ANALYSIS OF OBESITY LEVELS BASED ON EATING HABITS AND PHYSICAL CONDITIONS

STAT 421: PROJECT REPORT

NAME: Anupriya Halder
SEMESTER: 4
UNIVERSITY ROLL NUMBER: C91/STS/201006
REGISTRATION NUMBER: A01-2112-0786-17
PROJECT SUPERVISORS: Prof. Gaurangadeb Chattopadhyay and Prof. Bhaswati Ganguli
UNIVERSITY OF CALCUTTA

INTRODUCTION

- Obesity is defined as abnormal or excessive fat accumulation that may impair health. It has important consequences for morbidity, disability and quality of life.
- The objective of this study is to observe which social factors (like Gender and Age), eating habits (like consumption of high caloric food, vegetables, water intake etc.), smoking & drinking habits and movement habits (like physical activity, transportation used etc.) mostly impact obesity so that by taking appropriate measures obese conditions may be cured and also prevented.

THE DATASET

- **The dataset contains estimated obesity levels based on eating habits and physical condition of people from Mexico, Peru and Columbia. The dataset includes 2112 individuals aged 14 to 61. There are 17 attributes, namely**
- **Gender: Female / Male**
- **Age: numeric**
- **Height (in Meters): numeric;**
- **Weight(in Kg): numeric**
- **Family history of obesity : yes / no**
- **Frequent consumption of high caloric food : yes / no**
- **Frequency of consumption of vegetables : 1 = never, 2 = sometimes, 3 = always**
- **Number of main meals : 1=less than 3 meals, 2=more than 3 meals**
- **Consumption of food between meals : 1 = sometimes, 2 = frequently**
- **Smoke: yes / no**
- **Consumption of water : 1or 2 ;1 = less than 2 litres, 2 = more than 2 litres**
- **Consumption of Alcohol : 1 = never, 2 = frequently**
- **Calories consumption monitoring (CCM) : yes / no**
- **Physical activity frequency per week : 0 = none, 1 = 1 to 2 days, 2 = 3 to 5 days**
- **Time using technology devices a day (TUT) : 0 = 0-2 hours, 1 = more than 3 hours**
- **Transportation used : Automobile, public transportation, others (bike or walking).**
- **Obesity levels: Insufficient weight, Normal Weight, Level I Overweight, Level II Overweight, Type I Obesity, Type II Obesity, Type III Obesity (these categories are listed from lowest to highest body fat).**

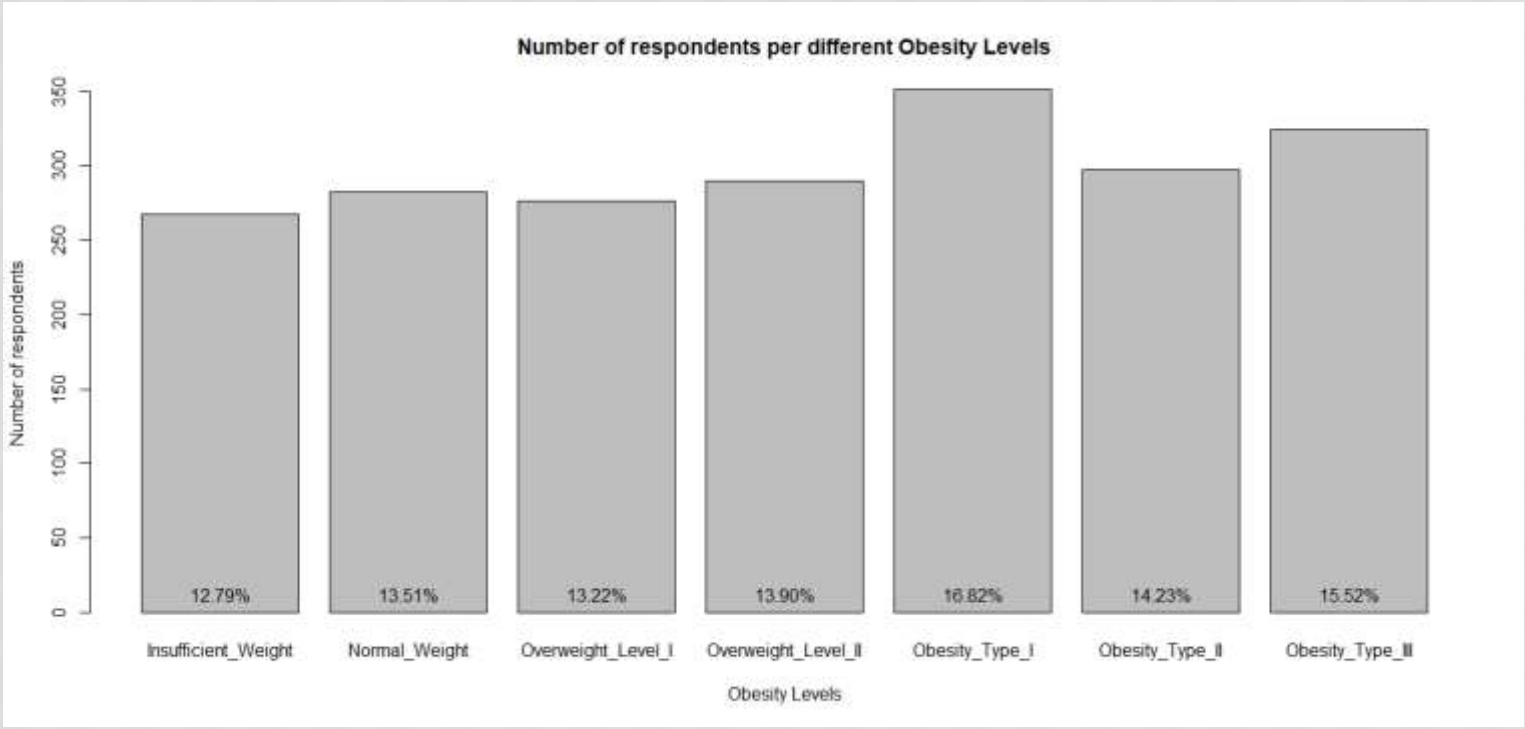
METHODOLOGY

- Exploratory data analysis of the dataset has been performed through Bar diagram, Pie chart and Scatterplots corresponding to the various factors.
- In order to explore which factors have largest influence on obesity levels, the Random Forest algorithm is implemented.
- To further investigate the impact of the different factors on the obesity levels, a proportional odds model has been fitted to the data.

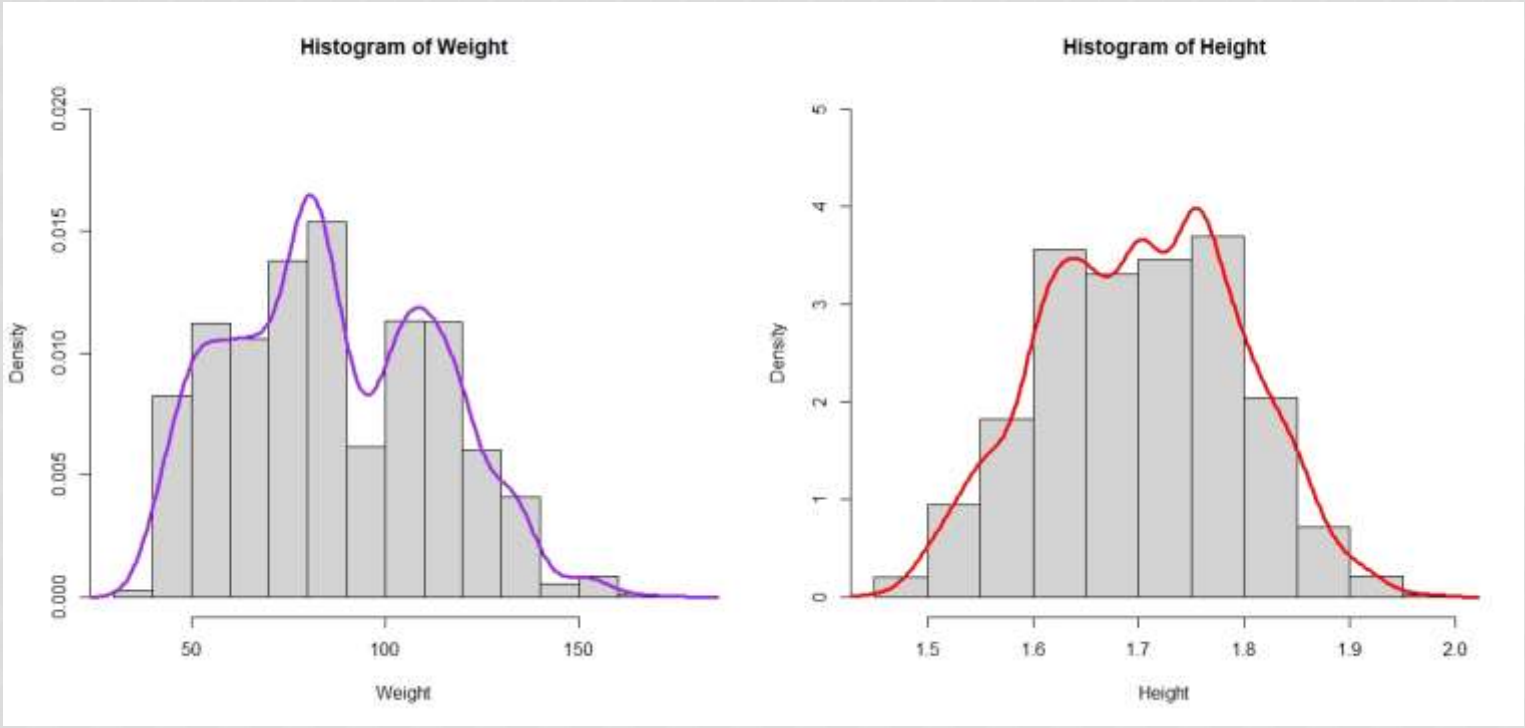
EXPLORATORY DATA ANALYSIS

- **To observe how the respondents are distributed over different obesity levels i.e., insufficient weight, normal weight,...etc**

The visualization shows that the most common obesity level is “Type I obesity” (no. Of respondents corresponding to this category is 351 (16.82%) and the least common level is “Insufficient weight” (with no. Of correspondent as 267, 12.79%). The number of respondents per category is more or less evenly distributed between the Obesity level categories and the immediate category preceding and following it.



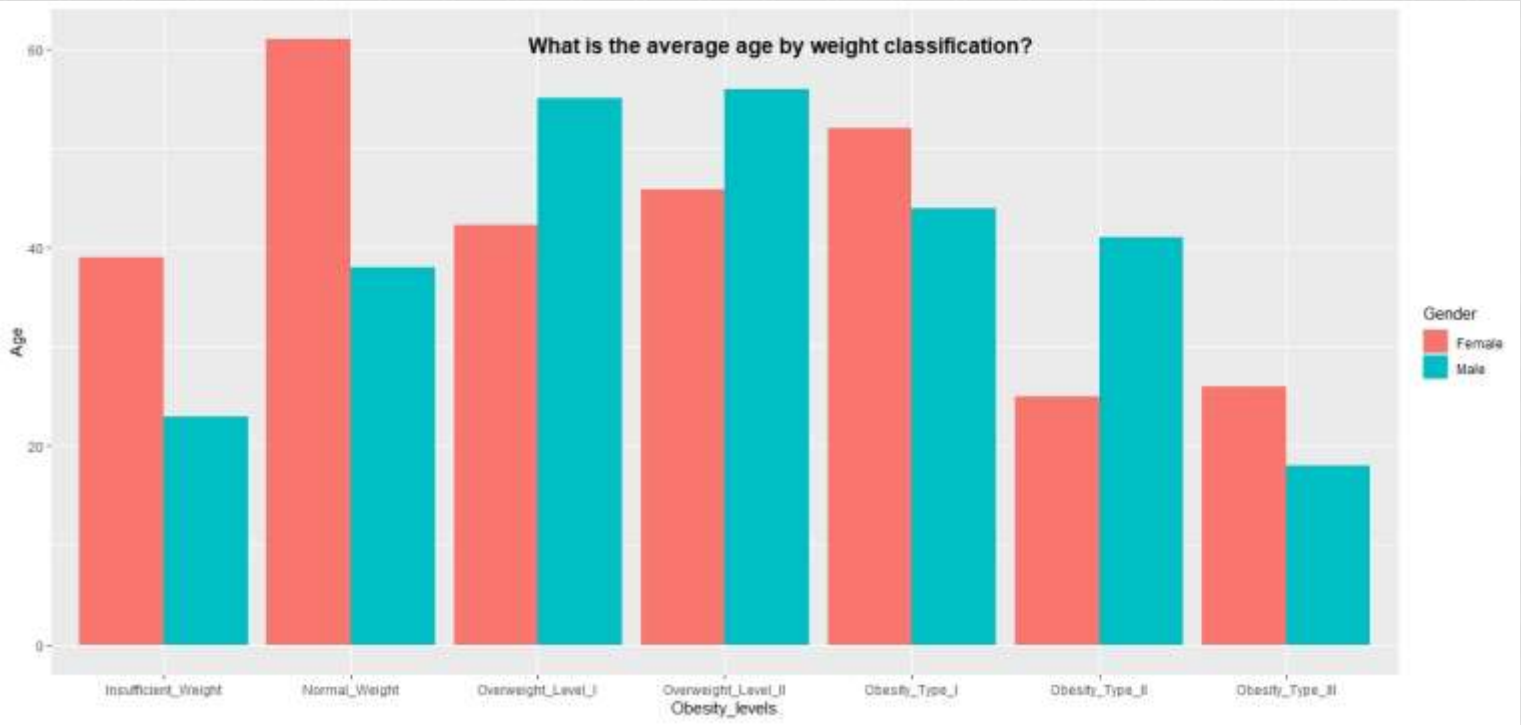
❖HOW ARE THE HEIGHTS AND WEIGHTS OF THE RESPONDENTS DISTRIBUTED?



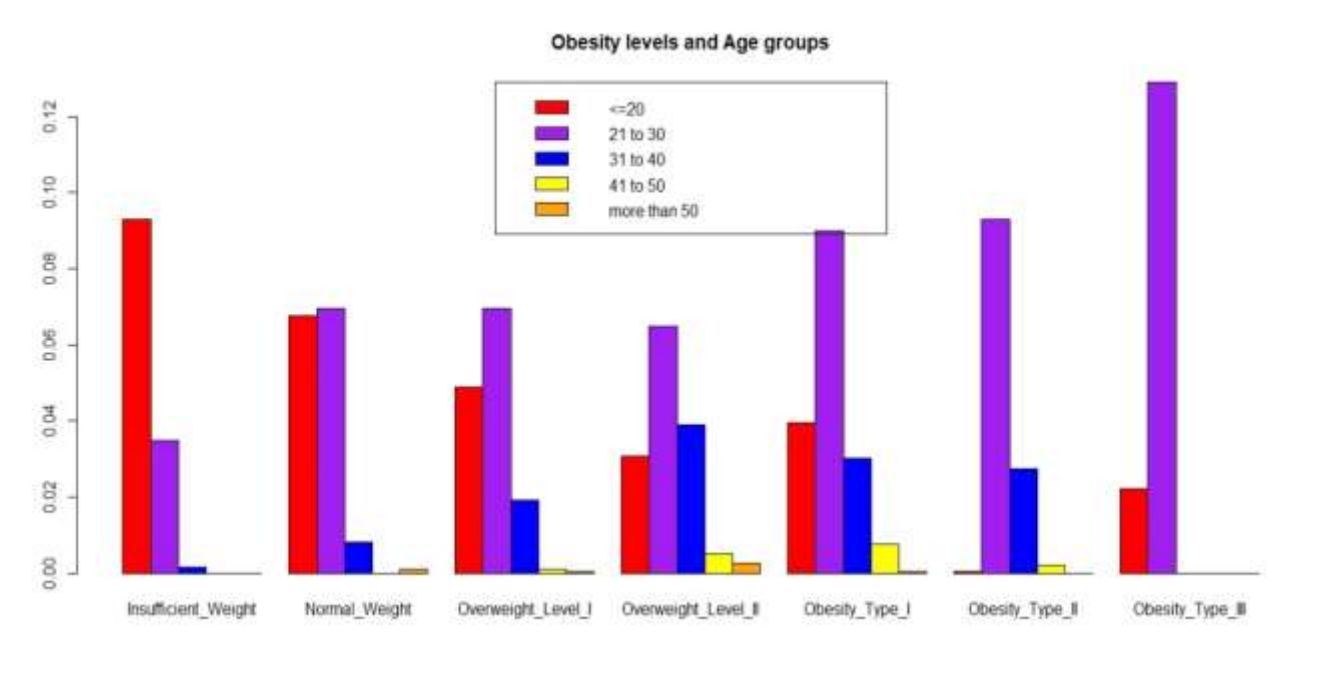
The weight data is almost bimodal and has an average around the 80kg mark, while the height data is more of symmetric and has an average around the 1.7 meters mark. Neither variable seem to be skewed.

❖What is the average age by weight classification?

To see whether Age had anything to do with whether an individual is overweight or obese bar diagrams are drawn. Also a breakdown between gender is included as gender may also have an impact on body fat. So, multiple barplot is obtained as below,

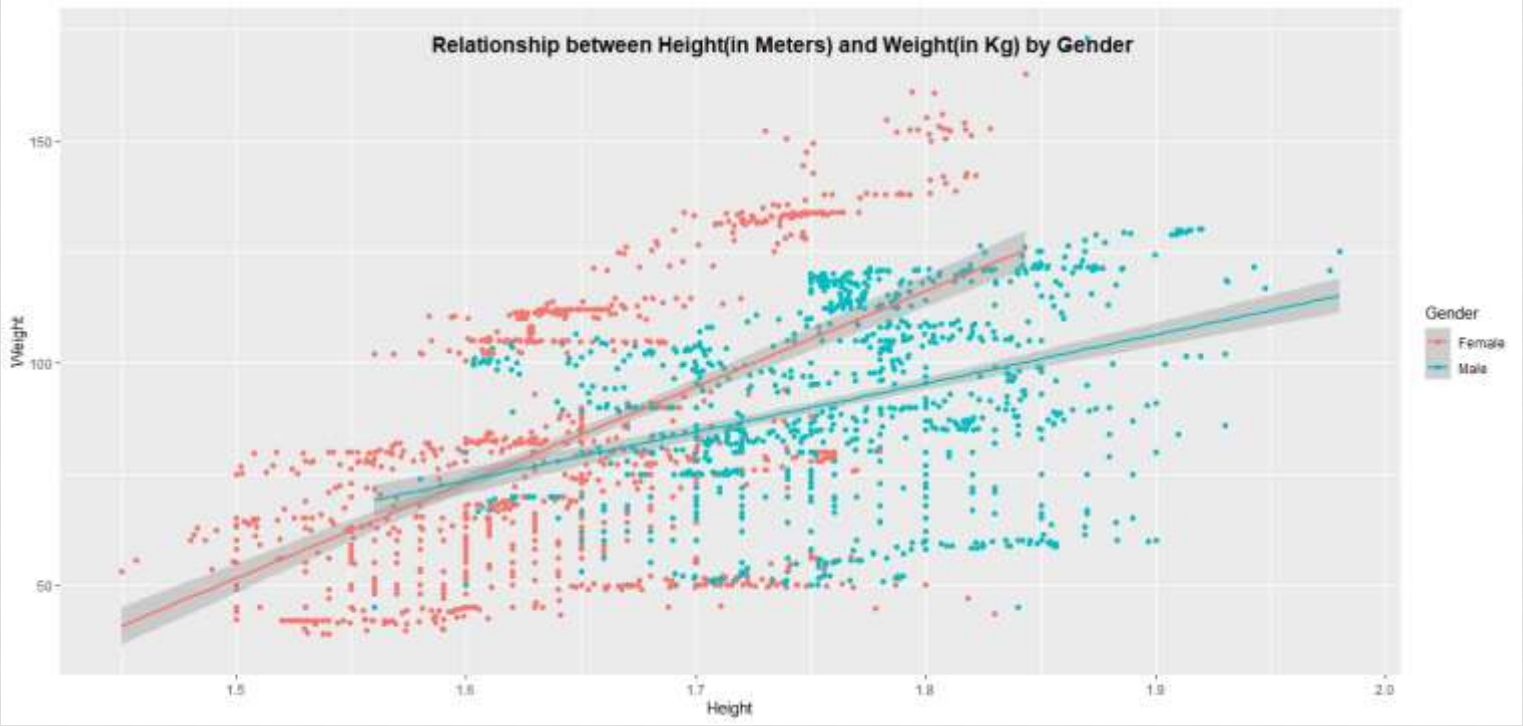


- **This graph implies that females tend to have a higher average age than males by a couple of years for each category besides overweight level I, overweight level II & type II obesity. To visualize more clearly the individuals can be grouped into age groups and the following barplot is obtained,**
- **It can also be seen that respondents with Insufficient weight, Obesity type II and Obesity type III mostly seem to be around age 20 years.**



❖WHAT IS THE RELATIONSHIP BETWEEN WEIGHT AND HEIGHT?

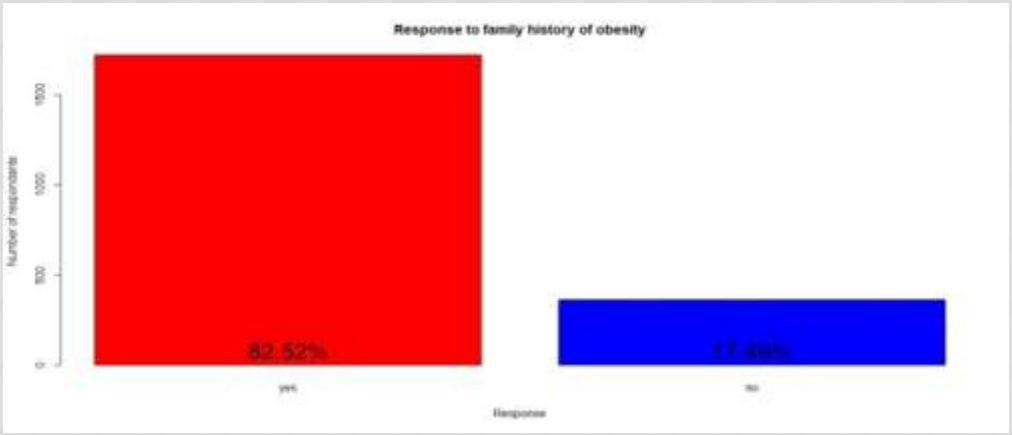
To visualize the linear relationship between the two variables Weight and Height (also as a function of a third variable, gender), through a regression,



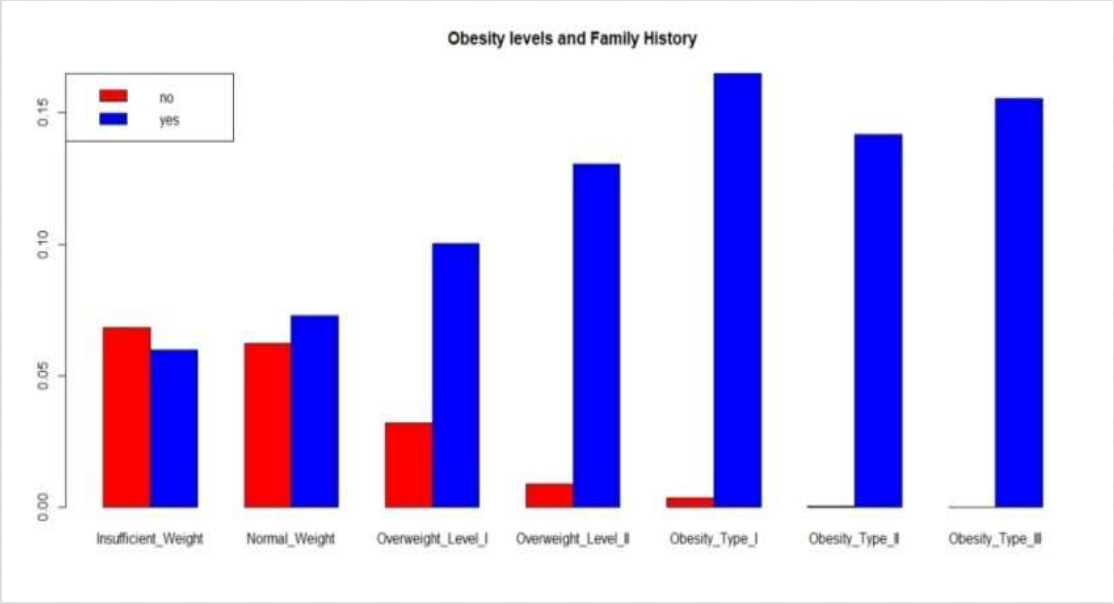
This graph shows that there is a positive association between the two variables for both female and male.

The regression line for female is slightly steeper than that of male.

FAMILY HISTORY OF OBESITY AND OBESITY LEVELS



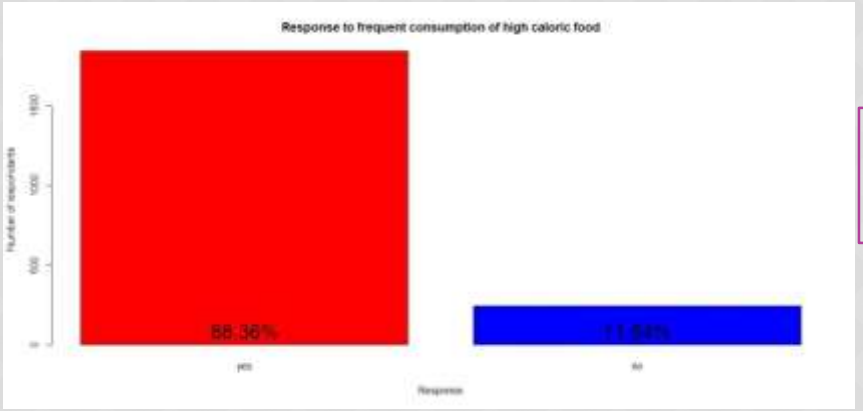
Majority of the individuals (around 82.52%) seem to have family history of obesity.



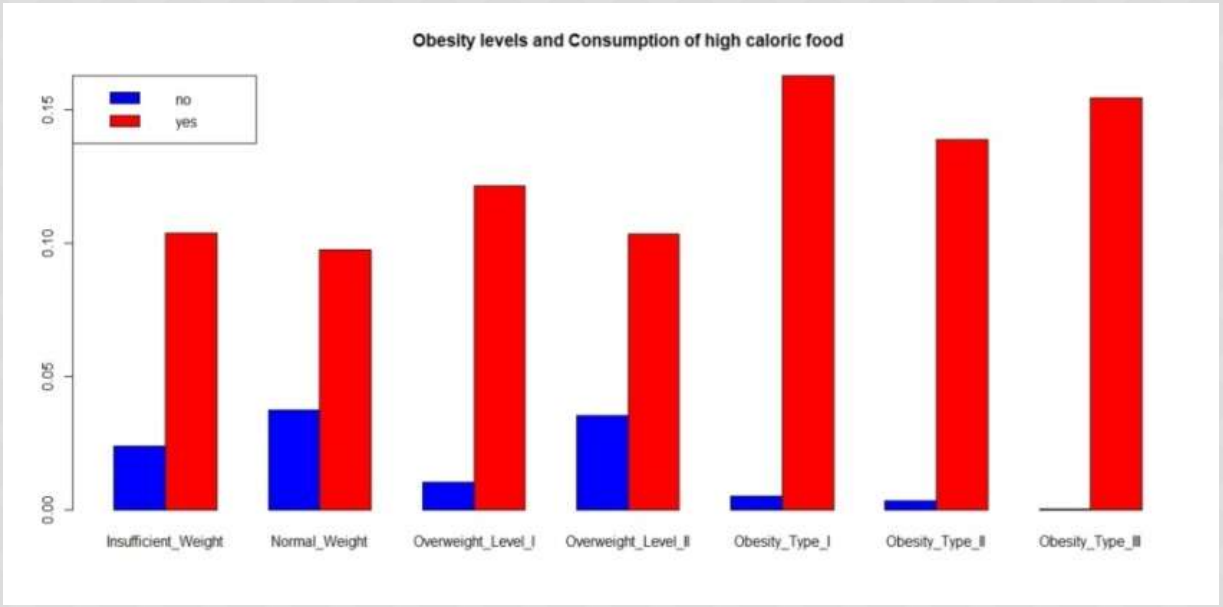
From the barplot it can be seen that people with evident family history of obesity are more prone to being overweight.

75.72% of the Overweight level I individuals, 93.80% of the overweight level II individuals, 98% of the obesity type I, 99.66% of the obesity type II & 100% of the obesity type III respondents have family history of obesity.

FREQUENT CONSUMPTION OF HIGH CALORIC FOOD



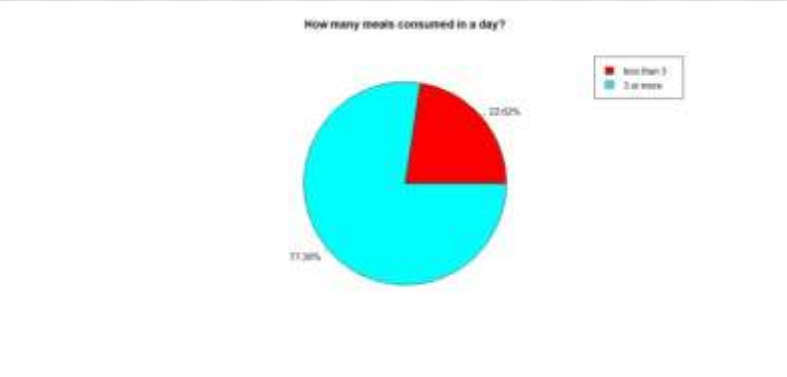
It seems that majority of the respondents (about 88%) consume high caloric food.



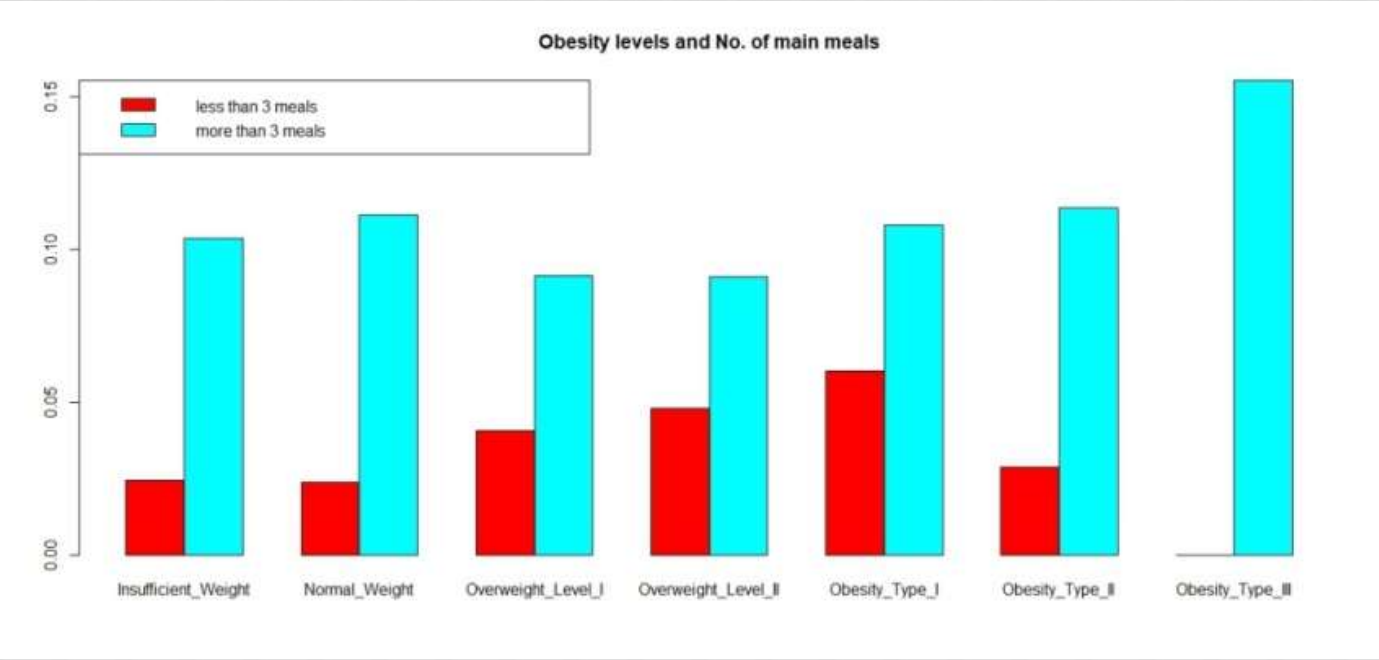
It seems that people consuming high caloric food are more prone to being obese.

96.87% of the respondents with obesity type I, 97.64% of the respondents with obesity type II, 99.69% of the respondents with obesity type III consume high caloric food.

NUMBER OF MAIN MEALS



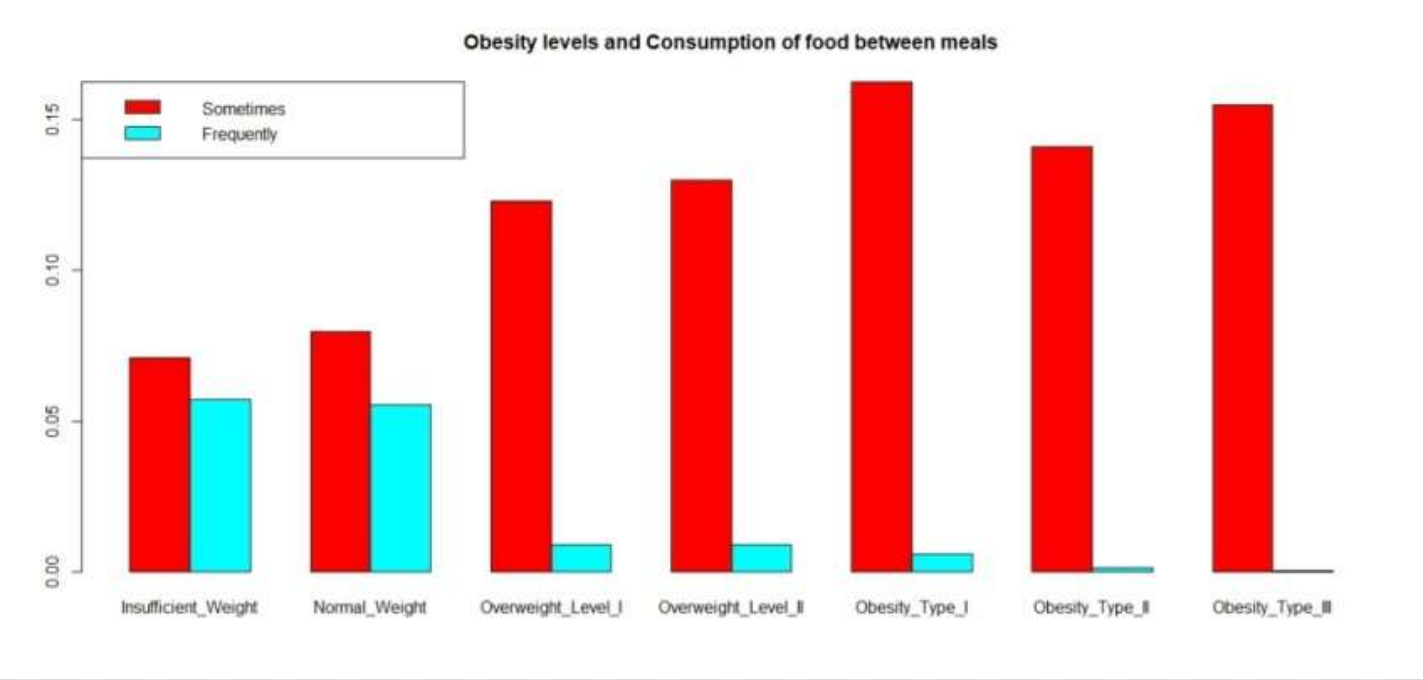
It seems that most of the individuals (around 77.38%) consume 3 or more meals in a day.



It is observed that people consuming more than 3 meals in a day tend to be more obese.

CONSUMPTION OF FOOD BETWEEN MEALS

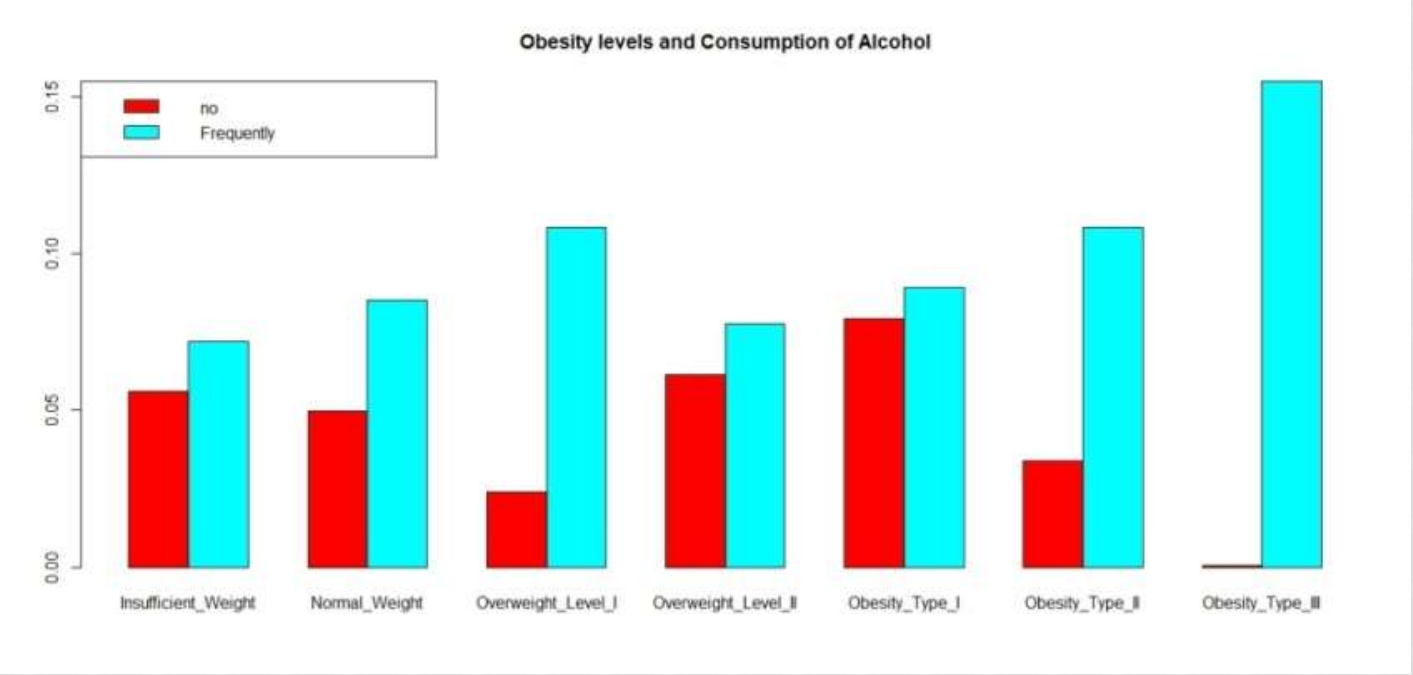
Most of the respondents (around 86.15%) consume food between meals sometimes.



It can be clearly observed that people consuming food between meals sometimes tend to be more obese.
Around 93% of the overweight level I individuals & 93% of the overweight level II individuals, 96.58% of the obesity type I, 98.98% of the obesity type II and 99.69% of the obesity type III respondents consume food between meals sometimes.

CONSUMPTION OF ALCOHOL:

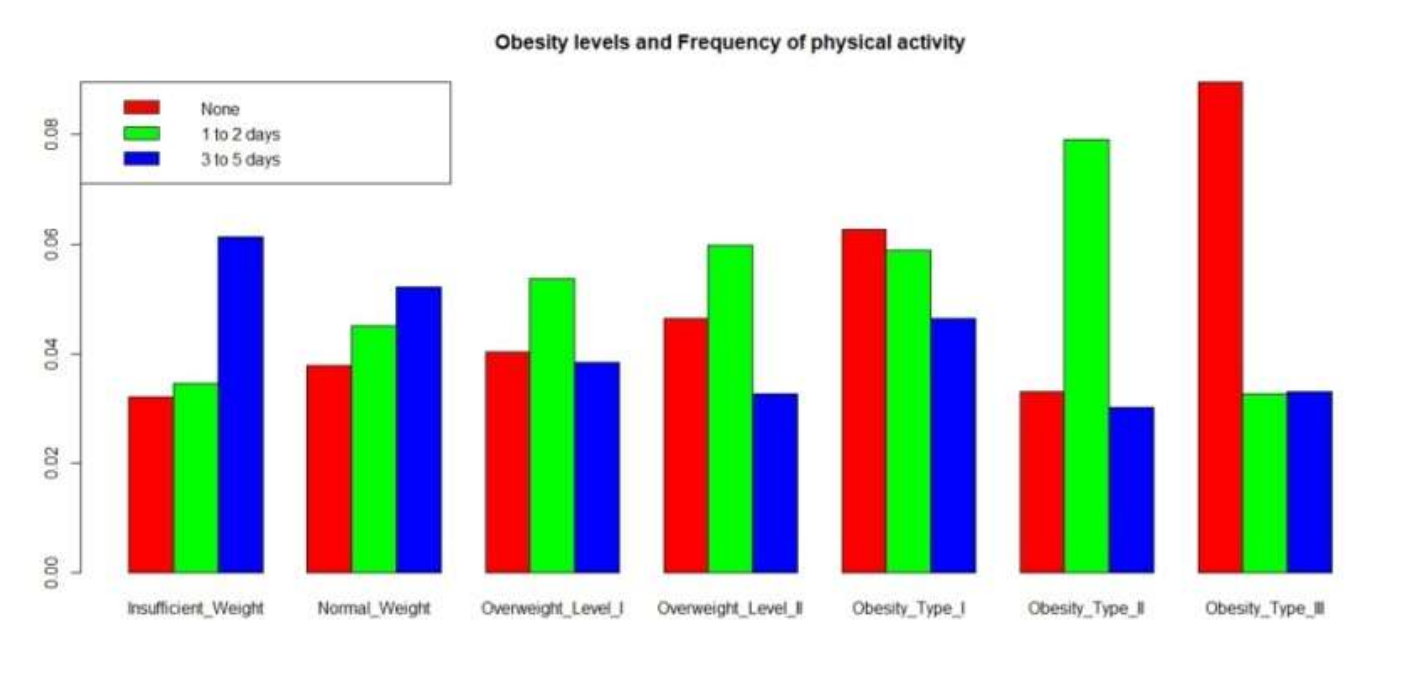
About 70% of the individuals frequently consume alcohol.



Individuals frequently consuming alcohol tend to more obese.
76.1% of the obesity type II & 99.7% of the obesity type III people frequently consume alcohol.

PHYSICAL ACTIVITY FREQUENCY PER WEEK

34.21% never exercise, while 36.37% exercise 1 to 2 days and 29.42% 3 to 5 days.



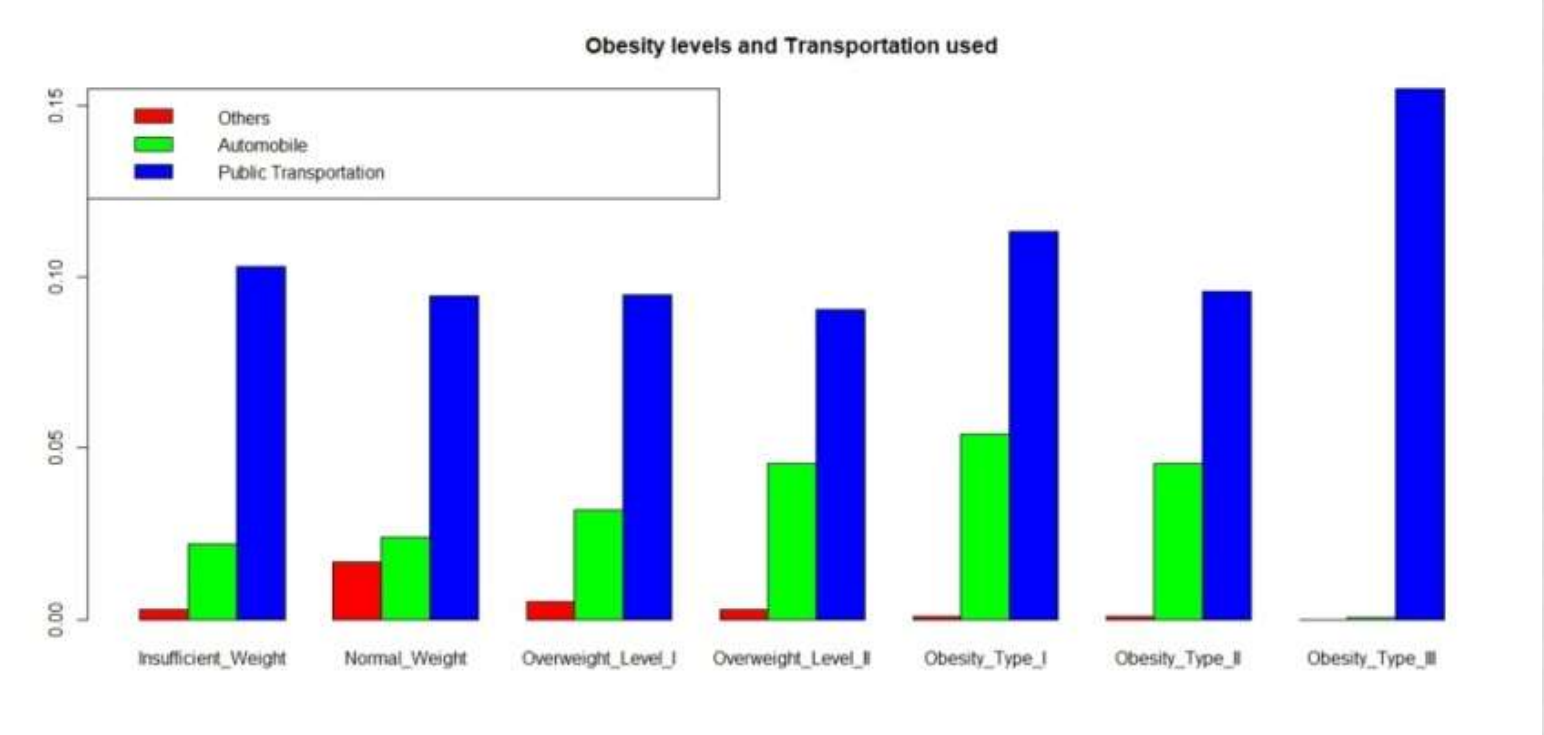
It seems that individuals who exercise 3 to 5 days in a week are least susceptible to being obese.

Individuals exercising 1 to 2 days tend to be more Overweight level I, level II and Obesity type II.

It can also be seen that people who never exercise are more vulnerable to Obesity type I (about 37.3% of the obesity type I people) & Obesity type III (about 57.7% of the obesity type III people).

TRANSPORTATIONS USED

Majority of the respondents, about 75%, uses public transportation. Very few individuals, only 2.97% of the individuals transport by walking or biking.



It can be clearly seen that people using public transportation are more prone to be obese compared to those people who are moving by walking or biking.

❖ANALYSIS

To begin with the analysis, a Random Forest model is built as the interest lies in finding the features that mostly impact obesity levels.

- **Variable Importance:**

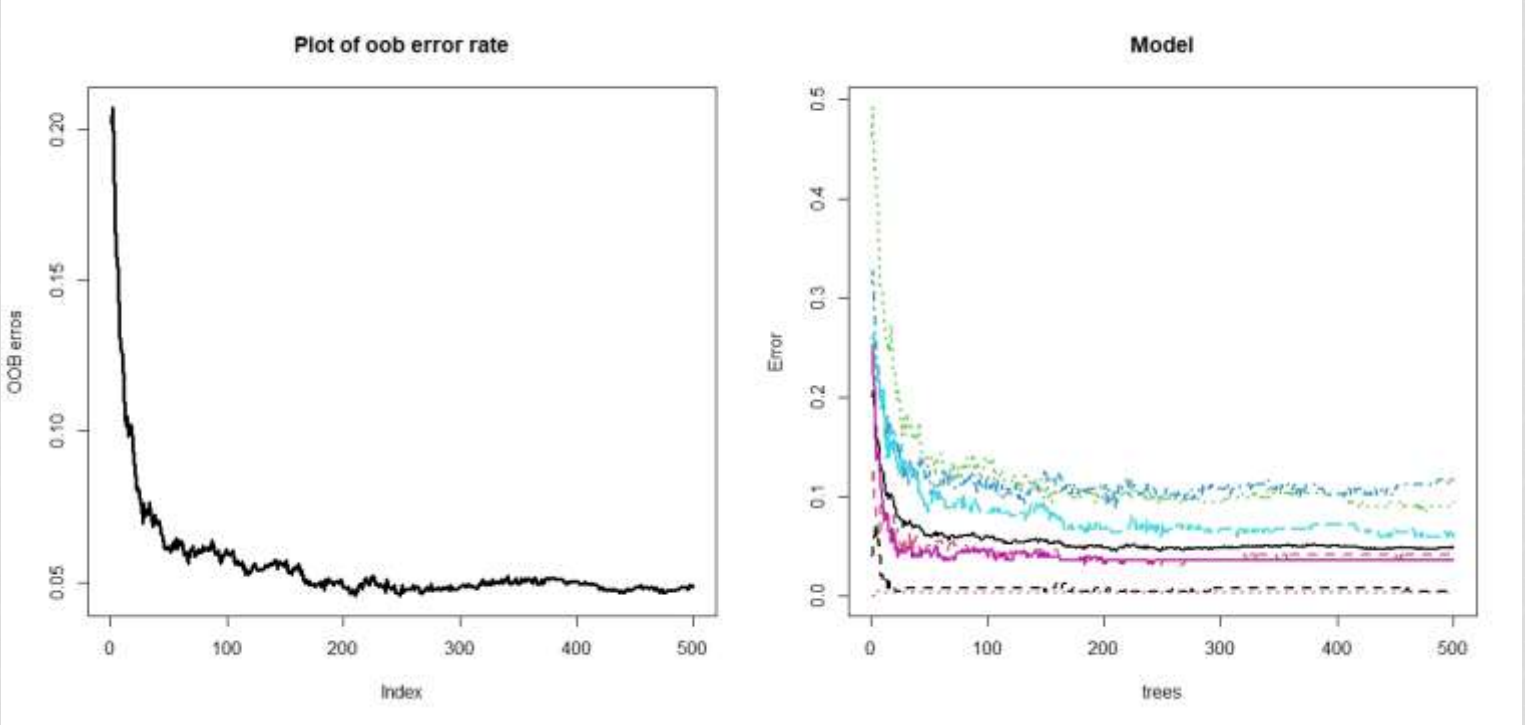
  To rank the importance of variables mean decrease accuracy or mean decrease gini can be used. Mean Decrease Accuracy table represents how much removing each variable reduces the accuracy of the model. Mean Decrease Gini gives the measure of how each variable contributes to the homogeneity of the nodes. The higher the value of mean decrease accuracy or mean decrease gini score, the higher is the importance of the variable in the model.

- **Out-of-Bag is equivalent to validation or test data. As the forest is built on training data, each tree is tested on the 1/3rd of the samples not used in building that tree. This is the out of bag error estimate - an internal error estimate of a random forest as it is being constructed.**

- **Performing random forest on the dataset:**

  To implement the algorithm on the dataset, we first split the dataset (after scaling) randomly into training and test set.

- **Applying random forest taking Obesity levels as the dependent variable and all the other variables as covariates, the following results are obtained.**
- **No. of variables tried at each split: 4**
- **We get the estimated OOB error rate as 4.85%.**

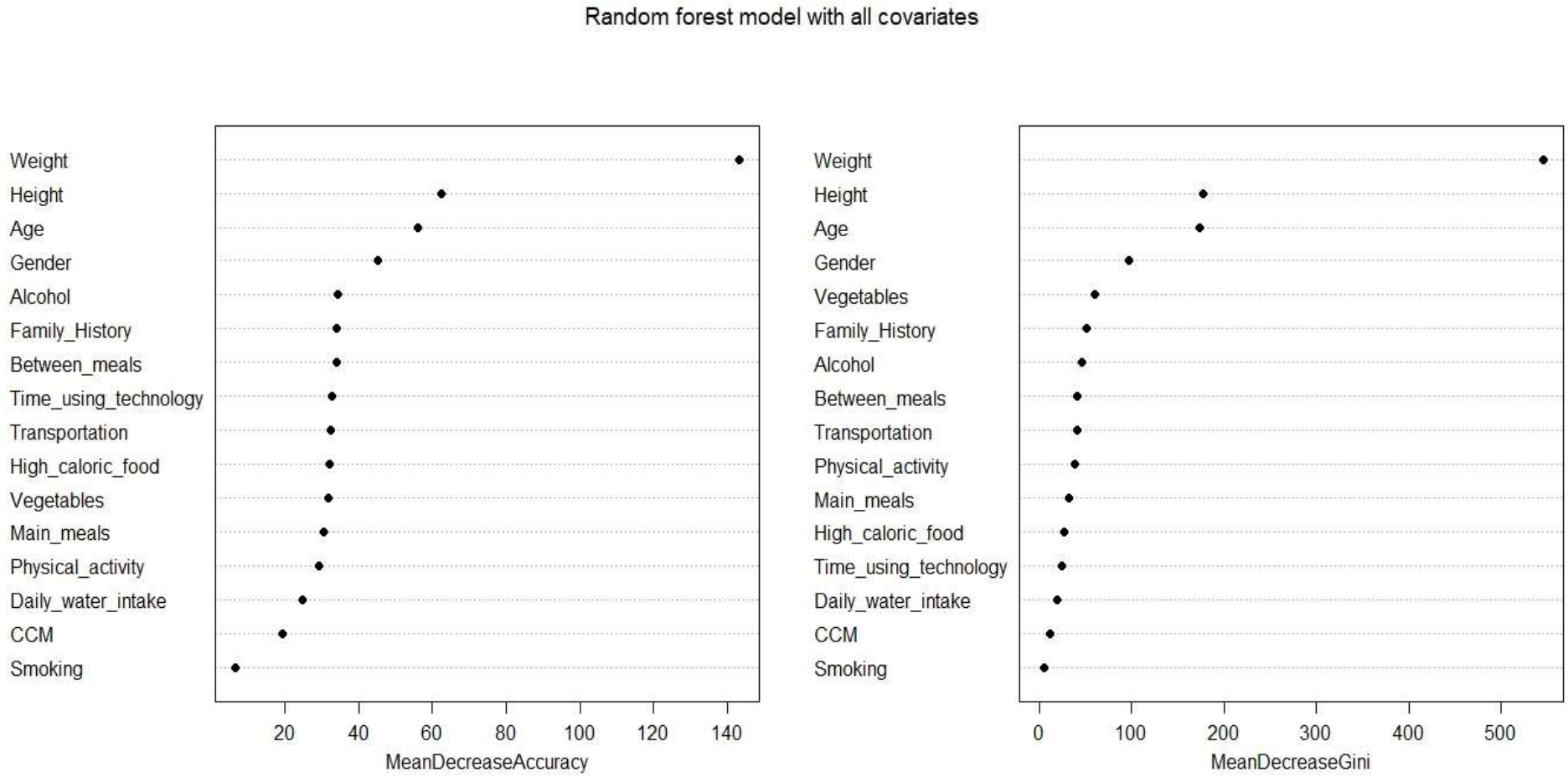


- **So, it is observed that error rate is stabilized with an increase in the number of trees.**

- **We get the accuracy as 0.9354.**
- We also get

Covariates	MeanDecreaseAccuracy	MeanDecreaseGini
Height	62.364	177.137
Weight	143.215	546.066
Gender	45.157	97.466
Age	56.102	173.942
Family_History	34.227	51.133
High_caloric_food	32.013	27.237
Vegetables	31.991	59.966
Main_meals	30.426	32.432
Between_meals	34.120	41.364
Smoking	6.591	4.691
Daily_water_intake	24.925	18.817
CCM	19.274	11.153
Physical_activity	29.169	38.712
Time_using_technology	32.684	24.881
Alcohol	34.549	46.237
Transportation	32.636	40.920

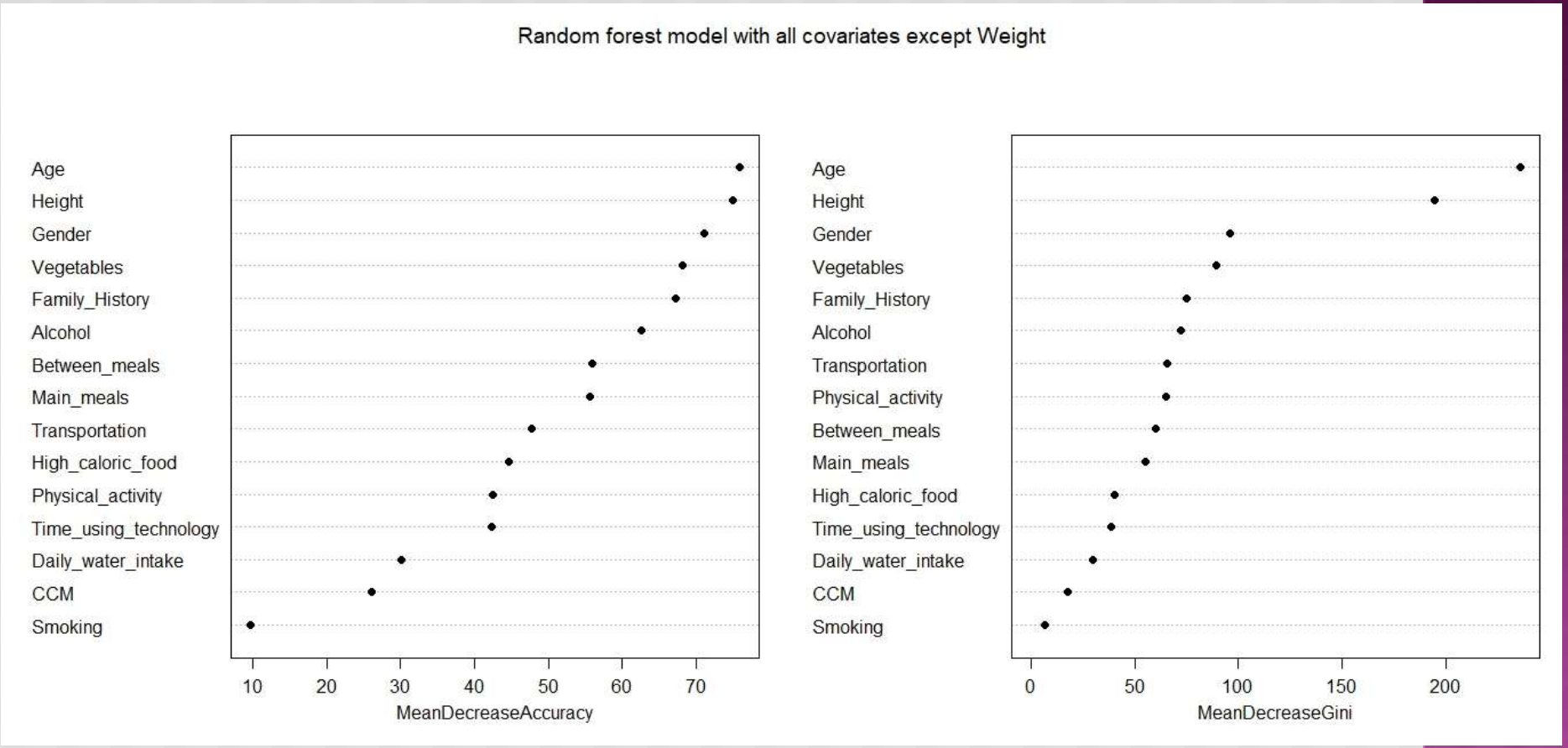
We thus get the following plot of important features



- **Comment:**
- **It is observed that Weight seems to be the most useful feature in predicting the obesity levels through the random forest model followed by height, age and gender.**
- **Without any tuning the model gives an accuracy of 93.54% when the covariate weight is included in the model. It is already seen that weight highly affects the obesity levels in an individual. So to analyze the other features more we proceed to execute the model after dropping the weight variable even though dropping it will lower the accuracy.**

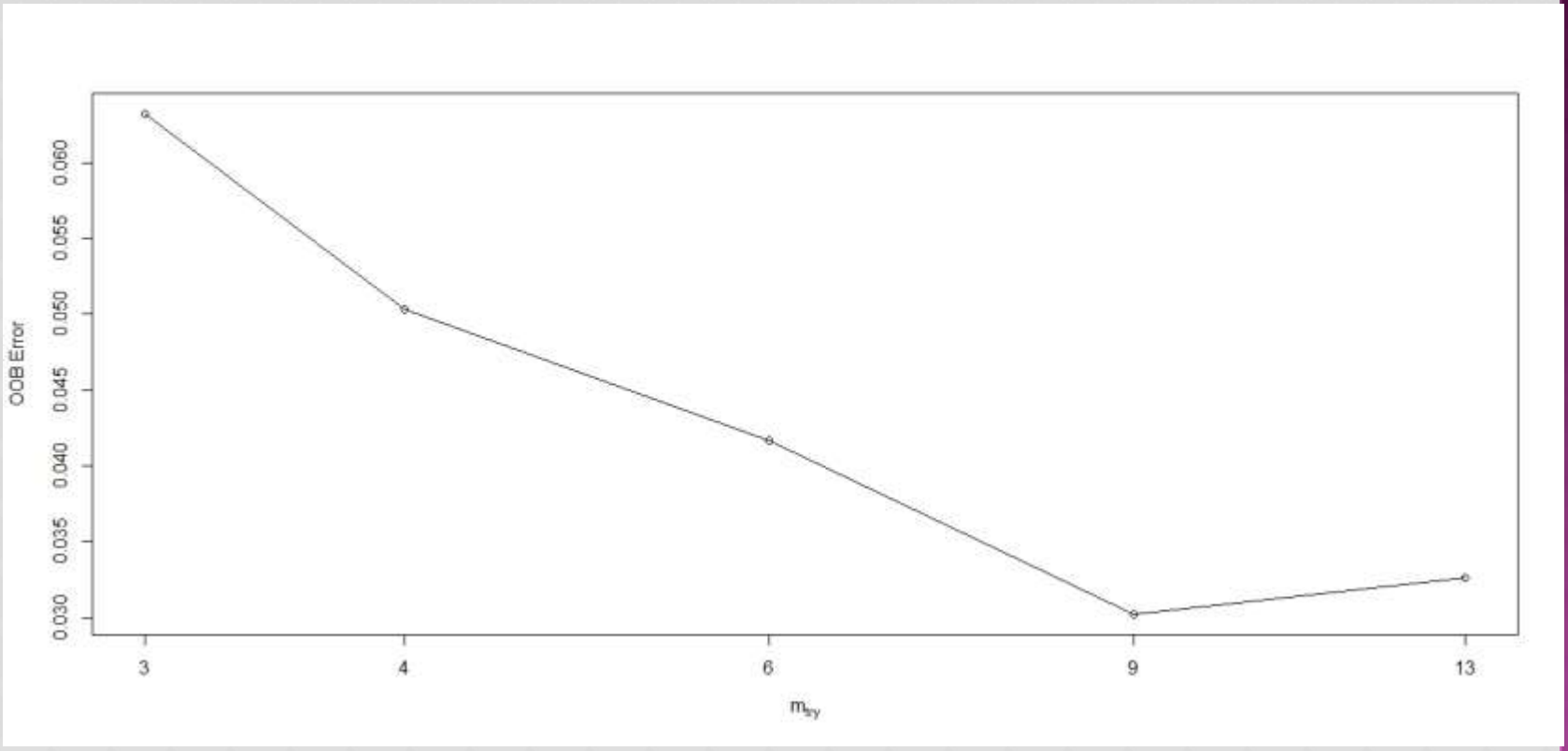
- **Now we implement the algorithm taking all the covariates except weight. The following results are obtained.**
- **No. of variables tried at each split: 3**
- **We get an estimate of OOB error rate as 18.57%.**
- **We get the accuracy as 0.7967.**

- **We thus get the following plot of important features**



- **<u>Comment:</u>**
- **Here Age seems to be the most useful feature in predicting the obesity levels through the random forest model followed by Height, Gender, Consumption of vegetables, Family history of obesity.**

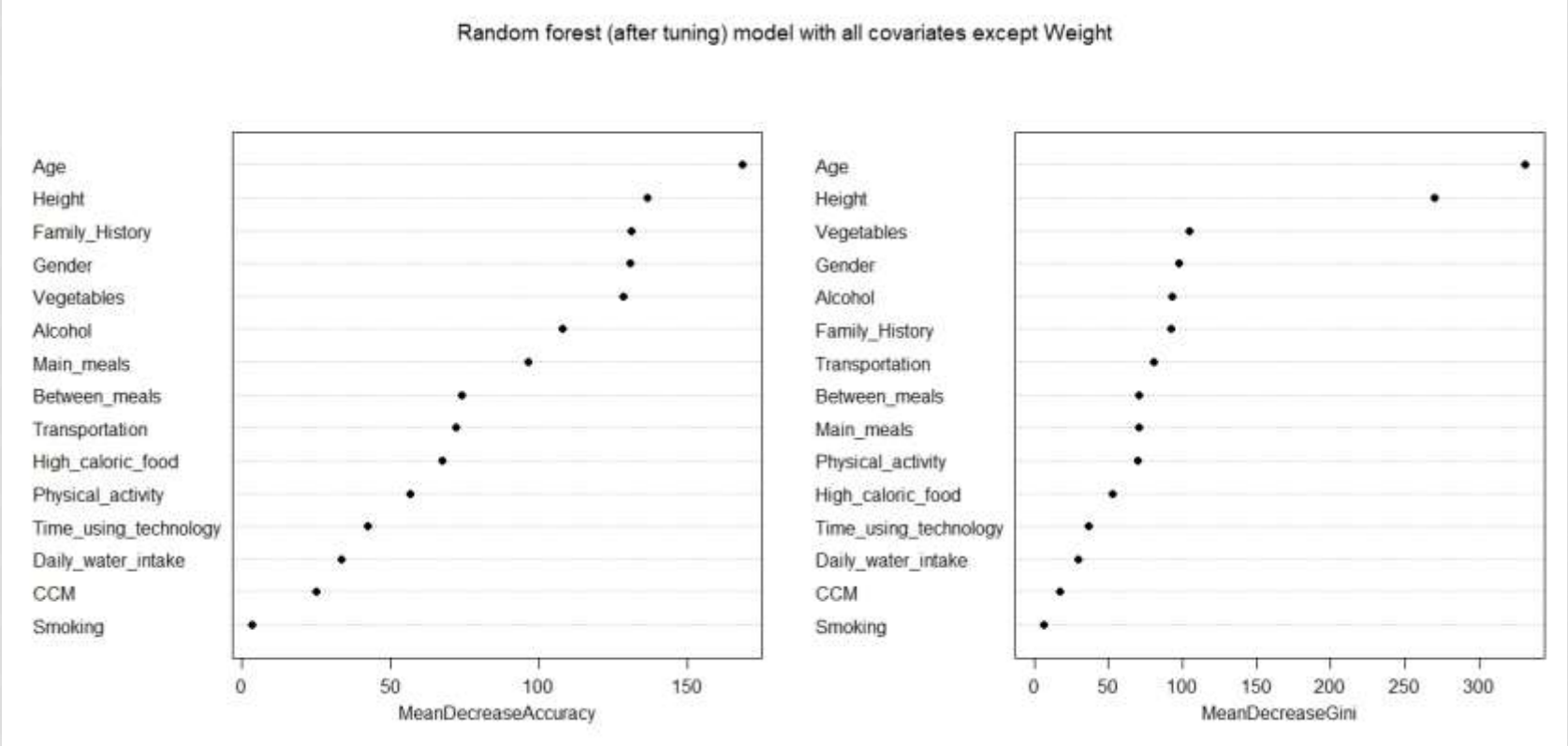
Now we proceed to tune our model.



mtry	OOBError
3	0.05941543
4	0.05031145
6	0.03929085
9	0.03114518
13	0.03449928

So, best m comes out to be 9.

- **Implementing the algorithm with all covariates except weight and taking number of trees 500 and number of variables tried at each split 9 we get the following results.**
- **The estimate of OOB error rate comes out as 16.84%**
- **We get an accuracy of 81.58%.**
- **We get the following plot**



Comment:

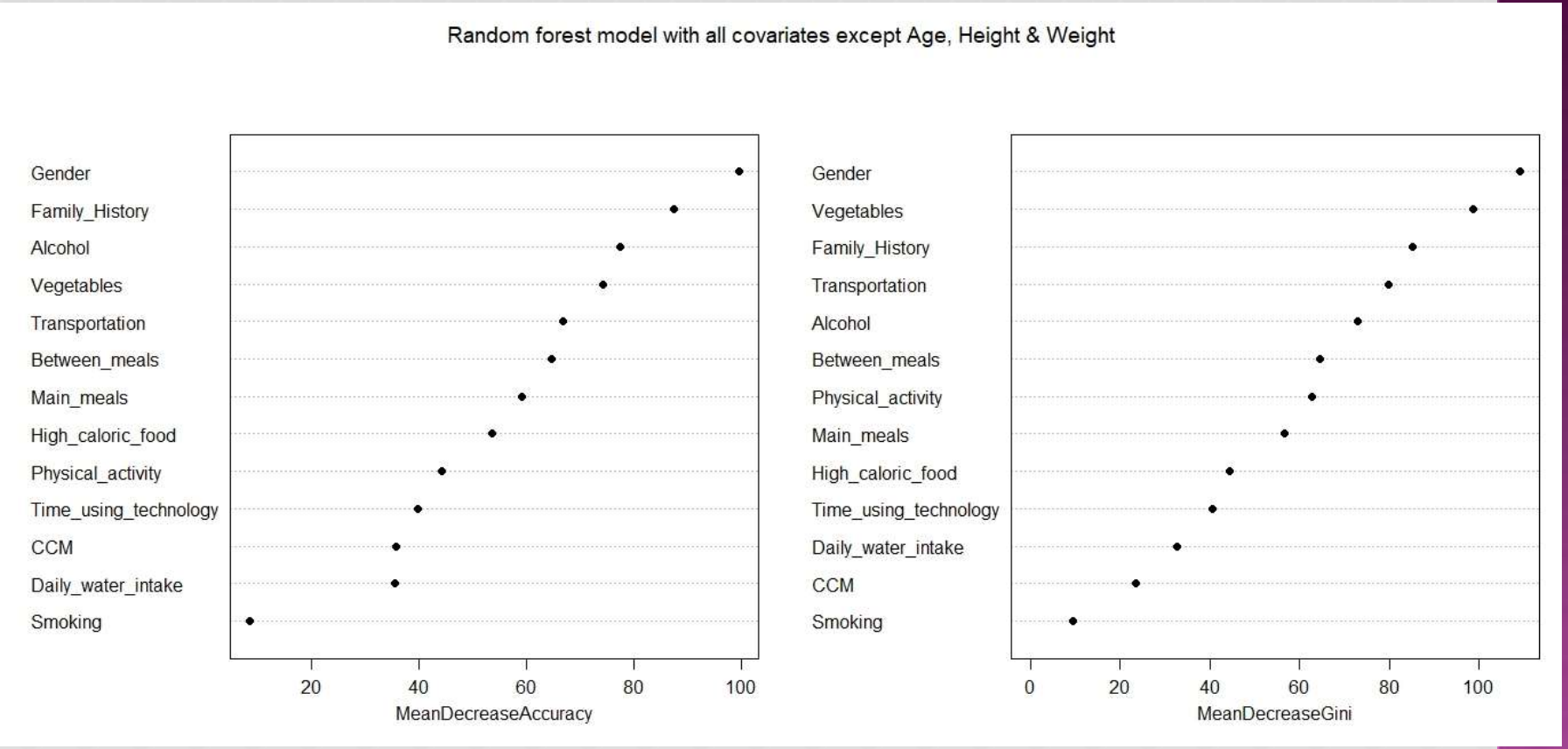
After tuning the model, it is found that, here also Age seems to be the most important feature affecting the obesity levels followed by Height, Family history, Gender, Consumption of vegetables, Consumption of alcohol.

Now to analyze the categorical features on the obesity levels we proceed to implement the model excluding the Weight, Height and Age.

We get the following result:

Covariates	MeanDecreaseAccuracy	MeanDecreaseGini
Gender	99.6724	109.210
Family_History	87.5990	85.214
High_caloric_food	53.5939	44.353
Vegetables	74.2531	98.597
Main_meals	59.2515	56.597
Between_meals	64.6369	64.589
Smoking	8.4156	9.379
Daily_water_intake	35.5155	32.800
CCM	35.6747	23.528
Physical_activity	44.3175	62.658
Time_using_technology	39.7095	40.514
Alcohol	77.4145	72.977
Transportation	66.7476	79.832

Thus the following plot of important features is obtained.



Comment:

It is observed that out of all the categorical covariates, Gender, Family history of obesity, Consumption of vegetables & alcohol, transportation used, frequency of consumption of food between meals, Number of main meals, Consumption of high caloric food and frequency of physical activity in a week seem to be the most useful features in predicting the obesity level of individuals.

While the other features such as smoking, calorie consumption monitoring are the least useful.

- **Here for the further analysis we want to model the probability of a particular obesity level given the various covariates. Thus, we are going to perform polytomous regression taking Obesity levels as our response variable In particular, the response variable is ordinal in nature. Hence, we would want to use Ordinal logistic regression also known as Proportional odds model.**

  Before fitting the regression we are going to check for the presence of any association between the covariates. It is observed that, the association between each pair of covariates is very low so that all of them can be included in our regression.

Proportional odds model:

Let Y be a categorical response with J categories. Here, J=7

$x = (x_1,...,x_{18})$ where,

x_1 : Height

x_2 : Age

x_3 = 1, if Male

0, if Female

x_4 = 1, if the individuals consume high caloric food

0, if the individuals do not consume high caloric food

x_5 = 1, if the individual consumes vegetables sometimes

0, otherwise

x_6 = 1, if the individual consumes vegetables always

0, otherwise

x_7 = 1, if the individual consume 3 or more meals in a day

0, otherwise

x_8 = 1, if the individuals consume food between meals frequently

0, otherwise

x_9 = 1, if the individual is a smoker

0, otherwise

x_{10} = 1, if daily water intake is more than 2 litres

0, otherwise

x_{11} = 1, if individual monitors calorie consumption

0, otherwise

x_{12} = 1, if individual exercises 1 to 2 days in a week

0, otherwise

x_{13} = 1, if individuals exercises 3 to 5 days in a week

0, otherwise

x_{14} = 1, if time using technology in a day is more than 3 hours

0, otherwise

x_{15} = 1, if consumes alcohol frequently

0, otherwise

x_{16} = 1, if transportation used is automobile

0, otherwise

x_{17} = 1, if public transportation

0, otherwise

x_{18} = 1, if the individual has family history of obesity and 0, otherwise

Let, $\pi_j(\underline{x})$ denote the probability of an individual whose covariate value is $\underline{x}$ lying in category j, j = 1 (1) J

$$\pi_j(\underline{x}) = P[Y = j \mid \underline{x}],\ j = 1\ (1)\ J \quad \text{such that} \sum_{j=1}^{J} \pi_j(\underline{x}) = 1$$

$$Y \sim Multinomial\ (1, \pi_1(\underline{x}), \pi_2(\underline{x}), \ldots, \pi_J(\underline{x}))$$

Model:

$$logit\left(P[Y \leq j \mid \underline{x}]\right) = \alpha_j + \underline{\beta}'\underline{x}, \quad j = 1,2,\ldots,J\text{-}1$$

$$i.e., log\left(\frac{P[Y \leq j \mid \underline{x}]}{P[Y > j \mid \underline{x}]}\right) = \alpha_j + \underline{\beta}'\underline{x}$$

- The $\{\alpha_j\}$ are increasing in j.
- The model has the same $\underline{\beta}$ for all j, i.e., the proportional odds model assumes that each explanatory variable exerts the same effect on each cumulative logit.

Fitting the model we get the following results:

Intercepts	Value	Standard error	t value	p-value
Insufficient_Weight|Normal_Weight	5.6869	1.0653	5.3383	0.000000
Normal_Weight|Overweight_Level_I	7.1222	1.0649	6.6880	0.000000
Overweight_Level_II|Overweight_Level_II	8.1961	1.0683	7.6721	0.000000
Overweight_Level_II|Obesity_Type_I	9.0810	1.0709	8.4794	0.000000
Obesity_Type_I|Obesity_Type_II	10.1612	1.0754	9.4484	0.000000
Obesity_Type_II|Obesity_Type_III	11.3943	1.0824	10.5270	0.000000

Coefficients	Value	Standard error	t value	p-value
Height	1.37685	0.635281	2.1673	0.030211
Age	0.12510	0.009389	13.3248	0.000000
GenderMale	-0.19852	0.113591	-1.7476	0.008052
Family_Historyyes	2.29964	0.129578	17.7472	0.000000
High_caloric_foodyes	0.58472	0.134738	4.3396	0.000014
VegetablesSometimes	-0.09958	0.192735	-0.5167	0.006054
VegetablesAlways	0.90622	0.195604	4.6329	0.000004
Main_meals 3 or more	0.53848	0.094533	5.6962	0.000000
Between_mealsFrequently	-2.32866	0.141677	-16.4364	0.000000
Smokingyes	-0.20456	0.276622	-0.7395	0.459601
Daily_water_intakeMore than 2 liters	0.40489	0.097534	4.1513	0.000033
CCMyes	-0.50194	0.209639	-2.3943	0.166508
Physical_activity1 to 2 days	-0.51570	0.099918	-5.1612	0.000000
Physical_activity3 to 5 days	-0.87693	0.110822	-7.9130	0.000000
Time_using_technologymore than 3 hours	0.12283	0.084810	1.4483	0.014753
AlcoholFrequently	0.56902	0.091616	6.2109	0.000000
TransportationAutomobile	0.70992	0.271552	-2.6143	0.008941
TransportationPublic_Transportation	0.96253	0.250837	3.8373	0.000124

The level of significance is taken as 0.05.

Comment:

From the p-values, it is observed that Smoking and Calorie Consumption Monitoring have p-value more than 0.05. Hence, Smoking and Calorie Consumption Monitoring are coming out as insignificant factors at 5% level of significance. This may be due to the fact that the percentages of respondents corresponding to these categories are small. Only 2.11% are smokers and 4.6% practice Calorie Consumption Monitoring.

All other variables are coming out to be significant at 5% level.

All the intercepts are coming out to be significant at 5 % level.

Thus the fitted equations come out as:

logit[P(Y≤1)] = 5.6869 + 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

logit[P(Y≤2)] = 7.1222+ 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

logit[P(Y≤3)] = 8.1961+ 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

logit[P(Y≤4)] = 9.0810+ 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

logit[P(Y≤5)] = 10.1612+ 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

logit[P(Y≤6)] = 11.3943+ 1.376 height + 0.125 age -0.19 gender_male+2.3 Family history + 0.585 High caloric food – 0.09 Vegetables_sometimes + 0.906 Vegetables_always + 0.54 Main meals – 2.33 Between meals + 0.40 Daily water intake – 0.52 Physical activity_1 to 2 days – 0.88 Physical activity_ 3 to 5 days + 0.12 Time using technology +0.57 Alcohol + 0.71 Transportation_automobile + 0.96 Public transportation

CONCLUSION:

- Obesity is a complex disease involving an excessive amount of body fat. It is a medical issue that increases the risk of other diseases and health problems.
- The interest of the study was investigating the relationship between body fat levels and various lifestyle habits and social factors, as the findings can be applicable to the health and wellbeing of many of us.
- From the study it is observed that factors apart from Weight & Height, Age, Gender, Family history of obesity, Consumption of high caloric food, Consumption of food between meals, consumption of alcohol, Transportation used, Number of main meals ,Physical activity & consumption of vegetables have large influence on an individual's Obesity levels. Factors that do not seem to be impactful were smoking, calorie consumption monitoring. People interested in the factors that influence obesity for their own health reasons can refer to the abovementioned prominent features.

- Other findings from the study-
- Young people ageing from 21 to 30 years are highly likely to be obese.
- Having a high energy intake and low energy expenditure attribute directly to obesity.
- It is advisable to limit consumption of high caloric food, consume 3 meals in a day and food between meals frequently to avoid overeating at once, consume adequate amount of vegetables, exercise 3 to 5 days in a week, along with moderation in drinking alcohol in order to treat obesity and maintain a healthy lifestyle.

THANK YOU