Capstone Project Submission

Instructions:

i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1. **Name: -** OM PRAKASH PRADHAN

   Email ID: - omprakashdoverr@gmail.com

   Contribution: -

   - Inspection of 'Zomato Restaurant names and Metadata' dataset
   - EDA of 'Zomato Restaurant names and Metadata' dataset
   - Feature selection
   - Feature engineering
   - Data preparation for clustering and sentiment analysis
   - Clustering algorithms and evaluation

2. **Name: -** RUCHIKA NAYAK

   Email ID: - nayakruchika1999@gmail.com

   Contribution: -

   - Inspection of 'Zomato Restaurant reviews' dataset
   - EDA of 'Zomato Restaurant reviews' dataset
   - Feature engineering
   - Feature selection
   - Data preparation for clustering and sentiment analysis
   - Classification algorithms for sentiment analysis

Please paste the GitHub Repo link.

GitHub Link:- https://github.com/Ruchika810/Clustering-and-Sentiment-Analysis

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your

conclusions. (200-400 words)

We are given with two datasets for the project of Zomato Restaurant and Sentiment analysis. The project is mainly focusing on analyzing the sentiments of the reviews given by the customer in the data and made some useful conclusion in the form of visualizations. And the visualizations is in the form of clustering.

First of all we have inspected the datasets i.e. in Zomato Restaurant names and Metadata dataset we have performed some changes like changed the datatype of Cost column, from object type to float type. Then removed the null values with help of null value treatment. Then we have performed the EDA on the dataset. With the help of horizontal bar plots we have shown top popular cuisines and popular collections.

Then we have inspected the other dataset i.e. Zomato restaurant and reviews. Here also we have changed the datatype of the rating column from object to integer/float. After that using null value treatment is performed. Then the ‘metadata’ column is separated into two separate column named ‘no. of reviews’ and ‘no. of followers’. Then EDA is performed on this dataset. With the help of horizontal box plots we have shown the top rated and worst rated restaurants, top reviewers, most reviewed and most followed restaurants.

After all these we have prepared data for clustering. For this we aggregated the ‘Rating’, ‘No of reviews’, ‘No of followers’ column to a single value for each restaurant by grouping them name wise. Then we have got some right skewed values so in order to normalize them we have performed square root transformation. Then we have got our final df with four columns of ‘Name’, ‘Mean Rating’, ‘Mean Followers’ and ‘Cost’.Then using different models like k-means, elbow method, hierarchical clustering we have got the optimal number of clusters.

After the clustering part we have done the sentiment analysis. For this we have merged the two dataset and taken some important columns like ‘name of restaurants’, ‘Reviews’, ‘Rating’ column. Then we have removed the stopwords and punctuations from the ‘review’ column. After that we have done stemming and vectorization. Then we have done train test split for classification algorithms. Then for sentiment analysis we have taken naïve bayes, logistic regression, xgboost and SVM.

From all these process we have concluded that in Clustering when two variables are taken the optimal number of clusters is 3 or 4. And when 3 number of variables were taken the optimal number of clusters is 4.

And in sentiment analysis we came to k now that logistic regression and SVM are the two most appropriate models with accuracy of nearly 87%.