Prediction of THE Price range & AVERAGE COST FOR TWO AT RESTAURANTS ASSOCIATED WITH ZOMATO Using machine learning

**Solanki Acharya**

02.11.2022

­­­­­

**Abstract:**

Many customers visit a restaurant based on food critics and reviews on websites such as Zomato.com. Restaurants strive at the initial stages of opening but their demand deteriorates after the initial hype. Further business for these restaurants is largely based on their reviews. What can the restaurant do to make its ratings better? Food taste is an obvious trigger to improve the ratings of a restaurant, but there are other factors that improve the ratings of a restaurant. We have used the data sets on Zomato Restaurants which were available to the public and on: <https://raw.githubusercontent.com//dsrscientist//dataset4//main//zomato.csv> & <https://github.com/dsrscientist/dataset4/blob/main/Country-Code.xlsx> . After combining these two the joint data set contain the restaurant’s ID, City, Country, the cuisine offered, Price Range, Average Cost for two, Currency, Aggregate rating, and some other relevant data. Zomato Data Analysis is one of the most useful analysis for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines. This paper aims a creating a prediction model for the Price Range and Average Cost for two and analyze them. Additionally, this analysis caters to the needs of people who are striving to get the best cuisine in the country and which locality of that country serves that cuisine with the maximum number of restaurants.

1. **Introduction and Problem Definition**

Zomato is a restaurant search and delivery network which provides food services for millions of users every month. Zomato lets users search restaurants, get recommendations, and add reviews, photos, and such. It also provides a glance overview of restaurants. There is online ordering in almost all serviceable areas. In the past few years, it was noticed that serval restaurants were comparatively creating an enormous amount of profit margin with the help of collaborating with Zomato and also based on the internet and techniques introduced in order to provide a clearer effective approach to deal with ongoing aspects. The essential goal is to oversee the machine calculations by following the day-by-day deals of different restaurants. This whole scenario likewise associates with the general methodologies dependent on appropriate extractions from Machine Learning Algorithms.

Machine learning sheds light on various domains unexplored by human analyses. It provides a viewpoint that is not visible in general. The prediction and classification models which took scientists decades to create are achieved in days. Data when available in huge amounts can be studied through machine learning algorithms to arrive at meaningful information.

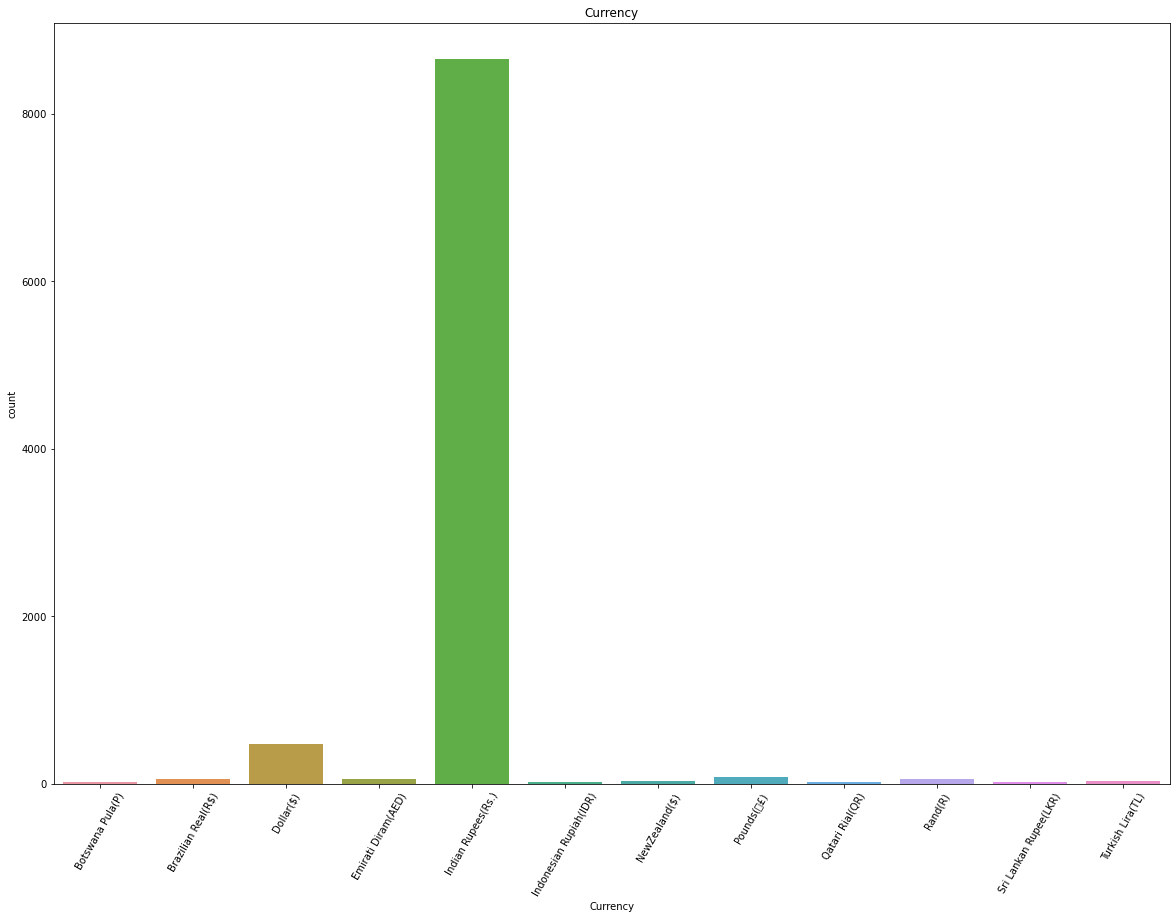
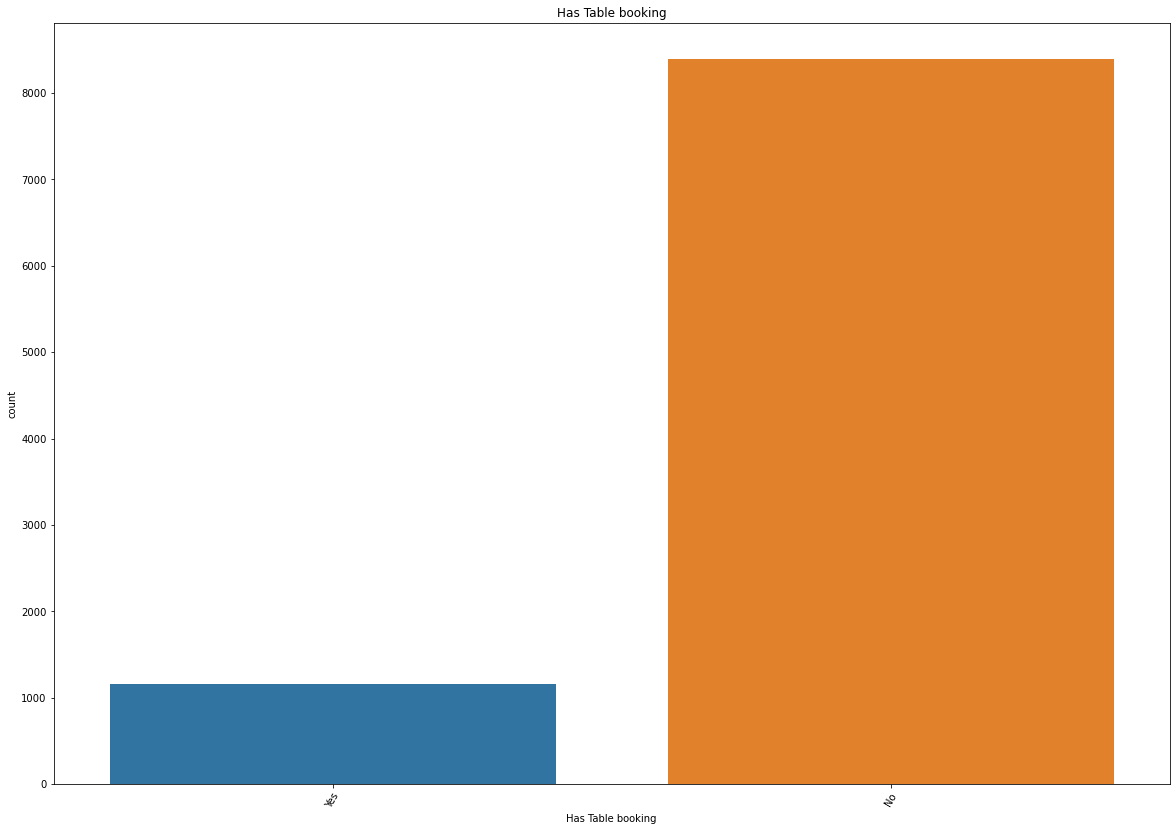
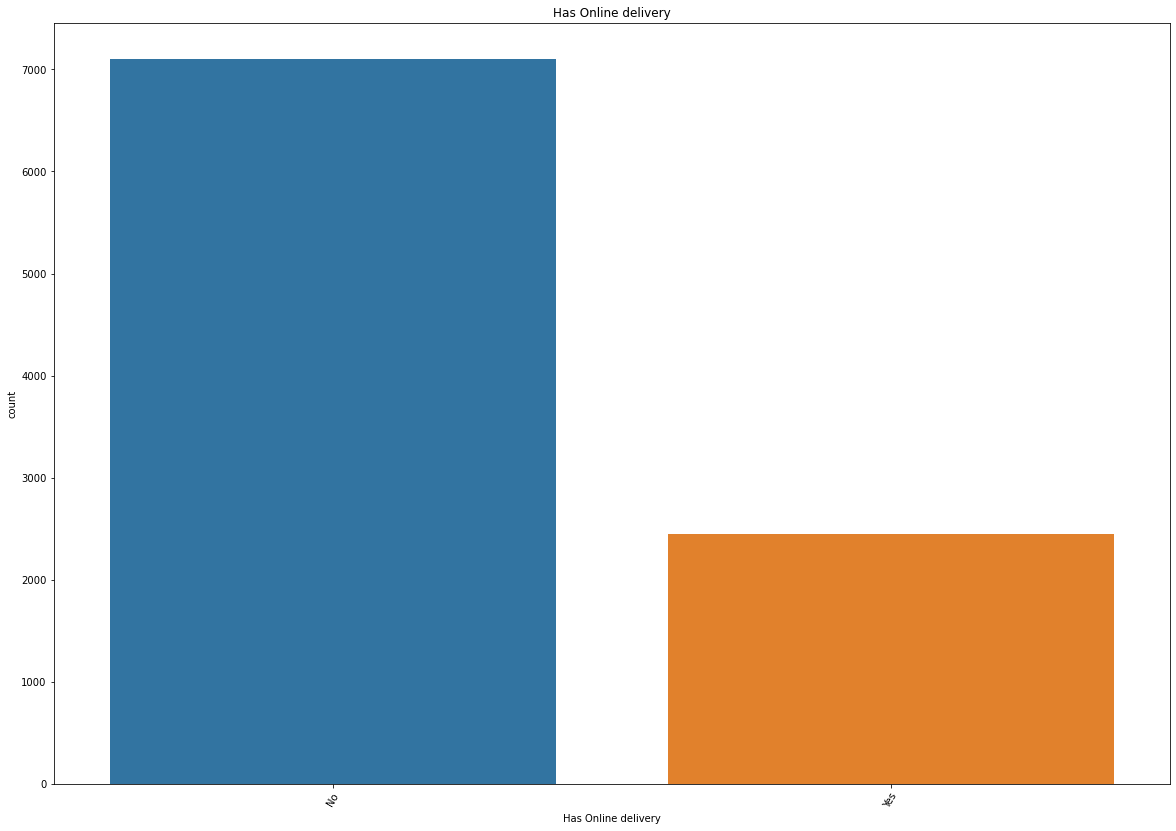
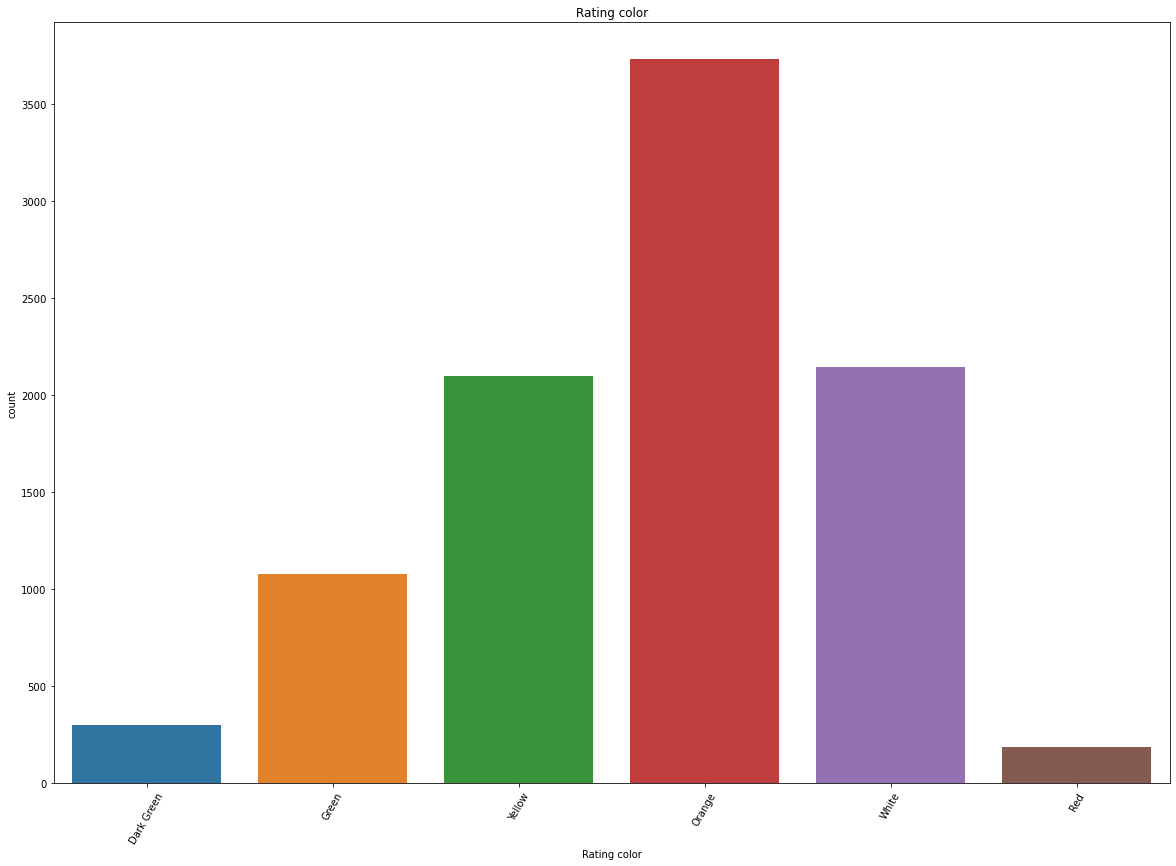
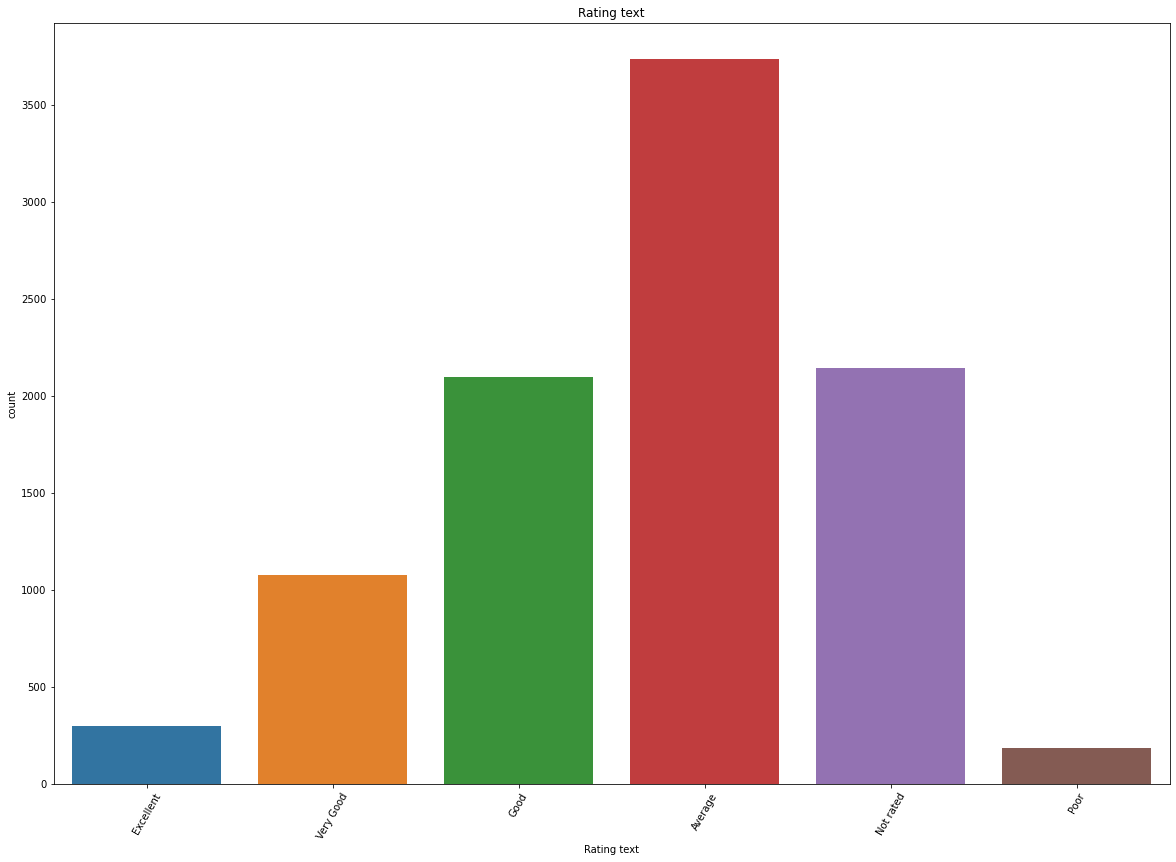
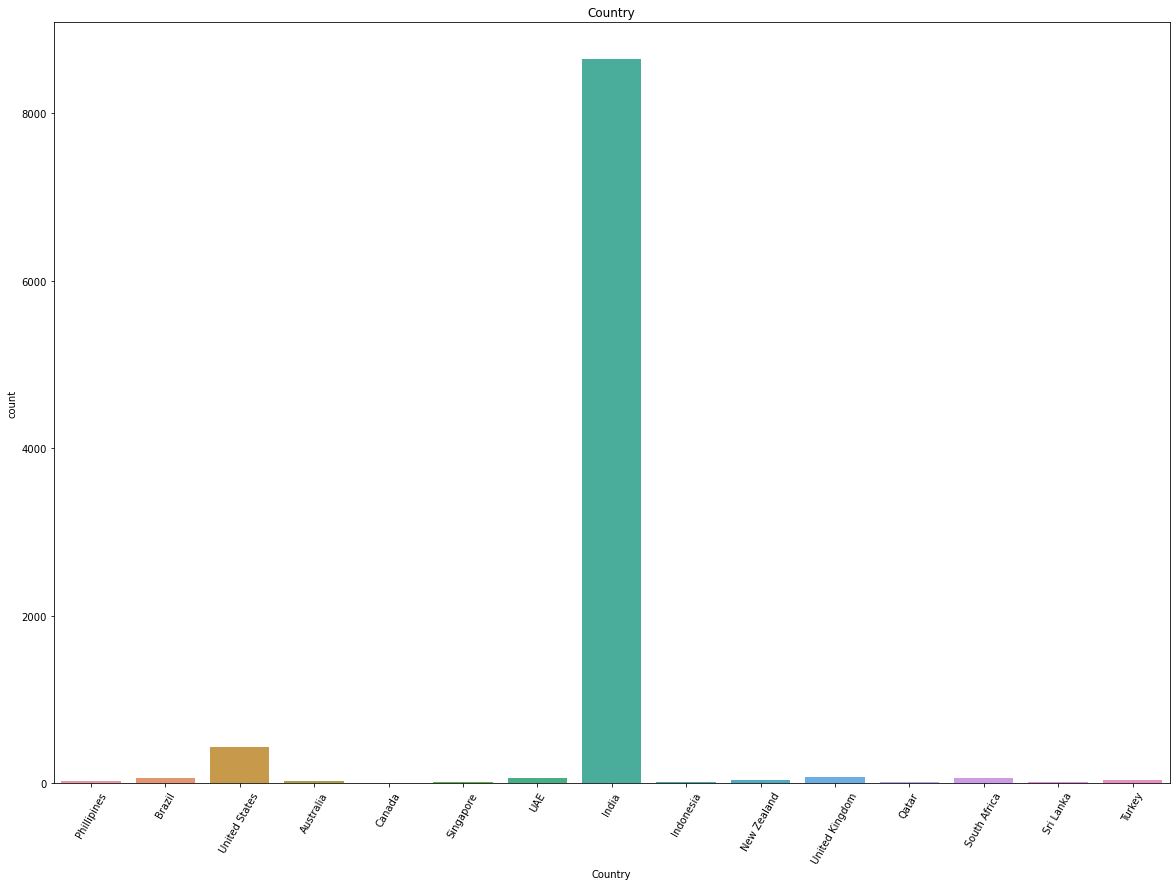
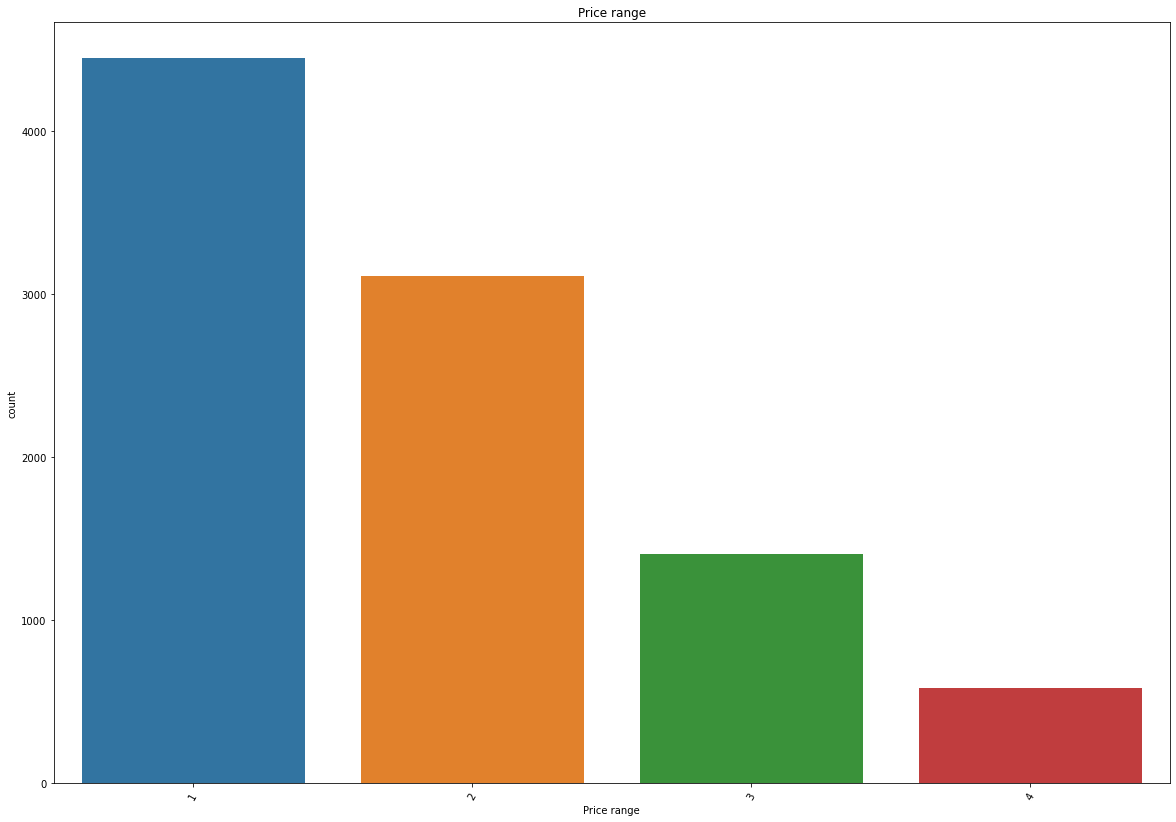
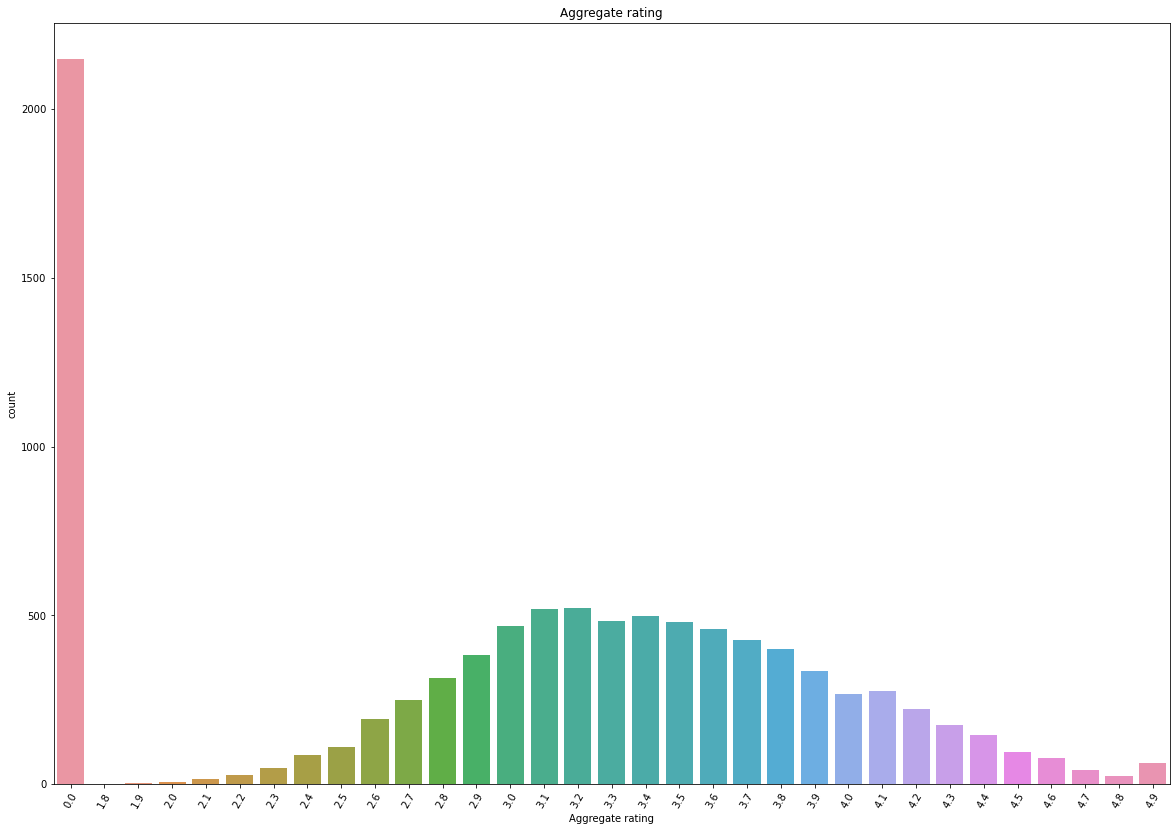
This is an exploratory and Predictive Data Analysis of the Zomato Dataset using Python. Here, we have looked into the data, visualize it, and made some conclusions. Ultimately, we have gone for the prediction of the price range and average cost for two persons, which were by nature classification and regression problems respectively. Ultimately, we have gone for prediction, checked the performances of our models, and at the last, we have discussed the results and made conclusions.

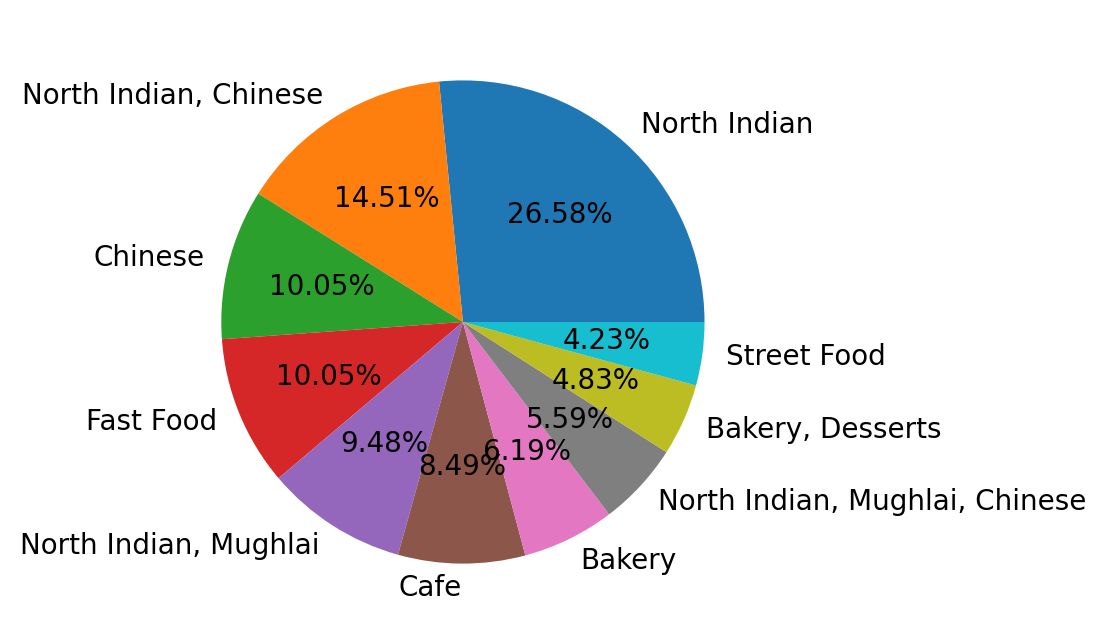
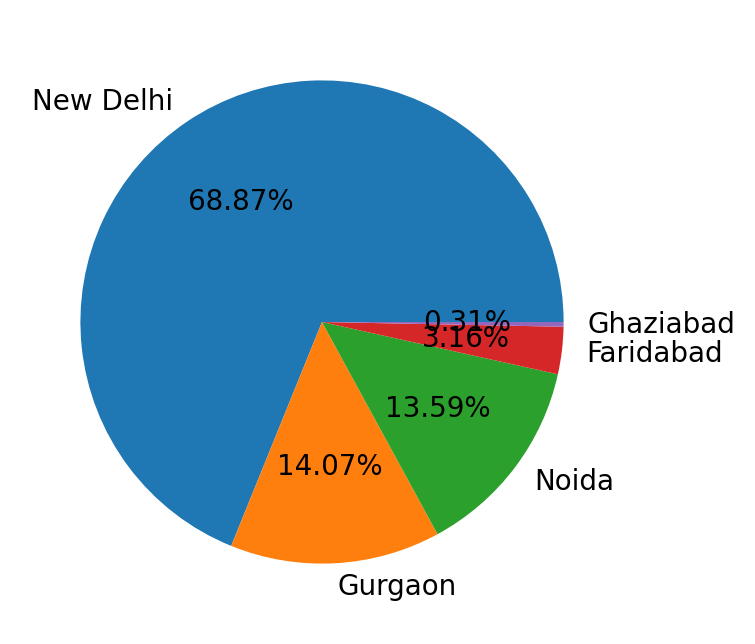
1. **Data Analysis and EDA**
   1. **DATA PRE-PROCESSING:**

Data pre-processing can be data cleaning or data transformation. The use of data pre-processing to improve the classification, regression, or prediction of the machine learning models is necessary to identify the features that would enable higher accuracy in classification. Classification and Regression accuracy are predominantly dependent on the proper representation of data. Correlation-based feature selection is used to reduce the number of features. This includes checking and replacing the NaN values, sorting out the relevant and irrelevant columns that are not dependent on the target columns, removing the skewness and outliers of the features, for the classification part class imbalance and for the regression part variance inflation factor checking, etc.

* 1. **Data Visualization:**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Through plots, we can predict some points directly although approximately. In our project, we have used many plots to predict some basic but useful predictions. We first tried to visualize our data with plots and graphs. Through this, we were able to make quick decisions and some basic predictions. We then checked for the correlation values of the variables using Python, and from the correlation table, we came to know about the correlations among the variables. Some of the important graphs are following: -

****

****

1. **EDA Concluding Remark**

Observations made from the data visualization are:

1. Indian city has the maximum number of Zomato restaurants. Zomato has its presence in 23 countries but the most important country is India and U.S.A comes next.

2. New Delhi has the highest number of restaurants associated with Zomato with a count of more than 5000. Gurgaon and Noida are behind New Delhi with a count of more than 1000 restaurants associated with Zomato.

3. Restaurants providing only North-Indian cuisines are the highest in number. Restaurants providing both Chinese and North Indian and restaurants providing only Chinese are behind the restaurants providing both North Indian.

4. Online deliveries are available in India and UAE.

5. Not rated count is quite high and the max no. of rating is between 2.5 to 3.4

6. Restaurants with ratings of 3.3 are the highest in number. Near about 30 restaurants are unrated. Maybe they are new. There are no restaurants having ratings less than 3, which may show that Zomato doesn’t collaborate with restaurants having ratings less than 3.

* Finally, we can actually conclude that Zomato is getting a large profit from India, which means the major business of Zomato is happening in India.

1. **Pre-Processing Pipeline**

* Removing Skewness:

In statistics, skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution in a given set of data. Skewness essentially measures the symmetry of the distribution.

A data transformation may be used to reduce skewness. A distribution that is symmetric or nearly so is often easier to handle and interpret than a skewed distribution. More specifically, a normal or Gaussian distribution is often regarded as ideal as it is assumed by many statistical methods. If a feature contains Skewness then there will be an asymmetry in the statistical distribution of that feature, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

In our project the features named 'Latitude', 'Longitude', 'Aggregate rating', 'Country Code', 'Votes', and 'Average Cost for two' contained skewness. So, we used Power Transformation to reduce the skewness.

* Checking Outliers:

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. The box plot is a useful graphical display for describing the behaviour of the data in the middle as well as at the ends of the distributions. The box plot uses the [median](https://www.itl.nist.gov/div898/handbook/eda/section3/eda351.htm) and the lower and upper quartiles (defined as the 25th and 75th [percentiles](https://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm)). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

In our project the features named 'Country Code', 'Longitude', 'Latitude', 'Average Cost for two', 'Price range', 'Aggregate rating', and 'Votes' contained outliers and we tried to reduce as much as possible by using the z-score method, but then our data loss percentage was 10.323526332321222% which is more than 10%, so we did not perform outlier removal.

* Handling Class Imbalance:

Our first problem was to predict the Price Range, which was an imbalanced classification-type machine-learning problem having four discrete outcomes. Classification predictive modelling involves predicting a class label for a given observation.

An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The distribution can vary from a slight bias to a severe imbalance where there is one example in the minority class for hundreds, thousands, or millions of examples in the majority class or classes.

Imbalanced classifications pose a challenge for predictive modelling as most of the machine learning algorithms used for classification were designed around the assumption of an equal number of examples for each class. This results in models that have poor predictive performance, specifically for the minority class. This is a problem because typically, the minority class is more important and therefore the problem is more sensitive to classification errors for the minority class than the majority class.

One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model. An improvement in duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective. The most widely used approach to synthesizing new examples is called the Synthetic Minority Oversampling Technique or SMOTE for short.

In the classification part of our project, we used this technique to solve the problem of class imbalance.

* Variance Inflation Factor:

Our second problem was to predict the Average cost for two people, which was a multiple-regression-type machine-learning problem having continuous outcomes. In this second part of our project, the problem of multicollinearity has been observed.

Multicollinearity creates a problem in the multiple regression model because the inputs are all influencing each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the regression model. In statistical terms, a multiple regression model where there is high multicollinearity will make it more difficult to estimate the relationship between each of the independent variables and the dependent variable. In other words, when two or more independent variables are closely related or measure almost the same thing, then the underlying effect that they measure is accounted for twice (or more) across the variables. It becomes difficult or impossible to say which variable is really influencing the independent variable.

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. [Multicollinearity](https://www.investopedia.com/terms/m/multicollinearity.asp) exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the [regression](https://www.investopedia.com/terms/r/regression.asp) results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

The VIF measures how much the behaviour of an independent variable is influenced, or inflated, by its interaction/correlation with the other independent variables and allows a quick measure of how much a variable is contributing to the [standard error](https://www.investopedia.com/terms/s/standard-error.asp) in the regression.

In our project, we successfully used VIF to solve the problem of multicollinearity between the features.

1. **Building Machine Learning Models**

Machine learning literally means, make the machine learn, machine learns by processing the data with various machine learning algorithms. There is no fixed algorithm to provide high accuracy this is called the No Free lunch theorem. For any application, it is important to apply a few machine learning algorithms to find out the best-suited model.

* 1. Classification Part (Prediction of Price Range):

Here we applied multiple classification machine learning techniques and ensemble methods like "Logistic Regression", "Decision Tree Classifier", "Random Forest Classifier", "Support Vector Classification", "KNeighbors Classifier", "Gradient Boosting classifier", "AdaBoost Classifier", and "Bagging Classifier". After analysing the accuracy and cross-validation scores of these methods we observed that Support Vector Classification performed the best and hence we used this machine learning model as our final classification prediction model for predicting the feature Price Range.

**Support Vector Classifier/ Support Vector Machine:**

Support Vector Machine (SVM) is a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification and regression challenges. However, it is mostly used in classification problems. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate N-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

So, the basic objective of the support vector machine algorithm is to find a hyperplane in N-dimensional space (N — the number of features) that distinctly classifies the data points. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e. the maximum distance between data points of both classes. Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Hyperplanes are decision boundaries that help classify the data points. Data points falling on either side of the hyperplane can be attributed to different classes. Also, the dimension of the hyperplane depends upon the number of features. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called Support Vectors. Support Vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyperplane.

***Methodology & Mathematical explanations of SVM:***

So, initially, we are given the data which is to be separated by the algorithm.

The data given for separating/classifying is represented as a unique point in a space where each point is represented by some feature vector x.

Let, x ∊ RD

RD here is a D-dimensional vector space.

Further, mapping the point on a complex feature space x,

Φ(x) ∊ RM

The transformed feature space for each input feature is mapped to a transformed basis vector Φ(x) can be defined as:

Φ(x): RD RM

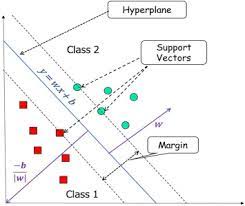
Our next job is to separate these points using a line and this is where the term decision boundary comes into the picture. The decision Boundary is the main separator for dividing the points into their respective classes.

The equation of the main separator line is called a hyperplane equation. Let us look at the equation for a straight line with slope m and intercept c.

The equation is: mx + c = y

We rename x with x1x1 and y with x2x2 and we get: ax1−x2+b=0

We can refer to the following picture (collected from Google image) for better understanding: -



Let, us define **x** = (x1, x2) and **w** = (a, −1)

we get:

w⋅ x+ b= 0w⋅ x+ b= 0

This equation is derived from two-dimensional vectors. But in fact, it also works for any number of dimensions. This is the equation of the hyperplane.

Classifier: Once we have the hyperplane, we can then use the hyperplane to make predictions. We define the hypothesis function H as:

H(xi)=+1, if w⋅x+b≥0

0, if w⋅x+b<0

The question then comes up as how do we choose the optimal hyperplane and how do we compare the hyperplanes.

Our best class separating hyperplane should be that one, for which the margin will be maximum. In other words, **“The goal is to maximize the minimum distance.”**

So, B= min|w⋅x+ b|, [for i=1...m]

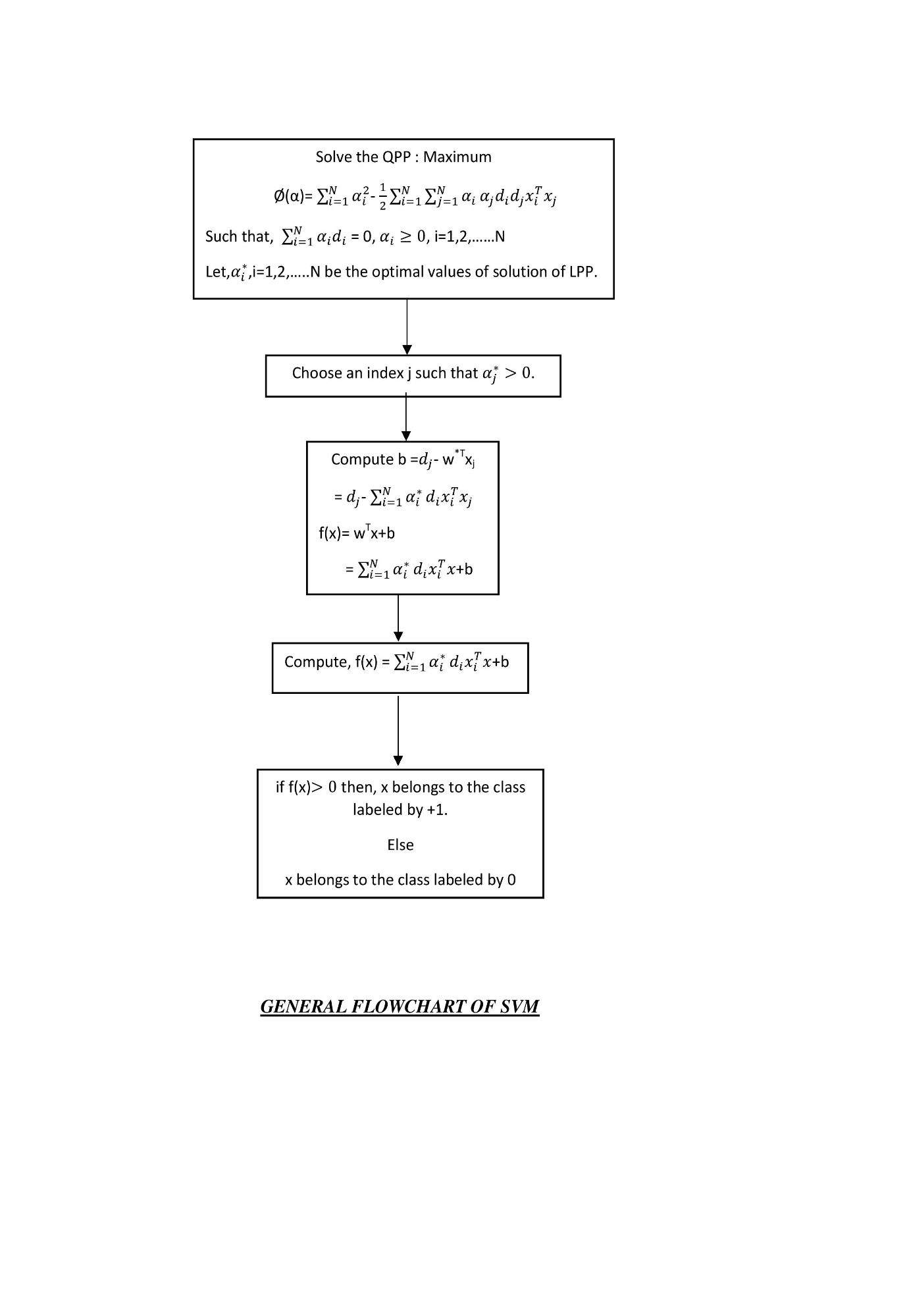
If we have k hyperplanes, each of them will have a Bi value, and we’ll select the hyperplane with the largest Bi value.

H=max {hi|Bi}, [for i=1...k]

So, from here after this whole method becomes nothing but a mathematical optimization problem.

SVM has a strong generalization performance and classification precision compared with other classification approaches.

The flowchart of general SVM is given below:

In the classification part of our project, SVM Classifier performed an accuracy level of 97% close to the actual values, which is a highly commendable!

* 1. Regression Part (Prediction of Average Cost of Two People):

Here we applied many of machine learning regression techniques and ensemble methods like Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, KNeighbors Regressor, Ridge, and Lasso. After analysing the R2 score and cross-validation scores of these methods we observed that Random Forest Regressor performed the best and hence we used this machine learning model as our final regression prediction model for predicting the feature Average Cost of Two People.

**Random Forest Regressor:**

Similar to SVM, Random Forest is also a supervised [machine learning algorithm](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2?utm_source=blog&utm_medium=understandingsupportvectormachinearticle) which can be used for both classification and regression challenges. Random Forest basically consists of multiple decision trees.

Decision trees learn how to best split the dataset into smaller and smaller subsets to predict the target value. The condition, or test, is represented as the “leaf” (node) and the possible outcomes as “branches” (edges). This splitting process continues until no further gain can be made or a pre-set rule is met, e.g., the maximum depth of the tree is reached.

Random forests (RF) construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As they use a collection of results to make a final decision, they are referred to as Ensemble techniques. Random-forest does both row sampling and column sampling with the Decision tree as a base. As we increase the number of base learners (k), the variance will decrease. Similarly when we decrease k, variance increases. But bias remains constant for the whole process. k can be found using cross-validation.

***Methodology & Mathematical explanations of Random Forest Regressor:***

The steps taken to implement a Random Forest:

**1**. Suppose there are N observations and M features in the training data set. First, a sample from the training data set is taken randomly with replacement.

**2**. A subset of M features is selected randomly and whichever feature gives the best split is used to split the node iteratively.

**3**. The tree is grown to the largest.

**4**. The above steps are repeated and prediction is given based on the aggregation of predictions from n number of trees.

**Train and run-time complexity:**

Training time = O(log(nd)\*k)

Run time = O(depth\*k)

Space = O (store each DT\*K)

As the number of base models increases, training run time increases so always use Cross-validation to find the optimal hyperparameter.

**Feature Importance**

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

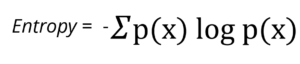
**Gini Index**

Gini index is the method to split out the data, it checks the impurity or purity of data it is used in the CART (Classification and Regression Tree) algorithm like Decision Tree. It creates a binary split and the CART algorithm uses it to create a binary split. An attribute is low Gini index is preferred as the root node.

**The formula to calculate the Gini index is:** Gini Index= 1- ∑jPj2

**Information Gain**

Information gain is calculated with the use of entropy in the data set and the attribute entropy, It gives us information about how much information a feature provides us with a class.



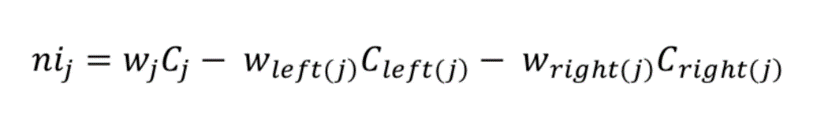
Entropy measures the impurity or randomness present in the given data. It is used to decide the root node in the decision tree to split out the data.

**Formula to calculate information gain:**

Information Gain= Entropy(S)- [(Weighted Average) \*Entropy (each feature)]

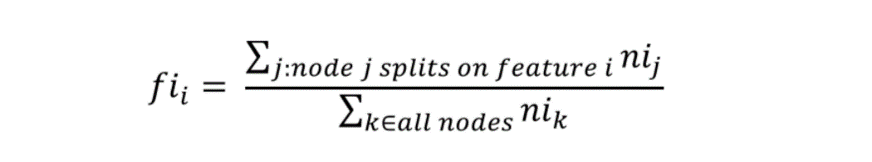
The highest the information gain we select that feature as the root node.

For each decision tree, the importance of a node is calculated using Gini Importance, assuming only two child nodes (binary tree):



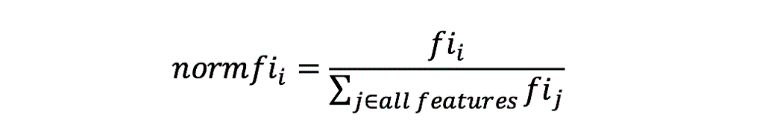
1. ni sub(j)= the importance of node j
2. w sub(j) = weighted number of samples reaching node j
3. C sub(j)= the impurity value of node j
4. left(j) = child node from left split on node j
5. right(j) = child node from right split on node j

The importance of each feature on a decision tree is then calculated as:

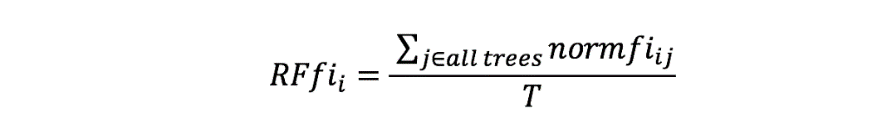


1. fi sub(i)= the importance of feature i
2. ni sub(j)= the importance of node j

These can then be normalized to a value between 0 and 1 by dividing by the sum of all feature importance values:



The final feature importance, at the Random Forest level, is it’s average over all the trees. The sum of the feature’s importance value on each tree is calculated and divided by the total number of trees:



1. RFfi sub(i)= the importance of feature i calculated from all trees in the Random Forest model
2. normfi sub(ij)= the normalized feature importance for i in tree j
3. T = total number of trees

So, in the random forest, we end up with trees that are not only trained on different sets of data but also use different features to make decisions.

In the regression part of our project, Random Forest Regressor performed an accuracy level of 91% close to the actual values, which is highly satisfactory!

1. **Concluding Remarks**

We have predicted the price range and average cost for two people when they have their meal in a Zomato-associated restaurant from the currency they use for payment. This project can be used to learn that although Zomato has a presence in 23 countries the most important country is India. Zomato will face ups and downs whenever the value of the Indian currency will change on an international scale. We came to know that Zomato-associated restaurants are categorized into Excellent, Very Good, Good, Average, Poor, and Not-rated according to the ratings received by the country. A customer can know about the average price range that a Zomato-associated restaurant should charge based on rating text. In the coming future, we can compare the difference between the average price for two based on the rating and also monitor a change in the average price for two when the value of the Indian currency alters in the international market.

1. **Acknowledgement**

I thank all of my supervisors in Mathematics and Data Science for providing support and tools for this research.