World Journal of Engineering Research and Technology
WJERT
www.wjert.org
SJIF Impact Factor: 5.924

PREDICTING THE RANKING OF RESTAURANT BASED ON FACILITIES AND FEATURES IT PROVIDES USING VARIOUS CLASSIFICATION MACHINE LEARNING ALGORITHMS

***[1]M. Nirmala, [2]V. Saravanan and [3]T. Seeniselvi**

[1]Assistant Professor, Department of Computer Applications Hindusthan College of Engineering and Technology.

[2]Professor, Department of Information Technology Hindusthan College of Arts and Science.

[3]Associate Professor, Department of Computer Science, Hindusthan College of Arts and Science.

Article Received on 12/07/2021 **Article Revised on 02/08/2021** **Article Accepted on 22/08/2021**

***Corresponding Author**
M. Nirmala
Assistant Professor, Department of Computer Applications Hindusthan College of Engineering and Technology.

ABSTRACT

Any entrepreneur wishing to start up a business in the restaurant industry can getter better insights about the industry through this data set analysis. The restaurant ranking system using the various facilities and features provided by the restaurant is used to predict the ranking based upon customer preferences and penchants by various classification models under Machine learning. This system helps the restaurant owners to know the facilities, features and amenites offered by them in due of attracting their customers thereby enhancing the productivity of the business. Various factors affecting the launching of restaurant in a specific city and the demand of opening a restaurant is increasing day by day. With the increase in demand it is very difficult for the upcoming new restaurants to get copeup with the established one. An analysis of the given restaurant data might bring a better insight and with derived conclusions, the rising demands for the upcoming resturant can be easily predicted.

INDEXTERMS: Restaurant Ranking, Customer Preferences, Classification Models, rising demands, ML Algorithms.

1. INTRODUCTION

The main objective of this study is to analyse the different restaurants registered in Zomato app/Website, according to their type of cuisines, location and facility. Any entrepreneur wishing to start up a business in the restaurant industry can getter better insights about the industry through this data set analysis. It can create a good domain knowledge about Restaurant Industry. Preferences of the customer is always accountable in any industry is solemnly proved. Various restaurants located at different location of Bangalore city have been only analysed and the scope lies with in the Bangalore city. The Bangalorian taste and preferences has been accounted only for this survey. Based upon their taste preferences, amenities of the restaurants, type of restaurant, Table booking services, preferences of dish, cost incurred for 2 people, menu choices and location have been considered in this project to explore more about restaurant industry.

2. RELATED WORK

Shina, Sharma S. and Singha A.[2] have used Random forest and decision tree to classifying restaurants into several classes based on their service parameters. Their results say that the Decision Tree Classifier is more effective than Random Forest.

K. C. U. Perera and H.A. Caldera[3] have used data mining techniques like Opinion mining and Sentiment analysis to automate the analysis and extraction of opinions in restaurant reviews.

Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga and Akila Pemasiri wrote a paper on[4] 'Eatery, a multi-aspect restaurant rating system' that identifies rating values for different aspects of a restaurant by means of aspect-level sentiment analysis. This research introduced a new taxonomy to the restaurant domain that captures the hierarchicak relationships among entities and aspects.

Neha Joshi wrote a paper in 2012[5] on A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City which aims to contribute to the limited research in this area and provide insight into the consumer decision making process specifically for the India foodservice industry. She did hypothesis testing using chisquare test.

Bidisha Das Baksi, Harrsha P, Medha, Mohinishree Asthana, Dr. Anitha C[6] wrote a paper that studies various attributes of existing restaurants and analyses them to predict an appropriate location for higher success rate of the new restaurant. The study of existing

restaurants in a particular location and the growth rate of that location is important prior to selection of the optimal location. The aim is to the create a web application that determines the location suitable to establish a new restaurant unit, using machine learning and data mining techniques.

According to[7], people rate a restaurant not only based on food but also on dine-scape factors such as facility aesthetics, lighting, ambience, layout, table setting and servicing staff. With the abundance of data available on the above factors, analysis on it can be done using various mathematical models and data analysis methods like multiple Regression, Neural Networks, Bayesian network model, Random forest, SVM and many more can be used in predicting potential revenue of restaurant depending on various factors.

3. RESEARCH METHODOLOGY

This is a Descriptive Research problem where the study of a Zomato data set has been explored and performed the Ranking of Restaurant based on facilities and features it provides using various Classification Machine Learning Algorithms.

The data collected from secondary data sources is in the form of.csv file are tabulated in the **Table 3: 1**.

Table 3: 1 Data Source Details.

Data Sources	**Zomato.csv**
Data Charecteristics	Multivariate
Number of Instances	51717
Number of Attributes	17
Attribute Type	Categorical and Numerical
Link	https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

PROPOSED SYSTEM

The proposed system states the prediction of the Restaurant ranking based on features and preferences of customers such as Online ordering, Table Booking, Location preferences, Menu Item, Dish Liked, approx cost for 2 people, restaurant type, cuisines etc. The data set features and description have been depicted.

Table 3: 2 Zomato Features and Description.

Feature	**Description**	**Type**	**Sub Type**
url	This feature contains the url of the restaurant on the Zomato website.	Categorical	Nominal

address	This feature contains the address of the restaurant in Bangalore	Categorical	Nominal
name	This feature contains the name of the restaurant	Categorical	Nominal
online_order	whether online ordering is available in the restaurant or not	Categorical	Nominal [yes/No]
book_table	table book option available or not	Categorical	Nominal [yes/No]
rate	contains the overall rating of the restaurant out of 5	Categorical	Ordinal
votes	contains total number of upvotes for the restaurant	Numerical	Discrete
phone	contains the phone number of the restaurant	Categorical	Nominal
location	contains the neighborhood in which the restaurant is located	Categorical	Nominal
rest_type	restaurant type	Categorical	Nominal
dish_liked	dishes people liked in the restaurant	Categorical	Nominal
approx_cost(for two people)	contains the approximate cost of meal for two people	Categorical	Nominal
reviews_list	list of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer	Categorical	Nominal
menu_item	contains list of menus available in the restaurant	Categorical	Nominal
listed_in(type)	type of meal	Categorical	Nominal
listed_in(city)	contains the neighborhood in which the restaurant is located	Categorical	Nominal
Cuisines	food styles, separated by comma	Categorical	Nominal

Machine Learning Pipeline

Machine Learning[1] workflow has various steps to be followed starting from Problem definition to Model Prediction. Various steps required to be followed before fitting the model are shown in the pipeline **Figure 3: 1**.

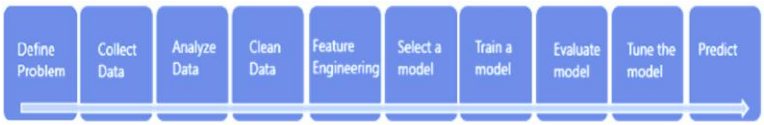
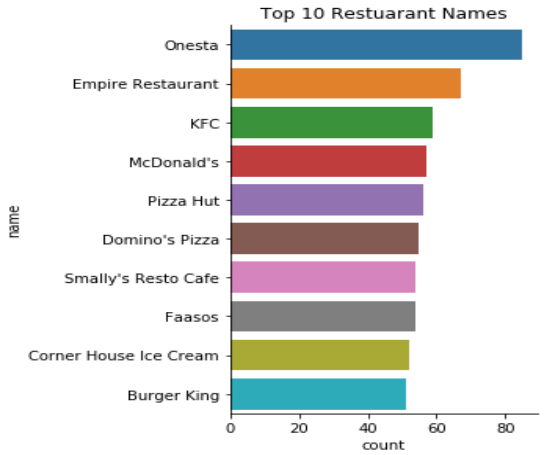
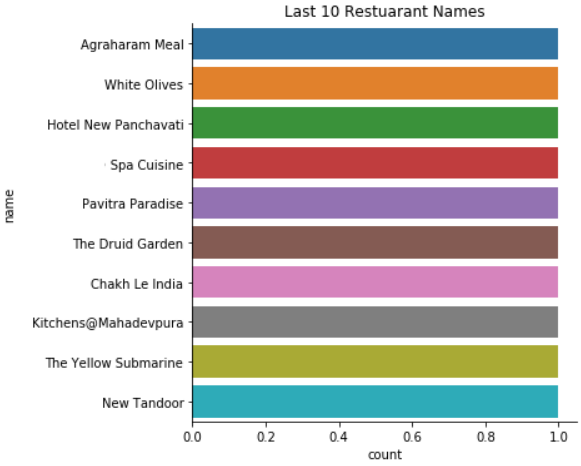


Figure 3: 1 Machine Learning Process Pipeline.

UNIVARIATE ANALYSIS – Individual Features / Variables.

The Univariate Analysis Data Visualization consists of single variable and it is a descriptive type of analysis and not infer its relationship with any other variables. In general count plot could be used for this analysis. It helps to potray the data and its respective patterns for the

user to get an bettter insight about the single variable and the graphical representation helps us to view maximum, minimum, mean values etc. The Univariate Analysis and its visualization inferences are described using charts.

Chart	Chart Values
 **Figure 3: 1 : Univariate Analysis – Top 10 Restaurant names**	Onesta 85 Empire Restaurant 67 KFC 59 McDonald's 57 Pizza Hut 56 Domino's Pizza 55 Smally's Resto Café 54 Faasos 54 Corner House Ice Cream 52
 **Figure 3. 2: Univariate Analysis – Last 10 Restaurant names**	Agraharam Meal 1 White Olives 1 Hotel New Panchavati 1 Spa Cuisine 1 Pavitra Paradise 1 The Druid Garden 1 Chakh Le India 1 Kitchens@Mahadevpura 1 The Yellow Submarine 1 New Tandoor 1
 **Figure 3. 3: Univariate Analysis –Restaurant Takes Online_order.**	Yes 16297 No 6749

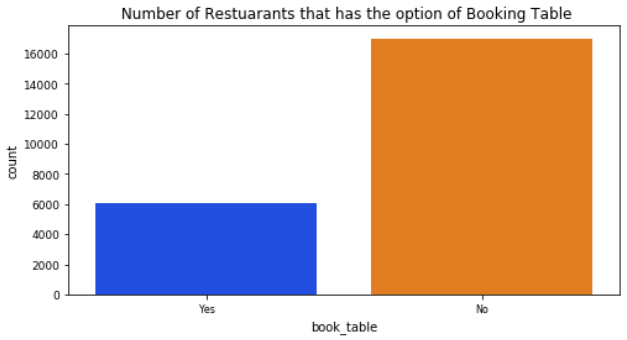


Figure 3. 4: Univariate Analysis –Restaurant Booking_table.

No	17005
Yes	6041

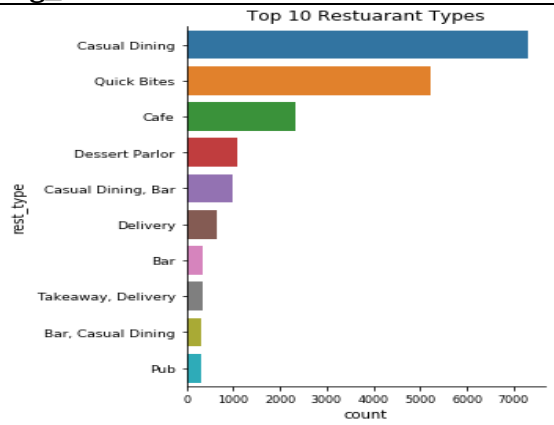


Figure 3. 5: Univariate Analysis –Top 10 Restaurant types

Casual Dining	5224
Quick Bites	2321
Café	1074
Dessert Parlor	980
Casual Dining, Bar	667
Delivery	357
Bar	341
Takeaway, Delivery	328
Bar, Casual Dining	325
Pub	325

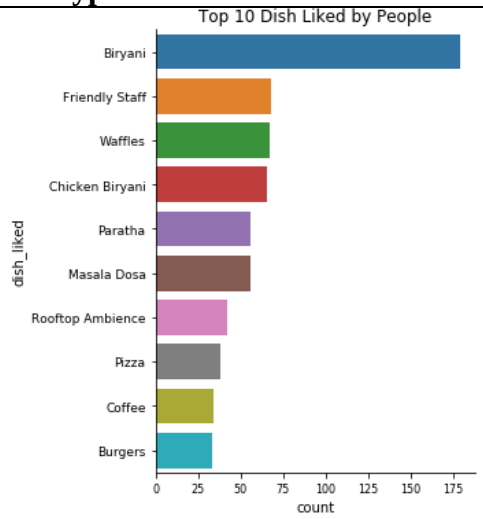


Figure 3. 6: Univariate Analysis – Top 10 Dish Liked by people.

Biryani	179
Friendly Staff	68
Waffles	67
Chicken Biryani	66
Paratha	56
Masala Dosa	56
Rooftop Ambience	42
Pizza	38
Coffee	34
Burgers	33

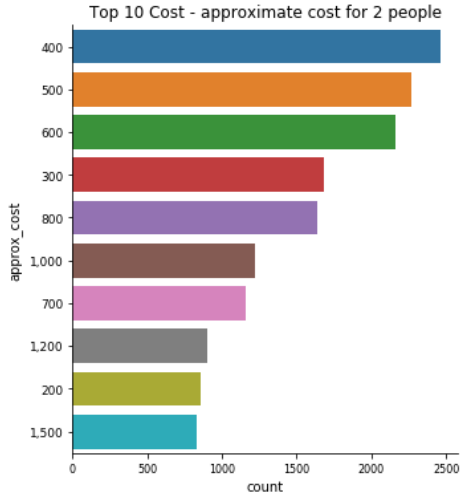


Figure 3. 7: Univariate Analysis – Top 10 Approximate cost for 2 people

400	2460
500	2269
600	2163
300	1682
800	1641
1,000	1218
700	1155
1,200	898
200	858
1,500	833

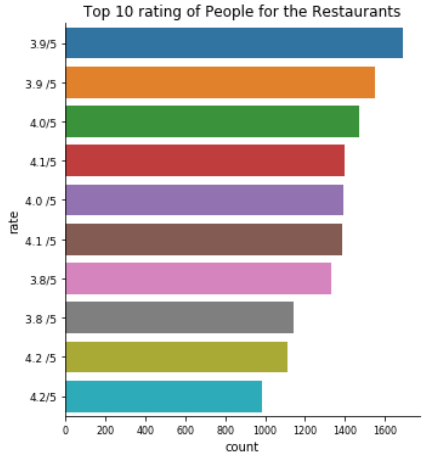


Figure 3. 8: Univariate Analysis – Top 10 Rating by people

3.9/5	1689
3.9 /5	1549
4.0/5	1469
4.1/5	1398
4.0 /5	1391
4.1 /5	1388
3.8/5	1329
3.8 /5	1144
4.2 /5	1115
4.2/5	982

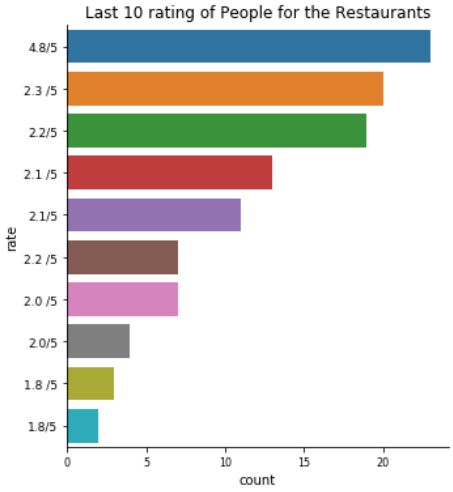


Figure 3. 9: Univariate Analysis – Last 10 Rating by people

4.8/5	23
2.3 /5	20
2.2/5	19
2.1 /5	13
2.1/5	11
2.2 /5	7
2.0 /5	7
2.0/5	4
1.8 /5	3
1.8/5	2

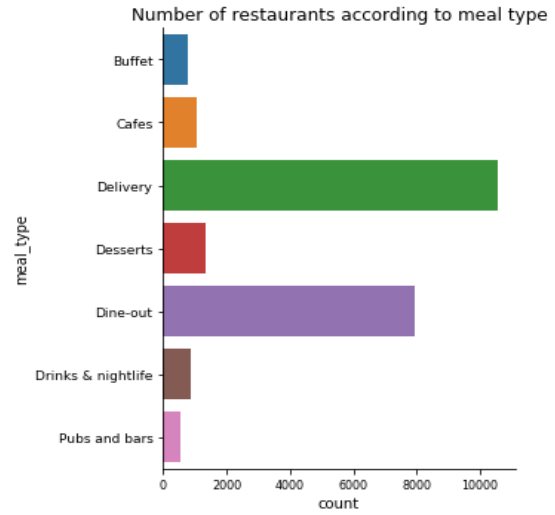


Figure 3. 10: Univariate Analysis - Restaurants count – meal_type.

Delivery	10575
Dine-out	7954
Desserts	1330
Cafes	1046
Drinks & nightlife	871
Buffet	753
Pubs and bars	517

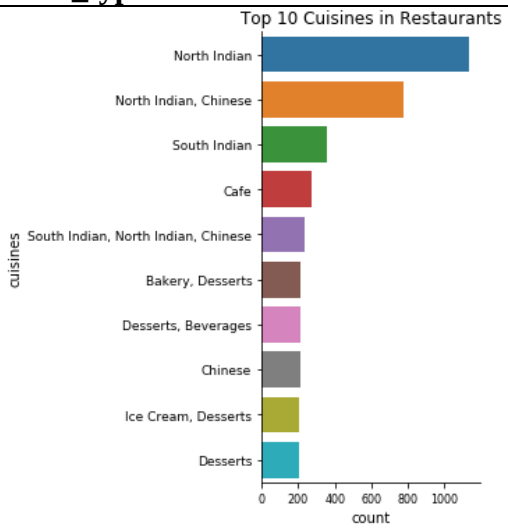


Figure 3. 11: Univariate Analysis –Top 10 Cuisines.

North Indian	1136
North Indian, Chinese	777
South Indian	359
Cafe	273
South Indian, North Indian, Chinese	233
Bakery, Desserts	215
Desserts, Beverages	214
Chinese	210
Ice Cream, Desserts	208
Desserts	204

Figure 3. 12: Univariate Analysis – Last 10 Cuisines.

Kerala, Biryani, Seafood, North Indian	1
Fast Food, North Indian, Continental	1
Desserts, North Indian, Chinese	1
Finger Food, Continental, Pizza, Asian, North Indian	1
Continental, Italian, Mediterranean, Salad	1
Cafe, Italian, Continental, European	1
Chinese, Vietnamese, Thai, Malaysian	1
North Indian, Mughlai, Lucknowi	1
Continental, Burger, Salad, Pizza, Beverages	1
South Indian, Asian	1

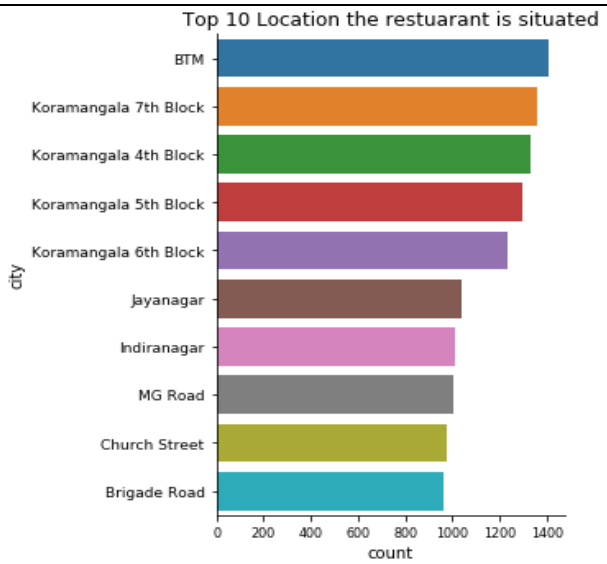


Figure 3. 13: Univariate Analysis – Top 10 Locations of Restaurant.

Koramangala 7th Block	1357
Koramangala 4th Block	1331
Koramangala 5th Block	1293
Koramangala 6th Block	1237
Jayanagar	1042
Indiranagar	1011
MG Road	1007
Church Street	979
Brigade Road	964

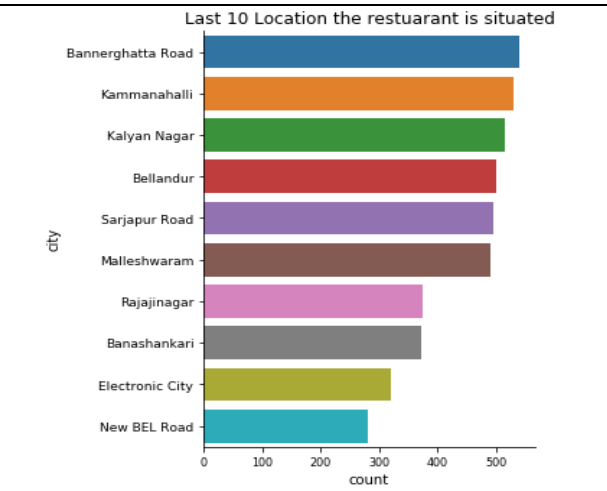


Figure 3. 14: Univariate Analysis – Last 10 Locations of Restaurant.

Bannerghatta Road	540
Kammanahalli	530
Kalyan Nagar	516
Bellandur	501
Sarjapur Road	496
Malleshwaram	492
Rajajinagar	374
Banashankari	372
Electronic City	320
New BEL Road	282

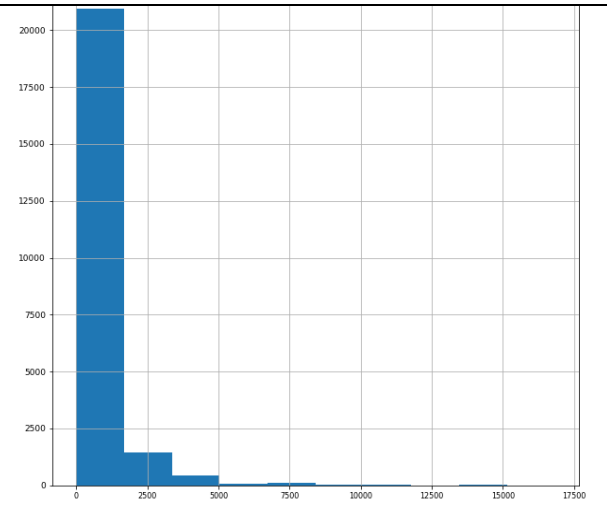


Figure 3. 15: Univariate Analysis – Voting for Restaurant.

Univariate Analysis –Report

Feature	Description
url	The URL column has been dropped
address	The Address column is not included for Modelling.
name	This top 10 restaurants names has been displayed with Onesta is the highest established restaurant at 85 places in bangalore and also the least count of 1 at a single location is bagged by many restaurant such as Whi;e olives, New Tandoor etc.
online_order	Various restaurant takes online ordering with a total count of 16297 and 6749 restaurants says a NO for Online_order
book_table	The book_table option has a low count for YES (6041) and for NO as 17005
rate	Almost people rate 3.9 on a total count of 3238 count.
votes	Number of people upvoted for the restaurant has been depicted by a Bar chart.
phone	The phone column has been dropped from Modelling
location	As the location and Listed_in(City) seems to be almost same and after analysis location has been dropped and Listed_in(City) is retained.
rest_type	Casual_dining has a count of 5224 and Quick Bites has a count of 2321 under the Rest_Type category
dish_liked	Biriyani is the most likely dish by many people.
approx_cost (for two people)	Top 10 approximate cost for 2 persons starts from Rs.400, then Rs.500, Then Rs.600 and so on
reviews_list	This column is dropped since it is almost similar to rate column.
menu_item	The Plotting is not done as it contains a lot of Textual Information.
listed_in(type)	Renamed as Meal_type and Delivery of Meal outside has the highest count of 10575 and Pubs and bars the least count of 517.
listed_in(city)	Top 10 Locationsof the restaurant situated in Bangalore. Koramangala has been listed for Top 4 Positions and New Bel road has the least count.
Cuisines	Top 10 cuisines with North Indian as the highest count of 1136 and the Least of 1 for south Indian Asian.

DATA PREPROCESSING

Removal of Null Values from the data set.

The total null values present in the data set are listed

Table 3: 3 Null Value Count.

url	0
address	0
name	0
online_order	0
book_table	0
rate	7775
votes	0
phone	1208
location	21
rest_type	227

dish_liked	28078
cuisines	45
approx_cost(for two people)	346
reviews_list	0
menu_item	0
listed_in(type)	0
listed_in(city)	0

There are totally 16 Categorical Column and 1 Numerical Column. In case of rate variable NAN columns are 7775, - is 69 and NEW is 2208. In the Rate column apart from NAN there are more inappropriate values such as "-"and "NEW". That should also be removed to make the rate column in a usable manner. Replace the NEW and – with NAN and then appropriately remove all the rows which contains a NAN in its column value.

Renaming Certain Columns for User Comfortability

The following columns are replaced by an other name for user comfortablity.

'approx_cost(for two people) renamed as ':'approx_cost'

'listed_in(city) renamed as 'city'

'listed_in(type)' renamed as 'meal_type'

After droppingcertian columns, renaming certain columns and fixing the NAN values, the total number of Rows are 23046 and the columns are reduced from 17 to 12.

Data Cleaning

In this certain noise data is removed and inconsistencies are fixed. In the target variable rate the data type is categorical and contains the value as 4.1/5 which includes / and spaces in between the values. The split command is used to split the /5 values from the rate column and the rate is converted to a float value.

The categorical column online_order and book_table are converted to Numerical values consisting of 1 for Yes and 0 for No.

The city column contains more values and so Label_encoder is used toconvert from categorical format to numerical representation.

The rest_type contains comma in between the textual format and the comma is replaced by space and Label Encoder is used for conversion to Numerical format.

The cuisines and meal_type columnhas been converted to Numerical values by Label Encoder.

The approx_cost contains comma in the numerical values and thus split replace function is used to replace the comma by space and then converted it to numerical quantity.

After convesrion from Categorical to Numerical and dropping the irelevant columns the total number of rows are 23046 and columns are 9.

DATA TRANSFORMATION

The data set is checked whether it is normalized or scale for the modelling operation. In this project the data is not scaled and thus needs scaling operation to be performed. By applying standard scaler methods the data is scaled.

The rate variable column which is the Target variable has to be converted into bins of 5 classes. So that the predicted value comes in the range of 0 to 4. The bins are created and grouping category is done.

MODEL TRAINING

The various proposed Model Building algorithms implemented in this project are

- Logistic Regression
- Random Forest
- K- Nearest Neighbout
- Support Vector Machine
- XG-Boost
- Decision Tree using Gini
- Decision Tree using Entropy

IV PERFORMANCE EVALUATION & EXPERIMENTAL RESULTS

Data Splitting

The data set is splitted in the ratio is 80:20 such that 80% is taken as training data and 20% as testing data. The random state is fixed in my project as 42.

IMPLEMENTATION REQUIREMENTS

JupyterLab is the next-generation user interface for Project Jupyter offering all the familiar building blocks of the classic Jupyter Notebook (notebook, terminal, text editor, file browser, rich outputs, etc.) in a flexible and powerful user interface. JupyterLab will eventually replace the classic Jupyter Notebook. Installation. JupyterLab can be installed using condo or pip. In

this system Anaconda3-2019.10-Windows-x86 64 has been installed with Jupyter Lab 1.1.4 which comes within it.

The following libraries are used for effective implementation.

Pandas is the most popular python library that is used for data analysis. It provides highly optimized performance with back-end source code is purely written in C or Python SciPy is an Open Source Python-based library, which is used in mathematics, scientific computing, Engineering, and technical computing.

Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shells, the Jupyter notebook, web application servers, and four graphical user interface toolkits.

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays.

Parameters For Model Fitting

Table 3: 4 Model Fitting Parameters.

Model Type	Parameters for Fitting the Model
Logistic Regression	solver='lbfgs',multi_class='auto', max_iter=2000
Random Forest	RandomForestClassifier(n_jobs=-1,random_state=0,criterion='gini', max_depth=7,)
KNN	KNeighborsClassifier(n_neighbors=7)
SVM	svm.SVC(kernel='rbf',gamma='auto')
XGBOOST	xgb.XGBClassifier(max_depth=10,learning_rate=0.1,n_jobs=0,n_estimatord=100, seed=10)
DECISION TREE – Gini	DecisionTreeClassifier(criterion="gini",random_state=100,max_depth=7, min_samples_leaf=5)
DECISION TREE - Entropy	DecisionTreeClassifier(criterion="entropy",random_state=100, max_depth=7, min _ samples_leaf=5)

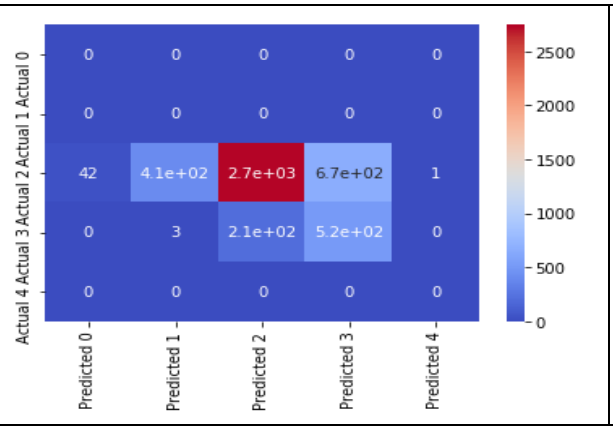
EXPERIMENTED RESULTS

Name of the Algorithm	Training Score	Testing Score
Logistic Regression	70.59014971	70.86767896
Random forest	77.24560642	76.13882863
SVM	74.06161857	73.79609544
KNN	80.28856585	73.62255965
DECISIONTREE (gini)	76.49164678	76.16052061
DECISION TREE (ENTROPY)	76.05771317	75.72668113
Xgboost	96.3169885	92.27765727

Confusion Matrix and Classification Report

Table 3: 5 Experimented Results.

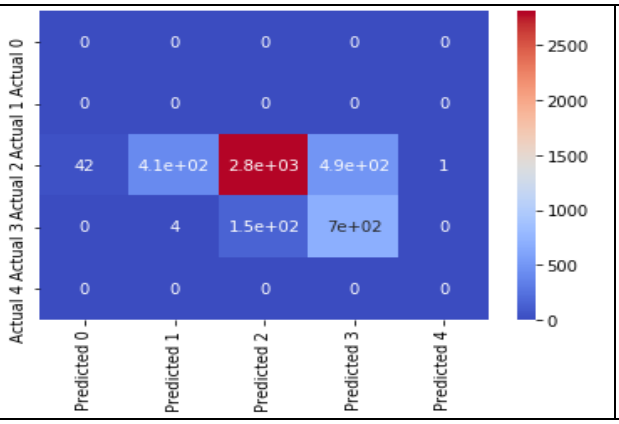
LOGISTIC REGRESSION



Classification Report Logistic Regression :

	precision	recall	f1-score	support
0	0.00	0.00	0.00	42
1	0.00	0.00	0.00	417
2	0.71	0.93	0.80	2957
3	0.71	0.44	0.54	1193
4	0.00	0.00	0.00	1
accuracy			0.71	4610
macro avg	0.28	0.27	0.27	4610
weighted avg	0.64	0.71	0.66	4610

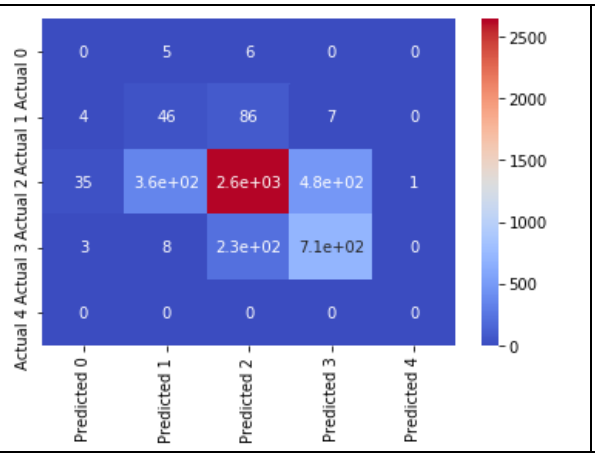
RANDOM FOREST



Classification Report Random Forest :

	precision	recall	f1-score	support
0	0.00	0.00	0.00	42
1	0.00	0.00	0.00	417
2	0.75	0.95	0.84	2957
3	0.82	0.59	0.68	1193
4	0.00	0.00	0.00	1
accuracy			0.76	4610
macro avg	0.31	0.31	0.30	4610
weighted avg	0.69	0.76	0.71	4610

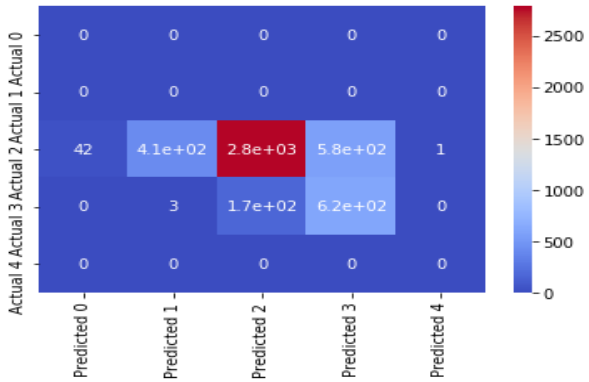
K-NEAREST NEIGHBOUR



**Classification Report KNN Algorithm :

	precision	recall	f1-score	support
0	0.00	0.00	0.00	42
1	0.32	0.11	0.16	417
2	0.75	0.89	0.82	2957
3	0.75	0.59	0.66	1193
4	0.00	0.00	0.00	1
accuracy			0.74	4610
macro avg	0.36	0.32	0.33	4610
weighted avg	0.71	0.74	0.71	4610

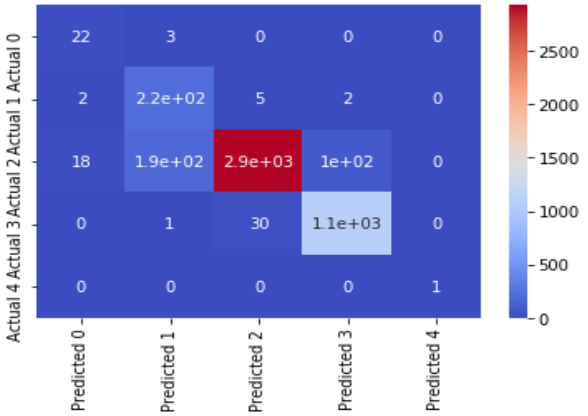
SUPPORT VECTOR MACHINES



**Classification Report SVM Algorithm :

	precision	recall	f1-score	support
0	0.00	0.00	0.00	42
1	0.00	0.00	0.00	417
2	0.73	0.94	0.82	2957
3	0.78	0.52	0.62	1193
4	0.00	0.00	0.00	1
accuracy			0.74	4610
macro avg	0.30	0.29	0.29	4610
weighted avg	0.67	0.74	0.69	4610

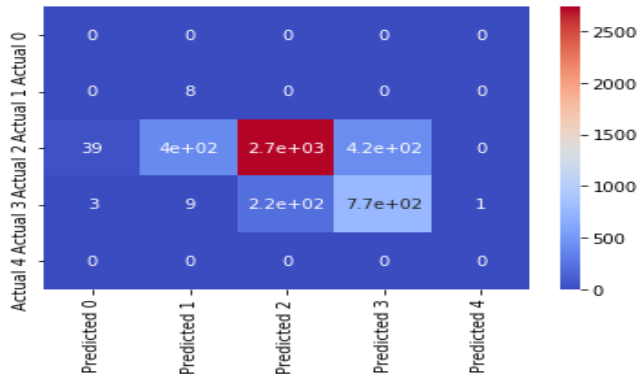
XGBOOST



**Classification Report XG BOOST Algorithm :

	precision	recall	f1-score	support
0	0.88	0.52	0.66	42
1	0.96	0.53	0.68	417
2	0.90	0.99	0.94	2957
3	0.97	0.91	0.94	1193
4	1.00	1.00	1.00	1
accuracy			0.92	4610
macro avg	0.94	0.79	0.84	4610
weighted avg	0.93	0.92	0.92	4610

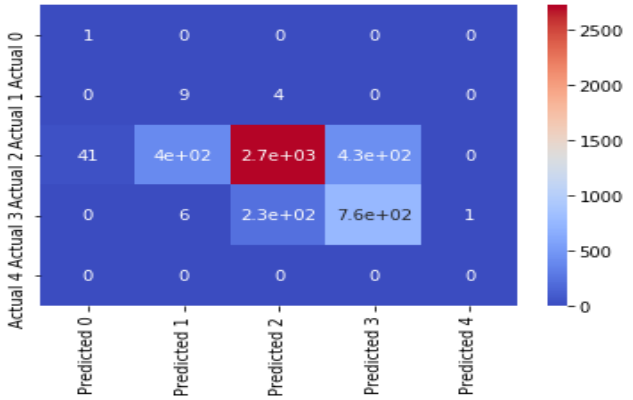
DECISION TREE USING GINI



**Classification Report DECISION TREE - GINI Algorithm

	precision	recall	f1-score	support
0	0.00	0.00	0.00	42
1	1.00	0.02	0.04	417
2	0.76	0.92	0.83	2957
3	0.76	0.65	0.70	1193
4	0.00	0.00	0.00	1
accuracy			0.76	4610
macro avg	0.50	0.32	0.31	4610
weighted avg	0.78	0.76	0.72	4610

DECISION TREE USING ENTROPY



**Classification Report DECISION TREE - ENTROPY Algorithm

	precision	recall	f1-score	support
0	1.00	0.02	0.05	42
1	0.69	0.02	0.04	417
2	0.76	0.92	0.83	2957
3	0.76	0.64	0.69	1193
4	0.00	0.00	0.00	1
accuracy			0.76	4610
macro avg	0.64	0.32	0.32	4610
weighted avg	0.75	0.76	0.72	4610

CONCLUSION & FUTURE SCOPE

The model deployment has been done for all the algorithms and the sample input has been given for evaluation. The sample input for evaluation is [1,1,775,29,947,800,0,1] and the model has predicted it as class.[3] in all the algorithms.

Code Sample

pred_new=SVM.predict ([[1,1,775,29,947,800,0,1]]) array([3])

The present study predicted the ranking of restaurant based on certain specific features and the project need to be expanded for unknown features and also study has to be carried out. The accuracy score could be increased by applying Hyperparameter optimization, Ensemble methods, Cross validation.

REFERENCES

1. Smola, A. S. (2008). Introduction to Machine Learning. *Cambridge University Press*.
2. Shina, Sharma, S. & Singha ,A. (2018). A study of tree based machine learning Machine Learning Techniques for Restaurant review. 2018 4th International Conference on Computing Communication and Automation (ICCCA DOI:/10.1109/CCAA.2018.8777649.
3. I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546. doi: 10.1109/CIAPP.2017.8167276.
4. Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga and Akila Pemasiri. Eatery – A Multi-Aspect Restaurant Rating System. Conference: the 28th ACM Conference.

5. Neha Joshi. A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City. Global Journal of Management and Business Research (2012).
6. Bidisha Das Baksi, Harrsha P, Medha, Mohinishree Asthana, Dr. Anitha C.(2018) Restaurant Market Analysis. International Research Journal of Engineering and Technology (IRJET).
7. Saravanan Mahalingam., Bhawana Jain., Mridula Sahay.: Role of Physical Environment (Dinescape Factors) Influencing Customers' Revisiting Intention to Restaurants. International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016.