Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name – Shri Prakash Yadav Email – ydvshri1412@gmail.com Contribution – Unsupervised Machine Learning (Customer Segmentation)
Please paste the GitHub Repo link.
Github Link:- **https://github.com/ydv1412/Customer_segmentation**
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

This was an unsupervised machine learning(clustering) project. In this project I have been provided with online retain dataset. This dataset consists of 541909 rows and 8 features. My task in this project was to perform EDA and group this dataset into optimum clusters based on features, using different clustering techniques.

I have divided this project into five major parts which are Exploratory Data Analysis, RFM analysis, handling null values, outliers' removal and dataset clustering.

In the Dataset Analysis part, I plotted different plots to see insights from dataset for example number of countries, number of customers from each country, amount spent by customers of each country etc. I plotted pie chart to see what percentages of the invoices get cancelled, what percentage of customers return to online retail store after purchasing once. I also plotted correlation matrix to see how all the features are dependent on each other.

In RFM analysis part the idea was to segment customers based on when their last purchase was, how often they've purchased in the past, and how much they've spent overall. RFM stands for recency, frequency, and monetary. For this I created three features monetary, frequency and recency and analyzed them using different plots.

In handling null values part to impute categorical features I used the most frequent value of the feature and for imputing numerical feature I used random sample from the feature.

In the outlier removal part firstly, I have standardized all numerical features. Then I removed the outlier points using interquartile range.

In the clustering part, I used three clustering techniques which are K-means, DBSCAN and hierarchical clustering to cluster our dataset (and RFM dataset that I have created). As the dataset was very large and with limited computational power, I computed the optimum number of k for centroid-based clustering using very small range of k. I used different k values to computed silhouette score and used knee method to find the optimum value of K. The best value of K that I found was three.