|  |  |
| --- | --- |
|  | Final Project |
|  |  |
|  | Christina Ling  IST 565 Data Mining  9/25/18 |

# Introduction

Nutrition and health are a concern and issue for many people. This issue affects everyone and does not discriminate against age, gender, race, sexuality, etc. This has been an issue for many years and continues to be an issue in the future. For some people nutrition and health can be a matter of life or death for them. For other it may be important to them in order to feel good about themselves. Many people are learning more about the importance of nutrition and health. When an individual needs to be more aware about their health, changes are required. One challenge to changing is finding the motivation.

One motivation can be feeling healthy. Often mood drives and encourages actions. There is the physicality of being healthy and there is the mental part. There is already a ton of research on the physicality; however very little can be found of the mental. It is important to identify what drive this healthy feeling. By figuring out what increase this feeling, people can change their habit to mentally feel healthier.

The first task to making someone feel healthier is first identifying what they can do. However, determining what impact emotions is not an easy task. It is difficult to determine what actions affect people. Since everyone is unique, one activity may be more effective on one individual than on another. While one technique may work on one person, it may not work at all on another. As a result, it is difficult to determine what impacts feeling healthy the most.

Feeling healthy is important to the following stakeholders:

* People who are dieting or are health advocates
* Nutritionist and Dietician
* Physical Trainers
* Psychologist

This analysis will determine the impact of feeling healthy. What impact feeling healthy: exercise, upbringing, or eating habits? Kinesiologist claim that exercise help releases chemicals in the brain that help people feel better. While psychologist may bring up nature versus nurture. Does people’s upbringing factor on feeling healthy? Parent teaches children many important habits and lesson. This may play a factor on a young adult’s habits moving forward. Lastly do people’s eating habits dictate how healthy they feel? Eating habits can be described as what is consumed to taking vitamins supplements to counting calories. This analysis will look at these three factors to determine what affects feeling healthy as a general statement instead of the individual level.

# Analysis

Data & Preprocess

The dataset is collected from the Kaggle’s Food Choices. The dataset is a compile collect of responses to a survey. The survey was given to 125 students at Mercyhurst University. The survey contains 54 questions ranging from their upbringing to cuisine choices to calories knowledge. Some questions askes the participant what they association with a given word – for example, what do they association with the word fries (McDonald’s fries or home fires). While other questions askes how likely they are to eat a certain type of cuisine.

The data represented with dummy variables – transformation was conducted to change it back into the answer. This allowed for two datasets to be created – one with the written answer and another with the dummy response.

While looking at the data, there were a few missing data, for some questions one or two has missing data – while others had no missing data. To fill the missing data, the median was used to replace the nulls.

# look at responses

unique(food.choices$cook)

#how many NaN?

length(which(is.na(food.choices$cook))) #3

# replace NaN with median

# get median

replace.number <- food.choices[-which(is.na(food.choices$cook)),

which(colnames(food.choices) == "cook")]

replace.number <- median(replace.number)

# replace NaN with median

food.choices[which(is.na(food.choices$cook)),

which(colnames(food.choices) == "cook")] = replace.number

# check work

unique(food.choices$cook

The data was subdivided into different categories. Each category was used for each analysis to determine which influences healthy feeling. The categories are exercise, upbringing, food intake, food association, health, and status. Table 1. lists the categories and which questions/columns it contained. For each sub-dataset, two datasets were created, one for discrete and one for numbers only. Another dataset was created for number only and another for discrete variables only.

Table 1. Sub-dataset and their content

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Exercise | Upbringing | Food Intake | Food Association | Health | Status |
| * Exercise * Healthy Feeling * Sports * Weight * Self-Perception of Weight | * Father’s Education * Healthy Feeling * Income * Mother’s Education * Parent Cook | * Comfort Food Reasons * Cook * Cuisine * Weekly Eating Out * Ethnic Food * Favorite Cuisine * Favorite Food * Daily Fruit Intake * Greek Food * Healthy Feeling * Indian Food * Italian Food * Pay for Meal * Persian Food * Thai Food * Daily Veggie Intake | * Breakfast * Coffee * Drink * Fries * Soup * Healthy Feeling | * Calories in Chicken * Count Daily Calories * Current Diet * Eating Changes * Detail Eating Changes * Healthy Feeling * Ideal Diet * Nutritional Check * Calories in Tortilla * Calories in Turkey Sandwich * Vitamins Supplements * Calories in Waffle | * GPA * Gender * Employment * Grade Level * Healthy Feeling * Life is Rewarding Feeling * Marital Status * Living On or Off Campus |

While the predicted variable, healthy feeling, was divided into three categories. Everyone has a different scale of feeling (on a scale of one to ten, how healthy do you feel). Therefore, instead of predicted the number, this analysis will round the estimate the feeling. The three categories will be called “unhealthy”, “average”, and “healthy”.

hf.discrete <- cut(food.choices$healthy\_feeling,

breaks=c(-Inf, 4.1, 7.1, Inf),

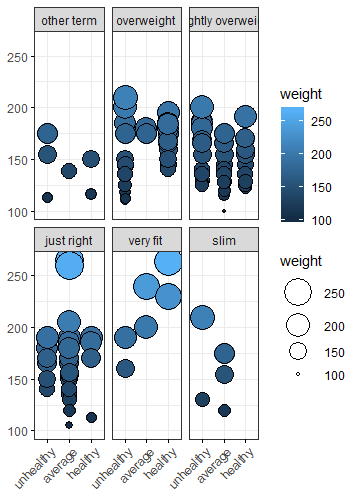
labels=c("unhealthy", "average", "healthy"))

Exploration and Visual

Table 2. displays the summary of the predicted variable, feeling healthy. The distribution is not uniform, it is heavily towards the unhealthy and average side.

Table 2. Distribution of Feeling Healthy

|  |  |
| --- | --- |
| Feeling Healthy Category | Count |
| Unhealthy | 48 |
| Average | 43 |
| Healthy | 34 |

Figure 1. Ballon plots

Left – Feeling Healthy vs. Gender & GPA; Right – Feeling Healthy vs. Weight and Self-Perception of Weight

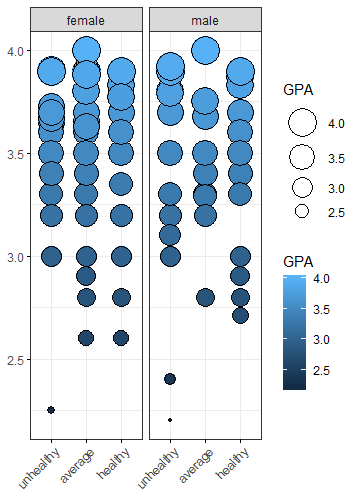


Figure 1. displays some balloon plots to show the relationship between feeling healthy and some variables. From the graph, there is no clear relationship between gender and feeling health. There is no pattern between GPA and feeling healthy. On the second balloon plot there are some interesting results.

#GPA/Gender and healthy feeling

ggballoonplot(status,

x="healthy\_feeling",

y="GPA",

size="GPA",

fill="GPA",

facet.by="gender",

ggtheme = theme\_bw())

#Weight/Self\_perception\_weight and healthy feeling

ggballoonplot(df,

x="healthy\_feeling",

y="weight",

size="weight",

fill="weight",

facet.by="self\_perception\_weight",

ggtheme= theme\_bw())

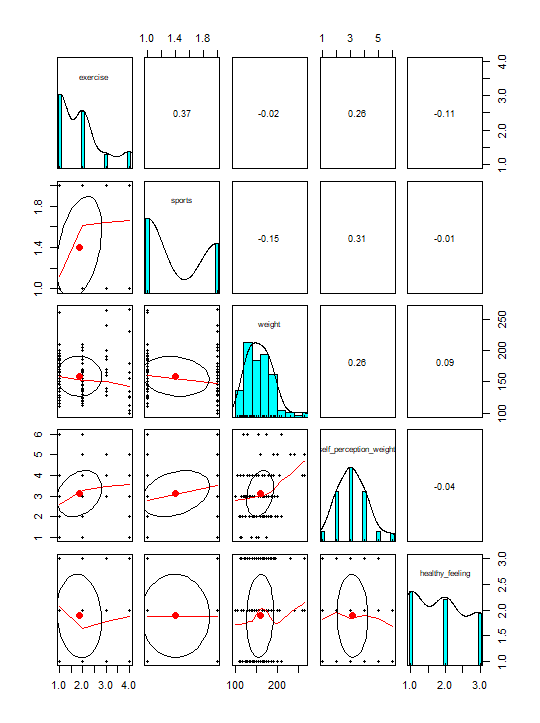
Unsurprisingly there is a relationship shown between self-perception of weight and feeling healthy. However, the relationship is not as expected; people who feel they are overweight or slightly overweight are more likely to feel healthy. While there are no cases of people who see themselves as fit that feel healthy. There is no clear relationship between weight and self-perception of weight or feeling healthy. The dataset did not contain height data, so it is difficult to determine their body mass index (BMI).

To explore this relationship further, Table 3. shows the self-perception of weight and how often those people exercise. The people who viewed themselves as overweight or slightly overweight are more likely to exercise. This could lead to a relationship between feeling healthy and exercise. Further exploration is required, but this is an excellent beginning.

Table 3. Self-Perception of Weight and Weekly Exercise Relationship.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | everyday | 2-3 times | 1 time | never |
| Other term | 3 | 2 | 0 | 1 |
| Overweight | 25 | 5 | 0 | 1 |
| Slightly overweight | 17 | 19 | 5 | 5 |
| Just right | 10 | 14 | 2 | 5 |
| Very fit | 0 | 1 | 4 | 1 |
| Slim | 2 | 3 | 0 | 0 |

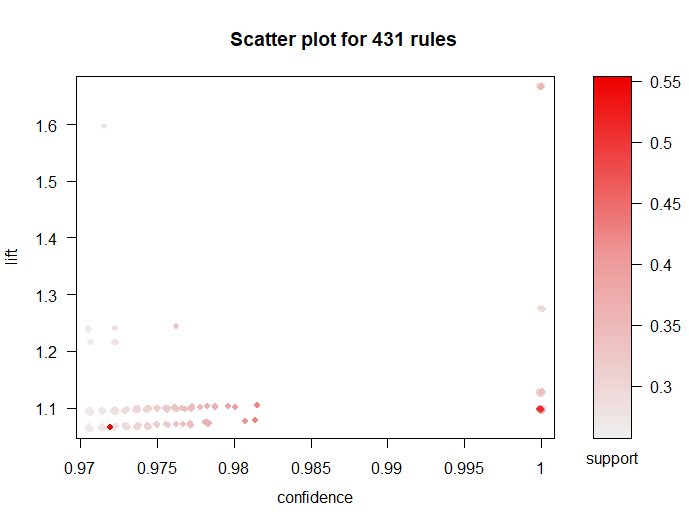
Figure 2. shows the relationships and distribution of exercise. It shows there are no strong correlations between the variables. Additionally, the weight distribution is slightly skewed towards the right.

Figure 2. Distribution and relationships in Exercise

pairs.panels(exercise.num)

Association Rule Mining

Association Rule Mining was used to look for pattern on what lead to a healthy, average or unhealthy feeling. For this method, many support and confidence levels were used. Originally tested was support = 0.1 and confidence = 0.7, this resulted in 997,031 rules. Since so many rules was created, support and confidence levels need to be increased. Eventually, support was increased to 0.25 and confidence was set to 0.97 – this resulted in 431 rules.

Figure 3. All rules created from Association Rules Mining

rule1 <- apriori(fc.discrete,

parameter = list(supp=0.25, conf=0.97))

rule1 <- sort(rule1, by="confidence", decreasing=TRUE)

inspect(rule1[1:20])

plot(rule1, measure=c("confidence", "lift"), shading="support")

However, this did not provide any insight into what creates a healthy feeling. So another set of rules was created with the right hand side set to healthy feeling. The originally settings of support = 0.1 and confidence = 0.7 did not provide any rules, so the parameters need to be lowers. Finally with support = 0.07 and confidence = 0.7, there were 81 rules created. Table 4. displays interesting and insightful rules.

Table 4. Interest Rules for Right Hand Side = Healthy Feeling

#healthy

healthy.rule <- apriori(fc.discrete,

parameter = list(supp=0.07, conf=0.7),

appearance = list(default="lhs", rhs="healthy\_feeling=healthy"),

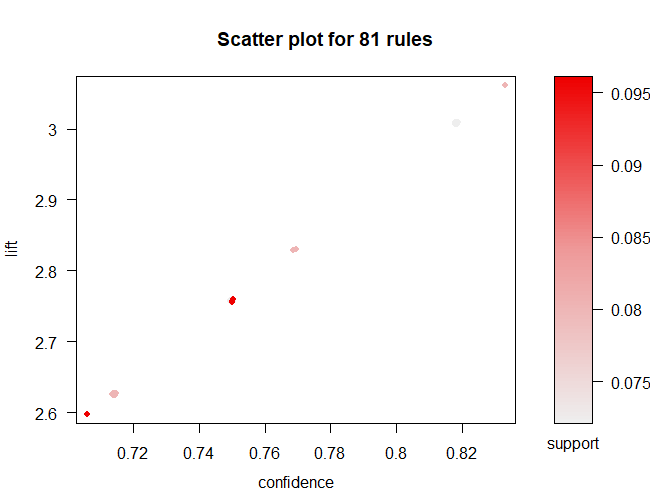
control=list(verbose=FALSE))

healthy.rule <- sort(healthy.rule, by="confidence", decreasing=TRUE)

inspect(healthy.rule[1:20])

plot(healthy.rule, measure=c("confidence", "lift"), shading="support") #81 rules

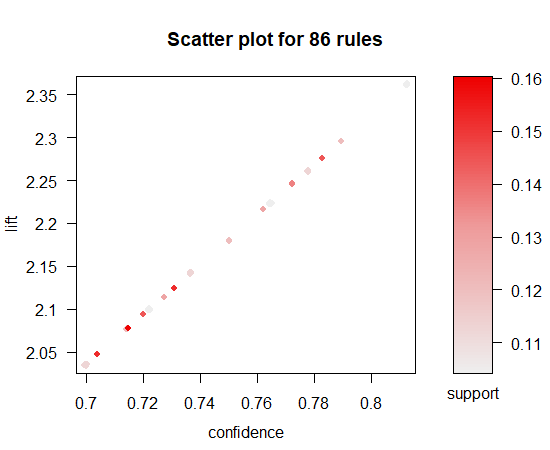
|  |  |  |  |
| --- | --- | --- | --- |
| LHS | Support | Confidence | Lift |
| Gender = male,  breakfast= cereal,  on\_off\_campus = on campus,  parents\_cook = almost everyday,  self\_perception\_weight = overweight | 0.080 | 0.833 | 3.064 |
| cook = not often,  employment = no,  exercise= everyday,  on\_off\_campus = on campus | 0.072 | 0.818 | 3.008 |
| gender = male,  on\_off\_campus = on campus,  parents\_cook = almost everyday,  self\_perception\_weight = overweight,  sports = yes | 0.072 | 0.818 | 3.008 |
| Cook = not often,  Employment= no,  Exercise = everyday,  Fries = McDonald’s fries,  On\_off\_campus = on campus | 0.072 | 0.818 | 3.008 |
| Gender = male,  On\_off\_campus = on campus,  Parents\_cook = almost everyday,  Self\_perception\_weight = overweight,  Sports =yes | 0.072 | 0.818 | 3.008 |

Figure 4. Rules created from Right Hand Side = Healthy Feeling

For average healthy feeling, the original parameters of support = 0.1 and confidence = 0.7 provided 86 rules. Therefore, no adjustments were required. Table 5. shows some interesting rules.

Table 5. Interesting Rules for Right Hand Side = Average Healthy Feeling

|  |  |  |  |
| --- | --- | --- | --- |
| Left Hand Side | Support | Confidence | Lift |
| Breakfast=cereal,  Self\_perception\_weight = just right,  Waffle\_calories=1315 | 0.104 | 0.813 | 2.362 |
| Breakfast=cereal,  Italian\_food = very likely,  Marital\_status=single,  Self\_perception\_weight = just right | 0.104 | 0.813 | 2.362 |
| Breakfast=cereal,  Calories\_day = moderately important,  cuisine=american,  self\_perception\_weight = just right | 0.104 | 0.813 | 2.362 |
| cuisine=american,  marital\_status = single,  self\_perception\_weight = just right | 0.104 | 0.765 | 2.22 |

Figure 5. Rules created from Right Hand Side = Average Feeling

#average

average.rule <- apriori(fc.discrete,

parameter = list(supp=0.1, conf=0.7),

appearance = list(default="lhs", rhs="healthy\_feeling=average"),

control=list(verbose=FALSE))

average.rule <- sort(average.rule, by="confidence", decreasing=TRUE)

inspect(average.rule[1:20])

plot(average.rule, measure=c("confidence", "lift"), shading="support") #86 rules

Similar to the average healthy feeling, the original parameters of support = 0.1 and confidence = 0.7 yielded 23 rules – so no adjustments were required. Table 6. shows some interesting rules created.

#unhealthy

unhealthy.rule <- apriori(fc.discrete,

parameter = list(supp=0.1, conf=0.7),

appearance = list(default="lhs", rhs="healthy\_feeling=unhealthy"),

control=list(verbose=FALSE))

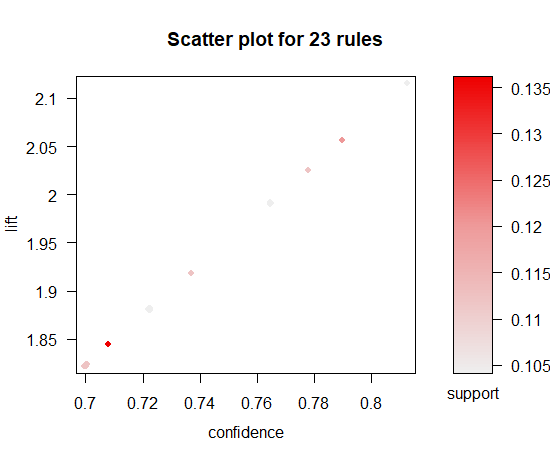
unhealthy.rule <- sort(unhealthy.rule, by="confidence", decreasing=TRUE)

inspect(unhealthy.rule[1:20])

plot(unhealthy.rule, measure=c("confidence", "lift"), shading="support") #23 rules

Table 6. Interesting Rules for Right Hand Side = Unhealthy Healthy Feeling

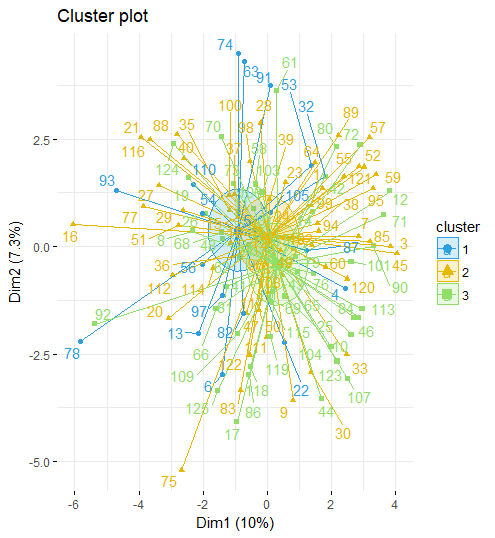
|  |  |  |  |
| --- | --- | --- | --- |
| Left Hand Side | Support | Confidence | Lift |
| cuisine = american,  fries = mcdonald’s fries,  mother\_education = college,  vitamins = yes | 0.104 | 0.823 | 2.116 |
| Fries = mcdonald’s fries,  mother\_education = college,  vitamins = yes | 0.120 | 0.789 | 2.056 |
| Calories\_scone = 420,  Fav\_food = cooked at home,  Fries = mcdonald’s fries,  On\_off\_campus = on campus,  Sports = yes | 0.104 | 0.722 | 1.881 |
| mother\_education = college,  vitamins = yes | 0.112 | 0.700 | 1.823 |

Figure 6. Rules created from Right Hand Side = Unhealthy Feeling

For the sub-dataset, none of them yielded any meaningful results. The only rules created were when support and confidence was low.

Clustering

For clustering, cluster group number was set to three since there are three levels of healthy feeling. Clustering was used to determine which variables created the best clustering. Five types of clustering techniques were used, k-means, EM clustering, HAC (Euclidean Method), HAC (Manhattan Method), and Cosine Similarity. Clustering for all data was used first.

Figure 6. K-means clustering for all data

### using all data

#k means, cluster 3, (low, average, high)

kmeans <- kmeans(food.choices.nolab, 3)

print(kmeans)

#results

table(food.choices.numbonly$healthy\_feeling, kmeans$cluster)

#bad results

#visual

fviz\_cluster(kmeans, data=food.choices.nolab,

palette= c("#2E9FDF", "#E7B800", "#92DE6A"),

ellipse.type= "euclid",

star.plot=TRUE,

repel= TRUE,

ggtheme = theme\_minimal())

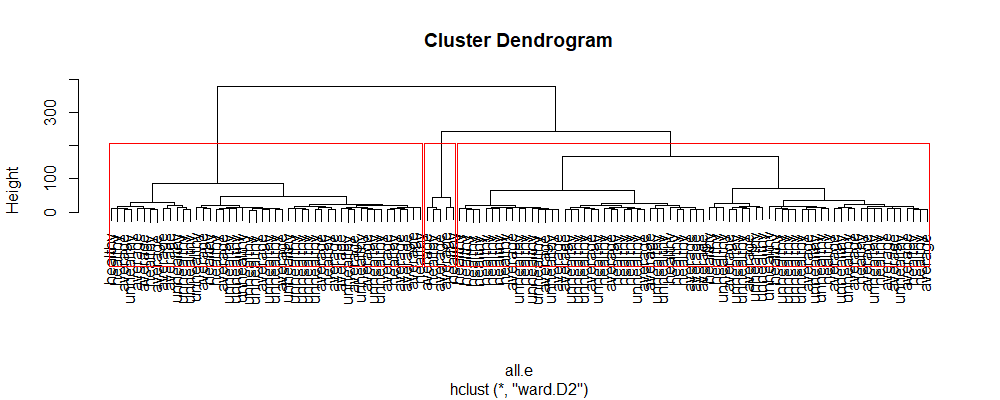
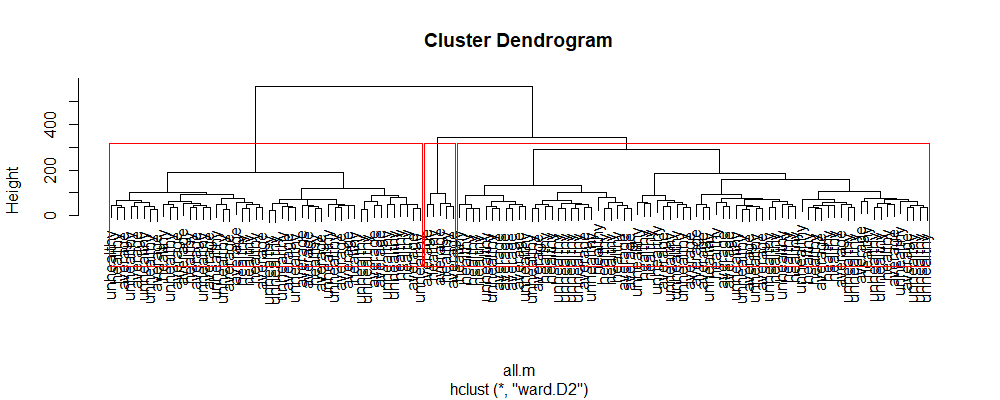


Figure 7. HAC Euclidean Method Clustering

Figure 8. HAC Manhattan Method Clustering

#Euclidean

all.e <- dist(food.choices.nolab, method="euclidean")

fit.all.e <- hclust(all.e, method="ward.D2")

plot(fit.all.e,

labels=food.choices.numbonly$healthy\_feeling)

rect.hclust(fit.all.e, k=3)

#Manhattan

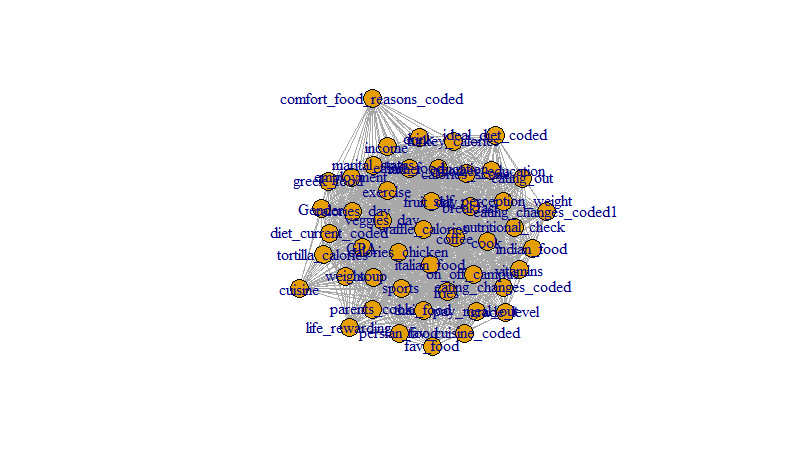
all.m <- dist(food.choices.nolab, method="manhattan")

fit.all.m <- hclust(all.m, method="ward.D2")

plot(fit.all.m,

labels=food.choices.numbonly$healthy\_feeling)

rect.hclust(fit.all.m, k=3)

Figure 9. Cosine Similarity Clustering

# all data

all.matrix <- as.matrix(food.choices.nolab)

cos\_sim.all.matrix <- cosine(all.matrix)

diag(cos\_sim.all.matrix) <- 0

head(cos\_sim.all.matrix)

#prune edges of the tree

edgeLimit <- 0.75

cos\_sim.all.matrix[(cos\_sim.all.matrix < edgeLimit)] <- 0

#make network

cos\_sim.networkall <- graph\_from\_adjacency\_matrix(cos\_sim.all.matrix,

mode="undirected",

weight=T)

plot(cos\_sim.networkall)

Four clustering models was applied to the exercise, upbringing, food intake, food association, health and status sub-dataset. The four models created are the Kmean, EM Clustering, HAC – Euclidean method and HAC – Manhattan method. This allows for comparison to see which effects the healthy feeling the most.

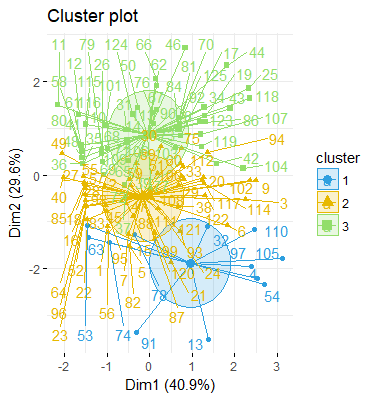
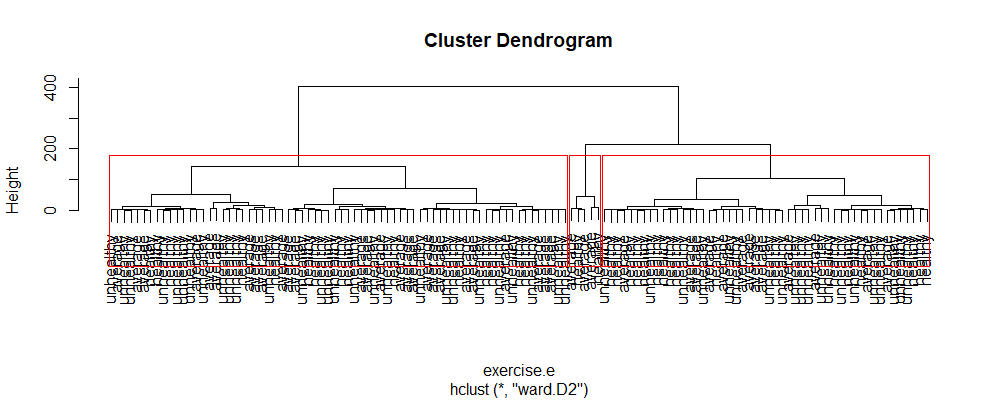
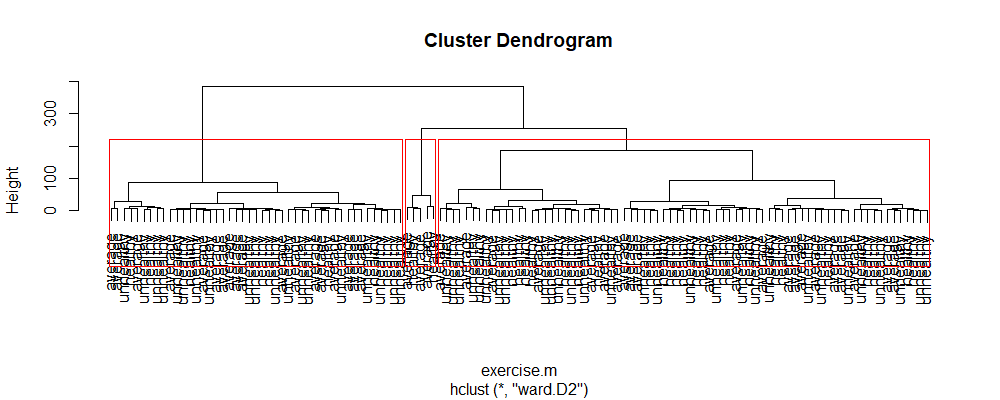
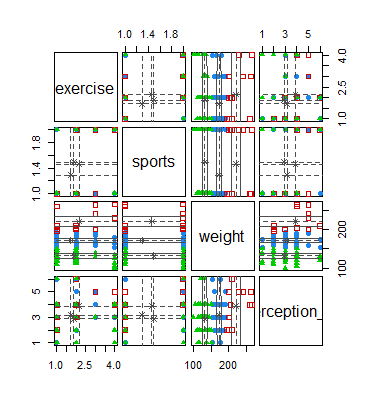
 Figure 10. K-means Clustering for Exercise Sub-Dataset

Figure 13. HAC – Manhattan Clustering for Exercise Sub-Dataset

Figure 12. HAC – Euclidean Clustering for Exercise Sub-Dataset

Figure 11. EM Clustering for Exercise Sub-Dataset

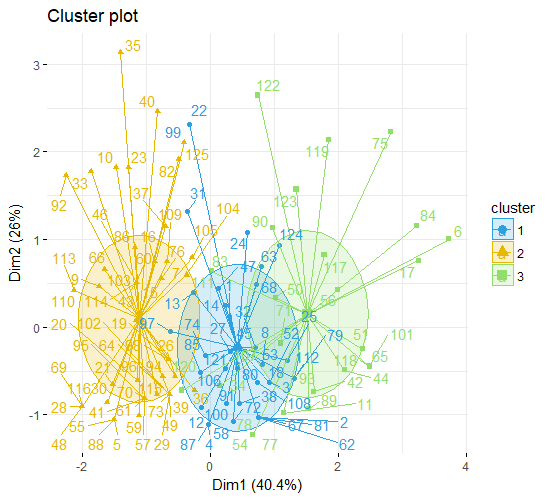
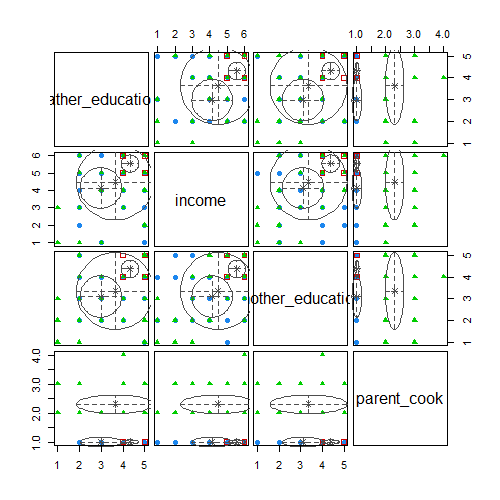
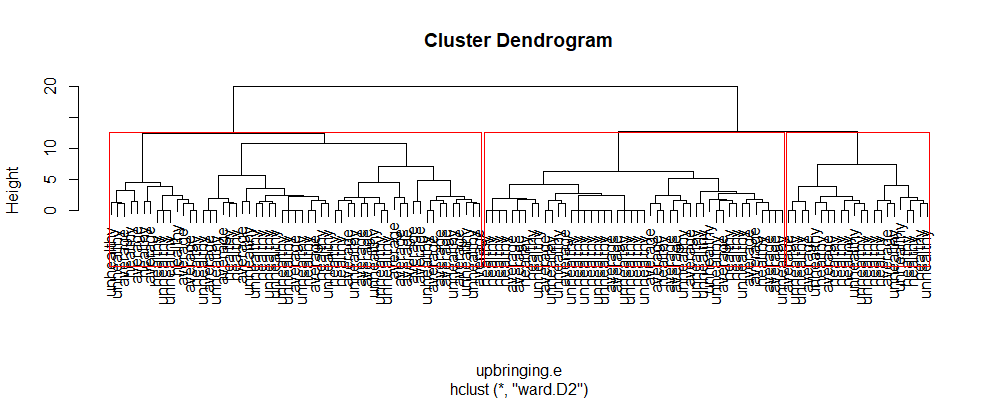
Figure 14. K-means Clustering for Upbringing Sub-Dataset

Figure 16. HAC – Euclidean Clustering for Upbringing Sub-Dataset

Figure 15. EM Clustering for Upbringing Sub-Dataset

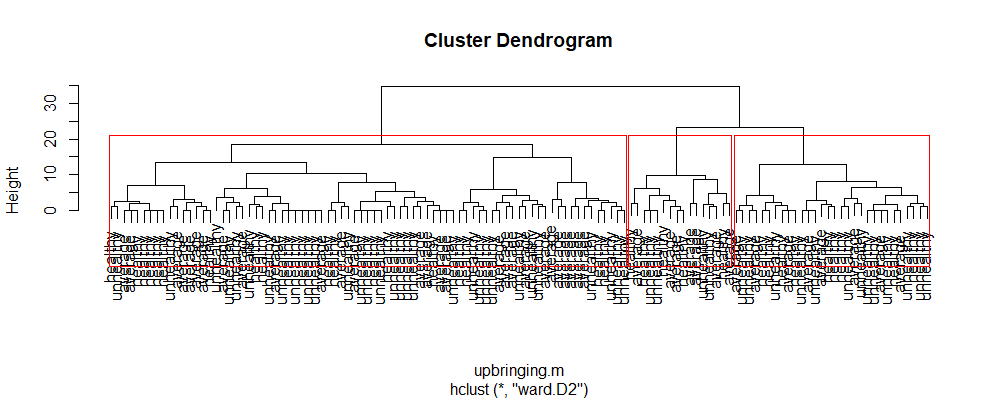


Figure 17. HAC – Manhattan Clustering for Upbringing Sub-Dataset

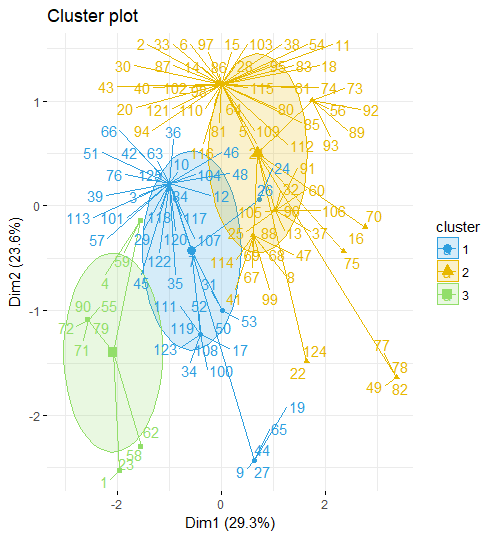
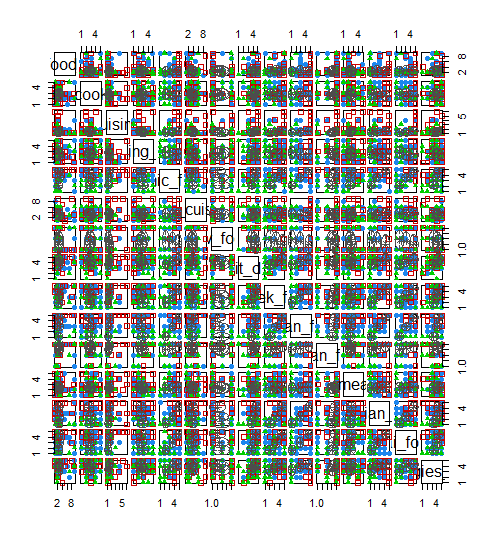
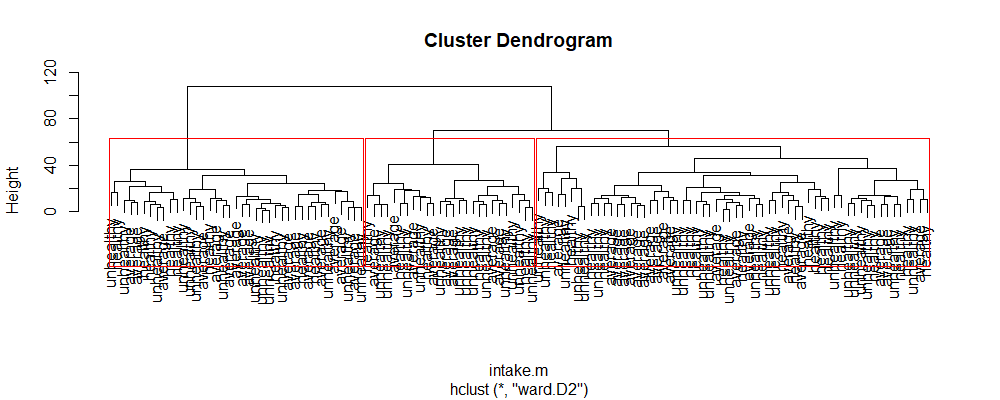
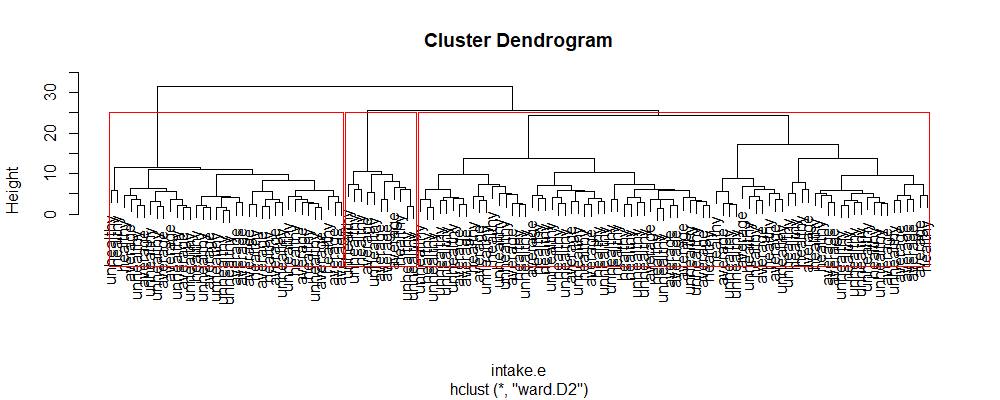
Figure 18. K-means Clustering for Food Intake Sub-Dataset

Figure 20. HAC – Euclidean Clustering for Food Intake Sub-Dataset

Figure 19. EM Clustering for Food Intake Sub-Dataset

Figure 21. HAC – Manhattan Clustering for Food Intake Sub-Dataset

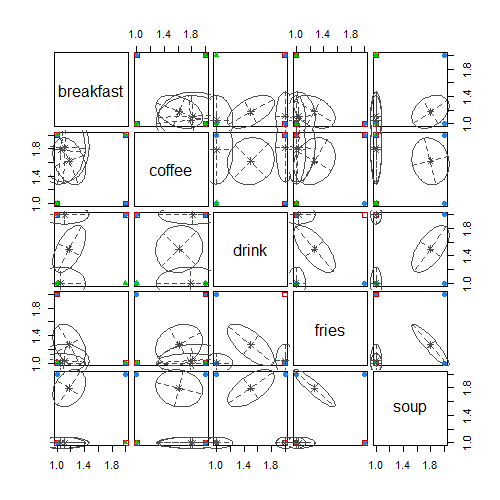
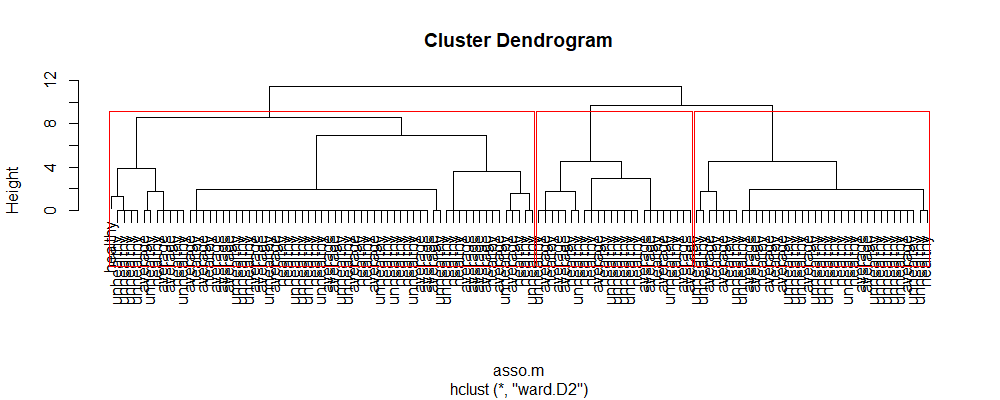
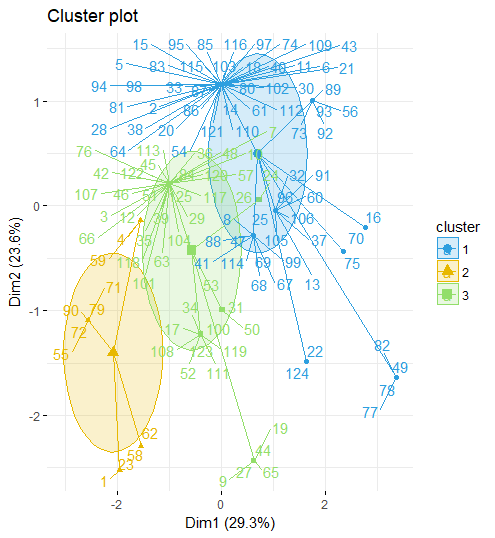
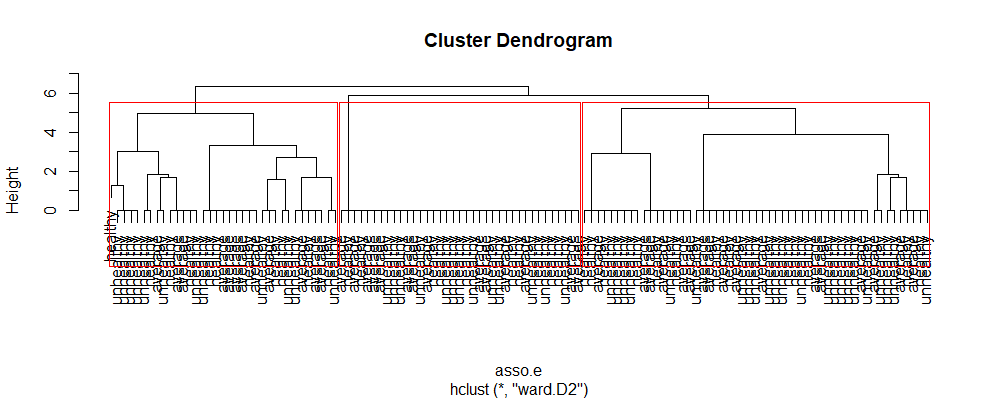
 Figure 22. K-means Clustering for Food Association Sub-Dataset

Figure 24. HAC – Manhattan Clustering for Food Association Sub-Dataset

Figure 23. HAC – Euclidean Clustering for Food Association Sub-Dataset

Figure 22. EM Clustering for Food Association Sub-Dataset

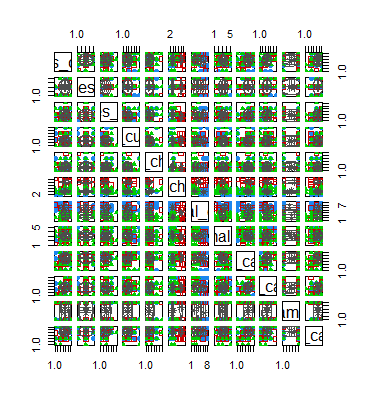
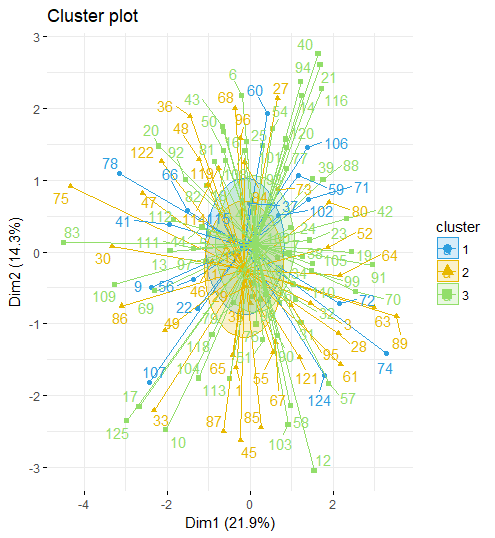
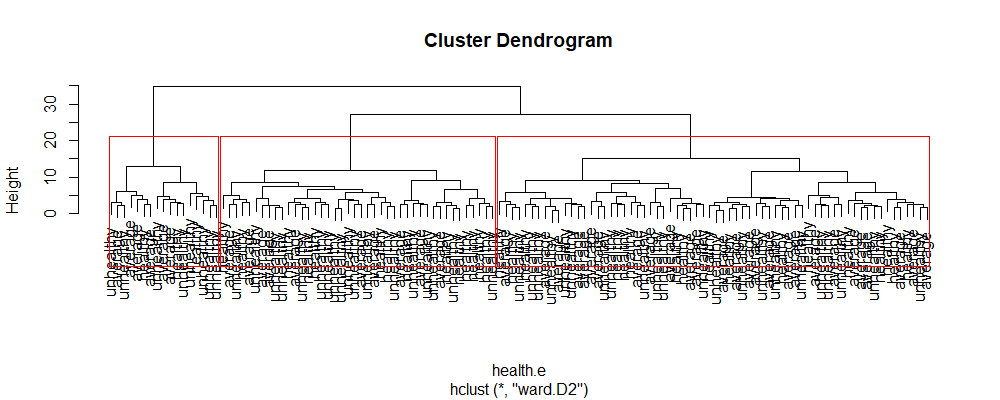
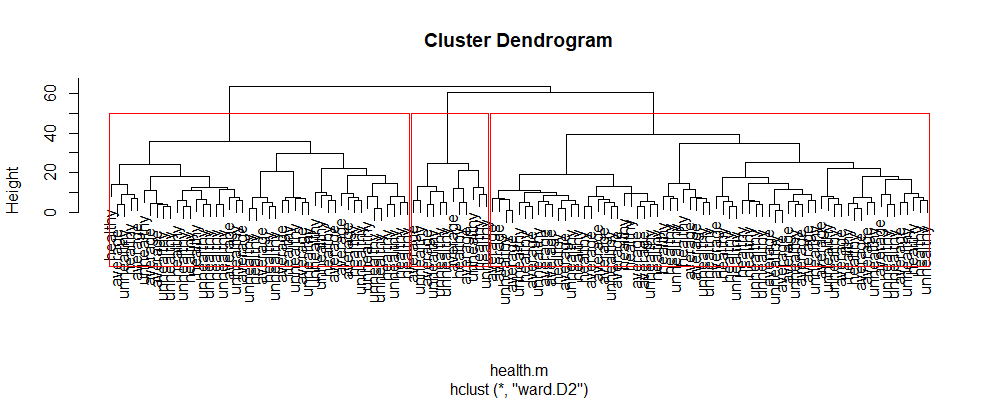
 Figure 25. K-means Clustering for Health Sub-Datasub

Figure 27. HAC – Euclidean Clustering for Health Sub-Dataset

Figure 28. HAC – Manhattan Clustering for Health Sub-Dataset

Figure 26. EM Clustering for Health Sub-Dataset

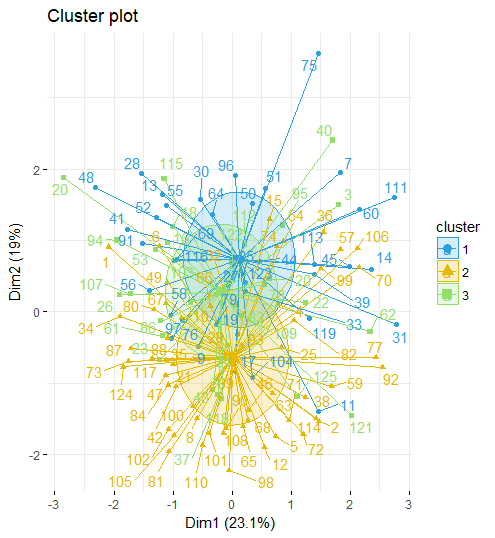
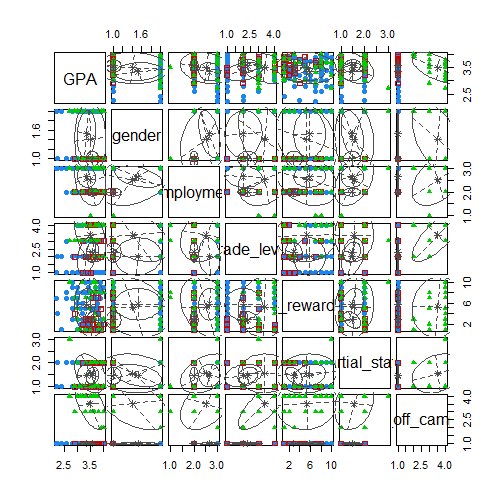
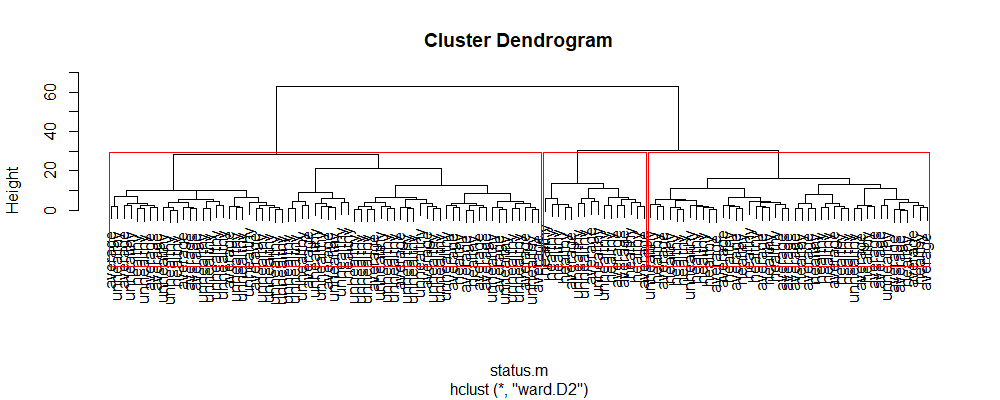
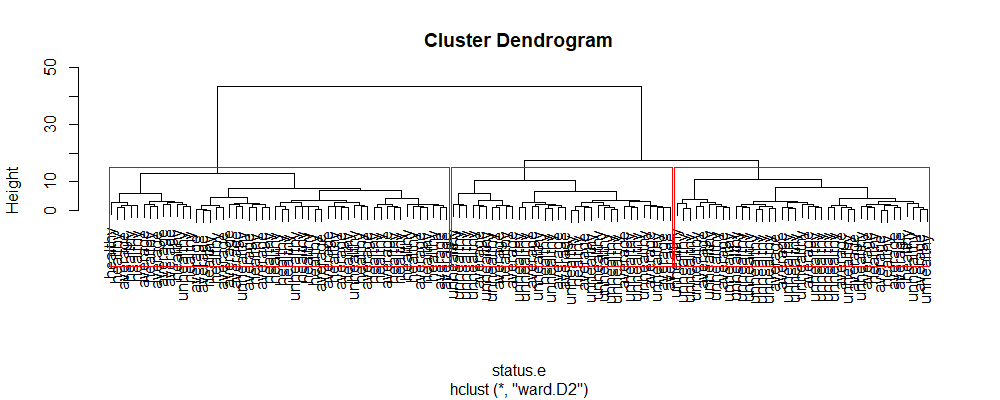
Figure 29. K-means for Status Sub-Dataset

Figure 30. EM Clustering for Status Sub-Dataset

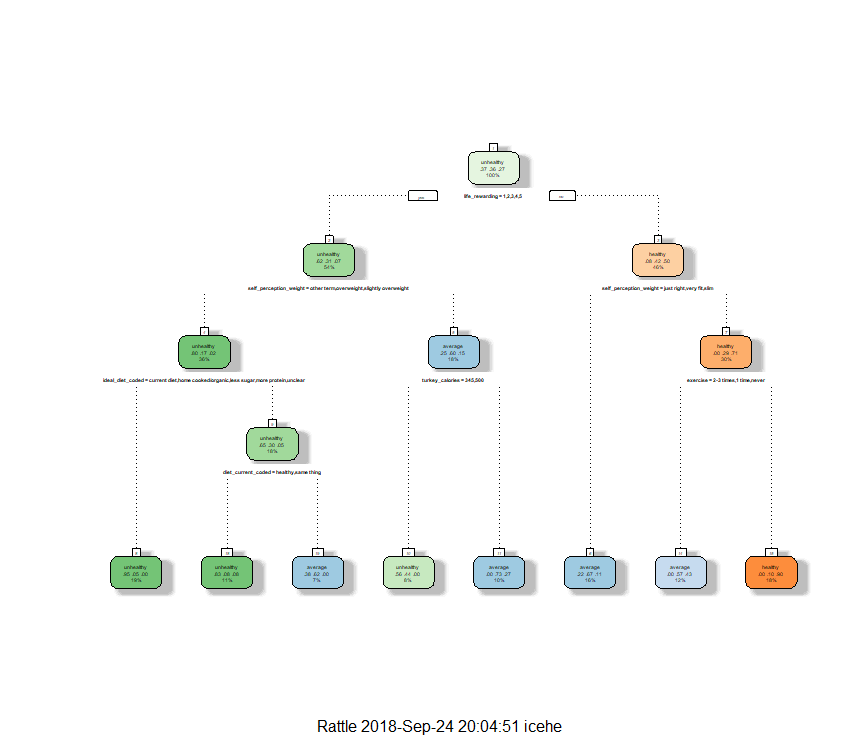
Figure 31. HAC – Euclidean Clustering for Status Sub-Dataset

Figure 32. HAC – Manhattan Clustering for Status Sub-Dataset

Decision tree

The decision tree model was used to determine what factors leads to a healthy feeling. It is also used to help predict how healthy an individual feel. To create the decision tree, the data was split into two set, a train and test dataset. Since the dataset is small, ten percent of the data was randomly selected to create the test dataset; leaving the remaining ninety percent for the train. This allows the most data to be used for training the model. For the decision tree model, all variables were converted into factors.

To test the Decision Tree Model, some adjustments need to be made to the test dataset. The healthy feeling variables needs to be removed and stored elsewhere. Afterwards, the model could use the test dataset to predict the healthy feeling. The predicted healthy feeling and the actual healthy feeling of the individual was then compared to each other.

Figure 33. Decision Tree using all Data.

Separate decision trees were created for each sub-dataset. The same process of creating a train and test dataset was applied. For the test dataset, the same randomly selected ten percent was used; leaving the remaining ninety percent for the train dataset.

# check response variables str for factor

str(fc.discrete$healthy\_feeling)

## randomly selection 10% for test data

n <- round(nrow(fc.discrete)/10)

s <- sample(1:nrow(fc.discrete), n)

#create test and training datasets

test <- fc.discrete[s,]

train <- fc.discrete[-s,]

# use all data

tree.fit\_all <- rpart(train$healthy\_feeling ~.,

data=train,

method="class")

summary(tree.fit\_all)

# predict test data

# remove healthy\_feeling from test

test\_label <- test[,which(colnames(test) == "healthy\_feeling")]

test <- test[,-(which(colnames(test) == "healthy\_feeling"))]

predict\_all <- predict(tree.fit\_all, test, type="class")

table(predict\_all, test\_label)

# visualize the tree

fancyRpartPlot(tree.fit\_all)

Figure 34. Decision Tree Using the Exercise (Left) and Upbringing (Right) Sub-Dataset

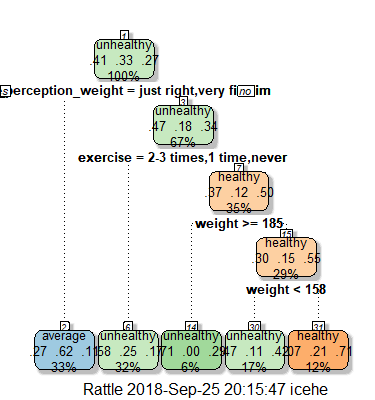
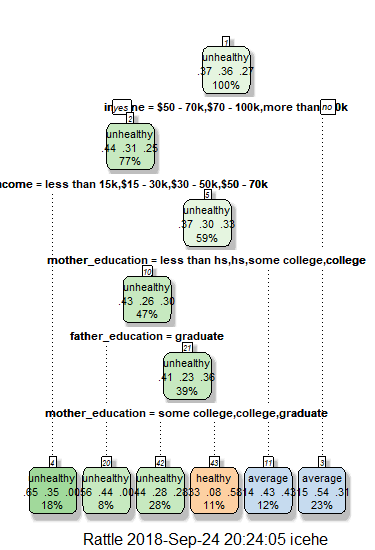
 

Figure 35. Decision Tree Using the Food Intake (Left) and Food Association (Right) Sub-Dataset

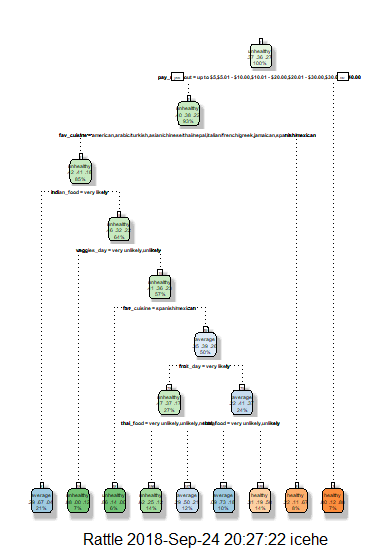
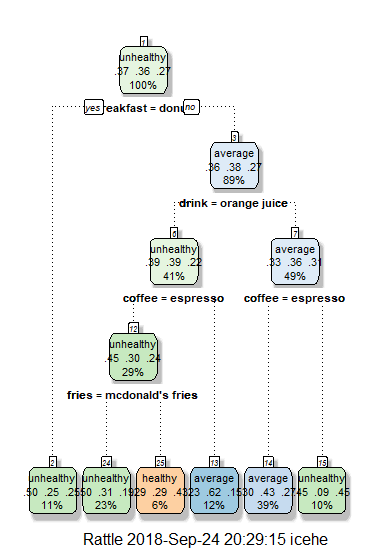
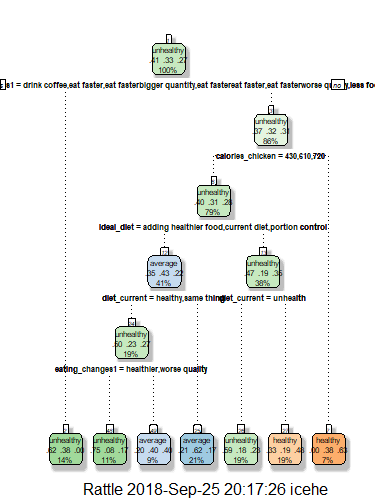
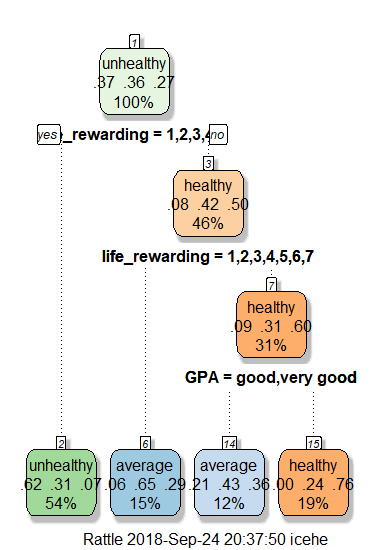


Figure 36. Decision Tree Using the Health (Left) and Status (Right) Sub-Dataset

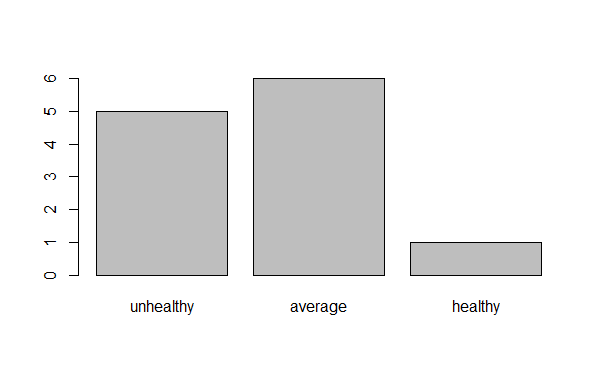
 

Naïve Bayes

For Naïve Bayes, the dataset containing the numbers only, or the dummy variables, was used. The dataset was split into two, a train and test dataset. Due to the small number of rows, ten percent was randomly selected for the test dataset; leaving the remaining ninety percent for the train dataset.

To use Naïve Bayes to predict the results of the test dataset, the healthy feeling had to be removed and stored separately. Afterwards removing the healthy feeling, the model used the test dataset to predict the healthy feeling. The model’s results and actual results were compared to each other.

Figure 37. Naïve Bayes Results Using all Data



## randomly selection 10% for test data

n <- round(nrow(food.choices.numbonly)/10)

s <- sample(1:nrow(food.choices.numbonly), n)

#create test and training datasets

test <- food.choices.numbonly[s,]

train <- food.choices.numbonly[-s,]

# check response variables str for factor

str(food.choices.numbonly$healthy\_feeling)

#set up model

nb\_all <- naiveBayes(healthy\_feeling ~.,

data=train,

na.action= na.pass)

#predictions

#remove label from test

test\_label <- test[,(which(colnames(test) == "healthy\_feeling"))]

test <- test[,-(which(colnames(test) == "healthy\_feeling"))]

nb\_predict\_all <- predict(nb\_all, test)

#look at results

print(nb\_predict\_all)

table(nb\_predict\_all, test\_label)

#visual

plot(nb\_predict\_all)

The same method and model were applied to the six sub-datasets. Ten percent of the data was randomly selected for the test dataset, leaving the remaining ninety percent for the train. This analysis allows for comparison to see the effect of each factor on feeling healthy.

Figure 38. Naïve Bayes Results Using Exercise (Left) and Upbringing (Right) Sub-Datasets

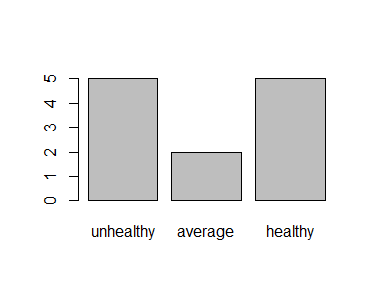
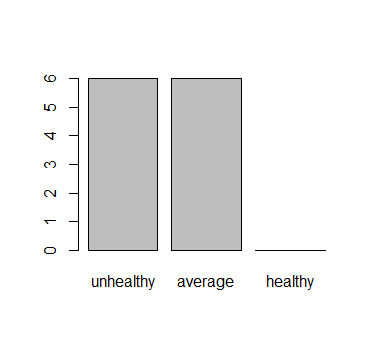
 

Figure 39. Naïve Bayes Results Using Food Intake (Left) and Food Association (Right) Sub-Datasets

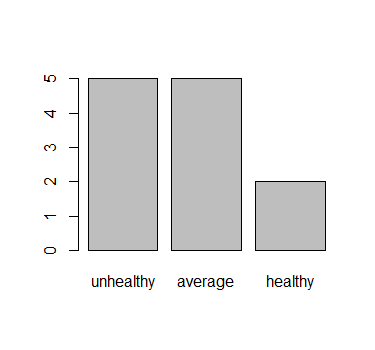
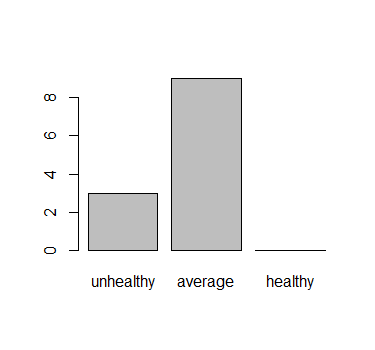
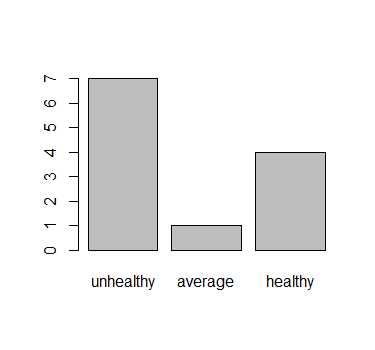
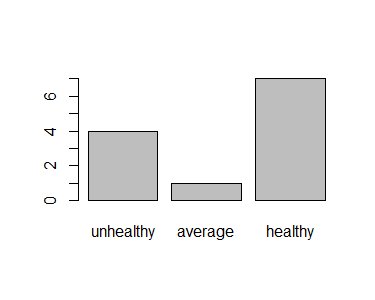
 

Figure 40. Naïve Bayes Results Using Health (Left) and Status (Right) Sub-Datasets



Support Vector Machine

For Support Vector Machine (SVM), three models were created: polynomial, linear and radial. The SVM requires a train to building the model and a test dataset to predict. Ten percent was randomly selected for the test dataset; leaving the remaining ninety percent for the training data. Ten percent was select because of the small number of rows inside the dataset, this allows fort the SVM to learn as much as possible from the train dataset.

In order for the SVM to predict, the healthy feeling had to be removed from the test dataset. The healthy feeling was stored separately from the test dataset. After using the model to predict, the results were compared to the actual healthy feeling of the individual.

# check structure of predicted variable

str(food.choices.numbonly$healthy\_feeling)

# change to ordered factor

food.choices.numbonly$healthy\_feeling <- as.factor(food.choices.numbonly$healthy\_feeling)

food.choices.numbonly$healthy\_feeling <- ordered(food.choices.numbonly$healthy\_feeling)

# randomly selection 10% for test data

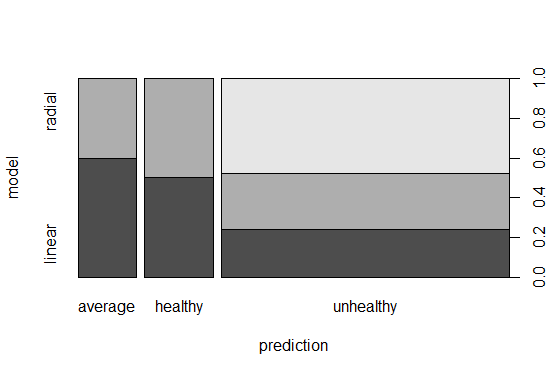
n <- round(nrow(food.choices.numbonly)/10)

s <- sample(1:nrow(food.choices.numbonly), n)

# create test and training dataset

test <- food.choices.numbonly[s,]

train <- food.choices.numbonly[-s,]

Figure 41. Results from SVM Using all Data

For each sub-dataset, three SVMs were created: polynomial, linear and radial. The same preprocessing method was applied. Ten percent of the data was randomly selected for the test dataset, leaving the remaining ninety percent for the train. This analysis allows for comparison to see the effect of each factor on feeling healthy.

# polynomial

svm\_fit\_all\_p <- svm(healthy\_feeling~., data=train, kernel="polynomial", cost=0.1,

scale=F)

print(svm\_fit\_all\_p)

#predictions

#remove label from test

test\_label <- test[,(which(colnames(test) == "healthy\_feeling"))]

test <- test[,-(which(colnames(test) == "healthy\_feeling"))]

pred\_all\_p <- predict(svm\_fit\_all\_p, test, type="class")

table(pred\_all\_p, test\_label)

## Linear

svm\_fit\_all\_l <- svm(healthy\_feeling~., data=train, kernel="linear", cost=0.1,

scale=F)

print(svm\_fit\_all\_l)

#prediction

pred\_all\_l <- predict(svm\_fit\_all\_l, test, type="class")

table(pred\_all\_l, test\_label)

## radial

svm\_fit\_all\_r <- svm(healthy\_feeling~., data=train, kernel="radial", cost=0.1,

scale=F)

print(svm\_fit\_all\_r)

#prediction

pred\_all\_r <- predict(svm\_fit\_all\_r, test, type="class")

table(pred\_all\_r, test\_label)

## visualize results

results <- c(pred\_all\_p, pred\_all\_l, pred\_all\_r)

results <- gsub(1, "unhealthy", results)

results <- gsub(2, "average", results)

results <- gsub(3, "healthy", results)

model <- c(replicate(12, "polynomial", simplify=T),

replicate(12, "linear", simplify = T),

replicate(12, "radial", simplify = T))

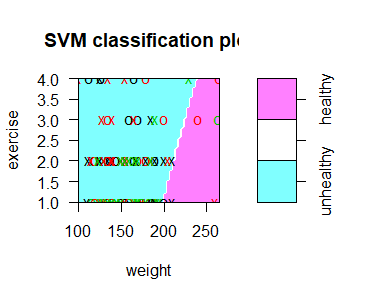
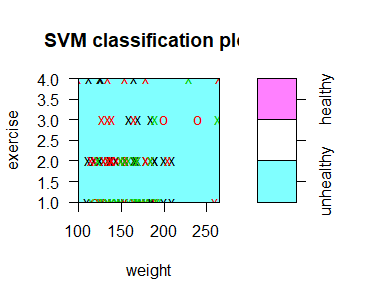
results <- data.frame(prediction=results,

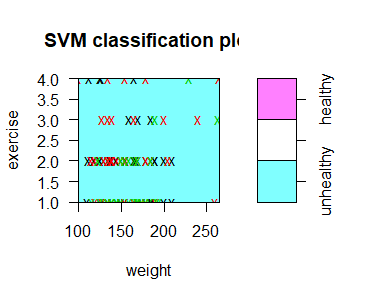
model)

plot(results)

Figure 42. SVM Model for Exercise Sub-Dataset;

Left – Polynomial; Right – Linear; Bottom - Radial



#visualize

plot(svm\_fit\_exercise\_l,

data=train,

exercise~weight)

plot(svm\_fit\_exercise\_l,

data=train,

exercise~weight)

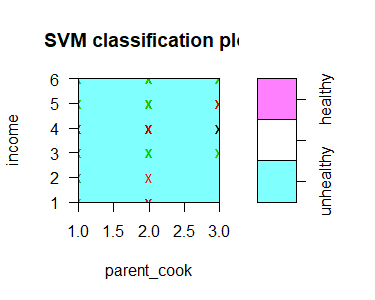
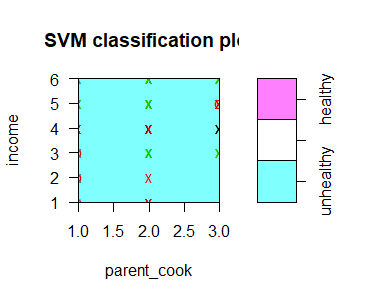
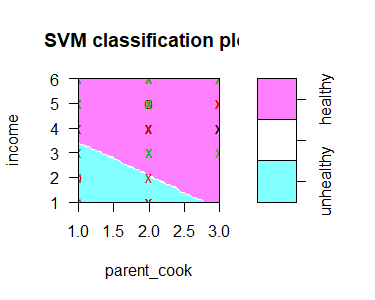
plot(svm\_fit\_exercise\_r,

data=train,

exercise~weight)

Figure 43. SVM Model for Upbringing Sub-dataset;

Left – Polynomial; Right – Linear; Bottom – Radial



#visualize

plot(svm\_fit\_upbringing\_l,

data=train,

income~parent\_cook)

plot(svm\_fit\_uprbinging\_l,

data=train,

income~parent\_cook)

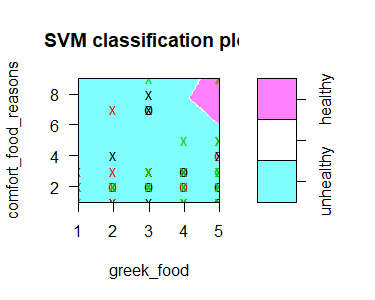
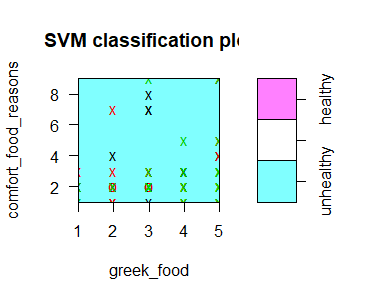
plot(svm\_fit\_upbringing\_r,

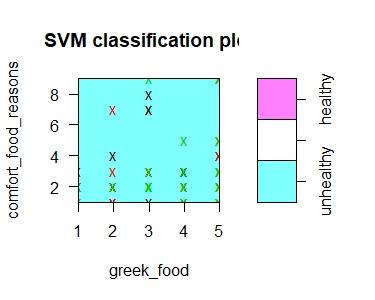
data=train,

income~parent\_cook)

Figure 44. SVM Model for Food Intake Sub-dataset;

Left – Polynomial; Right – Linear; Bottom – Radial



#visualize

plot(svm\_fit\_intake\_l,

data=train,

comfort\_food\_reasons~greek\_food)

plot(svm\_fit\_intake\_l,

data=train,

comfort\_food\_reasons~greek\_food)

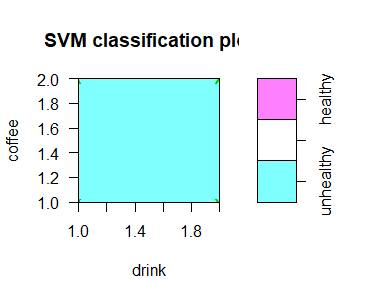
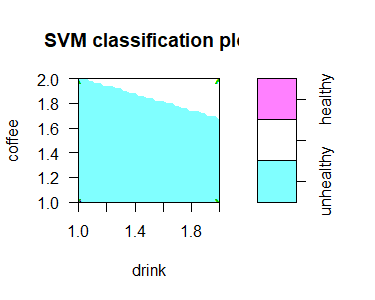
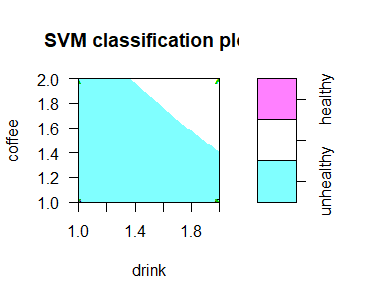
plot(svm\_fit\_intake\_r,

data=train,

comfort\_food\_reasons~greek\_food)

Figure 45. SVM Model for Food Association Sub-Dataset;

Left – Polynomial; Right – Linear; Bottom – Radial



#visualize

plot(svm\_fit\_asso\_l,

data=train,

coffee~drink)

plot(svm\_fit\_asso\_l,

data=train,

coffee~drink)

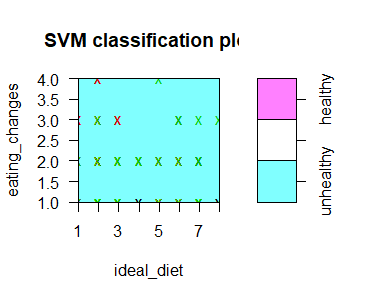
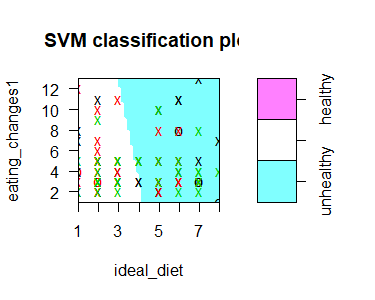
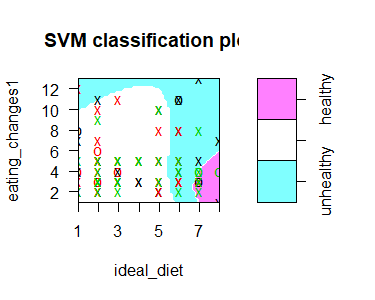
plot(svm\_fit\_asso\_r,

data=train,

coffee~drink)

Figure 46. SVM Model for Health Sub-Dataset;

Left – Polynomial; Right – Linear; Bottom – Radial



#visualize

plot(svm\_fit\_health\_l,

data=train,

eating\_changes1~ideal\_diet)

plot(svm\_fit\_health\_l,

data=train,

eating\_changes1~ideal\_diet)

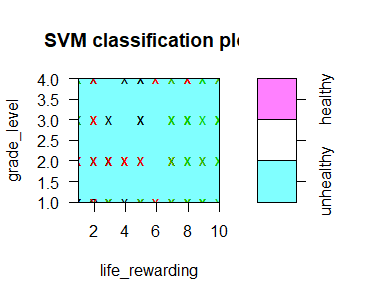
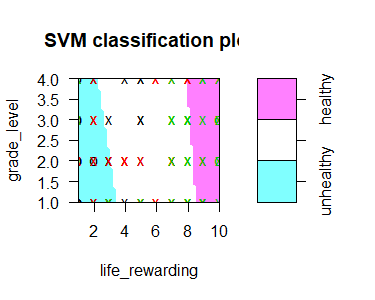
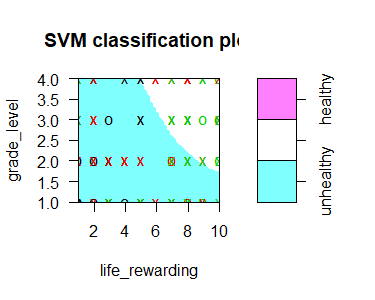
plot(svm\_fit\_health\_r,

data=train,

eating\_changes1~ideal\_diet)

Figure 47. SVM Model for Status Sub-Dataset;

Left – Polynomial; Right – Linear; Bottom – Radial



#visualize

plot(svm\_fit\_health\_l,

data=train,

eating\_changes1~ideal\_diet)

plot(svm\_fit\_health\_l,

data=train,

eating\_changes1~ideal\_diet)

plot(svm\_fit\_health\_r,

data=train,

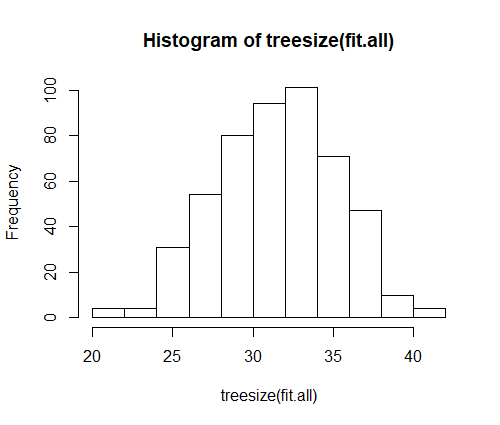
eating\_changes1~ideal\_diet)

Random Forest

Like the other models, some preprocessing was required. The dataset was split into two, one for training the model and the other for testing the model. The test dataset was created by randomly selecting ten percent of the data, leaving the remaining ninety percent for the train dataset. Since this is a small dataset, ten percent was choice to allow the train dataset to be as big as possible. This allows for the model to learn as much as possible.

For the test dataset, the healthy feeling, or the prediction variable, was removed and stored separately. After allowing the Random Forest to predict the healthy feeling, it was compared to the actual healthy feeling for accuracy.

Figure 48. Histogram of tree sizes from Random Forest Using all Data.



The same preprocess and method were applied to the six sub-datasets. The same random ten percent was selected for the test dataset. This allows Random Forest to determine which collection of variables has the biggest impact on feeling healthy.

# randomly selection 10% for test data

n <- round(nrow(food.choices.numbonly)/10)

s <- sample(1:nrow(food.choices.numbonly), n)

# create test and training dataset

test <- food.choices.numbonly[s,]

train <- food.choices.numbonly[-s,]

#set up RF

fit.all <- randomForest(healthy\_feeling~.,

data=train)

print(fit.all)

#predict

test\_label <- test[,(which(colnames(test) == "healthy\_feeling"))]

test <- test[,-(which(colnames(test) == "healthy\_feeling"))]

pred.all <- predict(fit.all, test)

table(pred.all, test\_label)

#visual treesize

hist(treesize(fit.all))

varImpPlot(fit.all)

Figure 49. Histogram of tree sizes from Random Forest Using Exercise Sub-Dataset

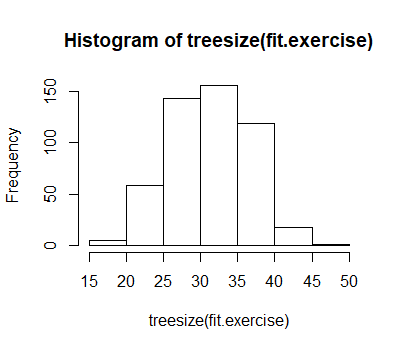


Figure 50. Histogram of tree sizes from Random Forest Using Upbringing Sub-Dataset

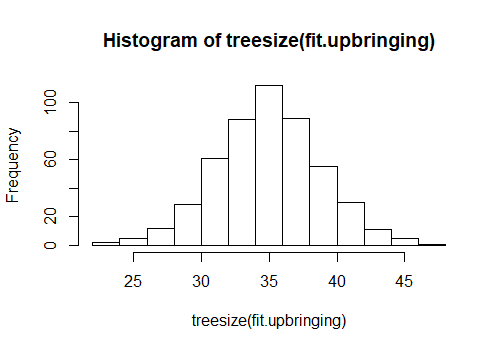


Figure 51. Histogram of tree sizes from Random Forest Using Food Intake Sub-Dataset

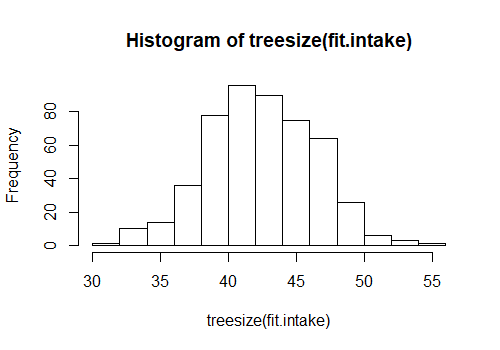


Figure 52. Histogram of tree sizes from Random Forest Using Food Association Sub-Dataset

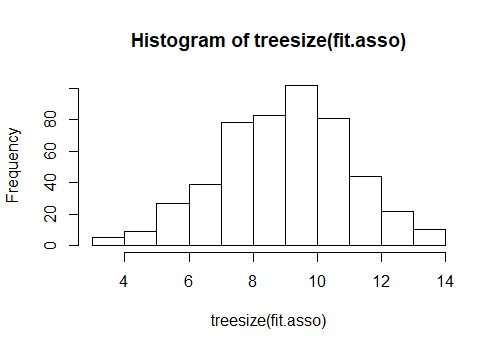


Figure 53. Histogram of tree sizes from Random Forest Using Health Sub-Dataset

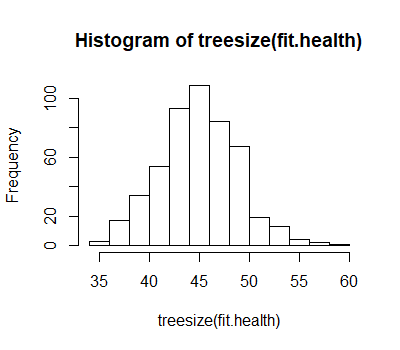
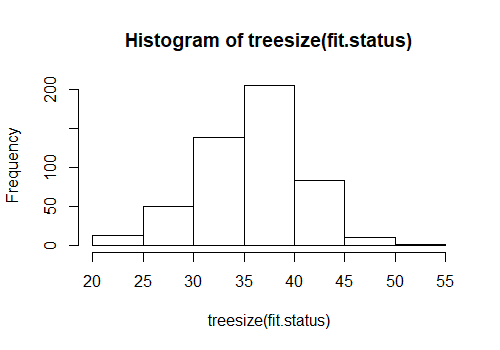


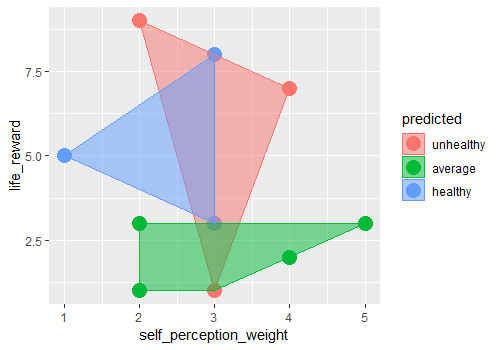
Figure 54. Histogram of tree sizes from Random Forest Using Health Sub-Dataset



K Nearest Neighbor

Like the other models, the dataset was required to be split into two datasets, one for training and one for testing the model. The test dataset was created by randomly selecting ten percent of the data, leaving the remaining ninety percent for the train dataset. However, unlike the other models, the healthy feeling label need to be removed from both the test and train dataset, since k Nearest Neighbor (kNN) is a lazy learner. All labels were removed and stored separately.

Figure 55. kNN Model Using all Data.



# randomly selection 10% for test data

n <- round(nrow(food.choices.numbonly)/10)

s <- sample(1:nrow(food.choices.numbonly), n)

# create test and training dataset

test <- food.choices.numbonly[s,]

train <- food.choices.numbonly[-s,]

k <- round(sqrt(nrow(food.choices.numbonly)))

#remove healthy feeling from test and train

train\_label <- train[,which(colnames(train)=="healthy\_feeling")]

train <- train[,-(which(colnames(train) == "healthy\_feeling"))]

test\_label <- test[,(which(colnames(test) == "healthy\_feeling"))]

test <- test[,-(which(colnames(test) == "healthy\_feeling"))]

# setup kNN

kNN\_all <- class::knn(train=train,

test=test,

cl=train\_label,

k=k,

prob=T)

print(kNN\_all)

#results

table(kNN\_all, test\_label)

CrossTable(x=test\_label,y=kNN\_all,prop.chisq=F)

The predictions from the model was compared to the actual labels of the healthy feeling for accuracy.

#visualize

plotdf <- data.frame(test,

predicted=kNN\_all)

plotdf <- data.frame(x=plotdf$self\_perception\_weight,

y=plotdf$life\_rewarding,

predicted=plotdf$predicted)

find\_hull <- function(df) df[chull(df$x, df$y),]

boundary <- ddply(plotdf, .variables="predicted", .fun=find\_hull)

ggplot(plotdf, aes(x, y, color=predicted, fill=predicted)) +

geom\_point(size=5) +

geom\_polygon(data=boundary, aes(x,y), alpha=0.5) +

scale\_x\_continuous("self\_perception\_weight") +

scale\_y\_continuous("life\_reward")

kNN was applied to the six different sub-datasets using the same preprocess and model.

Figure 56. kNN Model Using Exercise (Left) and Upbringing (Right) Sub-Dataset

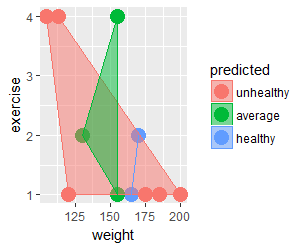
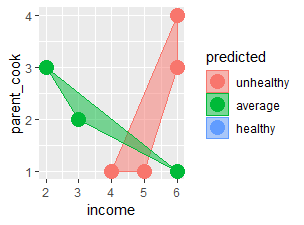
 

Figure 57. kNN Model Using Food Intake (Left) and Food Association (Right) Sub-Datasets

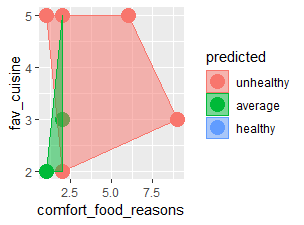
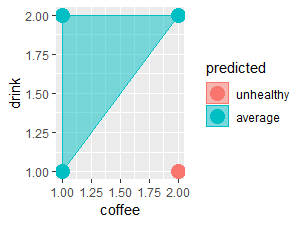
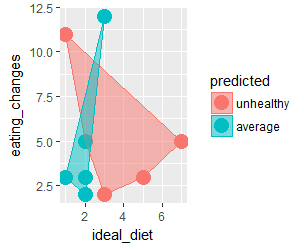
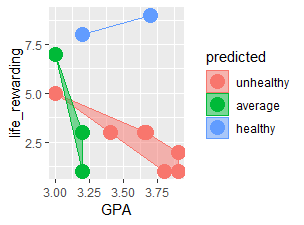
 

Figure 58. kNN Model Using Health (Left) and Status (Right) Sub-Datasets

# Results

The dataset itself provides some limitation and make skew the results. Since this survey was only conducted at Mercyhurst University, it is a limited sample and limits the diversity of the responses. Mercyhurst University in comprised of 58.2 percent white ethnicity. There is a slightly over ten percent of student combined in the Black or African American, Hispanic or Latino and Asian ethnicity. Additionally, majority of their students are female (55.2 percent). Lastly, most of their students comes from a Pennsylvania residency.

Figure 59. Mercyhurst University Diversity\*

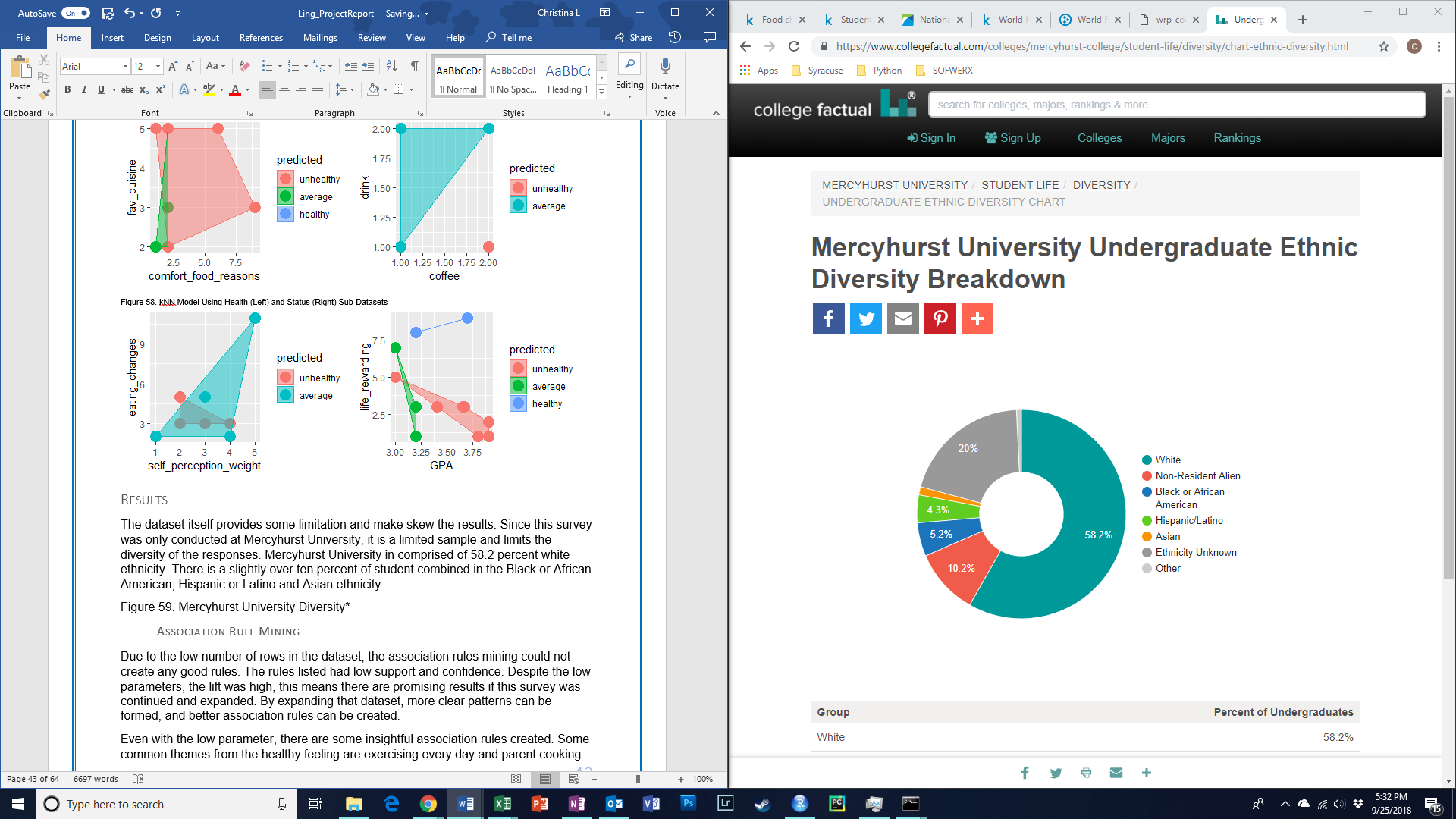


Figure 60. Mercyhurst University Gender Diversity\*

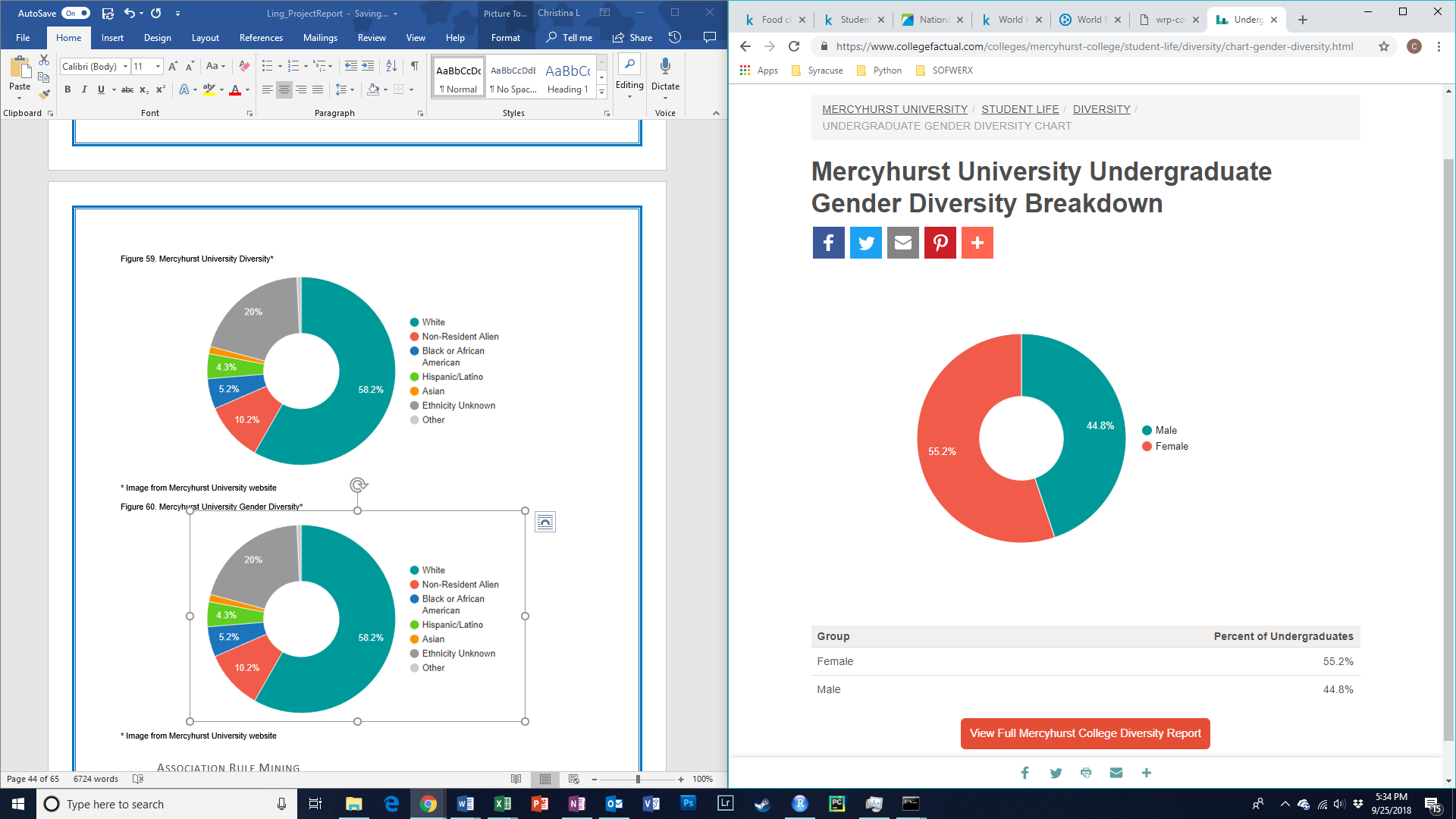


Figure 61. Mercyhurst University Residency Diversity\*



\* Images from Mercyhurst University website

Association Rule Mining

Due to the low number of rows in the dataset, the association rules mining could not create any good rules. The rules listed had low support and confidence. Despite the low parameters, the lift was high, this means there are promising results if this survey was continued and expanded. By expanding that dataset, more clear patterns can be formed, and better association rules can be created.

Even with the low parameter, there are some insightful association rules created. Some common themes from the healthy feeling are exercising every day and parent cooking almost every day. While there may seem to be a pattern with self-perception of being overweight, the data exploration showed that people who view themselves as overweight tend to work out more. There seem to be pattern with living on campus; however, majority of the student surveyed lives on campus (just under eighty percent of the data). So varied data would be required before making this conclusion.

Table 7. Summary of On\_Off\_Campus Variable.

|  |  |  |  |
| --- | --- | --- | --- |
| On Campus | Own House | Parents | Rent Off Campus |
| 98 | 2 | 9 | 16 |

On the other hand, for average healthy feeling, there are common pattern with self-perception of being just right in terms of weight. There is another common pattern with growing up eating American cuisine; however major of the student surveyed grow up eating American cuisine (just over eighty percent). So more varied data would be required before making this conclusion. While there may seem to be a pattern with association cereal with breakfast, majority of the responses was cereal (just under ninety percent). Before making this conclusion, more data would be required.

Table 9. Summary of Breakfast Variable.

|  |  |
| --- | --- |
| Cereal | Donut |
| 111 | 14 |

Table 10. Summary of Cuisine Variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| American | American Inspired International | Asian | Indian | Mexican | Other |
| 103 | 1 | 3 | 3 | 13 | 2 |

Surprisingly having a mother with college education (but not a master’s degree), lead to unhealthy feeling. There were also common themes of taking vitamins supplements. While there make seem to be a pattern with association McDonald’s Fries with fries, majority of the answers from the survey was McDonald’s Fries (over ninety percent). So more varied answers would be required before drawing this conclusion.

Table 10. Summary of Fries Variable.

|  |  |
| --- | --- |
| Home Fries | McDonald’s Fries |
| 11 | 114 |

Table 11. Summary of Pattern from Association Rules Mining and Healthy Feeling.

|  |  |  |
| --- | --- | --- |
| Healthy | Average | Unhealthy |
| * Exercise Every Day * Parent Cook Almost Every Day | * Self Perception of being Just Right | * Mother with a Bachelor’s Degree * Take Vitamin Supplements |

Clustering

Overall Clustering was not a useful method for this data. The method had difficulty clustering by healthy feeling. It is possible that it was clustering based from other close data points. The tables below show the results from K-means and EM clustering from each dataset.

Table 12. K-means Clustering (Left) and EM Clustering (Right) Results from all Data

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 8 | 19 | 21 |
| Average | 6 | 18 | 19 |
| Healthy | 5 | 19 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 1 | 47 | 0 |
| Average | 0 | 43 | 0 |
| Healthy | 0 | 33 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 21 | 19 | 8 |
| Average | 19 | 18 | 6 |
| Healthy | 10 | 19 | 5 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 19 | 5 | 24 |
| Average | 17 | 5 | 21 |
| Healthy | 19 | 3 | 12 |

Table 13. K-means Clustering (Left) and EM Clustering (Right) from Exercise Sub-Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 19 | 22 | 7 |
| Average | 10 | 20 | 13 |
| Healthy | 12 | 13 | 9 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 20 | 9 | 19 |
| Average | 19 | 5 | 19 |
| Healthy | 13 | 9 | 12 |

Table 14. K-Means Clustering (Left) and EM Clustering (Right) from Upbringing Sub-Dataset

Table 15. K-means Clustering (Left) and EM Clustering (Right) from Food Intake Sub-Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 23 | 20 | 5 |
| Average | 23 | 17 | 3 |
| Healthy | 15 | 17 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 18 | 17 | 13 |
| Average | 19 | 18 | 6 |
| Healthy | 18 | 10 | 6 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 13 | 19 | 16 |
| Average | 12 | 18 | 13 |
| Healthy | 9 | 17 | 8 |

Table 16. K-means Clustering (Left) and EM Clustering (Right) from Food Association Sub-Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 25 | 5 | 18 |
| Average | 23 | 3 | 17 |
| Healthy | 21 | 3 | 10 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 16 | 8 | 24 |
| Average | 9 | 6 | 28 |
| Healthy | 13 | 3 | 18 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 16 | 8 | 24 |
| Average | 9 | 6 | 28 |
| Healthy | 13 | 3 | 18 |

Table 17. K-Means Clustering (Left) and EM Clustering (Right) from Health Sub-Dataset

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 3 | 36 | 9 |
| Average | 12 | 17 | 14 |
| Healthy | 24 | 4 | 6 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | 1 | 2 | 3 |
| Unhealthy | 23 | 14 | 11 |
| Average | 26 | 9 | 8 |
| Healthy | 23 | 3 | 8 |

Table 18. K-Means Clustering (Left) and EM Clustering (Right) for Status Sub-Dataset

None of the dataset, all and sub-data, yielded in any useful clustering. There were no clusters with that was able to separate the healthy feeling by unhealthy, average and healthy.

Decision tree

The decision tree is more useful than clustering. Additionally, it provides useful insight in terms of what variables affect the healthy feeling. However, the visual is difficult to read and distinguish words. Some key factors are the feeling of life is rewarding, the individual’s self-perception of weight, how often they exercise in a week, their ideal diet, current diet code and the ability to guess how many calories are in a Panera Bread’s Roast Turkey and Avocado BLT. Some of these key variables support results from the association rule mining. There are common themes of self-perception of weight and exercising. There is only one path that lead to a healthy feeling and that is the following:

1. On a scale from one to ten, how much do you agree with the following statement: “I feel life is very rewarding!”? With a result greater than five.
2. A self-perception of weight being overweight
3. Exercising everyday

Table 19. Confusion Matrix for Decision Tree Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 5 | 0 | 0 | **5** |
| Average | 1 | 2 | 1 | **2** |
| Healthy | 0 | 0 | 3 | **3** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.83 | 1.00 | 0.75 |  |
| Recall | 1.00 | 0.50 | 1.00 |  |

Despite the small data to train the model, is has high accuracy and was an excellent model for predicting healthy feeling. Even though there is only one path to a healthy feeling, it predicted most of the healthy feeling individuals (missing only one).

In the Decision Tree using the exercise sub-dataset, there are some common themes that comes up. The variables self-perception of weight and exercise the common themes that reoccur from the models.

Table 20. Confusion Matrix for Decision Tree Using Exercise Sub-Dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 0 | 4 | 2 | **6** |
| Average | 1 | 2 | 2 | **5** |
| Healthy | 1 | 0 | 0 | **1** |
|  | **2** | **6** | **4** | **12** |
| Precision | 0.00 | 0.33 | 0.00 |  |
| Recall | 0.00 | 0.40 | 0.00 |  |

Despite the similar variables used, this model was not accurate. Exercise only does not play an important part of feeling healthy. Instead it is a combination of variables that play into feeling healthy.

Like association rules mining, there is a link between a mother’s with college degrees and feeling unhealthy. All the other variables are new to the analysis.

Table 21. Confusion Matrix for Decision Tree Using Upbringing Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 2 | 0 | 1 | **3** |
| Average | 3 | 2 | 2 | **7** |
| Healthy | 1 | 0 | 1 | **2** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.33 | 1.00 | 0.25 |  |
| Recall | 0.67 | 0.29 | 0.50 |  |

This model is more accurate than the decision tree made from the exercise sub-dataset. While individually upbringing has more of an impact on feeling healthy. Overall this decision tree is less accurate than the original. While upbringing has more impact than exercise, it does not have more impact than the combine variables.

Like the original decision tree, this visual is difficult to read. There are no common themes between this decision and other models

Table 22. Confusion Matrix for Decision Tree Using Food Intake Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 2 | 1 | **4** |
| Average | 5 | 0 | 3 | **8** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **6** | **2** | **1** | **12** |
| Precision | 0.17 | 0.00 | 0.00 |  |
| Recall | 0.25 | 0.00 | NA |  |

This decision tree is the less accurate so far. As a result, food intake does not play an important factor on feeling healthy according to decision trees.

At the bottom of the decision tree, the Fries factor that determines healthy from unhealthy. Fries was also a key factor in association rules mining. However, with so many responses lead toward McDonald’s Fries it is difficult to determine the impact of skewness. With the combine association rules and decision trees it is more likely to be factor in determining healthy feeling.

Table 23. Confusion Matrix for Decision Tree Using Food Association Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 1 | 2 | **6** |
| Average | 3 | 1 | 2 | **6** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.50 | 0.50 | 0.00 |  |
| Recall | 0.50 | 0.17 | NA |  |

While this is more accurate than food intake, this is still a low accurate model. Food association has more of an impact than food intake, however it is not a big impact on healthy feeling.

This decision tree is a little difficult to read. There are no common factors between this decision tree and other models.

Table 24. Confusion Matrix for Decision Tree Using Health Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 2 | 3 | 4 | **9** |
| Average | 0 | 2 | 0 | **2** |
| Healthy | 0 | 1 | 0 | **1** |
|  | **2** | **6** | **4** | **12** |
| Precision | 1.00 | 0.33 | 0.00 |  |
| Recall | 0.22 | 1.00 | 0.00 |  |

This is as accurate as food association; however, this does not necessarily mean that the two have the same impact on feeling healthy.

The original decision tree displays the first splitting being the feeling that life is rewarding, while on this decision tree it uses that variable twice immediately after one another. This shows the importance of feeling that life is reward.

Table 25. Confusion Matrix for Decision Tree Using Status Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 6 | 0 | 1 | **7** |
| Average | 0 | 2 | 2 | **4** |
| Healthy | 0 | 0 | 1 | **1** |
|  | **6** | **2** | **4** | **12** |
| Precision | 1.00 | 1.00 | 0.25 |  |
| Recall | 0.86 | 0.50 | 1.00 |  |

Of the sub-dataset, this decision tree was the most accurate. This proves that feeling that life in rewarding is a big factor on feeling healthy. Especially since this factor is used twice in this decision tree and is the top splitting factor in the original decision tree.

Naïve Bayes

Table 26. Confusion Matrix for Naïve Bayes Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 0 | 2 | **5** |
| Average | 3 | 2 | 1 | **6** |
| Healthy | 0 | 0 | 1 | **1** |
|  | **6** | **2** | **1** |  |
| Precision | 0.50 | 1.00 | 0.25 |  |
| Recall | 0.60 | 0.33 | 1.00 |  |

Overall this model performed at decent accuracy considering the small dataset. With more responses to the survey, this model could be improved to be more accurate.

Table 27. Confusion Matrix for Naïve Bayes Using the Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 0 | 2 | **5** |
| Average | 0 | 2 | 0 | **2** |
| Healthy | 5 | 0 | 0 | **5** |
|  | **8** | **2** | **2** | **12** |
| Precision | 0.38 | 1.00 | 0.00 |  |
| Recall | 0.60 | 1.00 | 0.00 |  |

This Naïve Bayes model is less accurate than the original model. So, exercise is not as impactful as a combination of variables. While it has a low accuracy, this model is still more accurate than other individual models. This shows the impact of exercise on feeling healthy

Table 28. Confusion Matrix for Naïve Bayes Using the Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 1 | 2 | **6** |
| Average | 3 | 1 | 2 | **6** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.50 | 0.50 | 0.00 |  |
| Recall | 0.50 | 0.17 | NA |  |

This has a lower accuracy than exercise and the combine. According to Naïve Bayes, upbringing does not have a big impact on feeling healthy. Other models previous examined stated that parents cooking (a variable under upbringing) is important to feeling healthy.

Table 29. Confusion Matrix for Naïve Bayes Using the Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 2 | 2 | **5** |
| Average | 3 | 0 | 2 | **5** |
| Healthy | 2 | 0 | 0 | **2** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.17 | 0.00 | 0.00 |  |
| Recall | 0.20 | 0.00 | 0.00 |  |

This is the less accurate model from Naïve Bayes. With this model and the decision tree model not being accurate, it shows that food intake does not play a big factor on feeling healthy.

Table 30. Confusion Matrix for Naïve Bayes Using the Food Association Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 2 | 0 | 1 | **3** |
| Average | 4 | 2 | 3 | **9** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **6** | **2** | **4** | **12** |
| Precision | 0.33 | 1.00 | 0.00 |  |
| Recall | 0.67 | 1.00 | 0.00 |  |

This is as accurate as the upbringing Naïve Bayes model. There is still a possibility that food association and upbringing has a small impact on feeling healthy. Association Rules Mining provides some using food association with feeling healthy.

Table 31. Confusion Matrix for Naïve Bayes Using the Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 0 | 1 | **4** |
| Average | 0 | 1 | 0 | **1** |
| Healthy | 5 | 1 | 1 | **7** |
|  | **8** | **2** | **2** | **12** |
| Precision | 0.38 | 0.50 | 0.50 |  |
| Recall | 0.75 | 1.00 | 0.14 |  |

This is as accurate as the exercise Naïve Bayes Model making it tied for second on the individual factors. The decision tree model of health was also decently accurate in comparison to the other sub-datasets. There is a possibility that health habits affect feeling healthy, further exploration is required.

Table 32. Confusion Matrix for Naïve Bayes Using the Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 6 | 0 | 1 | **7** |
| Average | 0 | 1 | 0 | **1** |
| Healthy | 0 | 1 | 3 | **4** |
|  | **6** | **2** | **4** | **12** |
| Precision | 1.00 | 0.50 | 0.75 |  |
| Recall | 0.86 | 1.00 | 0.75 |  |

This is the most accurate model of Naïve Bayes, even more than the original Naïve Bayes with all variables used. This result combined with the decision tree results shows the importance of feeling that life is rewarding. It is very likely that feeling that life is rewarding has an impact on feeling healthy.

Support Vector Machine

Table 33. Confusion Matrix for SVM – Polynomial Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 3 | 0 | **7** |
| Average | 1 | 0 | 1 | **2** |
| Healthy | 2 | 0 | 1 | **3** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.00 | 0.50 |  |
| Recall | 0.57 | 0.00 | 0.33 |  |

Table 34. Confusion Matrix for SVM – Linear Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 2 | 0 | **6** |
| Average | 2 | 1 | 0 | **3** |
| Healthy | 1 | 0 | 2 | **3** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.33 | 1.00 |  |
| Recall | 0.67 | 0.33 | 0.67 |  |

Table 35. Confusion Matrix for SVM – Radial Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 3 | 2 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.58 | NA | NA |  |

For the combined dataset, the Linear SVM model was the best model. This is more accurate than the Naïve Bayes Model, but less accurate than the Decision Tree. The SVM model still provided desent results considering the small dataset. The model could be improved with more response to the survey.

Table 36. Confusion Matrix for SVM – Polynomial Using the Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 3 | 0 | **4** |
| Average | 2 | 2 | 1 | **5** |
| Healthy | 0 | 0 | 3 | **3** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.33 | 0.40 | 0.75 |  |
| Recall | 0.25 | 0.40 | 1.00 |  |

Table 37. Confusion Matrix for SVM – Linear Using the Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 2 | 2 | **5** |
| Average | 2 | 3 | 2 | **7** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.33 | 0.60 | 0.00 |  |
| Recall | 0.20 | 0.43 | NA |  |

Table 38. Confusion Matrix for SVM – Radial Using the Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 5 | 4 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **3** | **5** | **4** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.25 | NA | NA |  |

For the Exercise sub-dataset, the Polynomial SVM was most accurate; although still less accurate than the combined data using the Linear SVM model.

Table 39. Confusion Matrix for SVM – Polynomial Using the Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 2 | 1 | **6** |
| Average | 1 | 1 | 0 | **2** |
| Healthy | 3 | 0 | 1 | **4** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.43 | 0.33 | 0.50 |  |
| Recall | 0.50 | 0.50 | 0.25 |  |

Table 40. Confusion Matrix for SVM – Linear Using the Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 2 | 1 | **4** |
| Average | 6 | 1 | 1 | **8** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** |  |
| Precision | 0.14 | 0.33 | 0.00 |  |
| Recall | 0.25 | 0.13 | NA |  |

Table 41. Confusion Matrix for SVM – Radial Using the Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 3 | 2 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** |  |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.58 | NA | NA |  |

While the Radial SVM was the most accurate, it was not a good model considering the precision and recall. The next best model was the Polynomial. This is less accurate than the Exercise Sub-dataset, which is a common theme between the models. Exercise might have a bigger impact on feeling healthy.

Table 42. Confusion Matrix for SVM – Polynomial Using the Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 2 | 1 | 1 | **4** |
| Average | 5 | 1 | 1 | **7** |
| Healthy | 0 | 1 | 0 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.29 | 0.33 | 0.00 |  |
| Recall | 0.50 | 0.14 | 0.00 |  |

Table 43. Confusion Matrix for SVM – Linear Using the Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 1 | 2 | **7** |
| Average | 3 | 2 | 0 | **5** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.67 | 0.00 |  |
| Recall | 0.57 | 0.40 | NA |  |

Table 44. Confusion Matrix for SVM – Radial Using the Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 3 | 2 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.58 | NA | NA |  |

Although the Radial has the highest accuracy of the three, it is not a good model when factoring in the precision and recall of the model. The Radial model only predicted feeling unhealthy. This is the first that the Food Intake Sub-dataset had decent model results. There is a possibility for a small impact on feeling healthy; however, this model could be an outlier. This will be considered moving forward.

Table 45. Confusion Matrix for SVM – Polynomial Using the Food Association Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 2 | 2 | **7** |
| Average | 4 | 1 | 0 | **5** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.43 | 0.33 | 0.00 |  |
| Recall | 0.42 | 0.20 | NA |  |

Table 46. Confusion Matrix for SVM – Linear Using the Food Association Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 3 | 2 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.58 | NA | NA |  |

Table 47. Confusion Matrix for SVM – Radial Using the Food Association Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 3 | 2 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.58 | NA | NA |  |

Like the other Radial Models, this had the highest accuracy; however, after considering the precision and recall, this is not a good model. Linear SVM has the same issues. The Polynomial Model does not have good accuracy. According to the SVM model, food association does not have a significant impact on feeling healthy.

Table 48. Confusion Matrix for SVM – Polynomial Using the Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 1 | 0 | **2** |
| Average | 2 | 4 | 3 | **9** |
| Healthy | 0 | 0 | 1 | **1** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.33 | 0.80 | 0.25 |  |
| Recall | 0.50 | 0.44 | 1.00 |  |

Table 49. Confusion Matrix for SVM – Linear Using the Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 0 | 1 | 1 | **2** |
| Average | 3 | 3 | 3 | **9** |
| Healthy | 0 | 1 | 0 | **1** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.00 | 0.60 | 0.00 |  |
| Recall | 0.00 | 0.33 | 0.00 |  |

Table 50. Confusion Matrix for SVM – Radial Using the Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 5 | 4 | **12** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **3** | **5** | **4** | **12** |
| Precision | 1.00 | 0.00 | 0.00 |  |
| Recall | 0.25 | NA | NA |  |

The Polynomial SVM Model is the best for the Health Sub-dataset. Considering the small dataset, it is decent accuracy, precision and recall. There is still a possibility for a small impact on feeling healthy.

Table 51. Confusion Matrix for SVM – Polynomial Using the Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 6 | 2 | 0 | **8** |
| Average | 1 | 1 | 1 | **3** |
| Healthy | 0 | 0 | 1 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.86 | 0.33 | 0.50 |  |
| Recall | 0.75 | 0.33 | 1.00 |  |

Table 52. Confusion Matrix for SVM – Linear Using the Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 2 | 0 | **9** |
| Average | 0 | 0 | 0 | **0** |
| Healthy | 0 | 1 | 2 | **3** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.00 | 1.00 |  |
| Recall | 0.78 | NA | 0.67 |  |

Table 53. Confusion Matrix for SVM – Radial Using the Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 7 | 2 | 0 | **9** |
| Average | 0 | 1 | 2 | **3** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 1.00 | 0.33 | 0.00 |  |
| Recall | 0.78 | 0.33 | NA |  |

The Linear SVM Model for the Status sub-dataset was the most accurate of all the SVM Models. However, this model did not predict any in the average feeling category. While the other SVM models did predict all categories. Even so, the other SVM Models for status is still more accurate than the other SVM Models.

Random Forest

Figure 62. Variables used in Tree from Random Forest Using All Data

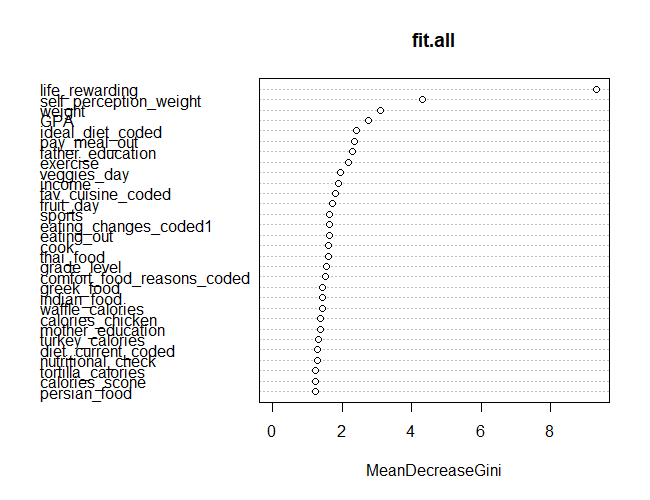


Table 54. Confusion Matrix for Random Forest Using All Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 5 | 2 | 0 | **7** |
| Average | 2 | 1 | 0 | **3** |
| Healthy | 0 | 0 | 2 | **2** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.71 | 0.33 | 1.00 |  |
| Recall | 0.71 | 0.33 | 1.00 |  |

This Random Forest model is accurate considering the small number of responses to the survey. Figure 62. displays the importance of feeling that life is reward and the self-perception of weight are on feeling healthy.

Figure 63. Variables used in Tree from Random Forest Using Exercise Sub-Dataset

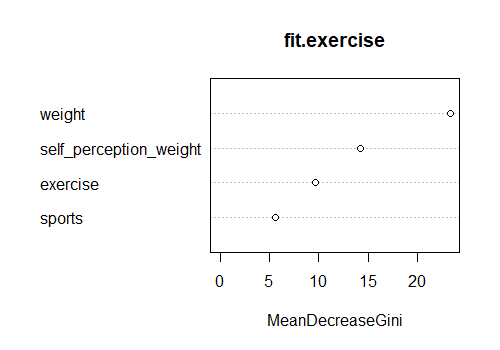


Table 54. Confusion Matrix for Random Forest Using Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 2 | 2 | 0 | **4** |
| Average | 1 | 1 | 1 | **3** |
| Healthy | 0 | 2 | 3 | **5** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.67 | 0.20 | 0.75 |  |
| Recall | 0.50 | 0.33 | 0.60 |  |

The Random Forest for the Exercise sub-dataset shows the important of weight and self-perception of weight on feeling healthy. Surprisingly the variable exercise was only used less than half the time. The model is decently accurate in comparison to the other models. It is likely the exercise has an impact on feeling healthy.

Figure 64. Variables used in Tree from Random Forest Using Upbringing Sub-Dataset

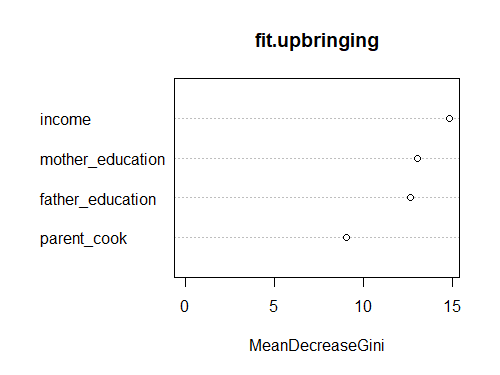


Table 55. Confusion Matrix for Random Forest Using Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 2 | 0 | **5** |
| Average | 4 | 1 | 0 | **5** |
| Healthy | 0 | 0 | 2 | **2** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.43 | 0.33 | 1.00 |  |
| Recall | 0.60 | 0.20 | 1.00 |  |

The Upbringing Random Forest model shows that income, mother’s education and father’s education are used the most in the model. The predict predicted with decent accurate. It is as accurate as exercise. It is likely that upbringing has some impact on feeling healthy.

Figure 65. Variables used in Tree from Random Forest Using Food Intake Sub-Dataset

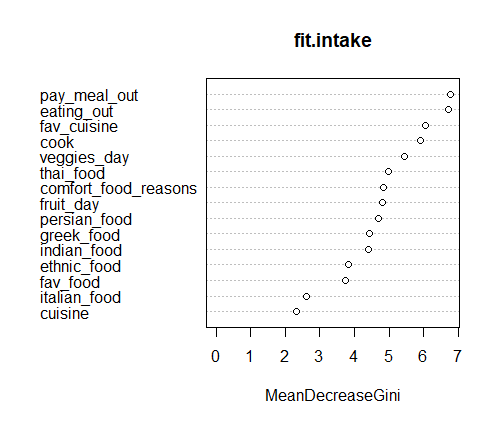


Table 56. Confusion Matrix for Random Forest Using Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 2 | 1 | **7** |
| Average | 2 | 1 | 1 | **4** |
| Healthy | 1 | 0 | 0 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.33 | 0.00 |  |
| Recall | 0.57 | 0.25 | 0.00 |  |

This is less accurate than the previous two models with sub-data. According to Random Forest, food intake has less of an impact on feeling healthy than upbringing and exercise. The variables used the most are how much individuals pay for a meal out and how often they eat out.

Figure 66. Variables used in Tree from Random Forest Using Food Association Sub-Dataset

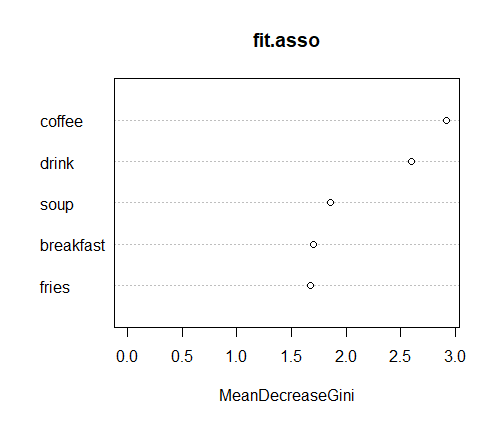


Table 57. Confusion Matrix for Random Forest Using Food Association Sub-Dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 2 | 1 | **6** |
| Average | 4 | 1 | 1 | **6** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.43 | 0.33 | 0.00 |  |
| Recall | 0.50 | 0.17 | NA |  |

This is not an accurate model. According to the Random Forest Model, food association does not have a big impact on feeling healthy. It had difficult distinguishing the healthy feeling using food association only. Surprisingly Figure 66. shows low values for breakfast and fries, which association rules mining showed these variables as important.

Figure 67. Variables used in Tree from Random Forest Using Health Sub-Dataset

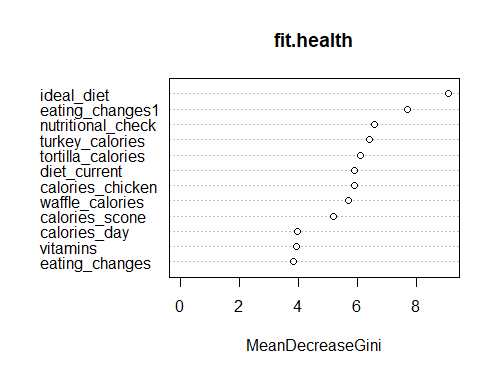


Table 58. Confusion Matrix for Random Forest Using Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 0 | 4 | 1 | **5** |
| Average | 3 | 0 | 2 | **5** |
| Healthy | 0 | 1 | 1 | **2** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.00 | 0.00 | 0.25 |  |
| Recall | 0.00 | 0.00 | 0.50 |  |

This is the less accurate of the Random Forest; additionally, the precision and recall are all low. The model predicted everywhere but the actual healthy feeling. It shows that these variables are not impactful on feeling healthy. Although other models show that there is some impact. There is still a possibility for a small impact.

Figure 68. Variables used in Tree from Random Forest Using Status Sub-Dataset



Table 59. Confusion Matrix for Random Forest Using Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 6 | 1 | 0 | **7** |
| Average | 1 | 2 | 1 | **4** |
| Healthy | 0 | 0 | 1 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.86 | 0.67 | 0.50 |  |
| Recall | 0.86 | 0.50 | 1.00 |  |

This is the most accurate of the Random Forest Models. Additionally Figure 68. shows the value of feeling that life is rewarding on feeling healthy. With so many models pointing to this relationship, there is a strong prediction factor on this variable.

K Nearest Neighbor

Table 60. Confusion Matrix for kNN Using all Data

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 0 | 2 | 1 | **3** |
| Average | 5 | 1 | 1 | **7** |
| Healthy | 2 | 0 | 0 | **2** |
|  | **7** | **3** | **2** |  |
| Precision | 0.00 | 0.33 | 0.00 |  |
| Recall | 0.00 | 0.14 | 0.00 |  |

Of all the models conducted, this is the less accurate model using all data. It nearly predicted everywhere but the actual healthy feeling.

Table 61. Confusion Matrix for kNN Using Exercise Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 1 | 3 | 2 | **6** |
| Average | 2 | 2 | 0 | **4** |
| Healthy | 0 | 0 | 2 | **2** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.33 | 0.40 | 0.50 |  |
| Recall | 0.17 | 0.50 | 1.00 |  |

In comparison to the last kNN model, using Exercise only data is a much better model. However, it is still far from being accurate.

Table 62. Confusion Matrix for kNN Using Upbringing Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 1 | 1 | **6** |
| Average | 2 | 2 | 1 | **5** |
| Healthy | 1 | 0 | 0 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.67 | 0.00 |  |
| Recall | 0.67 | 0.40 | 0.00 |  |

This is decently accurate, and considering the last few kNN models, this is the best one so far. Upbringing probability has a small impact on feeling healthy. More responses to the survey are required for further exploration.

Table 63. Confusion Matrix for kNN Using Food Intake Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 4 | 2 | 1 | **7** |
| Average | 2 | 1 | 1 | **4** |
| Healthy | 1 | 0 | 0 | **1** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.57 | 0.33 | 0.00 |  |
| Recall | 0.57 | 0.25 | 0.00 |  |

In comparison to the other kNN models, this is decently accurate. This is the second time that food intake returned decent results. There might be a small impact on healthy feeling; however, not as much as the other variables.

Table 64. Confusion Matrix for kNN Using Food Association Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 3 | 2 | 0 | **5** |
| Average | 4 | 1 | 2 | **7** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.53 | 0.33 | 0.00 |  |
| Recall | 0.60 | 0.14 | NA |  |

This is not an accurate model. Considering the past models and skewness in association rules mining, there is little impact food association has on feeling healthy.

Table 65. Confusion Matrix for kNN Using Health Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 0 | 2 | 3 | **5** |
| Average | 3 | 3 | 1 | **7** |
| Healthy | 0 | 0 | 0 | **0** |
|  | **3** | **5** | **4** | **12** |
| Precision | 0.00 | 0.60 | 0.00 |  |
| Recall | 0.00 | 0.43 | NA |  |

This is the less accurate kNN model so far. There is not much impact of healthy habits on feeling healthy. Most of the models using the Health Sub-Dataset did not result accurate predictions.

Table 66. Confusion Matrix for kNN Using Status Sub-Dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | Unhealthy | Average | Healthy |  |
| Predicted | Unhealthy | 6 | 1 | 0 | **7** |
| Average | 1 | 2 | 0 | **3** |
| Healthy | 0 | 0 | 2 | **2** |
|  | **7** | **3** | **2** | **12** |
| Precision | 0.86 | 0.67 | 1.00 |  |
| Recall | 0.86 | 0.67 | 1.00 |  |

This is the most accurate model of kNN. Considering all the other models, this dataset had a huge impact on feeling healthy. The variables that has the biggest impact in this sub-dataset is the life\_reward variable (the feeling that life is rewarding).

# Conclusion

Feeling healthy is important to everyone. Often people’s mood drives and encourages their actions. However, the challenge is how to feel healthy, it is difficult to determine what an individual can do to feel healthy. One of the biggest impacts on feeling healthy is feeling that life is rewarding. While exercising and upbringing has a factor on how healthy an individual feel; it is as impactful as feeling that life is rewarding.

Feeling healthy is important to the following people: people who are dieting or are health advocates, nutritionist and dietician, physical trainers, and psychologist. People who are dieting or are healthy advocates are concern about their health. By feeling healthy they are more likely to make healthier decisions; thus, they can fulfill the goals they set for themselves.

Nutritionist and dietician are also concerned with feeling healthy. They are concerned with what impacts feeling healthy. These are professional that people go to for advice about health. Since they provide a service, they are required to know as much as possible about their field. By knowing what makes people feel healthy, they can provide better service to their clients or customers.

Feeling healthy is also important to physical trainers for similar reasons as nutritionist/dieticians. Physical trainers also provide a service. Kinesiology explains that exercising makes people feel better. By exercise an individual can feel happier and healthy about themselves. It is important to physical trainers to explain to important of feeling healthy. By combining other activities with exercising, their clients can have better results.

Lastly is it important to psychologist, they are expert when it comes to feelings. So, know what impacts feeling health it important to them. People seek psychologist help when they need help. A psychologist may need to advise a client on how to feel healthier, so they can feel better about themselves. This can help push and motivate their clients.

However, these conclusions should be taken with caution. These conclusions are based on information that may be bias. More information would be required to be confidence about these conclusions. This analysis provides a strong foundation into exploring what impacts feeling healthy.