Modern methods of Data Analysis

Group project

Authors: Nikita Kharzheev, Alina Potemkina, Mikhail Sambuev

1. Problem statement

An analysis of global data estimates that the total number of children, adolescents and adults worldwide who are obese has exceeded one billion. The results of the study were published in the scientific journal The Lancet. For this reason, our main objective of the study is to analyze the relationship between eating habits and physical condition to further evaluate the level of obesity among people from Mexico, Peru and Colombia.

2. Summary of the dataset

Dataset Description: The dataset contains information on 2111 individuals from Mexico, Peru, and Colombia, aged between 14 and 61.

Attributes:

Eating Habits:

- Frequent consumption of high caloric food (FAVC)
- Frequency of consumption of vegetables (FCVC)
- Number of main meals (NCP)
- Consumption of food between meals (CAEC)
- Consumption of water daily (CH20)
- Consumption of alcohol (CALC)

Physical Condition:

- Calories consumption monitoring (SCC)
- Physical activity frequency (FAF)
- Time using technology devices (TUE)
- Transportation used (MTRANS)

DataFrame information:

```
<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
```

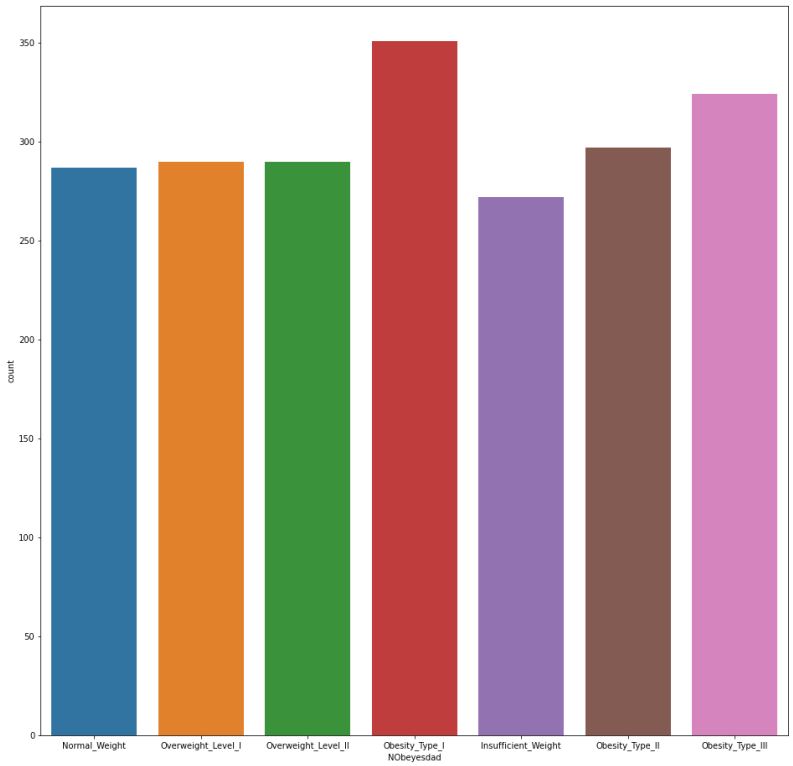
Categorical types of data

- Gender : ['Female', 'Male']
- family_history_with_overweight : ['yes', 'no']
- FAVC : ['no', 'yes']
- CAEC : ['Sometimes', 'Frequently', 'Always', 'no']
- SMOKE : ['no', 'yes']
- SCC : ['no', 'yes']
- CALC : ['no', 'Sometimes', 'Frequently', 'Always']
- MTRANS : ['Public_Transportation', 'Walking', 'Automobile', 'Motorbike', 'Bike']
- NObeyesdad : ['Normal_Weight', 'Overweight_Level_I', 'Overweight_Level_II', 'Obesity_Type_I', 'Insufficient_Weight', 'Obesity_Type_II', 'Obesity_Type_III']

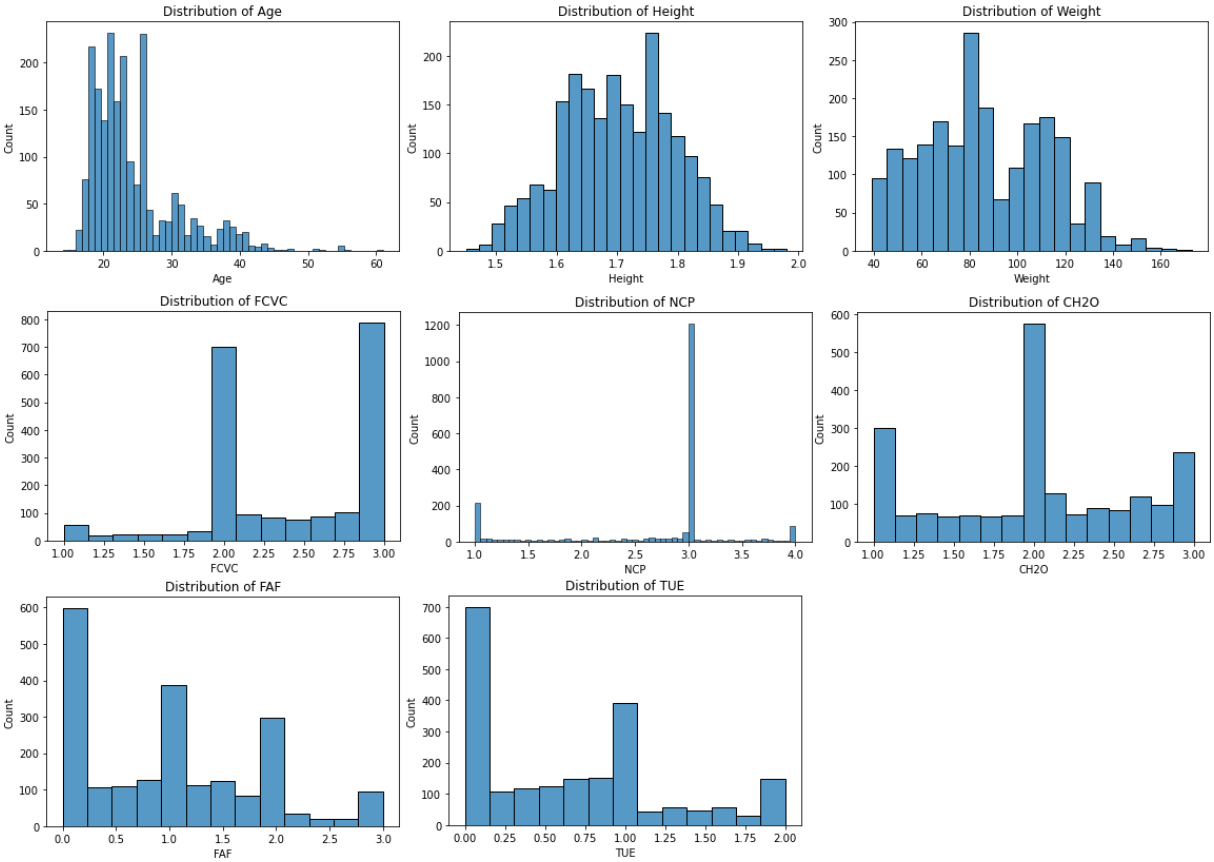
Numeric types of data

- Age, Height, Weight, FCVC, NCP, CH2O, FAF, TUE

Graph of the distribution of the number of people by degree of obesity



Column distribution



3. Methodology

In our research, we decided to take three methods: XGBoost, Random Forest and Logistic Regression. Because they provide a balanced approach to analyzing data and predicting obesity levels. XGBoost and Random Forest provide high accuracy and flexibility when working with large and complex data sets, while Logistic Regression provides interpretable results and is an excellent baseline model.

Preprocessing

We have two types of data: categorical and numeric. For the categorical data type, we did preprocessing. Namely, we created a function where two loops are processed. In the first loop, categorical data has binary variables. And in the second there are more than two variables. In this function we have converted categorical variables into numeric representations.

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	SMOKE	CH2O	...	CAEC_no	CALC_Always	CAL
328	1	19.000000	1.740000	74.000000	1	1	3.000000	1.000000	1	3.000000	...	0.0	0.0	
1651	1	25.027254	1.757154	112.200812	1	1	1.264234	3.000000	1	2.000000	...	0.0	0.0	
1980	1	25.795187	1.669039	104.593929	1	1	3.000000	3.000000	1	1.554925	...	0.0	0.0	
2003	1	26.000000	1.633195	111.883747	1	1	3.000000	3.000000	1	2.619517	...	0.0	0.0	
730	1	17.925497	1.829142	59.933015	1	1	2.860990	4.000000	1	2.000000	...	0.0	0.0	
207	1	30.000000	1.750000	73.000000	1	1	2.000000	3.000000	1	2.000000	...	1.0	0.0	
853	1	19.000000	1.779882	80.091886	1	1	1.078529	1.211606	1	2.568063	...	0.0	0.0	
647	1	20.744839	1.667852	49.803921	1	1	2.977018	3.193671	1	2.482933	...	0.0	0.0	
1327	1	29.389239	1.681855	90.000000	1	1	2.088410	2.644692	1	2.773236	...	0.0	0.0	
825	1	37.455752	1.508908	63.183846	1	1	2.048582	1.047197	1	2.000000	...	0.0	0.0	
989	1	20.392665	1.525234	65.220249	1	1	2.451009	3.000000	1	2.000000	...	0.0	0.0	
955	1	34.772902	1.675612	73.501233	1	1	3.000000	2.400943	1	1.509734	...	0.0	0.0	
689	1	16.834813	1.744020	50.000000	1	1	2.190050	3.420618	1	1.356405	...	0.0	0.0	
578	1	19.000000	1.530875	42.000000	1	1	2.844607	1.273128	1	1.695510	...	0.0	0.0	
281	1	18.000000	1.700000	55.000000	1	1	2.000000	3.000000	1	2.000000	...	0.0	0.0	
1803	1	26.000000	1.656320	111.933010	1	1	3.000000	3.000000	1	2.774014	...	0.0	0.0	
1700	1	36.839761	1.742850	106.421042	1	1	2.541785	2.902639	1	1.000000	...	0.0	0.0	
1984	1	20.781751	1.734092	125.117633	1	1	3.000000	3.000000	1	1.542490	...	0.0	0.0	
190	1	20.000000	1.600000	56.000000	1	1	2.000000	3.000000	1	2.000000	...	0.0	0.0	
2092	1	26.000000	1.626483	111.357062	1	1	3.000000	3.000000	1	2.619390	...	0.0	0.0	

4. Experiment setup and results; error analysis

In the first experiment we will take all the features.

We divided the data into training and test sets

Logistic Regression

We also defined these parameters for Gridsearch:

```
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],  # Regularization parameter
    'penalty': ['l1', 'l2']  # Penalty term
}
```

For validation we used 5 splits.

Random Forest Classifier

We also defined these parameters for Gridsearch:

```
param_grid = {
    'max_depth': [8, 10, 12],
    'max_leaf_nodes': [50, 75],
    'min_samples_split': [50, 75, 100, 150, 200]
}
```

For validation we used 5 splits.

XGBoost Classifier

We also defined these parameters for Randomizedsearch:

```
param_grid = {
    'learning_rate': [0.01, 0.1],  # Boosting learning rate
    'max_depth': [3, 5, 7],  # Maximum tree depth for base learners
```

'reg_alpha': [0.1, 0.5, 1] # L1 regularization term on weights
}

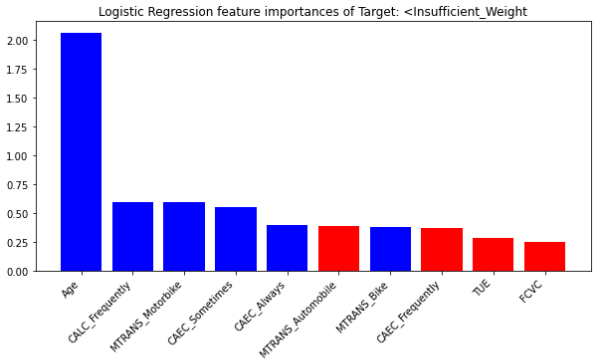
For validation we used 5 splits.

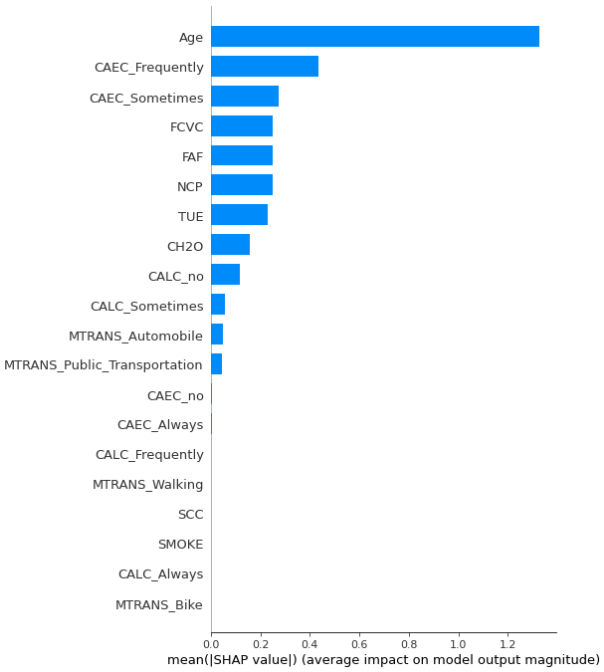
	Logistic Regression	Random Forest Classifier	XGBoost Classifier
Best parameters	{'C': 100, 'penalty': 'l1'}	{'max_depth': 12, 'max_leaf_nodes': 50, 'min_samples_split': 50}	{'reg_alpha': 0.1, 'max_depth': 5, 'learning_rate': 0.1}
CV F1_score	0.6982	0.8304	0.9469
Train F1_score	0.7278	0.886	0.9924
Test F1_score	0.7107	0.8526	0.9356

In the second experiment we will take all the characteristics except Weight and Height. Taking into account weight and height, you can calculate the body mass index, and this is an assessment of obesity.

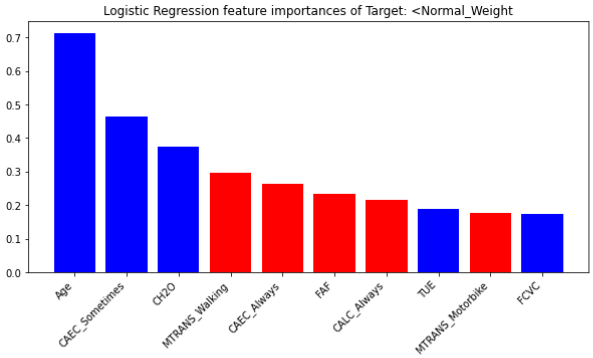
	Logistic Regression	Random Forest Classifier	XGBoost Classifier
Best parameters	{'C': 100, 'penalty': 'l1'}	{'max_depth': 10, 'max_leaf_nodes': 50, 'min_samples_split': 50}	{'reg_alpha': 0.5, 'max_depth': 7, 'learning_rate': 0.1}
CV F1_score	0.4914	0.6399	0.7414
Train F1_score	0.5205	0.7332	0.9476
Test F1_score	0.5015	0.666	0.7834

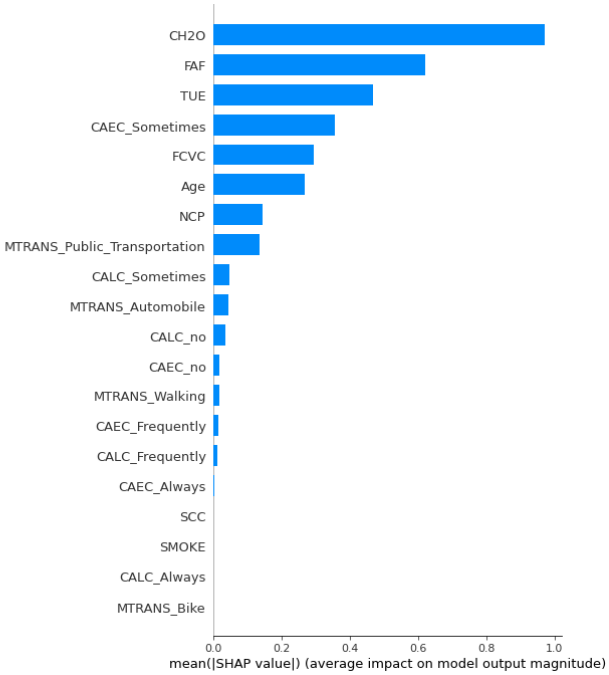
Feature importance for Insufficient_Weight



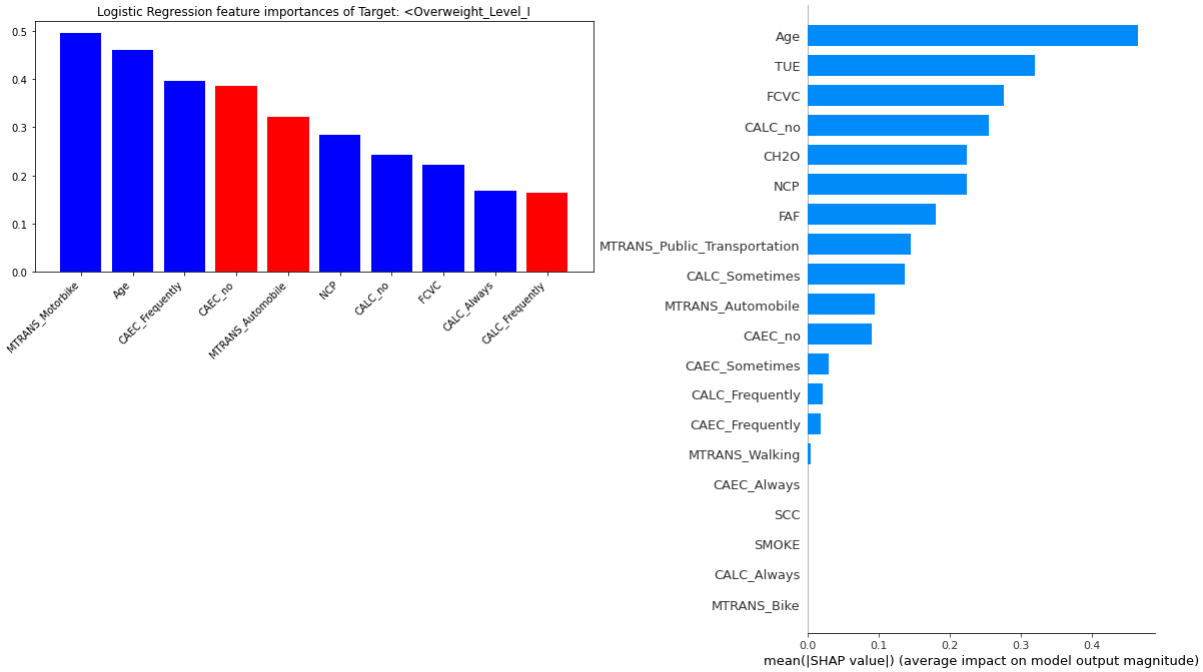


Feature importance for Normal_Weight

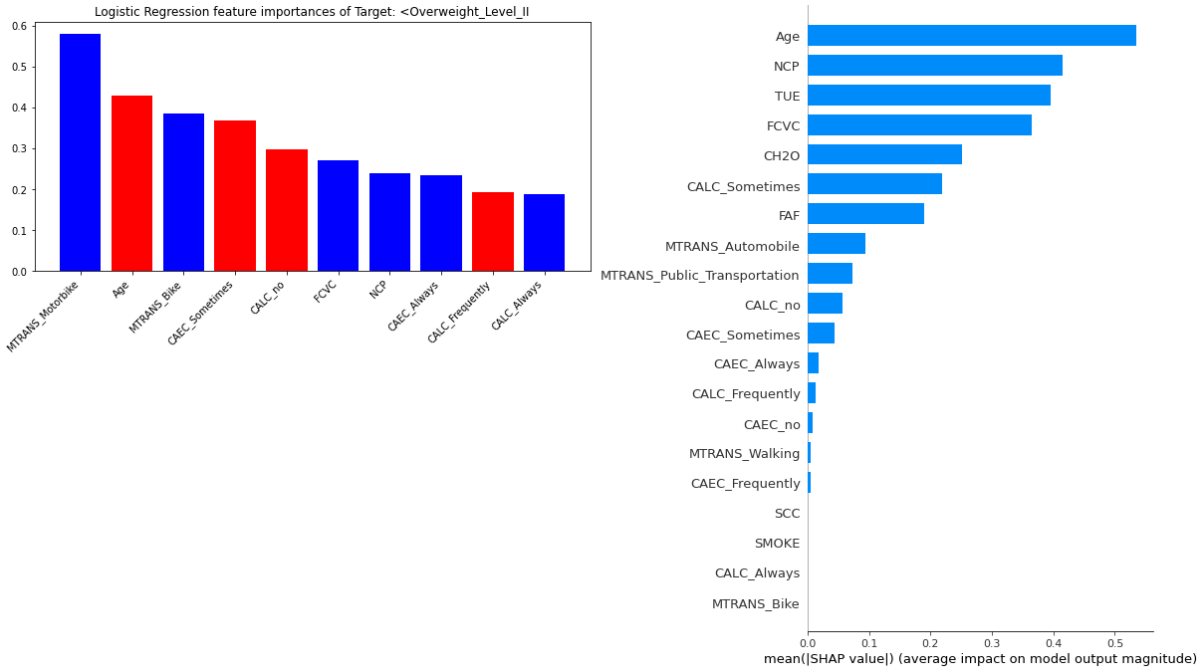




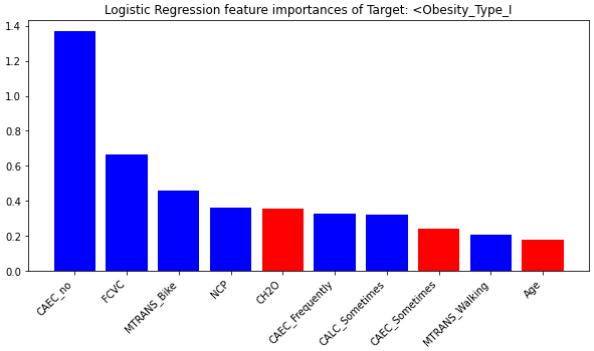
Feature importance for Overweight_Level_I

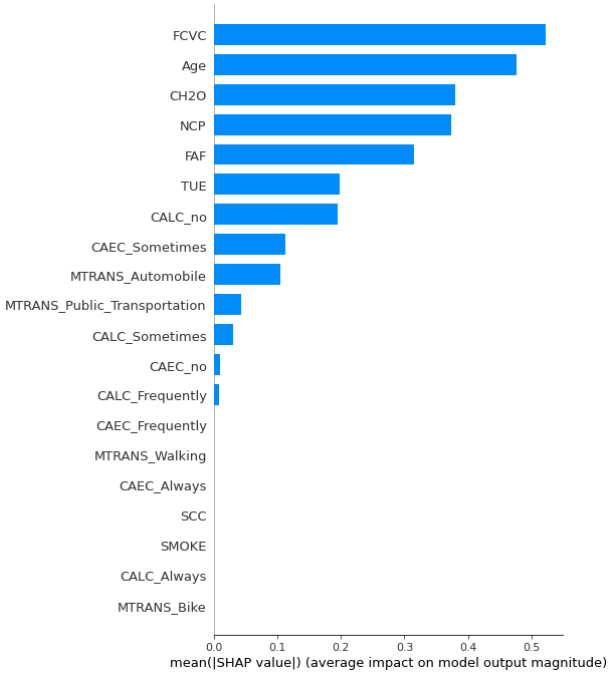


Feature importance for Overweight_Level_II

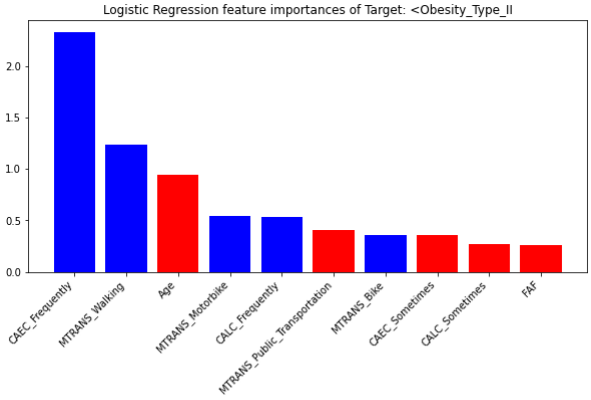


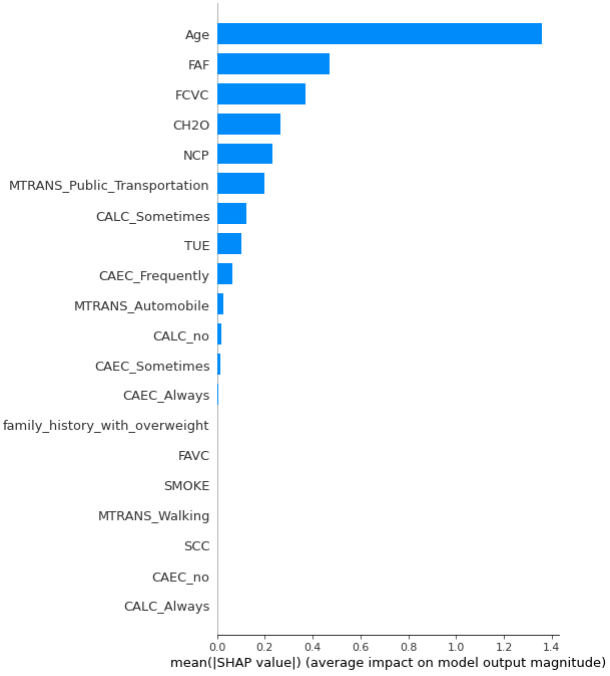
Feature importance for Obesity_Type_I



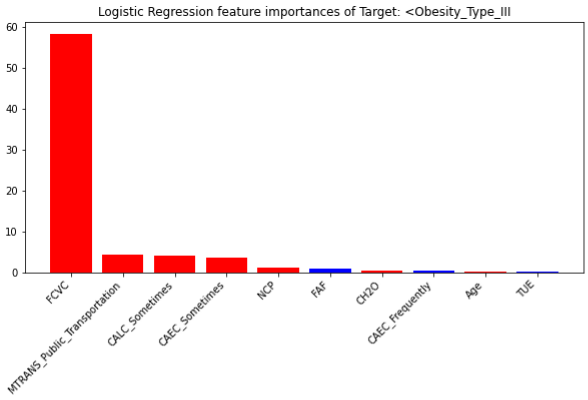


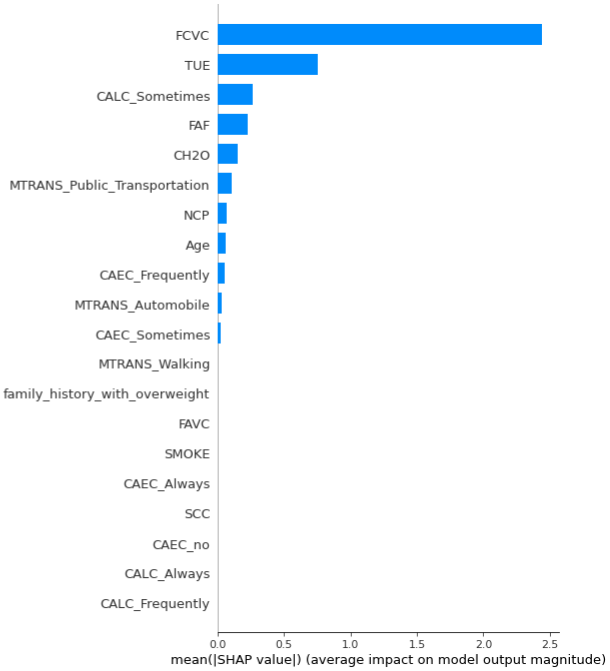
Feature importance for Obesity_Type_II





Feature importance for Obesity_Type_III





5. Discussion

For the first experiment we took in account all features including height and weight. So as a result we faced the problem named ‘data leakage’. In our case it means that to predict risk of obesity we look at current weight and height. It’s like our model work in such way:

So as we can see the scores for all models in the first experiment were highly overrated.

In the second experiment we dropped height and weight information to determine factors which have influence on obesity level. As a result we search for the feature importance to find factors that influence obesity risk. For feature importance we used two ways: coefficients of logistic regression and shap values for boosting. Result of the comparison is shown above.

For example related features for people with normal weight are:

- water consumption
- physical activity

And for the people with obesity III type:

- Frequency of consumption of vegetables

According to those results we can form some advices to maintain obesity risk.

6. Conclusion

In this work we have taken a look at the problem of obesity in the modern world. To do that we preprocessed the data, including categorical data type. Next, we chose 3 methods: XGBoost, Random Forest and Logistic Regression. We trained each type of model and obtained accuracy. As a result, these methods were compared and the results were interpreted. In addition we made some research for feature importance to summarize our conclusions.

Eat more vegetables and lead a healthy lifestyle to reduce your risk of obesity.