Machine Learning: Data-driven Customer Segmentation and Churn Prediction

***Abstract*—**

I. Introduction

In the modern business landscape, customer segmentation has emerged as a critical tool for organizations to effectively manage their customer base. By dividing customers into distinct groups based on shared characteristics and behaviors, businesses can tailor their marketing strategies, product offerings, and customer service initiatives to specific segments, leading to enhanced customer satisfaction, loyalty, and profitability.Machine learning can be used to segment customers into different groups based on their purchase history, demographics, and other factors. This information is used to personalize the customer experience, such as by sending targeted marketing messages and recommending products also other things.Customer segmentation is one of the best systems for data analysis project as well as it is a best for any industries for data analysis. With customer segmentation can helps any industry for their better business sta In machine learning clustering groups similar data points together without any prior knowledge of the data. This data is labeled and the algorithms learn to predict the labels for new data points.

Among the various techniques employed for customer segmentation, unsupervised learning algorithms, particularly k-means clustering, have gained prominence due to their ability to identify patterns and groupings in data without the need for predefined labels or target variables. In the banking sector, where customer churn, or the loss of valuable customers, poses a significant threat to profitability, k-means clustering has proven to be particularly effective in identifying at-risk customer segments.

In addition to unsupervised learning methods, supervised learning algorithms, such as support vector machines (SVM), random forests, and naïve Bayes classifiers, have also been widely adopted for customer segmentation tasks. These algorithms utilize labeled data to learn the relationships between customer characteristics and their propensity to churn, enabling them to predict which customers are most likely to leave the bank.Last but not least, the result of customer segmentation is applied to customer-churn model and gotten accuracy list of lost customer. Experiment proves that this method can obtain a satisfactory result of customer-churn [1].

The application of k-means clustering and supervised learning algorithms to customer segmentation in the banking sector offers several key benefits. Firstly, it allows banks to identify distinct customer segments based on their shared attributes and behaviors, providing a deeper understanding of their customer base. This knowledge can be utilized to develop targeted marketing campaigns, product offerings, and customer service strategies that resonate with specific segments, leading to improved customer engagement and retention. Secondly, customer segmentation enables banks to proactively identify at-risk customer segments, allowing them to implement intervention strategies to prevent churn. By understanding the common characteristics and behaviors of customers who are likely to leave, banks can address their concerns and address any underlying issues, potentially retaining valuable customers and minimizing churn rates.Finally, customer segmentation can be used to optimize resource allocation and marketing spending. By focusing their efforts on the most valuable and at-risk customer segments, banks can ensure that their resources are utilized effectively, maximizing the return on their marketing investments.

The choice of customer segmentation algorithm depends on the specific data available and the business objectives. K-means clustering is well-suited for exploratory data analysis and identifying patterns in unlabeled data, while supervised learning algorithms are more effective when labeled data is available and the goal is to predict customer churn or other specific outcomes.In summary, customer segmentation has become an essential tool for banks to manage their customer base effectively, enhance customer satisfaction and loyalty, and minimize churn rates. The application of unsupervised learning algorithms like k-means clustering and supervised learning algorithms like SVM, random forests, and naïve Bayes classifiers provides banks with valuable insights into their customer base and enables them to develop targeted strategies to retain and grow their customer relationships.

II. Related Works

Kansal et al. [2]The paper focused on the application of clustering algorithms to segment customers based on their behaviors and patterns so the authors implemented three clustering algorithms -k-Means, Agglomerative, and Meanshift. Last, compared the results obtained from each algorithm. Their data collection from a local retail shop consisting of two features- average number of visits to the shop and average amount of shopping done on a yearly basis .This resulted in the formation of five segments labeled as Careless, Careful, Standard, Target, and Sensible customers, with two additional clusters emerging from mean shift clustering labeled as-
1.High buyers and frequent visitors.
2.High buyers and occasional visitors.

Yadegaridehkordi et al. [3]The research collected data from the popular online tourism review website TripAdvisor by randomly selecting crawling 14,525 reviews and 50 eco-friendly hotels in Malaysia and ratings from 2018 to

979-8-3503-1600-1/23/$31.00 ©2023 IEEE

2019. Model they used k-means clustering, TOPSIS, and CART techniques to segment travelers, rank eco-friendly hotel criteria, and predict travelers based on past online reviews.The proposed method had three main stages: clustering the online reviews, ranking the eco-friendly hotels' features, and predicting travelers' choice preferences.The study used a text mining approach to find keywords in online reviews, but should consider using complementary methods .

Monil et al. [4]The paper Customer Relationship Management (CRM) in E-commerce Enterprises and the use of clustering analysis identifies different customer characteristics and apply marketing strategies accordingly. This paper's hybrid combination of clustering algorithms is discussed. The DBSCAN algorithm was mentioned which separates clusters of high density from clusters of low density.Hierarchical clustering was mentioned which analysis that builds a hierarchy of data points . CRM for enhancing customer service satisfaction and developing long-term customer-company relationships.

The paper by Dullaghan et al. [5] was focused dynamic customer segmentation analysis in the telecommunications industry using o machine learning techniques. The authors used the C.5 algorithm within naive Bayesian modeling to segment telecommunication customers based on their billing and socio-demographic aspects.Also they used Data mining techniques are employed to analyze customer behavior and create customer profiles.

In 2013, Smeureanu et al. [6] developed customer segmentation of private banking sector using machine learning Technique. This paper approaches two of the most popular machine learning techniques, Neural Networks and Support Vector Machines, and describes how each of these perform in a segmentation process.This model led to an overall detection rate of 98% on the training set and 97% on the test set.Both machine learning techniques performed well in the segmentation process.

The paper by Hung et al. [7] focused on customer Segmentation Using Hierarchical Agglomerative Clustering (HAC). It groups customers by different categories like age, location, spending habit and so on. It discovered that HAC had achieved a high degree of accuracy, correctly classifying over 90% of the customers. This study uses the hierarchical agglomerative clustering algorithm on a Credit cards dataset to perform customer segmentation.

In 2021, Shen [8] proposed a customer segmentation framework for e-commerce that utilized unsupervised machine learning techniques. The framework, in the past, generated behavioral features for each customer through the RFM model, extracted product categories from product descriptions using the TF-IDF method, and grouped customers into meaningful segments using the K-means clustering algorithm. After conducting an evaluation on a real-world dataset, the author discovered that the framework had successfully identified four distinct customer segments with an accuracy of 92.5%.

In 2020, Abidar et al. [9] proposed a customer segmentation approach that relied on machine learning algorithms, asserting that traditional methods were no longer adequate in the era of big data. They employed a four-step methodology to segment their customers: data collection, preprocessing, feature selection, and clustering. To assess the accuracy of their model, they used the Silhouette score and, in the past, achieved a commendable level of segmentation. Additionally, they discussed how their approach could be employed to develop targeted marketing campaigns.

Alghamdi's 2023 [10] article in the Arabian Journal for Science and Engineering introduces a hybrid method for customer segmentation in Saudi Arabian restaurants. The study, grounded in the dynamic field of customer segmentation, innovatively combines traditional clustering methods with advanced technologies like neural networks and optimization learning techniques. This approach, tailored to the specific context of Saudi Arabia, acknowledges the cultural and market intricacies of the region. Alghamdi's work contributes to the broader literature on customer segmentation in the restaurant industry, offering a nuanced and context-specific methodology with potential implications for enhancing marketing strategies and customer satisfaction.

Tabianan, Velu, and Ravi's 2022 article [11] in Sustainability introduces a K-means clustering approach for intelligent customer segmentation using purchase behavior data. This reflects a broader trend in leveraging data-driven methodologies for nuanced customer targeting. Emphasizing sustainability aligns with contemporary business interests. The work adds to existing literature exploring clustering techniques for dynamic customer behavior understanding. The application of the K-means algorithm showcases its potential in uncovering meaningful customer segments based on purchase behavior, contributing to ongoing discussions on intelligent customer segmentation strategies. Future research is expected to delve deeper into advanced techniques and industry-specific adaptations for effective customer engagement.

The article "Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering" by Nouri and Sabri (2022) [12] examines the efficacy of unsupervised machine learning techniques for customer segmentation. The authors compared the performance of clustering ensemble and spectral clustering methods and found that spectral clustering performs better in terms of accuracy and consistency. Overall, the research contributes to advancing the understanding of efficient customer segmentation through innovative unsupervised machine learning approaches.

The author's research [13] indicated that K-means clustering proved to be a more fitting algorithm for customer segmentation compared to hierarchical clustering, particularly in scenarios involving large and intricate datasets. Additionally, PCA served as an effective validation tool for both K-means and hierarchical clustering methods. These findings held significant implications for businesses. Utilizing K-means clustering allowed them to gain deeper insights into customer needs and preferences. This valuable information

could then be harnessed to create more precise marketing strategies and enhance the overall customer experience.

The authors introduced a novel approach to customer classification and segmentation using the RFM model [14], which considered the customer-product relationship. Their method achieved an impressive customer classification accuracy exceeding 97%. This approach offered businesses valuable insights into customer-product interactions, enabling them to create more precise marketing campaigns and enhance the overall customer experience.

The authors introduced a novel approach to customer churn prediction in e-commerce, employing K-means clustering and SVM algorithms [15], which yielded higher prediction accuracy compared to traditional models. This approach provided e-commerce businesses with the ability to accurately identify customers at risk of churning, enabling the development of targeted marketing campaigns for customer retention. As a result, substantial cost savings and enhanced customer loyalty were potential advantages for these businesses.

The authors presented a novel approach to comprehending and forecasting customer acceptance of circular business models [16] by integrating machine learning and simulation modeling. Their method accurately predicted customer preferences for various circular business models and revealed that a majority of customers preferred circular models over traditional linear ones. The approach offered several potential advantages for businesses and policymakers. Businesses could enhance their marketing strategies and create appealing circular product offerings by gaining insights into customer perceptions and evaluations of circular business models. Policymakers could utilize this approach to identify and promote successful circular business models within the market.

In 2013, an composition was published [17] in the International Journal of Computer Theory and Engineering that excavated into the innovative conception of client segmentation. The composition stresses the significance of dividing guests into specific groups grounded on identical characteristics in order to produce effective marketing strategies. The study uses complex clustering algorithms and data mining ways to produce homogeneous client orders, in turn streamlining business operations. Through this groundbreaking exploration, advanced logical tools are brought together with client relationship operation to increase the appreciation of the relationship between the two. The ultimate thing is to perfect strategies for reaching peak client satisfaction and ameliorate overall business issues.

This 2020 IEEE composition introduces a data- driven client segmentation strategy for power systems [18]. Penned by Yuan, Dehghanpour, Bu, and Wang, the study emphasizes guests' places in system peak demand. using a quantitative approach, the exploration likely proposes a system to classify guests grounded on their impact on peak demand. Published in the IEEE Deals on Power Systems, the composition contributes to optimizing power distribution strategies by acclimatizing approaches to different client parts, eventually enhancing the overall effectiveness of the power grid.

Published at the 2020 IEEE ICIIS Conference, Nandapala and Jayasena presented a practical approach to client segmentation using the K- means algorithm [19]. The paper fastening on artificial information systems is likely to bandy the use of K-means and its effectiveness in client segmentation. This work contributes precious perceptivity into the effectiveness of algorithms, which can lead to strategies for perfecting client value and service individualization The paper is the broader trend of clustering algorithms will be used to align practical business results.

III. Models and Methods

The overall methodology of the proposed work is shown in Figure 1

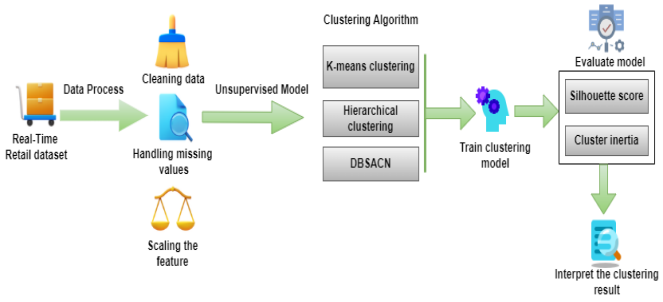

Fig. 1. Method

A. Data Collection

The churn.csv dataset from Kaggle is a publicly available dataset that contains information about a bank's customers and their propensity to churn, or leave the bank. The dataset includes a variety of customer attributes, such as demographic information, account information, and service usage patterns. This data is used to train machine learning models to predict which customers are most likely to churn, allowing the bank to take proactive measures to retain these valuable customers.This process provides banks with valuable insights into their customer base and enables them to implement effective customer retention strategies.The data set is a churn dataset figure**??**.

CustomerNum	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0

Fig. 2. Dataset read

B. Data Pre-Processed

Data preprocessing is a crucial step in data analysis and machine learning. It involves cleaning, transforming, and structuring raw data into a usable format for downstream tasks such as modeling, visualization, and statistical analysis. Preprocessing ensures data quality, consistency, and compatibility with the chosen analysis or modeling techniques

1) Dataset Cleaning: Data Even though there may be no missing values, data cleaning encompasses various steps to ensure data quality. This includes data type conversion, handling outliers, removal of duplicates, and normalization of numerical variables. These steps ensure that the data is consistent, standardized, and suitable for analysis and modeling.Data type conversion involves converting categorical variables into numerical representations, such as using one-hot encoding or label encoding. This allows machine learning algorithms to process and interpret categorical data effectively.Duplicate removal involves eliminating identical rows from the dataset.Duplicate removal involves eliminating identical rows from the dataset. Duplicate rows can inflate data volume, skew results, and hinder data analysis, so removing them ensures that each data point represents a unique observation The clean dataset figure8.

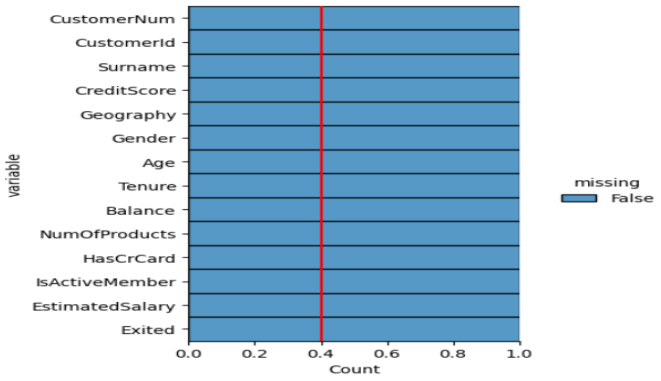

Fig. 3. Dataset read

2) Exploratory data analysis: Exploring the target variable ”Exited” reveals a clear imbalance between the two classes. The vast majority of customers (79.63%) have not yet exited the bank, while a relatively small proportion (20.37%) have churned. This imbalance is likely to pose a challenge for machine learning models, as they may struggle to identify the relatively small number of churned customers amidst the large majority of non-churned customers. Additionally, it is worth considering whether this imbalance is representative of the actual customer population or if it is due to some sampling bias. If the imbalance is indeed representative, then it may be necessary to employ techniques such as oversampling or undersampling to improve the model's performance.Figure8. This analysis suggests that a significant portion of customers churn within the first few months of opening an account. This could be due to a number of factors, such as dissatisfaction with the bank's services, a better offer from another bank, or a change in the customer's financial situation.Overall, the tenure graph provides valuable insights into the customer churn behavior of this bank. By understanding the distribution of customers by tenure, banks can develop strategies to reduce churn and improve customer satisfaction.Figure8.

C. Model Select

For this real-time dataset we choose unsupervised machine learning model because in this algorithm classified data via characteristics wise. If new update data will included in this

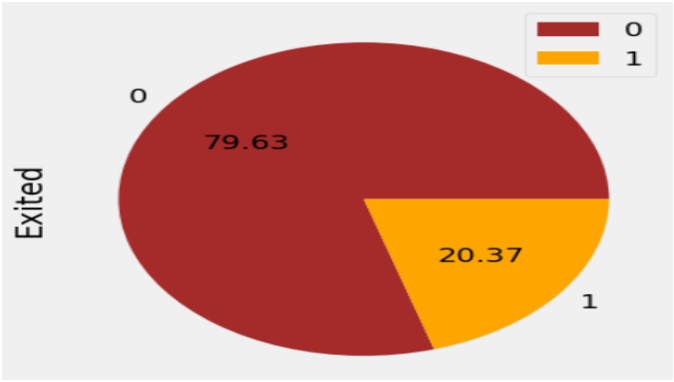

Fig. 4. Dataset read

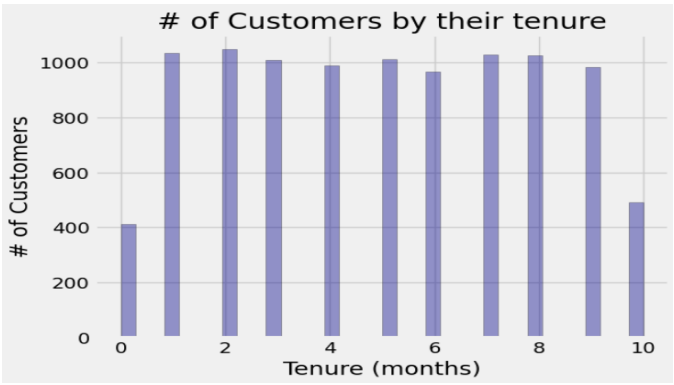

Fig. 5. Dataset read

unsupervised learning it will come same output as model train dataset.

D. Clustering

1.K-means clustering:

K-means clustering model using the Within Cluster Sum of Squares (WCSS) metric. WCSS measures how well data points fit their clusters, with lower values indicating a better fit. As the number of clusters (K) increases, WCSS decreases. At K=1, WCSS is high due to all points in one cluster. As K increases, WCSS decreases, but at K=3, the rate slows (elbow point), suggesting the optimal cluster number. Beyond K=3, WCSS decreases slowly, indicating little improvement in data fit, suggesting K=3 as the optimal number of clusters ,Figure

2.SVM:

SVM (Support Vector Machine) is a machine learning algorithm used for classification and regression. It finds a hyperplane in a high-dimensional space to separate different classes.Customer segmentation involves dividing a customer base into groups based on similarities such as demographics or behavior. It helps businesses tailor marketing strategies, personalize communication, allocate resources efficiently, improve retention, and gain insights into market trends.

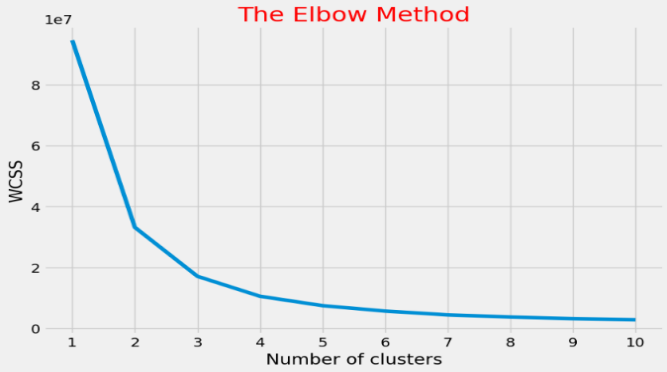

Fig. 6. KMeans

3.Random Forest

Random Forest is a powerful ensemble machine learning algorithm that combines multiple decision trees for enhanced accuracy. Its ability to handle complex datasets makes it valuable for customer segmentation. By analyzing diverse customer attributes, businesses can identify distinct groups and tailor strategies for personalized marketing and services. This ensures a more effective approach to meet the diverse needs of different customer segments, ultimately optimizing business performance and customer satisfaction. 4.Naive Bayes

Naive Bayes is a classification algorithm based on Bayes' theorem, assuming feature independence. It's valuable for customer segmentation by categorizing data like text, behavior, and demographics. This aids companies in tailoring strategies to specific customer groups efficiently, such as sentiment analysis in reviews or classifying customer behavior patterns. The algorithm's simplicity and effectiveness make it a useful tool in understanding and targeting distinct customer segments.

IV. Result Analysis

The goal of kmeans clustering is to find the clusters that minimize the within-cluster variance, which is the variance of the data points within each cluster.The output of kmeans clustering is a set of cluster labels, one for each data point. The cluster label for a data point indicates the cluster to which the data point belongs. In the example I provided, there are three clusters. The data points in cluster 0 have a credit score of 3 and a tenure of 2. The data points in cluster 1 have a credit score of 40 and a tenure of 10.This information can be used to make decisions about how to target different marketing campaigns. For example, you could target a campaign for financial products to people in cluster 0, as they have a lower credit score and may be more likely to be interested in such products.The target a campaign for luxury goods to people in cluster 1, as they have a higher credit score and may be more likely to be able to afford such goods.The specific way in which you use the output of kmeans clustering will depend on the specific problem you are trying to solve. However, the general idea is that you can use the cluster labels to group data points that are similar and make decisions about how to target different marketing campaigns to these groups. The cluster figure8.

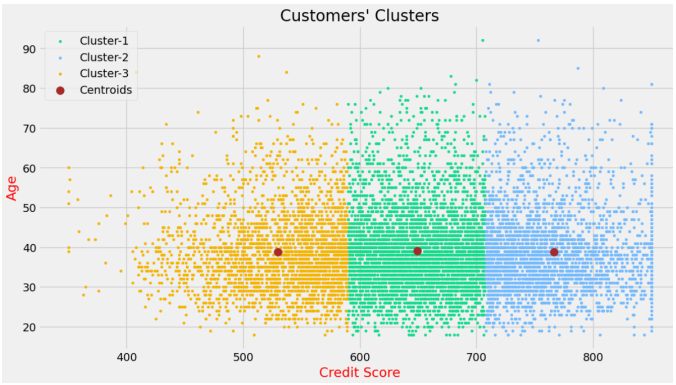

Fig. 7. Dataset read

For Churn prediction figure figure8.

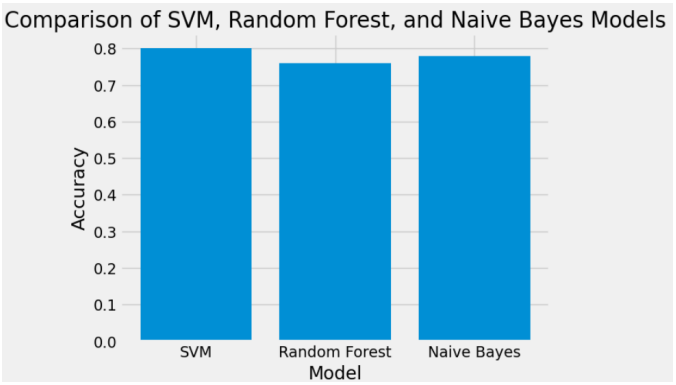

Fig. 8. Dataset read

V. Conclusion

References

[1] X. Zhang, G. Feng, and H. Hui, "Customer-churn research based on customer segmentation," in *2009 International Conference on Electronic Commerce and Business Intelligence*, 2009, pp. 443–446.

[2] T. Kansal, S. Bahuguna, V. Singh, and T. Choudhury, "Customer segmentation using k-means clustering," in *2018 international conference on computational techniques, electronics and mechanical systems (CTEMS)*. IEEE, 2018, pp. 135–139.

[3] E. Yadegaridehkordi, M. Nilashi, M. H. N. B. M. Nasir, S. Momtazi, S. Samad, E. Supriyanto, and F. Ghabban, "Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques," *Technology in Society*, vol. 65, p. 101528, 2021.

[4] P. Monil, P. Darshan, R. Jecky, C. Vimarsh, and B. Bhatt, "Customer segmentation using machine learning," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 8, no. 6, pp. 2104–2108, 2020.

[5] C. Dullaghan and E. Rozaki, "Integration of machine learning techniques to evaluate dynamic customer segmentation analysis for mobile customers," *arXiv preprint arXiv:1702.02215*, 2017.

[6] I. Smeureanu, G. Ruxanda, and L. M. Badea, "Customer segmentation in private banking sector using machine learning techniques," *Journal of Business Economics and Management*, vol. 14, no. 5, pp. 923–939, 2013.

[7] P. D. Hung, N. T. T. Lien, and N. D. Ngoc, "Customer segmentation using hierarchical agglomerative clustering," in *Proceedings of the 2nd International Conference on Information Science and Systems*, 2019, pp. 33–37.

[8] B. Shen, "E-commerce customer segmentation via unsupervised machine learning," in *The 2nd international conference on computing and data science*, 2021, pp. 1–7.

[9] L. Abidar, D. Zaidouni, and A. Ennouaary, "Customer segmentation with machine learning: New strategy for targeted actions," in *Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications*, 2020, pp. 1–6.

[10] A. Alghamdi, "A hybrid method for customer segmentation in saudi arabia restaurants using clustering, neural networks and optimization learning techniques," *Arabian Journal for Science and Engineering*, vol. 48, no. 2, pp. 2021–2039, 2023.

[11] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," *Sustainability*, vol. 14, no. 12, p. 7243, 2022.

[12] N. Hicham and S. Karim, "Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, 2022.

[13] A. Abdulhafedh, "Incorporating k-means, hierarchical clustering and pca in customer segmentation," *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021.

[14] M. A. Rahim, M. Mushafiq, S. Khan, and Z. A. Arain, "Rfm-based repurchase behavior for customer classification and segmentation," *Journal of Retailing and Consumer Services*, vol. 61, p. 102566, 2021.

[15] X. Xiahou and Y. Harada, "B2c e-commerce customer churn prediction based on k-means and svm," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, 2022.

[16] M. Lieder, F. M. Asif, and A. Rashid, "A choice behavior experiment with circular business models using machine learning and simulation modeling," *Journal of Cleaner Production*, vol. 258, p. 120894, 2020.

[17] K. R. Kashwan and C. Velu, "Customer segmentation using clustering and data mining techniques," *International Journal of Computer Theory and Engineering*, vol. 5, no. 6, p. 856, 2013.

[18] Y. Yuan, K. Dehghanpour, F. Bu, and Z. Wang, "A data-driven customer segmentation strategy based on contribution to system peak demand," *IEEE Transactions on Power Systems*, vol. 35, no. 5, pp. 4026–4035, 2020.

[19] E. Nandapala and K. Jayasena, "The practical approach in customers segmentation by using the k-means algorithm," in *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2020, pp. 344–349.