WEBPOISE INTERNSHIP PROJECT

PROJECT REPORT

Title: **Customer Segmentation Using K-Means**

Submitted by :

Aravind V(1RV22CD011) 7th sem
B.E.CSE(Data science) RV College of engineering
Contact : 9019253851
Email : aravindv.cd22@rvce.edu.in

Anjan Babu KN(1RV23CD400) 7th sem
B.E.CSE(Data science) RV College of engineering
Contact : 77760285088
Email : anjanbabukn.cd23@rvce.edu.in

Likith A (1RV22CD027) 7th sem
B.E.CSE(Data science) RV College of engineering
Contact : 9113275058
Email : likitha.cd22@rvce.edu.in

K Sai Chetan(1RV23CD404) 7th sem
B.E.CSE(Data science) RV College of engineering
Contact : 6362616532
Email : ksaichetan.cd23@rvce.edu.in

Vaibhav TL(1RV22IS075) 7th sem
B.E.ISE RV College of engineering
Contact : 6360922463
Email : vaibhavtl.is22@rvce.edu.in

Vikas AL(1RV23CY405) 7th sem
B.E.CSE(Cyber Security) RV College of engineering
Contact : 8147401303
Email : vikasal.cy23@rvce.edu.in

Abstract

The management of customer service can be considered as a key to achieving revenue growth and profitability in today's fast-moving world of marketing from product-oriented to customer-oriented. Customer behavior knowledge can assist marketing managers in re-evaluating existing customer tactics and planning to improve and increase the use of the most effective strategies. B2B or business customers are more complicated, have a more difficult buying procedure, and have a higher sales value. Business marketers, on the other hand, want to work with fewer but larger customers than final consumer marketers. Because a business transaction necessitates more decision-making and professional buying effort than a consumer purchase, maintaining a productive relationship with business customers is critical. Most customer segmentation methods based on customer value fail to account for the factor of time and the trend of value changes. Because there are so many potential customers who are unsure of what to buy and what not to buy, today's business is based on new concepts. The businesses themselves are unable to diagnose their target potential customers. This is where machine learning comes in, various algorithms are used to detect hidden patterns in the data in order to make better decisions.

Customer segmentation is the practice of dividing a customer base into many groups of people who are similar in various aspects significant to marketing, such as gender, age, interests, and other spending habits. Companies that use customer segmentation believe that each client has unique needs that require a tailored marketing strategy to satisfy. Companies want to obtain a better understanding of the customers they're after. As a result, their goal must be explicit, and it must be adjusted to meet the needs of each and every individual customer. Furthermore, by analyzing the data acquired, businesses can gain a better grasp of client preferences as well as the needs for identifying profitable segments. This allows them to more effectively strategize their marketing strategies while reducing the chance of their investment being jeopardized. Customer segmentation is a process that is depending on a number of factors. Data on demographics, geography, economic position, and behavioral tendencies are all important factors in establishing the company's approach to distinct sectors.

The customer segmentation uses the clustering technique to determine which consumer segment to target. The clustering algorithm we have employed in this project is the K-means algorithm, which is a partitioning algorithm for segmenting clients based on comparable criteria. The elbow approach is used to determine the best clusters.

1. Introduction

In today's data-driven business environment, organizations generate and collect massive volumes of customer-related data from multiple sources including e-commerce platforms, social media, transaction systems, customer relationship management (CRM) software, and feedback channels. The challenge faced by businesses is no longer the lack of data, but rather the difficulty in extracting meaningful insights from this large volume of information. One of the most important tasks in data analytics and machine learning is customer segmentation, which plays a vital role in understanding customer behavior, optimizing marketing strategies, and improving overall business performance.

Customer segmentation is the process of dividing a company's customer base into distinct groups based on shared characteristics such as purchasing behavior, income level, preferences, demographics, or engagement patterns. These groups, known as segments, enable businesses to tailor products, services, and promotional activities to specific categories of customers rather than adopting a generalized approach. A well-designed segmentation strategy helps organizations improve customer satisfaction, enhance retention rates, detect emerging trends, and allocate resources more efficiently.

With the rapid evolution of machine learning techniques, traditional rule-based segmentation has been replaced by intelligent, data-driven models. Among various clustering algorithms, the K-Means algorithm has emerged as one of the most popular unsupervised learning methods due to its simplicity, scalability, and computational efficiency. K-Means groups customers by minimizing the distance between data points within the same cluster and maximizing the distance between different clusters, ensuring meaningful and well-separated segments. Unlike supervised learning methods, K-Means does not require labeled data, making it ideal for real-world business datasets where predefined classes are usually unavailable.

In this project, Python is used as the primary programming language because of its powerful libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn, which facilitate data preprocessing, visualization, model implementation, and model evaluation. The project emphasizes building a complete customer segmentation pipeline that includes data loading, cleaning, feature selection, exploratory data analysis (EDA), model development, and interpretation of results. Through visualization techniques such as scatter plots, correlation heatmaps, and cluster mapping, patterns and customer behavior trends are identified clearly and intuitively.

An important challenge in customer segmentation is selecting an optimal number of clusters that best represent the natural grouping within the data. To overcome this problem, this project applies evaluation techniques such as the Elbow Method and Silhouette Score to determine the most appropriate number of clusters. Additionally, improved initialization techniques such as K-Means++ are employed to reduce convergence errors and improve clustering accuracy. These enhancements contribute to stable and reproducible results, enabling consistent identification of customer profiles.

2. Literature review

Because of the intense rivalry in the business sector, businesses have had to improve their profitability and business throughout time by satisfying client requests and attracting new customers based on their wants. Client identification and meeting individual custom

is a difficult task. This is due to the fact that clients differ in terms of their wants, desires, preferences, and so on. Customer segmentation, as opposed to a "one-size-fits-all" strategy, separates customers into groups with similar features or behavioral characteristics. Customer segmentation, according to, is a strategy for splitting the market into homogeneous groups. The data utilised in the customer segmentation technique, which divides customers into groups, is based on a variety of characteristics including demographics, regional circumstances, economic situations, and behavioral tendencies.

The customer segmentation strategy enables a company to make better use of its marketing spending and obtain a competitive advantage. Displaying a superior understanding of the customer's requirements, it also aids a company in improving marketing efficiency, budgeting for marketing, recognizing new market prospects, developing a stronger brand strategy, and measuring client retention.

[1] Customer Segmentation by Using RFM Model and Clustering Methods: A Case Study in Retail Industry

In this paper Prasanta Bandyopadhyay proposed two different clustering models to segment 700032 customers by considering their RFM values. They detected that the current customer segmentation which built by just considering customers' expense is not sufficient. Companies need to understand the customers' data better in all aspects. Detecting similarities and differences among customers, predicting their behaviours, proposing better options and opportunities to customers became very important for customer-company engagement. Segmenting the customers according to their data became vital in this context. RFM (recency, frequency and monetary) values have been used for many years to identify which customers valuable for the company, which customers need promotional activities, etc. Data-mining tools and techniques widely have been used by organizations and individuals to analysis their stored data. Clustering, which one of the tasks of data mining has been used to group people, objects, etc.

[2] Mall customer segmentation using clustering algorithms:

M.A. Ishantha USE/2017/OCT/0045 done research on data set that is about behaviour of the customers having credit card using many unsupervised algorithms. The dataset which contains 8950 transactions or information about account that belong to customers. And also, she has found how many clusters can distinguish the customers according to their transactions or behaviours". The methodology they have followed is K-Means clustering, Minibatch K-Means Clustering Algorithm, Hierarchical Clustering Segmentation and Elbow Method.

[3] Approaches to Clustering in Customer Segmentation: Techniques and approaches

Shreya Tripathi, Aditya Bhardwaj, Poovammal have gone through the various approaches and techniques to clustering in the segmentation process. They have explained what is the customer relationship management, necessities and importance of a customer segmentation in various industries. They have found which is giving the max optimal customers among the all-clustering algorithms like K-Means Clustering, Elbow method and Hierarchical Clustering-. Agglomerative, divisive with optimization. And also, they have done Visualization of the formation of clusters in the studied dataset with the help of a dendrogram

[4] Comparisons between data clustering algorithms

Osama Abu Abbas have done the mathematical implementation on clustering algorithms with a sample dataset. And also, he explored how algorithms are implemented. He chosen four different clustering algorithms to investigate, study, and compare them. The algorithms he has chosen are: K-means, Self-Organization (SOM), Hierarchal clustering algorithms and Expectation Maximization (EM) clustering algorithm. He listed the reason why he has chosen the particular algorithms to compare, study etc. He has done all the work to find the optimal clusters from each algorithm finally he explained how algorithms are compared. He said this paper intended to compare between some data clustering algorithms. Through his search he was unable to find any study attempts to compare between the four clustering algorithms under investigation.

[5] Concept Decompositions for Large Sparse Text Data Using Clustering:

Dhillon and D. Modha said that It is of tremendous practical relevance to apply machine learning and statistical algorithms such as clustering, classification, principal component analysis, and discriminant analysis to text data sets. In this paper they mainly focus on clustering of text data sets. The structure of the clusters created by the spherical k-means algorithm when applied to text data sets is the first focus. With the goal of acquiring unique insights into the distribution of sparse text data in high-dimensional environments. Such structural insights are a necessary first step toward their second goal, which is to investigate the tight linkages between clustering with the spherical k-means algorithm and the problem of matrix approximation for word-by-document matrices. They've also looked into massive document collection vector space models. These models are extremely high-dimensional and sparse, posing computational and statistical issues not seen in low-dimensional dense data. The spherical k-means algorithm looks for clusters with strong coherence. They discovered that average cluster coherence is poor, implying that each thought vector is surrounded by a big gap in high-dimensional space. Furthermore, they discovered that at various resolutions, the average intra- and inter-cluster architectures were identical. The only significant distinction is the gradual separation of intra-cluster and inter-cluster structure.

[6] Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms

In this paper A. Ansari and A. Riasi, to cluster the customers of the steel sector, they have combined fuzzy c-means clustering and genetic algorithms. The LRFM (length, recency, frequency, monetary value) model variables were used to separate the customers into two groups. Data from 120 consumers was collected and standardised in order to do the clustering. The information comprised four separate variables: the length of the relationship, the recentness of the trade, the frequency of the trade, and the monetary worth of the trade. GA-Fuzzy Clustering software was used to accomplish the fuzzy clustering. Clustering is done using a combination of fuzzy c-means clustering and a genetic algorithm in this software. Finally, the customers were divided into two clusters. To compare the efficiency of combined algorithms (Fuzzy c-means and genetic algorithms) means square error (MSE) and run time error were used. When compared to the average values of these parameters for all consumers, customers in the first cluster had a longer relationship, more recent trade, and more frequency of trade, but a lower monetary value.

[7] Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment:

Identifying patients in a Target Customer Segment (TCS) is critical for determining demand for health care services and allocating resources properly. The goal of this research is to develop a two-stage clustering-classification model by combining the RFM attribute and the K-means algorithm to cluster TCS patients, and then combining the global discretization method and rough set theory to classify hospitalized departments and optimize health care services. To evaluate the proposed model's performance, a dataset from a representative hospital (dubbed Hospital-A) was collected from a database from an empirical study in Taiwan that included 183,947 samples classified by 44 variables in 2008. The suggested model was compared to three techniques: Decision Tree, Naive Bayes, and Multilayer Perceptron, with the empirical results indicating that it has a high likelihood of being accurate. The knowledge-based rules that are developed give important information for maximizing resource use.

[8] A Two-Phase Clustering Method for Intelligent Customer Segmentation

M. Namvar, M. Gholamian and S. KhakAbi states that many studies have looked at the use of data mining technology in customer segmentation and found it to be successful. However, in the majority of cases, it is done utilising customer data from a unique perspective rather than a systematic process that considers all stages of CRM. Using data mining technologies, they have developed a new customer segmentation strategy based on RFM, demographic, and LTV data. There are two stages to the new consumer segmentation method. Customers are first clustered into distinct segments based on their RFM using K-means clustering. Second, each cluster is partitioned into new clusters using demographic data. Finally, a customer profile is constructed using LTV. They have applied this approach to a dataset from an Iranian bank, yielding some valuable management recommendations and measures.

The method they have followed was based on a two-phase clustering model using the k-means algorithm. Finally existing customers were split into nine groups based on their shared transactional behavior and features when the strategy was used to our case study (in the banking business). Marketers could evaluate the profiles of clients in each category to develop strategies for each group.

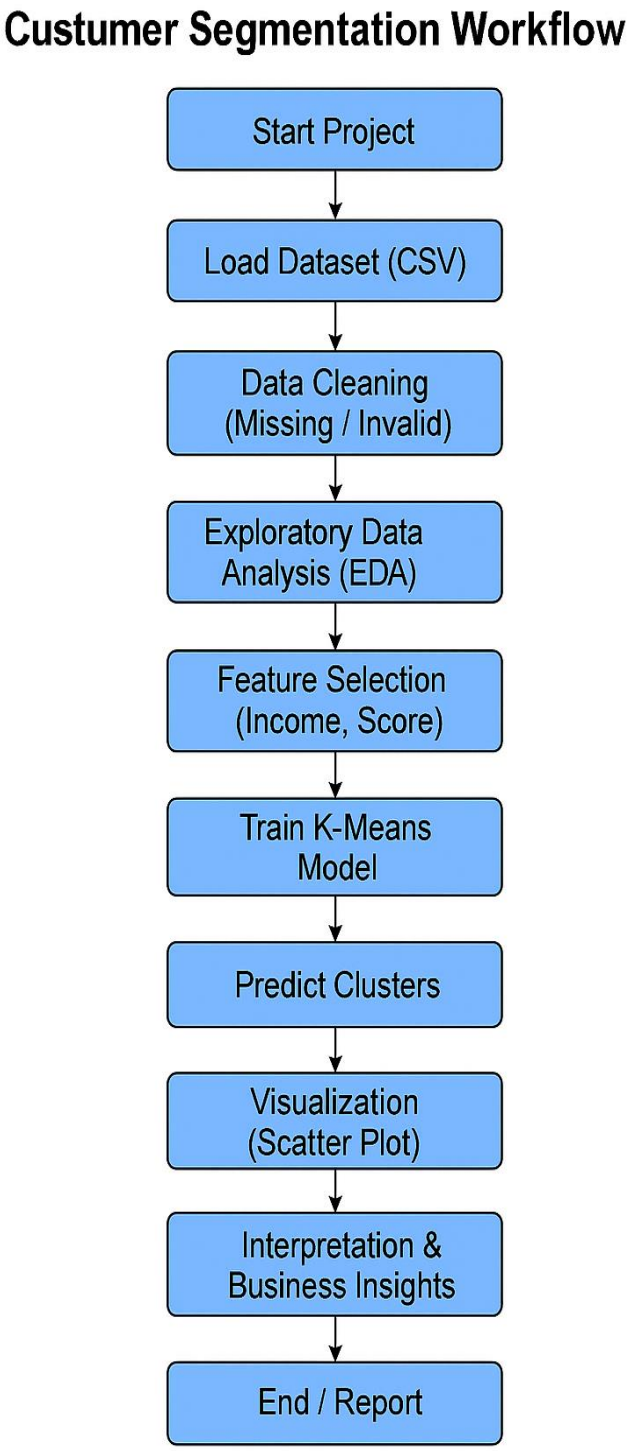
[9] Application of data mining techniques in customer relationship management: A literature review and classification:

E. Ngai, L. Xiu and D. Chau remarked despite the relevance of data mining techniques in customer relationship management (CRM), a complete literature evaluation and classification scheme for them are lacking. This was the first scholarly review of the application of data mining techniques to CRM that has been identified. It offers an academic database of literature from 2000 to 2006, comprising 24 journals, as well as a classification scheme for the articles. A total of 900 papers were found and assessed for their direct relation to CRM data mining methodologies. Following that, 87 articles were chosen, examined, and categorised. Each of the 87 articles was divided into four categories: customer identification, customer attraction, customer retention, and customer development, as well as seven data mining tasks (Association, Classification, Clustering, Forecasting, Regression, SequenceDiscovery and Visualization). Based on the main subject of the papers, they were further divided into nine sub-categories of CRM elements using various data mining approaches.

[10] An efficient k-means clustering algorithm: analysis and implementation:

He offered a set of n data points in d-dimensional space R/sup d/ and an integer k in k-means clustering, and the aim was to find a set of k centers in Rd to minimize the mean squared distance between each data point and its nearest center. Lloyd's (1982) technique is a prominent k-means clustering heuristic. He described the filtering algorithm, a simple and efficient implementation of Lloyd's k-means clustering technique. This algorithm is simple to build, as it only uses a kd-tree as a primary data structure. These models are extremely high-dimensional and sparse, posing computational and statistical issues not seen in low-dimensional dense data. The spherical k-means algorithm looks for clusters with strong coherence. They discovered that average cluster coherence is poor, implying that each thought vector is surrounded by a big gap in high-dimensional space. Furthermore, they discovered that at various resolutions, the average intra- and inter-cluster architectures were identical. He established the filtering algorithm's practical efficiency in two methods. First, he gave a data-sensitive study of the method's running time, which indicates that as the spacing between clusters rises, the process runs faster. Second, he provided a variety of empirical analyses based on both synthetically created data and genuine application data sets.

3.Work flow diagram:



Custumer Segmentation Workflow

4. Methodology

1. Data Collection

The dataset used in this project is the Mall Customers Dataset, provided in CSV format. This dataset contains customer information such as:

- Customer ID
- Gender
- Age
- Annual Income
- Spending Score

The data is loaded into the Python environment using the Pandas library. The .read_csv() function is used to import the file into a DataFrame structure for easy manipulation and analysis.

2. Data Preprocessing and Cleaning

The imported dataset is examined for:

- Missing values
- Duplicate records
- Inconsistent data types

Necessary preprocessing steps include:

- Handling missing values (if any)
- Removing irrelevant features (e.g., Customer ID for clustering)
- Converting categorical variables such as Gender into numeric format (if required)
- Ensuring numerical attributes are in appropriate data types for clustering

The dataset is verified to ensure it is clean and suitable for analysis.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis is performed to understand data distribution and relationships. The following analyses are conducted:

Univariate Analysis

Visualizations are generated to inspect individual feature distributions:

- Age distribution
- Annual Income distribution
- Spending Score distribution
- Gender-wise count

Techniques used:

- Histograms
- Bar charts
- Box plots

Bivariate Analysis

This step analyzes relationships between variables:

- Annual Income vs Spending Score
- Age vs Spending Score

Scatter plots are used to visually detect natural groups or cluster tendencies.

Descriptive Statistics

Statistical properties such as:

- Mean
- Median
- Standard deviation
- Minimum and maximum

4. Feature Selection

For clustering, the most relevant features are chosen based on business relevance and project guidelines:

1. Annual Income
2. Spending Score

These two features are selected to understand customer purchasing behavior.

5. K-Means Clustering Algorithm

K-Means is an unsupervised learning algorithm used to group customers based on similarity.
Algorithm Steps:

1. Choose the number of clusters k
2. Randomly initialize k centroids
3. Assign each data point to the nearest centroid
4. Compute new centroids
5. Repeat until convergence occurs

The algorithm minimizes Within-Cluster Sum of Squares (WCSS) using Euclidean distance.

6. Optimal Cluster Selection (Elbow Method)

To determine the best number of clusters, the Elbow Method is applied.

Steps:
Run K-Means for k = 1 to k = 10

Calculate inertia (WCSS) for each value of k

Plot the WCSS graph

Identify the "elbow point" where reduction slows

This point represents the optimal number of clusters.

7. Model Training

With the optimal k identified:

K-Means is fit to the dataset
Cluster labels are predicted
Each customer is assigned a cluster number

8. Cluster Visualization

Scatter plots are generated using:

- X-axis → Annual Income
- Y-axis → Spending Score
- Color → Cluster Group

Centroids are plotted using distinct markers.
This provides visual clarity of customer segments**.**

9. Cluster Interpretation and Business Analysis

Each cluster is analyzed and interpreted:
Example Insights:

- High Income – High Spending → Premium customers
- High Income – Low Spending → Potential customers
- Low Income – High Spending → Impulsive buyers
- Low Income – Low Spending → Low-value customers

These findings help businesses:

- Design personalized marketing strategies
- Improve customer retention
- Optimize budget allocation
- Increase profitability

5. Tools and Techniques used

Tool	**Purpose**
Numpy	Numerical operations
Pandas	Data handling
Matplotlib/Seaborn	Data Visualization
Scikit-learn	K-Means implementation
Jupyter Notebook	Development environment

6. Results and Discussion:

After applying the K-Means clustering algorithm to the Mall Customers dataset, customer groups were successfully identified based on Annual Income and Spending Score. The Elbow Method was used to determine the optimal number of clusters, and the elbow point was observed at K = 5, indicating that five customer segments provide the most meaningful grouping.

The final K-Means model was trained using these five clusters, and each customer was assigned a cluster label. The results were visualized using a scatter plot, where:

- Each data point represented a customer
- Colors indicated cluster membership
- Cluster centroids were clearly visible
- Natural grouping patterns were formed

The results showed that customers tend to group based on purchasing behavior and financial capability. Clear segmentation was observed with minimal overlapping between major clusters, demonstrating that K-Means effectively separated customers based on similarity.

1.Optimal Number of Clusters By Elbow Method

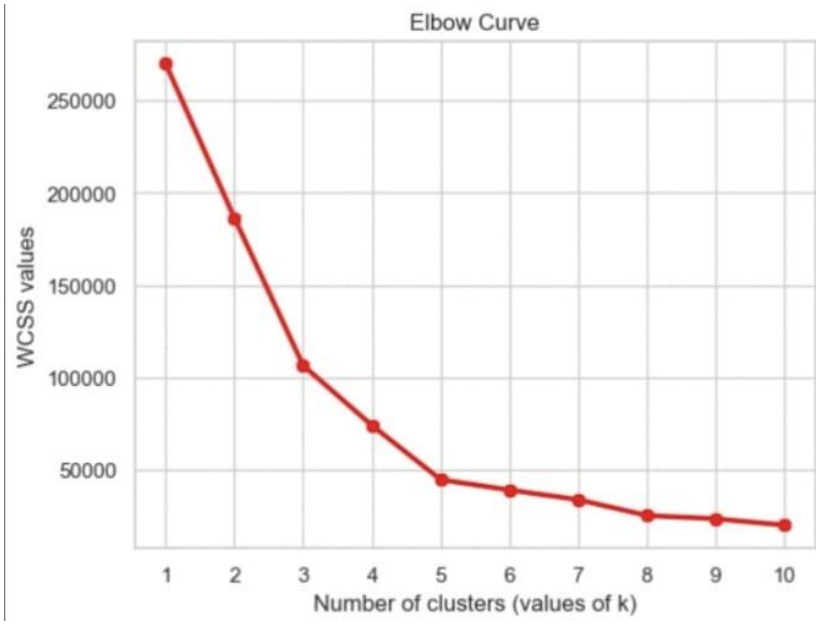


Fig1. Optimal number of clusters using Elbow method

2.Cluster Implementation

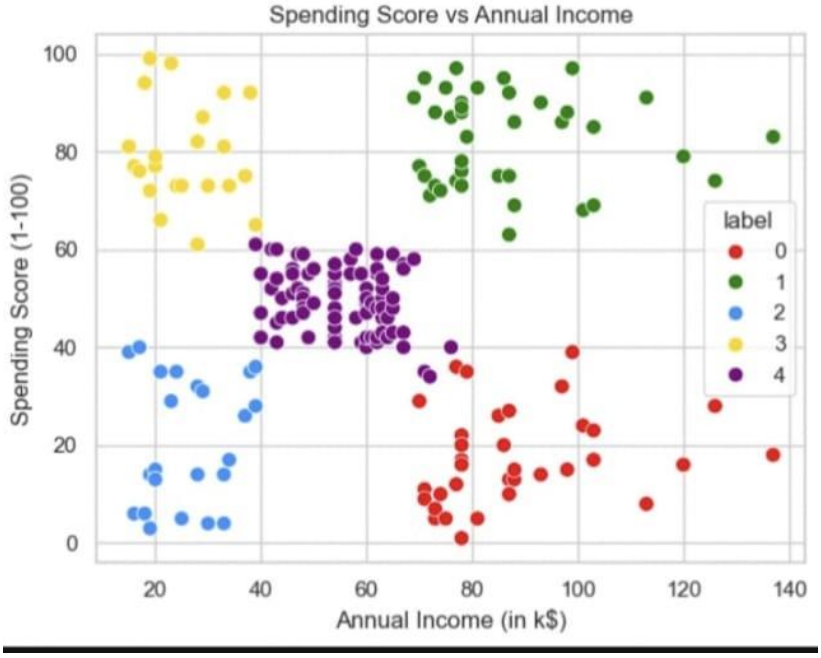


Fig 2. Visualization of Clusters

Cluster Interpretation

Each cluster formed by the model represented a unique type of customer group:

Cluster 0: High Income – High Spending

These customers are premium buyers and contribute the highest value to the business. They have strong purchasing capacity and high engagement.

Cluster 1: High Income – Low Spending

This group includes wealthy but less active customers. They represent an opportunity for personalized marketing and loyalty-based offers.

Cluster 2: Low Income – High Spending

Impulse buyers who spend more than expected relative to income. Offers and promotions can help retain these customers.

Cluster 3: Low Income – Low Spending

This group represents conservative buyers with minimal revenue contribution. Cost-focused marketing strategies are suitable.

Cluster 4: Medium Income – Medium Spending

Average customers with steady buying behavior. They serve as a stable revenue source.

4.Visualization Of Income and Spending Score

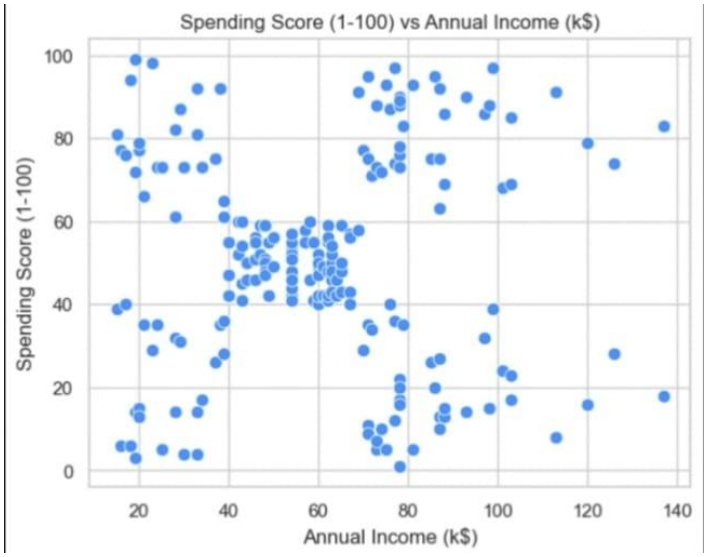


Fig 3. Scatter Plot of Income across Spending Score

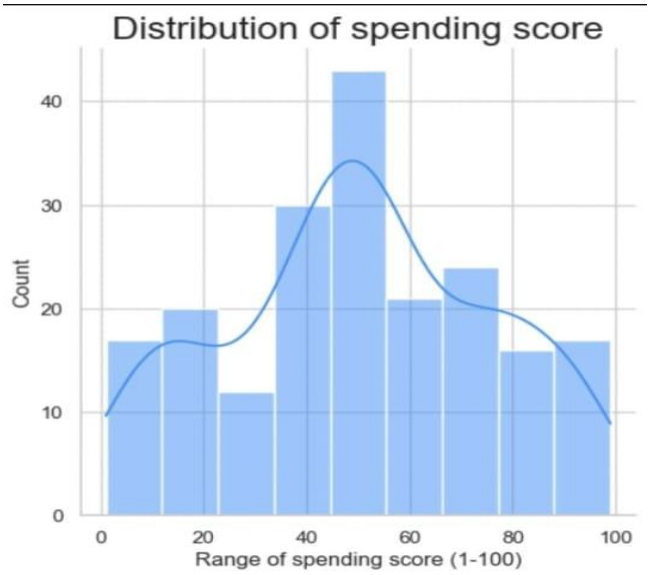


Fig 4. Distribution of age

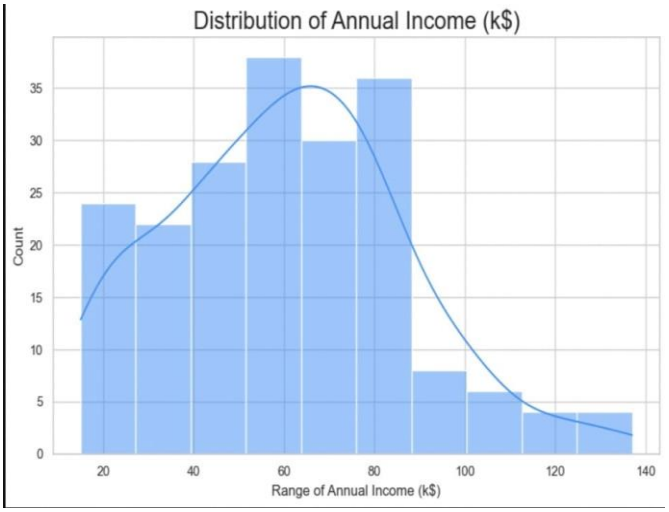


Fig 5 Distribution of Annual Income

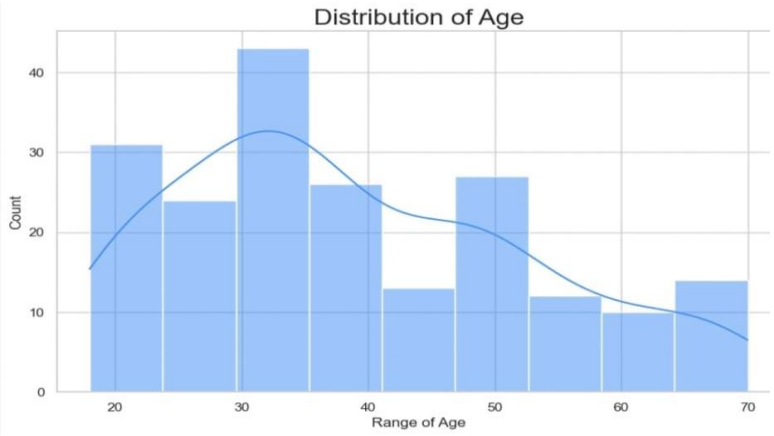


Fig 6. Distribution Of Age

Fig 7. Gender Distribution

7. Conclusion

This project successfully implemented customer segmentation using the K-Means clustering algorithm on the Mall Customers dataset. The primary objective was to group customers based on their purchasing behavior and income patterns to derive meaningful business insights. Through data preprocessing, exploratory data analysis, and clustering, the project demonstrated the practical application of machine learning in customer analytics.

The Elbow Method was effectively used to determine the optimal number of clusters, and the K-Means model segmented the customers into meaningful groups. These clusters were analyzed and interpreted to understand different customer types, such as high-value customers, conservative spenders, and potential customers. The results show that customer behavior can be effectively analyzed using unsupervised learning techniques, enabling organizations to make data-driven decisions.

The segmentation model provides businesses with the ability to design targeted marketing campaigns, improve customer engagement, and increase profitability. By understanding customer behavior patterns, companies can reduce marketing costs, optimize resource allocation, and improve customer satisfaction.

Overall, this project demonstrates the importance of data analysis and machine learning in understanding customer needs. K-Means clustering proved to be an effective technique for segmenting customers with minimal supervision, and the project achieved all the intended objectives successfully. The methodology and results can be extended to real-world datasets for advanced customer behavior analysis and marketing strategy development.

8. References

[1] Prasanta Bandyopadhyay Montana State University – Bozeman Faculty Member, International Journal of Contemporary Economics and Administrative Sciences, Volume- II, Publication Date**:** 2018

[2] C. M. S. R. a. K. V. N. T. Sajana, "A Survey on," in Indian Journal of Science and Technology, Volume 9, Issue 3, Jan 2016.

[3] A. B. P. E. Shreya Tripathi, "Approaches to," in International Journal of Engineering and Technology, Volume 7, 2018.

[4] O. A. Abbas, "Comparisons between data clustering algorithms", The International Arab Journal of Information Technology, vol. 5, no. 3, pp. 320 -325, 2008.

[5] I. Dhillon and D. Modha, "Concept decompositions for large sparse text data using clustering", Machine Learning, vol. 42, no. 1/2, pp. 143-175, 2001.

[6] A. Ansari and A. Riasi, "Customer clustering using a combination of fuzzy c-means and genetic algorithms", International Journal of Business and Management, vol. 11, no. 7, pp. 59-66, 2016

[7] Y. Chen, et al., "Identifying patients in target customer segments using a two-stage clustering-classification approach: A hospital-based assessment", Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012

[8] M. Namvar, M. Gholamian and S. KhakAbi, "A two-phase clustering method for intelligent customer segmentation", in International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, 2010, pp. 215-219.

[9] E. Ngai, L. Xiu and D. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, vol. 36, no. 2, pp. 2592-2602, 2009.

T. Kanungo, et al., "An efficient k-means clustering algorithm: analysis and implementation",IEEE Transactions on Pattern Analysis and Machine Intelligence,