# BUAN 6356.006 Business Analytics with R

Group Project Team 18 Members:

1. Abhishek Dubey
2. Aparna Mishra
3. Manmohan Dash
4. Palak Sharma

Under the guidance of Prof. Zhe Zhang

# Objective:

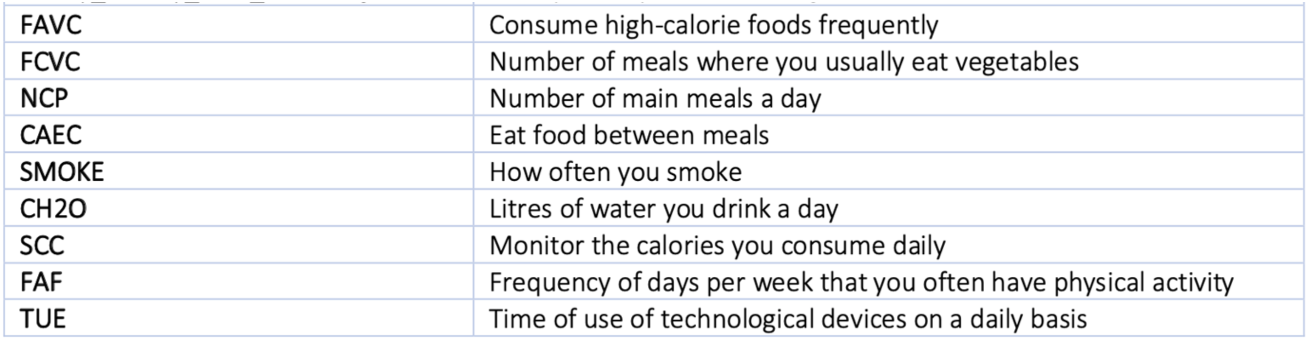
Obesity is a global health concern affecting millions of individuals worldwide. This project aims to leverage Business Intelligence (BI) techniques to estimate obesity levels in individuals from Mexico, Peru, and Colombia, based on their eating habits and physical condition. By analysing this dataset, we seek to gain valuable insights into the prevalence of obesity and its associated factors in these countries to develop targeted interventions.

# Insight Generation Points:

- Classification of individuals as ‘likely being obese / overweight’ based on their lifestyle choice
- Identifying the predominant lifestyle choices that majorly affect a person being obese or overweight
- Assessing and comparing the performance of the various classification models to come-up with the champion model to perform the task at hand

# Attribute Information

- The dataset has 3 numerical and 13 categorical attributes
- The “NObeyesdad” attribute contains BMI distributed into 7 categories[1]
- “Gender”, ”Age”, “Height”, “Weight” ,”family_history_with_overweight[2]” are the traits that an individual doesn’t have control over. These attributes aren’t lifestyle choices made by the individual and hence the first four, have been discarded from being used as variables in our models
- The lifestyle choice related attributes includes:

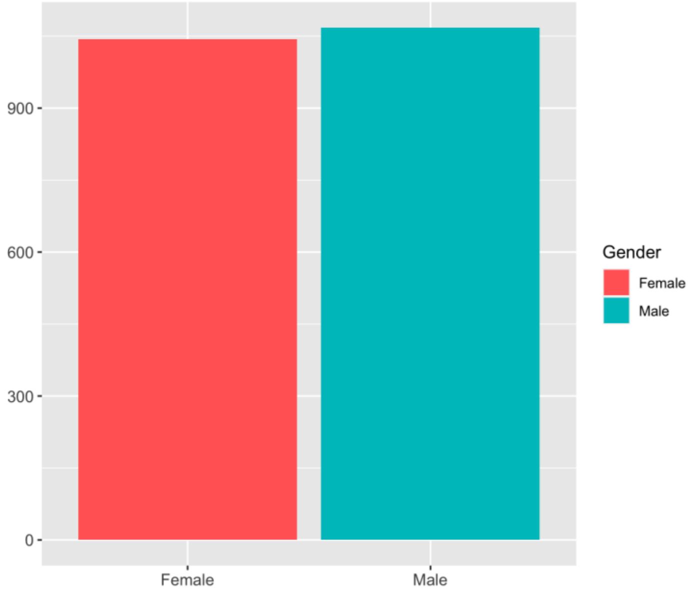


| | |
|---|---|
| FAVC | Consume high-calorie foods frequently |
| FCVC | Number of meals where you usually eat vegetables |
| NCP | Number of main meals a day |
| CAEC | Eat food between meals |
| SMOKE | How often you smoke |
| CH2O | Litres of water you drink a day |
| SCC | Monitor the calories you 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 |

**Note: 1. For ease of performing the analysis, the 7 (seven) categories have been discretized to 2 categories- ‘Non-overweight’ and ‘Overweight’**
**2. Although ‘family_history_with_overweight’ isn’t a lifestyle choice, but still we do include it in analysis heredity and genetics is an important factor in determining certain obesity conditions**

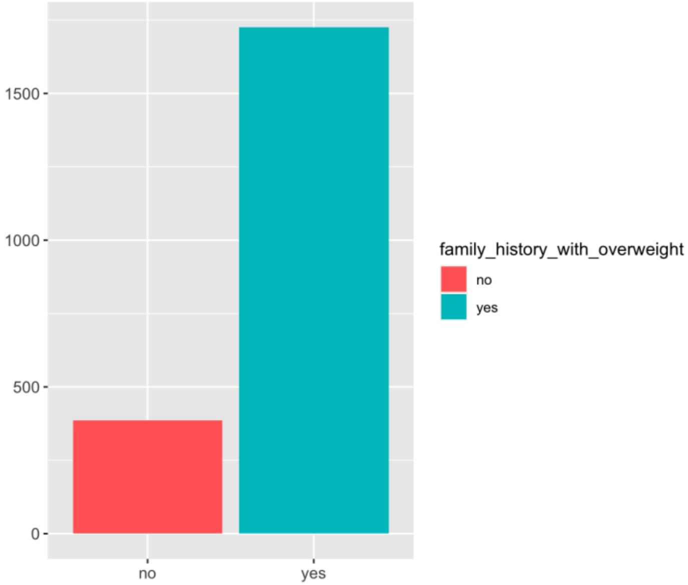
# Exploratory analysis

## Gender



The dataset exhibits gender balance, with almost equal representation of females and males

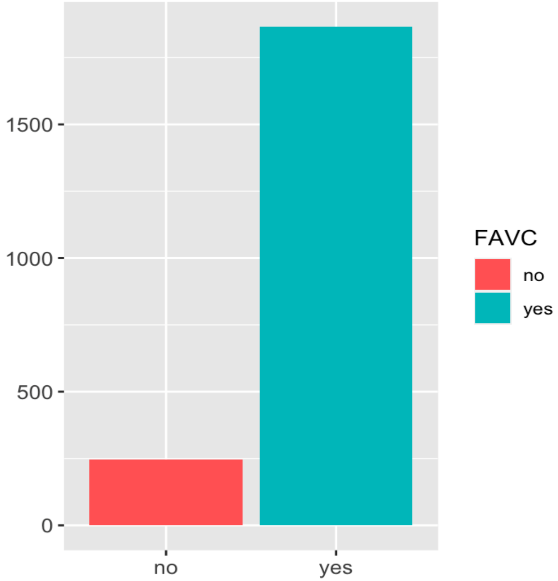
## Family history with overweight



The dataset has ~ 82% data pertaining to individuals with a family history of overweight

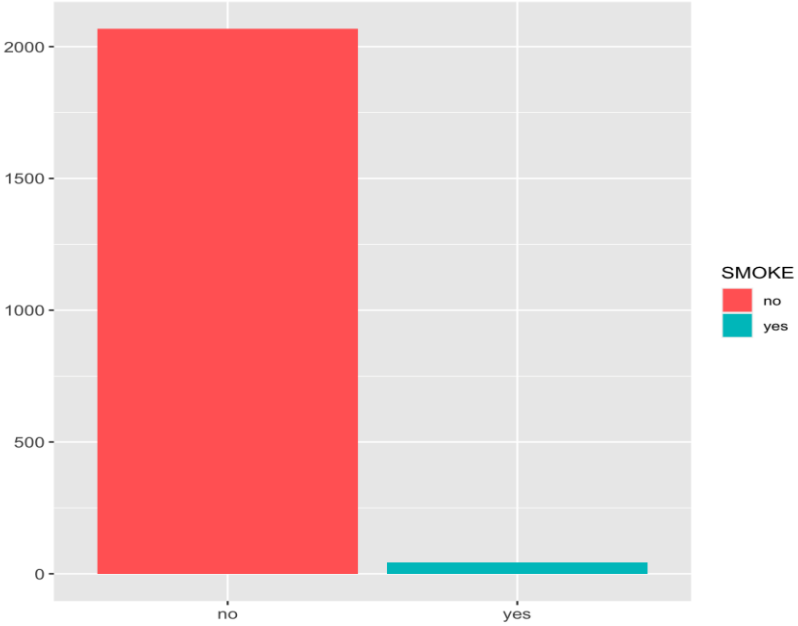
# Exploratory analysis

## High-calorie foods consumption



The dataset has ~ 88% individuals who consume high-calorie foods frequently

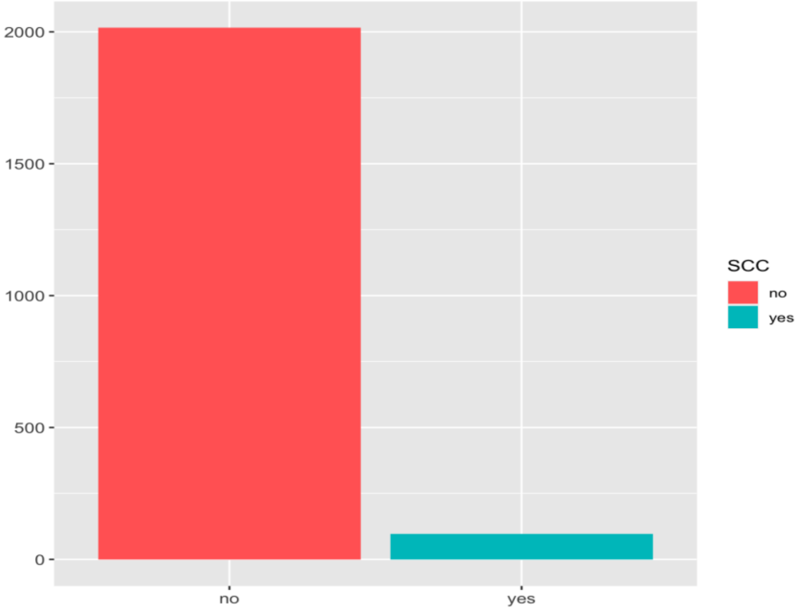
## Smoke frequency



The dataset has ~98% individuals who don't smoke

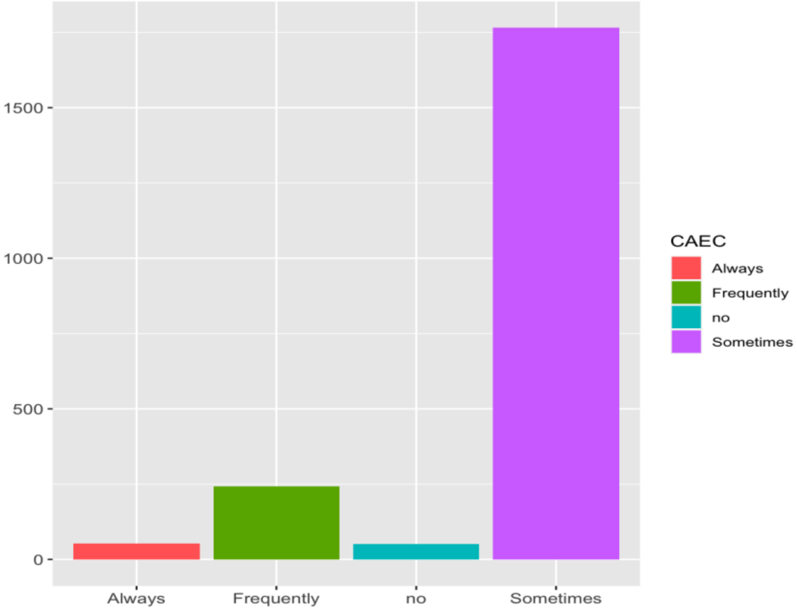
# Exploratory analysis

## Monitoring the calories consumption daily



The dataset has ~95% individuals who don't usually monitor the calories consumed daily

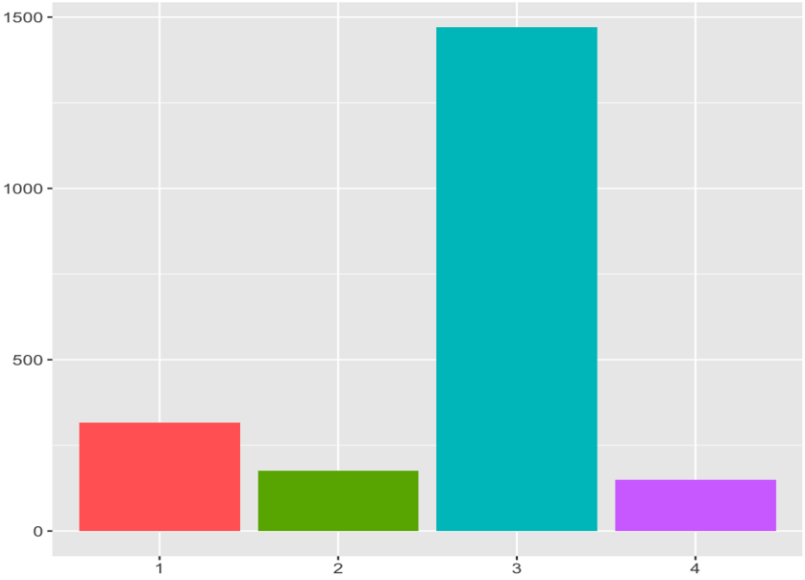
## Consumption of food between meals



The dataset has ~84% individuals who sometimes eat food between meals

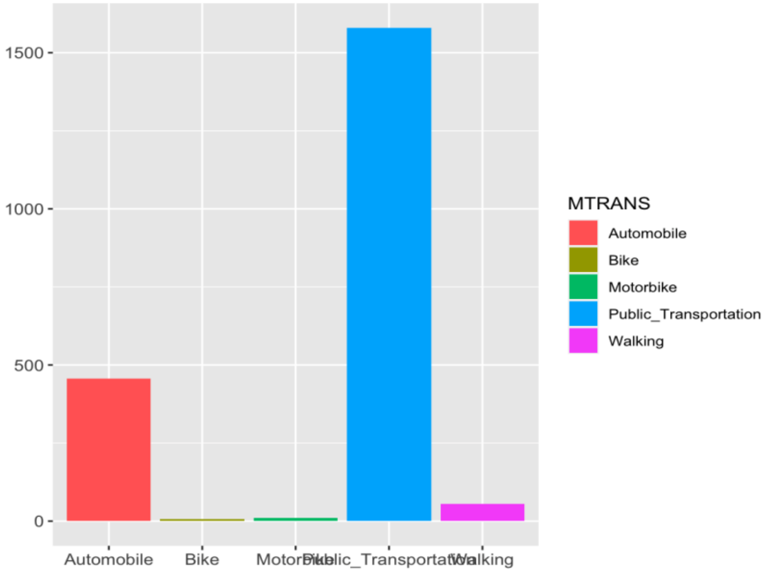
# Exploratory analysis

## Number of main meals a day



~50% individuals in this dataset are having three main meals a day

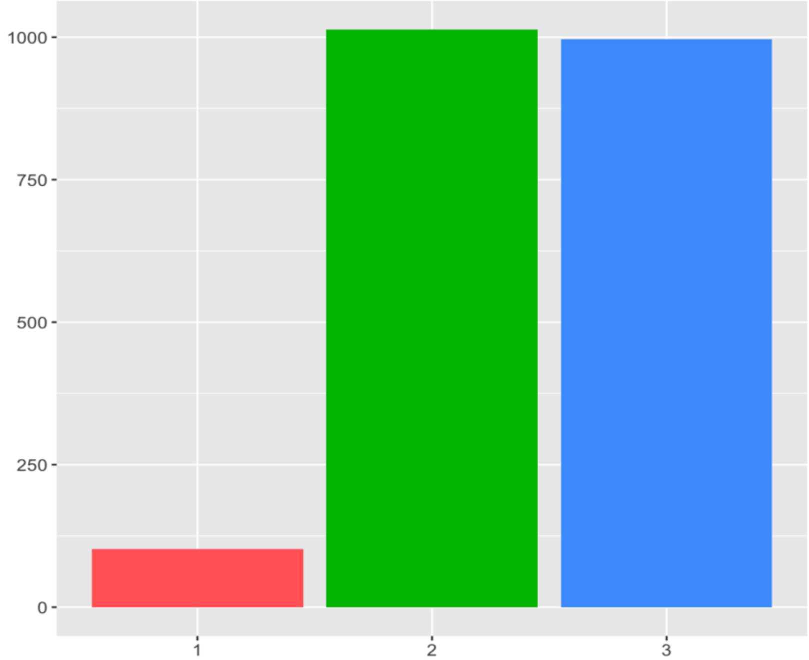
## Means of transportation



~75% individuals in this dataset uses public transportation

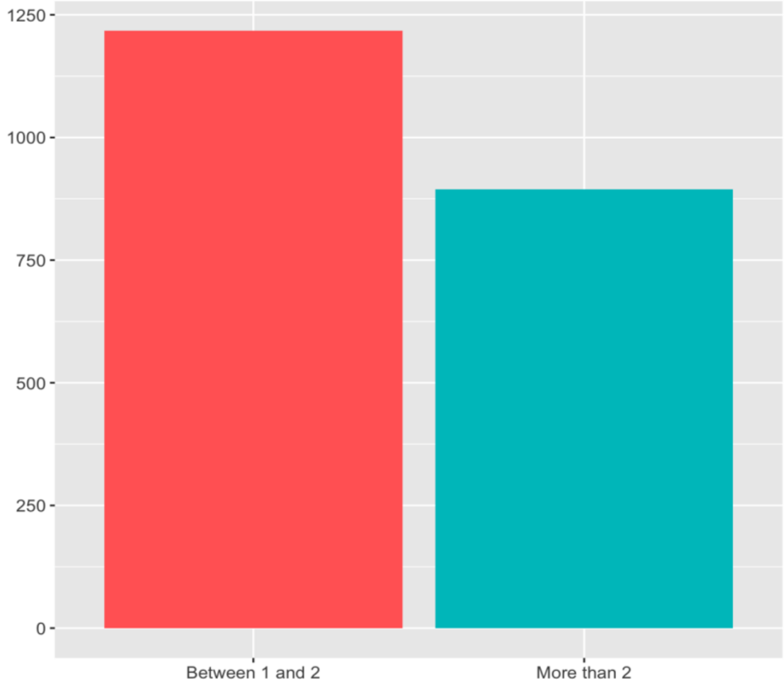
# Exploratory analysis

Number of meals where one eats vegetables



~95% individuals consumes two or more meals that include vegetables

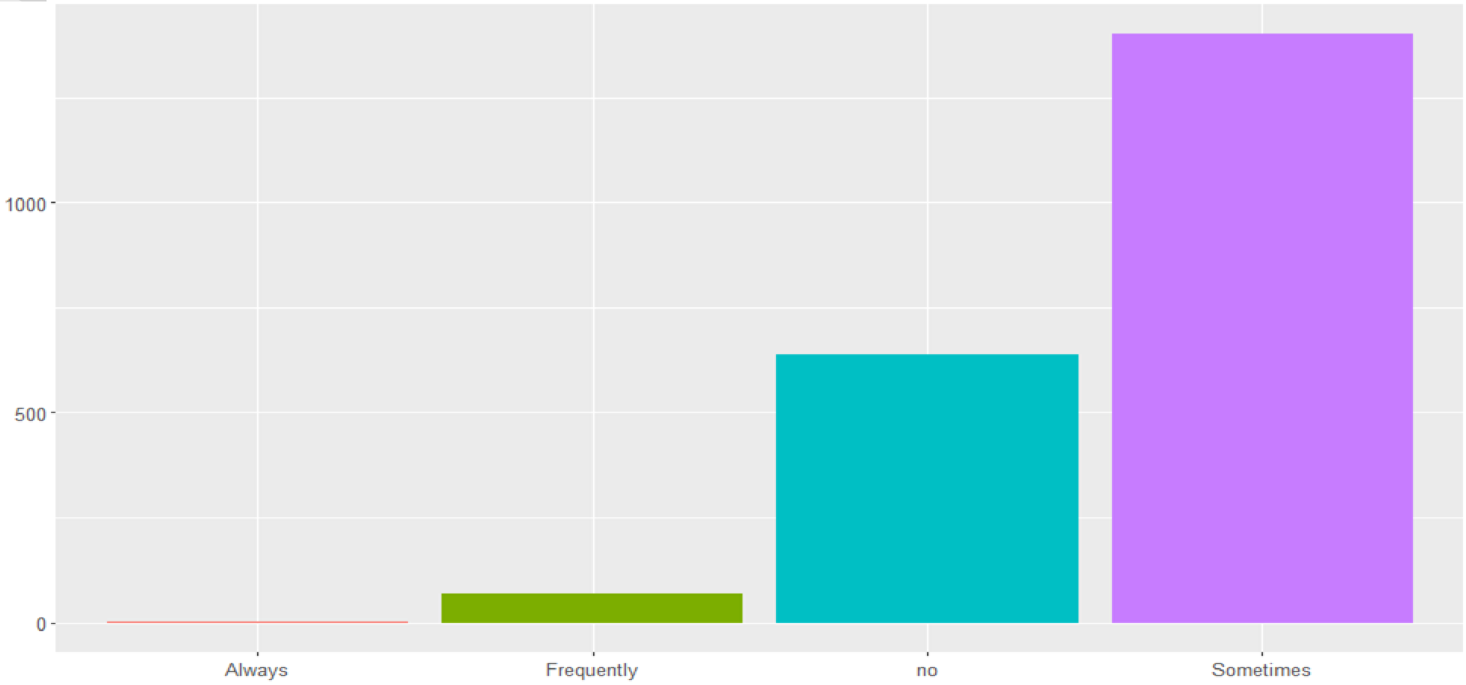
Liters of water per day



~60% of individuals in this dataset consume around 1-2 liters of water per day, while the remaining individuals consume more than 2 liters

# Exploratory analysis

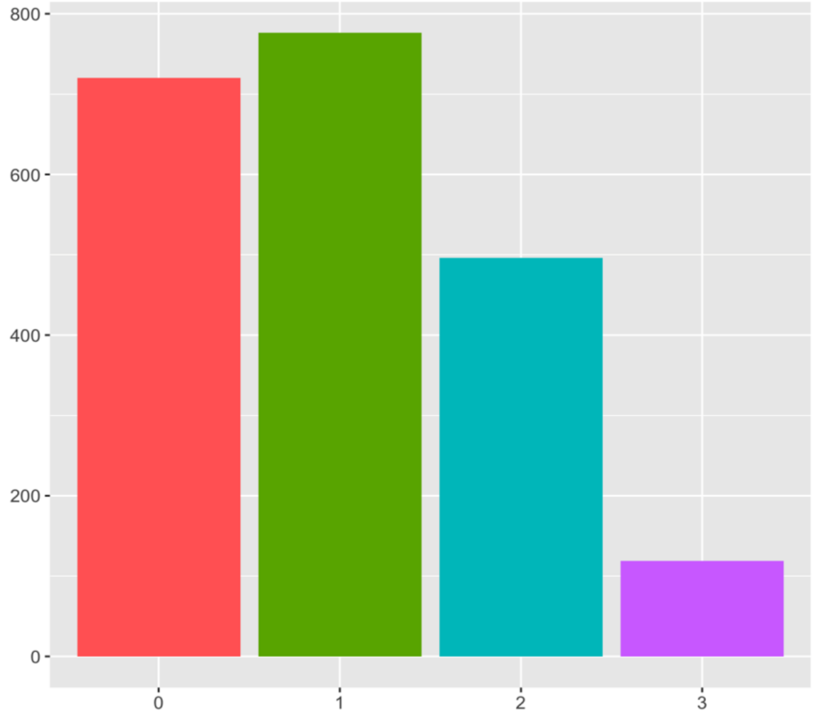
## Frequency of alcohol intake



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

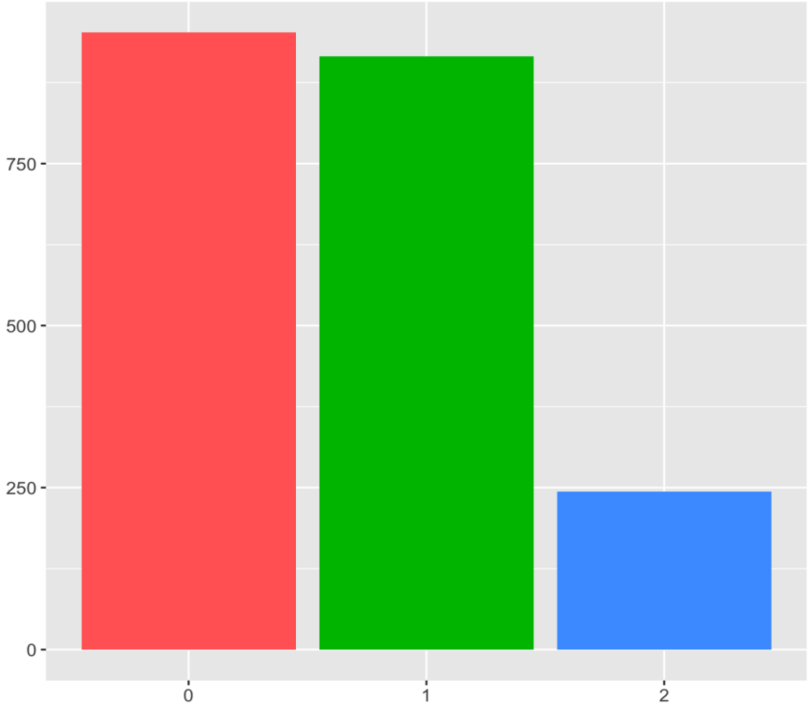
# Exploratory analysis

## Frequency of days of physical activity per week



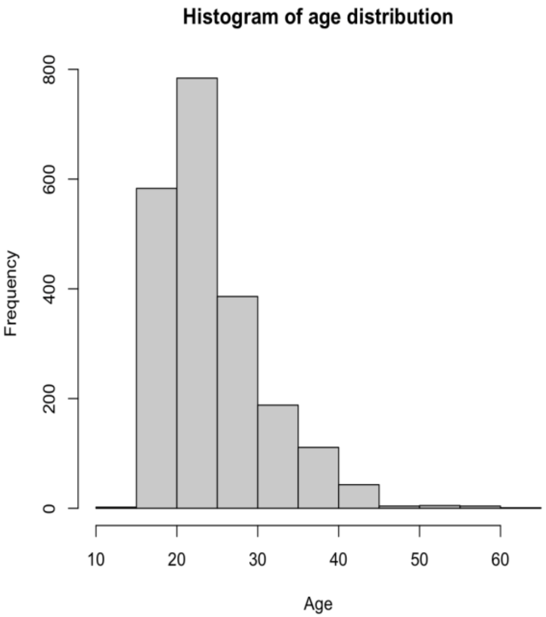
~94% of the individuals in the dataset engage in physical activity for a maximum of two days per week

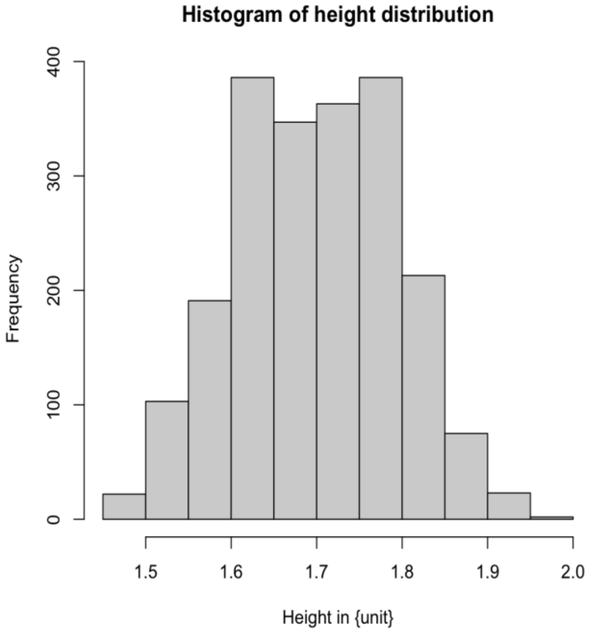
## Hours of use of technology devices on a daily basis

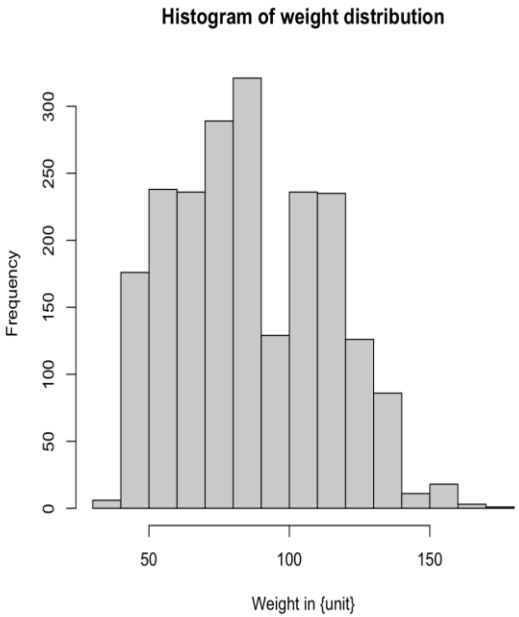


The dataset has ~88% individuals who uses 0-1 hours of technology devices on a daily basis

# Exploratory analysis

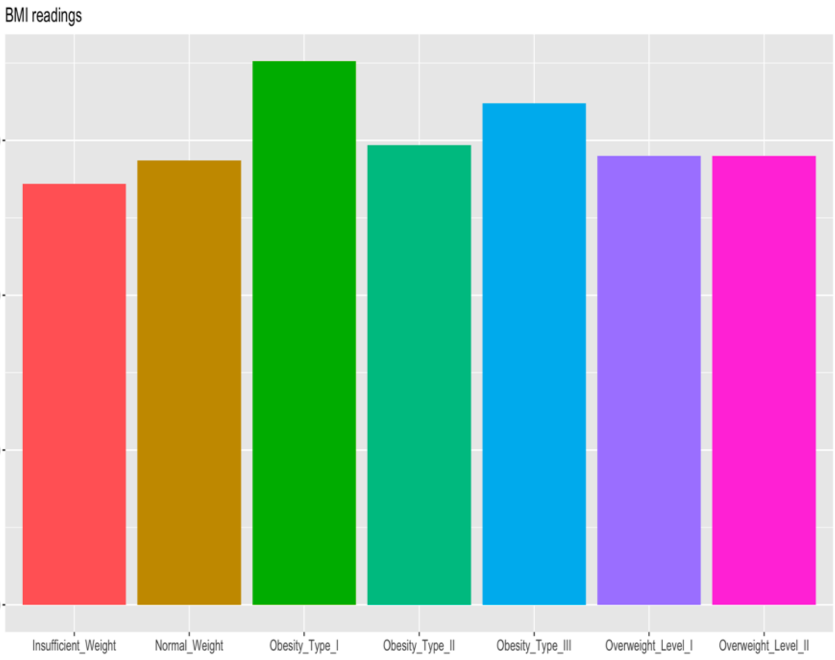

Histogram of age distribution


Histogram of height distribution

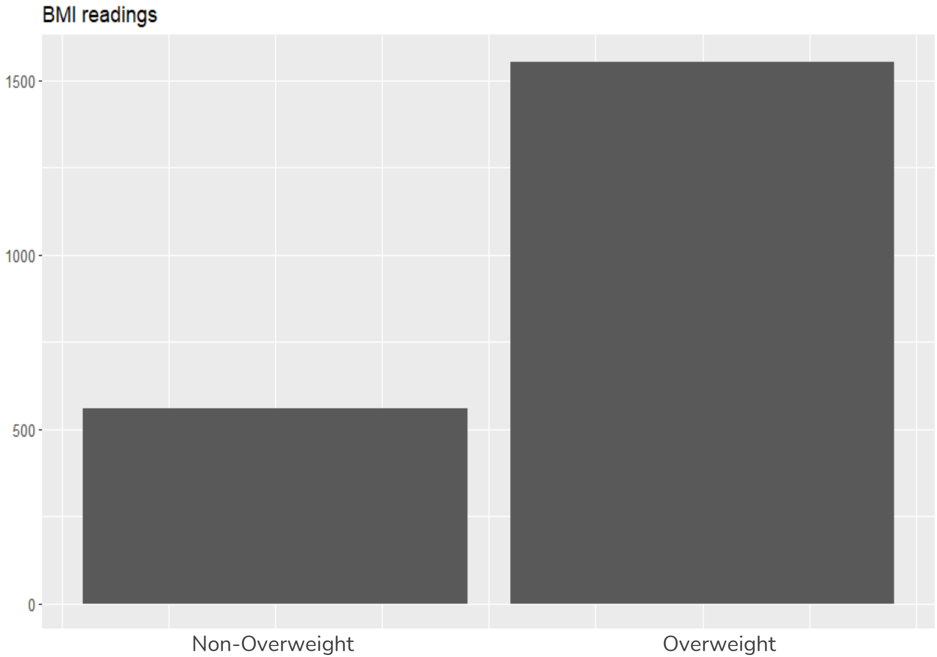

Histogram of weight distribution

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.



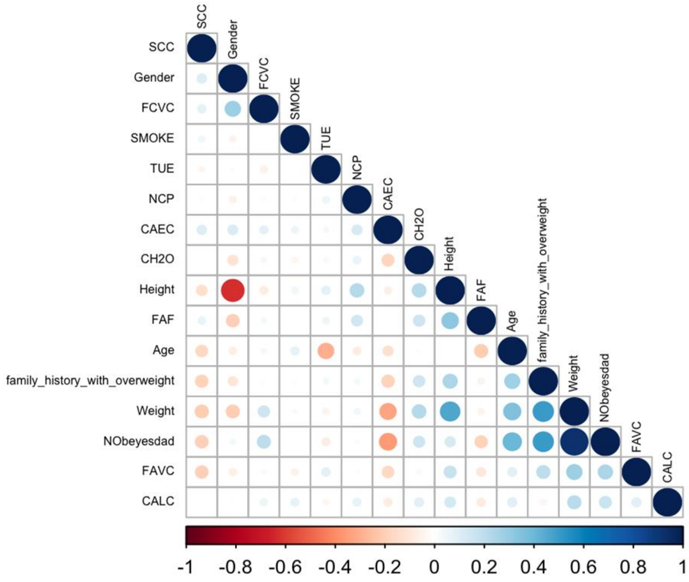
- The dataset is evenly balanced in terms of the BMI level, represented by the variable "NObeyesdad"

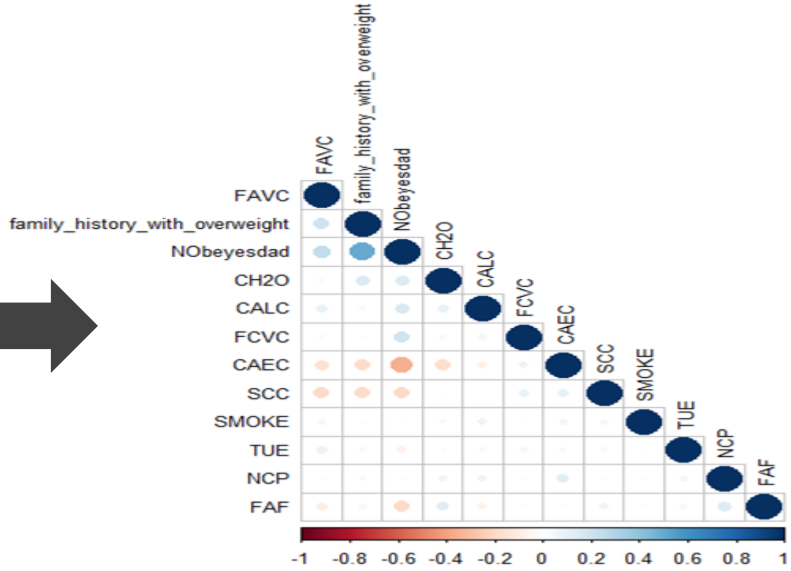


- For ease of performing the analysis, we convert the 7 categories into 2 categories- Overweight and Not overweight
- Upon this conversion, our dataset needs scaling as it is no longer balanced

**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





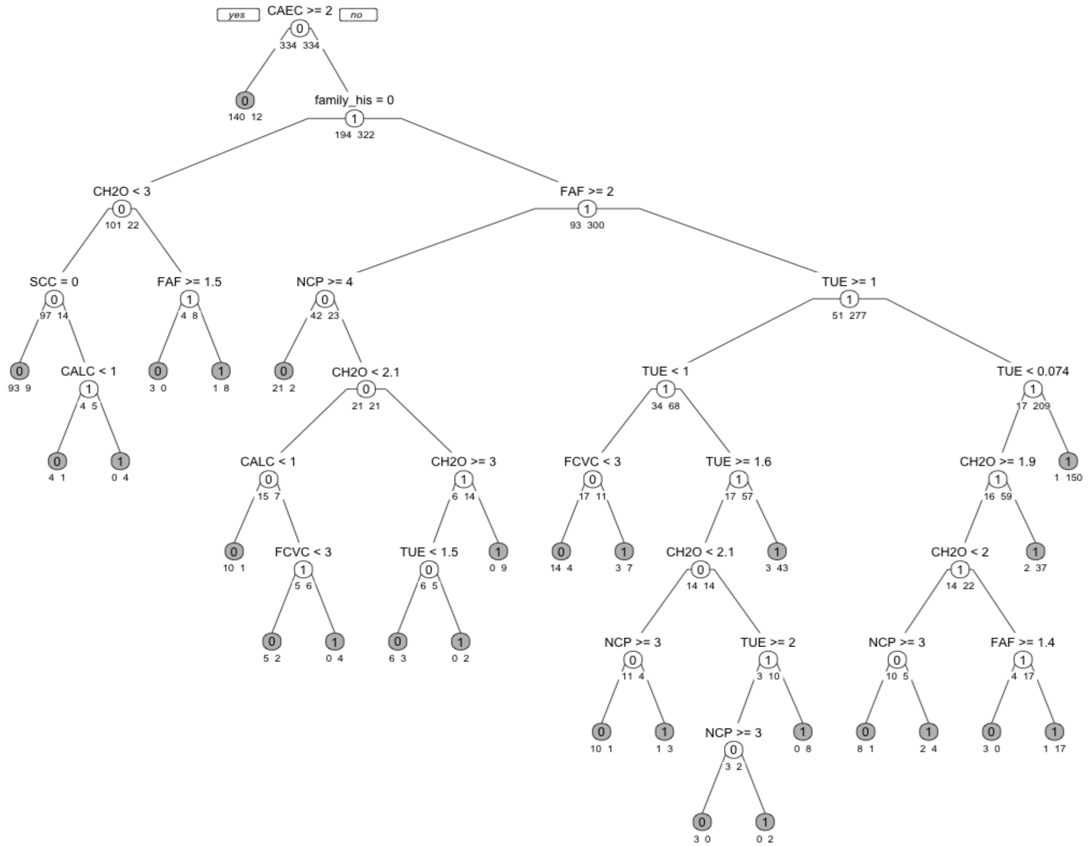
This is the correlation based on full dataset. We see high correlation between:

- 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)

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)

# Decision Tree Model



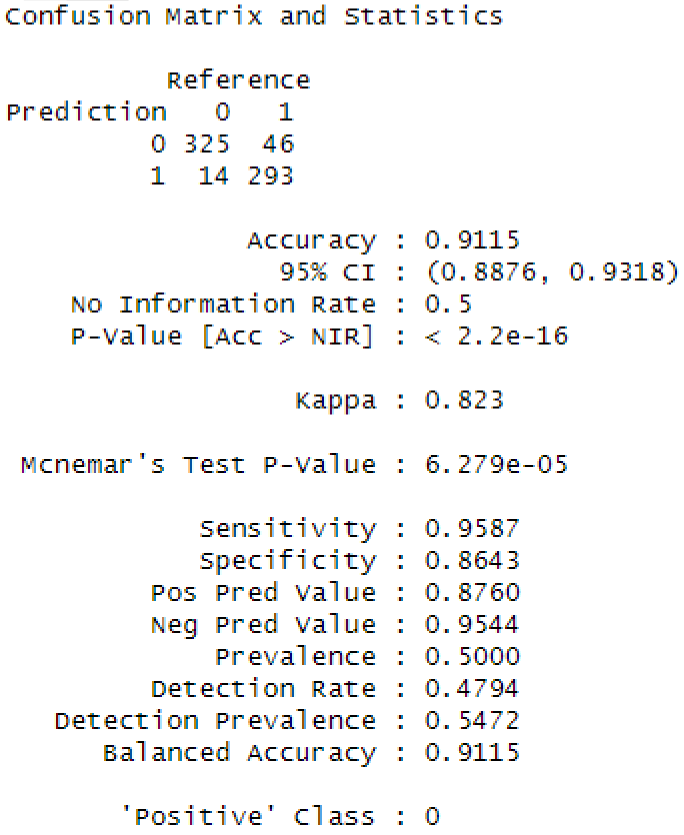
Split Based On:

- CAEC
- Family History
- CH2O
- FAF
- SCC
- CALC
- NCP
- TUE
- FCVC

Decision Tree Leaves: 27

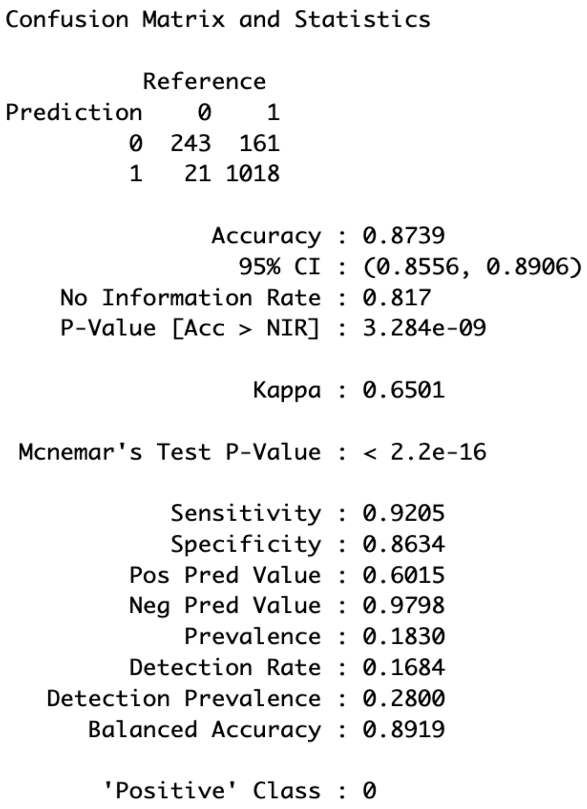
# Decision Tree Model

## Confusion Matrix for Training Dataset



```
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
```

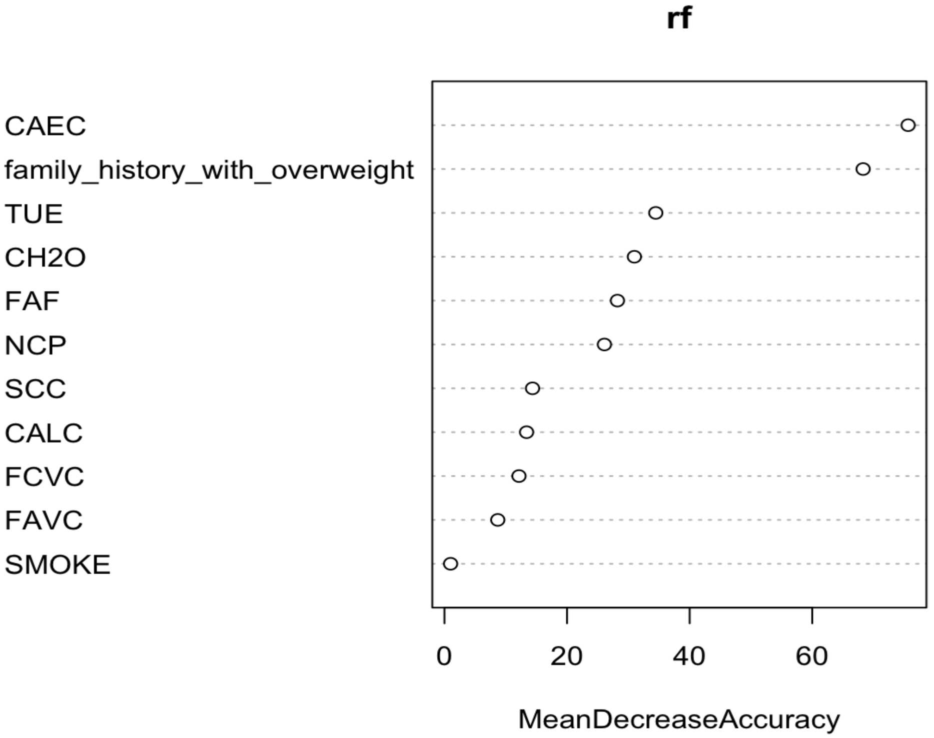
## Confusion Matrix for Validation Dataset



```
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
```

# Random Forest

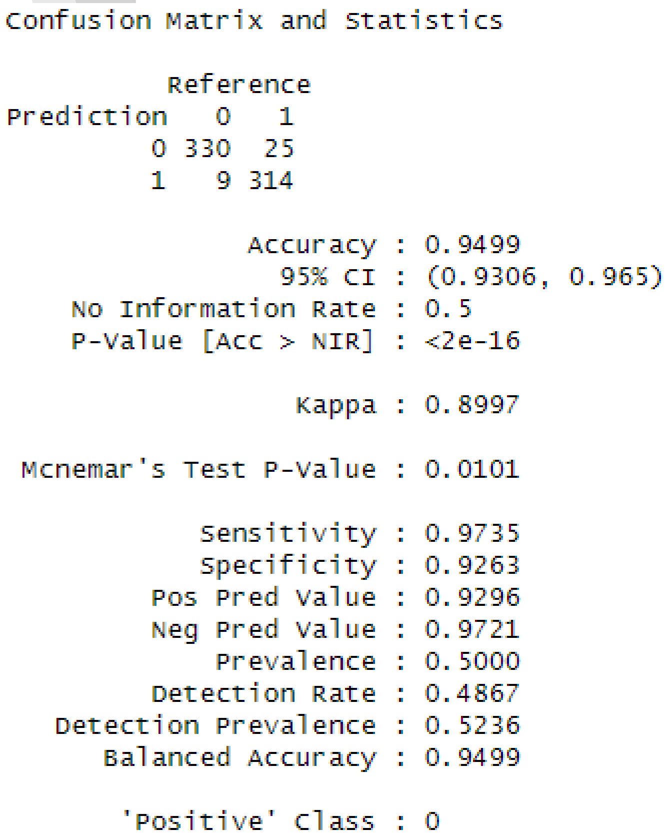
## Variable Importance Plot



CAEC (Individuals who consume food between meals) and Family history with overweight stand out as two most important attribute in our dataset

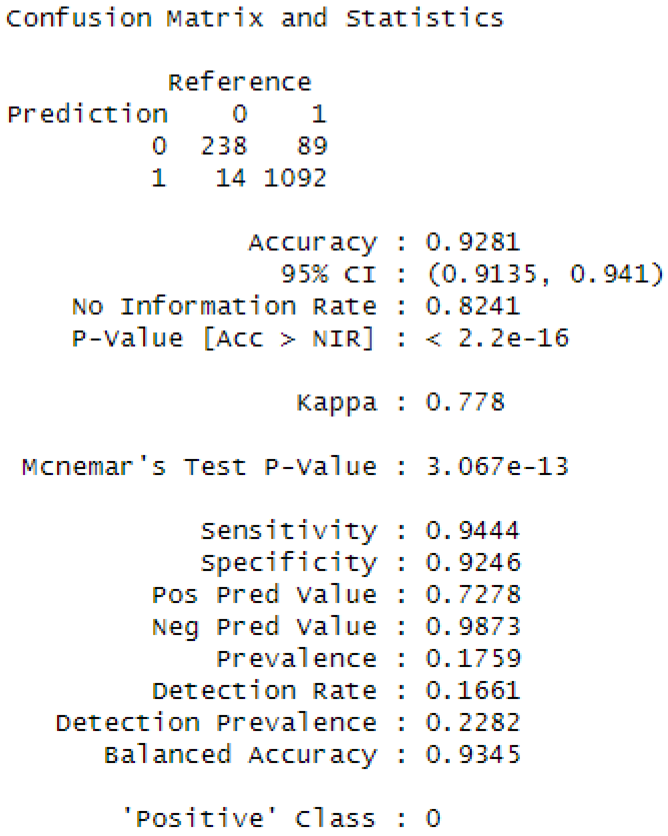
# Random Forest

## Confusion Matrix for Training Dataset



```
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
```

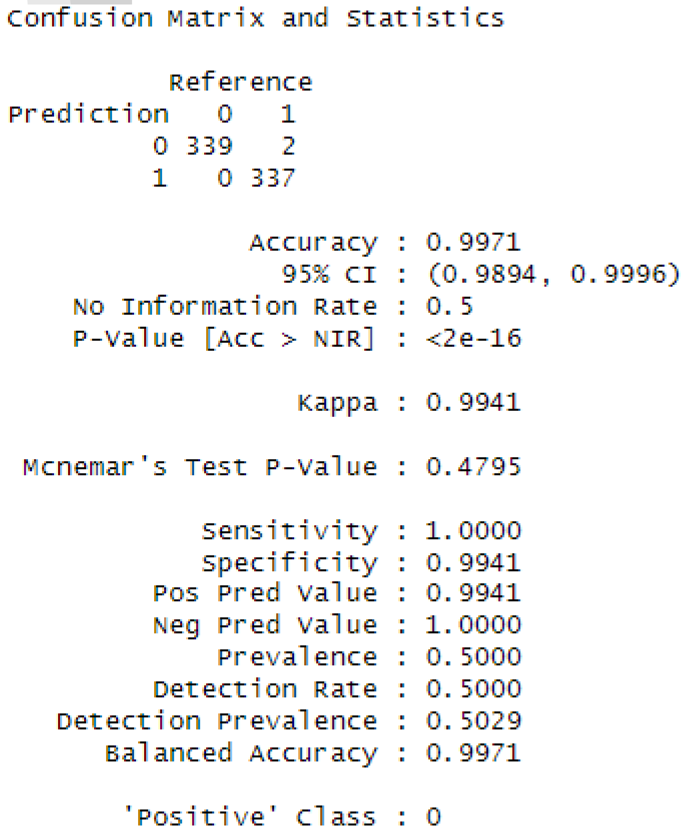
## Confusion Matrix for Validation Dataset



```
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
```

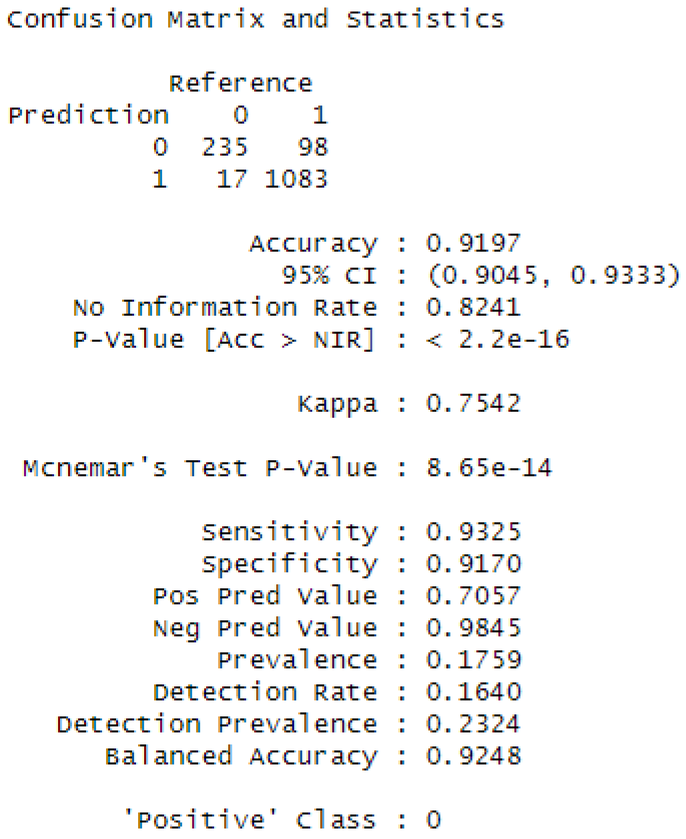
# Boosted Tree

## Confusion Matrix for Training Dataset



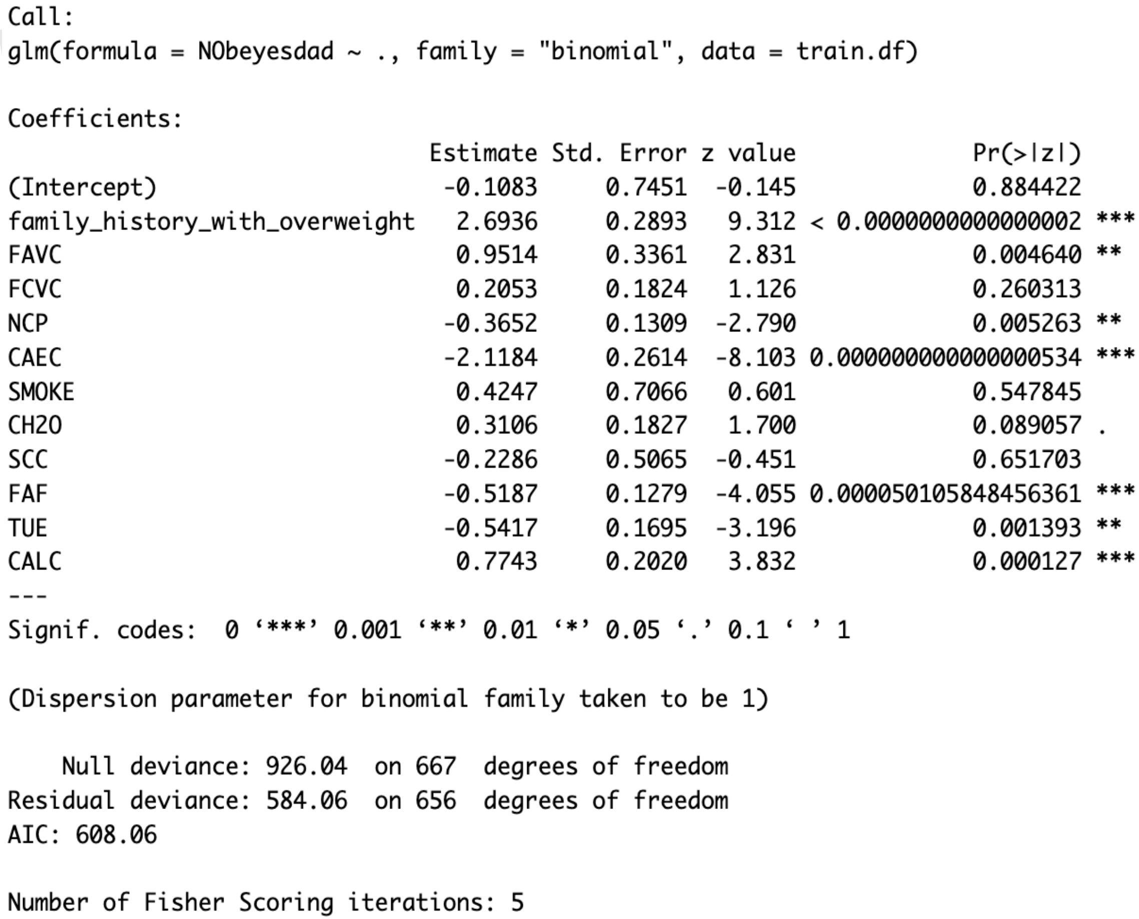
```
Confusion Matrix and Statistics

          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 for Validation Dataset



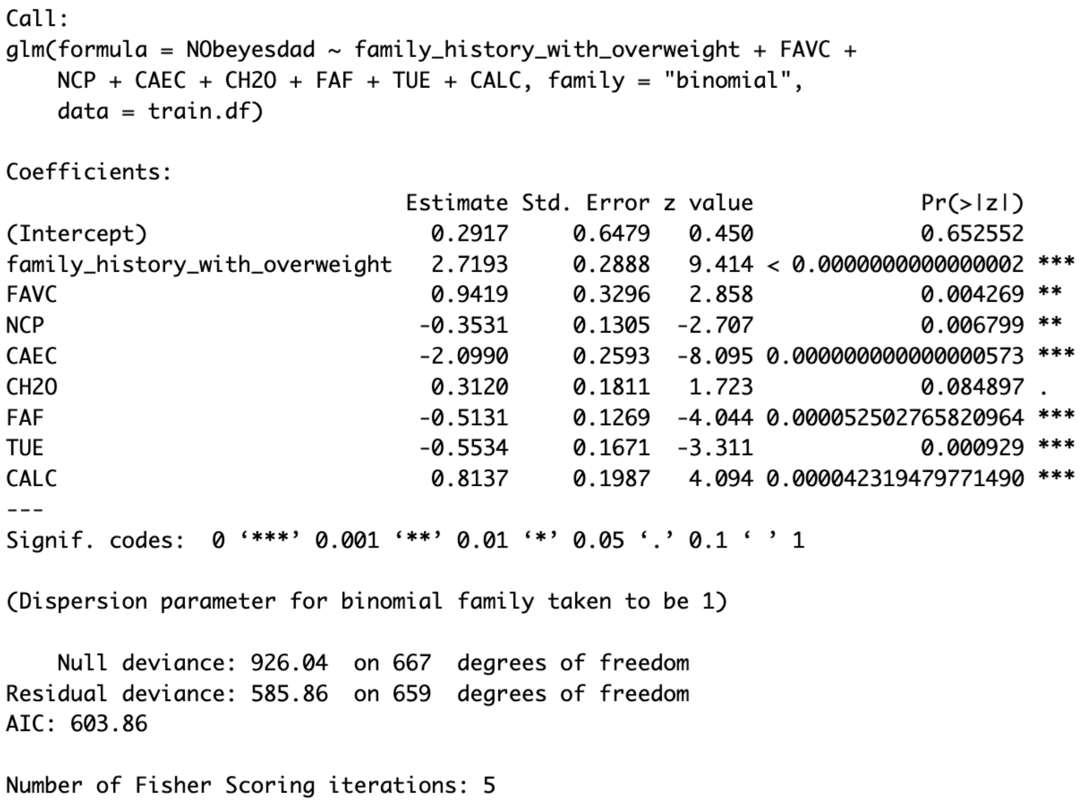
```
Confusion Matrix and Statistics

          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
```

# 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
```

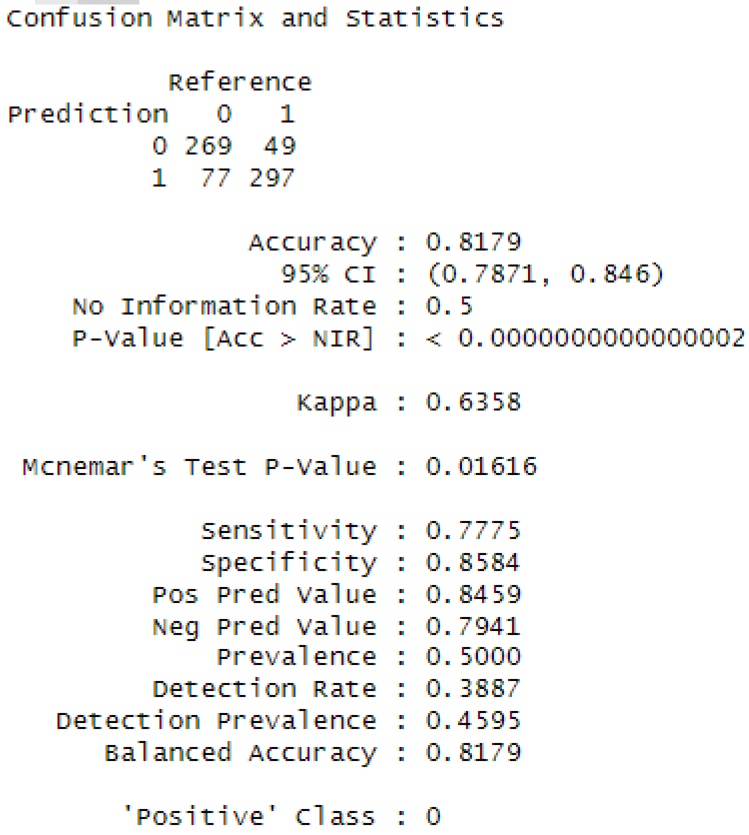
# Logistic Regression with Backward Elimination



```
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
```

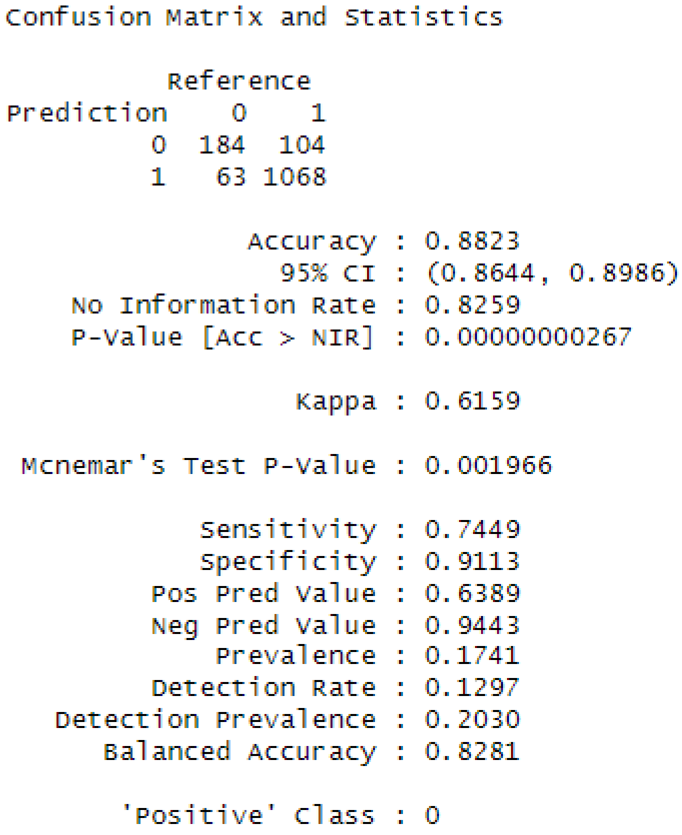
# Logistic Regression with Backward Elimination

## Confusion Matrix for Training Dataset



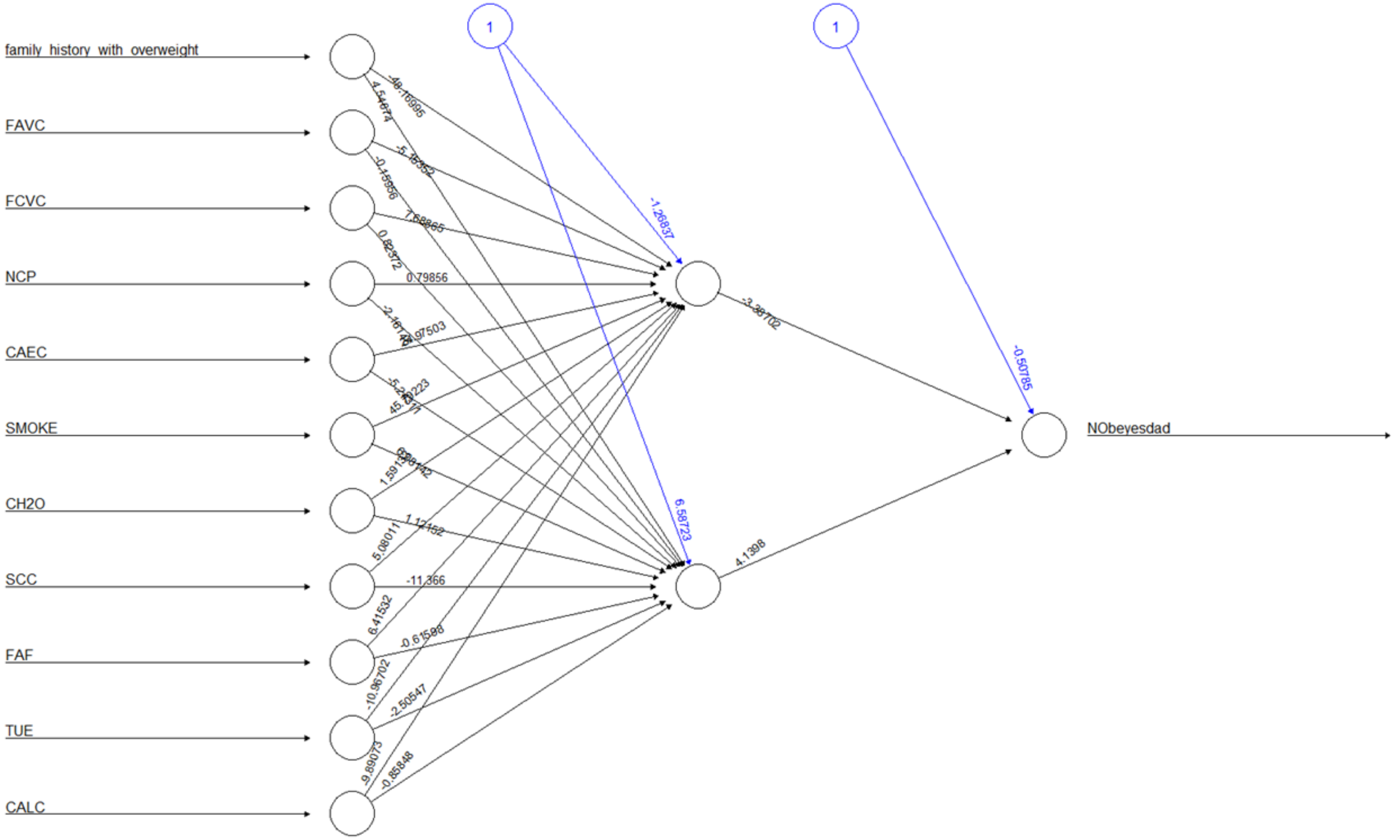
```
Confusion Matrix and Statistics

          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
```

## Confusion Matrix for Validation Dataset



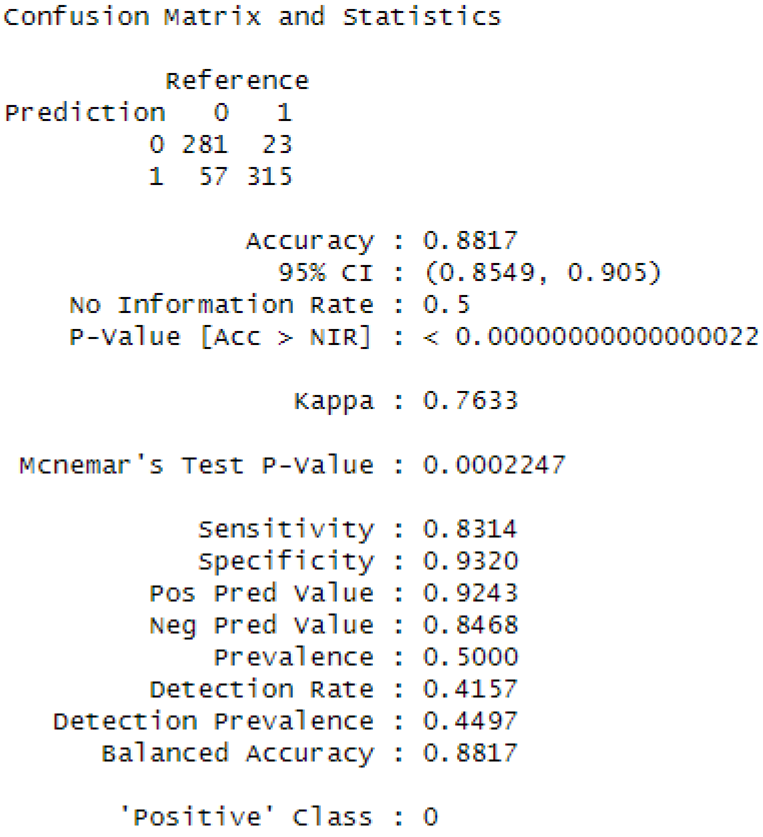
```
Confusion Matrix and Statistics

          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] : 0.0000000267

                  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
```

# Neural Network



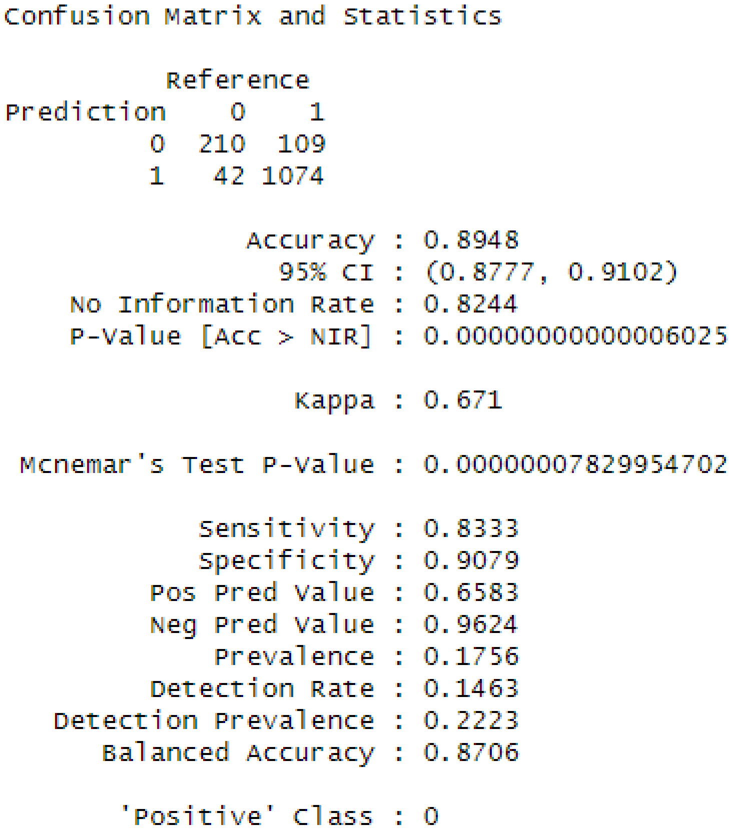
# Neural Network

## Confusion Matrix for Training Dataset



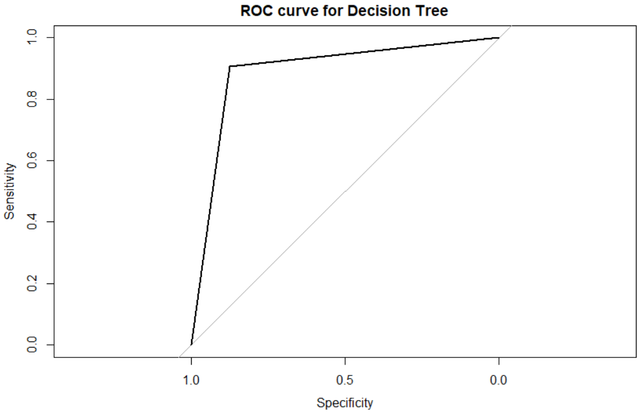
```
Confusion Matrix and Statistics

          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
```

## Confusion Matrix for Validation Dataset

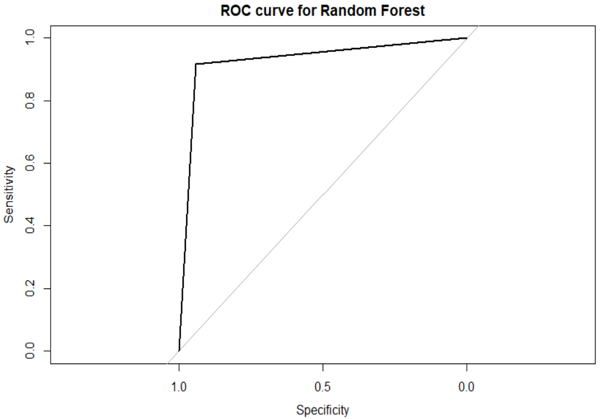


```
Confusion Matrix and Statistics

          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] : 0.0000000000006025

                  Kappa : 0.671

 Mcnemar's Test P-Value : 0.00000007829954702

            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
```

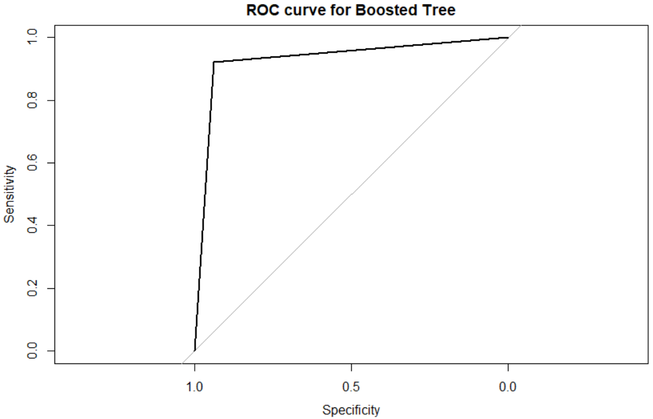
# ROC curve comparison



AUC: 0.8911

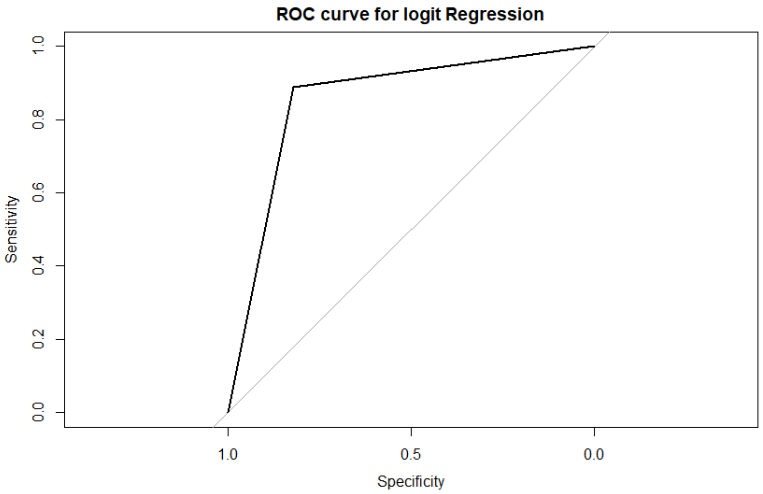


AUC: 0.9307

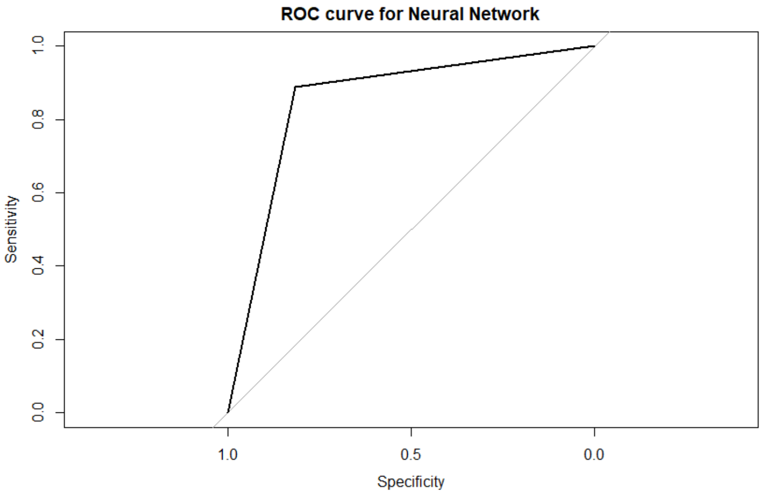


AUC: 0.9313

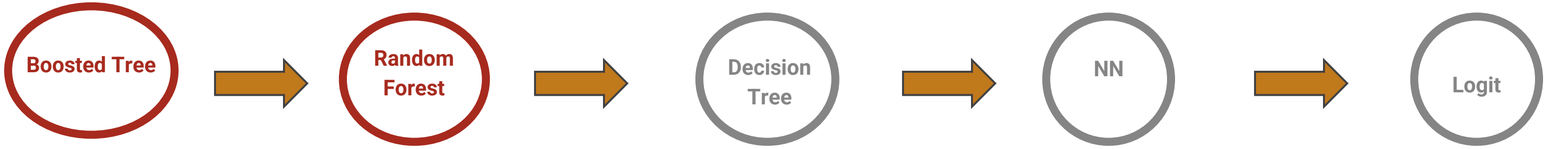
# ROC curve comparison



AUC: 0.8560



AUC: 0.8678



Boosted Tree → Random Forest → Decision Tree → NN → Logit

AUC: 0.9313

AUC: 0.9307

AUC: 0.8911

AUC: 0.8678

AUC: 0.8560

# Model Evaluation

After evaluating decision tree, random forest, boosted tree, logistic regression and neural network model, the random forest model had the best performance in terms of the accuracy rate of ~93% on validation dataset followed by boosted tree with accuracy rate of ~92%. However, the boosted tree model had a better performance according to the roc index having highest area under the curve of 0.9313. Hence, taking the cumulative effect of using accuracy along it with ROC curve’s area, we conclude that the boosted tree model is champion model for our dataset.

# Final Conclusion from The Data

**Major lifestyle decisions that affect a person being obese or not:**

- Food consumption pattern, frequency of use of technology devices on daily basis, water intake and physical activeness are some of the important lifestyle choice based attributes that affect a person being overweight or not
- Genetics, heredity or family history with obesity also plays an important factor

**Some important takeaways and special considerations:**

- Lifestyle trends and individual choices may vary over time and region wise, hence it’s advisable to regularly assess and update the model with these changes
- This analysis is for particular set of countries including Mexico, Peru and Colombia
- There may have been other confounding factors, and the model may not be representative of the whole population. This analysis is based on the given dataset and doesn’t generalise to the general population
- If a representative sample is obtained, this model can be helpful when deployed in healthcare facilities to assess the likelihood of a person being obese
- The model can also be helpful to device healthcare policies in order to check this menace of obesity and the risk of other health problems related to it

# Thank You