Python Machine Learning Final Project

分析健康人與病患的腸道菌相

資料集取自 Gutmeta 資料庫 https://gutmeta.deepomics.org/

IMLP 377 林仙雅

臨床意義

腸道菌相是指存在於人體腸道內的微生物群體，它們在人體中發揮著重要的生理和代謝作用，包括維持腸道免疫系統、合成維生素和消化食物等。適當的腸道菌相分佈有助於保持人體健康，而不良的腸道菌相分佈則與許多健康問題有關聯，如肥胖、糖尿病和心血管疾病等。

由於造成炎症性腸病（IBD）的確切原因仍然不明確,這是一種慢性腸道疾病，包括克羅恩病和潰瘍性結腸炎，症狀包括腹痛、腹瀉、便血等。最新研究顯示，腸道菌群的失衡和菌群的分佈改變可能是炎症性腸病的重要原因之一。因此,區分 IBD 和健康人的菌相分佈是一個非常具有意義且有挑戰性的問題。

程式所用到的模組

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score, recall_score, f1_score, accuracy_score
from sklearn.svm import SVC
from sklearn.feature_selection import VarianceThreshold

import warnings
warnings.filterwarnings('ignore')
```

資料數據

	Sample name	disease	k__Archaea	k__Archaea|p__Euryarchaeota	k__Archaea|p__Euryarchaeota|c__Methanobacteria	k__Archaea|p__Euryarcl
0	CA_C10001IS2006FE_t1M15	Health	0.0	0.0	0.0	
1	CA_C10001IS2009FE_t2M15	Health	0.0	0.0	0.0	
2	CA_C10001IS2012FE_t3M15	Health	0.0	0.0	0.0	
3	CA_C10002IS2023FE_t2M15	Health	0.0	0.0	0.0	
4	CA_C10002IS2026FE_t3M15	Health	0.0	0.0	0.0	
...	...	...	...	...	...	
20892	T2D-098	Diabetes Mellitus, Type 2	0	0	0	
20893	T2D-099	Diabetes Mellitus, Type 2	4.940769	4.940769	4.940769	
20894	T2D-100	Diabetes Mellitus, Type 2	0	0	0	
20895	T2D-101	Diabetes Mellitus, Type 2	0	0	0	
20896	T2D-104	Diabetes Mellitus, Type 2	0	0	0	

18726 rows × 2561 columns

總共 20897 筆資料, 顯示 20897 位檢測者的腸道中, 各菌株的百分比
2559 個特徵值 (菌株)
區分成 96 種健康狀態
本報告僅擷取其中健康人群 (12992 位) 與 IBD (2048 位) 患者做比較

比較不同數據分類法的準確度

LASSO回歸 (失敗…迴歸線完全不在訓練集附近)
隨機森林分類法 (Accuracy of testing data: 0.927527)
SVM分類法 (Accuracy of testing data: 0.870124, 且訓練集有過度擬合的現象)
先用方差門檻篩選特徵值, 再用 SVM (Testing accuracy: 0.8464)

目前看來, 隨機森林分類法可能是較適合的, 但仍須與 Gutmeta 資料庫中其他數據集多比較, 看看訓練結果是否可泛化