Minor Project Report: Consumer Segmentation using K-Means Clustering

-Deepak Balu M

1. Introduction

In today’s competitive market, understanding consumer behaviour is crucial for companies to offer personalized services and products. With the explosion of data, businesses are shifting from traditional marketing to data-driven strategies. Clustering is a major tool in data analytics, allowing businesses to group customers based on similar traits. This project explores the concept of consumer segmentation using K-Means clustering. Through this project, we aim to divide a group of consumers into smaller segments based on age, income, health awareness, and app usage time. By doing so, we can create targeted strategies and enhance customer satisfaction. Artificial Intelligence and Machine Learning (AI/ML) have become essential in simplifying complex tasks like these, providing more accurate and actionable insights.

2. Problem Statement

Businesses often struggle to address the diverse needs of their customer base. Without proper segmentation, marketing campaigns become generalized and less effective, resulting in wasted resources and missed opportunities. The lack of personalized engagement may lead to customer dissatisfaction. Our objective is to solve this problem by implementing a machine learning model that can categorize consumers into distinct groups based on similarities in their behaviour and characteristics. These clusters will help in devising focused marketing strategies, personalizing offers, and improving overall user engagement. The challenge lies in appropriately preprocessing data, selecting a suitable clustering algorithm, and interpreting the resulting segments meaningfully.

3. Objectives

The primary objective of this project is to implement the K-Means clustering algorithm to group consumers based on multiple parameters like age, monthly income, app usage time, and health awareness. Other objectives include visualizing the clusters, analysing the resulting groups, and building consumer personas based on the clustering output. Additionally, the project aims to design an interactive dashboard that displays segmentation results, allowing for better interpretation and future decision-making. This project also focuses on developing an understanding of machine learning workflows, feature scaling, and model evaluation in an applied real-world context.

4. Methodology

The methodology of this project follows a structured workflow starting with data collection, followed by preprocessing, applying K-Means clustering, and interpreting the results. Initially, synthetic data was created considering variables such as Age, Monthly Income, App Usage Time (in hours), and Health Awareness (in percentage). The data was then scaled using standardization techniques to ensure all features contribute equally to the model. An elbow method was used to determine the optimal number of clusters. After selecting the number of clusters, K-Means clustering was applied to segment the consumers. Post clustering, visualization tools like scatter plots and dashboards were used to understand the characteristics of each cluster. Finally, consumer personas were created based on cluster behaviour.

5. Tools and Technologies Used

For the execution of this project, several tools and technologies were utilized. Python was the primary programming language because of its extensive libraries for data science. The libraries used include Pandas for data manipulation, NumPy for numerical computations, Scikit-learn for machine learning algorithms, and Matplotlib and Seaborn for data visualization. Google Colab was used for coding and executing the project due to its ease of access and free GPU support. Streamlit was chosen for creating an interactive dashboard for visualizing the results. Together, these tools offered a complete ecosystem for building, evaluating, and showcasing the clustering project efficiently.

6. Data Collection

The dataset used in this project was synthetically generated to simulate a realistic consumer base. It contained four important features — Age, Monthly Income, App Usage Time in hours, and Health Awareness percentage. Generating synthetic data allowed us to control variability and simulate different consumer behaviours. The dataset consisted of approximately 100-200 entries to represent a medium-sized consumer group. Data integrity was ensured by making sure no null or inconsistent values were present. Additionally, the distributions of each feature were checked to ensure they resembled real-world scenarios, making the analysis more meaningful and relatable for business applications.

7. Data Preprocessing

Preprocessing is a critical step in preparing the dataset for machine learning. The first step involved checking for missing or inconsistent values, and ensuring that all data was clean. Next, feature scaling was applied using StandardScaler from Scikit-learn. This standardization step was necessary because features like income and age were on very different scales and could negatively affect the clustering results. Scaling transformed the data so that all features had a mean of 0 and a standard deviation of 1, allowing the K-Means algorithm to perform optimally. After preprocessing, the dataset was ready for model building.

8. Model Building

The K-Means clustering model was built using Scikit-learn. The elbow method was employed first to determine the optimal number of clusters by plotting Within-Cluster Sum of Squares (WCSS) against the number of clusters. Based on the elbow point observed from the graph, three clusters were selected. The model was then trained on the scaled dataset. After training, each consumer was assigned a cluster label indicating the group they belonged to. The model's output was stored in the data frame, which allowed further analysis and persona creation based on the cluster assignments.

9. Model Evaluation and Analysis

The model evaluation was done primarily through visualization. Scatter plots were generated to display the clusters, with colour coding for easy differentiation. The cluster centres were also visualized to understand the average characteristics of each group. An additional table showing average values of Age, Monthly Income, App Usage Time, and Health Awareness per cluster was created for deeper insights. It was observed that one cluster consisted of younger, tech-savvy consumers with high app usage but low health awareness, while another group showed older, more health-conscious consumers. The third group was a mix of young professionals with high income and moderate health awareness.

10. Consumer Persona Creation

Based on the cluster analysis, three consumer personas were created:

- **Persona 1**: Young Adventurers - These individuals are aged around 33 years, have a high income, use apps moderately, but have lower health awareness.
- **Persona 2**: Health Conscious Mid-Lifers - Aged around 45 years, these users have average income, high app usage, and strong awareness regarding health.
- **Persona 3**: Young Professionals - Around 34 years of age, with high income and high app usage time, but slightly lower health awareness compared to mid-lifers.

These personas help businesses personalize their services according to each group's preferences and habits.

11. Dashboard Creation

An interactive dashboard was created using Streamlit to visualize the clustering results. The dashboard displays the uploaded dataset, the clustered groups, and a scatter plot of different clusters. The user can interact with the dashboard, upload their own datasets, and view the segmentation live. This approach makes it easier for non-technical stakeholders to explore the results without digging into the code, promoting easier and faster decision-making. The dashboard also enhances the presentation quality of the project, making it more professional and insightful.

12. Challenges Faced

Throughout the project, several challenges were encountered. Determining the optimal number of clusters was initially confusing until the elbow method was clearly understood. Data preprocessing also posed challenges, especially with standardization and feature scaling. Another difficulty was creating consumer personas that accurately reflected the clusters. Finally, deploying the dashboard using Streamlit had a learning curve, particularly with setting up the correct environment. However, all challenges were overcome through research, trial and error, and persistence, resulting in a successful project outcome.

13. Conclusion

This project demonstrates how machine learning, specifically K-Means clustering, can be effectively applied to consumer segmentation. Through synthetic data generation, preprocessing, clustering, and dashboard creation, we learned the entire pipeline of a typical ML project. Consumer segmentation using data analytics can significantly improve marketing strategies and customer satisfaction by providing personalized experiences. The practical knowledge gained in this project will be beneficial in future data-driven projects and real-world business applications.

14. Future Enhancements

Future work on this project can include using more advanced clustering techniques like DBSCAN or Hierarchical Clustering for better results. Incorporating more consumer attributes like lifestyle preferences, spending behaviour, or geographic location can further enhance the accuracy of segmentation. A recommendation system could also be built on top of the clustering results to suggest personalized products or services to users. Furthermore, deployment of the dashboard online using cloud platforms like Heroku or AWS can make the project accessible to a wider audience.

15. References

- Scikit-learn Documentation
- Matplotlib and Seaborn Documentation
- Streamlit Official Documentation
- Free Aqua Project Resources (for reference designs)