Date a Scientist

Hello!

I am Alfonso Camblor

This is my capstone project for Machine Learning Fundamentals.

Repo: github.com/camblor/CodecademyMachineLearningFundamentals

1

Analysis

Do we see some correlation in the data?

Questions

Income by Education

As a first approach, I would like to know if I can predict income based on the education level of the person, as it is said that higher education gives you more income.

For this type of prediction, as it's a number prediction, multiple linear regression and KNeighbors regressor would be fine.

Job by Education + Income

Also, it makes sense to predict the job of a person if you know its education level and income.

As it is a classification problem, I'm going to try to predict it by using Support Vector Machines and KNeighbors classifier.

Age by Income

Finally, it is said older people have higher salary (usually because higher experience) so I would like to try to predict age based on income.

For this one, as it's a number prediction, we are using multiple linear regression and KNeighbors regressor.

Explanation of e_level column

Income - Education

In this table, we have classified education in levels:

0: No education

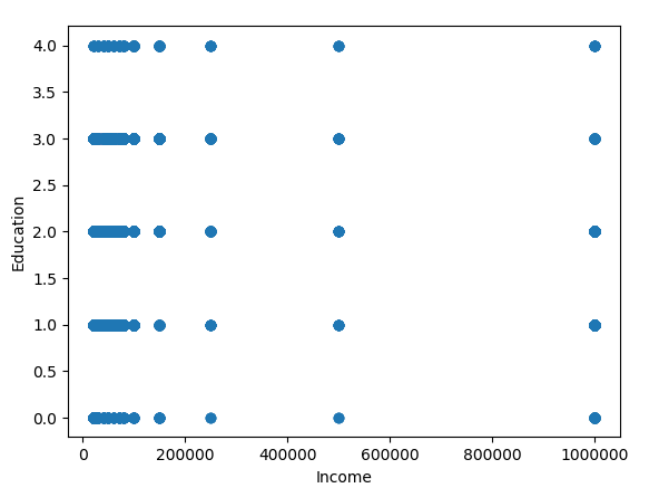
1: Education until bachelor.

2: Bachelor completed.

3: Master completed.

4: PhD and above.

This has been mapped and added to a new column called e_level.

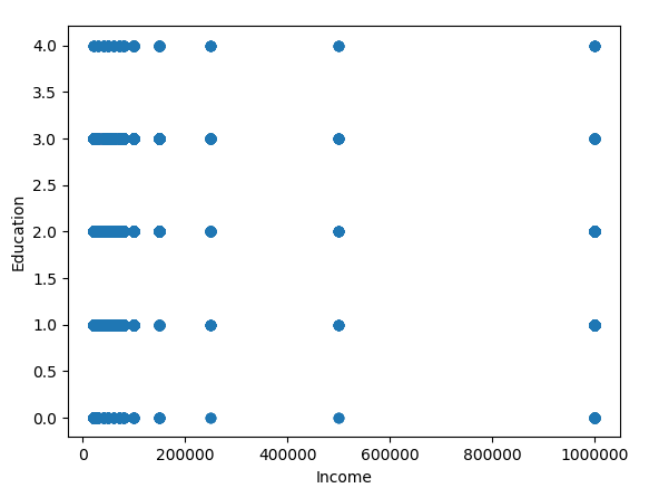


Data Exploration

Income - Education

As we can see, there is no correlation between education and income.

Anyways, we should plot this correlation with more data to confirm our conclusion. Remember that we have ignored 5/6 of the data to make this model.



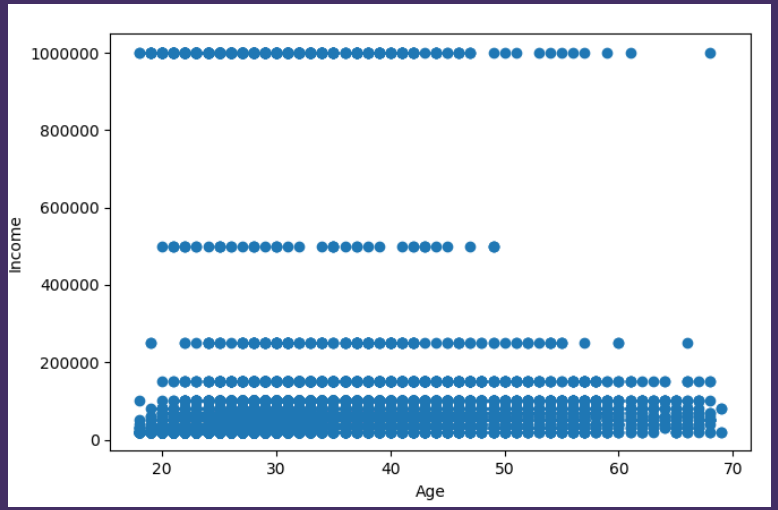
Data Exploration

Age - Income

As we can see, there are people of all ages with all kinds of income.

We now know age isn't that important in terms of income.

Also, we see most of the people under 200000$ income at all ages, and few people above that.



Explanation of Jobs column

Job - Education/Salary

Before analyzing, the labels we have used are:

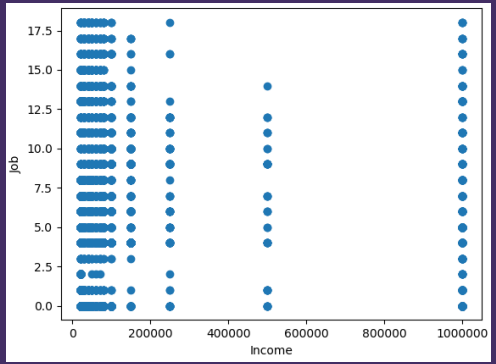
0: Unknown Fields

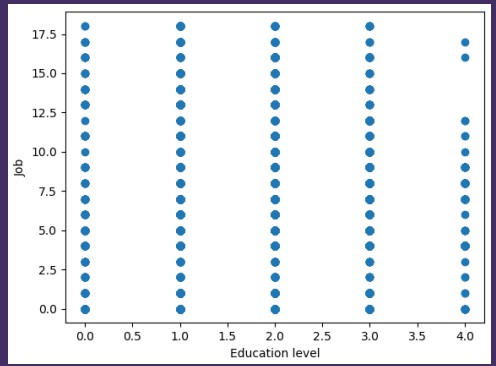
1: Student

2: Unemployed

3: Retired

Since label #4, every job field has its own label until label #18.





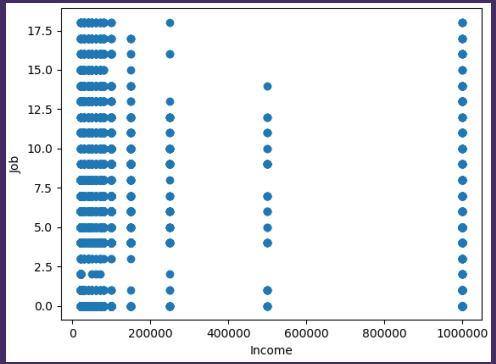
Data Exploration

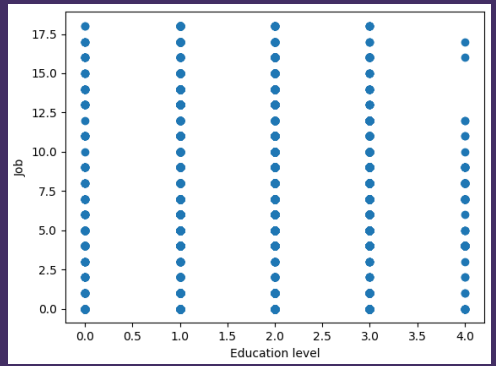
Job - Education/Salary

We might need to plot job with education and job with salary to see some correlation.

From separated data, we can see there are people winning a lot of money from every job and there are a people with every kind of studies working on every kind of jobs.

Job numbers are related with the field, one integer number for each.





2

Predictions

Can we get something new from our questions?

Can we predict Education with Income?

- Multiple Linear Regression score: -0.0003
- KNeighbors regressor score: -0.0269

After making the plot, we knew there wasn’t any correlation. That’s why we don’t get good scores. Anyways, we can see this as a proof to our first conclusion after visualizing the data.

Can we predict Income with Age?

- Multiple Linear Regression score: -0.00055
- KNeighbors regressor score: -0.21621

We saw in the plot it wasn’t any correlation. Now we have proved it with regression scores.

Can we predict job with Education and Income?

- Kneighbors Classifier score: 0.2511
- Kneighbors Classifier F1 score: 0.18
- Support Vector Machines score: 0.2718
- Support Vector Machines F1 score: 0.18

This scores mean that we aren't making nice predictions. We got 0.45 F1 Score in CS/engineering fields, which is our highest value. We have confirmed we can't predict Job by Education and Income.

Conclusions

Income by Education

After the data exploration and the regression, we now know there isn’t any correlation between this two features.

Removing all rows that had no income might have introduced some bias.

Job by Education + Salary

After the data exploration and the regression, we know that it’s difficult to predict the job knowing education and salary. There may be some variables that we haven’t think of.

Age by Income

After the data exploration and the regression, we can conclude with no correlation between age and income.

We already knew regression wouldn’t give us any answer since there was any visible correlation in the data.

Conclusions

- We can’t predict Income by Education
- We can’t predict Jobs by Education and Income
- We can’t predict Age by Income

All of our questions got negative answers. This means our models showed no prediction ability with poor correlation.

Also, we have seen that algorithms have score differences with exactly the same data.