Evaluation model of students learning outcome using k-means algorithm

Henderi*, Abas Sunarya, Zakaria, Sri Linggawati Nurmika, and Nur Asmainah

Master of Information Technology Study Program, Faculty of Science and Technology in Raharja University, Tangerang, Indonesia

*henderi@raharja.info

Abstract. The progress of students learning outcomes have to evaluated and make monitor. Student learning outcomes were indicated by score of the exam. Educational institutions can determine the ability of students' knowledge and competences based on student exam scores. For this reason, the goals of this research is develops an evaluation model of student learning outcomes. The model was developed using k-means algorithm and it has functions to analyze the student learning outcomes. The model was tested by conducted using 50 data exam of students, rapid manner software, and has implemented by the Python 3.0 programming language. The testing the evaluation model of student learning outcomes that were developed resulted: there are 3 clusters of student learning outcomes that are good, satisfying, and lacking. The tested results also showed that 40% of students had good learning outcomes, 44% of students were satisfactory, and 16% of other students had poor learning outcomes. The results of testing the student learning evaluation model can be used by the parties concerned to make decisions related to work programs or formulate steps that need to be taken to improve student learning outcomes.

1. Introduction

In Indonesia, the schools make classified students in category of graduation with differences for whom has good score. The most of schools setting a minimum set of grades that have to maintained to proceed to the next level. In some schools, the minimum requirement for students is 40%, while in others cases 33%. The value obtained in the exam is a general factor used for the assessment of academic development.

The traditional approach of grouping students based on average scores is difficult to find a comprehensive view of the states student's assessment. The grouping approach is difficult to find assessments in detail. A potential approach to getting valuable information about student's performance is data mining. The techniques of data grouping are used to find information, patterns and relationships of huge data helped in decision making. The data grouping methods, such as grouping, decision trees or association analysis are approach to find the main characteristics of student's performance.

This approach allows the use of characteristics for predictions in the future [7]. This paper discusses the application of the k-means algorithm to monitor the progress of student assessments in schools similar to those conducted by Abdurrahman [6]. However, in this study, student learning outcomes data sets are grouped using the k-means algorithm with the help of the RapidMiner Studio 9.2.1 and Python 3.0 application programs.

2. Method Overview

K-Means clustering produces some clusters. The K-Means method is numerical, unsupervised, not deterministic, and repetitive. This procedure follows a simple and easy way to classify data sets given through a number of clusters. This research was conducted through three stages, namely: preprocessing data, fixed data, and grouping data using the k-means algorithm. These stages are illustrated in Figure 1.

Figure 1. The stages of research

2.1. Selection of dataset

The data used in this study is the assessment of student learning outcomes. The data used is the value of the learning outcomes of 50 students as the initial dataset.

2.2. Data Preprocessing

The first, before conducting a cluster on student grade data, preprocessing data is first performed. The data preprocessing process is carried out through the stages of data integration, adding attributes, selecting data, converting data, replacing missing values, and z-score standardization.

2.3. K-means Clustering Algorithm

The grouping of data using K-means conducted after the data has processed through 6 phase od preprocessing. The statges of grouping data using the K-means algorithm are [8], [9]:

1. Select the number of clusters k
2. Initialize the cluster center randomly. The center of the cluster is given a random initial value.
3. Allocate all object data in the closest cluster. The similarity of the two data is determined by the distance between the two data. While the similarity of a data in a cluster is determined by the distance between the data and the center of the cluster. At this stage the calculation of the distance of each data to the center of the cluster.
4. Recalculate the cluster center with the current cluster membership. Cluster center is the average value of all objects / data in a cluster. However, if possible, you can also use a median cluster. As such, averages are not the only measure that can be used.
5. Repeat the previous cluster processing steps so that nothing changes.

The calculate of centroid cluster *i, vi,* using the Formula 1 [9]:

$$v_{ij} = \frac{\sum_{k=1}^{N_i} x_{kj}}{N_i} \quad (1)$$

In the Formula 1, Ni is the amount of data that is a member of the i-cluster. The grouping using k-means in this study was conducted by grouping data for three k clusters. The process of grouping have produced 3 clusters which will be reprocessed into 3 cluster to determine the students who need different guidance.

Analysis using k-means clustering is being carried out with the help of RapidMiner Studio. RapidMiner Studio is a free data mining software for academic and research purposes. The use of RapidMiner in this study is similar to that carried out by Mardilius [1].

Data analysis methods used are exploration, statistical learning, machine learning and database areas. RapidMiner Studio is a data science software platform developed by a company with the same name that provides an integrated environment for data preparation, machine learning, deep learning, text mining and predictive analysis. It is used for business and commercial use, also for research, education,

training, rapid prototyping, and application development and supports all steps in the machine learning process including data preparation, visualization results, model validation, and optimization.

3. Result and analysis

This model is applied to vocational students in ABC schools in Indonesia. The analysis was conducted based on the grades obtained by class X class 2018-2019 students from semester I to semester II. The number of students involved in the analysis is 50. The results generated using the RapidMiner Studio software for semester 1 to semester II for k = 3 (cluster) are shown in the overall results of the assessment by applying a deterministic model in the equation. Where the group assessment in each cluster size assessment by adding up the average score of individuals in each clusters.

K-Means algorithm results with 50 student data through iteration up to 12 times produce 3 clusters. The cluster one consists of 20 students with good grades, cluster two has 22 students with satisfactory results and carry out enrichment activities, and cluster three consists of 8 students with less remedial results. The enrichment activities are carried out with the aim of providing opportunities for students to deepen the mastery of subject matter related to the learning task being carried out in order to achieve an optimal level of development.

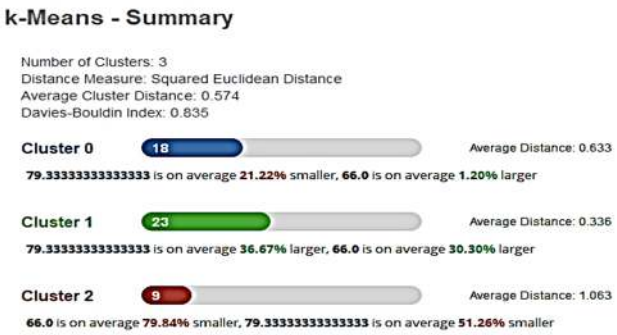

Figure 2. Results of clustering to 50 student's grade

The figure 2, showing there were 18 students in cluster 0 (good), there were 23 students in cluster 1 (satisfying), and there were 9 students in cluster 2 (less). Cluster 1 members are recommended to participate in enrichment activities, and cluster 2 members are recommended to attend additional and remedial tutoring activities. The results of this study are similar to Mungingsih, et al. [2], and Arora, et al. [3] which used the same algorithm. These results are also in line to the findings of Raval, et al. [4].

Based on the test of results, it was known that in the 11th iteration, the cluster position has not changed. Furthermore iterations are carried out only on unchanged iteration processes [10], namely on the 11th cluster. The results of clustering get the following information:

1. The first cluster can be interpreted as students who get good grades with a total of 20 students
2. The second cluster can be interpreted as a group of students with a satisfactory value of 22 people.
3. The third cluster can be interpreted as a group of students with a value of less than 8 people.

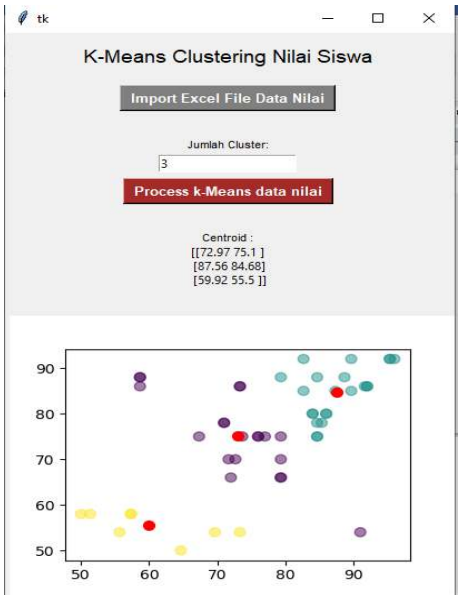

Figure 3. The results of implementation with Python

In the figure 3, the centroid value using Python 3.0 application and using the k-means algorithm produces the same results as inputting 3 clusters. This information a role in determining the reasons for the decline in student learning outcomes.

Coding using Phyton 3.0.

```
import tkinter as tk
from tkinter import filedialog
import pandas as pd
from pandas import DataFrame
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from matplotlib.backends.backend_tkagg import FigureCanvasTkAgg
from importlib import reload

#test
import numpy as np #untuk pembulatan
#from pprint import pprint
#
root= tk.Tk()
canvas1 = tk.Canvas(root, width = 400, height = 300,  relief = 'raised')
canvas1.pack()

label1 = tk.Label(root, text='K-Means Clustering Nilai Siswa')
label1.config(font=('helvetica', 14))
canvas1.create_window(200, 25, window=label1)

label2 = tk.Label(root, text='Jumlah Cluster:')
label2.config(font=('helvetica', 8))
canvas1.create_window(200, 120, window=label2)

entry1 = tk.Entry (root)
canvas1.create_window(200, 140, window=entry1)

def getExcel ():

  global df
  import_file_path = filedialog.askopenfilename()
  read_file = pd.read_excel (import_file_path)
  df = DataFrame(read_file,columns=['x','y'])
```

```
browseButtonExcel = tk.Button (text=" Import Excel File Data Nilai ", command=getExcel, bg='grey', fg='white', font=('helvetica', 10,
'bold'))
canvas1.create_window(200, 70, window=browseButtonExcel)

def getKMeans ():
   global df
   global numberOfClusters
   numberOfClusters =int(entry1.get())

kmeans = KMeans(n_clusters=numberOfClusters).fit(df)

    centroids = kmeans.cluster_centers_

   #test value
   #pprint(kmeans)
   #pprint(cluster_centers_)
   #a.round(decimals=2) ini contekannya
   #pprint(centroids.round(decimals=2)) //send decimals in 2 digit after comma

   label4 = tk.Label(root, text='Centroid :')
   label4.config(font=('helvetica', 8))
   canvas1.create_window(200, 220, window=label4)

   label3 = tk.Label(root, text= centroids.round(decimals=2))
   canvas1.create_window(200, 250, window=label3)

   figure1 = plt.Figure(figsize=(4,3), dpi=100)
   ax1 = figure1.add_subplot(111)
   ax1.scatter(df['x'], df['y'], c= kmeans.labels_.astype(float), s=50, alpha=0.5)
   ax1.scatter(centroids[:, 0], centroids[:, 1], c='red', s=50)
   scatter1 = FigureCanvasTkAgg(figure1, root)
   scatter1.get_tk_widget().pack(side=tk.RIGHT, fill=tk.BOTH)

processButton = tk.Button(text=' Process k-Means data nilai ', command=getKMeans, bg='brown', fg='white', font=('helvetica', 10,
'bold'))
canvas1.create_window(200, 170, window=processButton)

root.mainloop()
```

4. Conclusion

The k-means algorithm is proven to be used to evaluate student learning outcomes. The results of applying the K-means algorithm can help planners and academic leaders monitor student learning outcomes. The evaluation model developed in this study was instrumental in determining the reasons for the decline in student learning outcomes and the steps that need to be taken to improve it.

References

[1]. Mardalius 2018 Pemanfaatan Rapid Miner Studio 8.2 Untuk Pengelompokan Data Penjualan Aksesoris Menggunakan Algoritma K-Means, *Jurnal Sumatera Utara*.

[2]. Elly Muningsih, Sri Kiswati 2015. *Penerapan Metode K-Means untuk Clustering Produk Online Shop*, Yogyakarta.

[3]. Rakesh Kumar Arora, Dharmendra Badal 2013. *Evaluating Student's Performance Using k-Means Clustering*, India.

[4]. Unnati R. Raval, Chaita Jani 2016 *Implementing & Improvisation of K-means Clustering Algorithm*, India

[5]. Triyanna Widiyaningtyas, Martin Indra Wisnu Prabowo, and M. Ardhika Mulya Pratama 1986 *Implementation of K-Means Clustering Method to Distribution of High School Teachers*: Yogyakarta.

[6]. Ginanjar Abdurrahman 2007 *Clustering Data Ujian Tengah Semester Data Mining Menggunakan Algoritma K-Means*, Jember.

[7]. Ebel, R.L.,& Frisbie, D.A. 1991 *Essentials of educational measurement* (5th ed.), Engelwood Cliffs, New Jersey: Prentice Hall.
[8]. Maimon, O., and Rokach, L. 2010 *Data Mining and Knowledge Discovery Handbook*. London: Springer Science+Business Media.
[9]. Daniel T. Larose 2005 *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, Inc.
[10]. Oyelade, O. J., Oladipupo and O.O, Obagbuwa, I.C 2010 *Application of k-Means Clustering algorithm for prediction of Students' Academic Performance*, Nigeria