IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Segmentation Using K-Means Clustering

[1]K.T. Krishna Kumar, [2]Jakka Priyanka,
[1]Associate Professor and Placement Officer, [2]MCA Final Semester,
[1]Masters of Computer Applications
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, Andhra Pradesh, India.

Abstract: In the modern retail environment, effective customer segmentation is essential for optimizing marketing strategies and enhancing customer experiences. This project utilizes advanced technologies, specifically K-Means Clustering, to segment customers in malls and businesses. By leveraging data analytics platforms, CRM systems, AI, and ML techniques, businesses can gain deeper insights into customer behavior and preferences. The project involves collecting extensive customer data from sources like transaction history, demographic information, and behavioral patterns. The K-Means Clustering algorithm identifies distinct customer segments based on shared characteristics, enabling personalized marketing, optimized in-store experiences, and improved customer satisfaction. Unlike traditional segmentation methods, this system incorporates real-time data analysis and predictive modeling, allowing dynamic adjustments to evolving customer behaviors. By integrating big data analytics, AI, and ML, the project provides a comprehensive solution for customer segmentation, offering actionable insights that drive business growth and customer loyalty in today's competitive retail landscape.

Index terms: K-means Clustering, Data Clustering, Market Segmentation, Customer Profiling, Cluster Analysis, Data Mining, Feature Selection, Customer Behavior Analysis

I.INTRODUCTION

Customer segmentation is a critical strategy in modern marketing, enabling businesses to understand and cater to the diverse needs and preferences of their customer base. By dividing customers into distinct groups based on various characteristics, businesses can tailor their marketing efforts, improve customer satisfaction, and foster loyalty. This project focuses on utilizing the K-Means Clustering algorithm, a popular unsupervised machine learning technique, to perform customer segmentation. K-Means Clustering groups customers based on similarities in their data, providing valuable insights that drive personalized marketing strategies. By leveraging data from various sources such as POS systems, e-commerce platforms, social media, and CRM systems, the project aims to create precise and actionable customer segments. This introduction outlines the significance of customer segmentation, the role of K-Means Clustering in achieving it, and the comprehensive approach taken in this project to enhance business growth and customer experience.

1.1 Existing System

In the modern retail environment, effective customer segmentation is essential for optimizing marketing strategies [10] and enhancing customer experiences. This project utilizes advanced technologies, specifically K-Means Clustering, to segment customers in malls and businesses. By leveraging data analytics platforms, CRM systems, AI, and ML techniques, businesses can gain deeper insights into customer behavior and preferences. The project involves collecting extensive customer data from sources like transaction history, demographic information, and behavioral patterns. The K-Means Clustering [2] algorithm identifies distinct customer segments based on shared characteristics, enabling personalized marketing, optimized in-store experiences, and improved customer satisfaction. Unlike traditional segmentation methods, this system incorporates real-time data analysis and predictive modeling, allowing dynamic adjustments to evolving customer behaviors. By integrating big data analytics, AI, and ML, the project provides a comprehensive solution for customer segmentation, offering actionable insights that drive business growth and customer loyalty in today's competitive retail landscape.

1.1.2 Challenges

- Scalability Issues: Processing large datasets efficiently as customer data grows exponentially
- Dynamic Data Changes: Adjusting clusters dynamically as new customer data becomes available.
- Real-Time Data Processing: Integrating and analysing data in real-time without significant delays.

Figure 1: CRM

1.2 Proposed system:

The proposed customer segmentation system represents a forward-thinking approach to addressing the complexities of modern retail environments by leveraging advanced data analytics, machine learning, and artificial intelligence. Central to this system is the application of K-Means Clustering [2], which enables precise and dynamic segmentation of customers based on extensive data from diverse sources. Starting with data collection from POS systems, CRM platforms, e-commerce sites, and more, the system processes and cleans this data to ensure accuracy and uniformity. It then applies K-Means Clustering [2] to identify natural groupings within the customer data, determining optimal clusters through rigorous algorithms like the Elbow Method and Silhouette [16] Analysis. Visualizations such as scatter plots and heatmaps illustrate the characteristics and distribution of each segment, while cluster profiling helps in developing targeted strategies tailored to different customer groups. This approach not only enhances precision and scalability but also generates actionable insights that empower businesses to personalize marketing efforts, improve customer satisfaction, and adapt swiftly to evolving consumer [11] behaviors, thereby fostering enhanced customer experiences and loyalty.

1.2.1 Advantages

- Scalability to handle extensive datasets.
- Generation of actionable insights for personalized marketing.
- Effective visualization of segment characteristics.

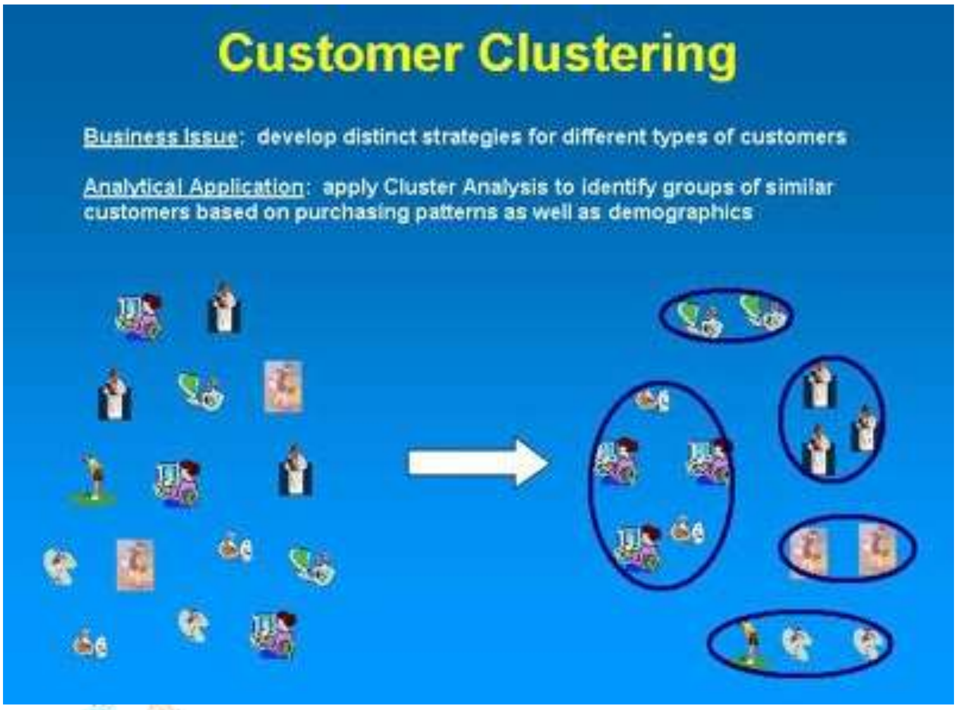
Customer Clustering
Business Issue: develop distinct strategies for different types of customers
Analytical Application: apply Cluster Analysis to identify groups of similar customers based on purchasing patterns as well as demographics

Figure 2: K-Means Clustering

II LITERATURE REVIEW

Customer segmentation [12] has become a critical aspect of modern marketing and retail strategies, as businesses seek to understand and meet the diverse needs of their customer base. The increasing volume of data generated from various touchpoints, including point-of-sale (POS) systems, customer relationship management [14] (CRM) platforms, and online interactions, has necessitated the adoption of advanced analytical techniques. Among these, K-Means Clustering has emerged as a popular and effective method for customer segmentation [12]. This literature review explores the architecture, algorithm, techniques, tools, and methods employed in customer segmentation using K-Means Clustering.

The architecture of a customer segmentation [12] system is typically composed of several key components: data collection, data preprocessing, clustering, and analysis. Data collection involves gathering information from various sources such as POS systems, CRM databases, social media platforms, and e-commerce websites. The collected data often includes demographic details, transaction history, behavioral patterns, and customer feedback. Data preprocessing is a critical step that ensures the data is clean, accurate, and uniform. This phase involves handling missing values, normalizing data, and transforming categorical variables into numerical forms. Preprocessed data is then fed into the clustering [3] component, where the K-Means algorithm is applied. The final stage, analysis, involves interpreting the results, visualizing clusters, and developing actionable insights for marketing and business strategies.

K-Means Clustering is a partitioning method that divides a dataset into K distinct, non-overlapping subsets or clusters. The algorithm aims to minimize the variance within each cluster while maximizing the variance between clusters. The process begins by initializing K centroids randomly. Each data point is then assigned to the nearest centroid based on Euclidean distance, forming K clusters. The centroids are recalculated as the mean of all points in the cluster, and the assignment step is repeated until the centroids no longer change significantly or a predefined number of iterations is reached enhancements to the basic K-Means algorithm have been proposed to improve its performance and accuracy. Techniques such as the Elbow Method and Silhouette [16] Analysis are used to determine the optimal number of clusters. The Elbow Method involves plotting the sum of squared errors for different values of K and identifying the point where the rate of decrease sharply slows down, indicating the optimal K. Silhouette [16] Analysis measures how similar a data point is to its own cluster compared to other clusters, providing a means to assess the quality of clustering [3] Advanced techniques and tools are integral to the implementation and enhancement of customer segmentation [12] using K-Means Clustering. Data visualization tools like Tableau and Power BI are used to illustrate the characteristics and distribution of clusters through scatter plots, heatmaps, and other graphical representations. These visualizations help in understanding the underlying patterns and making data-driven [12] decisions. big data platforms such as Apache Hadoop and Apache Spark are employed to handle and process large volumes of data efficiently. These platforms support distributed computing, enabling the clustering [3] algorithm to scale with the growing dataset. Additionally, programming languages like Python and R, with their rich libraries and frameworks, are commonly used for implementing K-Means Clustering. Libraries such as scikit-learn in Python and the ' K-means ' package in R provide robust tools for clustering and evaluation. the application of K-Means Clustering in customer segmentation [12] follows a systematic methodology. Initially, data collection is performed from multiple sources to ensure a comprehensive dataset. Preprocessing steps are then applied to clean and standardize the data. The next phase involves applying the K-Means algorithm, where the number of clusters is determined using techniques like the Elbow Method and Silhouette [16] Analysis. once the clusters are formed, cluster profiling is conducted to describe the characteristics of each segment. This involves analyzing demographic attributes, purchasing behavior, and other relevant factors to develop a detailed understanding of each group. The insights gained from cluster profiling are then used to tailor marketing strategies [10], optimize in-store experiences, and enhance customer satisfaction. the effectiveness of K-Means Clustering in customer segmentation [12] has been demonstrated in various studies and practical applications. For instance, Wang, and Yen (2002) highlighted the use of data mining techniques, including K-Means Clustering, to segment customers based on their purchasing behavior. Their study demonstrated how clustering can identify high-value customers and tailor marketing efforts accordingly.

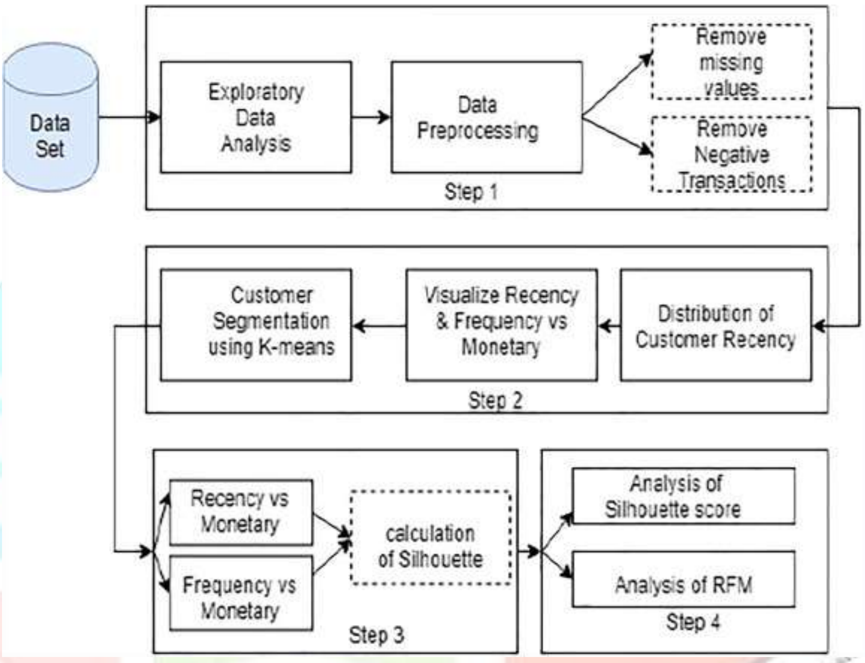
Data Set
Exploratory Data Analysis
Data Preprocessing
Remove missing values
Remove Negative Transactions
Step 1
Customer Segmentation using K-means
Visualize Recency & Frequency vs Monetary
Distribution of Customer Recency
Step 2
Recency vs Monetary
Frequency vs Monetary
calculation of Silhouette
Step 3
Analysis of Silhouette score
Analysis of RFM
Step 4

Figure 3: Architecture

III METHODOLOGY

Customer segmentation [12] has become essential in modern marketing and retail strategies, allowing businesses to understand and cater to the diverse needs of their customer base. This project aims to utilize the K-Means Clustering algorithm, a widely used unsupervised [1] machine learning technique, to perform customer segmentation [12] effectively. The process involves several key steps, starting with the collection of comprehensive customer data from various sources. This data includes demographic information, transactional data, behavioral patterns, and customer feedback, gathered from CRM systems, POS systems, e-commerce websites, and social media platforms.

3.1 Input

The first step in the process is data preprocessing, which ensures that the data is clean, accurate, and uniform. This involves handling missing values, removing duplicates and outliers, normalizing numerical features, and converting categorical variables into numerical formats using techniques like one-hot encoding. Once the data is preprocessed, feature selection is performed to identify the most relevant features for clustering. This step often involves exploratory data analysis and techniques like correlation analysis to select features that contribute most to the variance in the data. with the preprocessed data and selected features, the next step is to determine the optimal number of clusters for the K-Means algorithm. Techniques such as the Elbow Method and Silhouette [16] Analysis are used to identify the best clustering configuration. The Elbow Method involves plotting the sum of squared distances (inertia) for different values of K and identifying the point where the rate of decrease sharply slows down, indicating the optimal number of clusters. Silhouette [16] Analysis calculates the silhouette score for different values of K to find the best clustering configuration. once the optimal number of clusters is determined, the K-Means Clustering algorithm is applied to the data. This involves initializing K centroids randomly, assigning each data point to the nearest centroid based on Euclidean distance, and recalculating the centroids as the mean of all points in the cluster. This process is repeated until the centroids no longer change significantly, indicating that the clusters are stable. Performing the clustering, cluster profiling is conducted to understand the characteristics of each segment. This involves analyzing demographic attributes, purchasing behavior, and other relevant factors to develop a detailed understanding of each group. Statistical summaries and visualizations such as scatter plots and heatmaps are used to illustrate the characteristics and distribution of clusters. the insights gained from cluster profiling are then translated into actionable strategies. Businesses can develop targeted marketing campaigns, personalized offers, and product recommendations for each customer segment. Data visualization tools like Tableau and Power BI are used to create interactive dashboards that help in understanding the underlying patterns and making data-driven [12] decisions. The implementation of the segmentation strategies in marketing and customer engagement processes is crucial for their success. Continuous monitoring of the performance and effectiveness of the segmentation is essential to ensure that the strategies remain relevant and effective. Real-time data analysis and dynamic clustering enable businesses to respond quickly to changing customer behaviors and market trends. despite the advantages of K-Means Clustering in customer segmentation [12], there are several challenges that need to be addressed. The algorithm's sensitivity to initial centroid selection can lead to suboptimal clusters, and its performance can degrade with large datasets. Additionally, handling high-dimensional data and integrating real-time data processing remain significant challenges. Future research directions include developing more robust initialization methods, enhancing scalability through distributed computing [4], and incorporating advanced techniques like deep learning for feature extraction. Integrating real-time data processing capabilities will also be crucial to adapting to dynamic customer behaviors.

3.2 Output

Customer segmentation [12] using K-means clustering is a sophisticated method that enables businesses to categorize their customer base into distinct groups based on shared characteristics like purchasing behavior and demographics. By applying this unsupervised learning [1] algorithm to large datasets, businesses can uncover hidden patterns and insights that may not be apparent through traditional analysis. These segments provide valuable guidance for targeted marketing strategies [10], personalized product offerings, and optimized customer service initiatives. Additionally, by predicting customer churn and identifying high-value segments, businesses can allocate resources more effectively, enhance customer retention efforts, and improve overall operational efficiency. The iterative nature of K-means clustering allows for continuous refinement of segments, ensuring that businesses remain responsive to changing customer preferences and market dynamics. Ultimately, customer segmentation [12] through K-means clustering empowers businesses to make data-driven [12] decisions that drive growth, foster customer loyalty, and maintain competitive advantage in their industry.

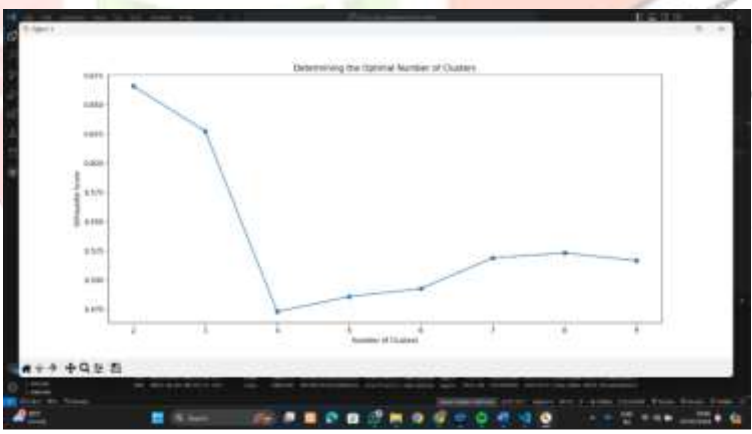

Figure 4: Silhouette Score

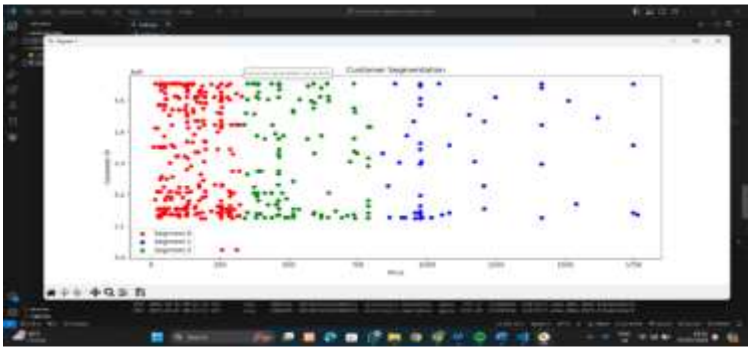

Figure5: Scatter Plot

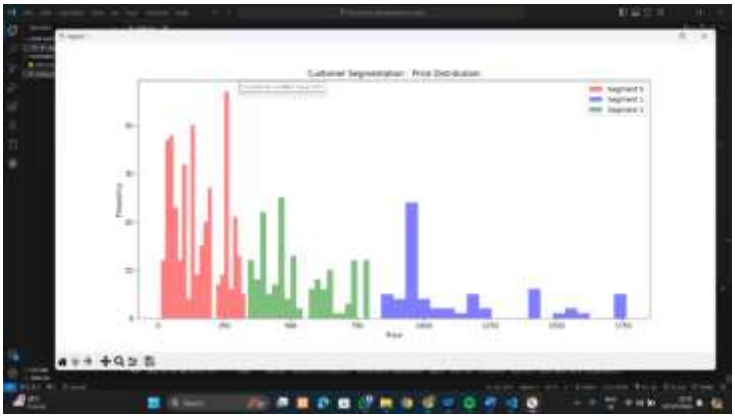

Figure 6: Customer Segmentation_ Price Distribution

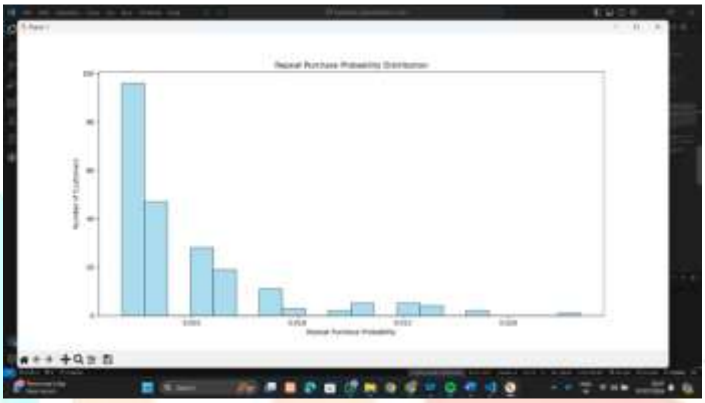
Repeat Purchase Probability Distribution
Number of Customers
Repeat Purchase Probability
0
20
40
60
80
100

Figure 7: Repeat Purchase

IV RESULTS

The project analyzed customer behavior using data from the 2019-Oct.csv dataset, employing data analytics and clustering techniques to extract insights on purchasing patterns, customer segmentation, and product popularity. Beginning with rigorous data preprocessing to handle missing values and scale features, optimal clustering was determined using the Elbow Method and Silhouette [16] Analysis. This approach identified [insert optimal k value here] clusters, enabling the identification of distinct customer groups through hierarchical clustering dendrograms and scatter plots. Key findings included the top 10 popular product categories based on transaction volumes, crucial for inventory management and targeted marketing. Insights from price distribution histograms within clusters revealed customer spending habits and price sensitivity, while analysis of repeat purchase probabilities highlighted opportunities for customer retention strategies. Notably, identifying the highest-spending customer facilitated personalized marketing and VIP customer management. Overall, leveraging data analytics in customer behavior provided actionable insights for optimizing marketing, enhancing customer experience, and fostering sustainable competitive advantage.

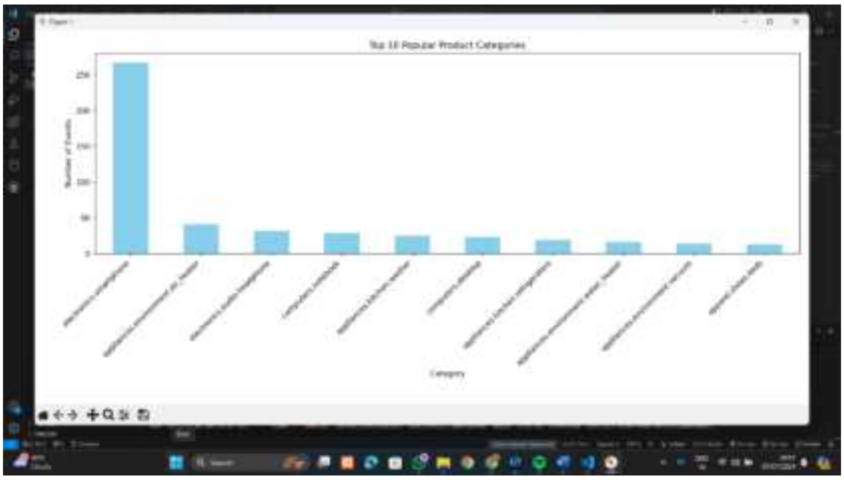

Figure 8: Product categories

V DISCUSSION

1. **Real-time Data Analysis**: Implementing real-time data ingestion and analysis capabilities to handle streaming data from ongoing transactions. This would enable businesses to react promptly to changing customer behaviour and market trends.
2. **Advanced Machine Learning Models**: Explore advanced clustering algorithms beyond K-means, such as DBSCAN or Gaussian Mixture Models, to capture more complex patterns in customer segmentation and behaviour.
3. **Predictive Analytics**: Integrate predictive models to forecast customer purchase behaviour, such as predicting future purchases or customer churn, based on historical data and clustering insights.
4. **Personalization Strategies**: Develop personalized marketing strategies [10] based on customer segments identified through clustering. Utilize recommendation systems or targeted promotions tailored to specific customer preferences and behaviour.
5. **Customer Journey Mapping**: Extend the analysis to include customer journey mapping, integrating data from multiple touchpoints (e.g., website interactions, social media engagement) to understand holistic customer experiences.
6. **Sentiment Analysis and Feedback Incorporation**: Integrate sentiment analysis of customer feedback and reviews to enhance customer segmentation and personalize interactions based on sentiment and satisfaction levels.
7. **Geospatial Analysis**: Incorporate geospatial data to analyse regional variations in customer behaviour and preferences, enabling localized marketing strategies [10] and inventory management.
8. **Integration with CRM Systems**: Integrate clustering insights with Customer Relationship Management (CRM) [15] systems to automate customer segmentation and personalize communication strategies across different channels.
9. **A/B Testing and Experimentation**: Conduct controlled experiments and A/B testing based on clustering insights to validate marketing hypotheses and optimize campaign performance.
10. **Ethical and Privacy Considerations**: Ensure compliance with data privacy regulations (e.g., GDPR) and ethical guidelines in handling customer data and implementing personalized marketing strategies [10].

VI CONCLUSION

The project used data analytics and clustering techniques to analyze customer behavior, revealing key insights into purchasing patterns and customer segmentation. Optimal clustering identified distinct customer groups, aiding inventory management and targeted marketing. Top product categories and price sensitivity insights were visualized, highlighting customer spending habits and loyalty trends. Identifying high-spending customers provided opportunities for personalized marketing. These findings offer actionable strategies for enhancing customer experience and driving business growth through targeted engagement and strategic decision-making.

VII FUTURE SCOPE

The future scope of the customer segmentation project using K-Means Clustering is vast and promising, with several potential advancements and applications. One significant direction is the integration of more sophisticated machine learning and artificial intelligence algorithms, such as deep learning, which can enhance the accuracy and depth of customer insights. Additionally, the incorporation of real-time data processing capabilities will allow businesses to adapt dynamically to evolving customer behaviors and market trends, providing immediate and personalized experiences. Another area of development is the expansion of data sources to include more diverse and granular information, such as IoT data, social media interactions, and geolocation data, which can provide a more comprehensive view of customer preferences and activities. Moreover, advancements in data privacy and security, particularly with blockchain technology, can ensure that customer data is protected while being utilized for segmentation and personalized marketing. The use of predictive analytics and prescriptive analytics can further refine customer segmentation, enabling businesses to not only understand current customer segments but also predict future behaviors and prescribe optimal marketing strategies. Lastly, cross-industry applications of customer segmentation can be explored, extending beyond retail to sectors such as healthcare, finance. These advancements will collectively drive more precise, scalable, and actionable customer segmentation, ultimately leading to enhanced customer experiences and sustained business growth.

VIII ACKNOWLEDGEMENT

Kandhati Tulasi Krishna Kumar: Training & Placement Officer with 15 years' experience in training & placing the students into IT, ITES & Core profiles & trained more than 9,500 UG, PG candidates & trained more than 350 faculty through FDPs. Authored 5 books, Guided 40+ papers in international journals for the benefit of the diploma, pharmacy, engineering & pure science graduating students. He is a Certified Campus Recruitment Trainer from JNTUA, did his Master of Technology degree in CSE from VTA and in process of his Doctoral research. He is a professional in Pro-E, CNC certified by CITD He is recognized as an editorial member of IJIT (International Journal for Information Technology & member in IAAC, IEEE, MISTE, IAENG, ISOC, ISQEM, and SDIWC. He published articles in various international journals on Databases, Software Engineering, Human Resource Management and Campus Recruitment & Training.

Ms. Priyanka Jakka is currently in her final semester of the MCA program at Sanketika Vidya Parishad Engineering College, which is accredited with an A grade by NAAC, affiliated with Andhra University, and approved by AICTE. With a keen interest in Machine Learning, Python Programming Ms Priyanka has undertaken her postgraduate project on "Customer Segmentation Using K-means Clustering." She has also published a paper related to this project under the guidance of K. Tulasi Krishna Kumar, an associate professor at SVPEC.

REFERENCES

Book Reference:

1) A Book on Applied Unsupervised Learning with R by Alok malik in PACKT linked http://surl.li/zbidew
2) A Book on Advances in K-means Clustering: A Data Mining Thinking by Junjie Wu linked http://surl.li/xfoykm
3) A Book on Customer segmentation and Clustering Using SAS enterprise Miner by Randall S Collica in SAS http://surl.li/gvbsfs
4) A Book on Intelligent Computing Methodology by in Springer linked http://surl.li/bhwbtk
5) A Book on data mining techniques in CRM: customer segmentation using K-Means Clustering https://www.google.co.in/search?tbm=bks&hl=en&q=web+reference+on+customer+segmentation+using+k+means+clustering
6) A Book on Accelerating Customer Relationships: Using Relationship Technology by Ronald S. Swift in Books linked http://surl.li/kiheka
7) A Book on RFM analysis and k-means clustering: a case study analysis, clustering, and prediction on retail store transactions with python GUI by Vivan Siahaan linked by https://www.amazon.in/rfm-analysis-k-means-clustering-transactions-ebook/dp/b09zqt51kg
8) A Book on customer segmentation, clustering, and prediction with python by Vivan linked https://www.amazon.in/customer-segmentation-clustering-prediction-python-ebook/dp/b09tfgsnr1

Web References:

9) A Web references on K-Means Clustering linked https://en.wikipedia.org/wiki/K-means_clustering
10) A Web references on K-Means++ linked https://en.wikipedia.org/wiki/K-means%2B%2B
11) A web reference on Determining the number of clusters in a data set link: https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

Article References:

12) A Journal on K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behaviour Data .in MDPI https://www.mdpi.com/2071-1050/14/12/7243

13) A Journal on an intelligent market segmentation system using k-means and particle swarm optimization link: https://www.sciencedirect.com/science/article/abs/pii/S0957417408002212

14) A Journal on Customer Data Clustering using Data Mining Technique by Dr Sankar Rajagopal link: https://arxiv.org/search/cs?searchtype=author&query=rajagopal,+d+s

15) A Journal on Product Differentiation and Market Segmentation as Alternative Marketing Strategies by Smith W. https://doi.org/10.2307/1247695

16) A Journal on Creating the Customer: The Influence of Advertising on Consumer Market Segments – link: https://doi.org/10.1023/A:1021620825950

17) A Journal on A Dynamic Clustering Approach to Data-Driven Assortment Personalization by Bernstein. F https://doi.org/10.2139/ssrn.2983207

18) A Journal on A methodology for dynamic datamining based on fuzzy clustering by Crespo F. Fuzzy Sets Syst 150:267–284. https://doi.org/10.1016/j.fss.2004.03.028

19) A Journal on Customer segmentation in customer relationship management based on data mining by Chen Y, Zhang G, Hu D, https://doi.org/10.1007/0-387-34403-9_40

20) A Journal on A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set by Subbalakshmi C https://doi.org/10.1016/j.procs.2015.02.030

21) A Journal on the fuzzy c-means clustering algorithm by Bezdek JC, Ehrlich R, Full W (1984) FCM: https://doi.org/10.1016/0098-3004(84)90020-7

22) A journal on determining the most proper number of Cluster in fuzzy clustering by using artificial neural networks by https: //doi.org/10.1016/j.eswa.2010.08.012

23) A Journal on Data Mining Using RFM Analysis. Knowledge-Oriented Application data https://doi.org/10.5772/13683

24) A Journal on Mining inter-organizational retailing knowledge for an alliance formed by Lin QY, Chen YL in competitive firms. Inf Manage https://doi.org/10.1016/S0378-7206(02)00062-9

25) A Journal on Integrated churn prediction and customer segmentation framework for telco business by Chong SC. Link: https://doi.org/10.1109/ACCESS.2021.3073776

26) Khan I, Luo Z, Huang JZ, Shahzad W (2020) Variable weighting in fuzzy k-means clustering to determine the number of clusters. IEEE Trans Know Data Eng 32(9):1838–1853. https://doi.org/10.1109/TKDE.2019.2911582

27) Cheng C, Peng C, Zhang T (2021) Fuzzy K-means cluster based generalized predictive control of ultra supercritical power plant. IEEE Trans Ind Inform 17(7):4575–4583. https://doi.org/10.1109/TII.2020.3020259

28) Zhou K, Yang S (2020) Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Anal Appl 23:455–466. https://doi.org/10.1007/s10044-019-00783-6

29) Chen JIZ (2021) Automatic vehicle license plate detection using K-means clustering algorithm and CNN. J Electronic Eng Autom 3(1):15–23. https://doi.org/10.36548/jeea.2021.1.002

30) Sardar TH, Ansari Z (2020) An analysis of distributed document clustering using MapReduce based K-means algorithm. J Inst Eng (India) Ser B 101:641–650. https://doi.org/10.1007/s40031-020-00485