2023 World Conference on Communication & Computing (WCONF)
Raipur, India. July 14-16, 2023

Predict Success of a Zomato Restaurant using Machine Learning

Sandeep Bhatia
School of Computing Science and Engineering
Galgotias University Greater Noida,
Uttar Pradesh, INDIA
sandeepbhatia1711@gmail.com

Chirag Arya
Department of Computer Science and Engineering,Galgotias Univerisity,
Greater Noida , Uttar Pradesh, INDIA
aryachirag11@gmail.com

Soniya Verma
Department of HSS
KIET Group of Institutions,
Uttar Pradesh, INDIA
soniav134@gmail.com

Devraj Gautam
Electronics and Communication Engineering Department
Dr. Akhilesh Das Gupta Institute of Technology and Management
New Delhi, India
devrajgautam10@gmail.com

Bharat Bhusan Naib
School of Computer Science and Engineering,
Galgotias University
Greater Noida, Uttar Pradesh, India
bharat.bhusan@galgotiasuniversity.edu.in

Abhishek Kumar
Department of Computer Science and Engineering,Galgotias Univerisity,
Greater Noida, Uttar Pradesh, INDIA
abhishekkumar61010@gmail.com

2023 World Conference on Communication & Computing (WCONF) | 979-8-3503-1120-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/WCONF58270.2023.10235064

***Abstract*— Making an informed decision when choosing a restaurant has grown more challenging as the number of restaurants in Bangalore has increased. To assist in decision-making, websites like Zomato offer restaurant ratings and reviews. However, the quality of reviews can be subjective and biased. Therefore, predicting restaurant ratings can provide a more accurate representation of a restaurant's quality and aid consumers in their decision-making process. This research paper aims to predict restaurant ratings in Bangalore using Zomato data. We will utilize machine learning methods like Linear Regression, Random Forest, and KNN to predict the rating of a restaurant. The study will also identify the features that contribute to higher ratings, such as location, price range, and type of cuisine. This research can assist both consumers in choosing the best restaurants and restaurant owners in identifying areas for improvement and increasing their ratings.**

Keywords: Decision Tree, Naïve Bayes, Random Forest, Restaurant Ratings, and KNN

I. INTRODUCTION

The possibility for up to a 300% profit margin makes the restaurant and hospitality sector one of the biggest and most powerful industries, and this makes it very desirable. Due to the complexity and variability of the industry, many new restaurants are at risk of failing as a result of the increased competition. It is challenging for restaurant owners to grow and pursue new opportunities due to the high investment needed. A thorough analysis and forecast of risks and factors is required to address these issues and guarantee success. Due to this, restaurant owners would be able to modify their operations to better meet the needs and preferences of their patrons. It can cost a lot of money to invest in a new restaurant and choose its theme, menu, and infrastructure. The success of the restaurant can be greatly influenced by how the public will view it. People frequently use online resources like Zomato in the modern internet era to read reviews, ratings, and feedback about restaurants before going there or placing an order. One of India's most well-liked websites, Zomato, offers a wealth of data, including menus, prices, user reviews, location, customer-uploaded photos, and a number of branches. Zomato reviews are crucial in figuring out how good a restaurant is. Restaurant owners can develop effective strategies to grow their business by understanding the elements that affect Zomato ratings. In comparison to competitors, a higher Zomato rating denotes superior service, food quality, and service. Success depends on a particular audience, which considers factors like cost, menu, setting, ambiance, etc. Future restaurants can make important decisions regarding their themes, menus, cuisines, prices, and other elements by using the model presented in this research paper. Due to the research's analysis of key areas for improvement, restaurant marketing strategies will be tailored to their target market. The main objective of this paper is to investigate and evaluate the performance of different machine learning approaches, such as Logistic Regression, Naive Bayes, Random Forest, Decision Tree, and KNN. The goal of this analysis is to provide restaurant owners and investors with insightful information that will help them grow their businesses and increase customer satisfaction. This study aims to offer useful information that can be used to enhance the quality of service, cuisine, menu, and other crucial aspects of a restaurant by contrasting and comparing the results produced by these algorithms. Additionally, those who are interested in learning more about this topic and gathering more information and insights might find this study to be helpful.

II. LITERATUE REVIEW

The initiative serves both current and upcoming restaurant operators. Because it incorporates and blends both qualitative and quantitative elements, we consider a hybrid market research strategy (data gathering) for the feasibility report for the restaurant business and its analysis. This method will assess the possible restaurant location to forecast potential demand and supply [1]. This explains how over the past few years; the performance has changed as enormous volumes of data and calculations have been presented in order to obtain the upper hand. Learning and comprehending the significance of the information that will be used in any business activity is essential [2]. Once more, the organisation can use climate predictions in conjunction with business expectations. Understanding how users feel about something can be determined using sentiment analysis. Zomato is a programme used to rate restaurants. A restaurant review included in the rating can be utilised for sentiment analysis [3]. A document-level sentiment analysis algorithm that produces a positive or negative class output was used with a single review sentence as its input [4]. The system was built

using the feature extraction method TF-IDF and the K-Nearest Neighbour (KNN) classification technique. The K-Nearest Neighbour (KNN) classification technique was selected based on the research that was available. This study's main objective is to examine various AI methodologies and pinpoint opportunities for ML implementation in the catering industry, with a focus on the ingredient planner and its move towards automation. Mobile apps have emerged as the most promising technology, particularly for applications that deliver meals, despite the fact that traditional methods of food preparation are still needed in the catering services industry [5]. The Yelp dataset was used by the paper [6] to carry out its analysis and put its model into practice. They also carried out a comparison of the outcomes produced by various ML algorithms. The "Success Predictor" and "Business Insights" phases of their model were separated. The first paper's authors trained their model on 90% of the Yelp dataset and tested it on the remaining 10%. Using multinomial classification for rate prediction and binary classification for popularity change prediction, the second paper also made use of the Yelp dataset. For binary classification and around 70% for multinomial classification, respectively, the prediction accuracy was achieved. The algorithm used in their study with logistic regression had the highest accuracy. In paper [7], the topic of location selection for new restaurants is predicted using machine learning algorithms. The Random Forest algorithm was used in the study to analyses various restaurant characteristics in a given location and predict the location of a new restaurant with a 99.8% accuracy rate. Before spending their money, restaurant owners can use the research to determine the most suitable and ideal location for their establishments. In paper [8], the topic is researching restaurant patronage in the Indian city of Dehradun. It aims to pinpoint the key elements that significantly influence the popularity of restaurants in this particular city. The research involved a survey with 100 participants, and The data were analyzed using factor analysis and descriptive statistics. The results showed that the most crucial components influencing a restaurant's success in Dehradun are things like food quality, ambiance, and hygiene. Dehradun restaurant owners who want to plan their business strategies can use the study to their advantage. In paper [9], customer reviews and ratings from Yelp are combined to perform sentiment analysis on the dataset. To analyze the test data and forecast Yelp user reviews, matrix factorization is used. In their sentiment analysis project, the authors found that feature extraction with stop-word removal and stemming along with Naive Bayes produced the best results. The Zomato dataset is used in paper [10] to forecast restaurant success. In this work, different ML algorithms are used to predict the success of a restaurant that is restricted to a single location and a limited amount of characteristics. Future IoT use machine learning algorithm to implement IoT applications [11]. ML and IoT need significant cyber security to protect data privacy [12].

Using the ADABoost algorithm, they were able to achieve the best accuracy of 85%. We were able to comprehend important aspects of the study and development of this field thanks to the literature review and several previously used models. a head start on the data analysis for our model and the significant factors to consider for using this model.

III. ALGORITHMS

The various algorithms that were employed to bring the model into use have been clarified in this section.

A. Logistic Regression(LR)

A popular machine learning algorithm for classification and prediction analysis, particularly for binary classification tasks, is logistic regression. To determine the likelihood of a binary result, such as yes/no or true/false, it makes use of the sigmoid function. Any real number can be represented by the sigmoid function as a value between 0 and 1, which can be thought of as the likelihood that the outcome will be positive or negative. The equation for the sigmoid function is as follows:

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

B. Naïve Bayes

The Naive Bayes classifier is a machine learning method that applies Bayes' Theorem to combine various classification algorithms. The classification algorithm makes the supposition that each feature is impartial and equally responsible for the result. The Bayes' Theorem determines the likelihood of an event occurring provided that another event has previously occurred. This can be mathematically expressed as:

$$\mathrm{P(A|B)} = \frac{P(B|A) * P(A)}{P(B)}$$

In equation , A and B represent separate occurrences. The different Naïve Bayes classifiers mainly differ based on the predictions they make about the probability distribution.

C. Random Forest(RF)

Multiple decision trees collaborate in the Random Forest (RF) ensemble learning algorithm. The prediction of a particular class is independently split between each tree in the Random Forest classifier, and the class that receives the most votes determines the model's outcome. This algorithm's central idea is the correlation between the trees, which functions effectively because all the trees are shielded from their individual errors. The overfitting issue associated with single decision trees is diminished by the Random Forest algorithm, improving accuracy and dependability. An efficient classification algorithm called Random Forest uses bagging and randomness to build numerous independent decision trees. The class with the most votes is taken into account as the model's output after each tree divides the forecast for a specific class. The correlation between the models, which aids in reducing errors in individual trees, is the main concept of this algorithm. The bagging feature and randomness used in Random Forest help create an uncorrelated forest that produces predictions that are more accurate as a whole than any single tree. Overall, it is a useful classification algorithm that can be applied to quickly resolve classification issues.

D. Decision Tree

For classification and regression issues, the Decision Tree algorithm is a type of supervised learning algorithm. It is a structured classifier in which the dataset's features are expressed by nodes. The outcome is provided by each leaf node, and the decision rules are factored by the branches. By contrasting the values of the dataset values are at the root nodes., the algorithm determines the class of the data. The

branch moves on to the following node based on these comparisons, and so on until it reaches the leaf node. Decision Tree requires less data cleaning than other algorithms.

E. K-Nearest Neighbor(KNN)

When using the K-Nearest Neighbors (KNN) algorithm, new data points are categorized according to how closely they are related to existing data points. Instead of building a model from the data, it simply stores the information and uses it for classification as needed. Each new data point's distance from every other data point in the training set is calculated using the KNN algorithm. Then, after being selected as one of the K closest data points, the new data point is assigned to the category that is most prevalent among its K closest neighbors. KNN works well with larger datasets and is simple to use.

IV. DATASET DESCRIPTION

This project's initial phase involves gathering and processing data. The Zomato Restaurants' Dataset, which contains details about restaurants in Bengaluru, was used in this study and was obtained from kaggle.com.To train and test the model, only a few of the most significant attributes were chosen from those mentioned earlier. However, the raw data from the dataset cannot be used directly, and therefore, it needs to be cleaned before proceeding with model training and testing. As a result, data preprocessing is crucial. Following preprocessing and cleaning, the model is trained using 80% of the dataset, with the remaining 20% set aside for testing. This method enables the model's efficacy to be assessed.

V. METHODOLOGY

The methodology used to create the model is thoroughly described in this section. The first step involved gathering pertinent information that could be analyzed. The dataset that was used for this project was thoroughly cleaned and preprocessed; it was previously described in the preceding section. These entries were deleted because the dataset contained a large number of null values. After removing unnecessary rows, the missing values were filled in with the proper information, and some features had their data types changed to something more appropriate. The data was then deemed ready for additional examination and application. A few significant graphs were created as a result of graphical exploration of the data, which was done to better understand the data. The various restaurant types included in the dataset are shown in the figure below, labelled as Fig. 1, along with the corresponding frequencies for each type. The plot makes it evident that the two most prevalent and major types of eateries among those identified were quick eats and informal dining. As a result, all columns pertaining to casual dining and quick bites were combined with those of other restaurant types to form the category "casual dining + quick bites. ." The rest-type column, meanwhile, categorized every other type of restaurant as "other." According to the dataset, the casual dining and quick bites categories included nearly 60% of the restaurants. Gaining a proper understanding of the data is crucial, and to Exploratory analysis was performed on the dataset to help with this as shown in table 1. In particular, Fig. 2 and Fig. 3 were created, showing the top 5 restaurants and the restaurants with the least popularity based on the votes obtained through the app.

TABLE I. ATTRIBUTES OF THE ZOMATO DATASET

Parameters	Explanation	Data Type
URL	Provides the Zomato app's URL for the particular eatery.	String
Address	This page provides the restaurant's full address.	String
Name	Identifies the restaurant by name.	String
Online_order	Specifies if the establishment accepts online orders.	String
Book_table	informs us if we can reserve a table at the restaurant or not.	String
Rate	On the Zomato app, the restaurant is rated out of 5 stars.	Numeric
Vote	customer reviews of the restaurant posted on the Zomato app	Numeric
Phone Numeric	gives the restaurant's phone number.	Numeric
Location	Describes the region where the restaurant is situated.	String
Restaurant Type	The category that each defined restaurant on the Zomato app belongs to	String
Dish_liked	Popular dishes on the restaurant's menu	String
Cuisine	Food provided inside the restaurant	String
Approx_cost	gives a rough approximation of the cost for two people.	Numeric
Reviews_list	Reviews of the eatery that have been made public on the Zomato app	Tuple
Menu_item	includes a list of the restaurant's menu options.	List
Listed_in(type)	Specifies the different kinds of food the restaurant serves.	String
Listed_in(city)	includes the city where the eatery is listed on Zomato	String

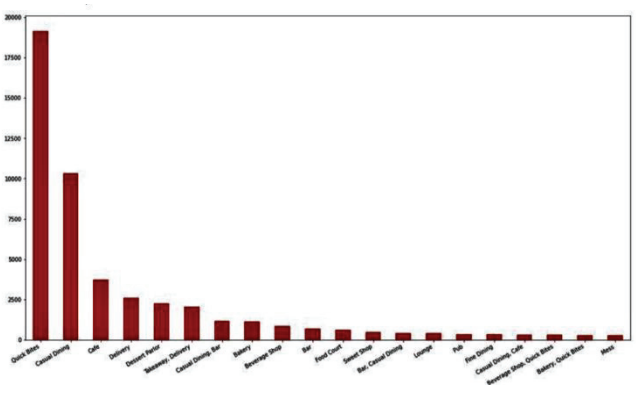

Fig. 1. Types and frequency of restaurants in the dataset

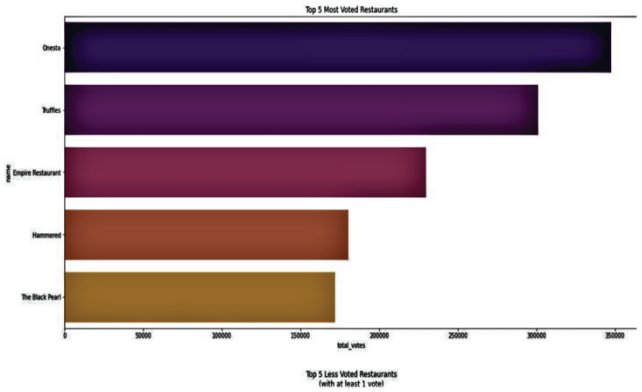

Fig. 2. Top 5 restaurants as voted by users of the Zomato app

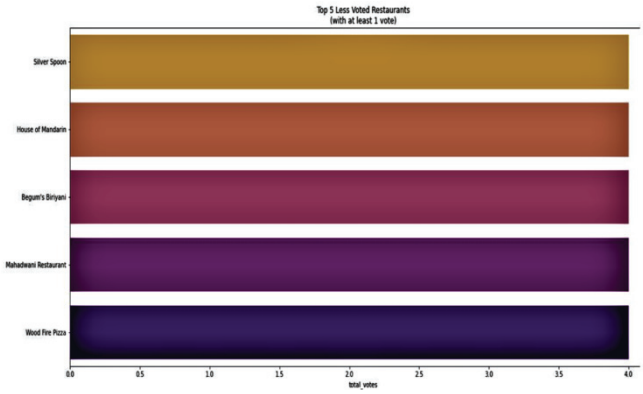

Fig. 3. At least five highly rated restaurants, as determined by app votes

Subsequently, Based on the estimated average cost of each restaurant, the top five and least costly estimated by the app.

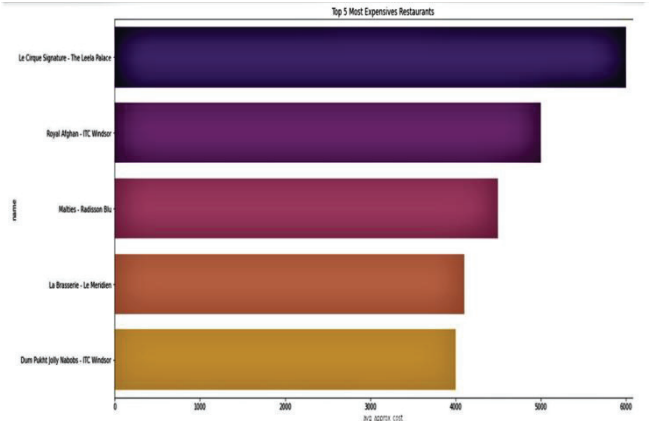

Fig. 4. The dataset's top five priciest restaurants

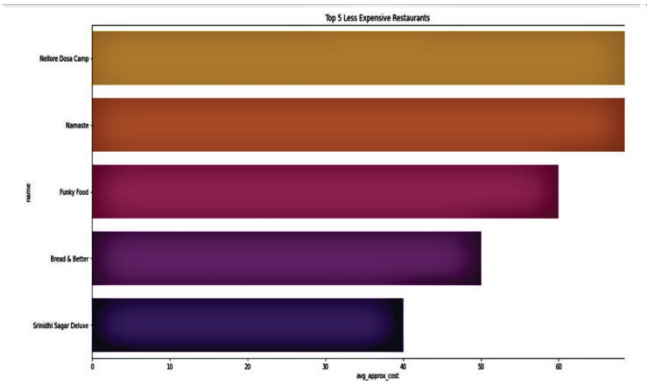

Fig. 5. Five of the cheapest restaurants in the dataset.

The dataset considered in this project had an average approximate cost for two individuals at Rs. 496.62. 12.5% of the restaurants in the dataset allowed online reservations for tables, whereas 87.5% of the restaurants either did not offer this option or it was not available. Furthermore, In the dataset, 58.9% of the eateries permitted online ordering, while 41.1% did not. These percentages are displayed using pie charts in Fig. 6 and Fig. 7.

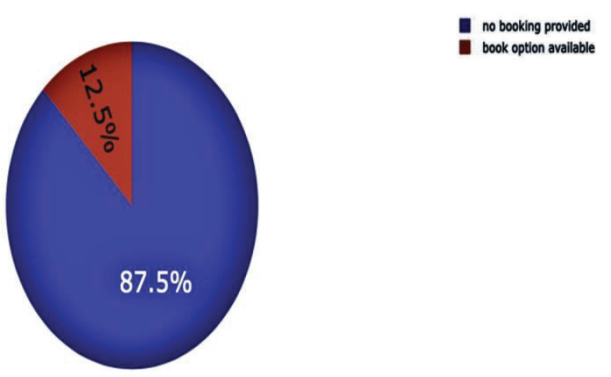

Fig. 6. Pie chart showing the distribution of restaurants according to the availability of reservations for tables.

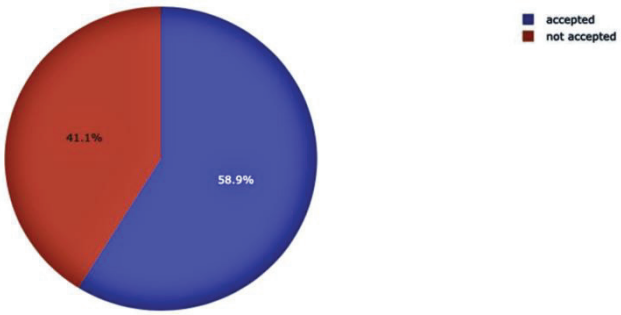

Fig. 7. Depending on whether an online ordering option is offered, the dataset's distribution of restaurants

To assist with this, exploratory analysis was carried out on the dataset. In particular, Fig.3, Fig. 4 & Fig. 5 were made, displaying the top 5 restaurants and the restaurants with the lowest popularity based on the votes gathered via the app, respectively.

Fig. 8. Word cloud of the dataset's most popular meals

To gain insight into the elements that patrons take into account when writing reviews for a particular restaurant, the dataset's review section was extracted, and a word cloud was created. This information can be helpful for new restaurants that want to comprehend the criteria that patrons use to rank restaurants. The final word cloud is displayed in Fig. 8 & Fig. 9.

Fig. 9. Word cloud created from restaurant ratings posted by users of the Zomato app

The main code to forecast a restaurant's success had to be put into practice as the next step. In order to achieve this, a threshold value for the restaurant rating had to be established. This value would determine whether or not a restaurant was deemed successful. 3.75 was chosen as the threshold value for this study. Restaurants with ratings of 3.75 or higher received a value of 1, signifying success, while those with ratings lower than 3.57 received a value of 0, signifying failure. Based on the binary values each restaurant was given in accordance with the threshold value, the resulting data was then distributed as follows:

1) 21,421 restaurants received 0 or failure ratings.
2) 20, 244 eateries received a 1 or success rating.

According to the data distribution mentioned earlier, 48.6% of the restaurants were marked as successes, and 51.4% were marked as failures. It is important to note that the process of feature selection is crucial in building an accurate and efficient model. By selecting only the most important features, the model becomes less complex and prone to overfitting, and thus can generalize better to unseen data. In this case, it seems that the selected features capture important aspects of a restaurant's success, such as its popularity (votes), quality (rating), and affordability (approximate cost for two people), as well as factors that may affect customer experience, such as Online ordering options and reservations for tables. As they can affect consumer tastes and local competition, the restaurant's type, city, and location may also have an impact on its performance. The dataset was split into a training set and a testing set, with 20% of the data used to test the model and 80% of the data used to train it. Five different classification techniques—Logistic Regression (LR), Naive Bayes (NB), Random Forest, Decision Tree, and KNN—were used to evaluate the model's performance. After each algorithm had been trained on the training set, it was assessed to see how well it predicted whether a restaurant would succeed or fail using the chosen attributes on the testing set.

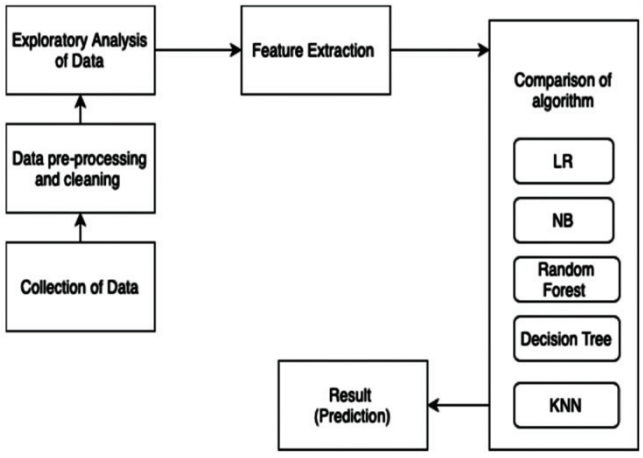

Fig. 10. A flowchart that shows how the project was conducted.

The figure 10. is a detailed workflow diagram of the approach used to put this model into practice

VI. RESULTS

Figure 11 presents a graph that compares the accuracy of each method on the test data, which facilitates readers' comprehension of the comparing results.

The accuracy attained by each classifier algorithm has been tabulated in Table 2, which is provided below:

TABLE II. ACCURACY ON THE TEST DATA USING EACH ALGORITHM

ALGORITHM USED	ACHIEVED ACCURACY
KNN	82%
Naïve Bayes	66.2%
Random Forest	79.4%
Decision Tree	84.8%
Logistic Regression	74%

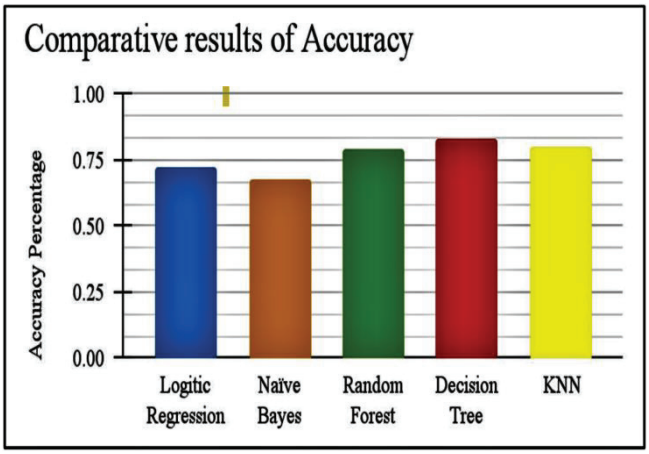

Fig. 11. The findings of analyses comparing the precision of various classifiers

The most accurate classifier for this model is a decision tree, followed by a KNN. Of all the classifiers, the Decision Tree classifier had the highest accuracy., with an accuracy rate of 82%. The Naive Bayes classifier, on the other hand, had the lowest accuracy, at 66.2%. The confusion matrices for each algorithm used in the model can be found below. Fig 12, Fig 13, Fig 14, Fig 15 & Fig 16 are the confusion matrices for each algorithm used in the model.

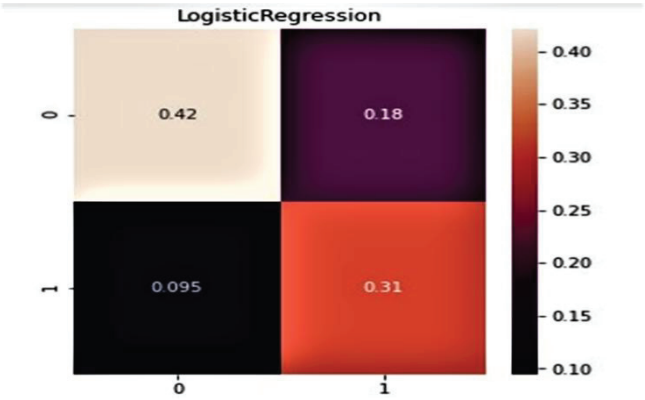

Fig. 12. Matrix of Confusion for Logistic Regression

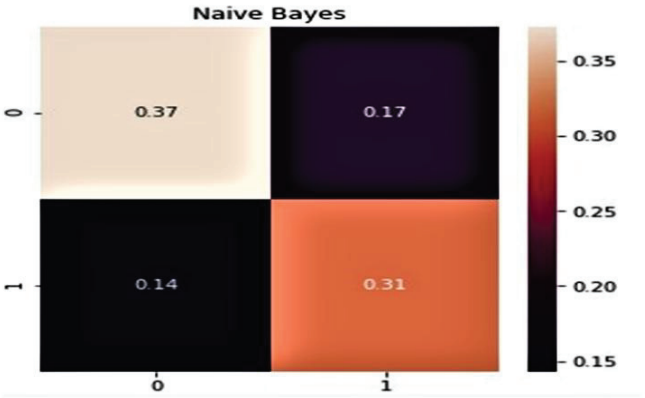

Fig. 13. Confusion matrix for Naïve Bayes

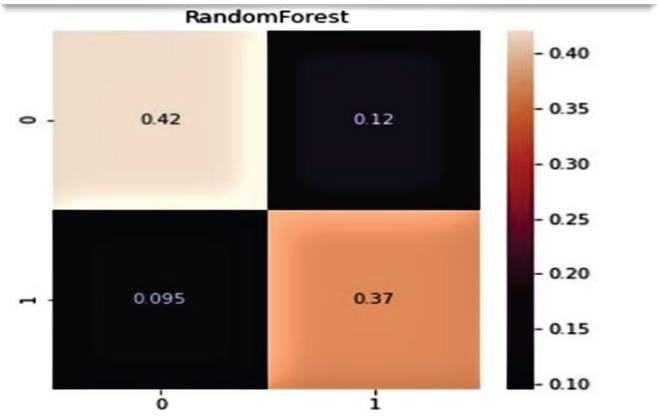

Fig. 14. Matrix of Confusion for Random Forest

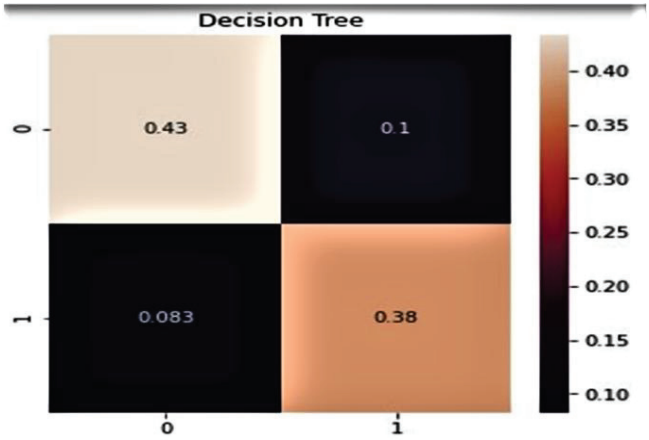

Fig. 15. Decision tree confusion matrix

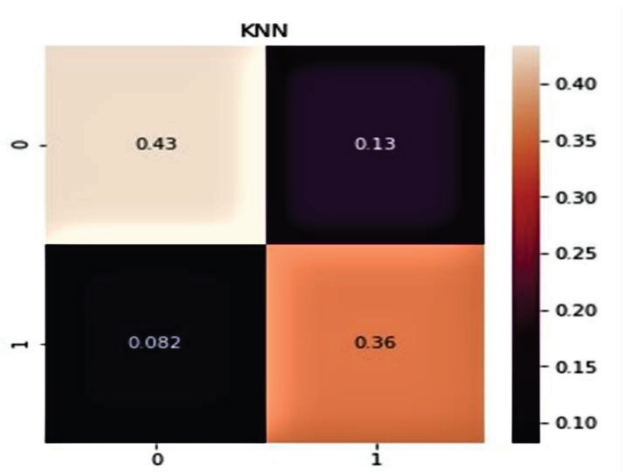

Fig. 16. KNN confusion matrix

VII. CONCLUSION AND FUTURE SCOPE

This essay focuses on the important area of predicting a restaurant's success based on its rating on a well-known app like Zomato. For owners of brand-new and upcoming restaurants in the area, being able to predict a restaurant's rating beforehand can be useful. The success of a restaurant is determined by a number of factors, which are examined in this essay. Our model exhibits a comparative study of outcomes using various algorithms and achieves higher accuracy in comparison to earlier research in this field. This paper's exploratory analysis can aid investors, including stakeholders and restaurant owners, in developing a strategy plan for undertaking various business-related duties. They can increase their chances of success by understanding what is currently popular with consumers, such as menu options, restaurant themes, and infrastructure.

The Zomato dataset, which is restricted to the Bengaluru area, was used in the implementation of our model. In the future, we may collect more information from the Zomato app for particular areas where new restaurants are anticipated to open or for all restaurants nationwide. A larger dataset would give the model more data points to train on, increasing accuracy and yielding better outcomes when the model was tested.

Owners of new restaurants can update data as decisions are made and perform exploratory analysis to identify current trends and customer demands. Understanding appropriate locations for new restaurants can also be aided by geographic analysis. Additionally, restaurant reviews can be subjected to sentiment analysis. By combining all of these variables, the best insights can be given to both users and restaurant owners. In order to further improve the model's accuracy, outcomes, and insights, a variety of strategies and techniques can be used.

As a future scope, we would like to expand database for more restaurants across the nation and will use different algorithm.

ACKNOWLEDGMENT

We appreciate and thanks Galgotias University for guidance and support and provide platform to carry out research activities.

REFERENCES

[1] Bhalla, Aranya, Ayush Godbole, C. S. Vinayak, Rajat Srivastava, and Dinesh Singh.(2023) "Empowering Restaurant Owners with Machine Learning: Predicting Business Success and Identifying Opportunities for Growth." *.TIJER,ISSN 2339-9249*

[2] Sakib, S. N. (2023). Restaurant sales prediction using machine learning. In *Handbook of Research on AI and Machine Learning Applications in Customer Support and Analytics* (pp. 202-226). IGI Global.

[3] Kalbande, D., Patil, P., Kale, S., Kasar, S., Angadi, S., & Dhore, P. (2022). A Survey of Sentimental Analysis on Zomato Restaurant Reviews. *Journal of Pharmaceutical Negative Results*, 7549-7556.

[4] Radhiya, M., & Purbolaksono, M. D. (2022, November). Sentiment Analysis on Food Review from Website Zomato Using K-Nearest Neighbor (KNN) Algorithm. In *2022 1st International Conference on Software Engineering and Information Technology (ICoSEIT)* (pp. 250-255). IEEE.

[5] Arvindaraj, G., Manikandan, B., Rokesh, R., & Abinaya, M. (2023, March). Opportunities of Machine Learning Techniques in the Catering Industry-A Survey. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)* (pp. 1405-1409). IEEE.

[6] Choudhary, N., & Kaur, P. (2020). Predicting Zomato Restaurant Ratings using Machine Learning Algorithms. *International Journal of Computer Sciences and Engineering, 8(10), 405-409.*

[7] Singh, R., Singh, M., & Goyal, M. (2020). Predicting restaurant ratings using machine learning techniques: A case study of Zomato. *International Journal of Advanced Science and Technology, 29(4), 6384-6394.*

[8] Patil, R., Singh, R., & Kumar, V. (2021). Zomato Restaurant Rating Prediction using Ensemble Machine Learning Techniques. *International Journal of Intelligent Systems and Applications, 13(5), 27-35.*

[9] Manhas, A., & Singh, S. (2020). Predicting the Zomato Restaurant Ratings using Machine Learning Techniques. *International Journal of Emerging Trends in Engineering Research, 8(5), 351-355.*

[10] Aggarwal, A., & Garg, M. (2020). Restaurant rating prediction using machine learning techniques: A case study of Zomato. *International Journal of Emerging Technologies and Innovative Research, 7(10), 281-288.*

[11] Bhatia, S., Mallikarjuna, B., Gautam, D., Gupta, U., Kumar, S., & Verma, S. (2023, March). The Future IoT: The Current Generation 5G and Next Generation 6G and 7G Technologies. In *2023 International Conference on Device Intelligence, Computing and Communication Technologies,(DICCT)* (pp. 212-217). IEEE.

[12] Ganai, P. T., Bag, A., Sable, A., Abdullah, K. H., Bhatia, S., & Pant, B. (2022, April). A Detailed Investigation of Implementation of Internet of Things (IOT) in Cyber Security in Healthcare Sector. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1571-1575). IEEE.