肺癌疾病因素探討

第六組
廖紫涵
許佩琪

指導教授 蔡孟勳

Outline

01 資料集介紹

資料集介紹

- 資料集名稱：Cancer Patients Data
- 資料集來源：Kaggle
- 資料筆數：1000筆
- 欄位數量：25

欄位名稱	欄位解釋	資料格式	欄位補充
Patient Id	Patient ID	ordinal	
Age	Age of Patient	int	
Gender	Gender of Patient	ordinal	
Air Pollution	Air pollution that each patient is exposed to	ordinal	
Alcohol use	Alcohol use of Patient	ordinal	
Dust Allergy	Severness of Patient's dust allergy	ordinal	
Occupational Hazards	Patient's occupational hazards	ordinal	Risks associated with working in specific occupations
Genetic Risk	Genetic Risk of Patient	ordinal	
Chronic lung disorder	Chronic lung disorder of patient	ordinal	e.g. Asthma
Balanced Diet	Balanced diet of patient	ordinal	

欄位名稱	欄位解釋	資料格式	欄位補充
Obesity	Whether or not the patient is obese	ordinal	
Smoking	Patient's smoking habits	ordinal	
Passive Smoker	Patient's smoking habits cont'd	ordinal	Secondhand smoke
Chest Pain	Patient's chest pain	ordinal	
Coughing of Blood	If patient coughs blood	ordinal	
Fatigue	Patient's fatigue	ordinal	
Weight Loss	If there was a significant weight loss	ordinal	
Shortness of Breath	Patient's experiences of shortness of breath	ordinal	
Wheezing	Patient's wheezing	ordinal	
Swallowing Difficulty	Patient's swallowing difficulty	ordinal	

欄位名稱	欄位解釋	資料格式	欄位補充
Clubbing of Fingers	Patient's clubbing of fingers	ordinal	A symptom of disease of lungs which cause chronically low blood levels of oxygen
Frequent Cold	Patient's frequent cold	ordinal	
Dry Cough	Patient's dry cough	ordinal	
Snoring	Patient's snoring habits	ordinal	
Level	Patient's level of cancer	categories	

02 研究動機與目的

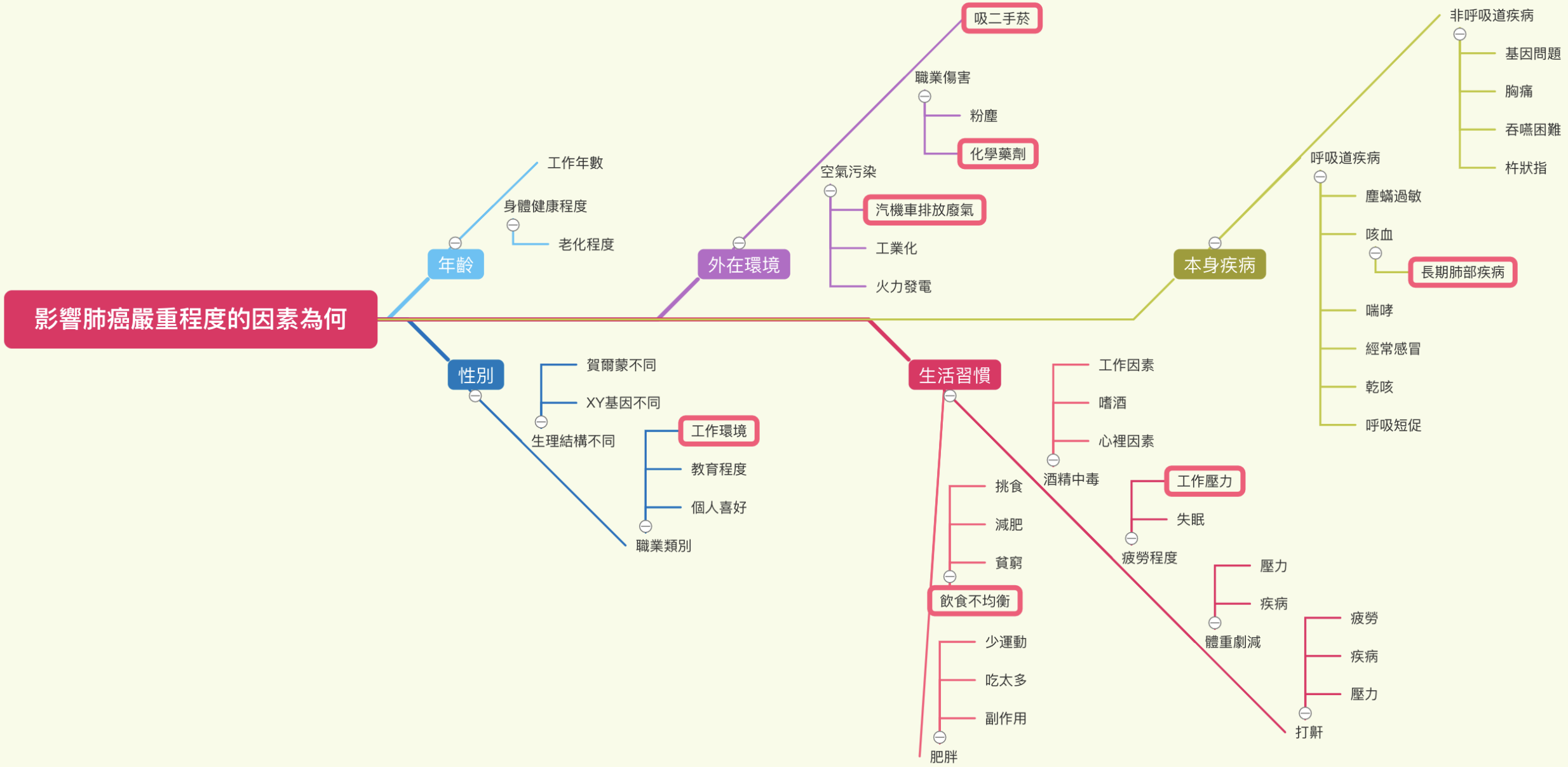
研究動機

排除今年肆虐全球的新冠肺炎，癌症是以往國人十大死因的第一位，其中又以「肺癌」死亡率最高，因此越來越多民眾開始思考如何預防癌症。我們所熟知的造成肺癌的因素有很多，例如吸菸或二手菸、長期吸入化學藥品或廚房油煙、飲食不均衡或家族遺傳因素等等...。為了建立民眾防癌觀念，台灣癌症基金會藉由媒體宣導健康飲食，並且加入國際抗癌聯盟，與美國癌症協會簽訂防癌共同宣言。許多醫師也提倡均衡飲食、規律運動、控制體重和定期篩檢，如果這五項皆做足了，推測可降低7成罹癌風險。

研究目的

既然造成肺癌的因素有那麼多種，其中影響最大的因素是什麼？大部分研究指出大約有85%肺癌患者罹癌原因是長期吸煙，而其他少數患者則是吸入二手菸或其他空氣髒污。但每個人的生活飲食習慣不同，加上近年來空氣污染越發嚴重，最直接影響罹癌的原因是否有所改變？因此，本研究目的在於找出影響肺癌嚴重程度的前幾項重要因素，以及不同年齡層影響原因是否有所不同，讓民眾能夠透過這幾項特徵，提高發現癌症的機率，提早治療。

魚骨圖


影響肺癌嚴重程度的因素為何
年齡
工作年數
身體健康程度
老化程度
性別
賀爾蒙不同
XY基因不同
生理結構不同
職業類別
工作環境
教育程度
個人喜好
外在環境
吸二手菸
職業傷害
粉塵
化學藥劑
空氣污染
汽機車排放廢氣
工業化
火力發電
生活習慣
工作因素
嗜酒
心裡因素
酒精中毒
工作壓力
失眠
疲勞程度
挑食
減肥
貧窮
飲食不均衡
少運動
吃太多
副作用
肥胖
壓力
疾病
體重劇減
疲勞
疾病
壓力
打鼾
本身疾病
非呼吸道疾病
基因問題
胸痛
吞嚥困難
杵狀指
呼吸道疾病
塵蟎過敏
咳血
長期肺部疾病
喘哮
經常感冒
乾咳
呼吸短促

03 文獻探討

文獻探討

Focus on lung cancer

John D. Minna,[1,3] Jack A. Roth,[2] and Adi F. Gazdar[1]

[1]Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas 75390
[2]Department of Thoracic and Cardiovascular Surgery, The University of Texas M.D. Anderson Cancer Center, Houston, Texas 77030
[3]Correspondence: john.minna@utsouthwestern.edu

Epidemiology and incidence statistics

Lung cancer is the most common form of cancer in the world (12.3% of all cancers), with an estimated 1.2 million new cases in 2000 (Parkin et al., 2001). Tobacco smoking is the most important cause of lung cancers with 80%–90% arising in cigarette smokers (Figure 1). There are major geographic, racial, and gender differences in incidence and some reports suggest that women may be at increased risk of lung cancer from exposure to tobacco smoke carcinogens. A lifetime smoker has a 20- to 30-fold increased risk of developing lung cancer compared to

specific carcinogens, especially the polycyclic aromatic hydrocarbons and the tobacco-specific nitrosamine 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK). Polymorphisms that reduce the activity of the glutathione-S transferase family, which inactivates the carcinogens, or that increase the activity of the P450 family, which activates them, may result in increased cancer susceptibility. The activated carcinogens bind to DNA, forming adducts, leading to mutations, especially G-to-T transversions, which may be repaired, lead to apoptosis, or persist. Molecular epidemiology has shown differences in smok-

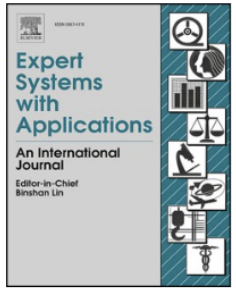
文獻探討

Expert Systems With Applications 164 (2021) 113981

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A k-NN method for lung cancer prognosis with the use of a genetic algorithm for feature selection

Negar Maleki [a], Yasser Zeinali [b], Seyed Taghi Akhavan Niaki [b,*,1]

[a] *Department of Industrial Engineering, Faculty of Engineering, University of Tehran, Iran*
[b] *Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran*

ARTICLE INFO

Keywords:
Lung cancer
Cancer staging diagnosis
Data mining

ABSTRACT

Lung cancer is one of the most common diseases for human beings everywhere throughout the world. Early identification of this disease is the main conceivable approach to enhance the possibility of patients' survival. In this paper, a k-Nearest-Neighbors technique, for which a genetic algorithm is applied for the efficient feature selection to reduce the dataset dimensions and enhance the classifier pace, is employed for diagnosing the stage

04 研究方法

資料統計

- 肺癌嚴重程度人數統計
- 年齡與肺癌嚴重程度人數統計
- 年齡與肺癌嚴重程度分配

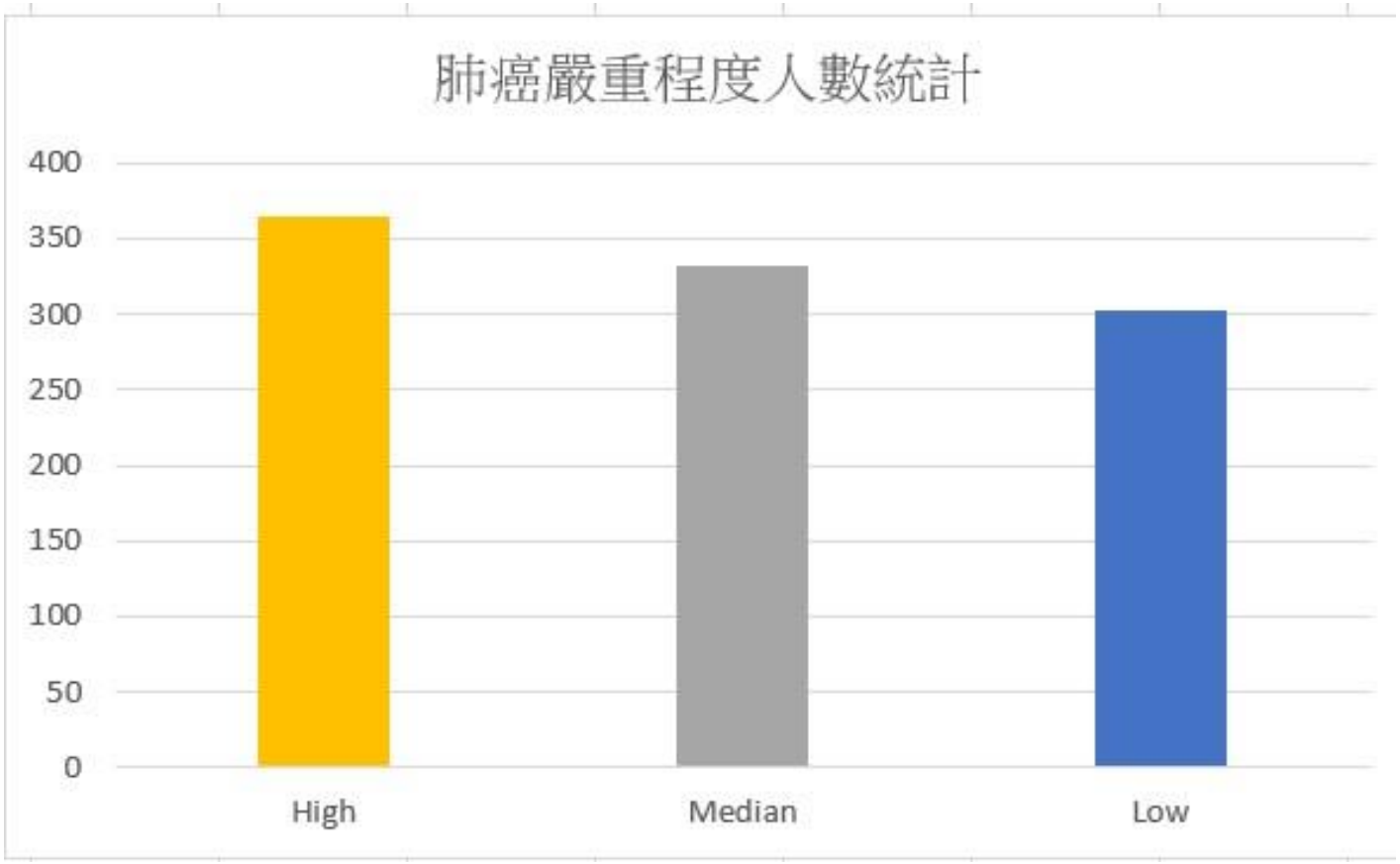
特徵選取

- ANOVA檢定、卡方檢定
- Heat Map
- 關聯式規則分析

模型預測

- Random Forest Classifier
- Naïve Bayes Classifier
- Support Vector Machine
- Logistic Regression

資料統計



High:365

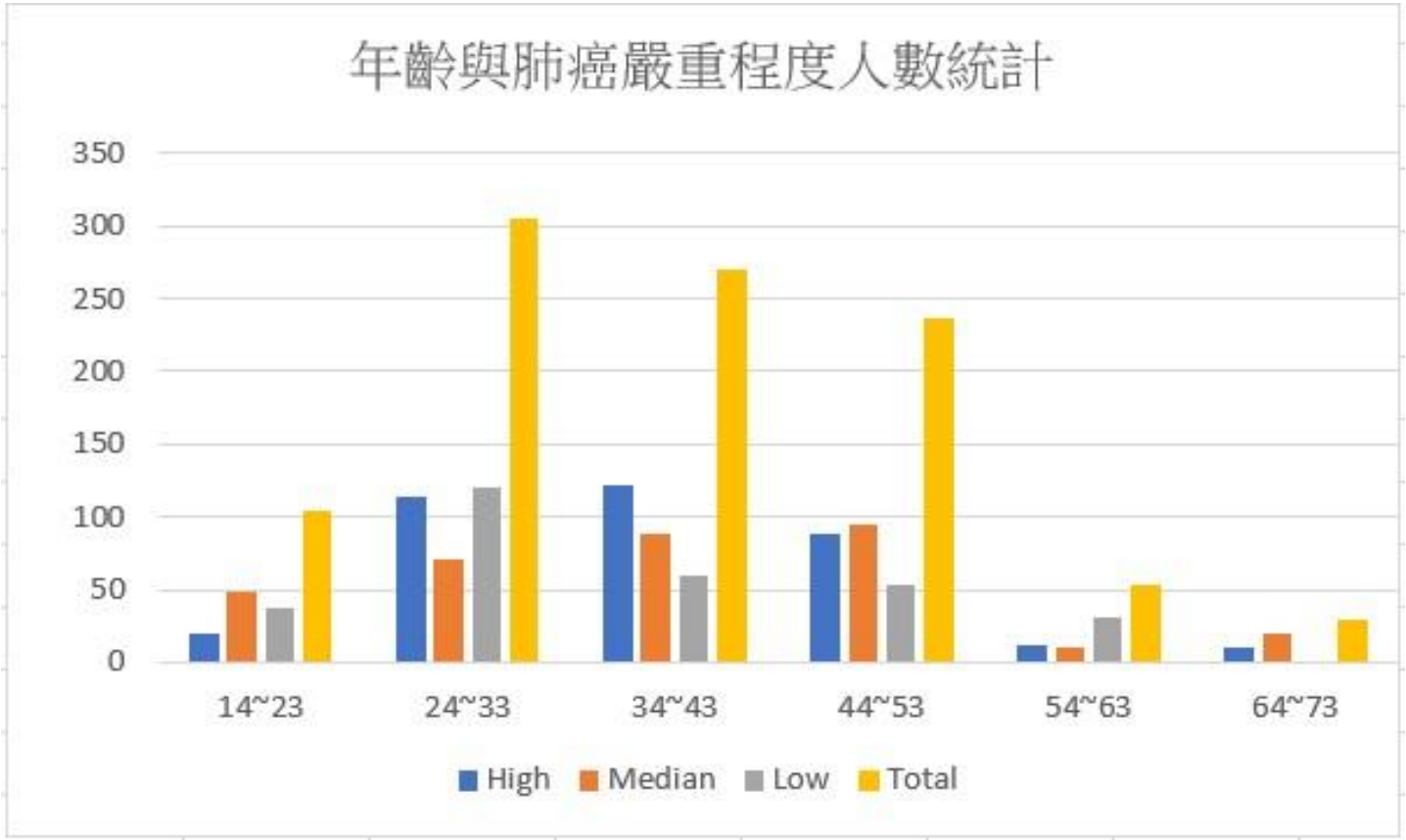
Medium:335

Low:303

Total:1000

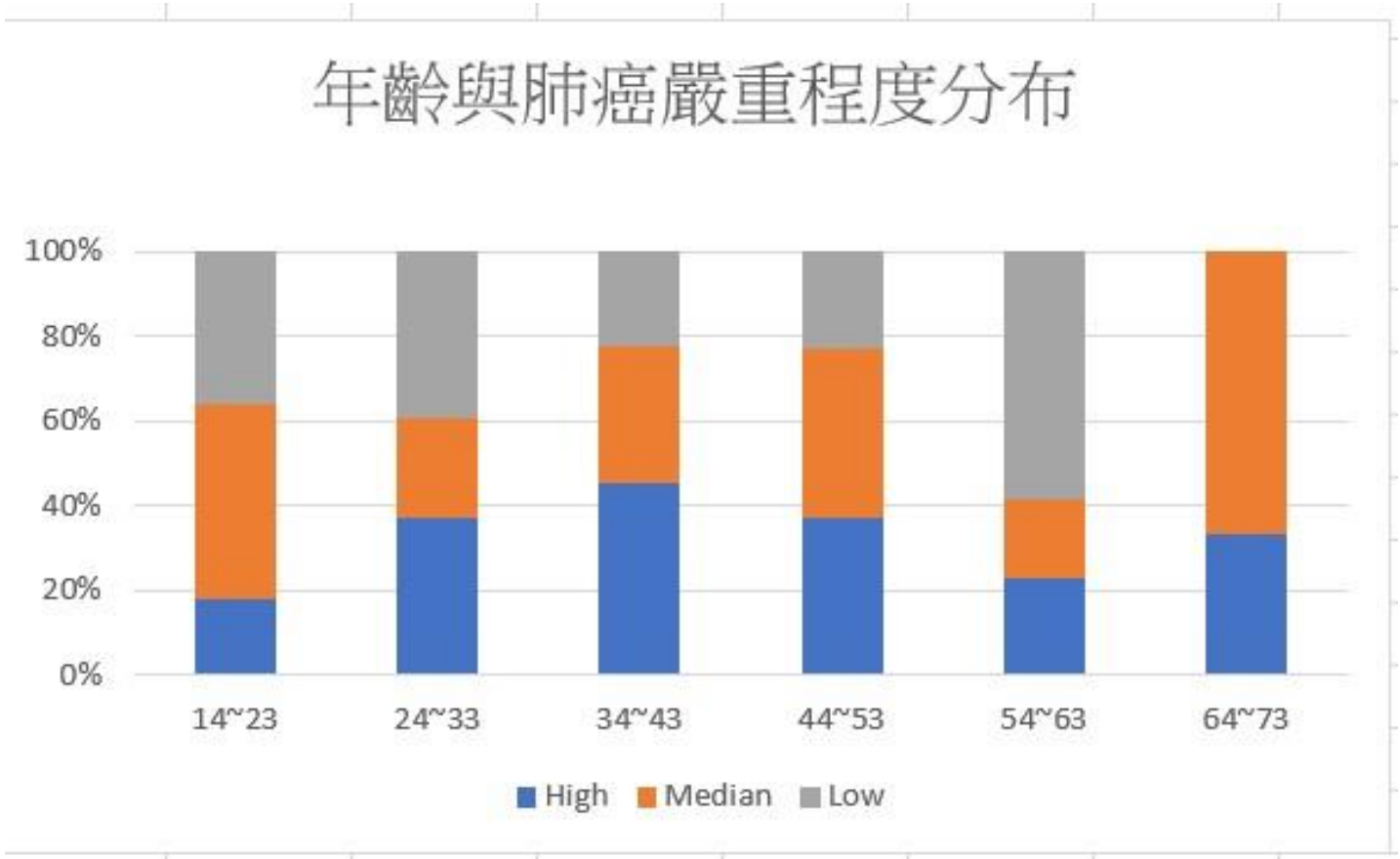
- 知識發現：資料分配平均

資料統計



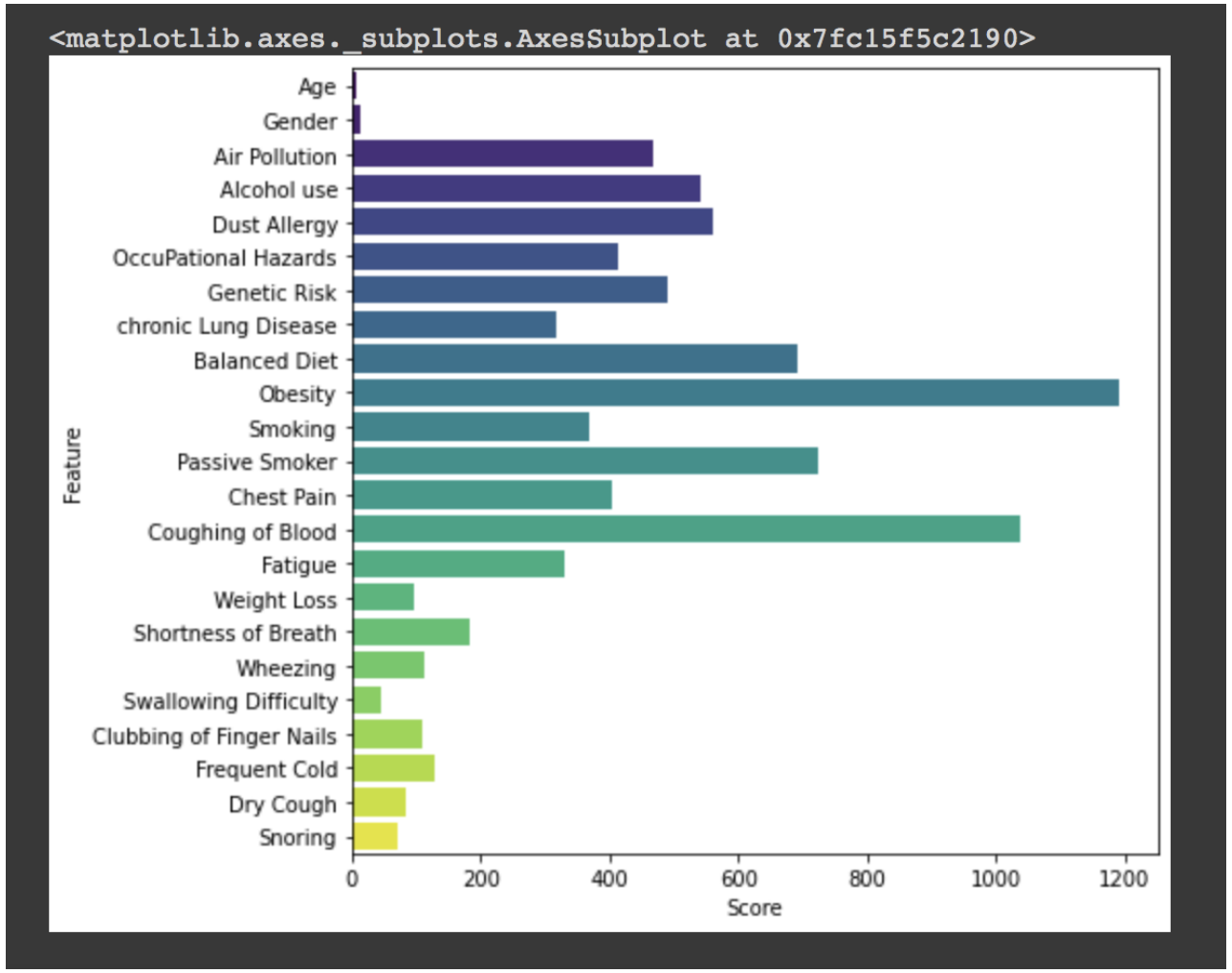
- 知識發現：資料集中得肺癌的患者大多集中在24-53歲。其中最多集中在24-33歲。

資料統計



- 知識發現：64-73歲族群當中肺癌嚴重程度都是中度或高度，可以推測肺癌嚴重程度可能與年齡有關

Feature Selection



- 利用特徵選取計算出各個特徵的重要性。
- 知識發現：Obesity、Coughing of Blood、Passive Smoker是前三個重要特徵。

Feature Selection

ANOVA

```
#ANOVA F檢定
skb = fs.SelectKBest(fs.f_classif,k=5)
skb.fit_transform(X, y)
print(skb.get_support())
```

卡方檢定

```
# 卡方檢定
chi = fs.SelectKBest(chi2,k=5)
chi.fit_transform(X, y)
print(chi.get_support())
```

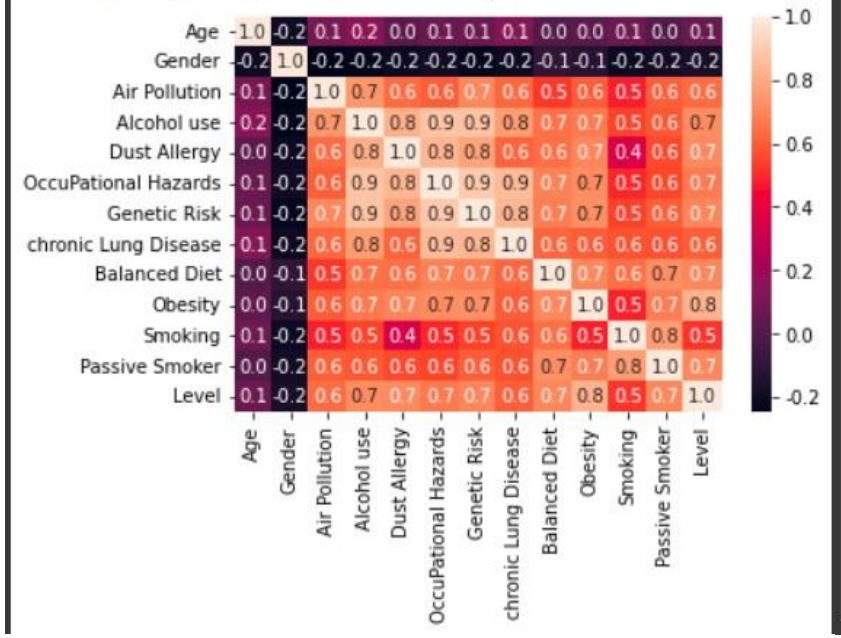
- 利用ANOVA檢定和卡方檢定選取特徵
- ANOVA選出最重要的特徵為Obesity
- 卡方檢定選出最重要特徵為Coughing of Blood
- 知識發現：不同檢定方式所選取出的特徵不同

Feature Selection

Heat Map

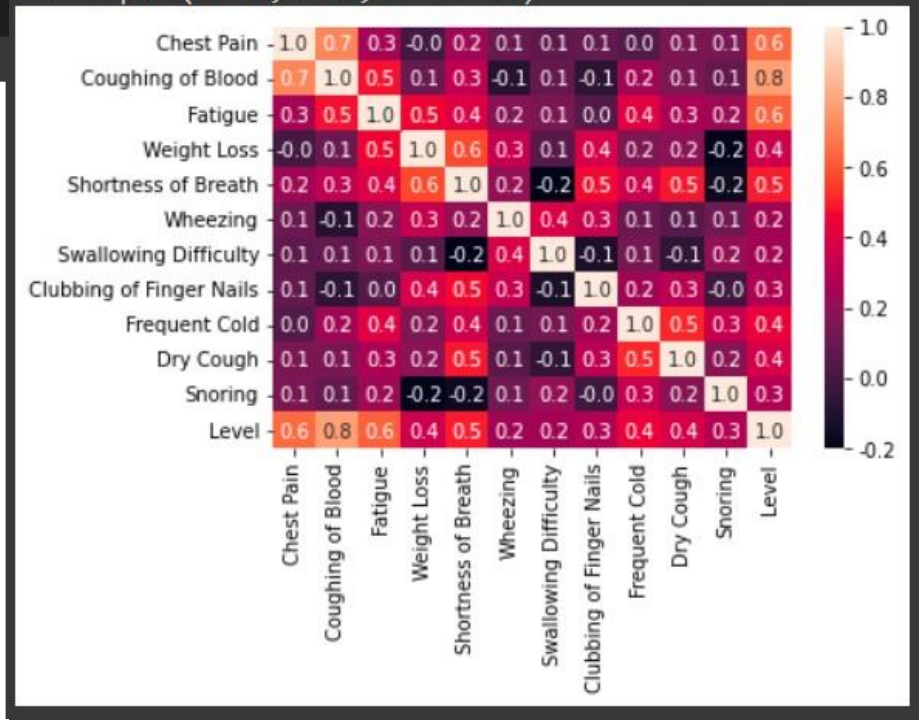
```
main1cor=main1.corr()

print(sns.heatmap(main1cor,annot=True,  fmt=".1f"))
```

```
AxesSubplot(0.125,0.125;0.62x0.755)
```




```
main2cor=main2.corr()
print(sns.heatmap(main2cor,annot=True,  fmt=".1f"))
```

```
AxesSubplot(0.125,0.125;0.62x0.755)
```



- 利用相關係數觀察各個特徵與結果的相關性。
- 知識發現：Obesity和Coughing of Blood分別與Level相關性高達0.8

關聯式規則分析

	support	itemsets	length
13	0.833333	(Coughing of Blood, Obesity, Balanced Diet)	3
14	0.833333	(Coughing of Blood, Passive Smoker, Balanced D...	3
15	0.833333	(Passive Smoker, Obesity, Balanced Diet)	3
16	0.833333	(Genetic Risk, Coughing of Blood, Passive Smoker)	3
17	0.833333	(Coughing of Blood, Obesity, Passive Smoker)	3

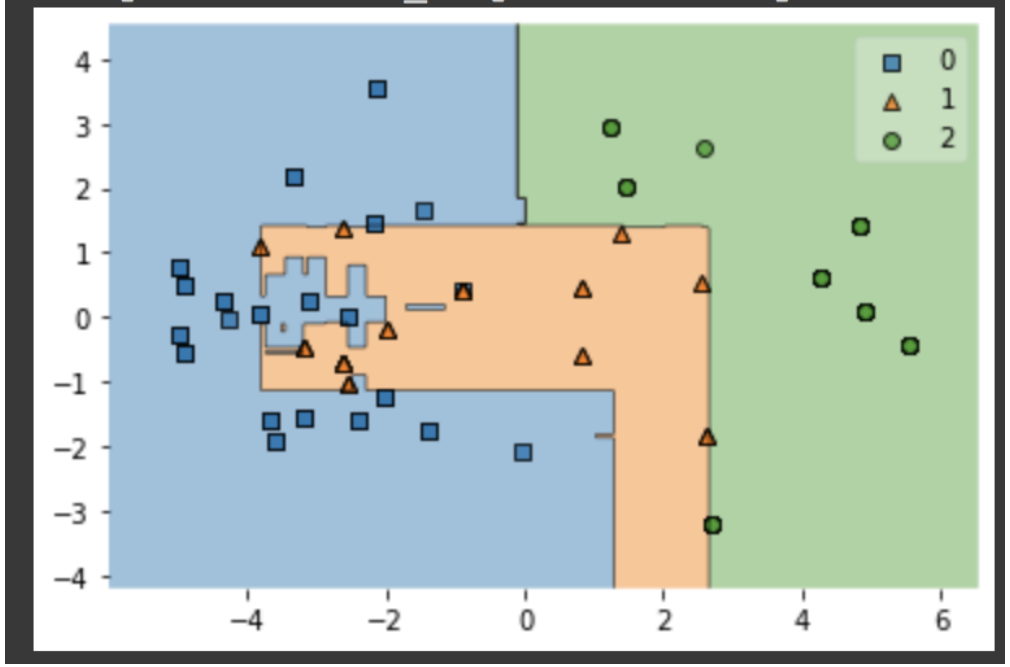
- 利用關連式規則分析找出各個特徵之間的關聯。
- 知識發現：可以看出以下這五組特徵組合的支持度高達0.8以上。

模型預測

Random Forest Classifier

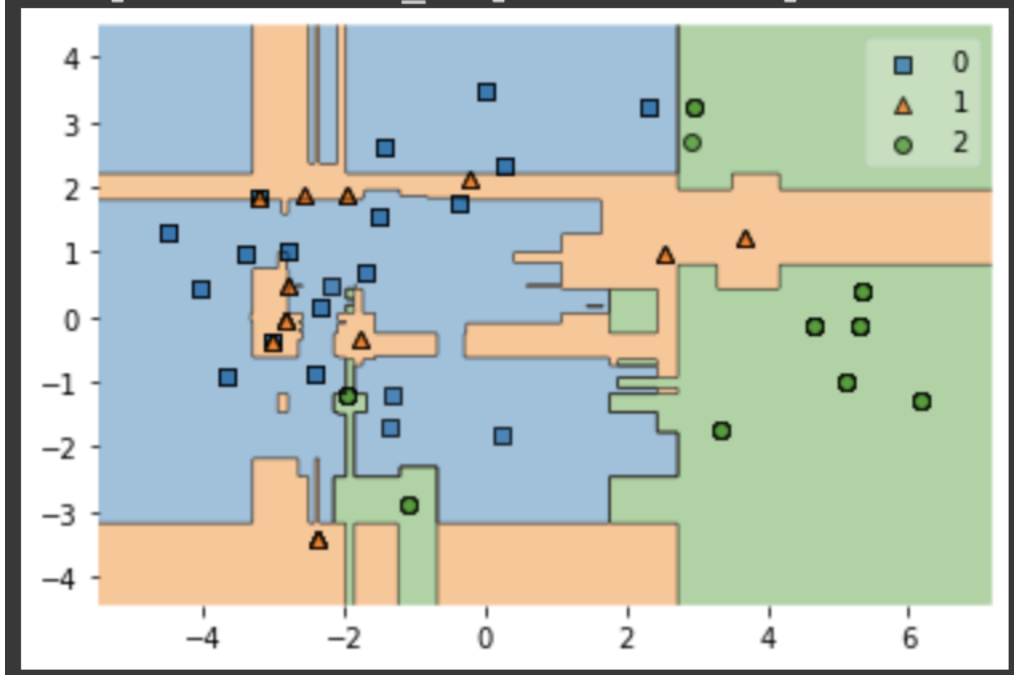
前三項重要特徵Obesity、
Coughing of Blood、Passive Smoker

```
Accuracy of train set ：0.9880
Accuracy of test set：0.9960
/usr/local/lib/python3.7/dist-packages/mlxter
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



原本設想的重要特徵Smoking、
Air Pollution、Passive Smoker

```
Accuracy of train set ：0.9667
Accuracy of test set：0.9760
/usr/local/lib/python3.7/dist-packages/mlxter
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



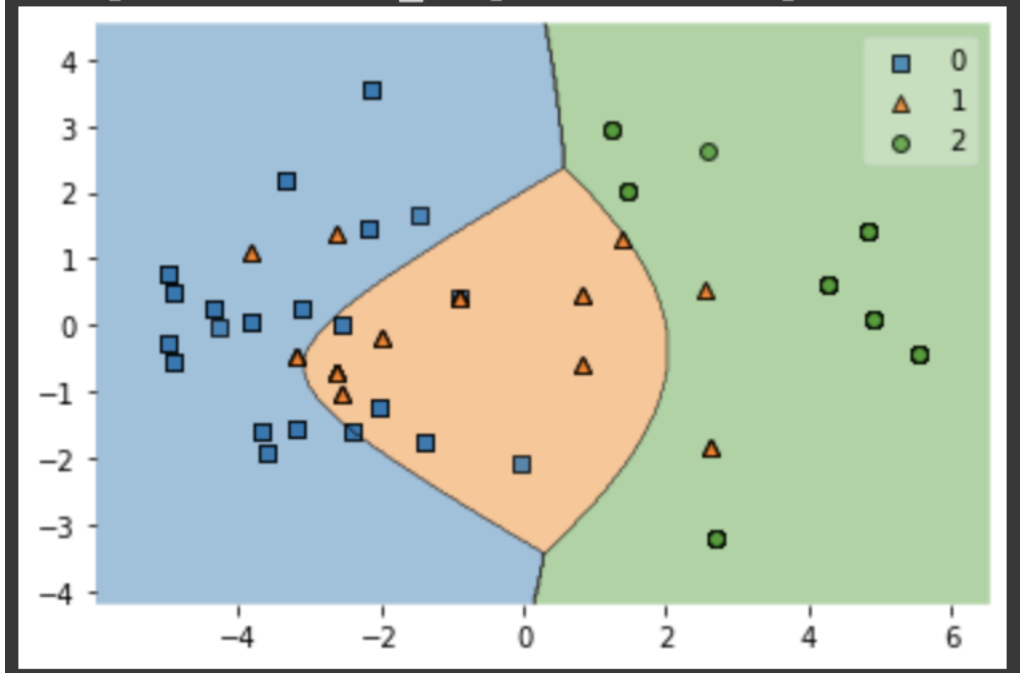
- 利用隨機森林預測。
- 知識發現：前三項特徵預測出的準確度高達99.6%

模型預測

Naïve Bayes Classifier

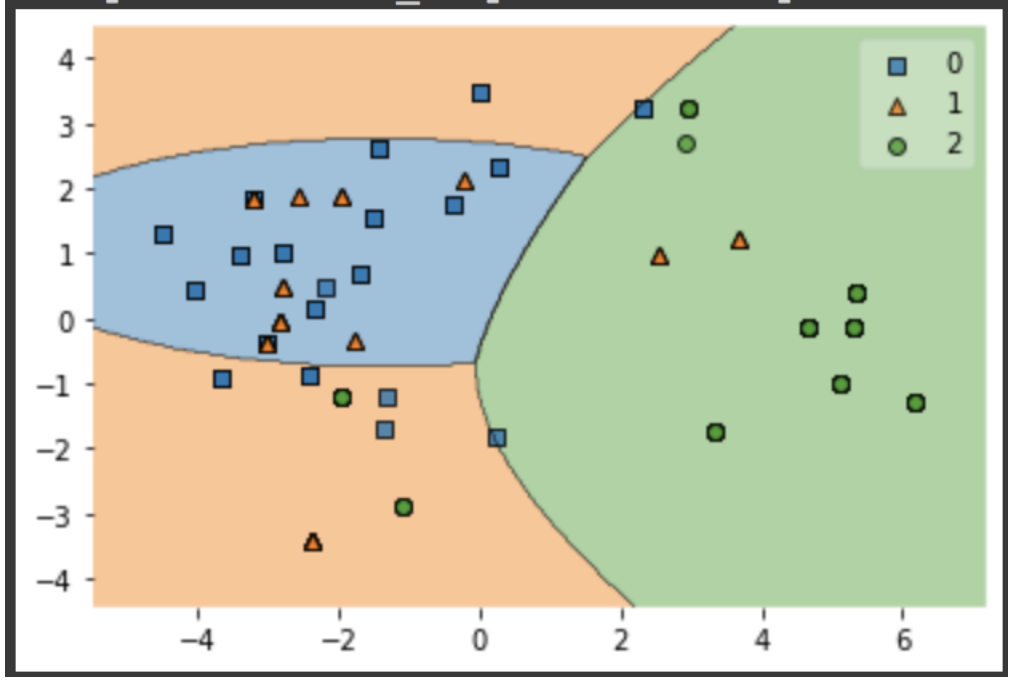
前三項重要特徵Obesity、
Coughing of Blood、Passive Smoker

```
Accuracy of train set  ：0.7933
Accuracy of test set：0.8520
/usr/local/lib/python3.7/dist-packages/mlxter
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x7
```



原本設想的重要特徵Smoking、
Air Pollution、Passive Smoker

```
Accuracy of train set  ：0.5547
Accuracy of test set：0.6040
/usr/local/lib/python3.7/dist-packages/mlxte
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



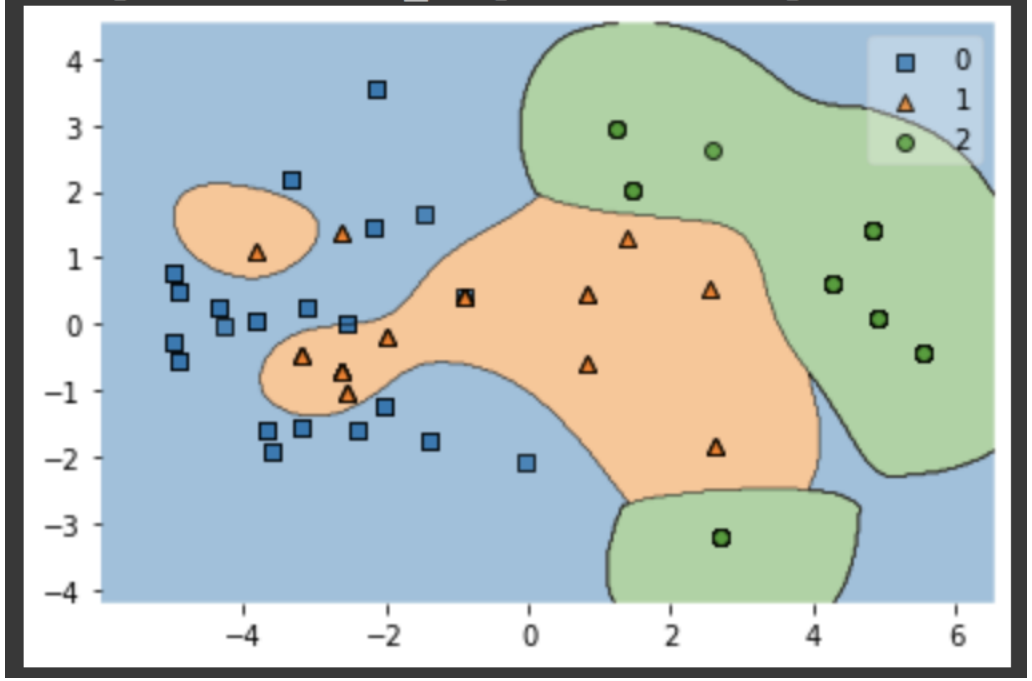
- 貝式分類預測。
- 知識發現：前三項特徵預測出的準確度比原本設想的特徵準確度高

模型預測

Support Vector Machine

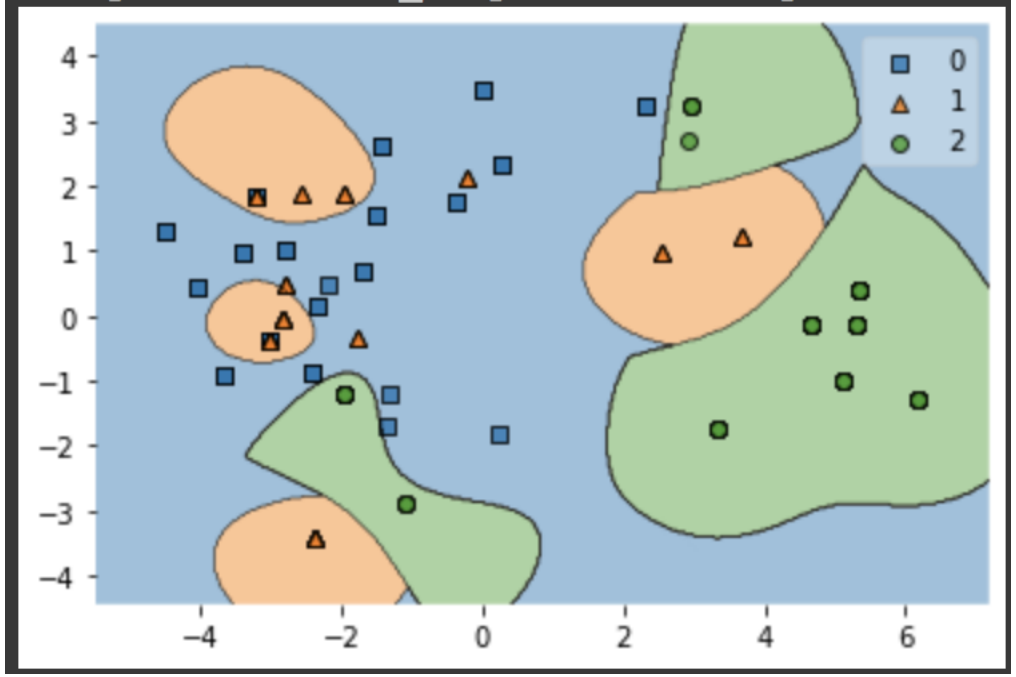
前三項重要特徵Obesity、
Coughing of Blood、Passive Smoker

```
Accuracy of train set ：0.9787
Accuracy of test set：0.9840
/usr/local/lib/python3.7/dist-packages/mlxten
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x7
```



原本設想的重要特徵Smoking、
Air Pollution、Passive Smoker

```
Accuracy of train set ：0.9413
Accuracy of test set：0.9320
/usr/local/lib/python3.7/dist-packages/mlxte
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



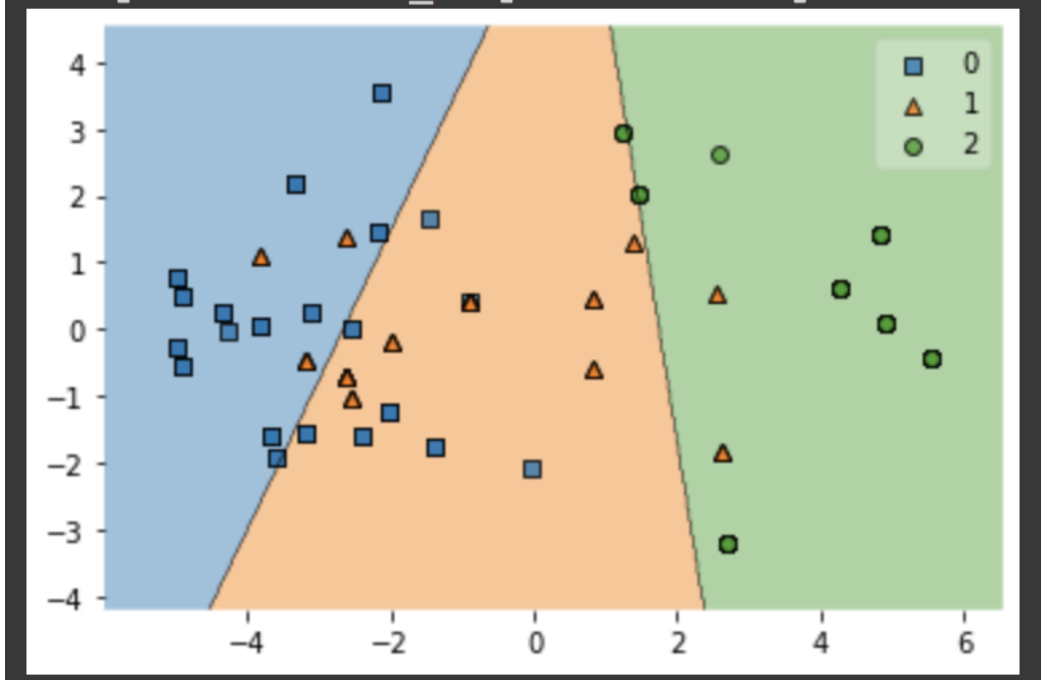
- SVM預測。
- 知識發現：前三項特徵預測出的準確度較高

模型預測

Logistic Regression

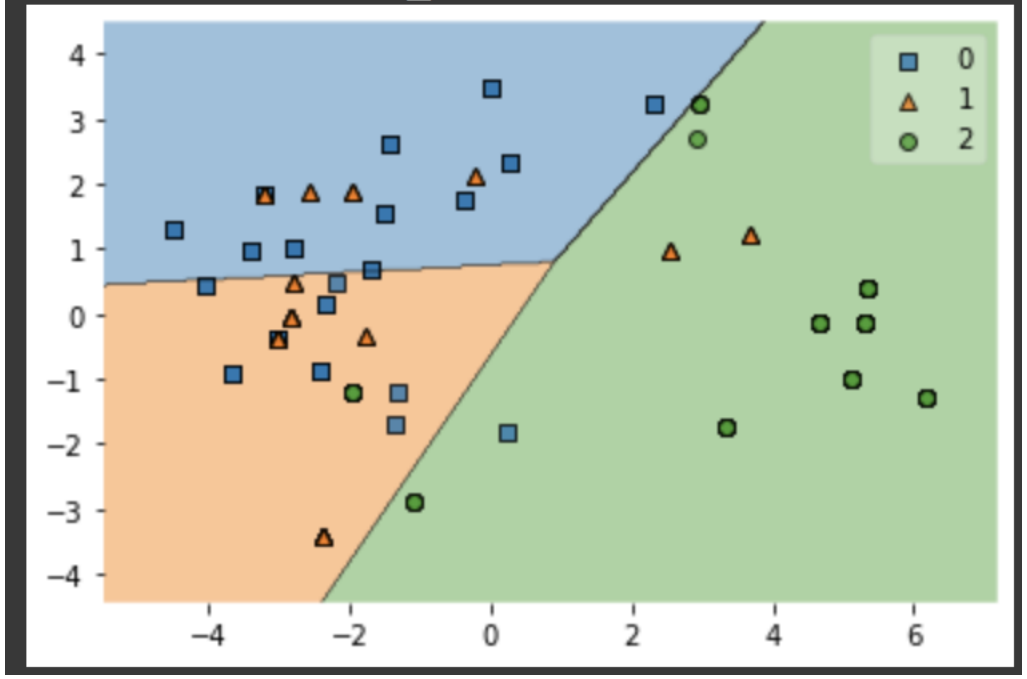
前三項重要特徵Obesity、
Coughing of Blood、Passive Smoker

```
Accuracy of train set ：0.7533
Accuracy of test set：0.8120
/usr/local/lib/python3.7/dist-packages/mlxten
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x7
```



原本設想的重要特徵Smoking、
Air Pollution、Passive Smoker

```
Accuracy of train set ：0.6293
Accuracy of test set：0.7400
/usr/local/lib/python3.7/dist-packages/mlxter
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```

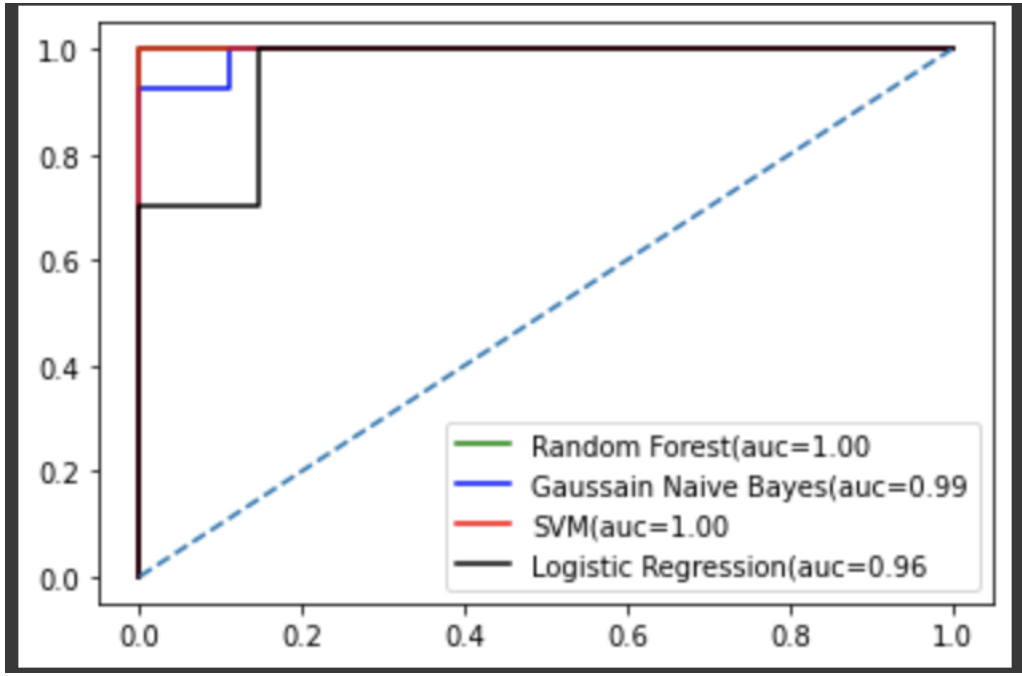


- 邏輯回歸預測。
- 知識發現：前三項重要特徵預測準確率較高

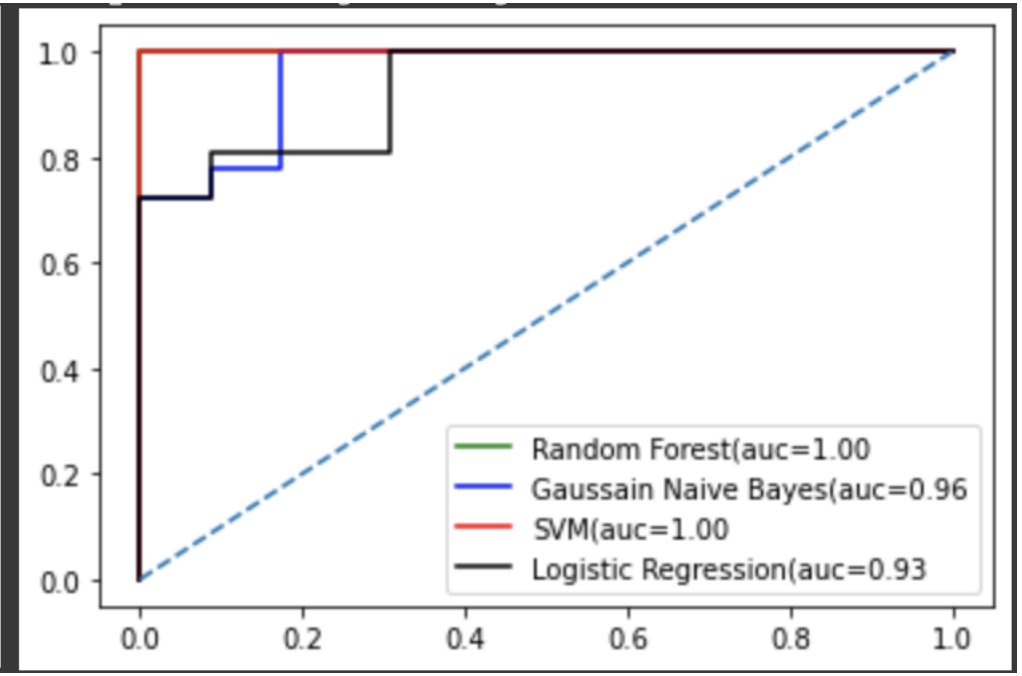
模型評估

ROC/AUC曲線

前三項重要特徵Obesity、
Coughing of Blood、Passive Smoker



原本設想的重要特徵Smoking、
Air Pollution、Passive Smoker



- 使用ROC/AUC曲線進行模型評估
- 知識發現：前三項特徵放入模型評估結果較好

05 結果與討論

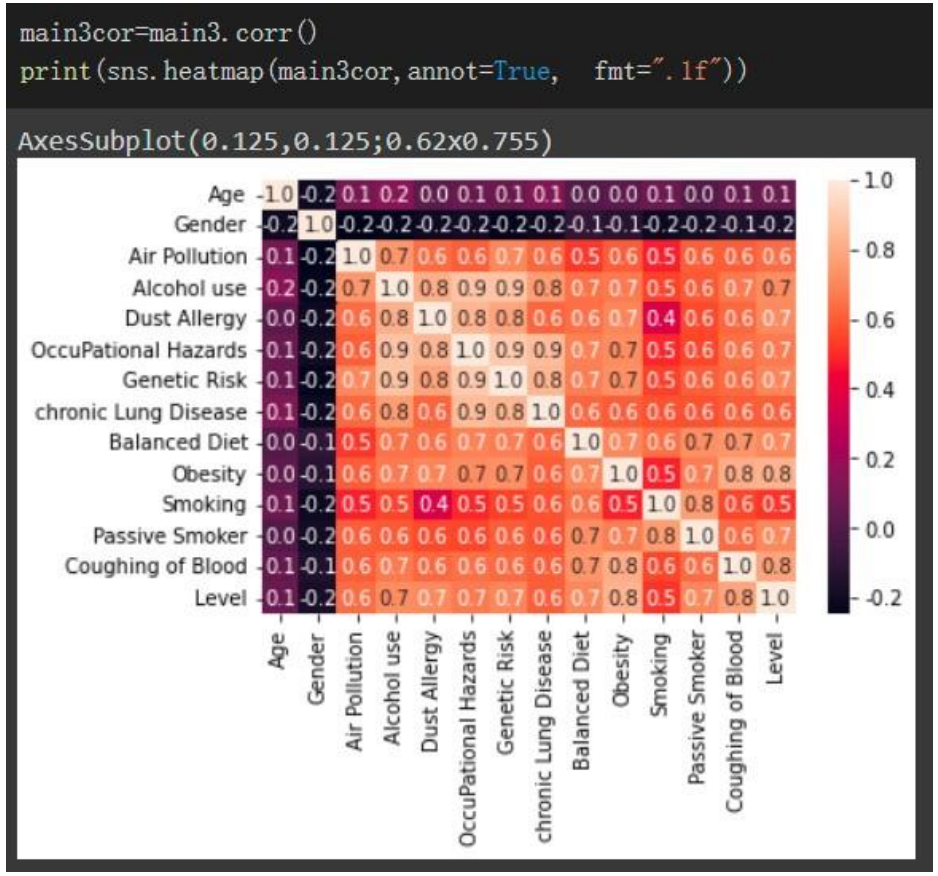
結果發現1

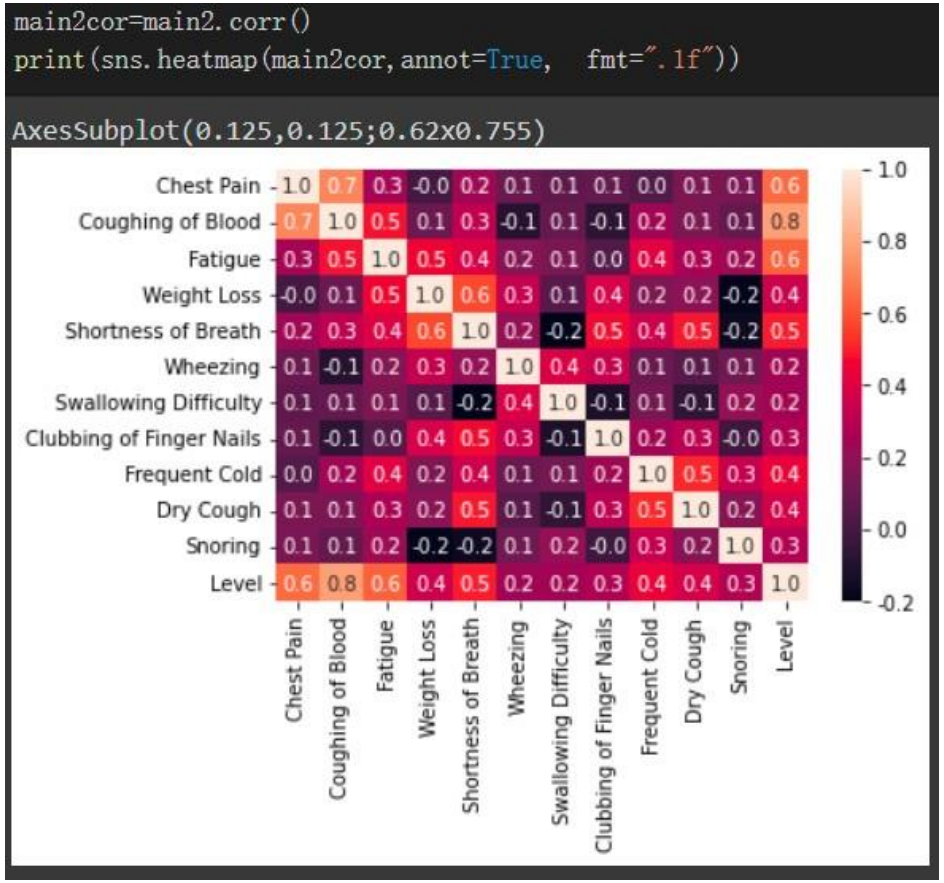
我們利用ANOVA檢定，根據不同年齡層選取以下這些特徵。我們發現各年齡層影響肺癌嚴重程度的特徵不同，其中影響34-53歲肺癌嚴重程度最主要的特徵為Obesity，且其餘特徵皆相同。

	第一特徵	第二特徵	第三特徵	第四特徵	第五特徵	第六特徵	第七特徵	第八特徵	第九特徵
14~23	Balanced Diet	Coughing of Blood	Passive Smoker	Chest Pain	Gender	Dust Allergy	Obesity	Genetic Risk	Alcohol use
24~33	Air Pollution	Passive Smoker	Obesity	Alcohol use	Smoking	Coughing of Blood	Genetic Risk	Shortness of Breath	Balanced Diet
34~43	Obesity	Coughing of Blood	Balanced Diet	OccuPational Hazards	Genetic Risk	Smoking	Passive Smoker	chronic Lung Disease	Dust Allergy
44~53	Obesity	Coughing of Blood	Balanced Diet	OccuPational Hazards	Genetic Risk	Smoking	Passive Smoker	chronic Lung Disease	Dust Allergy
54~63	Passive Smoker	Genetic Risk	Clubbing of Finger Nails	Alcohol use	Air Pollution	OccuPational Hazards	Shortness of Breath	Dust Allergy	Coughing of Blood
64~73	Shortness of Breath	Fatigue	Obesity	Chest Pain	Dry Cough	Coughing of Blood	Passive Smoker	Wheezing	Balanced Diet
Total	Obesity	Coughing of Blood	Passive Smoker	Balanced Diet	Dust Allergy	Alcohol use	Genetic Risk	Air Pollution	OccuPational Hazards

結果發現2

從Heat Map來看，可以看到與Coughing of Blood相關性高達0.7以上的特徵有Chest Pain、Alcohol Use、Balanced Diet、Obesity



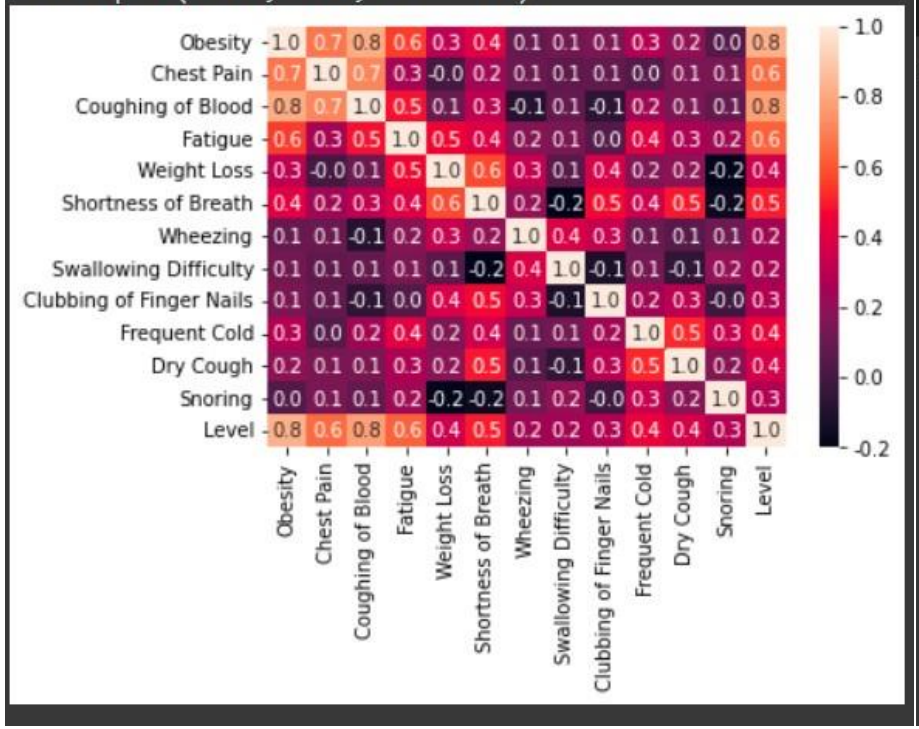


結果發現2(續)

從Heat Map來看，可以看到與Coughing of Blood相關係數0.7以上的特徵，與Obesity有重複的特徵有Chest Pain、Passive Smoker、Balanced Diet、Alcohol Use。可以推測這幾項特徵與判斷肺癌程度有很大關係。

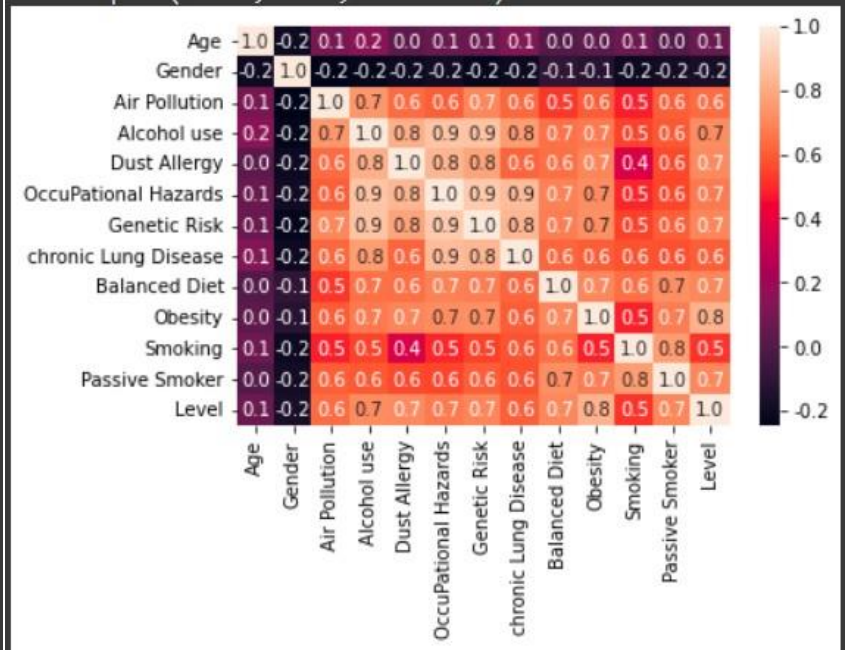
```
main4cor=main4.corr()
print(sns.heatmap(main4cor,annot=True,  fmt=".1f"))
```

```
AxesSubplot(0.125,0.125;0.62x0.755)
```



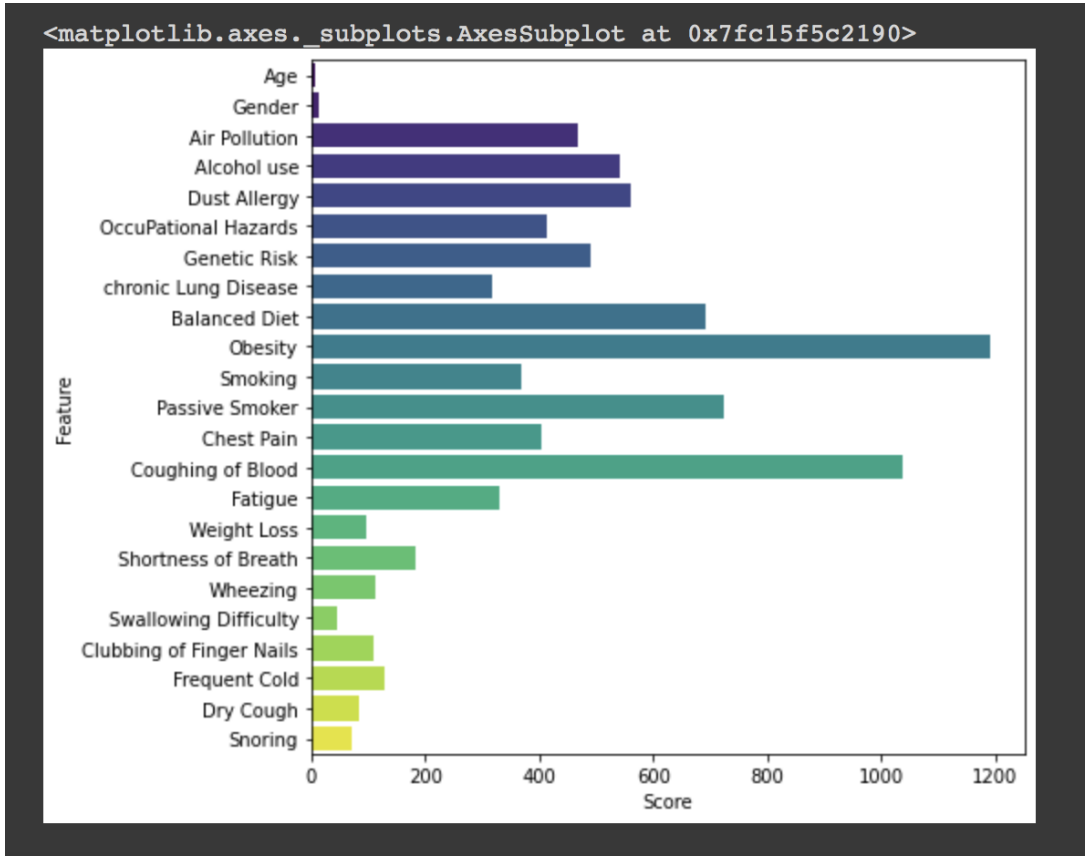

```
main1cor=main1.corr()

print(sns.heatmap(main1cor,annot=True,  fmt=".1f"))
```

```
AxesSubplot(0.125,0.125;0.62x0.755)
```



結果發現2(續)

利用SelectBest所畫出的長條圖，我們可以看出其重要特徵與利用Heat Map和ANOVA分析所選出的特徵大致相同。可以更加準確的知道影響肺癌嚴重程度的重要因素為Obesity、Coughing of Blood、Passive Smoker。

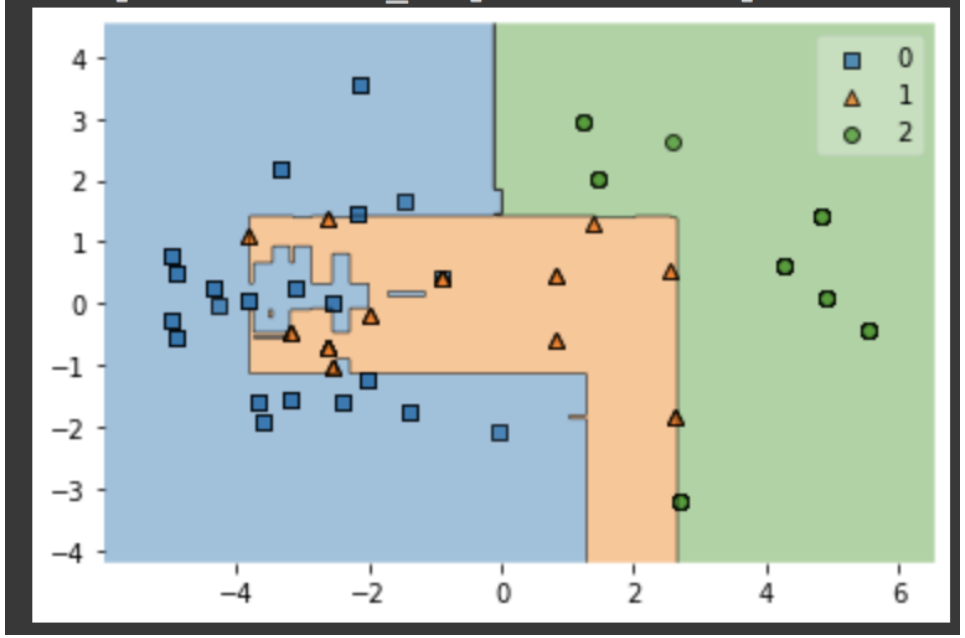


1. Obesity
2. Coughing of Blood
3. Passive Smoker
4. Balanced Diet
5. Dust Allergy

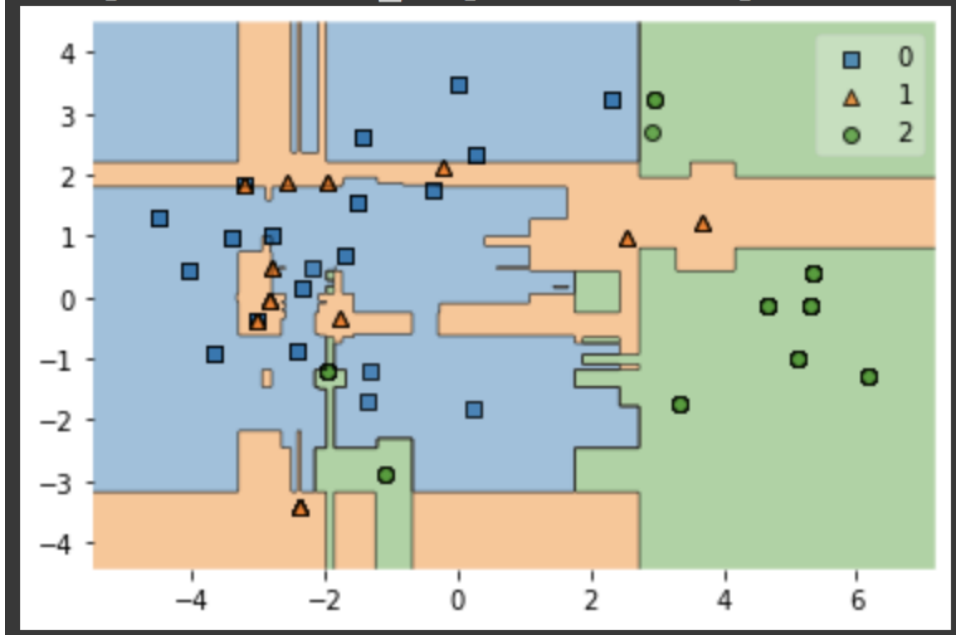
結果發現3

左邊是放入Obesity、Coughing of Blood、Passive Smoker預測出的準確度；右邊則是Smoking、Air Pollution、Passive Smoker 。利用隨機森林去預測，出來的結果是前三項重要特徵的準確度較高。

```
Accuracy of train set ：0.9880
Accuracy of test set：0.9960
/usr/local/lib/python3.7/dist-packages/mlxtend
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x
```




```
Accuracy of train set ：0.9667
Accuracy of test set：0.9760
/usr/local/lib/python3.7/dist-packages/mlxtend
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x
```



結果討論1

整體來說，影響肺癌嚴重程度的前三項重要因素為Obesity、Coughing of Blood、Passive Smoker，而各個年齡主要影響的特徵則如下圖。

Q：影響各年齡層肺癌嚴重程度的特徵為何會有這些差別？

我們可以再根據這些資料去做更細的分析，就目前結果看來我們可以針對不同年齡層提出不同的防止癌症更加惡化方式。
例如針對24-33歲的病人，我們可以提出戴口罩這個方式，以減少吸入更多髒空氣。

	第一特徵	第二特徵	第三特徵	第四特徵	第五特徵	第六特徵	第七特徵	第八特徵	第九特徵
14~23	Balanced Diet	Coughing of Blood	Passive Smoker	Chest Pain	Gender	Dust Allergy	Obesity	Genetic Risk	Alcohol use
24~33	Air Pollution	Passive Smoker	Obesity	Alcohol use	Smoking	Coughing of Blood	Genetic Risk	Shortness of Breath	Balanced Diet
34~43	Obesity	Coughing of Blood	Balanced Diet	OccuPational Hazards	Genetic Risk	Smoking	Passive Smoker	chronic Lung Disease	Dust Allergy
44~53	Obesity	Coughing of Blood	Balanced Diet	OccuPational Hazards	Genetic Risk	Smoking	Passive Smoker	chronic Lung Disease	Dust Allergy
54~63	Passive Smoker	Genetic Risk	Clubbing of Finger Nails	Alcohol use	Air Pollution	OccuPational Hazards	Shortness of Breath	Dust Allergy	Coughing of Blood
64~73	Shortness of Breath	Fatigue	Obesity	Chest Pain	Dry Cough	Coughing of Blood	Passive Smoker	Wheezing	Balanced Diet
Total	Obesity	Coughing of Blood	Passive Smoker	Balanced Diet	Dust Allergy	Alcohol use	Genetic Risk	Air Pollution	OccuPational Hazards

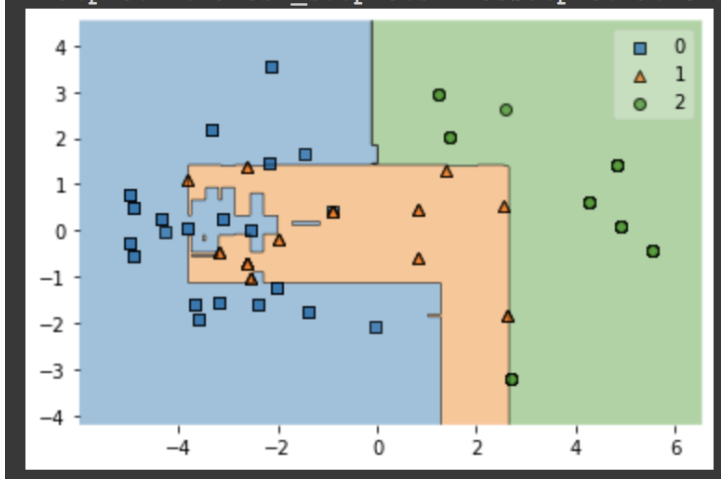
結果討論2

Q：透過ANOVA檢定選出的前三項重要特徵Obesity、Coughing of Blood、Passive Smoker所預測出的準確率高於我們原先設想的特徵Smoking、Air Pollution、Passive Smoker準確率，其原因為何？

其原因可能與資料集本身所蒐集的資料有關，也可能是顯示出Obesity、Coughing of Blood、Passive Smoker是影響肺癌嚴重程度最主要的特徵。

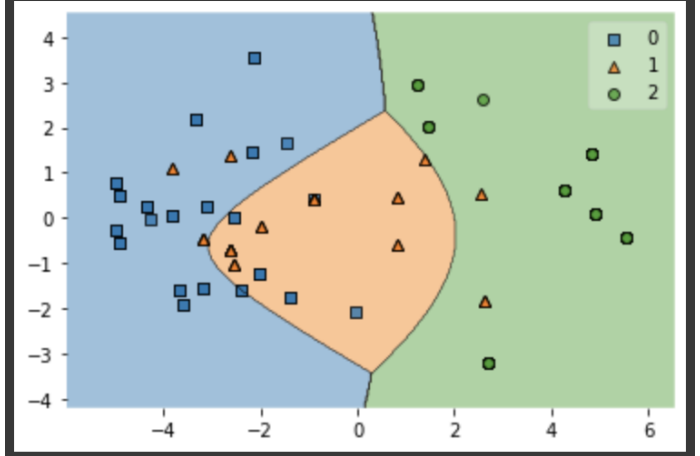
Random Forest Classifier

```
Accuracy of train set ：0.9880
Accuracy of test set：0.9960
/usr/local/lib/python3.7/dist-packages/mlxten
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



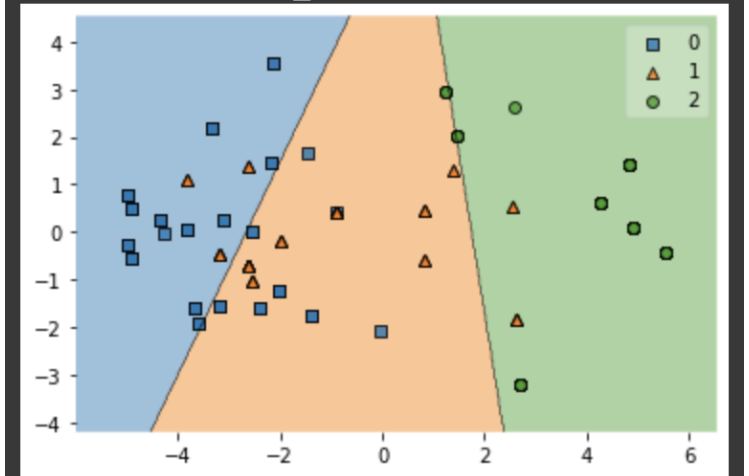
Naïve Bayes Classifier

```
Accuracy of train set ：0.7933
Accuracy of test set：0.8520
/usr/local/lib/python3.7/dist-packages/mlxten
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_mi
<matplotlib.axes._subplots.AxesSubplot at 0x
```



Logistic Regression

```
Accuracy of train set ：0.7533
Accuracy of test set：0.8120
/usr/local/lib/python3.7/dist-packages/mlxten
  ax.axis(xmin=xx.min(), xmax=xx.max(), y_min
<matplotlib.axes._subplots.AxesSubplot at 0x7
```



06 結論

結論

- 一開始我們認為影響肺癌的嚴重程度所呈現前幾項特徵是吸煙、空氣污染。
- 閱讀完文獻後，我們發現個人體質也是影響肺癌嚴重程度的重要特徵。
- 經過我們透過ANOVA作特徵選取，發現吐血、肥胖和吸二手菸是最重要特徵，與我們原先設想的不同。
- 不同年齡層影響肺癌嚴重程度的原因不同，因此在不同年齡時可以特別專注預防某些特徵的狀況。例如34-53歲的病人應特別注意肥胖問題；而54-63歲的病人則需特別注意吸二手煙的狀況。

07 參考資料

參考資料

- 特徵選取

https://machinelearningmastery.com/calculate-feature-importance-with-python/

- PCA

https://www.youtube.com/watch?v=HMOI_lkzW08&ab_channel=StatQuestwithJoshStarmer

- How to predict lung cancer levels! 100% Accuracy

https://www.kaggle.com/christopherwsmith/how-to-predict-lung-cancer-levels-100-accuracy


THANK YOU