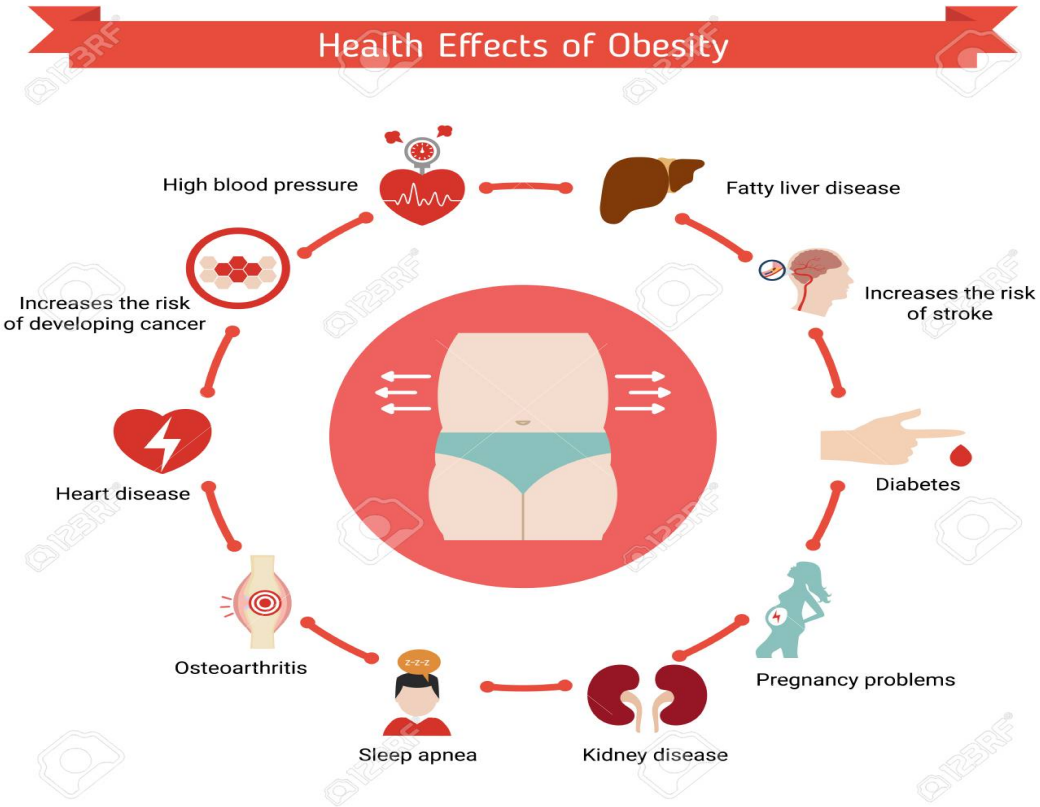
Classify Obesity Level Based on the Eating Habits and Physical Condition for Columbia, Peru and Mexico Adults



Vishalakshi Arumugam

The University of Texas San Antonio

DA 6813

Background

According to WHO, **"Nutrition is the intake of food, considered in relation to the body's dietary needs. Good nutrition – an adequate, well-balanced diet combined with regular physical activity – is a cornerstone of good health. Poor nutrition can lead to reduced immunity, increased susceptibility to disease, impaired physical and mental development, and reduced productivity."** Also, ***HIPPOCRATES,*** the Father of medicine says, **"Let food be thy medicine and medicine be thy food."** So, good nutrition indicates the right amount of nutrients for proper utilization for achieving the highest level of health. There is a strong relationship between nutrition and health. In order to live a best part of life we need to take care of ourselves. As we all know, health involves body, mind, emotion, spiritual, environment etc. But our physical body is driver to take care of other types of health. So, taking care of your physical health is important part of life.

But in this fast pace world, taking care of physical health is tossed which results in the epidemic health issue "Obesity". Obesity is getting increased day by day in children and adults with no signs of abating. At the same time, obesity is not only affecting the physical appearance of the body but also it impacts the functions of each and every part of the body which results in many chronic diseases such as hypertension, type 2 diabetes, coronary heart disease, stroke, gallbladder disease, osteoarthritis, sleep apnea, respiratory problems, dyslipidemia and endometrial, breast, prostate, and colon cancers. By analyzing physical condition and the eating habits of a person, we can identify which factors contributes the obesity more, so it helps to change the life style which is the first step in taking good care of the physical health.

Motivation

As we all know, obesity is a chronic health problem which in turn results in other dangerous health problem it is important to treat the obesity aggressively. Lifestyle changes remain the mainstay of treatment and are important for the long term maintenance of weight loss. Now a days there are many ways to bring the lifestyle changes and to track them in order to bring the weight to normal. This Project analysis mainly focuses on, which characteristics (Eating habits or physical condition) contribute more to the obesity levels so that it will helpful for the future data to get classified correctly. Based on the analysis result, we can recommend the patients to modify their life styles to keep their weight under control in order to avoid other associated dangerous disease.

This analysis mainly focuses on the below questions

- **Identify which obesity level has larger number of people?**

Based on the height and weight of the people, body mass index (BMI) is calculated and depends upon the BMI and comparison with WHO data, the people are classified to different

category of obesity level. From this analysis, we can get to know in which obesity level has larger number of surveyed people.

- **Which factors contribute more to the Obesity level Classification?**

The actual data arrives the label based on the BMI, but through this analysis, the people are classified not only based on BMI, also based on their eating habits and physical condition. Among these three group of features, the analysis helps to find out the contribution of each factor on classifying the people.

- **Compare the national average of obesity level with the analysis result**

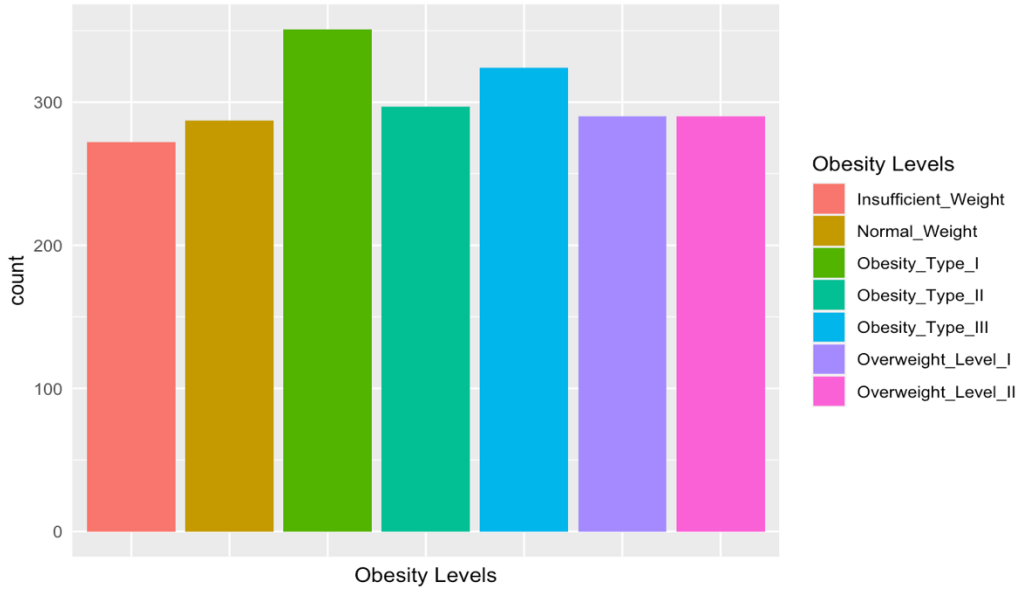
The national average obesity details are calculated based on different factors. Through this analysis the samples are classified based on eating habits and physical condition. Compare the result of obesity percentage with the national average in order to make sure the results are in sync with the national average or not.

- **Recommendation for the future data**

Through this analysis, we can come up with different models with different prediction result. After selecting the best model, we can recommend which factors contribute more for obesity and what needs to be changed in order to avoid the associated dangerous disease.

Description of the Data

Dataset consist of 2111 observations and 17 attributes which are obtained through the survey among the anonymous users from the countries Mexico, Columbia and Peru between the ages 14 and 61. In this dataset,77% of the data was generated synthetically using the Weka tool and the SMOTE filter, 23% of the data was collected directly from users through a web platform.so the dataset doesn't have any missing values and it is a balance dataset as shown below.



The 17 attributes are stated as below.

Attribute category	Attribute	Attribute Type
General (attributes to calculate body mass index)	Gender	Character
	Age	Number
	Height	Number
	Weight	Number
	Family History with overweight	Character
Eating Habits Attributes	Frequent consumption of high caloric food (FAVC)	Character
	Frequency of consumption of vegetables (FCVC)	Number
	Number of main meals (NCP)	Number
	Consumption of food between meals (CAEC),	Character
	Consumption of water daily (CH20)	Character
	Consumption of alcohol (CALC)	Character
Physical Condition Attributes	Calories consumption monitoring (SCC)	Character
	Physical activity frequency (FAF)	Number
	Time using technology devices (TUE)	Number
	Transportation used (MTRANS)	Character

For this analysis, the dataset is split into training and testing set in 70:30 combination. Th training of various machine learning method will be happened on the training dataset. Will use the test set to predict the result and based on the accuracy and other metric parameters will select the best model for final recommendation.

Proposed Analysis

Below is the proposed process of this analysis.
Logistic Regression is a parametric classification method in which is used to model the probability of a certain class or event existing based upon the independent variables. In Logistic Regression, we don't directly fit a straight line to our data like in linear regression. Instead, we fit a S shaped curve, called Sigmoid, to our observations. As the goal of analysis is to classify the obesity level based on the independent variables using the logic regression, will train the training dataset and compare the metrics whether it is a reasonable model. Also find out the parameter's contribution using various feature selection methods (such as step selection, VIF etc.) and optimization parameters in order to have a better metrics.
Though Logistic regression is an effective for classification parameter, it is worth to check the other machine learning methods with different optimization parameters, different feature selection in order to have a better model for predicting the obesity levels. Below are the different classification techniques that are going to be involved in this analysis
Random forest is a supervised learning algorithm used for classification and regression tasks. It is distinguished from decision trees by the randomized process of finding root nodes to split features. Random forest is efficient in handling missing values. Unless a sufficient number of

trees is generated to enhance prediction accuracy, the over fitting problem is a possible drawback of this algorithm.

SVM is a learning algorithm used in regression tasks. However, SVM is preferable in classification tasks. This algorithm is based on the following idea: if a classifier is effective in separating convergent non-linearly separable data points, then it should perform well on dispersed ones. SVM finds the best separating line that maximizes the distance between the hyperplanes of decision boundaries.

Linear Discriminant Analysis (LDA) is a simple and effective method for classification. It is a discriminant approach that attempts to model differences among samples assigned to certain groups. The aim of the method is to maximize the ratio of the between-group variance and the within-group variance. When the value of this ratio is at its maximum, then the samples within each group have the smallest possible scatter and the groups are separated from one another the most. LDA often produces robust, decent, and interpretable classification results.

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision rules are generally in form of if-then-else statements. The deeper the tree, the more complex the rules and fitter the model.

After comparing the metrics and selection the better model, validate the test data for the best result. Based on the result, visualize the metrics and also compare the percentage of each obesity levels to the national obesity level. After comparing the result recommend the lifestyle changes in order to reduce the obesity and the associated dangerous disease.

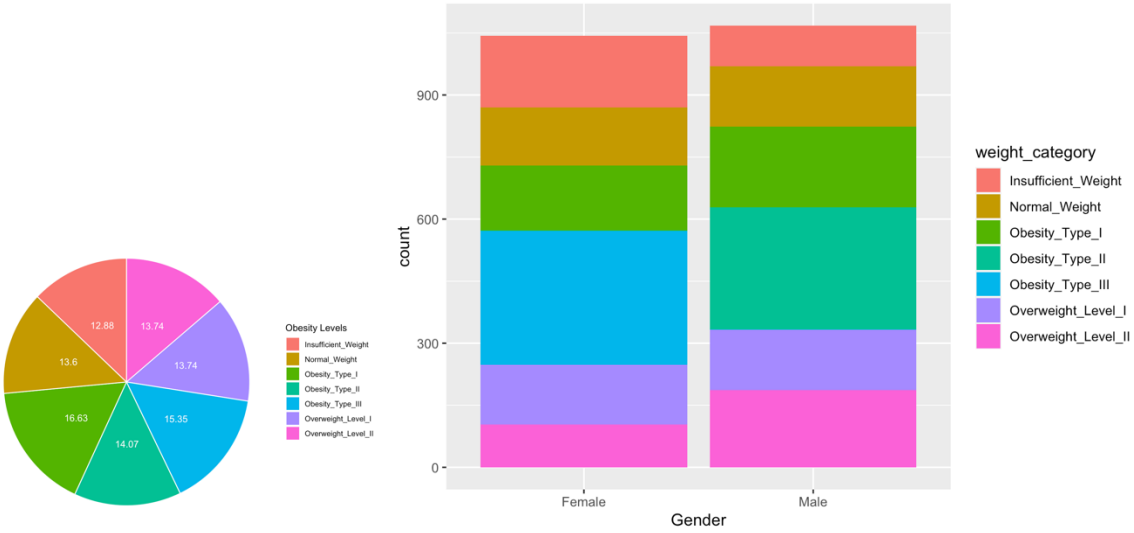
Data Pre-processing and Data Exploration

Before analyzing the data, the data needs to be preprocessed in such a way that whether any missing values exist, any character needs to be converted to factors, any new predictors need to calculated based on the existing predictors. For this data set, there exist no missing values. The dataset is more or less balanced as per the plot in the data description. Also, the following predictors Gender, Family History with overweight, Eating Habits attributes and Physical condition attributes are converted from characters to factors. The dataset is split with the ratio of 80% of data as training and 20% of data as testing data. Also, a new predictor “bmi” is created using the formula BMI= weight/height^2 in order to check whether any relationship exists.

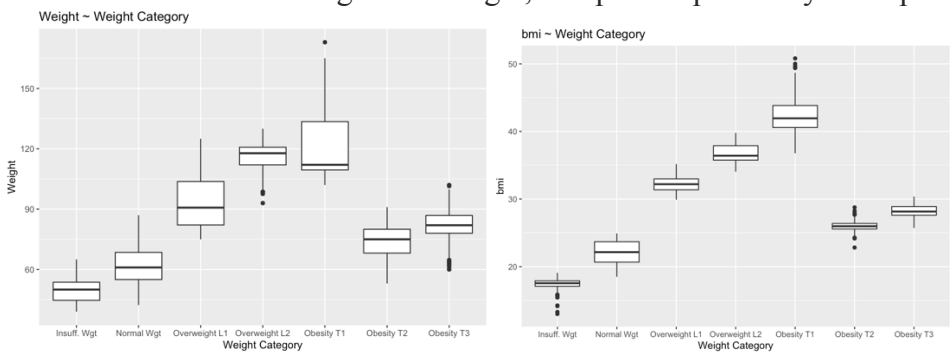
After preprocessing, Exploratory analysis is carried out in order to find out the relationship and distribution of predictors based on the response variable. As per the summary details, the observation is balanced on gender and weight category. The mean height is 1.7 meters, the mean weight is 87kg and mean bmi falls around 29.7 which is close to the average bmi level for obese category.

Exploratory Analysis is very interesting. Below are the findings. The distribution of weight category is almost are around 14% among all the categories. The distribution of observation is slightly high I. obesity type I with 16% and 15.35% on obesity type III as plotted below. Also,

there is no balanced data on the gender level. The data set contains female with more concentration of Obesity type I and III , whereas the male contains only the concentration of Obesity type I and II. The data set doesn't have balanced record on weight category on gender level.



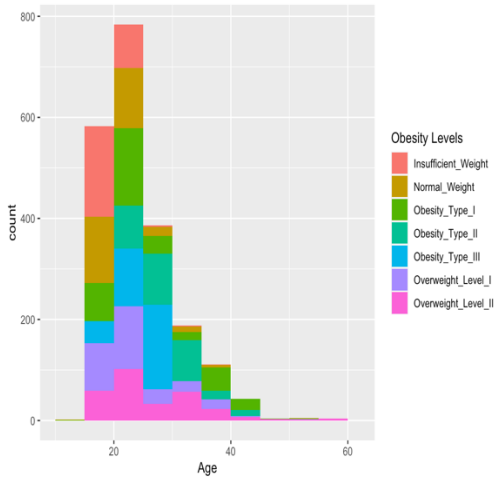
As per the Age, Height, weight and BMI Data Analysis, there is no clear relation between the weight category and weight/bmi. Till overweight II category the median weight is increased after that the median weight is lower than the overweight category. so, from this we can predict that obesity level is not classified based on weight, it depends upon many other parameters. Similarly, as weight, “bmi” doesn’t have no strong relationship on the weight category. Till overweight II category the median bmi is increased after that the median bmi is lower than the overweight category. At the same time, the average bmi level for the weight category is as below. but the data doesn't reflect the same.so this provides the strong evidence that the obesity level is not classified based on height and weight, it depends upon many other parameters.

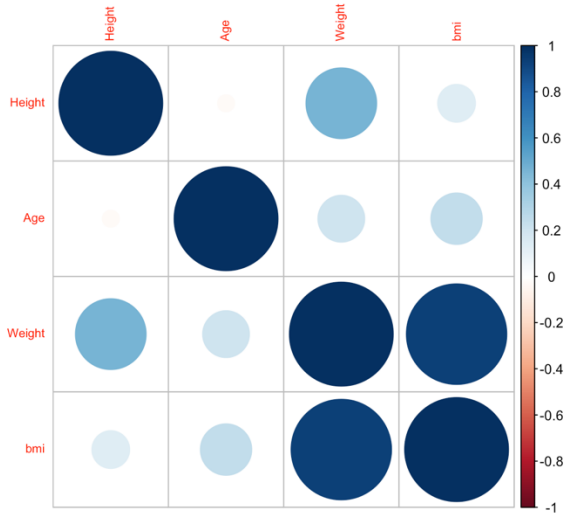


BMI	Weight Status
Below 18.5	*Underweight*
18.5–24.9	*Healthy*
25.0–29.9	*Overweight*
30.0 and Above	*Obese*

Based on the Age level Analysis, the count of obesity level people from obesity type I to III is more in the age range between 20 and 30, from 30 to 45 the data contains the people with more obesity level II and III which are at high risk of getting obesity associated health problems.

In this data set the continuous variables are Age, Height, weight and created predictor bmi. When we found the correlation there is no correlation between Age, height and weight, but there is a strong correlation between bmi and Height and Weight. So, when we use the predictors for modelling, Height, weight and bmi should not be used altogether.





Feature Selection

After the Exploratory data analysis, the main goal is to select which predictors are the best predictors to determine the obesity level. For that I have used the Stepwise Model selection and anova techniques to decide which parameters plays the significant role.

when execute the stepwise method for the whole dataset, The p *value (*1) of the Hosmer and lemeshow test is greater than the significant level (0.1), so we can conclude that the model by stepwise AIC method is adequate and we can conclude that ****Age, height, weight, eats_snacks, time_using_tech and exercises_often** are the best predictors to find out obesity level of an individual.

As per the Residual deviance difference from anova test, the addition of predictors such as Gender,Age,weight,eats_snacks,time_using_tech and method_trans to the null model reduces the deviance drastically, so the following predictors such as **Gender, Age, weight, eats_snacks, time_using_tech and method_trans** are considered as best predictors in estimating the obesity level of an individual.

```
Call:
glm(formula = weight category ~ Weight + Height + eats snacks
+
    time_using_tech + Age + exercises_often, family =
"binomial",
    data = obesity_data)


Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.01011   0.00000   0.00000   0.00000   0.01073

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         16652.55   73332.44   0.227    0.820
Weight                273.30    1118.12   0.244    0.807
Height             -18183.16   74342.47  -0.245    0.807
eats_snacksFrequently   -434.35   27419.61  -0.016    0.987
```

```
eats snacksno              188.92   59552.52   0.003    0.997
eats_snacksSometimes    -171.21   27379.73  -0.006    0.995
time using tech1         301.99    1245.96   0.242    0.808
time_using_tech2         394.87   19824.82   0.020    0.984
Age                      -15.77      74.77  -0.211    0.833
exercises_often1          15.63   14206.33   0.001    0.999
exercises_often2        -118.55     630.02  -0.188    0.851
exercises_often3        -381.02    7378.61  -0.052    0.959

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.6221e+03  on 2110  degrees of freedom
Residual deviance: 4.5920e-04  on 2099  degrees of freedom
AIC: 24

Number of Fisher Scoring iterations: 25
```

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  model obesity AIC$y, fitted(model obesity AIC)
X-squared = 5.071e-13, df = 8, p-value = 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: weight category

Terms added sequentially (first to last)


                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                            2110     1622.1
Gender                           1     25.4      2109     1596.6 4.601e-07 ***
Age                              1    269.1      2108     1327.5 < 2.2e-16 ***
Height                           1      0.2      2107     1327.3  0.622442
Weight                           1   1288.4      2106       38.9 < 2.2e-16 ***
family_history_with_overweight   1      0.7      2105       38.2  0.419733
eats_high_calor_food             1      0.6      2104       37.6  0.445912
eats_veggies                     2      1.5      2102       36.1  0.464086
num_meals                        3      6.8      2099       29.3  0.077428 .
eats snacks                      3     11.7      2096       17.5  0.008336 **
SMOKE                            1      0.0      2095       17.5  0.869068
drinks water                     2      0.1      2093       17.4  0.947304
counts_calories                  1      2.1      2092       15.3  0.149319
exercises often                  3      2.7      2089       12.6  0.443699
time_using_tech                  2     12.6      2087        0.0  0.001816 **
drinks_alcohol                   3      0.0      2084     3820.6  1.000000
method_trans                     4   3820.6      2080        0.0 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Since full dataset logistic regression variable selection and variable importance doesn't provide clear information about the best predictors, Split the data set based on gender and find out the best predictors to estimate the obesity level as per the gender and conducted Stepwise Model selection and anova techniques to decide which parameters plays the significant role.

```
Call:
glm(formula = weight_category ~ Weight + Height + drinks_water,
    family = "binomial", data = obesity data male)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.02180   0.00000   0.00000   0.00000   0.02505

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      82460     254568   0.324    0.746
Weight            1258       3388   0.371    0.710
Height          -86213     232193  -0.371    0.710
drinks_water2    -2890     124792  -0.023    0.982
drinks water3    -1406     128396  -0.011    0.991

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.5945e+02  on 1067  degrees of freedom
Residual deviance: 1.5719e-03  on 1063  degrees of freedom
AIC: 10.002

Number of Fisher Scoring iterations: 25
```

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  model obesity male AIC$y, fitted(model obesity male AIC)
X-squared = 2.7979e-10, df = 8, p-value = 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: weight_category

Terms added sequentially (first to last)


                                Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                            1067     659.45
Age                              1   226.00      1066     433.45 < 2.2e-16 ***
Height                           1    11.54      1065     421.92 0.0006826 ***
Weight                           1   412.10      1064       9.81 < 2.2e-16 ***
family_history_with_overweight   1     0.24      1063       9.58 0.6243477
eats_high_calor_food             1     1.20      1062       8.38 0.2738185
eats_veggies                     2     1.15      1060       7.23 0.5628614
num_meals                        3     7.23      1057       0.00 0.0649724 .
eats_snacks                      3     0.00      1054       0.00 1.0000000
SMOKE                            1     0.00      1053       0.00 1.0000000
drinks water                     2     0.00      1051       0.00 0.9999998
counts_calories                  1     0.00      1050       0.00 1.0000000
exercises_often                  3     0.00      1047       0.00 1.0000000
time_using_tech                  2     0.00      1045       0.00 0.9999999
drinks alcohol                   3     0.00      1042       0.00 1.0000000
method_trans                     4     0.00      1038       0.00 1.0000000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As per the Residual deviance difference from anova test, the addition of predictors such as Age, weight and Height to the null model reduces the deviance drastically, so the following predictors such as Age, weight and height are considered as best predictors in estimating the obesity level in male individual.

```
Call:
glm(formula = weight_category ~ bmi + eats_snacks + Age, family =
"binomial",
    data = obesity_data_female)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.078   0.000   0.000   0.000   1.589

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)          -177.7084  4206.6977  -0.042  0.96630
bmi                    10.4090     4.0324   2.581  0.00984 **
eats_snacksFrequently  -12.3725  4206.0268  -0.003  0.99765
eats snacksno          -10.9806  4206.1933  -0.003  0.99792
eats_snacksSometimes    -7.9234  4206.0281  -0.002  0.99850
Age                     -0.2697     0.1765  -1.528  0.12647
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 937.183  on 1042  degrees of freedom
Residual deviance:  15.582  on 1037  degrees of freedom
AIC: 27.582

Number of Fisher Scoring iterations: 20
```

```
Hosmer and Lemeshow goodness of fit (GOF) test

data:  model_obesity_female_AIC$y,
fitted(model_obesity_female_AIC)
X-squared = 0.00034063, df = 8, p-value = 1
```

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: weight_category

Terms added sequentially (first to last)


                                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                                             1042     937.18
Age                              1    93.54      1041     843.65  < 2e-16 ***
Height                           1     3.42      1040     840.23  0.06449 .
Weight                           1   815.61      1039      24.62  < 2e-16 ***
family_history_with_overweight   1     0.11      1038      24.51  0.74323
eats_high_calor_food             1     0.05      1037      24.46  0.82395
eats veggies                     2     2.57      1035      21.89  0.27702
num meals                        3     0.90      1032      21.00  0.82603
eats_snacks                      3     8.62      1029      12.38  0.03485 *
SMOKE                            1     0.03      1028      12.35  0.85845
drinks_water                     2     0.29      1026      12.06  0.86354
counts calories                  1     3.05      1025       9.00  0.08068 .
exercises_often                  3     0.95      1022       8.05  0.81258
time_using_tech                  2     8.05      1020       0.00  0.01785 *
drinks_alcohol                   2     0.00      1018       0.00  1.00000
method_trans                     3     0.00      1015       0.00  1.00000
bmi                              1     0.00      1014       0.00  1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As per the Residual deviance difference from anova test, the addition of predictors such as Age,weight,eats_snacks and time_using_tech to the null model reduces the deviance drastically, so the following predictors such as **Age, weight, eats_snacks and time_using_tech** are considered as best predictors in estimating the obesity level in female individual.

Since logistic regression p value is too high, variable importance value of full logistic regression doesn't give much clarity in deciding the significant value, I have used step wise model selection and anova inferential test for the full data set and data set with male and female individual separately in order to come with the reasonable predictors to estimate the obesity level. Based on the above analysis the predictors such as **Age, Gender, weight, height, eat_snacks, drinks_water, time_using_tech, method_trans and exercises_often** are the best parameters in estimating obesity level of an individual.

Modelling Techniques and Results

After selecting the Significant parameters, executed various modelling techniques for the full predictor variable, only significant predictors and model with bmi and other significant parameters except height and weight. The modelling techniques used are below.

- Logistic Regression
- K Nearest Neighbor
- Linear Discriminating Analysis
- Support Vector Machines
- Neural Network
- Random Forest
- Decision Tree
- Gradient Boosting Method

All the Method shows the average Accuracy percentage of 95% whereas For SVM Radial, Decision Tree and K nearest Neighbor are very low. I feel that is because of missing weight category observation altogether in the gender analysis. But need to check on that. Apart from that the accuracy level are reasonable with good f1 score and Recall metrics in deciding the obesity level. Below table represents the top 10 model with good accuracy.

Modelling Techniques	Accuracy
GBM_BMI	0.9747634
RF_BMI	0.9747634
RF_SIG	0.9637224
GBM_FULL	0.9605678
RF_FULL	0.9605678

LOG_SIG	0.9542587
LOG_BMI	0.9511041
LOG_FULL	0.9495268
GBM_SIG	0.9463722
LDA_BMI	0.9195584

Conclusion and Recommendation

As Per the Model comparison table, top 3 model are Random forest with significant parameters, Gradient Boosting method with significant parameters and bmi and Random forest with significant parameters and bmi with the accuracy of 97%.

So the model comparison accuracy results confirmed that the predictors such as **Age,(Height,Weight) or bmi, Gender, eats_snacks, drinks_water, time_using_tech, method_trans and exercises_often** are the best predictors in classifying the individual with accurate weight category.

Also , the recall value and balanced accuracy for all the models for the obesity class is almost higher than 98% , so it gives as strong evidence that the above predictors are the best predictors in classifying the weight category.

As per the coefficients, one can avoid getting into the obesity category by changing their lifestyle with the below changes.

- **Reduce the weight as per the Height.**
- **Avoid food between meals.**
- **Drink lots of water.**
- **Exercise often.**

References

Data:
https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+
Variable description:
https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub