Project Report

CZ4032: Data Analytics and Mining

Group: 04

Academic Year 2017/2018

Semester 1

Student Names:	Matriculation Number:	Contribution
Huang Jian Wei	U1521567A	20%
Yong Guo Jun	U1440217C	20%
Chong Ning Hui Cherrie	U1521041B	20%
Hsin Jia Jing	U1521040E	20%
Shannon Neo Si Lin	U1521821L	20%

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

NANYANG TECHNOLOGICAL UNIVERSITY

Table of Contents

1 Abstract

People who are born between the early 1980s and the early 2000s are considered Millennials. In recent years, the Millennials have over taken the previous generation (the Baby Boomer generation) as the largest generation[1]. As such, it is important to find out the spending habits of the Millennials as they now are now the largest group of consumers.

The data set "Young People Survey" documents the responses of Millennials participants, aged 15-30 to a series of questions. With this data set, we aim to find out the best and most accurate algorithm to classify individuals based on what they spend on. We also want to find out the different traits of individuals that will affect the classification results of their spending habits.

Firstly, using J48 as a baseline classifier, we eliminate categories of responses that are less relevant to improve the accuracy of the classifier. Thereafter, the different attributes are ranked to find out the relative importance of each input feature. More important attributes are extracted as subsets and used as the data for classifying.

Different algorithms such as Naïve Bayes, Random Tree, Random Forest, and SMO are then applied to the data set. We compare the accuracy of these different algorithms to find out which is the most accurate one for our data set.

2 Problem Description

2.1 Motivation

As the Millennials took over the Baby Boomers as the largest generation, they also become the largest consumer group. There is therefore advantage for companies and brands to make Millennials the target consumers of their products and services. The need to understand the spending habits of Millennials hence arises.

In our project, we aim to find out the best and most accurate algorithm to classify the spending habits of individuals based on the following 3 categories (Class Labels):

1) Partying and socializing
2) Gadgets
3) And those who are willing to pay more for good quality, healthy food

We will also find out the best different traits of individuals that will affect the classification results of the aforementioned categories.

We chose to classify individuals based on whether they spend on socializing and gadgets as there is a trend amongst the younger generation to spend more in this area[2]. Companies and brands can then capitalize on this for targeted marketing. The younger generation is also more health conscious and are willing to spend more on healthy food[3]. Hence, with a classifier for this, companies can also conduct targeted marketing.

[1] https://www.morganstanley.com/ideas/millennial-boomer-spending

[2] https://www.cnbc.com/2017/06/30/heres-how-millennials-spend-their-money-compared-to-their-parents.html

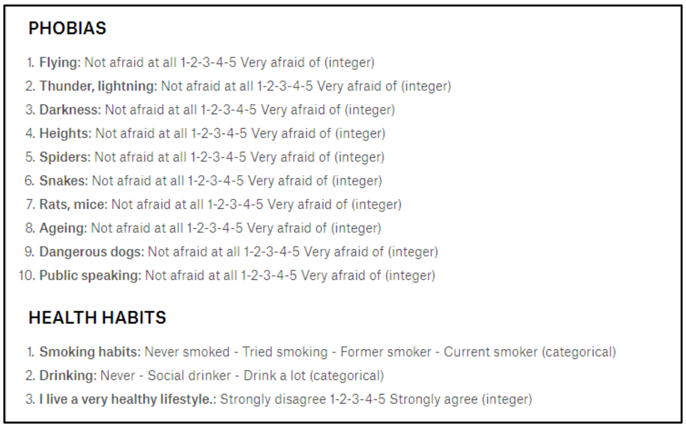
[3] https://www.huffingtonpost.com/elwood-d-watson/younger-consumers-are-tre_b_6632166.html

2.2 Overview of Dataset

Students of the statistic class at FESV UK were asked to invite their friends to participate in this survey. The survey consists of 150 questions or 150 attributes. These survey questions were categorized into the following eight categories:

- Music preferences (19 Questions)
- Movie preferences (12 Questions)
- Hobbies & interests (32 Questions)
- Phobias (10 Questions)
- Health habits (3 Questions)
- Personality traits, views on life, & opinions (57 Questions)
- Spending habits (7 Questions)
- Demographics (10 Questions)

In each category, there are different survey questions pertaining to the categories. The questions under each category make up the attributes for our dataset. The figure below is an example of the list of questions under the "Phobias" and "Health Habits" categories.



PHOBIAS

1. **Flying:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
2. **Thunder, lightning:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
3. **Darkness:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
4. **Heights:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
5. **Spiders:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
6. **Snakes:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
7. **Rats, mice:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
8. **Ageing:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
9. **Dangerous dogs:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
10. **Public speaking:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)

HEALTH HABITS

1. **Smoking habits:** Never smoked - Tried smoking - Former smoker - Current smoker (categorical)
2. **Drinking:** Never - Social drinker - Drink a lot (categorical)
3. **I live a very healthy lifestyle.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)

Figure 1.0 *Questions under Phobias and health habit category*

The full list of categories along with the questions (attributes) can be found in the appendix 9.1.

2.3 Problem Definition

Using the data set "Young people survey", we came up with 3 questions to be answered:

1) Do young people spend on partying and socializing?

The class label for this question will be based on the answer to the question "I spend a lot of money on partying and socializing." in the data set.

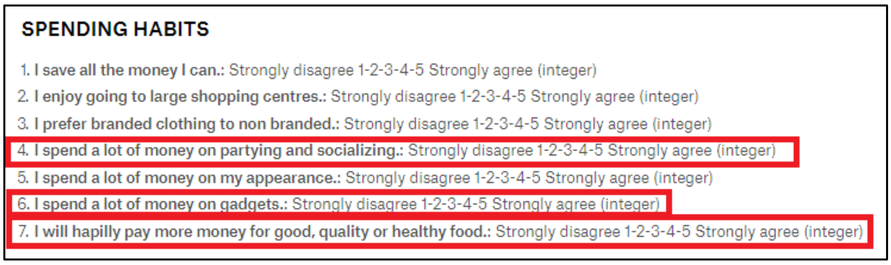
2) Do young people spend on gadgets?

The class label for this question will be based on the answer to the question "I spend a lot of money on gadgets." in the data set.

3) Do young people pay more for good quality and healthy food?

The class label for this question will be based on the answer to the question "I will happily pay more money for good, quality or healthy food." in the data set.

The project aims to find out the most accurate classifier for each question.



SPENDING HABITS

1. **I save all the money I can.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I enjoy going to large shopping centres.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. **I prefer branded clothing to non branded.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. **I spend a lot of money on partying and socializing.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. **I spend a lot of money on my appearance.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. **I spend a lot of money on gadgets.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. **I will hapilly pay more money for good, quality or healthy food.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)

Figure 2.0 *Questions that we will be using as our class labels for our classifier*

2.4 Related Work

While there are many other approaches to the dataset we have chosen, there are no related work or solutions to our questions as the questions are questions that we have come up with.

By coming up with our own questions, we are not focusing on the gaps based on other people's solution (hot solutions on Kaggle) but exploring the issues that other people have not solved. We believe this showcases our innovation and desire to innovate when it comes to data mining and problem creation.

This is also a unique dataset that is not chosen by other groups.

3 Approach

3.1 Methodology

Our data mining process is divided into several phases, starting from data preprocessing, data mining, postprocessing, and finally, knowledge discovery.

Data Preprocessing

Firstly, data cleaning is done to remove noise, inconsistent, and missing data. Then, data transformation is utilised such that the relevant data attributes are suitable for this project's binary classification task.

Next, feature selection, or data reduction, is vital to the project as the dataset involved is large, consisting of 150 attributes. Not only does computational cost increase polynomially as the number of input features increases, irrelevant input features may lead to overfitting or reduce the accuracy of the classification algorithms[4].

Thus, we explore ways to remove entire categories of features that are irrelevant to the classification of the 3 questions, as it is tedious to individually compute the relevance of every one of the 150

[4] http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

attributes. To identify the categories that are most relevant, the accuracy gain/loss when each category is removed from the dataset is computed. The results are used to narrow down the selection of input features.

Data Mining

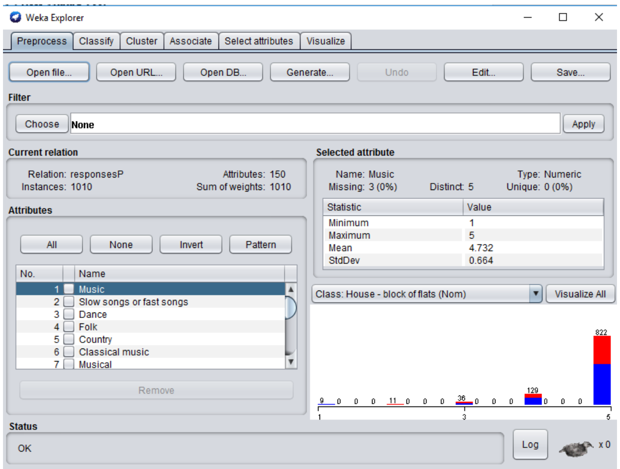
Now, using the reduced input features, classifiers for each of the 3 tasks are built using different algorithms. The performance of the classifiers are evaluated to determine the best algorithms for each question. The details of the algorithms used are further explained in 3.3.

Data Postprocessing

Lastly, the results are analyzed and evaluated using methods such as ROC curves, and any useful or interesting patterns are extracted.

3.2 Data Mining Tool

Weka is a collection of machine learning algorithm for data mining task. It also contains a wide range of tools for data-preprocessing, classification, regression, clustering, association rules, and visualization. Weka is our choice of tool for this project because there are also a wide-range of algorithms available for our use.



Filter: None

Current relation

Relation: responsesP Attributes: 150
Instances: 1010 Sum of weights: 1010

Selected attribute

Name: Music Type: Numeric
Missing: 3 (0%) Distinct: 5 Unique: 0 (0%)

Statistic	Value
Minimum	1
Maximum	5
Mean	4.732
StdDev	0.664

Attributes

No.		Name
1	□	Music
2	□	Slow songs or fast songs
3	□	Dance
4	□	Folk
5	□	Country
6	□	Classical music
7	□	Musical

Class: House - block of flats (Nom)

Status
OK

Figure 2.0 *Weka Interface with raw data input*

3.3 Algorithms

This project is a binary classification task for each of the three output features, and several data mining algorithms will be used to build classifiers and evaluated for their accuracy. The data mining algorithms used in this project are: Decision Tree, Naive Bayes algorithm, Random Forest, and Sequential Minimal Optimization (SMO).

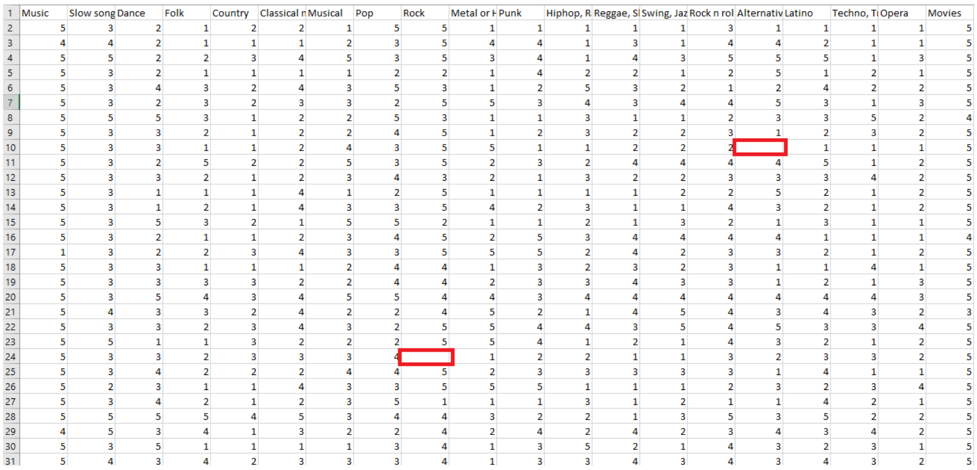
Algorithms	
Decision Tree	Decision tree is a predictive modelling approach where at each node of the tree, an attribute is selected which can most effectively split the samples into subsets. C4.5, implemented in Java as J48 in the Weka library, is an algorithm used to generate a decision tree and its splitting criterion is based on the attribute that can produce the highest information gain, or difference in entropy, which measures the homogeneity of a node. It is a fast and inexpensive classifier to construct.
Naive Bayes	The naive Bayes classifier is a probabilistic classifier which is based on the Bayes theorem. An advantage of this classifier is that it is robust to noise, as well as missing or irrelevant attributes. The main assumption that it holds, also one of its limitations, is that the attributes are independent of one another. It is implemented using the NaiveBayes class in Weka.
Random Forest	Random Forest is an ensemble learning method for classification and regression. An ensemble method predicts a class label by aggregating predictions made by a set of classifiers constructed from the training data. Random Forest operates by constructing multiple decision trees and aggregating their outputs, and corrects the tendency of decision trees to overfit to training data. It is implemented under the RandomForest class in Weka.
SMO	Sequential Minimal Optimisation (SMO), implemented in Weka under the SMO class, is an algorithm for training a support vector machine (SVM), a supervised learning model for classification and regression. It addresses the optimisation problem, or quadratic programming problem in SVMs.

4 Implementations

4.1 Pre-processing

A raw set of data usually has to undergo a process of pre-processing in order to ensure that the data is complete, consistent, and error free. In our project, our pre-processing involves data-cleaning, data transformation and data reduction.

Data Cleaning - We noticed that some of the values for the dataset are missing, and this may affect our classification accuracy. To resolve this, we can either fill up the empty column or ignore these missing values. In the end, we chose the former and wrote a script to scan the data row by row and check for any missing value. If there are missing values, we will remove the row so that our classification accuracy will be accurate.



1	Music	Slow song	Dance	Folk	Country	Classical n	Musical	Pop	Rock	Metal or H	Punk	Hiphop, R	Reggae, Sl	Swing, Jaz	Rock n rol	Alternativ	Latino	Techno, Ti	Opera	Movies
2	5	3	2	1	2	2	1	5	5	1	1	1	1	1	3	1	1	1	1	5
3	4	4	2	1	1	1	2	3	5	4	4	1	3	1	4	4	2	1	1	5
4	5	5	2	2	3	4	5	3	5	3	4	1	4	3	5	5	5	1	3	5
5	5	3	2	1	1	1	1	2	2	1	4	2	2	1	2	5	1	2	1	5
6	5	3	4	3	2	4	3	5	3	1	2	5	3	2	1	2	4	2	2	5
7	5	3	2	3	2	3	3	2	5	5	3	4	3	4	4	5	3	1	3	5
8	5	5	5	3	1	2	2	5	3	1	1	3	1	1	2	3	3	5	2	4
9	5	3	3	2	1	2	2	4	5	1	2	3	2	2	3	1	2	3	2	5
10	5	3	3	1	1	2	4	3	5	5	1	1	2	2	2		1	1	1	5
11	5	3	2	5	2	2	5	3	5	2	3	2	4	4	4	4	5	1	2	5
12	5	3	3	2	1	2	3	4	3	2	1	3	2	2	3	3	3	4	2	5
13	5	3	1	1	1	4	1	2	5	1	1	1	1	2	2	5	2	1	2	5
14	5	3	1	2	1	4	3	3	5	4	2	3	1	1	4	3	2	1	2	5
15	5	3	5	3	2	1	5	5	2	1	1	2	1	3	2	1	3	1	1	5
16	5	3	2	1	1	2	3	4	5	2	5	3	4	4	4	4	1	1	1	4
17	1	3	2	2	3	4	3	3	5	5	5	2	4	2	3	3	2	1	2	5
18	5	3	3	1	1	1	2	4	4	1	3	2	3	2	3	1	1	4	1	5
19	5	3	3	3	3	2	2	4	4	2	3	3	4	3	3	1	2	1	3	5
20	5	3	5	4	3	4	5	5	4	4	3	4	4	4	4	4	4	4	3	5
21	5	4	3	3	2	4	2	2	4	5	2	1	4	5	4	3	4	3	2	3
22	5	3	3	2	3	4	3	2	5	5	4	4	3	5	4	5	3	3	4	5
23	5	5	1	1	3	2	2	2	5	5	4	1	2	1	4	3	2	1	2	5
24	5	3	3	2	3	3	3	4		1	2	2	1	1	3	2	3	3	2	5
25	5	3	4	2	2	2	4	4	5	2	3	3	3	3	3	1	4	1	1	5
26	5	2	3	1	1	4	3	3	5	5	5	1	1	1	2	3	2	3	4	5
27	5	3	4	2	1	2	3	5	1	1	1	3	1	2	1	1	4	2	1	5
28	5	5	5	5	4	5	3	4	4	3	2	2	1	3	5	3	5	2	2	5
29	4	5	3	4	1	3	2	2	4	2	4	2	4	2	3	4	3	4	2	5
30	5	3	5	1	1	1	1	3	4	1	3	5	2	1	4	3	2	3	1	5
31	5	4	3	4	2	3	3	3	4	1	3	3	4	3	4	3	4	3	2	5

Figure 3.0 *Highlighted – empty data, rows will be removed*

Data Integration – As we only have one source for our data, data integration technique is not required for our pre-processing stage.

Data Reduction – Our full dataset contains 150 attributes. However, some attributes may be irrelevant and may even result in a loss of accuracy. Hence, in order to find the relevant attribute that co-relate to our question, we performed ranking and weighing of each category of attribute, also known as subset selection, using Weka. (Explained in 4.3)

Data Transformation - During our Pre-processing stage, we will be transforming the result attributes that are numerical (1 to 5) to nominal attributes (Yes or No) to suit our binary classification task.

The attributes that we have identified to transform are:

Attribute Name	Alias	Numeric (Before Transform)	Nominal (After Transform)
I spend a lot of money on party and socializing	Entertainment Spending	1-5	Yes/No
I spend a lot of money on gadget	Spend on Gadgets	1-5	Yes/No
I will happily pay more money for good, quality or healthy food	Spending on Healthy Eating	1-5	Yes/No

We chose to write a simple PHP script to perform the transformation. So first, the script will read the .xls file row by row and will check for the specific columns to be modified. We set a condition in our script that if the numerical value is < 3, it will be changed to a 'No' value else if numerical value is 3 or more, the original value will be change to a Yes'. This is so that there will only be two distinct result after our applying our classification algorithm.

4.2 Feature Selection - Ranking and Weighing of Attributes

Feature selection is the task where we select only features or attribute that is most relevant to our problem. This will in turn help us create a more accurate classifier, as many data mining algorithm do not perform well with large amounts of features or attributes. It also will help in dimension reduction as a large number of raw attributes can result in the curse of dimensionality. Dimensionality reduction not only speeds up the algorithm execution, but it will also improve the model performance. Hence feature selection technique must be applied before any algorithm can be performed on the dataset.

Weka supports the "Wrapper" method, which could involve a forward, backward or bi-directional search. However, the drawback of this method is that it involves a high computational cost, which our computer is unable to support. Therefore, we have decided to perform a filter selection by checking if the afore mentioned categories are favorable or detrimental to our algorithms.

We first run a J48 classifier with 10-fold cross validation on our full dataset with 150 attributes intact to get a baseline accuracy. We use J48 to rank because it is built on a concept of information gain where the amount of information contained in a dataset is measured. It also gives the idea of importance of an attribute in a dataset.

A 10-fold cross validation is used to avoid overfitting.

Overfitting refers to when an algorithm models the training data too well as the model learns the detail and noise from the training data and negatively impacts the performance of the model on new data sets. Hence to limit overfitting and to improve the performance of our algorithm, we use a resampling technique to estimate model accuracy.

We use the resampling technique 10-fold cross validation. This allows us to train and test our model 10 times on different subsets of training data and deduce an estimate of the performance of the algorithm.

The resultant values as shown in the below table represent the original accuracy.

J48 10-fold Cross Validation	Spend on partying and Socializing	Spend on gadgets	Pay more money for good quality, healthy food
Original Accuracy	66.3667%	58.4158%	75.1485%

Table 4.0 *Baseline accuracy of using J48 classifier with 10-fold cross validation on the full set of data*

We then perform filters for the 7 different categories, one at a time and take note the gain/loss for each category being removed. If there is a gain after removal, it means that a specific category is affecting the accuracy of the classifier and should be removed. Likewise, if removing result in a loss of accuracy, it will mean that the attribute is relevant and helpful in improving the accuracy of the classifier.

Beside these two cases, there are also cases where the accuracy did not change after removal. For these cases, we will take that the category is irrelevant remove it.

Category being removed	No. of Attributes in Category	Spend on partying and Socializing	Spend on gadgets	Pay more money for good quality, healthy food
Music Preference	19	62.4752% Loss 3.8915%	58.6139% Gain 0.1981%	75.1485% No change
Movie Preference	12	62.0792% Loss 4.5545%	58.8819% Gain 0.4661%	75.6436% Gain 0.4951%
Hobbies and Interests	32	65.1485% Loss 1.2812%	55.6436% Loss 2.7722%	76.9307 % Gain 1.7822%
Remove Phobias	10	62.4752% Loss 3.8915	59.3069% Gain 0.8911	71.7822% Loss 3.3622
Health Habits	3	63.6634% Loss 2.7033%	59.1089% Gain 0.6931%	71.7822% Loss 3.3663%
Personality traits, views on life, and opinions	57	64.0594% Loss 2.3073%	60.6931% Gain 2.2773%	72.7723% Loss 2.3762%
Removing demographics	10	65.1485% Loss 1.2181%	57.1287% Loss 1.2871%	73.8614% Loss 1.2871%
Final accuracy after removing		66.3667% Same	59.802% No. Of attributes left: 42	76.4356% No. Of attributes left: 80

Table 4.1 *Results of J48 classifier after removing certain category of attributes*

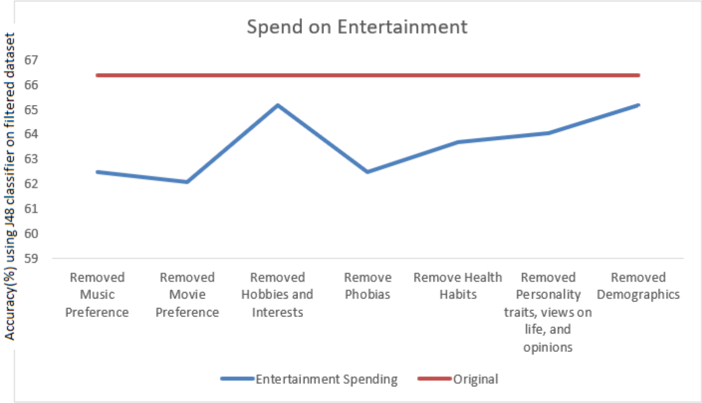


Figure 4.0 *Comparison of Accuracy of J48 classifier on "Entertainment Spending" when certain category of attributes was removed. As shown in the figure, the original accuracy without removing any of the attribute is higher when any of the attributes were removed. Hence, in this case, all attributes should be kept.*

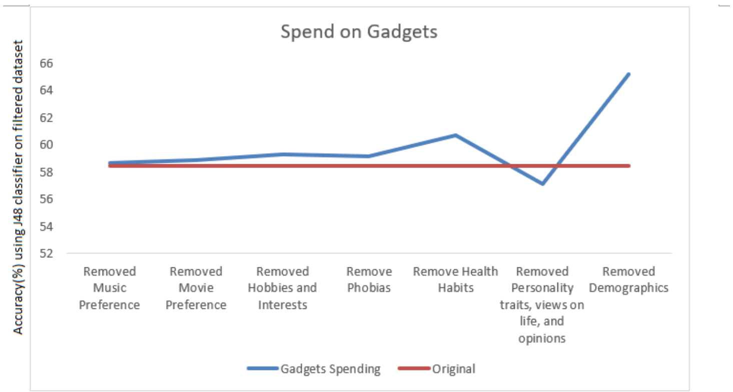


Figure 4.1 Comparison of *Accuracy of J48 classifier on "Gadget Spending" when certain category of attributes was removed. From our experiment, we concluded that the accuracy increases when all except personality, traits, views on life and opinions is removed. Hence, we will remove the rest.*

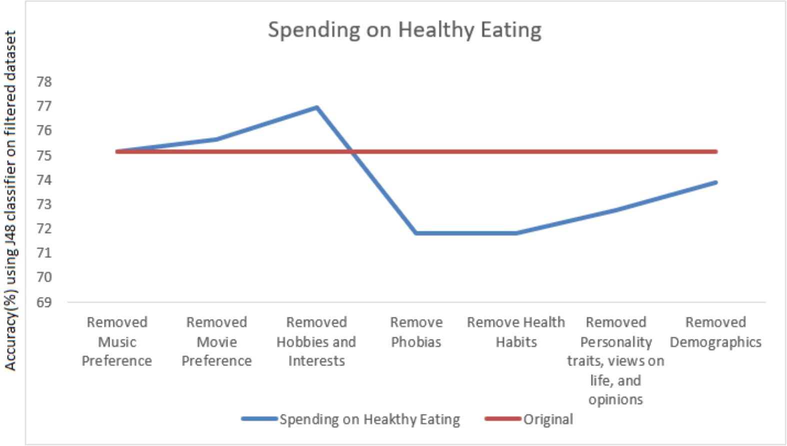


Figure 4.2 Comparison of *Accuracy of J48 classifier on "Spending on Healthy Eating" when certain category of attributes was removed. In this case, only music, movie preference and hobbies and interests reduce the accuracy of the classifier. They are the only categories that were removed in this instance.*

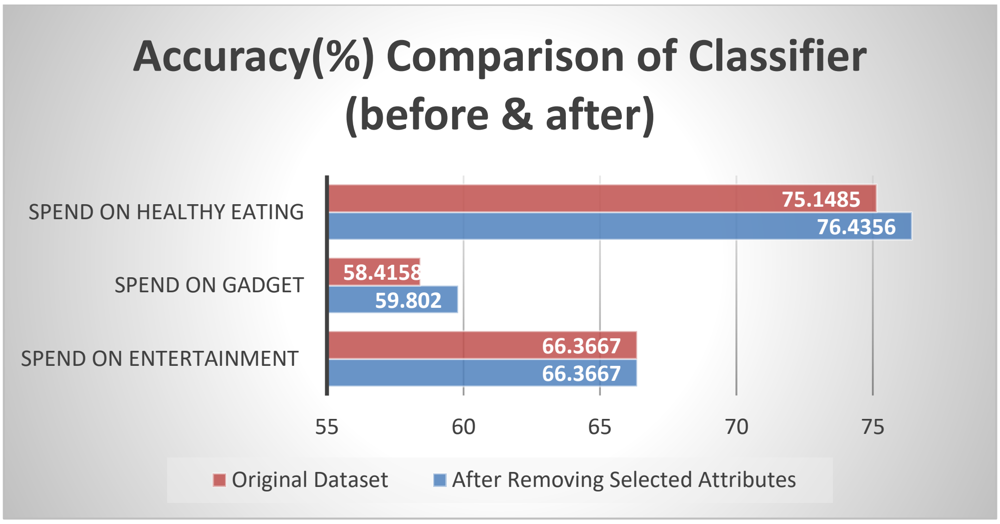
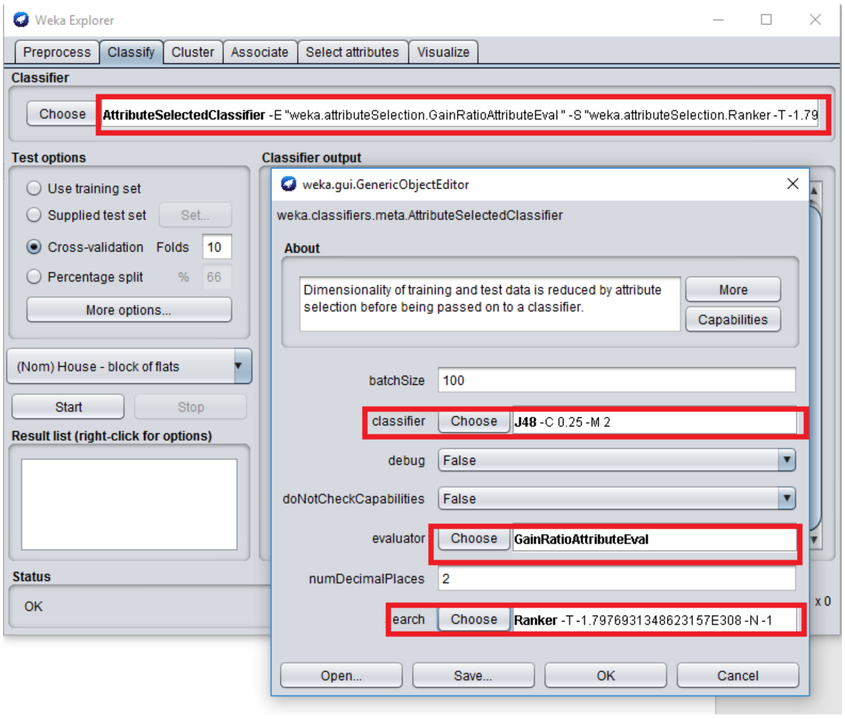


Figure 4.3 *Comparison of Accuracy before and after removing all irrelevant attributes*

After taking out the categories that reduce the accuracy we will rank the remaining attributes by using the "**Attribute Selected Classifier**" option with **J48 Classifier**, **Gain Ratio Evaluator** and **Ranker** search configurations.



Classifier: **AttributeSelectedClassifier** -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.79

○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 10
○ Percentage split   % 66

(Nom) House - block of flats

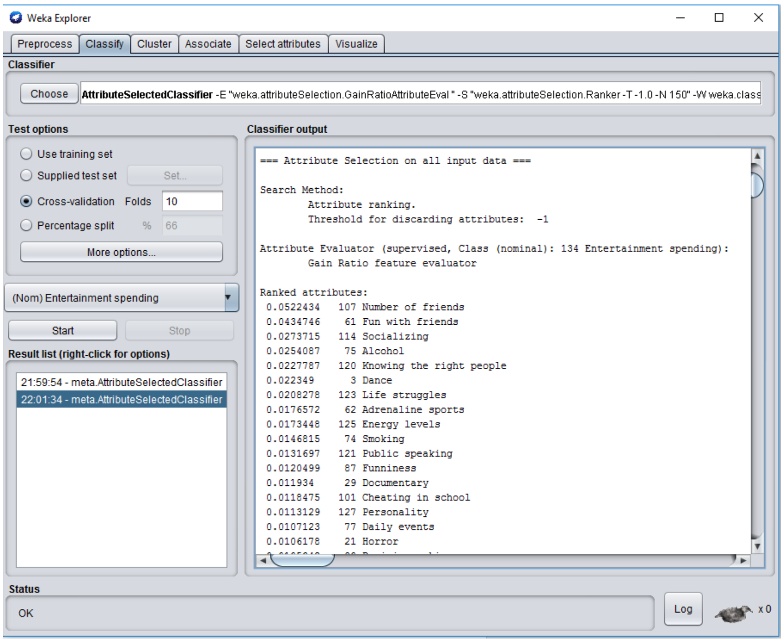
weka.classifiers.meta.AttributeSelectedClassifier

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier.

batchSize: 100
classifier: Choose **J48** -C 0.25 -M 2
debug: False
doNotCheckCapabilities: False
evaluator: Choose **GainRatioAttributeEval**
numDecimalPlaces: 2
earch: Choose **Ranker** -T -1.7976931348623157E308 -N -1

Figure 4.4 *Attribute Selected Classifier with specific configurations*

The classifier then ranked the remaining attributes, running on the dataset with 10-fold cross validation. An example of the output is as shown below.



Classifier: **AttributeSelectedClassifier** -E "weka.attributeSelection.GainRatioAttributeEval " -S "weka.attributeSelection.Ranker -T -1.0 -N 150" -W weka.class

○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 10
○ Percentage split   % 66

(Nom) Entertainment spending

21:59:54 - meta.AttributeSelectedClassifier
22:01:34 - meta.AttributeSelectedClassifier

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.
        Threshold for discarding attributes:  -1

Attribute Evaluator (supervised, Class (nominal): 134 Entertainment spending):
        Gain Ratio feature evaluator

Ranked attributes:
 0.0522434   107 Number of friends
 0.0434746    61 Fun with friends
 0.0273715   114 Socializing
 0.0254087    75 Alcohol
 0.0227787   120 Knowing the right people
 0.022349      3 Dance
 0.0208278   123 Life struggles
 0.0176572    62 Adrenaline sports
 0.0173448   125 Energy levels
 0.0146815    74 Smoking
 0.0131697   121 Public speaking
 0.0120499    87 Funniness
 0.011934     29 Documentary
 0.0118475   101 Cheating in school
 0.0113129   127 Personality
 0.0107123    77 Daily events
 0.0106178    21 Horror
```

Figure 4.5

By using the ranking tool in Weka, we can further remove attributes that have a ranking of less than 0 (deemed irrelevant or of no importance by the ranker). Any attributes with more than 0 ranking was kept. Finally, we are only left with 39 remaining attributes for "Entertainment spending", 20 attributes for "Spending on gadgets" and 27 attributes for "healthy eating".

An attribute with a higher-ranking value means that the attribute has a high relevance of aiding the classifier than the ones ranked lower.

Spending on Gadgets

S/N.	Attribute Name	Ranking	S/N.	Attribute Name	Ranking
1.	Internet	0.048265	11.	Physics	0.010177
2.	Gender	0.044769	12.	Economy Management	0.009044
3.	PC	0.04283	13.	Passive sport	0.007884
4.	Height	0.041364	14.	Fun with friends	0.007319
5.	Weight	0.037118	15.	Law	0.007202
6.	Cars	0.027727	16.	Village - town	0.005779
7.	Science and technology	0.026223	17.	Education	0.001522
8.	Adrenaline sports	0.02316	18.	House - block of flats	0.001363
9.	Reading	0.014137	19.	Left - right	0.001171
10.	Active sport	0.011745	20.	Only child	0.000825

Table 4.2 *Attributes ranked by Weka on Target – Spending on Gadgets*

The top 10 attributes are as such: Internet, Gender, PC, Height, Weight, Cars, Science & Technology, adrenaline sports and reading. From these attributes, we can intuitively see that the ranker is fairly accurate as "Internet", "PC", "Science and Technology" can be seen a huge factor influencing the surveyee's choice on spending on gadgets.

Spending on Healthy Eating

S/N.	Attribute Name	Ranking	S/N.	Attribute Name	Ranking
1.	Happiness in life	0.07842879	15.	Finding lost valuables	0.00735394
2.	Energy levels	0.0382917	16.	Giving	0.00728095
3.	Borrowed stuff	0.02982288	17.	Internet usage	0.0036321
4.	Parents' advice	0.02784256	18.	Punctuality	0.0030568
5.	Health	0.02677654	19.	Education	0.00182427
6.	Healthy eating	0.02539052	20.	Village - town	0.00180742

7.	Knowing the right people	0.01789287	21.	Only child	0.00167038
8.	Eating to survive	0.01306465	22.	Lying	0.00161619
9.	Workaholism	0.01247342	23.	Smoking	0.00113556
10.	Friends versus money	0.01213505	24.	Alcohol	0.00096499
11.	Mood swings	0.01151702	25.	House - block of flats	0.00042274
12.	Compassion to animals	0.01136035	26.	Left - right handed	0.00028654
13.	Interests or hobbies	0.01077628	27.	Gender	0.00000853
14.	Changing the past	0.00755887			

Table 4.3 *Attributes ranked by Weka on Target- Spending on Healthy Eating*

Similarly, like in Table 4.2, some of the most relevant factors are “Happiness in Life”, “Energy levels”, “Parents’ advice”, “Healthy”, “healthy eating” and “eating to survive”, which looks to be accurate.

Spending on Partying and Socializing

S/N.	Attribute Name	Ranking	S/N.	Attribute Name	Ranking
1.	Number of friends	0.0522434	21.	Politics	0.0099622
2.	Fun with friends	0.0434746	22.	Interests or hobbies	0.0098643
3.	Socializing	0.0273715	23.	Questionnaires or polls	0.0095454
4.	Alcohol	0.0254087	24.	Loneliness	0.0095428
5.	Knowing the right people	0.0227787	25.	Gender	0.0091218
6.	Dance	0.022349	26.	Achievements	0.0083755
7.	Life struggles	0.0208278	27.	Happiness in life	0.0082994
8.	Adrenaline sports	0.0176572	28.	Slow songs or fast songs	0.0082954
9.	Energy levels	0.0173448	29.	Giving	0.007639
10.	Smoking	0.0146815	30.	Active sport	0.0074208
11.	Public speaking	0.0131697	31.	Passive sport	0.0068915
12.	Funniness	0.0120499	32.	Punctuality	0.0052337
13.	Documentary	0.011934	33.	Internet usage	0.0033787
14.	Cheating in school	0.0118475	34.	Village - town	0.0026885
15	Personality	0.0113129	35.	Lying	0.0025241

16.	Daily events	0.0107123	36.	Education	0.0022403
17.	Horror	0.0106178	37.	Left - right handed	0.0006478
18.	Decision making	0.0105342	38.	Only child	0.0000843
19.	Thriller	0.0099826	39.	House - block of flats	0.0000156
20.	Flying	0.0099785			

Table 4.4 *Attributes ranked by Weka on Target- Spending on Partying and Socializing*

The more relevant attributes which co-relate to spending money on entertainment seems to be tied very closely to a person's social life, so as shown in the table, attributes that relates to friends and socializing activities are rated very highly.

5 Experimental Results and Analysis

5.1 Comparison Schemes

"Correctly Classified Instances" is used as a comparison scheme for results and analysis. It shows the percentage of test instances that were correctly classified over all the instances classified.

The reason why we chose to use correctly classified instances instead of the other values such as error values is because error rates are typically used for numeric prediction rather than classification. These values reflect errors in magnitude in numeric predictions which is not our primary concern as we are looking at binary classification.

However, using percentage of correctly classified instances alone is insufficient because it is not chance-corrected and insensitive to class distribution.

Hence, we also chose to use the Receiver Operating Characteristic (ROC) curve as a comparison scheme.

The ROC curve characterizes the trade-off between positive hits and false alarms. In general, at some threshold of algorithm, the higher the true positive rate and the lower the false positive rate, the better the performance of the algorithm.

In evaluating this binary classification system, it is essential to note that we do not take the above comparison schemes in an isolated manner. This is because different schemes offer different insights for comparison of results and for a fair comparison, we utilize both comparison schemes.

5.2 Results and Analysis

	Spend on partying and socializing	Spend on gadgets	Pay more money for good quality, healthy food
J48 (10-folds Cross Validation)	Correctly Classified Instances: 68.0198%	Correctly Classified Instances: 62.9703%	Correctly Classified Instances: 80.5941%
Naive Bayes	Correctly Classified Instances: 73.0693%	Correctly Classified Instances: 64.9505%	Correctly Classified Instances: 80%
Random forest	Correctly Classified Instances: 74.0594 %	Correctly Classified Instances: 67.2277 %	Correctly Classified Instances: 82.5743 %
SMO	Correctly Classified Instances: 73.9604 %	Correctly Classified Instances: 66.3366 %	Correctly Classified Instances: 82.6733 %

Legend
Green – Highest Accuracy for the specific spending
Yellow – Lowest Accuracy for the specific spending

From the results table above, we did a comparison for the different algorithms using subset of attributes.

For J48 algorithm, we can see that cross-validation folds for 10 folds had shown that it obtained the accuracy of correctly classified instances of 62.9703% for “Spending on gadgets” and accuracy of correctly classified instances of 68.0198% for “Spending on partying and socializing”, which are one of the lowest accuracy that can be obtained as compared across the different algorithm. However, for “Spending on healthy eating”, it has an accuracy of correctly classified instances of 80.5941% which is slightly better than Naïve Bayes algorithm by 0.5941% in this case.

For Naïve Bayes algorithm, we can see that the correctly classified instances across the three different kinds of spending indicated that “Spending for healthy eating” has an accuracy of correctly classified instances of 80% which is the lowest accuracy across all the different algorithm, while “Spending on gadgets” has an accuracy of correctly classified instances of 64.9505% and “Spending on partying and socializing” has an accuracy of correctly classified instances of 73.0693%, which appears to be mediocre.

For Random Forest algorithm, we analyzed and observed that the correctly classified instances across the three different kinds of spending indicated that “Spending for healthy eating” has an accuracy of correctly classified instances of 82.5743%, which is the second highest accuracy across all algorithms that we experimented with SMO being the highest, however, “Spending on gadgets” has the an accuracy of correctly classified instances of 67.2277%, which is the highest across all algorithms and “Spending on partying and socializing” has an accuracy of correctly classified instances of 74.0594%, which also constitutes to being the highest accuracy across all algorithms.

Lastly, for Sequential Minimum Optimisation (SMO) algorithm, we can see that the correctly classified instances across the three different kinds of spending indicated that “Spending for healthy eating” has the highest accuracy of correctly classified instances of 82.6733% across all algorithms that we experimented with a slight difference of 0.099% improved as compared to Random Forest algorithm, while “Spending on gadgets” has an accuracy of correctly classified instances of 66.3366% and “Spending on partying and socializing” has an accuracy of correctly classified instances of 73.9604%, which ranked second highest for these two spending.

Overall, from the comparison that we deduced above, Random Forest algorithm outperforms the rest of the algorithms with just 0.099% lesser accuracy compared to SMO under “Spending for healthy eating”, but still outperforms for “Spending on gadgets” with accuracy of 67.2277% and “Spending on partying and socializing” with accuracy of 74.0594%.

Similarly, our results can be shown using a Receiver Operating Characteristic curve (ROC curve). A ROC curve shows the tradeoff between sensitivity (true positive rate) and specificity (false positive rate): any increase in sensitivity will be accompanied by a decrease in specificity.

It is typically used to evaluate the performance of machine learning algorithms hence, we can use the ROC to represent the performance of our algorithms.

The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the classifier. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the classifier.

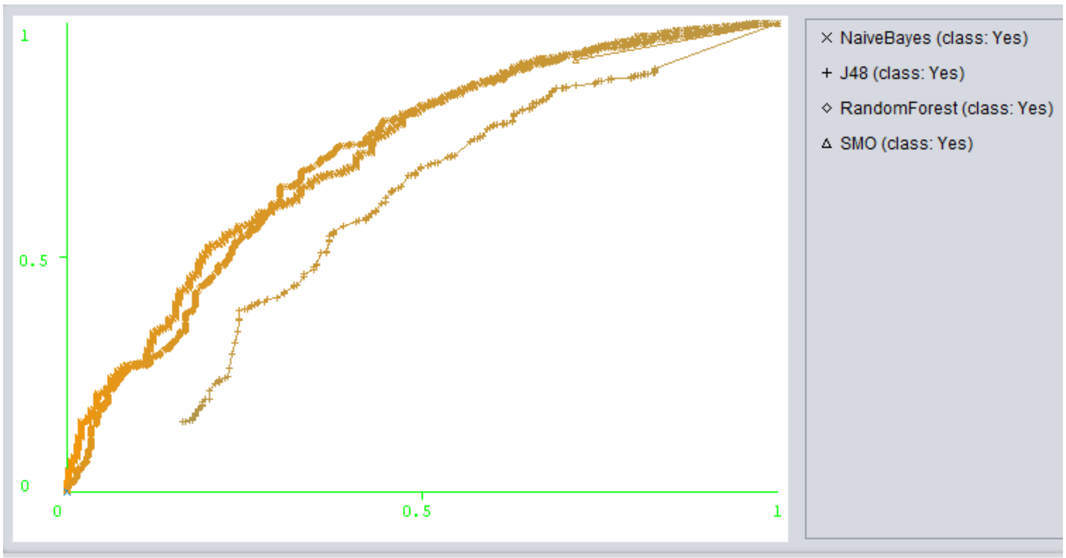


Figure 5.0 *ROC curve for class: partying & socializing*

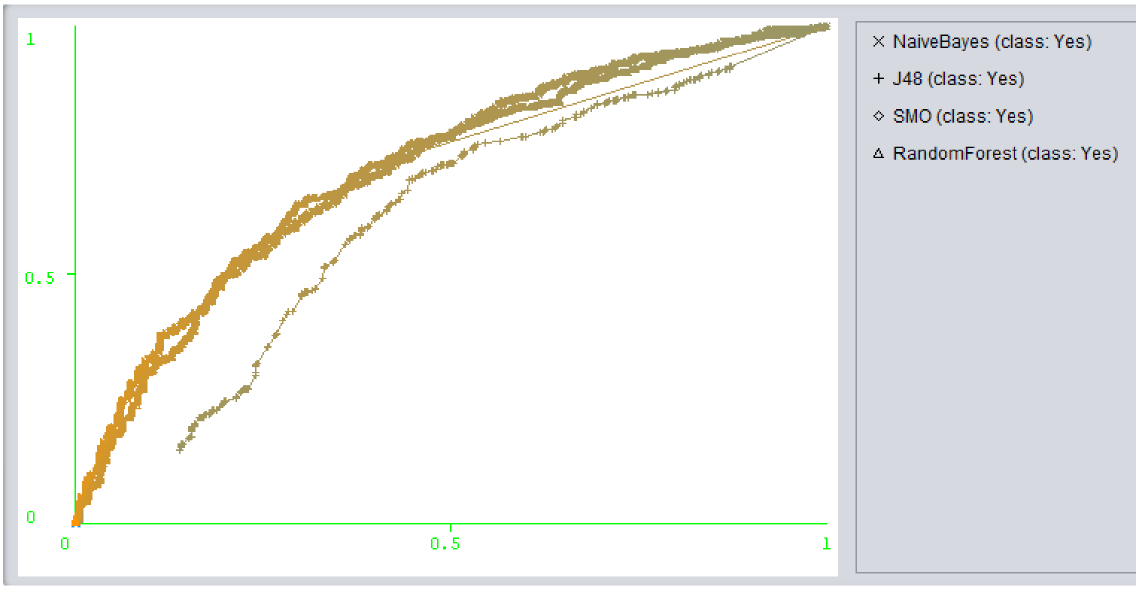


Figure 5.1 *ROC curve for class: spending on gadgets*

From Figure 5.0 and 5.1, as the ROC curve of the Random Forest algorithm hugs the left axis and the top border the most, it can be concluded that for people classifying people who spend on partying and socializing and to classify people who spend on gadgets, the Random Forest algorithm is the best.

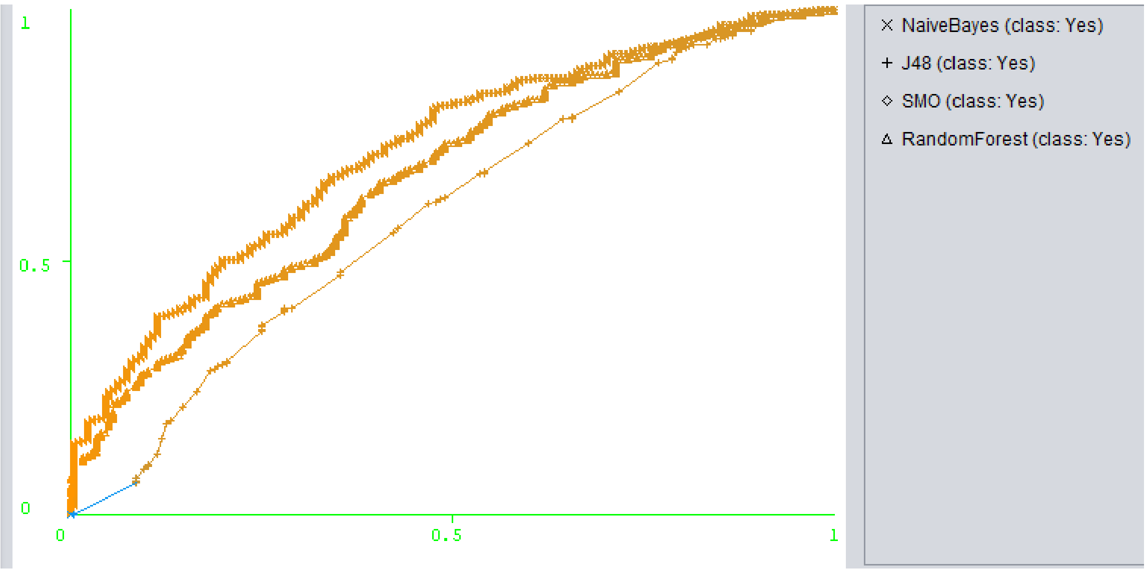


Figure 5.2 *ROC curve for class: willing to spend more on good quality, healthy food*

From Figure 5.2, as the ROC curve of the SMO algorithm hugs the left axis and the top border the most, it can be concluded that for people classifying people who are willing to spend more on good quality, healthy food, the SMO algorithm is the best for classifying.

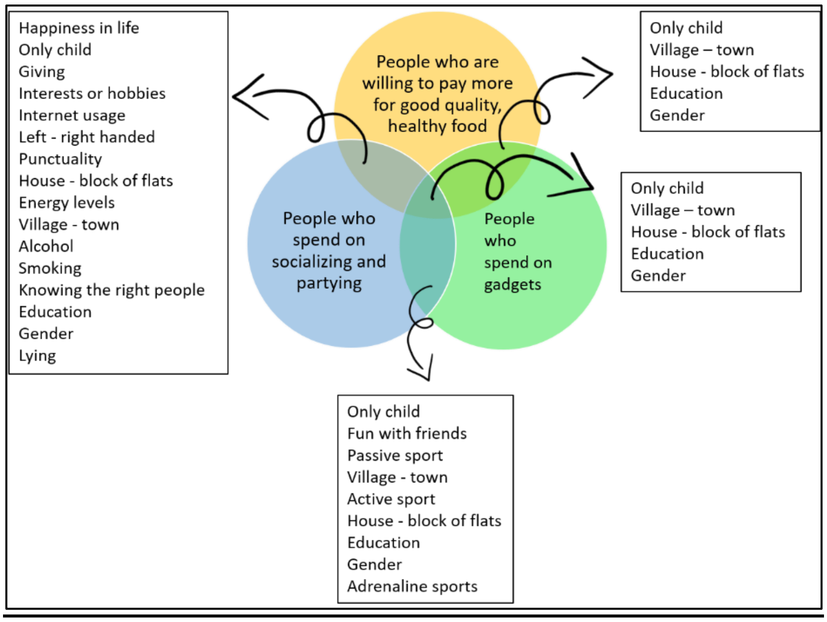


Figure 5.3 *Venn Diagram for similarities in attributes that predict spending habits of Millennials*

The Venn diagram above shows the common attributes that are used to predict the classes of the 3 different questions. By finding out these common attributes, we can path a way for future use of these results. For example, if a brand or company wants to classify people who spend on both gadgets and on socializing and partying, the company can easily narrow down the selection of attributes to the ones that are common between the two.

6 Discussion of Pros and Cons

In this section, we will discuss the pros and cons of the respective chosen algorithms:

Random Forest

Random Forest has shown overall highest accuracy amongst the 4 chosen algorithms in our results. It is also known to be one of the most accurate learning algorithms available and produce a highly accurate classifier. This is because the Random Forest algorithm can run efficiently on large data sets, give estimates of what variables are significant in the classification and generate an internal unbiased estimate of the generalization error during forest building. In addition, forests generated can also be saved for future use on other data.

Sequential Minimal Optimization (SMO)

SMO breaks up the large quadratic programming (QP) optimization problem into a series of small QP problems which are solved analytically and thus, avoids using a time-consuming numerical QP optimization as an inner loop. Each sub-problem is solved at a fast rate such that the overall QP problem is solved quickly. Hence, SMO's computation time is very efficient.

SMO does not require extra matrix storage which allows large SVM training problems to be able to fit inside main memory. Furthermore, SMO does not use matrix algorithm which causes it to be less susceptible to numerical precision problems.

Naïve Bayes

Naïve Bayes is easy to implement and robust to isolated noise points and irrelevant attributes. It also requires only a small training data set to estimate the necessary classification parameters.

Despite so, Naïve Bayes holds the assumption of class conditional independence which could result in accuracy loss. This is because in the real world, dependency do exist frequently among variables and the naïve independence assumption may not hold. This may be the reason why this algorithm did not perform well in our experiment as our attributes (the survey questions) might not be independent and hence, the independent assumption will not hold.

J48 Decision Tree

J48 Decision Trees are inexpensive to construct, easy to interpret and extremely fast at classifying unknown records.

However, J48 algorithm requires the entire data to fit in memory and thus prove to be unsuitable for large data sets which will need out-of-core sorting. Calculations can also get very complex especially if there are many uncertain values or when the outcomes are linked. It also tends to overfit, which might be why the performance was not as good.

7 Conclusions

7.1 Summary of project achievements

As the Millennials are now the largest group of consumers, it is important to find out the spending habits of them on different areas and the distinct traits of everyone who prioritize and utilize their monetary expenses on specific area through this project analysis.

Different data mining algorithms; Decision Tree, Naive Bayes algorithm, Random Forest, and Sequential Minimal Optimization (SMO) were adopted through this analysis, to capture the most accurate reading of our dataset.

To elaborate on the point mentioned about how we go about achieving the optimal accuracy of our algorithms, we firstly start off with 150 attributes on our full dataset to get a baseline accuracy of 66.3667% for “Spend on partying and socializing”, 58.4158% for “Spend on gadgets”, and 75.1485% for “Spend on healthy eating” using J48 classifier with 10-fold cross validation.

To better assess the relative importance of each attributes, we performed filters on the 7 different categories (Music Preferences, Movie Preferences, Hobbies and Interests, Phobias, Health Habits, Personality traits, views on life, and opinions, and Demographics) one by one, noting down the gain/loss and deduced that the final accuracy was 66.3667% for “Spend on partying and socializing” with same % intact, 59.802% for “Spend on gadgets” with 1.3862% improved, and 76.4356% for “Spend on healthy eating” with 1.2871% improved, after removing of selected attributes for these 7 categories.

Extensive effort had also been made to further eliminate attributes that have ranking of less than 0 (an attribute with a higher-ranking value means that the attribute has a higher predictive performance than the ones ranked lower), using the “Attribute Selected Classifier” option with J48 Classifier, Gain

Ratio Evaluator and Ranker search configurations, which resulted with 39 remaining attributes for "Spending on partying and socializing",", 20 attributes for "Spending on gadgets", and 27 attributes for "Spending on healthy eating".

With this, we further presented our results using Receiver Operating Characteristic curve (ROC curve) which is a plot of the true positive rate against the false positive rate for different threshold settings. It is also said that the closer the curve follows the left-hand border and the top border of the ROC space, the more accurate the classifier of the algorithms. Looking at the ROC curve that we had plotted, it is showing that our algorithms are leaning more towards the left-hand and the top border of the ROC space, which also implied and testified that the accuracy of the classifier of our algorithms hold.

Having the millennials as the largest group of consumers, the ability to accurately predict and identify factors related to their spending will provide ample opportunities for improving the quality of businesses to know where and which areas to focus on to attract them. From there, ideas can be further generated, and businesses are probable to prosper. Having said that, this project analysis will also allow people to better relate where are the millennials' interests leaning towards in the 21st century today.

7.2 Future Directions for improvements

Apart from classifying individuals according to their spending habits, from the attributes we have selected we can further analyze the value of the attributes (i.e. the numeric 1-5 value provided by the individuals). From that we can see how much the qualitative value of the attribute affects the result. For example, we can find the mean value of attribute "Fun with friends" to see whether if it's the people who like to have fun with friends (i.e. value >= 3) the ones who spend more on partying and socializing or those who do not like to have fun with friends (i.e. value < 3).

Also, we can expand our dataset by combining it with other additional dataset that also have information on young people. This way, we can have a more comprehensive dataset on this target group and we can hence use the data to learn more about them. In this dataset, as we are working with answers to survey questions as attributes, it is not very appropriate to create new attributes, however, with the combination of a new data set, we can then create new attributes. With a more comprehensive dataset on young people, we can mine for more data and can hence create more room for analyzation and discussion on topics regarding them.

More predictive questions can also be further drilled down/derived while doing this analytical report, for instance:

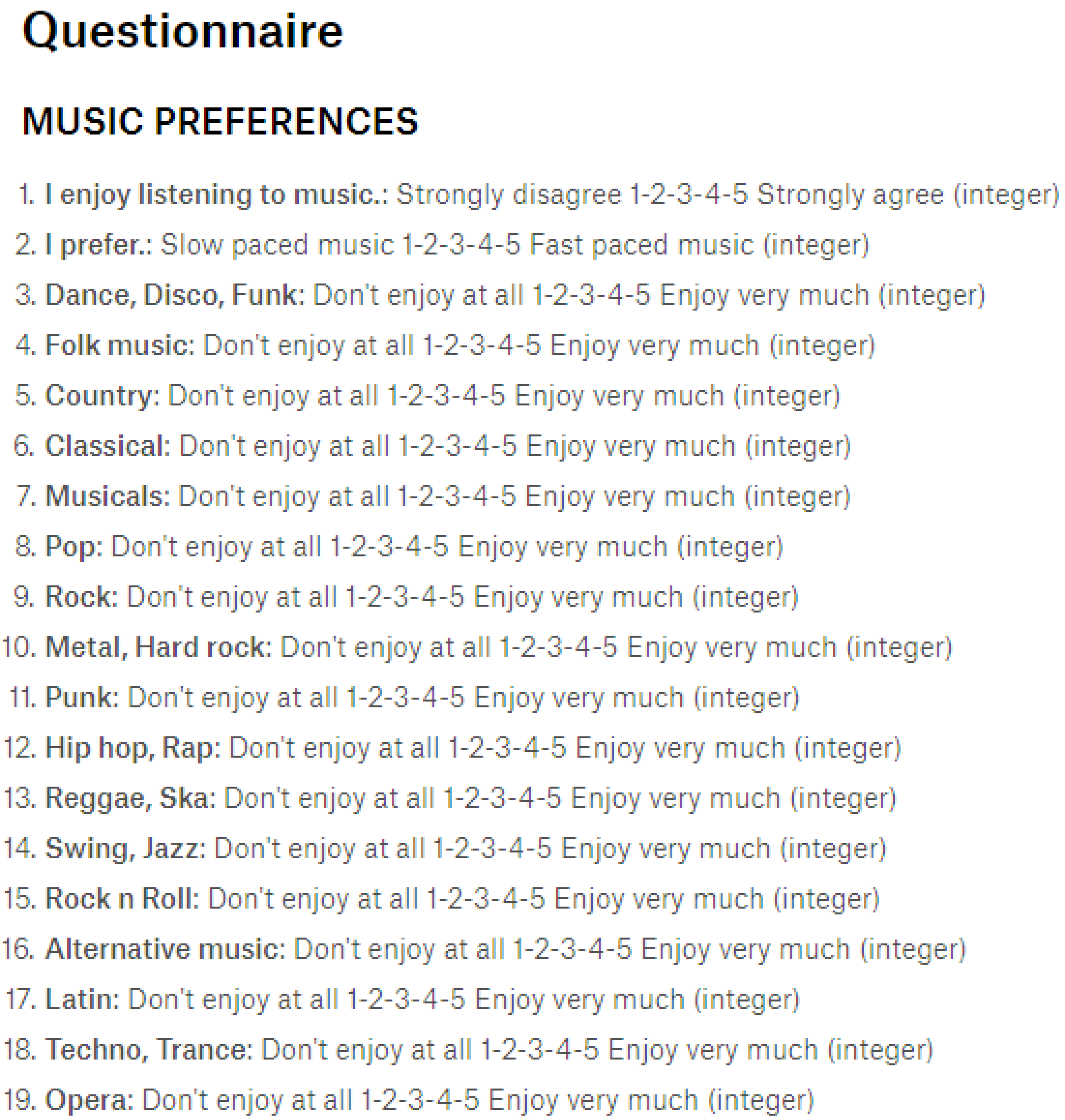
1) What are the specific kinds of entertainment the millennials are spending more on?
2) What are the specific kinds of gadgets the millennials are keen in?
3) What are the specific kinds of healthy food and habit the millennials are eyeing on?

With the specific activity or item, we can further compare it with another one, for example:

1) Why millennials preferred watching late night movies more than going night bars?
2) Why millennials preferred headphones over earpieces?
3) Why millennials preferred salads for meals over sandwiches for meals?

8 Appendix

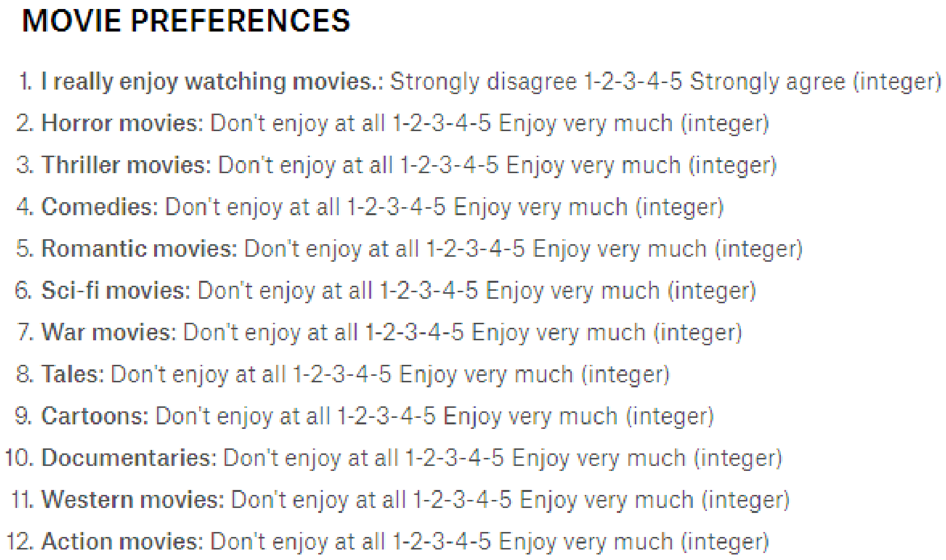
8.1 Dataset Reference



Questionnaire

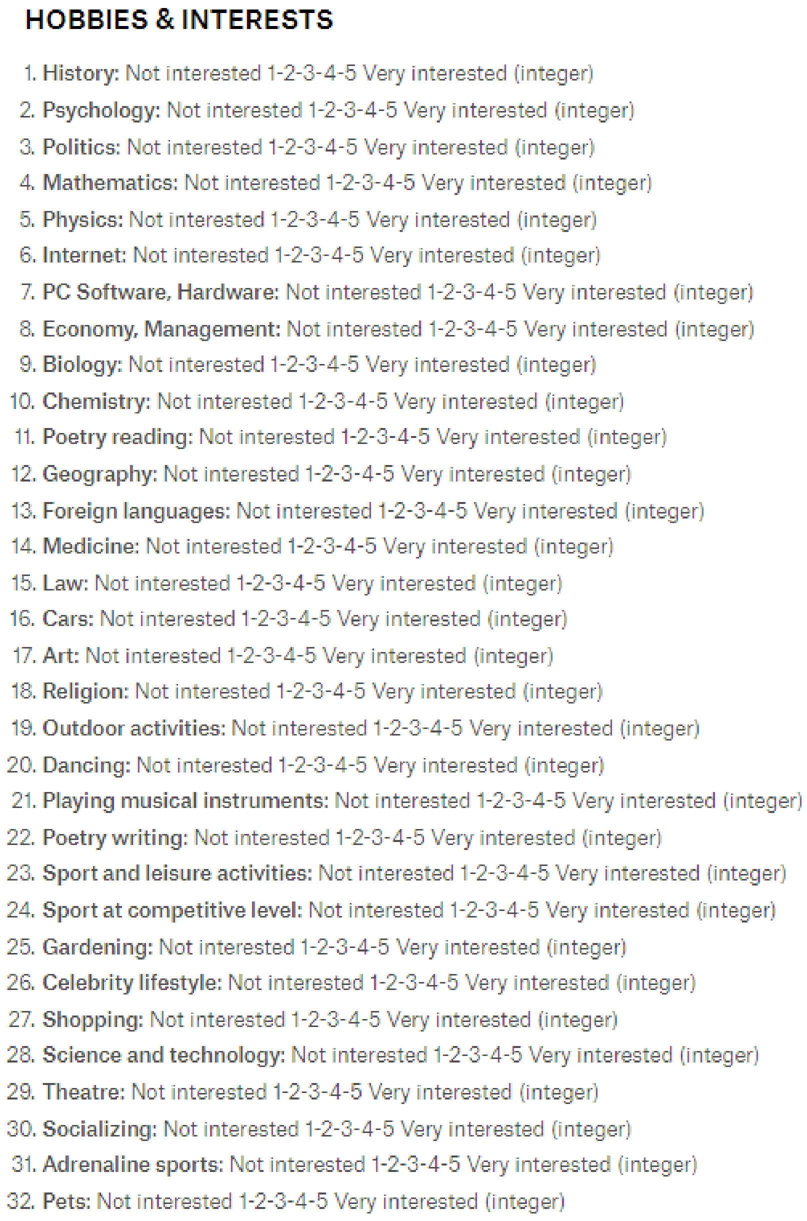
MUSIC PREFERENCES

1. **I enjoy listening to music.**: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I prefer.**: Slow paced music 1-2-3-4-5 Fast paced music (integer)
3. **Dance, Disco, Funk**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
4. **Folk music**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
5. **Country**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
6. **Classical**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
7. **Musicals**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
8. **Pop**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
9. **Rock**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
10. **Metal, Hard rock**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
11. **Punk**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
12. **Hip hop, Rap**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
13. **Reggae, Ska**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
14. **Swing, Jazz**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
15. **Rock n Roll**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
16. **Alternative music**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
17. **Latin**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
18. **Techno, Trance**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
19. **Opera**: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)



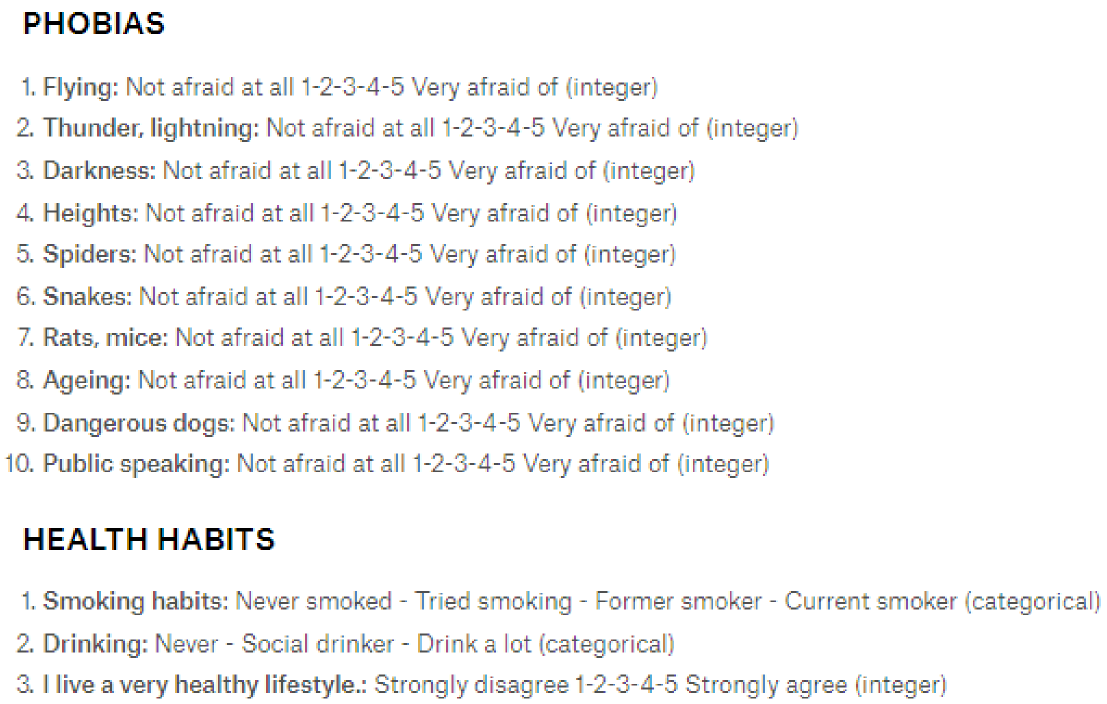
MOVIE PREFERENCES

1. **I really enjoy watching movies.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **Horror movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
3. **Thriller movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
4. **Comedies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
5. **Romantic movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
6. **Sci-fi movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
7. **War movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
8. **Tales:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
9. **Cartoons:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
10. **Documentaries:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
11. **Western movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
12. **Action movies:** Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)



HOBBIES & INTERESTS

1. **History:** Not interested 1-2-3-4-5 Very interested (integer)
2. **Psychology:** Not interested 1-2-3-4-5 Very interested (integer)
3. **Politics:** Not interested 1-2-3-4-5 Very interested (integer)
4. **Mathematics:** Not interested 1-2-3-4-5 Very interested (integer)
5. **Physics:** Not interested 1-2-3-4-5 Very interested (integer)
6. **Internet:** Not interested 1-2-3-4-5 Very interested (integer)
7. **PC Software, Hardware:** Not interested 1-2-3-4-5 Very interested (integer)
8. **Economy, Management:** Not interested 1-2-3-4-5 Very interested (integer)
9. **Biology:** Not interested 1-2-3-4-5 Very interested (integer)
10. **Chemistry:** Not interested 1-2-3-4-5 Very interested (integer)
11. **Poetry reading:** Not interested 1-2-3-4-5 Very interested (integer)
12. **Geography:** Not interested 1-2-3-4-5 Very interested (integer)
13. **Foreign languages:** Not interested 1-2-3-4-5 Very interested (integer)
14. **Medicine:** Not interested 1-2-3-4-5 Very interested (integer)
15. **Law:** Not interested 1-2-3-4-5 Very interested (integer)
16. **Cars:** Not interested 1-2-3-4-5 Very interested (integer)
17. **Art:** Not interested 1-2-3-4-5 Very interested (integer)
18. **Religion:** Not interested 1-2-3-4-5 Very interested (integer)
19. **Outdoor activities:** Not interested 1-2-3-4-5 Very interested (integer)
20. **Dancing:** Not interested 1-2-3-4-5 Very interested (integer)
21. **Playing musical instruments:** Not interested 1-2-3-4-5 Very interested (integer)
22. **Poetry writing:** Not interested 1-2-3-4-5 Very interested (integer)
23. **Sport and leisure activities:** Not interested 1-2-3-4-5 Very interested (integer)
24. **Sport at competitive level:** Not interested 1-2-3-4-5 Very interested (integer)
25. **Gardening:** Not interested 1-2-3-4-5 Very interested (integer)
26. **Celebrity lifestyle:** Not interested 1-2-3-4-5 Very interested (integer)
27. **Shopping:** Not interested 1-2-3-4-5 Very interested (integer)
28. **Science and technology:** Not interested 1-2-3-4-5 Very interested (integer)
29. **Theatre:** Not interested 1-2-3-4-5 Very interested (integer)
30. **Socializing:** Not interested 1-2-3-4-5 Very interested (integer)
31. **Adrenaline sports:** Not interested 1-2-3-4-5 Very interested (integer)
32. **Pets:** Not interested 1-2-3-4-5 Very interested (integer)

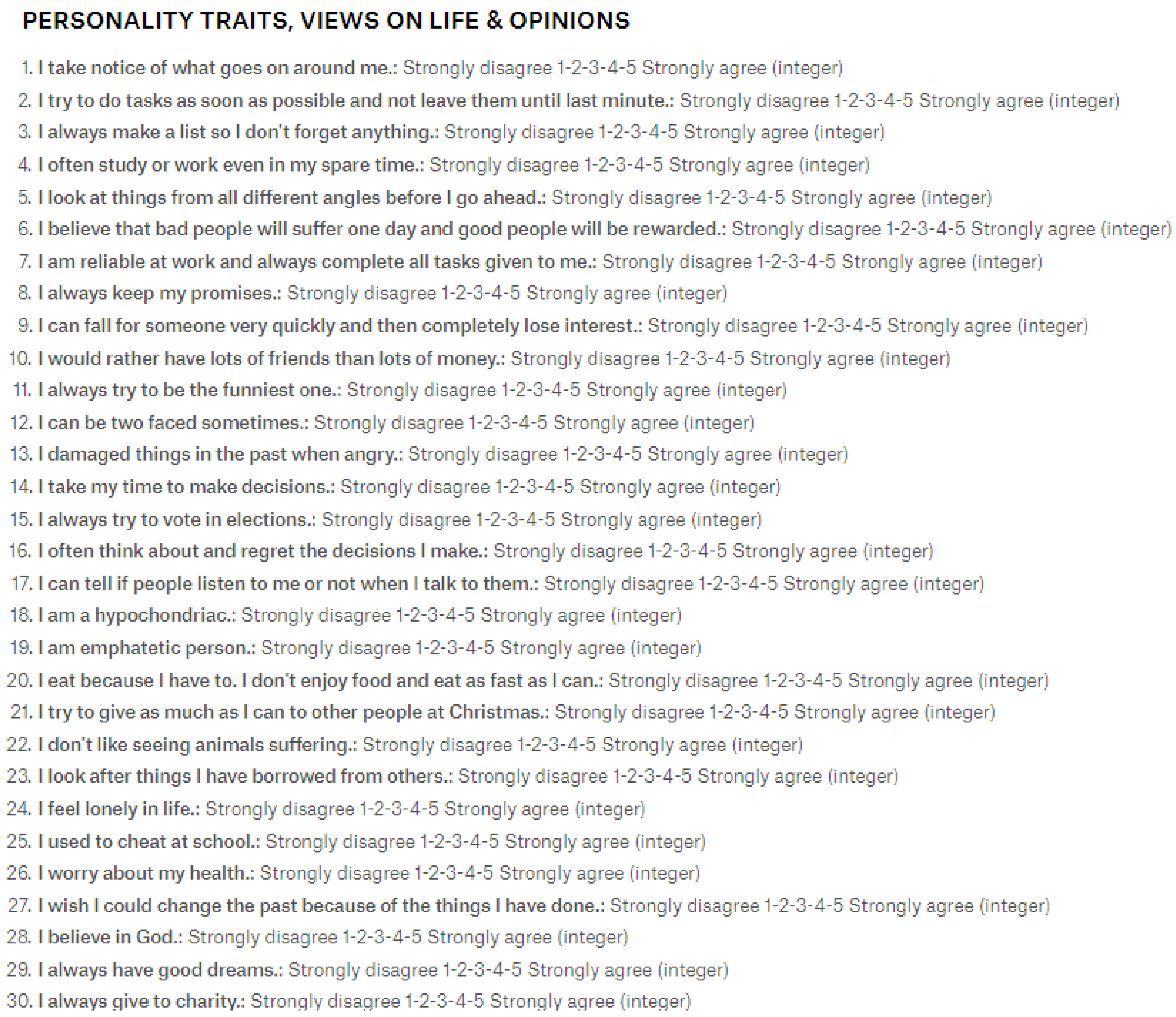


PHOBIAS

1. **Flying:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
2. **Thunder, lightning:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
3. **Darkness:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
4. **Heights:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
5. **Spiders:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
6. **Snakes:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
7. **Rats, mice:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
8. **Ageing:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
9. **Dangerous dogs:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)
10. **Public speaking:** Not afraid at all 1-2-3-4-5 Very afraid of (integer)

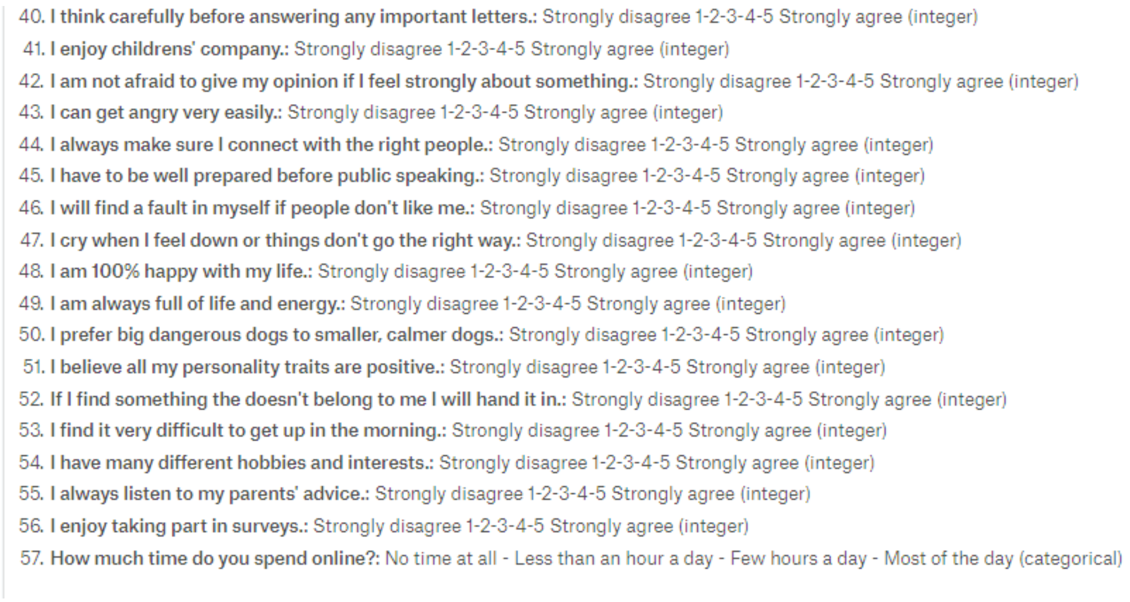
HEALTH HABITS

1. **Smoking habits:** Never smoked - Tried smoking - Former smoker - Current smoker (categorical)
2. **Drinking:** Never - Social drinker - Drink a lot (categorical)
3. **I live a very healthy lifestyle.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)

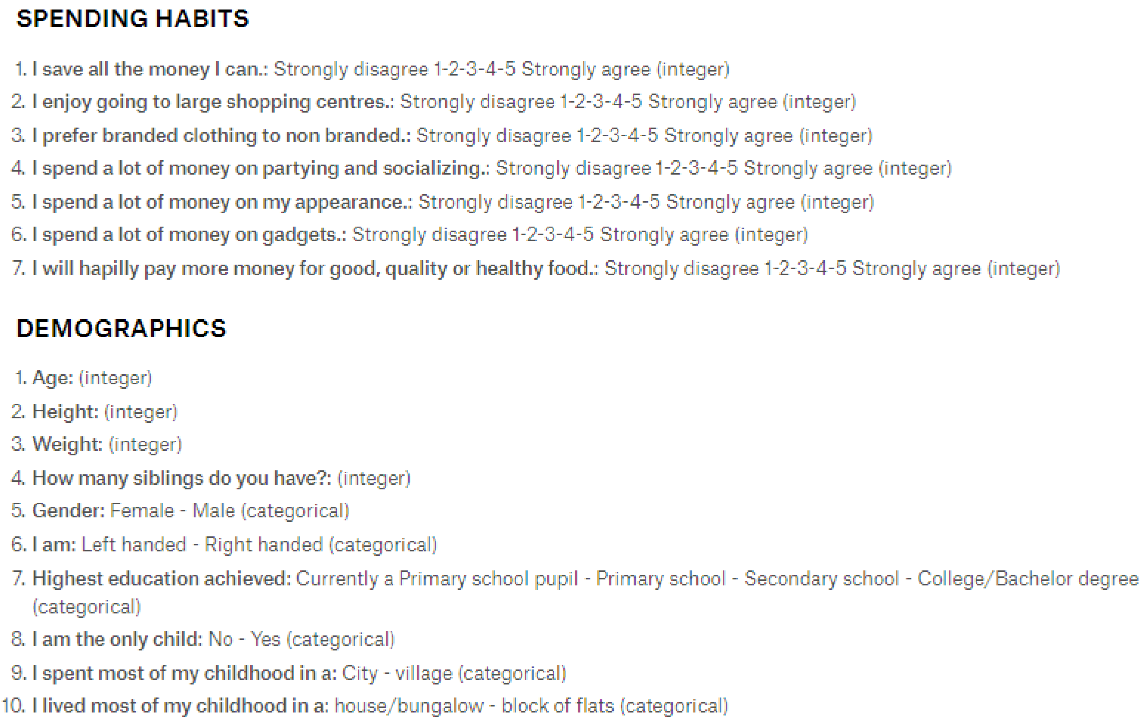


PERSONALITY TRAITS, VIEWS ON LIFE & OPINIONS

1. **I take notice of what goes on around me.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I try to do tasks as soon as possible and not leave them until last minute.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. **I always make a list so I don't forget anything.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. **I often study or work even in my spare time.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. **I look at things from all different angles before I go ahead.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. **I believe that bad people will suffer one day and good people will be rewarded.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. **I am reliable at work and always complete all tasks given to me.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
8. **I always keep my promises.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
9. **I can fall for someone very quickly and then completely lose interest.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
10. **I would rather have lots of friends than lots of money.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
11. **I always try to be the funniest one.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
12. **I can be two faced sometimes.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
13. **I damaged things in the past when angry.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
14. **I take my time to make decisions.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
15. **I always try to vote in elections.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
16. **I often think about and regret the decisions I make.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
17. **I can tell if people listen to me or not when I talk to them.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
18. **I am a hypochondriac.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
19. **I am emphatetic person.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
20. **I eat because I have to. I don't enjoy food and eat as fast as I can.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
21. **I try to give as much as I can to other people at Christmas.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
22. **I don't like seeing animals suffering.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
23. **I look after things I have borrowed from others.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
24. **I feel lonely in life.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
25. **I used to cheat at school.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
26. **I worry about my health.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
27. **I wish I could change the past because of the things I have done.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
28. **I believe in God.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
29. **I always have good dreams.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
30. **I always give to charity.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)



40. **I think carefully before answering any important letters.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
41. **I enjoy childrens' company.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
42. **I am not afraid to give my opinion if I feel strongly about something.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
43. **I can get angry very easily.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
44. **I always make sure I connect with the right people.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
45. **I have to be well prepared before public speaking.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
46. **I will find a fault in myself if people don't like me.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
47. **I cry when I feel down or things don't go the right way.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
48. **I am 100% happy with my life.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
49. **I am always full of life and energy.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
50. **I prefer big dangerous dogs to smaller, calmer dogs.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
51. **I believe all my personality traits are positive.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
52. **If I find something the doesn't belong to me I will hand it in.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
53. **I find it very difficult to get up in the morning.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
54. **I have many different hobbies and interests.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
55. **I always listen to my parents' advice.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
56. **I enjoy taking part in surveys.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
57. **How much time do you spend online?:** No time at all - Less than an hour a day - Few hours a day - Most of the day (categorical)



SPENDING HABITS

1. **I save all the money I can.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
2. **I enjoy going to large shopping centres.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
3. **I prefer branded clothing to non branded.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
4. **I spend a lot of money on partying and socializing.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
5. **I spend a lot of money on my appearance.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
6. **I spend a lot of money on gadgets.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)
7. **I will hapilly pay more money for good, quality or healthy food.:** Strongly disagree 1-2-3-4-5 Strongly agree (integer)

DEMOGRAPHICS

1. **Age:** (integer)
2. **Height:** (integer)
3. **Weight:** (integer)
4. **How many siblings do you have?:** (integer)
5. **Gender:** Female - Male (categorical)
6. **I am:** Left handed - Right handed (categorical)
7. **Highest education achieved:** Currently a Primary school pupil - Primary school - Secondary school - College/Bachelor degree (categorical)
8. **I am the only child:** No - Yes (categorical)
9. **I spent most of my childhood in a:** City - village (categorical)
10. **I lived most of my childhood in a:** house/bungalow - block of flats (categorical)

8.2 Source Codes

```
Preprocess.php

<?php

ini_set('max_execution_time', 300);
$filename = 'responses.csv';
$contents = file($filename);
$file = fopen("responses.csv","r");
$file2 = "responsesP.csv";
$file3 = fopen("responsesP.csv","w");
 foreach($contents as $line) {
if(!strpos($line,',,')){
$data = explode(",",$line);

        //Convert finance Row
        if($data[133]==null || $data[133]==1 || $data[133]==2)
        {
                $data[133]='No';
        }
        else if($data[133]==3 || $data[133]==4 || $data[133]==5 ){
                $data[133] = 'Yes';
        }

        //Convert Shopping centre Row
        if($data[134]==null || $data[134]==1 || $data[134]==2)
        {
                $data[134]='No';
        }
        else if($data[135]==3 || $data[135]==4 || $data[135]==5 ){
                $data[135] = 'Yes';
        }

        //entertainment spending
        if($data[136]==null || $data[136]==1 || $data[136]==2)
        {
        $data[136]='No';
        }
        else if($data[136]==3 || $data[136]==4 || $data[136]==5 ){
                $data[136] = 'Yes';
        }
        //spending on looks
        if($data[137]==null ||$data[137]==1 || $data[137]==2)
        {
                $data[137]='No';
        }
        else if($data[137]==3 || $data[137]==4 || $data[137]==5 ){
                $data[137] = 'Yes';
        }
        //spending on gadgets
        if($data[138]==null ||$data[138]==1 || $data[138]==2)
        {
                $data[138]='No';
        }
        else if($data[138]==3 || $data[138]==4 || $data[138]==5 ){
                $data[138] = 'Yes';
        }
        //spending on healthy eating
        if($data[139]==null || $data[139]==1 || $data[139]==2)
        {
                $data[139]='No';
        }
        else if($data[139]==3 || $data[139]==4 || $data[139]==5 ){

                $data[139] = 'Yes';

        }

        fputs($file3, implode($data, ','));

         }
 }

   ?>
```