Date-a-Scientist - Codecademy Pro Capstone

ZB

- at least two graphs containing exploration of the dataset
- a statement of your question (or questions!) and how you arrived there
- the explanation of at least two new columns you created and how you did it

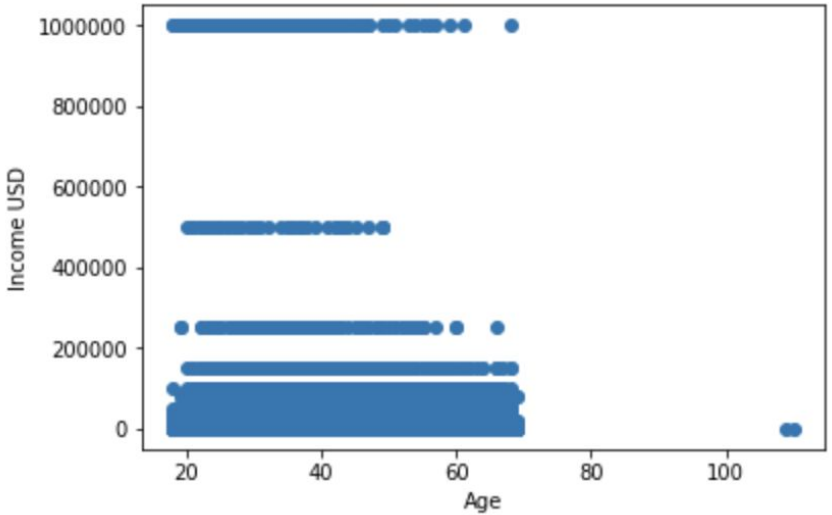
- the comparison between two classification approaches, including a qualitative discussion of simplicity, time to run the model, and accuracy, precision, and/or recall

- the comparison between two regression approaches, including a qualitative discussion of simplicity, time to run the model, and accuracy, precision, and/or recall

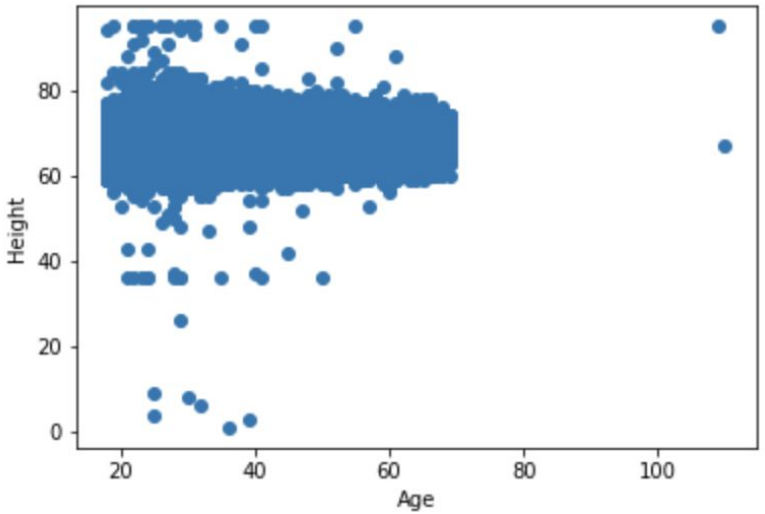
- an overall conclusion, with a preliminary answer to your initial question(s), next steps, and what other data you would like to have in order to better answer your question(s)

Table of Contents

Exploration of the Data Set



Scatter plot of Age against Specified Income



Scatter plot of Age against specified Height

Questions to possibly explore

Do heavy drug users and drinkers have better self-classified body types?

Is Age and Income linearly correlated?

Can I use Height and Age to predict average male and female heights in 20 years time?

Augmenting the dataset

Created new columns:

drug usage [0,1,2]

drinks usage: [0:11]

religion mapping: [0:44]

(see code)

Split the data 80:20 into training set and validation set randomly.

Classification Techniques. (1) MultinomialNB

Naive Bayes Classifier

- Training set (drug_usage, drink_usage), training labels (body type)
  - classifier.fit(training_counts, training_labels)

Validation set.

- prediction = classifier.predict(drug_drink_usage)
- probability = classifier.predict_proba(validation_counts)

Qualifying statement. Accuracy is average 0.55 . Time to run the model is long >3s. Precision 0.6.

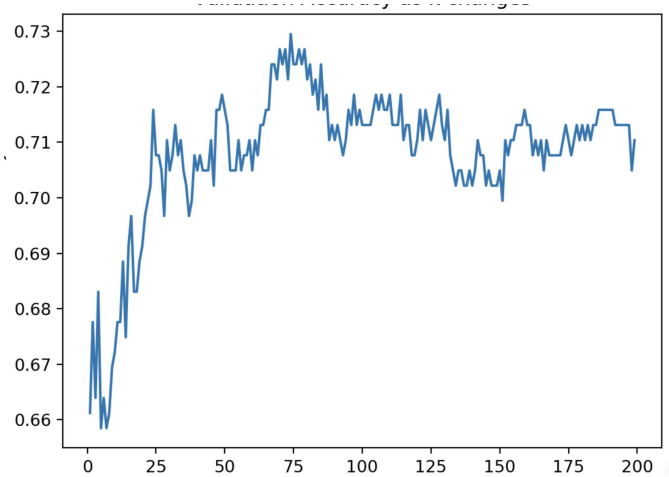
Classification Techniques. (2) KNeighborsClassifier

KNearest Neighbour Classifier:

- Training set (drug_usage, drink_usage), training labels (body type)
  - classifier = KNeighborsClassifier(n_neighbors = 10)
  - classifier.fit(habits, body_labels)

Validation set.

- prediction = classifier.predict(drink_drug_habits)



Qualifying statement. Accuracy is better 0.63 . Time to run the model is longer >10s. k is better at 75,

Regression Method Technique (1) MLR

Multiple Linear Regression

Features: Age. Income. Height.

Given 2 of these. Age and Income. Can you predict the height of member?

```
mlr = LinearRegression()
model=mlr.fit(age_train, income_train)
y_predict = mlr.predict(age_test)
```

Train score: 0.3725
Test score: 0.4051 Accuracy poor. Time to process Quick<1s. Precision low.

Regression Method Technique (2) KNeighborsRegressor

Scikit learn. KNeighborsRegressor.

Age. Income. Height.

regressor.fit(user_age_income, user_height)

regressor.predict(user_age_income_validation)

Train score: 0.4671
Test score: 0.5051 Accuracy better. Time to process slowers >3s. Precision average. Better than MLR.

Conclusions

Once outliers are removed. Age and Income is correlated.

But using age, income to predict height is inconclusive.

Using Drug Usage and Drinking Habits (once converted and normalized) can be used to predict body type.

Next Steps: I would want more data on weight, BMI to make the model more smoothed out.

Problem I have is also being able to take out default set answers by users that should be excluded.