

基於身體數值之 抽煙喝酒習慣預測

Team 19
Report Date: 12/27



Background

在一個追求幸福與長壽的世界中，理解影響公共健康的行為細節是至關重要的。吸煙和飲酒是兩種這樣的行為，深深地紮根於文化、社交和個人生活的各個方面。它們不僅影響個體在身體、心理和社交上，而且還對經濟產生深遠的影響，給全球的醫療系統帶來巨大的壓力。

我們希望透過 data mining 的方式，針對於揭示隱藏的模式和關聯，為公共健康干預提供信息以及增強個性化的健康建議幫助預測，對 data 做分析跟預測。

Problem Description

- Input:




Dataset包括人的各種身體指數, 例如 (身高, 體重, 視力, 血壓, 膽固醇指數等), 以及這個人的喝酒指數跟是否抽烟。

- Output:

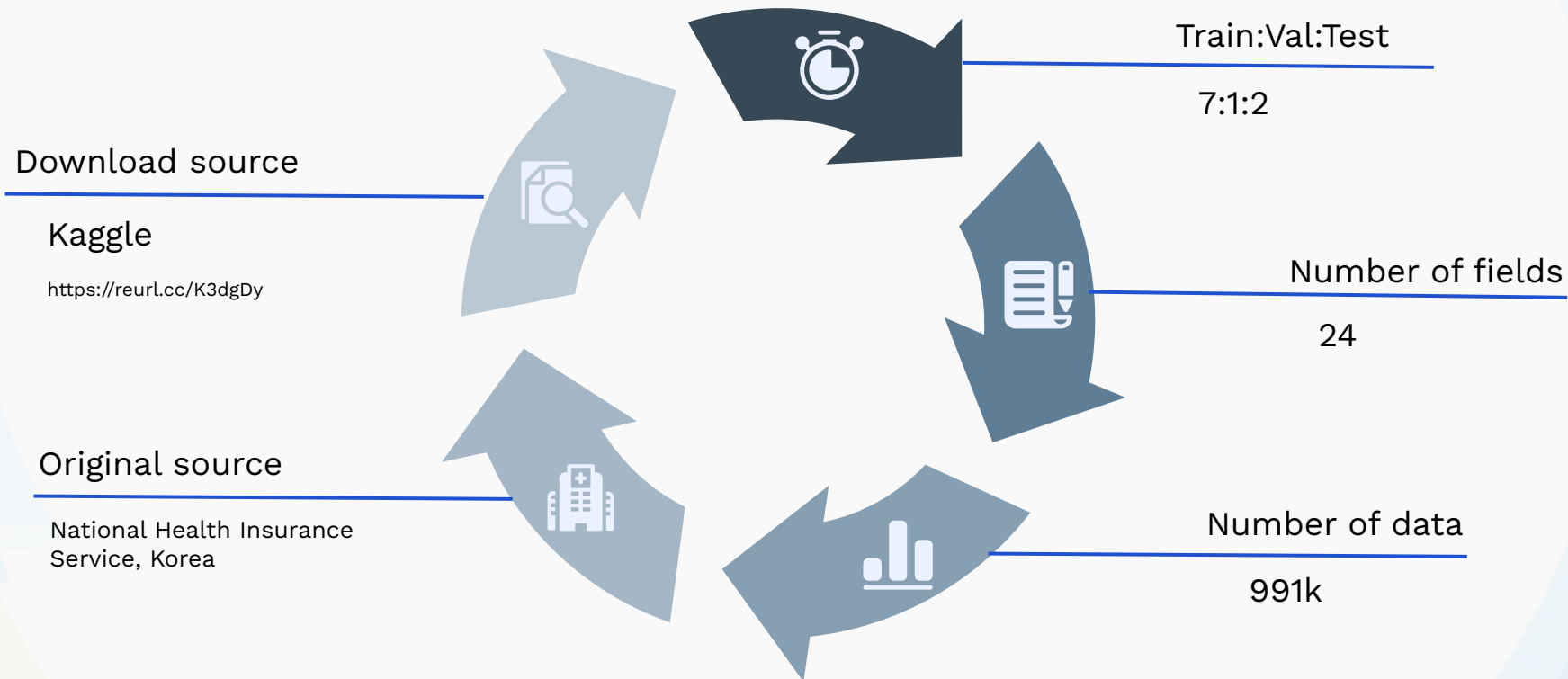
此人是否有烟癮跟他的喝酒指數, 並分析與這兩個預測目標相關的指數。

(訓練兩個模型, 分別預測煙癮以及喝酒, 再進一步分析兩者相關性)

Target Performance

- 藉由身體檢查數值預測一個人是否有煙癮及喝酒習慣
 - Baseline(Both drink & smoke):
 - Accuracy ≥ 0.70
 - F1 score ≥ 0.70
 - Expectation(Both drink & smoke):
 - Accuracy ≥ 0.75
 - F1 score ≥ 0.75
- 分析吸菸對身體造成的影響
- 分析各項數值的相關性
- reference1: [74% Accuracy](#)   [Prediction](#)  [EDA](#) | [Kaggle](#)
- reference2: [Prediction with 72 % accuracy for both smoke/drink](#) | [Kaggle](#)

Data Description



Data Description (Cont.)

Name(EN)	Name(CH)	TYPE	RANGE	MISSING	ADDITION
sex	性別	binary	x	None	男、女
age	年紀	categorical	20-85	None	各類間相差5歲
height	身高	categorical	130-190	None	各類間相差5公分
weight	體重	categorical	25-140	None	各類間相差5kg
waistline	腰圍	numerical	47-126	None	無
sight_left	左眼視力	numerical	0.1-2.06	None	無

Data Description (Cont.)

Name(EN)	Name(CH)	TYPE	RANGE	MISSING	ADDITION
sight_right	右眼視力	numerical	0.1-2.06	None	無
hear_left	左耳聽力	binary	x	None	1為正常、2為異常
hear_right	右耳聽力	binary	x	None	1為正常、2為異常
SBP	收縮壓	numerical	67-273	None	正常範圍 <120mmHg
DBP	舒張壓	numerical	32-185	None	正常範圍 <80mmHg
BLDS	血糖	numerical	25-852	None	正常範圍 70~100mg/dL

Data Description (Cont.)

Name(EN)	Name(CH)	TYPE	RANGE	MISSING	ADDITION
tot_chole	總膽固醇	numerical	30-354	None	正常範圍 130~200mg/dL
HDL_chole	高密度膽固醇	numerical	1-163	None	無
LDL_chole	低密度膽固醇	numerical	1-308	None	無
triglyceride	三酸甘油脂	numerical	1-760	None	無
hemoglobin	血紅素	numerical	1-25	None	無
urine_protein	尿蛋白	categorical	1-6	None	數字越小越正常

Data Description (Cont.)

Name(EN)	Name(CH)	TYPE	RANGE	MISSING	ADDITION
serum_creatinine	血清肌酸酐	numerical	0.1-2.06	None	無
SGOT_AST	麩胺酸轉氨酶 (AST)	numerical	1-201	None	無
SGOT_ALT	麩胺酸轉氨酶 (ALT)	numerical	1-145	None	無
gamma_GTP	谷氨酸轉肽酶	numerical	1-240	None	無
SMK_stat_type_cd	抽菸指數	categorical	1-3	None	1:從不 2:曾經但已戒菸 3: 成癮
DRK_YN	喝酒有無	binary	x	None	Y為有、N為無

Analysis Workflow

利用Apriori演算法找出相關性高的attribute, 再利用decision tree進行分群。

用各種evaluation metrics來衡量各種演算法, 來達成我們的target performance。

Data Processing
& Transformation

Feature
Engeering

Modelling

Evaluation

Knowledge

原始資料已經處理的相對乾淨, 沒有缺失資料, 只需對資料進行屬性的篩選和轉換。

嘗試其他演算法, 如

1. Random forest
2. KNN
3. Naive Bayes

等等來比較結果。

透過模型的各種結果, 希望可以找出一些有意義的pattern。

Data Processing
& Transformation

Feature
Enigneering

Modelling

Evaluation

Knowledge

Encoding



將類別型資料用 one
hot encoding 表示

Outlier



對於連續型資料用平
均以及表準差判斷
outlier, 並移除
outlier

Data Processing
& Transformation

Feature
Enigneering

Modelling

Evaluation

Knowledge

Apriori



將連續型資料分區
以利於進行mining
演算法

	sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right	SBP	...	LDL_chole	triglyceride	hemoglobin
0	Male	AGE: 20-35	Height: 161-170	Weight: 71-140	Waistline: >88	Sight left: 0.71-1.0	Sight right: 0.71-1.0	hear left1.0	hear right1.0	SBP: 68-120	...	HDL: 112-135	Tri: 74-106	hemo: >15.5
1	Male	AGE: 20-35	Height: 171-190	Weight: 71-140	Waistline: >88	Sight left: 0.71-1.0	Sight right: 1.1-1.2	hear left1.0	hear right1.0	SBP: 121-131	...	HDL: >136	Tri: 107-159	hemo: >15.5
2	Male	AGE: 36-45	Height: 161-170	Weight: 71-140	Waistline: >88	Sight left: 1.1-1.2	Sight right: >1.2	hear left1.0	hear right1.0	SBP: 68-120	...	HDL: 1-89	Tri: 74-106	hemo: >15.5
3	Male	AGE: 46-60	Height: 171-190	Weight: 71-140	Waistline: >88	Sight left: >1.2	Sight right: 1.1-1.2	hear left1.0	hear right1.0	SBP: >132	...	HDL: 90-111	Tri: 74-106	hemo: >15.5
4	Male	AGE: 46-60	Height: 161-170	Weight: 56-60	Waistline: 76-81	Sight left: 0.71-1.0	Sight right: 1.1-1.2	hear left1.0	hear right1.0	SBP: >132	...	HDL: 112-135	Tri: 74-106	hemo: 13.3-14.3

Data Processing
& Transformation

Feature
Engineering

Modelling

Evaluation

Knowledge

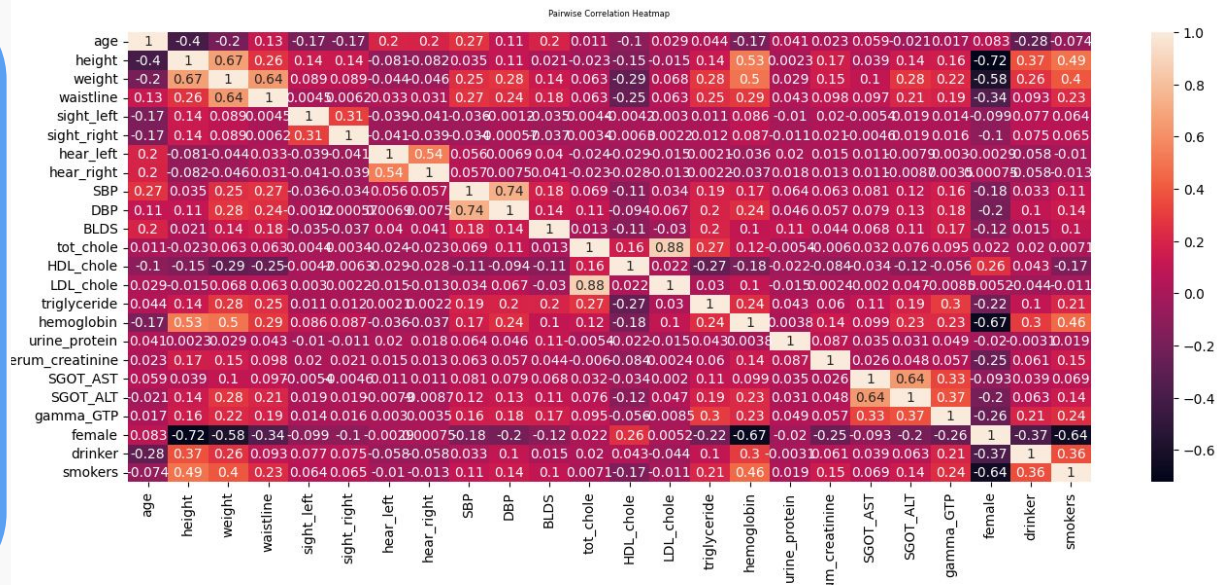
Apriori



利用Apriori演算法，嘗試找出重要的規則

```
('Rule: drkY', 0.49981338503408496)
('Rule: drkY -> Male', 0.3574100263681903)
('Rule: drkY -> Height: 161-170', 0.22380581552757564)
('Rule: smk3.0', 0.2158217211750489)
('Rule: drkY -> smk1.0', 0.21529415562275936)
('Rule: Sight right: 0.71-1.0 -> drkY', 0.205316811688351)
('Rule: drkY -> Sight left: 0.71-1.0', 0.20425663693604454)
('Rule: drkY -> Height: 161-170 -> Male', 0.19957512311544104)
('Rule: Male -> smk3.0', 0.19943188351998192)
('Rule: drkY -> Serum: 0.81-1.0', 0.19092022361516564)
('Rule: drkY -> Gamma: >40', 0.17998559534209044)
('Rule: drkY -> Serum: 0.81-1.0 -> Male', 0.17089694213725581)
('Rule: drkY -> hemo: >15.5', 0.16682470096212623)
('Rule: drkY -> AGE: 20-35', 0.16636371155983884)
('Rule: hemo: >15.5 -> drkY -> Male', 0.1654165145166269)
('Rule: drkY -> Male -> Gamma: >40', 0.16422823111204363)
('Rule: drkY -> smk3.0', 0.16298749377109506)
('Rule: drkY -> Weight: 61-70', 0.15338539722760772)
('Rule: drkY -> Male -> smk3.0', 0.15280134282077096)
('Rule: Weight: 71-140 -> drkY', 0.1520942234093848)
('Rule: drkY -> AGE: 46-60', 0.15175024663437386)
('Rule: drkY -> DBP: 32-70', 0.15018974202750604)
```

利用Pearson相關係數，找到可能和label相關的feature。



Data Processing
& Transformation

Feature
Enigneering

Modelling

Evaluation

Knowledge

Training set: 80% / Testing set: 20%

Random
Forest

LGBM

XGBoost

Catboost

Data Processing
& Transformation

Feature
Enigeering

Modelling

Evaluation

Knowledge



將模型結果交給專業人員, 進一步找出專業或更深層的相關的pattern

Evaluation Metrics

- Accuracy
- F1 score
- Precision score
- Recall score
- ROC curve

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Result Analysis

- Case : **Drink**

	Accuracy	F1 score	Precision	Recall
Random Forest	0.7355	0.7354	0.7358	0.7355
LGBM	0.7323	0.7296	0.7227	0.7366
XGboost	0.7281	0.7282	0.7245	0.7320
Catboost	0.7369	0.7317	0.7267	0.7368

Result Analysis

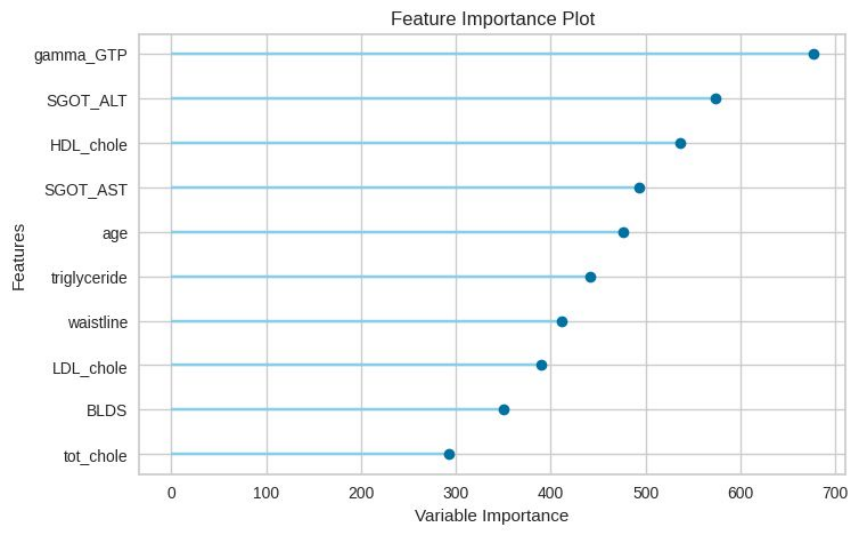
- Case : **Smoke**

	Accuracy	F1 score	Precision	Recall
Random Forest	0.7906	0.7597	0.6996	0.8311
LGBM	0.8132	0.7603	0.7065	0.8232
XGboost	0.8134	0.7595	0.7077	0.8209
Catboost	0.8143	0.7615	0.7088	0.8233

Result Analysis – Important Features

- Case : **Drink**

- gamma_GPT (谷氨酸轉肽酶)
- SGOT_ALT (麩胺酸轉氨酶(ALT))
- HDL_chole
- SGOT_AST (麩胺酸轉氨酶(AST))
- age
- triglyceride (三酸甘油脂)
- waistline
- LDL_chole
- BLDS (血糖)
- tot_chole



Discussion

- Case: **Drink**

- gamma_GPT, SGOT_ALT, SGOT_AST 皆為肝臟相關之酵素, 可以用來評估肝臟 的健康指標
- 過度喝酒極有可能進而導致肥胖, 便與 HDL_chole, LDL_chole, tot_chole (分別為膽固醇的指標)、三酸甘油脂、血糖以及腰圍都皆有相關

Result Analysis – Important Features

- Case : **Smoke**
 - Sex
 - Height
 - Hemoglobin (血紅素)
 - Gamma GTP (谷氨酸轉肽酶)
 - Serum creatinine (血清肌酸酐)
 - Weight
 - Triglyceride (三酸甘油酯)
 - Waistline (腰圍)
 - Age
 - SGOT ALT (麩胺酸轉氨酶)

Discussion

- Case: **Smoke**
 - 抽菸跟性別有非常大的相關性 (男性吸煙者 > 女性吸菸者)
 - 抽菸會使血液中一氧化碳濃度增加, 含氧量減少, 進而影響血紅素水平
 - 長期抽菸可能與肝臟損傷相關, 進而影響肝臟相關 酶的水平, 包括 Gamma GTP
 - 抽菸可能對腎臟功能 產生影響, 這可能會體現在 Serum creatinine (血清肌酸酐) 水平的變化上

Conclusion

- 我們的方法成功達到我們預設的 **Baseline**, 而在**expectation**的部分, 雖然**Drink**的部分並沒有成功達到原先預設的 值, 但在**Smoke**的部分卻有達標
- 藉此**project**, 我們探討了身體指數與抽菸喝酒的相關性, 分析了抽菸以及喝酒可能帶來的影響, 希望可以幫助到需要戒菸或戒酒的人

Thank you for your attention!

312554018 曾昱仁 312553046 何承原
312554011 謝翊庭 0812212 丁祐承

