# Programming assignment Week 4

## Question

Download O-A0038-003.xml. and convert then into classification and regression data set. Then train two simple machine learning model which are classification model and regression model.

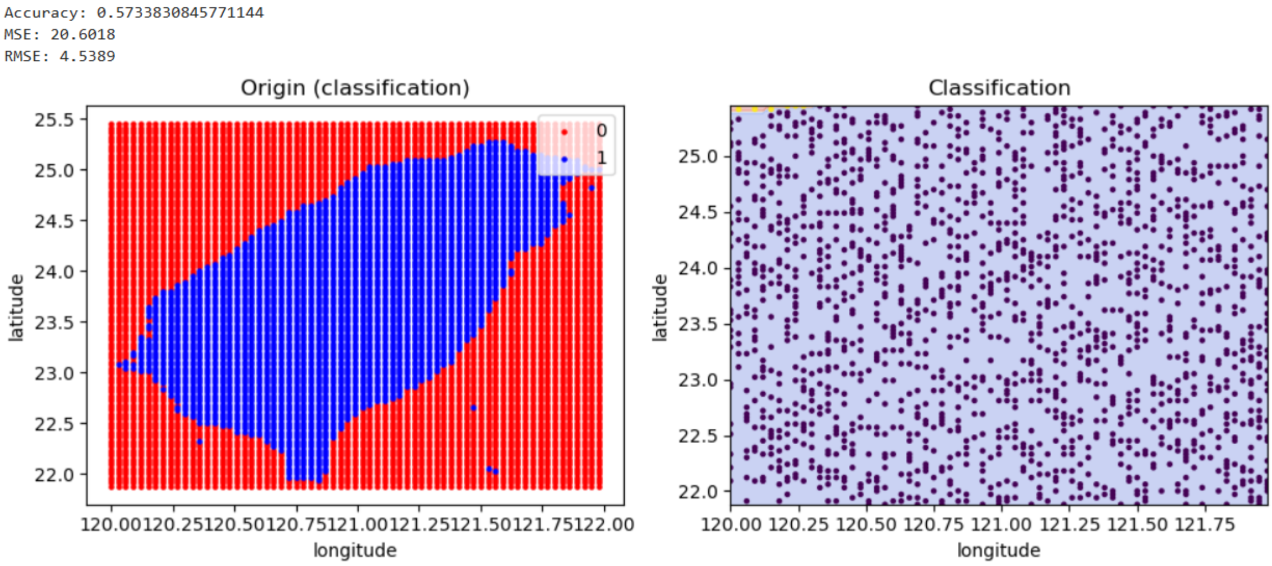
## About classification model

(This report used chatgpt for assistants to complete the code)
First, we convert data into classification data set called A. Since the output result is either 0 or 1. We decide to train it with logistic regression. That is, $z = b + w_1x_1 + w_2x_2$ and $p = \sigma(z)$. If $p > 0.5$, then output 1, otherwise output 0.

## About regression model

First, we convert data into regression data set called B. Then we train it with 3th-degree polynomial. That is, the function hypothesis is $h(x_1, x_2) = b + w_1x_1 + w_2x_1^2 + w_3x_1^3 + w_4x_2 + w_5x_2^2 + w_6x_2^3$. After that, we can predict the temperature at all area.

## Result



```
Accuracy: 0.5733830845771144
MSE: 20.6018
RMSE: 4.5389
```

Figure 1: Classification comparison

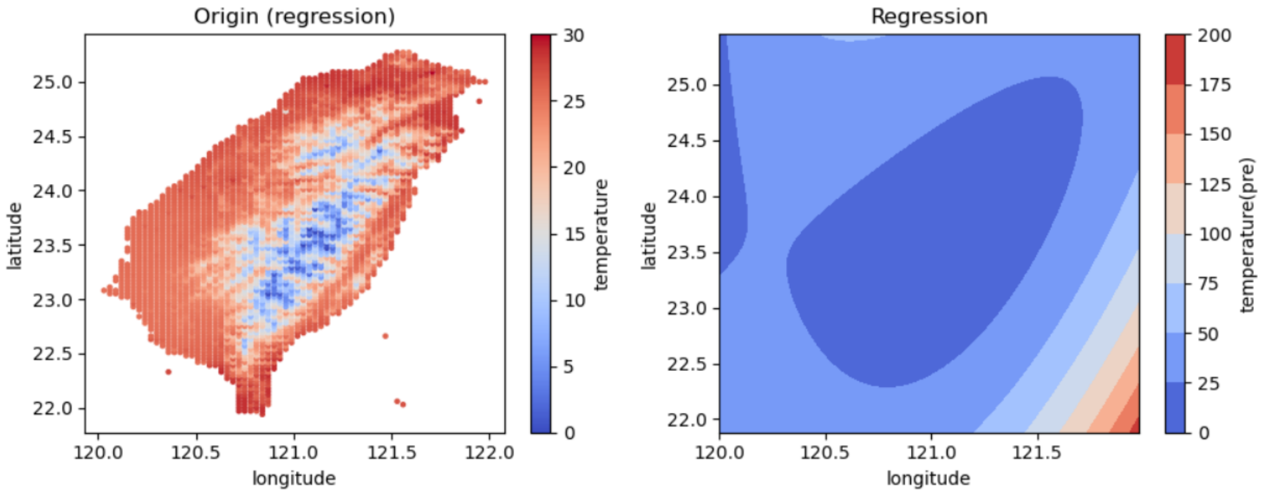


Figure 2: Classification comparison

## Analyze result

As above says, the accuracy is about 0.5734, which is extremely low. Therefore, it is not a good machine learning model. Thinking about the reason, I think it's because of the relation ship between real value's distribution function hypothesis. Most of valid value are gather in middle. However, logistic regression will separate data into two part with a line. As a result, no matter how it trains, it cannot still have a good classification. Observing the graph, we can notice that almost all points are predicted to 0, which proves it's not a proper hypothesis.

As for regression model, the mean squared error is 20.6018, RMSE is 4.5389. It means the average difference between each prediction and the true value is about 4.5389. It's also not a good machine learning model. Observing the graph, we can find the problem that the prediction of the range is high. Some of area even reach up to 200. So I think 3th-degree polynomial might not be a good hypothesis in this case since the temperature fluctuation is low in real value, but 3th-degree polynomial will increase rapidly outside valid data area.