International Journal of Artificial Intelligence,
ISSN 0974-0635; Int. J. Artif. Intell.
2025, Volume 23, Issue Number 1

Unveiling Latent Structures: Personalized Restaurant Recommendations via Machine Learning

Mohammad Khaki[1,2] and **Fereshte Dehghani**[3*]

[1] Department of Computer Engineering, University of Kashan, Iran;
[2] School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran;
E-mail: khaki_m@comp.iust.ac.ir

[3] Department of Computer Engineering, University of Kashan, Iran;
[*] Corresponding Author E-mail: fdehghani@kashanu.ac.ir

ABSTRACT

Social media provides a wealth of information for decision-making, especially for finding restaurants in unfamiliar areas. This study leverages Zomato restaurant data to unveil latent structures and hidden patterns within the data that influence customer ratings. By incorporating machine learning, we create a personalized recommendation system to guide users towards their ideal dining experience. To overcome the limitations of raw data, we employ a multi-step approach. Clustering algorithms unveil hidden patterns (latent structures) in user preferences and restaurant attributes. These structures improve the accuracy of our classification models, addressing the challenge of complex relationships between seemingly unrelated data points.
Our findings highlight the effectiveness of the Random Forest classification algorithm. Applied after the multi-step approach that unveils latent structures, it achieves a remarkable 90% accuracy rate. This success is demonstrably linked to uncovering hidden patterns in user preferences and restaurant attributes through clustering. These structures allow the Random Forest model to make more precise classifications, ultimately leading to a superior recommendation system. Notably, most errors involve misclassifications between similar restaurant types, which is acceptable in this context due to the inherent overlap in user preferences for these categories.

Keywords: Machine Learning, Random Forest, Zomato Restaurants Dataset, Customers Rating Prediction, Latent Pattern Recognition

Mathematics Subject Classification: 68T99

Computing Classification System: 10010147.10010257.10010321.10010336, , 10010147.10010257.10010258.10010260.10003697

1. INTRODUCTION

The rapidly growing online food scene, fueled by the rise of web-based applications and user-friendly smartphones, has witnessed a surge in online orders. This shift in consumer behavior necessitates businesses providing an online experience analogous to in-person interactions. Recommendation systems are vital in this digital landscape, personalize user experiences, and drive business success by making data-driven suggestions about various products (Konstan & Riedl, 2012).

One such prominent online service is food preparation (Dirsehan & Cankat, 2021), with restaurant classification becoming increasingly important for both users and businesses leveraging the Internet to expand their reach (AL-Bakri et al., 2021). The COVID-19 pandemic further accelerated this trend, with

a large segment of the population choosing online food delivery (Limsarun et al., 2021). While marketing research emphasizes the importance of studying social influence on food applications (Limsarun et al., 2021), this paper focuses on a different aspect: improving recommender system effectiveness. These factors extend beyond the core criteria of price, quality, and delivery services and encompass aspects like service speed, food portion size, lighting, and even the restaurant's atmosphere (Liu & Tse, 2018).

An analysis of Zomato's restaurant dataset reveals that user experience is primarily driven by food quality, regardless of additional services like delivery (Abdelhamied, 2011; Jalilvand et al., 2017). However, the presence of delivery can significantly impact the overall perception. Thus, this research presents an innovative method to predict user aggregate ratings by considering traditional factors and leveraging latent structures within restaurant features.

This paper proposes a novel approach to restaurant recommendation that leverages latent structure discovery. We move beyond explicit user ratings and discover hidden patterns influencing user satisfaction. Our method utilizes clustering algorithms to uncover these latent structures within restaurant features like food quality, price, and delivery service. This method enables us to capture the intricate relationships between these features and predict user ratings for new restaurants, even those lacking substantial reviews.

By incorporating latent structures, our method aims to improve the accuracy of restaurant recommendations significantly. This advancement has the potential to benefit both users by providing more personalized and relevant suggestions and businesses by enhancing their discoverability and customer satisfaction.

The subsequent parts of this article are structured as follows: Section 2 reviews existing recommender system techniques for restaurant classification. In Section 3, the proposed classification method is thoroughly examined. This method involves several crucial steps: data preprocessing to ensure high quality, uncovering latent structures through clustering to reveal hidden patterns, oversampling to correct class imbalances, feature selection to enhance model efficiency, and the utilization of machine learning methodologies for accurate predictions. Section 4 evaluates the approach's effectiveness, and Section 5 concludes the paper.

1.1. Related work

Several studies have explored factors influencing user satisfaction with online food delivery services and restaurant classification for recommendation purposes. Table 1 highlights a range of approaches used in restaurant recommendation systems, detailing the research question, methodology, dataset, and key findings of each study. Studies have explored various datasets, including user questionnaires (Cha & Seo, 2020; Haghighi et al., 2012), online review platforms (Banerjee & Chowdhury, 2021; Kumar et al., 2018), and restaurant listing websites with user ratings (AL-Bakri et al., 2021; Dixit et al., 2022; Priya, 2020; Zhang et al., 2011). This variety in data sources reflects the complexity of user preferences and the evolving nature of the online food industry. For example, a study in Korea (Cha & Seo, 2020) found that mobility and reliability are critical for user satisfaction with delivery apps, suggesting developers should prioritize these factors over providing more information.

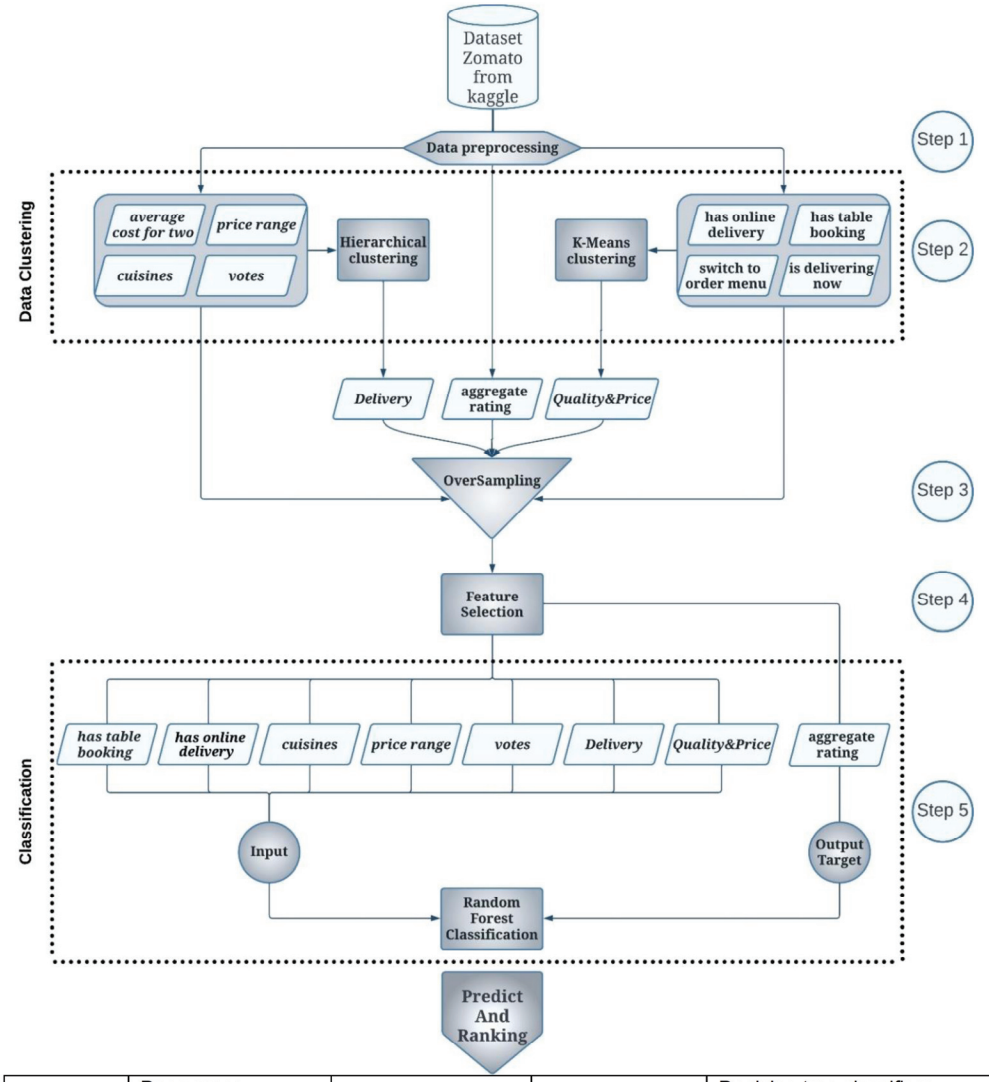
Shifting the focus to restaurant classification, prior work using predefined features and machine learning techniques has been explored. For example, a multi-label classification with Support Vector Machines (SVM) was employed to categorize restaurants based on features like "delivery service" and "classy ambiance" (AL-Bakri et al., 2021). This method achieved promising accuracy for restaurants without online delivery (AL-Bakri et al., 2021). Other studies explored various machine learning models for

restaurant classification, highlighting the importance of the chosen model's relationship with the features used (Dixit et al., 2022; Priya, 2020; Zhang et al., 2011). Additionally, research has examined user behavior analysis for recommendations (Luo & Xu, 2019).

However, a limitation of these existing works, as evident in Table 1, is their reliance on predefined features or user ratings. This approach neglects the underlying factors that truly shape user preferences. This article aims to fill this gap by introducing a novel method that leverages latent structures within restaurant features to improve recommendation accuracy.

Table 1. An Overview of Related Works

Ref No	*Research Focus*	*Methodology*	*Dataset*	*Key Finding*
(AL-Bakri et al., 2021)	Can predefined features be used to classify restaurants for recommendations?	Multi-label classification and Support vector machine (SVM)	Zomato	Identified "delivery service" and "classy ambiance" as classifiable features (accuracy ~88% for prediction)
(Cha & Seo, 2020)	What factors influence user satisfaction with online food delivery applications?	Structural Equation Modeling (SEM)	Questionnaire in Korea	Users prioritize mobility and reliability over extensive information within delivery applications.
(Haghighi et al., 2012)	What factors most influence customer satisfaction in restaurants?	Structural Equation Modeling (SEM)	Questionnaire in Iran	Food quality, service, ambiance, and perceived price fairness are essential for restaurant loyalty.
(Kumar et al., 2018)	Do methods to detect fake reviews improve the fairness and reliability of ratings?	Rev2	OTC, Alpha, Amazon, Flipkart, and Epinions	Rev2 system identifies fraudulent reviewers by analyzing review fairness, reliability, and product quality.
(Zhang et al., 2011)	Which classification model best predicts restaurant categories based on available data?	Naive Bayes, Support Vector Machine (SVM)	Open Rice	Machine learning model accuracy for Cantonese restaurant reviews depends on both the model and the type of features used (e.g., character pairs vs. single characters).
(Luo & Xu, 2019)	Can user location and behavior be used to categorize restaurant customers?	Recommender System	Yelp	Analyzing user reviews by sentiment towards different aspects can help identify helpful reviews.
(Dixit et al., 2022)	What classification system helps restaurant owners understand their customers for business growth?	XG-Boost	Zomato in India	An XGBoost model achieved high accuracy (98%) in understanding customer preferences, suggesting its value for restaurant optimization.
(Priya, 2020)	Is random forest regression the most effective technique for predicting restaurant customer behavior?	Random Forest	Zomato in India	Random forest regression may be suitable for predicting restaurant customer behavior based on a successful prior study.
(Banerjee & Chowdhury, 2021)	How can we develop a method to predict the impact of business issues on reviews while keeping the analysis efficient?	Random Forest	Zomato in India, And Yelp	Reviews from geographically diverse areas show a consistent relationship with user-based parameters using machine learning.



(Sharma & Singla, 2018)	Does cross-validation help pick the best classifier for restaurant service type?	Decision Tree	Zomato	Decision tree classifiers outperform random forest classifiers for restaurant service parameter classification (63.5% vs 56% accuracy).

2. THEORETICAL MODEL AND METHODOLOGY

Figure 1. Proposed framework

This section details the proposed framework for restaurant recommendation utilizing latent feature structures. To enhance classification performance, the framework Figure 1 leverages data clustering for latent structure discovery combined with feature selection. This process is achieved through five key steps: data preprocessing, clustering, oversampling, feature selection, and classification.

2.1. Data preprocessing - Step 1

The initial step involves preparing the Zomato dataset (MEHTA, 2017) for subsequent classification tasks. This dataset is a rich resource of information on restaurants, encompasses various features of numerous restaurants, and serves as the basis for the subsequent classification process. A subset of pertinent features is chosen based on their potential influence on user ratings. These features are then preprocessed to ensure compatibility with the chosen classification algorithms. Table 2 presents a comprehensive overview of the selected features, including their names, descriptions, and types. These selected features are subjected to several essential preprocessing steps:

1. **Unification of Units:** The 'Average cost for two' feature is standardized (e.g., USD) to ensure consistency across currency units.
2. **Encoding Categorical Features:** String values in the "Cuisines" feature are transformed using integer encoding. This technique assigns a unique integer value to each distinct cuisine type.
3. **Rounding and Scaling Numerical Features:** The "Aggregate rating" feature, representing user ratings, is rounded to integer values between 1 and 5. Specific criteria are applied to maintain the semantic meaning of the ratings:
   - If the Aggregate rating is greater than or equal to 4.8, it is assigned "5."
   - If the Aggregate rating falls between 3.3 and 4.8 (inclusive), it is assigned "4."
   - If the Aggregate rating falls between 2.3 and 3.3 (exclusive), it is assigned "3."
   - If the Aggregate rating falls between 1.3 and 2.3 (exclusive), it is assigned "2."
   - If the Aggregate rating is less than 1.3, it is assigned "1."
4. **Normalization:** Min-max normalization is applied to other numerical data to ensure a common scale (Equation (1)).

$$X_{std} = \frac{x - x.min}{x.\max - x.min} * (x.\max - \mathrm{x}.\min) + \mathrm{x}.min \qquad (1)$$

These preprocessing steps ensure that the selected features are appropriately formatted and ready for subsequent classification tasks.

Table 2. Description of Selected Features in the Zomato Dataset

Clustering method	Decision label	Feature name	Description	Data Type
Hierarchical	Delivery	**Has online delivery**	Online orders available or not	binary
		Has table booking	Is an online booking table available or not	binary
		Is delivering now	Ability to send food	binary
		Switch to order menu	Availability of food menu	binary
K-means	Quality and Price	**Cuisines**	Food diversity	string
		Average cost for two	The average cost of a meal for two people	number
		Price range	Range of food Price	number
		Votes	Number of comments	number
Response label		**Aggregate rating**	Average user votes out of five	decimal number
Auxiliary convert feature		**Currency**	Currency	string

2.2. Data Clustering - Step2

This step utilizes unsupervised learning techniques like k-means and hierarchical clustering to uncover hidden structures within the data. These algorithms group data points with similar characteristics, revealing potential classes or relationships that might be missed in raw data. This process improves classification by guiding algorithm selection and parameter tuning and providing new features based on the data's structure (Chawla et al., 2002).

This step utilizes unsupervised learning techniques like k-means and hierarchical clustering to uncover hidden structures within the data. These algorithms group data points with similar characteristics, revealing potential classes or relationships that might be missed in raw data. This process improves classification by guiding algorithm selection and parameter tuning and providing new features based on the data's structure.

This approach is grounded in a set of fundamental principles:

1. **Grouping Training Examples into Clusters:** The first step involves organizing the training examples into clusters based on inherent patterns within the data.
2. **Encoding Clusters as New Features:** Next, these clusters are encoded as new features, which serve as valuable input for the classification model.
3. **Utilizing the Trained Model:** Predictions are generated using the model that has been trained with these augmented features.

The underlying assumption of this step is that the classes in the dataset inherently reflect its natural groupings. However, this may not always be true. Therefore, it is worth investigating whether clustering can improve the classification process by uncovering the dataset's inherent structure. Ideally, if the structures discovered through clustering perfectly align with those indicated by the classes, classification becomes straightforward. Otherwise, the information about the "hidden" structure revealed by clustering can enhance the overall accuracy of classification models (Piernik & Morzy, 2021).

As depicted in Figure 1, the features are categorized into two groups based on the type of input data (AL-Bakri et al., 2021). The outcomes of the clustering processes, specifically achieved through hierarchical and K-means algorithms, result in two distinct clusters labeled "Delivery" and "Quality and Price," respectively. These resultant clusters are subsequently employed as novel attributes for classification purposes.

- The "Delivery" feature, representing the delivery quality of a restaurant, is characterized by five discrete values.
- The "Quality and Price" feature, encapsulating the cost-quality dynamics of food, spans a spectrum of thirteen discrete values.

The clustering results and the original features are combined to form the input for feature selection algorithms, a crucial step aimed at further improving classification results (Piernik & Morzy, 2021). A detailed exploration of the Hierarchical and K-means clustering approaches is provided in the subsequent section.

Hierarchical clustering approach:

Hierarchical clustering encompasses two primary types: agglomerative and divisive. In this study, a divisive hierarchical clustering approach is employed in a top-down fashion, with the root node initially encompassing the entire dataset. The hierarchy is constructed by recursively dividing each node,

denoted as X, into two child nodes, X1 and X2, where X = X1 ∪ X2 and X1 ∩ X2 = ∅ (Wei et al., 2019). This process continues until individual data elements are represented as singletons (Nielsen, 2016).

As illustrated in , given that the parameters related to "Delivery" are binary, values are assigned based on their presence or absence. The construction of this hierarchical tree is conducted manually, guided by correlation coefficients with the "Aggregate Rating" attribute. Consequently, the attribute "Has Table Booking" assumes the root position within the tree. The assignment of values to the tree's leaves is as follows:

1. If all variables are zero, the leaf corresponds to class one.
2. If "Has Table Booking" and "Has Online Delivery" are zero, and either "Delivering Now" or "Switch to Menu" is one, the leaf represents class two.
3. If "Has Table Booking," "Has Online Delivery," and "Delivering Now" are zero, one, and zero, respectively, the leaf corresponds to class two.
4. If "Has Table Booking," "Has Online Delivery," and "Delivering Now" are zero, one, and one, respectively, the leaf represents class three.
5. If "Has Table Booking," "Has Online Delivery," and "Delivering Now" are one, zero, and zero, respectively, the leaf corresponds to class three.
6. If "Has Table Booking," "Has Online Delivery," and "Delivering Now" are one, zero, and one, respectively, the leaf represents class four.
7. If all "Has Table Booking" and "Has Online Delivery" are one, and either "Delivering Now" or "Switch to Menu" is zero, the leaf corresponds to class four.
8. If all variables are one, the leaf represents class five.

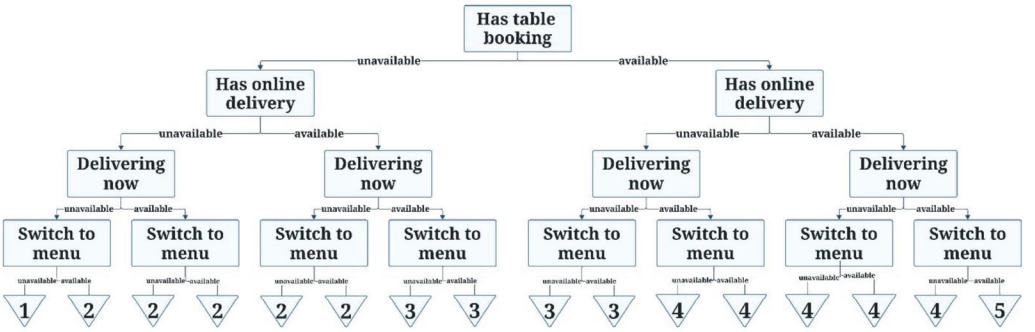


Figure 2. Hierarchical Clustering of "Delivery" Class

K-means clustering approach:

The K-means algorithm is a commonly employed clustering technique in the field of data mining. In K-means clustering, a set of n observations is partitioned into k distinct clusters, with each observation allocated to the cluster whose mean is closest to it (Moore, 2001). The significance of K-means lies in its dual utility for data analysts: firstly, it facilitates the identification of underlying simple patterns and effective handling of multi-class datasets (Yu et al., 2015), and secondly, it helps find the most suitable number of clusters, thus assisting in the final configuration of the clustering process.

In this study, the ELBOW method (Thorndike, 1953) is utilized to establish the optimal number of clusters, ultimately determining it to be 13 for the dataset. The ELBOW method, a common heuristic in mathematical optimization, involves an iterative calculation of cluster costs. It starts with an initial value of K=2 and increments it by one at each subsequent step. The cost experiences a significant reduction until a specific critical threshold of K is reached. Beyond this point, the cost stabilizes, maintaining a

relatively constant level. This threshold represents the most suitable cluster quantity, where further cluster augmentation would yield no significant improvements to data modeling.

Moreover, the "Quality and Price" feature with 13 possible distinct values and distances between each sample and all cluster centers is computed, resulting in the addition of 13 new features to the dataset for training. The results indicate that these new features have no impact and also cause additional processing time.

2.3. Oversampling - Step 3

Imbalanced datasets, where certain classes are underrepresented, can negatively impact the performance of classification models. To address this challenge, this step employs the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002). SMOTE increases the representation of underrepresented classes by generating synthetic data samples based on existing ones.

Figure 3 illustrates the impact of SMOTE on the user ratings dataset (Douzas & Bacao, 2018). Figure 3(A) shows the original imbalanced distribution, where class 4 (rating of 4) is dominant. Figure 3(B) depicts the dataset after oversampling using SMOTE. Here, the representation of underrepresented classes (like class 2) is increased, resulting in a more balanced distribution that can improve model performance during classification.

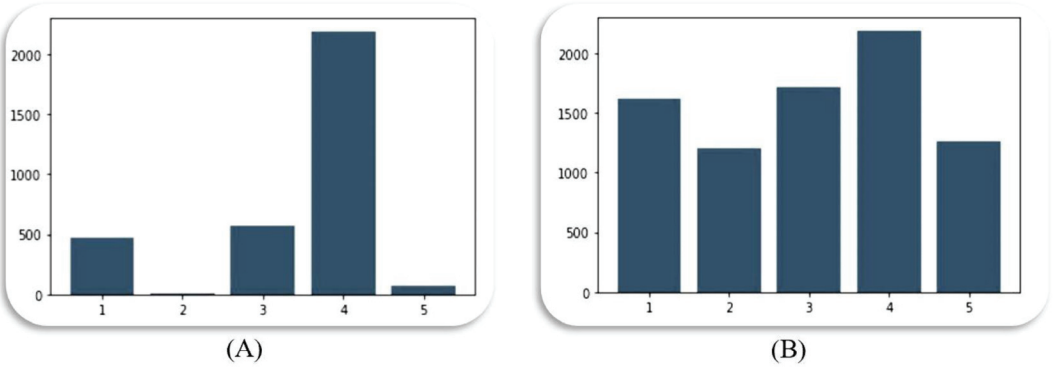


Figure 3. User Rating Distribution Before(A) and After(B) Oversampling with SMOTE

2.4. Feature Selection - Step 4

To pinpoint the most informative features for classification, this step employs the Sequential Feature Selection (SFS) algorithm (Liu, 2010; Solorio-Fernández et al., 2020). SFS iteratively adds features to the model based on their correlation with the existing set, ultimately selecting the most relevant ones for classification performance.

Initially, all features generated or used previously (including "Delivery" and "Quality & Price") are considered. SFS starts by including "Quality & Price" and "Delivery." In each subsequent step, it adds the feature that most improves the model's performance. This process continues until adding more features no longer benefits the model (Figure 4). As a result, SFS selects a subset of seven features out of the initial ten: "Has online delivery," "Has table booking," "Cuisine," "Price range," "Votes," "Delivery," and "Quality & Price."

Initially, all features generated or used previously (including 'Delivery,' 'Quality & Price,' 'Has online delivery,' 'Has table booking,' 'Is delivering now,' 'Switch to order menu,' 'Cuisines,' 'Average cost for two,' 'Price range,' and 'Vote' (AL-Bakri et al., 2021)) are considered. SFS SFS starts by including "Quality & Price" and "Delivery." In each subsequent step, it adds the feature that most improves the model's performance. This process continues until adding more features no longer benefits the model, as evident from the performance curve depicted in Figure 4. As a result, SFS selects a subset of seven features out of the initial ten variables, namely 'Has online delivery,' 'Has table booking,' 'Cuisine,' 'Price range,' 'Votes,' 'Delivery,' and 'Quality & Price.'

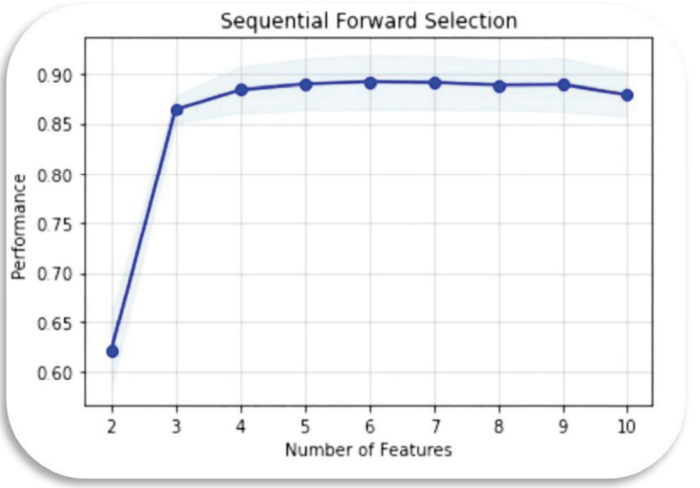


Figure 4. Feature Selection Performance

2.5. Classification - Step 5

This step utilizes various classification algorithms to predict restaurant classifications based on user ratings. The algorithms employed include Random Forest (Ho, 1995), Support Vector Machine (SVM) with different kernels (Cortes & Vapnik, 1995), Least Squares SVM (LS-SVM) (Suykens & Vandewalle, 1999), N-nearest Neighbors (Cover & Hart, 1967), Decision Tree (Wu et al., 2008), Naive Bayes (Hastie, Tibshirani, Friedman, & Friedman, 2009), and Gradient Boosting (Hastie, Tibshirani, Friedman, Hastie, et al., 2009). The classification algorithms utilize the newly created features ("Delivery" and "Quality & Price") alongside the features selected by the SFS algorithm (Section 2.4). Additionally, the "Aggregate Rating" serves as the class label for each data point.

A detailed evaluation of these algorithms is presented in the following section. However, it is important to highlight that Random Forest emerged as the best performer for restaurant classification, particularly concerning user ratings. This section can be further expanded upon after the evaluation to explain the specific advantages of Random Forest in this context (Breiman, 2001).

3. RESULTS

This analysis leverages a dataset of restaurant reviews obtained from the Kaggle platform (e.g., Zomato data), initially containing approximately 3,324 unique instances after removing duplicates. We employed a combination of feature selection and oversampling techniques to improve classification performance

for predicting user ratings. However, before exploring these techniques, we established a robust evaluation framework using F1-score, Cohen's Kappa Coefficient (Cohen, 1960), and ROC AUC. F1-score balances precision and recall for overall effectiveness, while Kappa considers chance agreement for a more robust measure (ranging from -1 to 1). Accuracy alone can be misleading for multi-class classification tasks because it does not account for chance agreement, especially with imbalanced data (Grandini et al., 2020). Cohen's Kappa offers a more robust solution by measuring the agreement beyond random chance, providing a clearer picture of a model's ability to distinguish between the various classes. ROC AUC indicates how well a model can distinguish between positive and negative categories. Subsequently, we employed data augmentation alongside feature selection. Oversampling, a technique that increases dataset size by creating new data points, significantly improved the F1-score. For instance, after dataset augmentation, the Random Forest algorithm's F1-score increased from 0.56 to 0.91.

In order to determine the most appropriate classification method for our framework, we conducted a comparative analysis of various options. Their model's performance was evaluated using the kappa score, accuracy, precision, and F1-score. The results are detailed in Table 3, with the best scores highlighted in bold. Based on this evaluation, Random Forest's effectiveness was confirmed as it outperformed other algorithms. Consequently, Random Forest will be the primary classifier used in the subsequent analysis.

Table 3. performance Comparison of classification algorithms

Supervised Classification Method	***Kappa Score***	***Accuracy***	***Precision***	***F1_score***
Random Forest	**87%**	**90%**	**91%**	**91%**
Gradient Boosting	**86%**	**88%**	**90%**	**89%**
Decision Tree	**83%**	86%	**88%**	85%
K Nearest Neighbors	70%	76%	76%	76%
LS SVM (Kernel: Gaussian)	57%	64%	60%	60%
SVM (Kernel: RBF)	55%	64%	65%	64%
SVM (Kernel: Linear)	40%	52%	54%	51%
Naive Bayes	45%	55%	64%	54%

Next, we assessed the generalizability and performance of the Random Forest model. While Clustering-augmented data may have influenced the topical scoring of restaurants, our primary focus here is classification accuracy.

The Random Forest model achieves a mean score of 88% across a 10-fold cross-validation, indicating good generalizability and avoiding overfitting the training data. To further analyze the model's performance breakdown, we will examine the confusion matrix presented in Table 4.

Table 4. Confusion Matrix for Random Forests Classification

Actual Values	Class 1	541	0	0	0	0
	Class 2	0	433	1	5	0
	Class 3	0	22	467	100	8
	Class 4	0	19	45	693	8
	Class 5	0	4	5	52	402
		Class 1	Class 2	Class 3	Class 4	Class 5
		Predicted Values				
Accuracy	**90%**					
F1-Score	**91.05%**					

The study reports a promising machine learning model based on a test with 2797 data points. The model achieved 90% accuracy, suggesting it effectively learned the underlying data patterns and can potentially generalize well to unseen data. Table 4 likely shows a confusion matrix, offering a more detailed model performance breakdown. Ideally, the diagonal of this matrix would have high values, indicating a large volume of accurate forecasts for each class. The text highlights that most errors involved the model picking a class adjacent to the correct one. This "almost right" scenario suggests the model understands the data well but might struggle to differentiate similar classes.

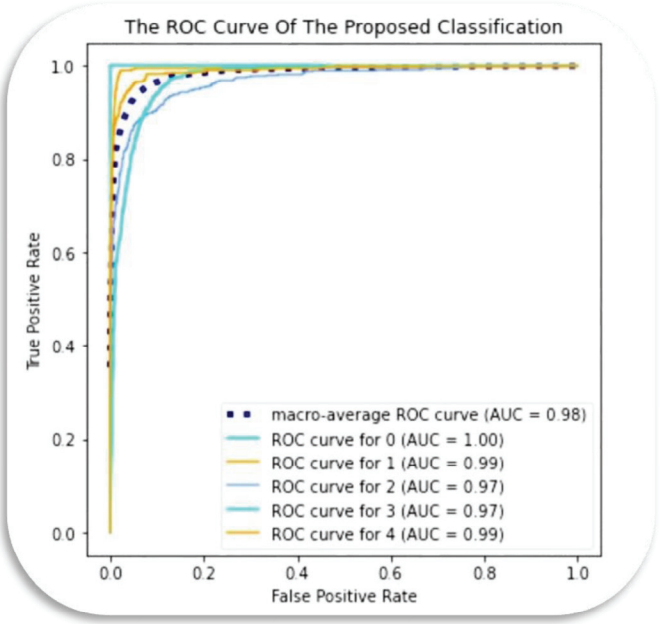
Figure 5 shows the model's performance when classifying data points near class boundaries. It reveals that many detection errors occur when the model predicts a data point as one of its adjacent classes. This can be readily understood considering the inherent similarity between neighboring classes, especially in the context of user recommendations.

However, a crucial aspect to consider is that these errors might not be entirely detrimental. Given the close relationship between the classes, if a data point is classified as a neighboring class, it still reflects a degree of correctness. This observation is reflected in the high accuracy improvement of approximately 98.8%. Overall, Figure 5 offers valuable insights into The model's capacity to differentiate between closely related classes, even when encountering occasional misclassifications that might be considered partially correct due to the inherent similarity between neighboring classes.

Figure 5. Impact of Near-Class Boundaries on Model Performance

Figure 6 showcases the model's performance using a Receiver Operating Characteristic (ROC) curve. The Area Under the Curve (AUC) is a key indicator of model effectiveness, is approximately 0.99 in this case. This high AUC signifies the model's strength in differentiating between positive and negative classes, demonstrating its potential for accurate prediction.

Figure 6. ROC Curve of the Novel Classification Model



4. DISCUSSION AND CONCLUSION

This research leverages machine learning techniques to extract latent patterns within user data, achieving superior accuracy in predicting restaurant rankings. Unlike prior studies (e.g., (AL-Bakri et al., 2021)), a key strength of this work lies in its ability to forecast user ratings for newly established restaurants. This capability is facilitated by uncovering latent data structures, or underlying patterns, within user ratings.

By employing K-means and hierarchical clustering algorithms, the data undergoes preprocessing and augmentation, revealing these latent structures. These structures improve the model's generalization capabilities, enabling it to adapt to new data and predict user preferences for unrated restaurants.

The proposed model's success hinges on the core principle of exploiting latent structures. This principle extends beyond predicting ratings for new establishments. Our approach delves deeper than typical ambiance evaluation by additionally predicting food delivery scores. Furthermore, it analyzes how user ratings in non-delivery restaurants, representing a distinct latent structure, influence the predicted outcomes. This multifaceted evaluation demonstrates the model's capacity to handle diverse user preferences.

Finally, the study incorporates restaurants from a global perspective, achieving superior ranking results compared to research confined to specific locations (e.g., (Moore, 2001; Nielsen, 2016; Piernik & Morzy, 2021)). This global scope underscores the generalizability of the findings, further solidifying the effectiveness of uncovering latent structures in user rating data.

This work offers valuable insights for both diners and restaurant owners. Diners can utilize the predicted rankings to make informed decisions, while restaurant owners gain a deeper understanding of customer preferences across various locations.

Future research directions include incorporating additional parameters like service speed and food volume to enhance model performance. Additionally, exploring how cultural variations influence dining experiences across different locations could provide even more comprehensive insights.

5. REFERENCES

Abdelhamied, H. H. (2011). Customers' perceptions of floating restaurants in Egypt. *Anatolia–An International Journal of Tourism and Hospitality Research*, *22*(01), 1-15.

AL-Bakri, N. F., Al-zubidi, A. F., Alnajjar, A. B., & Qahtan, E. (2021). Multi label restaurant classification using support vector machine. *Periodicals of Engineering and Natural Sciences*, *9*(2), 774-783.

Banerjee, A., & Chowdhury, T. (2021). Reviewing system using exploratory data analysis and ensemble machine learning algorithms. 2021 IEEE 2nd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET),

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cha, S.-S., & Seo, B.-K. (2020). The effect of food delivery application on Customer Loyalty in Restaurant. *Journal of Distribution Science*, *18*(4), 5-12.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, *20*(1), 37-46.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

Dirsehan, T., & Cankat, E. (2021). Role of mobile food-ordering applications in developing restaurants' brand satisfaction and loyalty in the pandemic period. *Journal of Retailing and Consumer Services*, *62*, 102608.

Dixit, A. K., Nair, R. R., & Babu, T. (2022). Analysis and Classification of Restaurants Based on Rating with XGBoost Model. 2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT),

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, *91*, 464-471.

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Haghighi, M., Dorosti, A., Rahnama, A., & Hoseinpour, A. (2012). Evaluation of factors affecting customer loyalty in the restaurant industry. *African Journal of Business Management*, *6*(14), 5039-5046.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Boosting and additive trees. *The elements of statistical learning: data mining, inference, and prediction*, 337-387.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Ho, T. K. (1995). Random decision forests. Proceedings of 3rd international conference on document analysis and recognition,

Jalilvand, M. R., Salimipour, S., Elyasi, M., & Mohammadi, M. (2017). Factors influencing word of mouth behaviour in the restaurant industry. *Marketing Intelligence & Planning*.

Konstan, J. A., & Riedl, J. (2012). Recommender systems: from algorithms to user experience. *User modeling and user-adapted interaction*, *22*(1), 101-123.

Kumar, S., Hooi, B., Makhija, D., Kumar, M., Faloutsos, C., & Subrahmanian, V. (2018). Rev2: Fraudulent user prediction in rating platforms. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining,

Limsarun, T., Navavongsathian, A., Vongchavalitkul, B., & Damrongpong, N. (2021). Factors Affecting Consumer's Loyalty in Food Delivery Application Service in Thailand. *The Journal of Asian Finance, Economics, and Business*, *8*(2), 1025-1032.

Liu, H. (2010). Feature Selection. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 402-406). Springer US. https://doi.org/10.1007/978-0-387-30164-8_306

Liu, P., & Tse, E. C.-Y. (2018). Exploring factors on customers' restaurant choice: an analysis of restaurant attributes. *British Food Journal*.

Luo, Y., & Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. *Sustainability*, *11*(19), 5254.

MEHTA, S. (2017). *Zomato Restaurants Data*. Kaggle. https://www.kaggle.com/datasets/shrutimehta/zomato-restaurants-data

Moore, A. (2001). K-means and Hierarchical Clustering. In: USA.

Nielsen, F. (2016). Hierarchical clustering. In *Introduction to HPC with MPI for Data Science* (pp. 195-211). Springer.

Piernik, M., & Morzy, T. (2021). A study on using data clustering for feature extraction to improve the quality of classification. *Knowledge and Information Systems*, *63*(7), 1771-1805.

Priya, J. (2020). Predicting restaurant rating using machine learning and comparison of regression models. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE),

Sharma, S., & Singla, A. (2018). A study of tree based machine learning techniques for restaurant reviews. 2018 4th International Conference on Computing Communication and Automation (ICCCA),

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2020). A review of unsupervised feature selection methods. *Artificial Intelligence Review*, *53*(2), 907-948.

Suykens, J. A., & Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, *9*, 293-300.

Thorndike, R. L. (1953). Who belongs in the family. Psychometrika,

Wei, W., Liang, J., Guo, X., Song, P., & Sun, Y. (2019). Hierarchical division clustering framework for categorical data. *Neurocomputing*, *341*, 118-134.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., & Yu, P. S. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*, 1-37.

Yu, Z., Wang, Q., Fan, Y., Dai, H., & Qiu, M. (2015). An improved classifier chain algorithm for multi-label classification of big data analysis. 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems,

Zhang, Z., Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications*, *38*(6), 7674-7682.