iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.37257

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Restaurant Success Rate Prediction

Apoorva Shete[1], Nikket Chandwani[2], Darakshan Khan[3]
[1]*Department of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College,* Mumbai, India
[2, 3]*Department of Computer Science Engineering, Thadomal Shahani Engineering College,* Mumbai, India

Abstract: Online reviews are becoming more and more vital when it comes to decision making of the consumer. Every business wants to succeed in the present and the upcoming future, thus good online reviews help the business grow. Especially in a restaurant business, its rating holds an important indicator to whether the restaurant is any good or not. It not only helps us understand the quality and services but also attracts more customers. Ordinarily a new restaurant owner would have to invest time in building a menu, setting the correct price range and making something unique to scope out their competition as well as improve their ratings on online services such as Zomato. Thus, this paper attempts to ease the efforts of new restaurant owners by predicting the success rate of a restaurant on the Zomato app.
Keywords: Restaurant Ratings, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree and KNN

I. INTRODUCTION

The restaurant and hotel business has been one of the most coveted businesses of all time with a profit margin of almost upto 300%. This makes it one of the most largest and influential industries and a people's favorite. With time, more and more restaurants are opening up in almost every nook and corner and many of them are at a risk of failure due to the high numbers of factors and variations that are involved in the restaurant business. The investment being huge in the business makes it more restraining to the restaurant owners to explore and expand their business and the risk of failure is extremely high and is a hindrance to the new ventures in this industry. With in depth analysis and calculated prediction of these risks and factors all these issues can be curbed and the success of a restaurant can be ensured. This will help the business and the restaurant owners to alter themselves as per the needs and likes of the customers.

Opening a new restaurant, setting it up and deciding the theme, cuisines and infrastructure, involves a significant amount of investment. Knowing in advance how a restaurant will be appreciated by the population can go a long way in the success of a restaurant. In this world and day of the internet era, one quickly runs to online platforms for reviews, rating and feedback of a particular place or restaurant before physically going there or ordering from there, one such platform is Zomato. It is currently one of the most popular platforms for finding restaurants in India. Their site provides the menu, the prices, the user reviews, the location, photos uploaded by customers, different branches, etc. Also, the ratings on Zomato help determine whether the restaurant is any good.

The ability to identify the factors that affect the ratings can help the restaurant owners to devise sensible strategies to improve the business. A higher rating on Zomato indicates that the restaurant has a better service, food quality and service compared to its competitors. With this in mind, the success also depends on a targeted audience where multiple other points come into play, such as the price, cuisine, location, ambience, etc. This research paper essentially provides a model that will help fledgling and upcoming restaurants in determining their theme, menus, cuisine, prices and several other influential factors. It will provide the restaurants with the analysis of key factors where they would need to make enhancements and would also curate themselves as per the target audience.

This paper therefore researches and compares several important ML algorithms namely, Logistic Regression, Naïve Bayes, Random Forest, Decision Tree and KNN. A comparative study of the results achieved using the various algorithms is also provided in this paper. The objective of this research is to provide efficient data to the owners and investors of new restaurants which will help them achieve higher standards in their business and gain customer satisfaction. This research will help people who want to dig deeper into this field and find more data and analysis.

The outline of this paper is as follows, significant research pertinent to this field has been discussed in Section 2. The algorithms used to implement this model and their important features have been discussed in Section 3. Section 4 gives a brief description of the dataset used for the training and testing of this model. Section 5 gives in depth information about the implementation techniques of this model. Section 6 and 7 give the experimental results and important conclusions about this model along with the future scope of this project.

II. LITERATURE REVIEW

This section reviews the previously implemented research and thesis in this field.

The paper [1] uses the Yelp dataset for conducting the research and implementing their model and a comparative study of results obtained using various ML algorithms is done. Here, they have essentially divided their model into two phases, the 'Success Predictor' and 'Business Insights'. They have used 90% of their data for training and 10% for model testing. In [2] too, the Yelp dataset has been used. This paper further implements multinomial classification as per prediction of the rate and binary classification considering the popularity change prediction. This model achieves a prediction accuracy of about 32% and multinomial accuracy of around 70%. They got the best accuracy from the Logistic Regression algorithm. In paper [3], different features about a restaurant in a specific area are studied and analyzed to predict suitable locations for success of upcoming restaurants. The location for a new restaurant is predicted with 99.8% accuracy using the Random Forest algorithm. This data helps owners decide the appropriate and best location for their restaurants before investing money. Paper [4] discusses and studies the restaurant related preferences of customers in the city of Dehradun. This contributes towards research related to restaurants limited to a particular city. This gives the user information about the significant factors that impact the restaurants in the city of Dehradun. Paper [5] too uses the Yelp data set to perform sentiment analysis using the reviews and ratings of customers both combined. Matrix factorization is used for test information analysis and prediction of the customer ratings on Yelp. They have concluded that the Naïve Bayes along with feature extraction with the removal and stemming of stop-words gave the best results in their sentiment analysis project. In paper [6], the Zomato dataset is used for predicting the success of a restaurant. This paper too uses and compares various Ml algorithms to predict the success of a restaurant limited to a specific area and a specific number of features. They have obtained the best accuracy using the ADABoost algorithm, 83%.

This literature research and the several previously implemented models helped us understand significant factors related to the study and development of this field. It helped us achieve a head start in analysing the data of our model and the important features to be considered for the implementation of this model.

III. ALGORITHMS

This section gives a brief description about the algorithms used to implement this model.

A. Logistic Regression (LR)

Logistic Regression is an important ML algorithm that is used for classification and prediction analysis. It is mostly utilized for the binary classifications. Logistic Regression or LR uses the sigmoid function on the data for predicted the result and for classification. The equation for sigmoid function is as shown,

$$f(X) = \frac{1}{1+e-x} \qquad (1)$$

This algorithm uses the concept of probability of the successful prediction of a classification issue. It usually works well with smaller datasets.

B. Naïve Bayes

The Naïve Bayes classifier is a combination of many classification algorithms that use the Bayes' Theorem. All of these classifier algorithms that make the NB algorithm share a general basis and every combination of two features classified are not codependent.

The most basic assumption of the NB algorithm is that every feature contributes independently and equally to the outcome. The Bayes' theorem gives us the probability of an event given that another event has already occurred. This is mathematically stated as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

In equation(2), A and B are independent events. The various NB classifiers differ mostly as per the assumptions made related to the distribution of probability.

C. Random Forest

This classifier basically has many decision trees that work together. Every independent tree in the RF classifier splits the prediction of a particular class, and the one having the most votes becomes the result of the model. The main concept and the key point used by this algorithm is the correlation between the models. This works extremely well because all the trees are protected from their independent errors.

The bagging feature and the randomness that builds each independent tree makes an uncorrelated forest that gives the most accurate predictions as a group than any of the individual trees. This is an extremely efficient and essential classification algorithm used to solve classification problems.

D. Decision Tree

Decision Tree is a Supervised learning algorithm, used for both classification and regression problems. It is a structured classifier that has nodes representing the features of a dataset. The branches factor the decision rules and the outcome is given by every leaf node. In this algorithm, the class of the data is predicted by comparing the values of the root nodes with that of the real dataset values. Based on these comparisons, the branch then goes to the next node. Again the comparison is made with other sub-nodes and so on. This process is continued till the leaf node of the is reached. As compared to the other algorithms, less data cleaning is required.

E. K-Nearest Neighbour (KNN)

The K-Nearest Neighbour or KNN algorithm presumes analogy among the new data and the existing data cases. It then places the new data in the category that is the most similar to all the existing categories. It does not learn directly from the dataset but stores the data and during classification it performs on the data. This algorithm is very easy to implement and is generally used for the training for larger datasets.

IV. DATASET DESCRIPTION

The initial step in this project is data collection and processing. The dataset used in this paper is the Zomato Restaurants' Dataset. It has been taken from kaggle.com[7] and gives information about restaurants in the city of Bengaluru. This dataset has different factors which can all collectively affect the rating of a restaurant on Zomato. In this dataset, there are 51,717 entries in this dataset and 17 different attributes. Table 1 shown below, gives the information about each attribute, in the order in which they are specified in the dataset.

Table I. Attributes of the Zomato Dataset

ATTRIBUTES	DESCRIPTION	DATA TYPE
URL	Gives the URL of the specific restaurant on the Zomato app	String
Address	Complete address of the restaurant is given here	String
Name	Gives the name of the restaurant	String
Online_order	Tells us if the restaurant accepts online orders	String
Book_table	Tells us whether table booking is available or not in the restaurant	String
Rate	The rating of the restaurant out of 5 on	Numeric

	the Zomato app is given	
Vote	Voting given by users for the restaurant on the Zomato app	Numeric
Phone	Gives the phone number of the restaurant	Numeric
Location	Specifies the area in which the restaurant is located	String
Restaurant Type	The type of restaurant each specified one falls into on the Zomato app	String
Dish_liked	Most liked dishes of the restaurant	String
Cuisine	Type of cuisine served in the restaurant	String
Approx_cost	Gives the approximate cost for two people	Numeric
Reviews_list	Reviews for the restaurant as posted on Zomato app	Tuple
Menu_item	Contains a list of menus as provided in the restaurant	List
Listed_in(type)	Specifies the types of meals served in the restaurant	String
Listed_in(city)	Gives the city under which the restaurant is listed on the Zomato	String

From the attributes mentioned above, only few of the most important features were selected to train and test the model. The data as given in the dataset is not ready to be used directly for model training and testing. It has to be cleaned and thus data preprocessing is essential.

After data preprocessing and cleaning, from the dataset, 80% is training data and 20% is testing data that has been used for evaluating the efficiency of this model.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

V. METHODOLOGY

This section gives a detailed explanation of the methodology followed to implement this model.

The initial step was collection of data that is appropriate for analysis. A detailed description of the dataset used in this project has been given in the previous section. This dataset required cleaning and preprocessing. Many of the entries in this dataset had null values, so such entries were dropped. After dropping all the duplicate rows, filling the columns of null values with appropriate data and converting the data type of a few features to a more suitable data type, the data was ready to be used for further analysis and implementation. In order to understand the data better, graphical exploration of the data was done and a few important graphs were plotted.

The Fig. 1 shown below shows us the different kinds of restaurants included in the dataset. It also shows us the frequency of each type of restaurant. From the plot it is clear that quick bites followed by casual dining were the two most important and frequent types of restaurants from all the ones listed. Thus in the dataset all the columns containing casual dining and quick bites along with other types were simply replaced by 'casual dining + quick bites' whereas all the other types of restaurants were simply replaced by 'other' in the rest_type column. In this dataset, almost 60% of the restaurants were under casual dining and quick bites.

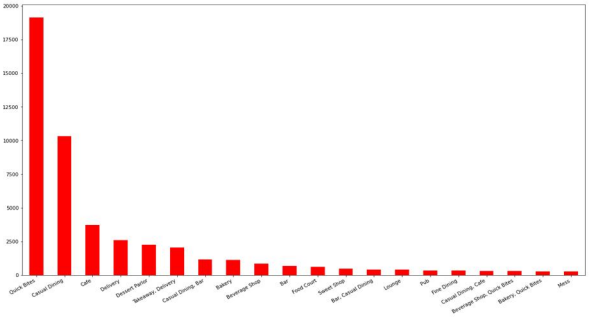


Fig. 1. Types of restaurants in the dataset and their frequency

It is of great importance to understand the data properly. For this, the following exploratory analysis was done on the data.

The top 5 most and least popular restaurants based on the votes on the app were plotted. This is shown by Fig. 2 and Fig. 3 below.

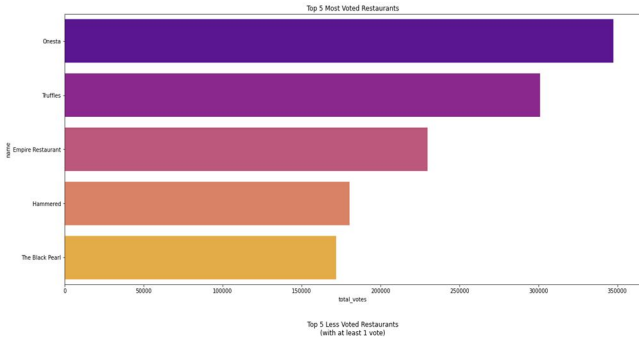


Fig. 2. Top 5 restaurants as per votes on the Zomato app

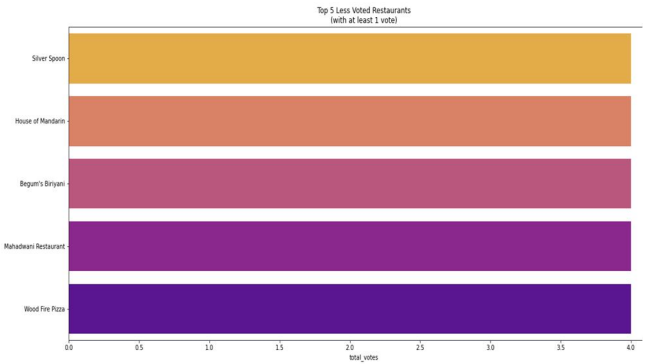


Fig. 3. Least 5 popular restaurants as per votes on the app

After this, five most and least expensive restaurants as per the approximate cost predicted by the app were plotted.

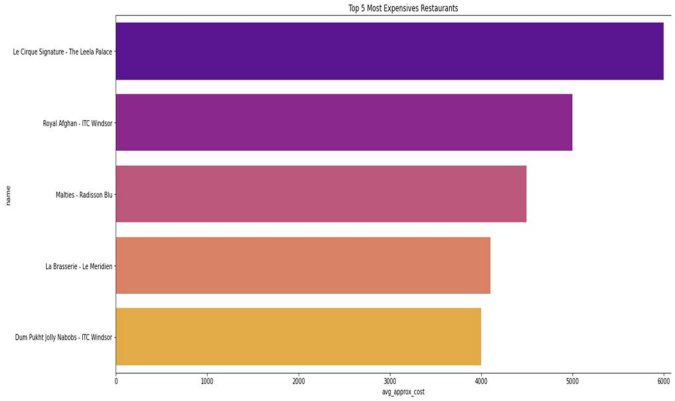


Fig. 4. Five of the most expensive restaurants in the dataset

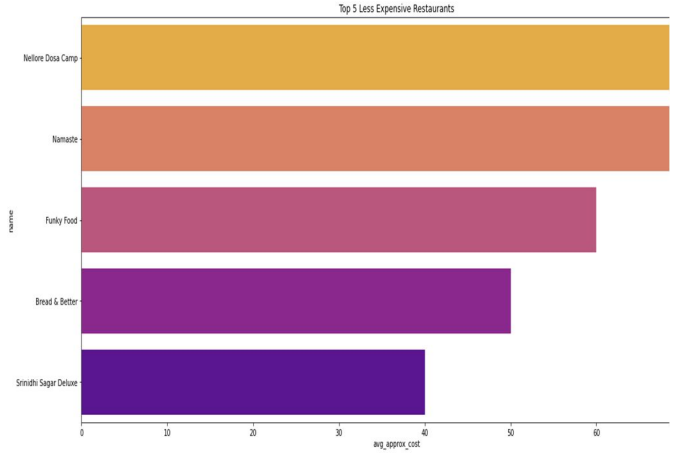


Fig. 5. Five of the least expensive restaurants in the dataset

The average approximate cost for two people for the restaurants considered in the dataset is Rs. 496.62.
Of all the restaurants in this dataset 12.5% offer an online table booking option, whereas for 87.5% of the restaurants this option is not available or is not provided. 58.9% of the restaurants from the dataset accept an online order and the remaining- 41.1% do not. This is shown in the pie chart in Fig. 6 and 7.

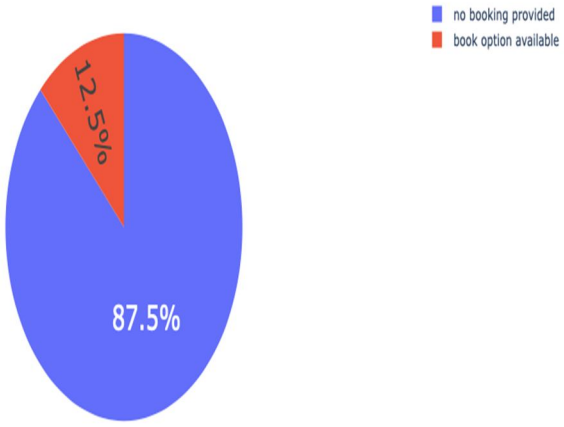


Fig.6. Pie chart to show distribution of restaurant depending on the availability of table booking option

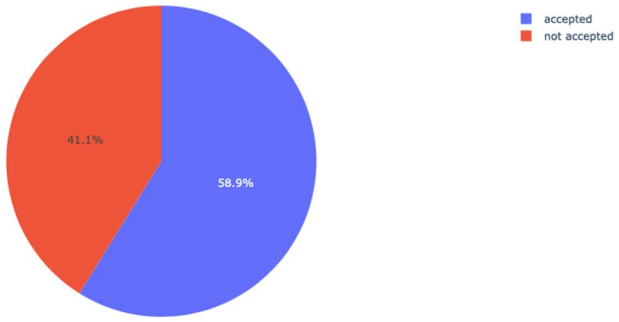


Fig. 7. Distribution of restaurants in the dataset depending on the availability of online ordering service

Further a wordcloud of the most preferred dishes by the customer was plotted with a view of understanding the most liked dishes by the people. This will help the new restaurants majorly in deciding the menu of their restaurant and the type of cuisine to be included in the menu. The Fig. 8 shows this wordcloud. From this, the dishes in major demand were easily recognized.

Fig. 8. Wordcloud of the most popular food dishes in the dataset

The reviews list was extracted and the wordcloud of the reviews was plotted as well. This will help the new restaurants in understanding what kind of factors are considered and taken into account by the people while writing a review for a particular restaurant. This wordcloud has been shown below in Fig. 9.

Fig. 9. Wordcloud of the reviews given for restaurants by the people on the Zomato app

After this step, the main code, i.e. the success predicted of a restaurant was implemented. For this it is essential to first decide a threshold value for the rating of restaurants. This threshold value is thus the deciding factor for the prediction of the success or failure of restaurants. The threshold value selected for the purpose of this paper was 3.75. Thus all the restaurants that had a rating value greater than or equal to 3.75 were given a value 1 indicating that the restaurant is a ‘success’ and all the restaurants having a rating value below 3.57 were marked as 0 indicating the ‘failure’ of the restaurant. The data was distributed as follows after giving binary values to each restaurant according to the threshold value:

1) 21,421 restaurants were marked as 0 or failure
2) 20, 244 restaurants were marked as 1 or success

In terms of percent values, 51.4% of the restaurants from the dataset were marked as failures and 48.6% of the entries were marked as a success.

The next step was feature extraction. This dataset contains 17 different attributes, some of which are not very important for the purpose of this research. Therefore, only the essential features were selected and extracted from the dataset. The selected features were: name of the restaurant, its votes and rating on the Zomato app, URL for that restaurant on the app and the approximate cost for two people as predicted by the Zomato app, online order availability, table booking option, location of the restaurant, type of the restaurant, and the city in which the restaurant is listed. This final dataset was now fed to the model. The data was split into 80-20 percent. Here, 80% of the data was used for training the model and 20% of the data was used for model testing. That is, from the dataset of above 51k entries, 33016 were used for model training and the remaining entries were used for testing of this model. Different algorithms were used for testing the accuracy of the model. The classifier algorithms chosen for this purpose were Logistic Regression (LR), Naïve Bayes (NB), Random Forest, Decision Tree and KNN. A detailed workflow diagram of the methodology followed for the implementation of this model has been given below.

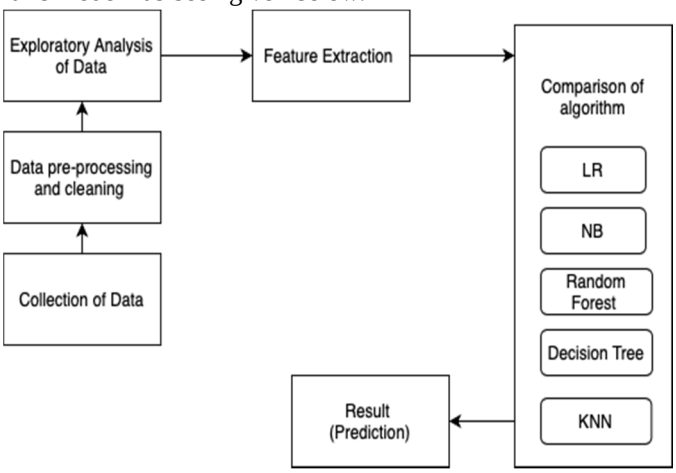


Fig. 10. Flowchart of the methodology followed for this project

VI. RESULTS

The accuracy achieved by each classifier algorithm has been written in a tabular form in Table 2 given below.

Table II. Accuracy achieved with each algorithm on the test data

ALGORITHM USED	ACCURACY ACHIEVED
Logistic Regression	72%
Naïve Bayes	67.6%
Random Forest	78.8%
Decision Tree	82.9%
KNN	80%

A comparative graph of the accuracies achieved from each algorithm on the test data has been shown below in Fig. 11 to make it easier for the reader to understand the comparative results.

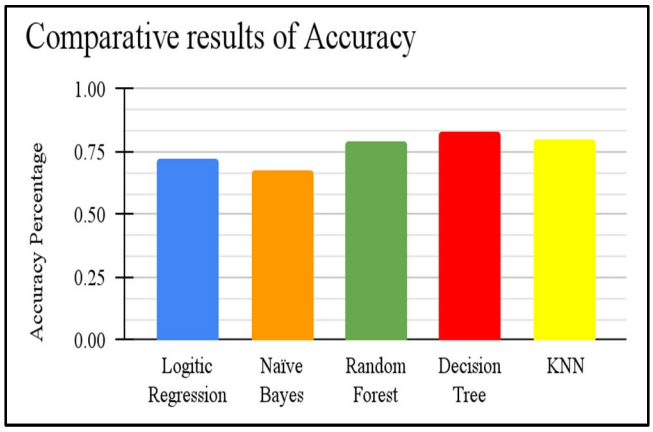


Fig. 11. Comparative results of accuracy obtained by using various classifiers

From the above comparisons it is clear that the best accuracy for this model is attained by using the Decision Tree classifier, followed by KNN. The accuracy obtained for the Decision Tree classifier was **82.9%,** the highest among all the classifiers. Whereas, the lowest accuracy was given by the Naive Bayes classifier, was the least **67.6%**.

The confusion matrices specific to each algorithm have been given below.

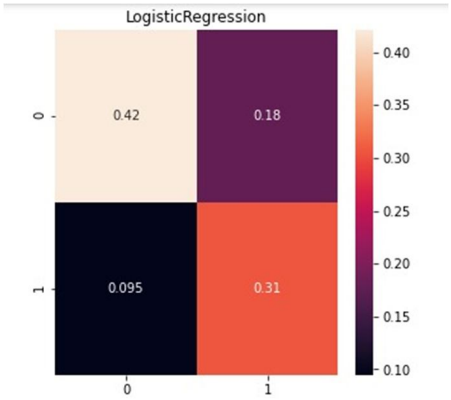


Fig.12. Confusion matrix for Logistic Regression

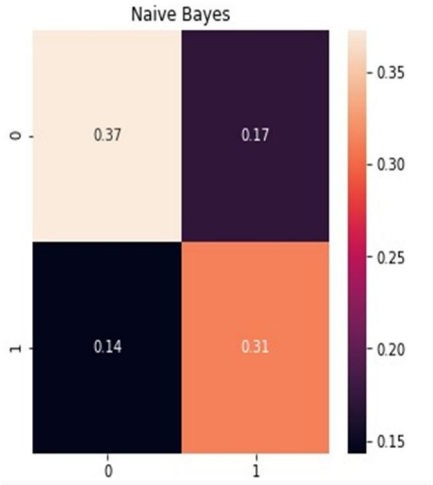


Fig.13. Confusion matrix for Naïve Bayes

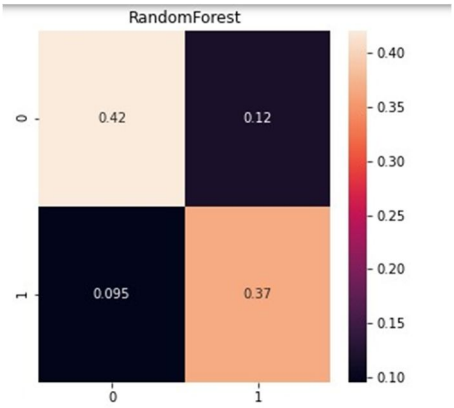


Fig.14. Confusion Matrix for Random Forest

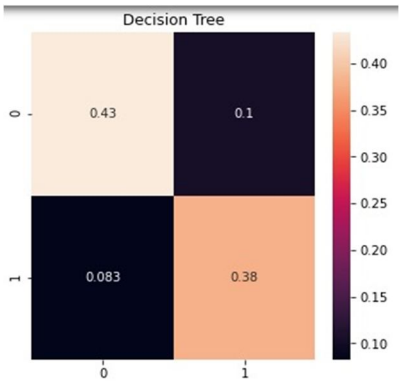


Fig.15. Confusion Matrix for Decision tree

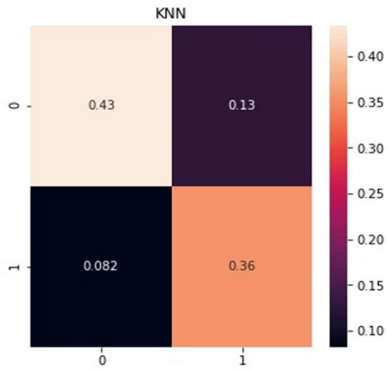


Fig. 16. Confusion Matrix for KNN

VII. CONCLUSION AND FUTURE SCOPE

The prediction of the success of a restaurant is an important field of research. If the rating a restaurant can achieve on an influential app like Zomato can be predicted before-hand, it will prove to be extremely beneficial for the owners of new and upcoming restaurants in town. This paper thus thoroughly studies the various features that are impactful in determining the success of a restaurant. As compared to the previous papers and research in this field, our model provides a higher accuracy than the previous models and also gives a comparative study of results using various algorithms. The exploratory analysis carried out on the data in this paper can help stakeholders and owners investing their money in opening new restaurants to cleverly plan and execute various aspects of their new venture. They can plan their menu, theme of the restaurant, infrastructure and various other features as per what is currently popular among the people.

For the implementation of our model we have used the Zomato dataset that is specific to the region of Bengaluru. In the future, we can collect more data from the Zomato app to regions specific to the location of the new restaurant or more data about all the restaurants from across the country. Larger dataset means more data for the model to better train itself and improved results and accuracy on the test data.

Once the owners have made decisions about their new restaurant, they can further update the data from time to time and carry out the same exploratory analysis to get to know what is popular and in demand then. More geographical analysis on the data can also help in understanding which locations are better suited for new restaurants. Sentiment Analysis can also be carried out on the restaurants reviews data and all these factors can be combined together to give the best insights to the user and owners of the new restaurants.

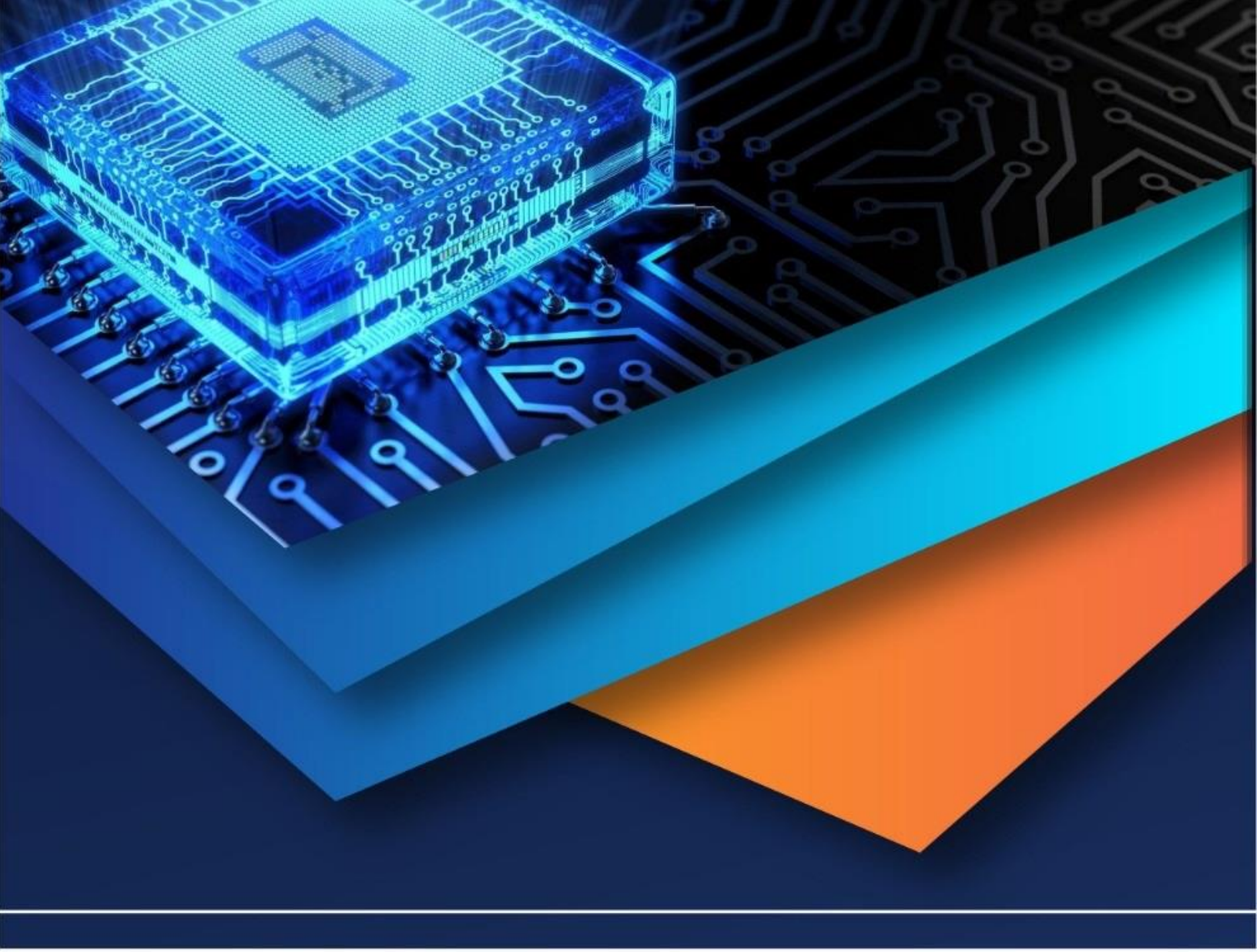
Thus there are several ways and methods using which this model can be further improved to give the best accuracy, results and insights to the user.

VIII. ACKNOWLEDGMENT

We would like to express our gratitude to Prof. Darakshan Khan for her immense support and guidance. We would also like to express our gratitude to our Head of Department, Dr. Tanuja Sarode, and our Principal, Dr. G.T. Thampi.

REFERENCES

[1] https://rafaelsilva.com/files/teaching/inf-553-fall-2018/017-predicting-success-upcoming.pdf
[2] http://cs229.stanford.edu/proj2017/final-reports/5244334.pdf
[3] https://www.irjet.net/archives/V5/i5/IRJET-V5I5489.pdf
[4] https://globaljournals.org/GJMBR_Volume12/5-A-Study-on-Customer-Preference.pdf
[5] Xu, Yun, Xinhui Wu, and Qinxia Wang. Sentiment Analysis of Yelps Ratings Based on Text Reviews. Stanford University. 2015.
[6] https://ijcat.com/archieve/volume8/issue9/ijcatr08091008.pdf
[7] A.Pant. “Introduction to Logistic Regression.” towardsdatascience.com. https://towardsdatascience.com/introduction-to-logistic-regression- 66248243c148
[8] “Naïve Bayes”, https://www.geeksforgeeks.org/naive-bayes-classifiers/
[9] “Random Forest Classifier”,https ://towardsdatascience.com/understanding-random-forest-58381e0602d2
[10] “Decision Tree”, https://www.javatpoint.com /machine-learning-decision-tree-classification-algorithm
[11] “KNN”, https://www.javatpoint.com/ k-nearest-neighbor-algorithm-for-machine-learning
[12] “Dataset”, https:// www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY