Machine Learning Assignment 2 Report

106070038 杜葳葳

一、 Model

- Preprocessing
  - fillna(): 針對有缺損的項補值
  - drop(): 把日期時間從表中刪除
  - replace(): 將 sex 和 ed_diagnosis 改為用數字表示
  - isnull().sum(): 確認所有項皆有值
- Validation
  - 為了使 Precision 和 Recall 最大化，在本次作業中使用 **F1-score** 來衡量 Model 的好壞，F1-score = 2 * Precision * Recall / (Precision + Recall)
  - Recall(召回率) = TP/(TP+FN)、Precision(準確率) = TP/(TP+FP)
  - 使用 10 個 fold 的 **Cross-validation**
- Model
  - **Strategies**
    - 刪除兩兩相關係數高的其中之一個變因：適合 SVM
    - 刪除|和 output 的 correlation|<0.15 的變因：適合 KNN
  - **KNN**
    - F1-score: 0.2852007255288146

      Parameter: {'n_neighbors': 4, 'weights': 'distance'}
  - **SVM**
    - 手動給參數

kernel	linear	poly	rbf	**sigmoid**
C=1	x	0.0299	0	0.0888
C=10	x	0.044215	0	0.21228
C=100	x	0.0664	0.07255	0.2096

    - **SVM (使用 Grid Search)**

      F1-score: 0.2042396982658671

      Parameter: {'C': 8, 'kernel': 'sigmoid'}
  - **Decision Tree (使用 Grid Search)**
    - F1-score: 0.4226740876725451

      Parameter: {'criterion': 'gini', 'splitter': 'best'}

- **XGBoost (使用 Grid Search)**
  - F1-score: 0.4266581105396895

    Parameter: {'max_depth': 5, 'n_estimators': 18}
- **Random Forest (使用 Grid Search)**
  - F1-score: 0.3992468651891402

    Parameter: {'criterion': 'gini'}
- **C-SVM**
  - 執行 ANOVA 挑選主要特徵，並且使用 C-SVM 來計算特徵的權重與預測
  - F1-score=0.417910447761194

二、 The importance of the attributes

- 計算各個因子的 correlation
  - 高度正相關(correlation > 0.7)

    pmhx_htn（高血壓）與 pmhx_chf（心臟衰竭）有相關係數高達 0.81，呈現高度正相關，推論有其中一個疾病指標的病人，很高的機率也有另一個。其他高度相關的是一些實驗室檢測的因子，下表為兩兩相關係數大於 0.7 的因子：

attribute1	attribute2	corr_value
pmhx_htn	pmhx_chf	0.81
lab_alt	lab_ast	0.91
lab_mch	lab_mcv	0.81
lab_hct	lab_rbc	0.88
lab_hct	**lab_hemoglobin**	**0.93**
lab_leukocyte	lab_neutrophil	0.81
lab_rbc	lab_hemoglobin	0.86

  - 高度負相關(correlation < -0.7)

    lab_lymphocyte_percentage（淋巴性白血球）和 lab_neutrophil_percentage（嗜中性白血球）呈現高度負相關。

attribute1	attribute2	corr_value
lab_lymphocyte_percentage	lab_neutrophil_percentage	-0.9

- 計算所有因子與 output 的 correlation

  總體來看相關性皆蠻低的，然而可以注意到**相關程度最高**的依序為 **age**、**lab_urea**（血清尿素氮）、和 **lab_neutrophil_percentage**（中性

粒細胞的比例），負相關度最高的依序為 **vitals_spo2_ed_first**（血氧飽和度）、**lab_lymphocyte_percentage**（淋巴性白血球比例）、**lab_prothrombin_activity**（凝血酶原活性）

attribute	corr
age	0.34
lab_urea	0.32
lab_neutrophil_percentage	0.24
lab_neutrophil	0.24
lab_crp	0.23
lab_ldh	0.22
lab_creatinine	0.21
lab_leukocyte	0.2
lab_mcv	0.17
lab_glucose	0.16
pmhx_dementia	0.15
lab_sodium	0.15
lab_rdw	0.15
lab_ddimer	0.13
pmhx_stroke	0.11
pmhx_ckd	0.11
lab_inr	0.11
pmhx_ihd	0.09
pmhx_copd	0.09
pmhx_diabetes	0.08
pmhx_activecancer	0.08
lab_ast	0.08
lab_mean_platelet_volume	0.08
lab_potassium	0.07
pmhx_htn	0.07
sex	0.07
pmhx_chf	0.07
pmhx_hld	0.06

attribute	corr
vitals_spo2_ed_first	-0.29
lab_lymphocyte_percentage	-0.24
lab_prothrombin_activity	-0.11
lab_rbc	-0.09
lab_hemoglobin	-0.07
ed_diagnosis	-0.07
lab_platelet	-0.06
PATIENT ID	-0.06
vitals_dbp_ed_first	-0.05
lab_lymphocyte	-0.05
lab_hct	-0.03
vitals_sbp_ed_first	-0.02
pmhx_asthma	-0.01
vitals_temp_ed_first	-0.0

lab_mch	0.06
pmhx_chronicliver	0.05
lab_alt	0.02
lab_aptt	0.02
vitals_hr_ed_first	0.01

三、 How to use the model file

- 讀入 fixed_test.csv、model、test_output_example.csv，**執行.ipynb 的最後五個 cell (從 Run the testing data 開始執行)**，如下圖所示：

Run the testing data

Decide to use XGBoost model

```
import pandas as pd
from sklearn import svm
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.pipeline import make_pipeline
from sklearn.metrics import f1_score
from sklearn.model_selection import train_test_split
from xgboost.sklearn import XGBClassifier
import joblib
from sklearn.model_selection import GridSearchCV
from sklearn import metrics
from xgboost.sklearn import XGBClassifier
```

四、 Summary

- Preprocessing
  - **以統計的觀點刪除 covariate**：最初想讓資料符合 iid 的假設，於是根據相關係數(包含 covariate 兩兩之間的 corr、y 和各個 covariate 之間的 corr)，刪除相關係數高的變因或刪除和 y 相關係數低的變因，但效果不佳，用 Grid Search 嘗試五種 model 的 **F1-score 最高皆卡在 0.4 上下**。
  - **執行 ANOVA 挑選主要特徵**：經過實驗發現，提取 15-20 個 feature 的效果較好
- Tune Parameter
  - 使用 **Grid Search**、**Cross Validation (切 10 個 fold)**
- Model Selection
  - F1-score 高低：XGBClassifier ≈ C-SVM > Decision Tree > Random Forest > KNN > SVM
  - **最後選擇 XGBClassifier、F1 score=0.57 的 model，在跑 model 前有加 ANOVA filter，取出 17 個較重要的 feature**