Reviewable project Date-A-Scientist

MACHINE LEARNING FUNDAMENTALS

PABLO FERNANDEZ

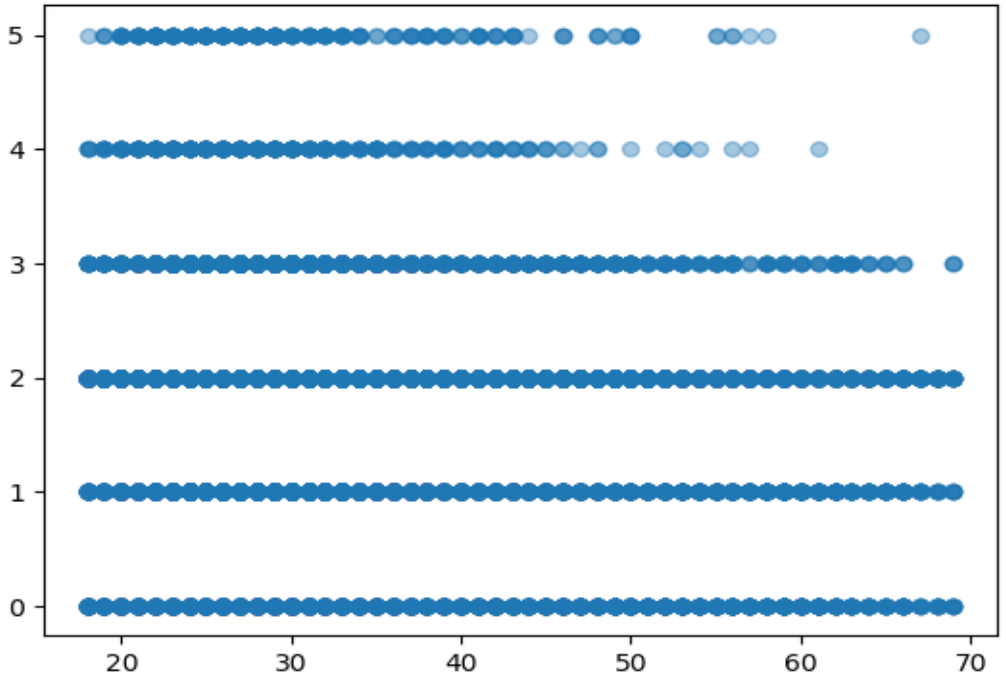
12/11/2018

Contents

- Exploration of the Dataset
- Question(s) to Answer
- Augmenting the Dataset
- Classification Approaches
- Regression Approaches
- Conclusions/Next steps

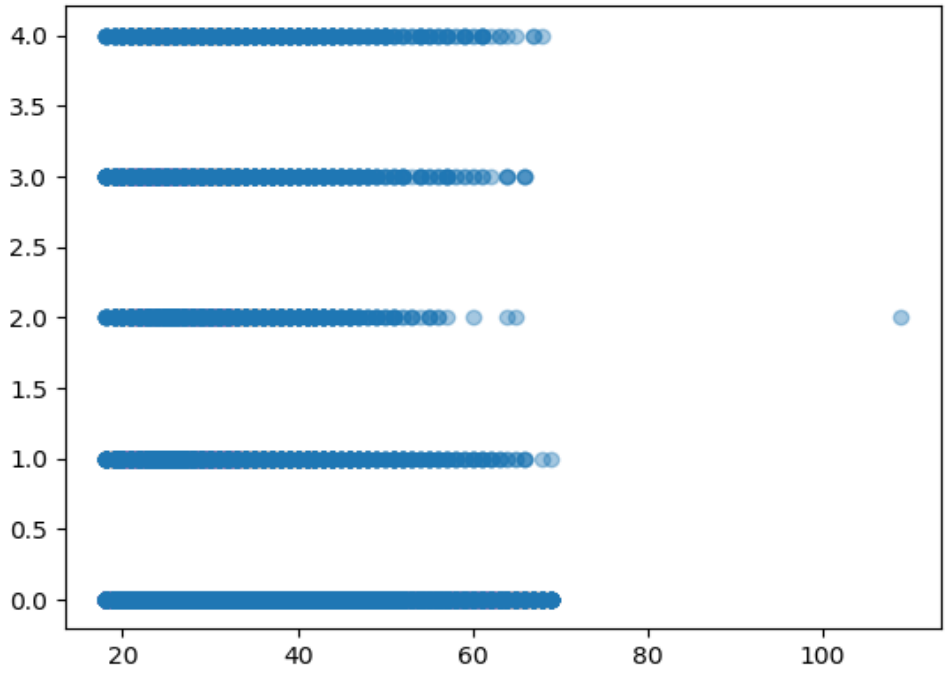
Exploration of the Dataset

- I have checked the relation betweern age and alcohol consumption giving no relevant relation. People of different ages consume different quantities of alcohol.



Exploration of the Dataset

- I have checked the relation betweern age and cigarettes consumption giving no relevant relation. People of different ages consume different quantities of cigarettes.



Question(s) to Answer

- I would like to check if we can predict the body type given the diet type, alcohol and smoking behaviour of the people, height and age.

Augmenting the Dataset

- I had to generate 4 new columns to answer my question.

```
drink_mapping = {"not at all": 0, "rarely": 1, "socially": 2, "often": 3, "very often": 4, "desperately": 5}
smokes_mapping = {"no": 0, "sometimes": 1, "when drinking": 2, "yes": 3, "trying to quit": 4}
drugs = {"never": 0, "sometimes": 1, "often": 2}
diet = {"mostly anything": 0, "anything": 0, "strictly anything": 0,
        "mostly vegetarian": 1, "mostly other": 2, "strictly vegetarian": 1,
        "vegetarian": 1, "strictly other": 2, "other": 2, "mostly vegan": 3,
        "strictly vegan": 3, "vegan": 3, "mostly kosher": 4, "mostly halal": 5,
        "strictly halal": 5, "strictly kosher": 4, "kosher": 4, "halal": 5}
```

Classification Approaches

KNeighbors

- Accuracy score: 0.2960658737419945
- Recall score: 0.2960658737419945
- Precision score: 0.2960658737419945

- Time to run the model: 3.6458871999999998

Classification Approaches

SVM

- Accuracy score: 0.2810612991765782 4
- Recall score: 0.2810612991765782 4
- Precision score: 0.2810612991765782 4

- Time to run the model: 127.52253

Regression Approaches

Multiple Linear Regression

- Accuracy score: 0.2194876486733760 3
- Recall score: 0.2194876486733760 3
- Precision score: 0.2194876486733760 3

- Time to run the model: 0.0799790000000000 1

Regression Approaches

K-Nearest Neighbors Regression

- Accuracy score: 0.296065873741994 5
- Recall score: 0.296065873741994 5
- Precision score: 0.296065873741994 5

- Time to run the model: 3.6339899

Conclusions/Next steps

- We can conclude that although the average score is around 0.26, well above 0.083 (probability of randomly guessing the body_type), is not enough to predict the body_type using features such as age, drink, smoke, etc.
- In the future we should augment the user data with more sports type information like trainings per week or sports habits so we can predict their body type with more precission.