FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 38, Nº 2, June 2025, pp. 355 - 374
https://doi.org/10.2298/FUEE2502355V

Original scientific paper

MACHINE LEARNING-DRIVEN STATISTICAL ANALYSIS OF INDIAN RESTAURANTS: INSIGHTS FROM THE ZOMATO DATASET

Ayushi Vaidhy[1], Deepak Batham[2], Rachit Jain[3], Amit Kumar Manjhwar[1]

[1]Dept. of Computer Science Engineering, Madhav Institute of Technology & Science, Deemed University, Gwalior, India
[2]Dept. of Electronics Engineering, Madhav Institute of Technology & Science, Deemed University, Gwalior, India
[3]Dept. of IT, Prestige Institute of Management & Research, Gwalior, India

ORCID iDs:	Ayushi Vaidhy	https://orcid.org/0009-0007-8701-9685
	Deepak Batham	https://orcid.org/0000-0003-4499-7239
	Rachit Jain	https://orcid.org/0000-0002-3001-2438
	Amit Kumar Manjhwar	https://orcid.org/0000-0002-9577-6295

Abstract. *Advances in technology and web applications, such as Zomato, have significantly transformed the restaurant industry by catering to diverse culinary preferences and offering a wide variety of food options to customers. This platform stores a vast amount of data that can be analyzed for valuable insights. The paper examines dining habits and restaurant performance through exploratory data analysis (EDA) and machine learning (ML) algorithms, helping customers find the best restaurants based on cost, ratings, location, food quality, and service. The study applies several ML models, including Linear Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), XGBoost, K-Nearest Neighbors (KNN), and LASSO to the Zomato dataset. The results are evaluated using metrics such as accuracy, mean absolute error (MAE), model fit time, and model prediction time. Among these models, DT and RF show the highest predictive accuracy, with RF achieving 97.86% and outperforming other algorithms. These findings provide restaurant owners with valuable insights to enhance customer satisfaction, optimize pricing, and improve service quality. The study also demonstrates the important role of ML in the restaurant industry and suggests future opportunities for integrating real-time data, deep learning models, and sentiment analysis to offer even more precise predictions and insights.*

Key words: *machine learning, EDA, Zomato, data analysis, accuracy*

Received October 29, 2024; revised February 17, 2025 and March 4, 2025; accepted March 9, 2025
Corresponding author: Rachit Jain
Dept. of IT, Prestige Institute of Management & Research, Gwalior, India
E-mail: rachit2709@gmail.com

1. INTRODUCTION

The global restaurant industries are undergoing a profound transformation driven by the technological advancements, shifting consumer expectations, multiple culinary choices, diverse food quality and tastes, and an unprecedented surge in data available through online platforms like YouTube, Facebook, WhatsApp, Instagram, and restaurant websites. This wealth of data, particularly customer reviews, is reshaping the restaurant landscape [1]. The restaurant industries are encountering unique challenges, but also huge opportunities, thanks to the fusion of data science and artificial intelligence (AI) tools, powered by machine learning (ML) algorithms. These technologies are revolutionizing how restaurants predict price, assess food quality, and understand customer preferences, serving as a key driver for sustainable growth [1-3].

ML, a subset of AI closely related to Deep Learning (DL), is transforming how restaurants navigate the complexities of modern consumer behavior. Fig. 1 illustrates the fundamental model and correlation of AI, ML, and DL, which utilize algorithms to analyze the vast datasets, identify patterns, making predictions that enable businesses to make informed, data-driven decisions. This technology plays a critical role in navigating the complexities of the modern restaurant industries. As consumer preferences constantly change and vast amounts of data are generated daily, traditional approaches of examination often fall short. ML allows for the efficient processing and analysis of this data, providing accurate predictions on customer behavior, food quality, pricing trends, and more. By integrating AI and ML into restaurant management, stakeholders can make better decisions, optimize operations, enhance customer satisfaction, and drive long-term growth. With the help of predictive models, restaurants can stay ahead of industry trends and maintain a competitive advantage.

In this evolving landscape, leading platforms such as Zomato, Swiggy, FreshMenu, Dunzo, Guruhub, EatSure, UberEats, Deliveroo, Domino's are playing crucial role, acting as a bridge between the customers and the vast tapestry of dining establishments [4]. Zomato is an aggregator of Indian restaurants and a food transportation company founded by Deepinder Goyal and Pankaj Chaddah in the year 2008. Since 2023, Zomato has expanded its offerings, providing food delivery option from partner restaurants, along with restaurant information, menus, and user reviews across more than 1,000 cities and towns in India [5]. The Zomato restaurant dataset is a treasure trove of culinary information that encapsulates the essence of the digital culinary revolution [6-7]. By analyzing these datasets, restaurant can gain valuable insights into key metrics such as the average cost of food for two people, emerging consumer trends and preferences, the best restaurants in different locations, food quality and taste, and service standards. Identifying patterns and correlations within this data can empower restaurant owners and stakeholders to anticipate and adapt to shifting customer demands, enabling them to stay competitive and relevant in an ever-changing industry [1].

In this paper, various restaurants data available on Kaggle, collected from the Zomato app, is used. This dataset contains a total of 2,11,944 entries with 26 distinct entities (See Table 1). Each entity serves a unique purpose in characterizing the restaurant details. This Zomato dataset is analyzed using EDA [8-10] under various evaluation parameters, including the top ten restaurant chains in India by the number of outlets and average ratings, the top five restaurants by establishment type, and the relationship between price range and restaurant ratings. Additionally, several ML algorithms are applied to the Zomato dataset, such as Linear Regression (LR) [11], Decision Tree (DT) [12], Random Forest (RF), Gradient Boosting (GB), XGBoost, K-Nearest Neighbors (KNN) [13], and

Least Absolute Shrinkage and Selection Operator (LASSO) [13-14]. The results are evaluated using metrics of Accuracy, Mean Absolute Error (MAE), Model Fit Time, and Model Prediction Time. Based on the simulation results on the Zomato dataset, the DT and RF algorithms exhibit the highest accuracy and the lowest MAE compared to the other algorithms. These findings can assist restaurant owners in making appropriate decisions that could potentially increase their earnings or profits.

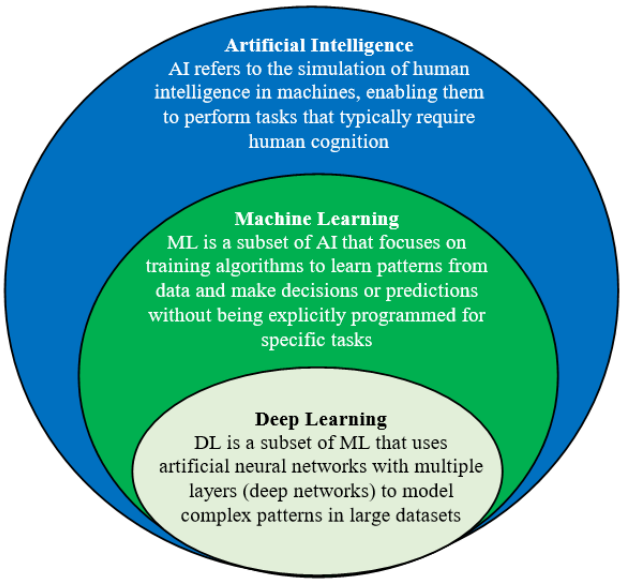


Fig. 1 Illustration of AI, ML and DL

Remaining paper is organized into six sections. Section 2 discusses the related work. Section 3 presents data description, pre-processing, and workflow. EDA and its results are covers in Section 4. Section 5 describes the ML algorithms used in this study. Section 6 presents the obtained results and their analysis. Finally, Section 7 concludes the research work.

2. Related Work

The present section discusses related work. Recently, several studies have focused on restaurant data analysis and price prediction. In [1], authors review various food recommendation systems (FRS) and analyze them using ML algorithms. The primary goal of this study is to bridge the gap between the development of FRS and other recommender systems. Out of 2738 studies, the authors selected only 67 high-quality studies for analysis. The study highlights that there is a wide range FRS, most of which are designed to offer non-personalized suggestions using ML techniques and content-based filtering. This information can guide future research in selecting appropriate strategies for developing FRS. In [4], authors applied LR technique to identify the best restaurant by location. This analysis is based on customer feedback obtained from the Zomato dataset, specifically satisfaction ratings. In [6], the authors used EDA tools to examine restaurants rating in Bangalore. In [15], the

authors used DT and RF algorithms on a sample of 8,500 data points to classify restaurants based on their service attributes. The results showed that the DT classifier, with 63.5% accuracy, outperformed the RF classifier, which had 56% accuracy. In [16], the authors examined various restaurants on the basis of food quality and services offered by the restaurant owners. The analysis revealed that the rating of a restaurant depends on many factors such as reviews, area, average cost of mile for two persons, votes, cuisines, and restaurant type.

Another important aspect to analyze restaurant is the offered food hygiene and its quality. Traditional quality measurement methods often require significant resources, skilled labor, and complex analytical techniques, making quick and cost-effective solutions essential. The food industry faces challenges in evaluating food quality due to the need for expensive equipment, intricate processes, and thorough analyses to ensure that the product sold to consumers are safe and of the highest quality. In this context, the authors in [17] review food quality assessment using traditional methods, advanced ML algorithms, and the Electronic Nose System (ENS). ENS technology is an innovative approach that can distinguish between different aromas using a variety of electronic sensors, showing promising results when applied to various food items. ML algorithms play a crucial role in analyzing the complex data collected by ENS, enabling accurate food identification and quality assessment based on distinct odors. This review investigates the combination of ENS and ML algorithms, suggesting a powerful non-destructive tool for food quality assessment that surpasses conventional, time-consuming analytical methods. In [18], the authors predicted restaurant rating and popularity using ML algorithms on the Yelp dataset. Logistic regression, naive Bayes, KNN, and Support Vector Machine (SVM) were used, with logistic regression achieving the best performance.

In [19], authors applied various ML algorithms to classify Indian dishes as vegetarian or non-vegetarian from images. Among these algorithms, the RF algorithm showing the highest accuracy compared to DT, KNN, and SVM. In [20], authors identified and categorized the Indian dishes from images by using deep learning model such as convolution neural network (CNN). In [21], authors used ML algorithms such as logistic regression and naive bayes, and deep learning models like CNN and bi-directional long short-term memory (Bi-LSTM) for sentiment analysis of restaurant reviews. Both CNN and Bi-LSTM model achieves the accuracy of 89% and 90%, respectively. In [22], authors calculate foods calories by traditional Indian food images, and classified using CNN model. In [23], authors classified Indian Food items among 5000 images, 15000 annotations with 30 food class by applying YOLO5, YOLO7 and YOLO8 algorithms. The algorithms are compared on the metric of accuracy, recall, and speed.

When setting up a new restaurant, it is important to conduct a basic survey covering market trend, customer preference, location, food choices, taste, and customer's spending capacity per person. Building on this, in [24] authors applied regression models to predict outcomes based on these easily controlled parameters before opening a new restaurant. The model's metrics were then compared to determine the optimal regression model for future predictions. It has been observed form this study that ML algorithms play a crucial role in analyzing datasets collected from various home delivery food apps and other platforms. These data can be used to assess the restaurant ratings, earnings, customer preferences, their needs at different time slots such as early morning (for breakfast), afternoon (for lunch), evening snacks, and dinner habits, drink choices, and more. In this paper, EDA and ML algorithms are applied to analyze the top restaurants, food choices, and established restaurants in India using the Zomato dataset.

3. Dataset Description, Pre-processing, and Work Flow

This section deals with dataset description, pre-processing, and flow of the presented work. Dataset description is presented in Section 3.1. Pre-processing of data such as data cleaning and handling of missing values are discussing in Section 3.2. Section 3.3 represents and discusses the flow diagram of present work.

3.1. Dataset Description

The foundation of this research is based on the Zomato restaurant dataset sourced from Kaggle, encompassing a total of 2,11,944 entries with 26 distinct columns. Each column serves a unique purpose in characterizing restaurants detail. Table 1 represents key entities and their description.

Table 1 Key entities and their description

S. No.	Entities	Entity Description
1	*res_id*	A unique identifier for each restaurant
2	*name*	Name of the restaurant
3	*establishment*	Type of establishment or restaurant category
4	*url*	Web URL associated with the restaurant
5	*address*	Physical address of the restaurant
6	*city*	The city where the restaurant is located
7	*city_id*	A numerical identifier for the city
8	*locality*	The specific locality or neighborhood within the city
9	*latitude*	The geographical latitude coordinate of the restaurant's location
10	*longitude*	The geographical longitude coordinate of the restaurant's location
11	*zipcode*	The postal code of the restaurant's location
12	*country_id*	The numerical identifier for the country (India in this case)
13	*locality_verbose*	A detailed description of the restaurant's locality
14	*cuisines*	The types of cuisines served by the restaurant
15	*timings*	Information about the operating hours of the restaurant
16	*average_cost_for_two*	The average cost for two people dining at the restaurant
17	*price_range*	A numerical indicator of the price range of the restaurant
18	*currency*	The currency used for pricing
19	*highlights*	Notable features and services offered by the restaurant
20	*aggregate_rating*	The overall rating of the restaurant
21	*rating_text*	A textual representation of the restaurant's rating (e.g., Excellent', Very Good'. etc.)
22	*votes*	The number of user votes or ratings received by the restaurant
23	*photo_count*	The count of photos associated with the restaurant
24	*opentable_support*	Indicates whether the restaurant supports reservations through open table
25	*delivery*	A binary indicator (0 or 1) for restaurant delivery service availability
26	*takeaway*	A binary indicator (0 or 1) for restaurant takeaway service availability

3.2. Data Cleaning and Handling of Missing Values

Ensuring data integrity is paramount for meaningful analysis. Several techniques have been employed in the data cleaning and handling of missing values.

- Removing Duplicates: All the duplicate entries based on 'res_id' have been removed or eliminated to maintain dataset consistency.
- Cleaning Establishment Values: Square brackets enclosing 'establishment' values have been removed for readability, and empty values were replaced with 'NA'.
- Handling Missing Values: Missing values in 'address' and 'timings' have been imputed with placeholder values ('Unknown' and 'Not available', respectively), while missing values in 'opentable_support' are filled with the default value '0'.

These steps made the dataset meaningful, robust, and free from inconsistencies and made it ready for exploration and predictive modelling. The updated dataset focuses on key attributes essential to facilitating a more concise and effective analysis. Initially, the dataset contained 2,11,944 entries. After handling missing values and feature selection, approximately 1,95,000 entries were retained for training and testing. A 90-10 split was applied, where 90% of the data was used for training, and 10% for testing.

3.3. Work Flow

Fig. 2 shows flow diagram of the present work. First collect the Zomato dataset from the Kaggle website. These data are pre-processed to avoid errors and null entity, then EDA is applying to the filtered dataset to extract the relevant features of restaurants. Now, select the relevant features and transform them for modelling. The ML algorithms used 90% data for training purpose and 10% data for testing. The performance of each algorithm is then evaluated and compared.

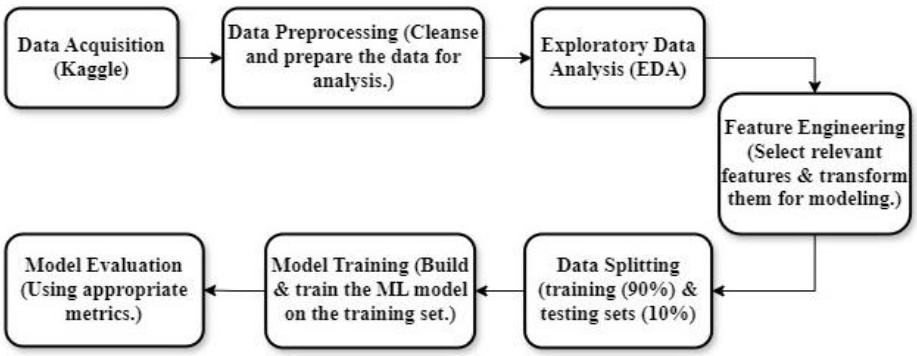


Fig. 2 Flow diagram of present work

4. EDA

The section provides a detailed description of the EDA tool and its analysis report. Data scientists use EDA as a fundamental and vital tool to examine, explore, and summarize dataset's key features, frequently using data visualization techniques [8], [9], [10]. It can help to identify patterns within data and determine errors and relationships among variables

used in the dataset. The purpose of exploratory analysis is to ascertain the validity of the data generated, thereby assisting stakeholders in obtaining answers to their inquiries regarding confidence intervals, categorical variables, and standard deviations. After EDA is finished, conclusions are made, and its features can be applied to more complex data analysis or modelling, such as machine learning [8], [9], [10]. In this paper, EDA is used for Zomato restaurant dataset to identify the patterns, relationships, key observations including harnessing various data visualization techniques to achieve desired business objectives and goals. The data visualization and interpretations for EDA is expressed in terms of top restaurants chain in India by the number of outlets, average rating, type of establishment, and the relationship between the price range and rating. The extracted results are shown in graph and discussed separately.

A. Top ten restaurant chain in India by number of outlets: Our exploration began with a horizontal bar chart, spotlighting Domino's Pizza as the leading chain with a substantial presence across India shown in Fig. 3(a). This dominance hints at successful expansion or strong brand recognition, presenting potential growth opportunities.

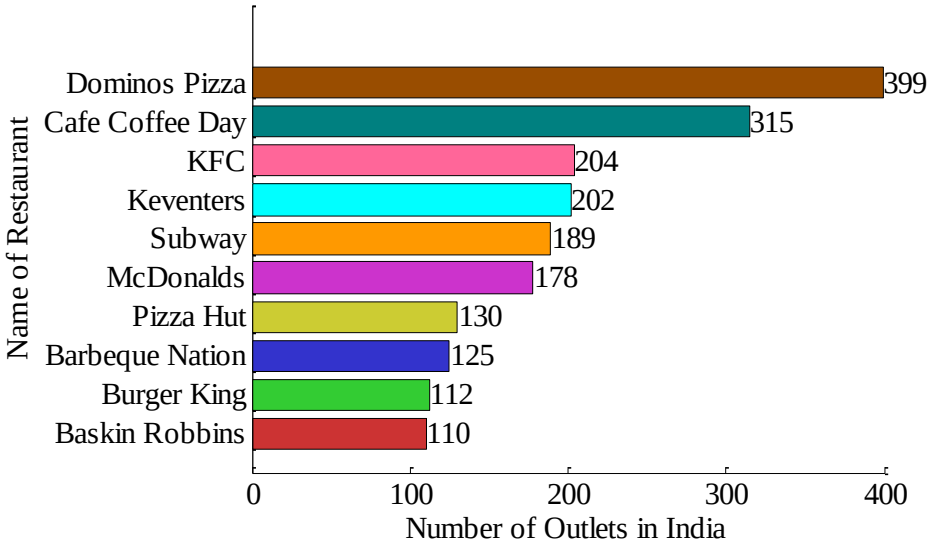
B. Top ten restaurant chain in India by average rating: Fig. 3(b) shows that Absolute Barbecues (AB's) emerged as the leader in average ratings, reflecting consistent customer satisfaction. High average ratings imply quality service, influencing customer choices and fostering loyalty.

C. Top five establishment type restaurants: Based on the EDA analysis shown in Fig. 4(a), Quick Bites took center stage, suggesting a prevalent establishment type in the Indian restaurant landscape. This insight informs strategic decisions for entrepreneurs and investors, recognizing the popularity of specific establishment types.

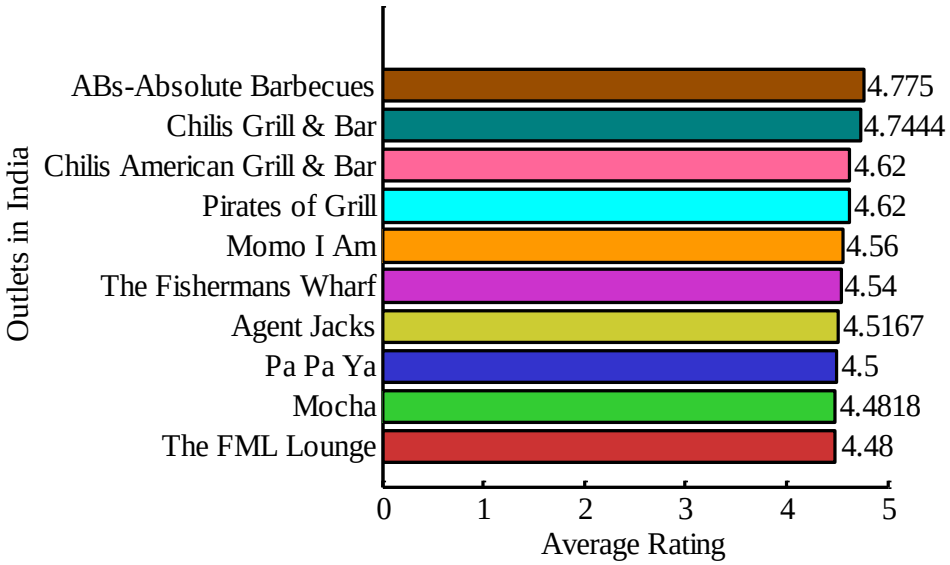
D. Relationship between price range and ratings: A boxplot unveiled the intricate relationship between price range and aggregate ratings shown in Fig. 4(b). Understanding this correlation aids in the formulation of effective pricing strategies, shedding light on how pricing impacts customer perceptions.

E. Other Observations from EDA

- Approximately 35% of Indian restaurants are a part of chain with dominant players like Domino's Pizza, Cafe Coffee Day, and KFC indicating a robust market presence.
- Barbecues and Grill food chains boast the highest average ratings, signaling positive customer sentiment. Quick Bites and Casual Dining establishments, on the other hand, lead in terms of the number of outlets.
- Bangalore emerges as the city with the highest number of restaurants, while Gurgaon boasts the highest-rated restaurants with an average rating of 3.83. Hyderabad takes the lead in critical votes.
- North Indian cuisine claims the top spot in preferences, closely followed by Chinese. Interestingly, international cuisines tend to receive higher ratings than local offerings.
- Gastro pubs, romantic dining, and establishments offering Craft Beer are particularly well-rated by customers. The majority of restaurants fall within the 3 to 4 rating range.
- The majority of restaurants fall within the budget-friendly category, with an average cost ranging between Rs. 250 to Rs. 800. Notably, higher average costs correlate with a higher likelihood of a restaurant securing a higher rating.

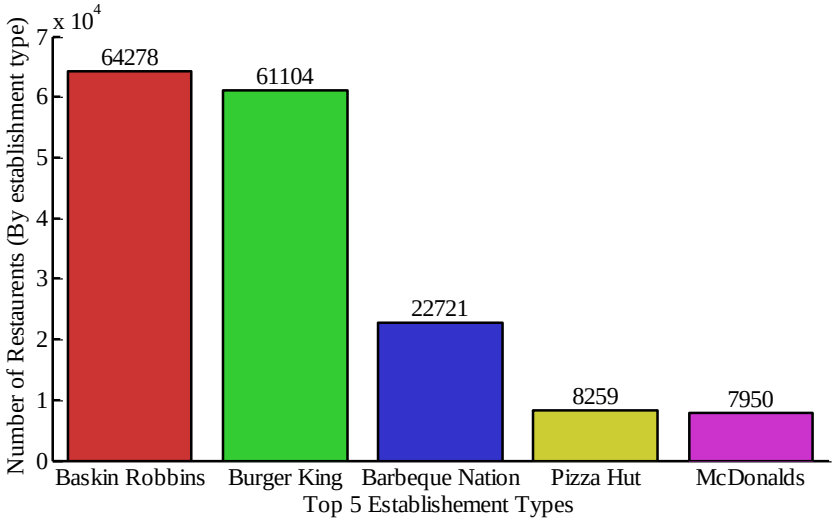


(a)

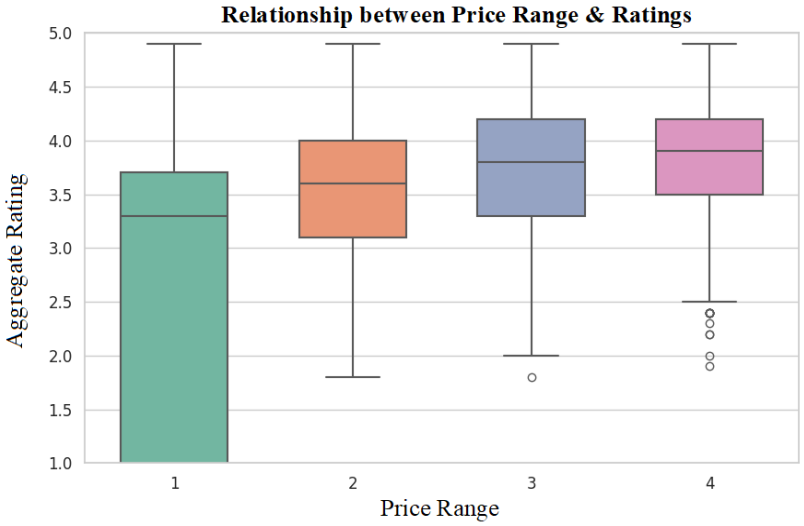


(b)

Fig. 3 Horizontal bar graph of (a) Top 10 restaurant chain in India by the number of outlets, and (b) Top 10 restaurant chain in India by average rating out of five



(a)



(b)

Fig. 4 (a) Vertical Bar graph shows the top five category wise established number of restaurants in India, (b) Boxplot shows the relationship between the price range and aggregate rating of restaurants in India

5. Machine Learning Algorithms

In ML, dataset is divided into two subsets; the training set, and the testing set. The training set is used to train the algorithms (i.e., models), while the testing set evaluates the algorithms performance. Here, 90% data is used for training, and the rest 10% data is used for testing purposes. This division ensures that the models generalize well to new, unseen data. Several ML models are available in the literature. In this paper, we have analyzed LR, DTR, RF, GB, XGBoost, KNN, and LASSO to predict best restaurants under different criterions such as restaurant rating, price ranges, type of establishment, food hygiene, service quality, and location. ML algorithms are evaluated on the basis of metrics of R-square (or accuracy (%)), MAE, model fit, and model prediction time. All such algorithms are discussed separately in detail.

5.1. Linear Regression (LR)

LR is a supervised learning technique that determines a linear relationship between x (input) and y (output) by predicting the value of the dependent variable (y) based on the provided independent variable (x). Thus, the term "linear regression," with y = mx + c serving as the LR hypothesis, is used. The line that fits our model the best is the regression line. In order to produce scientifically valid and dependable predictions, LR employs a well-established statistical process. The model can be trained rapidly, and the algorithm is simple to comprehend [11], [19]. The graphical representation of Linear Regression is provided in Fig. 5.

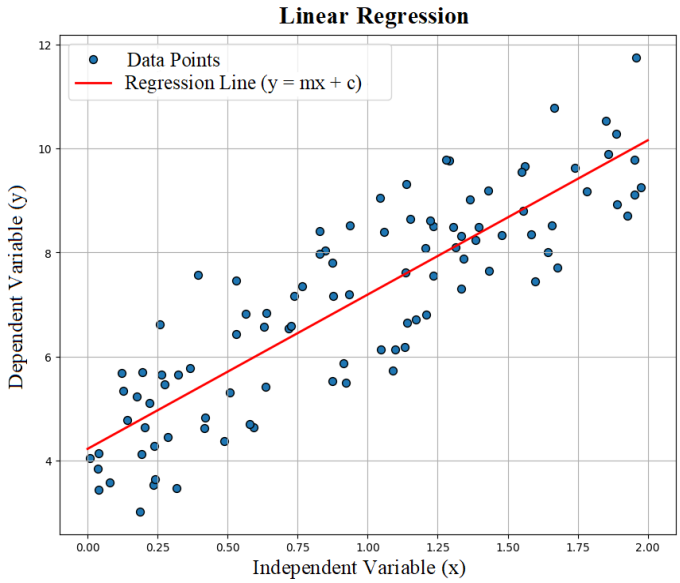


Fig. 5 Linear regression

5.2. Decision Tree (DT)

In DT [12], [15], data is structured as a tree, where each internal node represents a decision based on the input features. DT introduces non-linearity to the predictive process and captures intricate relationships between features and the target variable [12]. It offers flexibility in handling diverse data types, presenting a viable alternative for scenarios with complex non-linear patterns. The splitting process begins at the root node and proceeds along a branch tree to the leaf node, or terminal node, which houses the algorithm's prediction or result. DT employs a top-down methodology. A binary tree that shows how a decision node divides into two nodes according to certain conditions can be used to represent each subtree of the DT model. Regression trees are DTs in which the target variable or the terminal node is capable of accepting continuous values. The fundamental structure of the Decision Tree is presented in Fig. 6.

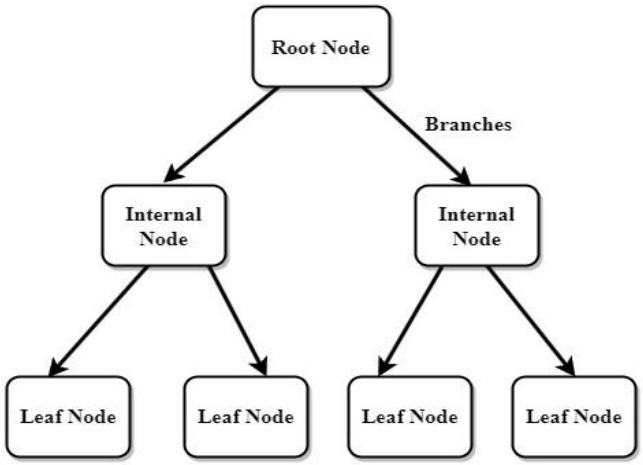


Fig. 6 Basic structure of decision tree

5.3. Random Forest (RF)

RF [6], [16] can handle both the regression and the classification problem by using several DTs and resampling methods known as aggregation, or bagging, and bootstrapping. Using bagging, the smaller models are combined to create an RF model, which produces a single prediction value. The fundamental concept is to integrate several DTs to determine the final result instead of depending only on individual DTs. RF model stands out as the accuracy leader, combining the merits of ensemble learning with feature importance analysis. By aggregating predictions from multiple correlated trees and introducing randomness in feature selection, RF excels in minimizing both bias and variance. Due to its versatility and ability to handle complex relationships make it the model of choice for robust predictions in the context diverse and dynamic Zomato restaurant dataset. This approach not only enhances predictive accuracy but also provides insights into the key factors influencing dining costs. After splitting the data, RF model is initiated to train the data. This is done with the help of RandomForestRegressor() module of scikit-learn. The fundamental working principle of Random Forest is illustrated in Fig. 7.

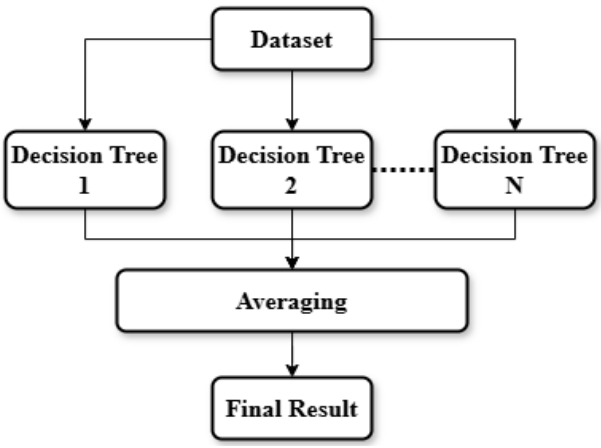


Fig. 7 Structural representation of the Random Forest

5.4. K-Nearest Neighbor (KNN)

KNN is an instance-based, supervised learning algorithm that is non-parametric. New data points are categorized according to the predominant class of their closest neighbors. The algorithm keeps track of every case that is accessible and categorizes newly discovered cases based on their feature space majority vote. The value of K represents the number of neighbors which is a crucial parameter that can significantly impact the model's performance. Graphical representation of the KNN model, showing classification into three clusters, is presented in Fig. 8.

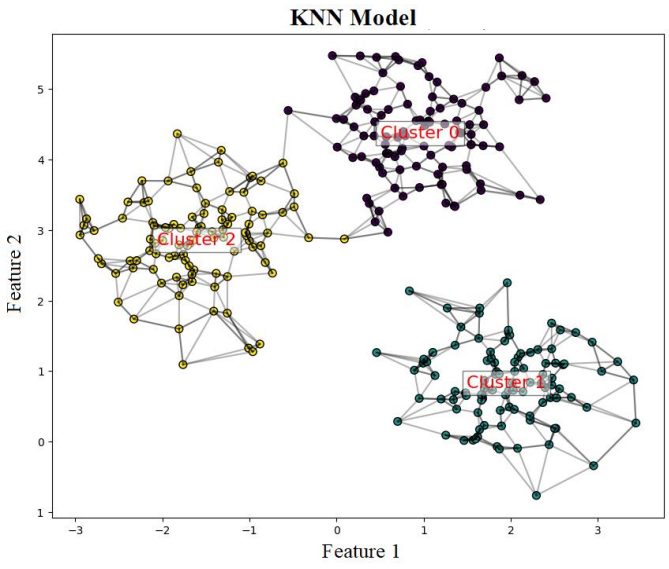


Fig. 8 Graphical representation of the KNN model

5.5. Gradient Boosting (GB)

The GB machine uses nonparametric regression and back-fittings to create predictive models. Rather of creating a single model, the GB creates an initial model and then iteratively fits additional models by minimizing the loss function in order to generate the best accurate model [14]. If you want to forecast a continuous value, like age, weight, or cost, you can apply regression with it. This is not the same as applying regression analysis. It differs slightly from the configuration utilized for classification. In GB, decision trees are employed as a weak learner. DTs convert the data into a tree representation to solve the ML problems. In tree format, a class label is shown by each leaf node and an attribute by each interior node. In general, the squared error serves as the loss function (especially for regression situations). In GB, differentiable loss function is required and the model uses same notations for the residuals as used in LR. A weak model that maps features to that residual is trained by GB regression. By adding the residual that a weak model predicted to the input of the current model, this technique gently nudges the model in the direction of the intended outcome. Performing these actions repeatedly will enhance the model's overall forecast. Fig. 9 presents the working mechanism of the Gradient Boosting algorithm. The following are the general procedures we use to apply GB Regression:

i. Pick a week learner
ii. Apply an additive framework
iii. Establish the loss function
iv. Minimize the loss function

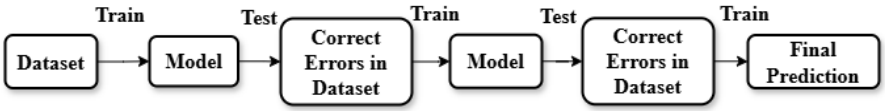


Fig. 9 Working mechanism of Gradient Boosting

5.6. Extreme Gradient Boost (XGBoost)

XGBoost is a supervised machine learning method for classification and regression. This approach, which builds on DTs, is superior than approaches like RF and GB. It uses a variety of optimization techniques and performs well with huge, complex datasets. An initial prediction is created before utilizing XGBoost to fit a training dataset. Using a similarity score for the residuals, a DT is constructed with the data. It is computed how similar the data in a leaf are, and how much more similar they become in the next split. To identify a feature and a threshold for a node, the gains are compared. The residuals are also used to calculate the output value for each leaf. The anticipated value and the observed values are used to compute residuals. Based on the observed and expected values, the residual is calculated. In classification, the log of chances and probabilities is usually used to calculate the values. The tree's output becomes the dataset's new residual, which is then utilized to build another tree. Until the residuals cease decreasing or after a predetermined number of repetitions, this process is repeated. Fig. 10 illustrates the structural framework of the XGBoost algorithm.

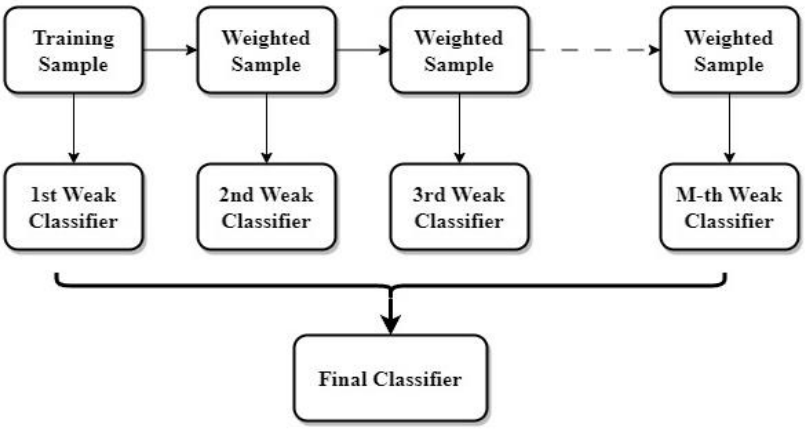


Fig. 10 Fundamental working principle of XGBoost

5.7. LASSO

Regression analysis techniques like LASSO (Least Absolute Shrinkage and Selection Operator) combine regularization with variable selection to improve predictability and interpretability [14]. It's a kind of linear regression where the coefficients of less significant features are encouraged to decline towards zero by including a penalty term in the loss function. A sparse model is the outcome, as these features are essentially eliminated from the model. Fig. 11 demonstrates the difference between LASSO and Linear Regression.

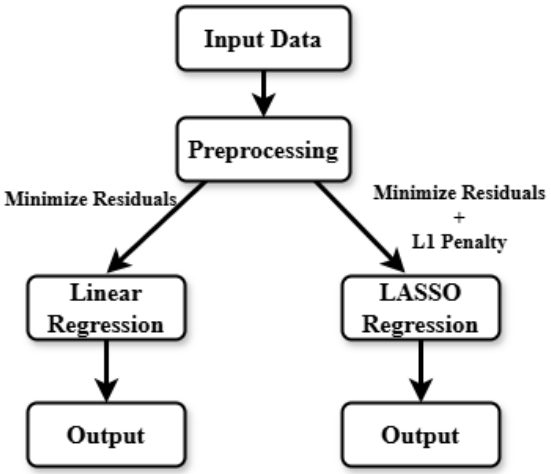


Fig. 11 LASSO vs Linear Regression

6. ML ALGORITHM RESULTS AND ANALYSIS

This section represents the results obtained by different ML algorithms and their analysis. The output metrics for evaluating ML algorithms are accuracy (R-Squared (%)), MAE, model fit time, and model prediction time [25-27]. The ML algorithms were developed and trained using Python programming language on Google Collab platform [28], which provides a cloud-based environment with GPU acceleration for faster model training. The details of the software tool are shown in Table 2.

Table 2 Details of software tool

Category	Details
Platform	Google Collab (Cloud-Based Environment)
Processor	Virtual Cloud-Based CPU/GPU
RAM	12 GB (Allocated by Google Collab)
Programming Language	Python
Libraries Used	Pandas (Data Manipulation), NumPy (Numerical Computations), Matplotlib & Seaborn (Visualization), Scikit-Learn (ML Algorithms), XGBoost (Advanced Boosting)

The performance of each ML algorithm heavily depends on the careful tuning of hyperparameters. Table 3 presents the specific hyperparameter configurations selected for each model, which were optimized to achieve the best performance in terms of accuracy, MAE, model fit time, and model prediction time. The hyperparameters were fine-tuned using grid search and cross-validation techniques to identify the best configuration for each algorithm.

Table 3 Hyperparameter configurations & values

Model	Key Hyperparameters	Optimal Values Selected
Linear Regression (LR)	None	Default
Decision Tree (DT)	Max Depth	15
	Min Samples Split	4
Random Forest (RF)	Number of Estimators	150
	Max Depth	20
	Min Samples Split	5
K-Nearest Neighbors (KNN)	Number of Neighbors (K)	7
Gradient Boosting (GB)	Learning Rate	0.05
	Number of Estimators	100
	Max Depth	4
XGBoost	Learning Rate	0.1
	Number of Estimators	150
	Max Depth	6
LASSO Regression	Alpha	0.01

The obtained numerical values of evaluating metrics for each ML algorithms are shown in Table 4. The first evaluating criterion is accuracy, defined as the closeness to the true value. Fig. 12(a) shows a bar graph of accuracy of each ML algorithms. The obtained accuracy is 63.51%, 97.71%, 97.86%, 81.31%, 79.63%, 89.48%, and 63.51% for LR, DT, RF, KNN, GB, XGBoost, and LASSO, respectively. Among these DT and RF show highest accuracy,

whereas LR and LASSO show minimal. DT algorithm show highest accuracy because it has better non-linearity handling capability, feature interactions, and offering flexibility for capturing complex relationship and complex data structures, which is why they achieve higher accuracy. However, RF regression algorithm shows standout performance, demonstrating superior predictive accuracy of 97.86% through ensemble learning and feature importance analysis. RF has the ability to mitigate over fitting, handle diverse data types, and identify influential factors makes it the optimal choice for analyzing the complex Zomato data. However, LR and LASSO show the lowest accuracy because they assume linear relationships between input features and the target variable. However, the data being analyzed, such as Zomato dataset, likely involves complex, non-linear patterns, that's the LR and LASSO are unable to capture effectively. Fig. 12(b) shows bar graph of mean absolute error in which DT and RF shows lowest value. Apart of that, the actual vs predicted graph of LR, DT, RF, KNN, GB, XGBoost and LASSO models are presented in Fig. 13 (a) to (g) respectively.

Table 4 Evaluated metric values for different ML algorithms

Algorithm Name	R-squared (Accuracy %)	Mean Absolute Error	Model Fit Time (s)	Model Prediction Time (s)
LR	63.51	199.2313	0.036169	0.006073
DT	97.71	22.89233	0.693119	0.010751
RF	**97.86**	**27.00114**	**35.41616**	**0.645274**
KNN	81.31	90.36253	0.350173	0.161622
GB	79.63	138.4035	12.22126	0.038875
XGBoost	89.48	107.3148	0.765353	0.034141
LASSO	63.51	198.7871	0.033882	0.006374

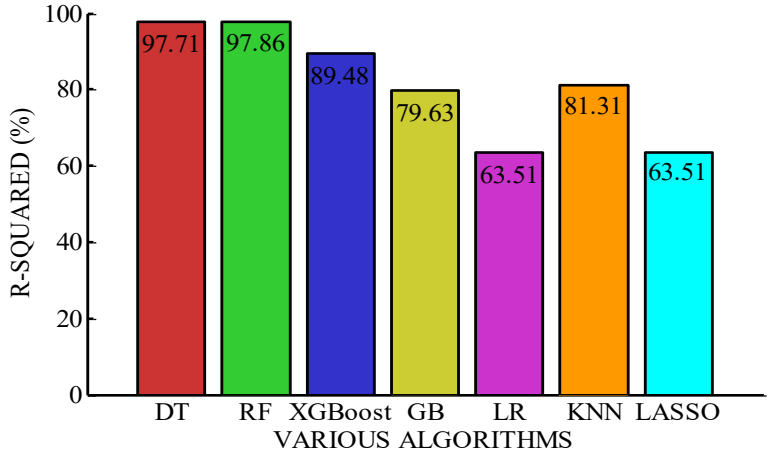
6.1. Hyperparameter Impact on Performance

DT algorithm achieved a remarkable accuracy of 97.71%, which can be attributed to the careful selection of hyperparameters, such as a maximum tree depth of 15 and a minimum sample split of 4. These parameters allowed the model to efficiently capture feature interactions and handle complex data structures present in the Zomato dataset.

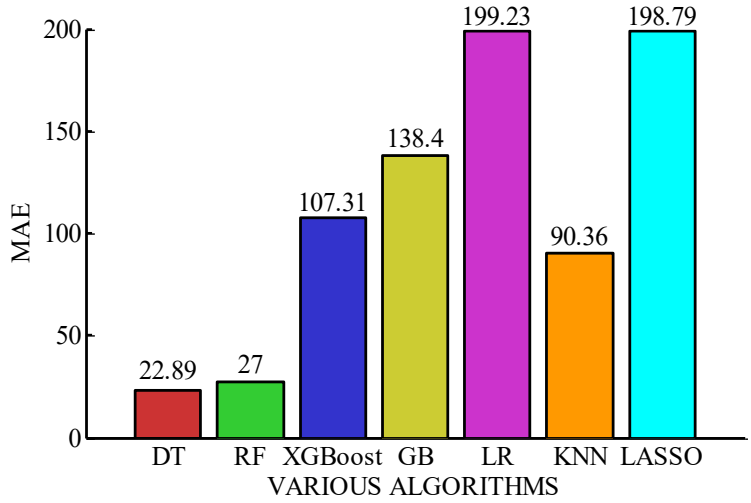
Similarly, RF algorithm achieved the highest accuracy of 97.86%. The ensemble learning technique employed by this model, combined with optimal hyperparameters, such as 150 estimators and a maximum depth of 20, ensured superior predictive performance. RF ability to mitigate overfitting and identify influential features further contributed to its standout performance.

On the other hand, simpler models like LR and LASSO exhibited lower accuracies (63.51%) due to their inability to capture the non-linear relationships inherent in the data. These models rely on linear assumptions, which do not align well with the complex patterns typical of restaurant and customer data in platforms like Zomato.

The importance of hyperparameter tuning is further demonstrated by the KNN model, which achieved an accuracy of 81.31% with an optimal neighbor value of 7. Models like GB and XGBoost achieved reasonably high accuracies of 79.63% and 89.48%, respectively, demonstrating the importance of parameters such as learning rate and depth for boosting models.

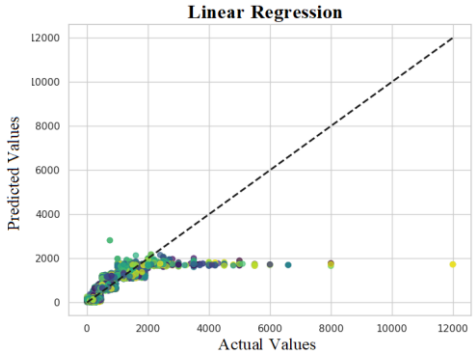


(a)

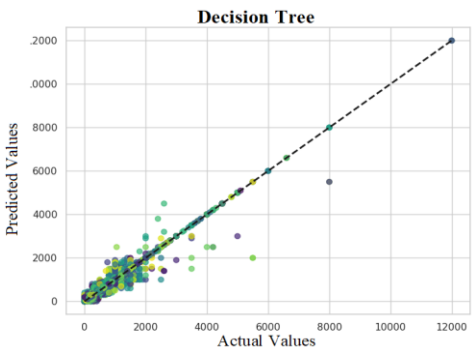


(b)

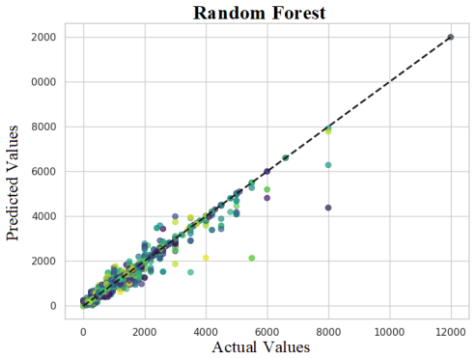
Fig. 12 Bar graph of (a) R-squared or Accuracy (%), (b) Mean Absolute Error of ML algorithms


Linear Regression
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

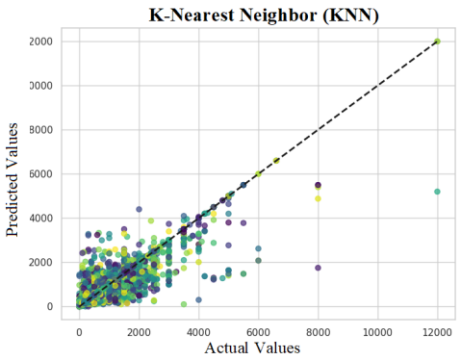
(a)


Decision Tree
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

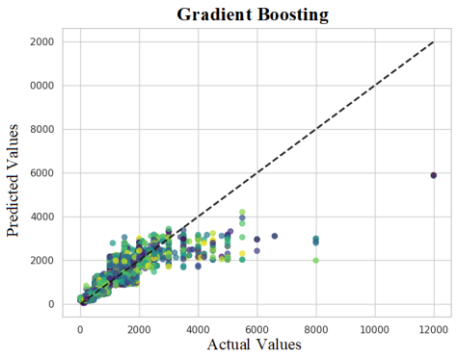
(b)


Random Forest
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

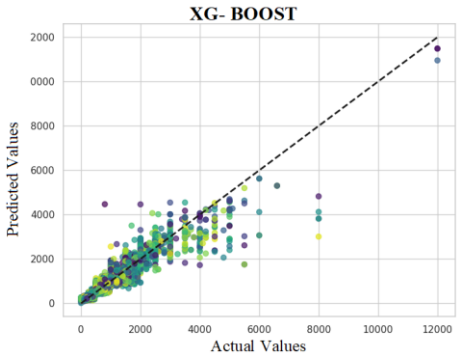
(c)


K-Nearest Neighbor (KNN)
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

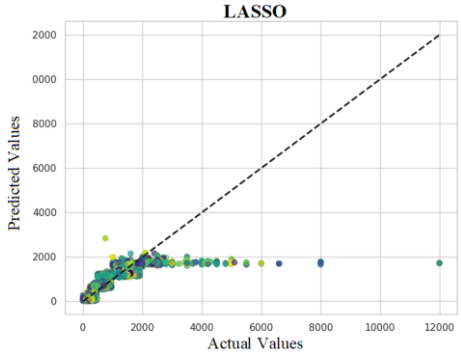
(d)


Gradient Boosting
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

(e)


XG- BOOST
Predicted Values
Actual Values
0
2000
4000
6000
8000
10000
12000

(f)



(g)

Fig. 13 Actual vs Predicted Values for (a) Linear Regression, (b) Decision Tree, (c) Random Forest, (d) KNN, (e) Gradient Boosting, (f) XGBoost, (g) LASSO

7. Conclusion

This research provides an in-depth analysis of the Zomato dataset using EDA and multiple ML algorithms. By evaluating various parameters such as restaurant types, price ranges, and ratings, the study highlights critical insights that can empower restaurant owners to take business decisions. The results demonstrate that DT and RF model significantly outperforms and achieved 97.71% and 97.86% accuracy, with minimum value of MAE. These models are particularly effective in handling non-linear relationships and complex interactions, making them ideal for the diverse and intricate data of the restaurant industries. The findings emphasize the importance of leveraging ML algorithms for restaurant data analysis, as they can reveal hidden patterns, predict customer preferences, and help optimize pricing strategies and operational efficiency. This has potential to drive sustainable growth and improve profitability for restaurants. Looking ahead, future research can explore the integration of deep learning techniques for more accurate predictions, incorporate real-time data for dynamic decision-making, and apply these models to other regions or global datasets for broader insights. Additionally, incorporating customer sentiment analysis from social media and review platforms could provide a more comprehensive view of consumer behavior, further enhancing decision-making for restaurant stakeholders.

References

[1] J. N. Bondevik, K. E. Bennin, O. Babur and C. Ersch, "A Systematic Review on Food Recommender Systems", *Expert Syst. Appl.*, vol. 238, p. 122166, 2024.

[2] D. Pandey and E. Swati, "Personalized dining experience: Leveraging machine learning for menu recommendations in foodtech application", *Int. J. Curr. Sci.*, vol. 14, no. 2, pp. 32-38, 2024.

[3] T. P. Armand, K. A. Nfor, J. I. Kim and H. C. Kim, "Applications of Artificial Intelligence, Machine Learning, and Deep Learning in Nutrition: A Systematic Review", *Nutrients*, vol. 16, p. 1073, 2024.

[4] S. S. Nidhi and R. S. Pandey, "Predicting Rating of Online Food Chain", *J. Global Res. Comput. Sci.*, vol. 14, no. 1, pp. 1-9, 2023.

[5] V. Khosla and S. Srinivasan, "Zomato co-founder Pankaj Chaddah quits as it shuffles top management", The Economic Times. Last Update: March 02, 2018, [Online]. Available at: https://economictimes.indiatimes.com/small-biz/startups/newsbuzz/zomato-co-founder-pankaj-chaddah-quits-as-it-shuffles-top-management/articleshow/63129470.cms

[6] N. Choudhary, V. Panwar, S. Mittal and G. Sahu, "Zomato Restaurants Data Analysis Using Machine Learning Algorithms", *J. Emerg. Technol. Innov. Res.*, vol. 8, no. 2, pp. 1435-1441, 2021.

[7] A. Panigrahi, A. Saha, A. Shrinet, M. Nauityal and V. Gaur, "A Case Study on Zomato-The online Foodking of India", *J. Manag. Res. Anal.*, vol. 7, no. 1, pp. 25-33, 2020.

[8] A. S. Rao, B. V. Vardhan and H. Shaik, "Role of Exploratory Data Analysis in Data Science", In Proceedings of the 6th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2021, pp. 1457-1461.

[9] IBM, *Exploratory Data Analysis*, [Online]. Available at: https://www.ibm.com/topics/exploratory-data-analysis

[10] GeekforGeeks, *What is Exploratory Data Analysis?*, 16 May 2024, [Online]. Available at: https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

[11] IBM, *Linear Regression*, [Online] Available at: https://www.ibm.com/topics/linear-regression#:~:text=Resources,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable

[12] The Click Reader, *Decision Tree Regression Explained with Implementation in Python*, Medium, 19 Oct. 2021, [Online]. Available at: https://medium.com/@theclickreader/decision-tree-regression-explained-with-implementation-in-python-1e6e48aa7a47

[13] S. A. Fitriani, Y. Astuti and I. R. Wulandari, "Least Absolute Shrinkage and Selection Operator (LASSO) and k-Nearest Neighbors (k-NN) Algorithm Analysis Based on Feature Selection for Diamond Price Prediction", In Proceedings of the International Seminar on Machine Learning, Optimization, and Data Science (ISMODE), Jakarta, Indonesia, 2022, pp. 135-139.

[14] IBM, *Lasso Regression*, 18 January 2024, [Online]. Available at: https://www.ibm.com/topics/lasso-regression

[15] Shina, S. Sharma and A. Singla, "A Study of Tree-Based Machine Learning Techniques for Restaurant Reviews". In Proceedings of the 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4.

[16] A. Kulkarni, D. Bhandari and S. Bhoite, "Restaurants Rating Prediction Using Machine Learning Algorithms", *Int. J. Comput. Appl. Technol. Res.*, vol. 8, no. 9, pp. 375-378, 2019.

[17] H. Anwar, T. Anwar and S. Murtaza, "Review on Food Quality Assessment Using Machine Learning and Electronic Nose System", *Biosens. Bioelectron.: X*, vol. 14, p. 100365, 2023.

[18] Y. Guo, A. Lu and Z. Wang, *Predicting Restaurants' Rating and Popularity Based on Yelp Dataset*, CS 229 Machine Learning Final Project.

[19] S. K. Naayak, M. Beura, M. Siddique and S. P. Mishra, "Analysis of Indian Food Based on Machine Learning Classification Models", *J. Sci. Res. Rep.*, vol. 27, no. 7, pp. 1-7, 2021.

[20] S. Chaudhary, A. Sharma and M. Dhankar, "Indian Food Recognition Using CNN", *Int. J. Res. Publ. Rev.*, vol. 5, no. 5, pp. 5092-5097, May 2024.

[21] R. N. Patil, Y. P. Singh, S. A. Rawandale and S. Singh, "Improving Sentiment Classification on Restaurant Reviews Using Deep Learning Models", In Proceedings of International Conference on Machine Learning and Engineering (ICMLDE 2023), Procedia Computer Science, 2024, vol. 235, pp. 3246-3256.

[22] B. Shah, P. Kanani, P. Joshi, G. Pandya, D. Kulkarni, N. Patil and L. Kurup, "Traditional Indian Food Classification Using Shallow Convolutional Neural Network", *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, pp. 769-774, 2024.

[23] R. Agarwal, T. Choudhury, N. J. Ahuja and T. Sarkar, "InidanFoodNet: Detecting Indian Food Items Using Deep Learning", *Int. J. Comput. Methods Exp. Meas.*, vol. 11, no. 4, pp. 221-232, Dec 2023.

[24] S. Somashekar and S. Mallesh, "Restaurant Rating Prediction Using Regression", In Proceedings of the 5th International Conference on Electronics, Communication and Aerospace Technology, Coimbatore, India, 2021, pp. 1139-1144.

[25] R. Jain, V. V. Thakare, and P. K. Singhal, "Enhancing Circular Microstrip Patch Antenna Performance Using Machine Learning Models", *FU: Elec. Energ.*, vol. 36, no. 4, pp. 589-600, 2023.

[26] R. Jain, V. V. Thakare and P. K. Singhal, "Design and Comparative Analysis of THz Antenna through Machine Learning for 6G Connectivity", *IEEE Lat. Am. Trans.*, vol. 22, no. 2, pp. 82-91, 2024.

[27] R. Jain, V. V. Thakare and P. K. Singhal, "Employing Machine Learning Models to Predict Return Loss Precisely in 5G Antenna", *Prog. Electromagn. Res. M*, vol. 118, pp. 151-161, 2023.

[28] Google, *Google Colaboratory*, [Online]. Avaliable at: colab.research.google.com. https://research.google.com/colaboratory/