110學年度第二學期光電與材料工程學系專題競賽

以Azure Machine Learning Studio分析台灣各縣市空氣品質與肺癌發生率的關係II

授課老師:胡光宇教授
專題編號:B37-110-2-011
組員:10737301曾智博
10737202黃靖淵

摘要

在臺灣癌症一直都位居十大死因之首，其中癌症又以氣管、支氣管和肺癌為死亡率最高的項目，本專題的分析資料來源主要來自台灣癌症發生率地圖所公布2007-2015年的肺癌發生率 與2013-2021年行政院環保署空氣品質監測年報為基礎，分別進行特徵篩選(Filter Based Feature Selection) 與使用何種演算法建立較為準確地預測模型。特徵篩選在全國方面臭氧與肺癌發生率最為相關，但在各縣市則是PM、SO_2、O_3最為相關，與前人張珀銀先生的研究一致，預測模型的演算法則是以Linear Regression 為最準確。透過找出對肺癌發生率有關的空氣汙染物，希望能降低此種汙染物的產生，進而大幅降低肺癌的發生，讓肺癌不再是台灣的國病。

研究方法

我們小組以 Microsoft Azure Machine Learning Studio，使用已經公布的台灣癌症發生率地圖所公布2007-2015年的肺癌發生率與2013-2021年行政院環保署空氣品質監測年報來分析空氣汙染與肺癌發生率的關係。先以特徵篩選(Filter Based Feature Selection)找出全國及各縣市空氣污染物中與肺癌最相關的。接著，嘗試建立以空氣品質來預測肺癌發生率的模型，使用6種迴歸關係演算法，經過訓練及驗證後，探討不同演算法的預測能力。

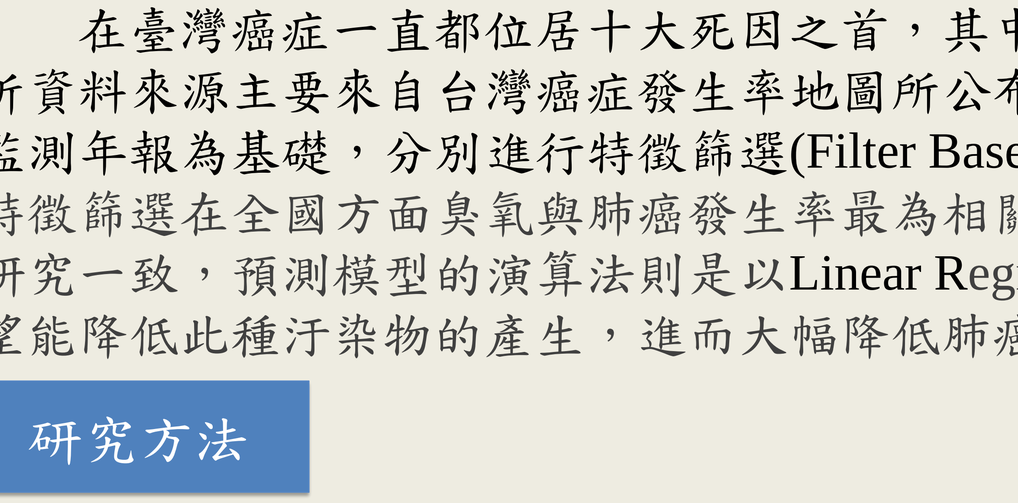

圖1.以Azure Machine Learning Studio特徵篩選流程圖

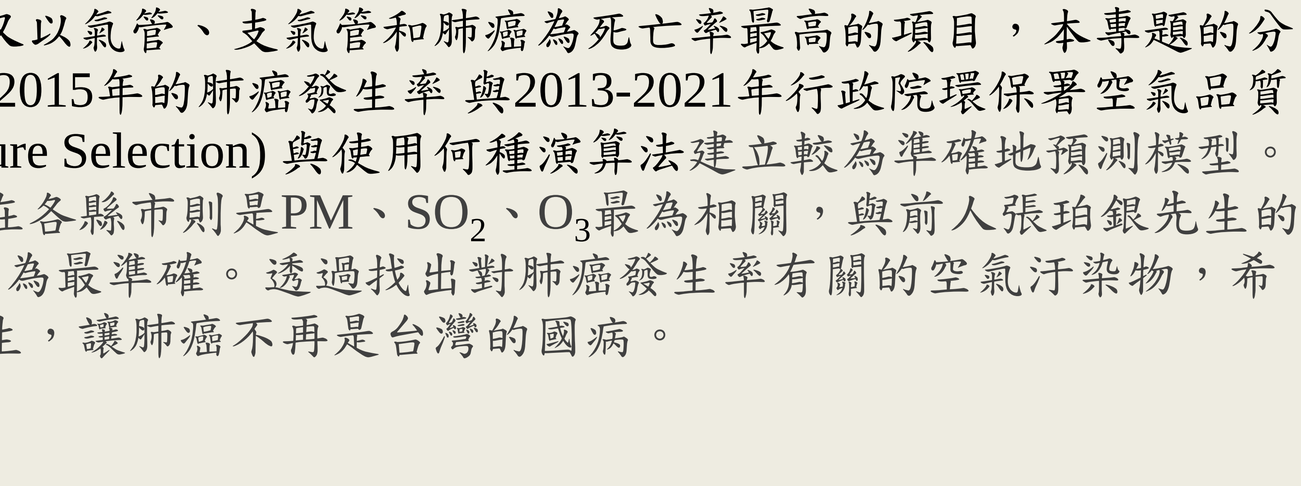

圖2.以Azure Machine Learning Studio計算回歸模型流程圖

研究成果

表1.全國特徵篩選結果

全國	lung cancer	O3,avg (ppb)	$O_{3,8hr}$ (ppb)	NO2 (ppb)	CO (ppm)	PM10 (μg/m3)	PM2.5 (μg/m3)	SO2 (ppb)
Pearson Correlation	1	0.387	0.210	0.168	0.161	0.020	0.058	0.075
Mutual Information	1	0.126	0.092	0.104	0.113	0.085	0.091	0.093
Kendall Correlation	1	0.081	0.121	0.073	0.070	0.014	0.034	0.041
Spearman Correlation	1	0.122	0.183	0.125	0.111	0.020	0.050	0.063

表2.六種回歸演算法判定係數比較

演算法	Negative Log Likelihood	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
Bayesian Linear Regression	169.672252	6.374059	7.242528	1.447675	2.054509	0.126647
Boosted Decision Tree Regression	Infinity	3.850283	4.833408	0.874475	0.915028	0.355426
Decision Forest Regression	52.534112	3.278873	4.165546	0.744697	0.679628	0.327602
Linear Regression	Infinity	3.201353	3.818845	0.727091	0.571205	0.550962
Neural Network Regression		3.218619	3.983184	0.731012	0.621424	0.37808
Poisson Regression		4.403665	5.052851	1.000159	1.000002	-0.000009

結論

1.在全國、桃園市和花蓮縣都是O_3與肺癌發生率較為相關，而新竹縣、雲林縣、嘉義縣和高雄市肺癌發生率中和PM較為相關，嘉義縣在9年的空氣監測年報中PM_{10}的平均也是全國最高的，台北市和嘉義市則是和SO_2較為相關，可以得到在台灣O_3、PM、SO_2與肺癌發生率有較大的相關性與先前張珀銀先生的研究一致。

2.利用迴歸分析建立模型來預測肺癌發生率方面，則是Linear Regression 的判定係數(Coefficient Of Determination) 0.550962值為最高，而判定係數大於0.5就是個不錯的迴歸模型，也代表此種演算法所做出的模型是較其他四種的準確，如果有往後其他年空氣汙染物資料可利用此演算法所做之模型來預測肺癌發生率。