**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。

1. **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：

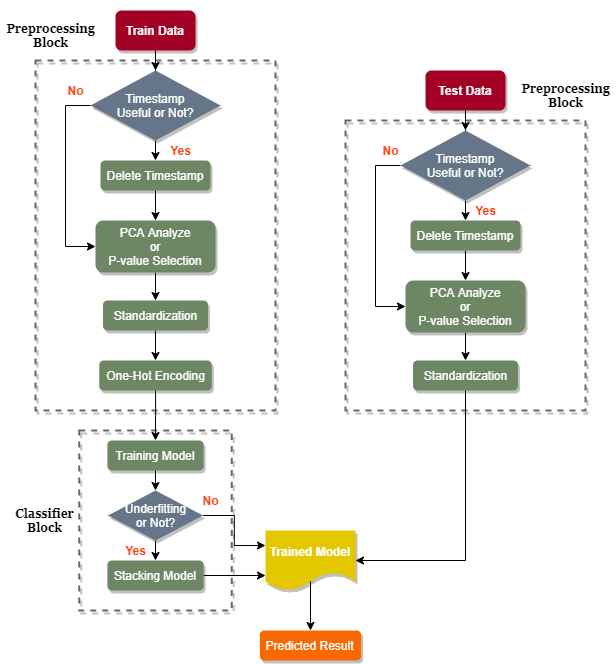
* **SVM多類分類器：**
* **直接法：**直接在目標函數上進行修改，將多個分類面的參數求解合併到一個最優化問題中。
* **間接法：**主要是通過組合多個二分類器來實現多分類器的構造，常見的方法有兩種。

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。

1. **Flowchart**

下圖為我們所規劃的flowchart：



* **References**

1. [Introduction to Data Preprocessing in Machine Learning](https://towardsdatascience.com/introduction-to-data-preprocessing-in-machine-learning-a9fa83a5dc9d)
2. [Feature selection — Correlation and P-value](https://towardsdatascience.com/feature-selection-correlation-and-p-value-da8921bfb3cf)
3. [淺談降維方法中的 PCA 與 t-SNE](https://medium.com/d-d-mag/%E6%B7%BA%E8%AB%87%E5%85%A9%E7%A8%AE%E9%99%8D%E7%B6%AD%E6%96%B9%E6%B3%95-pca-%E8%88%87-t-sne-d4254916925b)
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)](https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-%E5%A6%82%E4%BD%95%E5%9C%A8%E5%A4%9A%E9%A1%9E%E5%88%A5%E5%88%86%E9%A1%9E%E5%95%8F%E9%A1%8C%E4%B8%8A%E4%BD%BF%E7%94%A8%E7%94%A8%E4%BA%8C%E5%85%83%E5%88%86%E9%A1%9E%E5%99%A8%E9%80%B2%E8%A1%8C%E5%88%86%E9%A1%9E-multiclass-strategy-for-binary-classifier-b4e5017202ff)
5. [XGBoost – A Scalable Tree Boosting System](https://medium.com/@cyeninesky3/xgboost-a-scalable-tree-boosting-system-%E8%AB%96%E6%96%87%E7%AD%86%E8%A8%98%E8%88%87%E5%AF%A6%E4%BD%9C-2b3291e0d1fe)