Data Mining HW1 Report

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

我將 18 種污染物整理成時間序列，並使用一個 18×9 的滑動視窗來提取特徵，標籤則是下一個小時的 PM2.5，並將所有特徵標準化，missing value 補中位數。
在檢查所有特徵的相關係數後，我發現 **RAINFALL**、**RH**、**WIND_DIREC**、**WD_HR** 的相關係數絕對值都小於 0.1。
因此，我測試了三種不同的特徵集：

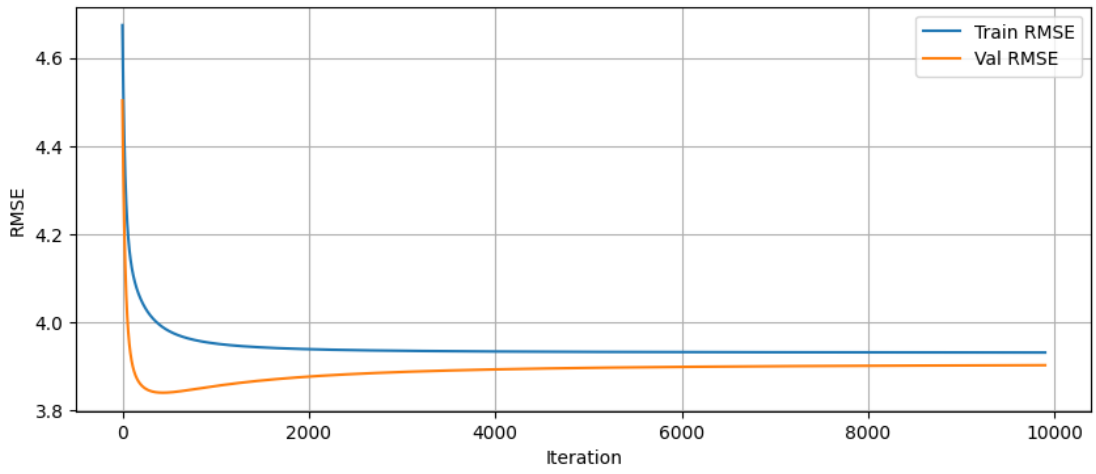
- **t1**：捨棄 {RAINFALL, RH, WIND_DIREC, WD_HR}
- **t2**：捨棄 {WIND_DIREC, WD_HR}
- **t3**：保留所有特徵（不捨棄）

結果的 RMSE 顯示，**t1 的表現最佳**，其次是 **t3**，而 **t2 的表現最差**。

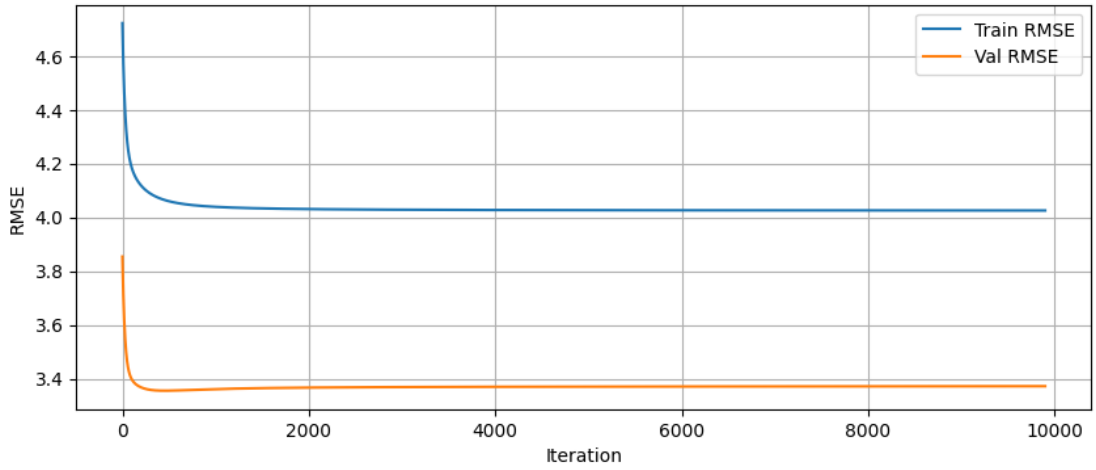
	Score
t1	3.71431
t3	3.72177
t2	3.73293

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

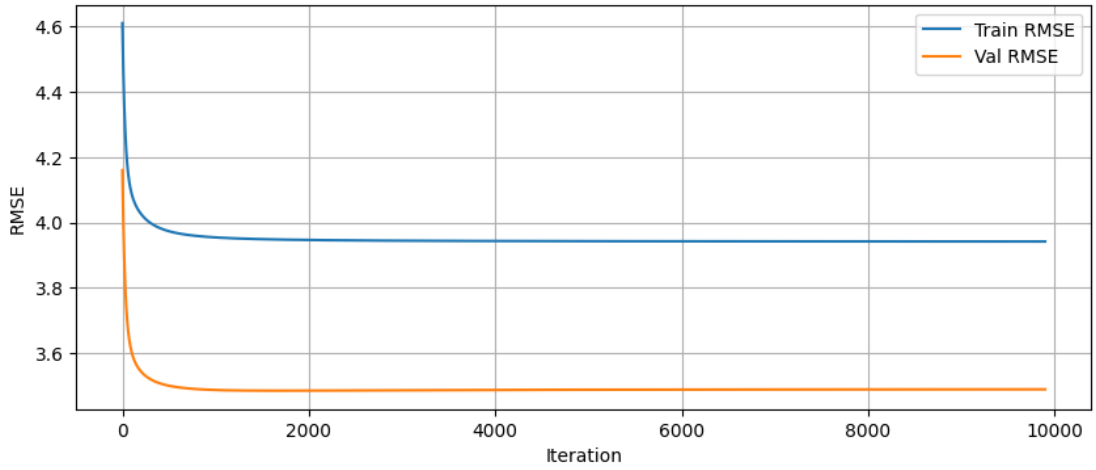
Train ratio = 0.7



Train ratio = 0.8



Train ratio = 0.9



Train ratio	Train RMSE	Val RMSE	原因
0.7	較低、平穩	稍低於 train	訓練樣本少，泛化估計準確但模型學不深
0.8	較高、平穩	顯著下降	訓練資料增加，variance 減少，bias 維持
0.9	平穩但不再下降太多	平穩但不再下降太多	模型泛化不一定變好，只是 Val 集變少導致估計噪音大

Train ratio = 0.9 在 Leaderboard 上分數最高

3. Discuss the impact of regularization on PM2.5 prediction accuracy.

適度的 λ 值則能抑制權重過度膨脹，降低共線性與雜訊的影響，並在訓練

過程中收斂更穩定。在這次作業中 λ=0.01 時在驗證集上有些微改善，但是在 Learderboard 上結果卻比不使用正則化來的差