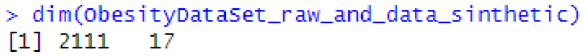
Nhi Ho: 2161095 | Timothy Lu: 2055688 | Kevin Tran: 2033078 | Kelsey Wong: 2042459

Obesity Level Predictions using Habitual and Physical Traits

1. Introduction

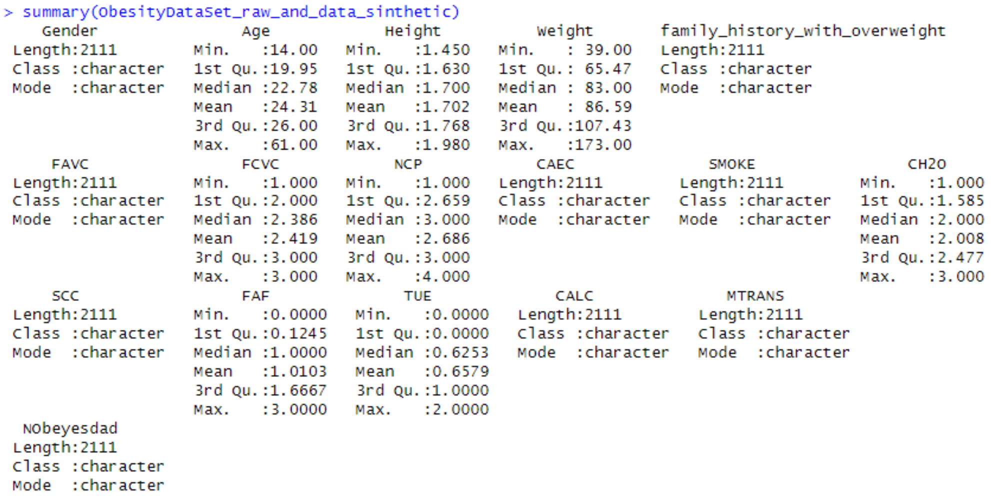
- Our goal is to predict obesity given the predictors of height, weight, number of main meals in a day, and physical activity levels. We believe that predicting obesity is important in the context of this data because obesity is a constant problem in modern society. The predictors we have chosen to help predict obesity are usually the factors that are most scrutinized by the common person. The relationships between these predictors are also the simplest ones for a regular person to draw observations from.
- The “Estimation Of Obesity Levels Based On Eating Habits and Physical Condition” dataset estimates obesity levels based on eating habits and physical condition reports from people between ages 14 and 61 in Mexico City, Mexico, Lima, Peru, and Barranquilla, Colombia. Approximately 23% of responses were obtained through an online web platform while the remaining responses were synthesized with machine learning processes. We decided to choose this dataset because obesity is a problem in the modern world that is continuing to trend upward in populations across the globe. Obesity is considered an epidemic by health professionals especially in the United States, which is close to Mexico where some of this data was collected. By looking at this dataset, we will be able to see the relationship between factors that can influence the likelihood of obesity in a person. This dataset was retrieved from the UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition).

- **Exploratory Data Analysis**



```
> dim(ObesityDataSet_raw_and_data_sinthetic)
[1] 2111   17
```

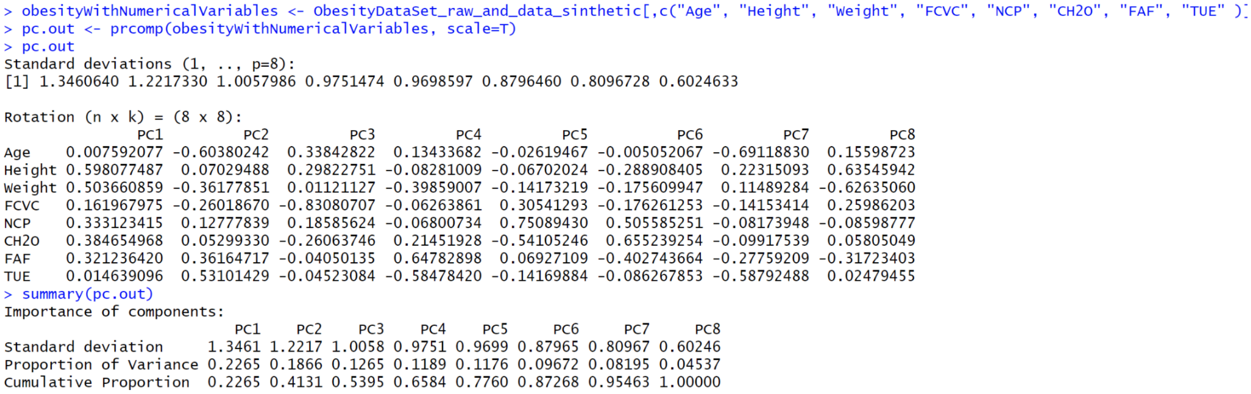
The data has 2111 observations with 17 predictors.



```
> summary(ObesityDataSet_raw_and_data_sinthetic)
    Gender               Age            Height          Weight       family_history_with_overweight
 Length:2111        Min.   :14.00   Min.   :1.450   Min.   : 39.00   Length:2111
 Class :character   1st Qu.:19.95   1st Qu.:1.630   1st Qu.: 65.47   Class :character
 Mode  :character   Median :22.78   Median :1.700   Median : 83.00   Mode  :character
                    Mean   :24.31   Mean   :1.702   Mean   : 86.59
                    3rd Qu.:26.00   3rd Qu.:1.768   3rd Qu.:107.43
                    Max.   :61.00   Max.   :1.980   Max.   :173.00
     FAVC              FCVC            NCP             CAEC              SMOKE                CH2O
 Length:2111        Min.   :1.000   Min.   :1.000   Length:2111        Length:2111        Min.   :1.000
 Class :character   1st Qu.:2.000   1st Qu.:2.659   Class :character   Class :character   1st Qu.:1.585
 Mode  :character   Median :2.386   Median :3.000   Mode  :character   Mode  :character   Median :2.000
                    Mean   :2.419   Mean   :2.686                                         Mean   :2.008
                    3rd Qu.:3.000   3rd Qu.:3.000                                         3rd Qu.:2.477
                    Max.   :3.000   Max.   :4.000                                         Max.   :3.000
     SCC                FAF              TUE              CALC              MTRANS
 Length:2111        Min.   :0.0000   Min.   :0.0000   Length:2111        Length:2111
 Class :character   1st Qu.:0.1245   1st Qu.:0.0000   Class :character   Class :character
 Mode  :character   Median :1.0000   Median :0.6253   Mode  :character   Mode  :character
                    Mean   :1.0103   Mean   :0.6579
                    3rd Qu.:1.6667   3rd Qu.:1.0000
                    Max.   :3.0000   Max.   :2.0000
  NObeyesdad
 Length:2111
 Class :character
 Mode  :character
```

This dataset measured data from people aged 14 to 61. The heights ranged from 1.45 meters to 1.98 meters while the weights ranged from 39 kilograms to 173 kilograms. It seems most surveyors ate an average of 3 meals per day with many of the responses reporting higher frequencies of vegetable consumption in their meals. Most of the subjects drank an average of 2 liters of water per day. The surveyors did a median of 1 day of physical activity during the week, and most responses reported a lower average usage of electronics that require little physical activity.

Performing principal component analysis:



```
> obesityWithNumericalVariables <- ObesityDataSet_raw_and_data_sinthetic[,c("Age", "Height", "Weight", "FCVC", "NCP", "CH2O", "FAF", "TUE" )]
> pc.out <- prcomp(obesityWithNumericalVariables, scale=T)
> pc.out
Standard deviations (1, .., p=8):
[1] 1.3460640 1.2217330 1.0057986 0.9751474 0.9698597 0.8796460 0.8096728 0.6024633

Rotation (n x k) = (8 x 8):
               PC1         PC2         PC3         PC4         PC5          PC6         PC7         PC8
Age    0.007592077 -0.60380242  0.33842822  0.13433682 -0.02619467 -0.005052067 -0.69118830  0.15598723
Height 0.598077487  0.07029488  0.29822751 -0.08281009 -0.06702024 -0.288908405  0.22315093  0.63545942
Weight 0.503660859 -0.36177851  0.01121127 -0.39859007 -0.14173219 -0.175609947  0.11489284 -0.62635060
FCVC   0.161967975 -0.26018670 -0.83080707 -0.06263861  0.30541293 -0.176261253 -0.14153414  0.25986203
NCP    0.333123415  0.12777839  0.18585624 -0.06800734  0.75089430  0.505585251 -0.08173948 -0.08598777
CH2O   0.384654968  0.05299330 -0.26063746  0.21451928 -0.54105246  0.655239254 -0.09917539  0.05805049
FAF    0.321236420  0.36164717 -0.04050135  0.64782898  0.06927109 -0.402743664 -0.27759209 -0.31723403
TUE    0.014639096  0.53101429 -0.04523084 -0.58478420 -0.14169884 -0.086267853 -0.58792488  0.02479455
> summary(pc.out)
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7     PC8
Standard deviation     1.3461 1.2217 1.0058 0.9751 0.9699 0.87965 0.80967 0.60246
Proportion of Variance 0.2265 0.1866 0.1265 0.1189 0.1176 0.09672 0.08195 0.04537
Cumulative Proportion  0.2265 0.4131 0.5395 0.6584 0.7760 0.87268 0.95463 1.00000
```

75% of the variance in the data can be determined based on the first 5 PCs.

In PC1 which explains the highest amount of variance for the dataset, both “Height” and “Weight” are the most positive, largest coefficients. In PC2 and PC3, the highest coefficients were for “Age” and “FCVC,” respectively. In PC4, “FAF” had the highest loading value, and in PC5, “NCP” had the highest loading value. These variables are therefore classified as an important contributor to that principal component. We chose to ignore “Age” and “FCVC” as predictors since these are less commonly attributable to obesity levels and were negatively correlated to PCs and focused on the remaining variables that were positively correlated with the PCs that accounted for a large portion of the data variance.

2. Methodology

- Our project requires methods for supervised learning as our goal is to build a statistical model to predict/estimate obesity levels based on multiple inputs. The models and methods we are using in order to predict obesity given the predictors: height, weight, number of main meals in a day, and physical activity levels are K-Nearest Neighbors (KNN) and support vector machines (SVM).
- KNN is a supervising learning machine that is used for classification and regression tasks (our project will be using KNN for classification). This method relies on the idea that similar data points tend to have similar labels or values. The KNN method calculates the distance from the input data point to the nearest neighbors in the training data set and then identifies the chosen value of K nearest neighbors closest to the input data point. This method then calculates the conditional probability for each classification of the K neighbors and assigns the highest probability / most common labels among the K neighbors to the input data point as the predicted label. The smaller the K is, the more flexible the method will be.
- SVM is a supervised learning method that solves the problem of enlarging a feature space to accurately classify classes using a support vector classifier where the decision boundary is linear or nonlinear, as simply enlarging a feature space would produce too many computations. By solving the support classifier optimization task using the inner products of the observations, SVM can replace these inner products with a generalization of the inner product that uses a function called a kernel. Kernels quantify the similarities

between the two observations. Using the linear kernels produces efficient linear decision boundaries, and using the polynomial and radial kernels produces efficient non-linear decision boundaries. A test observation is classified based on which side of the decision boundary it lies on. The SVM has an advantage over simply enlarging the feature space as it does not work in the enlarged feature space, leading to less computations as a whole. For our data analysis, we will be using the radial kernel to enlarge the predictor space. The function for a radial kernel is below.

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(-\gamma \sum_{k=1}^{p} (x_{ik} - x_{jk})^2), \gamma > 0$$

- Some advantages and disadvantages of each method are listed below.
  - Both KNN and SVM can work with linear and non-linear distributed data.
  - KNN is sensitive to the scale of data and outliers, so scaling and removing the irrelevant data points is necessary in order to get a better result. Meanwhile, SVM is good with outliers since it only uses relevant points to find the linear or non-linear separation.
  - KNN is less computationally intensive and easier to implement and understand compared to SVM. KNN classifies labels for input data simply based on the distance matrix, while SVM uses support vector classifiers to map observations to higher feature spaces so they can be categorized when the data is otherwise linearly inseparable.
  - SVM works well on a dataset that has many attributes.
  - KNN only needs to set K parameters while SVM needs to identify the kernel choices to be specified if classes are not linearly separated (gamma or degree values also need to be selected).

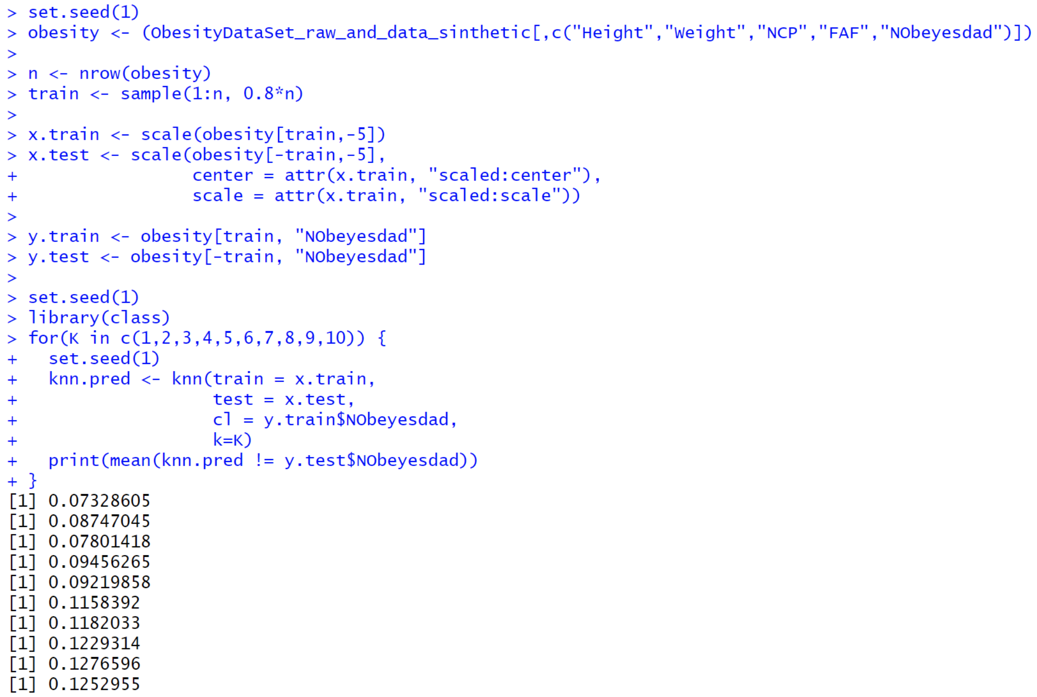
3. Data Analysis

(a) We chose to exclude the following predictors: gender, age, family_history_with_overweight, FAVC, FCVC, CAEC, SMOKE, CH2O, SCC, TUE, CALC, and MTRANS (see appendix). Our group determined that gender and age had no effect on obesity levels as these variables would least help determine overweight/obesity levels since an individual of any gender and any age may be obese depending on other factors. The other variables we chose not to account for did not seem like the first factors the common person would look to for the cause of their obesity.

As observed from the dataset, we can see all of the predictors were measured in different units. Plus the variable “Weight” has values that are significantly larger compared to others. As a result, scaling the data set is necessary in our case in order to achieve a more accurate result.

(b)

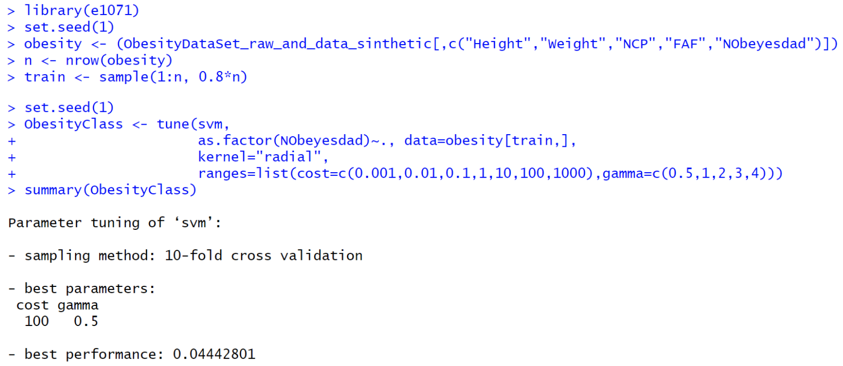
- KNN (Kelsey Wong and Nhi Ho)



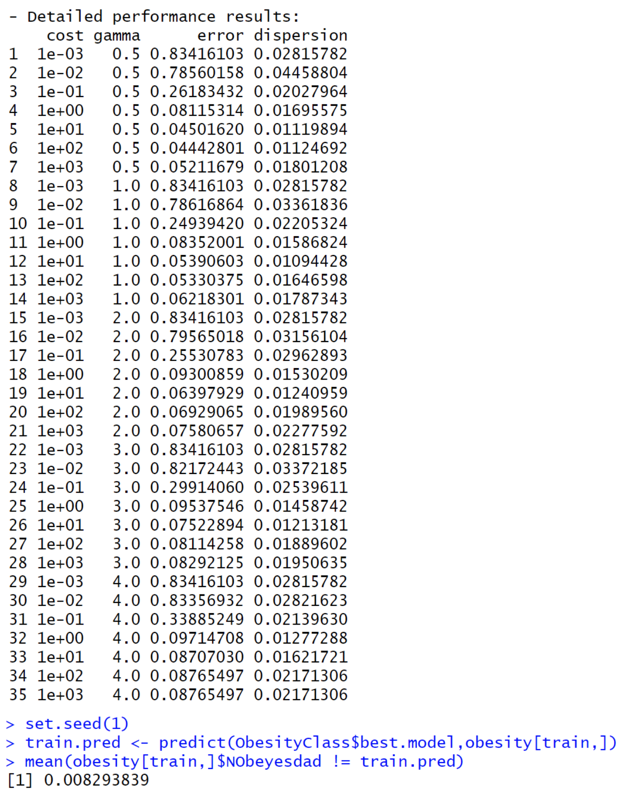
```
> set.seed(1)
> obesity <- (ObesityDataSet_raw_and_data_sinthetic[,c("Height","Weight","NCP","FAF","NObeyesdad")])
>
> n <- nrow(obesity)
> train <- sample(1:n, 0.8*n)
>
> x.train <- scale(obesity[train,-5])
> x.test <- scale(obesity[-train,-5],
+                 center = attr(x.train, "scaled:center"),
+                 scale = attr(x.train, "scaled:scale"))
>
> y.train <- obesity[train, "NObeyesdad"]
> y.test <- obesity[-train, "NObeyesdad"]
>
> set.seed(1)
> library(class)
> for(K in c(1,2,3,4,5,6,7,8,9,10)) {
+   set.seed(1)
+   knn.pred <- knn(train = x.train,
+                   test = x.test,
+                   cl = y.train$NObeyesdad,
+                   k=K)
+   print(mean(knn.pred != y.test$NObeyesdad))
+ }
[1] 0.07328605
[1] 0.08747045
[1] 0.07801418
[1] 0.09456265
[1] 0.09219858
[1] 0.1158392
[1] 0.1182033
[1] 0.1229314
[1] 0.1276596
[1] 0.1252955
```

k=1 produced the smallest test error rate, making it the optimal k value.
*We chose to use the validation set approach since our dataset had a large number of observations and Leave-One-Out Cross-Validation (LOOCV) is more computationally demanding.

- SVM (Timothy Lu and Kevin Tran)



```
> library(e1071)
> set.seed(1)
> obesity <- (ObesityDataSet_raw_and_data_sinthetic[,c("Height","Weight","NCP","FAF","NObeyesdad")])
> n <- nrow(obesity)
> train <- sample(1:n, 0.8*n)

> set.seed(1)
> ObesityClass <- tune(svm,
+                      as.factor(NObeyesdad)~., data=obesity[train,],
+                      kernel="radial",
+                      ranges=list(cost=c(0.001,0.01,0.1,1,10,100,1000),gamma=c(0.5,1,2,3,4)))
> summary(ObesityClass)

Parameter tuning of ‘svm’:

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
  100   0.5

- best performance: 0.04442801
```



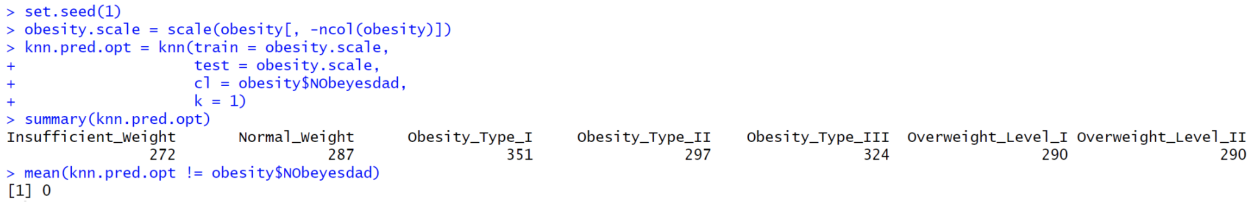
```
- Detailed performance results:
    cost gamma      error dispersion
1  1e-03   0.5 0.83416103 0.02815782
2  1e-02   0.5 0.78560158 0.04458804
3  1e-01   0.5 0.26183432 0.02027964
4  1e+00   0.5 0.08115314 0.01695575
5  1e+01   0.5 0.04501620 0.01119894
6  1e+02   0.5 0.04442801 0.01124692
7  1e+03   0.5 0.05211679 0.01801208
8  1e-03   1.0 0.83416103 0.02815782
9  1e-02   1.0 0.78616864 0.03361836
10 1e-01   1.0 0.24939420 0.02205324
11 1e+00   1.0 0.08352001 0.01586824
12 1e+01   1.0 0.05390603 0.01094428
13 1e+02   1.0 0.05330375 0.01646598
14 1e+03   1.0 0.06218301 0.01787343
15 1e-03   2.0 0.83416103 0.02815782
16 1e-02   2.0 0.79565018 0.03156104
17 1e-01   2.0 0.25530783 0.02962893
18 1e+00   2.0 0.09300859 0.01530209
19 1e+01   2.0 0.06397929 0.01240959
20 1e+02   2.0 0.06929065 0.01989560
21 1e+03   2.0 0.07580657 0.02277592
22 1e-03   3.0 0.83416103 0.02815782
23 1e-02   3.0 0.82172443 0.03372185
24 1e-01   3.0 0.29914060 0.02539611
25 1e+00   3.0 0.09537546 0.01458742
26 1e+01   3.0 0.07522894 0.01213181
27 1e+02   3.0 0.08114258 0.01889602
28 1e+03   3.0 0.08292125 0.01950635
29 1e-03   4.0 0.83416103 0.02815782
30 1e-02   4.0 0.83356932 0.02821623
31 1e-01   4.0 0.33885249 0.02139630
32 1e+00   4.0 0.09714708 0.01277288
33 1e+01   4.0 0.08707030 0.01621721
34 1e+02   4.0 0.08765497 0.02171306
35 1e+03   4.0 0.08765497 0.02171306

> set.seed(1)
> train.pred <- predict(ObesityClass$best.model,obesity[train,])
> mean(obesity[train,]$NObeyesdad != train.pred)
[1] 0.008293839
```

Optimal tuning parameters are cost = 100 and gamma = 0.5.

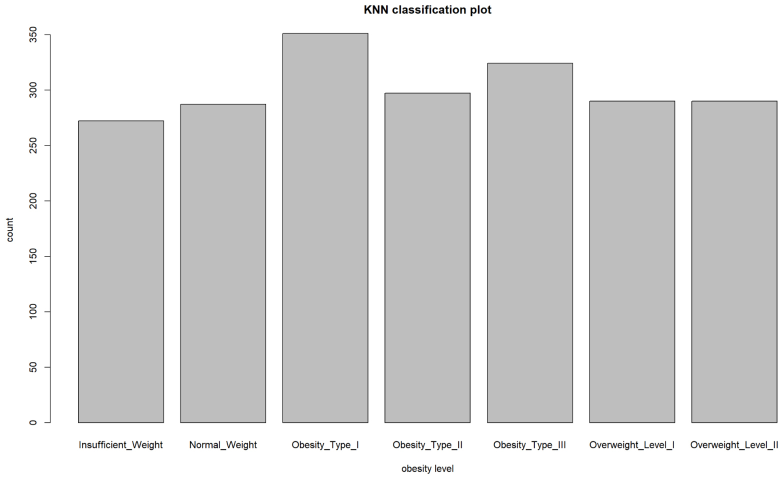
(c) In general, the SVM method produced a smaller test error rate when comparing both supervised learning techniques. KNN using k=1 produced a test error rate of 0.07328605 whereas SVM using cost = 100 and gamma = 0.5 produced a test error rate of 0.008293839.

(d)

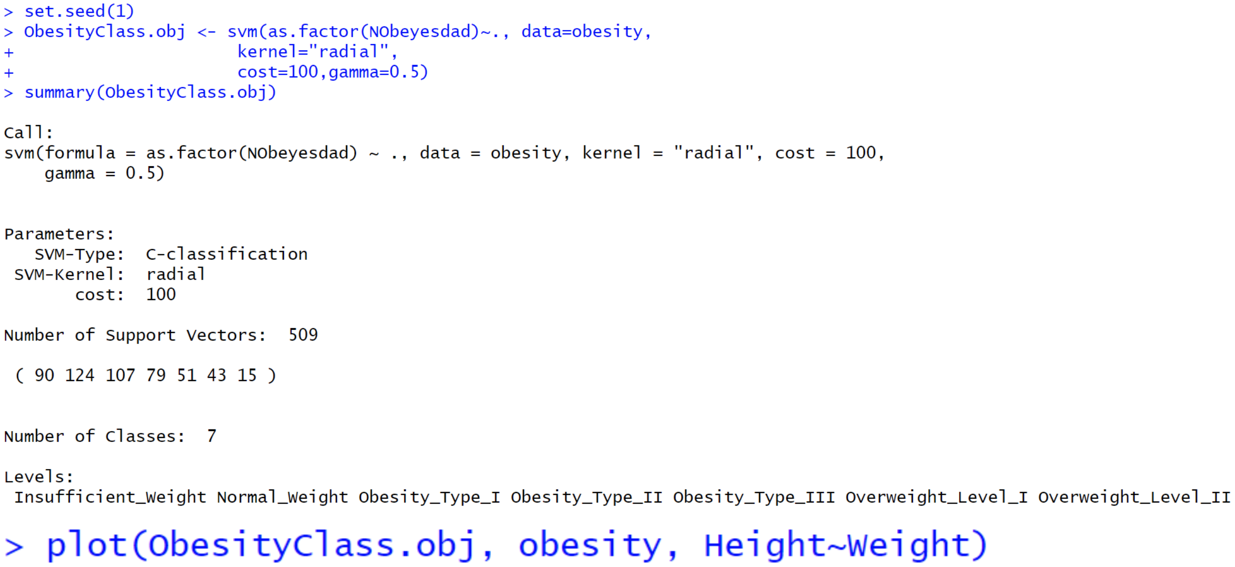
- KNN (Kelsey Wong and Nhi Ho)



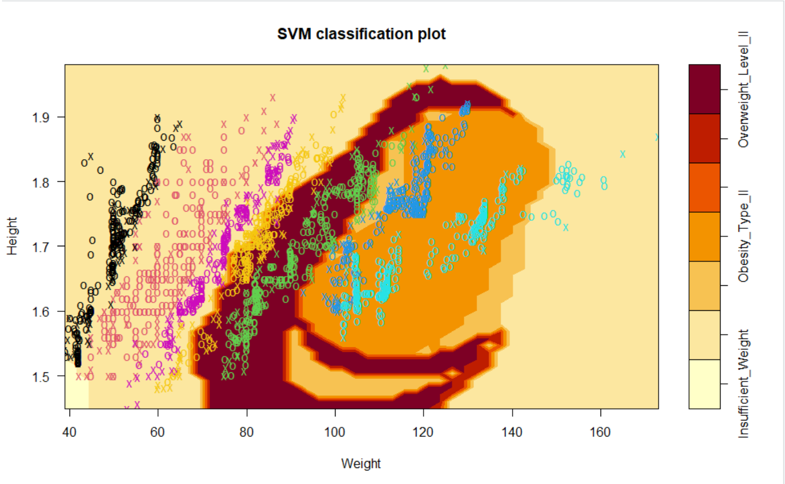
```
> set.seed(1)
> obesity.scale = scale(obesity[, -ncol(obesity)])
> knn.pred.opt = knn(train = obesity.scale,
+                    test = obesity.scale,
+                    cl = obesity$NObeyesdad,
+                    k = 1)
> summary(knn.pred.opt)
Insufficient_Weight       Normal_Weight      Obesity_Type_I     Obesity_Type_II    Obesity_Type_III  Overweight_Level_I Overweight_Level_II 
                272                 287                 351                 297                 324                 290                 290 
> mean(knn.pred.opt != obesity$NObeyesdad)
[1] 0
```

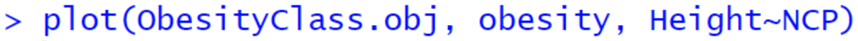


- SVM (Timothy Lu and Kevin Tran)

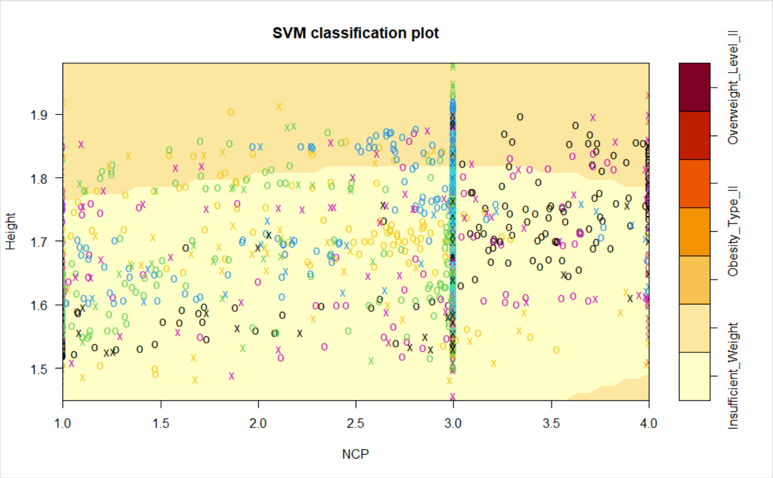


```
> set.seed(1)
> ObesityClass.obj <- svm(as.factor(NObeyesdad)~., data=obesity,
+                      kernel="radial",
+                      cost=100,gamma=0.5)
> summary(ObesityClass.obj)

Call:
svm(formula = as.factor(NObeyesdad) ~ ., data = obesity, kernel = "radial", cost = 100,
    gamma = 0.5)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  100

Number of Support Vectors:  509

 ( 90 124 107 79 51 43 15 )


Number of Classes:  7

Levels:
 Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweight_Level_I Overweight_Level_II
> plot(ObesityClass.obj, obesity, Height~Weight)
```

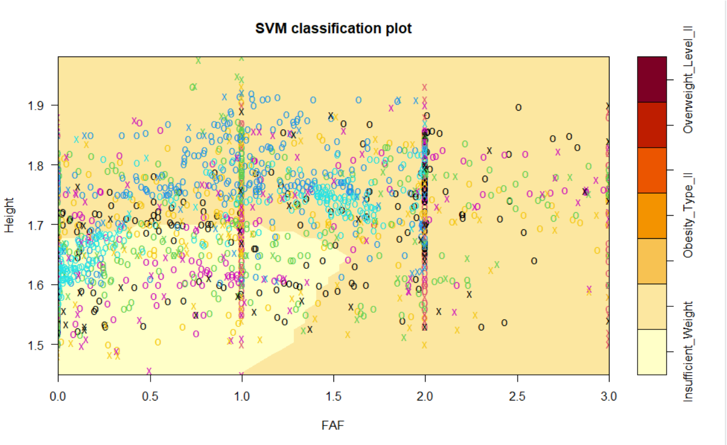




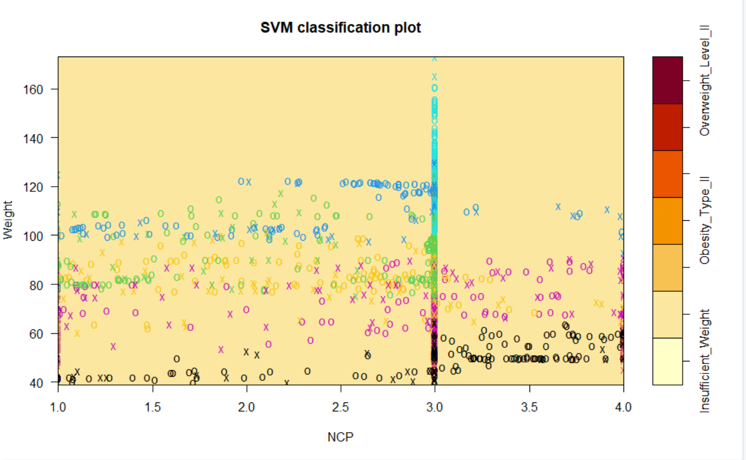
```
> plot(ObesityClass.obj, obesity, Height~NCP)
```



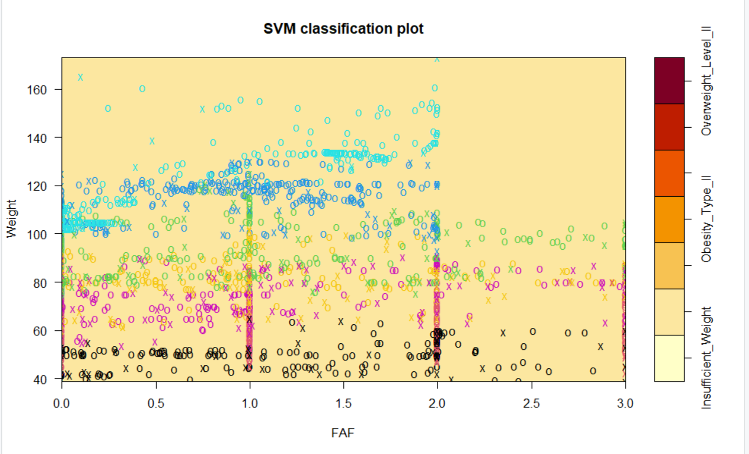
```
> plot(ObesityClass.obj, obesity, Height~FAF)
```



```
> plot(ObesityClass.obj, obesity, Weight~NCP)
```



```
> plot(ObesityClass.obj, obesity, Weight~FAF)
```



(e) Interpretation

- KNN (Kelsey Wong and Nhi Ho): From the report for KNN, the optimal k value is 1. This produced a model that is accurate in predicting the subject's weight level with a low training error rate and a low test error rate. After fitting the optimal value K = 1 to the whole data set, we get an error of 0%. This indicates that the chosen predictors and K = 1 yield a KNN model that predicted 100% accuracy for this data set. However, achieving 100% accuracy is not what data scientists are always looking for. In other words, it means that our model fits the training data very closely, which may result in overfitting. Therefore, if our model was given a different data set, it may not be able to perform as well as we saw it do with this particular one. In addition to that, KNN is sensitive to variable selections. Hence, if we choose other predictors instead of the ones that we did, K = 1 might not be the optimal value for our model anymore.
- SVM (Timothy Lu and Kevin Tran): From the report for SVM, the optimal cost and gamma values are 100 and .5 respectively, leading to the best compromise between bias and accuracy of our predictions. These two values yield the best model with a test error rate of 0.008293839 and a cross-validation error rate of 0.04442801. Seeing as there was a non-linear decision boundary, we utilized the radial kernel, rather than the linear kernel. This model is accurate in predicting the subject's weight level with a low test error rate, which is a respectable performance.

4. Conclusion

- Based on our findings, we were able to accurately predict a person's obesity level using the "Height", "Weight", "NCP", and "FAF" predictors (see appendix). The predictors we chose did indeed help to predict whether a person has obesity or not. However, it seems that while "Height" and "Weight" were helpful in predicting whether someone was classified as obese or not, "NCP" and "FAF" were not as influential in determining obesity level. This was determined from SVM plotted boundaries with respect to certain pairs of predictors. For example, the plot using the "Height" and "Weight" predictors clearly classified the observations into different classes, while the plot with respect to "Weight" and "NCP" seemed to imply that NCP is not as influential in predicting weight level. Additionally, our choice of predictors may have prompted K=1 to be the most optimal K value for the KNN technique. Using 1 as the K value for KNN makes our model more prone to make false predictions since the model only looks at one single nearest neighbor and causes our model to be more prone to the problem of overfitting. This means that our results may not be generalizable to other datasets. To improve upon our findings, we can retrain our models using different predictors to see how influential they all are in predicting obesity. For example, we can include more predictors in training our models, replace our current predictors with other predictors we think are most influential, or train multiple models using different predictors and compare them to each other.

5. References

http://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

6. Appendix

There are 17 variables in the “Estimation Of Obesity Levels Based On Eating Habits and Physical Condition” dataset:

- Gender: Gender of individual (male or female).
- Age: Age of individual.
- Height: Height of the individual in meters.
- Weight: Weight of the individual in kilograms.
- family_history_with_overweight: Overweight person in the family?
- FAVC: Does the individual often eat high-caloric foods?
- FCVC: Frequency of vegetable consumption in meals
- NCP: Number of main meals in a day
- CAEC: Frequency of food consumption between meals
- SMOKE: Does the individual smoke?
- CH2O: Measurement of daily water intake.
- SCC: Does the individual monitor their daily calorie intake?
- FAF: How many days a week the individual does physical activity
- TUE: How many hours a day the individual uses electronic devices that require little movement
- CALC: How often individual drinks alcohol
- MTRANS: Method of transportation
- NObeyesdad: Predicted obesity levels of the individual based on predictors