Estimation of obesity levels based on eating habits and physical condition

Profesor: Shafqat Shad
Students: Rabab Mohamed, Gabriel Palacios

Objective

According to the CDC, the prevalence of obesity is 42.5% among adults aged from 20 and older. Thus, we wanted to create a model that can predict the obesity levels of individuals based on eating habits and physical conditions.

Data Source

This dataset belongs to the UCI Machine Learning Repository. The data was gathered from individuals from Colombia, Mexico and Peru. Each entry describes the individual's eating habits and physical condition and is labeled with the class variable NObesity (Obesity Level). This data set used 7 different categories of obesity levels which are:

1. Insufficient Weight
2. Normal Weight
3. Overweight Level I,
4. Overweight Level II,
5. Obesity Type I,
6. Obesity Type II
7. Obesity Type III.

```
# Overview of the dataset's datatypes, number of columns and indexes.
obesity.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2111 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2111 non-null   object
 1   Age                             2111 non-null   float64
 2   Height                          2111 non-null   float64
 3   Weight                          2111 non-null   float64
 4   family_history_with_overweight  2111 non-null   object
 5   FAVC                            2111 non-null   object
 6   FCVC                            2111 non-null   float64
 7   NCP                             2111 non-null   float64
 8   CAEC                            2111 non-null   object
 9   SMOKE                           2111 non-null   object
 10  CH2O                            2111 non-null   float64
 11  SCC                             2111 non-null   object
 12  FAF                             2111 non-null   float64
 13  TUE                             2111 non-null   float64
 14  CALC                            2111 non-null   object
 15  MTRANS                          2111 non-null   object
 16  NObeyesdad                      2111 non-null   object
dtypes: float64(8), object(9)
memory usage: 280.5+ KB
```
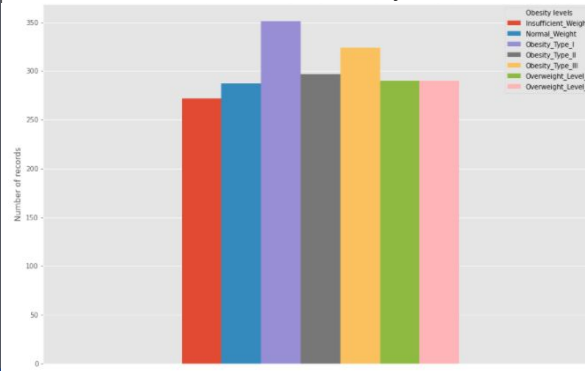
Data Visualization

BoxPlot of Obesity level and Age
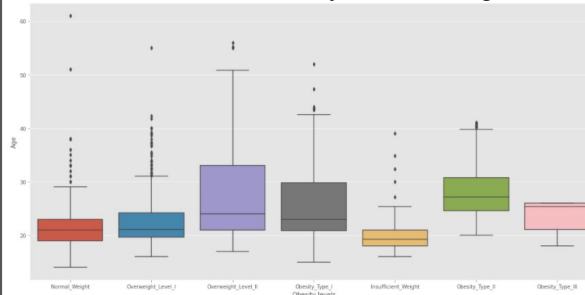
Distribution of obesity levels

Results

Confusion matrix

```
[[49  3  0  0  0  0  0]        [[52  0  0  0  0  0  0]
 [ 3 45  0  0  0 10  0]         [13 33  0  0  0  8  4]
 [ 0  0 58  1  2  0  1]         [ 0  0 60  1  1  0  0]
 [ 0  0  1 66  0  0  0]         [ 0  0  2 64  1  0  0]
 [ 0  0  0  0 69  0  0]         [ 0  0  0  2 67  0  0]
 [ 0  4  0  0  0 44  1]         [ 0  3  1  0  0 43  2]
 [ 0  0  3  0  0  2 61]]        [ 0  0  3  1  0  5 57]]
```
Tree based   KNN, K=2

Accuracy	Tree based	KNN, k=2
Training set	1.0	0.960
Testing set	0.936	0.889

Conclusion

- The model made with the highest accuracy label was trained with the tree based algorithm.
- Our model is able to predict obesity levels based on eating habits and physical conditions with an accuracy of 94%.