PREDICTING STUDENT DROPOUT SYSTEM FOR BASIC EDUCATION HIGH SCHOOL BY K-MEANS CLUSTERING

Ei Ei Thwe[1], Khin Sandar Myint[2], Soe Mya Mya Aye[3]

Abstract

The prediction of Basic Education High School students' dropout has been an important field for educational institutions. Recently, Educational Data Mining (EDM) has gained attention among educational researchers and information technology researchers. Developing a strategic plan helps every student to attend school with positive outcomes. Data mining technique called cluster can predict why students drop out. The proposed system of this study is to analyse the performance of data mining techniques and to predict students' dropout using the K-Means clustering algorithm. A pre-processing step for the student; parent; teacher survey data led to a higher level of accuracy due to data cleaning and data reduction using Principal Component Analysis. The proposed system processed the survey data of the Basic Education High School students in rural Pathein, Ayeyarwady Region.

Keywords: Educational data mining, K-Means clustering algorithm, Principal Component Analysis

Introduction

Education is a very important part in Human life. Education makes theory, practice, even moral. The prediction of students' dropout result has been an important field for Basic Education Institutions. Because it provides planning to support and develop any strategic programs that can improve student school attendance. In order to predict school dropouts, student achievement and failure, it is necessary to survey the factors influencing basic education. Educational Data Mining (EDM) is a developing research field, and many researchers are interested in it. The information collected by them must be collected and kept annually. The main reason for modelling this field is a large amount of spread information and it has on different platforms. The vast amount of data available from scatter information is a continuous challenge to seize value.

Predicting student outcomes after graduation is an important area for higher education institutions as they plan to expand and expand any strategic programs that may improve student academic performance. It can also affect the reputation of institutions regarding the quality of graduates.

Most completed studies use data mining or multi-attribute decision-making techniques to predict student achievement. Some techniques are C4.5, Decision Tree, Naïve Bayes, MADM and Support Vector Machine (E. Budiman et al, 2018). In their previous research, Student learning outcomes were predicted using C4.5 and Naive Bayesian methods. This study compares the performance of two methods in classifying students' graduation time into 3 classes.

K-means Clustering Algorithm

K-means algorithm is one of most important data clustering algorithms. Clustering can be understood as grouping. In this paper, K-means algorithm is used to cluster the courses into different groups according to the required learning skills. K-means algorithm was first raised by James MacQueen and Hugo Steinhaus. Generally, it can be separated into three main parts (A.Elizabeth et al, 2018).

(1) Centroids Initialization: To initialize the centroids by randomly choosing k observations from the dataset.

(2) Assignment Step: Allocate each observation to the nearest cluster (the distance between its mean and the observation is shortest compared with other clusters.

(3) Update Step: Recalculate the meaning to be the centroid of each new cluster. The pseudo code of the K-means algorithm as follows:

   (i) Randomly select two data points from students' survey dataset to serve as the initial centroids. These could be any two points in students' survey dataset.

   (ii) Calculate the distances between each observation and each centroid.

[1] Department of Computer Studies, University of Yangon
[2] Department of Computer Studies, University of Yangon
[3] Department of Computer Studies, University of Yangon

(iii) Allocate each observation to the nearest centroid.

(iv) Recalculate the meaning to be the centroid of each new cluster.

(v) Repeat 2 to 4 until convergence happens.

(vi) Repeat step 1 to step 5 with k from 2 to 6 and pick the best one according to the sum of distance between each observation and centroids.

K-Means is a non-hierarchical clustering method that attempts to partition data into defined clusters. Therefore, Data with similar characteristics are collected into the same cluster and others into another cluster (A. Sarker et al, 2018). The K-Means method is a prominent cluster analysis algorithm in data mining.

In the K-means method, the result obtained through some clustering test is the original cluster centre point (A. Sarker et al, 2018). Different from (C. Casuat, 2019), the advantages of K-means algorithm are faster convergence to distortion minimum and apprehending how many clusters in the dataset (M. Li et al, 2018).

Implementation of Predicting Student Dropout System

In the proposed system, methodology for data collection and techniques is firstly defined. Student Survey Dataset in the data pre-processing step includes the data cleaning and reducing the attributes using Principal Component Analysis (PCA) method. The Cluster Model Using K-Means step includes PCA to k-means algorithm and to determine the best number of clusters.

Methodology for Data Collection and Techniques

The two data mining techniques, K-means algorithm, and Principal Component Analysis (PCA) are described in this section. Figure 1 shows the steps of the proposed system to predict students' dropout.

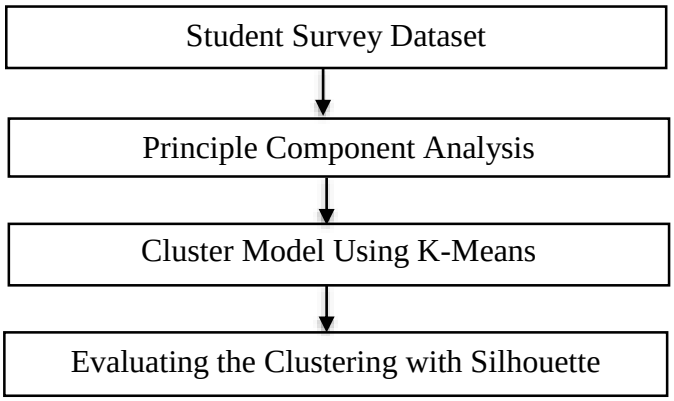

Figure 1 Proposed System Steps to Predict Students' Dropout

The 2015-2020 university entrance (10th grade) students list was gathered from the Office of the Director of Education, Pathein city, Ayeyarwady Region. The dataset comprises information from 4424 students across Basic Education High School in Ayeyarwady Region. Each student is represented by 20 features, including Name, Class, Age, Mother's Education, Father's Education, Mother's monthly average income, Father's monthly average income, Amount of Debt, Gender etc. Any deviations from the received data will be collected in the field in the townships. A difference between the percentage of students and the percentage of dropout students will be collected by township in the form of a questionnaire on mobile application. This difference will be analysed using MACHINE LEARNING MODEL and DROUPOUT rate.

Questionnaire format is mainly to collect information related to students, teachers, parents, and administrators. The cities that will collect data include Pathein city in Ayeyarwady Region. Knowledge Discovery in Databases (KDD) can help educational organizations by

turning their data into useful information. Organizations that take advantage of KDD techniques will find that they can improve education quality by using fast and better educational decision making. The data collection of the proposed system is based on cloud database and mobile android application in Figure 2.

Figure 2 Android Application of Data Collection

The android version survey apk installed on at least android 7 versions. Mobile phone applications use internet connection for data collection. Data collection mobile apk is installed on smartphones that provide a student dropout survey system. The student dropout survey app uses google cloud platform as shown in Figure 3.

Figure 3 Mobile Application

The attributes of student survey data for Predicting Student Dropout System is shown in Table 1.

Table 1 Attributes of Student Survey Data

No	Feature	Value range
၁။	အဓိပညာအရည်အချင်း	Degree
၂။	အဘပညာအရည်အချင်း	Degree
၃။	အဓိတစ်လဝင်ငွေ	Money amount
၄။	အဘဝင်ငွေ	Money amount
၅။	အကြွေးရှိပမာဏ	Money amount
၆။	ကျောင်းသားအသက်(နှစ်)	Age
၇။	ကျား/မ	Male/Female
၈။	ဆရာစာသင်ကြားမှုအခြေအနေ	1-5

၉။	မိသားစုမှာမိုခိုဘယ်နှစ်ယောက်ရှိလဲ	Count
၁၀။	စာလိုက်နိုင်မှုအခြေအနေ	1-5
၁၁။	မိသားစုအရေအတွက်	Count
၁၂။	သင်ယူတာနားလည်မှု	1-5
၁၃။	စာသင်ကြားရတာစိတ်ပါဝင်စားမှုရှိလား	1-5
၁၄။	မြို့/နယ်	1-2
၁၅။	ကျောင်းနှင့်အိမ်အကွာအဝေး	1-5
၁၆။	ကျန်းမာရေးအခြေအနေ	1-5
၁၇။	ကျောင်းသွားတက်ရသည့်အခြေအနေ	1-5
၁၈။	ကျောင်းသားGradeအဆင့်	A-E
၁၉။	ရှိ/ထွက်	0-1

Data Pre-processing

The data clean method is useful for cleaning datasets that have a missing value. It removes missing value attributes from the dataset. PCA is used to attribute reduction. Principal Component Analysis (PCA) is a feature selection method which is used to reduce missing value attributes. It is used to eliminate irrelevant attributes in predicting student's survey data. The selecting relevant and non-correlated attributes does not affect the information in the initial data set, then the prediction is developed using the K-means method to cluster the data set (C. Casuat, 2019), (M. Z. Nasution, 2018). It used to eliminate irrelevant attributes in predicting students' dropout data.

A correlation heatmap was designed to analyse the correlations between input features and output variables. The code in Python to create the heatmap of the correlation matrix is described in Figure 4.

```
principalDataframe = pd.DataFrame(data = principalComponents, columns = ["PC1", "PC2"])
targetDataframe = df[['Target']]

newDataframe = pd.concat([principalDataframe, targetDataframe],axis = 1)
print(newDataframe)

plt.scatter(principalDataframe.PC1, principalDataframe.PC2)
plt.title('PC1 against PC2')
plt.xlabel('PC1')
plt.ylabel('PC2')
plt.show()


fig = plt.figure(figsize=(8, 8))
ax = fig.add_subplot(1, 1,1)
ax.set_xlabel('PC1')
for i in range(1, 23)
```

Figure 4 Heatmap of the Correlation Matrix

Using the PCA method, selected relevant data and non-correlated attributes are without affecting the original information of the data set (C. Casuat, 2019), (M. Z. Nasution, 2018) as shown in Figure 5. There exists a medium correlation between important features and grade

points. Features such as Mother's qualification, Father's qualification, Mother's monthly average income, Father's monthly average income and health status have a strong correlation with grade points. As a result, strongly correlated features are considered in the model building due to their highest impact on student dropout outcomes. In the last stage of data pre-processing, the dataset was normalized using a standard scalar to eliminate the mean and scale it to unit variance.

Figure 5 Value of PCA Loading Factor for Each Variable.

Modelling for Proposed System

The K-means clustering method is used to predict student dropout in this proposed system. The algorithm of the K-means clustering method is shown in Figure 6.

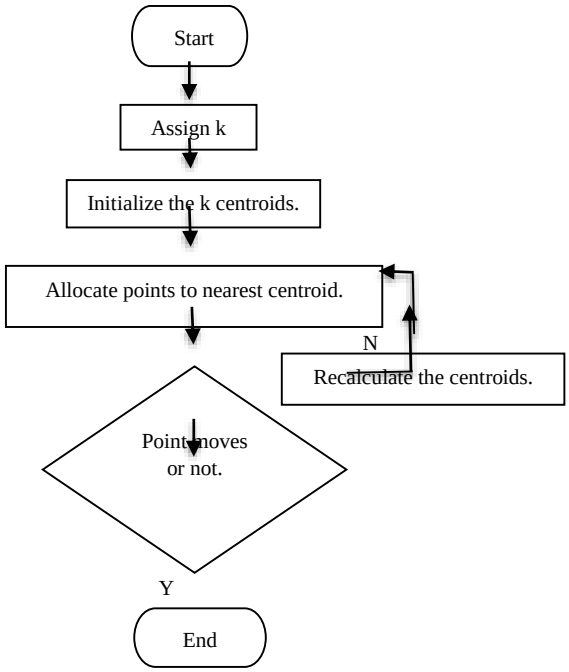

Figure 6 The Logic Flow of K-means Algorithm with a Specified K

The classification model describes these two: Naïve Bayes classifier has an accuracy rate of 60% and the Tree C4.5 method has an accuracy rate of 58.82% (M. E. Hiswati et al, 2018) over other methods. Data clustering aims to group similar data records together. Clustering is often confused as classification, but they both have different goals. Simply put, the inter-cluster distance needs to be reduced to achieve better clustering results, as in the K-means algorithm.

Evaluation

The evaluation is conducted using the Silhouette method and Principal Component Analysis (PCA) to calculate the prediction accuracy for the student survey dataset. Silhouette analysis can be used to study the separation distance between the resulting clusters. Silhouette Score is a metric to evaluate the performance of a clustering algorithm (P. Pannen et al, 2019) (Y. Pang et al, 2017). It uses compactness of individual clusters (***intra cluster distance***) and separation amongst clusters (***inter cluster distance***) to measure an overall representative score of how well our clustering algorithm has performed. The silhouette score is a valuable metric for objectively evaluating the effectiveness of K-means clustering, helping to determine the optimal number of clusters, and ensuring that the resulting clusters are meaningful and well-separated. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to assess parameters like number of clusters visually.

The proposed system approach has four important stages namely, student survey dataset processing, Principal Component Analysis (PCA), K-Means clustering and evaluation. All steps are important for predicting a student's dropout accuracy and performance.

Result and Discussion

The data pre-processing step carried out data cleaning and reduction data with the PCA method, so we obtained the relevant attributes to clustering the dataset using the K-means algorithm. The value of each PCA loading factor variable is presented in Figure 7.

Figure 7 Exact Match Ratio (EMR) in 6 models of cluster

Silhouette Plot of K-Means Cluster for k = 2, 3, 4, 5, and 6 are shown in Figure 8, 9, 10, 11, 12 respectively. Figure 13 shows Silhoutte Plot of K-Means Cluster for k=2 to k=6. For these following figures, YellowBrick-a machine learning visualization library in Python is used. Silhoutte score for clusters 2, 3, 4, 5, and 6 is shown in Figure 14 using line graph.

```
import matplotlib.pyplot as plt

def visualize_clusters(X, clusters):
    fig, axs = plt.subplots(1, len(clusters), figsize=(len(clusters) * 6, 6))
    for i, cluster in enumerate(clusters):
        km = KMeans(n_clusters=cluster, init='k-means++', n_init=10, max_iter=300, random_state=0)
        visualizer = SilhouetteVisualizer(km, colors='yellowbrick', ax=axs[i])
        visualizer.fit(X)
        visualizer.finalize()
        axs[i].set_title(f"{cluster} Clusters")

    plt.tight_layout()
    plt.show()

# Load your dataset and define the features
dataset = pd.read_csv("dataeieithwe11.csv")
features = ["Id","Mother's qualification", "Father's qualification", "Mother's occupation", "Father's occupation", 'Tuition fees up to
date', 'Gender', 'Age at enrollment', 'Unemployment rate', 'Inflation rate']
X = dataset.loc[:, features].values

# PCA transformation
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

# Define the cluster sizes you want to visualize
cluster_sizes = [2, 3, 4, 5, 6]
```

Ln 28, Col 1 1,054 characters 100% Windows (CRLF) UTF-8

K-means Clustering Coding

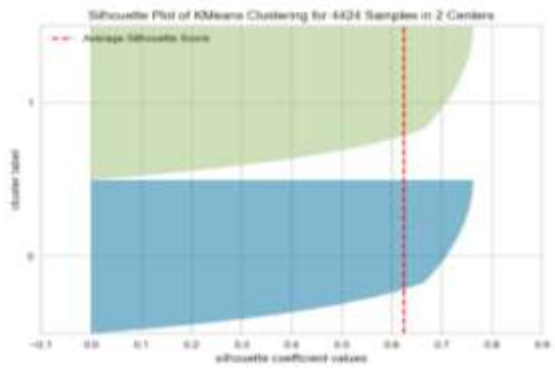

Figure 8 Silhouette Plot of K-Means Cluster for k = 2

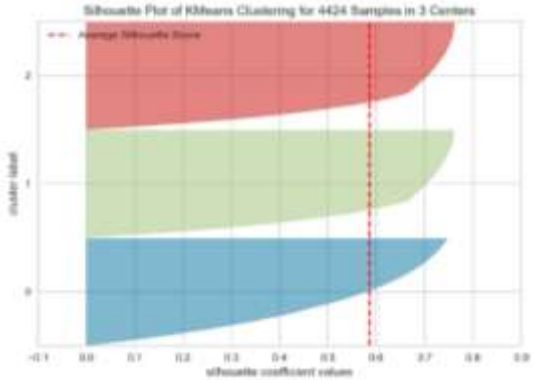

Figure 9. Silhouette Plot of KMeans Cluster for k = 3

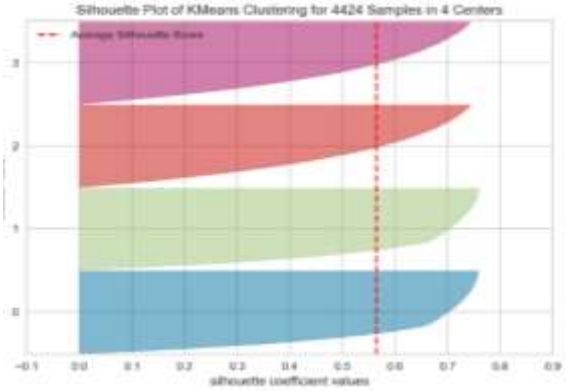

Figure 10. Silhouette Plot of K-Means Cluster for k=4

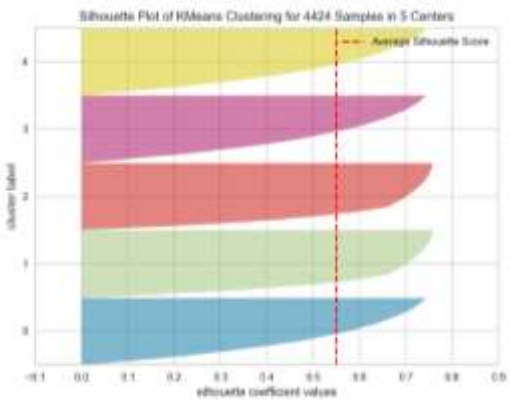

Figure 11. Silhouette Plot of K-Means Cluster for k = 5

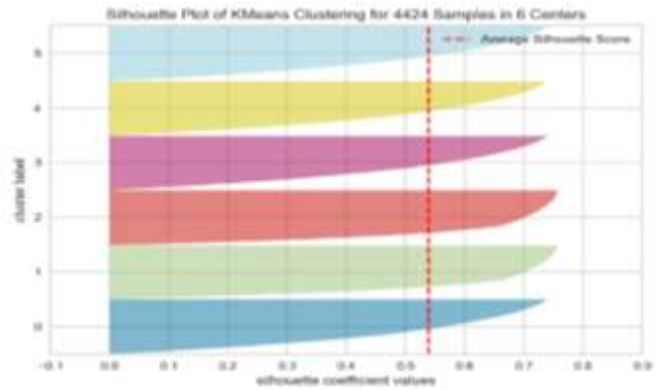

Figure 12. Silhouette Plot of K-Means Cluster for k = 6

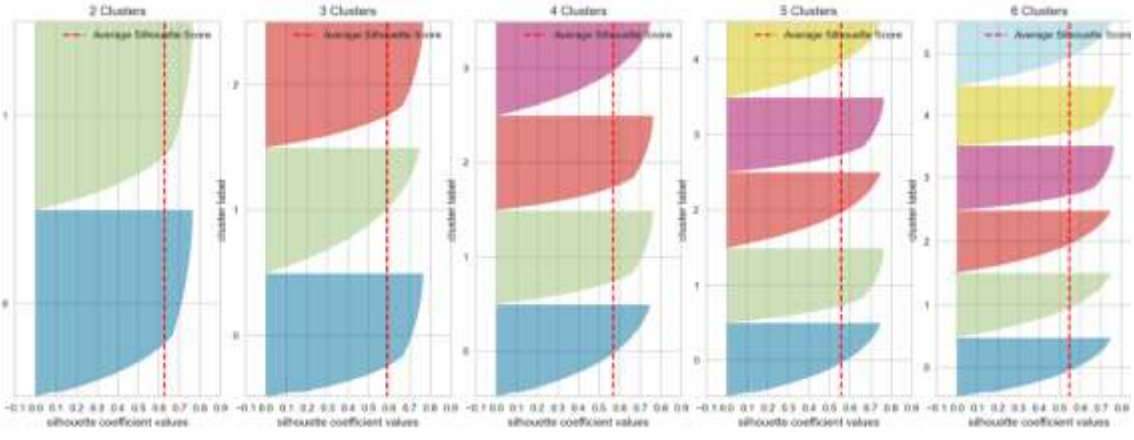

Figure 13. Silhoutte Plot of K-Means Cluster for k=2 to k=6

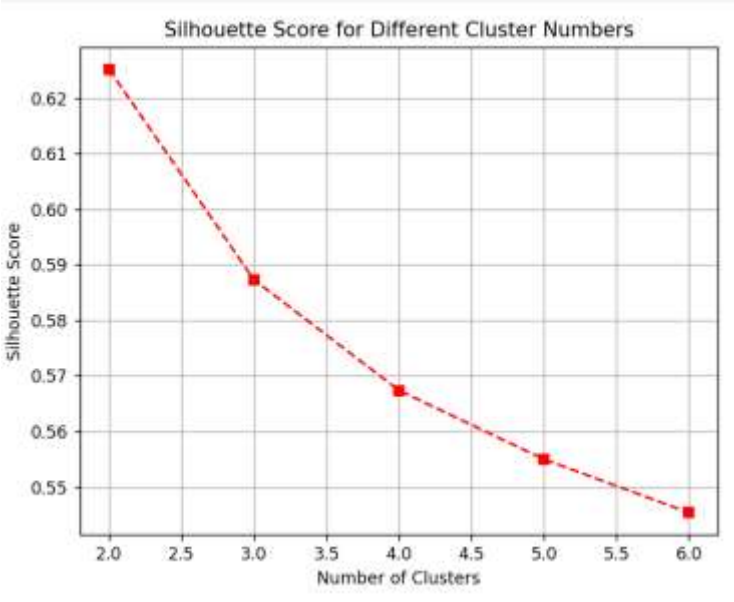

Figure 14. Silhoutte Score of K-Means Cluster for k=2 to k=6

The silhouette score for k = 2 is the highest in all clusters as shown in Figure 7. It is a proposed system except for this data set as shown.

The 2-cluster model obtained an accuracy rate of 62.4%. In the 3-cluster model, the accuracy rate is 58.5%. In the 4-cluster model, the accuracy rate is 56.4%. In the 5-cluster model, the accuracy rate is 55%. In the 6-cluster model, the accuracy rate is 53.9%.

To predict student dropout prediction the highest accuracy rate of K-Means clustering modelling with 2-cluster models use the student academic survey dataset in Table 1.

This shows that the most appropriate clustering of data to predict student dropout uses a 2-cluster model. The graph can illustrate the results of the study in Figure 7.

Conclusion

As the results of this proposed system obtained, the attributes are ages, school home travel time, teacher teaching ability, guardian and number of family members all have negative loading factor value in the PCA method, which means it all has a small correlation value to the prediction of the student dropout. This paper presented a comprehensive approach to predicting student dropout in Basic Education High Schools, utilizing K-means clustering and Principal Component Analysis (PCA). The integration of these techniques provides a refined and accurate prediction system, enabling educators to implement timely interventions.

Based on 2 cluster models it is obtained that the 2-cluster model is the best clustering with an accuracy rate obtained is 62.4%. As future work, it may explore the incorporation of additional data sources and advanced machine learning techniques to further improve the accuracy and applicability of the dropout prediction system.

Acknowledgement

The authors would like to thank professor Dr. Soe Mya Mya Aye, Head of Department of Computer Studies, University of Yangon for giving a chance to do this research and for encouraging completion in time. Also, I would like to thank my respectful Daw Khin Cho Myint, Principal of Bogalay Education Degree College for her co-operation and encouragement in this research.

References

A. Sarker et al, (2018), "Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm," Type Double Blind Peer Rev. Int. Res. J. Softw. Data Eng. Glob. J. Comput. Sci. Technol. C, vol. 18, no. 1.

A. Elizabeth et al, (2018), "Groundwater quality assessment: An improved approach to K-means clustering, principal component analysis and spatial analysis: A case study," Water (Switzerland), vol. 10, no. 4, pp. 1–21, doi: 10.3390/w10040437.

C. Casuat et al, (2019), "Predicting Students' Employability using Machine Learning Approach," ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci., doi: 10.1109/ICETAS48360. 2019. 9117338.

D. Konar et al, (2020), "Predicting Students' Grades Using CART, ID3, and Multiclass SVM Optimized by the Genetic Algorithm (GA): A Case Study," Recent Adv. Hybrid Metaheuristics Data Clust., pp. 85–99, doi: 10.1002/9781119551621.ch5.

E. Budiman et al, (2018), "Performance of Decision Tree C4.5 Algorithm in Student Academic Evaluation," Lect. Notes Electr. Eng., vol. 488, pp. 380–389, doi: 10.1007/978-981-10- 8276-4.

M. E. Hiswati et al, (2018), "Minimum wage prediction based on K-Mean clustering using neural based optimized Minkowski Distance Weighting," Int. J. Eng. Technol., vol. 7, no. 2, pp. 90–93, doi: 10.14419/ijet.v7i2.2.12741.

M. Li et al, (2018), "Application of CART decision tree combined with PCA algorithm in intrusion detection," Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS, vol. 2017-Novem, pp. 38–41, doi: 10.1109/ICSESS.2017. 8342859.

M. Z. Nasution et al, (2018), "PCA based feature reduction to improve the accuracy of decision tree c4.5 classification," J. Phys. Conf. Ser., vol. 978, no. 1, doi: 10.1088/1742-6596/978/1/012058.

P. Pannen et al, (2019), "Autonomous 5 higher education institutions in Indonesia," Gov. Manag. Univ. Asia Glob. Influ. local responses, p. 56.

Y. Pang et al, (2017), "Predicting students' graduation outcomes through support vector machines," Proc. - Front. Educ. Conf. FIE, vol. 2017-Octob, pp. 1–8, doi: 10.1109/FIE.2017.8190666.