Final Project Proposal

Team member 17

Deadline:　6/17

☐ Title

排煙脫硫警示預測

☐ Methods

1. Preprocessing

Data preprocessing 是 ML 中不可或缺的步驟，因為 data 的品質會直接影響我們 model 的學習能力，如果餵食太多沒用的資料只會導致 performance 變差，因此 data preprocessing 特別重要。我們預計可能會採用的方法如下：

- **刪除 Timestamp 項：**目前尚不知時間序是否會影響 performance，但目前先計畫刪除 timestamp 項。
- **PCA 主成分分析：**對資料求共變異數矩陣再進行奇異值分解，速度快，線性降維，特徵數量過多使用 PCA 能會造成降維後的 underfitting。
- **P-value selection method：**當資料中存在高度 linear dependent 的 feature 時，我們則捨去其中的一項，藉此來降維以增加 Model 的 performance。
- **Standardization：**將每個 feature 的 scale 做標準化，令其平均值變 0、變異數變 1，使資料得以符合 Normal distribution 而不偏向某處。
- **One-hot encoding：**因為類別是以 1、2、3、4、5 表示，數字有大小之分，但是類別並不分大小，所以我們將 class 項做 one-hot encoding，讓每個類別未來皆可獲得相同的 weight。
- **Our methods：**我們可能會對資料做其他數學上的運算與微調，以求獲得更準確的分類結果。

2. Classifier

做完資料前處理後，我們就將資料送進我們所選定的 Classifier 做訓練，目前尚未決定會使用哪個分類器，甚至有可能套用多個分類器以增加 Model 的複雜度，但我們目前有以下幾個人選：

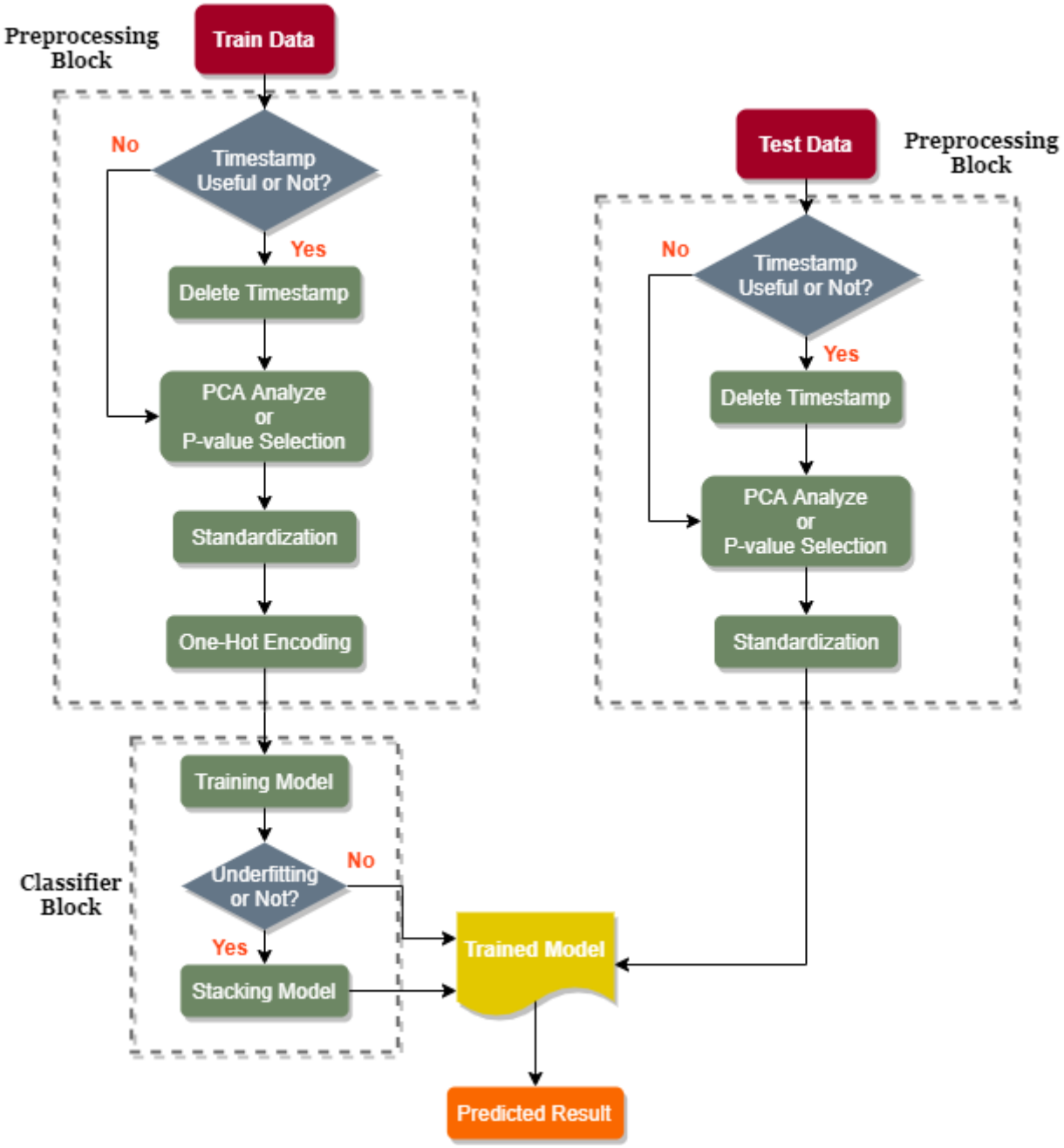
- **SVM 多類分類器：**
  - **直接法：**直接在目標函數上進行修改，將多個分類面的參數求解合併到一個最優化問題中。
  - **間接法：**主要是通過組合多個二分類器來實現多分類器的構造，常見的方法有兩種。
    (1) **一對多法（one-versus-rest）：**訓練時依序把某個類別的樣本歸為一類，其他剩餘的樣本歸為另一類。分類時將未知樣本分類為具有最大分類函數值的那類。
    (2) **一對一法（one-versus-one）：**在任意兩類樣本之間設計一個 SVM。當對一個未知樣本進行分類時，最後得票最多的類別即為該未知樣本的類別。

- **LogisticRegression 分類器：**根據現有資料對分類邊界建立迴歸公式，並以此分類。分類器中可調整不同 Penalty 正則化選擇參數(L1、L2)或優化算法選擇參數，常見分類器如下
  - **L1 Logistic**
  - **L2 Logistic (OvR)**
  - **L2 Logistic (Moltinomial)**

- **XGBoost：**是基於 Gradient Boosted Decision Tree (GBDT) 改良與延伸，被應用於解決監督式學習的問題，其特點主要有以下
  - **基於 Tree Ensemble 模型：**需要考慮多棵 Tree 的參數優化問題，但是我們卻無法一次訓練所有的 Tree，因此會透過 additive training 的方式，每一次保留原來的模型不變，並且加入一個新的函數至模型中，也就是說每一步皆會在前一步的基礎上增加一顆 Tree，以利修復上一顆樹的不足，有助於提升目標函數。
  - **XGBoost 有別於傳統的 GBDT：**選擇新增的樹、找到最好的樹以及減枝過程的方法差異。主要是透過貪心法，在樹的每層建構過程中優化目標函式的最大增益。

3. Flowchart

下圖為我們所規劃的 flowchart：



☐ References

1. Introduction to Data Preprocessing in Machine Learning
2. Feature selection — Correlation and P-value
3. 淺談降維方法中的 PCA 與 t-SNE
4. 機器學習:如何在多類別分類問題上使用二元分類器進行分類(Multiclass Strategy for Binary classifier)
5. XGBoost – A Scalable Tree Boosting System