# **1. Problem Definition**

The Zomato dataset provides a wealth of information about various restaurants worldwide. This dataset is a rich source for multiple analyses, including but not limited to predicting the price range of a restaurant, understanding customer preferences, and identifying factors that influence restaurant ratings. The diversity and depth of the dataset make it an excellent candidate for machine learning applications, which can provide valuable insights for both consumers and businesses.

In this project, our primary objective is to predict two crucial aspects of dining out using the Zomato dataset: the price range category of a restaurant and the average price for two people at a restaurant. The dataset includes various features such as restaurant details, location, cuisine, and pricing information, all of which can contribute to these predictions. This task can be divided into two main parts, each representing a different type of machine learning problem:

1. **Classification Problem**: The first task is to predict the price range category of a restaurant. The price range is a categorical variable that classifies restaurants into distinct pricing tiers (e.g., low, medium, high, very high). This classification can help users quickly identify which restaurants fit within their budget, making the decision-making process more efficient. For businesses, understanding what factors contribute to a restaurant's classification can inform pricing strategies and competitive positioning.
2. **Regression Problem**: The second task is to predict the exact average cost for two people at a restaurant. Unlike the price range, which categorizes restaurants into broad groups, the average cost for two is a continuous variable that provides a specific monetary value. Predicting this value can help users get a precise estimate of their potential dining expenses. For restaurant owners and managers, this prediction can help in pricing optimization, menu adjustments, and marketing strategies.

These predictions are valuable for several reasons. For consumers, they provide a way to make informed dining choices based on budgetary constraints and preferences. Users can filter and select restaurants that match their financial expectations without having to manually check each restaurant's pricing information. This enhances the user experience by saving time and effort.

For businesses, the ability to predict price range and average cost can offer strategic advantages. Restaurants can benchmark their pricing against competitors and adjust their prices to better meet market demand. They can also tailor their marketing campaigns to target specific customer segments more effectively. For instance, a restaurant that knows it is perceived as high-end can focus on marketing to affluent customers or highlight luxury aspects of their service.

Furthermore, understanding the factors that influence price range and average cost can provide deeper insights into market dynamics. For example, analyzing the impact of location, cuisine type, and customer reviews on pricing can reveal trends and preferences in different regions or demographics. This information can be crucial for new restaurant openings, menu changes, or service improvements.

Overall, this project aims to leverage the Zomato dataset to build predictive models that serve both consumers and businesses. By accurately predicting the price range and average cost for two, we can enhance user satisfaction and provide actionable insights for restaurant management. The combination of classification and regression techniques allows us to tackle different aspects of the pricing problem, making our approach comprehensive and robust.

In summary, our dual focus on classification and regression within the Zomato dataset not only aims to provide immediate practical benefits to users and businesses but also to contribute to the broader understanding of the restaurant industry's pricing strategies and consumer behavior. The insights gained from this analysis can drive smarter business decisions, better user experiences, and more effective market positioning in the highly competitive restaurant industry.

# **2. Data Analysis**

The Zomato dataset contains a wealth of information across various columns. Key features include restaurant name, country, city, address, localities, and cuisines offered. Additionally, it provides the average cost for two people, the currency used, and whether the restaurant offers table booking or online delivery services. The dataset also includes ratings, reviews, and more nuanced data points, which are essential for comprehensive analysis. Complementing this, the country code dataset provides a vital mapping between country codes and country names, enabling us to enrich the Zomato dataset with geographic context and facilitate deeper insights into the global distribution of restaurants and their pricing strategies.

**Key Columns**:

* Restaurant ID: Unique identifier for each restaurant.
* Restaurant Name: Name of the restaurant.
* Country Code: Country code where the restaurant is located.
* City: City where the restaurant is located.
* Address: Full address of the restaurant.
* Locality: Locality of the restaurant.
* Cuisines: Types of cuisines served.
* Average Cost for two: Average cost for two people dining.
* Currency: Currency used.
* Has Table booking: Indicates if the restaurant has table booking.
* Has online delivery: Indicates if the restaurant has online delivery.
* Aggregate rating: Overall rating of the restaurant.
* Rating color: Color representing the rating.
* Rating text: Text description of the rating.
* Votes: Number of votes received.

# **3. EDA Concluding Remarks**

Through exploratory data analysis (EDA), we gain insights into the dataset's structure and underlying patterns. Key findings from the EDA include:

* **Country Distribution**: The Zomato dataset provides a comprehensive overview of restaurants across multiple countries, with a predominant representation from India, accounting for 90.6% of the entries. This extensive representation underscores the significant presence and popularity of Zomato in the Indian market, making it a rich source of data for analyzing restaurant trends and consumer behavior in the region. The dataset also includes entries from other countries, though they constitute a smaller proportion, allowing for a diverse yet India-centric analysis.
* **Average Cost for Two**: the dataset reveals a wide range of values, reflecting the varying economic conditions and dining cultures across different countries with a predominant representation from Indonesia. In some countries, dining out is relatively inexpensive, whereas in others, it can be a more costly affair. This variability in average costs is indicative of different pricing strategies adopted by restaurants based on local economic factors and consumer purchasing power.
* **Ratings**: Most restaurants in the Zomato dataset exhibit good to excellent ratings, reflecting overall positive customer experiences. A significant portion of restaurants falls into the higher rating categories, indicating high levels of customer satisfaction. Fewer restaurants are found in the lower rating categories, suggesting that poorly rated restaurants are less common.
* **Online Delivery and Table Booking**: The availability of online delivery and table booking services varies significantly across different countries and regions. These services can impact both customer ratings and the overall cost of dining. The dataset reveals that about 34.5% of restaurants offer online delivery, making it a relatively common feature that caters to the convenience of customers who prefer dining at home. However, table booking is less prevalent, with fewer people opting for this service.
* **Price Range**: Most restaurants in the dataset fall into the lowest price category (Category 1), indicating that the majority of dining options are budget-friendly. This could be a reflection of the market demand for affordable dining experiences. Restaurants in this price range are likely to attract a broader customer base, including students, families, and individuals looking for value-for-money meals. The predominance of lower-priced restaurants also suggests a competitive market where affordability is a key factor in attracting customers.

EDA helps identify missing values, understand the distribution of numerical variables, and discover relationships between categorical variables and the target variable (price range).

# **4. Pre-processing Pipeline**

Before building machine learning models, we need to preprocess the dataset to ensure it is clean and suitable for analysis. The preprocessing steps include:

1. **Merging Datasets**: To enhance our dataset, we merge the Zomato dataset with the country code dataset. This step incorporates additional geographical information by linking country codes to their corresponding country names. This enriched dataset provides more context and can improve the predictive power of our models.
2. **Dropping Unnecessary Columns**: Not all columns in the dataset are relevant to our prediction tasks. For instance, Restaurant ID and Address do not provide meaningful information for predicting price range or average cost for two. Removing these irrelevant columns helps reduce noise and computational complexity, leading to more efficient and accurate models.
3. **Encoding Categorical Variables**: Machine learning algorithms require numerical input, so categorical variables must be converted into a numerical format. One common method is one-hot encoding, where each category is represented by a binary column. For example, the Country column, originally containing country names, would be transformed into multiple columns, each representing a country with binary values indicating the presence of that country.
4. **Feature Scaling**: Features with varying scales can adversely affect model performance, especially in algorithms like k-nearest neighbors or gradient descent-based methods. Normalizing or standardizing numerical features ensures that they are on a similar scale. Normalization scales the data to a range of [0, 1], while standardization centers the data to have a mean of 0 and a standard deviation of 1.

# **5. Building Machine Learning Models**

With the preprocessed data, we can now build and train machine learning models for our tasks.

**Classification Task: Predicting Price Range**

The first task is to predict the price range category of a restaurant, which is a classification problem. To achieve this, I built and evaluated multiple models, including:

* **Logistic Regression**: A straightforward model for binary and multi-class classification problems.
* **Decision Tree Classifier**: A tree-based model that splits the data into subsets based on feature values.
* **Random Forest Classifier**: An ensemble of decision trees that improves predictive performance by averaging multiple models.
* **Gradient Boosting Classifier**: Another ensemble method that builds trees sequentially to correct errors of previous models.
* **Support Vector Machine (SVM)**: A model that finds the hyperplane which best separates the classes.
* **MLP Classifier**: A neural network-based model with multiple layers.

After evaluating these models, the Decision Tree Classifier emerged as the best performer with an accuracy score of 0.9787610619469026. This high accuracy indicates that the model effectively captures the relationship between the features and the price range categories, making it suitable for our classification task.

**Regression Task: Predicting Average Cost for Two**

The second task is to predict the exact average cost for two people at a restaurant, which is a regression problem. For this, I built and tested several models, including:

* **Linear Regression**: A simple model that assumes a linear relationship between features and the target.
* **Ridge Regression**: A regularized version of linear regression that prevents overfitting by adding a penalty term.
* **Lasso Regression**: Another regularized model that can shrink coefficients of less important features to zero.
* **Decision Tree Regressor**: A tree-based model for regression tasks.
* **Random Forest Regressor**: An ensemble method that averages multiple decision trees to improve predictive accuracy.
* **Gradient Boosting Regressor**: An ensemble method that builds models sequentially, where each new model corrects errors from the previous ones.
* **Support Vector Regressor (SVR)**: A regression model based on SVM that tries to fit the best line within a certain threshold.
* **KNN Regressor**: A non-parametric method that predicts the target based on the k-nearest neighbors in the feature space.

Among these, the Gradient Boosting Regressor achieved the lowest Mean Squared Error (MSE) and the highest R² score, indicating superior performance in predicting the average cost for two people. This model's ability to sequentially improve and correct errors makes it highly effective for capturing complex relationships in the data.

# **6. Concluding Remarks**

In this article, we walked through the process of analyzing the Zomato dataset, starting from problem definition to building machine learning models. Through exploratory data analysis (EDA), we uncovered key insights about the dataset, such as the distribution of restaurants across countries, the range of average costs, and the prevalent cuisines. We then preprocessed the data to handle missing values, encode categorical variables, and normalize numerical features. Afterward, we built various machine learning models to predict the price range category of restaurants and the exact average cost for two people at a restaurant. Additionally, we performed hyper parameter tuning to optimize the model's performance.

### Key Takeaways:

* **Rich Information**: The Zomato dataset offers extensive details useful for various predictive tasks.
* **Importance of EDA and Preprocessing**: Proper exploratory data analysis and preprocessing are essential for understanding the data and preparing it for modeling.
* **Hyper parameter Tuning**: This step is crucial for refining models to achieve the best performance.

By following these steps, you can develop robust machine learning models that provide valuable insights and predictions based on the Zomato dataset, enabling better decision-making for both users and businesses.