Analysis of Credit Card Usage Against Business Segmentation Using Agglomerative Hierarchical Clustering

Nugrah Anggara Siregar[a], Arnita[b], Shabrina Prabudi[c] , Raudha Izmainy Nasution[d], Reza Nur Afdal [e]

[a b c d e]*Jurusan Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam,*
Universitas Negeri Medan, Indonesia
Corresponding Author:
[c]*shabrinaprabudi27@gmail.com*

ABSTRACT

This study analyzes credit card usage patterns and their impact on business segmentation using the Agglomerative Hierarchical Clustering (AHC) method. AHC was chosen because of its ability to group data in detail, especially for datasets with hierarchical relationships. The dataset includes balances, credit limits, monthly payments, and late history. The study aimed to identify high-risk credit card users in payments so that financial institutions can develop more effective risk management strategies. This study successfully identified customer groups with varying payment risks and offered solutions like debt consolidation and flexible payment programs. These findings contribute to the credit card industry in customer segmentation and credit risk management in the credit card industry.

Keywords : Agglomerative Hierarchical Clustering, Credit Card, Data Clustering, Consumer Behavior Analysis

INTODUCTION

This study examines credit card usage patterns and their impact on business segmentation. The Agglomerative Hierarchical Clustering (AHC) method was chosen because it produces accurate and detailed data segmentation, especially for datasets with structural and hierarchical relationships. This dataset includes customer information such as balance, credit limit, monthly payment, and late payment.

This study aims to help financial institutions develop better risk management strategies by identifying credit card users who are at high risk of payment difficulties. By analyzing customer behavior, this study aims to reduce credit risk while improving customer experience. This study, uses AHC to group credit card users according to their financial profiles. First, the dataset with various financial metrics is processed to ensure quality and consistency. This includes cleaning the data, using KNNImputer to handle missing values, and normalizing the data to address scale differences between variables.

The study found groups of customers with different payment risks. Solutions such as flexible payment programs and debt consolidation help high-risk customers reduce their risk of default. The results show that understanding how people use credit cards is essential to reducing credit risk and improving customer retention tactics. This study significantly contributes to credit risk management and customer segmentation in the credit card industry. It offers a new approach to customer segmentation using AHC and emphasizes the importance of understanding credit card usage patterns for effective risk management and customer-specific solutions. This study lays the foundation for future studies on credit risk management and customer-specific solutions.

LITERATURE REVIEW

Credit card usage has become a phenomenon that has increased significantly in recent decades due to technological developments and digitalization in the financial sector. With the development of the credit card business, the features offered are increasingly diverse, so that more and more consumers are interested in using credit cards (Akbar et al. 2023:9). Credit cards not only function as a payment tool that facilitates transactions but also stores valuable data on consumer shopping patterns and preferences. This data has great potential to help businesses understand consumer behavior, optimize marketing strategies, and identify different customer segments. In a business context, customer segmentation is an essential strategy that allows companies to classify customers into more homogeneous groups based on certain specific characteristics such as purchasing patterns, frequency of use, and credit risk levels.

In this study, data mining is used to perform clustering. Data mining is an interactive or automated process to find data patterns and predict future behavior based on these patterns (Fadliana & Rozi, 2015:40). One of the techniques used in multivariate statistical analysis is cluster analysis, which aims to group objects based on their characteristics (Kusumawardani et al. 2018:36). This study explores the potential of the agglomerative hierarchical clustering (AHC) technique in customer segmentation. AHC is considered more suitable for data that has with structural and hierarchical relationships, allowing for more detailed and accurate segmentation. AHC (Agglomerative Hierarchical Clustering) is a technique for exploring data by grouping it into several groups (clusters) (Syahara et al. 2024:314).

Hierarchical Clustering is a clustering technique that forms a hierarchy, resulting in a tree-shaped structure. The grouping process is carried out in stages or stages, and there are two main methods in the hierarchical clustering algorithm, namely agglomerative (bottom-up) and divisive (top-down) (Yulianti et al. 2023:308). AHC allows companies to understand customer characteristics better and group them based on more structured behavior.

Previous research using a cluster-based approach to credit card customer segmentation shows that customer segmentation is a process used by companies to group their customers based on similar characteristics (Alhamdani et al. 2021:77). Most studies tend to focus on simple clustering techniques or applications on a limited data scale. In addition, previous studies are limited in that they do not consider the combination of variables such as demographics, consumption patterns, and credit risk in segmentation.

Therefore, this study aims to fill this gap using a more comprehensive AHC approach to categorize credit card users based on various variables, including credit risk. The main objective of this study is to identify credit card users at high risk of experiencing payment difficulties and offer solutions such as debt consolidation and more flexible payment programs. This approach is expected to help financial institutions manage credit risk more effectively and proactively. Thus, this study not only contributes to the development of credit risk management strategies but also offers practical recommendations for credit card providers. These recommendations include providing incentives to users who make full payments and financial education for those who only pay the minimum amount, hoping to reduce the risk of default in the future.

METHODS

a) Research Stages

This research has a research framework that is structured based on the following stages.

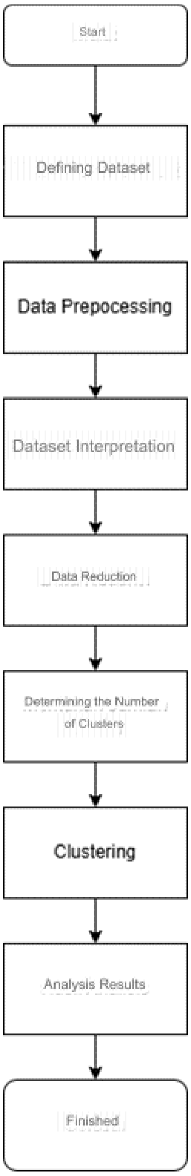


Figure 1. Research Stages

b) Defining DataSet

The dataset used is the result of data collection from around 9000 active credit card holders during the last 6 months in 2017. This dataset consists of 8950 entries with 17 columns, which include various information about customer credit card usage behavior. This data was obtained from Kaggle with the title Credit Card Dataset for Clustering.

Table 1. Data Description Statistics

	Count	Mean	Std	Min	25%	50%	75%	Max
BALANCE	8950.0	1564.474828	2081.531879	0.000000	128.281915	873.385231	2054.140036	19043.13856
BALANCE_FREQUENCY	8950.0	0.877271	0.236904	0.000000	0.888889	1.000000	1.000000	1.00000
PURCHASES	8950.0	1003.204834	2136.634782	0.000000	39.635000	361.280000	1110.130000	49039.57000
ONEOFF_PURCHASES	8950.0	592.437371	1659.887917	0.000000	0.000000	38.000000	577.405000	40761.25000
INSTALLMENTS_PURCHASES	8950.0	411.067645	904.338115	0.000000	0.000000	89.000000	468.637500	22500.00000
CASH_ADVANCE	8950.0	978.871112	2097.163877	0.000000	0.000000	0.000000	1113.821139	47137.21176
PURCHASES_FREQUENCY	8950.0	0.490351	0.401371	0.000000	0.083333	0.500000	0.916667	1.00000
ONEOFF_PURCHASES_FREQUENCY	8950.0	0.202458	0.298336	0.000000	0.000000	0.083333	0.300000	1.00000
PURCHASES_INSTALLMENTS_FREQUENCY	8950.0	0.364437	0.397448	0.000000	0.000000	0.166667	0.750000	1.00000
CASH_ADVANCE_FREQUENCY	8950.0	0.135144	0.200121	0.000000	0.000000	0.000000	0.222222	1.50000

CASH_ADVANCE_TRX	8950.0	3.248827	6.824647	0.000000	0.000000	0.000000	4.000000	123.00000
PURCHASES_TRX	8950.0	14.709832	24.857649	0.000000	1.000000	7.000000	17.000000	358.00000
CREDIT_LIMIT	8950.0	4494.178190	3638.702906	50.000000	1600.000000	3000.000000	6500.000000	30000.00000
PAYMENTS	8950.0	1733.143852	2895.063757	0.000000	383.276166	856.901546	1901.134317	50721.48336
MINIMUM_PAYMENTS	8950.0	854.290836	2351.606794	0.019163	166.269650	300.487478	804.414788	76406.20752
PRC_FULL_PAYMENT	8950.0	0.153715	0.292499	0.000000	0.000000	0.000000	0.142857	1.00000
TENURE	8950.0	11.517318	1.338331	6.000000	12.000000	12.000000	12.000000	12.00000

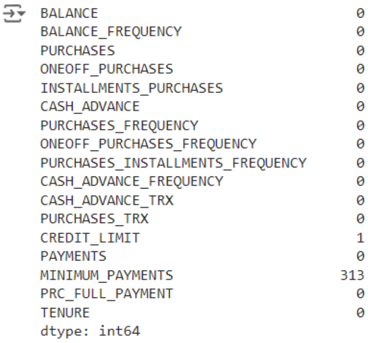
Each column in the table above plays an important role in understanding the habits and preferences of credit card users. Here are descriptions of some of the key columns:

1. BALANCE: Represents the remaining balance on a credit card, which gives an indication of the customer's use of funds.
2. BALANCE_FREQUENCY: Shows how often the balance is updated, which can illustrate the level of credit card activity.
3. PURCHASES: Total purchases made by the customer, important for understanding spending habits.
4. ONEOFF_PURCHASES: The largest purchase amount made in a single transaction, useful for identifying large purchases.
5. INSTALLMENTS_PURCHASES: Purchases made in installments, indicating a preference for installment payments.
6. CASH_ADVANCE: The amount of cash withdrawn as a loan through a credit card, relevant for measuring liquidity needs.
7. PURCHASES_FREQUENCY: The frequency of purchase transactions made, important for understanding how often the customer uses the credit card.

8. CREDIT_LIMIT: The credit limit given to the customer, useful for understanding available credit capacity.
9. PAYMENTS: The total amount of payments made by a customer, reflecting the customer's ability to pay bills.
10. MINIMUM_PAYMENTS: Describes the minimum payment made, useful for understanding debt repayment habits.
11. TENURE: The length of time a customer has used a credit card, which can help identify loyalty and long-term behavior.

c) Data Preprocessing

The first step of this study was to clean the dataset by removing irrelevant columns, such as unique customer IDs. This process is essential to ensure that the analysis is not disrupted by information that does not contribute significantly to the model.



```
BALANCE                               0
BALANCE_FREQUENCY                     0
PURCHASES                             0
ONEOFF_PURCHASES                      0
INSTALLMENTS_PURCHASES                0
CASH_ADVANCE                          0
PURCHASES_FREQUENCY                   0
ONEOFF_PURCHASES_FREQUENCY            0
PURCHASES_INSTALLMENTS_FREQUENCY      0
CASH_ADVANCE_FREQUENCY                0
CASH_ADVANCE_TRX                      0
PURCHASES_TRX                         0
CREDIT_LIMIT                          1
PAYMENTS                              0
MINIMUM_PAYMENTS                    313
PRC_FULL_PAYMENT                      0
TENURE                                0
dtype: int64
```

Figure 2. Missing Data Value Output

Data analysis shows that most variables in the dataset have complete data. However, there are missing values in the CREDIT_LIMIT and MINIMUM_PAYMENTS columns, indicating incomplete data for some customers regarding credit limits and minimum payments. To ensure better data quality, KNNImputer is used with the parameter n_neighbors=3 to fill in the missing values in the MINIMUM_PAYMENTS column. This method was chosen because it can provide more accurate estimates by considering information from the three nearest neighbors, resulting in more relevant imputations in accordance with existing data patterns. It is assumed that every credit card user must have a minimum payment transaction, estimated at around 0.019163. This approach can replace missing values using information from other samples with similar values on the related features. After the KNN stage is complete, the repaired dataset is ready to proceed to the next analysis stage.

d) DataSet Interpretation

This interpretation provides a comprehensive overview of the distribution pattern of each feature in the dataset. By describing the overall data structure, this analysis highlights significant features, helping to understand each feature's role and contribution to the dataset's overall characteristics. The interpretation for each feature based on its data distribution is as follows:

1. BALANCE: The majority of users have low balances, while some users have very high balances.
2. BALANCE_FREQUENCY: Most users regularly check their balances, with a frequency close to 1.

3. PURCHASES: Most users make small purchases, but there are some with large purchases.
4. ONEOFF_PURCHASES: Most users rarely make large purchases at one time, but there are some outliers with large purchases.
5. INSTALLMENTS_PURCHASES: Most rarely make installment purchases, but some users do so routinely.
6. CASH_ADVANCE: The majority do not use cash down payments, but some users take large down payments.
7. PURCHASES_FREQUENCY: Most users have a high frequency of routine purchases.
8. ONEOFF_PURCHASES_FREQUENCY : Most users rarely make one-time purchases.
9. INSTALLMENTS_PURCHASES_FREQUENCY : Most users rarely make installment purchases.
10. CASH_ADVANCE_FREQUENCY : Most users rarely use cash advances, but there are some users who use them frequently.
11. CASH_ADVANCE_TRX : Most users do not have any cash advance transactions.
12. PURCHASES_TRX : The majority of users make few purchases, with a few outliers who make many transactions.
13. CREDIT_LIMIT : Most users have low to moderate credit limits, with a few users having very high credit limits.
14. PAYMENTS : Most payments are made in small amounts, but there are outliers who make large payments.
15. MINIMUM_PAYMENTS : Most minimum payments are low, with a few users having high minimum payments.
16. PRC_FULL_PAYMENT : The majority of users rarely pay their full bill, only a small percentage do so consistently.
17. TENURE: The majority of users have a relationship duration of 12 months, indicating a predominance of long-term relationships.

e) Data Transformation

In data transformation, data normalization is achieved using RobustScaler. Data normalization using RobustScaler occurs when the data has outliers or extreme values that can affect model performance. RobustScaler is chosen because it is more resistant to outliers than other scalers. This is also because RobustScaler uses the median and interquartile range (IQR) to calculate the data scale, not the mean and standard deviation. The median is the middle value of the data unaffected by outliers, while the IQR (interquartile range) only includes data between the 25th and 75th percentiles. This way, the scale change is based on values that are more representative of the main primary data distribution and are not affected by extreme values. Next, data interpretation will be carried out again. Data interpretation after normalization ensures that the normalization process is successful and the data is ready for further analysis or modeling without distribution problems or the influence of outliers.

f) Data Reduction

In data reduction, there is the use of Principal Component Analysis (PCA) is used in data reduction because it can overcome various challenges when analyzing datasets with many variables. PCA is used to reduce the dimension of data, namely reducing the number of features or variables in a dataset without losing significant information. In this study, each row contains values for the original data's first (P1) and second (P2) principal components of each original data. By reducing the dimension of the data, the analysis can focus on the most significant features and ignore the less important ones. This makes the clustering process more efficient because the clustering algorithm can work with straightforward but still informative data. As a result, the clustering that is carried out becomes more accurate because the data has been simplified to highlight the most relevant information.

g) Determining the Number of Clusters

This study, it determines the optimal number of clusters using the elbow method and silhouette score algorithms.

1. Elbow Method

   The elbow method involves plotting the Within-Cluster Sum of Squares (WCSS) against the number of clusters. WCSS is used because of its simplicity and ability to measure clustering quality, so in the Elbow Method WCSS is involved to provide an intuitive way to determine the optimal number of clusters (Akbar et al. 2023:9).

2. Silhouette Score

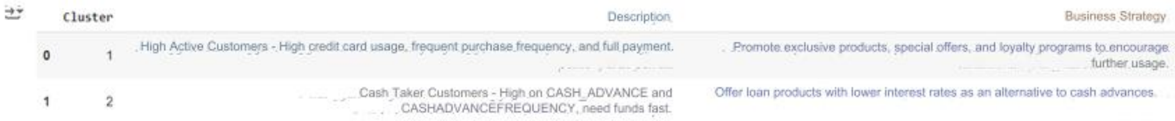
   The silhouette score measures how well objects are grouped in clusters. Its value ranges between -1 and 1, and higher values indicate that objects are closer to their own cluster than to neighboring clusters (Kusumawardani et al. 2018:36).

h) Clustering

The Agglomerative Hierarchical Clustering algorithm groups users based on similar characteristics. First, this algorithm produces a dendrogram that makes it easier for us to understand the data structure and visually see the grouping process. Second, this algorithm is flexible in determining the number of clusters. By looking at the dendrogram, we can choose the number of clusters that best suit the characteristics of our data. Third, this algorithm can handle various forms of distribution and different cluster sizes, making it suitable for user data with diverse usage patterns. Its implementation begins by determining the desired number of clusters and performing clustering using the Agglomerative method. The clustering results are then visualized using scatter plots and dendrograms to facilitate analysis. After that, we perform descriptive analysis on each cluster to understand the unique characteristics of each user group. This information is then used to develop more targeted business strategies, such as offering special promotions for users who are often late in paying or providing incentives for users who always pay on time.

i) Analysis Results

In this analysis results using Agglomerative Hierarchical Clustering algorithm two main customer segments were identified.



	Cluster	Description	Business Strategy
0	1	High Active Customers - High credit card usage, frequent purchase frequency, and full payment.	Promote exclusive products, special offers, and loyalty programs to encourage further usage.
1	2	Cash Taker Customers - High on CASH_ADVANCE and CASHADVANCEFREQUENCY, need funds fast.	Offer loan products with lower interest rates as an alternative to cash advances.

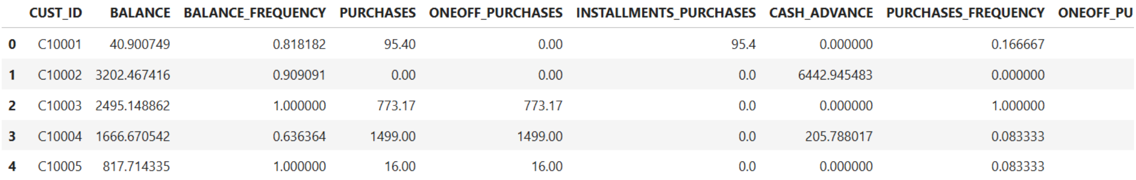
Figure 3. Analysis Results

The first segment is High Active Customers, characterized by intense credit card usage, frequent purchase frequency, and a tendency to pay off in full. Customers in this segment show strong loyalty and make optimal use of credit cards use credit cards optimistically. Therefore, the recommended business strategy for this segment is to promote exclusive products, offer special deals, and develop loyalty programs that can encourage them to increase their credit card usage further. The second segment is Cash Withdrawal Customers, characterized by high use of cash advance facilities especially in the CASH_ADVANCE and CASHADVANCEFREQUENCY variables, which indicate an urgent need for cash. For this segment, the business strategy that can be applied is to offer loan products with lower interest as an alternative to using cash advance facilities, so that customers get a more financially beneficial option.

RESULT

a) Data Description

The dataset used in this analysis consists of 8950 spanning 17 columns. After thorough checking, the values in this dataset are not missing, which means the data quality is maintained for further analysis. This allows for in-depth analysis without the need for imputation or handling missing values, which can often affect the final results.

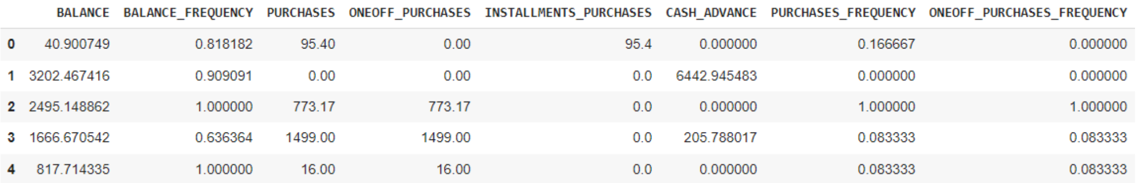


	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PU
0	C10001	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	
1	C10002	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	
2	C10003	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000000	
3	C10004	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083333	
4	C10005	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083333	

Figure 4. Data Description

b) Data Prepocessing

In the pre-processing stage, various steps are taken to ensure a clean dataset that is ready to be used in subsequent analysis. This process includes deleting columns, handling missing values with data normalization to ensure consistency across variables. The data that has been combined and cleaned is then processed so that it is ready to be used in the clustering process. Normalization is carried out to overcome differences in scale between variables, which is very important considering that the existing variables have different scales. Thus, the results of the analysis become more accurate and reliable. So that the results are obtained as in Figure 4.



	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY
0	40.900749	0.818182	95.40	0.00	95.4	0.000000	0.166667	0.000000
1	3202.467416	0.909091	0.00	0.00	0.0	6442.945483	0.000000	0.000000
2	2495.148862	1.000000	773.17	773.17	0.0	0.000000	1.000000	1.000000
3	1666.670542	0.636364	1499.00	1499.00	0.0	205.788017	0.083333	0.083333
4	817.714335	1.000000	16.00	16.00	0.0	0.000000	0.083333	0.083333

Figure 5. Preprocessing Result Data

c) Interpretasi Dataset

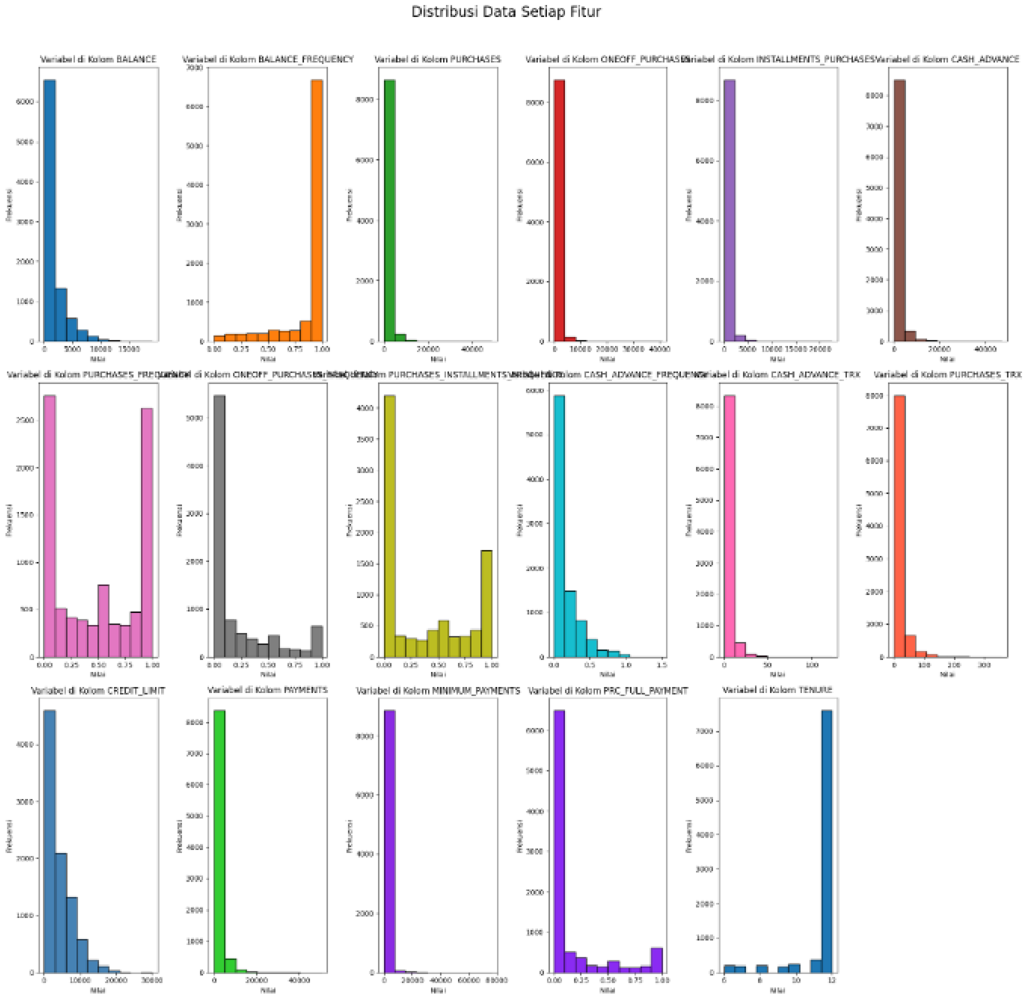


Figure 6. Data Distribution of Each Feature

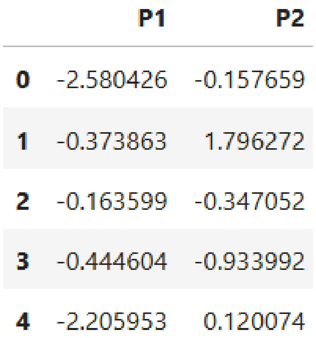
For this data analysis, it is important to understand the distribution of the dataset attributes. The following histogram shows the distribution of values for each of the 17 columns analyzed, providing an in-depth look at the characteristics of user behavior. By understanding this distribution, we can find common patterns and outliers that may affect the analysis results. A brief interpretation of each feature based on the observation of the histogram shown is given below:

Table 2. Histogram observation feature

BALANCE	Most users have low balances, with a few users having high balances.
BALANCE_FREQUENCY	Most users check their balance with high frequency, almost close to 1.
PURCHASES	Most users make small purchases, but there are some large purchases.
ONEOFF_PURCHASES	One-time purchases are rare in large amounts, the majority of users have low value.

INSTALLMENTS_PURCHASES	Most users rarely make purchases on installments, with only a few doing so regularly.
CASH_ADVANCE	Most users do not use cash advances, with only a few users having large down payments.
PURCHASES_FREQUENCY	The majority of users make purchases regularly with high frequency.
ONEOFF_PURCHASES_FREQUENCY	Most users rarely make a single purchase.
INSTALLMENTS_PURCHASES_FREQUENCY	Most users rarely make installment purchases.
CASH_ADVANCE_FREQUENCY	Most users rarely use cash advances.
CASH_ADVANCE_TRX	The majority of users do not have cash advance transactions.
PURCHASES_TRX	Most users make a small number of purchase transactions.
CREDIT_LIMIT	Most users have low credit limits, although some have very high credit limits.
PAYMENTS	Most payments were made in small amounts, with a few outliers making large payments.
MINIMUM_PAYMENTS	Most users make small minimum payments, although there are some who have large minimum amounts.
PRC_FULL_PAYMENT	Most users rarely pay their bills in full.
TENURE	The majority of users have a relationship duration of 12 months, indicating a long-term loyalty.

d) Data Reduction



	P1	P2
0	-2.580426	-0.157659
1	-0.373863	1.796272
2	-0.163599	-0.347052
3	-0.444604	-0.933992
4	-2.205953	0.120074

Figure 7. Data Reduction

The values in the table above are the results of the dimensionality reduction process, where the original dataset that has many features or variables is reduced to two main components, namely P1 and P2. Dimensionality reduction aims to simplify data without losing important information, so that data analysis and visualization become easier. The values in columns P1 and P2 describe the new representation of the data after the transformation. Each row shows the position of the observation in this new component space. For example, in the first row, the observation has a value of P1 = -2.580426 and P2 = -0.157659, which means that after the reduction process, this observation has a new representation that is dominant in the first component (P1), while the contribution from the second component (P2) is relatively smaller. Likewise, each other row shows how the observation is "projected" into these two main dimensions.

e) Determining the Number of Clusters

1. Elbow Method

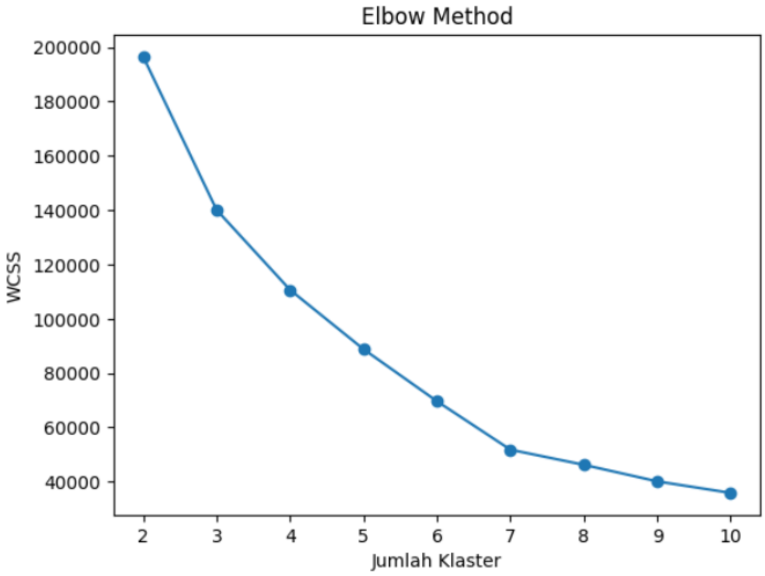


Figure 8. Elbow Method

The graph shown uses the Elbow Method to determine the optimal number of clusters in the Agglomerative Hierarchical Clustering algorithm. The horizontal axis shows the number of clusters, while the vertical axis depicts the Within-Cluster Sum of Squares (WCSS), which measures the variation within each cluster. In this graph, it can be seen that the WCSS decreases drastically when the number of clusters increases from 2 to about 4 clusters. This decrease indicates that in the early stages of clustering, merging clusters effectively reduces variation within clusters in the early stages of clustering. However, once the number of clusters reaches about 4 or 5, the decrease in WCSS begins to slow down significantly, forming an elbow point. This point indicates that adding clusters after that point does not significantly reduce WCSS, so the optimal number of clusters is around 4 or 5. Thus, this graph helps determine the optimal number of clusters, which produces efficient clustering results that minimize variation within each cluster.

2. Silhouette Score

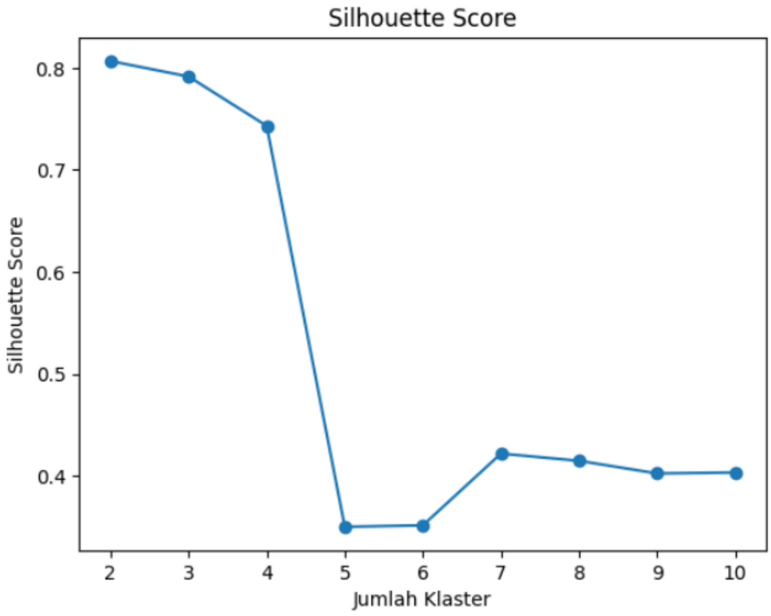


Figure 9. Silhouette Score

This graph illustrates the relationship between the number of clusters and the Silhouette Score, which is used to evaluate clustering quality. The x-axis represents the number of clusters (ranging from 2 to 10), while the y-axis shows the Silhouette Score (ranging from 0 to 0.8), where a higher value indicates better clustering quality. The results show that the best clustering occurs with 2 or 3 clusters, with a Silhouette Score approaching 0.8. After 4 clusters, the score drops sharply, indicating that increasing the number of clusters actually reduces the clustering quality. From 6 clusters onward, the score stabilizes around 0.4, reflecting poor quality.

f) Clustering

1. Plotting

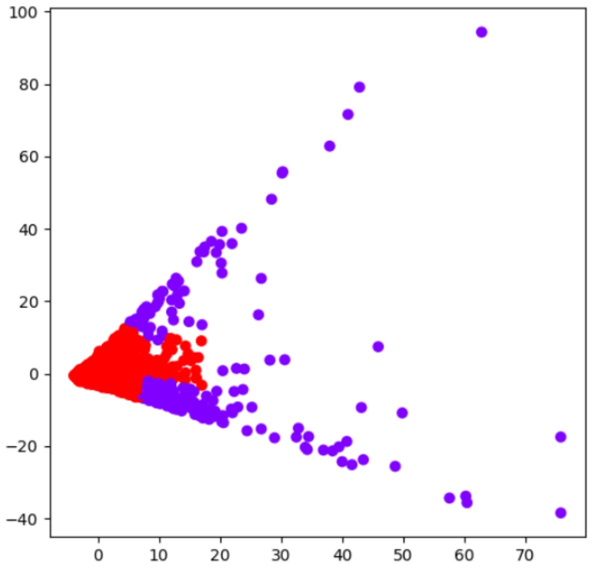


Figure 10. Plotting

In this graph, the red and purple points represent objects that have been grouped into different clusters. The red points represent one cluster, while the purple points represent the other cluster. This visualization shows the distribution of data where most objects are concentrated in denser areas near the x-axis, while some objects are spread farther along both axes, indicating variation in the data patterns.

2. Data Visualization

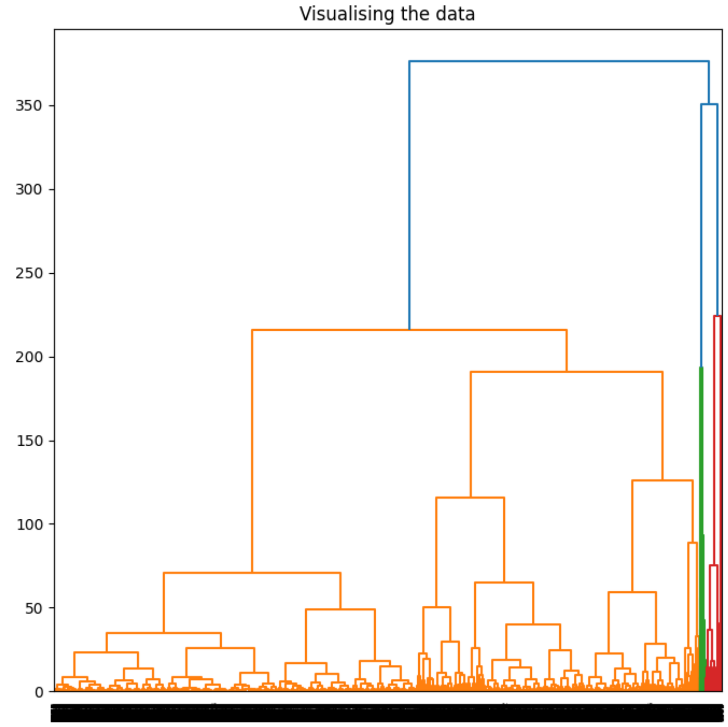
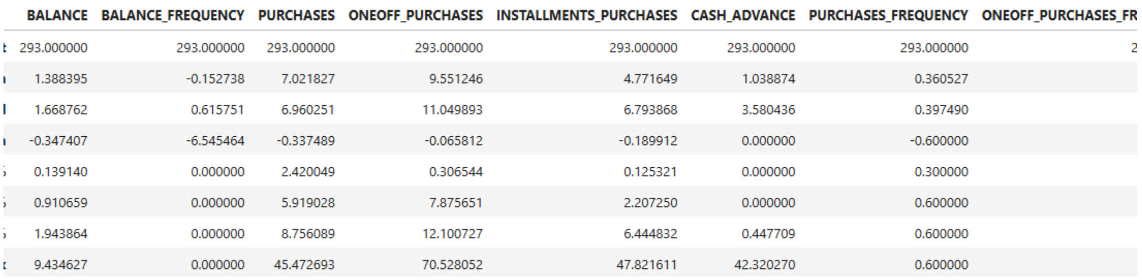


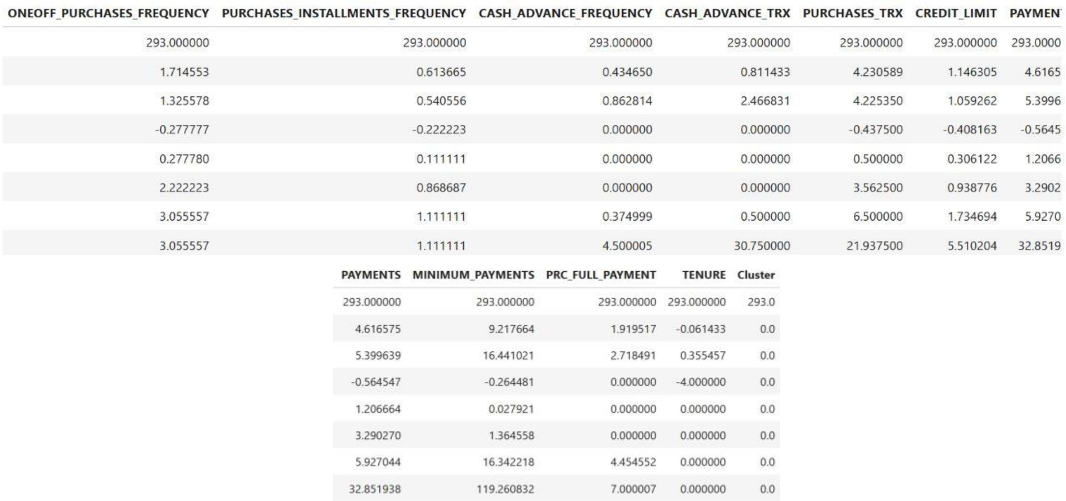
Figure 11. Data Visualization

We can observe how the data is gradually merged, and a cutting point at a certain height can be used to determine the optimal number of clusters. In this image, the ideal cutting potential is at a height of around 100 to 150, where several large branches are visible. This visualization helps in understanding the hierarchical structure of the data and how the groups are formed, while also making it easier to determine the appropriate number of clusters based on the differences between the data groups.

3. Business Strategy Preparation



	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FR
t	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	2
ı	1.388395	-0.152738	7.021827	9.551246	4.771649	1.038874	0.360527	
l	1.668762	0.615751	6.960251	11.049893	6.793868	3.580436	0.397490	
ı	-0.347407	-6.545464	-0.337489	-0.065812	-0.189912	0.000000	-0.600000	
;	0.139140	0.000000	2.420049	0.306544	0.125321	0.000000	0.300000	
;	0.910659	0.000000	5.919028	7.875651	2.207250	0.000000	0.600000	
;	1.943864	0.000000	8.756089	12.100727	6.444832	0.447709	0.600000	
(	9.434627	0.000000	45.472693	70.528052	47.821611	42.320270	0.600000	

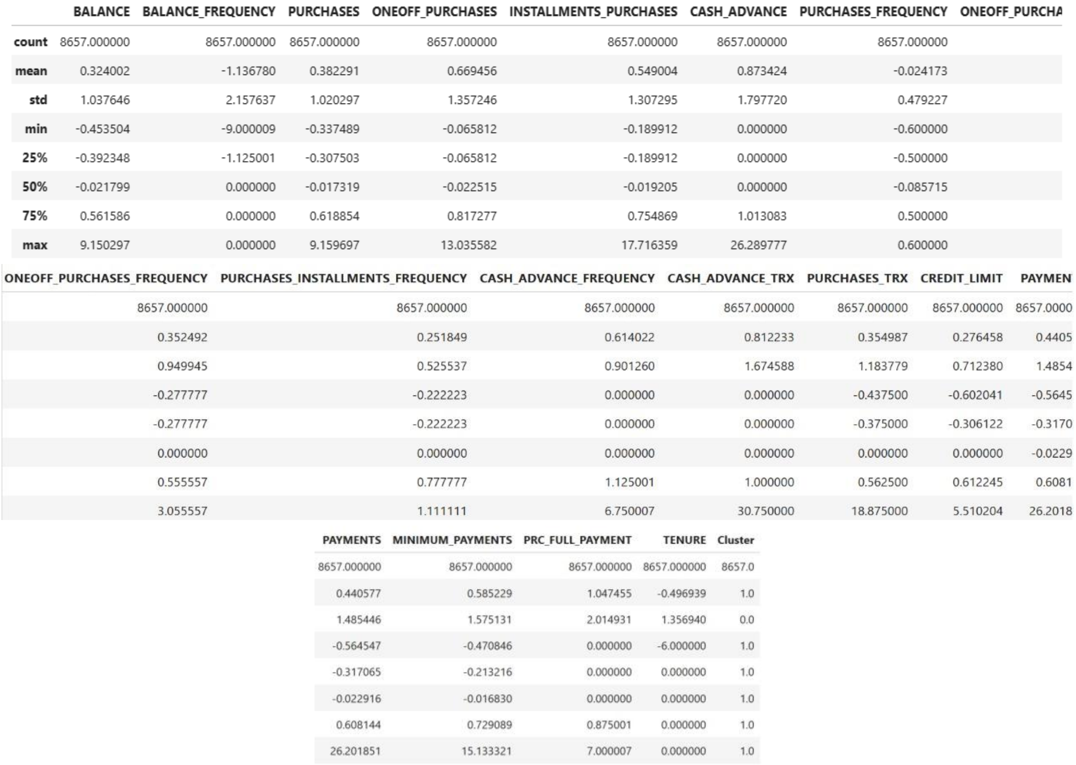


ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMEN
293.000000	293.000000	293.000000	293.000000	293.000000	293.000000	293.0000
1.714553	0.613665	0.434650	0.811433	4.230589	1.146305	4.6165
1.325578	0.540556	0.862814	2.466831	4.225350	1.059262	5.3996
-0.277777	-0.222223	0.000000	0.000000	-0.437500	-0.408163	-0.5645
0.277780	0.111111	0.000000	0.000000	0.500000	0.306122	1.2066
2.222223	0.868687	0.000000	0.000000	3.562500	0.938776	3.2902
3.055557	1.111111	0.374999	0.500000	6.500000	1.734694	5.9270
3.055557	1.111111	4.500005	30.750000	21.937500	5.510204	32.8519

PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE	Cluster
293.000000	293.000000	293.000000	293.000000	293.0
4.616575	9.217664	1.919517	-0.061433	0.0
5.399639	16.441021	2.718491	0.355457	0.0
-0.564547	-0.264481	0.000000	-4.000000	0.0
1.206664	0.027921	0.000000	0.000000	0.0
3.290270	1.364558	0.000000	0.000000	0.0
5.927044	16.342218	4.454552	0.000000	0.0
32.851938	119.260832	7.000007	0.000000	0.0

Figure 12. Cluster 1

Users in this cluster exhibit diverse usage patterns. Some have high balances (>290,000) with frequent purchases (PURCHASES_FREQUENCY close to 1), while others use their credit cards for one-time payments or installments, as well as cash withdrawals. This data can be leveraged for business strategies, such as offering special programs for users who actively use their credit cards for various needs.



	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHA
count	8657.000000	8657.000000	8657.000000	8657.000000	8657.000000	8657.000000	8657.000000	
mean	0.324002	-1.136780	0.382291	0.669456	0.549004	0.873424	-0.024173	
std	1.037646	2.157637	1.020297	1.357246	1.307295	1.797720	0.479227	
min	-0.453504	-9.000009	-0.337489	-0.065812	-0.189912	0.000000	-0.600000	
25%	-0.392348	-1.125001	-0.307503	-0.065812	-0.189912	0.000000	-0.500000	
50%	-0.021799	0.000000	-0.017319	-0.022515	-0.019205	0.000000	-0.085715	
75%	0.561586	0.000000	0.618854	0.817277	0.754869	1.013083	0.500000	
max	9.150297	0.000000	9.159697	13.035582	17.716359	26.289777	0.600000	

ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMEN
8657.000000	8657.000000	8657.000000	8657.000000	8657.000000	8657.000000	8657.0000
0.352492	0.251849	0.614022	0.812233	0.354987	0.276458	0.4405
0.949945	0.525537	0.901260	1.674588	1.183779	0.712380	1.4854
-0.277777	-0.222223	0.000000	0.000000	-0.437500	-0.602041	-0.5645
-0.277777	-0.222223	0.000000	0.000000	-0.375000	-0.306122	-0.3170
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.0229
0.555557	0.777777	1.125001	1.000000	0.562500	0.612245	0.6081
3.055557	1.111111	6.750007	30.750000	18.875000	5.510204	26.2018

PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE	Cluster
8657.000000	8657.000000	8657.000000	8657.000000	8657.0
0.440577	0.585229	1.047455	-0.496939	1.0
1.485446	1.575131	2.014931	1.356940	0.0
-0.564547	-0.470846	0.000000	-6.000000	1.0
-0.317065	-0.213216	0.000000	0.000000	1.0
-0.022916	-0.016830	0.000000	0.000000	1.0
0.608144	0.729089	0.875001	0.000000	1.0
26.201851	15.133321	7.000007	0.000000	1.0

Figure 13. Cluster 2

The image above presents descriptive statistics from cluster 2, resulting from the credit card user clustering analysis, covering variables such as BALANCE, PURCHASES (total purchases), ONEOFF_PURCHASES (one-time purchases), INSTALLMENTS_PURCHASES (installment purchases), and CASH_ADVANCE (cash withdrawals). The statistics include data count, mean, standard deviation, as well as minimum, maximum, and percentile values. The average balance (BALANCE) is 0.93, with a maximum balance of 9.15 and a minimum of -9.50. Installment purchases have a low average, while cash withdrawals show variability, with a maximum of 17.17. Purchase frequency reflects varied activity, with an average close to 0.

g) Analysis Result

1. Marketing and Sales

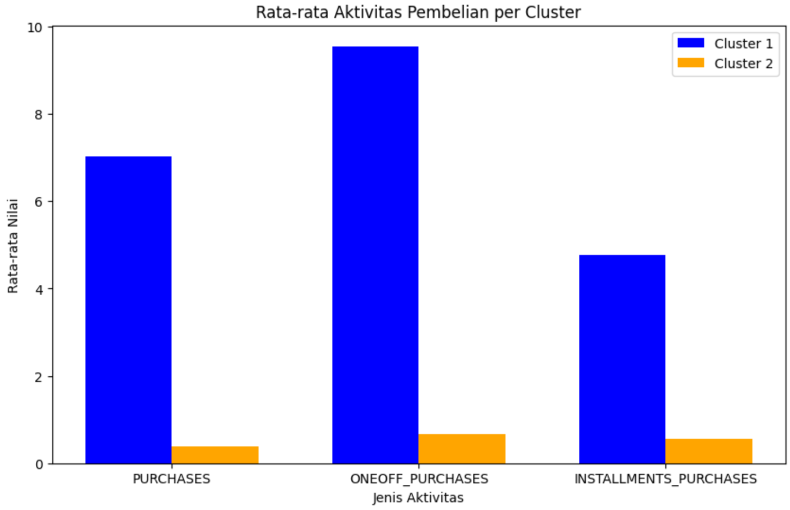


Figure 14. Marketing and Sales

Based on the Usage Activity Segmentation Graph above, this segmentation graph compares the average purchasing activity between two clusters (Cluster 1 and Cluster 2) across three categories: PURCHASES, ONEOFF_PURCHASES, and INSTALLMENTS_PURCHASES. Cluster 1 has a higher overall average purchasing activity compared to Cluster 2. Cluster 1 shows higher activity in the PURCHASES and ONEOFF_PURCHASES categories. Meanwhile, Cluster 2 has a higher average in the INSTALLMENTS_PURCHASES category. The business target is to focus on active users with promotional offers, loyalty programs, or additional products to enhance retention.

2. Risk and Financial Management

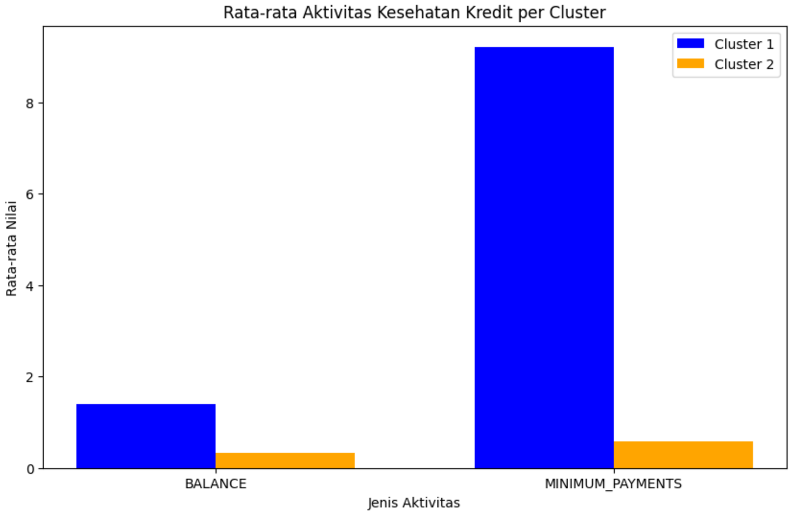


Figure 15. Risk and Financial Management

Based on the Credit Health Segmentation Graph, this segmentation graph compares the average credit health activity between two clusters (Cluster 1 and Cluster 2) across two categories: BALANCE and MINIMUM_PAYMENTS. Cluster 1 has a higher overall average in credit health activity compared to Cluster 2. Cluster 1 exhibits higher averages in both categories: BALANCE and MINIMUM_PAYMENTS. The purpose of this graph is to identify users who may be struggling to pay off debt and offer them debt consolidation or flexible payment programs.

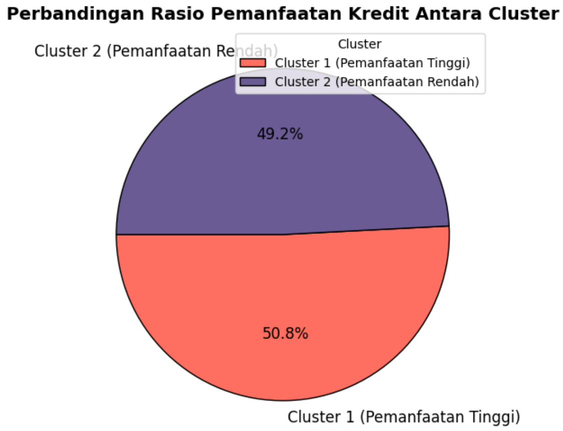


Figure 16. Average Credit Health Activity for Each Claster

Based on the Credit Health Segmentation Graph, Cluster 1 accounts for 50.8% of the total, indicating that customers in this cluster tend to use their credit cards more frequently or in larger amounts compared to their credit limits. Cluster 2 represents 49.2% of the total, suggesting that customers in this cluster tend to use their credit cards more cautiously or have higher credit limits relative to their credit card usage. The objective is to manage credit risk by reviewing credit limits and providing education on financial management for users with high utilization ratios. The application of this graph involves conducting credit limit reviews or issuing risk warnings to users who frequently utilize nearly their entire credit limit.

3. Customer Service

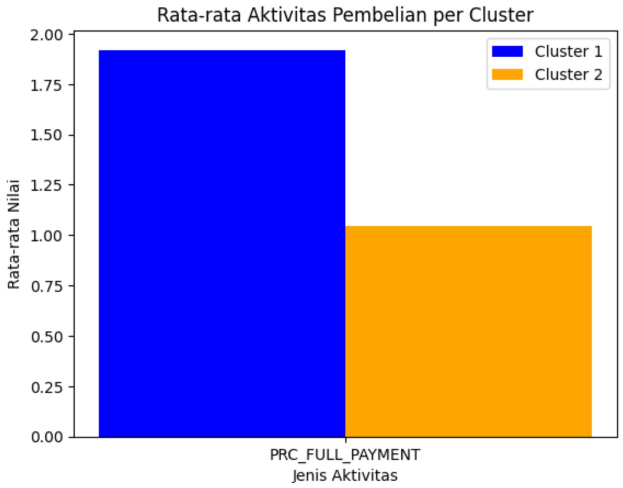


Figure 17. Average Credit Purhasing Activity for Each Claster

Based on the Credit Health Segmentation Graph, Cluster 1 has a higher average purchasing activity in the PRC_FULL_PAYMENT category compared to Cluster 2. The strategy is to encourage users to make full payments by offering interest discounts or reward points. For users who only make minimum payments, education or risk warnings should be provided.

CONCLUSION

This study has successfully applied the agglomerative hierarchical clustering technique to analyze credit card usage and identify different customer segments based on credit card usage behavior and credit risk. Using a comprehensive dataset that includes demographic variables, consumption patterns, and credit risk, this study has developed personalized business strategies for each customer segment. The findings suggest that by understanding credit card usage behavior, companies can take proactive steps to reduce credit risk and improve customer experience by offering programs that suit their needs by understanding credit card usage behavior. This study makes an essential contribution to credit risk management and customer segmentation in the credit card industry, and provides a basis for further research in this area.

REFERENCES

Agapito, G., Milano, M., & Cannataro, M. (2022). Python clustering analysis protocols for gene expression data sets. *MDPI*, 1-22.

Akbar, M. N., Salsabila, A., Asri, A. P., & Syawir, M. (2023). Analisis clustering untuk segmentasi pengguna kartu kredit dengan menggunakan algoritma K-Means dan Principal Component Analysis. *Journal of Artificial Intelligence & Data Science*, 3(1), 1-9.

Alhamdani, F. D., Dianti, A. A., & Azhar, Y. (2021). Segmentasi pelanggan berdasarkan perilaku pengguna kartu kredit menggunakan metode K-Means clustering. *JISKa*, 70-77.

Apfel, N., & Liang, X. (2024). Agglomerative hierarchical clustering selection: Validity of instrumental variables. *Applied Econometrics*, 1-19.

B., K., George, D. J., Manikandan, G., & Thomas, T. (2020). Comparative study of K-Means clustering and agglomerative hierarchical clustering. *International Journal of Emerging Trends in Engineering Research*, 8(5), 1600-1604.

Dwididanti, S., Anggoro, D. A., & Sutanto, M. H. (2022). Analisis perbandingan algoritme bisecting K-Means dan Fuzzy C-Means pada data pengguna kartu kredit. *Emitor: Jurnal Teknik Elektro*, 22(2), 110-117.

Fadliana, A., & Rozi, F. (2015). Penerapan metode agglomerative hierarchical clustering untuk klasifikasi kabupaten/kota di Provinsi Jawa Timur berdasarkan kualitas pelayanan keluarga berencana. 35-40.

Kusumawardani, Y., Hamzah, A., & Suraya. (2018). Perbandingan metode clustering menggunakan hierarchical clustering dan partitional clustering untuk mengelompokkan dokumen berita. *Jurnal Script*, 5(2), 23-36.

Munirsyah, M. A., Bijaksana, & Astuti, W. (2020). Developing synonym sets for English WordNet using the commutative agglomerative clustering method. *Jurnal Sisfokom (Sistem Informasi dan Komputer),* 9(2), 171-176.

Roux, M. (n.d.). Comparative study of divisive hierarchical clustering algorithms.

Simanjuntak, K. P., & Khaira, U. (2021). Pengelompokkan titik api di Provinsi Jambi dengan algoritma hierarchical clustering. *Malcom: Indonesian Journal of Machine Learning and Computer Science*, 1, 7-16.

Siswanto, & Syahrir, N. H. (2022). Agglomerative hierarchical clustering analysis in predicting antibacterial activity of compounds based on chemical structure similarity. *Jurnal Ilmu Matematika dan Terapan*, 16(4), 1441-1452.

Syahara, U., Kurniawati, E., Suhana, M. P., Anggraini, R., & Yandri, F. (2024). Penerapan metode agglomerative hierarchical clustering untuk klasifikasi habitat bentik di Desa Pengudang Kabupaten Bintan. *Insologi (Jurnal Sains dan Teknologi),* 3(3), 306-314.

Widyawati, S., Saptomo, W. L., & Utami, Y. R. (2020). Penerapan algoritma hierarchical clustering untuk segmentasi pelanggan. *JIS (Jurnal Ilmiah Sinus),* 18(1), 75-87.

Yulianti, D. I., Hermanto, T. I., & Defriani, M. (2023). Analisis clustering donor darah dengan metode agglomerative hierarchical clustering. *Resolusi: Rekayasa Teknik Informatika dan Informasi*, 3(6), 303-308.