Clustering Student Understanding Levels In Software Engineering Courses

Martini Dwi Endah Susanti[1]*, Rindu Puspita Wibawa[2]

[1] Informatics Department, Universitas Negeri Surabaya
[2] Informatics Department, Universitas Negeri Surabaya
Corresponding author. Email: martinisusanti@unesa.ac.id

ABSTRACT
The level of understanding in learning is one of the main things that influence the course of the process of learning activities. Software Engineering is a scientific discipline that addresses all aspects of software production starting from the early stages of system maintenance. In the Software Engineering course, each material is interrelated between one material and another. If students cannot understand the previous material, it will be difficult for them to understand the next material. Data mining technology can be used to understand some of the problems that arise in education management, including to analyze the level of students' understanding of certain subjects. This study aims to determine the level of understanding clusters of students in Software Engineering courses using the K-means Clustering method. The results of this study are that student data is clustered into 2 clusters, namely the GOOD and POOR clusters. Evaluation was carried out using the Elbow method and calculating the Silhouette Score. The optimal number of clusters obtained from the elbow method is 2 clusters with a silhouette score of 0.836.

Keywords: *clustering, k-means clustering, elbow method, silhouette score, data mining.*

1. INTRODUCTION

In the learning process, students' understanding of learning material is very important. The level of understanding in learning is one of the main things that influence the course of the process of learning activities. Not only is the willingness to learn from each student, but educators also play a very important role in providing learning materials that can be understood by each student and are supported by good learning methods and media as well. The emergence of the Learning Management System (LMS) which certainly makes it easier to deliver lecture material to increase student understanding of certain subjects. The existence of technology such as LMS makes it easier to convey material from instructors to students, but it is still difficult to predict student performance, it is difficult to know how much they understand the level of student understanding of a particular subject. Conducting an evaluation is one way to find out the learning outcomes of a student. Because the purpose of the evaluation itself is to assist students in knowing their learning performance. Thus, the performance and level of student understanding of a particular subject is expected to be seen from the assessment of tests and assignments [1].

Software Engineering is a course that studies software concepts and software engineering. Software Engineering is a scientific discipline that addresses all aspects of software production starting from the initial stages, namely communication, capturing requirements (analyzing user needs), specification (determining specifications of user requirements), design, coding, testing to maintenance (system maintenance after use) [2]. Being a software engineer is one of the jobs that awaits students after graduation. To become a software engineer requires several skills including one having to master the field of Software Engineering. In the learning process, concepts are things that must be learned, mastered and understood by a student. Software Engineering is a fairly complex subject with many abstract concepts for students to understand and learn. In Software Engineering subjects, each material is interrelated between one material and another. If students cannot understand the previous material, it will be difficult for them to understand the next material. So that makes students tend to have difficulty understanding Software Engineering lessons. For teachers of Software Engineering courses, this is a problem in providing material because students' understanding is uneven, this of course will lead to learning objectives that cannot be achieved in accordance with the semester study plan.

Data mining is a technique of digging up hidden or hidden valuable information in a very large data collection (database) so that an interesting pattern is found that was previously unknown. Data mining techniques are widely

Y. A. Yusran et al. (eds.), *Proceedings of the 2023 Brawijaya International Conference (BIC 2023)*, Advances in Economics, Business and Management Research 294,
https://doi.org/10.2991/978-94-6463-525-6_76

used in various fields, including education to discover new and hidden patterns from student or student data. Several methods that are often mentioned in the data mining literature include clustering, classification, association rules mining, neural networks, genetic algorithms and so on [3]. Data mining technology can be used to understand some of the problems that arise in education management, including to analyze the level of students' understanding of certain subjects. This analysis was carried out to group students according to their ability to understand and master the Software Engineering course material. Grouping the level of students' understanding using clustering techniques or methods is done by dividing student groups into subgroups called clusters. Clustering is a method in data mining that is used to analyze data so that it is more accurate when solving data grouping problems or dividing a set of data into subsets. The purpose of clustering is to group data into a group and then the relationship between members of the same cluster becomes stronger, while the relationship between members of different clusters becomes weaker. Objects in a cluster have similar characteristics but have different characteristics from objects in other clusters. So that clustering is used to determine unknown groups or clusters in a data [4].

This research aims to determine the level of understanding of student clusters in the Software Engineering course. The data mining method is used in this research to extract data and process the data so that it will produce the desired output, namely the level of student understanding of Software Engineering courses. The data used in this research is assessment data from students who have taken Software Engineering courses. The data mining method used in this research is K-means Clustering to determine student understanding level clusters based on students' level of understanding while studying Software Engineering courses for one semester. To evaluate and determine the optimal number of clusters using the *Elbow Method* and then the *Silhouette Score* will be calculated to find out the best model.

2. METHOD

The research methodology used is the concept of data mining using the K-Means Clustering algorithm which refers to the Cross Industry Standard Process (CRISP-DM). CRISP-DM is a framework in data mining which consists of 6 stages to identify data as an input in a process [5], where these stages include the stages as shown in Figure 1 below.

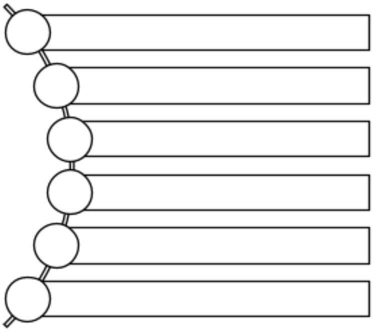


Figure 1 *CRISP-DM Method*

1.1. Business Understanding

The underlying problem in this study is the clustering or distribution of understanding levels of Informatics Engineering students at Surabaya State University for Software Engineering courses. The focus of clustering is grouping students into good, good enough, and not good clusters using the K-Means Clustering algorithm. It is hoped that the use of this algorithm will produce good quality decisions.

1.2. Data Understanding

This process includes data collection, data analysis and data visualization. The data to be processed is related to data on the level of understanding of Informatics Engineering students regarding Software Engineering courses. Where the criteria used are the results of Participation, Assignments assessment, Mid-Semester Evaluation and Final Semester Evaluation. The data used are 139 students taking the Software Engineering course in the even semester of 2023, Informatics Department and Informatics Education Department study programs.

1.3. Data Preparation

Data preparation is a process in data mining to prepare data so that the data can be processed optimally. The steps taken include checking the total data for blanks or missing values, removing duplicate data, checking inconsistent data and sorting data to be used in the next stage. This stage will also convert student letter grades into numbers so that they can be used in the k-means clustering stage.

1.4. Modelling

The modeling used in this study is the k-means clustering model to map data distribution points so that it can be known the number of students' understanding levels of the Software Engineering course. The number of *K* is obtained by the Elbow method to find the optimal *k*.

The K-Means algorithm is defined as an Unsupervised Learning method that has an iterative process in which the dataset is grouped into k number of predetermined non-overlapping clusters or subgroups, making the points in the cluster as close together as possible while trying to keep the clusters in that different space. allocate data points to clusters so that the sum of the squared distances between the cluster centroids and the data points is in the data points. At a minimum, at this position the centroid of the cluster is the average value of the data points in the cluster. The general k-means stages are [6]:

1. Determine the number of clusters (*K*)
2. Select a number of k objects randomly to be used as cluster centroid points
3. Determine the *k* centroid (midpoint or cluster center)
4. Group objects to the nearest centroid cluster based on Euclidean distance in (1):

$$d_{ik} = \sqrt{\sum_{i=1}^{n} (C_{ij} - C_{kj})^2} \quad (1)$$

where: *Cij* = cluster center and *Ckj* = data

5. Recalculate all centroid points with the equation (2)

$$\mu j(t + 1) = \frac{1}{Nsj}\sum_{jaj} xj \qquad (2)$$

where $\mu j(t + 1)$ = new centroid in iteration (t+1)

Nsj= data on cluster Sj

6. Repeat steps 3-5 so that the centroid point value no longer changes.

1.5. Evaluation

This stage evaluates or tests the clustering process carried out. The testing process is carried out using the Elbow method to determine the optimal number of clusters and calculate the *silhouette score* to evaluate whether the resulting k-means model is good enough.

The Elbow method is a method used to produce information in determining the best number of clusters by looking at the percentage of comparison results between the number of clusters that will form an elbow at a point [7]. This method provides ideas by selecting cluster values and then adding the cluster values to be used as a data model in determining the best cluster. And apart from that, the resulting calculation percentage becomes a comparison between the number of clusters added. The different percentage results for each cluster value can be shown using graphs as a source of information. If the value of the first cluster with the value of the second cluster provides a corner in the graph or the value experiences the greatest decrease, then the value of that cluster is the best. The elbow method graph can be seen in Figure 2. Elbow Method Algorithm for determining the K value in K-Means

1. Start
2. Initialize the K value
3. Increase the K value
4. Calculate the sum of squared errors for each K value
5. Look at the results of the sum of square error of the K value which has dropped drastically
6. Determine the K value in the form of an angle
7. Done

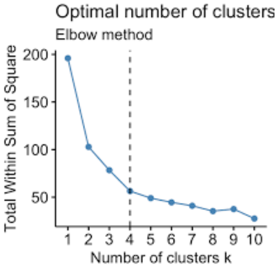


Figure 2 *Elbow Method*

Silhouette score or what is often called silhouette coefficient is a machine learning model measurement method that is able to measure the quality and strength of clusters, so that you can see how well the data is placed in a cluster [8].

$$silhouette\ score = \frac{(b-a)}{max(a,b)} \quad (3)$$

Formula (3) is the formula for the silhouette score. The *a* value is the average intra-cluster distance and the *b* value is the nearest cluster mean distance. The formula for the silhouette score is to divide the result of subtracting the value of the closest mean-cluster distance and the average intra-cluster distance by the maximum value of the average intra-cluster distance and the mean-cluster distance.

1.6. Deployment

At this stage, research results will be reported by compiling a final research report and writing scientific articles for publication at international conferences.

3. RESULT AND DISCUSSION

In this study, the data used were student assessment data in the Software Engineering course. Student assessment includes assessment of participation, assignments, Mid Semester Exam scores and Final Semester Exam scores. The number of data records is 139 which is the number of students taking the Software Engineering course. The following **Table 1** is the dataset used for research.

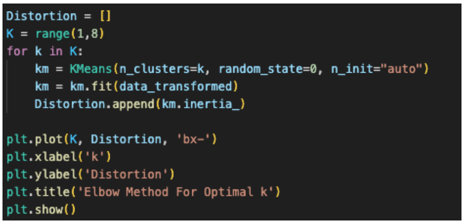
Table 1. Student Dataset

Part	Task	Mid-term	Final
80	54,5	80	80
90	79,5	79	85
80	83,25	91	75
.	.	.	.

.	.	.	.
.	.	.	.
80	78,25	84	85
80	73,75	65	78
80	82	91	75
0	0	0	0
80	80,75	72	85
90	82,5	92	85
80	74,25	80	75
80	38,75	67,5	72

Prior to data processing, data preprocessing is carried out in the form of checking the total missing value, so that the results of data processing will be more effective.

Data processing is carried out using Visual Studio Code tools with the Python programming language. Data that has been pre-processed is processed using the K-means clustering algorithm to determine data clustering. Data is processed using the Elbow method to determine the optimal number of clusters. Figure 3 is python source code that is used for the elbow method.



```
Distortion = []
K = range(1,8)
for k in K:
    km = KMeans(n_clusters=k, random_state=0, n_init="auto")
    km = km.fit(data_transformed)
    Distortion.append(km.inertia_)

plt.plot(K, Distortion, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('Elbow Method For Optimal k')
plt.show()
```

Figure 3 *Source code for elbow method*

The elbow method is one of the methods used to determine the optimal number of *n_clusters* clusters. This method works by running k-means for several k values, then using distortion or inertia as the y-axis and $k_1 \in n$_cluster clusters as the x-axis [9].

k

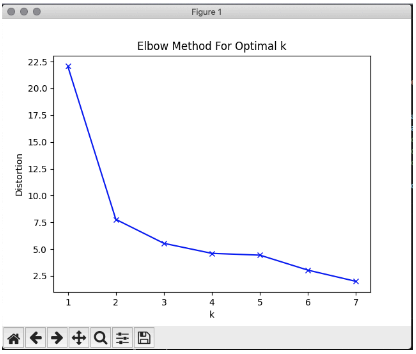


Figure 4 *Elbow method for optimal K*

Figure 4 is the optimal number of clusters for analyzing student understanding data in Software Engineering courses.

As evaluation material, tests were also carried out on various K values. The k values used for evaluation were 2, 3, 4, and 5. Table 2 below shows the results of the clusters performed on each k number.

Table 2. Result Of K-Means Clustering

Number of Cluster	**Cluster Name**	**Number of Members**
2 Cluster	Cluster 1	133
	Cluster 2	6
3 Cluster	Cluster 1	118
	Cluster 2	15
	Cluster 3	6
4 Cluster	Cluster 1	97
	Cluster 2	22
	Cluster 3	14
	Cluster 4	6
5 Cluster	Cluster 1	32
	Cluster 2	71
	Cluster 3	18
	Cluster 4	12
	Cluster 5	6

The data visualization in each cluster can be seen in Figure 5. When the number of clusters = 2, it is clear that the data is separate. Based on the figure, cluster 1 which is colored red is a GOOD cluster with centroids colored black and cluster 2 which is colored blue is a POOR cluster with centroids colored black.

When the number of clusters = 3, it is clear that the data is separated into 3 clusters. Based on the figure, cluster 1 which is colored red is a GOOD cluster with centroids colored black, cluster 2 which is colored green is a ENOUGH cluster with centroids colored black and cluster 3 which is colored blue is a LESS cluster with centroids colored black as seen on Figure 6.

Figure 7, when the number of clusters = 4, it can be seen that the data is separated into 4 clusters. But in the data visualization, it can be seen that there is a lot of overlapping data. So the resulting model is not good. Based on the figure, cluster 1 which is colored red is EXCELLENT cluster with centroids colored black, cluster 2 which is colored green is a GOOD cluster with centroids colored black, cluster 3 which is colored cyan is ENOUGH cluster with centroids given black and cluster 3 which is given a blue color is a POOR cluster with the centroid is given a black color.

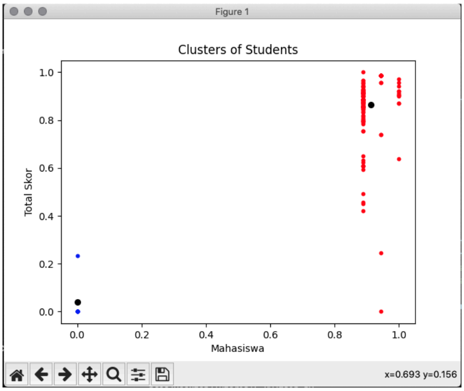


Figure 5 *Number of Cluster = 2*

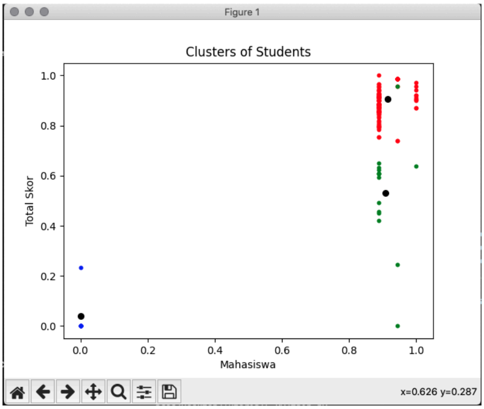


Figure 6 *Number of Cluster = 3*

And on Figure 8, when the number of clusters = 5, it can be seen that the data is separated into 5 clusters, but there is overlapping of the data so that the data does not appear to be separated properly. Based on the figure, cluster 1 which is colored red is a GOOD cluster with centroids colored black, cluster 2 which is colored green is a GOOD cluster with centroids colored black, cluster 3 which is colored cyan is a ENOUGH cluster with centroids given black, cluster 4 which is colored yellow is a UNGOOD cluster with centroids which are colored black and cluster 4 which is colored blue is a BAD cluster with centroids which are colored black.

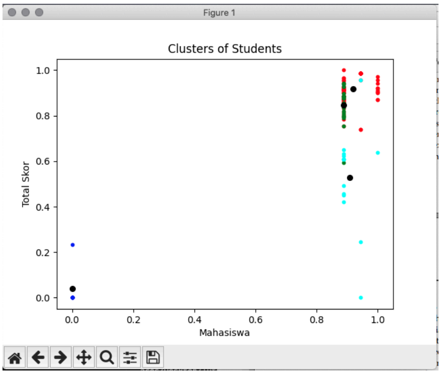


***Figure** 7 Number of Cluster = 4*

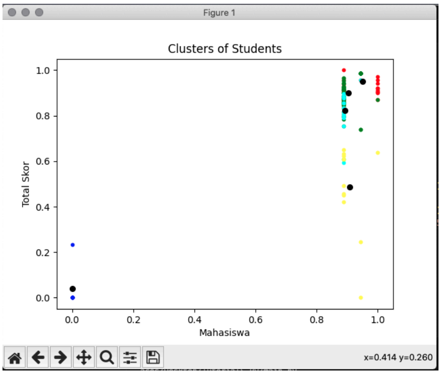


***Figure 8** Number of Cluster = 5*

Based on the application of the elbow method that has been carried out that the optimal cluster results are when k = 2, it can be concluded that the understanding data of students taking Software Engineering courses can be clustered into 2 clusters with the following description on Table 3. So that the results of student clustering will look like in Table 4.

Table 3. Result Cluster Data

Cluster	Description

Cluster 1 (133)	This cluster is data on the level of student understanding of the Software Engineering course with a GOOD rating
Cluster 2 (6)	This cluster is data on the level of student understanding of the Software Engineering course with a POOR rating

Table 4. Clustering Result

Student's Name	**Part**	**Task**	**Mid-term**	**Final**	**Cluster**
MOHAMMAD YUSRIL LUQMAN HAKIM	80	78,25	84	85	0
ACHMAD SYAHRUL RAMADHAN	80	73,75	65	78	0
DIAN NOVITASARI	80	82	91	75	0
HAQQANI FAWWAZ ALKAROMI	0	0	0	0	1
DINA AMILIA	80	80,75	72	85	0
FERDY SEPTIAWAN	90	82,5	92	85	0
AWWALIA AROFATUN NIKMAH	80	74,25	80	75	0
MOH. RIFKI DARMAWAN	80	38,75	67,5	72	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
HERMAN WORDAUN RAINHART RUMY	80	52,5	70	0	1
FATAHILLAH ASWAM QOWA'ID	0	20	0	0	1

In the data visualization image for cluster 2, it can be seen that the data is separated and well clustered. The k-means model is also evaluated by calculating the silhouette score. If we look at the example of the cluster results above by applying the elbow method, even without a model, we already know that the data has 2 groups. This is because we clearly see that one cluster is quite far from the other clusters as well as the distance between observations in one cluster to the centroid is quite close. The silhouette coefficient represents this evaluation in the numerical world where the value is between -1 to 1[10]. Silhouette coefficient can be calculated using (4).

$$\frac{b-a}{max(a,b)}, \text{ di mana } a := \text{jarak intra-cluster} \qquad b := \text{jarak inter-cluster} \tag{4}$$

Table 5 Silhouette Scores For Different K Values.

Number of Cluster (k)	Silhouette Score

2	0.836
3	0.599
4	0.409
5	0.203

Based on the silhouette score obtained in Table 5, it can be seen that the optimal number of clusters for research on students' understanding of data on Software Engineering courses is 2 clusters. The cluster is defined as a GOOD rating cluster and a POOR rating cluster.

4. CONCLUSIONS

The implementation of the k-means clustering model in clustering the level of student understanding of the Software Engineering course produces 2 optimal clusters based on the application of the Elbow method and also the calculation of the Silhouette Score. Data on students' understanding of the Software Engineering course are grouped into GOOD and POOR clusters.

This model is expected to be a reference for lecturers to develop the courses they teach so that they can increase students' understanding of the Software Engineering course. This course is important for students to understand, especially in the field of Informatics Engineering because this course is related to the skills possessed by students related to software development which will later be used as their provision in the world of work.

REFERENCES

[1] Paul, J., & Jefferson, F. (2019). A comparative analysis of student performance in an online vs. face-to-face environmental science course from 2009 to 2016. *Frontiers in Computer Science*, *1*, 7.

[2] Fernandes, J.M., Machado, R.J. (2016). Software Engineering. In: Requirements in Engineering Projects. Lecture Notes in Management and Industrial Engineering. Springer, Cham. https://doi.org/10.1007/978-3-319-18597-2_2

[3] Gorunescu, Florin. (2011). Data Mining: Concepts, models and techniques.

[4] Xiao-Xia Yin, Sillas Hadjiloucas, Yanchun Zhang, Min-Ying Su, Yuan Miao, Derek Abbott, Pattern identification of biomedical images with time series: Contrasting THz pulse imaging with DCE-MRIs. Artificial Intelligence in Medicine, Volume 67. 2016. Pages 1-23. ISSN 0933-3657. https://doi.org/10.1016/j.artmed.2016.01.005.

[5] Gunawan, Gunawan. (2021). DATA MINING USING CRISP-DM PROCESS FRAMEWORK ON OFFICIAL STATISTICS: A CASE STUDY OF EAST JAVA PROVINCE. Jurnal Ekonomi dan Pembangunan. 29. 183-198. 10.14203/JEP.29.2.2021.183-198.

[6] Zubair, Md & Iqbal, Asif & Shil, Avijeet & Chowdhury, Mohammad & Moni, Mohammad Ali & Sarker, Iqbal. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. Annals of Data Science. 1-20. 10.1007/s40745-022-00428-2.

[7] Tosida, Eneng & Wahyudin, Irfan & Andria, Fredi & Djatna, Taufik & Ningsih, Winda & Lestari, Desti. (2020). Business Intelligence of Indonesian Telematics Human Resource: Optimization of Customer and Internal Balanced Scorecards. Journal of Southwest Jiaotong University. 55. 10.35741/issn.0258-2724.55.2.7.

[8] Dinh, Duy-Tai & Fujinami, Tsutomu & Huynh, Van-Nam. (2019). Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. 10.1007/978-981-15-1209-4_1.

[9] Habib, A. B. (2021). Elbow method vs silhouette Co-efficient in determining the number of clusters.

[10] Verma, V., Markan, R., Khan, S., Brooks, L., Khandelwal, H., Dawar, V., ... & Paul, A. (2023). RWD12 Identifying Clinical Subgroups/Clusters of Alzheimer's Patients from Optum's De-Identified Market Clarity Database Using Machine Learning Techniques. *Value in Health*, *26*(6), S362.