Final Project Report

Estimation of Obesity Levels Based on Eating Habits and Physical Condition

(Submitted as the final project deliverable for Business Analytics with R)

Submitted by:

Aparna Mishra

Course: BUAN 6356.006

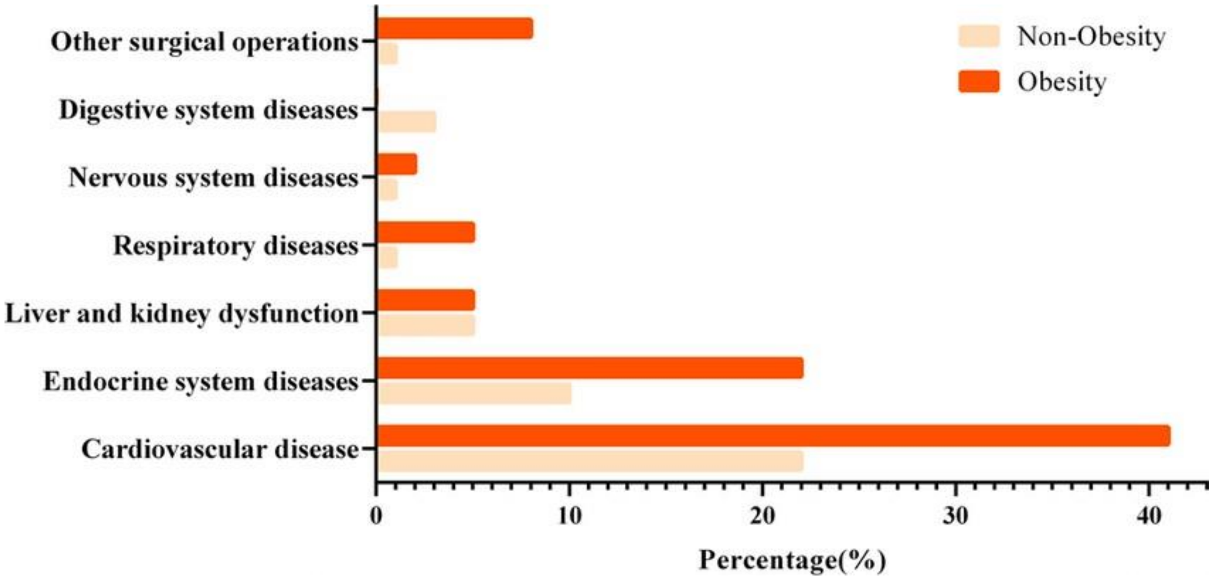
Under the guidance of

Professor Zhe Zhang

Executive Summary:

The global obesity epidemic has tripled since 1975, with over 1.9 billion adults affected, resulting in increased rates of noncommunicable diseases and nearly 5% of global annual deaths. It is crucial to tackle this widespread health concern as it significantly impacts healthcare systems and has a substantial effect on overall well-being.

The graph below demonstrates the comparison of basic diseases between obese and non-obese patients.



Graph Credits: ResearchGate

Our project aims to employ classification techniques to estimate obesity levels in diverse populations, providing insights and enabling targeted interventions. By analyzing various lifestyle factors, we seek to develop predictive models that can estimate an individual's risk of obesity. While acknowledging the dynamic nature of lifestyle trends and genetic influences, this project offers a proactive approach to combating obesity related issues.

In this project we have employed decision tree, random forest, boosted tree, logistic regression, and neural network models. We further evaluated their performance based on accuracy rates and roc curves to get the model that best captures the complexity correlations within the dataset. This model selection process will enhance the reliability and applicability of our obesity estimation tool, ensuring it aligns with the dataset's unique characteristics and contributes meaningfully to our goal of proactive health management.

Motivation/Background:

Obesity has become a worldwide epidemic, affecting a significant proportion of the population on a global scale. Obesity is a predominant risk factor for various chronic diseases like diabetes, cardiovascular disease, strokes, and some cancers. Obesity accounts for nearly 5% of all annual deaths globally. This project analyzed an obesity dataset of 2,111 individuals from Mexico, Peru and Colombia to develop predictive models estimating obesity levels based on factors like diet, lifestyle, demographics etc. By uncovering obesity correlates in this population, targeted interventions can be designed.

The outcomes of this project not only contributes to our understanding of obesity patterns in the studied regions but also offer tangible tools for individuals, healthcare providers and policymakers to take proactive measures. The goal is to empower people to make informed choices, leading to healthier lives and a collective effort to mitigate the global impact of obesity related issues.

Data description:

We have sourced this dataset from UC Irvine ML repository. The dataset contains 3 numerical and 13 categorical attributes. The "NObeyesdad" attribute contains BMI distributed into 7 categories. Following are the attributes:

Attribute	Description
Gender	Sex
Age	Age
Height	Height
Weight	Weight
family_history_with_overweight	Family history with overweight
FAVC	Consume high-calorie foods frequently
FCVC	Number of meals where you usually eat vegetables
NCP	Number of main meals a day
CAEC	Eat food between meals
SMOKE	How often you smoke
CH2O	Liters of water you drink a day
SCC	Monitor the calories you consume daily
FAF	Frequency of days per week that you often have physical activity
TUE	Time of use of technological devices on a daily basis
CALC	Frequency of alcohol intake.
MTRANS	Means of transportation that you use regularly
NObeyesdad	Body mass index

Data set source link:

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

Data Pre-processing and cleaning:

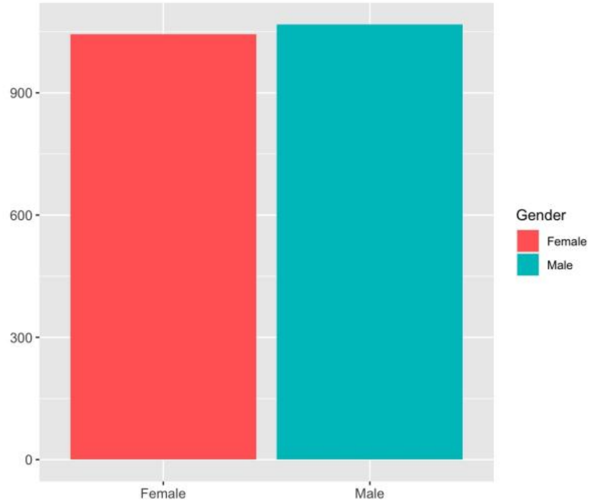
An essential phase in the data mining process is data preparation. It describes the processes of integrating, cleansing, and converting data to prepare it for analysis. Enhancing the quality of the data and tailoring it to the particular data mining task are the objectives of data preprocessing.

Our dataset consisted of binary, ordinal, non-ordinal variables with no missing values. There are 3 numerical and 13 categorical attributes. Among them, 'Gender', 'height', 'age', 'weight', 'family history with being overweight' are traits that an individual doesn't have control over. These attributes aren't lifestyle choices made by the individual. Moreover, 'Height' and 'Weight' can directly be used to calculate BMI. Hence, to have unbiased factors in the analysis, we discarded the first four noncontrollable attributes from being used as variables in our models. Even though 'family history with being overweight' is also a factor not controlled by an individual, genetics or heredity play quite a significant role in a person's chances of becoming overweight or not, and hence has been kept intact as one of the parameters in our analysis.

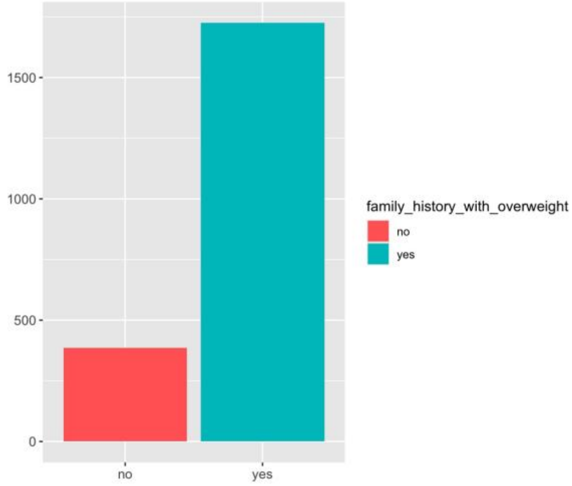
The target variable is NObeyesdad which reflects the BMI of the individual. The original dataset contains 7 categories in the target variable. For the sake of convenience and ease of performing the analysis, we binned them into two categories- 'Overweight' and 'Non-Overweight'. The former category consists of ‘Normal weight’ and ‘Insufficient weight’ categories; whereas the latter consists of ‘Obesity Type I’, ‘Obesity Type II’, ‘Obesity Type III’, ‘Overweight Level I’ and ‘Overweight Level II’.

Exploratory Analysis:

Gender

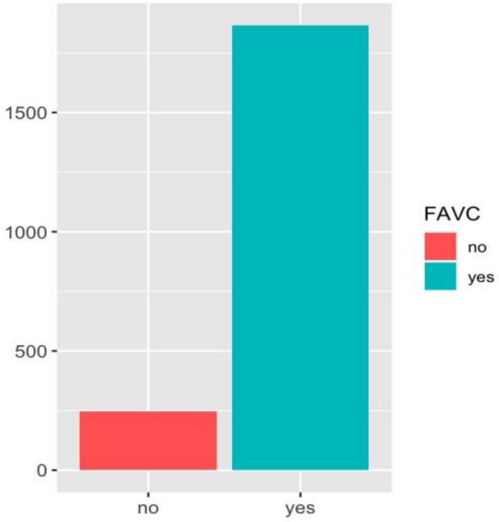


Family history with overweight

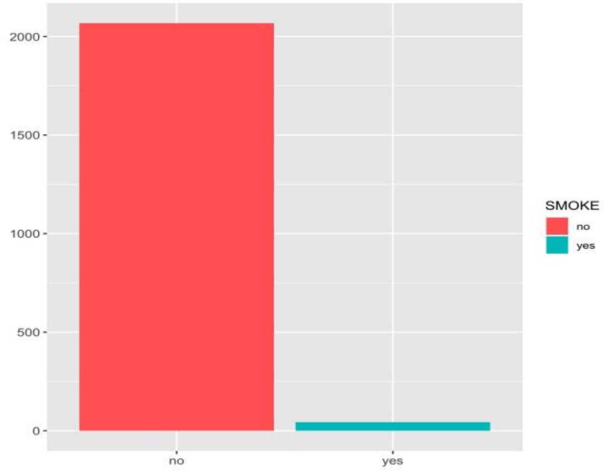


The dataset exhibits gender balance, with almost equal representation of females and males, and 82% individuals with a family history of overweight.

High calories food consumption



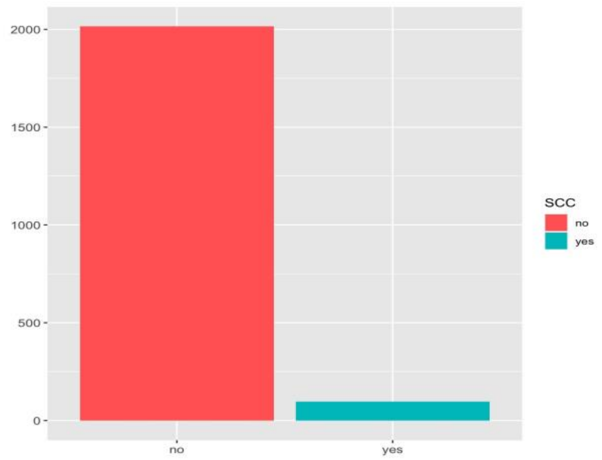
Smoke frequency

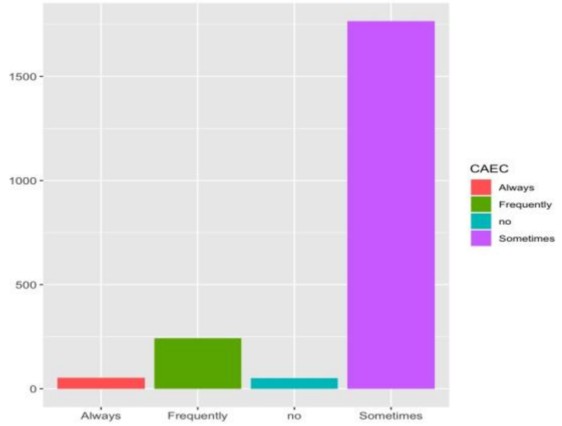


Nearly 88% of individuals consume high-calorie foods frequently and ~98% of individuals don't smoke.

Monitoring the calories consumption daily

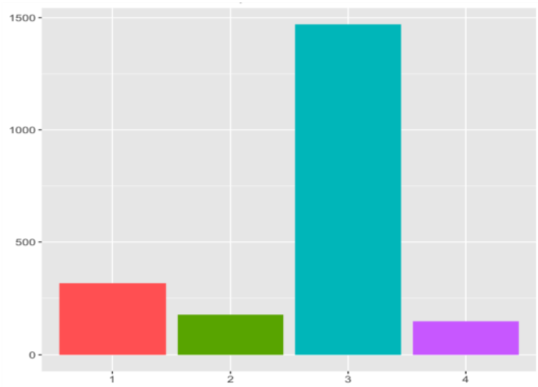
Consumption of food between meals



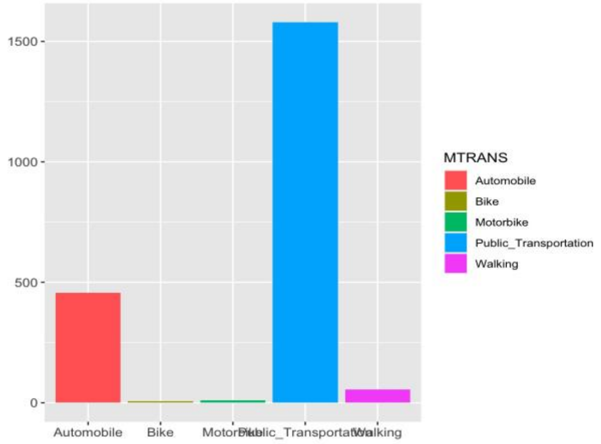


The dataset has ~95% individuals who don't usually monitor the calories consumed daily and ~84% individuals who sometimes eat food between meals.

Number of meals a day

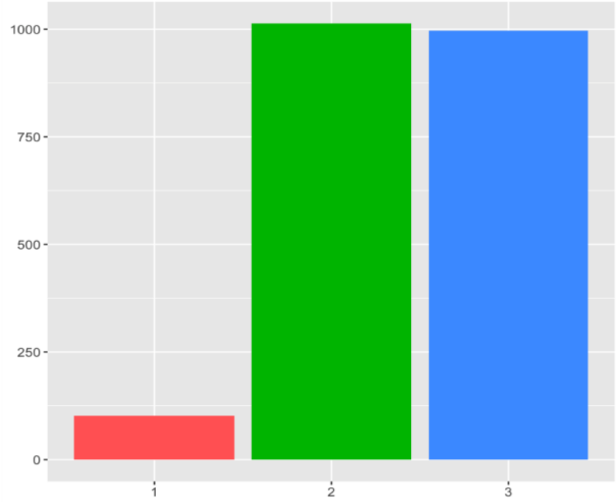


Means of Transportation

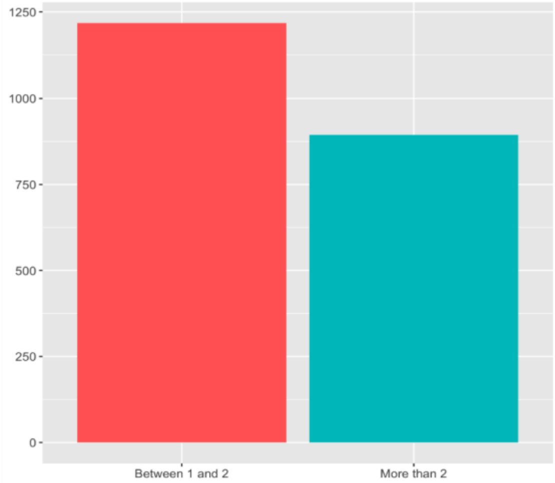


Nearly 50% individuals in this dataset are having three main meals a day and 75% individuals use public transportation.

Number of meals (eats vegetables)

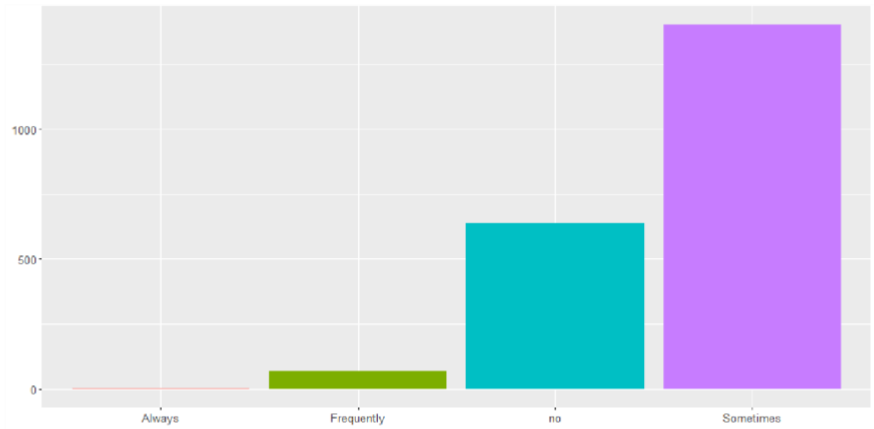


Liters of water per day



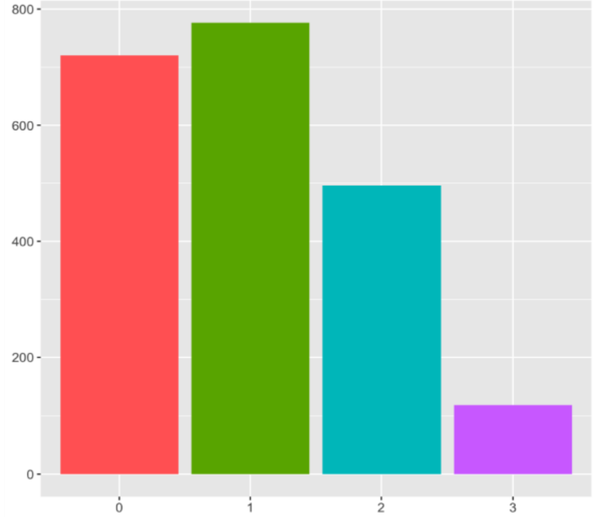
Nearly 95% individuals consume two or more meals that include vegetables and 60% of individuals in this dataset consume around 1-2 liters of water per day, while the remaining individuals consume more than 2 liters.

Frequency of alcohol intake



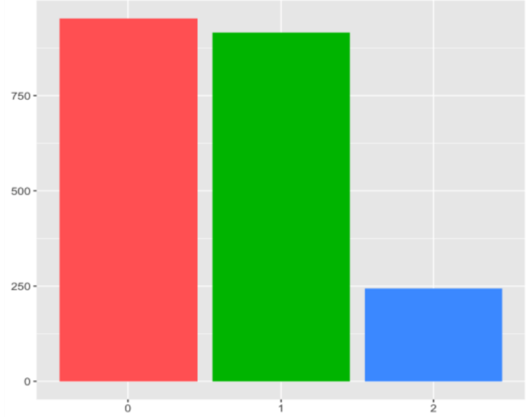
A majority of individuals drink alcohol occasionally or never drink at all and These individuals comprise about ~97% of the dataset.

Frequency of days of physical activity per week



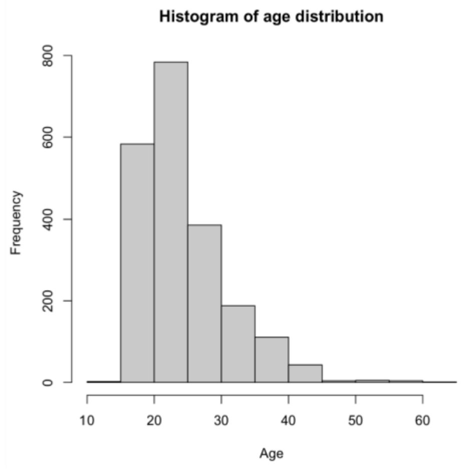
Nearly 94% of the individuals in the dataset engage in physical activity for a maximum of two days per week.

Hours of use of technology devices on a daily basis

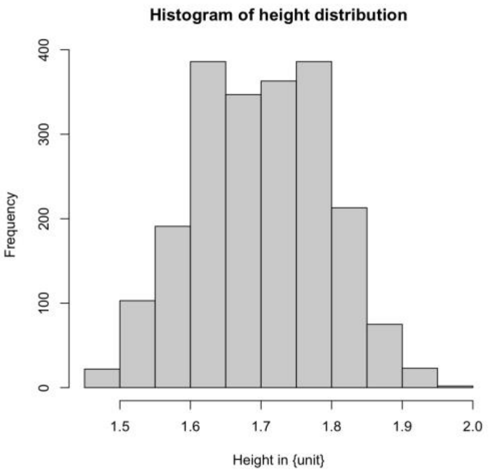


The dataset has ~88% individuals who use 0-1 hours of technology devices on a daily basis.

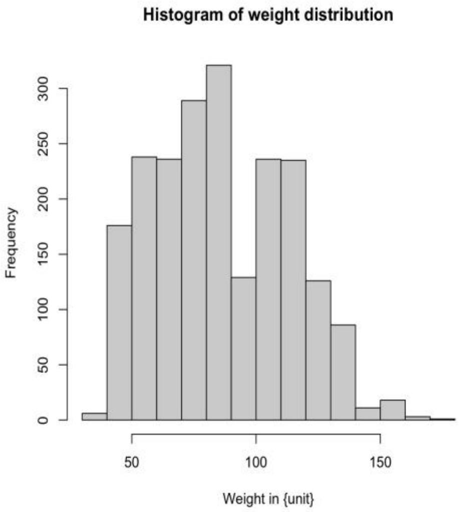
Histograms 1) Age distribution:



2) Height distribution:

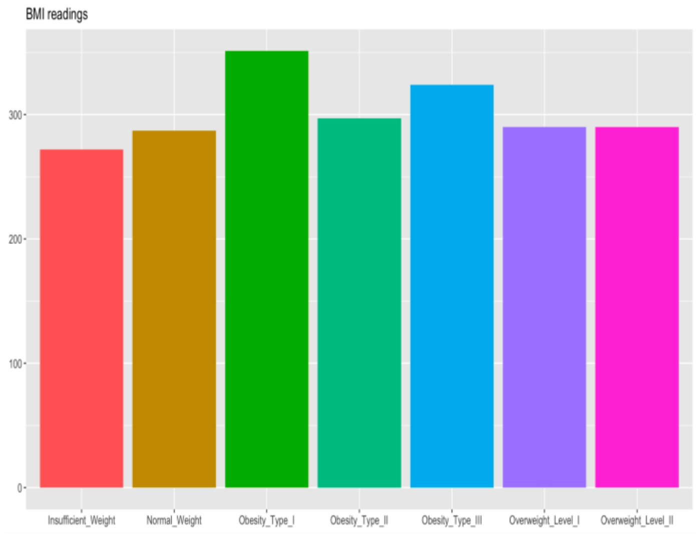


3) weight distribution:

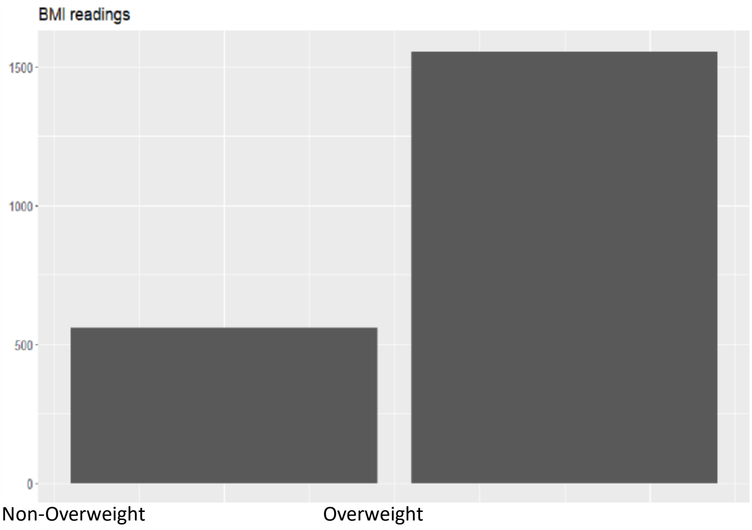


Within the dataset, there are three continuous variables: age, height, and weight:

- Respondents' ages range from 14 to 61, with the majority being relatively young; specifically, 75% of them are 26 years old or younger.
- Height data approximates a normal distribution.
- Weight exhibits a broader range, with an average weight of 87 kilograms.



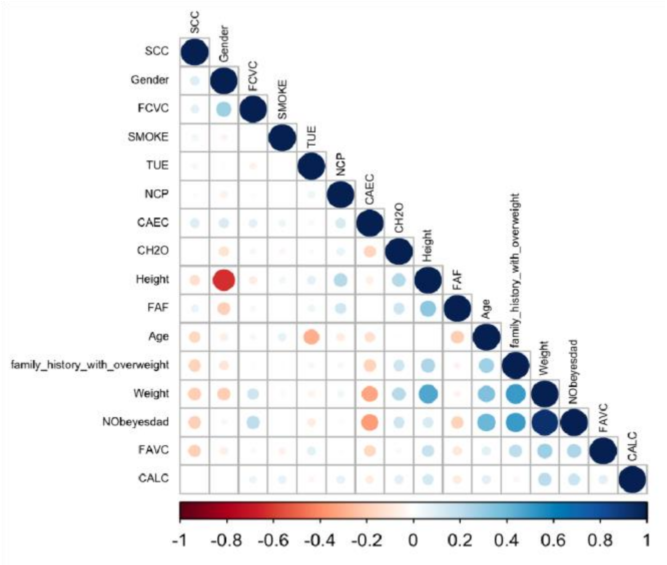
- The dataset is evenly balanced in terms of the BMI level, represented by the variable "NObeyesdad".



- For ease of performing the analysis, we convert the 7 categories into 2 categories- Overweight and Not overweight.
- Upon this conversion, our dataset needs scaling as it is no longer balanced.

Note: ‘Normal Weight’ & ‘Insufficient Weight’ constitute ‘Non-Overweight’ category, and ‘Obesity_Type_I’, ‘Obesity_Type_II’, ‘Obeity_Type_III’, ‘Overweight_Level_I’ and ‘Overweight_Level_II’ constitute ‘Overweight’ category in our analysis

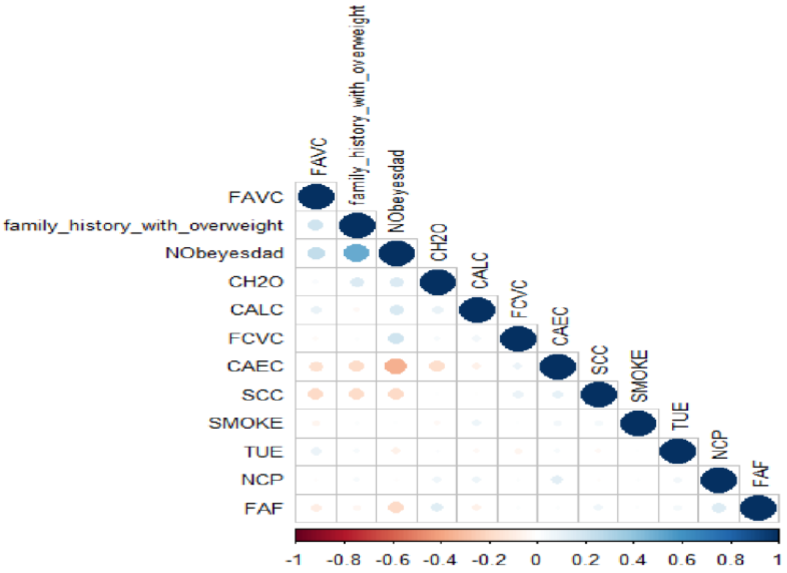
Correlation Matrix :



This is the correlation based on the full dataset. We see high correlation between:

- Height and Gender
- Weight and Height
- Weight and level of BMI (NObeyesdad)
- Family history with overweight and weight ● Family history with weight and BMI (NObeyesdad)

After removing required attributes



Upon removing the required attributes, we see high correlation between: ● NObeyesdad and Family history with weight

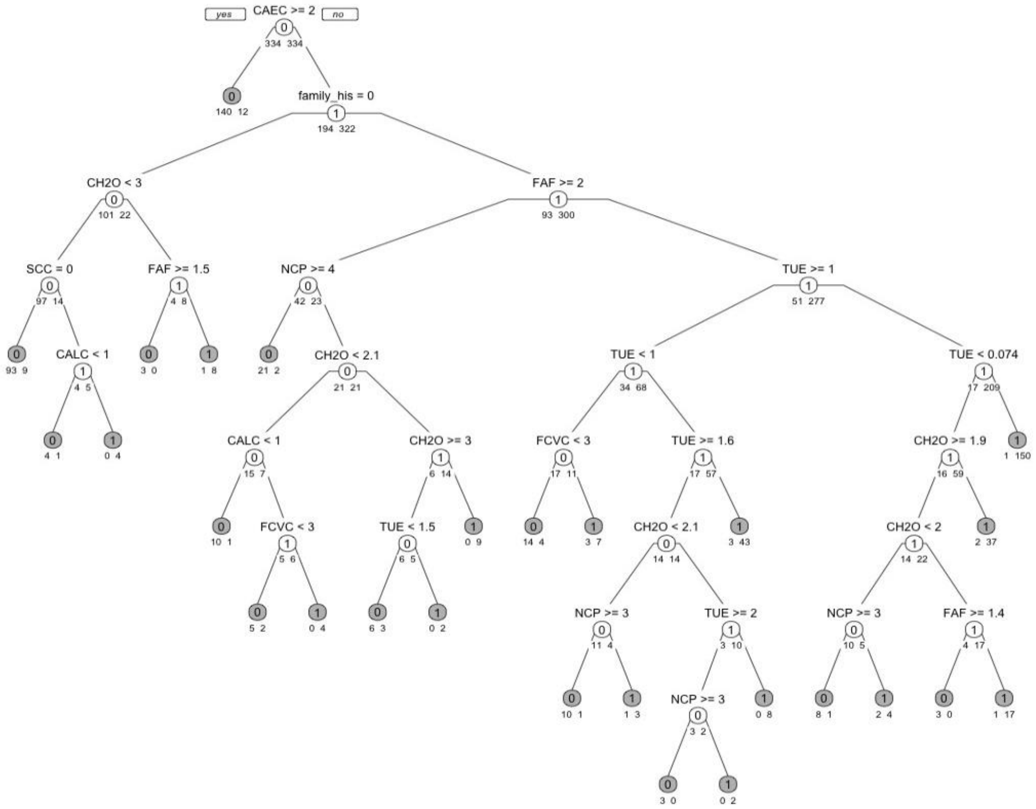
- NObeyesdad and CAEC (Individuals who consume food between meals)

BI models:

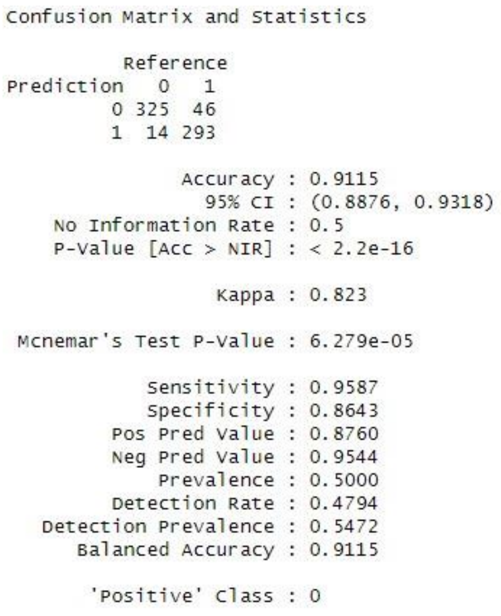
Decision Tree

Decision tree analysis is one of the most simple and intuitive classification techniques one can use to tackle classification problems. The process of building the model is fast and computation time is also comparatively quite less. The process takes just a single input variable at each node and makes a decision using certain criteria to recursively partition the tree into various splits till we hit the termination criteria.

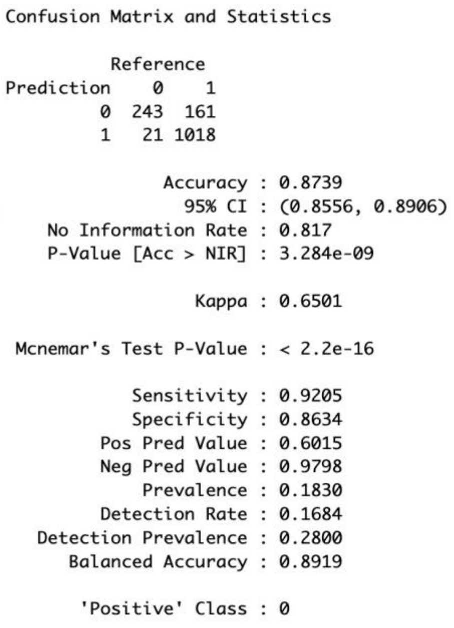
Post binning the output variable into just two separate categories, we make the problem statement, one of binary classification. The technique of 'post-pruning' was used on the developed decision tree to come up with the optimized decision tree, which averts much of overfitting. The final output in terms of the decision tree analysis renders a tree with 27 leaves.



Subsequent accuracy testing (using confusion matrix) was performed on training and validation dataset that rendered 91.15% accuracy on training dataset and 87.39% on validation dataset. The true positive rate (sensitivity) and true negative rate (specificity) as mentioned below show that the model works satisfactorily classifying the true positives and true negatives.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 325  46
         1  14 293

               Accuracy : 0.9115
                 95% CI : (0.8876, 0.9318)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.823

 Mcnemar's Test P-Value : 6.279e-05

            Sensitivity : 0.9587
            Specificity : 0.8643
         Pos Pred Value : 0.8760
         Neg Pred Value : 0.9544
             Prevalence : 0.5000
         Detection Rate : 0.4794
   Detection Prevalence : 0.5472
      Balanced Accuracy : 0.9115

       'Positive' Class : 0
```



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  243  161
         1   21 1018

               Accuracy : 0.8739
                 95% CI : (0.8556, 0.8906)
    No Information Rate : 0.817
    P-Value [Acc > NIR] : 3.284e-09

                  Kappa : 0.6501

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9205
            Specificity : 0.8634
         Pos Pred Value : 0.6015
         Neg Pred Value : 0.9798
             Prevalence : 0.1830
         Detection Rate : 0.1684
   Detection Prevalence : 0.2800
      Balanced Accuracy : 0.8919

       'Positive' Class : 0
```

Training Dataset **Validation Dataset**

Some limitations of using a decision tree analysis are as mentioned bellowed:

1. It is quite volatile and small changes in the input variable can render very different trees. It is highly sensitive to outliers and errors too.
2. It fails to consider the interaction among input variables.

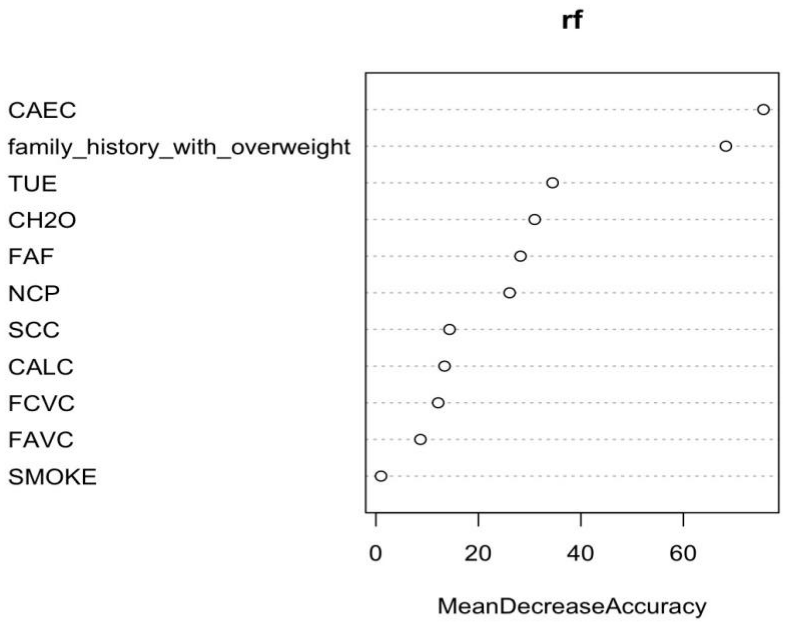
Some of these can be tackled, using ensemble methods of 'Random Forests' and ' 'Boosted Tree' analysis.

Random Forest:

Random forest is the first of the two ensemble methods we use in our Decision tree analysis to tackle problems related to overfitting and volatility. The basic logic here is to combine predictions from many different trees, which would result in better predictive performance than single trees. The model draws multiple bootstrap resamples of cases from the data, and each resample uses a random subset of predictors to produce a tree.

The downside to it is, we lose the rules we use to implement these trees since we are dealing with many trees, and not just one. Unlike simple decision tree analysis, we cannot intuitively state the splitting criteria rules to come up with the desired classifications.

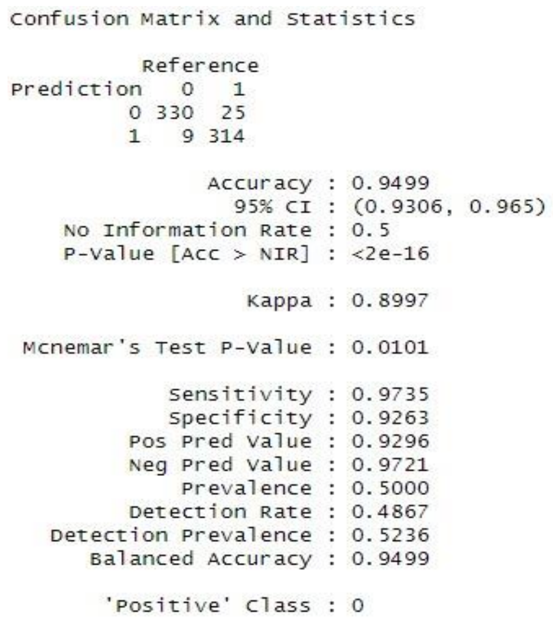
In the current analysis we make use of 500 trees for our random forest analysis. Using one of the helpful functionalities on R while modeling a random forest, we plotted the below variable importance plot, which gave us an idea of which variables have been predominantly viewed as an important factor by the model while designing the random forest.



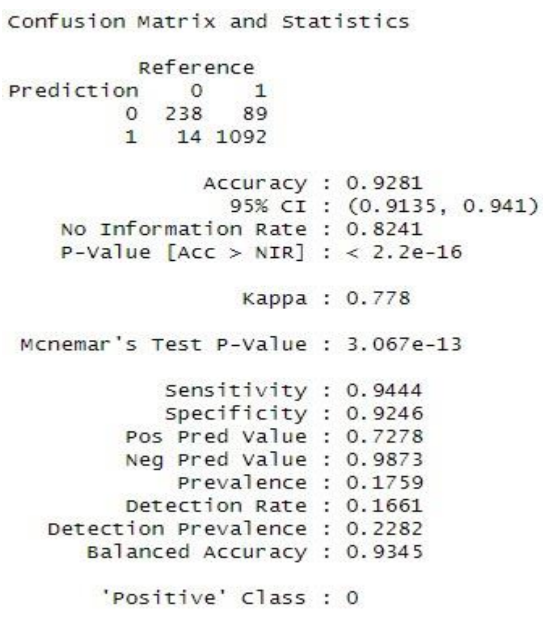
Variable Importance Plot

Using the variable importance plot, we deduce that CAEC (individuals who consume food between meals) and Family history with overweight (genetics) are some of the highly prominent factors that affect the chances of a person being overweight or not.

Post developing the model, accuracy testing using confusion matrix was performed on the training and validation dataset. The training dataset showed an accuracy of ~95%, while the validation dataset had an accuracy of 92.81%. The specificity and sensitivity numbers are in line to support of claim that this model works quite well in performing the said classification problem.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 330  25
         1   9 314

               Accuracy : 0.9499
                 95% CI : (0.9306, 0.965)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8997

 Mcnemar's Test P-Value : 0.0101

            Sensitivity : 0.9735
            Specificity : 0.9263
         Pos Pred Value : 0.9296
         Neg Pred Value : 0.9721
             Prevalence : 0.5000
         Detection Rate : 0.4867
   Detection Prevalence : 0.5236
      Balanced Accuracy : 0.9499

       'Positive' Class : 0
```

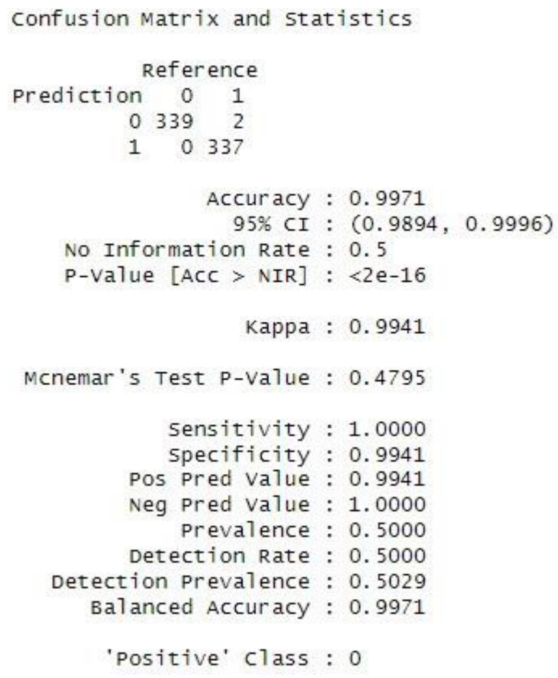


```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  238   89
         1   14 1092

               Accuracy : 0.9281
                 95% CI : (0.9135, 0.941)
    No Information Rate : 0.8241
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.778

 Mcnemar's Test P-Value : 3.067e-13

            Sensitivity : 0.9444
            Specificity : 0.9246
         Pos Pred Value : 0.7278
         Neg Pred Value : 0.9873
             Prevalence : 0.1759
         Detection Rate : 0.1661
   Detection Prevalence : 0.2282
      Balanced Accuracy : 0.9345

       'Positive' Class : 0
```

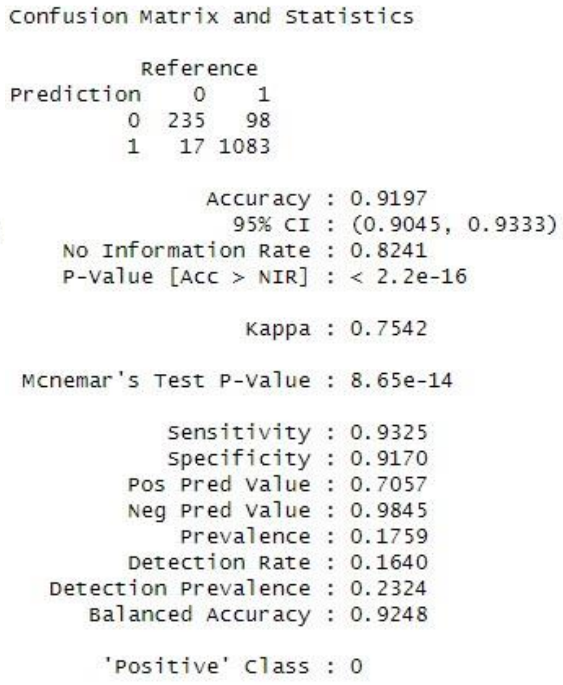
Training Dataset **Validation Dataset**

Boosted Tree

Just like random forest boosted tree analysis is an ensemble method, but it uses an iterative approach in which each successive tree focuses on correctly classifying the misclassified trees from the prior trees. We use the library 'adabag' on R studio to develop the model. Just like random forest, this model is difficult to imagine using graphics on R studio, as instead of a single tree, it is an ensemble of trees. After creating the model, we evaluate its accuracy by examining the training and validation datasets using a confusion matrix. The reported accuracy on the training dataset was 99.71% and that on the validation dataset was ~92% as mentioned in the below tables. The sensitivity and specificity numbers support the claim that this model works great in performing our classification job in hand.



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 339   2
         1   0 337

               Accuracy : 0.9971
                 95% CI : (0.9894, 0.9996)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9941

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 1.0000
            Specificity : 0.9941
         Pos Pred Value : 0.9941
         Neg Pred Value : 1.0000
             Prevalence : 0.5000
         Detection Rate : 0.5000
   Detection Prevalence : 0.5029
      Balanced Accuracy : 0.9971

       'Positive' Class : 0
```



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  235   98
         1   17 1083

               Accuracy : 0.9197
                 95% CI : (0.9045, 0.9333)
    No Information Rate : 0.8241
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7542

 Mcnemar's Test P-Value : 8.65e-14

            Sensitivity : 0.9325
            Specificity : 0.9170
         Pos Pred Value : 0.7057
         Neg Pred Value : 0.9845
             Prevalence : 0.1759
         Detection Rate : 0.1640
   Detection Prevalence : 0.2324
      Balanced Accuracy : 0.9248

       'Positive' Class : 0
```

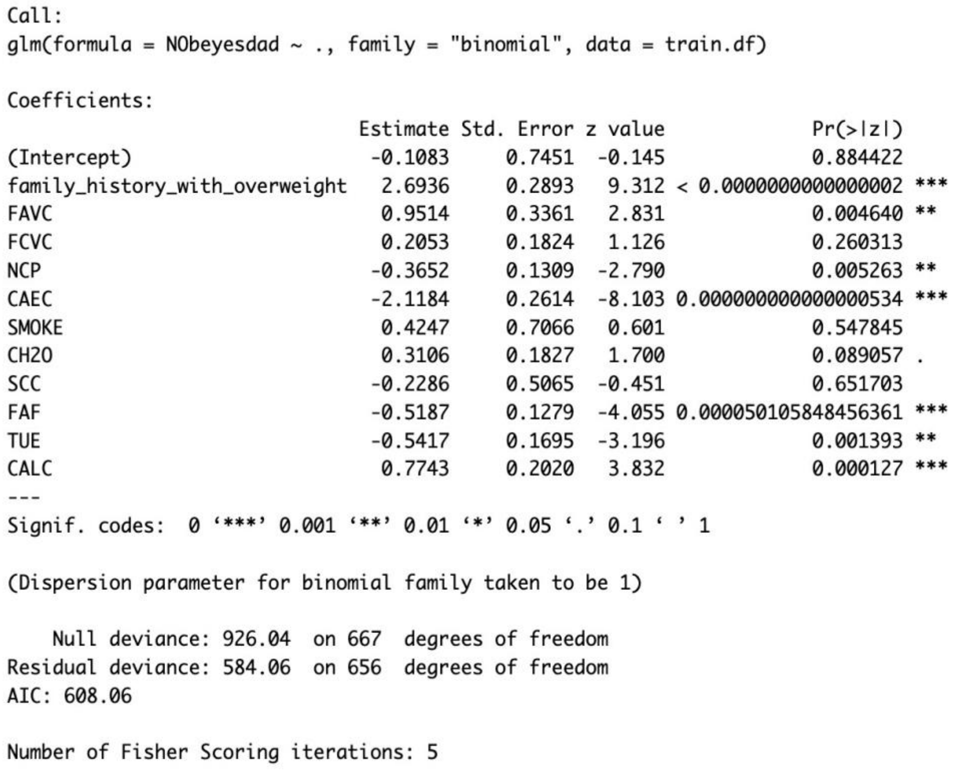
Training Dataset **Validation Dataset**

As mentioned above, none of the models involving decision trees used above could incorporate the interactions among the input variables. To enable us to tackle this problem, we then make use of Logistic regression which allows interaction relations in the input variables.

Logistic Regression

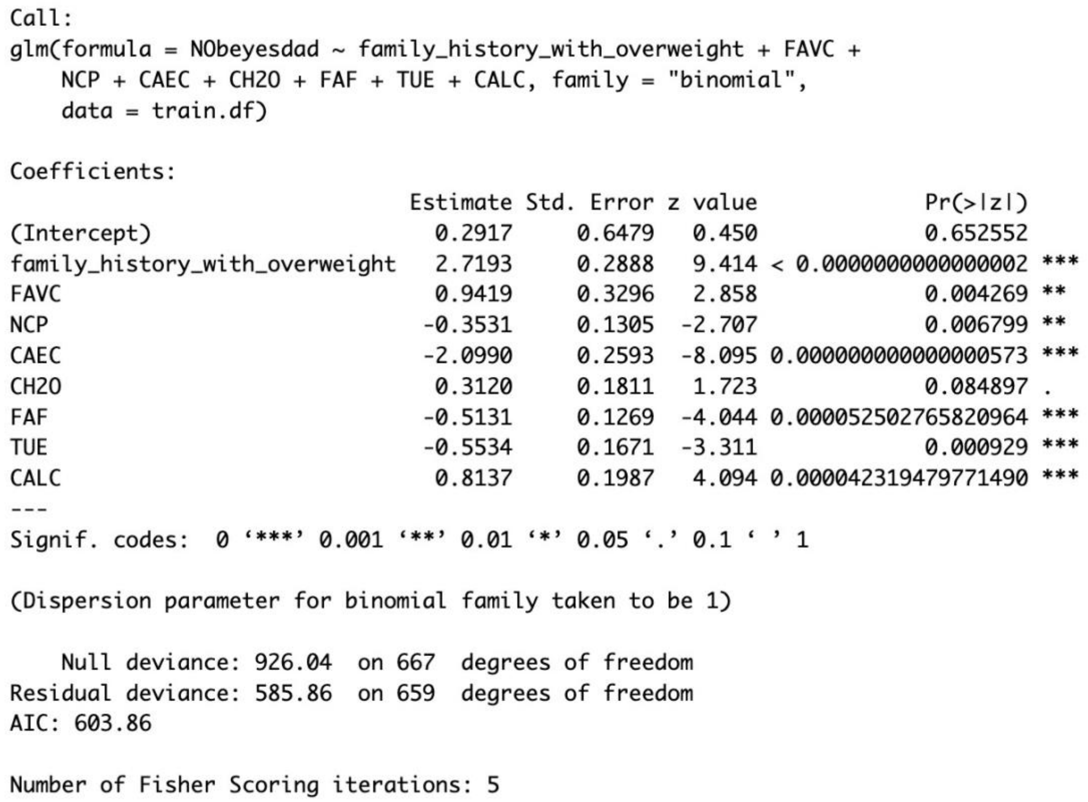
Our fourth classification model is the Logit model. Initially, we employed all available attributes to train the model. Subsequently, we implemented a backward elimination method to eliminate irrelevant attributes, removing three non-contributing attributes from the model. Summary of our full model and model with backward elimination is shown below:

Full Logistic Regression Model



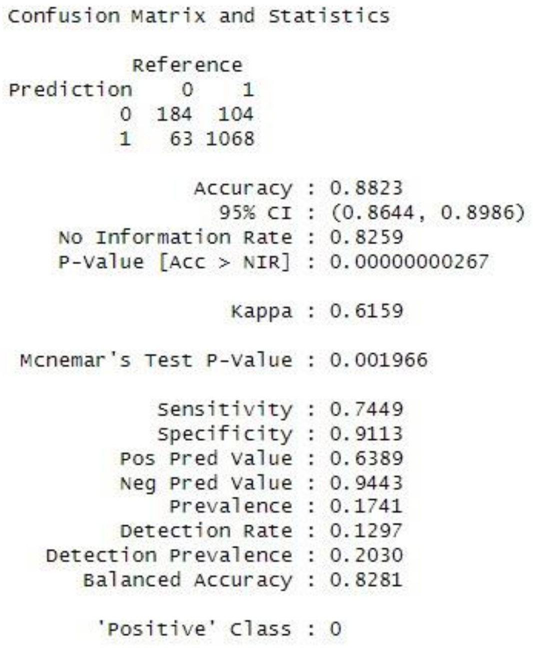
```
Call:
glm(formula = NObeyesdad ~ ., family = "binomial", data = train.df)

Coefficients:
                              Estimate Std. Error z value             Pr(>|z|)
(Intercept)                    -0.1083     0.7451  -0.145             0.884422
family_history_with_overweight  2.6936     0.2893   9.312 < 0.0000000000000002 ***
FAVC                            0.9514     0.3361   2.831             0.004640 **
FCVC                            0.2053     0.1824   1.126             0.260313
NCP                            -0.3652     0.1309  -2.790             0.005263 **
CAEC                           -2.1184     0.2614  -8.103 0.000000000000000534 ***
SMOKE                           0.4247     0.7066   0.601             0.547845
CH2O                            0.3106     0.1827   1.700             0.089057 .
SCC                            -0.2286     0.5065  -0.451             0.651703
FAF                            -0.5187     0.1279  -4.055 0.000050105848456361 ***
TUE                            -0.5417     0.1695  -3.196             0.001393 **
CALC                            0.7743     0.2020   3.832             0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 926.04  on 667  degrees of freedom
Residual deviance: 584.06  on 656  degrees of freedom
AIC: 608.06

Number of Fisher Scoring iterations: 5
```

Logistic Regression model with Backward Elimination



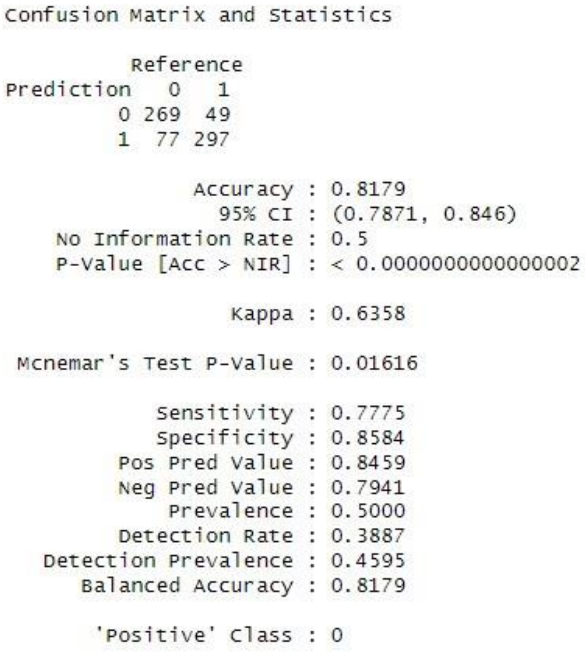
```
Call:
glm(formula = NObeyesdad ~ family_history_with_overweight + FAVC +
    NCP + CAEC + CH2O + FAF + TUE + CALC, family = "binomial",
    data = train.df)

Coefficients:
                              Estimate Std. Error z value             Pr(>|z|)
(Intercept)                     0.2917     0.6479   0.450             0.652552
family_history_with_overweight  2.7193     0.2888   9.414 < 0.0000000000000002 ***
FAVC                            0.9419     0.3296   2.858             0.004269 **
NCP                            -0.3531     0.1305  -2.707             0.006799 **
CAEC                           -2.0990     0.2593  -8.095 0.000000000000000573 ***
CH2O                            0.3120     0.1811   1.723             0.084897 .
FAF                            -0.5131     0.1269  -4.044 0.000052502765820964 ***
TUE                            -0.5534     0.1671  -3.311             0.000929 ***
CALC                            0.8137     0.1987   4.094 0.000042319479771490 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 926.04  on 667  degrees of freedom
Residual deviance: 585.86  on 659  degrees of freedom
AIC: 603.86

Number of Fisher Scoring iterations: 5
```

The confusion matrix of the validation dataset shows an accuracy of approximately 88%. An important observation is that the accuracy of the training dataset is lower than that of the validation dataset. This discrepancy could be attributed to an imbalanced distribution in our target variable within the dataset. To address this issue and create a balanced training dataset, we performed scaling, reducing the training dataset to 33% of the original dataset. This could explain why the accuracy appears higher in the validation dataset compared to the training dataset. The confusion matrix for training and validation dataset is shown below:



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  184  104
         1   63 1068

               Accuracy : 0.8823
                 95% CI : (0.8644, 0.8986)
    No Information Rate : 0.8259
    P-Value [Acc > NIR] : 0.00000000267

                  Kappa : 0.6159

 Mcnemar's Test P-Value : 0.001966

            Sensitivity : 0.7449
            Specificity : 0.9113
         Pos Pred Value : 0.6389
         Neg Pred Value : 0.9443
             Prevalence : 0.1741
         Detection Rate : 0.1297
   Detection Prevalence : 0.2030
      Balanced Accuracy : 0.8281

       'Positive' Class : 0
```

Training Dataset

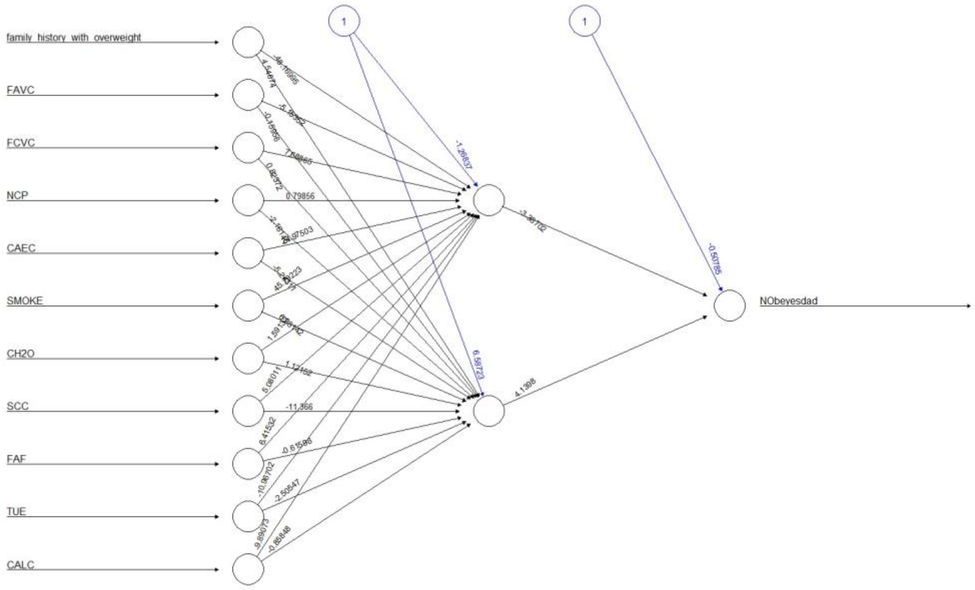


```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 269  49
         1  77 297

               Accuracy : 0.8179
                 95% CI : (0.7871, 0.846)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 0.0000000000000002

                  Kappa : 0.6358

 Mcnemar's Test P-Value : 0.01616

            Sensitivity : 0.7775
            Specificity : 0.8584
         Pos Pred Value : 0.8459
         Neg Pred Value : 0.7941
             Prevalence : 0.5000
         Detection Rate : 0.3887
   Detection Prevalence : 0.4595
      Balanced Accuracy : 0.8179

       'Positive' Class : 0
```

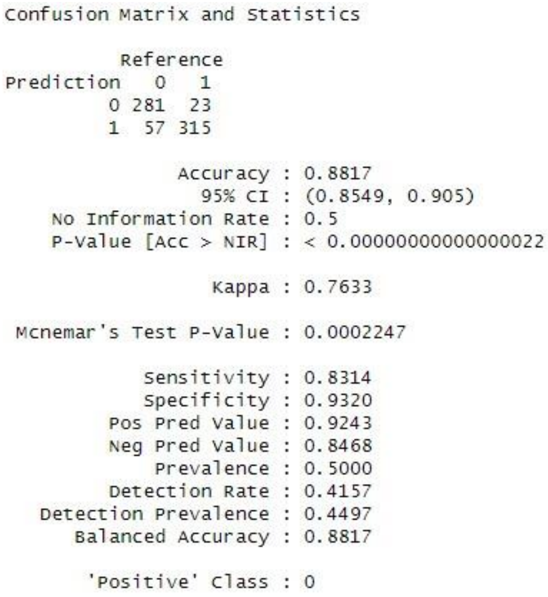
Validation Dataset

Neural Network

Our fifth classification model involves neural networks. We have used two hidden layers to run the model, enabling it to recognize more patterns. Increasing the size of the hidden layer introduces the risk of overfitting. It also increases the computational time. Consequently, after thorough evaluation, we have concluded that sticking with two hidden layers represents the optimal decision for our implementation. The neural network plot is shown below:

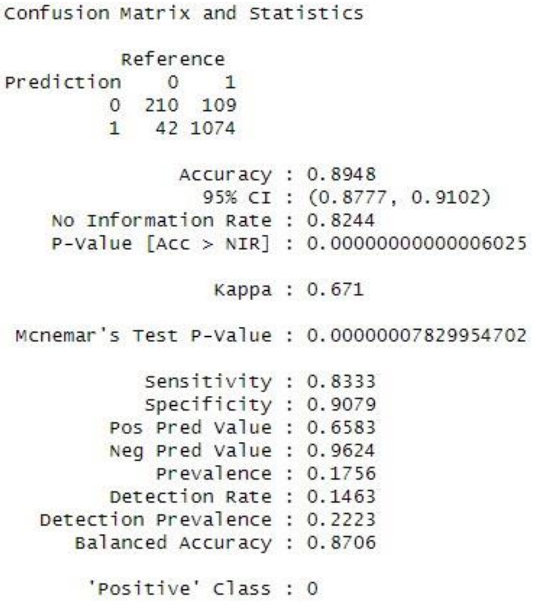


The confusion matrix of the training and validation datasets indicates an approximate accuracy of around 90% for the model. Moreover, both datasets' sensitivity and specificity values validate no noticeable overfitting issue. However, despite these observations, it is notable that this particular model falls short in terms of accuracy when compared to previous models, thereby indicating that it might not be the most optimal one among our existing models. The confusion matrix for training and validation dataset is shown below:



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 281  23
         1  57 315
                                               
               Accuracy : 0.8817               
                 95% CI : (0.8549, 0.905)      
    No Information Rate : 0.5                  
    P-Value [Acc > NIR] : < 0.00000000000000022
                                               
                  Kappa : 0.7633               
                                               
 Mcnemar's Test P-Value : 0.0002247            
                                               
            Sensitivity : 0.8314               
            Specificity : 0.9320               
         Pos Pred Value : 0.9243               
         Neg Pred Value : 0.8468               
             Prevalence : 0.5000               
         Detection Rate : 0.4157               
   Detection Prevalence : 0.4497               
      Balanced Accuracy : 0.8817               
                                               
       'Positive' Class : 0                    
```

Training Dataset



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  210  109
         1   42 1074
                                            
               Accuracy : 0.8948            
                 95% CI : (0.8777, 0.9102)  
    No Information Rate : 0.8244            
    P-Value [Acc > NIR] : 0.00000000000006025
                                            
                  Kappa : 0.671             
                                            
 Mcnemar's Test P-Value : 0.0000007829954702
                                            
            Sensitivity : 0.8333            
            Specificity : 0.9079            
         Pos Pred Value : 0.6583            
         Neg Pred Value : 0.9624            
             Prevalence : 0.1756            
         Detection Rate : 0.1463            
   Detection Prevalence : 0.2223            
      Balanced Accuracy : 0.8706            
                                            
       'Positive' Class : 0                 
```

Validation Dataset

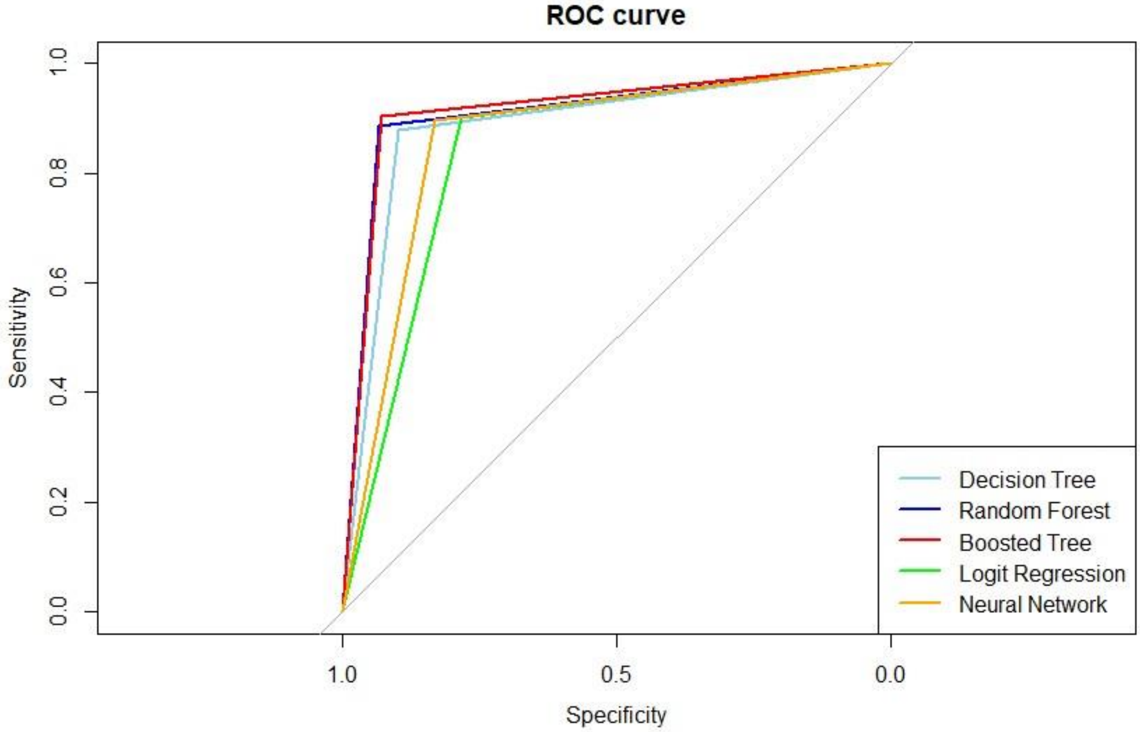
Model Evaluation

Post working on different classification models, it is important to evaluate the authenticity of the claim made by each model about correctly classifying various instances in our dataset. Even though accuracy along with sensitivity and specificity give us quite a good idea about how useful the model can be, accuracy alone is not always the best metric in evaluating and choosing the champion model. Sometimes accuracy can be misleading too.

So, we make use of the evaluation metric called ROC (Receiver Operating Characteristic) curve, to evaluate the models and come up with the champion model. This is in addition to the accuracy that we encountered in previous sections.

Classification Model	Area under curve
Boosted Tree	0.9313
Random Forest	0.9307

Decision Tree	0.8911
Neural Network	0.8678
Logistic Regression	0.865



Simply, by comparing the accuracies we see that Boosted tree and Random forest models show almost similar trends and perform a great task at classifying various instances in our dataset. Comparing the ROC curves of all the models along with the previous knowledge about the accuracy of each model, we see that Boosted tree stands out as the champion model, as it has higher area under the curve as compared to Random forest. Hence, taking the cumulative effect of accuracy and ROC curve's area, we conclude that for our dataset, it is 'Boosted tree' analysis that can best classify the dataset into 'Overweight' and 'Non-Overweight' categories.

Findings:

Some of the general findings and key takeaways from our analysis are outlined in the below mentioned points**:**

1. After performing classification analysis using different models (namely Decision Tree, Random Forest, Boosted Tree, Logit regression and Neural Network), we find that Boosted Tree Analysis stands out as the champion model.
2. Accuracy metric in conjunction with ROC curve area were used to evaluate the model performance.
3. Food consumption patterns, frequency of use of technology devices on a daily basis, water intake and physical activity are some of the important lifestyle choices that determine the possibility of a person being obese or not.
4. Genetics or heredity, even though not a lifestyle choice, does play an important role in determining the chances of an individual being obese.

Managerial implications:

Understanding the impact of lifestyle choices on obesity is crucial for designing effective interventions. Key lifestyle factors identified include food consumption patterns, daily use of technology devices, water intake, and physical activity. Moreover, genetic predisposition and family history also plays a significant role, highlighting the multifaceted nature of obesity.

It is imperative to mention that this analysis was performed for a particular set of people from the countries of Mexico, Peru and Colombia. There may have been various confounding factors and the model may not be representative of the whole population. Hence, there is a good probability that the model cannot simply be generalized to the wider population without evaluating the similarity with this sample set on which the model was run. What we can assert is that if a representative population is obtained by any means, this model can be quite helpful when deployed in healthcare facilities to assess the likelihood of an individual being overweight or not.

References:

1. Professor Zhe Zhang’s notes.
2. UCI ML repository - https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

Tools Used:

1. R Studio
2. MS Word
3. MS PowerPoint