JAYPEE INSTITUTE OF INFORMATION AND TECHNOLOGY

Department of Computer Science Engineering & IT

Introduction to Big Data and Data Analytics

Project Report
On
Customer Segmentation

Submitted By:
Shubham Garg
9919103057
F2

Submitted To:
Dr. Neeraj Jain

Problem Statement

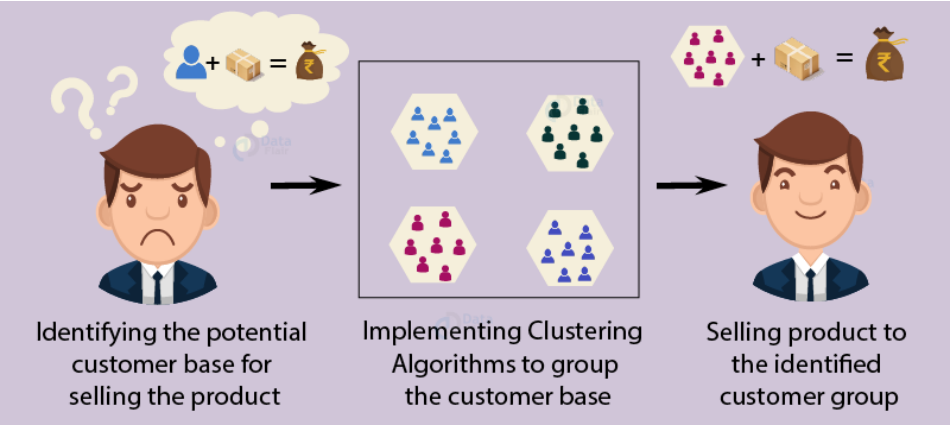
A supermarket Mall owner want to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.

In this project, We will be performing an unsupervised clustering of data on the customer's records from a Supermarket firm's database.

Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster. We will divide customers into segments to optimize the significance of each customer to the business.

Introduction

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this project, We will make use of K-means clustering which is the essential algorithm for clustering unlabeled dataset.



What is Customer Segmentation?

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioral patterns play a crucial role in determining the company direction towards addressing the various segments.

Detailed Design

Selection of Model

K-means clustering is an unsupervised learning algorithm, meaning that it is used for unlabeled datasets. In order to categorize this data on the basis of their similarity, we use the K-means clustering algorithm.

It is an unsupervised technique to group data in the order of their similarities. We then find patterns within this data which are present as k-clusters.These clusters are basically data-points aggregated based on their similarities.

Tools & Technologies Used

Python:We used the most common programming language that is used for building the machine learning model i.e., Python.

Jupyter Notebook: The IDE (Integrated Development Environments) that is used to build the unsupervised model is Jupyter Notebook.

Anaconda Command Prompt: It is used to launch Jupyter Notebook

Libraries used

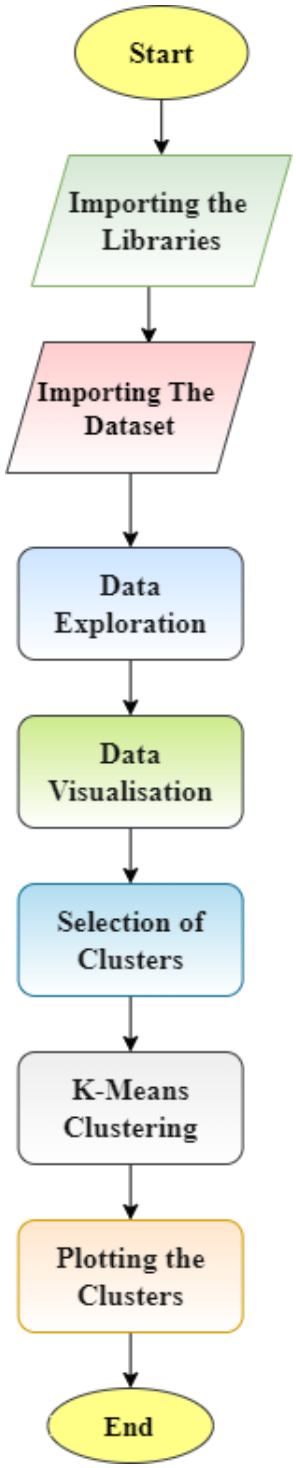
LIBRARY	DESCRIPTION
Pandas	The Pandas is one of the most important libraries while dealing with data, all the data analysis, and data manipulation can be done with the help of pandas
Numpy	All the mathematical functions can be computed through the NumPy library.
Seaborn	Seaborn library is used for visualization.
Matplotlib	Matplotlib is used for visualization

Dataset used

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

Implementation

Project Flowchart



DATA EXPLORATION

Steps involved:

- Importing all the necessary libraries for the study.

```
In [19]:
#import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

- Importing the data as a CSV file.

```
In [24]:
#Import the dataset
dataset = pd.read_csv('Mall_Customers.csv')
dataset.head()
```

Out[24]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

- Checking for all the information such as the number of columns and rows present in this dataset: There are 200 rows and 5 columns present.

```
In [25]:
dataset.shape
```

```
Out[25]: (200, 5)
```

- Types of variables: 1 categorical and 4 numerical.

```
In [27]:
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

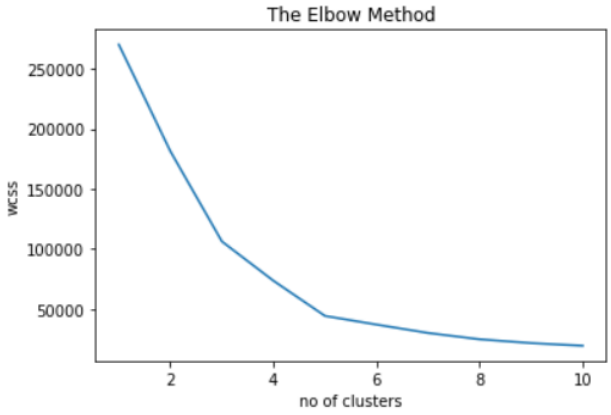
- Checking for the NULL or NAN value in the dataset. No Null Values in the dataset

```
In [23]:
dataset.isnull().sum()
```

```
Out[23]: CustomerID              0
Gender                  0
Age                     0
Annual Income (k$)      0
Spending Score (1-100)  0
dtype: int64
```

DATA VISUALISATION

Data visualization is the demonstration of taking (data) and setting it into a visual setting, for example, a guide or chart. Data visualizations make of all shapes and size data simpler for the human mind to understand, and visualization additionally makes it simpler to detect patterns, trends, and outliers in gatherings of data.



MODELLING APPROACH

K-Means Clustering Model

For better management and to create the homogenous groups of entities, most frequently used application is clustering. The clustering is a divide and conquer strategy which divides the dataset into a homogenous group. Clustering algorithms are unsupervised learning algorithms in which labels classes are not defined.

The function we used to define the clustering technique is K-Means. The K-Means uses the Euclidean distance to find the distance between the two or more observation. It takes the mean (average) of every sample in the dataset and fit them accordingly.

```
In [32]:  #Model Build
          kmeansmodel = KMeans(n_clusters= 5, init='k-means++', random_state=0)
          y_kmeans= kmeansmodel.fit_predict(X)

In [33]:  #Visualizing all the clusters

          plt.scatter(X[y_kmeans == 0, 0], X[y_kmeans == 0, 1], s = 100, c = 'red', label = 'Cluster 1')
          plt.scatter(X[y_kmeans == 1, 0], X[y_kmeans == 1, 1], s = 100, c = 'blue', label = 'Cluster 2')
          plt.scatter(X[y_kmeans == 2, 0], X[y_kmeans == 2, 1], s = 100, c = 'green', label = 'Cluster 3')
          plt.scatter(X[y_kmeans == 3, 0], X[y_kmeans == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
          plt.scatter(X[y_kmeans == 4, 0], X[y_kmeans == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
          plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow', la
          plt.title('Clusters of customers')
          plt.xlabel('Annual Income (k$)')
          plt.ylabel('Spending Score (1-100)')
          plt.legend()
          plt.show()
```

Experimental Result

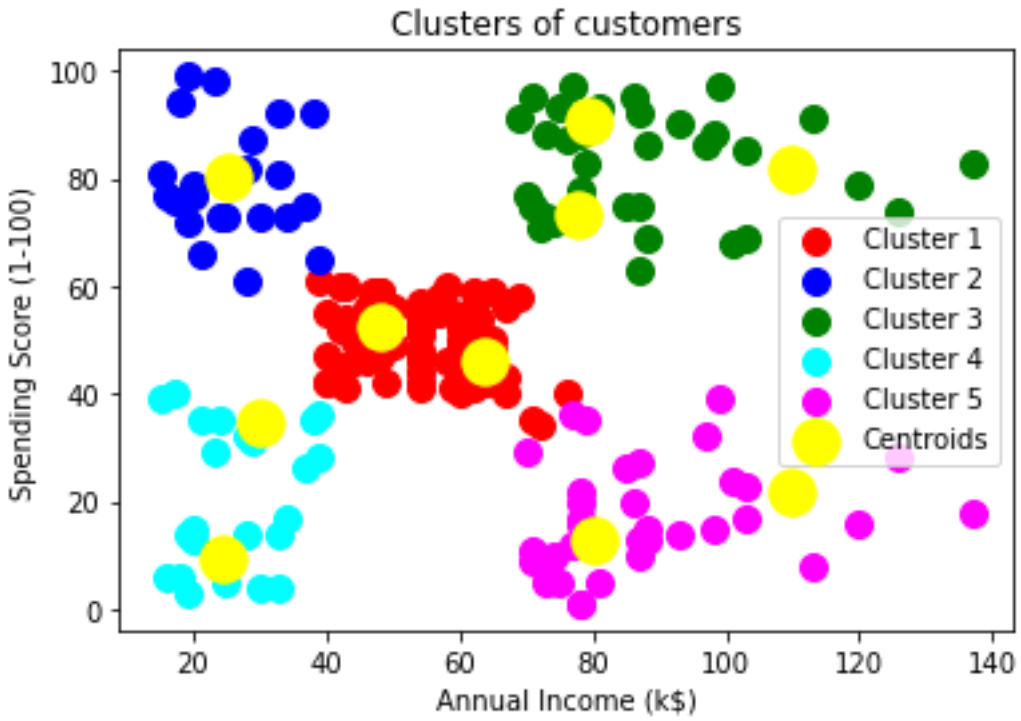
Cluster 1 (Red Color) -> earning high but spending less

Cluster 2 (Blue Color) -> average in terms of earning and spending

Cluster 3 (Green Color) -> earning high and also spending high [TARGET SET]

Cluster 4 (cyan Color) -> earning less but spending more

Cluster 5 (magenta Color) -> Earning less , spending less



Conclusion

This project has solved the problem statement to find the customers who can converge easily by clustering the applicable data.

Elbow curve method is used to identify the centroids of the every cluster and using that cluster was defined. The machine learning function that was used was the K-Means technique. Clustering helped to identify the customers who earn high as well as spend high depicted in cluster 3(Green Color) .

REFERENCES

[1]https://www.python.org

[2] https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[3] https://towardsdatascience.com/customer-segmentation-with-machine-learning-a0ac8c3d4d84

[4] https://data-flair.training/blogs/k-means-clustering-tutorial/