Python final project

Dataset : Obesity

The context

- This dataset is about health. It is a mix between real data, gathered from a survey, about public health in Mexico, Peru and Colombia, and artificially generated data, to complete it.
- The informations about each individual are : the age, height in meters, weight in kilograms, the family history with overweight, the habit of eating high caloric food, the habit of eating vegetables, the daily number of meals, the habit of eating in between meals, the smoking habit, the amount of water drank daily, if the person is monitoring their food amount of calories, the physical and technological activity times, the alcohol habits, and the preferred transportation method.

The Dataset

- This dataset is a mix of real data and generated data.
- It has over 2000 data points, quite evenly spread, with no null value, which makes it easier to interpret
- Some datapoints were created artificially. These points have been created to spread more evenly the categories, and make it easier for an AI to recognize the categories.

The Strangeness

- Some data was artificially created in the dataset. Why?
  - Because the dataset was too small for an AI to work properly
  - Because the results were not spread enough (a bias too heavy for one particular weight category)
- The dataset has 7 categories, when only 6 are defined in the article.
  - It most likely was done wilingly, with one category being split into two.
  - We will see which solution i came up with later on.

The data outside of the data :

- In the original paper, it was mentionned that the type of weight of a person is determined based on the following formula (MBI stands for Mass Body Index)

$$MBI=weight/height^2$$

- And the persons are in a category according to their MBI :

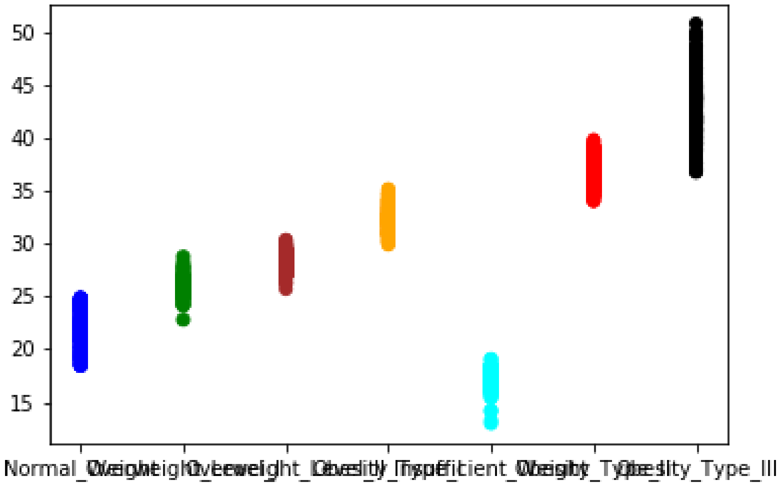
Minimum value (included)	Weight category	Maximum value (excluded)
0	Underweight	18.5
18.5	Normal	25
25	Overweight	30
30	Obesity I	35
35	Obesity II	40
40	Obesity III	∞

The problem

- Our task is to determine in which weight group a given person is.
- If possible, without using the height and weight.

Issue in the data

- As mentionned before, the dataset has split a category into two (2 different types of overweight, when only one is defined in the article)
- And some categories are not as separated as expected, since some datapoints are categorized as obesity type III when they are supposed to be categorize as type II, for example.



Solving the issue

- In light of this issue, I will use 3 model for this project :
  - One that is not an AI, and just calculates the MBI of the individual, given the height and the weight
  - Another one that will be trained without any data about height and weight, and that will use the given categories to classify.
  - And a last one, trained using the categories from the article, determined by the MBI, still without knowing the height and the weight of the individual.

Methods used

- For the first task, it will simply use the formula for the MBI, and give the category linked to the result.

- For the second and third tasks (later first and second models), since the problem at hand is a problem of classification among multiple categories, I have chosen to use the random forest algorithm, since it is a good AI model for such a task.

- To test both models, I split the data into training and test sets, with a ratio of around 70%-30%.

The results

- The non AI is, obviously, right 100% of the time, in real data, and not in the dataset (for the reason given earlier), since it is just a simple formula to apply.
- The first model, using the categories from the dataset, is right 95% of the time. It is a quite precise model.
- The second model is only right 72% of the time.
- The fact that the first model is better than the second is not a surprise. The artificial data has been added to better the AIs using the datasets.