Project Name: Classification model to analyze the customer's restaurant preferences

Team members: Meghana Anoop 1000981002

Saranya Ravichandran 1001160582

Objective:

The objective is to predict the user behavior of preferring Restaurants having parking facility over the ones that do not have the same.

Overview:

Yelp provides business information of a restaurants including many features like parking facility, the price range of the menu items, star rating and the check-ins etc. we intent to mine the data and classify it using the two major conditions, prices and parking facility. Depending on the number of check-ins recorded for businesses we can deduce the probability of a customer visiting a restaurant having parking facility against those who do not have. This task can help predict the probability of a user choosing among the two features - prices and parking. It can help business to decide, improve or alter their services to increase their popularity and sales.

What's the end product how is it displayed?

A tentative idea about the output website is that it should display the restaurants and their parking availability along with its price range and show in bar graphs what the predictions are. When the model is run on the test set during the demo we can show the probabilities of user choices.

Data mining tasks:

Classification is one of the important data mining tasks that can help us identify to which set of categories a new observation or data will belong. We are using this classification technique to deduce a probability problem.

Using Naive Bayes' we are going to build a classification model and predict whether a customer prefers a particular restaurant or not based on the two attributes, Parking and Price .We are going to use Naive Bayes' Classifier with WEKA tool to implement the model. The probability of a customer going to that restaurant is calculated by using the individual probability of whether the restaurant has parking facility or not as well as using the individual probability of the price range. Since parking is a discrete attribute and price is a continuous attribute, we calculate the probabilities using different Bayesian algorithm. Based on this we analyze how much impact will the parking and price have on the probability by which the customers prefer a particular restaurant.

WEKA is a machine learning tool with a collection of data mining tasks. We plan to split the dataset based on each city. Each dataset is split into training and test data which are converted to ARFF format. For each city we perform the Naive Bayes' classification in WEKA. Based on the total result, we also get an insight whether the probability of customers preferring a restaurant based on parking and price differs for each city.

Major Challenges and how are we dealing it?

Weka tool is a new area to venture into. It will pose as a challenge to use the tool. Alongside we need to classify the data using the Naïve Bayes classifier. Creating a website to enable displaying the results in bar charts is the final big milestone. Apart from that the data set that we are handling is too large, here comes in the time constraint. Creating an efficient algorithm to consume the least time possible is the final goal too.

The initial plan is to apply Bayes classifiers to the training set and create a model which can then be applied on the test set. Next we focus on using weka tool and integrating the Bayes algorithm in it if possible. The next big milestone is using google charts api to visualize the results using bar charts and display them on the website. As a counter to dealing with large data, we intend to concentrate on a part of the huge data, pertaining to a particular city area only.

Efficacy:

The efficacy of our solution can be analyzed by taking another dataset with Check in attribute. We calculate the total check in made for the restaurants in each city based on the parking, price attributes. A comparison is made between the probability of a person preferring a restaurant and the total check in made for that particular restaurant. If it is directly proportional, it suggests that our solution is efficient.

Dividing the task among the team:

While one can work on understanding weka tool and its integration the other can work on getting individual probability of attributes and building a model. Further the team can work together on visualization using google charts api.