Date a scientist

Machine Learning Fundamentals

Jacob Honoré

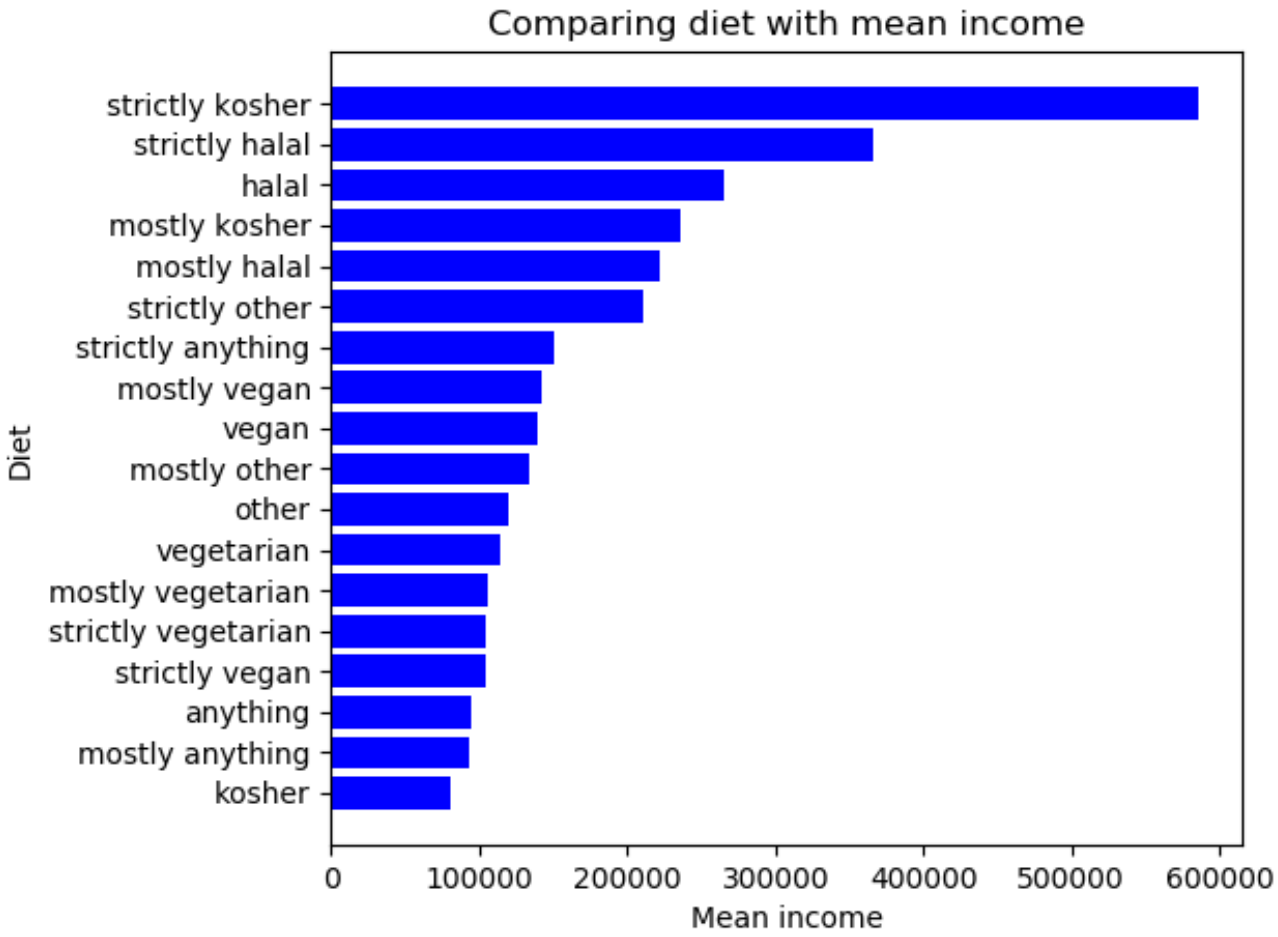
16-04-2019

codecademy

Table of Contents

Exploration of the Dataset

Comparing diet with mean income



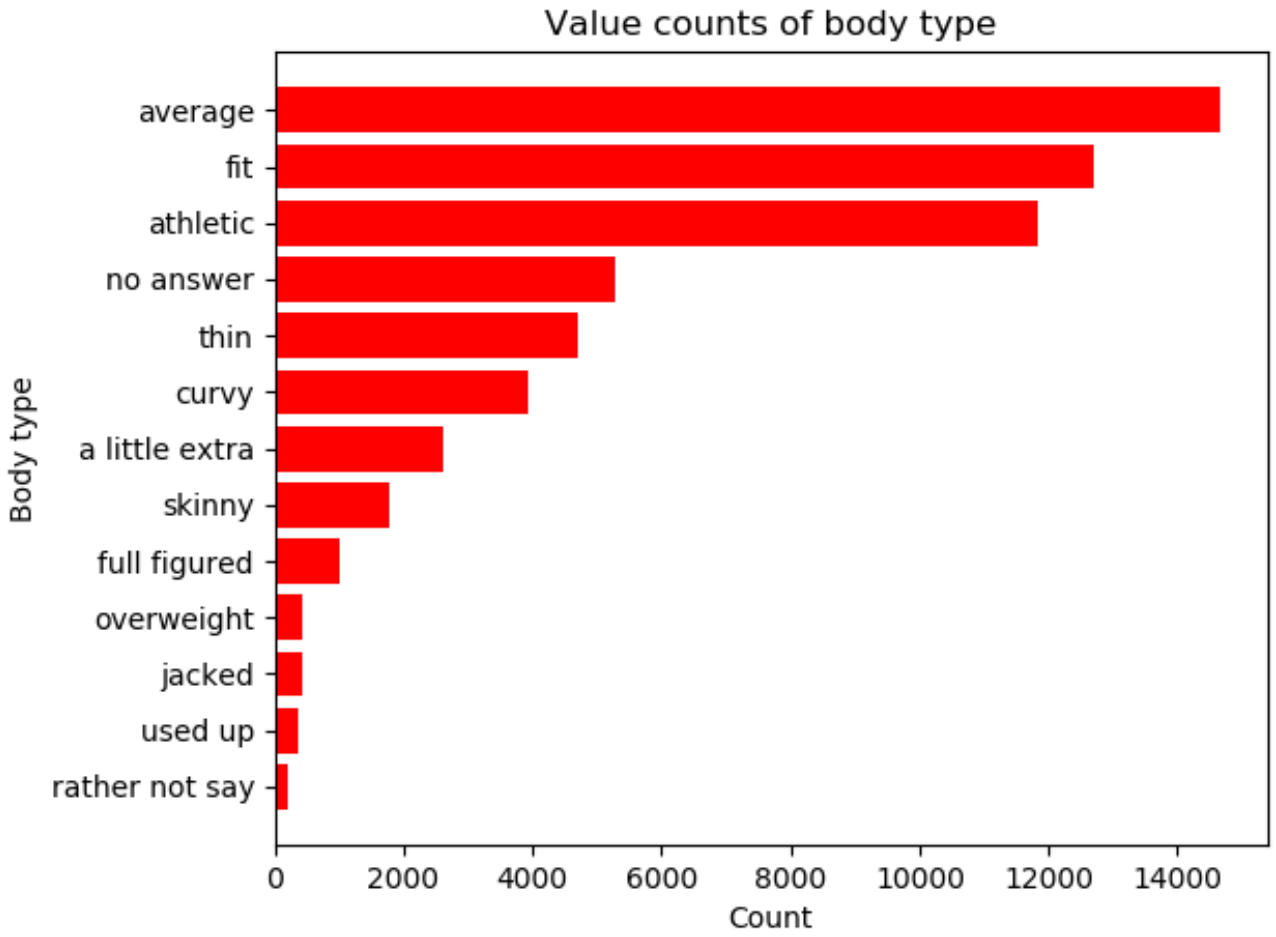
I wanted to explore if there was some correlation between diet and mean income.

I noticed during my exploration that there were very few ”strictly kosher” etc. So it should probably be normalized more. You could do this by mapping it into kosher/halal/other/anything/vegetarian. This mapping might show completely other results as the answers are too unbalanced now.

I made this graph by firstly excluding null answers and then grouping by diet to explore the mean income of each diet type.

Exploration of the Dataset

Value counts of body type



I wanted to explore what the body type column meant, and how many filled out data about this.

As shown, most people find themselves as average, fit or athletic. This shows that most people who are dating also keeps themselves fit. There might be a larger group of fit/average people because obese people are in the “no answer” category.

This graph has been made by firstly converting all null-answers into a new “no answer” category, then doing value count on the body types and sorting the graph to make it appear nicely.

It could be interesting to compare body type with diet.

Question to Answer

Out from my exploration of the dataset in the described graphs, I could not see any correlation between diet and income, by looking at the mean income of each diet type.

By looking at the different body types, I thought it would be a better idea to use this.

- Can I predict a user's diet by looking at body type and maybe other informations from their profile?

Augmenting the Dataset

Religion mapping

User	Religion containing	Religion seriousness
0	Laughing about it	0
1	Not too serious about it	1
2	No statement of seriousness	2
3	Somewhat serious about it	3
4	Very serious about it	4

I thought that religion seriousness can be used for predicting the diet of the user. To get a view of how serious the user is, I categorized it by looking at what the religion cells contained. I used this to create the ”Religion seriousness” column, as shown above a user is not categorised into 5 different categories, describing how serious they are about religion.

Augmenting the Dataset

Body type mapping

User	Body type	Body form
0	Overweight	0
1	Full figured	0
2	Used up	0
3	A little extra	0
4	Curvy	0
5	Average	1
6	Skinny	1
7	Fit	2
8	Athletic	2
9	Thin	2
10	Jacked	2

I thought that the body type would have impact on which diet the user is doing. The body types defined in the dataset is not very describable, so I chose to split it up in three categories, where the higher the number it is, the better the form.

The table to the left shows how the data has been mapped into the new “Body form” column.

Classification Approaches

Multinomial Naive Bayes Classifier

- Training runtime: 0.0591s
- Predicting runtime: 0.0025s
- Accuracy: 48,18%

K-Neighbors Classifier

- Training runtime: 0.0837s
- Predicting runtime: 0.3880s
- Accuracy: 48,18%

I do predictions on classification by using Naive Bayes Classifier vs. K-Neighbors Classifier. Both classifiers came up to the same accuracy, after predicting on a random training/validation set. The runtime however clearly shows that the Naive Bayes Classifier is both faster at training and predicting. It is significantly faster at predicting, which is why the preferred classifier would be Naive Bayes.

The results shows that I am on the right way, by using the select data to try to predict with classification.

Regression Approaches

Linear Regression

- Training runtime: 0.002s
- Predicting runtime: 0.0005s
- Accuracy: 00,3%

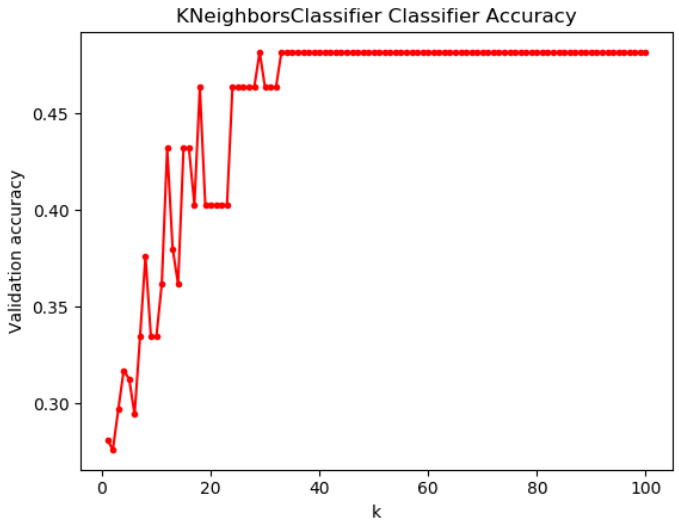
K-Neighbors Regressor

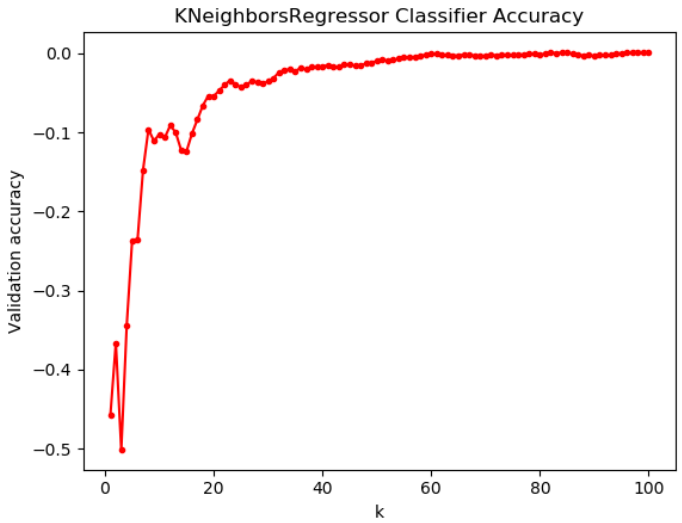
- Training runtime: 0.0532s
- Predicting runtime: 0.2637s
- Accuracy: 00,15%

The regression approaches clearly shows that regression is not the right way to answer the question. Both algorithms performs very poorly. Both alogorithms perform faster than the classification approaches.

The results shows that regression is not the right way to answer the question

Best K





The classifier performs OK from start and gets best when K is 29.
The regressor performs poorly and are actually making predictions that are completely wrong when K is lowest, it does however become better and better but will never be good enough for actual use.

Conclusions/Next steps

- The classification approaches are clearly best.
- Naive Bayes Classifier performs faster than K-Neighbors, but they show the same accuracy.
- Regression is not the right way to go for the question.
- I got promising results that shows that it might be possible to predict which diet a person is having, by using other data from the dataset.
- I would therefore go on with the Naive Bayes Classifier and look into which other data I could use to make it better and give better accuracy.