Predicting Overweight Based on Demographics, Family History, Habits and Physical Condition

October 28, 2020

I. Introduction

Today obesity has become a serious issue around the world. It not only affects a person's health, but also has a negative impact on our society causing billions of direct and indirect health care costs. This study uses a publicly available dataset "Estimation of obesity levels based on eating habits and physical condition Data Set" from the UCI Machine Learning Repository to explore possible factors that may lead to obesity. More specifically, it attemps to predict if a person is overweight or not given that person's demographic information, family history, habits and physical condition. Seven models are built using various machine learning algorithms, including logistic regression, k-nearest neighbor (KNN), and random forest algorithms. The performance of each model is evaluated based on their overall accuracy rate, sensitivity and specificity scores. Among all those models, the Random Forest Model (Model 6) has the best performance and therefore is selected as the final model to help us detect overweight.

II. Dataset and Data Exploration

1. Dataset

The data used in this study is the "Estimation of obesity levels based on eating habits and physical condition Data Set" published by Palechor and de la Hoz Manotas in 2019. It can be found from the UCI Machine Learning Repository here (https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+).

This particular dataset contains records from people that Palechor and de la Hoz Manotas surveyed in Mexico, Peru and Columbia. They collected family history, eating habits and physical condition in addition to surveyees' demographic information. Then they calculated the mass body index (MBI) for each individual using the following formula:

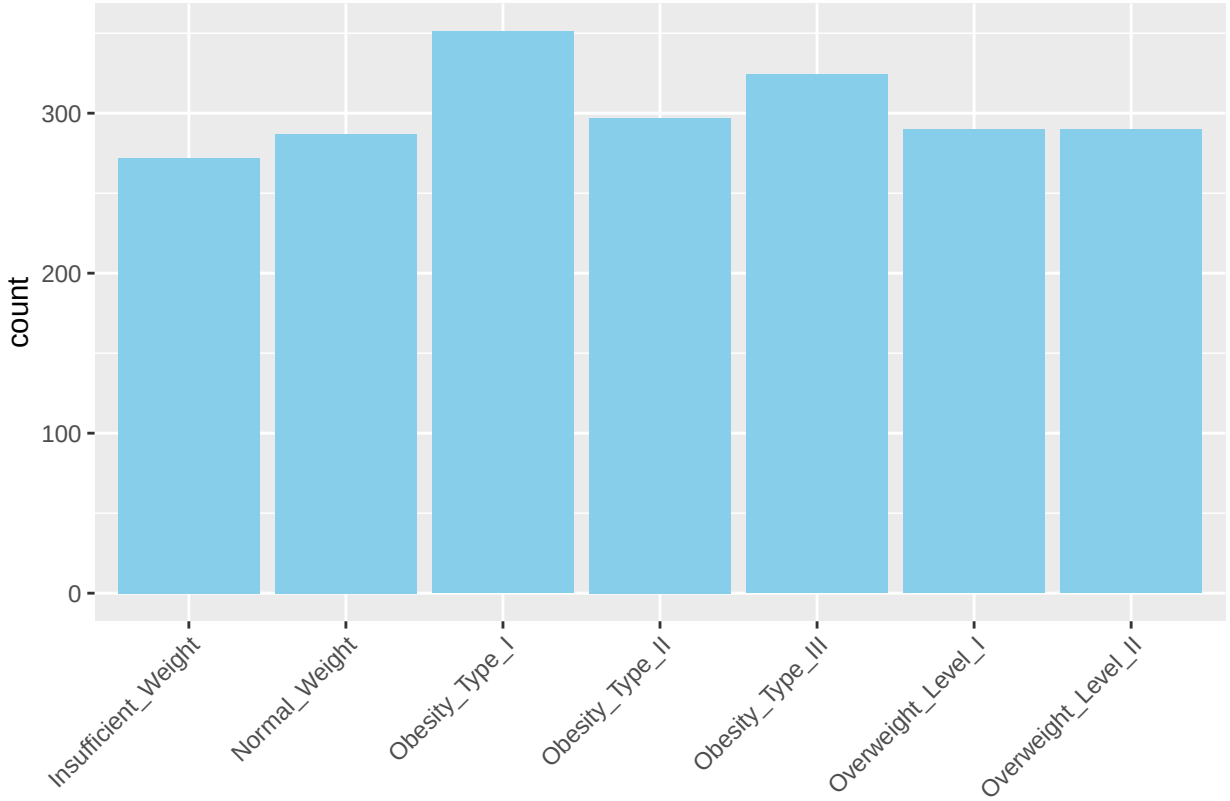
Mass body index = Weight/(Height*Height)

Palechor and de la Hoz Manotas compared the results to the WHO and Mexican data to determine each person's obesity level. Below is the defintion for the obesity level.

Obesity_Level	Definition
Underweight	Less than 18.5
Normal	18.5 to 24.9
Overweight	25.0 to 29.9
Obesity I	30.0 to 34.9
Obesity II	35.0 to 39.9
Obesity III	Higher than 40

Palechor and de la Hoz Manotas noticed that the categories of obesity levels in the survey data were unbalanced, so they generated synthetic data using the tool Weka and the filter SMOTE. The final distribution of each obesity level is shown below:

Figure 1.1 Distribution of Obesity Level



2. Data Exploration

Here is a snapshot of the dataset. We can see it contains 17 attributes, with each record representing a response to the survey (N=2,111).

Figure 1.2 Snapshot of the dataset

```
##    Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1: Female  21   1.62   64.0                            yes   no    2   3
## 2: Female  21   1.52   56.0                            yes   no    3   3
## 3:   Male  23   1.80   77.0                            yes   no    2   3
## 4:   Male  27   1.80   87.0                             no   no    3   3
## 5:   Male  22   1.78   89.8                             no   no    2   1
## 6:   Male  29   1.62   53.0                             no  yes    2   3
##         CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1: Sometimes    no    2  no   0   1         no Public_Transportation
## 2: Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3: Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4: Sometimes    no    2  no   2   0 Frequently               Walking
## 5: Sometimes    no    2  no   0   0  Sometimes Public_Transportation
```

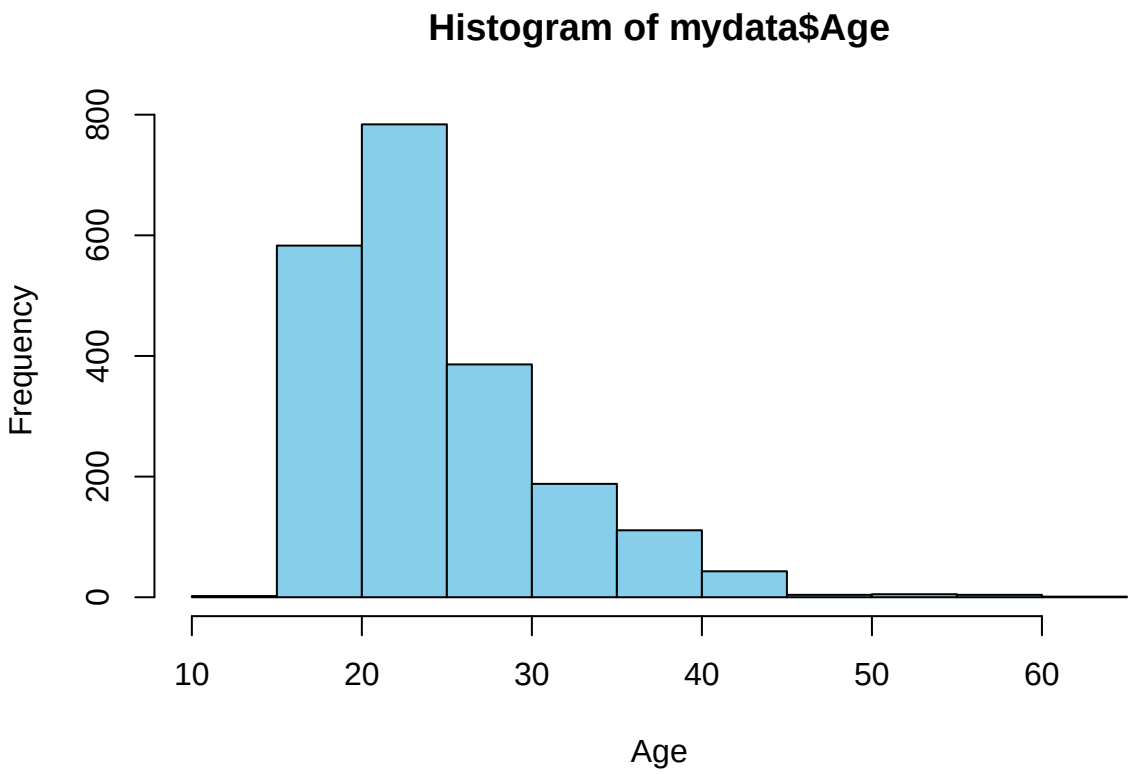
```
## 6: Sometimes    no    2  no   0   0  Sometimes            Automobile
##              NObeyesdad
## 1:       Normal_Weight
## 2:       Normal_Weight
## 3:       Normal_Weight
## 4:  Overweight_Level_I
## 5: Overweight_Level_II
## 6:       Normal_Weight
```

Males and Females are roughly equally distributed in the dataset with male representing about 51% of the population.

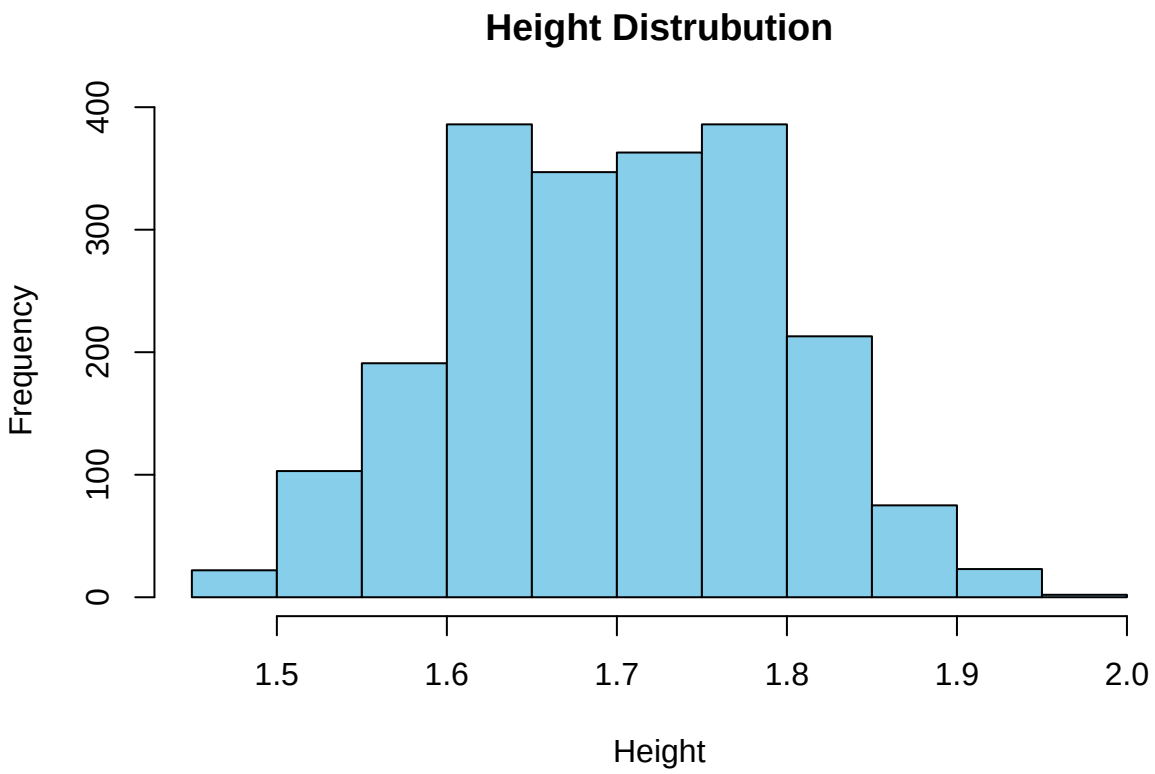
Figure 1.3 Gender Distribution

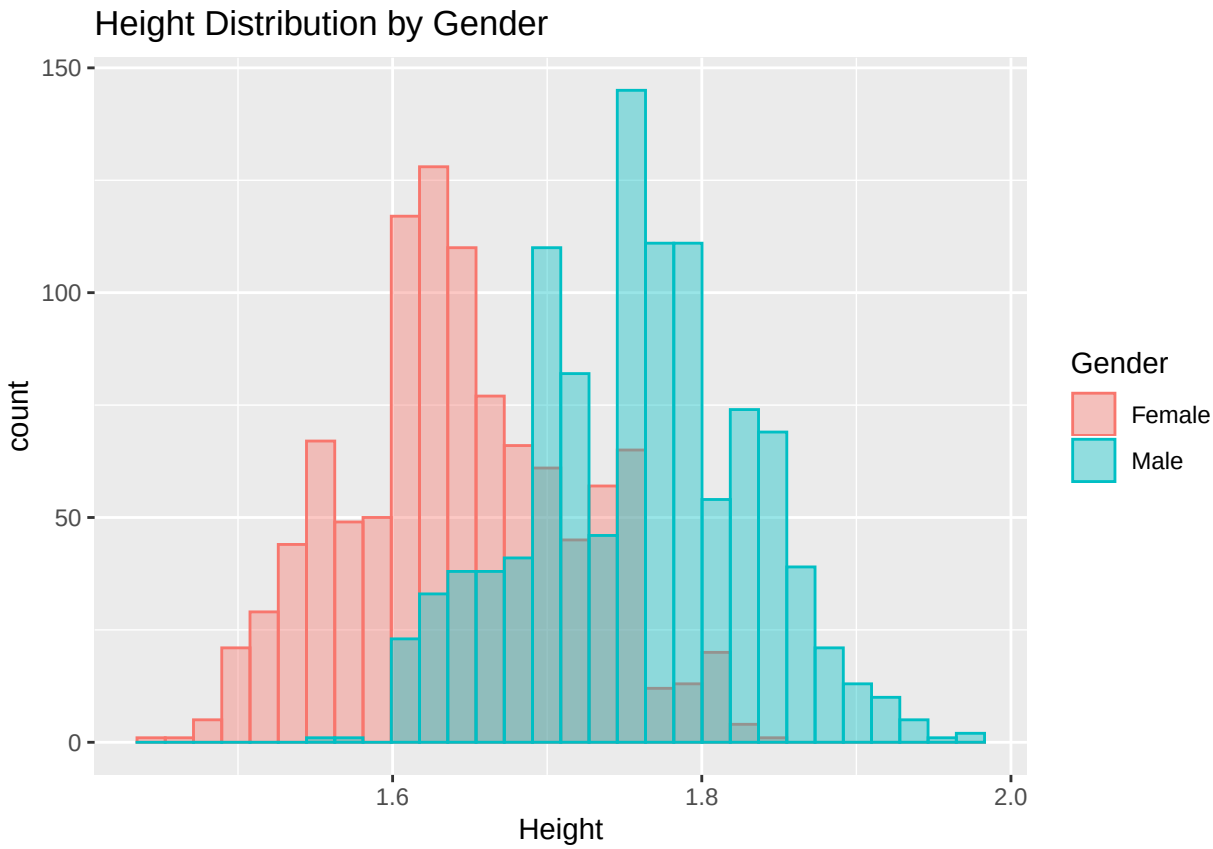
The age distribution of respondents is slightly skewed, with their age ranging from 14 to 61 (as shown in Figure 1.4).

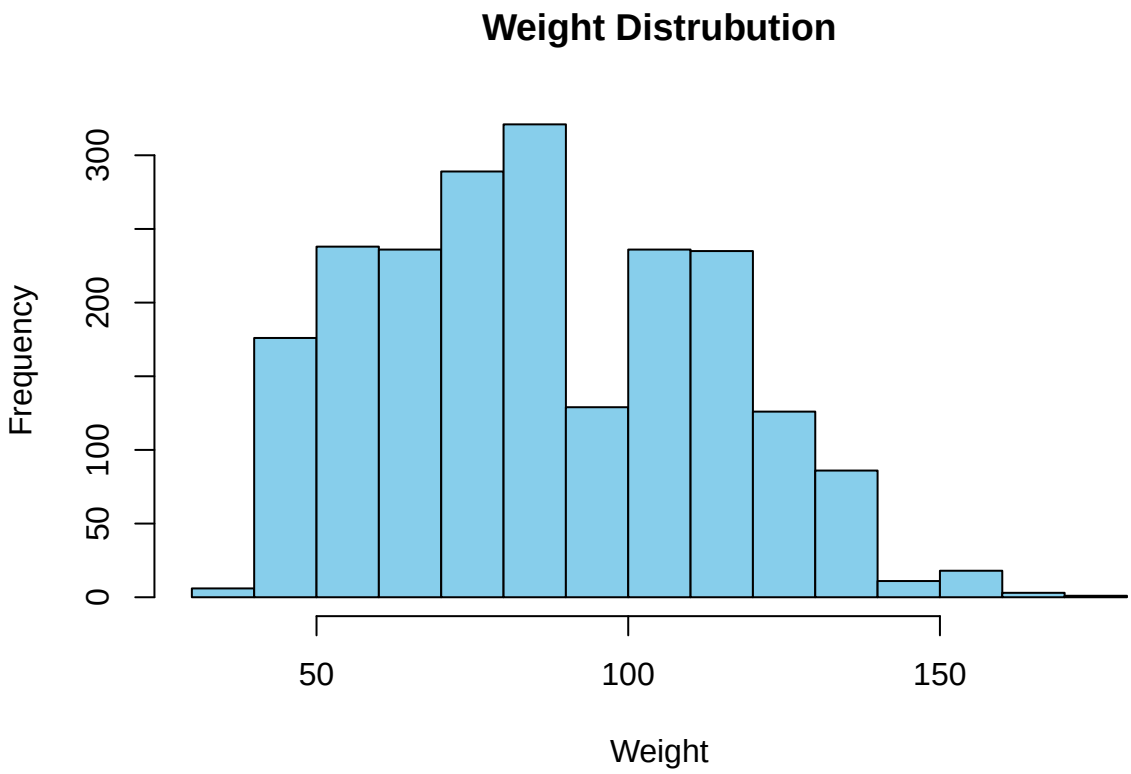
Figure 1.4 Age Distribution

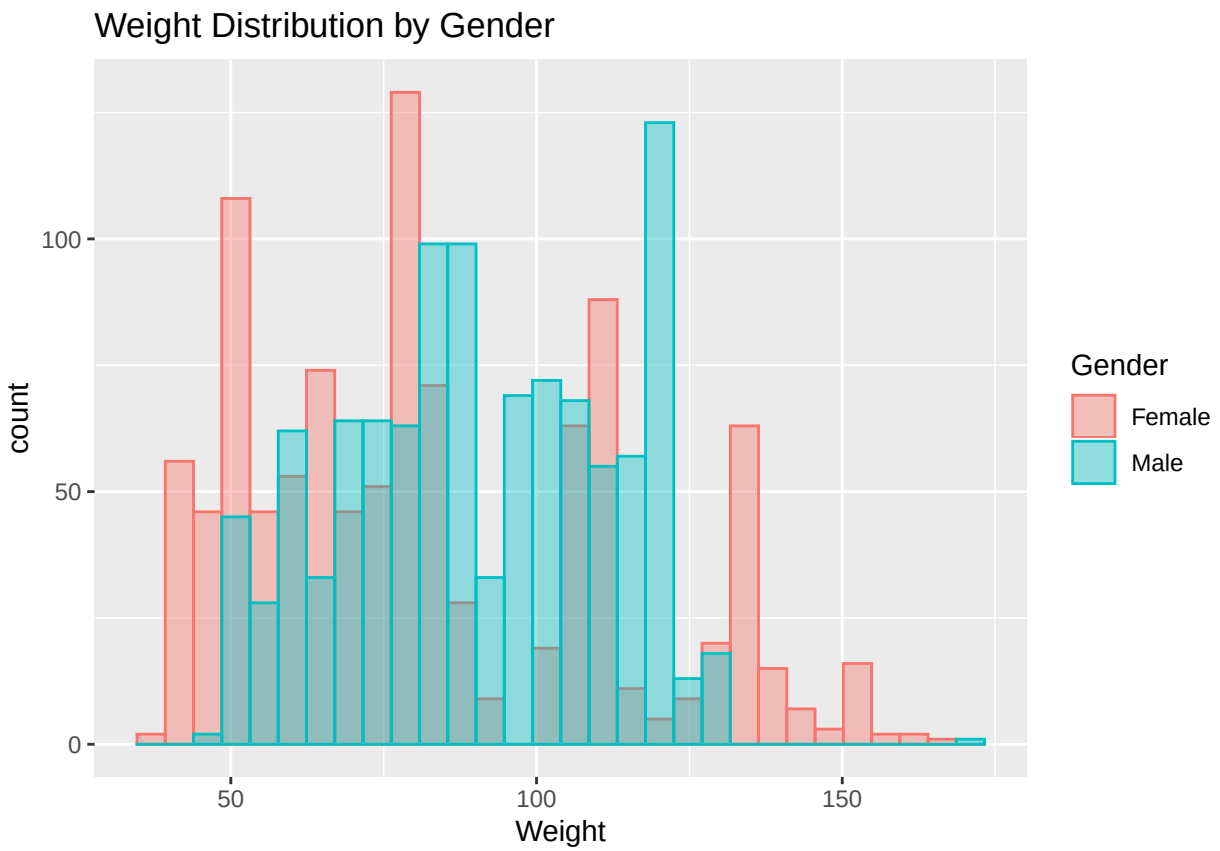


The distributions of height and weight as well as their distributions by gender are shown below. From the figures, we can see gender has an impact on height and weight, which is not very surprising. Generally speaking, males tend to be higher and heavier than females.


Height Distrubution
Frequency
0
100
200
300
400
1.5
1.6
1.7
1.8
1.9
2.0
Height


Height Distribution by Gender
150
100
50
0
count
1.6
1.8
2.0
Height
Gender
Female
Male


Weight Distrubution
Frequency
0
50
100
200
300
50
100
150
Weight



The remaining variables are related to family history, eating habits and physical condition. They are:

Variable	Definition
family_history_with_overweight	Has a family member suffered or suffers from overweight?
FAVC	Frequent consumption of high caloric food
FCVC	Frequency of consumption of vegetables
NCP	Number of main meals
CAEC	Consumption of food between meals
SMOKE	Do you smoke?
CH2O	Consumption of water daily
CALC	Consumption of alcohol
SCC	Calories consumption monitoring
FAF	Physical activity frequency
TUE	Time using technology devices
MTRANS	Transportation used

The majority of those variables are categorical or ordinal variables. To make it easier for model development, they were converted to binary variables (shown below) while continuous variables remained unchanged:

Original_Variable	New_Variable	Definition
Gender	female	1=female
family_history_ with_overweight	familyhistory	1=have a family member who suffered or suffers from overweight
FAVC	FAVC2	1=eat high caloric food frequently

Original_Variable	New_Variable	Definition
FCVC	N/A	N/A
NCP	N/A	N/A
CAEC	CAEC2	1=eat food between meals
SMOKE	smoking	1=smoke
CH2O	N/A	N/A
CALC	alcohol	1=drink alcohol
SCC	SCC2	1=monitor the calories
FAF	N/A	N/A
TUE	N/A	N/A
MTRANS	walkbike	1=use bike or walk

Similarly, the obesity level was also converted to a binary variable to indicate whether a person is overweight or not. The final dataset ("newdata") contains 15 variables. There are no missing values so no extra steps are needed to take care of that. The height and weight variables were excluded from the newdata set because as mentioned previously, the obesity level was determined by MBI, which was calculated directly from weight and height. Therefore, it was decided to remove the two variables to avoid contamination of models.

III. Methods and Analysis

1. Training and Test sets

The entire obesity dataset was randomly partitioned into two sets, the "traindata" set (80% of the whole dataset) and the "finaltest" set (20% of the whole dataset). The traindata set was used to develop models while the finaltest set was used to validate the final model and report its performance.

In addition, the traindata set was further split into two subsets, the train set and the test set, with the latter for model tuning and performance comparison. Again, the train set was created using 80% of the traindata set and the test set 20%.

Various ratios for splitting datasets are used in machine learning, ranging from 70% to 90%. The ratio of 80%/20% is most widely used because it is a common ratio as specified by the Pareto Principle. Therefore, this ratio is chosen in this study.

2. Model Development

Model 1: Baseline Model

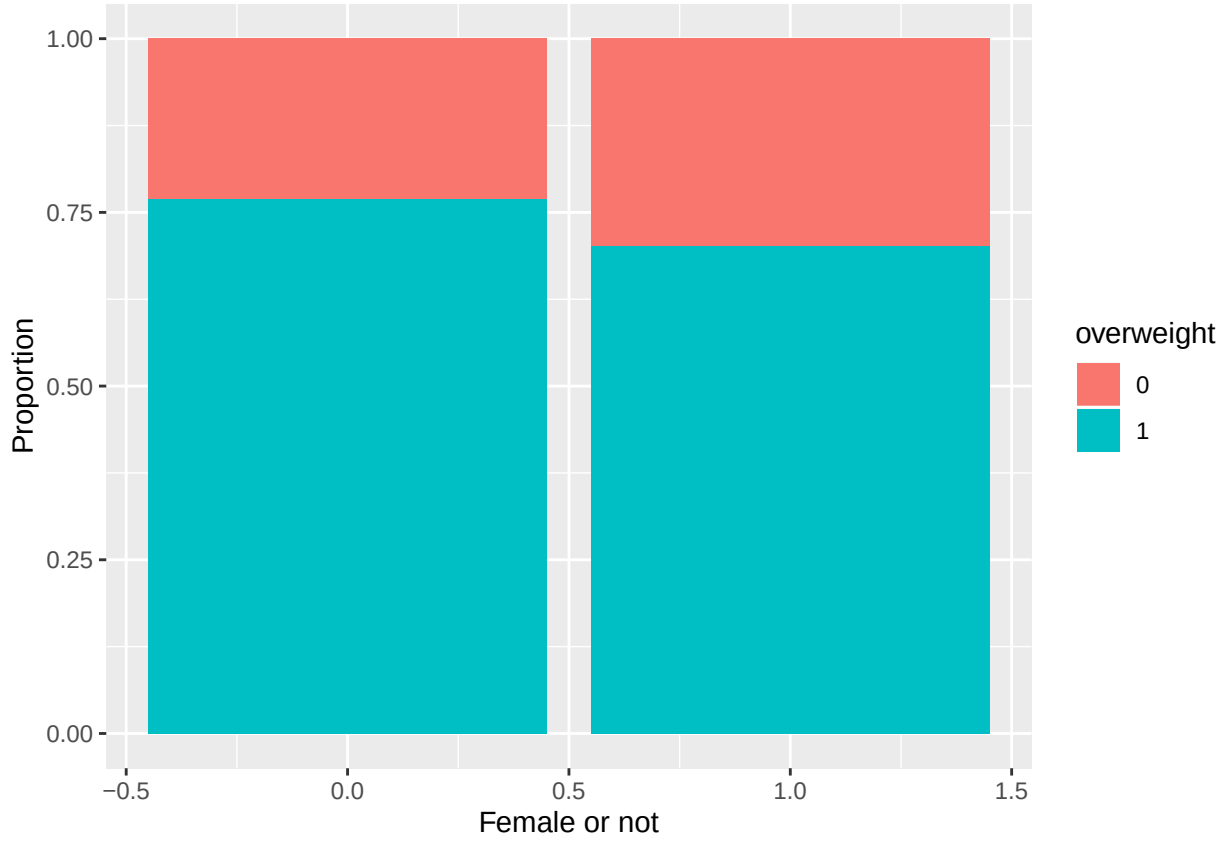
If we examine the train set, we can see that 74% of the records are overweight. The simplest model then is to predict everyone to be overweight, which will give us an accuracy rate of 73.5% on the test set. This serves as the benchmark to assess the performance of other models.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735

Model 2: Logistic Regression with Demographic Data

Obviously the previous model is not very accurate. It is not that helpful to simply predict everyone to be overweight, either. From the data exploration section, we notice the impact of gender on height and weight. So will gender help to improve the model? By plotting the following chart, we can see males are more likely to be overweight than females.

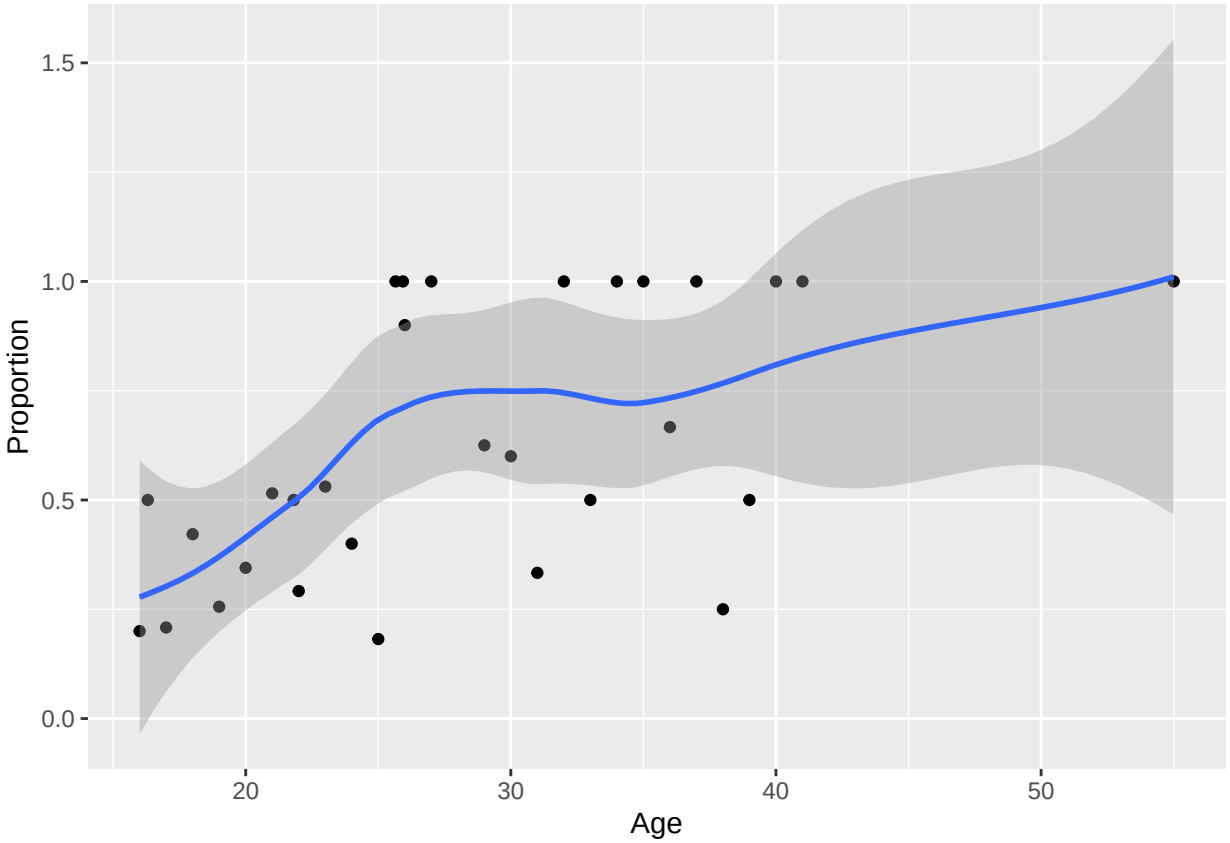
Figure 2.1 Proportion of Overweight by Gender



This trend is also confirmed by the chi-square test ($p<0.01$).

Another possible factor that may contribute to overweight is age. As we grow older, we may not be able to burn calories as effectively as we did when we were young, therefore leading to overweight. We can see this trend in Figure 2.2 below.

Figure 2.2 Proportion of Overweight by Age



Based on the analysis above, both gender and age seem to be good candidates for the second model. When selecting the algorithm, I decided to use logistic regression first. Logistic regression is a machine learing algorithm most commonly used for binary output. It is a good fit here because we want to predict if a person is overweight or not, which is a binary dependent variable.

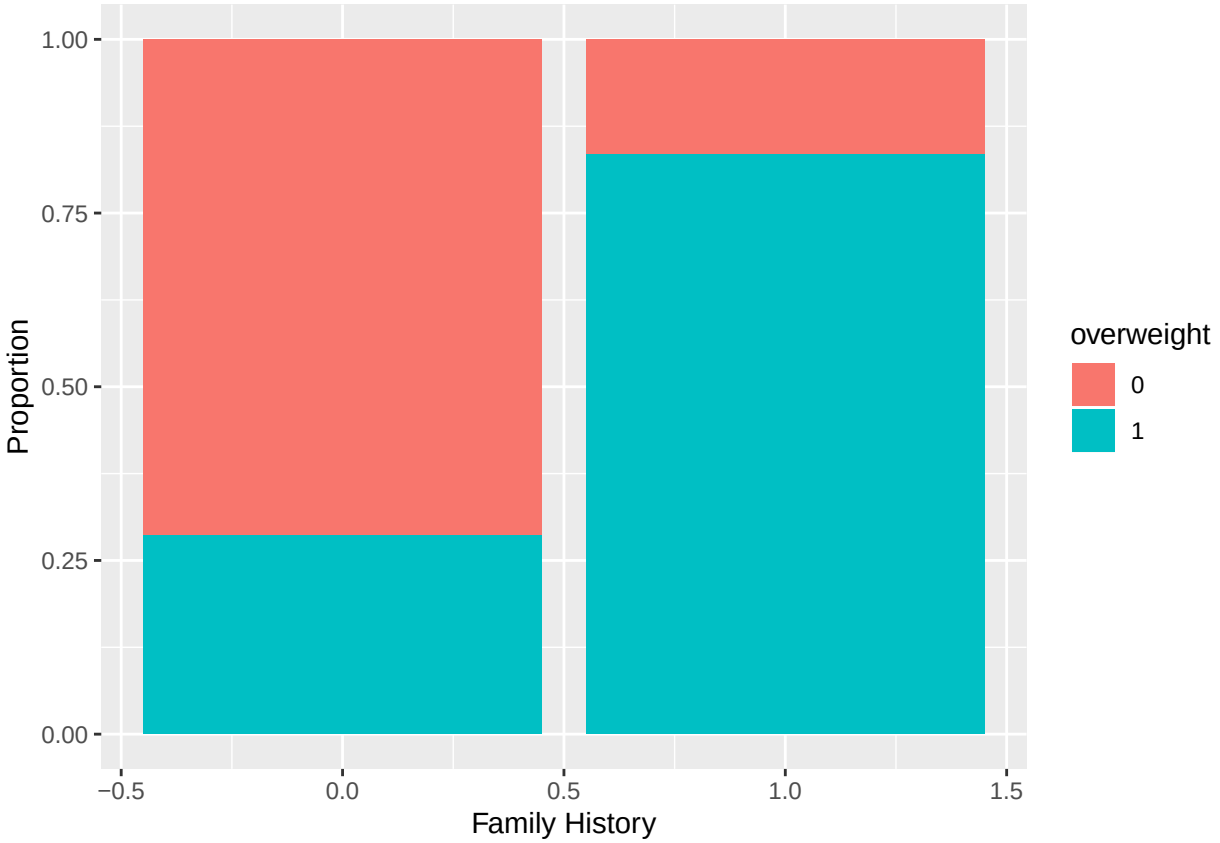
To summarize, the second model is a logistic regression model with gender and age. When it was run on the test set, this model yielded an accuracy rate of 72.3%. This rate is slightly lower than the first model. Apparantely gender and age alone are not as useful as I initally thought.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723

Model 3: Logistic Regression with Demographic Data and Family History

As we know, family history often plays a critical role in a person's health. This has been examined in numerous studies, and its importance is also confirmed in this study. Figure 2.3 clearly tells us people with a family history of overweight are much more likely to suffer the same problem.

Figure 2.3 Proportion of Overweight by Family History



Given the pattern obversed here, the family history variable was added to the third model. As a result, the accuracy rate rose to 83.2%, a pretty big improvement compared to the previous models.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832

Model 4: Logistic Regression with Demographic Data, Family History, Habits and Physical Condition

It is common sense that how we eat, what we eat, and whether we have physical activities are likely to affect our health as well as our weight. Therefore it seems natural to include those variables to help us predict overweight more accurately. Figure 2.4 and 2.5 show a sample variable of habits and physical condition respectively, and their impact on overweight is quite clear here.

Figure 2.4 Proportion of Overweight by Frequent Consumption of High Caloric Food

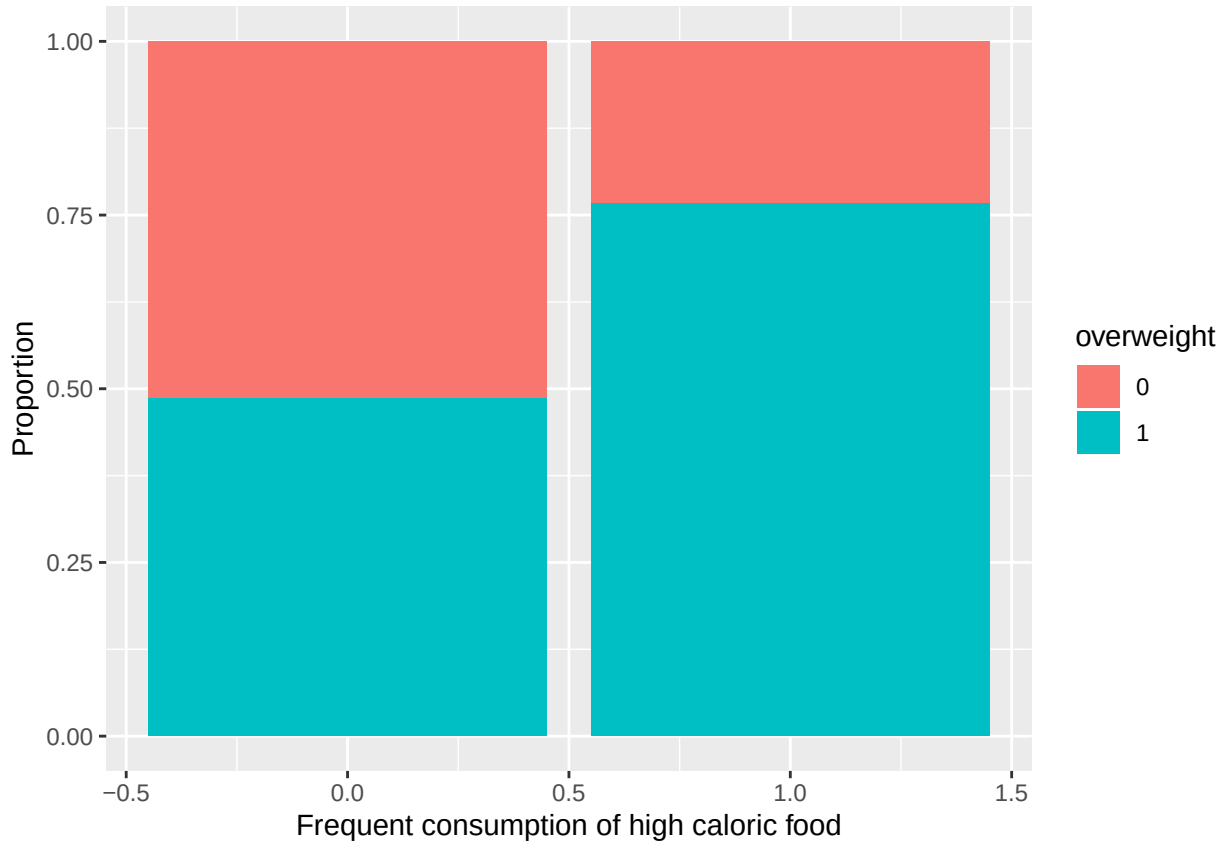
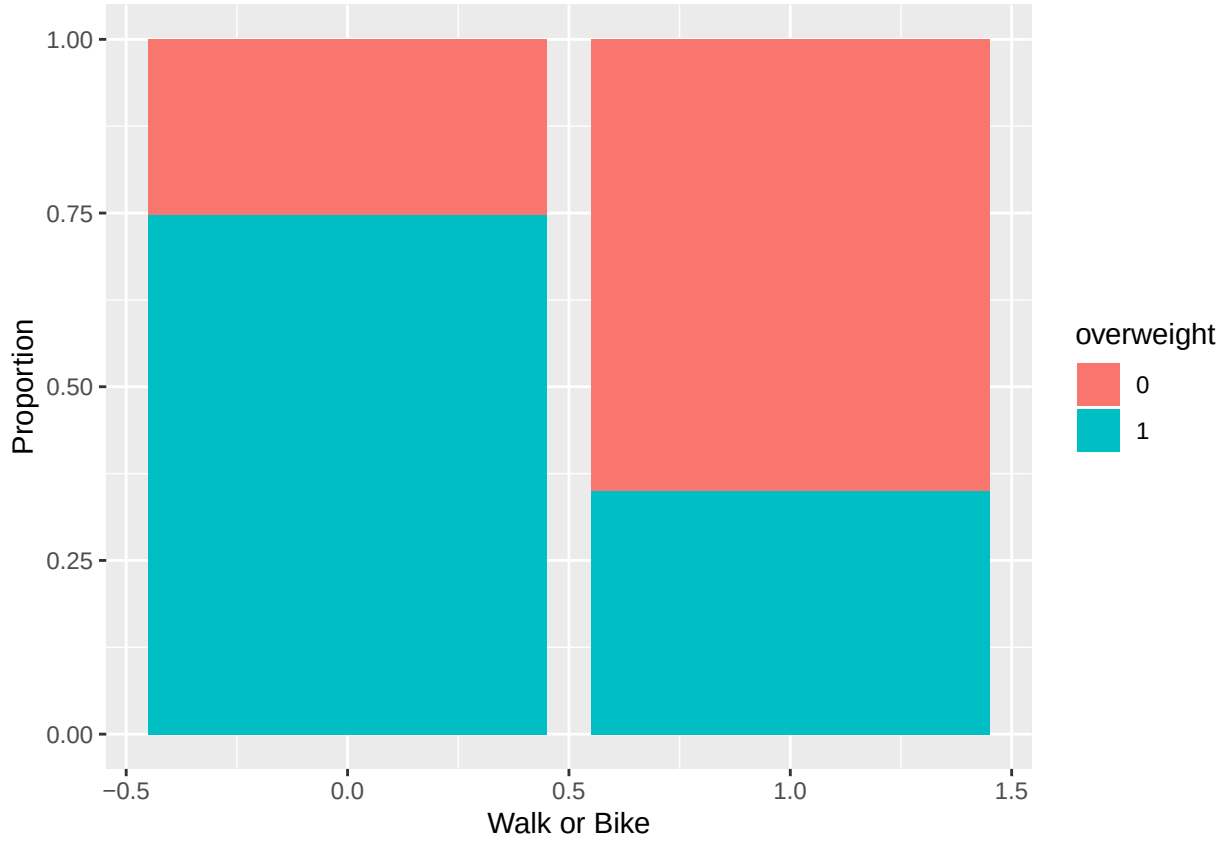


Figure 2.5 Proportion of Overweight by Transportation Used



Before those new variables were added to the model, a correlaton matrix was run to inspect any signifcant correlations among them. As we can see from Figure 2.6, no strong correlations were detected that might raise the alarm.

Figure 2.6 Correlation Matrix

	FAVC2	FCVC	NCP	CAEC2	smoking	CH2O	SCC2	FAF	TUE	alcohol	walkbike
FAVC2	1.000	-0.035	-0.024	0.055	-0.058	-0.016	-0.234	-0.103	0.079	0.140	-0.210
FCVC	-0.035	1.000	0.042	0.105	0.024	0.073	0.086	0.033	-0.085	0.080	0.032
NCP	-0.024	0.042	1.000	0.164	0.012	0.045	-0.006	0.140	0.032	0.075	0.035
CAEC2	0.055	0.105	0.164	1.000	0.025	-0.177	-0.043	-0.034	0.113	-0.057	-0.025
smoking	-0.058	0.024	0.012	0.025	1.000	-0.040	0.033	-0.001	0.020	0.063	-0.026
CH2O	-0.016	0.073	0.045	-0.177	-0.040	1.000	0.008	0.178	-0.005	0.084	0.036
SCC2	-0.234	0.086	-0.006	-0.043	0.033	0.008	1.000	0.107	0.013	-0.012	0.095
FAF	-0.103	0.033	0.140	-0.034	-0.001	0.178	0.107	1.000	0.024	-0.132	0.135
TUE	0.079	-0.085	0.032	0.113	0.020	-0.005	0.013	0.024	1.000	-0.070	0.030
alcohol	0.140	0.080	0.075	-0.057	0.063	0.084	-0.012	-0.132	-0.070	1.000	-0.049
walkbike	-0.210	0.032	0.035	-0.025	-0.026	0.036	0.095	0.135	0.030	-0.049	1.000

Model 4 was run on the test set using variables of demographics, family history, habits and physical condition. It reported an accuracy rate of 83.5%, again an improvement over previous models.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.835

Model 5: KNN Model

The k-nearest neighbor (KNN) algorithm is a non-parametric machine learning algorithm that is often used for classification predictive modeling. It assumes similar things exist in close proximity and estimates the conditional probabilities based on the idea of similarity. The KNN method seems appropriate for our data and therefore is implemented in Model 5.

Based on the analysis of previous models, variables related to demographics, family history, habits and physical condition all contribute to prediction. As a result, all those variables are included in this model.

In KNN modeling a critical step is to decide the number of neighbors (K). There is no standard formula for K. Smaller values can be noisy while large values may increase bias. Usually we find out the best value of K by trial and error. Because we have an even number of classes here, an odd number of K is preferred to avoid a tie. Additionally, literature on KNN suggests to choose the value of sqrt(N) where N stands for the size of the training dataset, which is approximately 37 in this case. Consequently, the model was tuned using K values from 3 to 37 and it turned out a K value of 3 produced the best results. The KNN model was then applied to the test set and yielded an accuracy rate of 86.7%.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.835
Model 5	KNN Model	0.867

Model 6: Random Forest Model

The random forest algorithm is another powerful machine learning algorithm. It consists of individual decision trees that operate as an ensemble. Each individual tree gives a class prediction and the random forest method merges them together to get a more accurate outcome. Because of its simplicity and diversity, random forest is widely used in machine learning applications and it is chosen to be implemented in this model.

A key value in random forest is the number of variables available for splitting at each tree node (mtry). Different values may affect the performance of the model. Again there is no standard formula for this parameter, and it can vary from one dataset to another. A common practice is to try different values and compare the results. On the other hand, some literature suggests to use the square root of the number of predictors, which is about 4 here. Therefore, this particular parameter was tuned up to 7 to allow more

trials. The results indicated that the value 4 gave the best performance. The model was then run on the test set using this parameter and yielded an accuracy rate of 89.7%.

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.835
Model 5	KNN Model	0.867
Model 6	Random Forest Model	0.897

Model 7: Ensemble Model

In Model 6, the random forest algorithm operates as an ensemble. In this model, a customized ensemble is created using results of the previous 3 models (Model 4 through 6).

Here is how it works: the prediction from each model is treated as a vote and the one with the most vote wins and becomes the final prediction. When this model was run on the test set, it yielded an accuracy rate of 88.2%

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.835
Model 5	KNN Model	0.867
Model 6	Random Forest Model	0.897
Model 7	Ensemble Model	0.882

IV. Results

Altogether seven models were built on the train set and their performace on the test set is displayed in the table below.

Table 1. Model Performance by Overall Accuracy Rate

Model	Name	Accuracy_Rate
Model 1	Baseline Model	0.735
Model 2	Logistic Regression with Demographic Data	0.723
Model 3	Logistic Regression with Demographic Data and Family History	0.832
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.835
Model 5	KNN Model	0.867
Model 6	Random Forest Model	0.897
Model 7	Ensemble Model	0.882

The accuracy rate is measured as the proportion of records that are correctly classified as either overweight or not overweight compared to their original value in the test set. According to this measure, Model 6 (Random Forest Model) has the highest accuracy rate, therefore produces the best performance among all the models.

In theory, an ensemble combines decisions from multiple models to improve the overall performance. We can see Model 7 (Ensemble Model) does do a better job than Model 1 through 5. However, it is not as good as Model 6 (Random Forest Model) in this study.

In addition to the overall accuracy rate, sensitivity and specificity are additional measures to evaluate predictive models. Sensitivity is the proportion of observed positives that are predicted to be positive. In this case, it tells us among all the respondents who are overweight from the test set, what percentage each model is able to predict them correctly. Specificity is the proportion of observed negatives that are predicted to be negatives. Again in this case, it indicates how many out of those who are not overweight are classified as such by each model. Ideally we want both to be high, but the two measures are inversely proportional, meaning that as one increases, the other decreases and vice versa. In reality, the threshold to assess each of those two measures often varies on a case-by-case basis as people value each indicator differently under different circumstances. So it is important to select a model that has a good balance between the two.

Table 2 presents both sensitivity and specificity of each model on the test set. While Model 6 (Random Forest Model) does not have the highest sensitivity score, it has the highest specificity score and performs better than other models when both measures are taken into consideration.

Table 2. Model Performance by Sensitivity and Specificity

Model	Name	Sensitivity	Specificity
Model 1	Baseline Model	1.000	0.000
Model 2	Logistic Regression with Demographic Data	0.944	0.111
Model 3	Logistic Regression with Demographic Data and Family History	0.956	0.489
Model 4	Logistic Regression with Demographic Data, Family History, Habits and Physical Condition	0.944	0.533
Model 5	KNN Model	0.936	0.678
Model 6	Random Forest Model	0.932	0.800

Model	Name	Sensitivity	Specificity
Model 7	Ensemble Model	0.952	0.689

Based on the results of overall accuracy, sensitivity and specifity of each model, Model 6 is chosen as the final model for this study. It is trained on the entire "traindata" set, and then applied to the final validation set (the "finaltest" set). The results show an overall accuracy rate of 92.0%, with sensitivity of 96.1% and specificity of 80.4%.

Final_Model	Accuracy_Rate	Sensitivity	Specificity
Random Forest Model	0.92	0.961	0.804

V. Conclusion

Obesity is a global health problem that affects millions of people. It is a major risk factor for diseases like diabetes and stroke. This study examines this particular issue by exploring a public dataset and identifying factors that can help us detect overweight. Seven models are discussed and analyzed in the study, and the Random Forest Model is chosen as the final model based on its accuracy rate, sensitivity and specificity scores. This model is tested on the final test set and is able to predict overweight correctly 92% of the time given a person's demographics, family history, habits and physical condition. This surpasses the results of previous studies mentioned by Palechor and de la Hoz Manotas, which ranged from 73% to 85%. With the help of this model, we can more accurately identify people who are likely to be overweight. We can then help those people to address this problem early on to avoid future medical complications that it may cause.

1. Limitations

One limitation of this study is that the dataset used consists of both actual data (responses collected from a survey) and data generated synthetically using the Weka tool and the SMOTE filter by Palechor and de la Hoz Manotas. The latter is about 77% of the whole set, which could potentially affect both the performance and validity of the final model.

Another limitation of the study is that it only focuses on predicting overweight or not. The original data contains various obesity levels, ranging from overweight to obesity III. Those different levels are not examined separately in this study.

2. Future Work

As mentioned in the Limitations section, different obesity levels are not differentiated in the final model. One possible area for future study is to examine those different levels and explore if individual factors may explain them separately. In addition, if new survey data will be collected in the future, the model can be re-trained on the actual data without synthetic data. As a result, it may improve its performance and potentially provide more accurate predictions.

References

Palechor, F. M., & de la Hoz Manotas, A. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. Data in Brief, 104344.