Understanding University Students' Fast Food Consumption Behavior and Associated Health Concern

SUBMISSION DATE: AUGUST 03, 2018

SUBMITTED BY

Md. Ridowan Chowdhury (14341010)
Md. Maruf Rahman (14301036)
Department of Computer Science and Engineering

Supervisor
Samiul Islam
Lecturer
Department of Computer Science and Engineering

Co-supervisor
Dipankar Chaki
Lecturer
Department of Computer Science and Engineering

Declaration

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references.

Signature of Supervisor

Samiul Islam

Lecturer

Department of Computer Science and Engineering

BRAC University

Signature of Co-Supervisor

Dipankar Chaki

Lecturer

Department of Computer Science and Engineering BRAC University

Signature of Authors

Md. Ridowan Chowdhury

(14341010)

Md. Maruf Rahman

(14301036)

ACKNOWLEDGEMENT

Thanks to Almighty Allah, the creator, and the proprietor of this universe for providing us direction, guidance and confidence to finish this research within time.

We would like to offer our gratitude to our thesis supervisor Mr. Samiul Islam for being kindly patient with us throughout this one year and guiding us how should we approach to the problem. His important suggestions and time had been of a tremendous value to our work.

We are also thankful to our thesis co-supervisor, Mr. Dipankar Chaki for his support and time. Lastly, we would like to give our genuine appreciation to our parents who always supported us in our hard times and our friend Mr. Razaul Haque Subho for his active support and encouragement that helped us to complete this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIRGURES

LIST OF EQUATIONS

ABSTRACT

The aim of this research is to investigate the potential measurement of fast food consumption behaviour and the health hazard factor associated with it. Fast food consumption is getting more popular in Bangladesh, and we intend to capture the impact on the young generations. For this, we have drawn some questionnaires, gathered responses and tried to figure out the insightful information from this survey analysis using data-driven methods. A total of 170 university going students, of whom 122 were male (71.76%) and 48 were female (28.23%) constitute the sample of this research. We have analyzed the data with correlation analysis and chi-squared test to understand the behaviour of the features. Furthermore, we have used the K-means clustering algorithm to group students among their preferences while choosing a restaurant. In addition, we have used supervised machine learning models, Gaussian Naive Bayes, decision tree classifier (CART), Random forest classifier and Logistic regression to predict student's fast food consumption rate where Naive Bayes performed best with 79.4% accuracy. The result draws a conclusion on university student's health status and finds potential insights to fast food business.

CHAPTER 1

1.1 Introduction

Consumption of Fast food and takeaway among the young generation is no longer restricted in developed countries, it has spread to the developing countries as well [1]. In parallel with the quickly creating innovation, eating propensities additionally experience changes. The investigation of how customers pick among items has been perceived as a basic region of promoting research for the most recent couple of decades [2]. The significance of these restaurant attributes is ultimately evaluated by the customer's mind for selecting the restaurant and the fast-food. The motive may be credited via the enlargement of mindfulness, development of education, improvement of Information technology, and extension of networks [1]. Several other factors such as tradition, social class, group, age, gender also tend to increase or decrease the selection of restaurant and fast-food.

The subculture of speedy food consumption has diversified among university students and is indeed a great concern [1]. In spite of the fact that nourishment is critical for all sections of the public, it is of a distinctive significance for university students. Individuals, who pick up autonomy in this period, begin to choose their eating inclinations, to eat out additional often and to get affected by their friend network more. In this manner, they tend to expand those nourishments that are regarded as undesirable, for example, fizzy beverages and fast food more. Thus, it creates a profitable business opportunity to fast food business owners.

In the competitive market of the restaurant business, customer segmentation is a crucial part to target consumers. Food quality and food type are primary variables that influence consumer preferences in the fast food business, secondary variables are restaurants' decorations and atmosphere [8]. So, food quality and atmosphere related features can predict a restaurant's success if proper customer segmentation is available.

The outcome of this research is the identification of the factors that influence young customer's fast food preferences in Bangladesh. We have particularly chosen university students as a sample out of the population of the young generation. Based on some online research, we categorized important questionnaires that have significance in choosing a restaurant and food they prefer to eat. Then we conducted a survey among 170 university going students and try to understand their preferences and perspective. Further, over fast food consumption comes with

potential health risks. Therefore, we also tried to collect their health-related issues for relating to fast-food consumption. Furthermore, we developed a cluster model to categorize the young customer based on their preference. However, the survey analysis was conduct with a very small sample space and the purpose of this research was to figure out the pattern out of these sample space and fit into some of the existing models to get some positive result out of it.

1.2 Problem Statement

From the fast food consumption of a society, we can easily predict the food culture of that society. Moreover, the food culture effects directly the physical health and well-being. Unfortunately, there is not enough relevant research on fast food consumption pattern of the young generation of Bangladesh.

1.3 Aim of Research

Fast food consumption is getting more popular among young generations. However, over fast food consumption can arise health issues that cannot be ignored. The impact takes time, so when the impact is noticed, it may be already late. The aim of this study is to check whether the fast food consumption rate in Bangladesh is getting higher and whether it is already showing any indications of it. Further, fast food consumption is related to the restaurant business. We aim to find and analyze the indicators influence fast food business and attract people in choosing a restaurant.

1.4 Thesis Overview

Remaining part of this report is designed like this, chapter 2 consists of the literature review where we have demonstrated the background study that we have done for this thesis. Chapter 3 is the section where we discussed the research methodology of the work. It also contains some exploratory data analysis along with the graphical representation. Chapter 4 we attempt to predict fast food consumption rate using a different model. Chapter 5 represents a research hypothesis and findings. Chapter 6 contains future work and conclusion.

Our contribution on this research is finding insights and information on university students' fast food consumption behavior that may help to realize university students' current health status and may helpful to potential fast food business.

CHAPTER 2

2.1 Background Study

This part of this report is dedicated to study several types of research that have been done related to young generations' food behaviour and forecasting restaurants success measures based on consumer preferences.

A research attempted to develop restaurants recommendation system based on ratings of real users [4]. For this task, they chose to achieve user ratings from the Zagat survey and Google places. The research attempted to predict Zagat ratings based on Google Places ratings [4]. They developed a model with three parts. Firstly, an inner product model helps to predict the ratings of a user, then a Gaussian model to make the models consistent with one another [4]. Finally, latent variables have discussed by a Gaussian prior. They achieved minimum RMSE up to 1.144. In addition to, they analyzed how prices vary with the effect of food, décor and service and how cuisine types effects prices. They also proved a hypothesis that the more a user gives ratings, the higher is his reliability [4].

Another study conducted in Saudi Arabia with an intention to understand the factors that lead to high fast-food consumption rate [5]. The approached to collect data with a survey in the primary health care centres and analyzed the results with statistical tools. They concluded that education is independently correlated with fast-food consumption rate [5]. In addition, they found out that people choose to go to fast-food restaurants when they have limited time, or they want to change their taste or routine.

According to D. Fried, M. Surdeanu, S. kobourov, M. Hingle and D. Bell, overweight rate, the rate of diabetes, political views, and home topographical location and other information can be predicted from the food-related tweets [6]. For this, they have used the language-based model. Topic modelling has further improved the analysis. They have used the textual features of the tweets to understand associations between the dialect of nourishment, geographic district, and group attributes. Over 8 months, they have collected a corpus of meal-related tweets. To aggregate, annotate, and query these tweets they have created a system so that they can come up with a predictive system and visualization. After cleaning the data, they got some hidden info about people in groups like their percentage of being overweight, the percentage of diagnosed with diabetes. This language-based model has more improved result than previous systems and with NLP (topic modelling), they have more improvements. To examine the

effectiveness, they have used special textual features to understand relations between foods and geographic locations and community characteristics. They showed that representations of the dialect of sustenance over land or fleeting measurements could tell about the importance of certain meals in different regions and migration patterns in the USA and all over the world. They have extracted data through Twitter public streaming API2 containing hashtags of meals. Twitter provides data regarding the location, ethnicity, gender-related tweets. They filtered their desired tweets according to tweet hashtags, which are meal related. Then they used the topic model in order to mine the tweets.

A study attempts to find a dietary pattern and fast food consumption behaviour with a survey on children. They concluded that children have a better diet and energy intake in a day without a fast food than a day with fast food [11].

To track customer preferences for the food business, another research has been done using the twitter data on Indonesian people [7]. The main purpose was to find the food trend in a certain area of interest. The proposed approach was a bag of words model to extract features and then k-means clustering and simple additive weighting to rank the trends. Further, the results were compared with the sales record of some restaurants. The model reached an accuracy of 72.75%.

To evaluate daily menus at student's restaurant and to report dietary habits and other health related behavior, a designed self-administered was conducted to Croatian University students based on gender [17]. The paper was carry out by I.C.Baric, Z.Salatic, and Z.Lukesic and the aim was to highlight Nutritive value of meals, nutritive status and dietary habits of Croatian University Students. The examined subjects were 2075 universities students and data were accentuated concerning gender (Male 47.5%, Female – 52.5%). Random sampling chose menus and through food consumption table, nutritive value was calculated. This study showed that meals offered at students' restaurants are adequate. The energy fraction of food groups was calculated, and foods were grouped according to the Euro code Food Coding System. Several outcomes were concluded from that survey which differs gender. For example, Male exercise more than Female (4.4 hlweek versus 1.6 hlweek; $P < 0.05$). A higher level of Females (29.8%) than Male (17.2%) smoked cigarettes. For alcohol consumption, it was the other way around 88.9 and 84.8% of Male and Females, separately. A sum of 80.4% of understudies were very much supported.

Similar work was seen in “Gender differences in health habits and in motivation for a healthy lifestyle among Swedish university students” supervised by M.I.K von Bothmer and B.Fridlund where the purpose was to investigate gender differences in student’s health habit and motivation for a healthy lifestyle[18]. Probability Systematic stratified sample from each department at small university comprised the sample of students. Data Collection for this study was done by creating questionnaire. Number of health protestations estimated self-evaluated health, where good wellbeing was characterized as having less than three-health dissensions amid the most recent month. The Study showed several analysis and conclusion by computing on habits related to smoking, alcohol consumption, food habit, physical activities and stress. For example, Female students had more advantageous propensities identified with alcohol consumption and sustenance yet were more stressed. Male students demonstrated an abnormal state of overweight and heftiness and were less keen on nourishment counsel what's more; health upgrading exercises. Descriptive statistics were used to illustrate preliminary information. Chi-squared statistic was used as a test of independence between groups, and Phi-coefficient or Cramér’s V index was used as a measure of association to quantify the strengths of the relationships.

2.2 Concept Learning

Supervised Learning Model

Supervised learning models are a predictive model where the model is trained with a set of features along with a target feature. In our research, we attempt to work with a Naive Bayes classifier, decision tree classifier, random forest classifier and logistic regression classifier.

Naive Bayes Classifier

Naïve Bayes Classifier is a collection of classification algorithms based on “Bayes Theorem” which works with the independence assumptions between predictors. It works based on certain condition. Data from the dataset are already been predefined as train value. When test value is given to the model, it compares with the training value, finding the probability and then using classification it tells which category the test value falls under. Bayes theorem works just well in finding the probability of the outcome. Naïve Bayes classifier assumes that the effect of the value of a predictor(x) on a given class (c) is independent of the values of another predictor. This assumption is called Conditional independence [12].

$$P(c \vee x) = \frac{P(x|c)P(c)}{P(x)} \text{.....(1)}$$

Where, P *(x|c) is the Posterior Probability of target given predictor*

P(c) is the prior belief/Probability of target

P (x|c) is the likelihood and this is Gaussian because this is a normal distribution.

P(x) is the prior probability of predictor.

Since we are dealing with the Continuous value, we wanted to use Gaussian naïve Bayes. Gaussian Naive Bayes deals with the continuous data that assumes that the data is distributed as per Gaussian distribution. Following the property, we use Bayes theorem to find out the probability of an event.

Decision tree classifier

Decision Trees are a type of supervised Machine learning where data splits continuously based on certain parameters. It builds a classification model in the form of a structure. The dataset is broken down into smaller subset until it reaches the end. The tree is overviewed with decision nodes and leaf nodes.

The CART is used by beginning at the root node of the tree where all node receives a list of rows as input and root will receive the entire training set. Each node will ask a true or false question for one of the features and in response to that question, the data is partitioned into two subsets. The subset becomes the input of the two-child node. The goal is to produce the purest possible distribution of the level of each node. To build an effective tree, it is important to understand which question needs to ask and where it needs to ask. To unmixed, the level we need to quantify the number of questions. The uncertainty of a single node hence can be obtained by a metric called Gini impurity. Purity means having a maximum of one class when we do split. Gini index indicates how the classification spits with respect to the class [13].

Random forest classifier

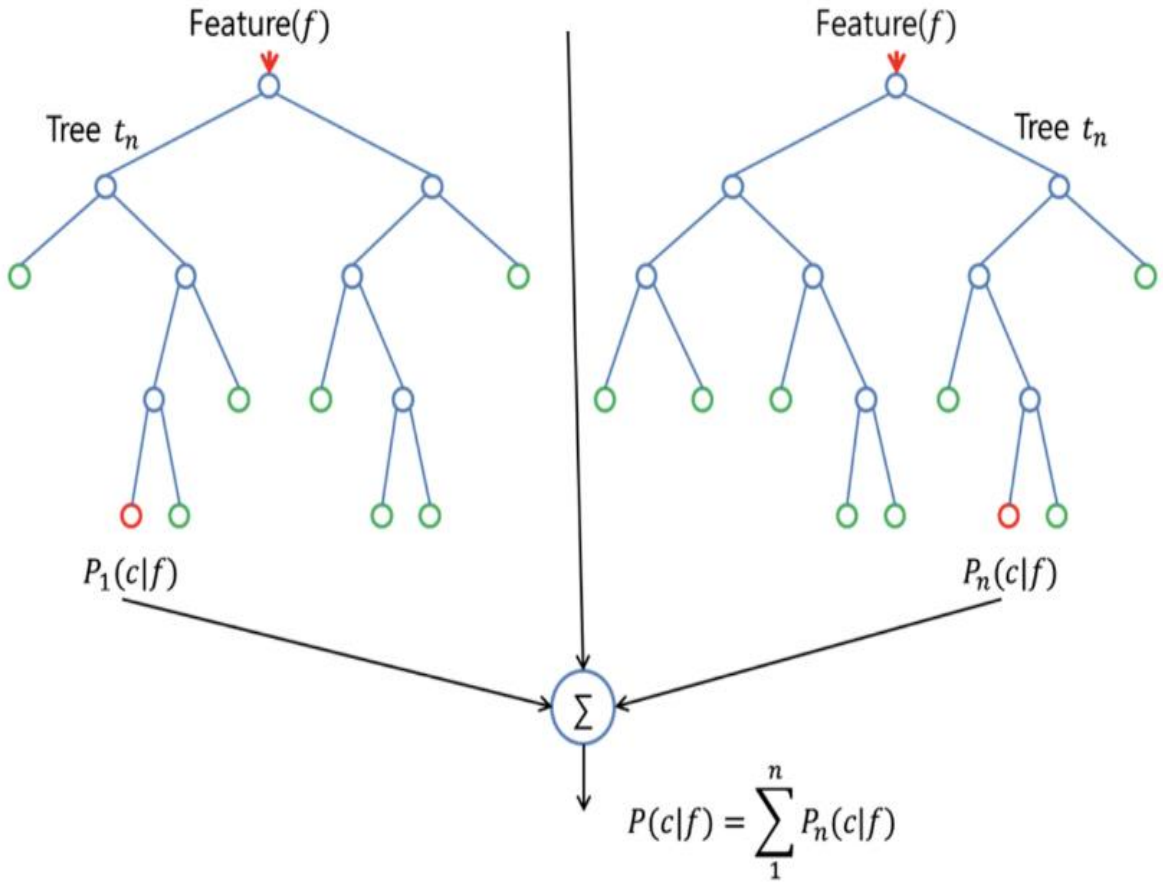


Fig 1: Random Forest Classifier

Random Forest algorithm is a supervised classification algorithm. It creates the forest according to the number of trees and makes it random precisely the random forest builds multiple decision trees and merge them together to get a more accurate and stable prediction. Random forest algorithm ensures the randomness of the model [14]. More trees in the forest build robustness of the Forest, additionally, they provide high accuracy results. Random Forest and decision tree share almost the same hyperparameters as decision tree classifier. However, it does not need to run a combined decision tree with bagging classifier. Unlike Decision tree, Random forest searches the best features among a random subset of features. The algorithm for splitting the node chooses particularly only a random subset of the features. This shows the results diversity which outcomes as a better model [15].

Logistic regression

Logistic Regression is the go-to method for Binary classification that gives discrete binary outcome between 0 and 1. Logistic Regression works by estimating the probabilities of the relationship between one or several Independent variables with one dependent variable. The

probability values therefore transformed into binary values to make an actual prediction [14]. Based on multiple assumption that is used in logistic regression, the following mathematical formula can be expressed as,

$$P = \frac{\exp(\beta 0 + \beta 1.\chi 1 \ldots \ldots + \beta n \chi n)}{1 + \exp(\beta 0 + \beta 1.\chi 1 \ldots \ldots + \beta n \chi n)} \ldots\ldots (2)$$

2.3 Statistical analytical tools

- Microsoft Excel
- Spyder (Python)
- Scikit-Learn
- SPSS
- Rapid Miner

For all the frequency measuring and testing, drawing bar chart, we have used Microsoft Excel. Scikit-Learn tools were used in Spyder (python) for predictive analysis and Cluster analysis.SPSS tool was used for Chi-squared value testing and cluster analysis, and for optimizing features and classification Rapid miner was used.

CHAPTER 3

3.1 Primary List of Features and Data Collection

At first, we make a primary feature list that can relate to our work based on similar research surveys conducted on other countries in the world and ground truth. The features are divided into two parts. One attempts to find the features that can capture the common health issues and, the other attempts to capture fast food consumption habits and influences behind them. This allows us to create 48 question questionnaires to conduct the survey. A number of 170 university students age ranged from 19 to 25 attend the cross-sectional questionnaire survey where 71.7% are males and 28.3% are females. Some part of the survey data is given in the table 1.

3.2 Data Description

Table 1: Snapshot of the Survey responses

Do you feel healthy?	Your height (inches)? (hint: 1 ft = 12 inches)	Your weight (kg)?	Do you have numbness or tingling in your arm?	Do you feel fatigued after eating?	Do you often feel "older" than you should for your age?	Do you catch colds or the flu easily?	Do colds, flu, or other infections tend to linger in your system for more than 5 days?	Do you take any medicine?
Yes	68	68	No	No	No	No	No	Yes
Yes	65	88	No	No	No	No	No	No
Yes	67	80	No	No	No	No	No	No
Yes	71	64	No	Sometimes	No	Yes	No	Sometimes
Yes	73	100	Yes	Sometimes	No	Sometimes	Sometimes	Yes
No	67	70	No	Yes	Sometimes	Sometimes	No	Yes
Maybe	64	60	No	o	Yes	Yes	No	Sometimes
Yes	72	70	No	No	No	Yes	Yes	No
No	70	57	No	Sometimes	No	Sometimes	No	No
No	79	101	No	Sometimes	Yes	Yes	No	No
Yes	67	60	No	No	No	Yes	No	Sometimes
Yes	68	74	No	No	No	No	No	Yes
Yes	6	70	No	Yes	Yes	Yes	Yes	No
Maybe	62	41	No	Sometimes	No	No	Sometimes	Sometimes
Maybe	72	65	Sometimes	Sometimes	No	No	No	No
No	65	60	No	No	No	No	No	No
No	68	89	No	No	No	No	Yes	No
Yes	62	53	No	No	No	No	No	No

Table 2: Data Description

Attribute	Description
1. Sex	Categorical value to Numeric value
2. Age	Student's Age(numeric : Range 18-30)
3. Fatigued	Categorical value to Numeric value
4. Red_meat_habit	Categorical value to Numeric value
5. Catch cold	Categorical value to Numeric value
6. numbness	Categorical value to Numeric value
7. Take medicine	Categorical value to Numeric value
8. Drink tea(per day)	Categorical value to Numeric value
9. Method_of_cooking	Numeric(No of times per day)
10. Select_of_ingredients	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
11. Food presentation	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
12. Traditional symbols	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
13. Environment	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
14. Price	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
15. Service	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
16. Convenience	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
17. Elegance	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
18. Lively	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
19. Decoration	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
20. Casual	Numeric (rate 1-5, 1= 0 likely, 5= extremely likely)
21. Restaurant_visit/15days	Numeric(No of visit per 15 days)

The dataset comprises health-related features and restaurant attribute related features. Some health-related features are numbness, fatigue after eating, catch colds or flu easily, take medicine, smoking habit etc. The respondents also responded that what restaurant related features influence them most on the scale of 1-5. Some of the features are decoration, a method of cooking, environment, price etc. In addition to, students responded to their average spending per visit, and average visiting frequency per 15 days. In addition, students rated social media influence, influence from the offer, promotion, advertisement etc.

3.3 Data Cleaning and Preprocessing

Target based encoding is an enumeration of a categorical variable via target. In this method, we replace the categorical variable with just one new numerical variable and replace each category of the categorical variable with its corresponding probability of the target. We appraise target-based encoding using for categorical to numerical process. From the dataset, we can observe from the Gender section that there are 122 male and 48 females among total 170 responses to the questionnaires. Therefore, we used 122/170 = 0.72 as male value and used 48/170 = 0.28 as female value. In this way, we converted each categorical attribute's value to its corresponding probability of the target.

CHAPTER 4

4.1 Research Methodology

To complete the research work, we created a survey questionnaire and conducted a survey between 170 students and these have been used to perform data analysis and develop machine-learning models.

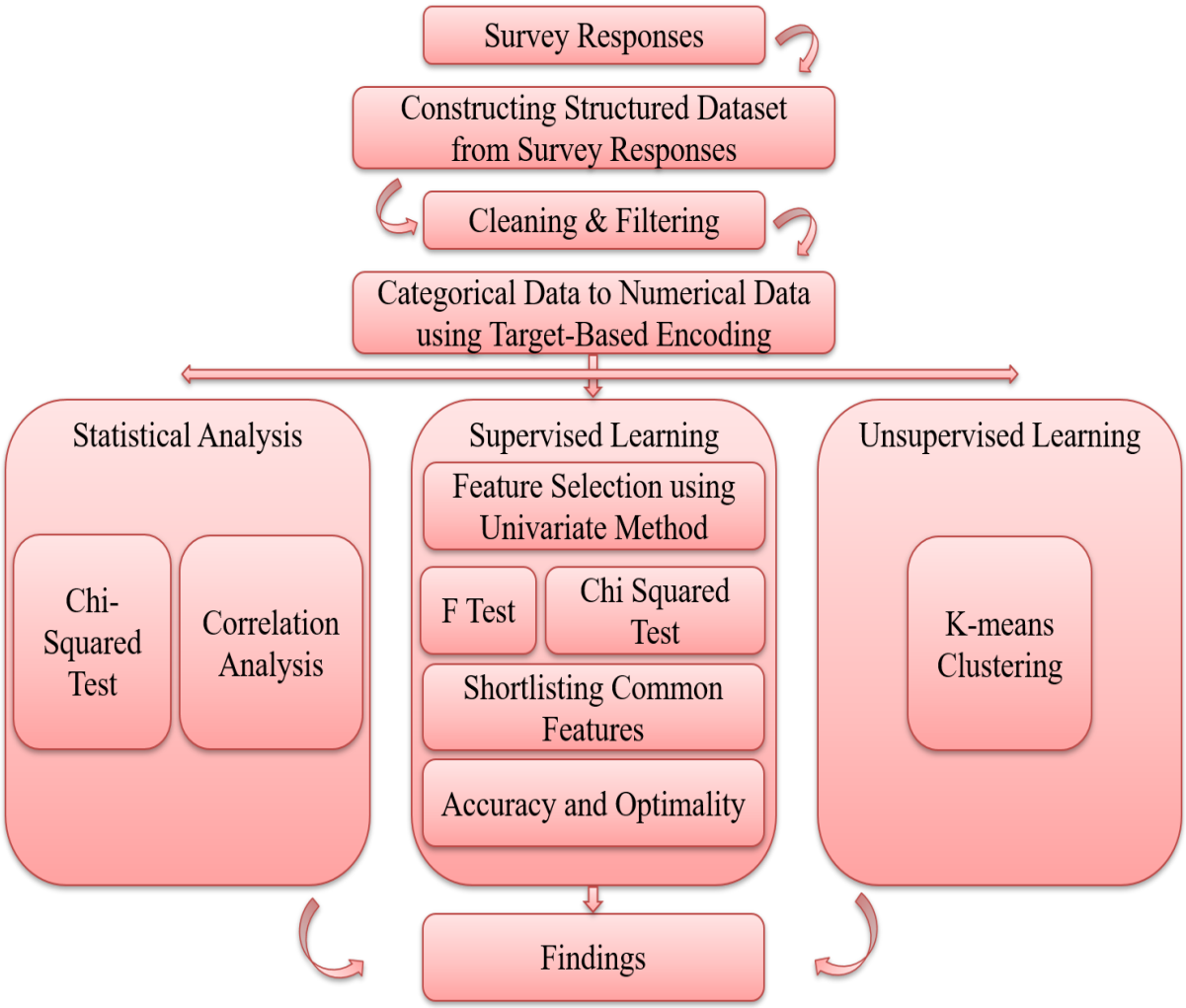


Fig 2: Research Methodology

Fig.2 demonstrate the basic workflow of our research. First, we assembled questionnaires from different articles, newspapers, websites and ground truth. Several research articles have been picked; relevant ones are thoroughly studied after the idea generation stage to form a questionnaire that is a composition of all the necessary parts pointed out by the preceding researchers of this track and the key questions that we believe more impactful in then conduct a survey analysis among a group of students. After collecting all the responses, the dataset was

accumulated after filtering and cleaning it. Target based encoding was used to convert categorical value to numerical value. The dataset was then used to find out frequency and Chi-square value to understand the significance of each of the factor. Apart from that, K-means clustering was also used to learn students' preference for choosing the restaurant. Further machine learning models are applied after a univariate feature selection process to predict fast food consumption rate.

4.2 Statistical Analysis

Table 3: Frequency Table

	Male=122		**Female=48**		**Total=170**	
	N	%	N	%	N	%
Fast food consumption frequency/15 days						
1 time	40	32.78	12	25.0	52	30.59
2 times	26	21.31	7	14.58	33	19.41
3 times	10	8.19	7	14.58	17	10.0
4 times	9	7.37	5	10.42	14	8.23
5 times	6	4.91	5	10.42	11	6.47
More than 5 times	27	22.13	12	25.0	39	22.94
Rarely	4	3.27	0	0.00	4	2.35
Chi_2=6.58, p_value=0.36>0.05						
Average spending per visit						
Less than 200 taka	31	25.41	8	16.66	39	22.94
200-400 taka	60	49.18	25	52.08	85	50.0
400-600 taka	23	18.85	9	18.75	32	18.82
600-800 taka	5	4.1	2	4.17	7	4.12
More than 800 taka	3	2.46	4	8.33	7	4.12
Chi_2=4.09, p_value=0.39>0.05						

Issues	**Male=122**				**Female=48**				**Total=170**			
Health-related issues	yes	No.	sometimes	%yes	Yes	No.	Sometimes	%yes	Yes	No	Sometimes	%yes
Numbness or tingling in arm	4	99	19	3.28	2	40	6	4.16	6	139	25	3.53
Fell fatigued after eating	15	68	39	12.29	11	25	12	22.92	26	79	51	15.29
Catch colds easily	32	71	19	26.23	17	23	8	35.42	40	94	27	23.53
Unable to lose weight	22	100	--	18.03	13	35	--	7.65	35	135	--	20.58
Smoking habit	26	87	9	21.31	2	43	3	4.16	28	130	12	16.47
Red mead habit	88	16	18	72.13	36	5	7	75.0	124	21	25	72.94
Prefer more chili	62	38	22	50.81	25	16	7	52.08	87	54	29	51.18
Prefer more salt	19	79	24	15.57	7	36	5	14.59	26	115	29	15.29
Take medicine	21	66	35	17.21	19	19	10	39.58	40	85	45	23.53
Average sleep at night												
0-3 hours	8			6.56	1			2.08	9			5.29
3-5 hours	53			43.44	16			33.33	69			40.59
5-7 hours	50			40.98	26			54.16	76			44.71
7-9 hours	10			8.19	4			8.33	14			8.23
More than 9 hours	1			0.82	1			2.08	2			1.17
BMI												
Underweight	12			9.84	6			12.50	18			1059
Healthy weight	73			59.84	24			50.0	97			57.05
Overweight	23			18.85	13			27.08	36			21.18
Obese	14			11.47	5			10.41	19			11.18

Chi-squared test is a statistical test to check whether two categorical variables are statistically independent. It is generally used to evaluate test of independence when using a bivariate table. Cross tabulation presents the distribution of two categorical variable simultaneously. By comparing the observed pattern of response, test of independence evaluates if there exist any correlation between two variables [16]. The calculation of Chi-squared test can be evaluated by the following formula:

$$\chi 2 = \frac{\sum(\mathrm{O}-\mathrm{E})2}{E} \quad \text{.....(3)}$$

O = Observed frequency

E = Expected under null hypothesis and computed by,

$$E = \frac{\text{row total} * \text{column total}}{Esample\ size} \quad \text{.....(4)}$$

In our dataset, we performed a chi-squared test to find whether the fast food consumption rate and average spend on fast food between male and female are independent of one another. In the following table, we see that the chi-squared test result on fast food consumption rate is, chi-square = 6.58, and p=0.36. In this case, the degree of freedom is 6, and the critical value for the degree of freedom of 6 is 12.592, which is greater than chi-square value 6.58. In this, we cannot reject the null hypotheses and conclude that fast food consumption rate does not depend on gender.

In addition to average spending on the internet has chi value 4.09, where the degree of freedom is 4. The critical value for the degree of freedom 4 is 9.88 that is also greater than the chi value. Therefore, we can say that the null hypothesis cannot be rejected, so, average spending does not depend on gender. Hence, we can conclude that fast food consumption behaviour does not depend on gender.

From the table 3, we can also see that different health-related features have a relatively small percentage of positive response. The numbness has 3.53%, fatigued after eating has 15.2%, catch colds easily has 23.5%, unable to lose weight has 20.5%, the smoking habit has 16.4 and takes medicine has 23.5%. Further, only 6.46% sleeps more or less than average at night and 11.18% experience obesity. To conclude that, we can say, fast food habit is more or less the same for both gender and health-related issues do not depend on gender.

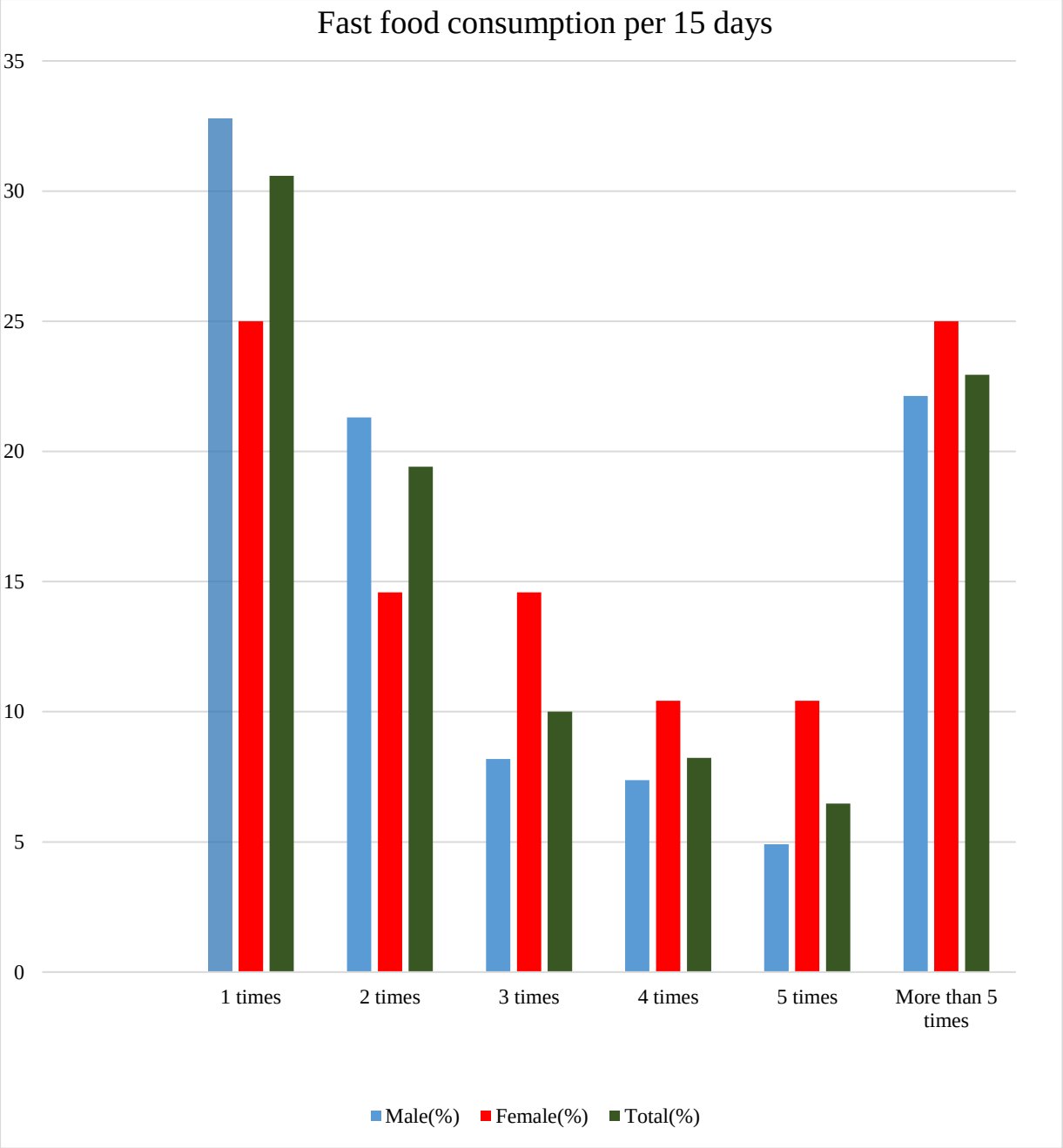


Fig 3: Fast food restaurant visiting ratio of the responders

We can observe how much money the population would like to spend while visiting the restaurant. The percentage is almost half of the population who would like to stick around 200 BDT to 400 BDT and significantly both male and female's percentage is the same (male= 49.18%, female-52.08%). Apart from that, very few portions of the population like to go above 800 BDT. Among below 200-taka category male (25.51), the percentage is slightly better than female (16.66%), which means they probably like to have the street food or budget meal more. People who like to spend around 400- 600 BDT are 18.82% and 4.12 percent above 600 BDT.

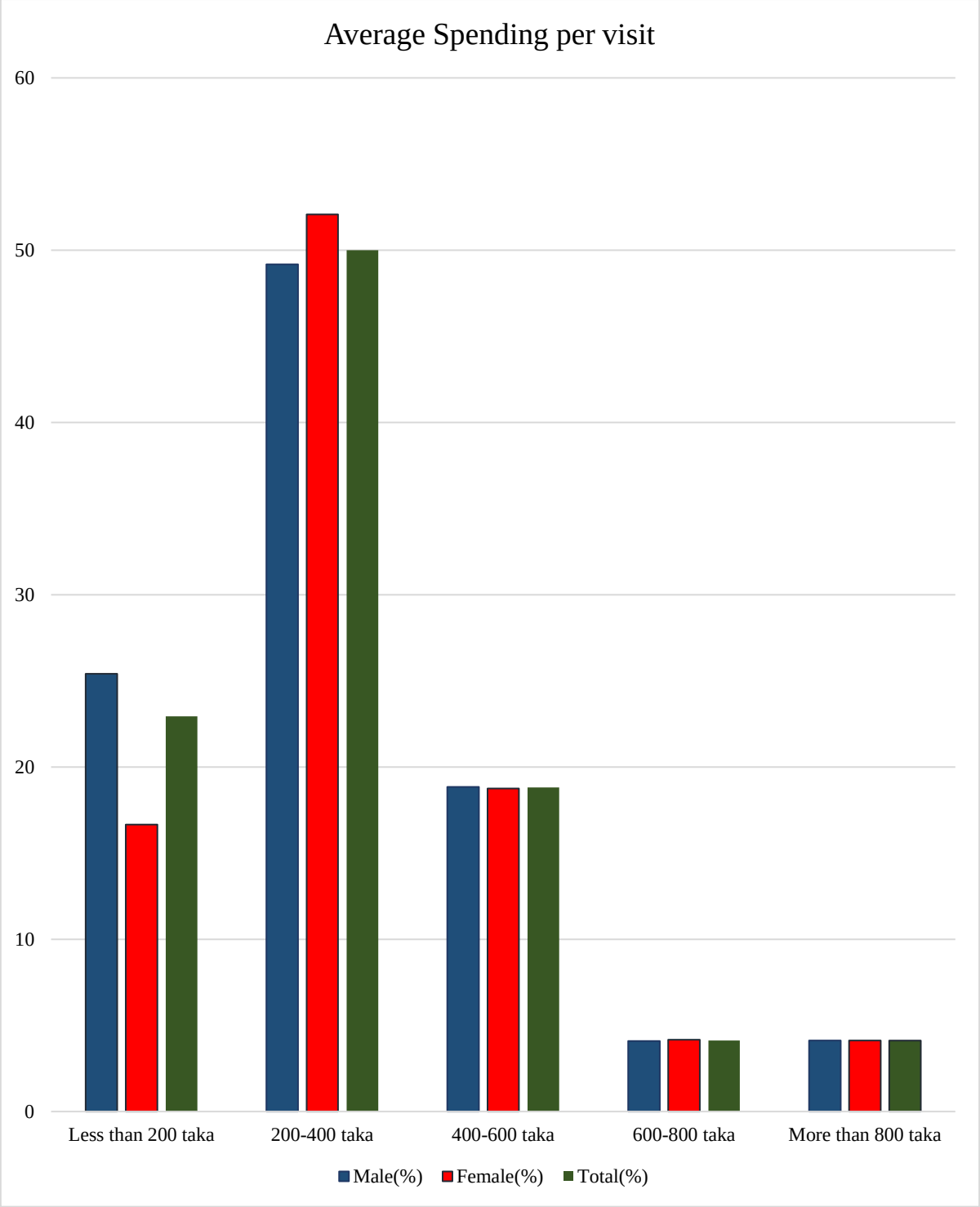


Fig.4. Spending ratio between the genders

Throughout the survey analysis, we tried to identify does fast food consumption has anything to do with health-related issues. Based on questionnaire's we have levelled out Numbness, fatigued after eating, catch a cold easily, unable to lose weight, smoking habit, red meat habit, more chilli, more salt, take medicine (fig 4). Numbness or tingling in arm has significantly quite a low percentage (3.53% in total).

Among all the questionnaires we identified Red meat habit has the most expected choice in total population with a percentage of 72.94. Both Male (72%) and Female (75%) prefer to have red meat. About 51.18% of the total population prefer to have more chilli in their food. Among 20.58% of the population suffers to lose weight and a few portions of them belong to the female compared to male. Smoking habits are significantly higher for male (21.31) than female (4.16%). Consequently, people are less likely to add more salt to their food and the percentage shows only 15.29.

The Average sleep at night also differs from person to person, although there is around 45 % of the population who likes to sleep for 5 to 7 hours at night. The table also shows that there are least of the population who barely gets more than 9 hours to sleep. There are also 5.29% in total who sleep around 0 to 3 hours at night. People who sleep for around 3 to 5 hours also have a very good portion (40.59%) in total.

Body mass index is also an important factor to analyze the health condition of the sample. Among all the population there is 10.59% population who feel they are underweight and 57.05% feel they are healthy and have no issue. There are 21.18% who feel overweight and female tends to have more overweight compared to male. Lastly, there is the least percent of the population who suffers from obesity.

4.3 Correlation Analysis

Correlation is a statistical measure that describes the association between random variables. We used the Pearson correlation coefficient formula, which uses raw data and the means of two variables, X and Y.

In our dataset, we have used Pearson's algorithm, which commonly uses statistics to measure the relationship between two variables. Pearson's algorithm R range in value from negative one to positive one. The further the correlation from zero, the stronger the correlation is. There are three types of correlation. Weak correlation ($R< 70$), Moderate correlation ($0.3 < P < 0.7$) and Strong Correlation ($r > 0.70$).

Table 4: Correlation Table

Category (Sample)	**Feature1**	**Feature2**	**Pearson's Correlation**	**Interpretation**
Overweight & Obesity	Tea/Coffee-day	Smoking-habit	0.58	Moderate Correlation
	Catch_cold	Take medicine	0.42	Low-moderate Correlation
Eats Fast-food or Both	Red-Meat_Habit	Restaurant_ Visit/15days	0.36	Low-moderate Correlation
Eats Homemade Food	Fatigued after Eating	Catch_Colds	0.46	Low-moderate Correlation
	Tea/Coffee-day	Restaurant_ Visit/15days	0.38	Low-moderate Correlation
Pays less than or equals to 400 Taka	Method_of_cooking	Selection of Ingredients	0.67	High-moderate correlation
	Environment	Service	0.66	High-moderate correlation
	Service	Price	0.71	Strong Correlation
	Decoration	Service	0.59	Moderate correlation
Visits less than or equals to 2 times per 15 days	Price	Service	0.71	Strong Correlation
	Method_of_cooking	Selection_of_ ingredients	0.66	High moderate Correlation
	Service	Environment	0.65	High-Moderate Correlation
Visits more than 2 times per 15 days	Method_of_cooking	Selection_of_ ingredients	0.73	Strong Correlation
	Service	Environment	0.59	Moderate correlation
Pays more than 400 Taka	Method of cooking	Selection of_ ingredients	0.75	Strong Correlation
	Food_ Presentation	Environment	0.55	Moderate Correlation
Male Preference	Service	Environment	0.68	High-moderate Correlation
	Method of cooking	Selection_of_ ingredients	0.67	High-moderate Correlation
Female Preference	Method_of_cooking	Selection_of_ ingredients	0.75	Strong Correlation
	Elegance	Decoration	0.66	High-moderate Correlation

We attempt to find a correlation between different kinds of features available on our dataset. We have collected survey of 170 people where all of them are students; based on their nature we have added several questionnaires. From the responses, we categorized our features and based on certain criteria, and later we have figured out the relationship between them. From table 4 we can see that those who have Obesity and Overweight category there is a moderate relationship between tea-coffee/day and smoking_habit (p=0.58), take_medicine and catch_cold. The correlation between Price and Service (p=0.71) were showing relatively better for those who spend less than or equals to 400 BTD. Those who visit less than or equals to two times per 15 days also focus on the amount of price and the service quality (p= 0.71) so we can observe a good correlation between them. Apart from that Method_of_cooking and Selection_of_ingredients also showing moderate relationship for those who spend less than or equals to 400 takas (p= 0.67) and those who visit less than or equals to two times per 15 days. Consequently, those who visit more than two times per 15 days look for method_of_cooking and Selection_of_ingredients (p=0.73). Method_of_cooking and Selection_of_ingredients also have a strong correlation for those who pay more than 400 takas. Surprisingly Red_meat_habit and restaurant_visit/per 15 days have a low-moderate relationship (p=0.36) for those who love to eat outside or both homemade and outside. However, when we categorized between male and female preference, we see Female look for Method_of_cooking and Selection_of_ingredients (p=0.75), Elegance and Decoration (p= 0.66) where Male look for Service and environment (p=0.68), Method_of_cooking and Selection_of_ingredients.

The result shows the moderate relationship between attributes, which indicates not a very significant relationship. However, in the most category, we observed some of the attributes, which shows great results like Method_of_cooking and Selection_of_ingredients, Service and Price.

4.4 Clustering

Clustering is an unsupervised approach to categorize unlabeled data. The K-means clustering algorithm is an efficient and commonly used algorithm to do such clustering [8]. In our survey dataset, we applied the K-means clustering algorithm to learn students' preferences over choosing a food place. On the survey questionnaire students were asked, what do they look for when they seek authenticity in a food place? They rated on a scale 1-5 over the parameters of cooking standard, decorations, environment, price, traditional symbols, convenience etc. They

were further asked to rate their preferred atmosphere on a food place over the parameters of casual, lively, decorated and elegance. Thus, the responses allowed us to create an 11 features dataset to train an unsupervised model with a k-means clustering algorithm. As the model needs to specify the number of centroids explicitly, we choose to find the optimal k value with the elbow method. The elbow method is a visual method where cost vs k graph shows a plateau in the optimal k-value [9]. The elbow method suggested choosing k=3. The model used PCA (principal component analysis) to reduce dimensions to two axes. The model gave us the graph shown in fig 5:

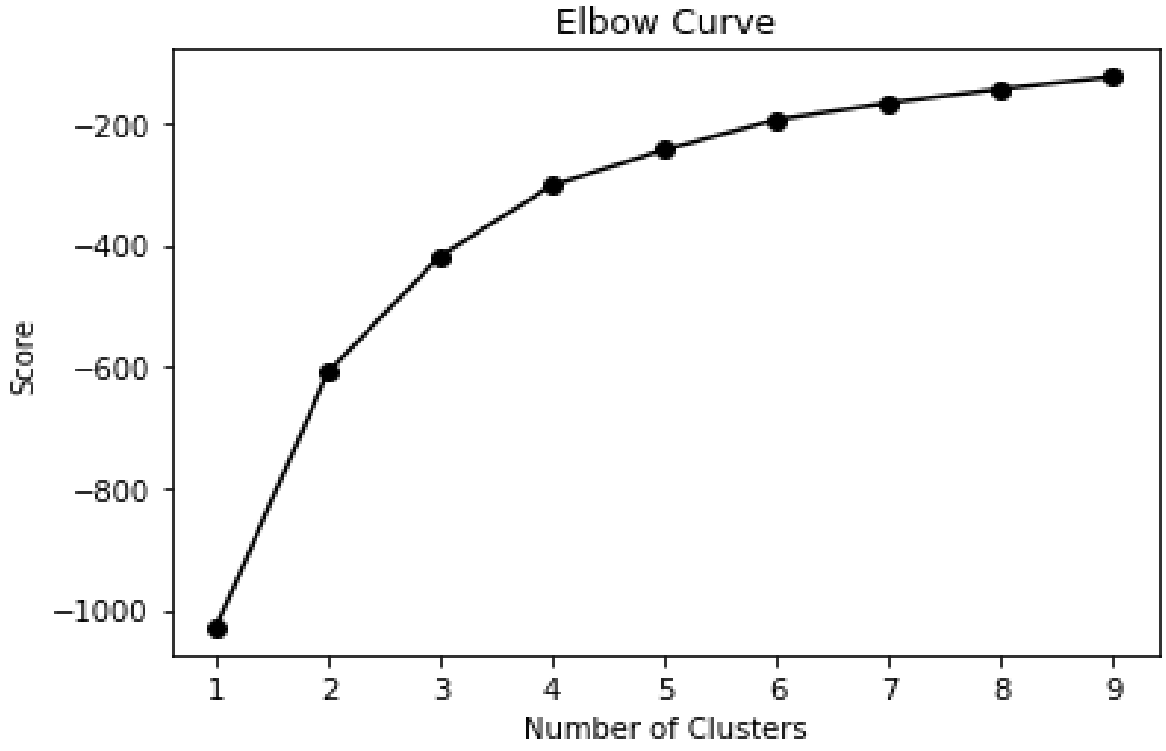


Fig 5: Elbow Curve

The three clusters: cluster 0, cluster 1 and cluster 2 have 27%, 30%, and 42% instances respectively. We took a restaurant visit per 15 days and average spending per visit as profiling attributes. The table shows that cluster0 are the population who visits and pays more, cluster2 population visits and pays a little lower than that, and cluster1 population visits and pays the least among them.

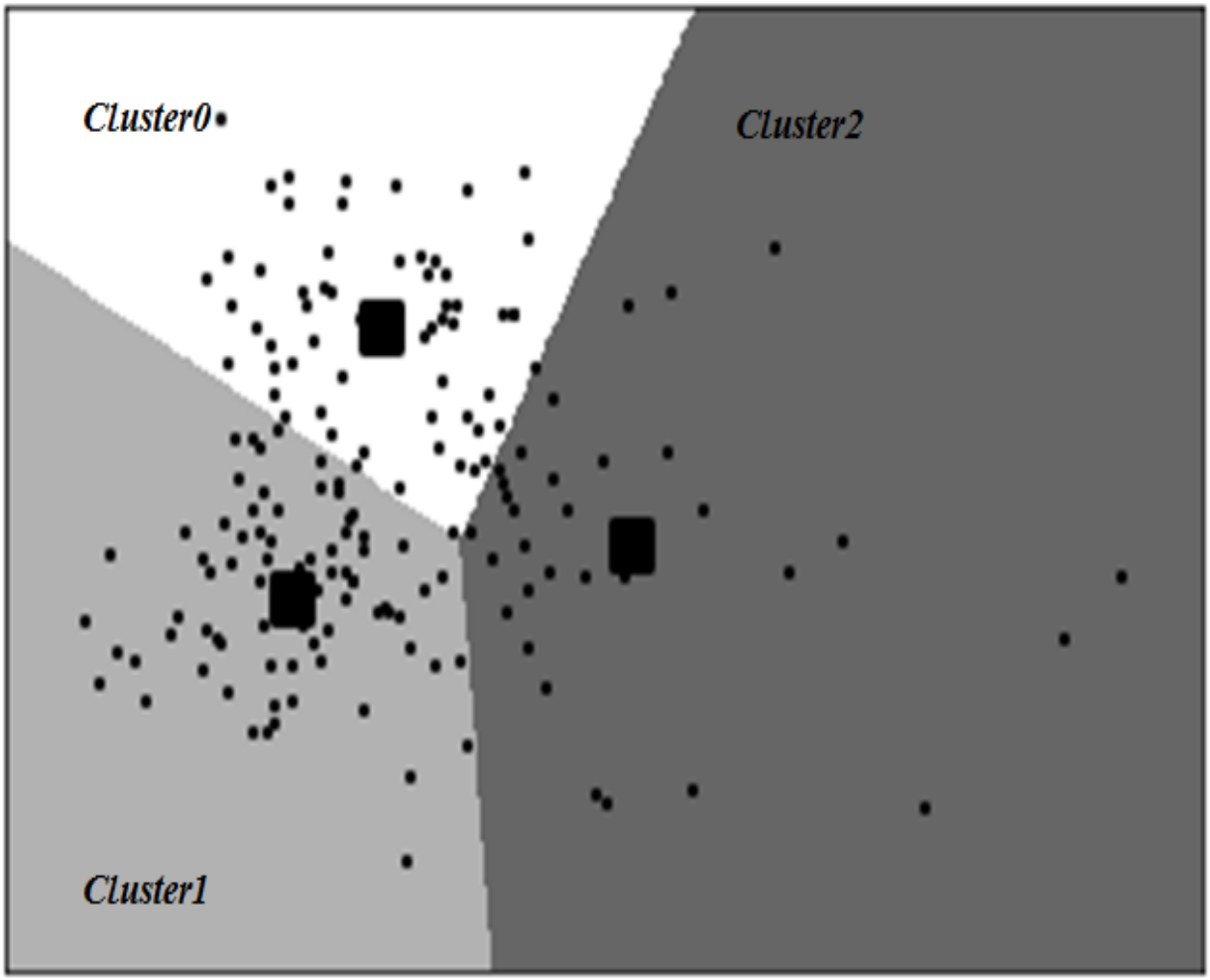


Fig 6: Scatter plot of clusters

Table 5: Cluster Findings

Cluster	Method of Cooking	Selection of ingredients	Food presentation	Traditional Symbols	Environment	Price	Service	Convenience	Elegance	Lovely	Casual	Decorations	Visit/15days	Average spend per visit
Cluster 0	1.80	2.26	3.71	2.17	4.67	4.84	4.71	4.39	4.32	3.85	3.3	4.54	3.24	460.87
Cluster 1	2.80	2.78	2.94	1.73	3.46	3.76	3.40	3.55	2.92	2.94	2.9	2.87	2.76	396.15
Cluster 2	4.20	4.06	4.34	3.33	4.73	4.52	4.62	4.15	3.97	4.06	3.9	4.13	3.06	441.66

From the table 5, the three clusters and with a mean of all features of cluster members have been shown. We see that cluster 0 and cluster 2 are the almost the same people who rated environment, price, elegance, price, service decorations and convenience highly. The

difference is that cluster 2 people additionally rated method of cooking and liveliness highly which cluster 0 people did not. Their profiling attributes also almost similar to cluster 0 people tend to visit and spend little more. Further, the cluster1 are the people who rated averagely almost all the features although they are 30% of the population. Their profiling attributes are a little lower than other clusters.

4.5 Supervised Learning

This chapter attempts to find whether supervised learning models can predict a student's fast food consumption frequency. To make the task simpler, we turn the student's fast food restaurant visiting rate per 15 days a binary classification task. The two classes are:

- Students who visit a restaurant more than thrice per 15 days (40% of instances)
- Students who visit a restaurant less than thrice per 15 days (60% of the instances)

Further, supervised learning models, decision tree classification (CART), random forest classification, logistic regression and Naive Bayes has been used to do this task. To evaluate the performance of the models, we analyzed three evaluation metrics,

- Accuracy: For a binary classification task, accuracy can be measured
  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ ……...….(5)
  It ranges from 0 to 1, with accuracy value 1 means all predictions are correct.
- Precision: It is the number of true positives divided by a total positive number. The value defines as the exactness of a model.
  $\text{Precision} = \frac{TP}{TP+FP}$……..…(6)
  It ranges from 0 to 1, with a precision value 1 means the prediction has no false positive.
- Recall: Recall is the sensitivity of the true positive rate of the data.
  $Recall = \frac{TP}{TP+FN}$……..(7)
  It ranges from 0 to 1, with recall value 1 means that the prediction has no false negative.

(TP= True Positive, TN= True Negative, FP=False Positive, FN=False Negative)

Prediction

Prediction model tends to be highly dependent on the particular problem to be solved. Considering all the relevant features of the dataset, the supervised learning model's results are:

Table 6: Comparison of accuracy between machine learning models

Model	Accuracy (%)	Precision (%)	Recall (%)	Runtime (ms)
Naive Bayes	64.7	65.5	90.5	55
Decision tree	64.7	68.8	81	530
Random forest	61.8	61.8	99.4	8000
Logistic regression	52.9	59.3	76.2	2000

The results showed that Naive Bayes, decision tree and random forest performed average, while logistic regression performed poorly. Decision tree and Naive Bayes reached the highest accuracy of 64.7%. Nevertheless, considering the balance between precision and recall, the decision tree is a slightly better option.

Feature selection

Feature selection is a process to find the most relevant features that contributing to the model's accuracy. A shortlist of features shorten runtime and increases accuracy. From the previous section, we see that all the tested models performed averagely in predicting the student's restaurant visiting frequency. To increase accuracy, feature selection methods can be applied.

Univariate feature selection

Univariate is a feature selection method that returns a shortlist of features based on different statistical scoring function. It is a pre-processing step before the model receives its features. To perform univariate, we choose F-test and chi-squared test as a scoring function.

- Chi-squared test: It performs a chi-squared test on the non-negative features and removes the features that are independent of the target label, and so has no impact on

the classification task. A small value of chi-squared indicates a strong correlation between the features.

- F-test: It scores and ranks the features by performing an ANOVA (analysis of variance) F-test between a training feature and target feature. The ANOVA F-value can be calculated as,

$$F = \frac{explained\ varience}{unexplained\ variance} \quad \text{.....(8)}$$

Shortlist of features

The univariate feature selection method is applied to the dataset using the scikit learn feature selection module. Two scoring functions, chi-squared and f-test is used as the parameter to find two sets of features.

The suggested features are shown in table 7:

Table 7: Selected Features by Univariate methods

F-Test	**Chi-squared Test**
Spending Time on the Internet	Spending Time on the Internet
Avg Expense per Visit	Avg Expense per Visit
Influenced by Offer	Influenced by Offer
Restaurant Environment	Restaurant Environment
Influenced by Price	Influenced by Price
Visit through Advertisement	Visit through Advertisement
Visit through Offer	Visit through Offer
Visit through social media	Visit through social media
Elegance	Elegance
Decoration	BMI
Food presentation	Visit through Friends
Method of cooking	Lively

From table 7, we observe that there are 9 features that are common in both the table which suggest that these 9 features may have significance in finding an optimal result for the given

model. The rest of the features are subtracted providing that when using or operation for both of the test's feature, they provide less score. So, from two of these univariate tests (F-test & Chi-Squared Test), we performed and operation which gives us 9 significant features. These 9 features are further used to train the supervised learning models. The features are:

Table 8: Final List of Features

Number	Selected Features
1	Spending Time on the Internet
2	Avg Expense per Visit
3	Influenced by Offer
4	Restaurant Environment
5	Influenced by Price
6	Visit through Advertisement
7	Visit through Offer
8	Visit through social media
9	Elegance

Training model with limited features

The model is trained again with the new list of features. The limited number of features increased the accuracy of all the models except the decision tree model.

Table 9: Comparison of accuracy between machine learning models with selected features:

Model	Accuracy (%)	Precision (%)	Recall (%)	Runtime(ms)
Naive Bayes	79.4	76.9	95.2	21
Decision tree	52.9	60.9	66.7	2000
Random forest	64.7	65.5	90.5	21000
Logistic regression	70.6	70.4	90.5	254

Table 09 shows the respective outcomes of each model. From this table, we observe that among all the models Naive Bayes has the best accuracy for the new optimized 9 features. It shows an accuracy of 79.4% and it works faster and better than other models.

CHAPTER 5

Throughout this research, we attempt to test out hypotheses. The results are shown in the following section.

5.1 Students' Current Health Status

One of the main purposes of this research was to understand university student's health status in Bangladesh, whether fast food consumption started affecting their health or not. Analyzing the data, we find no strong evidence of that. From the figure 7,8,9 & 10, we see that none of the health-related features (numbness, fatigued after eating, catch colds or flues easily, unable to lose weight, smoking habit, take medicine) has been experienced positively by a large percentage of students. The numbness has 3.53%, fatigued after eating has 15.2%, catch colds easily has 23.5%, unable to lose weight has 20.5%, the smoking habit has 16.4 and takes medicine has 23.5%. Further, only 6.46% sleeps more or less than average at night and 11.18% experience obesity Therefore, it can be said that fast-food consumption has its own demerits, but it has not reached an alarming state in Bangladesh.

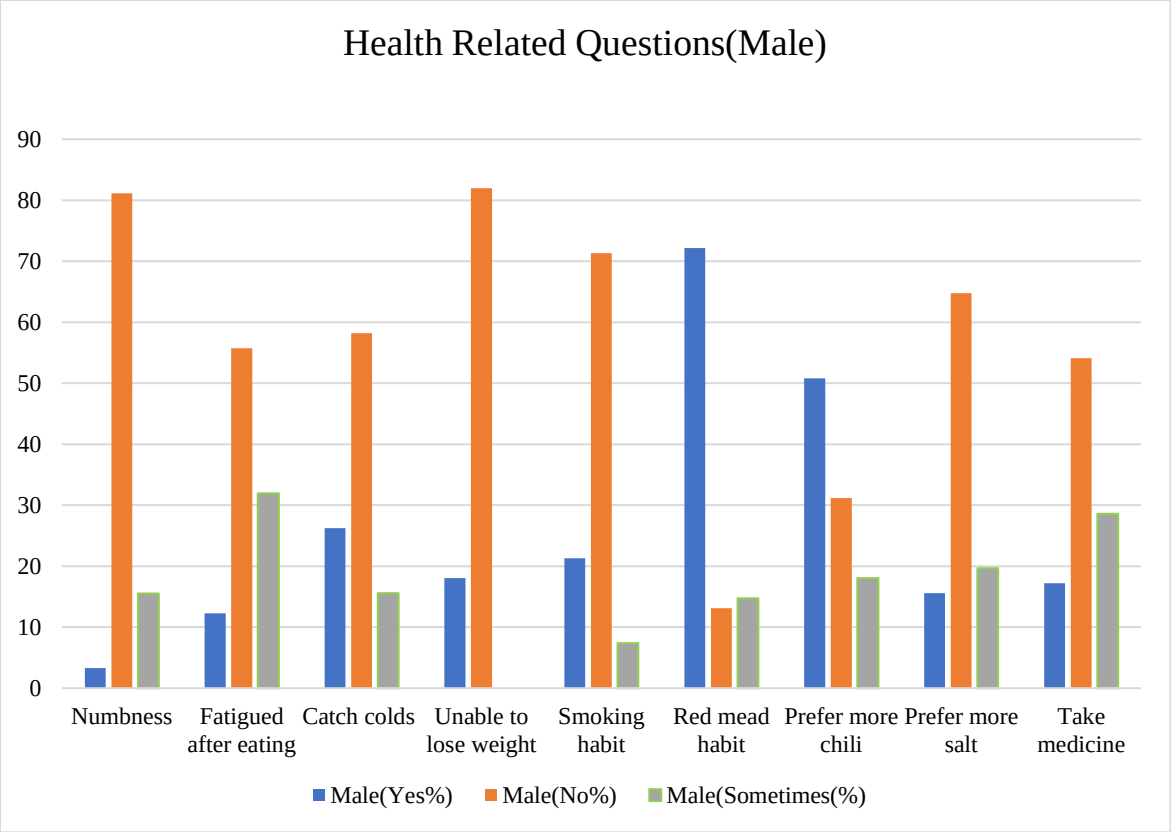


Fig 7: Health-Related answers between all the male responders

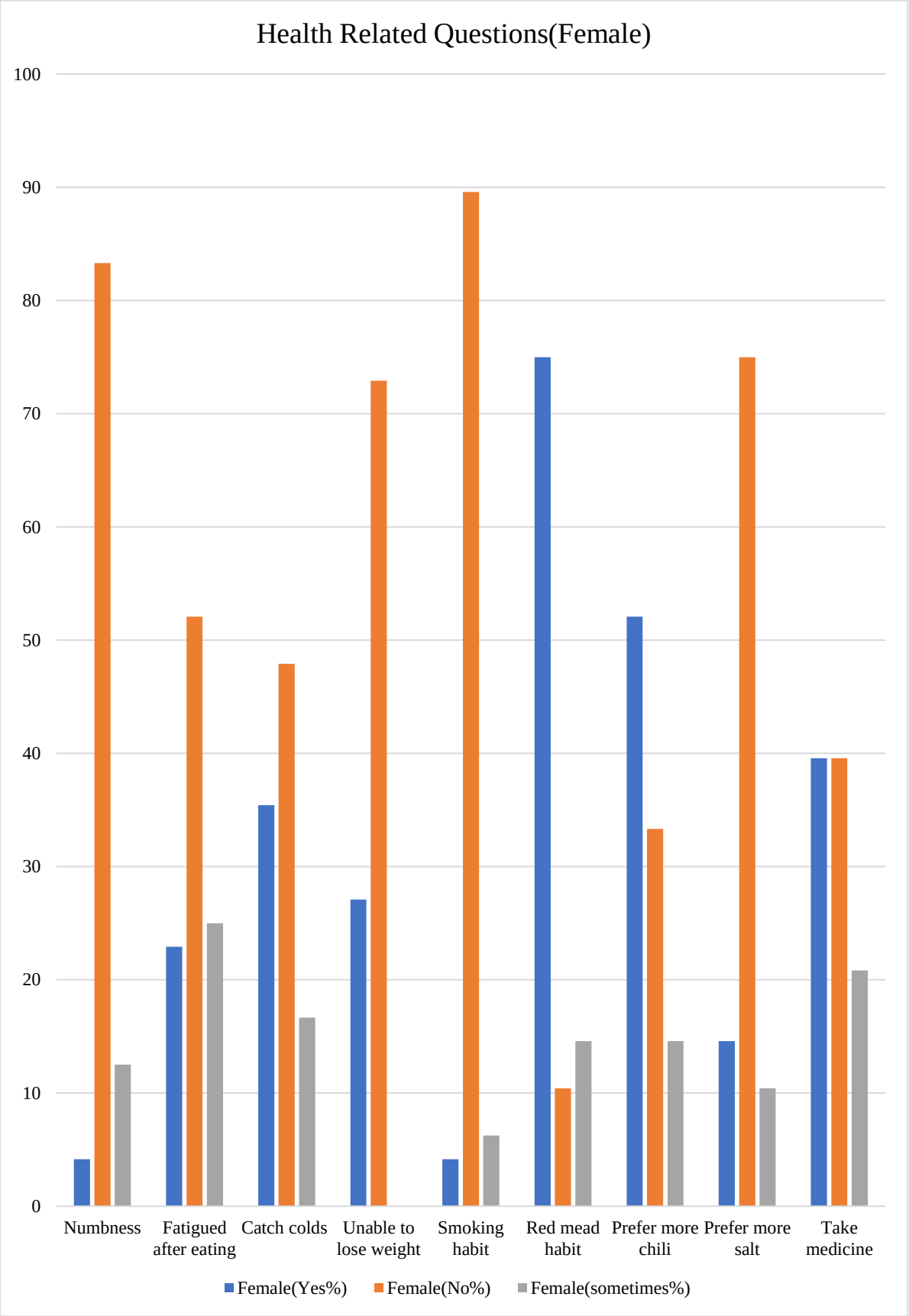


Fig 8: Health-Related answers between all the female responders

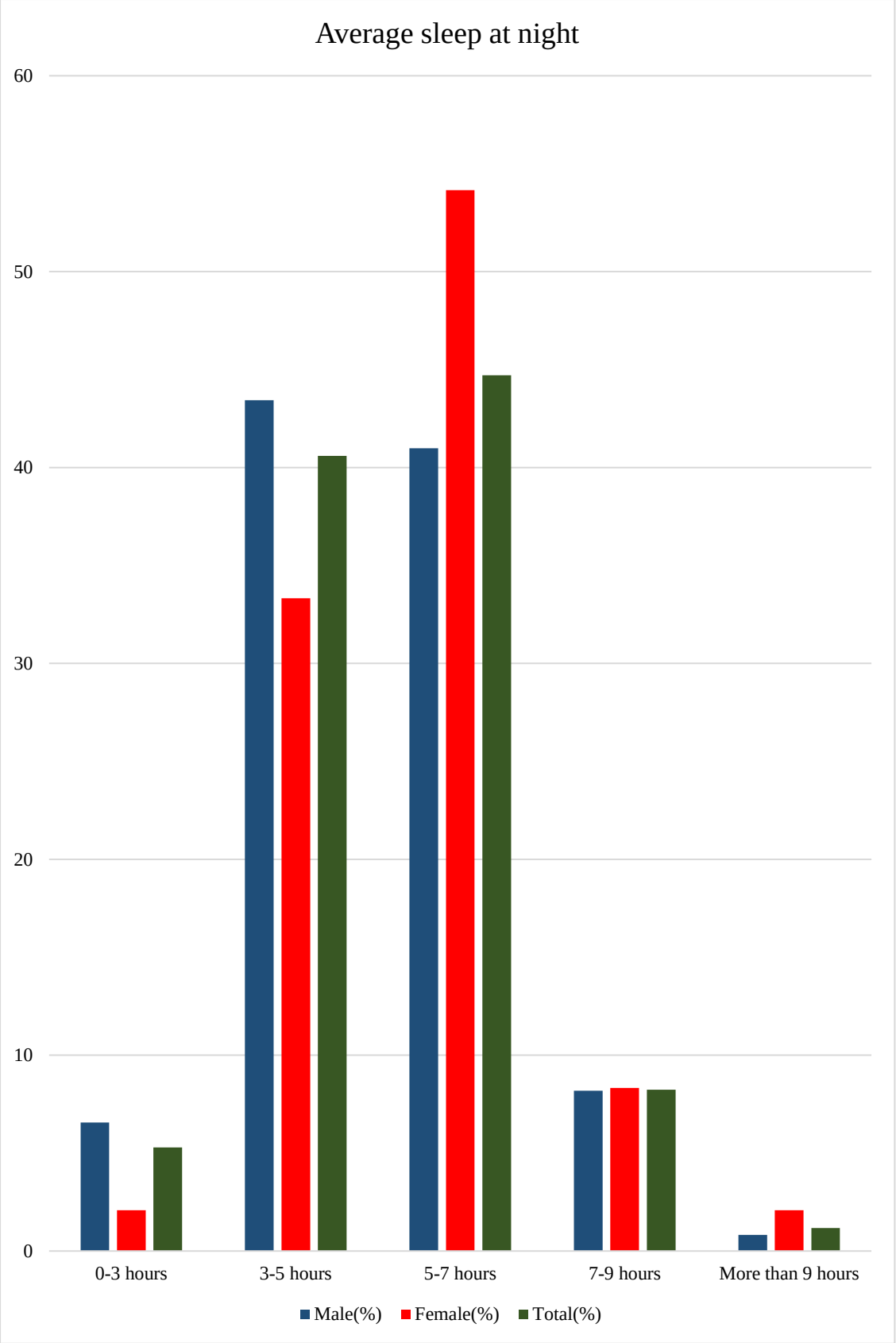


Fig 9: Average sleep at night among all the genders

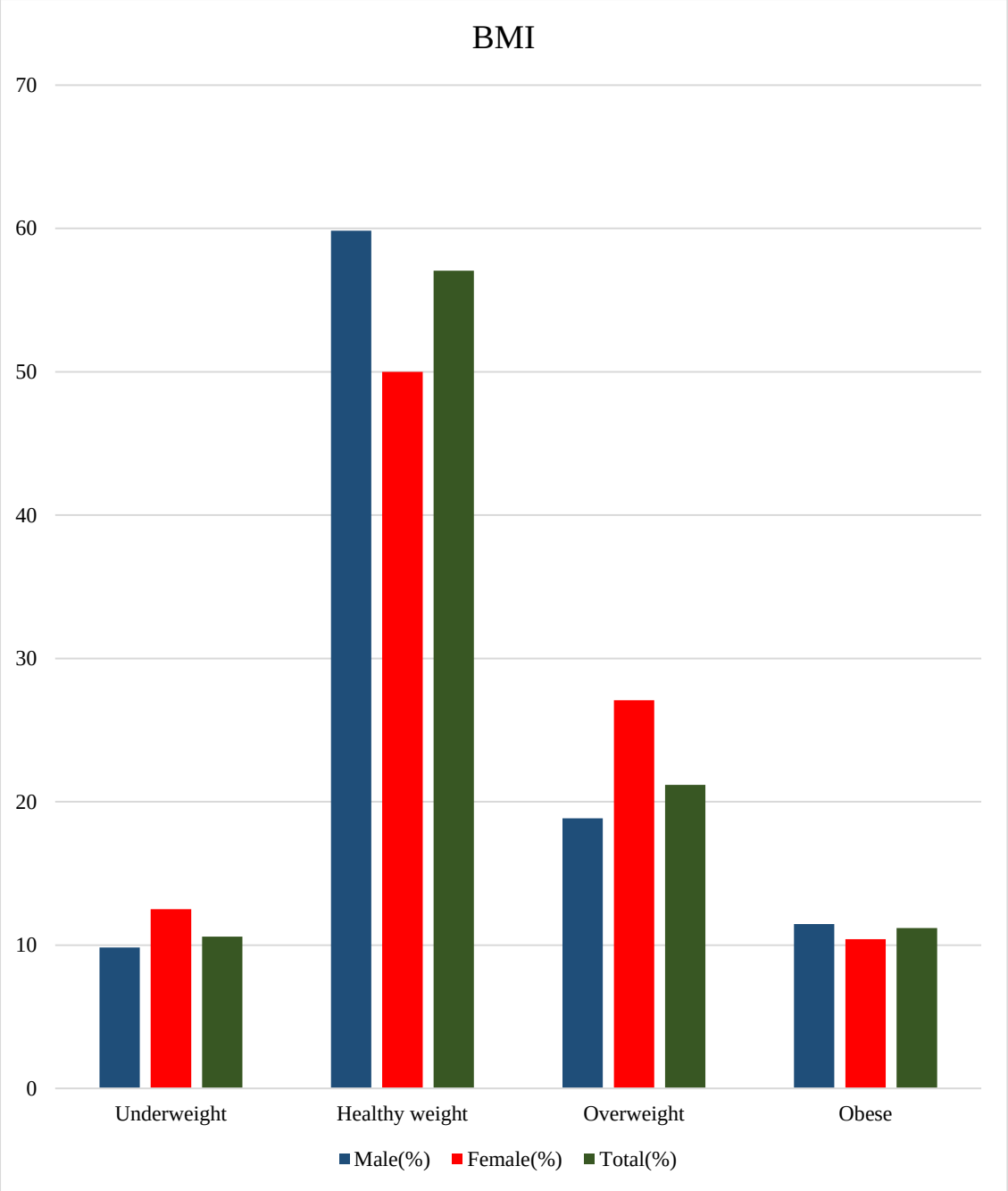


Fig 10: Body mass index ratio

5.2 Price matters more than a restaurant's attributes

Students are influenced by the price of the food more than the restaurant's food and environment related attributes, like decoration, food standard, ambience, food presentation, smoking zone etc. In section 4.5, in predicting, a student's fast food restaurant visit frequency, we see that, most of the restaurant's attribute related features (food presentation, method of cooking, ingredients of food, convenience, smoking zone, decoration, environment etc.) are not strong predictors except elegance and environment of restaurant. But elegance and

environment have a weak positive correlation with the target feature. On the other hand, price rating of a restaurant and average spending or budget are strong predictor in predicting a student's fast food restaurant visiting frequency. In addition to, these features have positive correlation with these features. Which means that, the more budget someone have, he/she tends to more visit a restaurant (table 10).

Table 10: Correlation between selected features and target feature

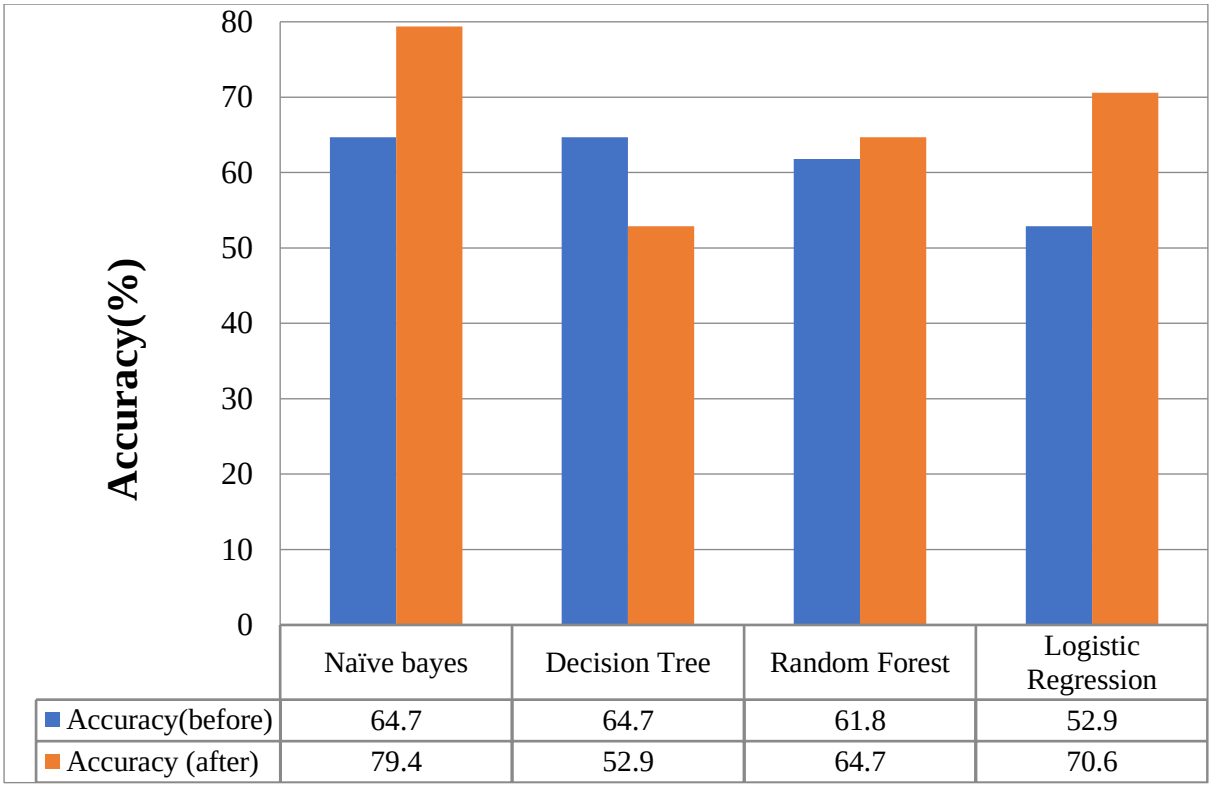
Number	Selected Features	Pearson's Correlation (Target)
1	Spending Time on the Internet	0.232
2	Avg Expense per Visit	0.167
3	Influenced by Offer	0.159
4	Restaurant Environment	0.192
5	Influenced by Price	0.203
6	Visit through Advertisement	0.317
7	Visit through Offer	0.361
8	Visit through social media	0.389
9	Elegance	0.143

5.3 Internet, social media, advertisement and offer

In section 4.5, we see that time spent on internet, social media promotions and offers from a restaurant are strong predictors in predicting a student's fast food consumption rate. Social media marketing is getting more popular these days, and in this restaurant business sector, it is attracting more student customers. In addition to, different discount offers offered by restaurants also influences students (table 10).

5.4 Prediction task

In chapter 4, we attempt to predict student's fast food consumption rate by making it a binary classification task using popular machine learning models. With Gaussian Naive Bayes classifier, we could reach up to 79.4% accuracy which is quite acceptable considering the size of the dataset and the task. These conclude that among our initial choice of features has some relation with university students' fast food consumption behavior (fig 11).



	Naïve bayes	Decision Tree	Random Forest	Logistic Regression
Accuracy(before)	64.7	64.7	61.8	52.9
Accuracy (after)	79.4	52.9	64.7	70.6

Fig 11: Predicting food consumption behavior

5.5 Other findings

From k-means clustering experiment with the restaurants food and environment related features and sampling them with different type of restaurant user was not quite good. The mean of the sampling features has no significant variance. Thus, it cannot capture different type of users effectively. Also, with chi-squared test, we see that, fast food consumption behavior (visiting frequency and average spend per visit) do not depend on sex.

In a summary we can say that, fast food consumption behavior among Bangladeshi students has not reached to a high rate, also its impact on health has not reached to a high rate. Fast food consumption behavior does not depend on sex. In addition to, students are more likely to be influenced by internet, social media, advertisement and price rather than different restaurants' food and ambience related features.

CHAPTER 6

6.1 Conclusion

The principal motivation behind this examination venture is to play out an exploration on youthful age's fast-food propensity, wellbeing status, and making expectations for nourishment on the setting of Bangladesh. We played out certain examination on our review dataset to pick up bits of knowledge on youthful understudy's well-being status and their fast-food utilization propensity. In light of the survey, we can watch that since there is altogether almost no relationship between the health-related traits, so we can accept the health risk of the students is low. Distinctive informative information investigations incorporate connection, disperse plots discovered a few bits of knowledge from the information, although inadequacy of tests and clamour confined us to make any solid case.

6.2 Future Work

There are several tasks that we have to complete in the further part of our research. Firstly, the dataset is very small, so we have to manage more responses to our survey. Secondly, we have to remove outliers and noise from the dataset to train our models best fit for the machine learning algorithms. Further we have to consider other ways of fast food consumption. In addition to, we could not find any direct linear relationship between the health-related features. Therefore, we have to look for non-linear relationships between them. Besides, more clustering algorithms have to be tested to find the best model that best fits the data.

REFERENCES

[1] N. Islam & G. M. S. Ullah, "Factors Affecting Consumers Preferences On Fast Food Items In Bangladesh," *Journal of Applied Business Research (JABR)*, vol. 26, no. 4, 2010.

[2] Kakkar, P. and R. J. Lutz , "Toward a Taxonomy of Consumption Situations," in Combined Proceedings, Series 37, ed. E. M. Maze, Chicago: American Marketing Association, pp. 206-210, 1975

[3] S. W. Wibowo, & M. Tielung, "ANALYTICAL HIERARCHY PROCESS (AHP) APROACH ON CONSUMER PREFERENCE IN FRANCHISE FAST FOOD RESTAURANT SELECTION IN MANADO CITY". Jurnal EMBA, Vol.4 No.2 June 2016, Hal. 022-028,2016

[4] Tan C, Chi E.H, David H, and Smola A.J, Instant foodie: predicting expert ratings from grassroots, Proceedings of the 22nd ACM international conference on Information & Knowledge Management(2013).

[5] Mandoura N, Al-Raddadi R, Abdulrashid O, Shah H. B. U, Kassar S.M, Hawari A.R.E, & Jahhaf J.M, Factors Associated with Consuming Junk Food among Saudi Adults in Jeddah City.Cureus,open access journal(2017).

[6] Fried D, Surdeanu M, Kobourov S, Hingle M, Bell D. Analyzing the language of food on social media. In: Big Data (Big Data), 2014 IEEE International Conference on. IEEE; 2014. p. 778–783.

[7] Mihuandayani, Ramandita H.D, Setyanto A, Sumafta I.B, Food Trend Based on Social Media for Big Data Analysis Using K-Mean Clustering and SAW:A Case Study on Yogyakarta Culinary Industry, 2018 International Conference on Information and Communications Technology (ICOIACT),2018

[8] S. Auty, "Consumer Choice and Segmentation in the Restaurant Industry", The Service Industries Journal Volume 12, 1992 - Issue 3, 2006Kanungo T, Mount D.M., Netanyahu N.S, C.D., Piatko C.D, Silverman R, Wu A.Y., An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence(2002).

[9] Kodinariya T.M, & Makwana P.R, Review on determining number of Cluster in K-Means Clustering, International Journal of Advance Research in Computer Science and Management Studies(2013).

[10] B. Shanthy A., Steven L. Gortmaker, PhD; Cara B. Ebbeling, PhD; Mark A. Pereira, PhD; and D. S. Ludwig, MD, PhD,"Effects of Fast-Food Consumption on Energy Intake and Diet Quality Among Children in a National Household Survey",PEDIATRICS Vol. 113 No. 1,,2004

[11] I,Rish, "An empirical study of the naïve Bayes classifier", IJCAI 2001 workshop on empirical methods in artificial intelligence,2001.

[12] Usama M.FayyadKeki B.Irani,"On the handling of continuous-valued attributes in decision tree generation",Machine Learning,January 1992, Volume 8,Issue 1,pp87-102,1992.

[13] P. Gupta, “Decision Trees in Machine Learning – Towards Data Science,” *Towards Data Science*, 17-May-2017. [Online]. Available: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052. [Accessed: 23-Jul-2018].

[14] N. Donges, “The Random Forest Algorithm – Towards Data Science,” *Towards Data Science*, 22-Feb-2018. [Online]. Available: https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd. [Accessed: 23-Jul-2018].

[15] “The Logistic Regression Algorithm,” *machinelearning-blog.com*, 23-Apr-2018. [Online]. Available: https://machinelearning-blog.com/2018/04/23/logistic-regression-101/. [Accessed: 23-Jul-2018]

[16] “Using Chi-Square Statistic in Research,” *Statistics Solutions*. [Online]. Available: http://www.statisticssolutions.com/using-chi-square-statistic-in-research/. [Accessed: 23-Jul-2018].

[17] I. C. Barić, Z. Šatalić, and Ž. Lukešić, “Nutritive value of meals, dietary habits and nutritive status in Croatian university students according to gender,” *International Journal of Food Sciences and Nutrition*, vol. 54, no. 6, pp. 473–484, 2003.

[18] M. I. K. V. Bothmer and B. Fridlund, “Gender differences in health habits and in motivation for a healthy lifestyle among Swedish university students,” *Nursing and Health Sciences*, vol. 7, no. 2, pp. 107–118, 2005.