Feature Analysis and Predictive Modeling with Obesity Data

Declan Riddell

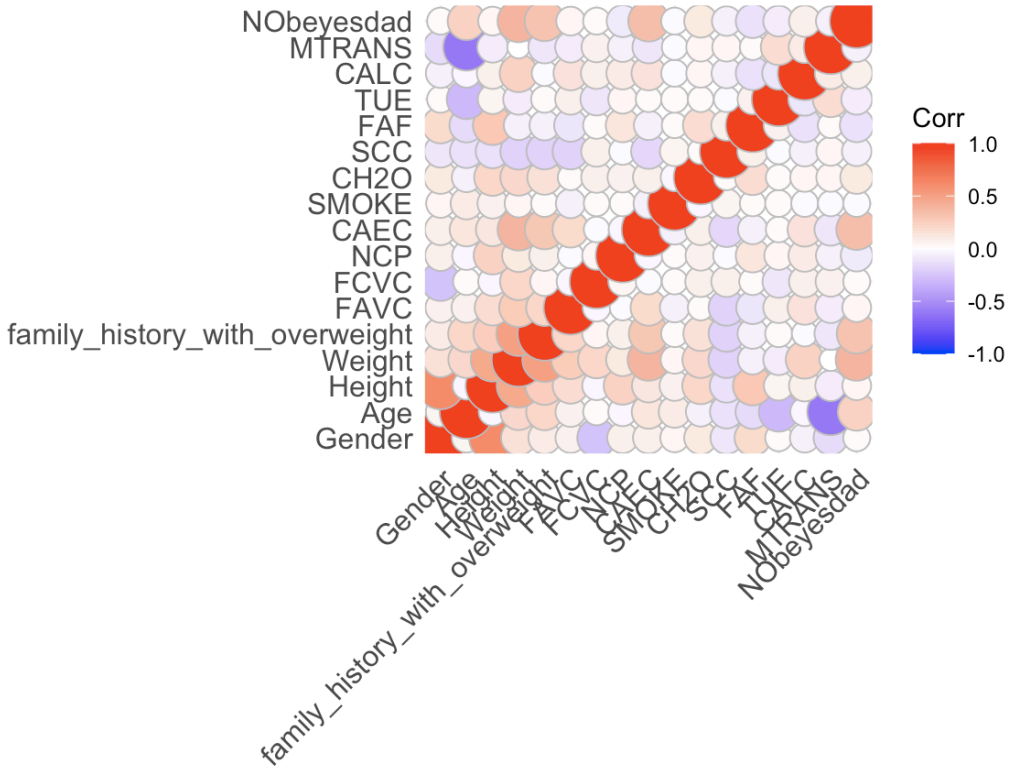
Abstract:

My dataset is a 16 attribute data set with about 2100 rows with no missing values. The data were gathered from an online survey. The dataset provides information on obesity levels with data regarding their diet and physical activity. Contained in the dataset are basic qualifier variables like gender, age, and height. Along with those, there are some interesting fields that I am curious to see if they have a correlation with the level of obesity. Examples would be use of technological devices, smoking, or whether the person monitors their calories on a daily basis. There are some interesting hypotheses you can base off of social norms, for example someone who uses their phone more or plays more video games would be much more likely to be obese based on the implication that they are not very physically active. Or perhaps someone who monitors their calories on a daily basis would be more likely to not be obese, as that type of behavior is usually associated with someone who might be very into lifting weights or maintaining a certain level of physique.

Methods:

The analysis of the dataset was done through the use of the programming language R. The first steps taken in the analysis were to transform the categorical variables such as family_history_with_overweight, and SMOKE(whether the patient smokes cigarettes or not), into numerical variables that were easier to perform different

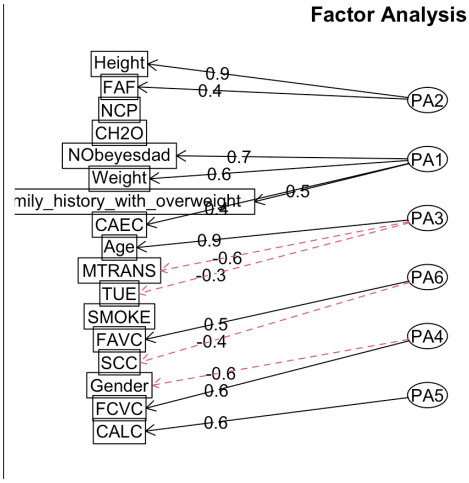
statistical tests on. These variables were coded into numerical values where the number would correspond to a specific categorical value, such as the number 1 for male, or the number 2 for female. This was done utilizing a function I created in R that would take the field which was provided as a string, turn it into a factor, and then into a numeric. There were a total of 9 variables that had to be encoded: FAVC, CAEC, SMOKE, SCC, CALC, MTRANS, Gender, family_history_with_overweight, and the target NObeyesdad. Once these were encoded, I was able to begin the feature analysis of the data set, starting with the correlation plot of the entire dataset.

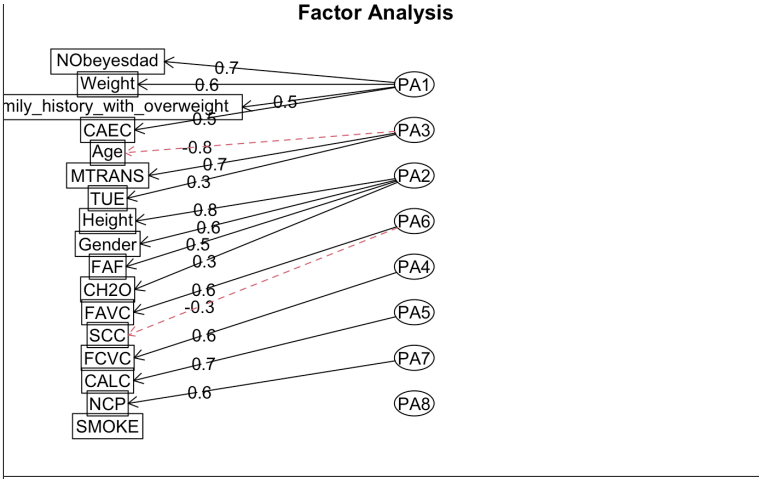


Here you can see that the dataset is pretty highly correlated. This can be confirmed with PCA analysis. Normally, when using PCA the goal is to reduce the dimensionality by forcing a higher amount of variance into a smaller subset of components.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11
Standard deviation     1.6897 1.3948 1.28394 1.15479 1.07979 1.01897 1.00279 0.94309 0.91495 0.89261 0.87280
Proportion of Variance 0.1679 0.1144 0.09697 0.07844 0.06859 0.06108 0.05915 0.05232 0.04924 0.04687 0.04481
Cumulative Proportion  0.1679 0.2824 0.37936 0.45780 0.52639 0.58746 0.64662 0.69893 0.74818 0.79504 0.83985
                          PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     0.80719 0.76758 0.74154 0.62386 0.56945 0.46730
Proportion of Variance 0.03833 0.03466 0.03235 0.02289 0.01908 0.01285
Cumulative Proportion  0.87818 0.91284 0.94519 0.96808 0.98715 1.00000
```

The PCA results for this dataset confirm that the dataset is highly correlated as you can see the proportion of the variance is very small for even the first few components. This data is still useful in determining the level of factors we can use to find the optimum set of features. Generally, you want the smallest amount of factors that still represent the largest amount of variance of the dataset. Because PCA proved ineffective in



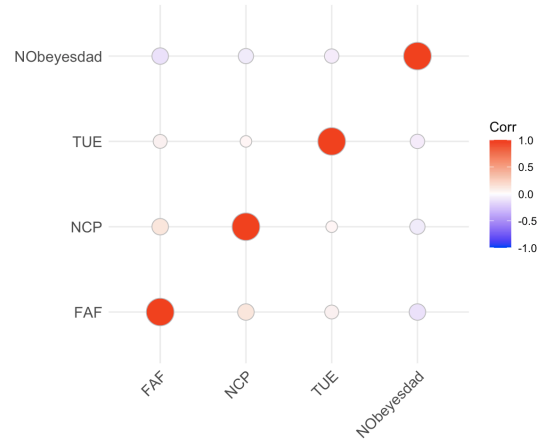


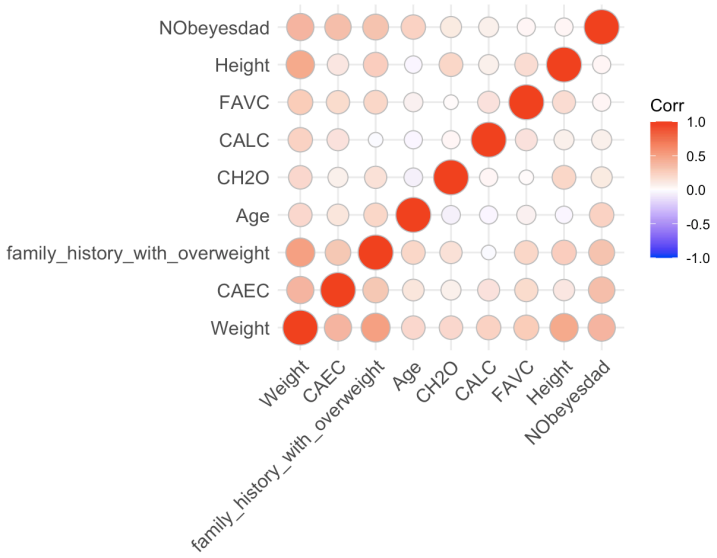
compressing the dimensionality of the data, I chose to go with a number of components based on the proportion of the variance.

Above are my Factor Analysis tests that I conducted. Based on the desire to maintain at about 70% of the variance, I chose to continue with 8 factors, although you can see that the 8th component does not have any significant correlation with the principal components. This could mainly be due to the fact that the dataset itself is quite small at just over 2100 rows.

I chose to use 3 classification techniques in evaluating the dataset for predicting obesity levels: KNN, Random Forrest, and Logistic Regression. For each of these techniques I conducted 4 tests each with different amounts and levels of variable correlation: Full, Highest Correlated Subset(HCS), Lowest Correlated Subset(LCS), Factor Level Subset(FS). The HCS and LCS both consisted of the 3 top variables for their categories respectively. The three highest correlated variables to the target NObeyesdad are Weight, CAEC, and family_history_with_overweight. The three lowest correlated variables are FAF, NCP, and TUE. The next five highest correlated variables, which are included in the FS subset are Age, CH20, CALC, FAVC, and Height.


NObeyesdad
family_history_with_overweight
CAEC
Weight
Weight
CAEC
family_history_with_overweight
NObeyesdad
Corr
1.0
0.5
0.0
-0.5
-1.0

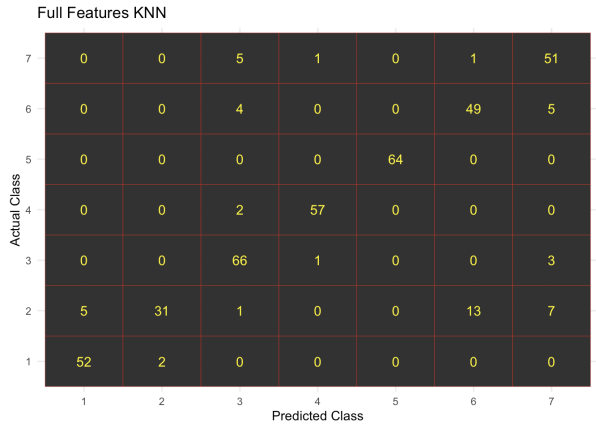

NObeyesdad
TUE
NCP
FAF
FAF
NCP
TUE
NObeyesdad
Corr
1.0
0.5
0.0
-0.5
-1.0


NObeyesdad
Height
FAVC
CALC
CH2O
Age
family_history_with_overweight
CAEC
Weight
Weight
CAEC
family_history_with_overweight
Age
CH2O
CALC
FAVC
Height
NObeyesdad
Corr
1.0
0.5
0.0
-0.5
-1.0

Accuracy %	K-Nearest Neighbor	Random Forest	Logistic Regression
FULL	*88%*	*92%*	*96%*
HCS	*79%*	*48%*	*48%*
LCS	*25%*	*25%*	*25%*
FS	*79%*	*97%*	*97%*
MEAN	*68%*	*66%*	*67%*

You can see that the KNN model had a peak performance with the FULL set of features with an 88% accuracy rate. The RF, and LR models had their peak performances with the FS subset. My hypothesis for why the LCS models all performed equally at 25% would be due to the size of the sample. With the dataset being only 2111 observations, there is not as much for these models to learn from as you would normally desire. For each of the models I created a Confusion Matrix and used the predicted values against the actual values from a train/test split in the dataset. Each confusion matrix for the peak performances of the models are as follows:

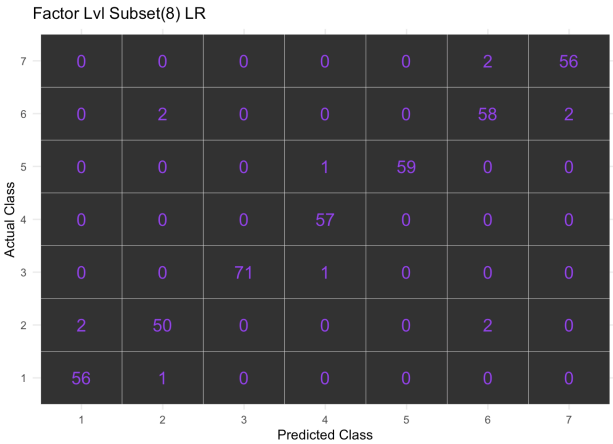
KNN -


Full Features KNN
Actual Class
Predicted Class
7 0 0 5 1 0 1 51
6 0 0 4 0 0 49 5
5 0 0 0 0 64 0 0
4 0 0 2 57 0 0 0
3 0 0 66 1 0 0 3
2 5 31 1 0 0 13 7
1 52 2 0 0 0 0 0
1 2 3 4 5 6 7

RF -


Factor Lvl Subset(8) RF
Actual Class
Predicted Class
7 0 0 0 0 0 2 56
6 0 2 0 0 0 58 2
5 0 0 0 1 59 0 0
4 0 0 0 57 0 0 0
3 0 0 71 1 0 0 0
2 2 50 0 0 0 2 0
1 56 1 0 0 0 0 0
1 2 3 4 5 6 7

LR -


Factor Lvl Subset(8) LR
Actual Class
Predicted Class
7 0 0 0 0 0 2 56
6 0 2 0 0 0 58 2
5 0 0 0 1 59 0 0
4 0 0 0 57 0 0 0
3 0 0 71 1 0 0 0
2 2 50 0 0 0 2 0
1 56 1 0 0 0 0 0
1 2 3 4 5 6 7

Conclusion:

I am skeptical of the RF data because of some of the issues I ran into when creating and running the models. Because of the small sample size, I was unable to generate a desirable number of trees for the model to make predictions, having to go as low as 15 trees in order to get a proper set of predictions. This indicates an area for improvement without too much analysis. The other models would surely benefit greatly from an increased sample size as well.Based on the data I would assert Logistic Regression as the ideal technique to build a predictive model for this dataset. Although the average accuracy is the highest with KNN I have some doubts about the efficacy of the model for this dataset based on the discrepancy between the FULL and FS performance. Because of the high correlation within the dataset, the FS subset becomes more important, as we can confirm with the performance of the other two models, that there is likely some redundancy in the features. I would hypothesize that with more data the mean accuracy of each of the models would increase.

After this analysis, I have come to the conclusion that these models would likely see an increase in performance with a larger sample size. I have also concluded that this dataset does not answer any major questions on its own. After analyzing the dataset and generating a model with a high accuracy, I would say that this could better serve in conjunction with data relating to Cardiovascular Diseases. CVD is the main cause of mortality globally and developing models to help detect early stages of CVD linked with obesity could be valuable.