IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.x, No.x, July xxxx, pp. 1~5
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: 10.22146/ijccs.xxxx

Optimizing Clustering Models Using Principle Component Analysis for Car Customers

Agnes Riska Savira*[1], Amril Mutoi Siregar[2], Dwi Sulistya Kusumaningrum[3], Yana Cahyana[4]
[1,2,3,4] Department of Computer Science University of Buana Perjuangan Karawang, Indonesia
e-mail: ***[1]if20.agnessavira@mhs.ubpkarawang.ac.id**, [2]amrilmutoi@ubpkarawang.ac.id, [3]dwi.sulistya@ubpkarawang.ac.id , [4]yana.cahyana@ubpkarawang.ac.id

Abstrak

Dalam dunia bisnis yang kompetitif, perusahaan secara strategis memanfaatkan data pelanggan untuk mencapai tujuan, sehingga memerlukan pemahaman komprehensif tentang beragam sifat, perilaku, dan kebutuhan pelanggan. Segmentasi pelanggan, sebuah strategi penting, memerlukan pengelompokan individu berdasarkan berbagai karakteristik. Algoritme K-Means yang banyak digunakan konektivitas pengelompokan data pelanggan karena kemudahan implementasinya dalam Pembelajaran Mesin. Namun, tantangan muncul pada data berdimensi tinggi, sehingga mendorong perlunya pengurangan dimensi. Analisis Komponen Utama (PCA) muncul sebagai metode yang efektif untuk komunikasi data sekaligus meminimalkan kehilangan informasi. Penelitian sebelumnya menekankan keberhasilan PCA dalam meningkatkan efisiensi analisis dan pengelompokan. Penelitian ini berkontribusi dengan mengintegrasikan PCA ke dalam clustering K-Means untuk menganalisis segmen pelanggan di sebuah perusahaan mobil. Hal ini memberdayakan perusahaan untuk menarik pelanggan baru, menerapkan pemasaran yang ditargetkan, memahami hubungan pelanggan-perusahaan, dan meningkatkan profitabilitas yang diharapkan. PCA, yang mempertahankan 75% variasi dengan 3 komponen utama, mendahului implementasi K-Means setelah normalisasi. Evaluasi menggunakan Metode Elbow dan Silhouette Score mengidentifikasi delapan cluster yang optimal. Model K-Means pasca-PCA dengan pemilihan cluster optimal menghasilkan Skor Silhouette sebesar 0,7789.

Kata kunci*— K-Means, PCA, Segmentasi Pelanggan, Pembelajaran Mesin*

Abstract

In the competitive business world, companies strategically utilize customer data to achieve goals, requiring a comprehensive understanding of various customer traits, behaviors and needs. Customer segmentation, an important strategy, requires grouping individuals based on various characteristics. The K-Means algorithm is widely used for customer data grouping connectivity because of its ease of implementation in Machine Learning. However, challenges arise in high-dimensional data, prompting the need for dimensionality reduction. Principal Component Analysis (PCA) is emerging as an effective method for data communication while minimizing information loss. Previous research emphasizes the success of PCA in improving analysis and clustering efficiency. This research contributes by integrating PCA into K-Means clustering to analyze customer segments in a car company. This empowers companies to attract new customers, implement targeted marketing, understand customer-company relationships, and increase expected profitability. PCA, which preserves 75% of the variation with 3 principal components, precedes the implementation of K-Means after normalization. Evaluation using the Elbow and Silhouette Score Method identified eight optimal clusters. The post-PCA K-Means model with optimal cluster selection produces a Silhouette Score of 0.7789.

Keywords— *K-Means, PCA, Customer Segmentations, Machine Learning*

Received June 1st,2012; Revised June 25th, 2012; Accepted July 10th, 2012

1. INTRODUCTION

In facing product competition, companies need to utilize customer data information to achieve predetermined goals. Companies need to have a deeper understanding of the characteristics, behavior and needs of diverse customers. One approach that can be taken is to segment customers. Customer segmentation is the process of dividing a customer base or data into groups of individuals who have similar characteristics or behavior [1] .

In the context of Machine Learning, there is an algorithm that is very useful for grouping customer data, namely K-Means. This algorithm is commonly used in the grouping process because of its ease of implementation [2]. However, some clustering algorithms such as K-Means often face problems when applied to data with high dimensions or features. Some of the problems that arise include decreased classification accuracy, poor cluster quality, and long computing times. To maintain optimal algorithm performance, one approach that can be taken is to carry out dimension reduction. Dimensionality reduction can be done through two methods, namely feature selection and feature extraction. PCA is a method of simplifying a data set by carrying out linear transformations to form a new coordinate system with minimum variance. PCA is used to reduce data dimensions with a very small risk of losing information [3].

Previous research entitled "Application of Principal Component Analysis (PCA) for Dimensional Reduction in the Clustering Process of Agricultural Production Data in Bojonegoro Regency" utilized a combination of the K-Means algorithm and PCA techniques. The research results show that the application of PCA in reducing the dimensions of agricultural production data in Bojonegoro Regency can increase the efficiency of analysis and clustering results. By reducing the dimensions of the data, analysis can be carried out more efficiently, and grouping results become more accurate and easier to interpret [4] . Another research conducted by [5], entitled "Business Marketing Analysis with Data Science: Customer Personality Segmentation based on the K-Means Clustering Algorithm" also combines K-Means and PCA on customer data to segment customer characteristics.

In this study, we detail the steps that include data collection, data analysis, data preprocessing, algorithm implementation, result evaluation, and overall analysis. We use features such as gender, marital status, age, graduation status, profession, work experience, spending score, and family size. By undertaking these stages, our aim is to provide deep insights into customer characteristics and behaviors, as well as to establish a strong foundation for the development of more targeted and effective marketing strategies.

Reviewing several previous studies, the PCA method is useful in extracting significant information. PCA makes an important contribution by reducing the dimensions of complex data to smaller ones, allowing K-Means to operate efficiently on new data sets [6] . Based on the problems that have been described, this research makes a contribution by applying the Principal Component Analysis (PCA) technique to optimize the K-Means clustering method. By doing this, companies can attract new customers, implement appropriate marketing strategies, understand the relationship between customers and the company, and increase the company's expected profitability.

2. METHODS

This section will explain the flow of the research methods carried out. The following figure 1 is a diagram of the method flow in the research.

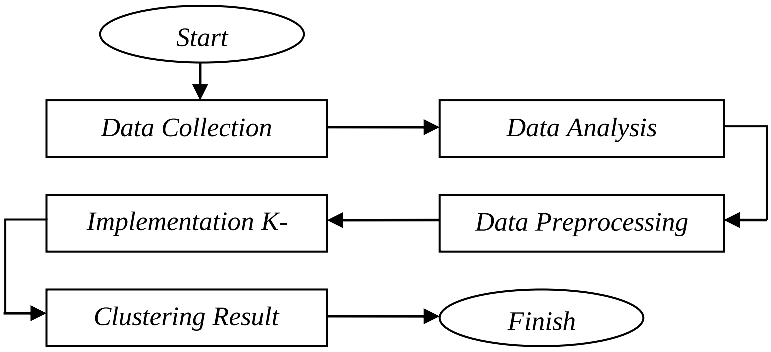


Figure 1 Stages of Research Method Flow

Research starts with data collection, gathering relevant information for analysis. Following this, data analysis identifies patterns and trends. Subsequently, data preprocessing is essential to enhance data quality and prepare it for further analysis. We will elaborate on each preprocessing method, including noise removal, handling missing values, and data transformation. These techniques aim to ensure clean, accurate, and consistent data for more valid and reliable results. The K-Means algorithm is then implemented as the primary step in clustering analysis, grouping data with similar characteristics. Evaluation of clustering results provides deeper insights.

2.1 Data Collection

At the data collecting stage, data related to the research will be used. The dataset used comes from car company customer data. This data was accessed on September 6, 2023 at 19:00 WIB via the Kaggle website at https://www.kaggle.com/datasets/vetrirah/customer/data. This dataset consists of 10695 rows and has 11 attributes.

2. 2 Data Analysis

In the second stage, data understanding or initial analysis is carried out to identify the data used, such as the attributes and variables contained in the dataset [7]. Data analysis is also part of this data processing process, with the aim of extracting important information contained in the data to provide deep insights into customer characteristics and behavior, as well as for developing more targeted and effective marketing strategies.

2. 3 Data Preprocessing

The initial process in data processing in exploratory data analysis is called pre-processing, which aims to produce data in an appropriate format and ready for use at the next stage. The purpose of pre-processing is to make it easier to use data when modelling [8].

Data preprocessing starts with an 11-feature raw dataset. Focusing on clustering methods, 3 irrelevant features are eliminated initially, including those serving as segmentation targets and unique values unrelated to modeling. Subsequently, missing values and duplicates are managed for data cleanliness. Addressing outliers follows to enhance data quality, succeeded by data transformation and Principal Component Analysis (PCA) for dimension reduction. Below illustrates two stages in preprocessing.

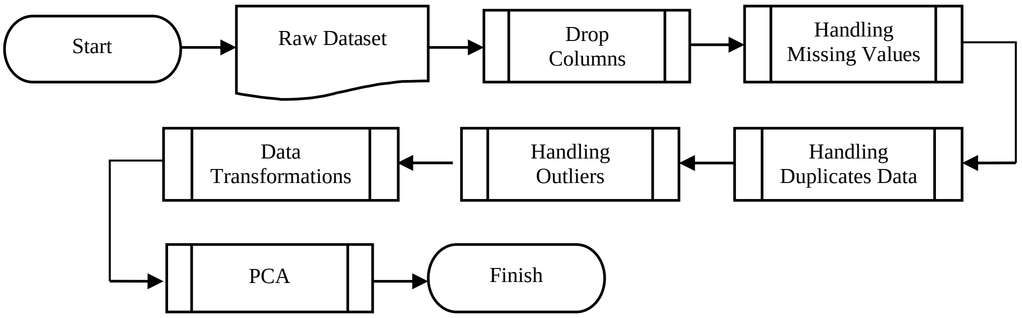


Figure 2 Preprocessing Stages

2. 4 Implementation of K-Means Clustering

The subsequent step involves implementing the K-Means Clustering algorithm, an unsupervised learning method, to create a machine learning model. Here, clustering results are assessed by segmenting customers into specific groups. Customer segmentation begins with data preparation, followed by selecting data for clustering and determining the number of clusters using the elbow method. Segmentation entails dividing the dataset into clusters based on similar attributes, placing each data point at the cluster center according to its characteristics [9]. This step aims to identify similar groups within the data, simplifying analysis and interpretation.

2. 5 Clustering Result

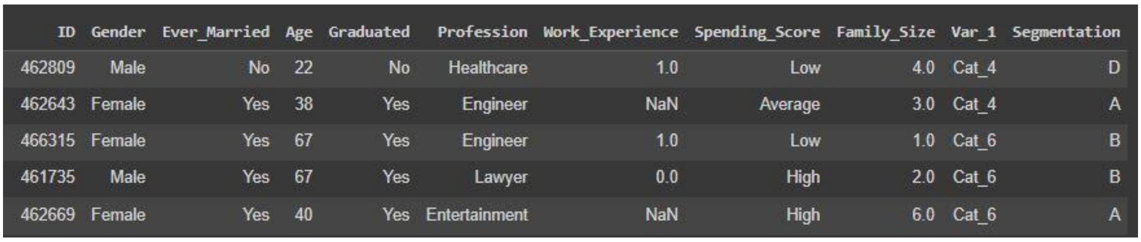
Through the cluster grouping stage, customer segmentation results can be identified using the K-Means method. The output comprises groups of customers contributing to sales, segmented based on their purchasing behavior patterns. The analysis delves into various aspects such as the proportion of customer clusters, age distribution, purchasing power, professions, marital status, and detailed characteristics of each cluster. This comprehensive analysis offers insights into demographic diversity, spending behaviors, and professional distributions within different customer groups.

3. RESULTS AND DISCUSSION

This chapter is an explanation and analysis of the steps outlined in the research method.

3.1 Data Collection

The dataset used in this research is secondary data taken from the Kaggle website uploaded by Vetrirah. This dataset was downloaded on September 6, 2023 at 19:00 WIB. The data taken consists of 10695 rows and has 11 attributes, namely ID, Gender, Ever_Married, Age, Graduated, Profession, Work_Experience, Spending_Score, Family_Size, Var_1, Segmentation. Figure 3 is an example of the data used in this research.



ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A

Figure 3 Data Collection Sample

3. 2 Data Analysis

At this stage, data analysis will be carried out using visualization to facilitate understanding of the data set, determining analysis features, identifying data correlations, and other aspects.

3. 2. 1 Age and Purchasing Power

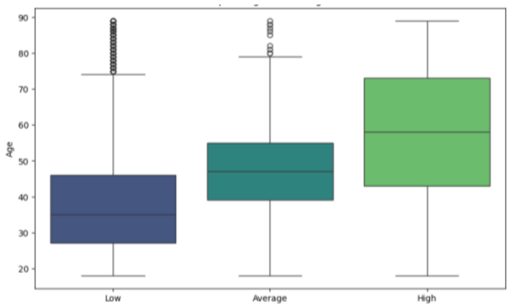


Figure 4 Visualization of Age and Purchasing Power

In Figure 4, it can be seen that the higher the age, the higher the purchasing power. In the low purchasing power category, the average age is between 25 - 45 years, then for the medium purchasing power category, it is 40-45, while in the high purchasing power category the age group is dominated by 45 - 70 years.

3. 2. 1 Purchasing Power and Family Size

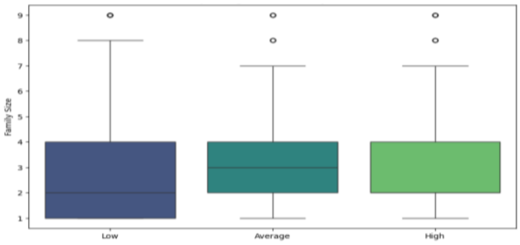


Figure 5 Visualization of Purchasing Power and Family Size

Figure 5 shows that customers who have a low level of purchasing power usually have around 1-4 family members, while customers who have medium and high spending have an average of 2-4 family members.

3. 3 Data Preprocessing

Preprocessing of data begins with the original dataset containing 11 features. With a focus on clustering techniques, the first step involves eliminating 3 features irrelevant to clustering. These include attributes for segmentation and unique values not contributing to modeling. Following feature removal, data cleaning addresses missing values and duplicates. Subsequently, outlier detection enhances data quality. Lastly, data transformation, such as scaling or format adjustment, is conducted, along with Principal Component Analysis (PCA) to reduce dimensions without losing vital information, enhancing the efficiency of the clustering process.

3. 3. 1 Drop Columns

Drop Columns is a process in data processing that involves deleting certain columns from the dataset [10]. In this research, the columns that were dropped were "ID", "Var_1", and "Segmentation", these three columns are not relevant to the clustering method. Columns with irrelevant attributes may lead to overfitting in the model and adding unnecessary complexity to it. After irrelevant columns are removed, there are 8 features used for this research.

3. 3. 2 Handling Missing Values

Handling missing values is a process in data processing that includes strategies for overcoming or managing, missing or empty values in a dataset. Handling missing values is critical in machine learning because the presence of missing values can affect the accuracy and validity of analysis results [11]. In this study there were 1799 data identified as missing values, these data will then be deleted (Drop). If not removed, missing values may introduce bias into the outcomes of the research. The number of data after completing this stage was 8896 data.

3. 3. 3 Handling Duplicates Data

Handling duplicate data is a process in data processing that involves identifying and handling duplicate data in a dataset. Handling duplicates is important to ensure data integrity, prevent bias in analysis, and increase the accuracy of interpreted results [12]. After checking, it was found that 1859 rows of data were identified as duplicate data. Therefore, it is necessary to drop duplicate data. This step is important because duplicate data can "mislead" algorithms into thinking there is more data than there actually is, potentially leading to inaccurate final results.

3. 3. 4 Handling Outliers

Handling outliers is a process that involves identifying, evaluating, and handling extreme values or outliers in a dataset. Outliers are observations that are statistically significant different from the bulk of the data [13]. Checking for outliers in this study uses a boxplot from the matplotlib library which is shown in Figure 6.

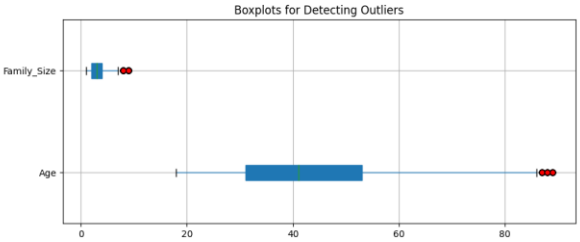


Figure 6 Boxplot Of Checking Outliers In Column Numbers

Based on Figure 6, it is shown that each numerical column has different outliers. Outliers in each column are marked with red dots that are outside the range of that column. After that, outliers were dropped using the Interquartile Range (IQR) method. The results of this stage can be seen in Figure 7.

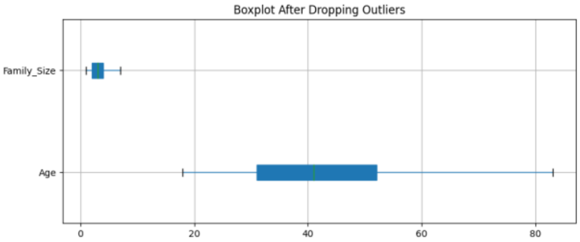


Figure 7 Boxplot after Dropping Outliers

3. 3. 5 Data Transformation

Data transformation is the process of manipulating or changing the structure or characteristics of data from its initial format or distribution into a form that is more suitable for certain analysis or modeling [14]. In this research, data transformation was carried out by converting categorical data using label encoding and ordinal encoding methods. Then perform feature scaling on the numerical data using the MinMaxScaler method to normalize the data to be in the range 0 – 1.

3. 3. 6 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that aims to convert data that has many dimensions into a representation that has lower dimensions while retaining as much of the original variance as possible [15]. Figure 8 is the result of this method for finding optimal principal components.

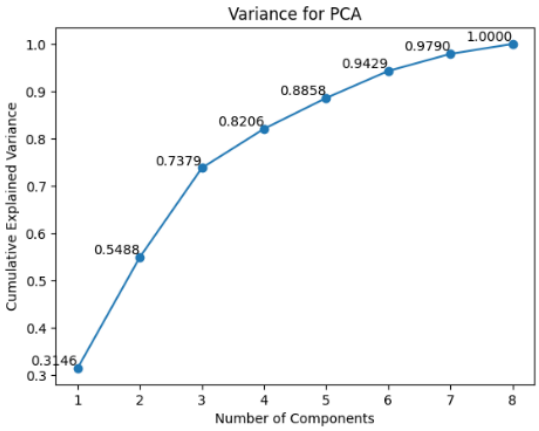


Figure 8 Find the Optimal Principal Component

To reduce the dimension of the data, PCA in this study uses 3 main components (Principal Components) or n_components = 3. By selecting the 3 main components that have the most influence, we can change the initial data into a new feature container that is simpler, but still able to represent large data variations, and retains about 75% of the total data variations.

3. 4 Implementation of K-Means Clustering

In this section, the researcher explains the clustering results by applying the K-Means algorithm. This algorithm has the ability to identify patterns in unlabeled data sets by grouping data points based on similar attributes in the dataset [16]. Previously, in the preprocessing process, data standardization was carried out using MinMaxScaler to normalize the data into the 0-1 range. Next, we utilize PCA to identify crucial variables in the dataset by choosing the number of components that explain around 75% of the data variation.

Before employing the K-Means algorithm on PCA-transformed data, it is crucial to conduct elbow method and silhouette score analyses to ascertain the ideal number of clusters. The elbow method involves iteratively applying K-Means with varying cluster numbers, measuring the inertia value or within-cluster sum of squares (WCSS) for each iteration. WCSS assesses how closely data points in a cluster approach the centroid, and plotting WCSS against cluster numbers can produce a curve resembling an elbow [17]. The optimal number of clusters is determined at the point where the decrease in WCSS becomes insignificant, often observed around 8 clusters. The outcomes of this stage are depicted in Figure 9.

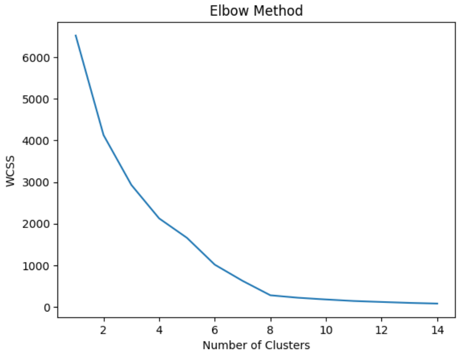


Figure 9 Elbow Method Result

Additionally, silhouette scores can provide additional insights. Silhouette score measures the extent to which a data point corresponds to the assigned cluster compared to other clusters. Silhouette score values range from -1 to 1, where higher values indicate that objects are better placed within their cluster and more separated from neighboring clusters [18]. The highest silhouette score value was obtained with 8 clusters. This can be observed in Figure 10.

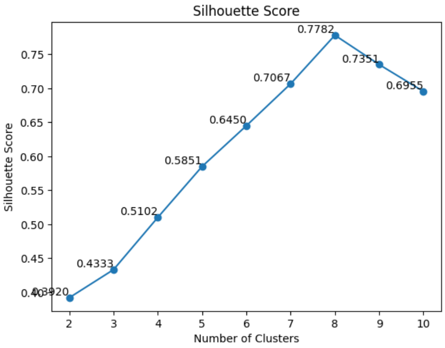
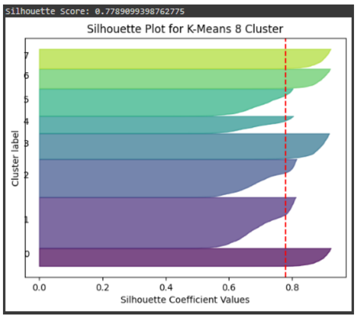


Figure 10 Comparison of Silhouette Score Values for Each Cluster

Based on the visualization of the two methods used to determine the optimal k value in the K-Means algorithm, the elbow and silhouette score methods produce the same optimal k value, namely k = 8 or 8 clusters. So the researchers decided to use k = 8. The silhouette score using 8 clusters produced the highest score, namely 0.7789. The following Figure 11 is a plot of the silhouette score values using 8 clusters.



```
Silhouette Score: 0.7789099398762775
```

Silhouette Plot for K-Means 8 Cluster
Cluster label
Silhouette Coefficient Values
7
6
5
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8

Figure 11 Silhouette Score Plot for 8 Clusters

3. 5 Clustering Result

Next, the results of customer clustering will be analyzed into several parts, customer segmentation results can be identified using the K-Means method. The output of this grouping is in the form of groups of customers who contribute to sales, divided into certain segments according to their purchasing behavior patterns.

3. 5. 1 Proportion of Customer Clusters

After analysis, it was found that Cluster 0 was the largest cluster (24.1%) or around 1/4 of all customers. Followed by Cluster 2 at 17.7%, Cluster 6 at 12.3%, and Cluster 4 at 12.2%. Meanwhile, other clusters show smaller percentages, Cluster 5 at 6.9%, Cluster 1 at 8.7%, Cluster 3 at 8.7%, and Cluster 7 at 9.3%. These result can be seen in Figure 12.

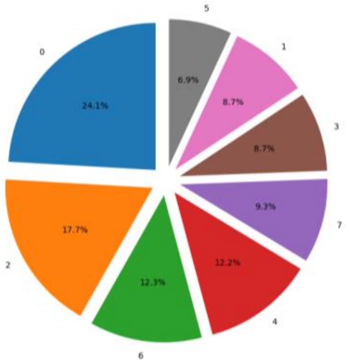


Figure 12 Proportion of Customer Clusters

3. 5. 2 Age Distribution

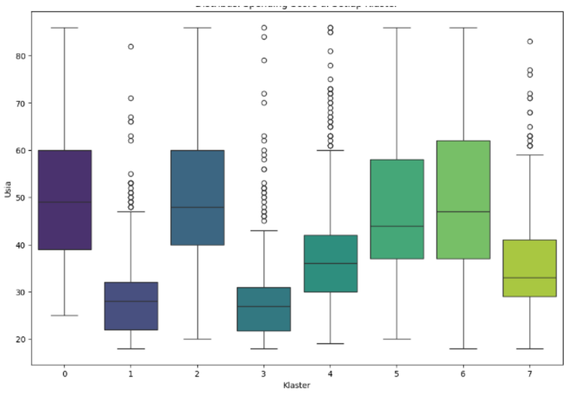


Figure 13 Age Distribution of Each Cluster

Figure 13 shows eight groups based on age range. Clusters 1 and 3 have the youngest ages, while Cluster 6 has the oldest ages. Clusters 0, 2, 4, 5, and 7 have almost the same characteristics in the adult age range.

3. 5. 3 Purchasing Power

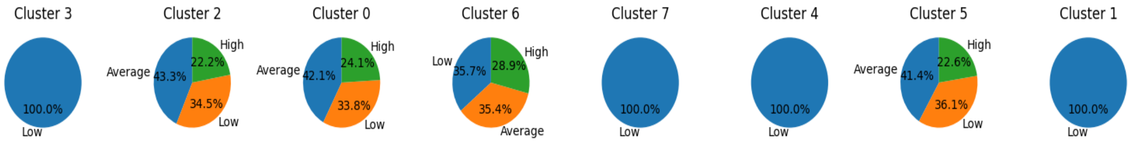


Figure 14 Percentage of Purchasing Power Each Cluster

Figure 14 reveals that four clusters (1, 3, 4, and 7) exhibit homogeneous purchasing power, indicating that all members within these clusters possess low purchasing power. The other four clusters (0, 2, 5, and 6) show variations in purchasing power, with varying proportions in each category.

3. 5. 4 Profession

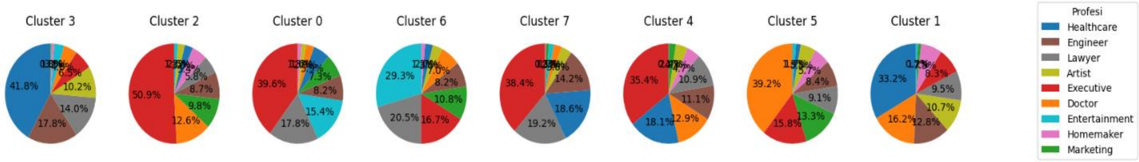


Figure 15 Distribution Of Professions For Each Cluster

Figure 15 contains eight clusters, with Cluster 3, Cluster 7, and Cluster 1 predominantly featuring professions in the health sector (Healthcare), including professionals like engineers and lawyers. On the other hand, Cluster 2 and Cluster 5 are characterized by professions in the executive (Executive), Doctor, and Marketing fields. Professions in the fields of Executives, Lawyers, and Entertainment primarily occupy Cluster 0 and Cluster 6. Interestingly, Cluster 4, even though it places the Executive profession at the top, the proportion is not much different from the Health and Doctor professions. This analysis provides an overview of the distribution of professions within each cluster, with differences in focus and proportions that are interesting to note.

3. 5. 5 Marital Status

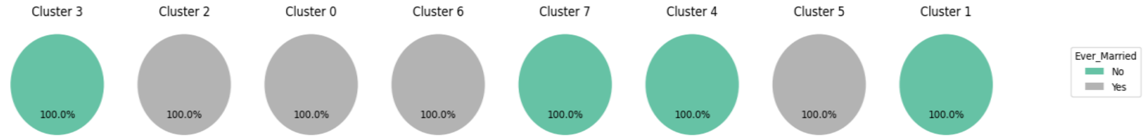


Figure 16 Distribution of Marital Status for Each Cluster

Figure 16 shows that each cluster falls into the Married/Not Married category. Four clusters, namely 1, 3, 4, and 7, are included in the Not Married category. Meanwhile, the other four clusters, namely 0, 2, 5, and 6, are included in the Married category.

3. 5. 6 Proportion of Customer Clusters

Next, an analysis is carried out per cluster based on the results of the previous analysis.

• Cluster 0 is characterized by individuals aged 39-60 with spending scores that tend to be low. The majority are single, and this cluster is dominated by executives, lawyers and entertainment professionals. This group appears to be composed of cautious shoppers, perhaps individuals who prioritize financial stability or have spending habits related to their profession.
• Cluster 1, which mostly consists of individuals aged 22-32, has generally low spending scores, and the majority are unmarried. Professions in the health and arts sectors appear to dominate. This group may reflect individuals who are watching their spending, perhaps driven by early career stages or artistic pursuits, with possibly higher levels of singleness.
• Cluster 2, with an age range of 40-60, shows diverse spending patterns, including 34.54% low, 43.26% medium, and 22.2% high. The majority consist of married individuals, with a predominance of executives, doctors, and marketers. This group shows a mix of spending behavior, indicating a variety of financial priorities, perhaps with a balance between responsible spending and greater financial capability.
• Cluster 3 mainly consists of individuals aged 22-31 with low spending scores and the majority are married. Health professionals dominate this cluster. Individuals in this group may be characterized by a conservative approach to spending, a focus on basic needs, and may reflect the financial stability associated with the medical profession.
• Cluster 4, with individuals aged 30-43, shows uniformly low spending scores, with the majority unmarried. Executive dominance with similar professions in the health sector and doctors. This group likely consists of career-focused individuals, perhaps saving more for personal goals or professional investments, given low spending tendencies.
• Cluster 5, dominated by individuals aged 38-58, displays varied spending patterns, with a prevalence of married individuals. Explained by doctors, executives, and marketing professionals. This group may include financially diverse households, a mix of conservative and moderate spenders, reflecting a range of financial priorities and responsibilities.
• Cluster 6, which includes individuals aged 38-62, displays varying spending patterns, including 35.74% low, 35.38% medium, and 28.88% high. The majority are married, and this cluster is dominated by executives, lawyers, and entertainment professionals. These diverse spending behaviors may indicate a mix of financial goals or priorities within this group of experienced professionals.
• Lastly, Cluster 7, which consists of individuals aged 29-41 with low expenditure scores and the majority of whom are unmarried, is dominated by health professionals. This group most likely consists of individuals who prioritize financial prudence, perhaps due to early or mid-career stages, and have not yet taken on the additional financial responsibilities associated with marriage.

Clustering results provide an overview of the diversity of demographic characteristics, spending patterns and professions in different customer groups. This analysis can provide companies with in-depth insights to better understand their target market. With a deeper understanding of consumer preferences and behavior in each group, companies can develop more targeted and effective marketing strategies. This allows companies to tailor their products, services and marketing communications to the specific needs and preferences of each customer group, thereby increasing the likelihood of success in reaching and satisfying diverse market needs.

4. CONCLUSIONS

In this study, we aim to utilize K-Means and Principal Component Analysis (PCA) methods for customer segmentation. The dataset undergoes preprocessing steps before entering the PCA and Clustering stages using K-Means. Data is collected from a secondary source, specifically a dataset from Kaggle, containing 10695 rows and 11 attributes. The analysis involves several stages, beginning with data exploration through visualizations to understand relationships between variables such as age, purchasing power, and family size. Subsequently, data preprocessing is conducted to ensure data quality, including removing irrelevant columns, handling missing values, and addressing duplicates. This ensures analysis accuracy and prevents biases from incomplete or duplicated data. Following preprocessing, the K-Means clustering algorithm is applied to identify patterns and segment customers based on their characteristics.

Prior to clustering, data standardization and dimensionality reduction using PCA are performed to optimize the clustering process. PCA is employed to reduce the dimensionality of the data, retaining approximately 75% of the data variation by selecting three principal components. After data normalization and identification of important variables through PCA, the K-Means algorithm is implemented. For evaluation, the Elbow Method and Silhouette Score are used to determine the optimal number of clusters, with both methods yielding the same optimal number of eight clusters.

The K-Means model, following PCA and optimal cluster selection, achieves a Silhouette Score of 0.7789, indicating its effectiveness, as a score close to 1 signifies an improved model. Clustering analysis results provide insights into various customer segments, including proportions, age distribution, purchasing power, professions, and marital status. This data equips businesses with valuable information to understand their target market better and create tailored marketing strategies to address diverse customer needs. For instance, clusters with low purchasing power may be targeted with more affordable product offerings, while premium products can be tailored for clusters with high purchasing power. Understanding each cluster's characteristics enables companies to develop products or services that precisely meet the target market's needs.

This research aims to assist companies in understanding the characteristics and needs of their customers better, enabling the development of more effective and personalized marketing strategies. However, it is important to note the limitation of this research lies in the suboptimal quality of the dataset, with some data containing noise that could impact the research results. Future research could expand by incorporating more diverse data and improving the overall quality of the dataset.

REFERENCES

[1] N. H. Harani, C. Prianto, and F. A. Nugraha, “Segmentasi Pelanggan Produk Digital Service Indihome Menggunakan Algoritma K-Means Berbasis Python,” *J. Manaj. Inform.*, vol. 10, no. 2, pp. 133–146, 2020, doi: 10.34010/jamika.v10i2.2683.

[2] A. T. Widiyanto and A. Witanti, “Segmentasi Pelanggan Berdasarkan Analisis RFM Menggunakan Algoritma K-Means Sebagai Dasar Strategi Pemasaran (Studi Kasus PT Coversuper Indonesia Global),” *KONSTELASI Konvergensi Teknol. dan Sist. Inf.*, vol. 1, no. 1, pp. 204–215, 2021, doi: 10.24002/konstelasi.v1i1.4293.

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

[4] D. Hediyati and I. M. Suartana, “Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten

Bojonegoro," *J. Inf. Eng. Educ. Technol.*, vol. 5, no. 2, pp. 49–54, 2021, doi: 10.26740/jieet.v5n2.p49-54.

[5] M. Harahap, Y. Lubis, and Z. Situmorang, "Analisis Pemasaran Bisnis dengan Data Science : Segmentasi Kepribadian Pelanggan berdasarkan Algoritma K-Means Clustering," *Data Sci. Indones.*, vol. 1, no. 2, pp. 76–88, 2022, doi: 10.47709/dsi.v1i2.1348.

[6] S. Dwididanti and D. A. Anggoro, "Analisis Perbandingan Algoritma Bisecting K-Means dan Fuzzy C-Means pada Data Pengguna Kartu Kredit," *Emit. J. Tek. Elektro*, vol. 22, no. 2, pp. 110–117, 2022, doi: 10.23917/emitor.v22i2.15677.

[7] N. Khairu Nissa, Y. Nugraha, C. F. Finola, A. Ernesto, J. I. Kanggrawan, and A. L. Suherman, "Evaluasi Berbasis Data: Kebijakan Pembatasan Mobilitas Publik dalam Mitigasi Persebaran COVID-19 di Jakarta," *J. Sist. Cerdas*, vol. 3, no. 2, pp. 84–94, 2020, doi: 10.37396/jsc.v3i2.77.

[8] N. Y. Aswad, "Clustering Algoritma K-Means Pengadaan Barang Non Medis Di Rumah Sakit Jantung Hasna Medika Cirebon," *J. Data Sci. dan Inform.*, vol. 2, no. 1, pp. 6–14, 2022.

[9] A. Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 20–28, 2023, doi: 10.58602/jaiti.v1i1.22.

[10] T. Tommy and A. M. Husein, "Model Prediksi Prestasi Mahasiswa Berdasarkan Evaluasi Pembelajaran Menggunakan Pendekatan Data Science," *Data Sci. Indones.*, vol. 1, no. 1, pp. 14–20, 2021, doi: 10.47709/dsi.v1i1.1168.

[11] T. F. Johnson, N. J. B. Isaac, A. Paviolo, and M. González-Suárez, "Handling missing values in trait data," *Glob. Ecol. Biogeogr.*, vol. 30, no. 1, pp. 51–62, 2021, doi: 10.1111/geb.13185.

[12] P. Arsi, R. Wahyudi, and R. Waluyo, "Optimasi SVM Berbasis PSO pada Analisis Sentimen Wacana Pindah Ibu Kota Indonesia," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 2, pp. 231–237, 2021, doi: 10.29207/resti.v5i2.2698.

[13] T. Nyitrai and M. Virág, "The effects of handling outliers on the performance of bankruptcy prediction models," *Socioecon. Plann. Sci.*, vol. 67, no. August, pp. 34–42, 2019, doi: 10.1016/j.seps.2018.08.004.

[14] E. P. Cynthia and E. Ismanto, "Metode Decision Tree Algoritma C.45 Dalam Mengklasifikasi Data Penjualan Bisnis Gerai Makanan Cepat Saji," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 3, no. July, p. 1, 2018, doi: 10.30645/jurasik.v3i0.60.

[15] A. S. Ritonga and I. Muhandhis, "Teknik Data Mining Untuk Mengklasifikasikan Data Ulasan Destinasi Wisata Menggunakan Reduksi Data Principal Component Analysis (Pca)," *Edutic - Sci. J. Informatics Educ.*, vol. 7, no. 2, 2021, doi: 10.21107/edutic.v7i2.9247.

[16] A. Sulistiyawati and E. Supriyanto, "Implementasi Algoritma K-means Clustring dalam Penetuan Siswa Kelas Unggulan," *J. Tekno Kompak*, vol. 15, no. 2, p. 25, 2021, doi: 10.33365/jtk.v15i2.1162.

[17] K. D. Ramgude and N. R. Rajhans, "K-means clustering for optimization of spare parts delivery," *Manag. Sci. Lett.*, vol. 13, no. 4, pp. 235–240, 2023, doi: 10.5267/j.msl.2023.6.004.

[18] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal. DSAA 2020*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.