# Game and Application Purchasing Patterns on Steam using K-Means Algorithm

Salman Fauzan Fahri Aulia[1]*, Yana Aditia Gerhana[2], Eva Nurlatifah [3]
Faculty of Science and Technology [1], [2], [3]
Universitas Islam Negeri Sunan Gunung Djati Bandung
Bandung, Indonesia
salman.fazzz@gmail.com[1], yanagerhana@uinsgd.ac.id[2], evanurlatifah@uinsgd.ac.id[3]

***Abstract*— Online games are visual games that utilize the internet or LAN networks. With the growth of the gaming industry, platforms like Steam offer a wide variety of games, making it challenging for users to decide which game to play. This study employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology to address this issue by understanding user preferences. The k-means algorithm clusters game data based on similar characteristics, helping users and developers identify the most popular game types. Data sourced from Kaggle, obtained through the Steam API and Steamspy, consists of 85,103 entries. A normalization process is applied to enhance calculation accuracy. The elbow method determines the optimal number of clusters, resulting in three clusters from the k-means algorithm. The evaluation includes the silhouette coefficient, which measures the proximity between variables, and precision purity, which compares labels by assigning a value of 1 (actual) or 0 (false). The study finds an average silhouette coefficient of 0.345 and a precision purity value of 0.734, indicating that the k-means algorithm performs optimally based on the precision purity metric. The findings reveal that free-to-play games are the most popular among users, while the "Animation & Modelling" category is the most expensive based on price comparisons.**

***Keywords*— Clustering, CRISP-DM, Game, K-Means, Purity, Silhouette Coefficient**

## I. INTRODUCTION

Online games are a type of visual game that can be played on a gadget or computer by utilizing an internet network or LAN [1]. The game industry has a significant role in many countries' economic development [2]. The video game industry focused on selling games as complete products, and the emergence of third-party services like rentals and resale markets expanded the overall value proposition for gamers. This suggests the industry could benefit from exploring its service offerings beyond the core game product [3].

Steam is one of the largest digital distribution platforms in the gaming industry, developed by Valve Corporation [4]. This platform provides various services to users, such as buying, downloading, discussing, and sharing games. Its success is proven by the availability of more than 12,000 games in its database and the use of approximately 555 million users [5]. This success has led to difficulties for users due to the large number of game choices available and different preferences regarding the cost and quality of the games played [6].

The problem of determining game choices can be overcome with the K-Means clustering algorithm, which is useful for finding patterns in data with similar characteristics [7]. This clustering functions to find out user preferences have different interests and behaviors. By analyzing the characteristics of user groups, it can provide appropriate recommendations and improve marketing and sales [8]. Clustering in this game data, from sales data, is used to determine what types of games are most in demand by users based on the characteristics of the categories and genres of the most played games.

The research uses k-means clustering to categorize students into different engagement levels in an e-learning environment, analyzing metrics like logins and assignment submission times. Three experiments were conducted with varying numbers of clusters in each experiment, and an evaluation was carried out using the silhouette coefficient on the model. The silhouette coefficient results in the first experiment with two-level clustering were 0.700, then three-level clustering 0.598, and five-level clustering 0.380. These results indicate that changes in clusters will affect the silhouette coefficient results [9].

Based on research using k-means on 550 student data, the Davies Bouldin method results were 0.769 from 3 clusters obtained using the elbow method [10]. Further research compared the K-Means and K-Medoids algorithms with the results of the k-means Davies Bouldin index being better, namely 0.134 and 0.277 compared to k-medoids of 0.523 and 0.496 [11]. Further research compared the hierarchical clustering and k-means algorithms with the results of k-means getting a consistent number of clusters with 3 clusters, and the silhouette coefficient k-means result of 0.289 is better than the hierarchical clustering of 0.311 [12]. Further research compared the clustering algorithms, which showed k-means had better performance with a consistent number of clusters and a mean silhouette coefficient k-means of 0.716 higher than DBSCAN of 0.296 and hierarchical clustering of 0.301 [13]. Further research compared DBSCAN and Affinity Propagation (AP) with 3 clusters. The DBSCAN silhouette score result was 0.499, which was better than AP's 0.699 because of its ability to group data density. However, the davies Bouldin index results showed that AP was superior in recognizing data patterns [14]. Further research conducted an evaluation using purity, getting very good results of 95% because the use of this purity is very suitable for data in the form of categories or labels [15].

The problem of users determining the type of game on the

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

Steam platform has never been studied before, making it a novelty to use the k-means algorithm to group game data based on similar characteristics so that user preferences can be identified. Grouping game types based on the characteristics of categories and genres can provide recommendations for types of games to be played by users and for types of games that can increase marketing and sales from game developers.

## II. RESEARCH METHODS

This research uses game data from the Steam platform via the Steam API and the Steamspy website. One method used in data mining is the Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of six stages and was carried out in this study [16]. The CRISP-DM method has the following flow:

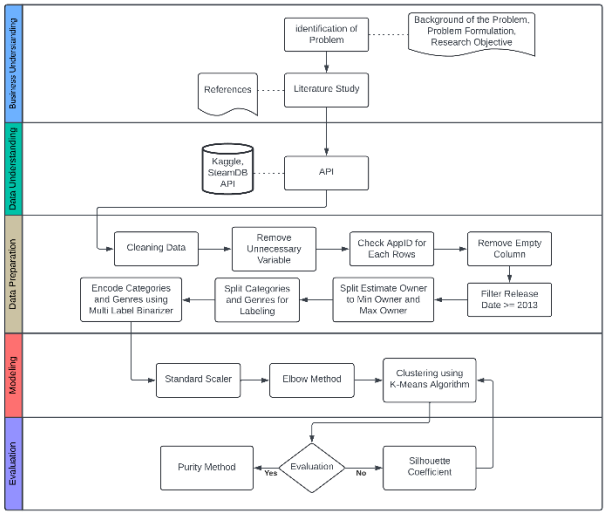


*Fig 1. CRISP-DM*

The stages that will be carried out are business understanding, data understanding, data processing, modeling, and evaluation. The six stages of CRISP-DM will be explained as follows:

### *1) Business Understanding*

Business Understanding, which explains the process of analyzing the model that will be created next. Understanding the business process that is created helps us understand the research's purpose. The process is explained as follows.

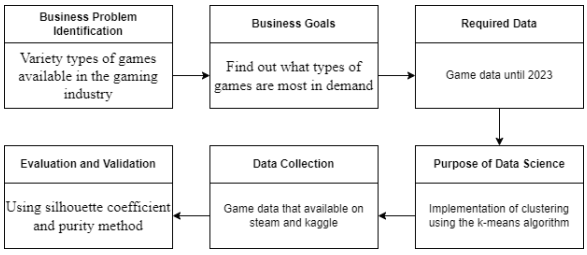


*Fig 2. Business Understanding*

Business problems occur due to the large number of game choices available in the gaming industry; the business objective of the problem is to find out the types of games that are most in demand. The data required is game data until 2023. The process of solving the problem above involves the implementation of clustering using the k-means algorithm to find the types of games that are most in demand by grouping data. The data was obtained from Kaggle and Steamspy, totalling 85103. The evaluation and validation process is needed to determine whether the use of the k-means algorithm works optimally or not.

### 2) *Data Understanding*

Data Understanding is the process of understanding data in research. The process of data understanding is explained as follows.


*Fig 3. Data Understanding*

Understanding the data is necessary for future research because it can facilitate the implementation process later. Data understanding begins by determining what data will be used and then how to collect data that will be used in the study. After that, cleaning the data by normalizing it is necessary to facilitate the modelling process using the k-means algorithm.

The data used is game sales data on the Steam platform obtained from Kaggle and Steamdb API. The data is game sales data until 2023 with a total of 85103 data and contains 36 variables. Of the 36 variables, nine relevant variables were selected for use in the study, which were then normalized to help improve accuracy and make it more accessible during the modeling process using the k-means algorithm.

### *3) Data Preparation*

Data preparation is done to facilitate the model to perform clustering to obtain optimal data results so as to obtain an optimal level of accuracy. The process of data preparation is explained as follows.

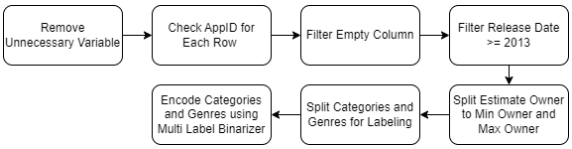


*Fig 4. Data Preparation*

This stage prepares the data by performing a data normalization process to increase accuracy. Of the 36 variables available in this game data, nine were selected for use and carried out at the remove unnecessary variable stage. The selection of the nine variables, namely AppID, Name, Release date, Estimated owners, Price, Positive, Negative, Categories, and Genre, is considered necessary because each of these variables is interrelated with game sales data and is very influential in the application of clustering using the k-means algorithm.

The next stage is carried out to reduce irrelevant data by filtering the existing data. After the filter stage, the next step is to separate the Estimate owner variable, which is the user range data, into two new variables: min owner and max owner. Then, separate the Categories and Genre variables for the labeling process so that the contents of the categories and genre variables become more accessible to read, which will then be encoded using a multi-label binarizer.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

*4) Modeling*

Modeling is a stage to implement the K-Means algorithm on the available data. The K-Means algorithm is used to form clusters by maximizing the similarity of each data to group game data. The stages or flow of this modeling are explained as follows.

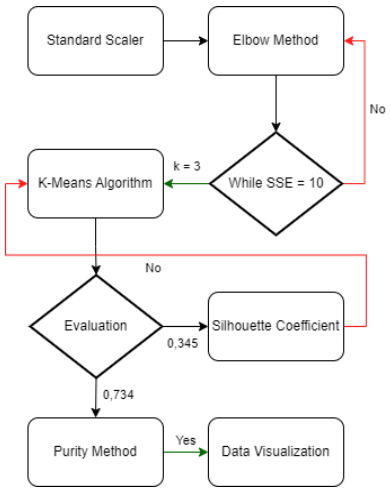


*Fig 5. Modeling*

This modeling stage is carried out using Jupyter Notebook. Jupyter Notebook is the most widely-used system for interactive literate programming. It was designed to make data analysis more accessible to document, share, and reproduce [17]. Jupyter notebooks combine code, documentation, and various cell types to offer a user-friendly environment for both learning and professional use. This unique structure promotes the creation of easily shared, self-contained code projects, fostering collaboration and knowledge dissemination [18].

The first process is to perform a standard scaler to calculate the mean and standard deviation of the nine variables used. Then, it is repeated ten times on the Sum Squared Error (SSE), and the results of the Sum Squared Error (SSE) are depicted on a graph using the elbow method to determine the optimal number of clusters. After obtaining the optimal number of clusters, the k-means algorithm implementation process can be carried out. The results of the modeling using the K-Means algorithm will later be evaluated to determine the performance results of using the K-Means algorithm.

*5) Evaluation*

Evaluation is a stage to evaluate the model that has been created. This evaluation stage is carried out using the silhouette coefficient and purity to determine the performance results of the K-Means algorithm [19]. The evaluation process using the silhouette coefficient is carried out by calculating the closeness between each variable, and then, the purity method is carried out by comparing each label by setting a value of 1 or true on the same label and 0 or false on different labels.

## III. RESULT AND DISCUSSION

The data in this study comes from data on Kaggle obtained from the Steam API and the Steamspy website. The data used is game data on the Steam platform until 2023, totalling 85103. Game data consists of 36 variables (columns), namely:

1. Name : Nama masing-masing *game*
2. Release date : Tanggal rilis
3. Estimated owners : Estimasi jumlah pengguna
4. Peak CCU : Jumlah pengguna dalam satu waktu
5. Required age : Minimal usia
6. Price ($) : Harga dalam dollar / usd
7. DLC count : Jumlah konten tambahan
8. About the game : Deskripsi singkat *game*
9. Supported languages : Bahasa yang dapat digunakan
10. Full audio languages : Audio yang dapat digunakan
11. Reviews : Ulasan pengguna
12. Header image : Gambar header atau poster *game*
13. Website : Alamat web resmi *game*
14. Support url : Alamat bantuan / dukungan *game*
15. Support email : Email bantuan / dukungan *game*
16. Windows / Mac / Linux : Sistem operasi yang didukung
17. Metacritic score : Nilai ulasan metacritic
18. User score : Nilai pengguna
19. Positive : Ulasan positive
20. Negative : Ulasan negative
21. Score rank : Peringkat berdasarkan skor
22. Achievements : Jumlah pencapaian
23. Recommendations : Jumlah rekomendasi pengguna
24. Notes : Catatan
25. Average playtime forever : Rata-rata waktu bermain
26. Average playtime two weeks : Rata-rata waktu bermain seminggu
27. Median playtime forever : Nilai tengah waktu bermain
28. Median playtime two weeks : Nilai tengah waktu bermain seminggu
29. Developers : Pembuat *game*
30. Publishers : Penerbit *game*
31. Categories : Kategori *game*
32. Genres : Jenis *game*
33. Tags : Penanda
34. Screenshots : Gambar *gameplay*
35. Movies : Video *gameplay*

*Fig 6. Data Variable*

To simplify the implementation process, the available game data will need to be cleaned so that it is ready to be processed at the time of implementation.

The first stage eliminates variables that will not be used to facilitate the implementation process later. After this process, nine variables were obtained that will be used in the implementation process.

1. AppID : Kode identitas masing-masing game
2. Name : Nama masing-masing game
3. Release date : Tanggal rilis
4. Estimated owners : Estimasi jumlah pengguna
5. Price ($) : Harga dalam dollar / usd
6. Positive : Ulasan positive
7. Negative : Ulasan negative
8. Categories : Kategori game
9. Genres : Jenis game

*Fig 7. Cleaned Data*

The second stage ensures that each data has an AppID, and a checking process is carried out. It is known that the number of rows in the game data before checking is 85103 rows.

| | AppID | Name | Release date | Estimated owners | Peak CCU | Required age | Price | DLC count | About the game | Supported languages | ... | Average playtime two weeks | Median playtime forever | Median playtime two weeks | Developers | Publishers |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 20200 | Galactic Bowling | Oct 21, 2008 | 0 - 20000 | 0 | 0 | 19.99 | 0 | Galactic Bowling is an exaggerated and stylize... | ['English'] | ... | 0 | 0 | 0 | Perpetual FX Creative | Perpetual FX Creative |
| 1 | 655370 | Train Bandit | Oct 12, 2017 | 0 - 20000 | 0 | 0 | 0.99 | 0 | THE LAWI Looks to be a showdown atop a train.... | ['English', 'French', 'Italian', 'German', 'Sp... | ... | 0 | 0 | 0 | Rusty Moyher | Wild Rooster |
| 2 | 1732930 | Jolt Project | Nov 17, 2021 | 0 - 20000 | 0 | 0 | 4.99 | 0 | Jolt Project: The army now has a new robotics ... | ['English', 'Portuguese - Brazil'] | ... | 0 | 0 | 0 | Campião Games | Campião Games |
| 3 | 1355720 | Henosis™ | Jul 23, 2020 | 0 - 20000 | 0 | 0 | 5.99 | 0 | HENOSIS™ is a mysterious 2D Platform Puzzler w... | ['English', 'French', 'Italian', 'German', 'Sp... | ... | 0 | 0 | 0 | Odd Critter Games | Odd Critter Games |
| 4 | 1139950 | Two Weeks in Painland | Feb 3, 2020 | 0 - 20000 | 0 | 0 | 0.00 | 0 | ABOUT THE GAME Play as a hacker who has arrang... | ['English', 'Spanish - Spain'] | ... | 0 | 0 | 0 | Unusual Games | Unusual Games |

5 rows × 39 columns

*Fig 8. Data Before Check AppID*

There is a difference between the data before and after the AppID check, which initially had 85103 rows; now it has 80385.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

| | AppID | Name | Release date | Estimated owners | Price | Positive | Negative | Categories | Genres |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20200 | Galactic Bowling | Oct 21, 2008 | 0 - 20000 | 19.99 | 6 | 11 | Single-player,Multi-player,Steam Achievements,... | Casual,Indie,Sports |
| 1 | 655370 | Train Bandit | Oct 12, 2017 | 0 - 20000 | 0.99 | 53 | 5 | Single-player,Steam Achievements,Full controll... | Action,Indie |
| 2 | 1732930 | Jolt Project | Nov 17, 2021 | 0 - 20000 | 4.99 | 0 | 0 | Single-player | Action,Adventure,Indie,Strategy |
| 3 | 1355720 | Henosis™ | Jul 23, 2020 | 0 - 20000 | 5.99 | 3 | 0 | Single-player,Full controller support | Adventure,Casual,Indie |
| 4 | 1139950 | Two Weeks in Painland | Feb 3, 2020 | 0 - 20000 | 0.00 | 50 | 8 | Single-player,Steam Achievements | Adventure,Indie |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 85098 | 2669080 | Mannerheim's Saloon Car | Jan 2, 2024 | 0 - 0 | 0.00 | 0 | 0 | Single-player,Tracked Controller Support,VR Only | Adventure,Simulation |
| 85099 | 2736910 | Beer Run | Jan 3, 2024 | 0 - 0 | 0.00 | 0 | 0 | Single-player | Casual,Indie |
| 85100 | 2743220 | My Friend The Spider | Jan 4, 2024 | 0 - 0 | 0.00 | 0 | 0 | Single-player | Adventure,Simulation |
| 85101 | 2293130 | Path of Survivors | Jan 8, 2024 | 0 - 0 | 3.99 | 0 | 0 | Single-player,Steam Achievements,Partial Contr... | Action,Casual,Indie,RPG,Simulation |
| 85102 | 2738840 | The Night Heist | Jan 5, 2024 | 0 - 0 | 9.99 | 0 | 0 | Single-player,Steam Achievements,Full controll... | Casual,Indie |

80385 rows × 9 columns

*Fig 9. Data After Check AppID*

The third stage is filtering to determine whether the data used still contains empty rows from the estimated owners, positive, and negative variables. If the empty data has been removed, the results can be more accurate and optimal.

| | AppID | Name | Release date | Estimated owners | Price | Positive | Negative | Categories | Genres |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 20200 | Galactic Bowling | Oct 21, 2008 | 0 - 20000 | 19.99 | 6 | 11 | Single-player,Multi-player,Steam Achievements,... | Casual,Indie,Sports |
| 1 | 655370 | Train Bandit | Oct 12, 2017 | 0 - 20000 | 0.99 | 53 | 5 | Single-player,Steam Achievements,Full controll... | Action,Indie |
| 3 | 1355720 | Henosis™ | Jul 23, 2020 | 0 - 20000 | 5.99 | 3 | 0 | Single-player,Full controller support | Adventure,Casual,Indie |
| 4 | 1139950 | Two Weeks in Painland | Feb 3, 2020 | 0 - 20000 | 0.00 | 50 | 8 | Single-player,Steam Achievements | Adventure,Indie |
| 5 | 1469160 | Wartune Reborn | Feb 26, 2021 | 50000 - 100000 | 0.00 | 87 | 49 | Single-player,Multi-player,MMO,PvP,Online PvP,... | Adventure,Casual,Free to Play,Massively Multip... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 85077 | 2704060 | Ant Farm Simulator | Jan 5, 2024 | 0 - 20000 | 0.99 | 1 | 1 | Single-player | Casual,Indie,Simulation,Early Access |
| 85079 | 2645600 | The Holyburn Witches | Jan 5, 2024 | 0 - 20000 | 2.99 | 1 | 3 | Single-player | Casual,Indie,Early Access |
| 85083 | 2464700 | Digital Girlfriend | Jan 5, 2024 | 0 - 20000 | 3.74 | 8 | 7 | Single-player | Adventure,Casual,Indie |
| 85085 | 2602790 | Above the Hill | Jan 5, 2024 | 0 - 20000 | 8.49 | 2 | 1 | Single-player,Steam Achievements | Adventure,Indie |
| 85094 | 2345080 | Cats VS Ghosts | Jan 2, 2024 | 0 - 20000 | 0.99 | 1 | 0 | Single-player | Casual,Indie |

62947 rows × 9 columns

*Fig 10. Data After Remove Empty Column*

The fourth stage reduces the amount of game data so that the implementation's results will be more relevant. Only the latest game data is used.

| | AppID | Name | Release date | Estimated owners | Price | Positive | Negative | Categories | Genres |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 655370 | Train Bandit | 2017-10-12 | 0 - 20000 | 0.99 | 53 | 5 | Single-player,Steam Achievements,Full controll... | Action,Indie |
| 3 | 1355720 | Henosis™ | 2020-07-23 | 0 - 20000 | 5.99 | 3 | 0 | Single-player,Full controller support | Adventure,Casual,Indie |
| 4 | 1139950 | Two Weeks in Painland | 2020-02-03 | 0 - 20000 | 0.00 | 50 | 8 | Single-player,Steam Achievements | Adventure,Indie |
| 5 | 1469160 | Wartune Reborn | 2021-02-26 | 50000 - 100000 | 0.00 | 87 | 49 | Single-player,Multi-player,MMO,PvP,Online PvP,... | Adventure,Casual,Free to Play,Massively Multip... |
| 6 | 1659180 | TD Worlds | 2022-01-09 | 0 - 20000 | 10.99 | 21 | 7 | Single-player,Steam Achievements,Steam Cloud | Indie,Strategy |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 85077 | 2704060 | Ant Farm Simulator | 2024-01-05 | 0 - 20000 | 0.99 | 1 | 1 | Single-player | Casual,Indie,Simulation,Early Access |
| 85079 | 2645600 | The Holyburn Witches | 2024-01-05 | 0 - 20000 | 2.99 | 1 | 3 | Single-player | Casual,Indie,Early Access |
| 85083 | 2464700 | Digital Girlfriend | 2024-01-05 | 0 - 20000 | 3.74 | 8 | 7 | Single-player | Adventure,Casual,Indie |
| 85085 | 2602790 | Above the Hill | 2024-01-05 | 0 - 20000 | 8.49 | 2 | 1 | Single-player,Steam Achievements | Adventure,Indie |
| 85094 | 2345080 | Cats VS Ghosts | 2024-01-02 | 0 - 20000 | 0.99 | 1 | 0 | Single-player | Casual,Indie |

61232 rows × 9 columns

*Fig 11. Data After Filter Release Date*

The fifth stage requires separating the estimate owner variable into new columns or variables, namely min owner and max owner, to simplify the calculation process.

| | AppID | Name | Release date | Price | Positive | Negative | Categories | Genres | min_owners | max_owners |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 655370 | Train Bandit | 2017-10-12 | 0.99 | 53 | 5 | Single-player,Steam Achievements,Full controll... | Action,Indie | 0 | 20000 |
| 3 | 1355720 | Henosis™ | 2020-07-23 | 5.99 | 3 | 0 | Single-player,Full controller support | Adventure,Casual,Indie | 0 | 20000 |
| 4 | 1139950 | Two Weeks in Painland | 2020-02-03 | 0.00 | 50 | 8 | Single-player,Steam Achievements | Adventure,Indie | 0 | 20000 |
| 5 | 1469160 | Wartune Reborn | 2021-02-26 | 0.00 | 87 | 49 | Single-player,Multi-player,MMO,PvP,Online PvP,... | Adventure,Casual,Free to Play,Massively Multip... | 50000 | 100000 |
| 6 | 1659180 | TD Worlds | 2022-01-09 | 10.99 | 21 | 7 | Single-player,Steam Achievements,Steam Cloud | Indie,Strategy | 0 | 20000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 85077 | 2704060 | Ant Farm Simulator | 2024-01-05 | 0.99 | 1 | 1 | Single-player | Casual,Indie,Simulation,Early Access | 0 | 20000 |
| 85079 | 2645600 | The Holyburn Witches | 2024-01-05 | 2.99 | 1 | 3 | Single-player | Casual,Indie,Early Access | 0 | 20000 |
| 85083 | 2464700 | Digital Girlfriend | 2024-01-05 | 3.74 | 8 | 7 | Single-player | Adventure,Casual,Indie | 0 | 20000 |
| 85085 | 2602790 | Above the Hill | 2024-01-05 | 8.49 | 2 | 1 | Single-player,Steam Achievements | Adventure,Indie | 0 | 20000 |
| 85094 | 2345080 | Cats VS Ghosts | 2024-01-02 | 0.99 | 1 | 0 | Single-player | Casual,Indie | 0 | 20000 |

61232 rows × 10 columns

*Fig 12. Data After Separating Estimate Owner Variable*

The sixth stage requires separating or changing the form of the variable categories and genres into an array.

| | AppID | Name | Release date | Price | Positive | Negative | Categories | Genres | min_owners | max_owners |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 655370 | Train Bandit | 2017-10-12 | 0.99 | 53 | 5 | [Single-player, Steam Achievements, Full contr... | [Action, Indie] | 0 | 20000 |
| 3 | 1355720 | Henosis™ | 2020-07-23 | 5.99 | 3 | 0 | [Single-player, Full controller support] | [Adventure, Casual, Indie] | 0 | 20000 |
| 4 | 1139950 | Two Weeks in Painland | 2020-02-03 | 0.00 | 50 | 8 | [Single-player, Steam Achievements] | [Adventure, Indie] | 0 | 20000 |
| 5 | 1469160 | Wartune Reborn | 2021-02-26 | 0.00 | 87 | 49 | [Single-player, Multi-player, MMO, PvP, Online... | [Adventure, Casual, Free to Play, Massively Mu... | 50000 | 100000 |
| 6 | 1659180 | TD Worlds | 2022-01-09 | 10.99 | 21 | 7 | [Single-player, Steam Achievements, Steam Cloud] | [Indie, Strategy] | 0 | 20000 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 85077 | 2704060 | Ant Farm Simulator | 2024-01-05 | 0.99 | 1 | 1 | [Single-player] | [Casual, Indie, Simulation, Early Access] | 0 | 20000 |
| 85079 | 2645600 | The Holyburn Witches | 2024-01-05 | 2.99 | 1 | 3 | [Single-player] | [Casual, Indie, Early Access] | 0 | 20000 |
| 85083 | 2464700 | Digital Girlfriend | 2024-01-05 | 3.74 | 8 | 7 | [Single-player] | [Adventure, Casual, Indie] | 0 | 20000 |
| 85085 | 2602790 | Above the Hill | 2024-01-05 | 8.49 | 2 | 1 | [Single-player, Steam Achievements] | [Adventure, Indie] | 0 | 20000 |
| 85094 | 2345080 | Cats VS Ghosts | 2024-01-02 | 0.99 | 1 | 0 | [Single-player] | [Casual, Indie] | 0 | 20000 |

61232 rows × 10 columns

*Fig 13. Data After Separating Variable Into An Array*

Multilabel binarization is a classification stage where data is filled out with input "1," which means true, and input "0," which means false. With this process, it will be easier to identify the contents of the variable categories and genres [20].

| | AppID | Name | Release date | Price | Positive | Negative | min_owners | max_owners | Captions available | Co-op | ... | Short | Simulation | Software Training | Sports | Strategy | Ti |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 50034 | 218700 | Angelica Weaver: Catch Me When You Can | 2013-01-07 | 13.99 | 92 | 39 | 20000 | 50000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | |
| 32729 | 211280 | Primal Fears | 2013-01-08 | 9.99 | 178 | 124 | 50000 | 100000 | 0 | 1 | ... | 0 | 0 | 0 | 0 | 0 | |
| 27769 | 219200 | Droid Assault | 2013-01-09 | 12.99 | 296 | 65 | 50000 | 100000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | |
| 36859 | 220090 | The Journey Down: Chapter One | 2013-01-09 | 6.99 | 1070 | 156 | 100000 | 200000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | |
| 52140 | 215710 | Fieldrunners 2 | 2013-01-10 | 9.99 | 376 | 65 | 100000 | 200000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 1 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 84981 | 2710320 | THE DESCENT | 2024-01-05 | 6.79 | 8 | 1 | 0 | 20000 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | |
| 85029 | 2736710 | TOON | 2024-01-05 | 1.19 | 1 | 0 | 0 | 20000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | |
| 84993 | 2642700 | 鬼打墙 | 2024-01-05 | 7.64 | 42 | 26 | 0 | 20000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | |
| 84838 | 2633500 | Dragon Throne: Battle of Red Cliffs | 2024-01-06 | 9.99 | 4 | 2 | 0 | 20000 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 1 | |
| 85058 | 2738420 | Backrooms: Eight Levels | 2024-01-06 | 0.74 | 1 | 0 | 0 | 20000 | 0 | 0 | ... | 0 | 1 | 0 | 0 | 0 | |

*Fig 14. Multilabel Binarization Process*

A standard scaler is one of the data normalization methods that calculates the average value and stores the raw data of each data. This method can maintain the consistency of the numeric characteristics of the data set [21].

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

```
Means of the scaled features (should be close to 0):
[ 1.48532581e-17  0.00000000e+00 -1.62457511e-18  4.64164317e-18
 -1.06757793e-17 -2.22798872e-17 -3.13310914e-18 -5.10580748e-18
  1.29966009e-17  4.92014176e-17  0.00000000e+00 -2.41365445e-17
  1.25324365e-17 -1.29966009e-17 -1.85665727e-18 -3.77133507e-17
 -7.89079338e-17  4.27031171e-17 -9.79386708e-17 -1.39249295e-18
 -3.89898026e-17 -5.01297462e-17 -6.68396616e-17  7.42662906e-17
  5.75563753e-17 -3.71331453e-18  5.47713894e-17 -3.82935561e-17
  5.56997180e-18  9.19045347e-17 -1.00259492e-16 -1.10471107e-16
  4.13106242e-17 -2.22798872e-17  1.48532581e-17  5.56997180e-18
 -7.42662906e-17  1.92628191e-17  7.42662906e-18 -1.48532581e-17
 -1.25324365e-17  1.85665727e-18 -1.85665727e-18  8.12287554e-19
  1.62457511e-18 -5.94710531e-17 -1.06757793e-17 -4.64164317e-19
  5.56997180e-18  1.15112751e-16 -2.32082158e-18  8.12287554e-19
 -5.56997180e-17 -3.24915022e-18  8.12287554e-19  9.84028351e-17
  2.08873942e-18  1.94949013e-17 -5.89488682e-17 -8.35495770e-18
  2.43686266e-18 -9.28328633e-18 -5.45393072e-18 -5.56997180e-18
  2.64573660e-17  1.57815868e-17  8.12287554e-19  5.77884574e-17
  1.02116150e-17  3.01706806e-18  3.75973096e-17  8.12287554e-19
 -1.06757793e-17 -1.29966009e-17  1.20682722e-17  1.16041079e-17]
Standard deviations of the scaled features (should be close to 1):
[1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.
 1. 1. 1. 1.]
```

*Fig 15. Standard Scaler Process*

The elbow method, which is based on the sum square error calculation, is used to determine the optimal number of clusters using visual techniques[22].

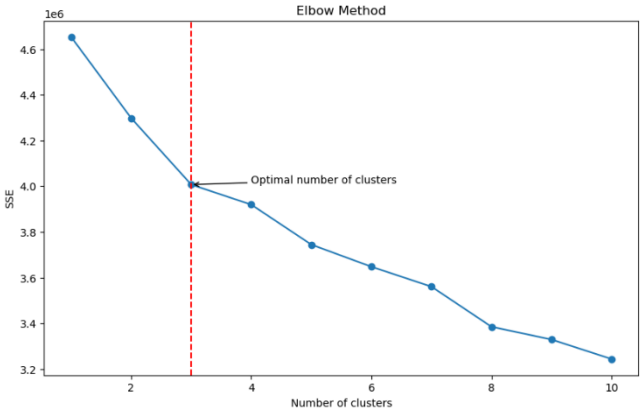


*Fig 16. Result of Elbow Method*

The results of the elbow method calculation show that the optimal number of clusters is 3. The decrease in SSE value after cluster 3 is slight, so the reason for determining cluster 3 as the optimal cluster is that the more clusters there are, the longer the time required.

After obtaining the optimal number of clusters using the elbow method, the data will be grouped based on the characteristics of each data using the K-Means algorithm to facilitate the analysis process of the data used. From these results, a heatmap correlation diagram is obtained to determine the relationship between each variable.

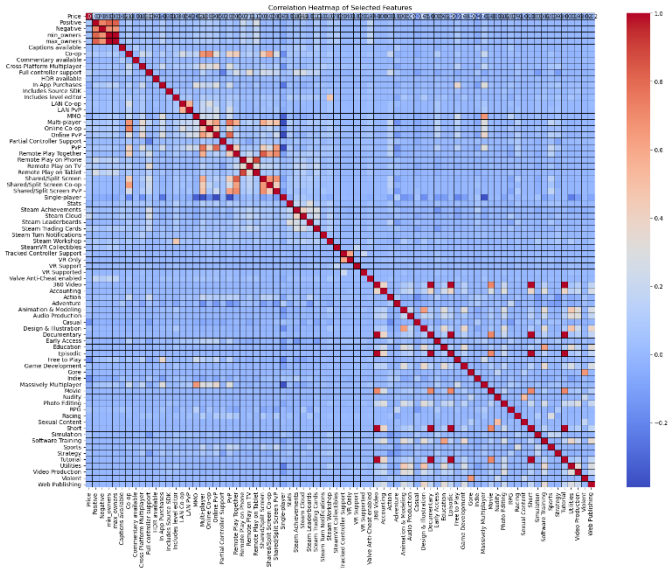


*Fig 17. Heatmap Correlation Diagram*

The first evaluation is done by calculating the average between each cluster based on their similar characteristics [23]. From the calculations that have been carried out, the silhouette average calculation value is 0.345. The silhouette coefficient's result should be nearest to +1, but the evaluation result in this research is 0,345. This is because the high-dimensional data or the cluster was not well separated, and the silhouette coefficient gave a misleading result.

The second evaluation is done by calculating as a point the number of data according to the cluster to the total data owned [24]. After the data is grouped using clustering based on similar characteristics, the data will then be labeled to facilitate the calculation process. The calculation results of the purity method are a precision of 0.734, a recall of 0.270, and an F1 Score of 0.381. The precision result of 0.734 indicates that the cluster contains the appropriate data, but some data needs to be in the right place. The recall result of 0.270 suggests that data that should have been entered into the cluster was missed during the clustering process. The F1 score result of 0.381 indicates a balance between precision and recall because the precision result is higher than recall. The interpretation results of the purity method can be seen from the high accuracy and low recall, which means that the clustering results are accurate, but many missing data that needs to be entered.

The results of implementing the K-Means algorithm can be visualized to determine market segmentation based on user preferences.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

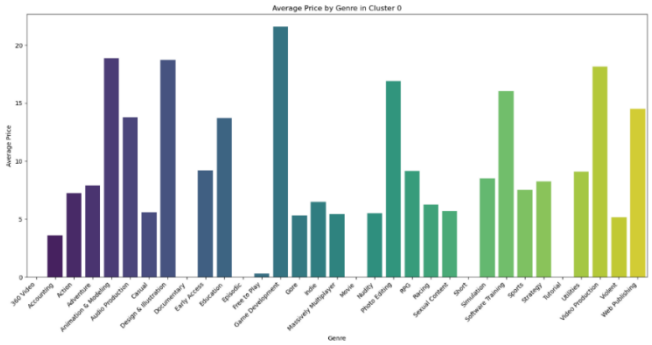


*Fig 18. Visualization Price and Genre Cluster 0*

The visualization explains the comparison between genres and the average price in each cluster. The first visualization in cluster 0 explains that the highest average price is the game development genre with a price of 25 dollars ($).

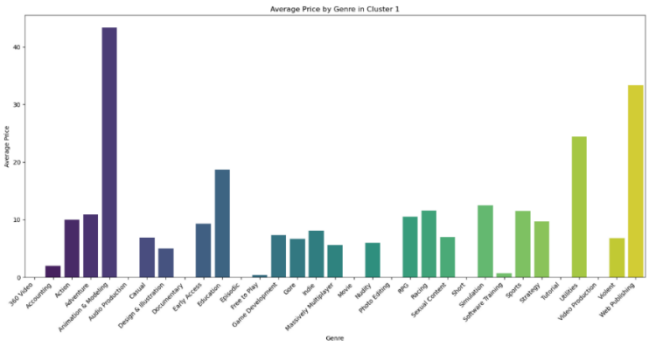


*Fig 19. Visualization Price and Genre Cluster 1*

The second visualization in cluster 1 explains that the highest average price is animation & modeling with a price of 45 dollars ($).

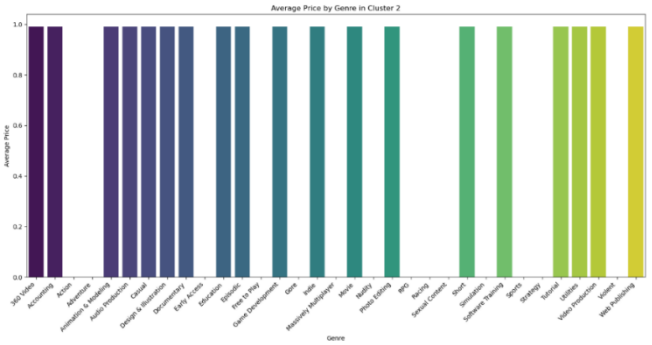


*Fig 20. Visualization Price and Genre Cluster 2*

The third visualization in cluster 2 explains that the average price in this cluster is 1 dollar ($). From the three visualizations of the comparison between the average price and genre, it is known that the Steam platform also sells applications with the highest prices, such as game development, animation, and modelling, followed by games such as action, adventure, RPG, and others.

The next visualization explains the comparison between genres and the average user in each cluster.

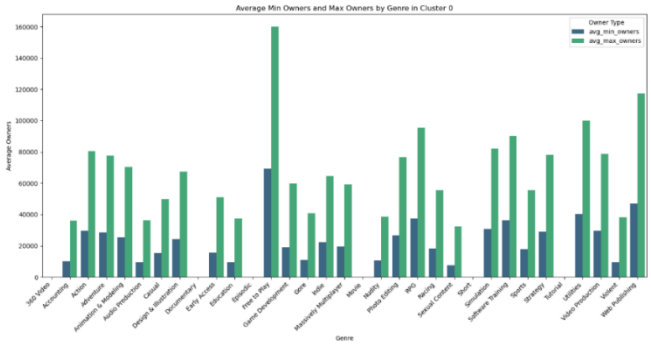


*Fig 21. Visualization Estimate Owner and Genre Cluster 0*

The first visualization explains the comparison of the average estimate owner and genre, which shows that the genre with the largest average number of max owners is free to play, with a total of 160,000 and an average number of min owners of 70,000. Other genres, such as action, adventure, RPG, simulation, and others, have a fairly large average number of max owners.

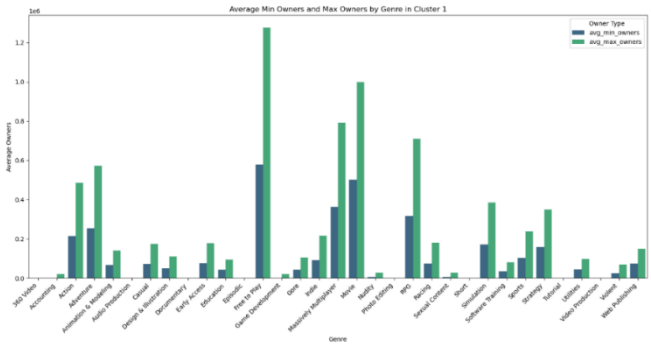


*Fig 22. Visualization Estimate Owner and Genre Cluster 1*

The second visualization in cluster 1 shows that the free-to-play genre has an average number of max owners of 1.4 and an average number of min owners of 0.6.

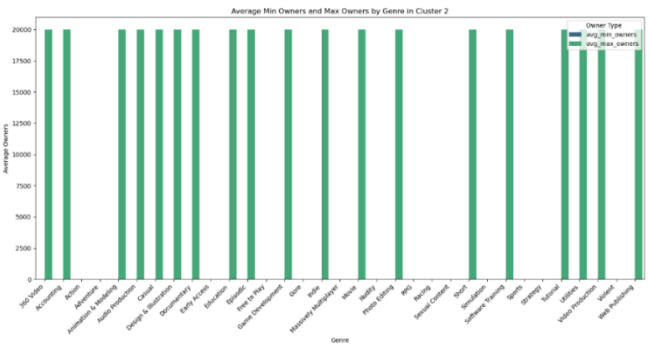


*Fig 23. Visualization Estimate Owner and Genre Cluster 2*

The third visualization in cluster 2 shows many genres that have an average max owner of 20,000. From the visualization results, it is known that the visualization results of the comparison of genres and estimate owners in cluster 0 are clusters with the highest average min owner and max owner dominated by free-to-play game sales, then in cluster 1 is a cluster with the lowest average min owner and max owner

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

which is also dominated by free-to-play game sales, and in cluster 2 is a cluster with a medium average min owner and max owner dominated by application sales. From the visualization results of comparing genres and prices, it is known that cluster 0 is a cluster with a medium average price dominated by application sales followed by game sales. Cluster 1 is a cluster with the highest average price dominated by application sales, and cluster 2 is a cluster with the lowest average price with a comparison of game and application sales equivalent.

The data visualization results between Price comparison with genres differ from the comparison of estimate owners with categories or genres. The comparison of estimate owners with genres is dominated by free-to-play because everyone can use it for free, then the comparison of Prices with genres is dominated by simulation games or applications and developers who have high selling prices so that only a few users have the game or application. Games or applications that can be played for free or free-to-play usually have paid content, which makes users have to spend money to get the goods or content even though the game or application can be played for free. Games or applications with a high selling price also have paid content, but users prefer to use free games or applications only to pay once to get paid content.

## IV. CONCLUSION

The results of this study showed that there was a difference between the evaluation results using the silhouette coefficient and purity. The silhouette coefficient is less suitable for the type of game data with a label with a silhouette average calculation value of 0.345 because the high-dimensional data or the cluster was not well separated, and the silhouette coefficient gave a misleading result. Then, the results of the calculation using purity get pretty good results, namely, a precision value of 0.734, a recall value of 0.270 and an F1 score of 0.381, which shows that the use of this purity is very suitable for evaluating the type of data that has a label. From the visualization results obtained, we can see that the clustering results in the study are divided into 3 clusters; the visualization results between estimate owner and genre show that the data is divided based on the number of users, which are divided into many, medium and few.

The visualization results between price and genre show that the data is divided based on price, divided into high, medium, and low. By knowing user preferences from the visualization results that have been done, it can be seen that sales with the highest estimate owner are sales of free-to-play games, and sales with the highest price are sales of simulation applications. From the clustering results, it can help users determine what type of game can be played based on recommendations from the clustering results. For game developers, it can also help determine what kind of game can be developed based on recommendations from the clustering results to increase sales profits.

The results of the purity method show high precision values, low recall values , and low f1 score values, which means that the results of k-means clustering show a high level of accuracy, but many data that should be in the cluster are lost or not used, causing low recall and f1 score values. The silhouette coefficient results are similar to the f1 score results from the purity method. Therefore, according to the researcher, the evaluation process using the silhouette coefficient has good performance results even though it produces low values because there is data in the missing cluster. The missing data can be caused because the process during data cleaning has a lot of empty data but is not used because the data deletion process is not carried out and only the filter process is carried out. Therefore, for further research development on clustering in this game sales data, the data cleaning process can be carried out manually or using different data cleaning stages to ensure the data can be used properly. The use of game sales data on other platforms, such as the Epic Games Store or other game sales platforms, can produce different results in predicting the types of games that are most in demand.

## REFERENCES


[1] R. Machfiroh, A. Rahmansyah, and A. Budiman, "The Effect of Massively Multiplayer Online Game on Player Behaviour," in *Journal of Physics: Conference Series*, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1764/1/012081.

[2] D. Bai, L. Chen, Z. Shang, Y. Wang, and G. Guan, "E-sports Industry, Video Game Industry and Economic Growth: An Empirical Research in China," 2022. doi: 10.2139/ssrn.4074000.

[3] C. H. Cheng and S. F. Huang, "A novel clustering-based purity and distance imputation for handling medical data with missing values," *Soft comput*, vol. 25, no. 17, pp. 11781–11801, Sep. 2021, doi: 10.1007/s00500-021-05947-3.

[4] S. Wijaya, N. Nur Setyo, and W. Nur Azizah, "Potential Analysis And Supervision Of VAT On The Utilization Of Digital Contents (Case Study: Steam Platform," *Dinasti International Journal Of Digital Business Management*, vol. 1, no. 3, 2020, doi: 10.31933/dijdbm.v1i3.238.

[5] Z. Wang, V. Chang, and G. Horvath, "Explaining and Predicting Helpfulness and Funniness of Online Reviews on the Steam Platform," *Journal of Global Information Management*, vol. 29, no. 6, 2021, doi: 10.4018/JGIM.20211101.oa16.

[6] I. Busurkina, V. Karpenko, E. Tulubenskaya, and D. Bulygin, "Game Experience Evaluation. A Study of Game Reviews on the Steam Platform," in *Digital Transformation and Global Society*, D. A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova, I. Musabirov, and S. Pashakhin, Eds., Cham: Springer International Publishing, 2022, pp. 117–127. doi: 10.1007/978-3-030-93715-7.

[7] K. P. Sinaga and M. S. Yang, "Unsupervised K-means clustering algorithm," *IEEE Access*, vol. 8, pp. 80716–80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

[8] B. Zhang, L. Wang, and Y. Li, "Precision Marketing Method of E-Commerce Platform Based on Clustering Algorithm," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/5538677.

[9] A. Moubayed, M. Injadat, A. Shami, and H. Lutfiyya, "Student Engagement Level in an e-Learning Environment: Clustering Using K-means," *American Journal of Distance Education*, vol. 34, no. 2, pp. 137–156, Apr. 2020, doi: 10.1080/08923647.2020.1696140.

[10] L. Pamungkas, N. A. Dewi, and N. A. Putri, "Classification of Student Grade Data Using the K-Means Clustering Method," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 1, pp. 86–91, Feb. 2024, doi: 10.32736/sisfokom.v13i1.1983.

[11] S. Ariska, D. Puspita, and I. Anggraini, "Comparison Of K-Means and K-Medoids Algorithm for Clustering Data UMKM in Pagar Alam City," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 13, no. 2, pp. 193–199, Jun. 2024, doi: 10.32736/sisfokom.v13i2.2090.

[12] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.

[13] S. D. K. Wardani, A. S. Ariyanto, M. Umroh, and D. Rolliawati, "Comparison of K-Means, Db Scanner & Hierarchical Clustering Method


p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024

Results for Market Segmentation Analysis," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 2, p. 191, Sep. 2023, doi: 10.26798/jiko.v7i2.796.

[14] D. P. Agustino, I. G. B. A. Budaya, I. G. Harsemadi, I. K. Dharmendra, and I. M. S. A. Pande, "Comparison of the DBSCAN Algorithm and Affinity Propagation on Business Incubator Tenant Customer Segmentation," *Jurnal Sisfokom (Sistem Informasi dan Komputer)*, vol. 12, no. 2, pp. 315–321, Jul. 2023, doi: 10.32736/sisfokom.v12i2.1682.

[15] M. Sarkar, ✉ Aisharyja, R. Puja, and F. R. Chowdhury, "Optimizing Marketing Strategies with RFM Method and K-Means Clustering-Based AI Customer Segmentation Analysis," *Journal of Business and Management Studies*, 2024, doi: 10.32996/jbms.2024.6.2.5.

[16] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[17] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire, "Understanding and improving the quality and reproducibility of Jupyter notebooks," *Empir Softw Eng*, vol. 26, no. 4, Jul. 2021, doi: 10.1007/s10664-021-09961-9.

[18] S. Chandel, C. B. Clement, G. Serrato, and N. Sundaresan, "Training and Evaluating a Jupyter Notebook Data Science Assistant," 2020. [Online]. Available: www.aaai.org

[19] C. H. Cheng and S. F. Huang, "A novel clustering-based purity and distance imputation for handling medical data with missing values," *Soft comput*, vol. 25, no. 17, pp. 11781–11801, Sep. 2021, doi: 10.1007/s00500-021-05947-3.

[20] D. Solunke, G. Deshmukh, S. Wagh, A. Agrawal, and I. Priyadarshini, "Unlocking The Genres: Multilabel Anime Genre Predicition," *International Research Journal of Modernization in Engineering Technology and Science* , vol. 05, no. 04, 2023, [Online]. Available: www.irjmets.com

[21] F. Aldi, F. Hadi, N. A. Rahmi, and S. Defit, "Standardscaler's Potential In Enhancing Breast Cancer Accuracy Using Machine Learning," *Journal of Applied Engineering and Technological Science*, vol. 5, no. 1, pp. 401–413, 2023, doi: 10.37385/jaets.v5i1.3080.

[22] H. Humaira and R. Rasyidah, "Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm," in *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*, European Alliance for Innovation n.o., Mar. 2020. doi: 10.4108/eai.24-1-2018.2292388.

[23] D.-T. Dinh, T. Fujinami, and V.-N. Huynh, "Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient," in *Knowledge and Systems Sciences*, J. Chen, V. N. Huynh, G.-N. Nguyen, and X. Tang, Eds., in Communications in Computer and Information Science, vol. 1103. Singapore: Springer Singapore, 2019, pp. 1–17. doi: 10.1007/978-981-15-1209-4.

[24] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris, and S. Mirjalili, "Multi-verse optimizer: Theory, literature review, and application in data clustering," in *Studies in Computational Intelligence*, vol. 811, Springer Verlag, 2020, pp. 123–141. doi: 10.1007/978-3-030-12127-3_8.

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v13i3.2214, Copyright ©2024
Submitted : July 11, 2024, Revised : August 25, 2024, Accepted : September 6, 2024, Published : November 15, 2024