Customer Segmentation using K-Means Clustering Algorithm

Abstract:

This code demonstrates customer segmentation using K-Means clustering algorithm in Python. The dataset used in this project is the Mall Customers dataset, which contains information about customers' annual income and spending scores. The aim is to identify different segments of customers based on their spending behaviour.

Table of Contents:

Introduction:

In marketing, customer segmentation is a crucial task that involves dividing a customer base into groups of individuals who share similar characteristics. This allows businesses to target their customers with personalised marketing strategies and improve customer satisfaction. One of the most popular methods of customer segmentation is clustering, which involves grouping customers based on similar behaviour patterns. K-Means clustering is a widely used algorithm for this purpose, as it is simple and efficient.

The purpose of this project is to demonstrate customer segmentation using K-Means clustering algorithm in Python. The dataset used in this project is the Mall Customers dataset, which contains information about customers' annual income and spending scores. The aim is to identify different segments of customers based on their spending behaviour.

Existing Method:

Various clustering algorithms have been used for customer segmentation, such as K-Means, Hierarchical Clustering, and DBSCAN. K-Means is a popular choice due to its simplicity and efficiency. It is an unsupervised learning algorithm that clusters data points into k clusters based on their proximity to the cluster centroids.

Hierarchical Clustering:

The existing method uses hierarchical clustering with the AgglomerativeClustering algorithm to cluster customer data based on their annual income and spending score. The goal is to identify distinct groups of customers based on their purchasing behavior in order to better understand and target their needs.

The first step of the proposed method is to load the data into a pandas DataFrame and extract the relevant columns as features for clustering. In this case, the 'Annual Income (k$)' and 'Spending Score (1-100)' columns are chosen as they are likely to be strong predictors of customer behaviour.

Once the data is loaded and the features are extracted, the next step is to standardise the data using the StandardScaler. This step is important as it helps to ensure that each feature is given equal weight in the clustering process, regardless of its scale or units.

After standardising the data, the AgglomerativeClustering algorithm is applied to perform clustering. In this case, the algorithm is set to produce 5 clusters and uses 'euclidean' as the distance metric and 'ward' as the linkage criterion. The 'euclidean' distance metric is a popular choice as it is easy to compute and interpret, while the 'ward' linkage criterion is a hierarchical clustering method that seeks to minimise the variance between clusters.

Once the clustering is complete, a dendrogram is plotted to visualise the results. A dendrogram is a tree-like diagram that displays the hierarchical relationships between

clusters. It can be useful in identifying the optimal number of clusters as well as any sub-clusters within each cluster.

Finally, the cluster labels are added back to the original data set so that each customer can be assigned to a specific cluster. This allows businesses to gain a deeper understanding of their customer base and tailor their marketing strategies to better meet the needs and preferences of each cluster.

Overall, the proposed method is a powerful technique for analysing customer data and identifying distinct groups of customers based on their purchasing behaviour. By clustering customers based on their annual income and spending score, businesses can gain valuable insights into their customer base and develop targeted marketing strategies that are more likely to resonate with each group.

Proposed Method with Architecture:

1. In this project, we use the K-Means clustering algorithm to cluster customers based on their annual income and spending scores. The algorithm is implemented in Python using the scikit-learn library. The architecture consists of the following steps:

2. Data loading: The Mall Customers dataset is loaded using the pandas library.

3. Data preprocessing: The dataset is checked for null values, and the features to be used for clustering (annual income and spending scores) are selected.

4. Determining the optimal number of clusters: The elbow method is used to determine the optimal number of clusters.

5. Applying K-Means clustering: K-Means clustering algorithm is applied with the optimal number of clusters.

6. Visualising the clusters: The clusters are visualised using scatter plots with different colours for each cluster.

Methodology:

The methodology consists of the following steps:

- Load the Mall Customers dataset: The dataset is loaded using the pandas library, and the first five rows of the dataset are printed using the "head()" method.

- Check for null values: The dataset is checked for null values using the "isnull().sum()" method

.

- Select the features to be used for clustering: The features "Annual Income (k$)" and "Spending Score (1-100)" are selected for clustering, and they are stored in a separate variable "X".

- Determine the optimal number of clusters: The elbow method is used to determine the optimal number of clusters. This involves applying K-Means clustering for different values of "k" and calculating the Within-Cluster-Sum-of-Squares (WCSS) for each value of "k". WCSS is the sum of the squared distances between each data point and its assigned centroid. The optimal value of "k" is the value after which there is no significant decrease in WCSS. The WCSS values for different values of "k" are stored in a list "wcss", and a line plot is used to visualise the elbow curve.

- Apply K-Means clustering: K-Means clustering algorithm is applied with the optimal number of clusters determined in the previous step. The "fit_predict()" method is used to fit the data and predict the cluster labels for each data point. The cluster labels are stored in a separate variable "Y".

- Visualise the clusters: The clusters are visualised using scatter plots with different colours for each cluster. The "scatter()" method is used to plot the data points, and the "scatter()" method is called separately for

- After designing and testing our model, it is now time to implement it. We can do this by using our model to cluster the customer data based on their Annual Income and Spending Scores. The output will be the five clusters we identified, along with their corresponding centroids.

- We can then use this information to make business decisions. For example, we can use this model to create targeted marketing campaigns for each cluster. For example, we can create a marketing campaign that focuses on luxury items for customers in the higher income and higher spending cluster, while creating a campaign that focuses on deals and discounts for customers in the lower income and lower spending cluster.

Proposed method -2

The objective of this analysis is to segment mall customers based on their annual income and spending score, using a clustering algorithm. This will help the mall management team to better understand their customer base and target them more effectively with marketing and promotional activities.

Data Exploration:

The dataset contains 200 entries, with five features: CustomerID, Gender, Age, Annual Income, and Spending Score. We first explored the data to understand the distribution of the features and their relationship with each other. We found that the majority of the customers were in the age range of 20-50 years, and there was an almost equal distribution of male and female customers. We also found that the annual income of customers ranged from $15k to $137k, with an average of $60k, and the spending score ranged from 1 to 99, with an average of 50.

Methodology:

We used the K-means clustering algorithm to segment customers into groups based on their annual income and spending score. We first determined the optimal number of clusters using the elbow method, which involved running the K-means algorithm for a range of cluster numbers and plotting the within-cluster sum of squares (WCSS) against the number of clusters. The point at which the WCSS starts to level off is the optimal number of clusters. In this case, we determined that three clusters were optimal.

We then ran the K-means algorithm with three clusters and visualized the results using a scatter plot. We plotted each customer's annual income against their spending score, with each point colored based on the cluster they belonged to. We also plotted the cluster centers as yellow dots.

Results and Interpretation:

We found that the customers could be segmented into three distinct groups based on their annual income and spending score. The first group, Cluster 1, had a low annual income and low spending score, indicating that they were likely budget-conscious shoppers who visited the mall for essentials. The second group, Cluster 2, had a high annual income and high spending score, indicating that they were likely affluent shoppers who visited the mall for luxury items. The third group, Cluster 3, had a moderate annual income and moderate spending score, indicating that they were likely value-conscious shoppers who visited the mall for a mix of essentials and luxury items.

Conclusion

In conclusion, the method using annual income and spending score as features is more effective for clustering mall customers based on their purchasing behaviour than the method using gender and age. The results obtained from the second method show that customers can be divided into 5 clusters based on their purchasing behaviour, which can be useful for targeted marketing and promotions. Cluster 1 and 5 customers are potential high-value customers, while Cluster 2 and 4 customers can be targeted for promotions to increase their spending. Cluster 3 customers are average spenders, and can be retained through loyalty programs or personalised offers. Overall, this analysis provides valuable insights into customer segmentation for mall management Overall, customer segmentation is a powerful tool for businesses to better understand their customers and make data-driven decisions. By using a clustering algorithm like K-Means, we can create distinct customer segments that can help us make better business decisions and ultimately drive growth.