HW2 – Time Series Regression

本次作業將使用新竹地區2018年10~12月之空氣品質資料，進行時間序列分析&迴歸預測pm2.5值。 （繳交期限：11/01 23:55）

本資料集(如附檔)包含AMB_TEMP, CH4, CO, NMHC, NO, NO2, NOx, O3, PM10, PM2.5, RAINFALL, RH, SO2, THC, WD_HR, WIND_DIREC, WIND_SPEED, WS_HR共18種屬性(汙染物)之逐時資料。

第一行之"00"~"23"代表小時。

表示儀器檢核為無效值，* 表示程式檢核為無效值，x 表示人工檢核為無效值，NR 表示無降雨，空白 表示缺值。

(資料來源: https://taqm.epa.gov.tw/taqm/tw/YearlyDataDownload.aspx)

作業要求：

請使用10和11月資料當作訓練集，12月之資料當作測試集，

將前六小時的汙染物數據做為特徵，第七個小時的pm2.5數據為預測目標

使用兩種模型 Linear Regression 和 Random Forest Regression 建模並計算MAE

1. 資料前處理

a. 取出10.11.12月資料

b. 缺失值以及無效值以前後一小時平均值取代 (如果前一小時仍有空值，再取更前一小時)

c. NR表示無降雨，以0取代

d. 將資料切割成訓練集(10.11月)以及測試集(12月)

e. 製作時序資料: 將資料形式轉換為行(row)代表18種屬性，欄(column)代表逐時數據資料

**hint: 將訓練集每18行合併，轉換成維度為(18,61*24)的DataFrame(每個屬性都有61天*24小時共1464筆資料)

2. 時間序列

a. 取6小時為一單位切割，例如第一筆資料為第0~5小時的資料(X[0])，去預測第6小時的PM2.5值(Y[0])，下一筆資料為第1~6小時的資料(X[1])去預測第7 小時的PM2.5值(Y[1]) *hint: 切割後X的長度應為1464−6=1458

b. X請分別取

1. 只有PM2.5 (e.g. X[0]會有6個特徵，即第0~5小時的PM2.5數值)

2. 所有18種屬性 (e.g. X[0]會有18*6個特徵，即第0~5小時的所有18種屬性數值)

c. 使用兩種模型 Linear Regression 和 Random Forest Regression 建模

d. 用測試集資料計算MAE (會有4個結果，2種模型*2種X資料)