Multi Class Prediction of Obesity Risk

Project Overview

- Developing a model which will accurately predict obesity risk among individuals based on varios lifestyle factors

Dataset

- **id**: A unique identifier for each individual in the dataset.
- **Gender**: The individual's gender, indicating whether they are male or female.
- **Age**: The age of the individual, representing their age in years.
- **Height**: The height of the individual, typically measured in meters.
- **Weight**: The weight of the individual, typically measured in kilograms.
- **family_history_with_overweight**: Indicates whether there is a family history of overweight for the individual (yes/no).
- **FAVC**: Stands for "Frequency of consuming high caloric food," representing how often the individual consumes high-calorie foods (yes/no).
- **FCVC**: Stands for "Frequency of consuming vegetables," representing how often the individual consumes vegetables.
- **NCP**: Stands for "Number of main meals," indicating the number of main meals the individual consumes daily.
- **CAEC**: Stands for "Consumption of food between meals," representing the frequency of consuming food between meals.
- **SMOKE**: Indicates whether the individual smokes or not (yes/no).
- **CH2O**: Represents the amount of water consumption for the individual.
- **SCC**: Stands for "Calories consumption monitoring," indicating whether the individual monitors their calorie consumption (yes/no).
- **FAF**: Stands for "Physical activity frequency," representing the frequency of the individual's physical activities.
- **TUE**: Stands for "Time using technology devices," indicating the amount of time the individual spends using technology devices.
- **CALC**: Stands for "Consumption of alcohol," representing the frequency of alcohol consumption.
- **MTRANS**: Stands for "Mode of transportation," indicating the mode of transportation the individual uses.
- **NObeyesdad**: The target variable, representing the obesity risk category of the individual. It has multiple classes such as 'Overweight_Level_II', 'Normal_Weight', 'Insufficient_Weight', 'Obesity_Type_III', 'Obesity_Type_II', 'Overweight_Level_I', and 'Obesity_Type_I'.

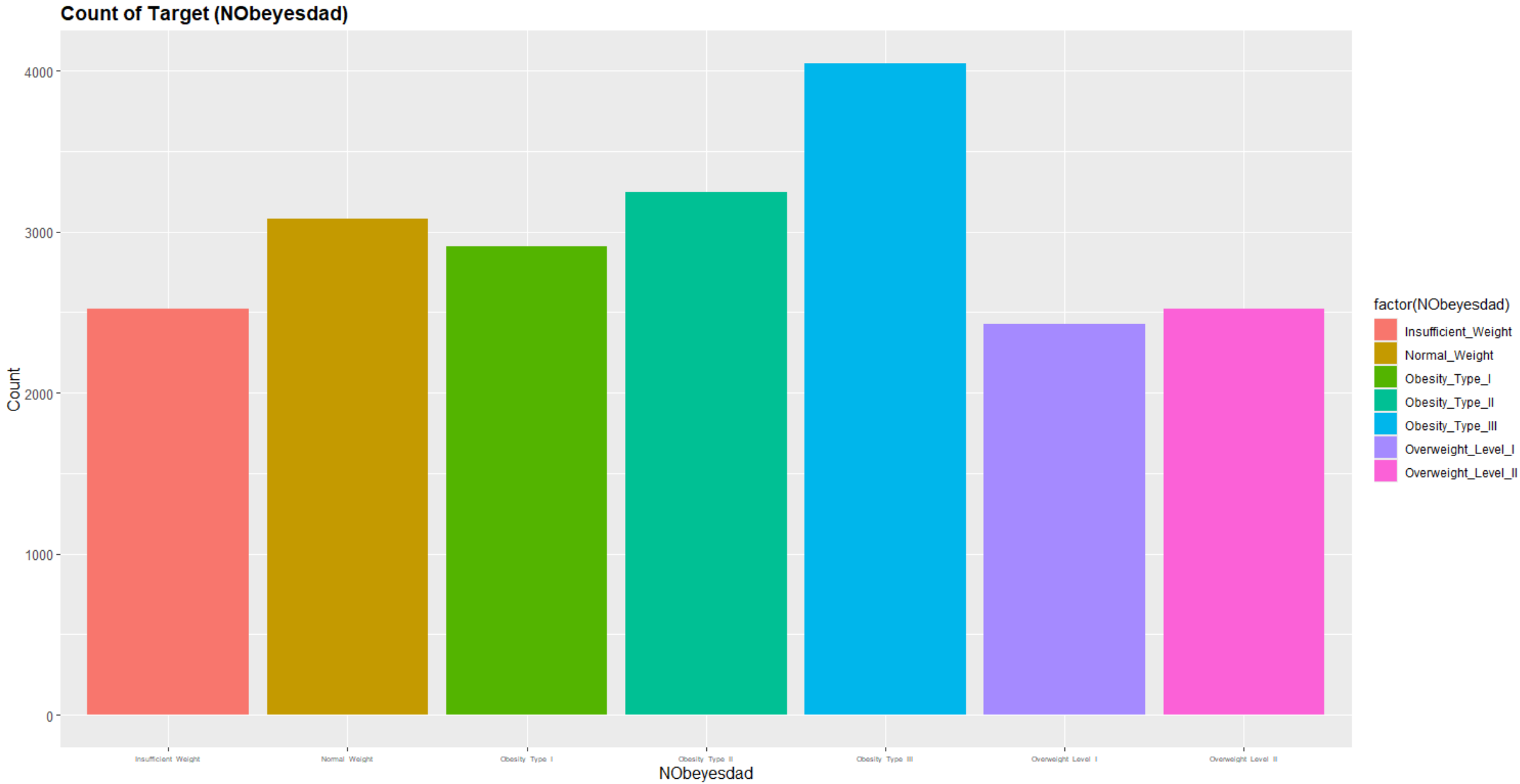
Pre-Processsing

Overviewing the Dataset

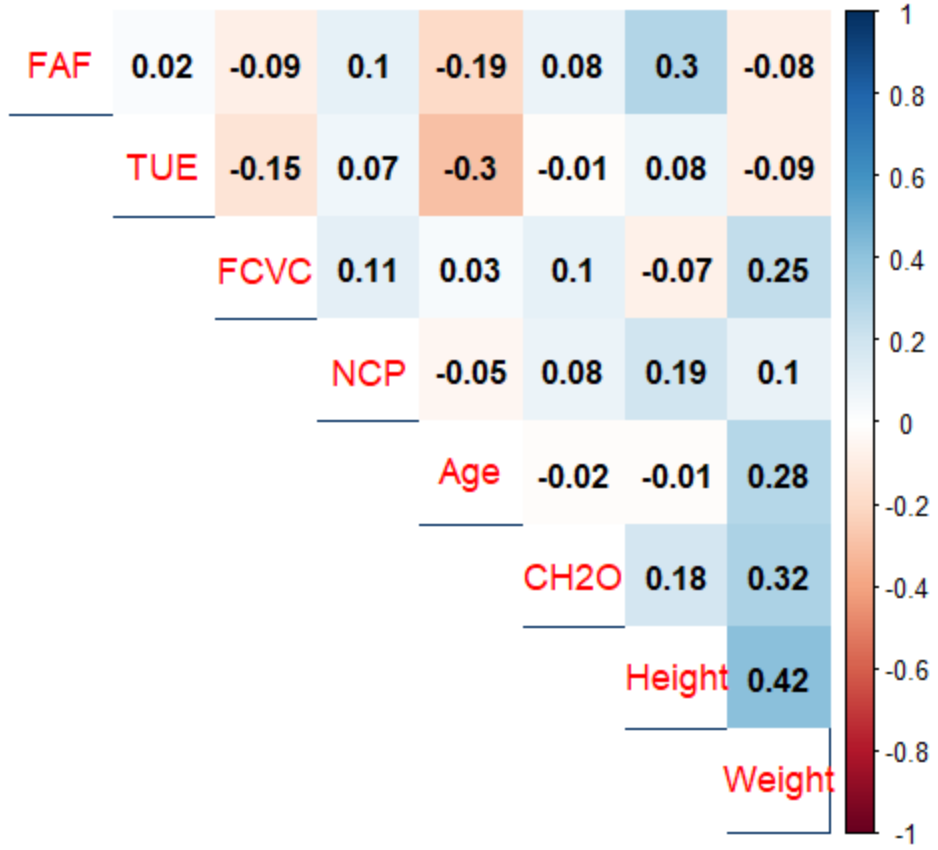
Missing Values
and duplicates

Encoding all variables into
numerical variables

Feature Engineering


Count of Target (NObeyesdad)
Count
4000
3000
2000
1000
0
Insufficient Weight
Normal Weight
Obesity Type I
Obesity Type II
Obesity Type III
Overweight Level I
Overweight Level II
NObeyesdad
factor(NObeyesdad)
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II

Correlation between numerical variables



Missing Values and Duplicates

```
##                        Gender                                   Age
##                             0                                     0
##                        Height                                Weight
##                             0                                     0
## family_history_with_overweight                                FAVC
##                             0                                     0
##                          FCVC                                   NCP
##                             0                                     0
##                          CAEC                                SMOKE
##                             0                                     0
##                          CH2O                                   SCC
##                             0                                     0
##                           FAF                                   TUE
##                             0                                     0
##                          CALC                                MTRANS
##                             0                                     0
##                    NObeyesdad
##                             0
```

```
# duplicates across train/ test
sum(
    duplicated(rbind(df_train[, -17],df_test))
    )
```

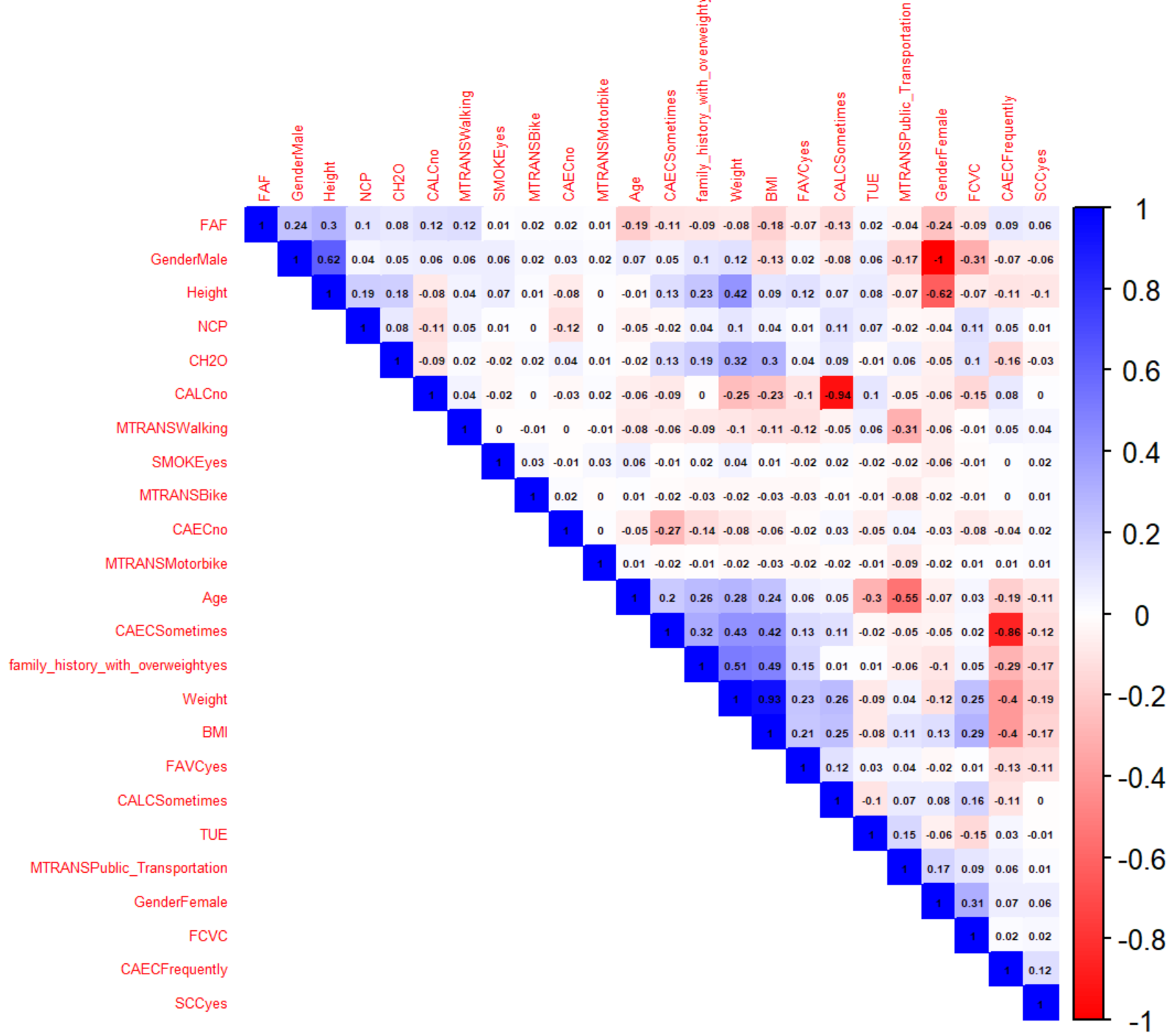
```
## [1] 0
```

```
# null values by column
colSums(is.na(df_train))
```

Adding BMI column adjusted by age and gender

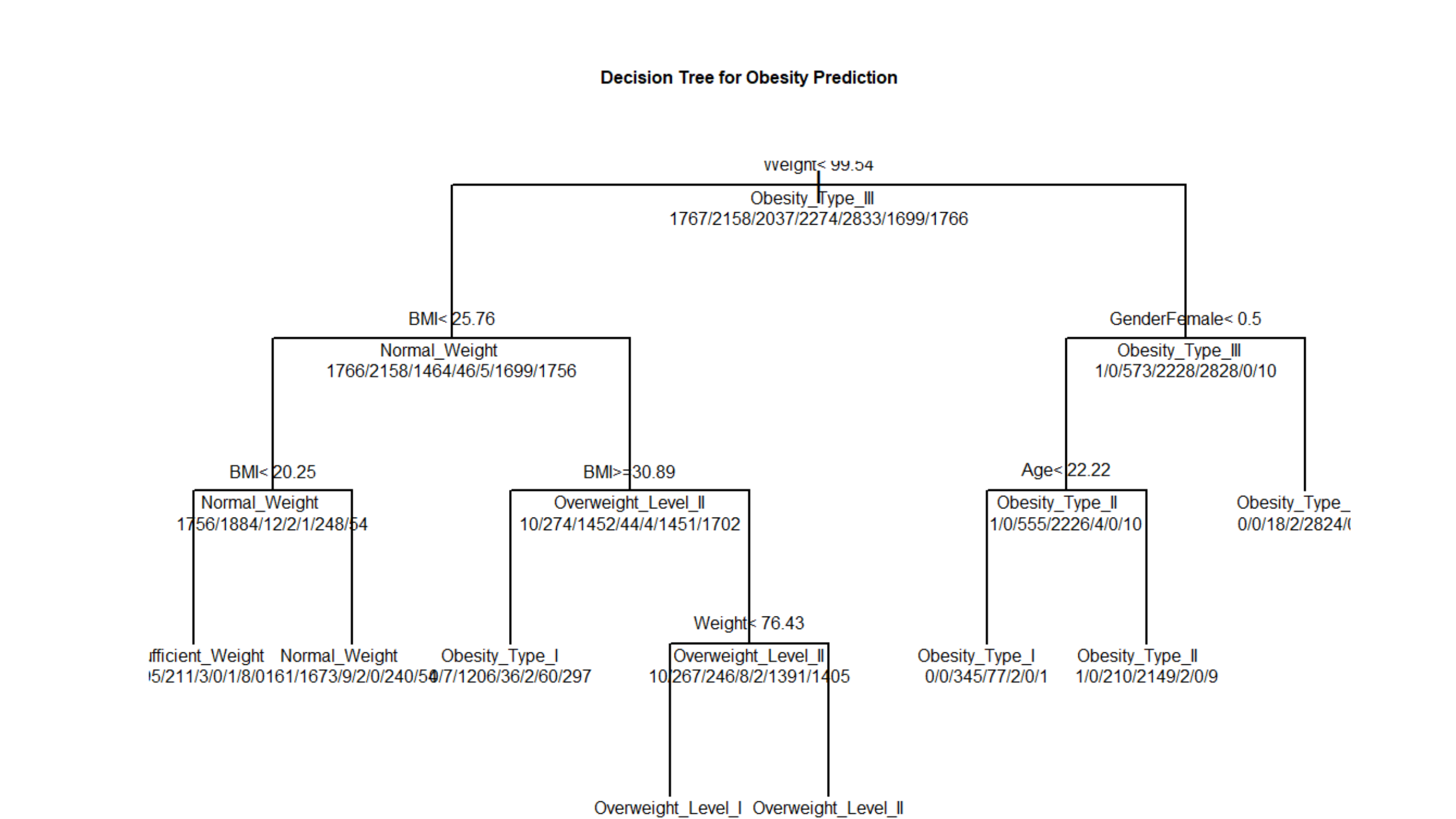
```
calculate_BMI <- function(weight, height, age, gender_male) {
 bmi <- weight / (height ^ 2)

 # Adjust BMI for age and gender
 if (gender_male == 1) {
  if (age < 18) {
   bmi <- bmi * 1.1
  } else if (age >= 18 & age <= 24) {
   bmi <- bmi * 1.05
  }
 } else {
  if (age < 18) {
   bmi <- bmi * 1.15
  } else if (age >= 18 & age <= 24) {
   bmi <- bmi * 1.08
  }
 }
```

```
GenderFemale                          : num [1:20758] 0 1 1 1 0 0 0 0
GenderMale                            : num [1:20758] 1 0 0 0 1 1 1 1
Age                                   : num [1:20758] 24.4 18 18 21 31
Height                                : num [1:20758] 1.7 1.56 1.71 1.
Weight                                : num [1:20758] 81.7 57 50.2 131
family_history_with_overweightyes: num [1:20758] 1 1 1 1 1 1 1 1
FAVCyes                               : num [1:20758] 1 1 1 1 1 1 1 1
FCVC                                  : num [1:20758] 2 2 1.88 3 2.68
NCP                                   : num [1:20758] 2.98 3 1.41 3 1.
CAECFrequently                        : num [1:20758] 0 1 0 0 0 0 0 0
CAECno                                : num [1:20758] 0 0 0 0 0 0 0 0
CAECSometimes                         : num [1:20758] 1 0 1 1 1 1 1 1
SMOKEyes                              : num [1:20758] 0 0 0 0 0 0 0 0
CH2O                                  : num [1:20758] 2.76 2 1.91 1.67
SCCyes                                : num [1:20758] 0 0 0 0 0 0 0 0
FAF                                   : num [1:20758] 0 1 0.866 1.468
TUE                                   : num [1:20758] 0.976 1 1.674 0.
CALCno                                : num [1:20758] 0 1 1 0 0 0 0 0
CALCSometimes                         : num [1:20758] 1 0 0 1 1 1 1 1
MTRANSBike                            : num [1:20758] 0 0 0 0 0 0 0 0
MTRANSMotorbike                       : num [1:20758] 0 0 0 0 0 0 0 0
MTRANSPublic_Transportation           : num [1:20758] 1 0 1 1 1 1 0 0
MTRANSWalking                         : num [1:20758] 0 0 0 0 0 0 0 0
NObeyesdad                            : Factor w/ 7 levels "Insufficie
BMI                                   : num [1:20758] 28.3 25.3 18.5 4
```


FAF
GenderMale
Height
NCP
CH2O
CALCno
MTRANSWalking
SMOKEyes
MTRANSBike
CAECno
MTRANSMotorbike
Age
CAECSometimes
family_history_with_overweightyes
Weight
BMI
FAVCyes
CALCSometimes
TUE
MTRANSPublic_Transportation
GenderFemale
FCVC
CAECFrequently
SCCyes
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Models used to compute output

- Decision tree
- Logistic regression
- Random forest


Decision Tree for Obesity Prediction
Weight< 99.54
Obesity_Type_III
1767/2158/2037/2274/2833/1699/1766
BMI< 25.76
Normal_Weight
1766/2158/1464/46/5/1699/1756
GenderFemale< 0.5
Obesity_Type_III
1/0/573/2228/2828/0/10
BMI< 20.25
Normal_Weight
1756/1884/12/2/1/248/54
BMI>=30.89
Overweight_Level_II
10/274/1452/44/4/1451/1702
Age< 22.22
Obesity_Type_II
1/0/555/2226/4/0/10
Obesity_Type_
0/0/18/2/2824/(
Weight< 76.43
Overweight_Level_II
10/267/246/8/2/1391/1405
Obesity_Type_I
0/0/345/77/2/0/1
Obesity_Type_II
1/0/210/2149/2/0/9
Overweight_Level_I
Overweight_Level_II

Decision tree

- Confusion Matrix and Statistics
-
- Reference
- Prediction Insufficient_Weight Normal_Weight Obesity_Type_I Obesity_Type_II Obesity_Type_III Overweight_Level_I Overweight_Level_II
- Insufficient_Weight 673 75 0 0 0 6 0
- Normal_Weight 77 753 3 0 0 87 16
- Obesity_Type_I 0 1 643 52 0 21 117
- Obesity_Type_II 0 0 92 920 0 0 7
- Obesity_Type_III 0 0 10 1 1211 0 1
- Overweight_Level_I 5 92 29 0 1 563 144
- Overweight_Level_II 1 3 96 1 1 51 471

Overall Statistics
Accuracy : 0.8409
95% CI : (0.8316, 0.8499)
No Information Rate : 0.1949
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.8134
Mcnemar's Test P-Value : NA

Logistic Regression

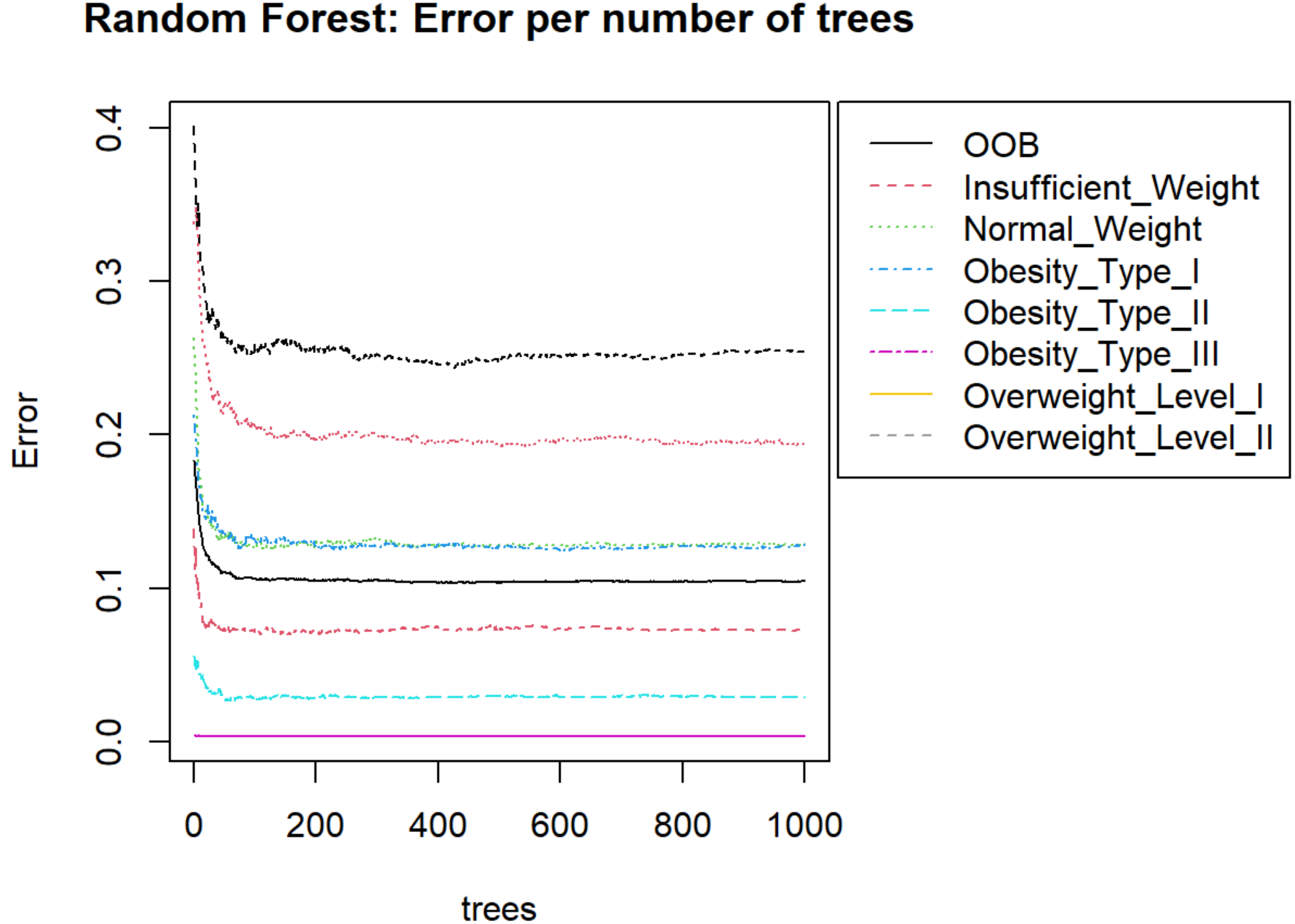
Prediction	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	693	97	0	0	0	8	0
Normal_Weight	59	757	2	0	0	74	16
Obesity_Type_I	0	0	714	37	1	13	79
Obesity_Type_II	0	0	47	928	1	0	6
Obesity_Type_III	0	0	6	0	1210	0	1
Overweight_Level_I	2	62	27	0	1	513	103
Overweight_Level_II	2	8	77	9	0	120	551

Overall Statistics Accuracy : 0.8621
95% CI : (0.8533, 0.8706)
No Information Rate : 0.1949
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.8383
Mcnemar's Test P-Value : NA

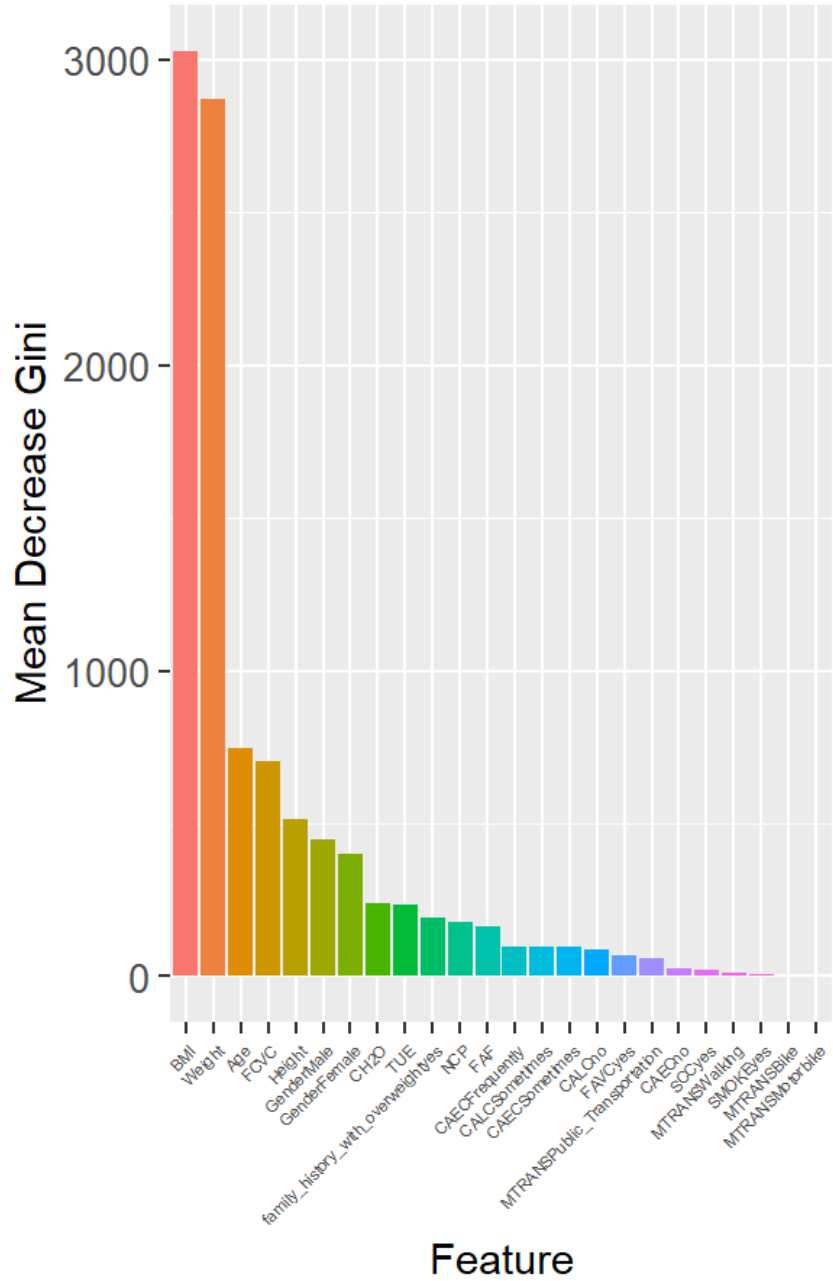
Random Forest

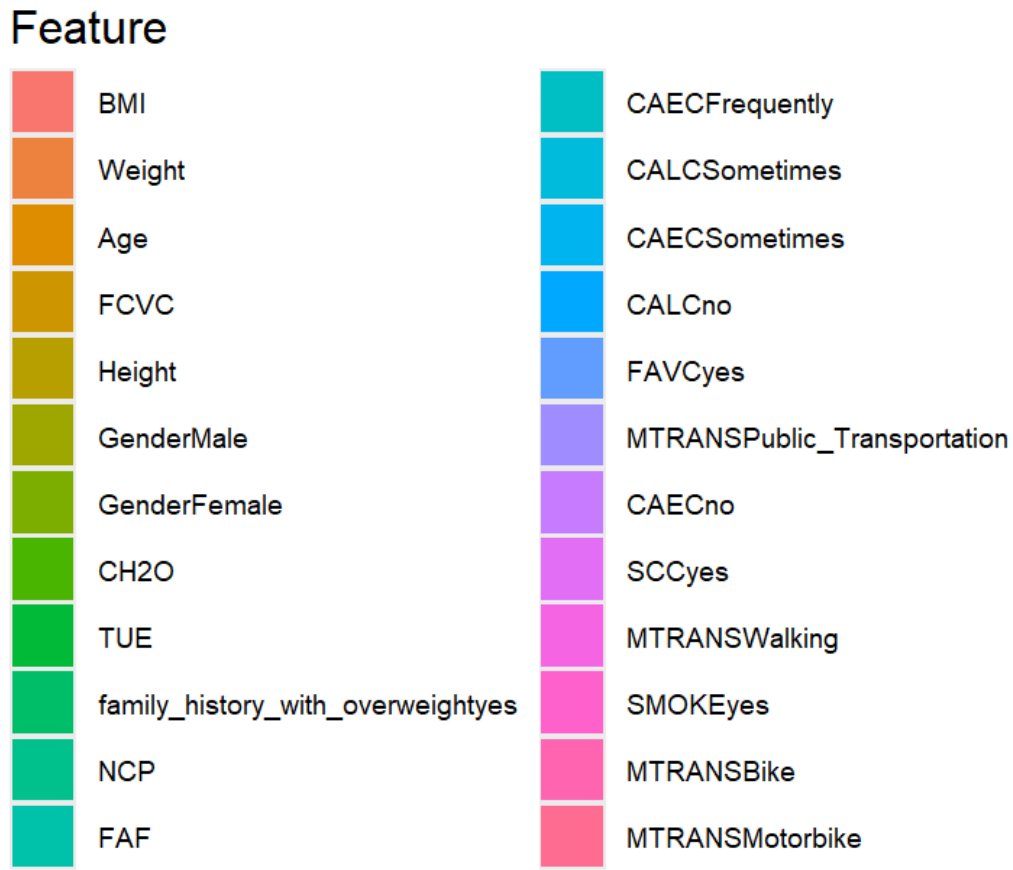
Prediction	Insufficient_Weight	Normal_Weight	Obesity_Type_I	Obesity_Type_II	Obesity_Type_III	Overweight_Level_I	Overweight_Level_II
Insufficient_Weight	697	41	0	0	0	6	0
Normal_Weight	53	825	4	0	1	85	17
Obesity_Type_I	0	1	774	14	1	18	48
Obesity_Type_II	0	0	30	951	0	0	7
Obesity_Type_III	0	0	4	1	1211	0	0
Overweight_Level_I	5	44	11	0	0	530	50
Overweight_Level_II	1	13	50	8	0	89	634

Overall Statistics
Accuracy : 0.9026
95% CI : (0.8965, 0.9113)
No Information Rate : 0.1949
P-Value [Acc > NIR] : < 2.2e-16
Kappa : 0.8875
Mcnemar's Test P-Value : NA


Random Forest: Error per number of trees
Error
0.0
0.1
0.2
0.3
0.4
0
200
400
600
800
1000
trees
OOB
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II

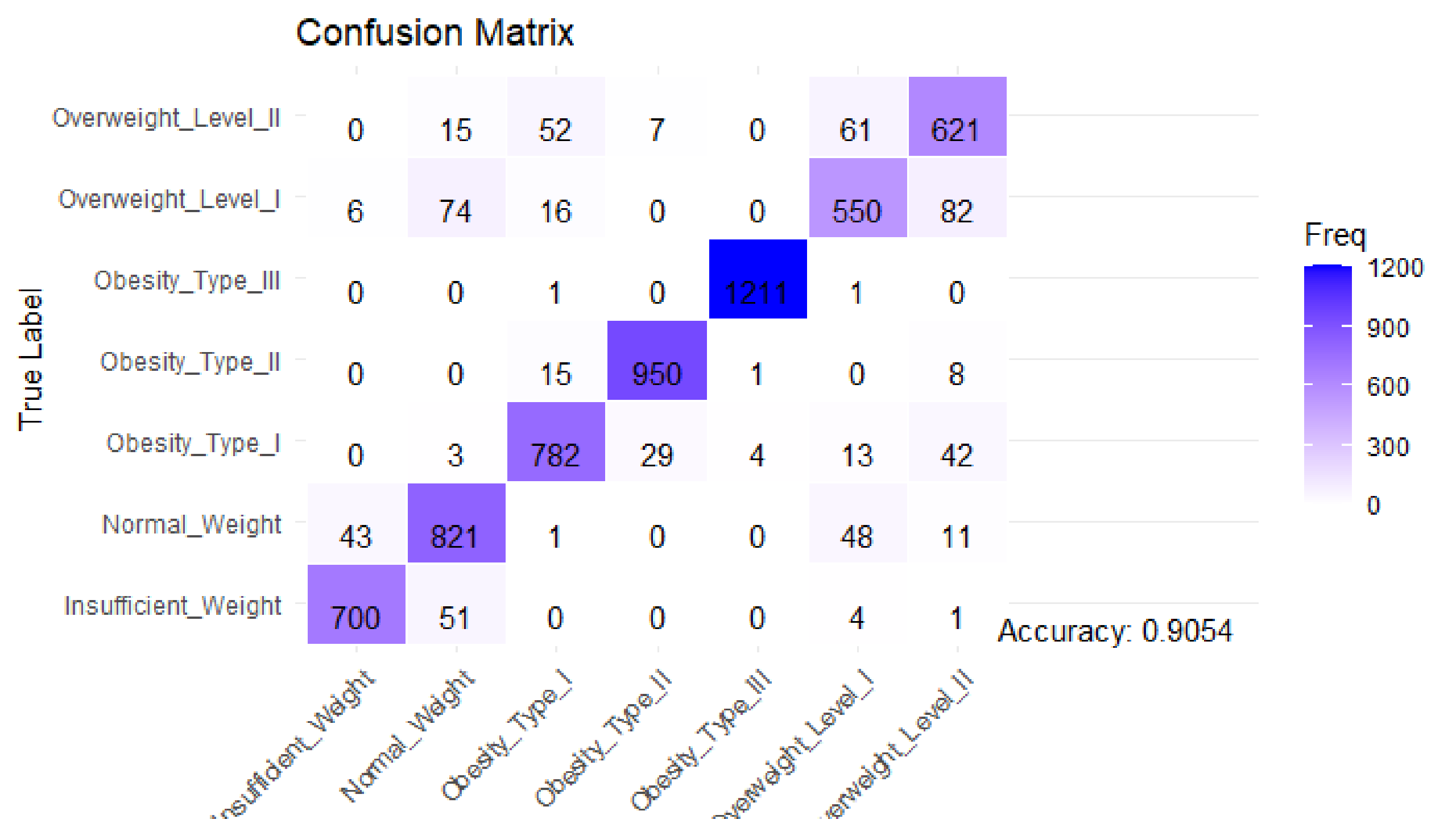
Feature Importances from Random Forest Model

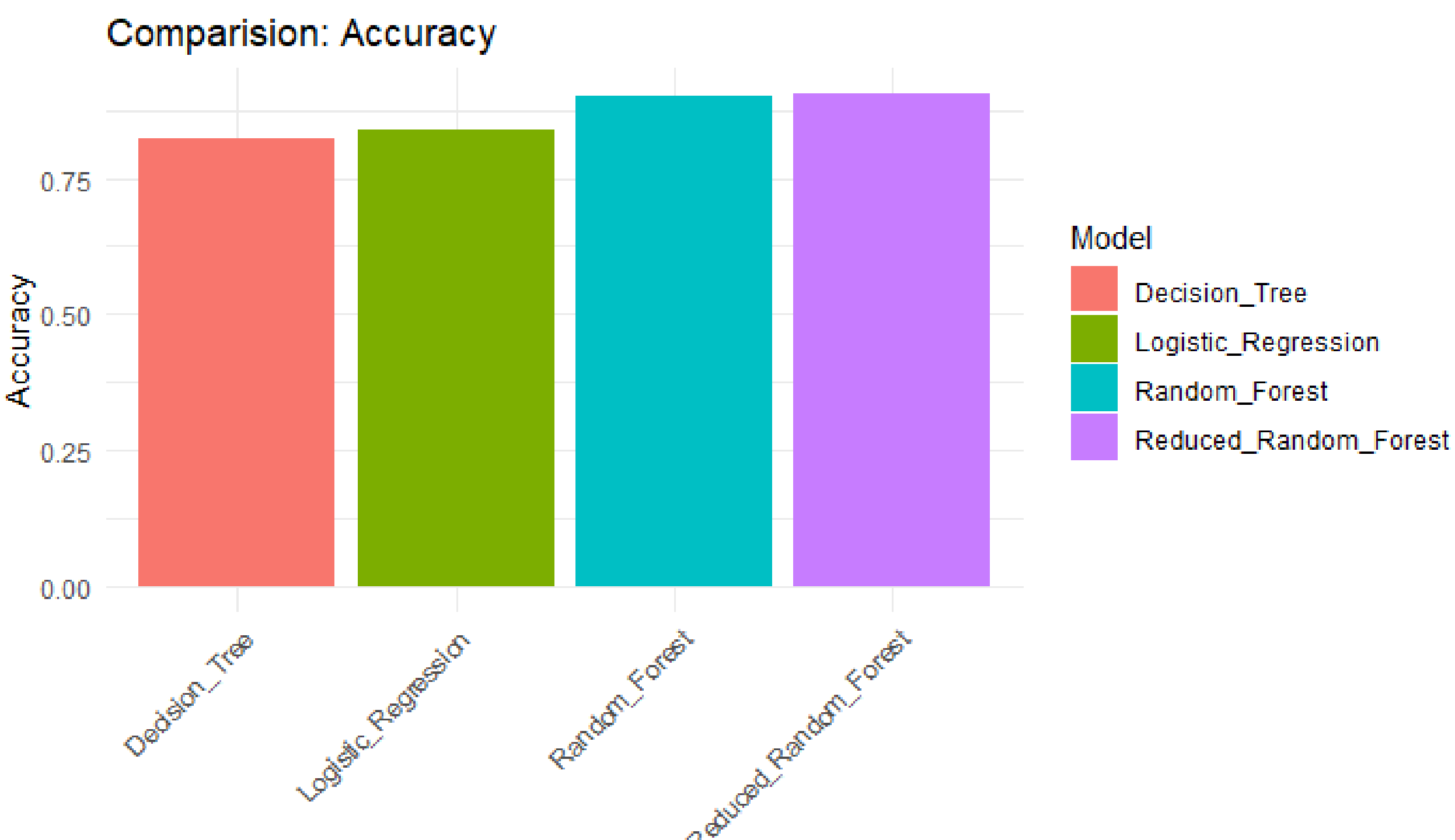




Improving the random forest model

- Deleting features with low importance and high correlation


Confusion Matrix
True Label
Overweight_Level_II
Overweight_Level_I
Obesity_Type_III
Obesity_Type_II
Obesity_Type_I
Normal_Weight
Insufficient_Weight
0 15 52 7 0 61 621
6 74 16 0 0 550 82
0 0 1 0 1211 1 0
0 0 15 950 1 0 8
0 3 782 29 4 13 42
43 821 1 0 0 48 11
700 51 0 0 0 4 1
Insufficient_Weight
Normal_Weight
Obesity_Type_I
Obesity_Type_II
Obesity_Type_III
Overweight_Level_I
Overweight_Level_II
Freq
1200
900
600
300
0
Accuracy: 0.9054


Comparision: Accuracy
Accuracy
0.75
0.50
0.25
0.00
Decision_Tree
Logistic_Regression
Random_Forest
Reduced_Random_Forest
Model
Decision_Tree
Logistic_Regression
Random_Forest
Reduced_Random_Forest