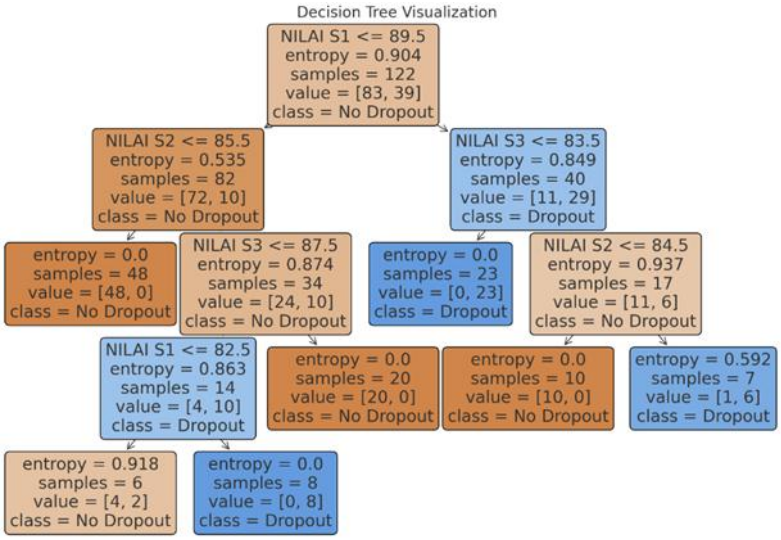

Figure 7. Decision Tree Model

Each node indicates a division rule, while a leaf indicates the final prediction for "No Dropout" or "Dropout". Darker colors indicate the more dominant nodes in the classification.

The results of this study show that the main factors that affect the risk of dropout in SMK Islam Nusantara include the socio-economic condition of the family, the distance from home to school, and the academic achievement of students. Using the K-Means algorithm for clustering, students were grouped based on the similarity of their profiles, which was then used to predict the risk of dropout using the C4.5 algorithm. The clustering analysis identified that groups of students with lower family income and lower academic achievement had a higher risk of dropout.

These results show that in SMK Islam Nusantara, economic and academic factors have a significant role in determining whether students will continue their education or not. The distance from home to school also plays an important role, especially for students who live far from school, which can affect attendance levels and learning motivation. Therefore, effective interventions, such as providing financial support, academic guidance programs, or more affordable transportation, can be a solution to reduce the dropout rate at SMK Islam Nusantara.

The prediction system developed can be implemented as a tool for schools to identify students who have the potential to experience dropouts early. As such, schools can design more targeted intervention programs to help students overcome the challenges they face. This prediction model is not only relevant for SMK Islam Nusantara but can also be adapted and applied in other vocational high schools in Indonesia that are facing similar challenges.

CONCLUSION

Overall, this study succeeded in providing a clearer picture of the profile of students who are at high risk of dropping out at SMK Islam Nusantara. By using data mining and machine learning approaches, this study provides practical recommendations that can be applied by schools in an effort to reduce the dropout rate and improve the quality of vocational education at SMK Islam Nusantara.

REFERENCES

Amalia, N. L. R., Supianto, A. A., Setiawan, N. Y., Zilvan, V., Yuliani, A. R., & Ramdan, A. (2021). Student Academic Mark Clustering Analysis and Usability Scoring on Dashboard Development Using K-Means Algorithm and System Usability Scale. *Jurnal Ilmu Komputer Dan Informasi*, *14*(2). https://doi.org/10.21609/jiki.v14i2.980

Daulay, S., Apriani, W., & Perwira, Y. (2022). Application Of Data Mining For Prediction Of Students Out Of College With The Method Algorithm C4.5. *Jurnal ICT : Information and Communication Technologies*, *13*(1). https://doi.org/10.35335/jict.v13i1.28

Guntara, M., & Suprawoto, T. (2022). Drop Out Student Clusterization Using the k-Medoids Algorithm. *JTKSI (Jurnal Teknologi Komputer Dan Sistem Informasi)*, *5*(1), 61–66.

Haji, U. (2024). PREDIKSI TINGKAT KELULUSAN MAHASISWA S1 UNIVERSITAS AN NUUR DENGAN METODE DECISION TREE C4. 5. *Julia: Jurnal Ilmu Komputer An Nuur*, *4*(1), 21–29.

Kassim, M., & Sukor Shaari, A. (2014). THE IMPACT OF QUANTUM TEACHING STRATEGY ON STUDENT ACADEMIC ACHIEVEMENTS AND SELF-ESTEEM IN INCLUSIVE SCHOOLS School of Education and Modern Languages Universiti Utara Malaysia. *Malaysian Journal of Learning and Instruction*, *11*.

Kim, D., & Kim, S. (2018). Sustainable education: Analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability (Switzerland)*, *10*(4). https://doi.org/10.3390/su10040954

Liang, J., Li, C., & Zheng, L. (2016). Machine learning application in MOOCs: Dropout prediction. *ICCSE 2016 - 11th International Conference on Computer Science and Education*. https://doi.org/10.1109/ICCSE.2016.7581554

Marhayati, M., Fa'ani, A. M., Ruhmanasari, S. U., & Faridah, S. (2023). Application of K-Means Cluster Analysis for Grouping State Islamic University in Indonesia based on the Readiness Indicators for World Class University (WCU). *CAUCHY: Jurnal Matematika Murni Dan Aplikasi*, *8*(2). https://doi.org/10.18860/ca.v8i2.18046

Markos, V., Kocsis, Z., & Dusa, Á. R. (2019). Different Forms of Civil Activity and Employment in Hungary and Abroad, and the Development of Student Drop-out. *Central European Journal of Educational Research*, *1*(1). https://doi.org/10.37441/cejer/2019/1/1/3342

Mduma, N., Kalegele, K., & Machuve, D. (2019). Machine learning approach for reducing students dropout rates. *International Journal of Advanced Computer Research*, *9*(42). https://doi.org/10.19101/ijacr.2018.839045

Muhamad, M., Windarto, A. P., & Suhada, S. (2019). PENERAPAN ALGORITMA C4.5 PADA KLASIFIKASI POTENSI SISWA DROP OUT. *KOMIK (Konferensi Nasional Teknologi Informasi Dan Komputer)*, *3*(1). https://doi.org/10.30865/komik.v3i1.1688

Mustakim, M., Aini, D. N., Batubara, A. U., Erkamim, Moh., & Legito, L. (2024). Fuzzy Clustering-Based Grouping for Mapping the Distribution of Student Success Data. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *3*(2), 366–372. https://doi.org/10.57152/malcom.v3i2.1227

Santosa, R. G., Lukito, Y., & Chrismanto, A. R. (2021). Classification and Prediction of Students' GPA Using K-Means Clustering Algorithm to Assist Student Admission Process. *Journal of Information Systems Engineering and Business Intelligence*, *7*(1). https://doi.org/10.20473/jisebi.7.1.1-10

Tanjung, R. R., & Ramadhani, A. (2023). PENERAPAN POPULASI DAN SAMPEL UNTUK MENGETAHUI HUBUNGAN SOSIAL ORANGTUA DENGAN ANAK PUTUS SEKOLAH PADA JENJANG SMA. *Jurnal Bakti Sosial*, *2*(2), 144–154.

Wuni, S., Siregar, A. M., & Kusumaningrum, D. S. (2020). K-Means Clustering untuk mengelompokan Tingkat Putus Sekolah Jenjang SMP di Indonesia. *Scientific Student Journal for Information, Technology and Science*, *1*(1), 47–54.

Yi, H., Zhang, L., Yao, Y., Wang, A., Ma, Y., Shi, Y., Chu, J., Loyalka, P., & Rozelle, S. (2015). Exploring the dropout rates and causes of dropout in upper-secondary technical and vocational education and training (TVET) schools in China. *International Journal of Educational Development*, *42*. https://doi.org/10.1016/j.ijedudev.2015.04.009