Report on Peer-graded Assignment: Course Final Project

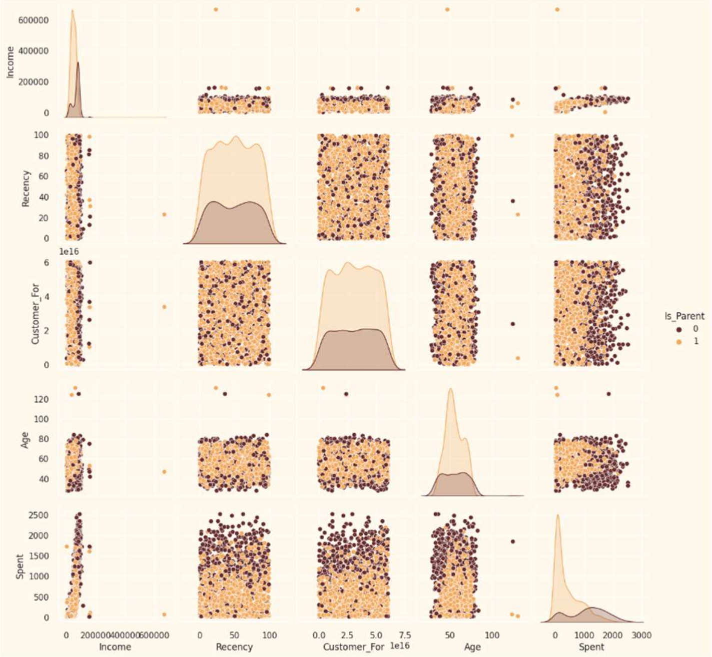
About the dataset:

The dataset contains the customer records from the grocery store database. Originally , the dataset has 2400 rows and a total of 29 feature columns . The 29 attributes can be divided into 4 different subsets .

Customer's Information	Products	Place	Promotion
ID	MntWines	NumWebPurchaces	NumDealsPurchases
Year_Birth	MntFruits	NumCatalogPurchases	AcceptedCmp1
Education	MntMeatProducts	NumStore Purchases	AcceptedCmp2
Marital_Status	MntFishProducts	NumWebVisits Month	AcceptedCmp3
Complain	MntSweetProduct		AcceptedCmp4
Income	MntGoldProds		AcceptedCmp5
Kidhome			Response
Teenhome			
Dt_Customer			
Recency			

Data Cleaning :

- By using the data.info() method , we realized that the “Income” feature has null values . To eliminate these , we use the dropna() function . Now our new data set has 2216 rows with no null values .
- We also notice that the Dt_Consumer is not parsed as DateTime . So using the pandas function , we convert it into DateTime format.
- We also realise that not all the features are given in the necessary format . Like we are given age in the year of birth format and not as Age. So we first recognise all such features and engineer to find newer features .(listed in the python file attached below)
- Now we plot and check all the features by using “pairplot”. Following is the result



- From the pairplot , we also detect the presence of outliers which can critically affect our algorithm. Hence , we remove them by manual checking …

Data Preprocessing:

- Firstly , we have two columns who are of type object . To convert them , we use the LabelEncoder() .
- Secondly , we have many features which are not in the required scale. So we use the StandardScaler() and convert them. The observation of many clustering algorithms depends on the scale of the data so this step is extremely crucial .
- Lastly , we use dimensionality reduction . The current data has too many features. Hence, we use Principle Component Analysis(PCA) and reduce the number of features to 4 columns. In this , we use the KernelPCA and GridSearchCV to obtain the optimum number of features

Objective of the Analysis:

In this project, I will be performing an unsupervised clustering of data on the customer's records from a groceries firm's database. Customer segmentation is the practice of separating customers into groups that reflect similarities among customers in each cluster. I will divide customers into segments to optimize the significance of each customer to the business. To modify products according to distinct needs and behaviours of the customers. It also helps the business to cater to the concerns of different types of customers. Through the analysis , we understand which clustering method is the best for the current dataset and can be used for future classification on test-data .We choose the scoring metric as “Silhouette Score” and “Davies-Bouldin Index”.

About the Classification Models:

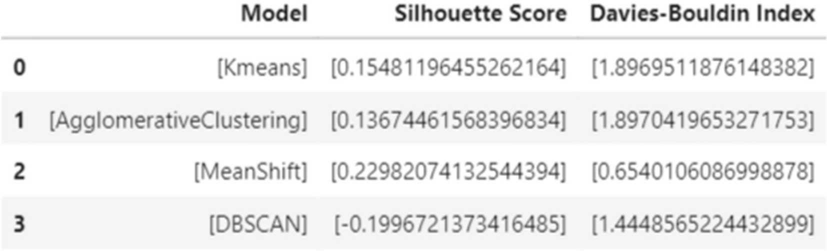
1. KMeans Clustering
2. DBSCAN
3. Agglomerative Clustering
4. Mean Shift Clustering

Best Score : Agglomerative Clustering

KeyFindings :

From the analysis, we see that models with low Silhouette Score and high Davies-Bouldin Index are the best models. In our case, MeanShift produces a high Silhouette Score and a low Davies-Boulidin index. So we prefer AgglomerativeClustering over it as it produces a optimal result.

1. From the ElbowPoint method , we get the optimal number of clusters as 4.
2. We also generated a negative silhouette score for DBSCAN . This indicates that a point is already assigned in the wrong cluster



	Model	Silhouette Score	Davies-Bouldin Index
0	[Kmeans]	[0.15481196455262164]	[1.8969511876148382]
1	[AgglomerativeClustering]	[0.13674461568396834]	[1.8970419653271753]
2	[MeanShift]	[0.22982074132544394]	[0.6540106086998878]
3	[DBSCAN]	[-0.1996721373416485]	[1.4448565224432899]

Further Suggestions:

The notebook I have used is attached below . It can be implemented on various similar datasets with same feature engineering on the data .

1. ipynb Notebook :
   https://github.com/darKKnight14110/Customer-Segmentation---Clustering-Methods/blob/main/customer-segmentation-classification.ipynb
2. Dataset :
   https://www.kaggle.com/datasets/mohandeep20/marketingcampaign