CLUSTERING OF POPULAR SPOTIFY SONGS IN 2023 USING K-MEANS METHOD AND SILHOUETTE COEFFICIENT

Nur Rohman[1*]; Arief Wibowo[2]

Magister Ilmu Komputer[1], Sistem Informasi[2]
Universitas Budi Luhur, Jakarta, Indonesia[1,2]
https://www.budiluhur.ac.id/[1,2]
2211601667@student.budiluhur.ac.id[1*], arief.wibowo@budiluhur.ac.id[2]
(*) Corresponding Author

Abstract—*The rapid advancement of technology and globalization in this era has brought about comprehensive and easily accessible music streaming services, one of which is Spotify. According to Kompas.com, Spotify has experienced a rise in subscribers up to 130 million, as a platform that offers various features besides music streaming. Spotify also provides a better user experience and has the ability to compete with other music streaming platforms. The mission of this research is to classify popular Spotify song data in 2023, which can aid in a deeper understanding of listener preferences or music trends. Based on the test results, there were 2 clusters obtained with cluster 0 containing 863 data and cluster 1 containing 90 data. From the testing results conducted in the K-Means analysis, a Silhouette Coefficient of 0.81 was obtained, which falls into the category of Strong Structure. From these results, it can be suggested that cluster formation was done very well to provide more personalized and relevant music recommendations to Spotify platform users. By understanding the preferences and patterns of listeners revealed through clustering, streaming services can enhance user experience by providing more tailored content.*

Keywords: *clustering, data mining, k-means, silhouette coefficient, spotify.*

Abstrak—*Pesatnya perkembangan teknologi dan globalisasi pada era ini menghadirkan layanan streaming music yang lengkap dan mudah diakses, salah sastunya adalah Spotify. Dilansir dari Kompas.com spotify mengalami kenaikan pelanggan hingga 130juta pelanggan, sebagai platform yang mempunyai fitur berbagai macam selain music streaming, Spotify juga memberikan pengalaman pengguna yang lebih baik dan memiliki kemampuan untuk bersaing dengan platform streaming musik lainnya. Misi dari penelitian ini adalah untuk menggolongkan data lagu-lagu pupuler Spotify pada tahun 2023 yang dapat membantu dalam pemahaman lebih lanjut tentang preferensi pendengar atau tren musik. Berdasarkan hasil pengujian yang dilakukan, cluster yang didapatkan sebanyak 2 cluster dengan cluster 0 berisi 863 data sedangkan cluster 1 berisi 90 data. Dari hasil pengujian yang dilakukan dalam analisis K-Means, diperoleh Silhouette Coefficient sebesar 0,81 yang masuk dalam kategori Struktur Kuat, Dari hasil ini, dapat disarankan bahwa pembentukan cluster dilakukan dengan sangat baik untuk memberikan rekomendasi musik yang lebih personal dan relevan kepada pengguna platform Spotify. Dengan memahami preferensi dan pola pendengar yang terungkap melalui pengelompokan, layanan streaming dapat meningkatkan pengalaman pengguna dengan memberikan konten yang lebih disesuaikan.*

Kata Kunci: *klasterisasi, data mining, k-means, silhouette coefficient, spotify.*

INTRODUCTION

The rapid development of technology and globalization has brought about increasingly comprehensive and accessible music streaming services across various platforms. One of them is Spotify. Spotify offers various services such as digitally listening to music using an internet connection. With Spotify, we can easily enjoy music by connecting to the digital internet, listening to music anywhere and anytime, and of course, using portable devices. The activity of listening to music seems to have become a habit that is difficult to ignore or forget (Navisa, Hakim, & Nabilah, 2021). Spotify is a large platform with a substantial user

base. Certainly, analysis is needed to enhance and strengthen competitiveness with other platforms (Privandhani, 2022). In this research process, the researcher utilized public data by conducting clustering of popular songs in 2023 based on the artist and frequently played tracks. This study was conducted because users tend to listen to music based on their favorite artists.

Clustering method, is an unsupervised approach where the nature of each cluster is not predetermined. This process is based on the similarity of attributes within a group. (Ramadhani et al., 2022). According by (Nisa & Yustanti, 2021) Clustering or clustering is a data analysis method that aims to form data groups (clusters) based on similarity of characteristics among the members of each group. The main goal of clustering is to produce groups that have high similarity within the group and, at the same time, have significant differences between the groups. This process can help understand patterns that may exist in the data and facilitate further analysis.

In this research, the researcher utilizes the K-Means algorithm to process data, and this process is carried out using Google Colab and the Python programming language. According by (Aji et al., 2023) The K-Means algorithm is one of the clustering algorithms. The K-Means algorithm is used to cluster data, observations, or cases based on similarities in what is being studied. K-means clustering is a non-hierarchical cluster analysis technique that attempts to divide existing objects into one or more clusters or groups based on their properties. In the K-Means method, accuracy towards object sizes is very high; this algorithm is relatively concise and efficient when handling a large number of objects.

In the previous research conducted by (Hasyim & Muafi, 2022) in determining the promotion strategy for a program implemented with the K-Means clustering technique, the most effective promotion strategy to target new family planning participants is to focus on the most popular family planning programs It is concluded that the solution is to have a certified BPPKB team with associated costs. Additionally, it might also be appropriate to implement field-based promotions for new family planning participants by adapting the use of the promotional mix. In the research (Wahyudi et al., 2023), utilizing the K-Means method and Davies-Bouldin Index, three attributes and 793 data formed three clusters. Further validation was conducted using the Davies-Bouldin Index (DBI), which resulted in a DBI value of 0.679 for the first cluster, 0.816 for the second cluster, and 0.837 for the third cluster. The lower the DBI value, the higher the quality. Therefore, from the DBI values obtained, it can be concluded that the first

cluster with 128 products and a DBI value of 0.679 falls into the category of very popular. In the previous research, the K-Means method was also employed to classify the top 50 Spotify songs from 2010 to 2019 (Musyarofah et al., 2022). In another study conducted by (Triyandana et al., 2022), he K-Means method and DBI (Davies-Bouldin Index) method were used. The case study involved clustering food and beverage menus based on sales levels, resulting in 3 clusters with a model accuracy value of -0.457. The study conducted by (Ramadhani et al., 2022) involves categorizing information to identify several disaster-prone areas in the Purbalingga region. Based on data analysis, there are five groups of disaster-prone areas in Purbalingga Regency with different risk levels: very high, high, moderate, low, and very low. The classification of these areas can serve as a proactive measure against potential disasters, enabling effective ongoing prevention efforts to minimize the impact of disasters on the community. In this research, the analysis results were obtained from the circular cluster diagram, bar graph cluster, and coordinate points related to the valence of songs.

This research aims to determine the optimal number of clusters for grouping songs on the Spotify application in 2023 using the silhouette coefficient as an evaluation method. Additionally, this research will conduct the grouping of popular songs in the Spotify application in 2023 to help further understand listener preferences or music trends, with clusters of popular and unpopular songs. By analyzing groups of popular songs, we can identify emerging music trends and patterns. This information can be used by music producers, record labels, and DJs to make better decisions regarding marketing strategies and music production. Furthermore, based on the results of this research, it is hoped that stakeholders in the music industry can make better decisions regarding music development, promotion, and distribution. They can use the information they obtain to optimize their music catalogs and maximize the impact of their promotions. Based on the good accuracy values, the method used employs clustering techniques with the K-Means method. The data processing steps carried out in this research are conducted using the Python programming language.

MATERIALS AND METHODS

In this research, the researcher employs the data mining technique known as CRISP-DM (Cross-Industry Standard Process for Data Mining). This is a technique that utilizes a data mining process model commonly used by researchers to solve problems. The research process follows the six

phases of CRISP-DM, as described by (Fransiska et al., 2022), as shown in Figure 1.
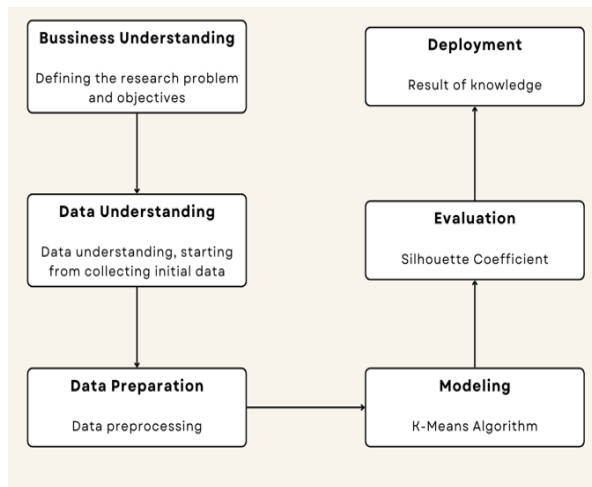
Source: (Fransiska et al., 2022)
Figure 1. Schema Research CRISP-DM

Business Understanding

Business Understanding is the first level of CRISP-DM and is a crucial component. In this stage, the researcher defines the problem of the data mining object and determines the research goals (Fransiska et al., 2022). Business understanding can also be interpreted as a comprehensive understanding of the research scope to meet the overall project and business needs or specific objectives and goals. According by (Dhewayani et al., 2022) Business Understanding is a process such as setting business objectives, understanding the situation and conditions being investigated, and establishing research goals with the aim of solving problems through data mining.

Data Understanding

Data Understanding is the initial stage in understanding data, which involves studying and describing the data, identifying constraints related to data quality, and searching for the required data as initial hypotheses. According by (Dhewayani et al., 2022) Data understanding is the initial step where you consider the data you will use, gather initial data, and assess the quality of your data. When understanding the data, each feature is analyzed through a descriptive process(Nasari & Am, 2023).

Data Preparation

In this stage, data preparation includes selecting, integrating, and cleaning the data to be used. For example, selecting tables, records, and attributes that have been collected to perform data grouping and selection into pre-determined clusters, as well as processing the data for use in the modeling stage. According by (Dhewayani et al., 2022) Data preparation is a process that occurs after data collection.

In this stage, data goes through a series of processes, including identification, data selection, data cleaning, and data transformation. The activities performed in the Data Preparation stage include Data Selection, Data Cleaning, and Data Transformation, according to (Fahmi et al., 2021).

A. Data Selection
   Data Selection is the stage where data is chosen, and attributes are selected based on the goals of data mining.
B. Data Preprocessing
   This process involves cleaning the data by handling outliers, noisy data, and missing values. The goal of this phase is to ensure data quality.
C. Data Transformation
   This process involves grouping attributes into new data, followed by data integration and transformation according to its purpose, then processed in the data mining process.

Modelling

This process begins with the application of modeling techniques and data mining algorithms suitable for the research. K-Means is used in the modeling stage. K-Means is a relatively simple and fast non-hierarchical clustering technique.

In the K-Means method, K-Means divides the existing data into one or more groups (clusters) and places data with similar characteristics in the same cluster, while data with different characteristics are placed in different clusters. (Pratama et al., 2022).

Evaluation

In this step, the focus is on carrying out a quality modeling stage. This step is also conducted to assess the effectiveness of the model used and whether it aligns with the existing K-Means standards (Fransiska et al., 2022). In the Evaluation stage, the researcher uses the Silhouette Coefficient method, which serves to measure the accuracy and quality of the obtained clusters in the modeling stage.

The Silhouette Coefficient is useful for assisting in choosing the optimal number of clusters in the K-Means algorithm. The Silhouette Coefficient values range from -1 to 1, with the guideline that if the positive value approaches 1, it can be concluded that objects in the data are in suitable clusters and have good distance from other clusters. Conversely, if the value is -1, the data objects have been placed in less suitable clusters. The data in Table 1 reflects the accuracy level in measuring the Silhouette Coefficient.

Table 1. The Standard Value of Silhouette Coefficient

Silhouette Coefficient	Standard value
0.7 < Silhouette Coefficient ≤ 1.0	Strong Structure
0.5 < Silhouette Coefficient ≤ 0.7	Medium Structure
0.25 < Silhouette ≤ 0.5	Weak Structure
Silhouette ≤ 0.25	No Structure

Source: (Fransiska et al., 2022)

This evaluation stage assesses whether the applied modeling is appropriate and suitable for this research case and whether it achieves the desired goals.

The results of the evaluation are used to decide whether to proceed to the next steps or to restart if the goals are not achieved.

Deployment

The final stage in the CRISP-DM method is the Deployment stage, which is conducted to generate knowledge or information that can be presented in the form of creating applications or simple reports.

RESULT AND DISCUSSION

Based on the results of the clustering research using the CRISP-DM (Cross-Industry Process for Data Mining) method. The CRISP-DM Data Mining research method is a combination of qualitative and quantitative methods to describe information from the research subjects and then provide prescriptive recommendations (Asyraf & Prasetya, 2023). And using the Python programming language, supported by Google Colab tools, and covering the discussions as follows:

Bussiness Understanding

Spotify is one of the music streaming services that can be listened to and played anywhere. Currently, Spotify has approximately 130 million users who use it as a music streaming service. However, this streaming service faces numerous competitors, and there is a likelihood that the user base of around 130 million could decline if the service is not improved. Spotify can conduct an analysis based on popular songs or music by utilizing data mining.

Data Understanding

The first phase of data understanding involves recognizing and comprehending the data we possess, as well as analyzing it to discover potential information and actions that can be taken from the data. The dataset used in this research was obtained from the website Kaggle.com. This dataset consists of 953 entries collected in the year 2023 and is available in CSV (comma-separated values) format.

This dataset contains 24 relevant attributes used in the research. Refer to Figure 2 for details.

Source: (Research Results, 2024)
Figure 2. Information Data Spotify

Data Preparation

In this phase, the researcher carries out data processing from the Knowledge Discovery in Database (KDD) phase, such as data cleaning, data integration, data selection, and data transformation. In this stage, the 953 data from the Kaggle.com website will be processed through several steps before entering the clustering phase. The steps are as follows:

A. The first stage of data preparation is to determine which data will be processed from the obtained dataset. Not all data will be processed; columns that will not be used in the Modeling stage are removed, as shown in Figure 3.

Source: (Research Results, 2024)
Figure 3. Information Attribute Data

B. The second stage involves processing data with missing values by removing empty values in attributes where such gaps exist. However, based on Figure 4, it is apparent that the data to be used does not have any missing or null values. Therefore, no values are deleted in this process.

Source: (Research Results, 2024)
Figure 4. Finding Empty Or Null Values

C. The third stage involves data transformation. In this stage, the researcher performs data transformation through data normalization or Min-Max scaling. Figure 5 shows the results of the normalization process.

Source: (Research Results, 2024)
Figure 5. Information Normalization Data

Modelling

The next step is the modeling stage, where the researcher determines the number of clusters to be used in the data clustering process using the K-Means method. The researcher uses the Elbow method to determine the number of clusters, and the results can be seen in Figure 6.
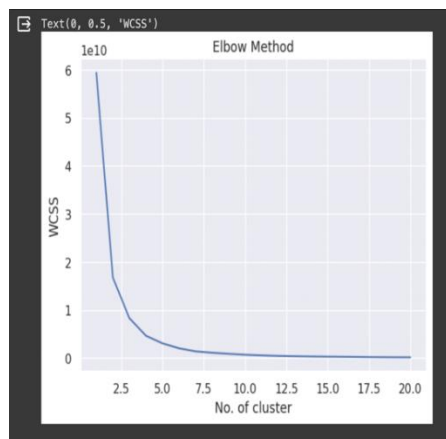
Source: (Research Results, 2024)
Figure 6. Grapich Cluster Elbow Method

From the results of the Elbow method documented in Figure 6, it can be concluded that from the processed and analyzed data, 2 clusters are optimal. In this method, the optimal cluster is the point forming the elbow.

Evaluation

In this stage, the researcher employs an approach with the Silhouette Coefficient technique, which is used to measure or test the quality of the previously obtained clusters in the Modeling stage. The researcher measures and tests 2-5 clusters. Table 2 shows the accuracy results for each cluster.

Table 2. Results of the Silhouette Coefficient Values

No	Cluster	Accuration Silhouette Coefficient
1	2	0.8113997870257713
2	3	0.7431462113579907
3	4	0.6924926306048924
4	5	0.6701590066697268

Source: (Research Results, 2024)

From Table 2, it can be concluded that cluster 2 from the K-Means process obtains the highest silhouette value compared to clusters 3, 4, or 5, which is 0.81. Based on Table 1, cluster 2 falls into the Strong Structure criteria. Silhouette Coefficient results greater than 0.7 indicate that the quality assessment of the Spotify popular song clustering in 2023 using the K-Means algorithm shows excellent quality for cluster 2.

Deployment

The deployment stage is carried out to generate information or knowledge related to the research that has been conducted, based on the results of the previous stages. Based on the K-Means clustering results, out of 953 Spotify songs data in the year 2023, cluster 0 contains 863 data, while cluster 1 contains 90 data as shown in Figure 8 for cluster visualization. It can be analyzed that cluster 0 has a better distribution compared to cluster 1. Validation testing was then conducted

using the Silhouette Coefficient method, and the testing resulted in a value of 0.81, as seen in Figure 7.

```
[ ] from sklearn.metrics import silhouette_score

    silhouette_score(x, features)

    0.8113997870257713
```
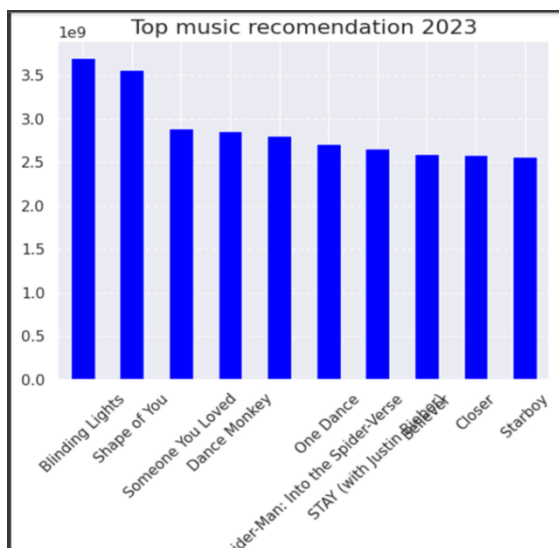
Source: (Research Results, 2024)

Figure 7. Display of the Cluster Model Results
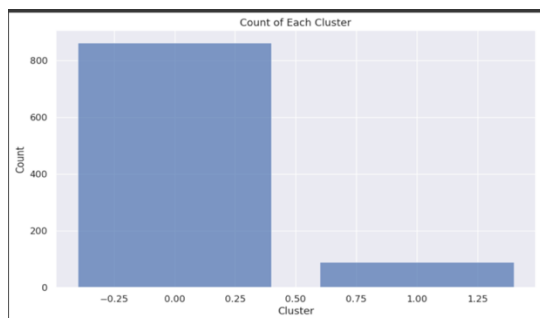
Source: (Research Results, 2024)

Figure 8. Cluster Visualization Display

The bar graph in Figure 9 shows that the songs "Blinding Lights" by The Weeknd and "Shape of You" by Ed Sheeran dominated the index of the most played songs in 2023. Meanwhile, the songs "Dance Monkey" by The Weeknd and "One Dance" by Drake, WizKid, Kyla also ranked in the top five positions. Additionally, "STAY" (with Justin Bieber), "Believer" by Imagine Dragons, "Closer" by The Chainsmokers, Halsey, and "Starboy" by The Weeknd, Daft Punk were in the last positions, making them among the songs that were difficult to play in 2023.

Source: (Research Results, 2024)

Figure 9. Chart of Popular Songs

CONCLUSION

In this study, the researcher successfully applied the K-means clustering method and the CRISP-DM method to test the data. This process involves six phases: business understanding, data understanding, data preparation, modeling, evaluation and deployment. It then classified Spotify's popular songs data for the year 2023 using Google Colab as the development environment and Python as the programming language. The study used 17 attributes from the dataset acquired from Kaggle.com. Based on the Elbow method analysis, the optimal number of clusters (K) found was two. Cluster formation with the K-Means method resulted in cluster 0 with 863 popular songs and cluster 1 with 90 less popular songs. The evaluation of clustering quality with the Silhouette Coefficient method produced a value of 0.81, which approaches the maximum value of 1. This indicates that the clustering performed is of very high quality. The conclusion of this study shows the effectiveness of the K-Means method in classifying popular songs on Spotify and its potential to provide more targeted recommendations to users. The suggestions for this research have significant potential, such as utilizing the latest Spotify data, clustering based on genres or sub-genres, and implementing Spotify data using the K-Medoids algorithm. This is because the K-Medoids algorithm is suitable for addressing the weaknesses of K-Means. K-Medoids is sensitive to noise and outliers, making it a better choice for certain datasets. Therefore, incorporating these suggestions could enhance the robustness and accuracy of the research findings.

REFERENCE

Aji, N. S., Natsir, F., & Istianah, S. (2023). *Penentuan Penjualan Barang Berdasarkan Pengelompokan Produk dengan K-Means Clustering Metode CRISP-DM.*

Asyraf, H., & Prasetya, E. (2023). Implementasi Metode CRISP DM dan Algoritma Decision Tree Untuk Strategi Produksi Kerajinan Tangan pada UMKM A. *Jurnal Media Informatika Budidarma*, *8*, 94–105. https://doi.org/10.30865/mib.v8i1.7050

Dhewayani, F. N., Amelia, D., Alifah, D. N., Sari, B. N., & Jajuli, M. (2022). Implementasi K-Means Clustering untuk Pengelompokkan Daerah Rawan Bencana Kebakaran Menggunakan Model CRISP-DM. *Jurnal Teknologi Dan Informasi*, *12*. https://doi.org/10.34010/jati.v12i1

Fahmi, R. N., Jajuli, M., & Sulistiyowati, N. (2021). Analisa Pemetaan Tingkat Kriminalitas Di Kabupaten Karawang Menggunakan

Algoritma K-Means. *Journal of Information Technology and Computer Science (INTECOMS)*, *4*(1). Retrieved from www.pasundanekspres.co

Fransiska, N., Anggraeni, D., & Enri, U. (2022). Pengelompokkan Data Kemiskinan Provinsi Jawa Barat Menggunakan Algoritma K-Means dengan Silhouette Coefficient. *Jurnal Teknologi Informasi Komunikasi*, *9*, 29–35. https://doi.org/10.38204/tematik.v9i1.921

Hasyim, F., & Muafi. (2022). *Implementasi Data Mining Dalam Menentukan Strategi Promosi Program KB Menggunakan Algoritma K-Means Clustering*. *3*(1). Retrieved from https://ejournal.unuja.ac.id/index.php/core

Musyarofah, U. L., Alima, S. N., & Kartika, D. S. Y. (2022, September). KLASIFIKASI TOP 50 SPOTIFY TAHUN 2010-2019 MENGGUNAKAN METODE K-MEANS CLUSTERING. In Prosiding Seminar Nasional Teknologi dan Sistem Informasi (Vol. 2, No. 1, pp. 215-220).

Nasari, F., & Am, A. N. (2023). Implementasi K-Medoids Clustering Dalam Pengelompokkan Harga 8 Jenis Minyak Goreng. *SINTECH JOURNAL*, *6*. Retrieved from https://doi.org/10.31598

Navisa, S., Hakim, L., & Nabilah, A. (2021). Komparasi Algoritma Klasifikasi Genre Musik pada Spotify Menggunakan CRISP-DM. In *Jurnal Sistem Cerdas*.

Nisa, C., & Yustanti, W. (2021). Studi Perbandingan Algoritma Klastering Dalam Pengelompokan

Persediaan Produk (Studi Kasus : Subdirektorat Perencanaan Sarana Prasarana Dan Logistik PTN X). *JEISBI*, *02*.

Privandhani, N. A. (2022). Clustering Pop Songs Based On Spotify Data Using K-Means And K-Medoids Algorithm. Jurnal Mantik, 6(2), 1542-1550.

Ramadhani, D. I., Damayanti, O., Thaushiyah, O., & Kadafi, A. R. (2022). Penerapan Metode K-Means Untuk Clustering Desa Rawan Bencana Berdasarkan Data Kejadian Terjadinya Bencana Alam. *JURIKOM (Jurnal Riset Komputer)*, *9*(3), 749. https://doi.org/10.30865/jurikom.v9i3.4326

Pratama, E. F. A., Khairil, K., & Jumadi, J. (2022). Implementasi Metode K-Means Clustering Pada Segmentasi Citra Digital. Jurnal Media Infotama, 18(2), 291-301.

Triyandana, G., Putri, L. A., & Umaidah, Y. (2022). Penerapan Data Mining Pengelompokan Menu Makanan dan Minuman Berdasarkan Tingkat Penjualan Menggunakan Metode K-Means. In *Journal of Applied Informatics and Computing (JAIC)* (Vol. 6). Retrieved from http://jurnal.polibatam.ac.id/index.php/JAIC

Wahyudi, T., Sa'adah, N., & Puspitasari, D. (2023). Penerapan Metode K-Means Pada Data Penjualan Untuk Mendapatkan Produk Terlaris di PT. Titian Nusantara Boga. *Jurnal Sains Dan Teknologi*, *5*(1), 228–236. https://doi.org/10.55338/saintek.v5i1.1379