**Credit Card Clustering: An Examination of Consumer Segmentation**

Mr. Shubham Sonake1 , Dr. A.D. Sawarkar2.

*1*Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded.

[*sonakeshubham1817@gmail.com*](mailto:sonakeshubham1817@gmail.com)

*2*Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded.

[*adsawarkar@sggs.ac.in*](mailto:adsawarkar@sggs.ac.in)

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

**Abstract:** Credit card customer segmentation plays a key role in understanding customer behavior and tailoring marketing strategies. In this study, we aimed to cluster credit card customers based on their transactional patterns to gain insights into distinct customer segments. We utilized a dataset containing features such as purchase behavior, credit limits, payments, and balances. First, we pre-processed the data by handling missing values, scaling features, and addressing skewness. We then employed unsupervised clustering algorithms including K-means, DBSCAN, and Agglomerative Clustering to identify customer clusters. Using elbow method and silhouette scores were utilized in order to establish the ideal number of clusters and assess cluster quality. Next, we employed techniques for dimensionality reduction such as t-SNE and PCA to visualize high-dimensional data in lower dimensions. Our analysis turned up several notable customer segments characterized by different spending patterns and credit utilization behaviors. This study enhances comprehension of credit card customer behavior and can guide personalized marketing strategies tailored to different customer segments. The results emphasize how crucial consumer segmentation is to credit card business's ability to make wise business decisions and maintain client relationships.

**Keywords:** Credit card, customer segmentation, clustering, transactional patterns, marketing strategies.

**1. Introduction**

In today’s data-driven environment, especially in credit card and consumer analytics, leveraging advanced clustering and machine learning techniques to extract patterns and insights from vast amounts of data holds significant promise. As businesses strive to better understand their customers' behavior and preferences, to enable more effective target marketing, the main goal of clustering approaches is to identify unique client types and divide the consumer base into groups with similar profiles [[1]](#one). Physical and virtual fraud are the two categories into which theft of credit cards is divided.

The term "virtual fraud" refers to fraud committed by using someone else's card details online or through other means, whereas "physical fraud" refers to fraud committed using a stolen card [[2]](#two). This study delves into the implementation of a clustering algorithm tailored for credit card customers, utilizing consumer data to identify distinct customer segments characterized by their income levels, credit behavior, and payment patterns. By categorizing customers into meaningful groups, financial institutions can gain valuable insights into their clientele, enabling them to tailor their services and marketing strategies more effectively.

Machine learning is becoming an indispensable instrument for businesses, helping them to solve a range of business problems by identifying trends, connections, and correlations in data. Machine learning applications include real-time fraud detection, focused customer outreach, adherence to compliance, risk assessment, improved customer service, and study of purchase patterns [[3]](#three). Utilizing sophisticated techniques such as DBSCAN or hierarchical clustering, the research aims to unveil homogeneous customer groups with similar characteristics [[4]](#four). Customer segmentation in international banks and other organizations has been the subject of numerous studies, which have highlighted its significance for customer recruitment and management of customer relationships [[5]](#five).

The wide range of individual client preferences is a significant marketing problem for financial services. It is rare for banks to be able to tailor their services to meet the individual needs of every customer. Banks need to employ "market segmentation" tactics to get around this barrier. By using the abundance of consumer data at their disposal, they can focus on and sell them to  clients in a way that is similar to the tactics used by internet companies and networking platforms [[6]](#six).By exploring these segments, financial institutions can enhance customer service by offering personalized experiences, streamline risk assessment processes by identifying high-risk groups, and boost customer retention efforts by addressing the specific needs and preferences of different customer segments. The paper follows a structured approach, beginning with an introduction to the methodology employed. This section highlights the clustering algorithms utilized and outlines the data preprocessing techniques applied to ensure the quality and relevance of the analysis [[7]](#seven). By providing transparency into the methodology, the paper sets the stage for the subsequent analysis and findings.

The study then goes on to highlight important findings from the clustering analysis. The research clarifies the underlying patterns and traits influencing consumer behavior within each cluster by carefully analysing and interpreting the identified customer groups [[8]](#eight). Mobile financial services come with opportunities and hazards, as well as benefits and drawbacks [[9]](#nine). Additional details regarding the significance of these findings for financial institutions are provided in the discussion section.

This study's result emphasizes how important it is intended for use clustering algorithms in credit card analytics in order to obtain important data from client’s data. Financial institutions should have a better understanding of their clients by using cutting-edge approaches and tactics, which will improve customer-centric strategies and help them come to wiser conclusions.

**2. Literature Review**

Credit card transactions aimed at preventing fraud have been the subject of extensive research, with various approaches and techniques from Duman and Ozcelik [[10]](#ten) using genetic design and dispersion search solutions to overcome fraud detection challenges. Reilly and Ghosh [[11]](#eleven) conducted an assessment of Mellon Bank's viability that demonstrated the efficiency of neural networks in reducing total fraud losses by 20% to 40% Aleskerov et al. [[12]](#twelve) proposed CARDWATCH, a neural network-based data mining system especially made to identify credit card fraud, which presents a different method to address the issue.

Quah and Sri Ganesh [[13]](#thir_fout) developed a technique for detecting fraud in real time that provided outlier analysis using self-organizing mapping (SOM) and classification algorithms Panigrahi et al. [[14]](#thir_fout) proposed a fraud detection scheme with four sequential links using Bayesian learning to detect suspicious behavior. Srivastava et al. [[15]](#fifteen_nineteen) introduced a System relying on concealed Markov models (HMMs) trained on common cardholder actions for classification.

Maes and others. [[16]](#fifteen_nineteen) described a passive artificial intelligence in its Bayesian belief networks (BBN) and fraud detection system neural networks (ANN) and. Li and Liu [[17]](#fifteen_nineteen) introduced a game theory-based method for predicting attacks on protected IDS systems, demonstrating the versatility of game theories by Stolfo and and colleagues. [[18]](#fifteen_nineteen) used a combination of base classifiers such as ID3, CART, and Bayes RIPPER to enable the exchange of the fraudulent transaction information between financial institutions and to provide better classifier selection for meta-study purposes of various types. Vasta et al. [[19]](#fifteen_nineteen) developed a model to depict the engagement between an assailant and an FDS as a typical match involving a pair of participants attempting to optimize its benefit.

3. **Methodology**

**3.1 Dataset Overview**

The dataset that was used was acquired from Kaggle [[20]](#twenty_twenthree). Dataset used in this study consists of usage behavior data collected from approximately 9000 active credit card holders over a period of 6 months. This information is arranged according to the consumer and includes eighteen behavioral factors that represent different facets of credit card processing. The credit card dataset comprises the following variables.

* CUST\_ID: Credit card holder identifier (segmented).
* BALANCE: The balance in the account for the purchase
* BALANCE\_FREQUENCY: How often the balance is updated, with scores between 0 and 1(1 = frequently updated, 0 = rarely updated).
* PURCHASES: The total amount of account purchases
* ONEOFF\_PURCHASES: Maximum number of simultaneous purchases
* INSTALLMENTS\_PURCHASES: Number of purchases in installments
* CASH\_ADVANCE: Cash advance by the user
* PURCHASES\_FREQUENCY: Frequency of purchases, with scores vary on on a 0–1 scale (1 = frequently, 0 = rarely).
* ONEOFFPURCHASESFREQUENCY: One-off purchase frequency (1 = frequent, 0 = rare).
* PURCHASES INSTALLMENTS FREQUENCY: Partial purchase frequency (1 = frequent, 0 = rare).
* CASHADVANCEFREQUENCY: Frequency of cash advances
* CASHADVANCETRX: Number of transactions with "Cash in Advance".
* PURCHASES\_TRX: Number of purchase transactions made
* CREDIT\_LIMIT: Support the credit card limit for the user
* PAYMENTS: The amount the user pays
* MINIMUM\_PAYMENTS: The minimum amount the user pays
* PRCFULLPAYMENT: A percentage of the total amount paid by the user
* TENURE: When the credit card service is activated for the user

**3.2 Data Preprocessing**

Data preprocessing is an important step in preparing a dataset for modeling and analysis. Initially, we identified and controlled for the dataset's missing values to ensure data integrity [[21]](#twenty_twenthree). It removed rows that contained missing values, which are informational pieces that are not listed or do not exist for specific attributes. Next, we selected the subset of required columns such as 'BALANCE', 'PURCHASES' and 'CREDIT\_LIMIT' for further analysis.

These scores offer important perceptions of customer behavior and the credit card usage patterns,which are essential for segmentation and analysis. To deal with the skewed distribution of some numerical features, we used a log transformation [[22]](#twenty_twenthree). Skewness refers to the unequal distribution of data points. Taking the logarithm of these factors in addition to a small constant (0.1) we aimed to normalize the distribution and make the data more usable for analysis and sampling In addition, we removed unnecessary columns such as 'PURCHASES\_INSTALLMENTS\_FREQUENCY', 'ONEOFF\_PURCHASES', and 'CASH\_ADVANCE\_FREQUENCY'.

These entries were considered less relevant to the classification analysis we were performing, allowing us to focus on the most relevant features and gain actionable insights from the dataset This preprocessing ensures that the dataset is prepared has been appropriately modified, and optimized for subsequent research projects [[23]](#twenty_twenthree).

**3.3 Model Selection**

Here, in this study focusing on credit card usage behavior and customer segmentation, the clustering algorithms—KMeans, DBSCAN, and AgglomerativeClustering—were selected using specific criteria developed for dataset characteristics and research objectives

**Kmeans** was chosen because of its simplicity and efficiency in identifying circular groups, namely the concept of discrete, well-defined consumer segments driven by behavioral variables such as frequency of purchase and the number of displays corresponds to This format is computationally efficient and ideal for medium data sets [[24]](#last).

**DBSCAN** was used to overcome challenges from data sets exhibiting different densities and irregular clusters, commonly found in viable data The strength of DBSCAN resides in its capacity to detect outliers and spontaneous noise, which is important very much to detect possible fraud or unusual communication patterns embedded in data [[25]](#last).

Furthermore, a hierarchical clustering technique called **Agglomerative Clustering** adapted to analyze hierarchical patterns of consumer behavior and identify branded groups [[26]](#last).This hierarchical strategy this horizontal scale provides insights into broader consumer groups and more nuanced segments, facilitating broader surveys of credit card usage

Using these clustering techniques, this study aims to gain insights appropriate for targeted marketing strategies and fraud detection Unique strengths of each algorithm—KMeans provides clear customer segments are defined, DBSCAN for handling complex data structures, and Agglomerative Clustering for hierarchical clustering— on the underlying patterns and structures give this holistic approach credit ् by transaction It provides a strong analysis of it informs strategic decision-making regarding trading and associated risk management.

**3.4 Evaluation Metrics**

The clustering algorithm provides a quantitative measure to assess quality of clusters produced by the algorithm for clustering and helps to compare different clustering results or select the ideal quantity of clusters.

* Silhouette Score
* The most widely used analysis metric is the silhouette score [[27]](#last), and is shown that s(i). The score of each data point's silhouette, which reflects the best fit, is calculated on an intra-cluster basis

s(i) =

here,

a(i) represents mean distance between i and every additional data point from the same cluster (intra-cluster distance).

b(i) is mean distance between i and each and each data point in the nearest, surrounding cluster (inter-cluster distance).

* The Silhouette Score ranges from -1 to +1:
* A score close to +1 indicates that data point has been well-matched to its own cluster and poorly-matched to neighbouring clusters, suggesting a good clustering.
* A score close to 0 indicates that particular data point is close to the cluster-to-cluster decision boundary.
* A score close to -1 suggests that data point may be assigned to wrong cluster.
* Davies-Bouldin Index (DBI)

The quality of clustering can be assessed using the Davies-Bouldin Index (DBI) [[28]](#last). It is determined by comparing each cluster's average similarity ratio to that of its closest comparable cluster. ratio of intra-cluster to inter-cluster distances is used in order to calculate the similarity.

DBI =

Here

There are k clusters in total.

The cluster i’s centroid is located at a, and distance between each point in cluster i and its centroid is denoted by c.

* Calinski-Harabasz Index (CHI)

The Variance Ratio Criterion, or Calinski-Harabasz Index (CHI), evaluates the effectiveness of a clustering is done by taking the ratio of the total between- and within-cluster dispersions [[29]](#last).

CHI = \*

Here

A = is the spread between the cluster centroid measured by the trace between the group dispersion matrix.

B is a trace that indicates the dispersion inside clusters within the cluster dispersion matrix.

n istotal amount of datapoints.

K are clusters in total.

4. **Experimental Setup**

Used scikit-learn to develop Python for model development using clustering models such as KMeans, DBSCAN, and Agglomerative Clustering. The experimental work was performed on a Windows XI operating system, using a computing environment consisting of an Intel i7 processor with a clock speed of 4.7 GHz and 16 GB RAM This hardware configuration provided sufficient computing resources to efficiently execute clustering algorithms and consume data is properly managed to establish during the model implementation and analytical phases

KMeans, DBSCAN, and Agglomerative Clustering were selected for this study in accordance with their unique characteristics and suitability for different data collection and clustering tasks:

* KMeans: This particular algorithm was chosen because its simplicity and effectiveness in detecting spherical clusters. KMeans is suitable for datasets with well divided clusters and spherical. It also augments centralized data, making it useful for this analysis.
* DBSCAN: DBSCAN was chosen to handle pieces of data with varying density and irregular clusters. It can detect distant objects and noise automatically, making it robust to noise and able to detect clusters with any size without assuming prior awareness of quantity of clusters
* Agglomerative clustering: This hierarchical clustering method was selected so as to examine the hierarchical structure of consumer behavior and identify nested clusters. Agglomerative Clustering is useful when the underlying data structures can be grouped into clusters, allowing for detailed analysis of credit card processing patterns

The dimension reduction process focused on optimizing the dimensions of each sample to be able to achieve the best clustering performance. For KMeans, the elbow method was employed to find out the optimal number of clusters, while for DBSCAN parameters such as epsilon (ε) and minimum number of samples (MinPts) were adjusted. Dimensionality reduction using t-SNE helped to visualize the clustering results in low-dimensional space, which aided cluster interpretation.

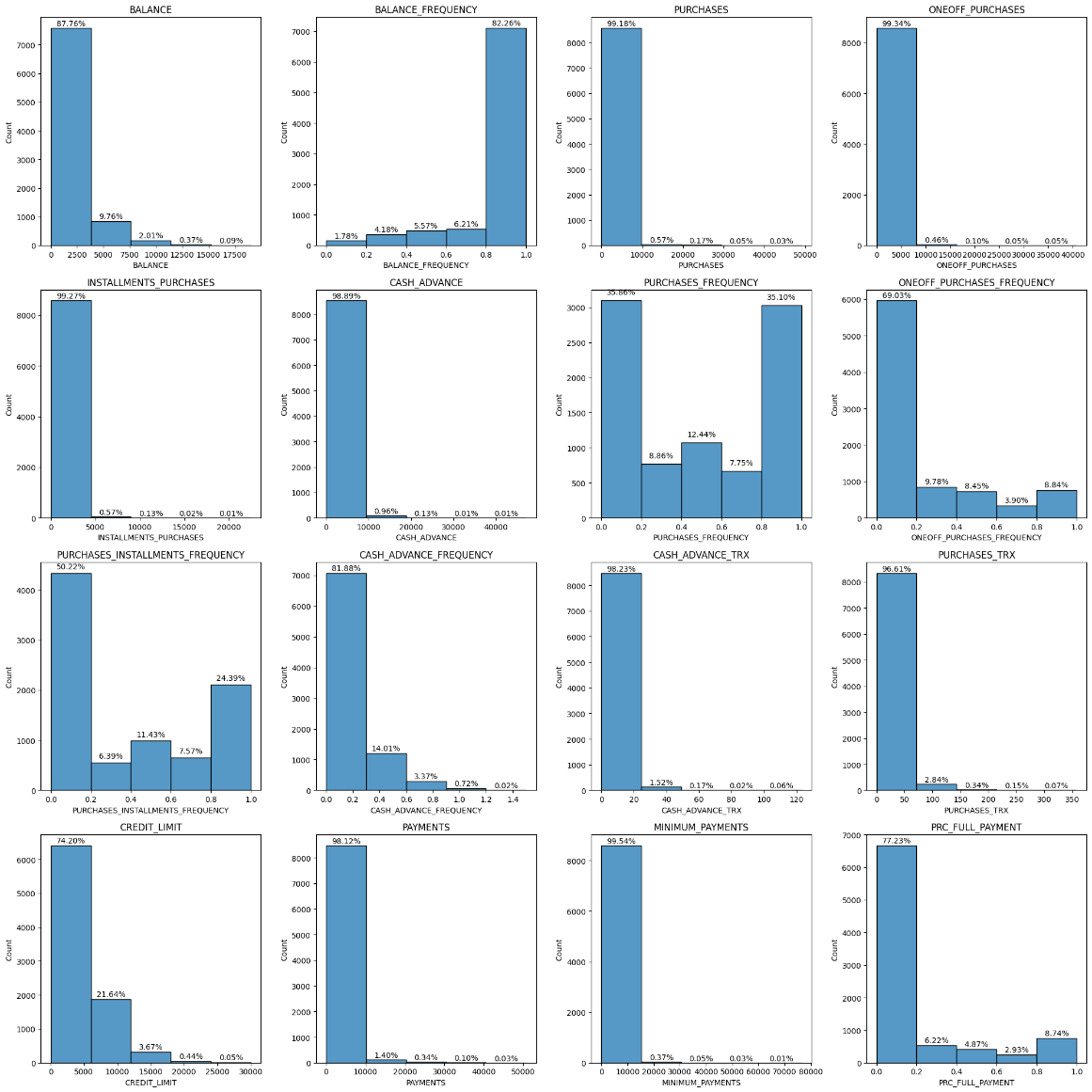
Silhouette scores were calculated to validate the clustering results and check for robustness to k.This comprehensive approach ensured that clustering models were properly selected and validated for a meaningful view of credit card usage behaviour.

Davies-Bouldin Index values were computed to access the average similarity ratio of every cluster with the cluster that best resembles it. This thorough strategy made sure that clusters were compact and well-separated, providing a meaningful view of credit card usage behavior.

Calinski-Harabasz Index values were determined to assess the ratio of between-cluster dispersion to within-cluster dispersion. This robust evaluation ensured that clustering models were optimally structured for an insightful analysis of credit card usage patterns.

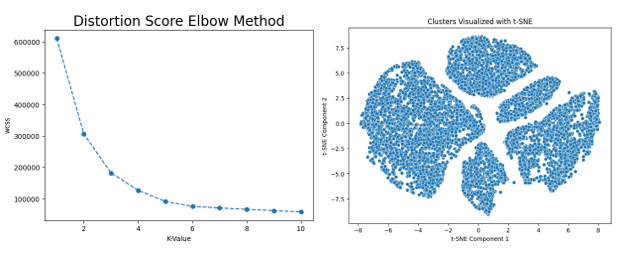
**5. Results and Discussions**

The results from the clustering experiments demonstrate the efficacy of different algorithms in partitioning the dataset into meaningful clusters.

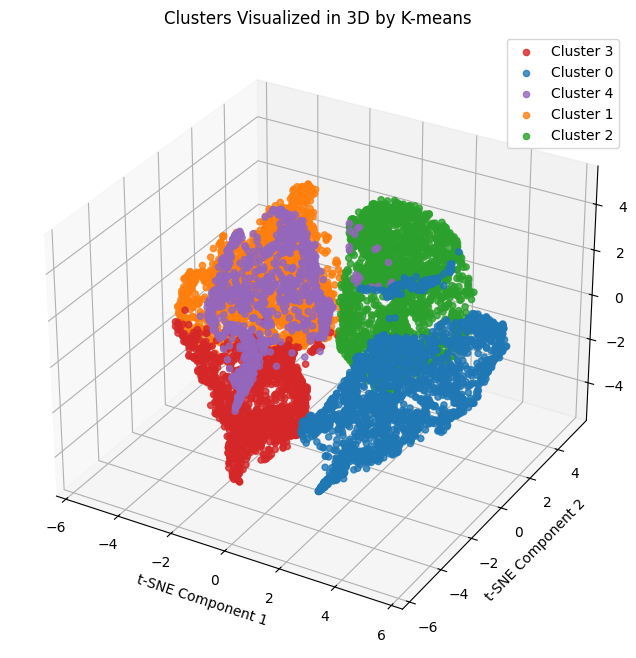
A research using a histogram to analyze client account balances found that 87.76% of users maintain balances under $3750, indicating prudent financial management and thrifty spending. Customers usually update their account balances on a regular basis, showing that they are proactive in monitoring their credit card usage and financial status. 99 percent of consumers make purchases, with the largest single transaction being around $10,000, indicating a preference for moderate spending patterns.

With 56% of customers not liking frequent installment payments and 24% indicating a fondness for them, there is a wide variety of preferences visible when it comes to installment purchases. The use of cash advances is generally discouraged, as 95% of consumers do not regularly use them. 99% of clients have, nevertheless, completed at least one cash advance transaction; the volume of these transactions ranges from 0 to 50, indicating infrequent use or unexpected financial demands.

96% of clients have completed purchase transactions, with volumes ranging from 0 to 70, demonstrating a range of purchasing activity. 74% of consumers have credit card limits that don't go over $6000, indicating a preference for moderate credit limits. Payment behavior reveals that 98% of consumers have made fewer than $10,000 in total payments, demonstrating cautious payment behavior and good credit management. The majority of consumers keep their credit card debt within reasonable bounds, with minimum installments of less than $15,000. Nonetheless, 74% of consumers would rather make partial payments or use credit cards frequently rather than pay off the entire amount owed.

Initially, K-means clustering was applied using the elbow method, revealing the ideal quantity of clusters (K) of 5on the basis of combined squared errors (WSS) of the group

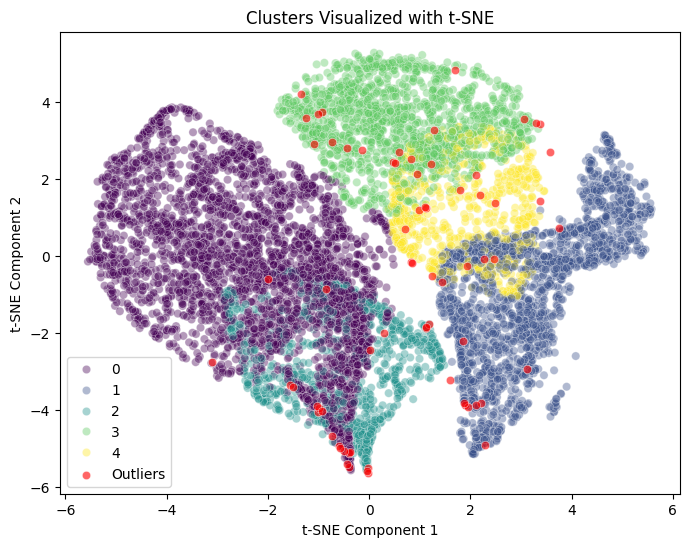
*Fig 1. Elbow method*

Subsequently, dataset was grouped using K-means with K = 5, identifying distinct clusters with meaningful separation validated by the silhouette score

*Fig 2. Clusters Visualised in 3D*

.

In addition, DBSCAN used specific parameters (epsilon=2.5, min\_samples=5), identifying clusters and outliers in the data set in excess. The calculated silhouettes reconfirmed the clusters generated by DBSCAN.

 *Fig 3. Clusters Visualised*

**Formulas to calculate Evaluation Scores:**

* Silhouette Score

s(i) =

* Davies-Bouldin Index (DBI)

DBI =

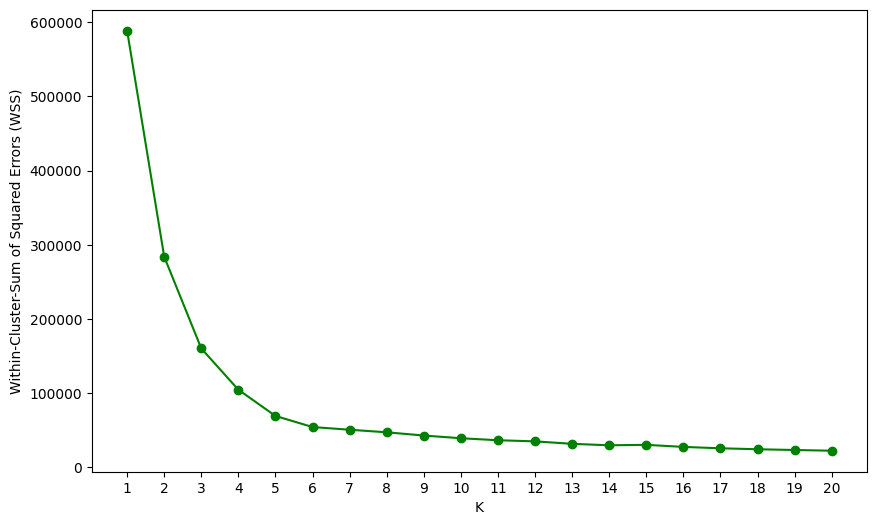
* Calinski-Harabasz Index (CHI)

CHI = \*

In addition, agglomerative clustering with K = 5 was used to reveal hierarchical cluster structures in the dataset, which were assessed by silhouette scoring.

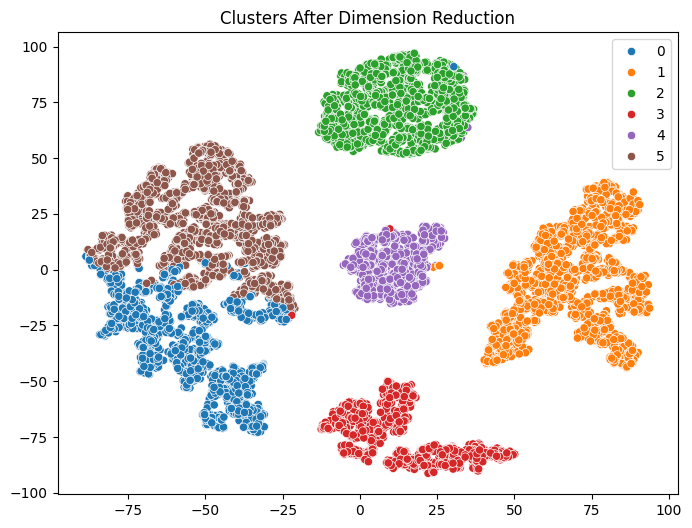
|  |  |  |  |
| --- | --- | --- | --- |
| Models | K-Means | DBSCAN | Agglomerative Clustering |
| Silhouette Scores | 0.433 | 0.546 | 0.549 |
| DBI score | 0.86 | 1.58 | 0.714 |
| CHI score | 9696.03 | 9568.81 | 12294.05 |

To refine the analysis, principal component analysis (PCA) was performed to reduce the dataset's dimensionality while retaining 95% of the variance. The reduced dataset was then subjected to K-means Clustering, confirming the ideal quantity of clusters as 6 based on post-dimensional reduction of WSS



*Fig 4. Elbow method after dimension reduction*

Comparative analysis of cluster results confirmed that K-means and aggregate clustering perform well in dividing the data set into distinct and overlapping clusters DBSCAN showed robustness in detecting outliers and clusters as they sort, it was very beneficial for noisy data sets.



*Fig 5 Clusters after dimension reduction*

Interpretation of the findings revealed that the data were carefully divided into logical groups, consistent with the research objectives of identifying underlying patterns and patterns. This insight into the data distribution facilitates subsequent analysis and decision making.

After applying the dimension reduction, the scores are

|  |  |  |  |
| --- | --- | --- | --- |
| Models | K-Means | DBSCAN | Agglomerative Clustering |
| Silhouette Scores | 0.46 | 0.462 | 0.549 |
| DBI score | 0.84 | 0.841 | 0.714 |
| CHI score | 12180.34 | 12180.34 | 12294.05 |

In our analysis, the Agglomerative Clustering model fared better than the other two clustering models. Its higher performance was demonstrated by a range of evaluation measures, demonstrating how well it handled the clustering tasks and how robust it was. Because of this, agglomerative clustering is the best option for our requirements in data analysis.

6. **Conclusion:**

In conclusion, clustering analysis using K-means, DBSCAN, and agglomerative clustering, combined with dimensionality reduction via PCA provided valuable insights into the underlying structure of data set Ne revealed internal separation revealed by K-means and agglomerative clustering Co showed strong performance in identifying different clusters, while DBSCAN effectively handled outliers and noise in data set The results were it fits well with the research objective of revealing hidden patterns and patterns in the data, providing actionable insights for subsequent research and decision-making processes. From our perspective, these results can inform future research and applications of data classification and pattern recognition

**7. References**

[1] Tripathi, S. et al., “Approaches to Clustering in Customer Segmentation.,” International Journal of Engineering & Technology, pp. 802-807, 2018.

[2] Online fraud is twelve times more common than offline fraud, June 20, 2007<http:// sellitontheweb.com/ezine/news0434.shtml>

[3] In the International Journal of Computer Technology and Application, Bhambri's article "Implementation of Data Mining Techniques for Strategic CRM Issues" was published in 2011.

[4] Chaudhry et al., "A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective." 15.9 Symmetry (2023): 1679.

[5] Hassan, "Data mining techniques for customer segmentation and profiling in retail banks." 2020.

[6] An Introduction to Banking Principles, Strategy, and Risk Management, by M. Choudhry, John Wiley & Sons, Ltd. and Moorad Choudhry, United Kingdom, 2011.

[7] "Automatic clustering algorithms: a systematic review and bibliometric analysis of relevant literature" was written by Absalom E. Ezugwu and colleagues. 2021; Neural Computing and Applications 33: 6247–6306.

[8] Michalakopoulos, Vasilis, et al. "A machine learning-based framework for clustering residential electricity load profiles to enhance demand response programs." Applied Energy 361 (2024): 122943.

[9] A Swot analysis of mobile money services using evidence from Bukavu was conducted by Kulondwa and Lukogo in their article "Financial inclusion in Africa through Mobile Money Services," which was published in the International Journal of Commerce and Finance in 2019.

[10] M.H. Ozcelik and E. Duman, “Scatter search and genetic algorithm for detecting credit card fraud,” Expert Systems with Applications 38 (2011), 13057–13063.

[11]The paper "Credit card fraud detection with a neural network" was published in 1994 by D.L. Reilly and S. Ghosh in the Proceedings of the Annual International Conference on System Science, pages 621–630.

[12]A neural network-based database mining system for credit card fraud detection, called CARDWATCH, was developed by B. Freisleben, E. Aleskerov, and B. Rao. It was presented at the Computational Intelligence for Financial Engineering conference in 1997, with pages 220–226.6.

[13] Quah, J. T. S. and Srinagesh, M., "Computational intelligence for real-time credit fraud detection," Expert Systems with Applications, 35, pp. 1721–1732, 2008.

[14] A., Sural, Panigrahi, S., Kundu, & Majumdar, A., "Dempster-Shafer theory and bayesian learning for credit card fraud detection through fusion approach," Information Fusion, pp. 354–363, 2009.

[15] "Credit Card Fraud Detcetion using Hidden Markov Model," IEEE Transactions on Reliable and Secure Computing, Vol. 5, pp. 37-48, 2008, A. Kundu, A. Shrivastava, S. Sural, and A.K. Majumdar.

[16] The paper "Credit card fraud detection using Bayesian and neural networks," published in the Proceedings of the First International NAISO Congress on Neuro Fuzzy Technologies in 2002, was written by K. Tuyls, S. Maes, B. Vanschoenwinkel, and B. Manderick.

[17] "Distributed data mining in credit card fraud detection," W. Fan, A.L. Prodromidis, P.K. Chan, and S.J. Stolfo, Proceedings of the IEEE Intelligent Systems, pp. 67–74, 1999.

[18] In the Proceedings of the IEEE International Conference on e-Technology, e-Commerce and e-Service, pp.177181, 2004, Chiu and Tsai C. "A web services-based collaborative scheme for credit card fraud detection."

[19] V. Vatsa, S. Sural, and A.K. Majumdar, "A game-theoretic approach to credit card fraud detection," in Lecture Notes in Computer Science, vol. 3803, pp. 263–276, 2005, Proceedings of the International Conference on Information Systems Security.

[20] https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

[21] Wei, Runmin, et al. "Missing value imputation approach for mass spectrometry-based metabolomics data." Scientific reports 8.1 (2018): 663.

[22] To transform or not to transform: an estimation of log models by Willard G. Manning and John Mullahy. (2001) Journal of Health Economics 20.4: 461-494.

[23] Francisco Herrera, Julián Luengo, García, and Salvador. Preprocessing data before mining it. Vol. 72. Springer International Publishing, Cham, Switzerland, 2015.

[24] Naldi, Murilo Coelho, et al. "Efficiency issues of evolutionary k-means." Applied Soft Computing 11.2 (2011): 1938-1952.

[25] Omer Karal, Ali, and Degirmenci. "Efficient density and cluster based incremental outlier detection in data streams." Science of Information 607 (2022): 901-920.

[26] Anthman, Bouguettaya et al. "Efficient agglomerative hierarchical clustering." 42.5 (2015): 2785–2797 in Expert Systems with Applications.

[27] "Silhouette analysis for performance evaluation in machine learning with applications to clustering," by Shutaywi, Meshal, and Nezamoddin N. Kachouie. Entropy 23.6: 759 (2021).

[28] Jumadi Dehotman Sitompul, Bernad, Opim Salim Sitompul, and Poltak Sihombing. "Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm." Journal of Physics: Conference Series. Vol. 1235. No. 1. IOP Publishing, 2019.

[29] Gustriansyah, Rendra, Nazori Suhandi, and Fery Antony. "Clustering optimization in RFM analysis based on k-means." Indonesian Journal of Electrical Engineering and Computer Science 18.1 (2020): 470-477.