

Behavioural Analysis of Factors Influencing Online Food Ordering and Its Relation to Obesity

S. M. Fazle Rabby Labib^{a,*}, Fahmida Zaman Achol^a, Md. Aqib Jawwad^a, Dr. Md. Golam Rabiul Alam^a, Dr. Md Iftekharul Mobin^c and Dr. Ashis Talukder^b

^aBRAC University, Kha 224 Bir Uttam Rafiqul Islam Avenue, Merul Badda, Dhaka 1212, Bangladesh

^bUniversity of Dhaka, Nilkhet Rd, Dhaka 1000, Bangladesh

^cAmerican International University Bangladesh, Kuratoli., Dhaka 1229, Bangladesh

ARTICLE INFO

Keywords:

Online Food Ordering
Obesity
Behavioural Analysis
Machine Learning
Customer Segmentation
Hypothesis Testing


ABSTRACT


This research endeavors to unravel the determinants shaping individuals' decisions to engage with online food ordering (OFO) services and the subsequent intensity of their ordering habits. Additionally, it explores whether a correlation exists between OFO and obesity, encompassing diverse consumer demographics. Drawing insights from an online questionnaire survey with 343 participants aged 16–30, our study employs a multifaceted approach, incorporating Ordinary Least Squares, Decision Tree, and Random Forest methodologies. The identified influential factors span impulsive decision-making, ordering convenience, variety of options, and the allure of promotional offers and discounts. Statistical analyses reveal nuanced patterns, indicating a weak positive relationship between obesity and the frequency of online food ordering (p -value = 2.43×10^{-7} , coefficient = 0.28). To analyze consumer behavior, our study utilizes diverse classifiers such as Logistic Regression, Naive Bayes, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Machine, and K-Nearest Neighbors. Notably, Random Forest emerges as a robust classifier, achieving an accuracy of 81% in classifying frequent users of OFO. Customer segmentation through K-Means clustering uncovers varied preferences. Users with infrequent OFO habits often perceive the services as expensive, while frequent users express concerns about the nutritional content of the ordered food. Chi-squared tests highlight significant associations, indicating that tech savvy correlates with the perception of OFO services as user-friendly. Moreover, globally, individuals leading hectic lifestyles often exhibit impulsive decision-making tendencies, with the affordability of online food ordering (OFO) services closely tied to ordering behavior shaped by promotional offers and discounts. These findings, accompanied by developed prediction models and a customer segmentation approach, contribute valuable insights to the understanding of universal consumer behavior in an ever-evolving, digitalized food landscape.

1. Introduction

Online Food Ordering (OFO) Services have globally transformed meal accessibility, offering unparalleled convenience compared to traditional dining. However, this convenience raises health concerns, contributing to major issues such as high blood pressure, obesity, diabetes, cardiovascular diseases and various types of cancer. The World Health Organization (WHO) [1] notes a tripling of obesity rates since 1975, emphasizing its urgency as a global public health problem. During the COVID-19 pandemic, the surge in OFO services has redefined how we procure meals globally, including in Bangladesh. Platforms like foodpanda, HungryNaki, and Pathao Food have become integral to modern food consumption. Despite their widespread use, the factors [2] influencing individuals to adopt these services and their potential contribution to the global rise in obesity [3] remain largely unexplored.

*Corresponding author: S. M. Fazle Rabby Labib, Phone: +880 1623925656. Department of Computer Science & Engineering, Brac University, Kha 224 Bir Uttam Rafiqul Islam Avenue, Merul Badda, Dhaka 1212, Bangladesh

 s.m.fazle.rabby.labib@g.bracu.ac.bd (S.M.F.R. Labib); fahmida.zaman.achol@g.bracu.ac.bd (F.Z. Achol); md.aqib.jawwad@g.bracu.ac.bd (Md.A. Jawwad); rabiul.alam@bracu.ac.bd (Md.G.R. Alam); iftekhar.mobin@aiub.edu (M.I. Mobin); ashis@du.ac.bd (A. Talukder)

 www.bracu.ac.bd (S.M.F.R. Labib); www.bracu.ac.bd (F.Z. Achol); www.bracu.ac.bd (Md.A. Jawwad); www.bracu.ac.bd (Md.G.R. Alam); www.aiub.edu (M.I. Mobin); www.du.ac.bd (A. Talukder)

ORCID(s): 0009-0008-5397-3881 (S.M.F.R. Labib); 0009-0001-6119-2179 (F.Z. Achol); 0009-0001-0043-5798 (Md.A. Jawwad); 0000-0002-9054-7557 (Md.G.R. Alam); 0000-0002-3065-2486 (M.I. Mobin); 0000-0003-2991-9136 (A. Talukder)

Bangladesh, as a valuable test case, provides insights into how OFO services impact diverse cultural and geographical contexts. Understanding these dynamics on a global scale is essential for devising effective strategies to address health implications associated with the widespread adoption of online food ordering services.

The linkage between the widespread popularity of Online Food Ordering (OFO) services and the rising prevalence of obesity raises significant concerns. According to the National Health Service (NHS) [4], obesity is attributed to excessive calorie intake and inadequate physical activity. OFO platforms, often offering highly processed, calorie-rich food options with limited nutritional information, may contribute to this severe trend.

This research aims to address this knowledge gap by investigating the factors influencing OFO service usage and its potential correlation with the increasing rates of obesity. Our qualitative survey concentrates on the Bangladeshi population, serving as a case study scenario. The difficulty arises from the limited availability of comprehensive data on this matter, both on a global scale and specifically within Bangladesh. Thus, this research aims to address the following questions:

1. What factors influence users' decisions to order food through online services?
2. Does the intensity of ordering correlate with the increase in obesity rates
3. Does the intensity of ordering correlate with the increase in obesity rates
4. Can users be effectively classified and segmented into multiple clusters based on their behaviour?

Within the framework of this research, we curate a pioneering dataset focusing on Bangladeshi customers engaging with Online Food Ordering (OFO) services. Our objective is to explore the determinants influencing the adoption of OFO services and investigate potential correlations between Body Mass Index (BMI) and monthly order frequencies. Additionally, we craft a predictive model designed to classify users exhibiting frequent online food ordering behaviors [5]. Applying advanced customer segmentation techniques, we categorize the user base into two distinct clusters, facilitating a detailed and nuanced analysis. To further deepen our insights, we undertake hypothesis testing to unveil relationships between various factors. Its primary contributions can be summarized as follows:

1. We identify the factors driving the adoption of OFO services among Bangladeshi consumers, utilizing techniques such as Decision Tree, Random Forest, and Ordinary Least Squares to assess feature importance.
2. We conduct hypothesis testing to determine the relation among influencing factors using the Chi-squared test of independence.
3. We have found that even though it is very weak, there exists a correlation between obesity and orders per month. We also conducted a correlation analysis between other BMI categories and orders per month as well as between physical activity, order history, and orders per month. For this, we used Pearson Correlation, Spearman's Rank Correlation, and Point Bi-serial Correlation.
We explore the correlation, albeit weak, between BMI and monthly order frequencies using various correlation analyses such as Pearson, Spearman's Rank, and Point Bi-serial correlations. Additionally, we examine the correlations among BMI categories, physical activity, order history, and order frequencies.
4. We develop a model to classify frequent users of OFO services, employing classifiers such as Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Gradient Boosting Machine, and K-Nearest Neighbors.
5. We apply customer segmentation techniques, specifically K-Means, and PCA, to find user clusters, offering insights into consumer behaviour.
6. We create a novel dataset of Bangladeshi OFO service users, a significant addition to the Bangladeshi research landscape.

2. Literature Review

The surge in online food ordering (OFO) has reshaped consumer habits, which brings us to explore its multifaceted impact. Here, we examine various studies related to OFO, encompassing its influence on consumer behaviour, health implications like obesity, and the factors driving its popularity.

Kale *et al.* [6] highlight convenience and flexibility as prime reasons behind the preference for using OFO which is supported by Alagoz *et al.* [7] who emphasize the significance of an easy ordering process, particularly among university students. Similar results can be found in the study conducted by Vinish *et al.* [8] which indicates convenience of placing orders, the quality and availability of food, restaurant reviews, offers and discounts, quick delivery, and a wide choice of restaurants are crucial for customer satisfaction. On the other hand, Parameshwaran *et al.* [9] and Anam

et al. [10] delve into the influence of advertising, social factors, and the shift to online platforms, accentuated during the COVID-19 pandemic.

In examining health ramifications, Chatterjee *et al.* [11] use machine learning methods to explore obesity risk factors. They emphasize the need to understand the lifestyle choices of consumers. Prentice *et al.* [12] shed light on the role of energy-dense food in obesity, stressing its prevalence in fast food offerings and its adverse impact, especially among children. Meanwhile, studies by Harahap *et al.* [13] and Kurniawati *et al.* [14] emphasize the correlation between online food ordering, dietary choices, and obesity risks among university students, offering insights into nutritional aspects and consumption patterns, albeit with differing conclusions. Whereas Harahap *et al.* suggest OFO and fast food consumption lead to overconsumption which contributes to obesity, Kurniawati's study reveals the nutritional content of the meals ordered online does not establish a direct link between OFO and the nutritional state of college students. This study by Dana *et al.* [15] highlights that younger users with higher BMI tend to use OFO services.

Studies conducted by Hong *et al.* [16] and Ayubi *et al.* [17] analyze customer intentions, behavioural aspects, and influences across demographics and regions. Similarly, Jahidi *et al.* [18] and Inthong *et al.* [19] concur, highlighting the potential variation in motivations across different regions.

Despite extensive research, gaps persist, particularly regarding the Bangladeshi perspective. Existing studies primarily hail from diverse geographical locations, with datasets unavailable for future research. This gap underscores the need for localized studies in Bangladesh to understand cultural nuances and their influence on OFO dynamics. In contrast to prior methodologies employing behavioural intention models like TAM [20] and UTAUT [21] for data analysis, our approach leverages machine learning techniques.

In summary, while existing studies provide invaluable insights into OFO's global impact, there's a clear need for localized research in Bangladesh. Understanding how OFO affects Bangladeshi consumers, considering cultural, socioeconomic, and dietary differences, will be crucial for informing policies and practices in this evolving landscape.

3. Methodology

In figure 1, we can see the top-level overview of our proposed system. We start by collecting responses from the survey and constructing our dataset. We do data pre-processing and test for validity and reliability. Then we do correlation analysis, find factors that influence the use of OFO services, create a prediction model, apply customer segmentation to our data and do some hypothesis testing. Finally, we show our findings.

3.1. Data Collection

To investigate the factors that influence the decision to order food online among individuals in Bangladesh, a survey is conducted. The survey questions are divided into two parts. One is the demographic data and the other is questions related to online food ordering and the factors influencing it. There were in total 25 questions in the questionnaire. The survey is done anonymously to protect the respondent's privacy. The survey is done through Google Forms and is designed to take the smallest amount of time possible. To accomplish that, we first share the form with our peers and take feedback from them. We incorporate the feedback and make it public. The questions ranged from text-based, multiple-choice questions (MCQ), checkboxes, and Likert scale questions.

Throughout the data collection period, we received several queries regarding why a respondent needs to log in to their email to respond to our questions. We explained to them that this is necessary to protect the survey from getting spam and random responses. Another concern we faced, was about the user being reluctant to share their personal information like height and weight we assured them that none of the answers were connected to their emails. We also made sure to not force anyone to share their information and only took responses if they gave it willingly.

3.2. Data pre-processing

After collecting the survey responses, the column names are modified for improved readability and ease of use. The height variable, which was initially recorded in feet and inches, gets converted to meters for consistency. Using the converted height and weight variables, the BMI of each respondent is calculated. A new variable, BMI_CAT, is created to store the categorical values of BMI, including "Underweight", "Normal", "Overweight" and "Obese". To maintain the focus on the target population of individuals aged 16-30, any responses from individuals outside of this age range are removed, as about 98% of the data consists of individuals within this age group. The Likert scale data obtained from the questions related to various factors are encoded into ordinal numerical values for ease of analysis. This conversion allowed for a more accurate statistical analysis of the collected data. We remove any responses where

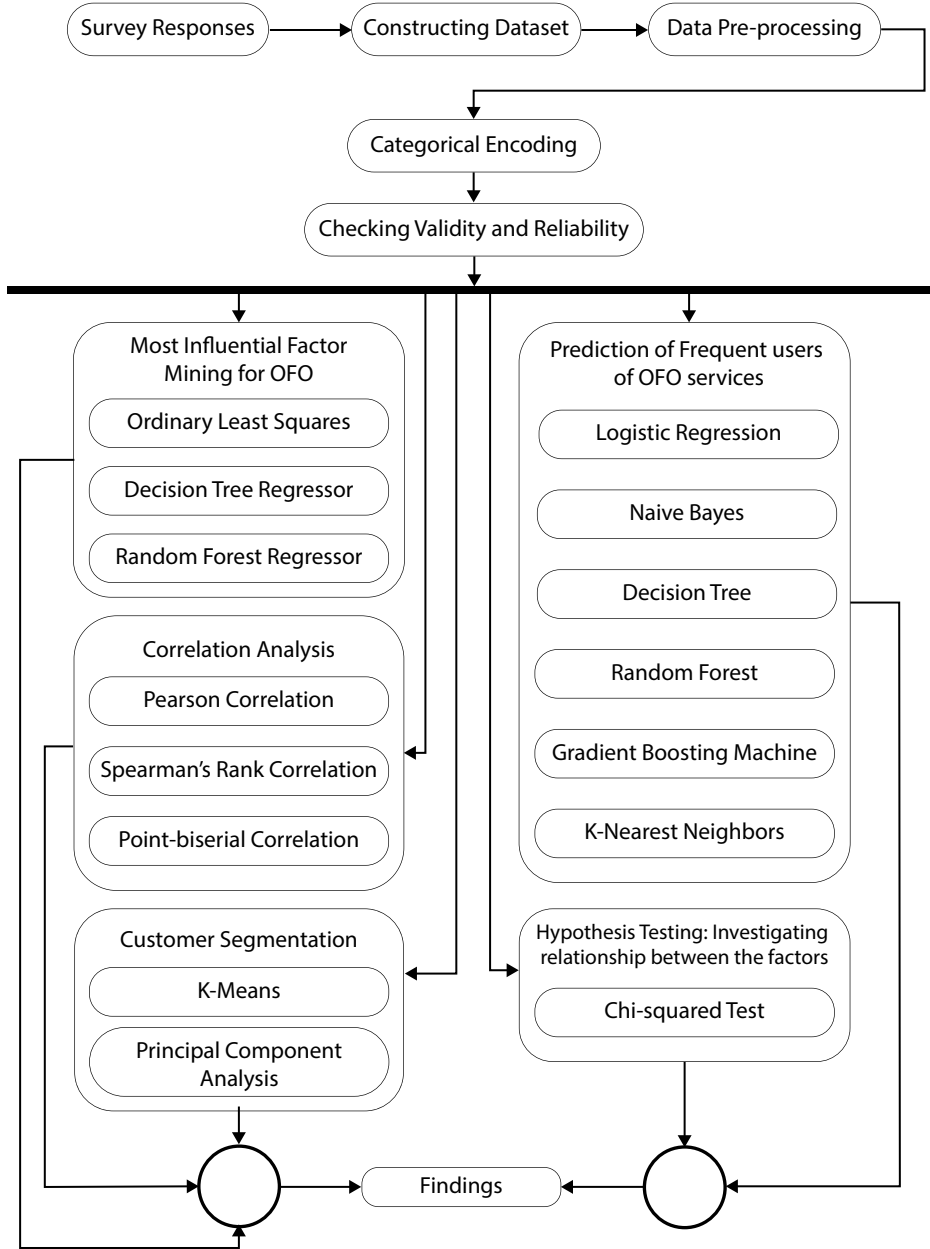


Figure 1: Top level overview of the proposed system.

orders per month are less than 1 since that information is not useful for our research problem. After pre-processing we ended up with valid 331 responses which we used to conduct the rest of our research. A sample set of the dataset after pre-processing can be seen in table 2

3.3. Dataset Validity and Reliability

Before applying all of the discussed methods we use Cronbach's Alpha [22], which is a way of assessing reliability by comparing the amount of shared variance, or covariance, among the items making up an instrument to the amount of overall variance. This is done to test the validity and reliability of our dataset. The idea is that if the instrument is reliable, there should be a great deal of covariance among the items relative to the variance.

Table 1
Questionnaire Description.

Attributes	Data Type
Gender	Categorical Value
Age	Numeric Ranges
Weight	Numeric Value (Kilogram)
Height	Numeric Value (Feet and Inches)
Educational qualification	Categorical Value
Employment status	Categorical Value
Financial dependency	Categorical Value
Marital status	Categorical Value
Physical activity	Numeric Range: 1-7 (per week)
Preferred OFO apps	Categorical Value
OFO apps use duration	Categorical Value
Orders per month	Numeric Value
Ordering time	Categorical Value
Types of food ordered	Categorical Value
Tries out new technologies	Likert Scale
Makes impulsive decisions	Likert Scale
Too busy to cook	Likert Scale
Does not like to cook	Likert Scale
Ordering online is easy	Likert Scale
Variety of options	Likert Scale
Ordering is inexpensive	Likert Scale
Promo codes and discounts	Likert Scale
Food is delivered quickly	Likert Scale
Food is safe and hygienic	Likert Scale
Food is nutritious	Likert Scale

After applying Cronbach's Alpha on our dataset we get the value of 0.89. According to this paper by Taber [23] a Cronbach's alpha value higher than 0.70 is considered acceptable and above 0.80 and above is considered better. Therefore we can safely say that our dataset is valid and reliable.

3.4. Most Influential Factor Mining for OFO

To find factors that are influencing a person to use OFO, we use Ordinary Least Squares (OLS) [24], Decision Tree Regressor [25] and Random Forest Regressor [26].

Ordinary Least Squares: This is a linear least squares method that chooses unknown parameters in a linear regression. This method tries to find the line that fits the data points the best by reducing the total amount of error between the predicted values and the actual values of the dependent variable. The model takes the form of an equation:

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \quad (1)$$

where y is the dependent variable, x_1, x_2, \dots, x_n are the independent variables, and $b_0, b_1, b_2, \dots, b_n$ are the coefficients of the equation. The goal of OLS is to find the values of the coefficients that minimize the difference between the predicted values of y (based on the equation) and the actual values of y . A system of linear equations is solved to find the coefficients. The matrix representation of this system is $X'Xb = X'y$, here X is the matrix of independent variables, b is the vector of coefficients and y is the vector of the dependent variable. These coefficients can tell us how strongly and in which way the independent variable affects the dependent variable. The one with the biggest coefficient is the most important factor in influencing the dependent variable.

Decision Tree: This method uses a tree-shaped model to make decisions and predict outcomes. The tree is created by splitting the data into smaller groups based on the values of the independent variables. This process is repeated recursively until we have subsets that can predict the consequences of each decision. At each internal node of the tree,

Table 2

Sample of the Dataset after pre-processing. Due to its length, the dataset is divided into 5 subtables.

Gender	Age	Weight	Height	BMI	BMI_CAT	Education	Employment_Status
Female	16-30	80	1.63	30.27	Obese	Undergraduate	Prefer not to say
Male	16-30	90	1.75	29.30	Overweight	Undergraduate	Seeking opportunities
Male	16-30	75	1.68	26.69	Overweight	Undergraduate	Prefer not to say
Male	16-30	78	1.78	24.67	Normal	Undergraduate	Prefer not to say
Male	16-30	98	1.78	31.00	Obese	Undergraduate	Part-time

(a)

Financial_Dependency	Marital_Status	Physical_Activity	Ordering_History	Ordering_Time
Partially dependent	Single	3	Less than 3 months	Evening Snacks, Dinner
Fully Dependent	Single	0	More than 6 months	Lunch, Dinner, Midnight snacks
Fully Dependent	Single	0	More than 6 months	Evening Snacks
Fully Dependent	Single	5	More than 6 months	Lunch, Evening Snacks, Dinner
Partially dependent	Single	1	Less than 3 months	Evening Snacks, Dinner

(b)

Food_Types	New_Tech	Impulsive_Decision	Busy_To_Cook
Chinese, Pizza, Pasta, Fried chicken, French fries, Milkshake	Agree	Strongly Agree	Disagree
Indian, Burger, Pizza, Pasta, Fried Chicken, Cake	Strongly Agree	Neutral	Neutral
Burger, Pizza, Fried chicken, French fries	Neutral	Neutral	Disagree
Burger, Pizza, Biryani/Khichuri/Tehari	Neutral	Agree	Neutral
Pizza, Fried chicken, French fries, Biryani/Khichuri/Tehari	Agree	Agree	Neutral

(c)

Dont_Like_To_Cook	Ordering_Easy	Options_To_Choose	Ordering_Inexpensive	Promo_Discounts
Disagree	Strongly Agree	Agree	Neutral	Strongly Agree
Disagree	Agree	Strongly Agree	Disagree	Agree
Disagree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree
Neutral	Agree	Agree	Agree	Strongly Agree
Disagree	Strongly Agree	Strongly Agree	Strongly Agree	Strongly Agree

(d)

Quick_Delivery	Food_Safety	Nutritious	OFO_APPS	Order_Frequency	Order_Per_Month
Neutral	Disagree	Strongly Disagree	foodpanda, Pathao	≤ 5	4
Agree	Neutral	Neutral	foodpanda, Pathao	≤ 5	2
Agree	Neutral	Neutral	foodpanda	≤ 5	3
Neutral	Agree	Neutral	Pathao	≤ 5	5
Neutral	Neutral	Neutral	foodpanda	≤ 5	1

(e)

a decision rule is applied to the data, and the tree branches accordingly. Once the tree is built, we use it to identify the most important variables for ordering food online by looking at the variables that are used in the decision rules at the top of the tree.

Random Forest: This is an ensemble method that merges the predictions of multiple decision trees. To use this method, we first divide the data into different subsets. Then, we train multiple decision trees on each subset. Each tree

makes its predictions based on the data it was trained on. Finally, we take the average of all the predictions from these trees to get the final result. Random forest is more robust to overfitting than a single decision tree, and it also allows us to identify the most important variables by looking at the variables that are used in the decision rules of the majority of the trees.

The most important equation for Random Forest is the feature importance formula which is used to calculate the importance of each feature. The feature importance of a feature is calculated as the average decrease in impurity across all decision trees in the forest. The impurity can be calculated using the Gini Impurity.

$$Gini = 1 - \sum_{i=1}^m p_i^2 \quad (2)$$

We use the 11 factors which are liking to try new technologies, making an impulsive decision, being too busy to cook, disliking cooking, finding ordering online to be easy, having a variety of options, finding ordering to be inexpensive, because of promotional offers and discounts, quick delivery, food safety and nutritious value. We use orders per month as the dependent variable. We calculate the mean and standard deviation of each factor. Then apply OLS, Decision Tree Regressor, and Random Forest Regressor. After that, we use the r-squared value to measure the goodness of fit of our regression models. The r-squared values for OLS, Decision Tree, and Random Forest are 0.64, 0.01 and 0.04. Which means OLS is a better approach for our data.

3.5. Hypothesis Testing: Investigating the relationship between the factors

In our study, we also conduct some hypothesis [27] testing using the chi-squared test [28]. We use this test since the factors are categorical values and we want to see if the factors are independent of each other.

Chi-Squared Test: This test is used to determine if there is a significant difference between an observed distribution and a theoretical distribution. The test statistic is calculated using the following equation:

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (3)$$

Where:

O = observed frequency

E = expected frequency

The degrees of freedom for the test is calculated as:

$$df = (\text{number of rows} - 1) * (\text{number of columns} - 1) \quad (4)$$

The degrees of freedom (df) adjust for the information lost due to fixed marginal constraints within the contingency table. This adjustment ensures the proper chi-square distribution for calculating the p-value, henceforth controlling Type I error and preventing erroneous conclusions about the significance of observed associations between the variables.

The p-value is then looked up in a chi-squared distribution table with the corresponding degrees of freedom. If the calculated p-value is less than the chosen significance level (usually 0.05), then the null hypothesis (that there is no significant difference between the observed and theoretical distributions) is rejected, and it is concluded that there is a significant difference between the two. We use the chi-squared test of independence as the factors have ordinal values. So, something like Pearson's correlation coefficient might not be the best measure of testing here as it assumes that the data is normally distributed and continuous.

We test the following hypotheses for our study,

People who are open to experimenting with new technology such as OFO services, may find it to be effortless to place an order. These platforms offer an easy-to-use and intuitive interface which simplifies the ordering process, thus allowing users to place orders quickly without any sort of hassle. Moreover, these applications often provide the option

Table 3
Results of Hypothesis Testing.

Serial	P-Value	Decision
H1	< 0.001	Accepted
H2	0.114	Rejected
H3	< 0.001	Accepted
H4	< 0.001	Accepted

to track orders in real-time, personalized recommendations. All of these features can enhance the overall ordering experience and make it more convenient for the users to use. Based on this argument we come up with the following hypothesis:

H1: There is a significant relationship between trying out new technology and thinking that using OFO service is easy.

People who have busy schedules may be more likely to make impulsive decisions when it comes to using OFO services. Due to their busy lifestyle, they may not have the energy to shop for ingredients or cook at home. Therefore, they may turn to quick and easy options such as ordering takeout. Based on this argument we come up with the following hypothesis:

H2: There is a significant relationship between making impulsive decisions when using OFO service and being too busy to cook.

People who do not like to cook may appreciate the variety of options that are available on the OFO platforms. For them, the thought of preparing meals at home can be unappealing or they might prefer to have a variety of options to choose from when ordering food. Based on this argument we come up with the following hypothesis:

H3: There is a significant relationship between people not liking to cook and finding there are many options to choose from when using OFO service.

Promotional offers or discounts can make the use of OFO services more cost-effective for users. These offers can take various forms such as coupons, loyalty programs, and special deals. These can be used to reduce the cost of items, delivery fees or sometimes even the entire order. By making use of these discounts, customers can save money on their orders. Therefore, this leads them to believe that ordering food online is inexpensive. Based on this argument we come up with the following hypothesis:

H4: There is a significant relationship between thinking OFO service is inexpensive and ordering food because of promo codes and discounts.

3.6. Correlation Analysis

To examine the relationship between BMI (x) and the frequency of ordering food online (y), we apply several correlation analysis methods. These are Pearson Correlation [29] where we use the actual BMI value of a respondent, Spearman's Rank Correlation [30] where the BMI categories are used and Point-biserial Correlation [31] where we convert BMI categories to binary categories.

Pearson correlation: This method measures the linear relationship between two continuous variables. The process involves calculating the means (\bar{x} , \bar{y}) and standard deviations of each variable, followed by computing the product of the deviations of each variable from its mean for each data point. The following equation is then used to calculate Pearson's correlation coefficient, r :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

The resulting value of r ranges from -1 to 1, here -1 indicates a perfect negative linear relationship, 0 indicates no linear relationship, and 1 indicates a perfect positive linear relationship. To evaluate the significance of the correlation

coefficient, a p-value is determined using the correlation coefficient, degrees of freedom ($df = n-2$), and a significance level of $\alpha = 0.05$ is used usually. The table of critical values is used to find the p-value. The correlation coefficient is assumed as zero by the null hypothesis, implying that there is no correlation between the two variables. The alternative hypothesis is that the correlation coefficient is not zero, implying that there is a correlation between the two variables. If the p-value is less than α , we get enough evidence to accept the alternative hypothesis and reject the null hypothesis. This method is applicable when both variables are continuous and assumes the data follow a normal distribution. It also assumes the observations are independent of each other.

Spearman's rank correlation coefficient: This is a non-parametric way of measuring correlation. With this, we can measure how well the relationship between two variables can be described by using a monotonic function. It is similar to Pearson correlation but instead of measuring linear association, this evaluates rank order similarity between variables.

In this method, each variable is assigned ranks based on its values. The smallest value receives a rank of 1, the second smallest a rank of 2 and so on. Then it determines the difference between ranks for each paired observation and these differences represent the change in the variables' relative positions. Each difference obtained in the previous step is then squared to stop the positive and negative differences cancelling each other out. Finally, the coefficient is calculated using the following formula:

$$\rho = 1 - \frac{(6 * \text{sum of squared differences})}{n * n^2 - 1} \quad (6)$$

The results will range from -1 to 1, where a value of +1 indicates that as one variable increases, the other variable also increases consistently. On the other hand, a value of -1 means that as one variable increases, the other variable decreases consistently. A coefficient of 0 suggests no monotonic relationship between the variables. This approach only works when the variables are ordinal or ranked. This method is also less sensitive to outlier data.

Point-biserial correlation: This approach measures the strength and direction of association between a continuous variable and a binary variable.

At first, the data is divided into two groups based on the binary variable. One group represents the presence of the characteristic while the other group represents the absence of it. Then, the mean value of the continuous variable is calculated for each group separately. After that, the standard deviation of the whole dataset is calculated, disregarding the grouping which represents the variability of the continuous variables across the entire dataset. The mean of the group representing the absence of the characteristic is then subtracted from the mean of the group representing the presence of it. We divide the difference by the standard deviation. We then get the point-biserial correlation coefficient which quantifies the strength and direction of the relationship between the continuous and binary variable.

The coefficient value ranges from -1 to +1. A positive value indicates a positive relationship, meaning higher values of the continuous variable tend to be associated with the presence of the characteristic. A negative value indicates a negative relationship, where higher values of the continuous variable tend to be associated with the absence of the characteristic. A value of 0 indicates no relationship between the continuous variable and the binary variable. This approach assumes that the continuous variable follows a normal distribution or at least approximate normality. Skewed or variable with outliers may affect the accuracy. The observations should also be independent of each other. Larger sample sizes tend to yield more accurate estimates.

For Pearson correlation, we use the BMI with continuous values and orders per month which also consists of continuous values. When using Spearman's rank correlation coefficient we use the categorical values of BMI instead and for Point-biserial correlation, we divide the BMI categories into binary groups such as not obese and obese, not normal weight and normal weight and apply the method.

We also apply Pearson and Spearman's correlation on physical activity and orders per month. Treating the values for physical activity as continuous for Pearson and categorical for Spearman's.

3.7. Prediction of Frequent users of OFO services

To classify frequent users of OFO services we use several methods such as Logistic Regression [32], Naive Bayes [33], Decision Tree Classifier [25], Random Forest Classifier [26], Gradient Boosting Machines (GBM) [34] and K-Nearest Neighbour (KNN) [35].

Table 4

Pearson and Spearman Correlation Test Results of BMI vs Orders Per Month.

Test	P-value	Coefficient	Relationship
Pearson	2.43 ⁻⁷	0.28	Weak Positive
Spearman	0.003	0.15	Very Weak Positive

Logistic Regression: This is a commonly used statistical model for tasks that involve classifying things into two categories. The main aim is to predict the probability of an instance belonging to a specific class. Unlike linear regression, which predicts continuous values, it is specifically designed to predict the probability of an event happening. Instead of providing a direct numerical output, this method calculates the likelihood of an event occurring based on the input variables. This makes it suitable for tasks where we are interested in determining the likelihood or probability of a specific outcome.

Logistic Regression models the relationship between the predictor variables and the output class probabilities using a logistic function or sigmoid function. The sigmoid function is defined as:

$$S(z) = \frac{1}{1 + e^{-z}} \quad (7)$$

Here, z represents a linear combination of the input features and model parameters.

The model is trained by finding the optimal values for the model parameters which minimize the error between the predicted probabilities and the actual class labels in the training data. This is typically done using optimization algorithms such as Newton's method or Gradient Descent. The objective is to minimize the log loss of the predicted probabilities or maximize the likelihood of it. Once the model is trained and the optimal parameters are obtained, we can use the model to make predictions on new, unseen data. The sigmoid function is applied to the linear combination of the input features and parameters to obtain the predicted probability. A common threshold (0.5) is then used to classify the instance into one of the two classes based on the predicted probability. If the predicted probability is above the threshold, it is classified as class 1; otherwise, it is classified as class 0.

Naive Bayes: This algorithm is built upon Bayes' theorem and assumes that given the class label, the features are conditionally independent of each other.

This algorithm starts by calculating the prior probabilities of each class in the dataset. The prior probability of a class is the probability of encountering that class in the dataset without considering any feature information. For each feature in the dataset, this method calculates the conditional probability of observing that feature given a specific class label. This is done by counting the occurrences of each feature in the training samples belonging to a particular class. Once the prior probabilities and conditional probabilities are computed, Bayes' theorem is applied to calculate the posterior probability of each class given the observed features. Finally, the classifier predicts the class label for a new, unseen sample by selecting the class with the highest posterior probability.

The algorithm for Decision Tree and Random Forest classifier works in the same as the regressor approach but the classifier is more suitable for categorical or discrete class labels.

Gradient Boosting Machines: This algorithm combines multiple weak predictive models (usually decision trees) to create a better predictive model. It is an ensemble learning method that sequentially builds an additive model by minimizing a predefined loss function using gradient descent.

A weak learner which usually is a decision tree with a small depth which is also called a decision stump, is fitted to the training data. The tree is trained to minimize the loss function with respect to the target variable. The predictions of the weak learner are then subtracted from the true values of the target variable to obtain the residuals. The residuals represent the errors or the parts of the target variable that the model has not yet captured. Another weak learner is then fitted to the residuals obtained from the previous step. The goal is to find a new model that can capture the remaining patterns in the data that the previous model missed. The new weak learner is now added to the ensemble by combining it with the previous models. To combine the models, a weight is assigned to each weak learner, this is also known as the learning rate. The weight determines the contribution of each model to the final prediction. The process of calculating residuals, fitting the next learner and updating of the model is repeated iteratively for a specified number of times or

Table 5

Results of feature selection.

Chi-squared Test	ANOVA F-value
Gender	Gender
BMI	BMI
Employment Status	Employment Status
Financial Dependency	Financial Dependency
Physical Activity	Marital Status
Ordering History	Physical Activity
Impulsive Decision	Ordering History
Busy To Cook	Impulsive Decision
Dont Like To Cook	Busy To Cook
Ordering Easy	Dont Like To Cook
Options To Choose	Ordering Easy
Ordering Inexpensive	Options To Choose
Promo Discounts	Promo Discounts
Quick Delivery	Quick Delivery
Food Safety	Food Safety
Nutritious	Nutritious

until a predefined stopping criterion is met. In each iteration, a new weak learner is fitted to the negative gradients of the ensemble model built so far. The final prediction is gained by summing the predictions of all the weak learners with their respective weights. The model assigns higher weights to the more accurate learners, which allows it to give more importance to their predictions.

K-Nearest Neighbour: KNN is an algorithm that does not rely on any assumptions about the underlying data distribution. The fundamental idea behind KNN is to classify or predict the target variable of a new data point based on the majority vote or the average of the target variables of its K number of nearest neighbours in the feature space.

At first, the value of K is figured out, which represents the number of nearest neighbours to consider for classification or regression. The optimal value of K can be found using techniques such as cross-validation or grid search. To classify or predict a new data point, the distance between that point and every other data point in the training data is calculated. The distance can be calculated using different metrics such as Manhattan distance and Euclidean distance. The distances are then sorted in ascending order and the K data points are selected with the shortest distances as the nearest neighbors. Then, the class label is determined that occurs most frequently among the K nearest neighbours. This can be done by counting the occurrences of each class label and selecting the one with the highest count. Once the class label or predicted value has been determined, it is assigned as the output for the new data point.

We divide the orders per month variable into two categories. One with less or equal to 5 orders per month which covers 56% of the sample and another with more than 5 orders per month which covers the rest of 44% of the sample. We then use this as the target label against all the features available in our dataset. We use univariate methods such as the Chi-squared test and ANOVA F-value to do some feature selection.

From table 5, we can see that 15 out of the 16 selected features are common for both methods. Therefore, we use these 15 to build our prediction models.

We use several evaluation metrics like accuracy score, precision, recall, and F1-score to find the model with the best results.

3.8. Customer Segmentation

Customer segmentation is a technique in data analysis to group customers based on their similarities and differences. One way of doing this is using the K-Means clustering algorithm [36], which is an unsupervised machine learning algorithm that aims to divide a given dataset into K distinct clusters, where each data point goes to the cluster with the nearest mean value. Principal Component Analysis (PCA) [37] is a technique that can be used to reduce the dimension of the dataset.

Table 6

Comparison of results of evaluation metrics for K-Means clustering with and without PCA.

Metric	With PCA	Without PCA
Silhouette Coefficient \uparrow	0.43	0.21
Calinski-Harabasz Index \uparrow	258.27	88.82
Davies-Bouldin Index \downarrow	0.87	1.71

K-Means: The algorithm works by assigning data points iteratively to the cluster with the closest centroid and then updating the centroids to be the average of the data points in each cluster. This process continues until the centroids no longer change significantly.

The number of clusters K is usually chosen by using the Elbow method or silhouette scores. Then the the centroids of the clusters are initialized. This is done randomly or by using some other method, such as k-means++. Each data point then gets assigned to the cluster with the closest centroid. The centroids get updated to be the average of the data points in each cluster. This is repeated until the centroids no longer change significantly.

Principal Component Analysis: This is a widely used statistical technique for dimensionality reduction and data analysis. It aims to find the directions or principal components in which the data varies the most and represents the data in a lower-dimensional space while preserving the essential information.

At first, the dataset is standardized to have a unit variance and zero mean. This step ensures that all features are on a similar scale and prevents variables with large variances from dominating the analysis. Then the covariance matrix of the standardized data is computed. The covariance matrix indicates the relationships between different features and measures how they vary together. The value in the i^{th} row and j^{th} column of the covariance matrix represents the covariance between the i^{th} and j^{th} features. An eigendecomposition is performed on the covariance matrix to find its eigenvalues and corresponding eigenvectors. The amount of variance explained by each principal component is represented by the eigenvalues. The eigenvectors represent the directions of these components. The eigenvalues indicate the relative importance of each principal component. PCA typically sorts the eigenvalues in descending order and selects the top- k eigenvectors corresponding to the largest eigenvalues, where k is the desired number of dimensions in the reduced space. The selected eigenvectors form a new coordinate system. The original data is projected onto this new coordinate system to obtain the reduced-dimensional representation. This projection involves multiplying the standardized data matrix by the eigenvector matrix, resulting in a transformed dataset with reduced dimensions.

For this, we use only the 11 factors we developed that influence OFO to do customer segmentation. We use the K-Means [36] algorithm to create the segmentation. We apply K-Means using both Principal Component Analysis (PCA) [37] and one without using it. Before applying PCA, we use Kaiser-Meyer-Olkin test (KMO) [38] and Bartlett's Test of Sphericity [39] to check if the data is suitable to apply PCA. The KMO test measures the proportion of variance in each variable that is explained by other variables, while Bartlett's Test confirms whether the correlations between variables are significantly different from zero. These tests help us avoid applying PCA to data where variables are entirely unrelated or where common factors are minimal, leading to unreliable results. The results of the KMO test give us 0.85 which makes it sufficient to perform factor analysis according to this paper [40]. The result of Bartlett's Test of Sphericity gives us a p-value of 1.06^{-228} which is lower than the α value of 0.05 and a high chi-squared value of 1301.43. Therefore, we can apply PCA to our data.

We use the Elbow method and silhouette score to find an optimal number of clusters. From figures 2 and 3 we can see that the optimal number of clusters is 2.

We use evaluation metrics such as Calinski-Harabasz Index [41] and Davies-Bouldin Index [42] to choose the one with best performance. The Calinski-Harabasz Index prioritizes tight, well-separated clusters, while the Davies-Bouldin Index penalizes clusters with high internal variation or proximity. Analyzing both helps identify the optimal cluster number (k) balancing compactness and separation in our data. In table 6, we can see that applying PCA gives us a comparatively better evaluation score.

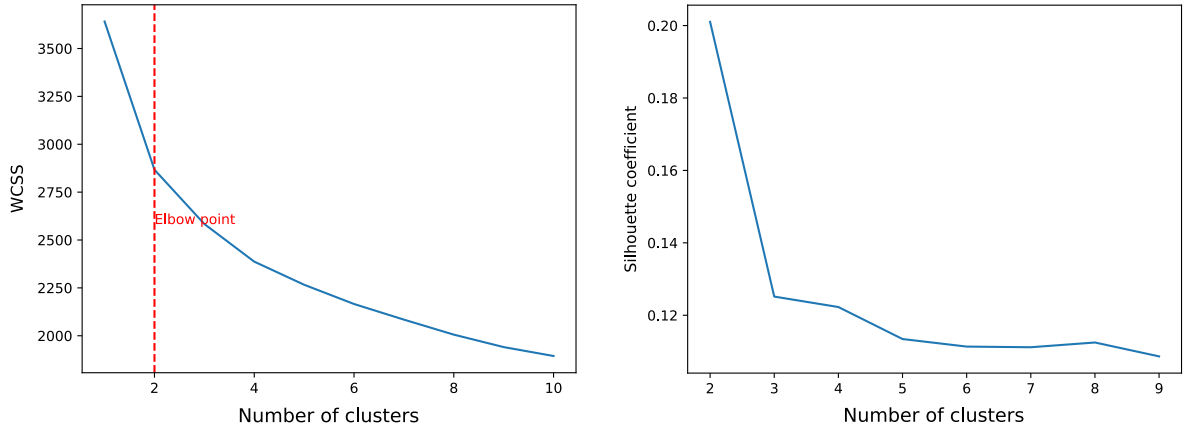


Figure 2: Elbow and Silhouette plot to determine the optimal number of clusters before applying PCA. WCSS stands for Within-Cluster Sum of Square.

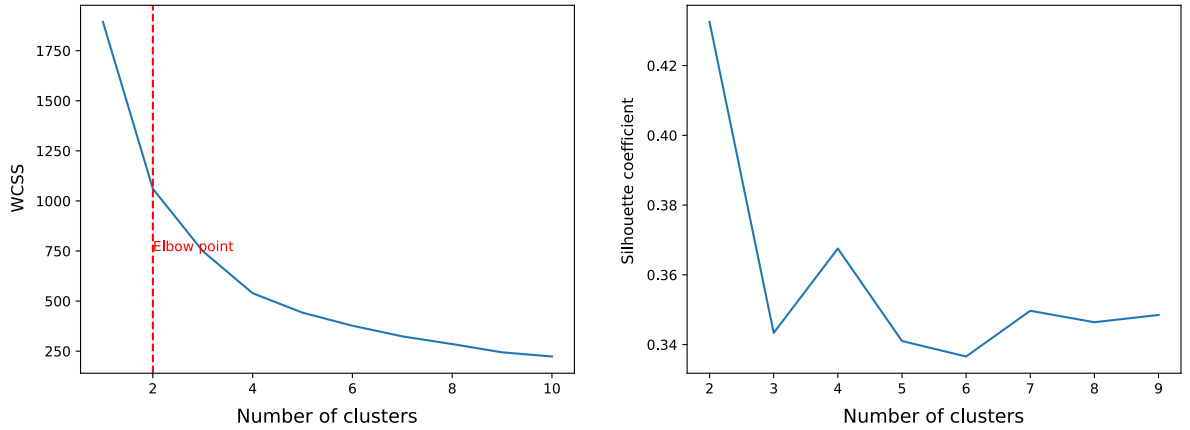


Figure 3: Elbow and Silhouette plot to determine the optimal number of clusters after applying PCA. WCSS stands for Within-Cluster Sum of Square.

4. Empirical Analysis

Now, we use the 331 valid responses out of the 343 responses we receive, to show some statistical description of the dataset and do some exploratory data analysis:

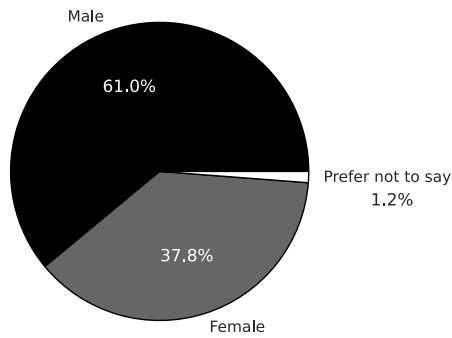


Figure 4: Distribution of Gender.

According to figure 4, most of the participants are male (61.03%). We have received 125 responses from female participants (37.76%) and 4 participants preferred not to disclose their gender (1.21%).

Analyzing the frequency distribution of BMI, it is evident that the majority of participants (58.3%) fall into the category of normal BMI, with 193 calculated instances. The second-largest group (25.7%) corresponds to participants with an overweight BMI, totalling 85 instances. Additionally, nearly 28 participants (8.5%) have a BMI in the obese range. Furthermore, 25 participants (7.6%) have a BMI indicating being underweight, as illustrated in Figure 5.

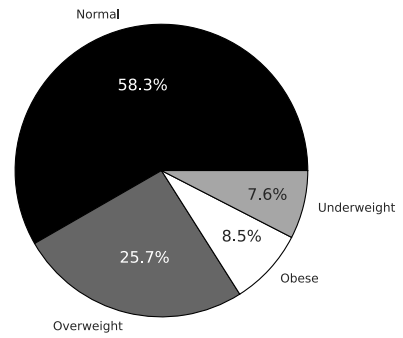


Figure 5: Distribution of BMI.

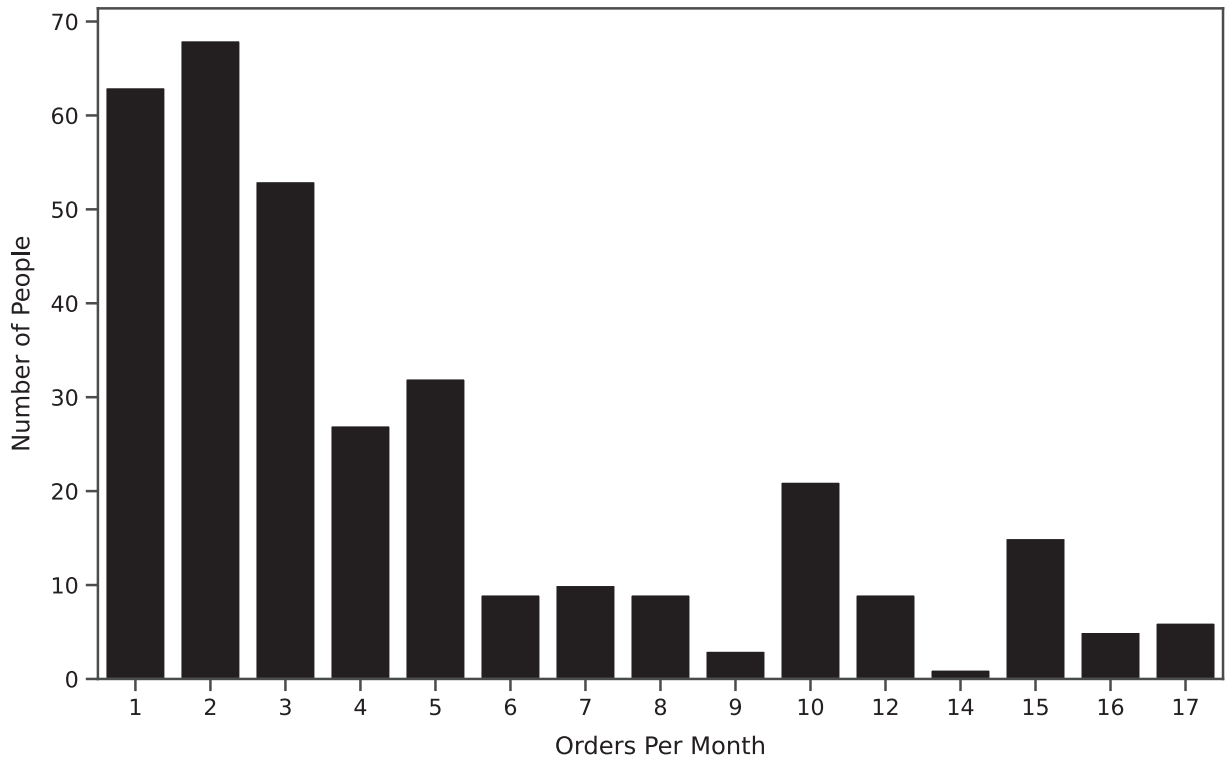


Figure 6: Distribution of Online Food Orders Per Month.

From Figure 6 we can see the number of foods ordered online per month by our survey participants. Here the range of ordering online is from 1 to 17. The highest number of ordering food in a month is 3 according to the chart. After that 2 and 1 are respectively the second and third highest numbers.

Behavioural Analysis of Factors Influencing Online Food Ordering and Its Relation to Obesity

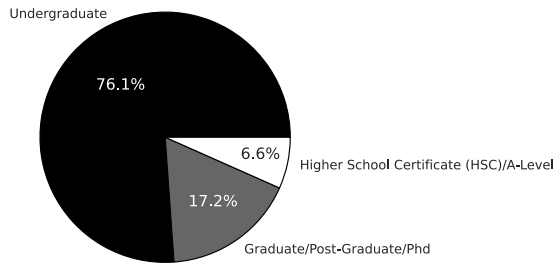


Figure 7: Distribution of Educational Qualification.

Figure 7 shows that our online food ordering survey participants were mostly undergraduate students (76.13%), with 252 responses. We also had 57 responses (17.22%) from Graduate / Post-Graduate / Ph.D. students. The fewest responses (6.65%) came from high school students, with only 22 responses.

From Figure 8, we see that 146 participants (44.11%) are seeking career opportunities, while 62 participants (18.73%) are full-time employed. A significant number of participants (58, or 17.52%) did not mention their employment status. Furthermore, 65 participants (19.64%) are doing part-time jobs.

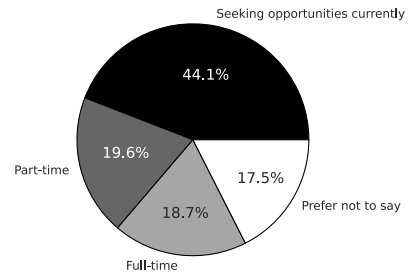


Figure 8: Employment Status.

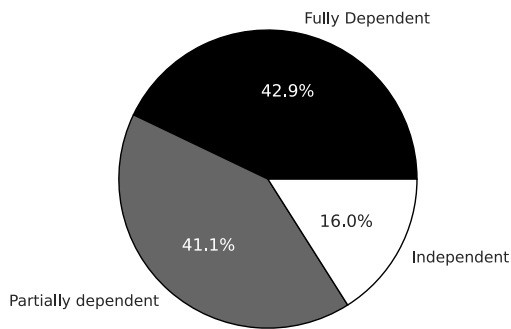


Figure 9: Distribution of Financial Dependency.

Inspecting the figure 9, we see that 142 participants are fully dependent (42.90%). The number of partially dependent (41.09%) participants is 136. Lastly, 53 people responded to their status as independent (16.01%).

In figure 10, we can see that 304 individuals are single (91.84%). We also have 18 participants who are married (5.44%) and only 9 respondents did not mention their marital status (2.72%).

Figure 11 shows the distribution of respondents' physical activity over a week. Over 160 participants reported not doing any physical activity, while fewer than 60 engaged in it 2 to 3 times per week. Only 20 or fewer reported doing it 4 to 7 times per week, and less than 40 did it once.

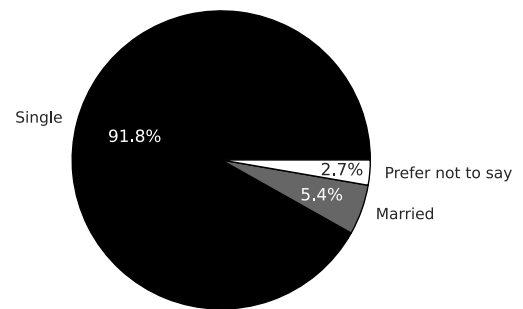


Figure 10: Marital Status.

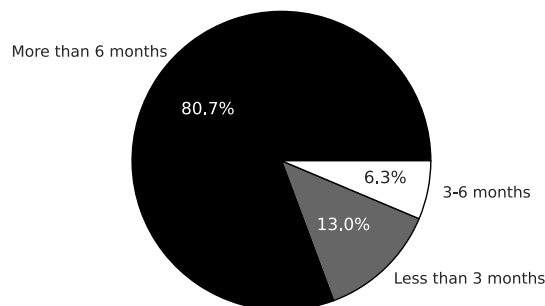


Figure 12: Distribution of OFO History.

In figure 12, we can see that 262 participants have been using OFO services for more than 6 months (80.66%). Moreover, we can notice that 43 participants have used it for less than 3 months (12.99%). Furthermore, 21 participants have been using it for 3 to 6 months (6.34%).

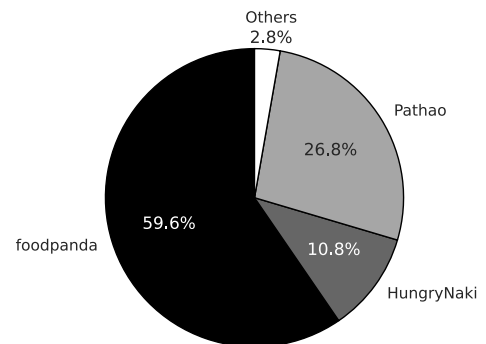


Figure 13: Distribution of OFO Apps.

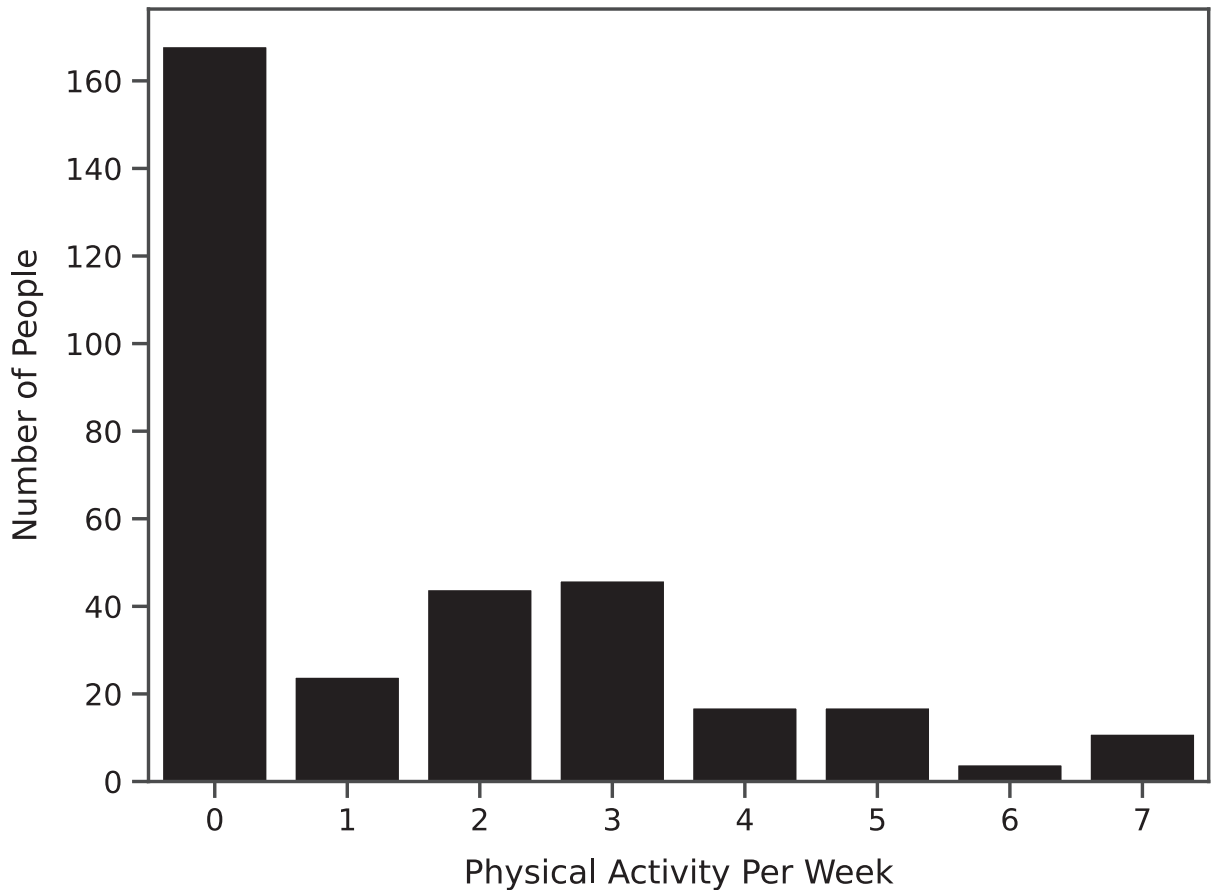


Figure 11: Distribution of Physical Activity.

Most of the survey participants (59.57%) are using foodpanda and the number is 302. The second highest (26.82%) used app by our survey participants is Pathao where the number of users is 136. Thirdly, 55 people are using Hungrynaki (10.85%). Lastly, we can see that 14 survey participants are using other apps (2.76%).

Figure 14 gives us an idea of when people tend to order food. Firstly, most of the participants (38.22%) are ordering food for evening snacks and the number is 227. Secondly, 176 participants voted for online food ordering time as dinner (29.63%). Thirdly, 138 people responded to lunchtime for ordering food online (23.23%). Moreover, The number of ordering food online for midnight snacks (6.06%) is 36. Furthermore, 17 people selected breakfast for ordering food online (2.86%).

From figure 15, we can see that the majority tend to order burgers (267) and pizzas (249) when ordering food online, which consist of high amounts of saturated fat and salt. Salads (19) are among the lowest to be ordered. There were some food types with only one entry, we dropped those since those are negligible.

The figure 16 gives us an idea of the importance of each factor on a Likert scale. Firstly, we found that 100 people agreed and 125 responded as neutral to being inclined to try out new technologies. Secondly, 120 people voted neutral and 73 agreed to make impulsive decisions to order food online. Thirdly, 97 people responded neutral, 86 agreed and 77 disagreed with being too busy to cook food themselves and leaning towards OFO services. There is a match with 83 numbers of responses where each is allocated for disagree and neutral where the factor is people don't like to cook. A huge number of participants agreed that ordering food online is easy, where 152 people agreed and 102 strongly agreed. Again, lots of people agreed that they like having a variety of options to choose their desired food from OFO platforms, where 139 people agreed and 114 strongly agreed. In the case of choosing to order food online due to being inexpensive, 110 participants responded as neutral and 98 people disagreed. 117 people agreed that promotional

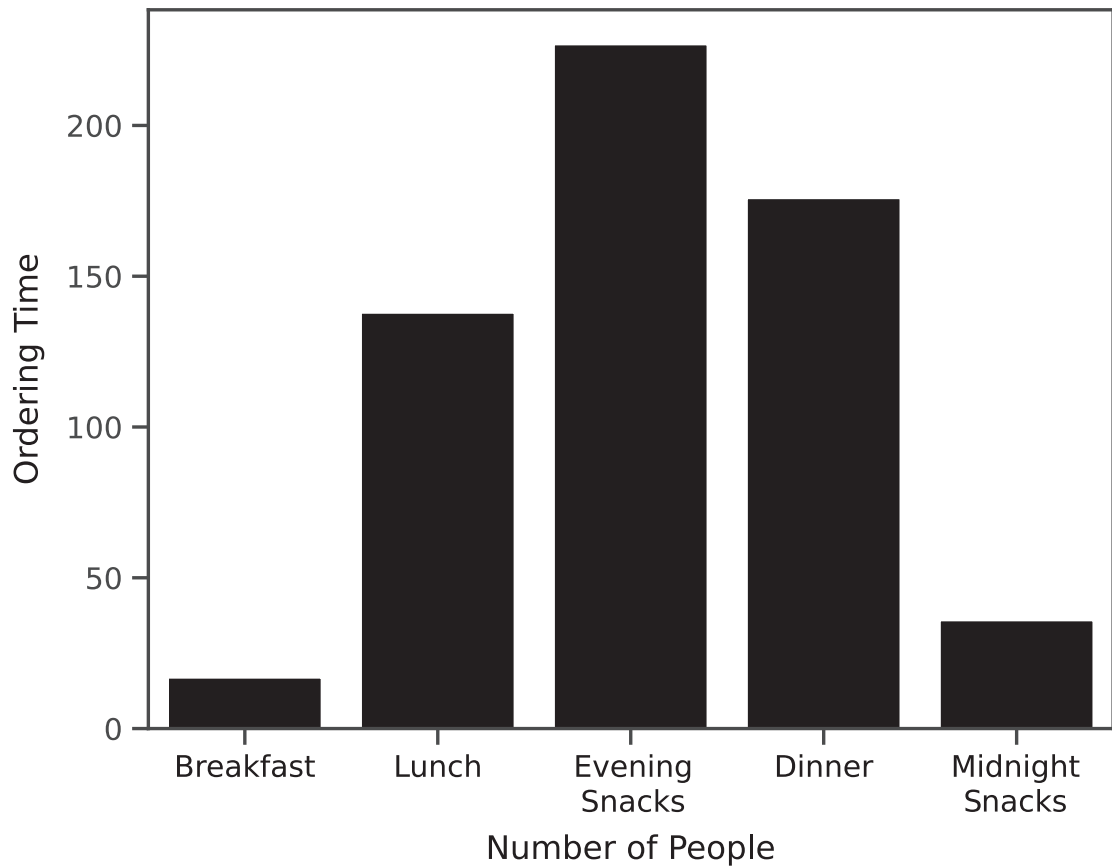


Figure 14: Distribution of Online Food Ordering Time.

discounts convince them to order food online whereas 86 responded to neutral. A big portion of our survey participants agreed that ordering food online allows them to get their food quickly and the response number is 120, besides 116 people voted neutral. 174 of the participants stayed neutral when they were asked if they think their food is safe and hygienic and 78 participants disagreed with it. Lastly, very few survey participants believe that their online food ordered item is nutritious with 118 people responding as neutral, 106 disagreed and 63 participants strongly disagreed with the factor.

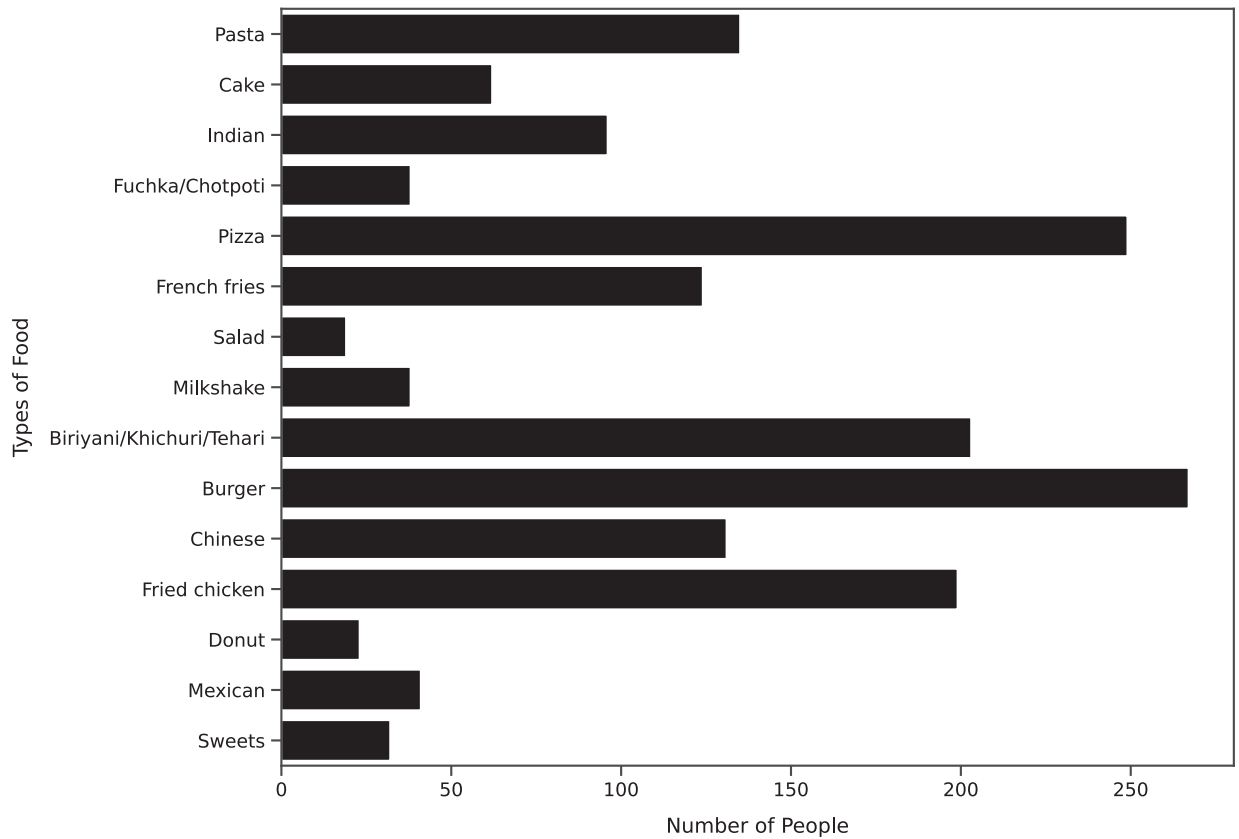


Figure 15: Distribution of type of foods ordered through online platforms.

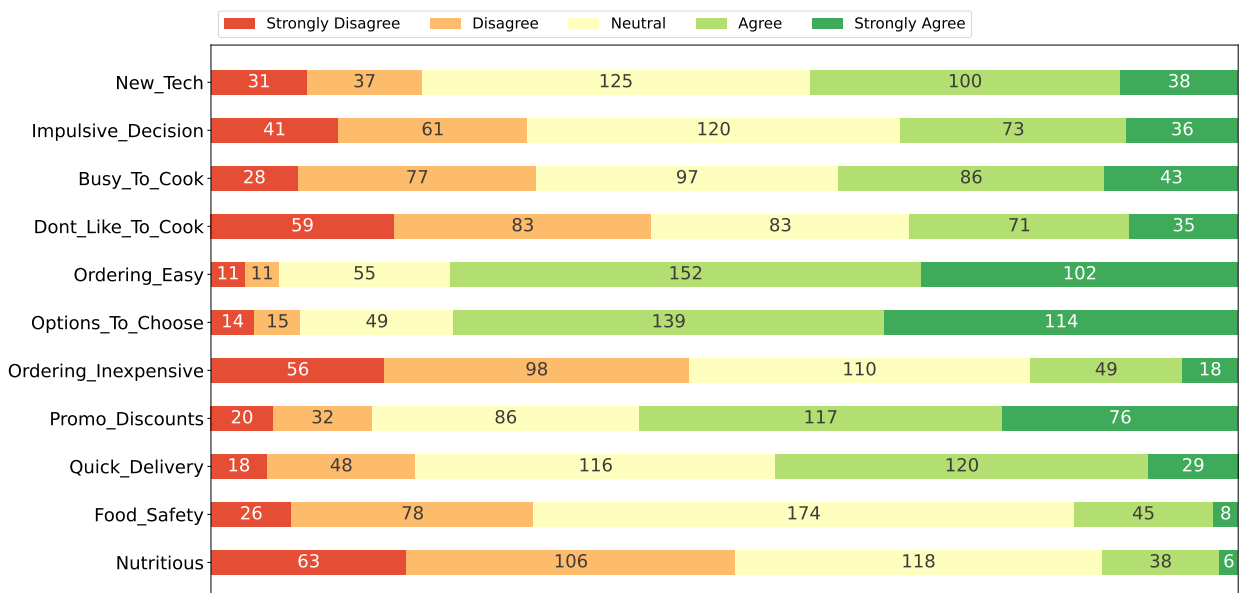


Figure 16: Distribution of Factors that Influence Online Food Orders.

5. Results

5.1. Results of Most Influential Factors Mining

The results of the experiments which are done to find the most influencing factors are shown in table 7 gives us an idea of which factors influence a person to use OFO services.

The mean values based on solely the responses from the respondents suggest that the most important factors in influencing people's decision to order food online are the convenience of ordering (mean score of 2.98), the variety of options available (mean score of 2.98), and promotional offers and discounts (mean score of 2.60).

5.2. Results of Hypothesis Testing

The results of the hypothesis study point out that the first hypothesis, H1 is accepted and the null hypothesis is rejected as the p-value is < 0.001 , which is less than ($\alpha = 0.05$). This means that people who tend to try out new technologies also feel that ordering food online is convenient and easy.

For the second hypothesis, H2 is rejected as the p-value (0.1143) is higher than the α (0.05). This means we could not find any significant relationship between being too busy to cook and making impulsive decisions when ordering food online.

The third hypothesis, H3 is accepted and the null hypothesis is rejected as the p-value (< 0.001) is lower than the α (0.05). This means there is a significant relationship between having many options to choose from when ordering food online and not liking to cook.

The fourth hypothesis, H4 is accepted and the null hypothesis is rejected as the p-value (< 0.001) is less than the α (0.05). This means that there exists a significant relationship between finding ordering food online to be inexpensive and using promo codes and discounts when ordering food online.

5.3. Results of Correlation Analysis

The results from Pearson correlation and Spearman's rank correlation which was done between BMI and orders per month, give us a p-value of 2.43×10^{-7} and 0.003 respectively, which suggests that there exists a statistically significant positive correlation between the two variables since both these values are less than 0.05. The coefficient values are 0.28 and 0.15 respectively which suggests there exists a weak positive correlation and a very weak positive correlation [43].

The results shown in table 8 suggest that due to insufficient p-value, any relation between Underweight and orders per month is rejected, normal weight has a weak negative relationship, overweight has a very weak positive relationship and obese has a weak positive relationship.

As seen in table 9, we can see that there exists a very weak negative relation between physical activity and orders per month which suggests that people who tend to order more may exercise less. However, the results from Spearman's test are rejected as the p-value is 0.085, which is greater than the α value of 0.05.

The results of correlation analysis between ordering history and orders per month as seen on table 10 show that people who have been using OFO for more than 6 months tend to order more. The coefficient value of -0.23 suggests that there exists a weak relationship between ordering frequency and people who have been using this type of service for less than 3 months. However, due to the p-value (0.66) being higher than the α value of 0.05, any relationship between people who have been ordering for 3-6 months is rejected.

5.4. Results of the Prediction of Frequent Users of OFO services

The results of the prediction model can be seen in table 11.

We can see that out of 6 models, Random Forest performs the best with 81% accuracy. In comparison, Naive Bayes performed worst with 68% accuracy. Random Forest also has the best precision score of 0.74 which means it has a high level of precision in identifying positive instances. Additionally, Random Forest correctly identified 64% of the positive instances as seen in the Recall column.

5.5. Results of Customer Segmentation

In table 13, we have the mean values of each cluster. We show the values of both with PCA applied and without it. Cluster 0 contains respondents who tend to order less and Cluster 1 represents the people who order comparatively more. We can see that people of cluster 0 find food delivered by OFO services to not be inexpensive with the mean value for 'Ordering Inexpensive' as 0.83 and 0.97. Whereas, people belonging to cluster 1 tend to find the food from the OFO services to lack nutritious values with the mean value for 'Nutritious' as 1.64 and 1.66.

Table 7

Influential Factors in Online Food Orders: Comparing Mean, Standard Deviation, OLS, Decision Tree, and Random Forest Values.

Factors	Mean	SD	OLS	DT	RF
New Tech	2.23	1.09	0.39	0.01	0.22
Impulsive Decision	2.01	1.16	0.90	0.34	0.22
Busy To Cook	2.12	1.16	0.18	0.05	0.10
Don't Like To Cook	1.82	1.25	0.33	0.01	0.13
Ordering Easy	2.98	0.95	0.65	0.11	0.08
Options To Choose	2.98	1.03	0.15	0.01	0.06
Ordering Inexpensive	1.62	1.10	0.27	0.22	0.07
Promo Discounts	2.60	1.12	0.33	0.01	0.09
Quick Delivery	2.28	1.00	0.11	0.01	0.05
Food Safety	1.79	0.86	-0.40	0.01	0.05
Nutritious	1.45	0.98	-0.32	0.28	0.09

Table 8

Results of Point-biserial Correlation on BMI vs Orders Per Month.

BMI	P-value	Coefficient	Relationship
Underweight	0.84	-0.01	Rejected
Normal	1.97^{-5}	-0.23	Weak Negative
Overweight	0.04	0.04	Very Weak Positive
Obese	1.13^{-10}	0.34	Weak Positive

Table 9

Results of Correlation analysis between physical activity and orders per month.

Test	P-value	Coefficient	Relationship
Pearson	0.026	-0.12	Very Weak Negative
Spearman	0.085	-0.10	Rejected

Table 10

Results of Point-biserial Correlation on Ordering History and Orders Per Month.

Ordering History	P-value	Coefficient	Relationship
< 3 months	1.40^{-5}	-0.23	Weak Negative
3 - 6 months	0.66	0.02	Rejected
> 6 months	< 0.0001	0.18	Very Weak Positive

Table 11

Comparison of results among prediction models before feature selection.

Model	Accuracy	Precision	Recall	F1-Score
LR	0.78	0.65	0.57	0.58
NB	0.66	0.54	0.54	0.53
DT	0.59	0.46	0.46	0.46
RF	0.79	0.73	0.56	0.55
GBM	0.71	0.51	0.50	0.50
KNN	0.80	0.72	0.64	0.66

Table 12

Comparison of results among prediction models after feature selection.

Model	Accuracy	Precision	Recall	F1-Score
LR	0.78	0.64	0.55	0.54
NB	0.79	0.70	0.58	0.59
DT	0.68	0.55	0.56	0.56
RF	0.81	0.74	0.61	0.63
GBM	0.76	0.61	0.56	0.56
KNN	0.74	0.56	0.53	0.52

Table 13

Comparison between cluster means between the clusters created with and without PCA. Here the factors are New Tech (NT), Impulsive Decision (ID), Too Busy To Cook (BC), Don't Like to Cook (DLC), Ordering is Easy (OE), Options to Choose (OC), Ordering is Inexpensive (OI), Promotional Offers and Discounts (POD), Quick Delivery (QD), Food Safety (FS), Food is Nutritious (FN).

Factors	With PCA		Without PCA	
	Cluster 0	Cluster 1	Cluster 0	Cluster 1
NT	1.45	2.47	1.45	2.48
ID	1.19	2.25	1.19	2.27
BC	1.47	2.31	1.47	2.34
DLC	1.08	2.04	1.08	2.04
OE	1.95	3.29	1.95	3.38
OC	1.75	3.35	1.75	3.44
OI	0.83	1.86	0.83	1.90
POD	1.53	2.92	1.53	3.02
QD	1.29	2.59	1.29	2.65
FS	1.17	1.98	1.17	2.02
FN	1.29	1.64	1.29	1.66

6. Discussions

The results of OLS show that the most influencing factors are making impulsive decisions (coefficient value of 0.90), and convenience of ordering (coefficient value of 0.65). On the other hand, food safety (coefficient value of -0.40) and nutritious value of the food (coefficient value of -0.36) have a negative impact. The results of the most influential factor mining support some of the findings from previous works. The high coefficient value on impulsive decision supports [7] as they also find that hedonic motivation is one of the important factors. The positive impact of ease of convenience as found by [6, 7, 8] is supported by our results as well. We find that food safety is a major concern as it has a negative impact when using OFO services, this is also supported by [16].

Additionally, we can see that the results of the hypothesis testing show that hypotheses, H1, H2 and H3 were accepted and H2 were rejected. This suggests some of the factors are not necessarily independent of each other.

The results of the Pearson Correlation suggest that when we treat BMI as a continuous value, there exists a weak positive correlation between BMI and orders per month. When treating BMI as a categorical value, the results of the Spearman Correlation show that there exists a weak positive correlation. Although, this does not indicate that there is necessarily any causation here. Moreover, when we treat the BMI values as binary categories it shows that there exists a very weak positive and a weak positive correlation for overweight and obese respectively. This supports the research conducted by Dana *et al.* [15]. However, people with normal BMIs show a weak negative relationship with orders of food per month. So, even though there exists a positive relationship between being overweight or obese and ordering a lot of food per month, it is however very weak. These findings can be used in future studies to combine with other factors to see if any causality exists.

We can see in table 12, that after applying feature selection the performance of Naive Bayes, Decision Tree, Random Forest and Gradient Boosting Machine improves. Although the performance of the KNN decreases and that of Logistic Regression stays the same as before. This can also be seen in figure 17.

Moreover, the results of customer segmentation show that respondents in both clusters show the tendency to order food because it is convenient, there are a variety of options to choose from, and due to the availability of promotional offers and discounts. However, we can see that people of cluster 0 find food delivered by OFO services to not be inexpensive. Whereas, people belonging to cluster 1 tend to find the food from the OFO services to lack nutritious values.

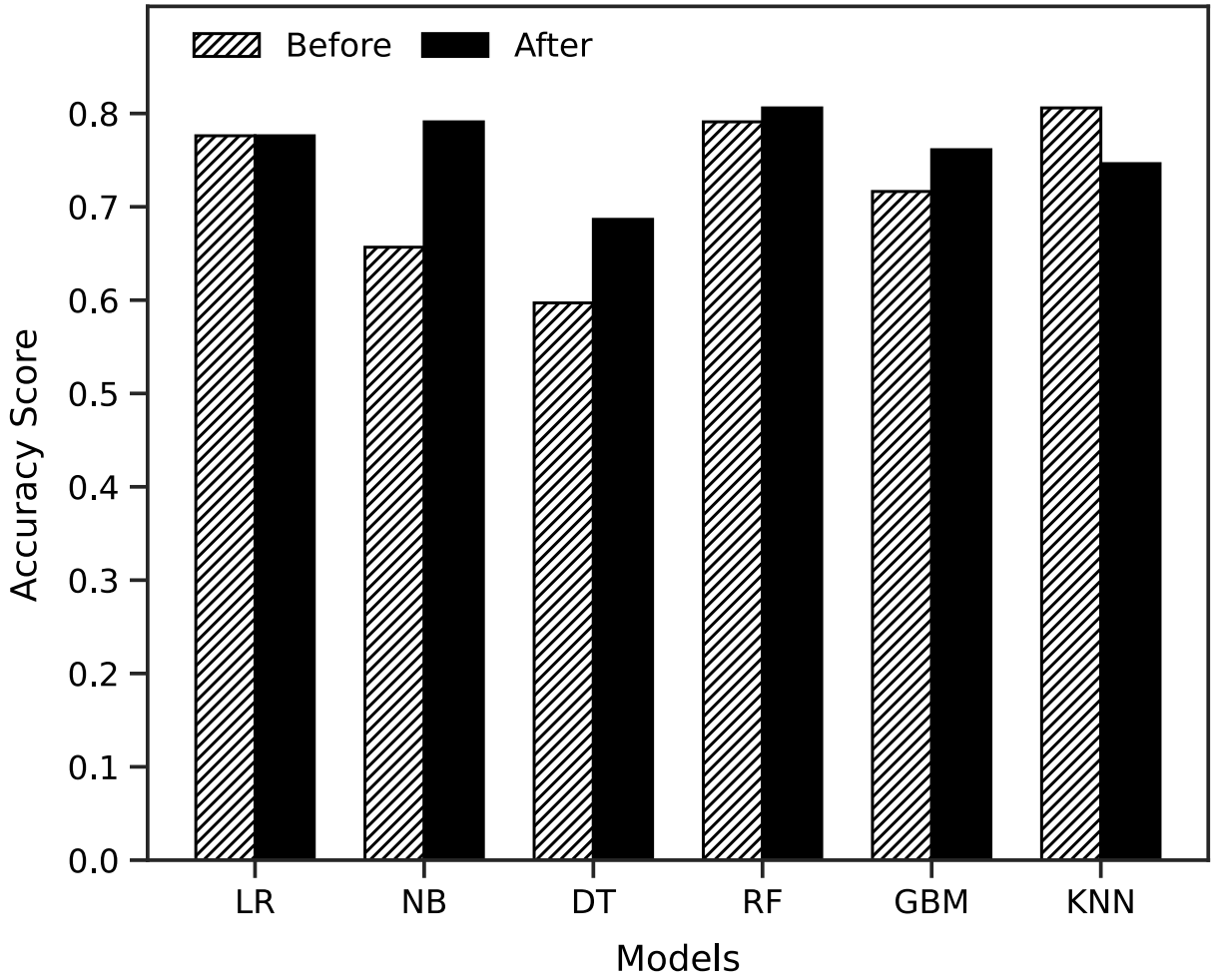


Figure 17: Comparison between results of prediction models before and after applying feature selection.

7. Conclusion

This research delves into the determinants influencing the utilization of online food ordering (OFO) platforms on a global scale, with a specific focus on the context of Bangladesh. We have generated a unique dataset capturing the behaviours of OFO service users, a noteworthy contribution considering the limited existing data, particularly within the Bangladeshi context. Our findings unveil a subtle positive correlation between Body Mass Index (BMI) and monthly order frequencies, suggesting that individuals with higher BMIs tend to place orders more frequently. It is crucial to highlight that correlation does not imply causation. The study identifies impulsive decision-making, ordering convenience, and the allure of promotional offers and discounts as pivotal factors driving the frequency of online food ordering. Furthermore, we achieve an 81% accuracy in classifying frequent users of OFO services. This research extends beyond regional boundaries, considering a densely populated country Bangladesh as a Test-case scenario, offering insights into the nuanced dynamics that shape online food consumption behavior on a global scale.

Dataset availability and usage policy

The data underpinning the findings of this study are accessible from the corresponding author upon reasonable request.

Funding:

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRedit authorship contribution statement

S. M. Fazle Rabby Labib: Conceptualization, Formal Analysis, Investigation, Methodology, Software Visualization, Writing - original draft. **Fahmida Zaman Achol:** Conceptualization, Formal Analysis, Investigation, Methodology, Visualization, Writing - original draft. **Md. Aqib Jawwad:** Conceptualization, Formal Analysis, Investigation, Methodology, Writing - original draft. **Md. Golam Rabiul Alam:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – review & editing. **Md Iftekharul Mobin:** Supervision, Writing – review & editing. **Ashis Talukder:** Supervision & Funding acquisition.

References

- [1] W. H. Organization, Obesity and overweight, <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>, 2021.
- [2] C. A. Dakin, G. S. Finlayson, R. J. Stubbs, Can eating behaviour traits be explained by underlying, latent factors? an exploratory and confirmatory factor analysis, *Appetite* (2024) 107202.
- [3] G. Weydmann, P. M. Miguel, N. Hakim, L. Dubé, P. P. Silveira, L. Bizarro, How are overweight and obesity associated with reinforcement learning deficits? a systematic review, *Appetite* (2023) 107123.
- [4] NHS, Obesity causes, <https://www.nhs.uk/conditions/obesity/causes/>, 2023.
- [5] C. Screti, K. Edwards, J. Blissett, Understanding family food purchasing behaviour of low-income urban uk families: An analysis of parent capability, opportunity and motivation, *Appetite* (2024) 107183.
- [6] G. Kale, S. Chourishi, The customer perception on online food ordering and its significance, *Alochana Chakra Journal* (2020).
- [7] S. M. Alagoz, H. Hekimoglu, A study on tam: Analysis of customer attitudes in online food ordering system, *Procedia - Social and Behavioral Sciences* 62 (2012) 1138–1143.
- [8] V. P. P. Pinto, I. Thonse Hawaldar, S. Pinto, Antecedents of behavioral intention to use online food delivery services: an empirical investigation, *Innovative Marketing* 17 (2021) 1–15.
- [9] R. Parameswaran, K. Krishnasamy, A study on factors influencing online consumer buying behavior, *International Journal of Scientific & Technology Research* 9 (2020) 3620–3625.
- [10] S. Anam, D. Golam Yazdani Showrav, M. A. Hassan, A. Chakrabarty, Factors influencing the rapid growth of online shopping during covid-19 pandemic time in dhaka city, bangladesh, *Academy of Strategic and Management Journal* 20 (2021) 1–13.
- [11] A. Chatterjee, M. W. Gerdes, S. G. Martinez, Identification of risk factors associated with obesity and overweight—a machine learning overview, *Sensors* 20 (2020) 2734.
- [12] A. M. Prentice, S. A. Jebb, Fast foods, energy density and obesity: a possible mechanistic link, *Obesity Reviews* 4 (2003) 187–194.
- [13] L. A. H. Harahap, E. Aritonang, Z. Lubis, The relationship between type and frequency of online food ordering with obesity in students of medan area university, *Britain International of Exact Sciences (BioEx) Journal* 2 (2020) 29–34.
- [14] N. D. Kurniawati, S. N. Cahyaningsih, A. S. Wahyudi, The correlation between online food ordering and nutritional status among college students in surabaya, *Indonesian Journal of Community Health Nursing* 6 (2021) 70.
- [15] L. M. Dana, E. Hart, A. McAleese, A. Bastable, S. Pettigrew, Factors associated with ordering food via online meal ordering services, *Public Health Nutrition* 24 (2021) 5704–5709.
- [16] C. Hong, H. H. Choi, E.-K. C. Choi, H.-W. D. Joung, Factors affecting customer intention to use online food delivery services before and during the COVID-19 pandemic, *Journal of Hospitality and Tourism Management* 48 (2021) 509–518.
- [17] E. Martha, D. Ayubi, B. Besral, N. D. Rahmawati, A. P. Mayangsari, Y. Sopamena, M. Astari, R. S. Zulfa, Online food delivery services among young adults in depok: Factors affecting the frequency of online food ordering and consumption of high-risk food, *Mapping Intimacies* (2021).
- [18] I. Jahidi, N. A. Ruyani, D. P. Alamsyah, The study of consumer behavior on online food ordering system (go-food) in the metropolitan city (2022).
- [19] C. Inthong, T. Champahom, S. Jomnonkwao, V. Chatpattananan, V. Ratanavaraha, Exploring factors affecting consumer behavioral intentions toward online food ordering in thailand, *Sustainability* 14 (2022) 8493.
- [20] F. D. Davis, Perceived usefulness, perceived ease of use, and user acceptance of information technology, *MIS Quarterly* 13 (1989) 319–340.
- [21] V. Venkatesh, M. G. Morris, G. B. Davis, F. D. Davis, User acceptance of information technology: Toward a unified view, *MIS Quarterly* 27 (2003) 425–478.
- [22] C. G. Forero, Cronbach's Alpha, Springer Netherlands, Dordrecht, 2014, pp. 1357–1359. URL: https://doi.org/10.1007/978-94-007-0753-5_622. doi:10.1007/978-94-007-0753-5_622.
- [23] K. S. Taber, The use of cronbach's alpha when developing and reporting research instruments in science education, *Research in Science Education* 48 (2017) 1273–1296.
- [24] B. Zdaniuk, Ordinary Least-Squares (OLS) Model, Springer Netherlands, Dordrecht, 2014, pp. 4515–4517. URL: https://doi.org/10.1007/978-94-007-0753-5_2008. doi:10.1007/978-94-007-0753-5_2008.

- [25] J. Fürnkranz, Decision Tree, Springer US, Boston, MA, 2010, pp. 263–267. URL: https://doi.org/10.1007/978-0-387-30164-8_204. doi:10.1007/978-0-387-30164-8_204.
- [26] T. K. Ho, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, volume 1, IEEE, 1995, pp. 278–282.
- [27] M. M. Stefanczyk, A. Zielińska, Are cooks more disgust sensitive? preliminary examination of the food preparation hypothesis, *Appetite* 192 (2024) 107117.
- [28] K. Pearson, X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (1900) 157–175.
- [29] D. Freedman, R. Pisani, R. Purves, Statistics (international student edition), Pisani, R. Purves, 4th edn. WW Norton & Company, New York (2007).
- [30] C. Spearman, Spearman Rank Correlation Coefficient, Springer New York, New York, NY, 2008, pp. 502–505. URL: https://doi.org/10.1007/978-0-387-32833-1_379. doi:10.1007/978-0-387-32833-1_379.
- [31] L. J.M, The expected value of a point-biserial (or similar) correlation, 2008. URL: <https://www.rasch.org/rmt/rmt221e.htm>.
- [32] D. R. Cox, The regression analysis of binary sequences, *Journal of the Royal Statistical Society: Series B (Methodological)* 20 (1958) 215–232.
- [33] G. I. Webb, Naïve Bayes, Springer US, Boston, MA, 2010, pp. 713–714. URL: https://doi.org/10.1007/978-0-387-30164-8_576. doi:10.1007/978-0-387-30164-8_576.
- [34] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Annals of statistics* (2001) 1189–1232.
- [35] A. Mucherino, P. J. Papajorgji, P. M. Pardalos, k-Nearest Neighbor Classification, Springer New York, New York, NY, 2009, pp. 83–106. URL: https://doi.org/10.1007/978-0-387-88615-2_4. doi:10.1007/978-0-387-88615-2_4.
- [36] X. Jin, J. Han, K-Means Clustering, Springer US, Boston, MA, 2010, pp. 563–564. URL: https://doi.org/10.1007/978-0-387-30164-8_425. doi:10.1007/978-0-387-30164-8_425.
- [37] I. Jolliffe, Principal component analysis, in: *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, 2011, pp. 1094–1096. URL: https://doi.org/10.1007/978-3-642-04898-2_455. doi:10.1007/978-3-642-04898-2_455.
- [38] H. F. Kaiser, A second generation little jiffy, *Psychometrika* 35 (1970) 401–415.
- [39] M. S. Bartlett, Tests of significance in factor analysis., *British journal of psychology* (1950).
- [40] A. Zbiciak, T. Markiewicz, A new extraordinary means of appeal in the polish criminal procedure: the basic principles of a fair trial and a complaint against a cassatory judgment, *Access to Justice in Eastern Europe* 6 (2023) 1–18.
- [41] C. T. Harabasz, M. Karoński, A dendrite method for cluster analysis, in: *Communications in Statistics*, volume 3, 1974, pp. 1–27.
- [42] D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE transactions on pattern analysis and machine intelligence* (1979) 224–227.
- [43] Z. Jaadi, Everything you need to know about interpreting correlations, 2019. URL: <https://towardsdatascience.com/everything-you-need-to-know-about-interpreting-correlations-2c485841c0b8>.