題目：Patient Data Train and Test Set

學號：10446001

姓名：李玉棋

- 摘要

資料分析的領域中，使用機器語言分析大量的數據，已經是一件習以為常的方式，分析結果經常用於研究對未來的預測和決策，是一個非常實用的工具，本研究主要以決策樹進行分析，決策樹廣泛使用於醫學領域，利用分析結果警惕大眾盡量避免一些不好的嗜好、飲食習慣，像是吃油炸食物罹癌機率會提高，或是生活習慣不正常可能會造成身體一些負擔等等，這些種種的研究報告，可以讓大眾做為參考依據，照顧好自己的身體。

- 介紹（研究背景及研究目的）

研究背景：

高血壓素來被視為「沈默殺手」，因沒有明顯症狀而容易被輕忽，一旦太過嚴重，不但容易造成心血管疾病、腦中風、糖尿病和腎臟病等，甚至可能導致猝死，輕忽不得，所以需要好好預防，免除於這些疾病上身。

研究目的：

依據歷年的分析報告，可見高血壓的患者越來越多也逐漸年輕化，若持續下去，國人的健康令人擔憂。以下運用”train_ajEneEa”資料集做研究分析，目的是希望能從分析結果找到可以預防的方式，使用決策樹進行分析。

- 資料集介紹（含資料特徵）及資料集來源

一、資料採集

資料集來自於病人資料的數據，資料集筆數共為 43401 筆，經數據整理後，再以決策樹方式進行分析。

二、資料特徵

自變數共有十二項，依據所需的欄位進行整理，移除不適合的資料，資料整理後剩下 29072 筆，作為分析依據。

```
In [1]: import pandas as pd
        df = pd.read_csv("train_ajEneEa.csv",encoding = "big5").
        dropna(subset=['id','gender','age','hypertension','heart_disease','bmi','smoking_status','avg_glucose_level'])
        df.head()
```

Out[1]:

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
6	52800	Female	52.0	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
7	41413	Female	75.0	0	1	Yes	Self-employed	Rural	243.53	27.0	never smoked	0
8	15266	Female	32.0	0	0	Yes	Private	Rural	77.67	32.3	smokes	0

- 資料預處理

```
In [2]: #missing data
total = df.isnull().sum().sort_values(ascending=False)
percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data.head(20)
```

Out[2]:

	Total	Percent
stroke	0	0.0
smoking_status	0	0.0
bmi	0	0.0
avg_glucose_level	0	0.0
Residence_type	0	0.0
work_type	0	0.0
ever_married	0	0.0
heart_disease	0	0.0
hypertension	0	0.0
age	0	0.0
gender	0	0.0
id	0	0.0

- 機器學習方法進行研究

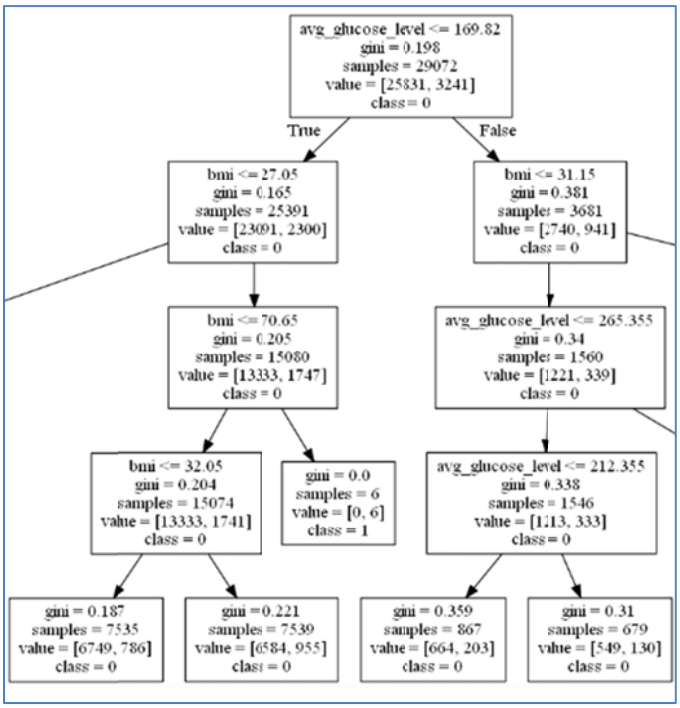
運用機器學習-決策樹方式，探討身體質量指數（bmi）、平均葡萄糖指數（avg_glucose_level）、高血壓（hypertension）、抽菸狀況（smoking_status）。度量方法分為吉尼不純度和熵兩種模型進行研究 …

一、吉尼不純度（Gini）：
運用吉尼不純度的度量方法，依據資料數據的身體質量指數（bmi）和平均葡萄糖指數（avg_glucose_level）預測高血壓（hypertension），class 為 0 表示沒有高血壓、class 為 1 表示有高血壓。從預測結果可以分析出 bmi 指數過高或是 bmi 指數和葡萄糖指數均過高的情況下，導致高血壓的風險提升。

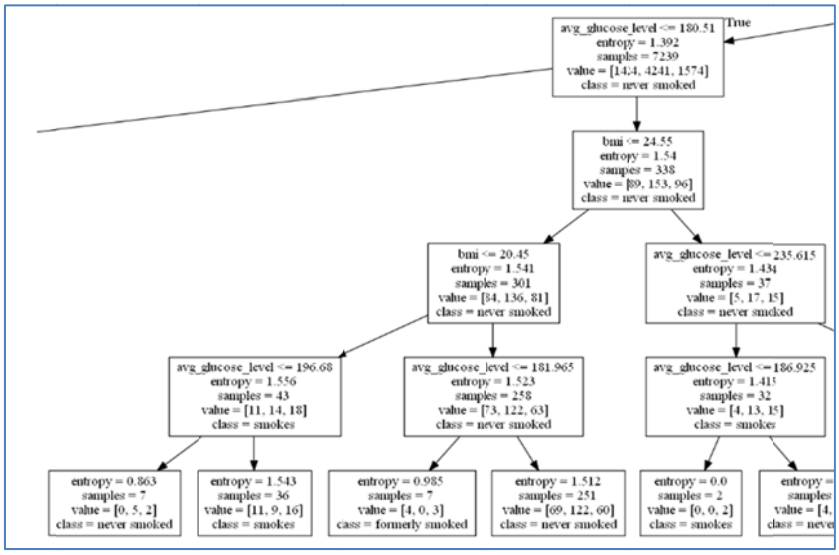
二、熵（Entropy）：
運用熵的度量方法，依據資料數據的身體質量指數（bmi）和平均葡萄糖指數（avg_glucose_level）探討抽菸狀況（smoking_status）。在研究的結果中，能看出抽菸習慣的人，葡萄糖指數會偏高。

- 研究結果及討論(含模型評估與改善)

在吉尼不純度分析中，由葡萄糖指數 169.52 分為兩類，在 bmi 大於 70.65 中 class 則屬於 1，表示可能罹患高血壓的風險極高。



在熵分析中，可以看出 bmi 值在 20 左右、葡萄糖指數大於約 196.68 的情況下，可能是有抽菸的習慣所造成葡萄糖指數偏高。



- 結論

從以上分析中，可以發現抽菸對高血壓的影響極大，因為菸的尼古丁成分會引起小動脈的持續性收縮，小動脈壁的平滑肌變性，血管內膜漸漸增厚，形成小動脈硬化，更促進了高血壓的進一步惡化。一旦成了癮君子，可能會導致一些疾病的產生、傷害自己的身體，後果不堪設想…；分析數據中從 bmi 指數可以發現，bmi

指數越高，高血壓的風險提升，所以我們要控制好自己的體重，才不會惹病上身。

- 參考文獻

高血壓年輕化：
https://www.toplhealth.com/Article/57687

衛生福利部國民健康署
https://www.hpa.gov.tw/Pages/Detail.aspx?nodeid=1135&pid=2978

Patient Data Train and Test Set
https://www.kaggle.com/asaumya/patient-data-train-and-test-set/kernels