Application Cluster Analysis On The Google Play Store Using The K-Means Method

Hastri Cantya Danahiswari
Computer Science Faculty
UPN "Veteran" East Java
Surabaya, Indonesia
hastricantyad@gmail.com

Yovan Febriawan Nurpratama
Computer Science Faculty
UPN "Veteran" East Java
Surabaya, Indonesia
yovebz2502@gmail.com

Dhian Satria Yudha Kartika
Computer Science Faculty
UPN "Veteran" East Java
Surabaya, Indonesia
dhian.satria@upnjatim.ac.id

***Abstract*—Implementation of data mining can be used to identify information that will be useful for several parties. There are various methods in data mining, one of the methods used in clusters is the K-Means algorithm. These clusters can be used for android developers in identifying what applications need to be improved and developed to make it better for android users. The results showed that there were two clusters that had different averages. The first cluster is defined as an application that is less attractive to users due to several factors, while for the second cluster it is defined as an application that the user is interested in, caused by the application offering the features that the user needs, is informative, does not require costs and can function properly.**

Keywords—cluster application; google play store; k-means algorithm; data mining

I. INTRODUCTION

For Android users, the Google Play Store is a digital service created by Google to provide users with finding and downloading various applications developed using the Android SDK. Google Play Store was released on October 22, 2008 by Google. In that year, the Apple Store also appeared and since its release, both platforms have provided more than 1 million applications. There are more than 4.5 million apps on the Google Play Store. On the Google Play Store, each app is divided into unique categories. These applications are paid and free ones, so each application has a different rating or assessment based on the user experience using the application. Categories make it easier for users to find the application they need. In addition to providing applications, the Google Play Store also offers digital media such as movies, books, and television. All services on the Google Play Store are free and paid.

Processes that use computational, statistical and statistical techniques machine learning to extract and identify useful information related to large databases are the definition of data mining. In data mining, a K-Means method is implemented for data clusters[5].

In this study, the author uses the K-Means method to use the Google Play Store application dataset from the Kaggle site. The K-Means method was applied to group data based on the clusters that had been created. These clusters can be used for android developers in identifying what applications need to be improved and developed to make it better for android users.

II. METHODOLOGY

The figure below shows the flow of the method implemented based on the K-Means algorithm.

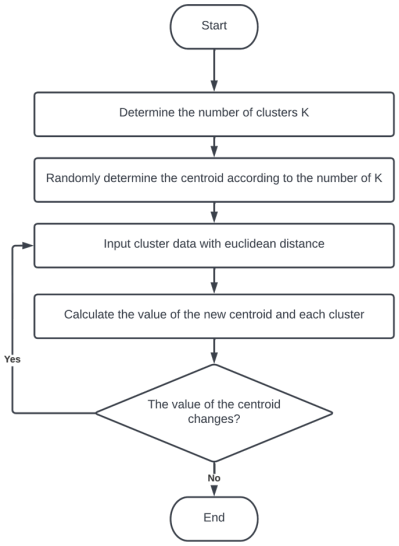

Fig. 1. Flowchart K-Means

A. Data Mining

According to Suntoro, data mining is a process to generate usefully and original information from an extensive database so that it needs to be extracted so that it becomes new information and can help in decision making. Meanwhile, according to Abdurrahman, the data mining process is finding meaningful relationships, patterns and trends by examining extensive data in storage using pattern recognition technology, such as statistics and mathematics[3].

B. Data Mining

This study uses the Google Play Store application dataset. This dataset was taken on the Kaggle website. The available dataset has 1 million Google Play Store application data with 11 attribute columns. Furthermore, the data that has been tidied up by removing the missing value will be normalized using the MaxMin Scaling method. Two attributes are used for the K-Means model: user rating and the number of downloads[9].

C. Clustering with K-Means

Centroid is used to calculate the number of clusters based on the calculation of the distance data. This addition is done by calculating the distance based on each value in the data with its centroid distance[1]. Data adjacent to the centroid will be entered with the centroid and then entered into the cluster with the centroid value.

D. K-Means Algorithm

According to Sandi, K-Means Cluster Analysis is a non-hierarchical cluster analysis method to partition objects into clusters or groups of things based on their characteristics so that objects with the same features will be grouped in the same set with objects that have different characteristics will be grouped into other clusters[10]. According to Yudi Agusta, the basic algorithm for clustering using the K-Means method will be carried out based on the following steps:

- First, determine the number of clusters.
- Second, allocating data into clusters randomly.
- Third, calculate the centroid or the average value of the data in each cluster.
- Fourth, allocate each data to the nearest centroid or average.
- Fifth, go back to step 3.

If it is found that there is data that moves clusters or there is a change in the centroid value, then there is data whose weight is above a predetermined threshold. Determining the K-Means value is done by assessing the number of clusters, namely K Clusters. Then, randomly select the centroid according to the number of sets. Next, enter each data into the centroid with the closest distance to the Euclidean distance. Thus, a cluster will be formed according to the distance between the data and the centroid[4]. The cluster that has been formed then recalculates the centroid value. Then, the formed cluster data is entered again into the nearest centroid cluster. Repeat the last step until the centroid value is unchanged and stable.

III. Results and Discussion

A. Dataset Exploration

```
   Rating  Installs
0       0        10
1       4      5000
2       0        50
3       5        10
4       0       100

(167000, 2)

Rating      int64
Installs    int64
dtype: object
```

Fig. 2. Dataset Exploration

The Kaggle site provides data that will be implemented in the research in the form of a description of the application on the Google Play Store. Implemented attributes in this study are Ratings and Install, where the type of both attributes is an integer. The dataset has about 167,000 application data, as shown above.

B. Visualization of Possible Clusters

This stage is used to check between each attribute against the presence of cluster possibilities. Visualization can be done in a variety of methods[8]. 2D graphic shapes or 3D in cartesian diagrams is one of them. Here is the data visualization between the Rating and Installs attributes.

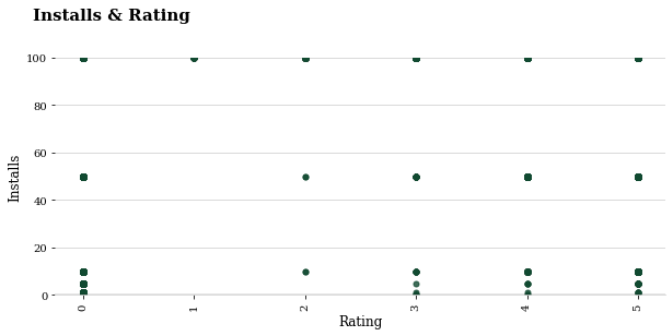

Fig. 3. Visualization of Possible Clusters

C. Determining the Most Optimal K Amount

The number of K or the number of clusters is the foremost resource for the K-Means algorithm. The Elbow method is one of the methods to determine the optimal number of K for the dataset[6]. The following graph shows the elbow method for the Rating and Installs attributes.

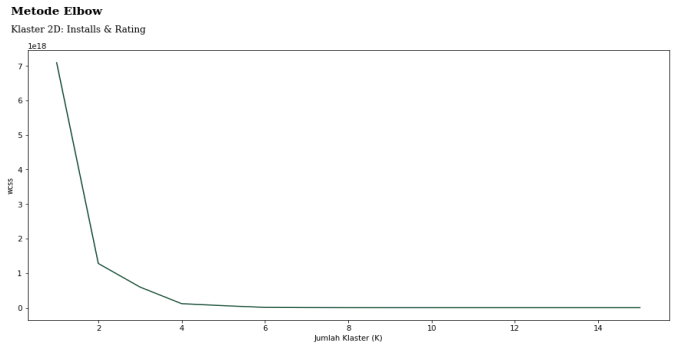

Fig. 4. Determination of K Amount Using Elbow Method

D. Cluster Validation

The silhouette Coefficient is one method to verify that it is as good as the clusters formed from the analyzed dataset. The technique is created from a distance between clusters and entities within the cluster[2]. In addition to verification, this method can also be used to select the most optimal number of K. The lowest range on this method is -1, while the highest is +1. The better the data grouping of a cluster if it leads towards the value of 1[7]. The value of 0 represents the distance between little clusters and the importance of 0 means that the cluster is not well-formed. Here are the results of cluster validation on the Rating and Installs attributes.

```
Jumlah klaster = 2 nilai rata-rata silhoutte = 0.9996717611595615
Jumlah klaster = 3 nilai rata-rata silhoutte = 0.9995557009163731
Jumlah klaster = 4 nilai rata-rata silhoutte = 0.9978605544368447
Jumlah klaster = 5 nilai rata-rata silhoutte = 0.9971576160087995
```

Fig. 5. Cluster Validation

The picture above shows that the silhouette value is the largest owned by the cluster with the number of K = 2, which is 0.99. The deal is almost close to the value of 1 so that their grouping in the cluster becomes better. A Sum of K = 2 is a K value implemented in research.

E. Assign Cluster Results and Centroids to Each Cluster

	Rating	Installs	Cluster
0	0	10	0
1	4	5000	0
2	0	50	0
3	5	10	0
4	0	100	0
...	...	...	...
166995	5	5000	0
166996	0	50	0
166997	4	100000	0
166998	0	10	0
166999	0	50	0

[167000 rows x 3 columns]

	Cluster	Rating	Installs
0	0	2.2	109311.1
1	1	4.4	727272727.3

Fig. 6. Assign Cluster and Centroid Results

By entering the value of K = 2, it can be concluded that there are 2 clusters in the dataset Google Play Store by using the Rating and Installs attributes. The centroid of each cluster can be seen in the picture above.

F. Visualization of Cluster Results

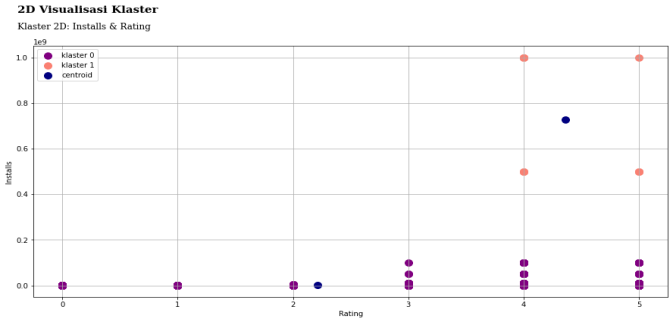

Fig. 7. Cluster Visualization

In the picture above, a cluster is formed with the number of K = 2 between the Rating and Installs attributes. The analysis of these results begins with two segments as follows.

- Cluster 0: applications in this segment are at the bottom of the number of downloads and have a low average rating of 2.2.
- Cluster 1: applications in this segment are in the top position in the number of downloads and have a high average rating of 4.4.

IV. CONCLUSION

The results showed that there were two clusters that had different averages. The first cluster is defined as an application that is less attractive to users due to several factors, for example, an application that does not run according to its function, applications that cannot be opened, containing inappropriate information and others. So it has a low average of reviews and download counts. While for the second cluster it is defined as an application that the user is interested in, caused by the application offering the features that the user needs, is informative, does not require costs and can function properly. So it has a high average of reviews and many downloads. With this study, it is hoped that application developers can improve the features of applications that are still not in demand by users and still maintain the quality of the application that already has a good image in the eyes of the user.

REFERENCES

[1] Bozanta, A., & Co., M. K-Means vs. Fuzzy C-Means : A comparative analysis of two popular clustering techniques on the featured mobile applications benchmark, 2018.

[2] Effendi, J., & M. Jorgi, R. Application cluster analysis on the google play store using the k-mean method, 2018, 4(1), 978–979.

[3] Febrianti, F., Hafiyusholeh, M., & Asyhar, AH. Comparison of iris data clustering using k-means and fuzzy c-means methods. Journal of Mathematics "MANTIK,"2016, 2(1), 7.

[4] Martin, W., Sarro, F., Jia, Y., Zhang, Y., & Harman, M.. A survey of app store analysis for software engineering. IEEE Transactions on Software Engineering, 2017, 43(9), 817–847.

[5] Putra, RR, & Wadisman, C. Data mining implementation of potential customer selection using k-means algorithm, 2018, 1, 72–77.

[6] Sandi, TAA, Raharjo, M., Putra, JL, & Ridwan, R. Customer loyalty clustering with rfm (Recency, Frequency, Monetary) and k-means models. Journal of Pilar Nusa Mandiri, 2018, 14(2), 239.

[7] Sangani, C., & Ananthanarayanan, S. Sentiment analysis of app store reviews. Technical Report, Stanford University., 1– 5.Setiawan, R, 2013.

[8] Application of data mining using k-means clustering algorithm to determine new student promotion strategy (Case Study: Polytechnic Lp3i Jakarta). J. Lantern Ict, 3(1), 2016, 76–92.

[9] Wira, SH, Fahmi, MJ, Rahmatullah, S., & Gata, Windu. Analysis of application clusters in the app store using the k-means method. informatics ferris wheel, 2020, 8(2) 86-90.

[10] MA Syakur, BK Khotimah, EMS Rochman, and BD Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile clusters," in IOP Conference Series: Materials Science and Engineering, 2018, vol. 336, no. 1, p. 12017.