Classification：

- Logistic regression

  accuracy: 0.59

  precision: 1.0

  recall: 0.59

  f1: 0.7421383647798743

  conf_matrix:

  [[0 0 0 0 0 0]

   [0 0 0 0 0 0]

   [0 0 0 0 0 0]

   [3 3 16 59 6 13]

   [0 0 0 0 0 0]

   [0 0 0 0 0 0]]

  Category 4 was predicted as Category 4 59 times, which performed poorly out of the total 100 times.

  Logistic regression assumes a linear relationship between features and target variables. If the relationships in your data are non-linear, logistic regression may not be able to effectively capture and model these relationships.

- Support vector machine

accuracy: 0.79

precision: 1.0

recall: 0.79

f1: 0.8826815642458101

conf_matrix:

[[0 0 0 0]

 [4 79 16 1]

 [0 0 0 0]

 [0 0 0 0]]

Category 2 was predicted as Category 2 79 times, which performed not very bad of the total 100 times.

Support vector machines can effectively process high-dimensional data and find optimal classification boundaries

- Decision tree

  accuracy: 0.68

  precision: 1.0

  recall: 0.68

  f1: 0.8095238095238095

  conf_matrix:

[[0 0 0 0 0]

 [0 0 0 0 0]

 [1 19 68 8 4]

 [0 0 0 0 0]

 [0 0 0 0 0]]

Category 3 was predicted as Category 3 68 times, which performed not very bad of the total 100 times.

The decision tree performs well, probably because it can handle nonlinear relationships and different types of features, and has a better ability to capture complex relationships between data features.

- Multi-layer perceptron

  accuracy: 0.7

  precision: 1.0

  recall: 0.7

  f1: 0.8235294117647058

  conf_matrix:

  [[0 0 0 0]

   [15 70 12 3]

   [0 0 0 0]

   [0 0 0 0]]

Category 2 was predicted as Category 2 70 times, which performed not very bad of the total 100 times.

Multilayer perceptrons can capture complex nonlinear relationships in data through nonlinear activation functions and learn features through hierarchical structures to adapt to complex tasks.

- Random Forest

  accuracy: 0.87

  precision: 1.0

  recall: 0.87

  f1: 0.9304812834224598

  conf_matrix:

  [[0 0 0 0]

   [9 87 3 1]

   [0 0 0 0]

   [0 0 0 0]]

  Category 2 was predicted as Category 2 89 times, which performed good of the total 100 times.

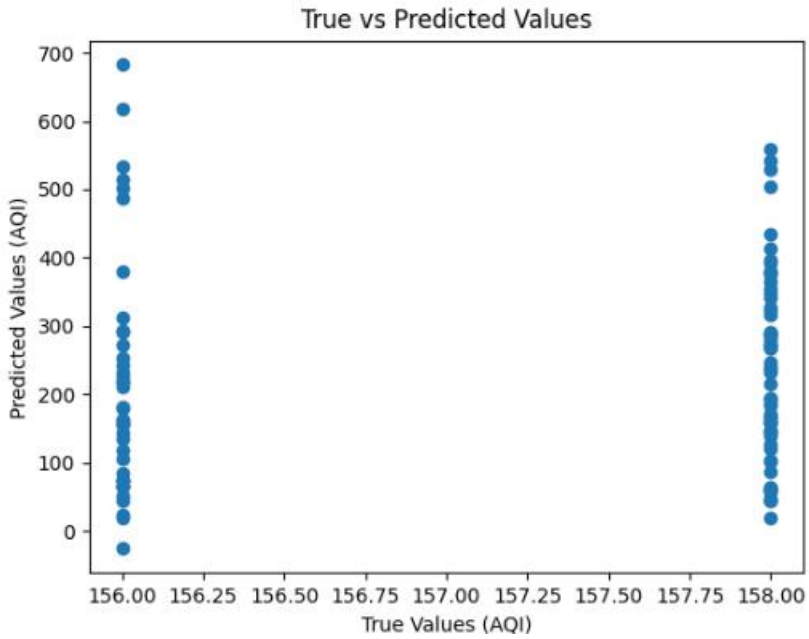
  By integrating the results of multiple decision trees, Random Forest can effectively reduce overfitting while capturing complex patterns and nonlinear relationships in the data.

Regression：

- Linear regression

  Best cross-validation MSE: 14392.945721832948

  Test set MSE: 27864.683167889798



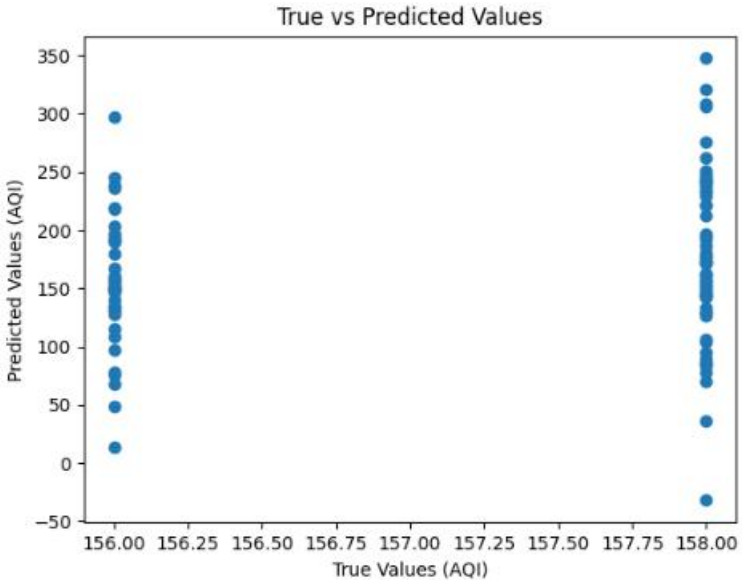
I think this scatterplot performs very poorly because of my data. His data has no diagonal shape at all.

I think one of the reasons for the poor performance is because of my data, and the second reason is that Linear regression is good at finding linear relationships between data. And my data doesn't have a strong linear relationship.

- Polynomial regression

  Best cross-validation MSE: 7900.74992141812

  Test set MSE: 4450.43253416713



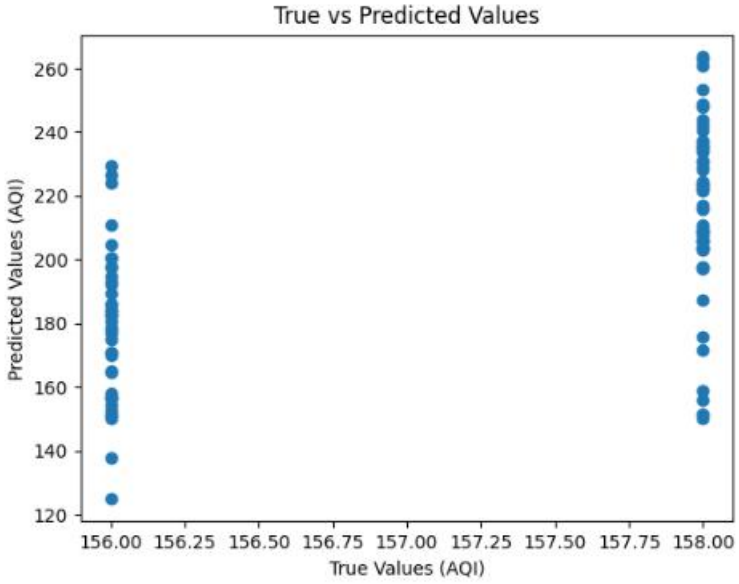
I think this scatterplot performs very poorly because of my data. His data has no diagonal shape at all.

Similarly, Polynomial regression is good at finding linear relationships between data. And my data doesn't have a strong linear relationship.

- Support vector machine

  Best cross-validation MSE: 9329.149707326504

  Test set MSE: 2808.0916957623967



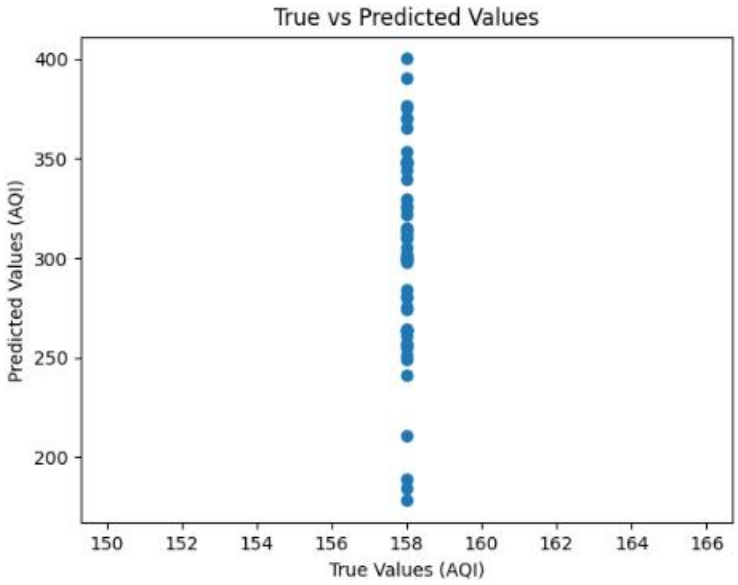
This performance may be slightly better because there are more points on the left concentrated around 158.

Support vector machines are good at finding boundaries between data. However, maybe regression is not a good way to handle my data

- Multi-layer perceptron

  Best cross-validation MSE: 6850.496492833417

  Test set MSE: 23054.558747068608



This scatter plot performs very poorly. I think this is due to the poor computing power of my computer, which prevents me from calculating too much data and image analysis. Therefore, the accuracy is very poor.

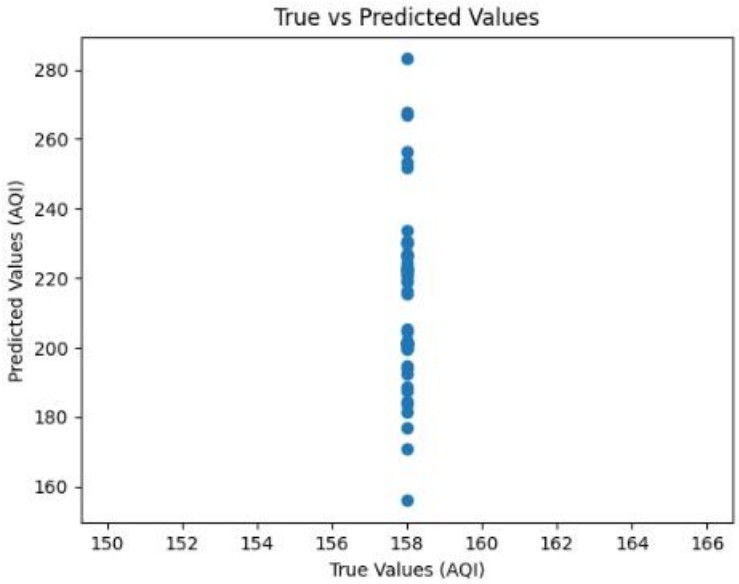
Actually, I think Multi-layer perceptron should have good results because Multi-layer perceptron can capture complex nonlinear relationships in data. However, due to my computer performance issues, the accuracy rate is very low.

- Random Forest

  Best cross-validation MSE: 8556.682082352942

  Test set MSE: 3794.4599520000006

  Test set R^2 score: 0.0



Again, the scatterplot representation is poor, with almost no outcome predicted correctly.

In Classification, Random Forest performs very well. I think this model is suitable because it combines multiple decision tree models. However, due to the performance of my computer, I keep shrinking the amount of data as well as the image size. Resulting in poor results.

Normal Question

- How long did this assignment take you?

  2 day

- Whom did you work with, and how?

Emily, we talk about how can we get good data set.

- Which resources did you use?

Chat gpt and data set from www.kaggle.com.

- A few sentences about:

First of all, I think it is difficult to find data. Since I've never tried training a model with images, I don't know what format the correct dataset should be. The dataset formats on www.kaggle.com are all kinds of weird, for example, many of them only have pictures, which makes me wonder if my idea of processing the data myself is wrong. Not only that, when calculating regression, it requires very high computer performance. The normal size of my picture is 224*224. However, even if I take the picture to 16*16, it will still cause the computer to crash. At the same time, the max iteration of some models must be adjusted. This resulted in me constantly compressing the image quality and sample count. The accuracy rate is very low. Even so, this assignment gave me a general understanding of all commonly used ML models. Even for many hyper parameters, I still need to further understand and master them. But I'm already familiar with the steps of calling the model. In addition, searching for data on the website also exercised my ability to be independent. I think in the future I can create the csv file myself and

analyze it. I think this assignment is quite difficult. Because for people who have never been exposed to ML, all steps need to be explored by themselves. I don't have any suggestions for changes, but maybe it would be better to provide a database that applies to all students?