PYTHON FOR DATA ANALYSIS

Obesity Data Set

DIA4 – LIMNAVONG Thomas, KULUMBA Francis

SOMMAIRE

PROBLEM DEFINITION

Python for Data Analysis

DIA4 – LIMNAVONG Thomas, KULUMBA Francis

Problem

Data coming from a survey of 3 differents countries

Colombia

Peru

Mexico

Based on their eating habits and physical condition.

Problem

The data contains 17 attributes and 2111 records, the records are labeled with the class variable NObesity (Obesity Level), that allows classification of the data using the values of Insufficient Weight, Normal Weight, Overweight Level I, Overweight Level II, Obesity Type I, Obesity Type II and Obesity Type III.

Data Set

17 x 21111

ESILV
ENGINEERING SCHOOL
DE VINCI PARIS

Category	Feature Name	Description	Variable Type
Target Variables	Nobesity	Based on BMI	Categorical
Eating Habits	FAVC	Frequent consumption of high caloric food	Categorical
Eating Habits	FCVC	Frequency of consumption of vegetables	Ordinal
Eating Habits	NC¨P	Number of main meals	Ordinal
Eating Habits	CAEC	Consumption of food between meals	Ordinal
Eating Habits	CH20	Consumption of water daily	Ordinal
Eating Habits	CALC	Consumption of alcohol	Ordinal
Physical Condition	SCC	Calories consumption monitoring	Categorical
Physical Condition	FAF	Physical activity frequency	Ordinal
Physical Condition	TUE	Time using technology devices	Ordinal
Physical Condition	MTRANS	Transportation used	Categorical
Physical Condition	SMOKE	Smokes Yes or No	Categorical
Responder Characteristics	Family History with Overweight	Yes or No	Categorical
Responder Characteristics	Gender	Gender Male or Female	Categorical
Responder Characteristics	Age	Age in years	Integer
Responder Characteristics	Height	Height in meters	Float
Responder Characteristics	Weight	Weight in kilograms	Float

DATA DISTRIBUTION AND EXCEPTIONS

Python for Data Analysis

DIA4 – LIMNAVONG Thomas, KULUMBA Francis

Synthetic Data

77% of the data was generated synthetically with WEKA, 23% of the data was collected directly from users through a web platform.

In order to avoid imballenced data. In fact, th population's BMI is not equally distributed. So is the data collected. Thus they might end up with more people with a normal weight and very few people with Obesity Type III.
Computing a model out of this kind of data creates biais where the model is very precise but with a bad recall (it is mostly trained to recognize people with normal weight).

Synthetic Data

Different ways to generate sythetic data:

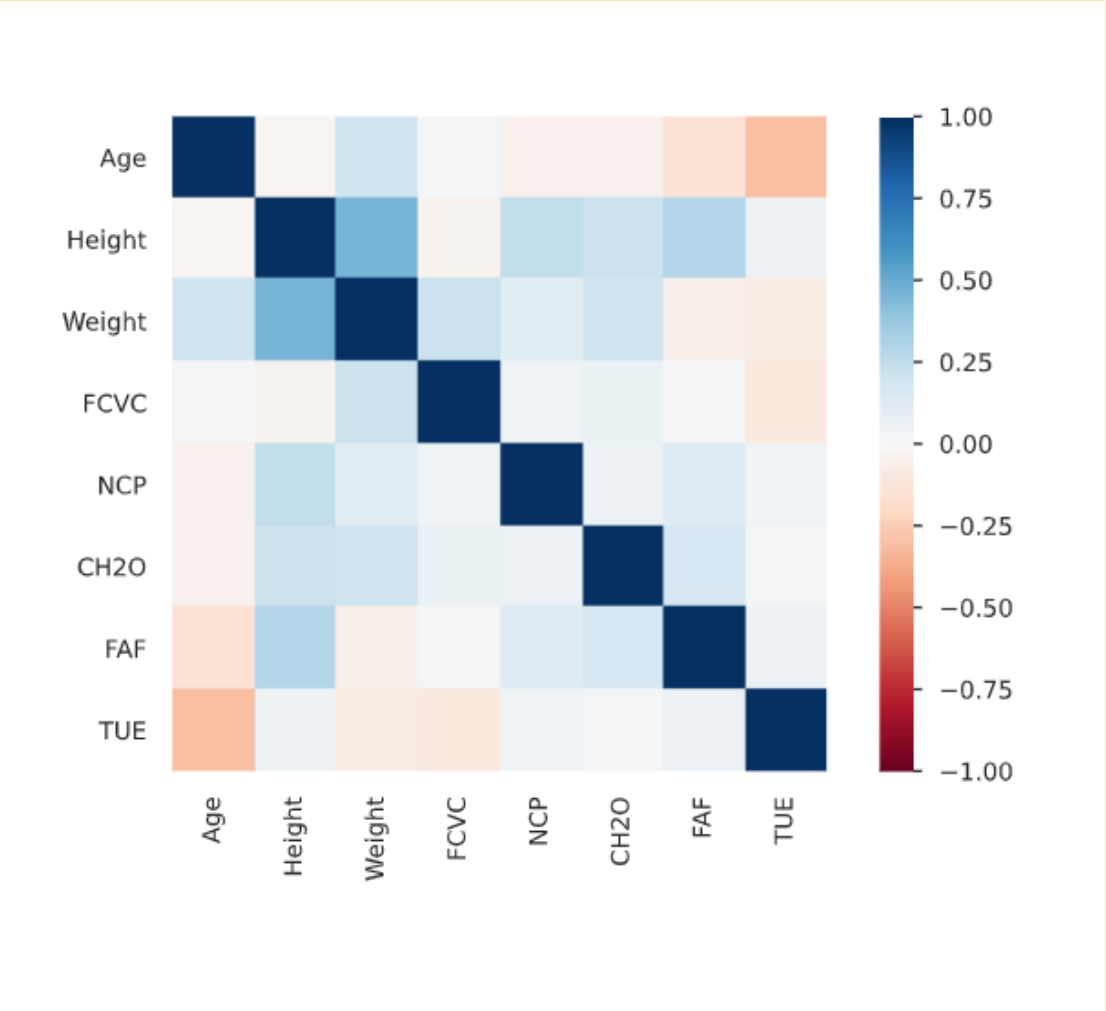
- Taking random observations for each dimensions within a target and mix them up to create a new data point.
- For continuous variable, taking random a number between the minimum and the maximum observation within a target.
- Creating it manually by field knowledge.

Missing Values/Duplicates

No missing values to handle here

Only 4 duplicates rows which is a few and possible in real world, we keep them.

Linear Combinations



Height and Weight are strongly correlated (Pearson correlation)

However our target is based on BMI calculation

BMI = Weight/Height²

So our target is a linear combination of age and weight, thus we delete those 2 dimensions.

MACHINE LEARNING

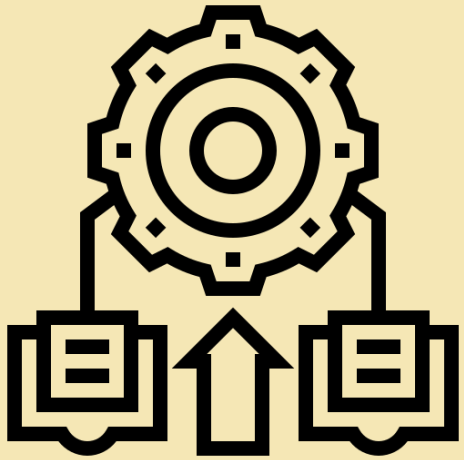
Python for Data Analysis

DIA4 – LIMNAVONG Thomas, KULUMBA Francis

Best Algorithms

There are 17 features, no need to reduce the dimensionality.

There are a lot of categorical features, thus our first thought was to apply SVM since the data point should have good defined distance within each features.
We don't have a lot of features, so decision trees and random forest should also do great.



Results

Random forest

```
Random Forest:
----------------------------------------------------------------
Accuracy: 0.82177
Accuracy w/Scaled Data (ss): 0.82177
Accuracy w/Scaled Data (mm): 0.82177

Classification Report (mm):
                    precision    recall  f1-score   support

Insufficient_Weight       0.85      0.87      0.86        92
      Normal_Weight       0.60      0.69      0.64        77
     Obesity_Type_I       0.85      0.80      0.82       114
    Obesity_Type_II       0.90      0.94      0.92        85
   Obesity_Type_III       0.99      0.99      0.99        92
 Overweight_Level_I       0.79      0.71      0.75        89
Overweight_Level_II       0.76      0.74      0.75        85

           accuracy                           0.82       634
          macro avg       0.82      0.82      0.82       634
       weighted avg       0.83      0.82      0.82       634

                    ----------------------------------
```

Results

Decision tree

```
Decision Tree:
----------------------------------------------------------------
Accuracy: 0.76183
Accuracy w/Scaled Data (ss): 0.76025
Accuracy w/Scaled Data (mm): 0.75394

Classification Report:
                     precision    recall  f1-score   support

Insufficient_Weight       0.82      0.82      0.82        92
      Normal_Weight       0.53      0.52      0.52        77
     Obesity_Type_I       0.80      0.72      0.76       114
    Obesity_Type_II       0.86      0.87      0.87        85
   Obesity_Type_III       0.98      0.99      0.98        92
 Overweight_Level_I       0.72      0.67      0.70        89
Overweight_Level_II       0.60      0.72      0.65        85

           accuracy                           0.76       634
          macro avg       0.76      0.76      0.76       634
       weighted avg       0.77      0.76      0.76       634

                                 ----------------------------------
```

Results

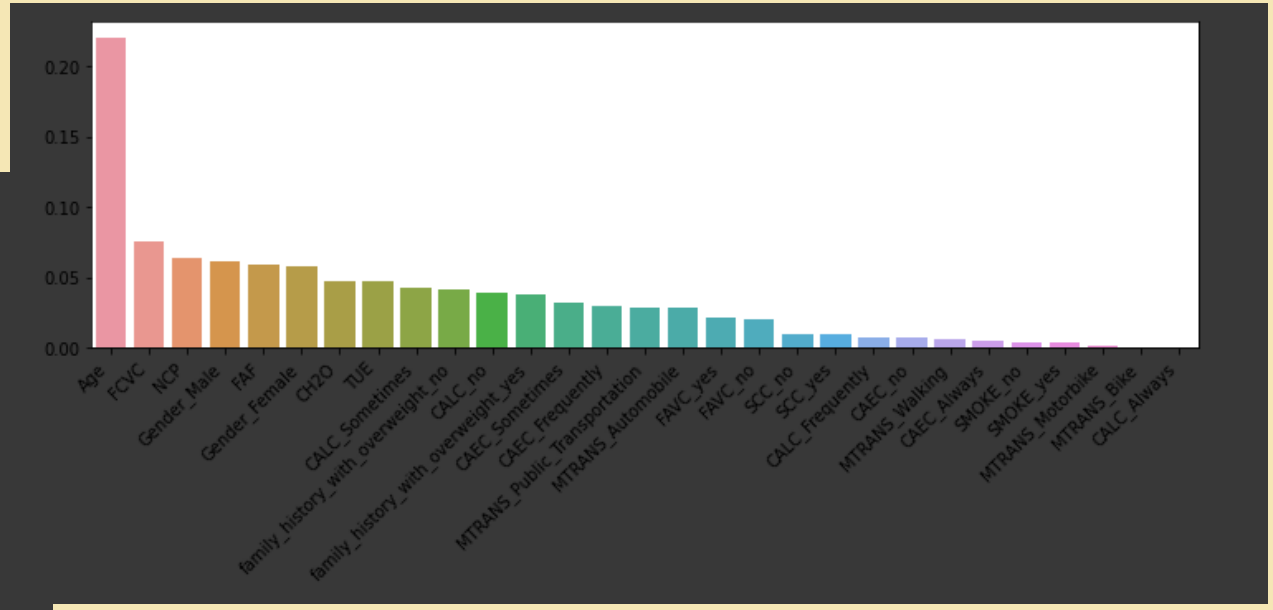
Support Vector Machin

```
SVM:
------------------------------------------------------------
Accuracy: 0.47319
Accuracy w/Scaled Data (ss): 0.71609
Accuracy w/Scaled Data (mm): 0.71609

Classification Report (mm):
                    precision    recall  f1-score   support

Insufficient_Weight       0.79      0.80      0.80        92
      Normal_Weight       0.54      0.66      0.60        77
     Obesity_Type_I       0.65      0.58      0.61       114
    Obesity_Type_II       0.70      0.98      0.82        85
   Obesity_Type_III       0.98      0.98      0.98        92
 Overweight_Level_I       0.64      0.49      0.56        89
Overweight_Level_II       0.70      0.54      0.61        85

          accuracy                           0.72       634
         macro avg       0.71      0.72      0.71       634
      weighted avg       0.72      0.72      0.71       634

------------------------------------
```

Results

Top features by contribution

```
Top 10 Features:
Variable: Age                          Importance: 0.22
Variable: FCVC                         Importance: 0.07
Variable: NCP                          Importance: 0.06
Variable: FAF                          Importance: 0.06
Variable: Gender_Female                Importance: 0.06
Variable: Gender_Male                  Importance: 0.06
Variable: CH2O                         Importance: 0.05
Variable: TUE                          Importance: 0.05
Variable: family_history_with_overweight_no Importance: 0.04
Variable: family_history_with_overweight_yes Importance: 0.04

Bottom 10 Features:
Variable: CAEC_Always                  Importance: 0.0
Variable: SMOKE_no                     Importance: 0.0
Variable: SMOKE_yes                    Importance: 0.0
Variable: CALC_Always                  Importance: 0.0
Variable: MTRANS_Bike                  Importance: 0.0
Variable: MTRANS_Motorbike             Importance: 0.0
Variable: CAEC_no                      Importance: 0.01
Variable: SCC_no                       Importance: 0.01
Variable: SCC_yes                      Importance: 0.01
Variable: CALC_Frequently              Importance: 0.01
```



BIBLIOGRAPHY

Python for Data Analysis

DIA4 – LIMNAVONG Thomas, KULUMBA Francis

Bibliography

F. Mendoza Palechor, A. de la Hoz Manotas, "*Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico*", https://www.sciencedirect.com/science/article/pii/S2352340919306985?via%3Dihub

F. Mendoza Palechor, A. de la Hoz Manotas, "Estimation of obesity levels based on eating habits and physical condition Data Set", https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+