IS 603 Type II Final Report

University Restaurants Recommendation System

Project Members:

Abdullah Almohammadi

Pooja Gond

Olumuyiwa Somide

Professor: Dongsong Zhang
IS 603 - Decision Making Support System
Department of Information Systems
University of Maryland, Baltimore County

Content

INTRODUCTION

For university students, food has always been a hot topic as they stand in a line to get their meals in campus restaurants with strong expectations. Nevertheless, sometimes, they are just confused by the variety of dishes served by those restaurants in campus and they are hesitant their choices which might not be based on nutritional facts. However, when it comes to selecting and choosing a specific restaurant or dish based on decision bases, it has always been hard. One reason is that every individual has his/her own criteria for judging a restaurant based on different factors. Those factors contribute to success of a restaurant like quality of service, ambience of place, variety of dishes, waiting time, portion of food , and price these parameters constitute important objective factors based on which decision making can be done on restaurant selection. However, apart from these, there are also subjective factors like Word – of – mouth which forms an important base of information for any restaurant selection. Therefore, to meet the students demand every campus should have good food services. In our project, we propose a system that helps students make informed decisions based upon three aspects; quality, price and calories. Students on campus are uncertain in their food choices for different reasons, such as quality and price. Therefore the proposed system will help students to make an informed and wise decision in selecting a restaurant based on important aspects, such as quality, calories, and price on the university campus.

PROBLEM STATEMENT

To help selecting a restaurant based on different aspects, such as quality, calories, and price in the university campus. Also, analyzing and evaluating the meals taken at the restaurant so that student have balanced and healthy meals.

OBJECTIVE OF THE ANALYSIS

Student often share their opinions and their feedback on campus food pages about restaurants that they have visited. Which gives the team the opportunity to use those reviews in the proposed system to make a student able to choose a restaurant in the university using the food quality which was given by the students who have been to those restaurants. Also, there are two factors which are prices and calories will assist the team to make the system more useful for students in their food choices.

EXISTING APPLICATIONS

There are a few applications which are used to find a restaurant based on the reviews and quality of the food service. Also, there are some applications that their selection are not only based on the reviews of the restaurants but also the price. There are some of practical applications out on web which can cater to the needs of increasing user base in suggesting best restaurants in a particular locality as and when needed. Few of these applications as are listed as below:

Yelp.com: Yelp is an American multinational corporation mainly dealing with published crowd-sourced reviews about local business including restaurants. It is available in almost all the metropolitan cities with active and enthusiastic crowd using this website, yelp.com has been renowned source of restaurant recommendations based on user preferences and interests. It has features like involving geospatial location based recommendation, user reviews based on their past experience, restaurant information and working hours, specialties which is a good factor in choosing a restaurant [1].

TripAdvisor.com: TripAdvisor is a food and restaurant recommendation engine. Explore restaurants and reserve tables online. Download maps, reviews, and your saves for over three hundred cities worldwide in your phone for free are avoided using expensive data roaming plans while you travel. The application uses the location to send a personalized hotel, restaurant and vacation rental recommendations as well as remember the places a person visited via our Travel Timeline feature [2].

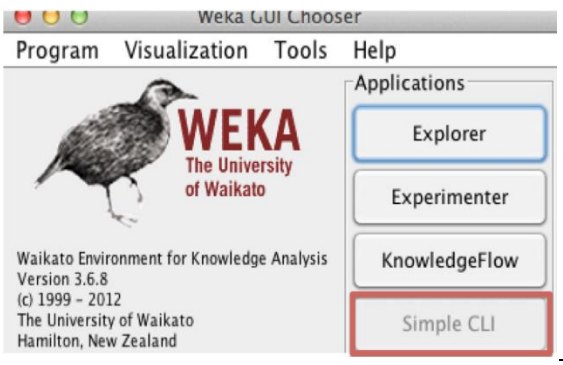
OVERVIEW

DATA PREPROCESSING

.CSV Files: A CSV is a comma separated values file. It saves data in a table structured format. CSV file look like a garden-variety spreadsheet and with a .csv extension. CSV files can be used with any spreadsheet program, such as Microsoft Excel, Open office, Calc., or Google Spreadsheets. They are different from other spreadsheet file types in which it only has a single sheet in a file, it cannot save cell, column, or row styling, and also cannot save the formulas. [7] A comma-separated values (.CSV) file stores tabular data (numbers & text) in a plain text. Each line of the CSV file is in a form of data record. Each record consists of one or more fields, separated by commas as columns. The use of the comma as a field separator is the source of the name for its file format. [8]

TOOL USED

Weka:



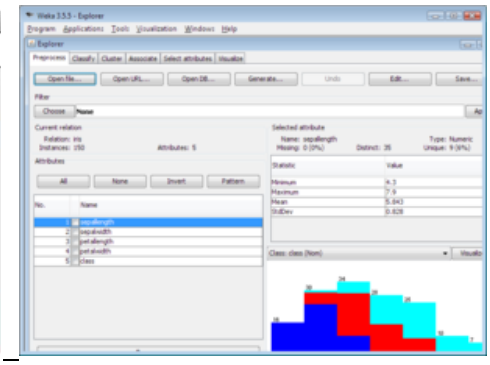


Figure 1: Weka Tool

Weka 3.8.0 (Waikato Environment for Knowledge Analysis) Weka is defined as a collection of machine learning algorithms for data mining tasks. It is based on java, the algorithms can be applied directly to the data sets which the team has done or can be called from a java class. Weka is a popular suite of machine learning software written in Java, which was developed at the University of Waikato, New Zealand. Weka contains tools for supporting several standard data mining tasks, more specifically, data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is also compatible for developing new machine learning schemes [4].

SYSTEM IMPLEMENTATION

DATA PROCESS MODEL

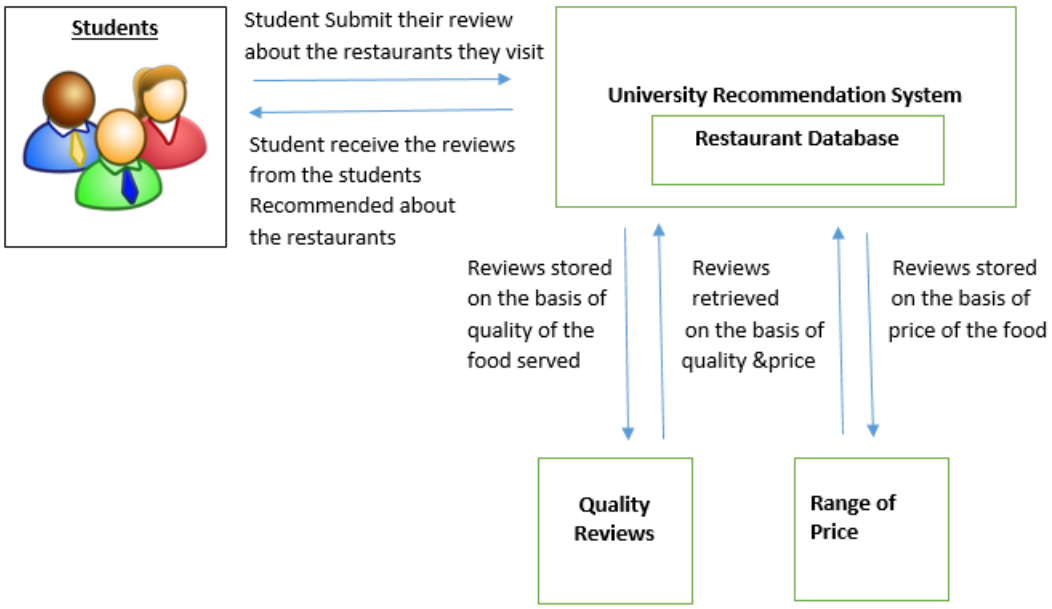


Figure 2: Data Process Model

The student who have visited the restaurant give the review about the restaurant they visit on-campus then submit their review on website. The dataset we receive is from the website itself we create an excel file and csv file using the given data. A database is generated on the basis of the data received.

Quality Review given by the students was done on this basis:

Upscale - Offer high quality cuisine at a high end price. They offer full service and have a high quality of ambience.

Mid-scale – The full offer meals at a medium price that customers observe as good value. They can be of full service, pounds or limited service with customers order at the counter and having their food brought to them or using the self-service.

Destitute -For fast-food restaurants. The food offered is of limited menus that is prepared quickly but are not good according to the reviews of students. They usually have driven through windows and take-out. They may also be self-service outfits.

Price Range was decided on this basis:

Low- The food served at the restaurant with the price less than 4$ were considered as low range.

Medium - The food served at the price between 5$ to 7$ were considered to be medium range.

High- Food which was more than 8$ were considered to be High.

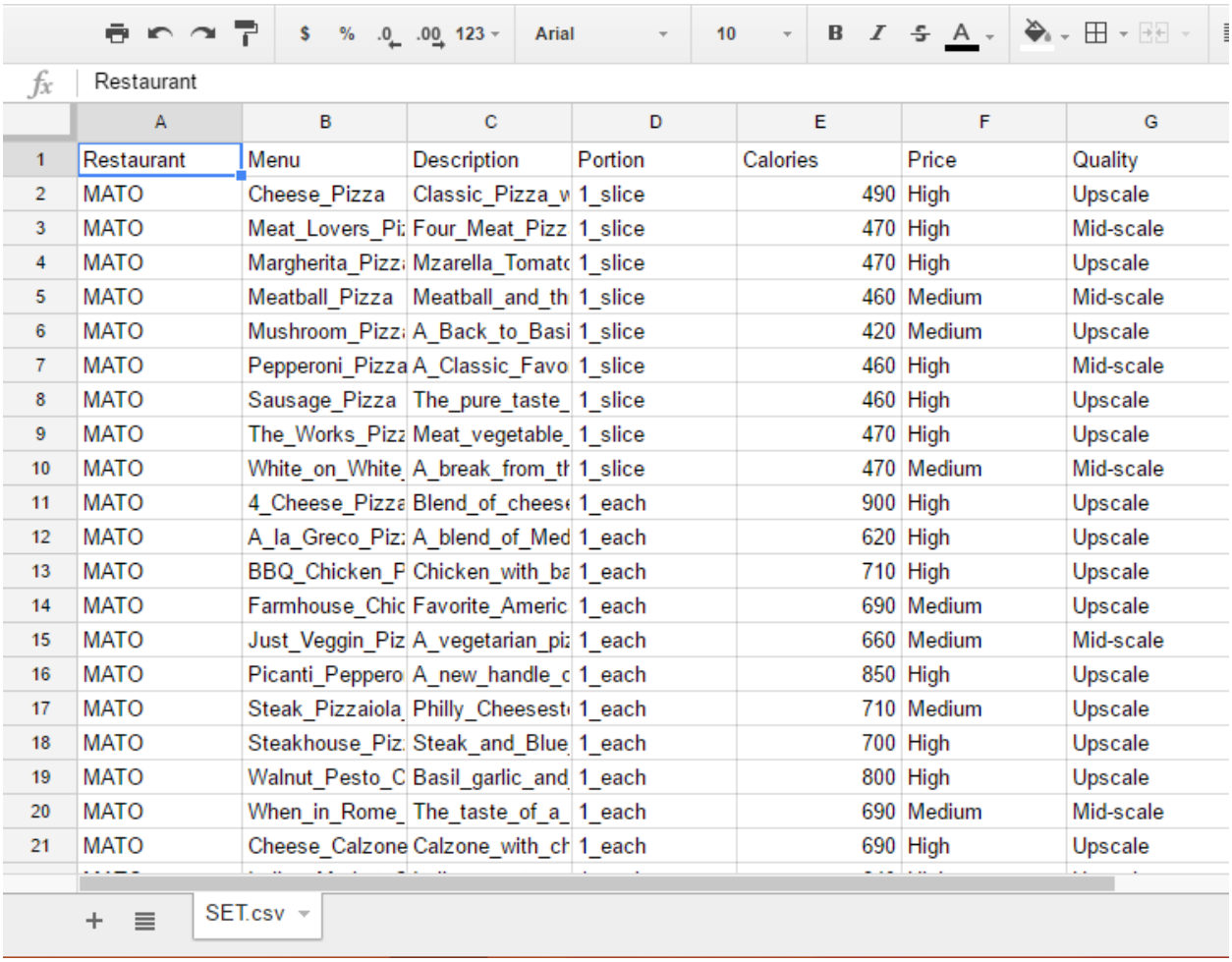
On the basis of this the dataset was created accordingly.

DATASET

The source of our dataset is Dine On-Campus [3]. The dataset that we have obtained was raw and was for seven restaurants including more than 500 items. The attributes of the dataset are restaurant name, menu, description, portion, calories, price and quality. This made adjustments on two columns; price and quality. The changes were that the price of each item was converted to high, medium and low based on the price itself. $1-$4 is consider low, $4-7$ is considered medium and above $7 is consider high. Also, the team has done some adjustments the reviews that were

given by the students into three categories Destitute, Mid-Scale and Upscale. In addition, rows of missing values were deleted.

Here is a screenshot of the dataset of University of Maryland- Baltimore County that we have edited.



Restaurant

	A	B	C	D	E	F	G
1	Restaurant	Menu	Description	Portion	Calories	Price	Quality
2	MATO	Cheese_Pizza	Classic_Pizza_w	1_slice	490	High	Upscale
3	MATO	Meat_Lovers_Pi:	Four_Meat_Pizz	1_slice	470	High	Mid-scale
4	MATO	Margherita_Pizz:	Mzarella_Tomatc	1_slice	470	High	Upscale
5	MATO	Meatball_Pizza	Meatball_and_th	1_slice	460	Medium	Mid-scale
6	MATO	Mushroom_Pizz:	A_Back_to_Basi	1_slice	420	Medium	Upscale
7	MATO	Pepperoni_Pizza	A_Classic_Favo	1_slice	460	High	Mid-scale
8	MATO	Sausage_Pizza	The_pure_taste_	1_slice	460	High	Upscale
9	MATO	The_Works_Pizz	Meat_vegetable_	1_slice	470	High	Upscale
10	MATO	White_on_White_	A_break_from_th	1_slice	470	Medium	Mid-scale
11	MATO	4_Cheese_Pizza	Blend_of_cheese	1_each	900	High	Upscale
12	MATO	A_la_Greco_Piz:	A_blend_of_Med	1_each	620	High	Upscale
13	MATO	BBQ_Chicken_P	Chicken_with_ba	1_each	710	High	Upscale
14	MATO	Farmhouse_Chic	Favorite_Americ	1_each	690	Medium	Upscale
15	MATO	Just_Veggin_Piz	A_vegetarian_piz	1_each	660	Medium	Mid-scale
16	MATO	Picanti_Peppero	A_new_handle_c	1_each	850	High	Upscale
17	MATO	Steak_Pizzaiola_	Philly_Cheesest	1_each	710	Medium	Upscale
18	MATO	Steakhouse_Piz	Steak_and_Blue_	1_each	700	High	Upscale
19	MATO	Walnut_Pesto_C	Basil_garlic_and_	1_each	800	High	Upscale
20	MATO	When_in_Rome_	The_taste_of_a_	1_each	690	Medium	Mid-scale
21	MATO	Cheese_Calzone	Calzone_with_ch	1_each	690	High	Upscale

SET.csv

Figure 3: Data Set

ALGORITHM USED

ID3 (ITERATIVE DICHOTOMISER 3) ALGORITHM

It is used to make a decision tree from a dataset and is characteristically used in the machine learning and natural language processing domains. The ID3 algorithm starts with the set as the root node. On each repetition of the algorithm, it repeats through every unused attribute of the set and calculates the entropy of that attribute. It then selects the attribute which has the largest information gain value. [4]

1. Calculate the entropy of all attribute using the data set given
2. Divided the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Make a decision tree node containing that attribute.
4. Recurrence on subsets using remaining attributes.

C4.5 ALGORITHM

C4.5 is a classification algorithm, which produces a Decision tree using the Training data and smears it for Classification.

Handling both continuous and distinct attributes - In order to handle continuous attributes, C4.5 generates a threshold and then separates the list into those whose attribute value is above the threshold and those that are fewer than or equal to it.

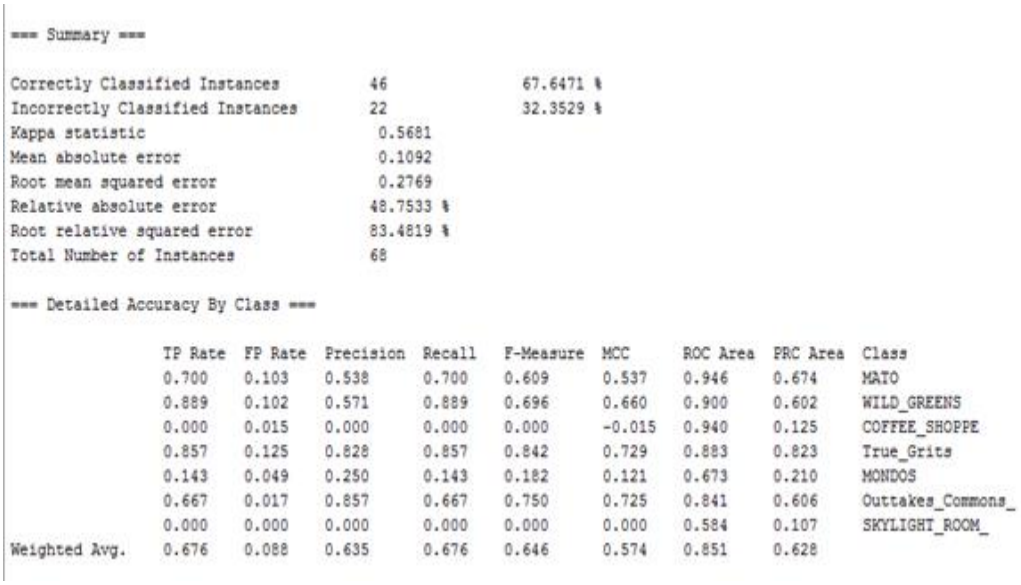
J48 DECISION TREE

Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. It is an open source Java implementation of the C4.5 decision tree algorithm. The basic idea is the division of the data into range based on the attribute values for that item that are found in the training sample.

While building a tree, J48 algorithm, ignores the missing values that is the value for that item can be predicted based on what is known about the attribute values for the other archives. Decision Tree is a classification technique which is used for data mining can be used on existing data as a test set and the resulting decision tree could maximize to create a plan for new restaurant and stream lining operations of existing restaurants in university restaurants.

Decision tree signifies a decision situation and hence helps in communication. The branches of a tree show all the factors within the analysis that are considered pertinent to the decision. For example, we can see the price factor was used in any of the branches because people are sensitive to money when it comes to better dining experience.

The Decision tree technique can be used to identify the influence of change on result, if one of the attribute value is changed. Henceforth it allows businesses to classify the factors that are more complex and less sensitive. This kind of sensitivity analysis is difficult to do in other modeling environments. Decision tree permits for forward and backward calculation paths to happen and thus the choice of the correct decision to take is made robotically. Since the moving parts / risk and success factors in restaurant business are narrow and business is small scale, decision tree is quick, easy and precise technique for decision making.



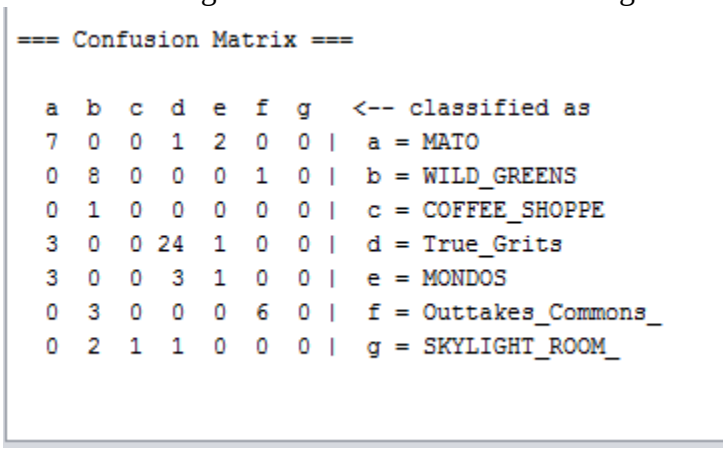
```
=== Summary ===

Correctly Classified Instances          46               67.6471 %
Incorrectly Classified Instances        22               32.3529 %
Kappa statistic                          0.5681
Mean absolute error                      0.1092
Root mean squared error                  0.2769
Relative absolute error                 48.7533 %
Root relative squared error             83.4819 %
Total Number of Instances               68

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0.700    0.103    0.538      0.700    0.609      0.537    0.946     0.674     MATO
               0.889    0.102    0.571      0.889    0.696      0.660    0.900     0.602     WILD_GREENS
               0.000    0.015    0.000      0.000    0.000      -0.015   0.940     0.125     COFFEE_SHOPPE
               0.857    0.125    0.828      0.857    0.842      0.729    0.883     0.823     True_Grits
               0.143    0.049    0.250      0.143    0.182      0.121    0.673     0.210     MONDOS
               0.667    0.017    0.857      0.667    0.750      0.725    0.841     0.606     Outtakes_Commons_
               0.000    0.000    0.000      0.000    0.000      0.000    0.584     0.107     SKYLIGHT_ROOM_
Weighted Avg.  0.676    0.088    0.635      0.676    0.646      0.574    0.851     0.628
```

Figure 4: J48 Decision Tree Testing Dataset Screenshot



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  7  0  0  1  2  0  0 |  a = MATO
  0  8  0  0  0  1  0 |  b = WILD_GREENS
  0  1  0  0  0  0  0 |  c = COFFEE_SHOPPE
  3  0  0 24  1  0  0 |  d = True_Grits
  3  0  0  3  1  0  0 |  e = MONDOS
  0  3  0  0  0  6  0 |  f = Outtakes_Commons_
  0  2  1  1  0  0  0 |  g = SKYLIGHT_ROOM_
```

Figure 5: J48 Confusion Matrix

NAIVE BAYES

Naive Bayes requires you build a classification by hand. There is no way to just toss a bunch of tabular data at it and then have it choose the best features it will use to classify. Selection of which features matter is up to us. Decisions trees will pick the best features from tabular data. If there were a way for Naive Bayes to choose features we would be getting close to using the same techniques

that make decision trees would work like that. Give this fact that means you might need to combine Naive Bayes with any other statistical techniques to help guide us towards what features best classify and that could be using decision trees. Naive Bayes will response as a continuous classifier. There are techniques to adapt it to definite prediction however they will answer in terms of probabilities like (E 90%, F 5%, G 2.5% H 2.5%) Bayes can perform fairly well, and it does not over fit nearly as much so there is no need to prune or process the network that makes them simpler algorithms to implement. However, they are harder to debug and understand because it's all probabilities getting multiplied 1000's of times so you have to be careful to test its doing what you expect. Naive Bayes does fairly well when the training data does not contain all possibilities so it can be very decent with low amount of data. Decision trees work better with a lot of data compared to Naive Bayes.

```
=== Summary ===

Correctly Classified Instances          50               73.5294 %
Incorrectly Classified Instances        18               26.4706 %
Kappa statistic                          0.6528
Mean absolute error                      0.1098
Root mean squared error                  0.2268
Relative absolute error                 49.0259 %
Root relative squared error             68.3812 %
Total Number of Instances               68

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.800     0.086     0.615       0.800   0.696      0.643    0.960     0.672     MAIO
               1.000     0.068     0.692       1.000   0.818      0.803    0.983     0.812     WILD_GREENS
               0.000     0.045     0.000       0.000   0.000      -0.026   0.910     0.143     COFFEE_SHOPPE
               0.893     0.050     0.926       0.893   0.909      0.848    0.954     0.955     True_Grits
               0.286     0.049     0.400       0.286   0.333      0.275    0.871     0.447     MONDOS
               0.667     0.000     1.000       0.667   0.800      0.796    0.987     0.936     Outtakes_Commons_
               0.000     0.016     0.000       0.000   0.000      -0.031   0.828     0.270     SKYLIGHT_ROOM_
Weighted Avg.  0.735     0.049     0.737       0.735   0.725      0.682    0.946     0.788
```

Figure 6: Naïve Bayes Testing Dataset Screenshot

```
=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
  8  0  0  0  2  0  0 |  a = MATO
  0  9  0  0  0  0  0 |  b = WILD_GREENS
  0  1  0  0  0  0  0 |  c = COFFEE_SHOPPE
  1  0  0 25  1  0  1 |  d = True_Grits
  4  0  0  1  2  0  0 |  e = MONDOS
  0  3  0  0  0  6  0 |  f = Outtakes_Commons_
  0  0  3  1  0  0  0 |  g = SKYLIGHT_ROOM_
```

Figure 7: Naïve Bayes Confusion Matrix

IMPLEMENTATION OF DECISION TREE

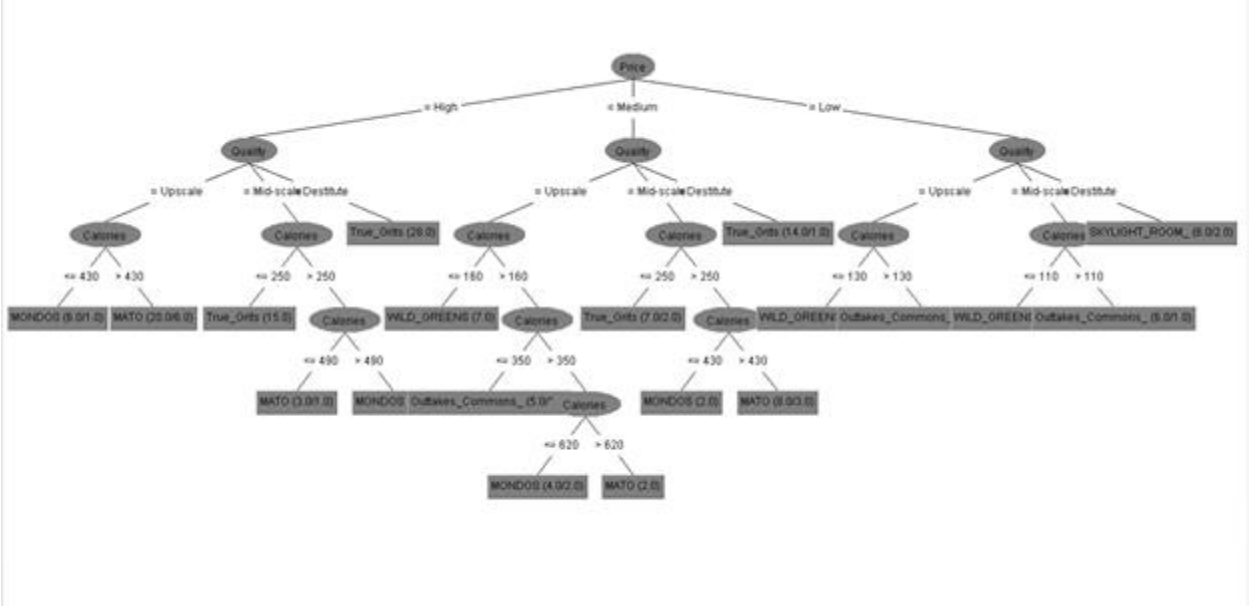


Figure 8: Decision Tree

The J48 Decision Tree is self-explanatory, it gives a better picture for the Student to see which restaurant to go for as they have multiple options they can select the restaurant given on the leaf node on the basis of different attributes node they go through.

CONSOLIDATED RESULTS

	J48 Decision Tree	Naive Bayes
Testing Accuracy	41	50
Accuracy	60.2941 %	73.5294 %
Mean absolute error	0.1326	0.1098
Testing Percentage Split	40	40

The training data set percentage split was kept at 60% and the Testing split was 40 %. The accuracy obtained was 60.2941% in the case of J48 Decision Tree and 73.5294% in the Naive Bayes which proves when we Naive Bayes it gives a better result than the J48 Decision Tree.

LIMITATIONS

1. No friendly user interface: The user using this system is underwriter. Weka does not provide friendly user interface for user who is not used to with it.

2. Compensating factors are not considered: Consider an example where calories intake of a person is less than specified amount. So, in such cases, according to the system rules, process may not get followed up. It is difficult to consider compensating factors because they purely depend on human judgments.

3. Change in reviews may result difficulties in training data set: If a student changes his rating or review about the restaurant or gives a wrong quality review which makes it good restaurant rather than bad would give different result.

4. Time constraints limited the scope of the project. We initially wanted to incorporate Body Mass Index (BMI) quantitative inputs variables and selections from the user that would yield customized outputs, specified to the user of the system. However, due to time constraints we decided to include the BMI as a future work scope to embark on. The classifier algorithms used for our project presents a problem in that it was a low precision to recall ratio received.

CONCLUSION

In this project, we proposed a system that can be used help students make informed decisions based upon three criteria's: quality reviews, pricing and caloric content. Our system analyzes and evaluates the menu items extracted from campus based restaurants, then it recommends to students, balanced and healthy meals options to choose from. Using a supervised machine learning classification algorithm, such as Naive Bayes and or J48, Our system would help students to make an informed and wise decisions in selecting a restaurant based on important aspects, such as quality, calories, and price on the university campus.

Upon analysis of our J48 decision tree, our system recommends the "Wild Greens" restaurant as the most moderate and cost effective priced dining option with an upscale quality review. Similarly, the "Mondos" restaurant was recommended as the most expensive (high cost) restaurant with an upscale quality review that offers the most beneficial health conscious meals.

In this project, our consolidated results from our dataset shows that Naive bayes was more efficient and accurate than the J48 algorithm. Therefore, the Naive bayes would be the ideal algorithm to use for a project such as ours in helping users make the best decisions.

FUTURE SCOPE

In this research paper, we used supervised machine learning classification algorithms. However, we would like to consider incorporating some degree of Heuristics.

In the future we would aim to incorporate a rule based classifier that would recommend the healthiest and most optimally rated restaurant choice, based on three conditional statements. These three conditional rules would have two constants, one of which is the quality review that would be set at an “upscale” value and secondly, the caloric rating, set at a “low” value range. These statements would only differ in terms of the pricing variable. For example, the first rule would have the price variable set as “expensive”, the second rule would be “average”, and the third rule would have the value of price set as “cheap”. The resultant output to the user would be three recommended restaurants that the user can make an informed decision about. The recommended restaurants can be sorted and listed from the healthiest choice to the least. This would then be presented to the user of the system. The recommendation classifier would be a Boolean type that denotes true or false for a recommended restaurant.

In addition, we would like to expand our future scope to include a body mass index (BMI) as an additional constraint, so that the resultant restaurant recommendation can be more so specified to

each user of the system. The BMI feature would be integrated with the existing project scope and would require the systems interface to prompt for more input variables from the user. The BMI is calculated from the user's input of height in inches and body weight in pounds (Lbs.) using metric units. Along with the BMI, a body fat chart would be used to determine the fitness level and percentage of body fat, calculated by circumference measures of the forearm, wrist, waist, and hip, along with the weight and gender type.

REFERENCES

[1] *YELP*. (2016, 05 05). Retrieved from YELP: www.yelp.com

[2] *Trip Advisor*. (2016, 05 05). Retrieved from Trip Advisor: www.tripadvisor.com

[3]*Dine on Campus*. (2016, 05 05). Retrieved from Dine on Campus: www.dineoncampus.com/umbc

[4] *Weka*. *ID3_algorithm. C4.5 algorithm* (2016, 05 06). Retrieved from Weka: http://www.cs.waikato.ac.nz/ml/weka/, https://en.wikipedia.org/wiki/ID3_algorithm, https://en.wikipedia.org/wiki/C4.5_algorithm

[5] *Abhinav Soni* (2015).Predictive Success of a Restaurant Using Data Science, Data Analytics Page 1-6. Retrieved from http://www2.rmcil.edu/dataanalytics/v2015/papers/Predictive_Success_of_a_Restaurant_Using_Data_Science.pdf

[6] *J48 Decision Tree*. (2016, 05 05)Data Mining with R. Retrieved from http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree

[7] *Unknown* (2015). What is a csv file and how do I save my spreadsheet as one. Bigcommerce. Retrieved from: https://support.bigcommerce.com/articles/Public/What-is-a-CSV-file-and-how-do-I-save-my-spreadsheet-as-one

[8] *Y. Shafranovich* (October 2005). Definition of the CSV Format. Common Format and MIME Type for Comma-Separated Values (CSV) Files. IETF. p. 1. RFC 4180.

Retrieved from: http://www.ietf.org/rfc/rfc4180.txt#page-1

[9] *Hemant Kumar Singh* (2015, 05). Classification of Restaurants. Hospitality Blog. Retrieved from: http://fnbclasses.blogspot.com/2010/05/classification-of-restaurants.html