Capstone Project_

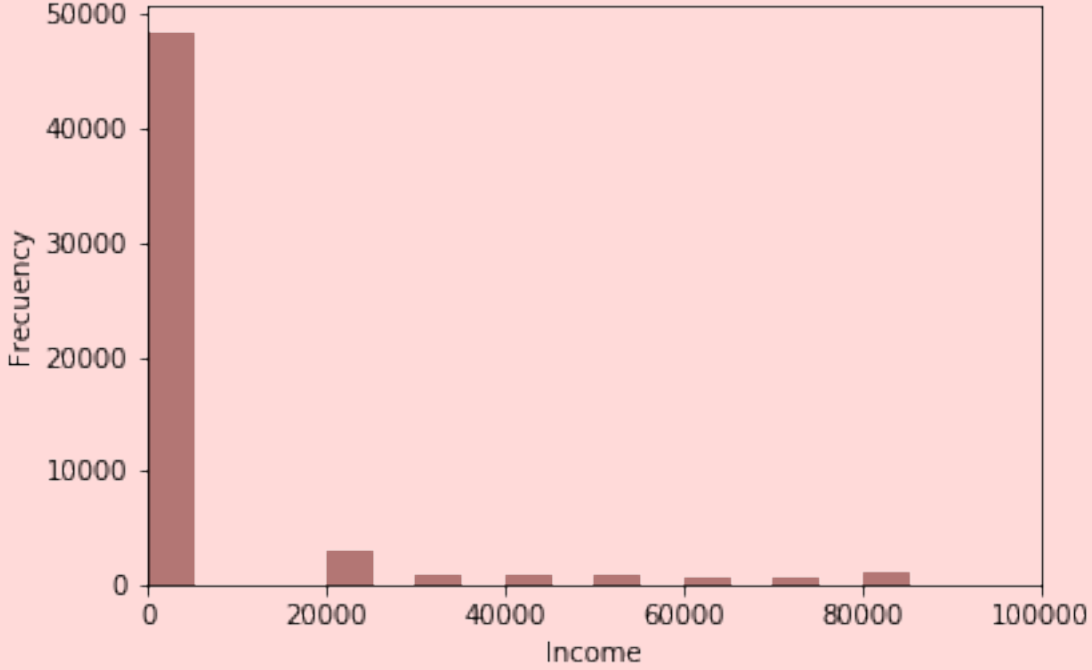
Machine Learning Fundamentals

M.Àngels Jover

22.04.2019

Table of Contents_

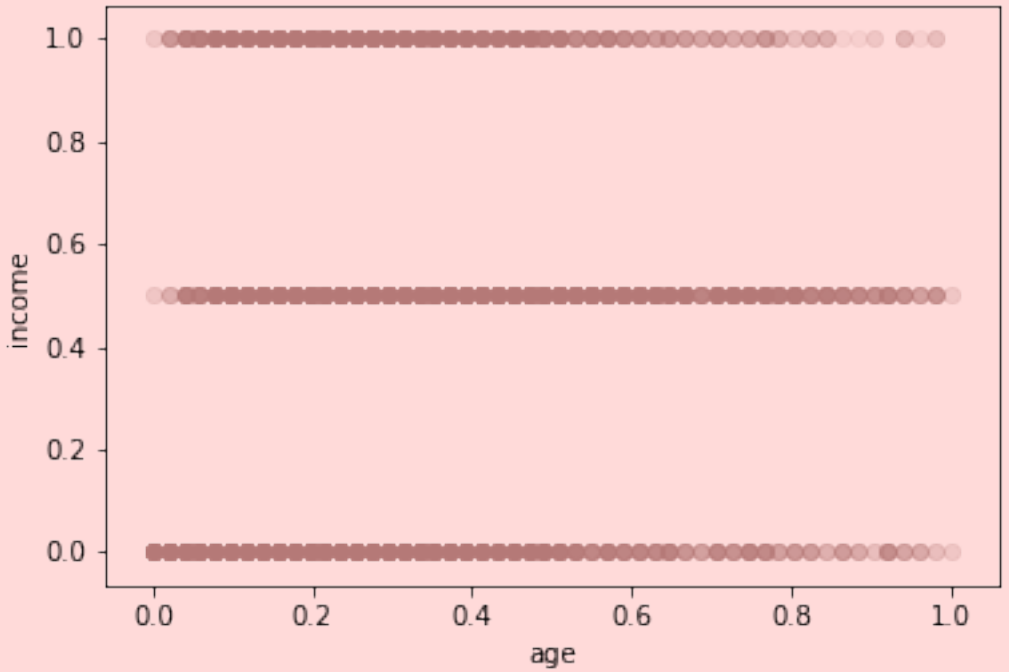
Exploration of the Dataset Graph #1_



I have plotted an histogram to show the frequency of the income among the profiles.

I have realized that the majority of the people either didn't want to respond to this question or didn't have any income. Most likely the first.

Exploration of the Dataset Graph #2_



I have assumed that:
< 40K$ is low income
<100K$ is middle income
>100K$ is high income

I understand this may vary between countries.

I wanted to see if after removing the -1 income I was able to find a correlation between income and age, this is why I have plotted this graph.

The insights are clear:

1. Low income decreases with age.
2. High income is more frequent between young and middle age people. Makes sense as this is when people reach full potential.
3. Most people in old ages have middle income.

Questions to Answer_

Being a parent increases the chance to drink, smoke & take drugs?

I am formulating this question because I am not a mother and every time I ask someone about parenthood I always get the same answer: “At the end, pays off” which for me is not positive at all, it implies a lot of negativity in it.

As their answers are not clear to me, I want to investigate though data whether exists a relation between being a parent and drinking, smoking and taking drugs, maybe to cope with this new responsability…?

Can I predict whether a person drinks knowing sex, if smokes or takes drugs?

I wanted to answer this question because at first glance it makes sense that at least smoking, taking drugs and drinks might be related.

Can I predict the income knowing job, offspring and whether drinks, smokes & takes drugs?

I have choosen to anwer this question because I think that some of them are factors that might have an impact on income.

Augmenting the Dataset_

offspring_cat	offspring_code
No	0
Yes	1
No	0
Yes	1

The columns on the left are the ones that I have created in order to be able to start with the analysis. There are two kinds of columns:

- The first one (offspring_cat) is a column I have created in order to convert the "offspring" column in a binary variable where I have catgorized as Yes when they have kids and No whether they don't have kids. I have created this column with an if statement over offspring column.

- The second one (offspring_code) is the conversion of the offspring_cat column to a numeric mapping.

I have repeated this with the columns drugs, smokes and drinks.

Classification Approaches_

Being a parent increases the chance to drink, smoke & take drugs?

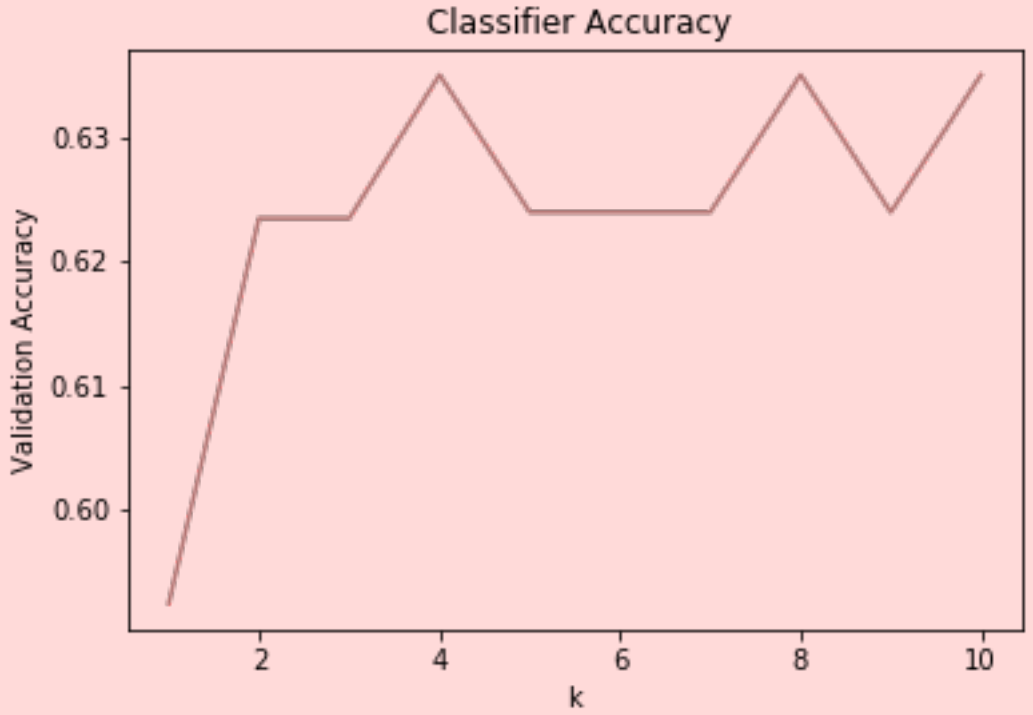
Native Bayes

After performing Native Bayes algorithm I have ended up with an accuracy score of 68% which has really impressed me. Too bad that precision, recall and f1 scores were 0 ☹. After checking the confusion matrix I see that the problem is that always predicts Yes.

KNeighbors

After performing Kneighbors I end up with an accuracy socre of 63%, recall score of 0.094, precision score of 0.29 and f1 score of 0.14, all pretty low, in line with Native Bayes.

In the graph on the left we can see how the best k is around 4.



Between the two models I would prefer KNeighbors over Native Bayes. Although it takes a lot more time to run the model, as finding k takes some time to run. I think KNeighbors is more accurate because finding k gives you more information regarding the model. Also, in this specific case, in the confusion matrix we can check that Kneighbors has predicted both outputs, which with Naives Bayes only predicted “Yes”.

Classification Approaches_

Can I predict whether a person drinks knowing sex, if smokes or takes drugs?

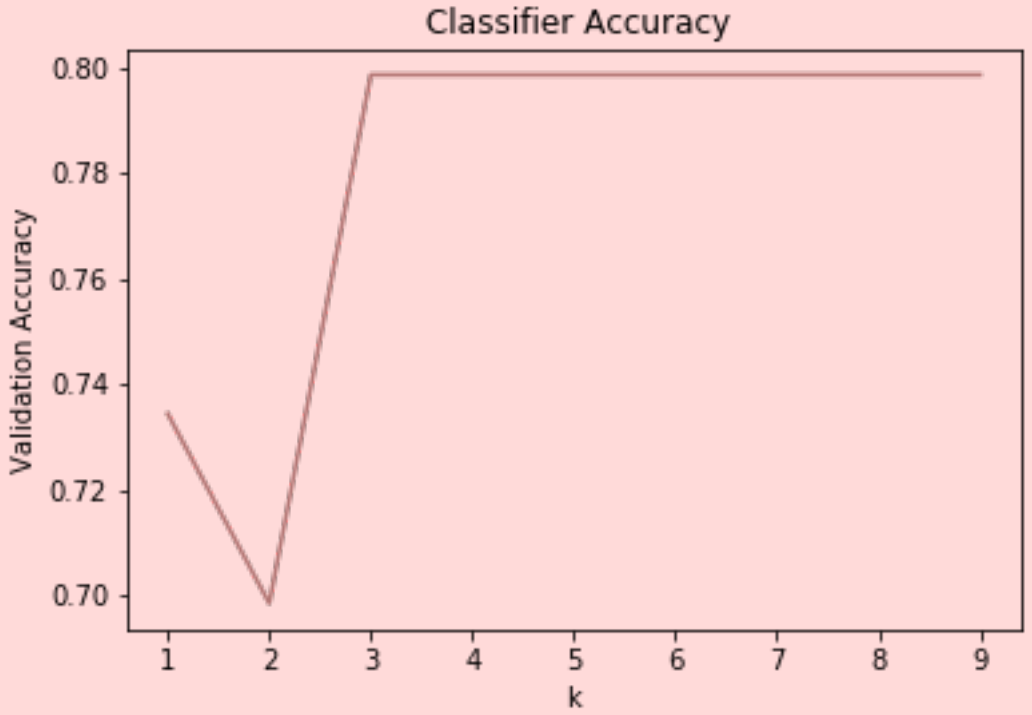
Native Bayes

After performing Native Bayes algorithm I have ended up with lovely accuracy score of 80%, a recall score of 1, a precision score of 80% and a f1 score of 0.89. The bad news came with the confusion matrix, were it was pretty clear that the model only had predicted “Yes” ☹.

KNeighbors

After applying Kneighbors I got the exact same figures than in Native Bayes, as per scores an also in the confusion matrix.

According to the graph, 3 is the best K.



Regression Approaches _

Can I predict the income knowing job, offspring and whether drinks, smokes & takes drugs?

K-Nearest Regression

With K-Nearest Regression I get an score of 0.81 when I take in to account the variables sex, job, and the following binary ones: smokes, drugs, drinks and offspring.

If I add the variable age the score goes down to 0.33.

Multiple Linear Regression

I have performed Multiple Linear Regression on the same variables, and what it seems interesting to me is that in this case scores better if I include “age” in the analysis. However, the score it’s pretty bad 0.012 for train score and 0.011 for test score.

I don’t trust any of the two models, I don’t think I have enough data to perform this analysis properly. The time to run for both models are similar, and If I had to choose one, I would choose Multiple Linear Regression as I really like the .coef_ feature, as gives a lot of info in what matters and what not.

Conclusions/Next steps_

I have really enjoyed a lot this practice and exploring the data set.

I don't feel confident with any insights obtained in the asked questions, however I have found interesting insights when I was playing with the data as the relation with income and age, showed in graph 2, that women are more likely to drink and also that the people who drinks have twice as many possibilities to take drugs that people who don't.

In the analysis I have tried to answer more questions that the exposed in this presentation, however I didn't find any interesting enough in order to add them.

I can conclude that in order to obtain a better and more accurate analysis I would need to spend more time digging in the data and, as I am getting insights, try to broaden the analysis further than the original questions. Those will be my next steps ☺.

Thank you