Capstone Project Submission

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| |  |  |  | | --- | --- | --- | | **NAMES** | **E-MAIL** | **CONTRIBUTION** | | Sunil Kumar | 14bbt1019@gmail.com | Entire Project work. | |
| **Please paste the GitHub Repo link.** |
| **GitHub link:-** **https://github.com/Sunilkumar17-design/Unsupervised\_ML-Capstone\_Project-.git** |
| **Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)** |

**Summary of work**

**Started with data loading and importing the libraries and then started with exploring the data and did some of the visualization to see the common pattern in the data and looked into columns and rows. It was seen that there were few missing values present as per the ratio, imputed the values accordingly. I have seen that I have almost everything in the form of words which I converted accordingly. Found some of the string and removed it. Found 4 years were unique.  Explored the restaurant with respect to the rating received. Found some of the top restaurants for the ratings and year wise popularity and saw the ratings trend with respect to year and seems to be growing over the year. Checked the restaurant prices which were around 2800, 2500/- and being 1600/- top 10th number. Similarly explored the least costly restaurant around 350/- to 150 as top 10th.Also checked the overall cost. During the clustering of cuisines I have seen that North India, continental, chinese,seafood, European, fast food etc. were popular.**

**Also, I have explored and seen restaurants weekly open time. After all these exploration I have merged the dataframe of reviews. Found the north Indian restaurant being the highest number of 61 while top 10, 10th was bakery of 7 numbers. Finally started with natural language processing for removing the unwanted words or punctuations present in the data, tried removing emojis and some of the stopwords. I then divided the sentiments into 0 and 1 values where 0 being the negative comments or comments under 3.5 rating and 1 being good comments or comments above 3.5 rating for the restaurant.**

**Converted the words to lowercase, and removed the spaces and special characters etc.**

**Finally started with the Modeling made use of Bag of words,TF-IDF and Naïve bayes multinomial classification, Decision tree, Random Forest, K-mean clustering, and Logistic regression. However I have tried with different weights for the model at different times in iteration but only I have good findings for the Random forest classifier and for logistic regression using cross validation hyperparameter tuning have given better results with best parameters to use.**