Unleashing Customer Insights through K-Means Clustering for Enhanced Retail Decision-Making

Georgina Asuah[1], Lemdi Frank Prikutse[2]

[1] *Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana. georgina.asuah@stu.ucc.edu.gh*

[2] *Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana. lemdi.prikutse@stu.ucc.edu.gh*

Abstract: The modern retail industry now has access to vast volumes of data thanks to rising standards, automation, and technology, but the commercial decision-making process has become complicated. The utilization of Data Mining technologies for retail businesses has become indispensable for making decisions concerning sales, profit, customer satisfaction, and reduced cost. This study's foundation is segmentation principles using K-Means Algorithm in RapidMiner. This research adds to the development of useful insights into the future of Data Mining and its applications in the retail business. The results obtained from the survey indicate how retail businesses can make more informed decisions on how to keep their already customers satisfied and happy as well as how to alter factors to be more attractive to other customers.

Keywords: Data Mining, K-Means Clustering, Retail Industry, Segmentation

1. INTRODUCTION

Businesses are collecting data at a never-before-seen rate via social media, using the Internet and mobile phones. Organizations are now recognizing the benefits of using data mining. Data mining has shown to be a useful technique such that, it allows companies to uncover potentially useful information from massive amounts of data, therefore, making them gain a competitive edge. Exploring new big data investigative tools, such as Data Mining tools, is critical for the retail industry, which aspires to uncover important information from a massive amount of data to improve strategic management and consumer happiness. Today's retailers are looking for more effective marketing campaigns as they operate in a dynamic and competitive world where globalization and competition expand [1]. Customers' data collected by businesses is one of their most valuable assets because valuable information about customers can be extracted from the data, which can significantly impact how any corporate organization manages its operations and engages with both existing and potential clients, providing it a competitive edge over its rivals. Data Mining could be used to extract useful information about customers from these complex, multidimensional data.

Data mining is a collection of automated methods for finding hidden or previously unidentified pieces of information in massive databases using a variety of criteria. With the use of this newly discovered knowledge, firms may make critical business decisions that will provide them a competitive advantage [2] in the areas of decision support, prediction, forecasting, and estimating.

Mining tools are now inextricably linked to risk management and organizational decision-making [3]. A detailed examination of the various retail functional areas reveals that Business Intelligence (BI) performs a significant part in nearly all of them. It can give businesses fresh and often astonishing insights into consumer behavior, helping them to better meet their constantly changing demands. BI can assist retailers in identifying their top suppliers and figuring out what differentiates them from the competition on the supply side. It can assist retailers acquire a better insight into their inventory, and how it moves, as well as improve storefront operations with improved category management. BI by providing a variety of studies and reports, enhances internal organizational support tasks such as finance and human resource management [4]. Data Mining is useful in various sectors but in this paper, data mining application in the retail sector is covered.

2. REVIEW OF LITERATURE

Due to the importance of the topic, academics have built numerous analytical models throughout the years to investigate the issue of data mining in the retail industry and its advantages. The relevant research in the literature employs both shallow learning techniques and deep learning approaches. In a paper written by [3], a Clustering algorithm was used to segregate customer profiles where they discovered that customer data in the retail industry is kept and used to make informed decisions. The findings revealed that profits for businesses are raised, but also data mining issues in terms of determining how likely customers will choose products recommended to them and how preferable the platform or store is with the focus on customer interests are presented.

According to [5], data mining has arisen as a method for identifying trends to improve tactics and judgments. They reviewed the basic duties involved in data mining, as well as the use of data mining in various industries, with a focus on the retail business, to demonstrate how data mining can be used to enhance marketing campaigns. The use of data mining in the retail industry and its benefits and drawbacks for customers are highlighted specifically in [2]. Some of the customer-based retail businesses making use of customer data were considered like Master Card, Walmart, Just for Feet, and Burger King, and how well it is benefiting them. Their paper delves into the sorts of data required for data mining applications in customer-based enterprises, such as customer data: 1) demographics; 2) economic status; 3) geographic details.

High-quality, usable data that can be used by businesses, such as that provided by business intelligence (BI) solutions like data warehousing, data mining, and OLAP, is valued in the information economy [4]. The objective is to find general business intelligence solutions for the business sector. According to the authors, business intelligence (BI) can give businesses fresh and often astonishing knowledge of consumer behavior, enabling them to better respond to their customers' constantly shifting demands. BI can assist retailers in identifying their top vendors and figuring out what makes them stand out from the competitors on the supply side. It can assist retailers acquire a better insight into their inventory, and how it moves, as well as improve storefront operations with improved category management. BI by providing a variety of studies and reports, enhances internal organizational support tasks such as finance and human resource management. [6] sets out an overview of Business Intelligence, as well as the development and use of a BI system in the retail industry, as well as the fundamental technology of Business Intelligence. Business domain and dimension design, ETL tool design, and Data Presentation middleware design and the main innovation. were the system's main features. Retailers are acknowledged for innovation. People who use business information to gain a sustainable competitive edge are the most inventive retailers of today. The knowledge gained through evaluating a significant amount of data will aid in achieving each aspect of the organization's objectives. [7].

In Customer relationship management and data mining in organized banking and retail, a critique of the concept is presented by [8]. Machine Learning methods such as classification, decision trees, and clustering were also studied, as well as Data Mining applications in banking and retail industries. They claim that data mining may help retailers make proactive decisions by providing data on customer buying habits and preferences, product sales trends, supplier-delivery efficiency, seasonal changes, peak customer traffic hours, and other information. Time series data mining is specifically applied in [9] in a real-world setting. Provided with retail grocery store chain data by Dunnhumby, the authors employed a time series data mining process called dynamic time warping (DTW) to analyze the data. Retail marketers are capable of making well-informed decisions on advertising and promotion strategies by identifying products with similar purchase histories and transactional commonalities.

In the retail industry, data mining techniques are widely used. Businesses can use this process to find groups of observations and records that have similar buying habits. A study done by [10] specifically identifies marketing, fraud detection, risk management, and customer acquisition and retention as significant data mining applications. Data mining in the area of marketing allows a company to go through massive amounts of client data to target the correct customers. Data mining supports risk management by estimating the number of customers who will be lost to competitors, allowing the company to plan. The relevance of data mining for fraud detection is key because it allows an organization to avoid losing large sums of money by discovering fraudulent activities. Data mining aids customer acquisition and retention, it enables an organization to identify new customers while maintaining the ones.

[11] offer a way for extracting customer visit categories from basket sales data using the business analytics method. According to the authors, a client's visit is defined by the purchased product categories in the basket and the shopping

purpose or goal that inspired the visit. They also offered a semi-supervised feature selection technique that uses product taxonomy as an input and outputs customized categories. The practicality of the technique was demonstrated clearly by using a real-life example incorporating a well-known European fast-moving consumer goods (FMCG) store.

In recent years, deep learning approaches have drawn a lot of attention due to their capacity to create better and more accurate predictions regardless of the data's complexity and dimensionality. In this paper, a revolutionary deep learning technique is described for forecasting sales in the fashion industry, which projects the sales of new individual products in subsequent seasons [12]. The authors compared the sales estimates acquired through deep learning techniques against a variety of shallow approaches, including Linear Regression, Decision Trees, Support Vector Machines, and Random Forest. Although it was discovered that the deep learning model performed well at predicting sales in the fashion retail market, it did not outperform some of the shallow algorithms, such as Random Forest, for all the assessment criteria examined.

The banking sector has adopted Data Mining in the majority of their systems employing multiple DM methods, with clustering and classification demonstrating adequate practicality and attractiveness, primarily for fraud detection, customer relationship management (CRM), and risk management. [13].

According to the RFM model (Recency, Frequency, and Monetary), [14] deployed K-Means concepts for dataset segmentation where they implemented and applied the scientific technique on real-time transactional and retail datasets to help increase sales and profits for businesses, as well as provide highly advanced insights on customer purchasing behavior and routines. Supporting this, [15] specified the RFM pattern and establish a new technique for discovering a thorough collection of RFM patterns that can replicate the RFM-customer pattern collection without any need for buyer identification data. Authors offer an RFM-pattern-tree tree structure for compressing and storing the whole transactional databases and RFMP-growth, a pattern growth-based technique for discovering all RFM-patterns in an RFM-pattern-tree. The suggested technique is efficient and can adequately determine most RFM-customer patterns, according to the results of their studies.

For informed decisions concerning risks and their management in the retail sector, a DM-based framework is developed for the Detection of risk indicators, collecting and storing of risk data, the transformation of the risk management challenges into DM issues, data processing using DM methodologies, and interpretation of findings to propose intelligent risk reduction techniques [16]. A case study relying on semi-structured interviews, dialogues, and focus group research is used to validate the concept. The evidence found shows how DM can assist you in finding hidden and relevant information in unstructured risk data to make better risk management judgments.

[17] proposes a comprehensive data mining method in order to create marketing approaches as well as forecast demand for refurbished goods in the Indian market. Optimization methods were employed to examine real-world datasets from three randomly chosen e-commerce websites to validate their scheme.

3. EVALUATION PARAMETERS

This chapter explains the proposed analytics strategy, which uses clustering methods to achieve customer segmentation, as well as the exploratory design for the study's anticipated outcome.

3.1 Selection of Clustering algorithm

A cluster can be considered a theoretically significant collection of items with comparable features. Customers can be segmented through clustering for further study. Customer segmentation is among the functionalities of K-means, according to the literature review [18]. A simulated partition clustering method known as the K-Means algorithm determines the number of clusters that the user specifies based on the matching centroids. K-Means performs well with large datasets and is computationally efficient when compared to other clustering methods. It also reduces the amount of data that is misclassified. This work uses the K-Means algorithm in RapidMiner.

3.2 Selection of Number of Clusters

The elbow approach was used to determine how many clusters to use for the k-means clustering. The primary goal of k-means clustering is to construct clusters with a minimum total Within-Cluster Sum of Squares (WCSS). The clustering's compactness is determined by the total WCSS. The appropriate number of clusters can be determined using the Elbow approach by considering the total WCSS, which is dependent on the number of clusters [19]. Using the chosen dataset, a WSS curve was drawn for various K values. As illustrated below, the plotted curve resembled an elbow. Typically, the position of a bend in the plot is used as a guide to determine the optimal number of clusters, beyond which the addition of another cluster does not significantly improve the overall WSS. The curve determined that five clusters were the ideal number.

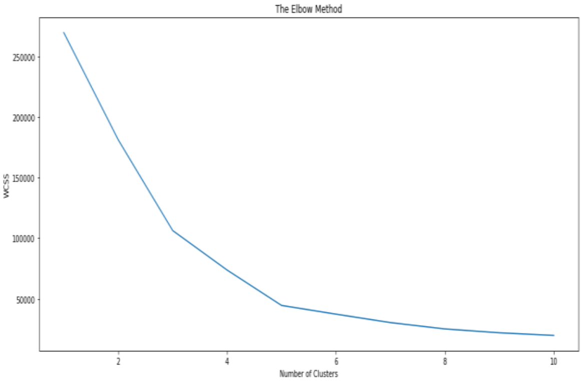


Fig 1. Number of Clusters Determination

3.3 Proposed Methodology

The proposed approach can be broadly divided into 4 steps as shown in Figure 2. The details of the steps are elaborated below:

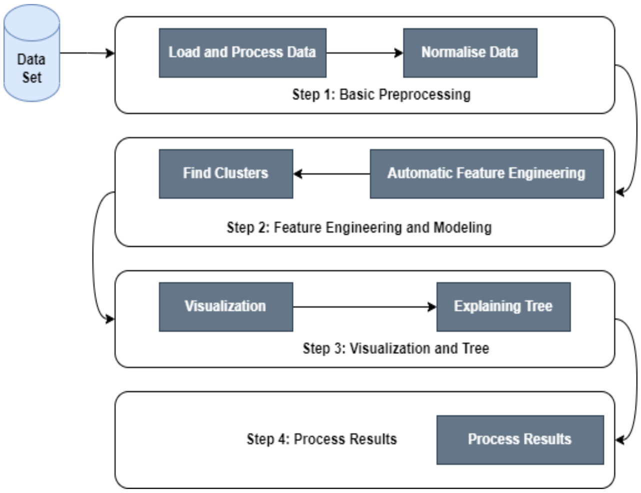

Data Set
Load and Process Data
Normalise Data
Step 1: Basic Preprocessing
Find Clusters
Automatic Feature Engineering
Step 2: Feature Engineering and Modeling
Visualization
Explaining Tree
Step 3: Visualization and Tree
Step 4: Process Results
Process Results

Fig 2: Steps of Proposed Model.

3.3.1 Basic Preprocessing

In this step, the dataset is loaded and some basic preprocessing tasks are performed. Afterward, it delivers all labeled data points as well as unlabeled ones for which the model should be applied later. The data is then normalized and the normalization model is remembered so that we can later transform the data back.

3.3.2 Feature Engineering and Modeling

Automatic feature selection is performed if desired. This is a unique approach using multi-objective optimization and the concept of information preservation. The actual clustering is performed on the transformed data in this step.

3.3.3 Visualization and Tree

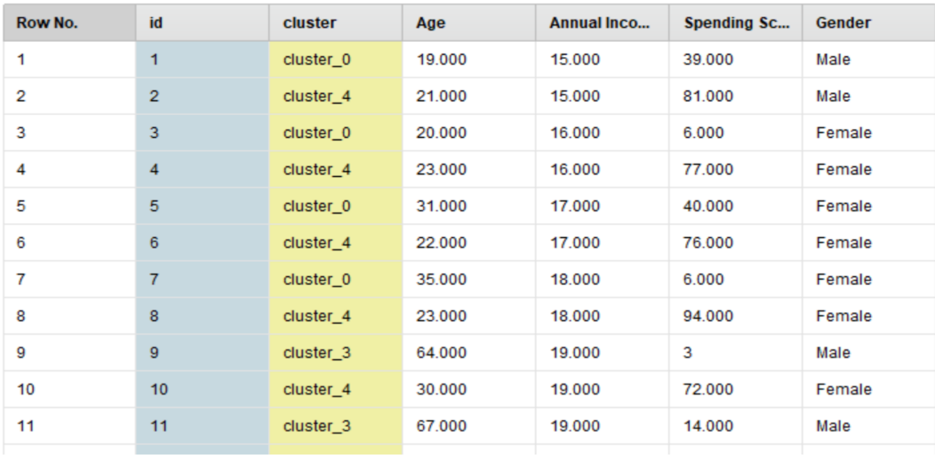
Creation of the visualizations for the cluster model is created in this step. De-normalization of the data and creation of a decision tree, explaining which data points belong to which cluster is also performed.

3.3.4 Process Results

This step helps in processing the results for display. We recall all the clusters, data, optimized features, decision trees, and annotations generated.

3.4 Experiments, Findings, and Discussions

The proposed method was evaluated using a fictitious dataset of client transactions collected in a mall over some time, Kaggle's repository. The dataset contains 200 entries of consumer purchase information with the following attributes; Customer ID, Gender, Age, Annual Income (k$), and Spending Score (1-100). After preprocessing the data, apply customer segmentation analysis and K-Means clustering.



Row No.	id	cluster	Age	Annual Inco...	Spending Sc...	Gender
1	1	cluster_0	19.000	15.000	39.000	Male
2	2	cluster_4	21.000	15.000	81.000	Male
3	3	cluster_0	20.000	16.000	6.000	Female
4	4	cluster_4	23.000	16.000	77.000	Female
5	5	cluster_0	31.000	17.000	40.000	Female
6	6	cluster_4	22.000	17.000	76.000	Female
7	7	cluster_0	35.000	18.000	6.000	Female
8	8	cluster_4	23.000	18.000	94.000	Female
9	9	cluster_3	64.000	19.000	3	Male
10	10	cluster_4	30.000	19.000	72.000	Female
11	11	cluster_3	67.000	19.000	14.000	Male

Fig 3: Clustered Data.

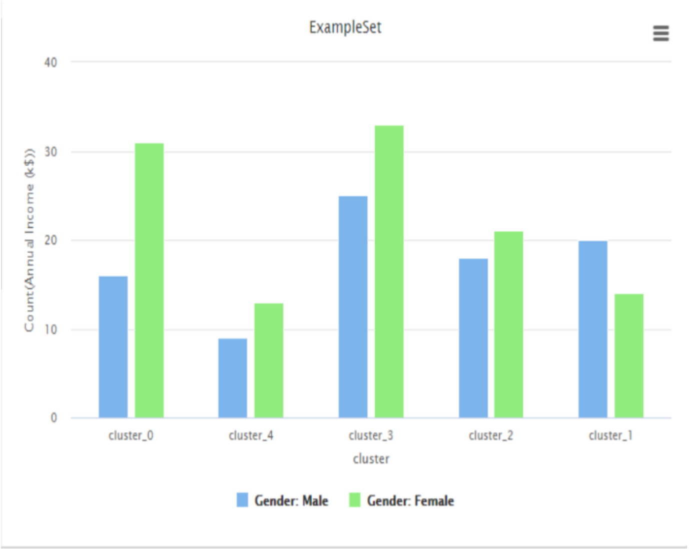


Fig 4: Visualization of Annual Income based on Clusters colored by Gender.

Figure 4 shows the amount of income earned annually by males and females in different clusters. Both males and females who earn the highest are in cluster 3. Likewise, those who earn less in both genders are in cluster 4.

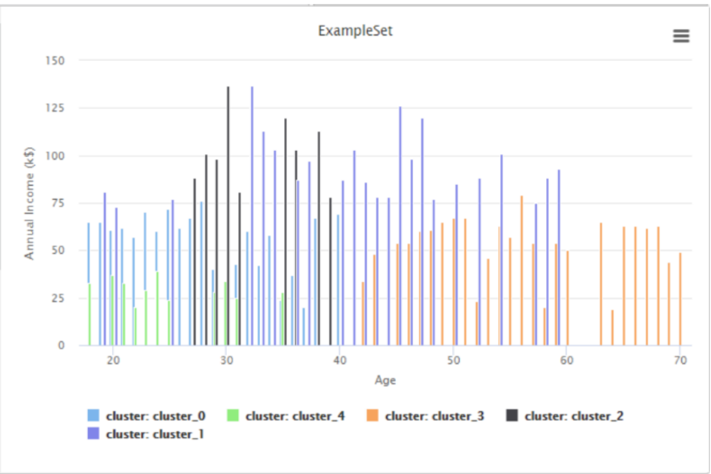


Fig 5: Visualization of Annual Income based on Age.

Figure 5 represents the annual income in terms of age. It can be seen that the people who earn more fall in the range of 27 years to 59 years of age.

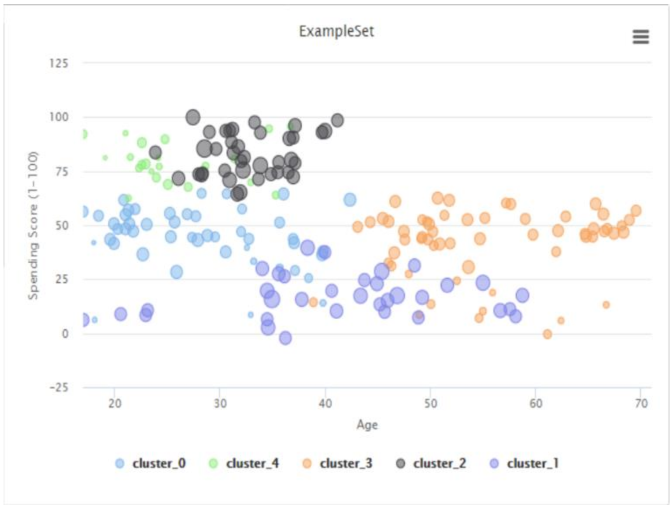


Fig 6: Visualization of Spending Score based on Age.

Figure 6 shows the Spending Score according to age. From this figure, it can be noted that customers who spend more come in the age range of 17 to 40 years of age compared to others.

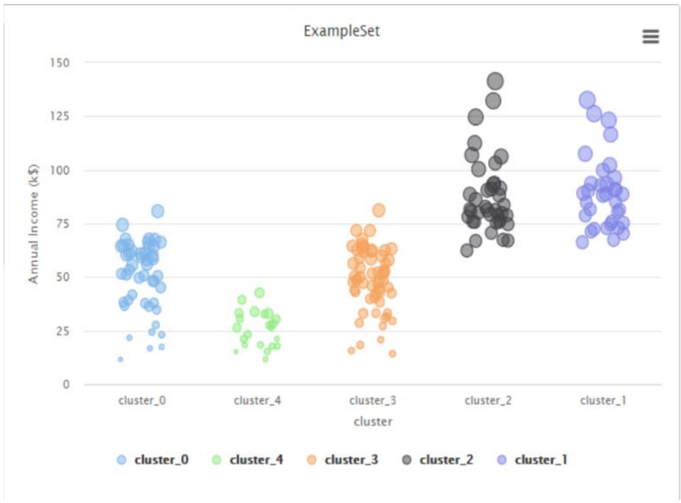


Fig 7: Visualization of Clusters by Annual Income.

Figure 7 gives a general overview of every cluster based on their annual income. Based on this we can see that customers in clusters 2 and 1 earn the most as against customers in other clusters. Customers in cluster 4 earn the least.

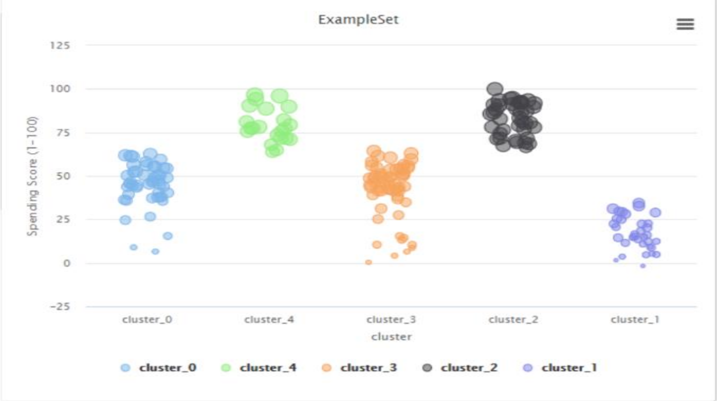


Fig 8: Visualization of Clusters by Spending Score.

Figure 8 also displays an overview of the spending habits of customers in different clusters. Customers in clusters 2 and 4 spend the most. Customers in cluster 1 spend the least.

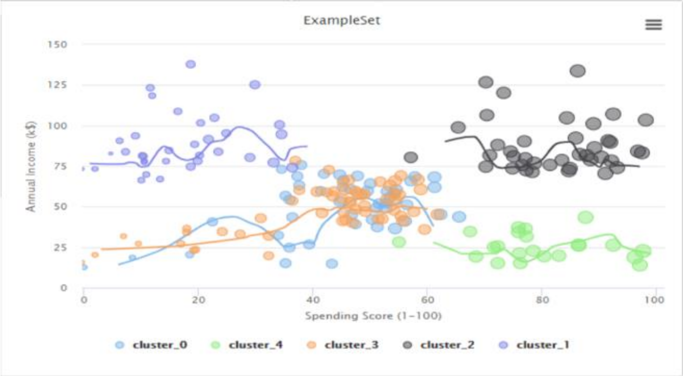


Fig 9: Visualization of Clusters; Annual Income and Spending Score.

Figure 9 represents the clusters of spending scores against annual income. The regression interpolation used is loess.

Cluster 4 shows consumers with little income but high spending scores; these are consumers who enjoy buying items more frequently despite their limited income. Possibly it is due to these clients being happy with what the mall offers. These customers will not be lost, even if the mall chooses not to pursue them.

People in cluster 2 have high incomes and high spending rates, which is advantageous for malls as these people are their main source of revenue. These patrons may have been enticed by the mall's amenities as a result they may be regular mall shoppers.

Interestingly, people in Cluster 1 have high incomes but low expenditure scores. They can be the patrons who are unsatisfied with the mall's offerings. These are perhaps the mall's main goals because of their large potential for sales. Consequently, mall officials will attempt to provide additional amenities to cajole these customers and satisfy their expectations.

Whereas customers in cluster 0 and cluster 3 are averaging in their income and spending.

4. CONCLUSION AND FUTURE SCOPE

The overall purpose of this study is to probe deeply into and grasp consumers' buying habits, allowing merchants to give them suitable services that are personalized to their needs. The use of clustering methods to identify consumer subgroups has been recommended as a business analytics method. Data from shopping malls were examined and groupings of customers with specific information and buying behaviors were created. Then, consumer segments were created, and shopping behaviors were assigned to every segment. Another big hurdle for both offline and online businesses is customer retention. Finally, we can derive from the machine learning approach that, to boost the mall's profitability, the mall's authorities must focus their attention on customers in clusters 1 and 0, while simultaneously maintaining their standards to retain those in clusters 2 and 4 happy and fulfilled.

The comprehensive study of clients from many walks of life will be the focus of future research in this area. Other business parameters, including the sort of items bought, and the most effective sales strategy used during a certain event, might be researched to build beneficial company enhancements. Retailers will be able to increase sales by offering discounts and establishing distinctive strategies that will give them a competitive advantage over their rivals thanks to such advancements and talks in this field.

5. REFERENCES

[1] Onestepretail, “6 Biggest Challenges Retailers Face Today,” *One Step Retail Solut.*, vol. 85027, no. May, pp. 2–26, 2018.

[2] S. R. Ahmed, “Applications of Data Mining in Retail Business,” 2004.

[3] P. Agarwal, “Benefits and Issues Surrounding Data Mining and its Application in the Retail Industry,” vol. 4, no. 7, pp. 1–5, 2014.

[4] H. Li, “Applications of data warehousing and data mining in the retail industry,” *2005 Int. Conf. Serv. Syst. Serv. Manag. Proc. ICSSSM’05*, vol. 2, pp. 1047–1050, 2005, doi: 10.1109/ICSSSM.2005.1500153.

[5] B. M. Ramageri, “ROLE OF DATA MINING IN RETAIL SECTOR,” vol. 5, no. 01, pp. 47–50, 2013.

[6] T. Gang and C. Kai, “The Research & Application of Business Intelligence System in Retail Industry,” no. September, pp. 87–91, 2008.

[7] I. J. Of, “RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS SURVEY OF BUSINESS,” vol. 4, no. 9, pp. 13–17, 2016.

[8] P. S. Raju, V. R. Bai, and G. K. Chaitanya, “Data mining : Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries,” pp. 2650–2657, 2014.

[9] B. Anderson, J. M. Hardin, C. College, and B. Administration, “Time Series Data Mining :,” vol. 1, no. December, pp. 51–68, 2014, doi: 10.4018/ijban.2014100104.

[10] P. Taylor, A. M. Hormozi, and S. Giles, “DATA MINING : A COMPETITIVE WEAPON FOR BANKING AND,” no. August 2014, pp. 37–41, 2006, doi: 10.1201/1078/44118.21.2.20040301/80423.9.

[11] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, “Retail business analytics: Customer visit segmentation using market basket data,” *Expert Syst. Appl.*, vol. 100, pp. 1–16, 2018, doi: 10.1016/j.eswa.2018.01.029.

[12] A. L. D. Loureiro, V. L. Miguéis, and L. F. M. da Silva, “Exploring the use of deep neural networks for sales forecasting in fashion retail,” *Decis. Support Syst.*, vol. 114, pp. 81–93, 2018, doi: 10.1016/j.dss.2018.08.010.

[13] H. Hassani, X. Huang, and E. Silva, “Digitalisation and big data mining in banking,” *Big Data Cogn. Comput.*, vol. 2, no. 3, pp. 1–13, 2018, doi: 10.3390/bdcc2030018.

[14] P. Anitha and M. M. Patil, “RFM model for customer purchase behavior using K-Means algorithm,” *J. King Saud Univ. - Comput. Inf. Sci.*, 2020, doi: 10.1016/j.jksuci.2019.12.011.

[15] Y. H. Hu and T. W. Yeh, “Discovering valuable frequent patterns based on RFM analysis without customer identification information,” *Knowledge-Based Syst.*, vol. 61, pp. 76–88, 2014, doi: 10.1016/j.knosys.2014.02.009.

[16] M. Er Kara, S. Ü. Oktay Fırat, and A. Ghadge, “A data mining-based framework for supply chain risk management,” *Comput. Ind. Eng.*, vol. 139, 2020, doi: 10.1016/j.cie.2018.12.017.

[17] Dr. V. Suma, “Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics,” *J. Soft Comput. Paradig.*, vol. 2, no. 3, pp. 153–159, 2020, doi: 10.36548/jscp.2020.3.002.

[18] D. Arunachalam and N. Kumar, “Benefit-based consumer segmentation and performance evaluation of clustering approaches: An evidence of data-driven decision-making,” *Expert Syst. Appl.*, vol. 111, pp. 11–34, 2018, doi: 10.1016/j.eswa.2018.03.007.

[19] Trupti M. Kodinariya and Dr. Prashant R. Makwana, “Review on determining the number of Cluster in K-Means Clustering,” *Int. J. Adv. Res. Comput. Sci. Manag. Stud.*, vol. 1, no. 6, pp. 90–95, 2013.

DOI: https://doi.org/10.15379/ijmst.v10i1.2616