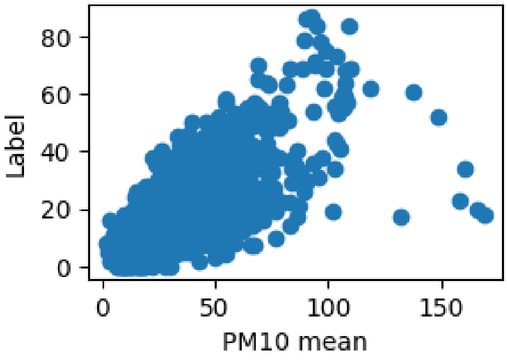
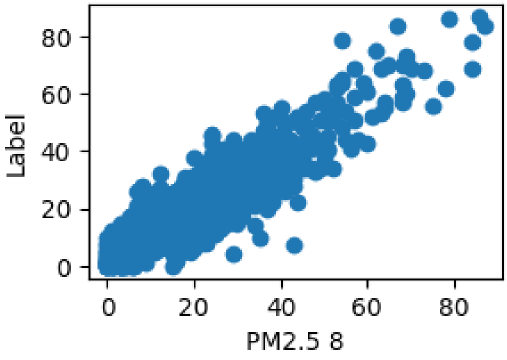
Data Mining Project1 Report

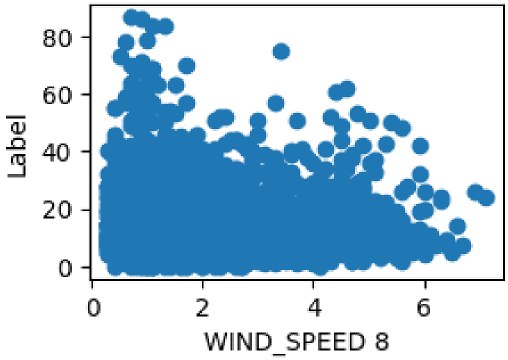
Student ID: 110550014 Name: 吳權祐

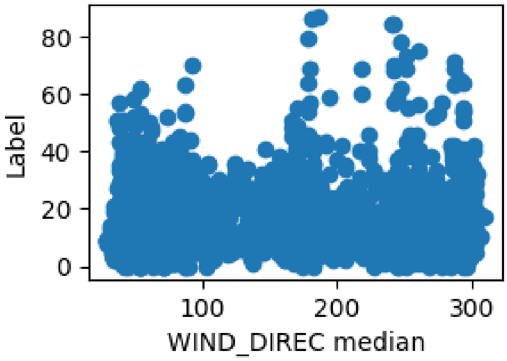
1. **How do you select features for your model input, and what preprocessing did you perform ?**

   I select features by examining their correlation with the label in training data from the visualization. These features in my model could be based on statistics such as *mean, std, max, min, median,* or specific time-based data from individual items. I performed preprocessing by handling empty data. If there was missing data, I found the two closest values on the left and right sides and then used linear interpolation to estimate the missing value.



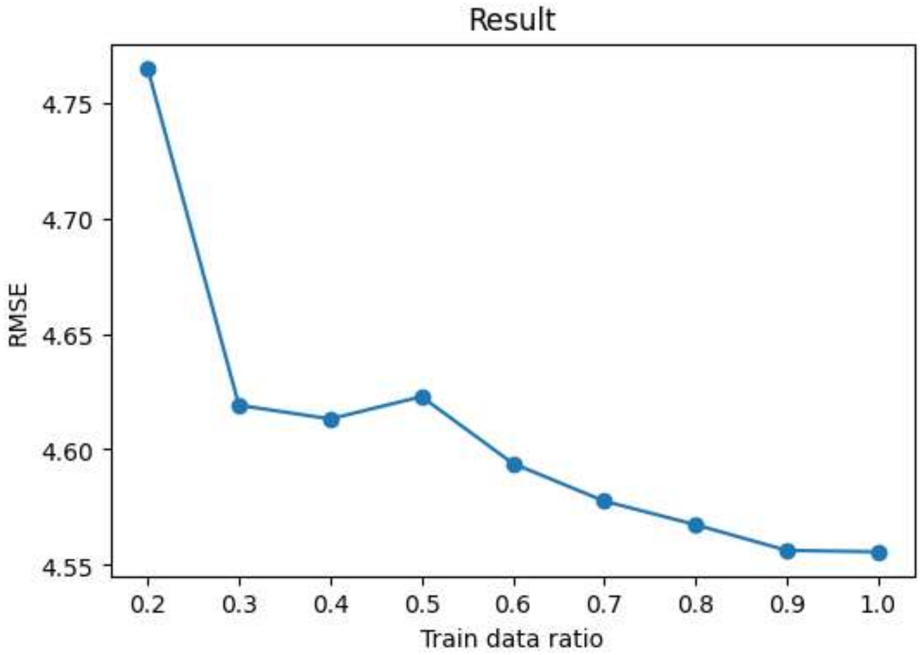




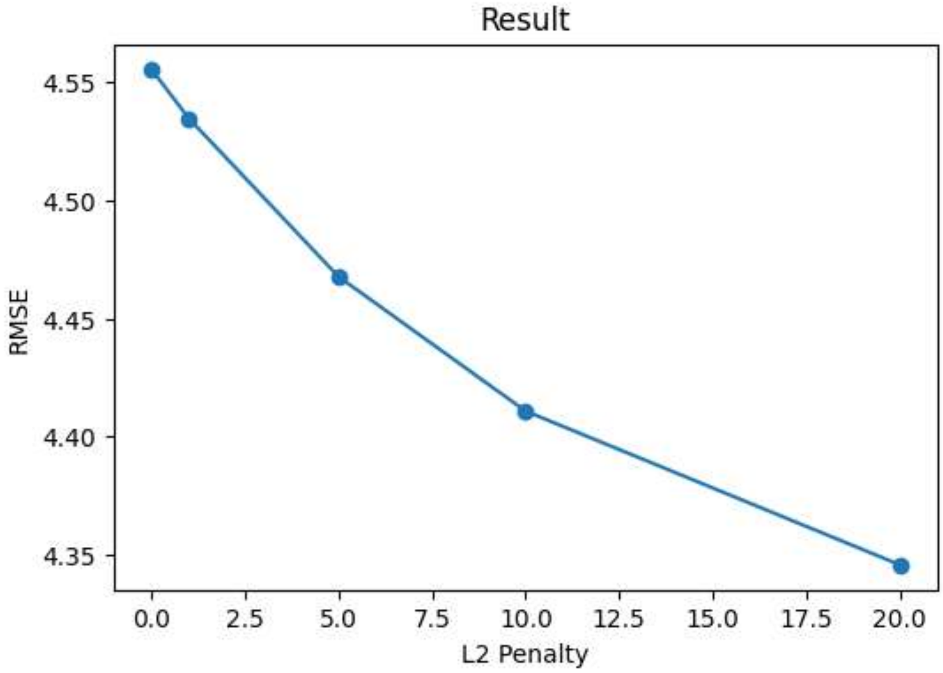


2. **Compare the impact of different amount of training data on the PM2.5 prediction accuracy. Visualize the results and explain them.**

   I initially divided the dataset into a training set (85%) and a validation set (15%) to test performance of my model. And it seems that the rate of improvement in model performance slows down when the training data size increases, but it has not yet started overfitting. I believe this is because the number of features is large enough to prevent the model from overfitting.



3. **Discuss the impact of regularization on PM2.5 prediction accuracy.**
   I performed ridge regularization in my gradient descent linear regression algorithm. And I observed that when the number of features increasing, it shows the importance of regularization, the accuracy of the prediction got improved. The following figure shows the effect of the L2 penalty with the size of model input being 41. It shows that when the L2 penalty increasing, the performance of model getting improved.


Result
RMSE
L2 Penalty