Competency test clustering through the application of Principal Component Analysis (PCA) and the K-Means algorithm

Raditya Danar Dana, Dodi Soilihudin, Ryan Hamonangan Silalahi, D.A Kurnia, Umi Hayati
STMiK IKMI Cirebon, Cirebon, Indonesia

*radith_danar@yahoo.com

Abstract. The implementation of the Competency Test at the LSP institution in higher education is an effort to ensure that students have abilities according to predetermined competency standards. Education providers are required to always strive to improve the quality and quality of education with the aim that the student's academic performance will always improve. From the results of observations made in the research location, it was found a problem with the high number of failures in the implementation of the competency test. This study aims to conduct cluster analysis of the data from the implementation of competency tests using Machine Learning techniques through the application of Principal Component Analysis (PCA) and K-Means Algorithm, through several stages in the form of data collection, data cleaning, data transformation, data modeling and experimentation. This study resulted in grouping the results of competency tests which were divided into 3 clusters, namely cluster 1 as much as 38%, cluster 2 as much as 32% and cluster 3 as much as 30%.

1. Introduction
A Professional Certification Institute (LSP) at a university was established as an effort to ensure that students have the ability according to predetermined competency standards. The results of a competency test activity are one of the indicators that can be used as benchmarks for the success of the learning process. So that the urgency of the importance of efforts to analyze student academic performance obtained through the implementation of competency tests. This is very important to make it easier for the management to determine a policy such as whether or not an enrichment activity is needed for students whose competencies have not met standards and also to determine in what semester a competency needs to be taught to students [1][2][3][4].

Machine Learning technology has proven its use in solving various problems in various fields of life, one of which is the education domain. Through machine learning technology, data patterns can be identified and processed into very valuable information as a basis for determining effective efforts to improve the quality of education[5][6][7][8].

In this study, the data clustering process will be carried out based on the characteristics of the data obtained from the results of the implementation of the competency test at the Professional Certification Institute in Higher Education, namely STMIK IKMI Cirebon. Data clustering will be carried out based on several main attributes, namely "Job Assessment", "Competency Test Scheme", "Competency Test Results" and "Study Program", "Year of Force", based on this information a
clustering process will be carried out to obtain new information, in the form of a relationship between attributes so that it will be easier to analyze student competency results.

The need for cluster analysis efforts in the implementation of competency tests at Professional Certification Bodies is due to a problem in the form of implementing competency tests that have not yet enforced the minimum semester level requirements that are allowed to take competency tests. So that all semester levels can follow the implementation of the same competency test scheme. So it is feared that the process of implementing the competency test will not be optimal because the student concerned has not received the material or otherwise has received the material but it has been too long so that there is a potential for students to forget the material

Cluster analysis will be carried out using Machine Learning Technology through Clustering techniques with the application of Principal Component Analysis (PCA) on the K-Means Algorithm. The K-Means method is widely used in clustering techniques, K-Means has advantages in computation speed and relatively easy process steps [9][10][11]. However, the K-Means algorithm has several weaknesses in the form of its inability to handle data that has storage (noisy data and outliers). Besides, it has a dependence on the choice of the number of clusters and centroid initialization which is done randomly. The application of PCA can overcome the weaknesses in the K-Means algorithm because the centroid selection process becomes more optimal and streamlines processing time by reducing data attributes. Dimensional reduction can eliminate irrelevant features and reduce noise. PCA can reduce high data dimensions to lower data dimensions with very little risk of losing information [12][13]

2. Methodology
This research was conducted using an experimental research method involving several phases of development, evaluation and problem-solving projects divided into several stages as follows

![Research Methods](image)

2.1 Data Collection
The data were obtained from the implementation of competency tests at the Professional Certification Institute (LSP) which was attended by 472 test participants. Below is the data attribute data used
a) Test Participant Name, namely the full name of the assessor or competency test participant
b) Occupation, namely occupation where an assessor has the possibility of not undergoing activities only as a student, some of which are while working. This attribute has several possible values, namely as a student, employee and entrepreneur. This attribute is considered important because it could be that an assessor's activities and activities will have an impact on their academic performance
c) Competency Test Scheme, this attribute has 2 types of competency test schemes offered, namely "Database Design for the Web" and "Network Security Monitoring"
d) Competency Test Results, this attribute has 3 possible values, namely "Competent" and "Not yet Competent"
e) Study Program, this attribute has 3 possible values, namely "Informatics Engineering", "Information Management" and "Computerized Accounting"
f) Year, this attribute has 4 possible values namely 2015, 2016, 2017 and 2018

2.2 Data Preprocessing
Data preprocessing is carried out to ensure that all data has been filled or there is no null data through the data cleaning process. Pre-processing is also intended to convert data that is still categorical to be converted into numeric data because clustering computation can only process numerical data. The process of converting label data to numeric is done by creating a dummy table that generates new
attributes for a number of possible values that appear by converting them to a value of 0 if not selected and 1 if selected [14].

2.3 Data Modelling
The data modeling process begins with determining the optimal number of clusters using the elbow method

![Inertia Plot per k](image)

**Figure 2. Methods Optimal Number of Clusters**

Through the visualization process generated using the Elbow method, it is obtained an illustration that the optimal number of clusters is 3, because starting from 4 onwards does not provide significant changes.

2.4 Experiment
The following are the steps in carrying out clustering using the K-Means algorithm

a) Determine the number of clusters

b) Select the object that will serve as the centroid point of each cluster

\[ \overline{v}_{ij} = \frac{1}{N_i} \sum_{k=0}^{N_i} x_{kj} \]  

(1)

c) Group objects to the nearest centroid based on Euclidean Distance

\[ D_{ij} (x_2 - x_1) = \sqrt{\sum_{j=1}^{p} (x_{2j} - x_{1j})^2 } \]  

(2)

d) Recalculate all centroid point values

e) Repeat step 2 again until the centroid point value does not change

The process of calculating the K-Means method is done by finding the minimum distance between the object and each centroid point of each cluster [15]. The general PCA algorithm is as follows:

a) Calculate the covariance matrix using the equation

\[ \text{Cov}(xy) = \frac{\sum_{i=1}^{n} x_{i}y_{i}}{n} = (x)(y) \]  

(3)

b) Calculate the eigenvalues using the equation

\[ [A - \lambda I] = 0 \]  

(4)

c) Calculate the eigenvector using the equation

\[ [A - \lambda I][X] = 0 \]  

(5)

d) Determine the new variable (principal component) by multiplying the original variable by the eigenvector matrix

3. Result and Discussion
The following below is an interpretation of the characteristics of the cluster formed
The PCA calculation process and the K-Means algorithm produce the following clusters

| No | Cluster   | Number Of Item |
|----|-----------|----------------|
| 1  | Cluster 1 | 121            |
| 2  | Cluster 2 | 102            |
| 3  | Cluster 3 | 95             |
| 4  | Total Items | 318          |

a) 38% of the number of new student admissions applicants are grouped into "Cluster 1" with the average distance between the object point and cluster 1 is 3.5663
b) 32% of the new student admissions applicants are grouped into "Cluster 2" with the average distance between the object point and cluster 2 is 2.6724
c) 30% of the number of new admissions applicants are grouped into "Cluster 3" with the average distance between the object point and cluster 3 is 2.9312

![Figure 3. Cluster Visualization](image)

The following below is an interpretation of the characteristics of the cluster formed

a) Cluster 1, this cluster can be called a cluster of assessments that get Competent results. This cluster is dominated by assessments from the Informatics Engineering Study Program in 2017 which took part in the "Database Design for the Web" scheme test and assessments from the Informatics Engineering Study Program year 2018 which took the "Network Security Monitoring" Scheme.
b) Cluster 2, this cluster can be called a cluster of assessments originating from the Informatics Engineering Study Program year 2017 which participated in the test of the "Network Security Monitoring" scheme with almost even results between competent and not yet competent.
c) Cluster 3, this cluster is named as a cluster of assessments that get the results not yet competent, this cluster is dominated by those from the Computerized Accounting Study Program which takes the Database Design Competency Test Scheme for the Web, an assessment from the Informatics Management study program that participates in the Schema test. "Database Design for the Web" and assessments originating from the Informatics Engineering study program in batch 2018 which took the Database Design scheme for the Web "

Based on the characteristics of the clusters formed, a policy recommendation can be taken by the college management as follows:
a) We recommend that the Competency Test Scheme "Database Design for the Web" is only tested on the assessments originating from the Informatics Engineering Study Program at least in 2017 or 6th semester who have mastered web application development
b) Strengthening the abilities and competencies of students from the Computerized Accounting and Information Management Study Program regarding Web Application Development material through enrichment activities
c) Especially for assessments originating from the Informatics Engineering study program in 2017, competence should be strengthened in the field of Computer Network Security through enrichment activities because the processes that obtain Competent results are not more than 50%

4. Conclusion
From the results of experiments that have been carried out, it can be concluded that the application of Principal Component Analysis (PCA) on the K-Means algorithm can optimize the clustering process in determining the initial centroid point of the K-Means algorithm. Three optimal numbers of clusters were formed in grouping data on the results of the implementation of competency tests at professional certification bodies, namely "Cluster 1", "Cluster 2" and "Cluster 3". With the cluster analysis using Machine Learning technology in the implementation of competency tests at the Professional Certification Institute, it can make it easier for management to determine policies to improve student academic competence and performance.

References
[1] Pemerintah Republik Indonesia, “Undang-Undang Republik Indonesia No.13 Tahun 2003 tentang Ketenagakerjaan,” in Undang-Undang Ketenagakerjaan, no. 1, 2003, pp. 1–34.
[2] Pemerintah Republik Indonesia, Peraturan Pemerintah Nomor 10 tahun 2018 Tentang Badan Nasional Sertifikasi Profesi, no. 10. 2018.
[3] Badan Nasional Sertifikasi Republik Indonesia, Peraturan Badan Nasional Sertifikasi Profesi Nomor : 09/BNSP.301/XI/2013. 2013.
[4] A. S. Nagesh and C. V. S. Satyamurty, “Application of clustering algorithm for analysis of Student Academic Performance,” Int. J. Comput. Sci. Eng., vol. 6, no. 1, pp. 381–384, 2018, doi: 10.26438/ijcse/v6i1.381384.
[5] K. Sya’iyah, H. Yuliansyah, and I. Arfiani, “Clustering student data based on K-means algorithms,” Int. J. Sci. Technol. Res., vol. 8, no. 8, pp. 1014–1018, 2019.
[6] O. Somantri, S. Wiyono, and D. Dairoh, “Metode K-Means untuk Optimasi Klasifikasi Tema Tugas Akhir Mahasiswa Menggunakan Support Vector Machine (SVM),” Sci. J. Informatics, vol. 3, no. 1, pp. 34–45, 2016, doi: 10.15294/sji.v3i1.5845.
[7] S. Asmiatun and N. Wakhidah, “Identifikasi Pengelompokkan Kondisi Permukaan Jalan Menggunakan Algoritma K-Means,” J. Pengemb. Rekayasa dan Teknol., vol. 14, no. 1, p. 17, 2019, doi: 10.26623/jprt.v14i1.1215.
[8] D. Sudrajat, R. D. Dana, N. Rahantingsih, A. R. Dikananda, and D. A. Kurnia, “Clustering student’s satisfaction in complex adaptive blended learning with the six value system using the K-means algorithm,” Univers. J. Educ. Res., vol. 7, no. 9, pp. 1990–1995, 2019, doi: 10.13189/ujer.2019.070920.
[9] S. Madhukumar and N. Santhiyakumari, “Evaluation of k-Means and fuzzy C-means segmentation on MR images of brain,” Egypt. J. Radiol. Nucl. Med., vol. 46, no. 2, pp. 475–479, 2015, doi: 10.1016/j.ejrm.2015.02.008.
[10] T. Velmurugan, “Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data,” Appl. Soft Comput. J., vol. 19, pp. 134–146, 2014, doi: 10.1016/j.asoc.2014.02.011.
[11] F. Y. Bisilisn, Y. Herdiyeni, and B. I. B. P. Silalahi, “Optimasi K-Means Clustering Menggunakan Particle Swarm Optimization pada Sistem Identifikasi Tumbuhan Obat Berbasis
Citra K-Means Clustering Optimization Using Particle Swarm Optimization on Image Based Medicinal Plant Identification System,” *Ilmu Komput. Agri-Informatika*, vol. 3, no. 2002, pp. 38–47, 2014.

[12] C. Zhu, C. U. Idemudia, and W. Feng, “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques,” *Informatics Med. Unlocked*, vol. 17, no. January, p. 100179, 2019, doi: 10.1016/j.imu.2019.100179.

[13] A. Jamal, A. Handayani, A. A. Septiandri, E. Ripmiatin, and Y. Effendi, “Dimensionality Reduction using PCA and K-Means Clustering for Breast Cancer Prediction,” *Lontar Komput. J. Ilm. Teknol. Inf.*, no. December, p. 192, 2018, doi: 10.24843/lkjiti.2018.v09.i03.p08.

[14] Z. Cebeci and F. Yildiz, “Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures,” *J. Agric. Informatics*, vol. 6, no. 3, pp. 13–23, 2015, doi: 10.17700/jai.2015.6.3.196.

[15] T. Bücker, “Customer Clustering in the Insurance Sector by Means of Unsupervised Machine Learning,” 2016.