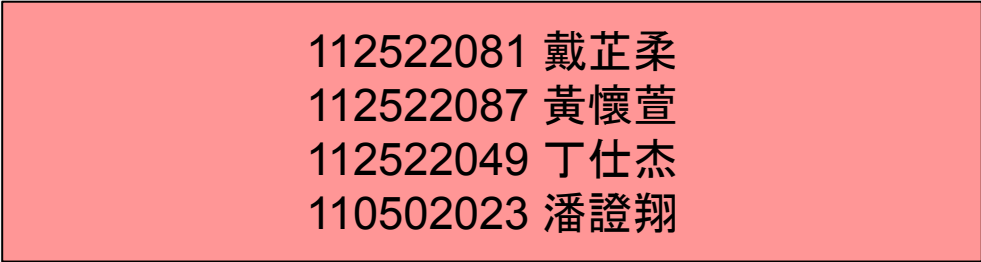




G12



Diabetes Health Indicators



112522081 戴芷柔
112522087 黃懷萱
112522049 丁仕杰
110502023 潘證翔



資料集介紹

資料前處理

特徵說明



CDC Diabetes Health Indicators External

Linked on 9/25/2023

The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. The 35 features consist of some demographics, lab test results, and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes, is pre-diabetic, or healthy.

Dataset Characteristics

Tabular, Multivariate

Subject Area

Health and Medicine

Associated Tasks

Classification

Feature Type

Categorical, Integer

Instances

253680

Features

21

資料清理、處理缺失值

整數型特徵

BMI, GenHlth, Age, Education...

二元分類特徵

HighBP, HighChol, Smoker, Stroke...

By. UC Irvine Machine Learning Repository



BMI(體重指數) 0 - 100

GenHlth(自我報告健康狀)

1 = 優, 2 = 非常好, 3 = 良好,
4 = 一般, 5 = 差

MentHlth(心理健康):

有幾天心理健康狀況不佳的 1 - 30 day

PhysHlth(身體健康)

有無感冒或是受傷的情況 1~30 day

Age(年齡):

年齡區間, 分類 1 - 13 (18 ~)

Education(教育程度)

1: 未受教育 OR 幼稚園

2: 小學 ~ 國中

3: 國中畢業

4: 高中畢業

5: 大學 or 五專

6: 大學畢業

Income(收入水平)

1: 低於\$10,000

2: \$10,000 ~ \$14,999

3: \$15,000 ~ \$19,999

4: \$20,000 ~ \$24,999

5: \$25,000 ~ \$34,999

6: \$35,000 ~ \$50,000

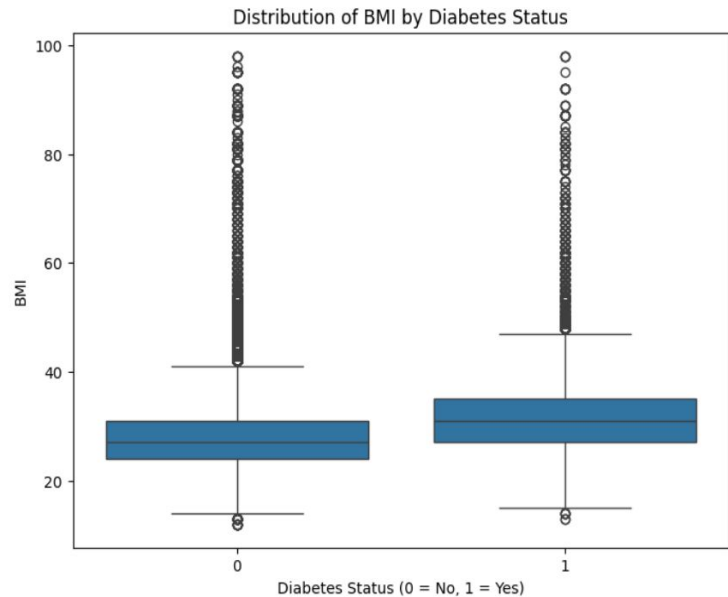
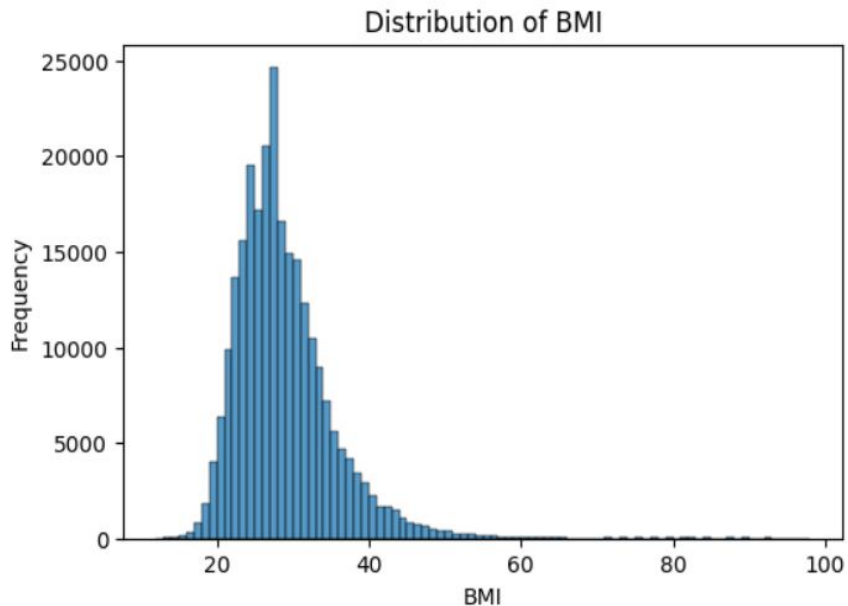
7: 高於\$75,000

▶▶▶ Data

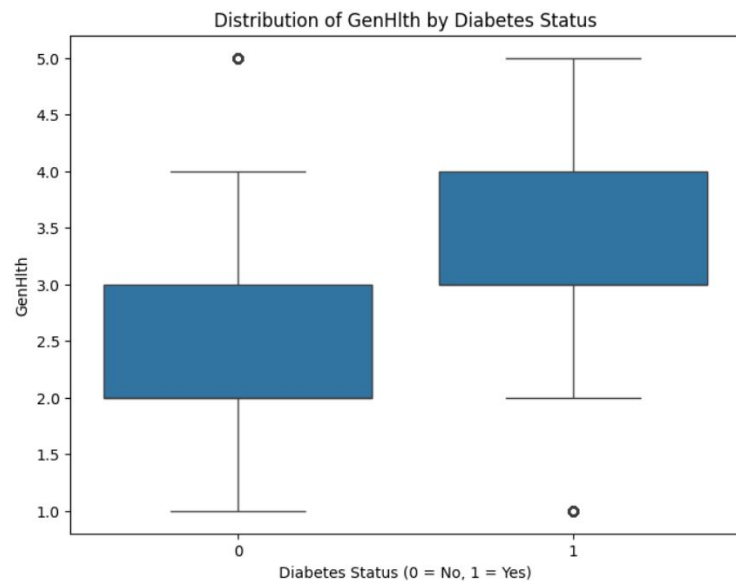
二元分類特徵 (Binary Features)

- **HighBP** (是否有高血壓)
- **HighChol** (是否有高膽固醇)
- **CholCheck** (近五年有無進行膽固醇檢查)
- **Smoker** (一生至少吸過100支煙)
- **Stroke** (是否有中風)
- **Heart Disease or Attack** (是否有冠狀動脈心臟病或心肌梗塞)
- **PhysActivity** (過去30天內的不包括工作的體力活動)
- **Fruits** (每日水果攝入)
- **Veggies** (每日蔬菜攝入)
- **HvyAlcoholConsump** (成年男性每週飲酒超過14杯, 成年女性每週飲酒超過7杯)
- **AnyHealthcare** (是否有任何形式的醫療保障)
- **NoDocbcCost** (曾因費用原因未看醫生)
- **DiffWalk** (感覺行走或爬樓梯困難)
- **Sex** (性別)
 - 0: 女性, 1: 男性

BMI

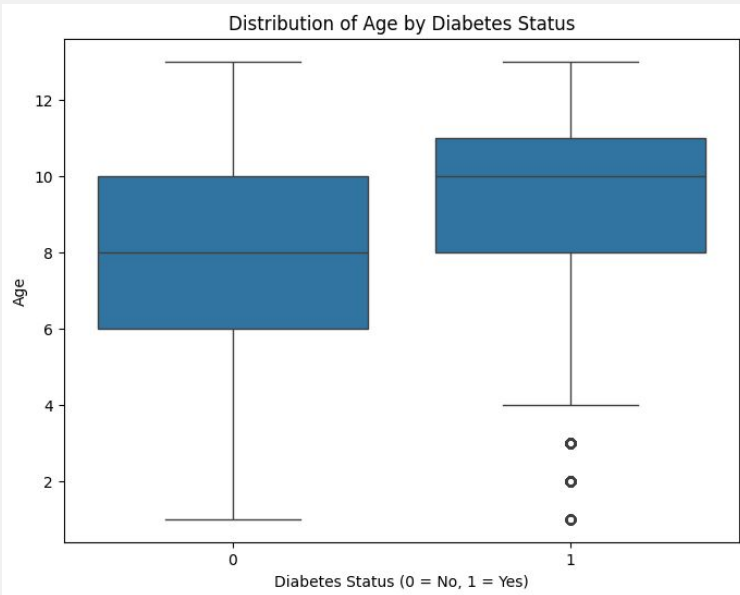


健康自評



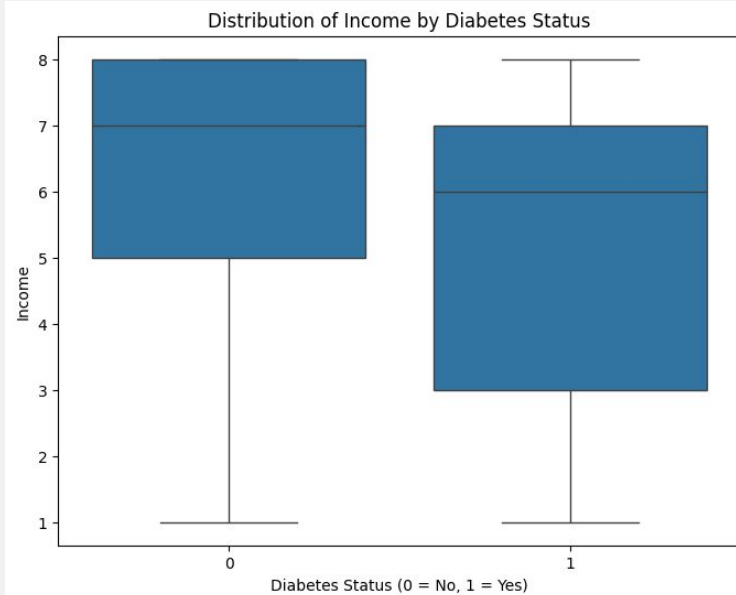
1 = 優
2 = 非常好
3 = 良好
4 = 一般
5 = 差

年齡



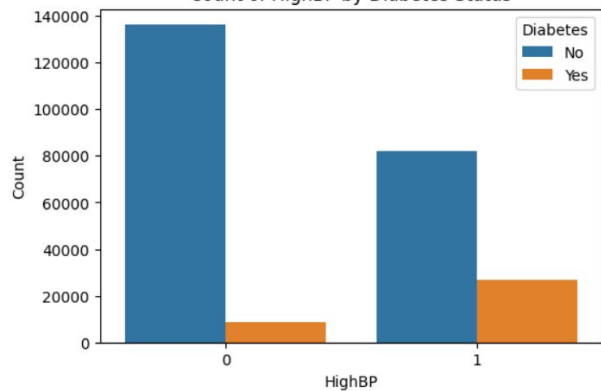
1 = 18 ~ 24
2 = 25 ~ 29
3 = 30 ~ 34
4 = 35 ~ 39
5 = 40 ~ 44
6 = 45 ~ 49
7 = 50 ~ 54
8 = 55 ~ 59
9 = 60 ~ 64
10 = 65 ~ 69
11 = 70 ~ 74
12 = 75 ~ 79
13 = 高於80

收入

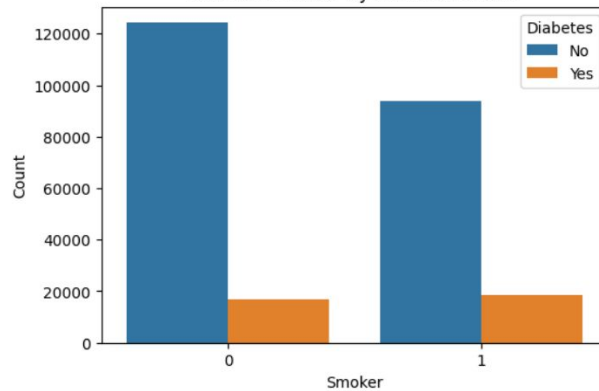


1 : 低於\$10,000
2 : \$10,000 ~ \$14,999
3 : \$15,000 ~ \$19,999
4 : \$20,000 ~ \$24,999
5 : \$25,000 ~ \$34,999
6 : \$35,000 ~ \$49,999
7 : \$50,000 ~ \$75,000
8 : 高於\$75,000

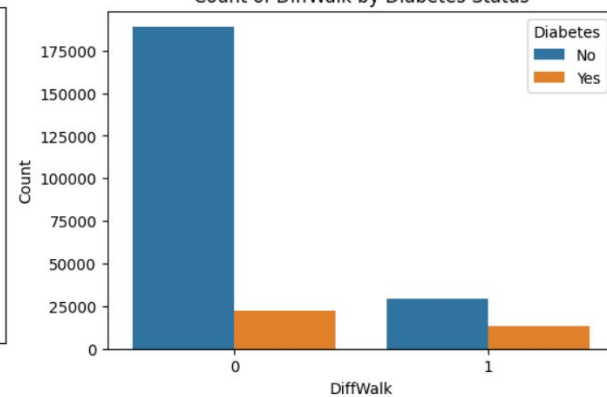
Count of HighBP by Diabetes Status



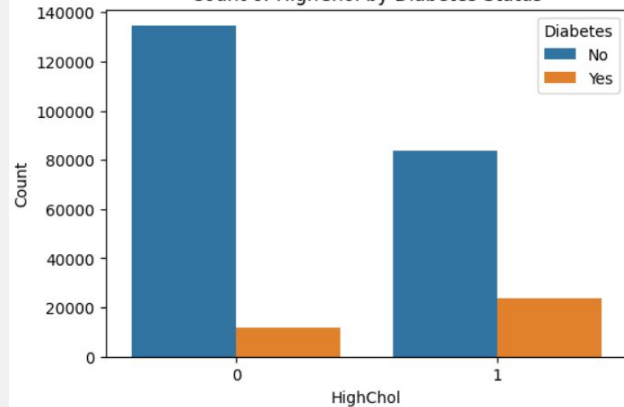
Count of Smoker by Diabetes Status



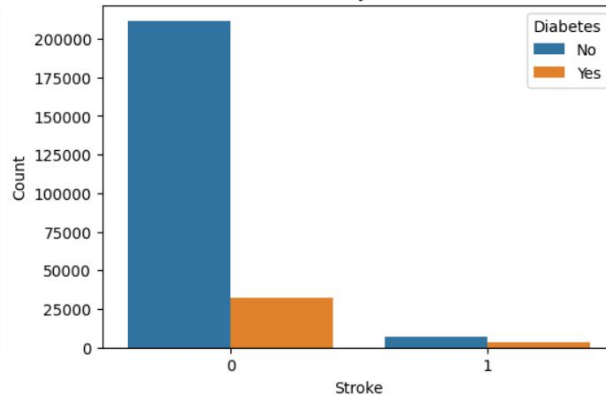
Count of DiffWalk by Diabetes Status



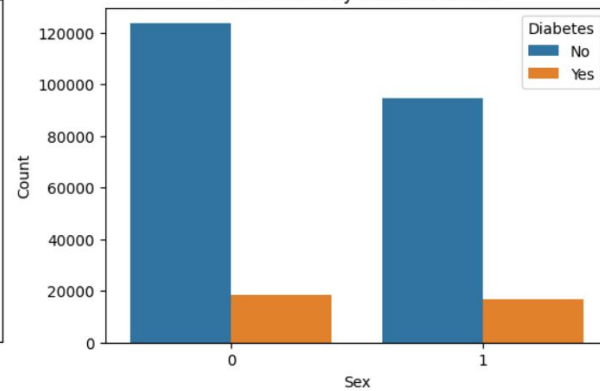
Count of HighChol by Diabetes Status



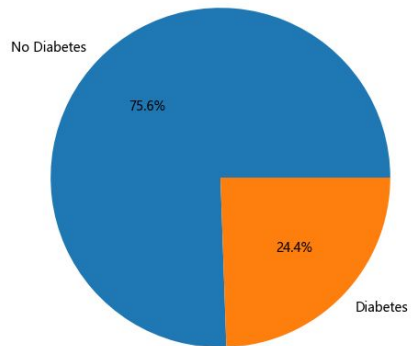
Count of Stroke by Diabetes Status



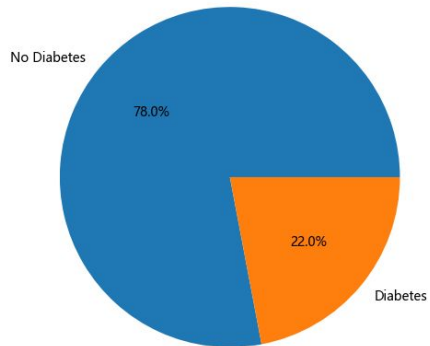
Count of Sex by Diabetes Status



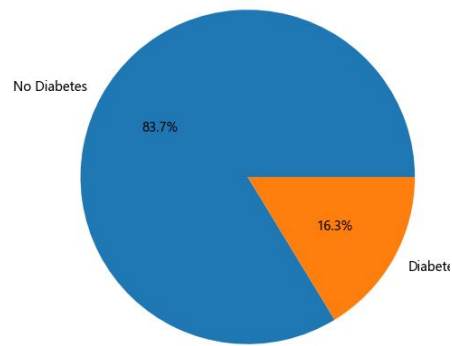
Diabetes Distribution for HighBP = 1



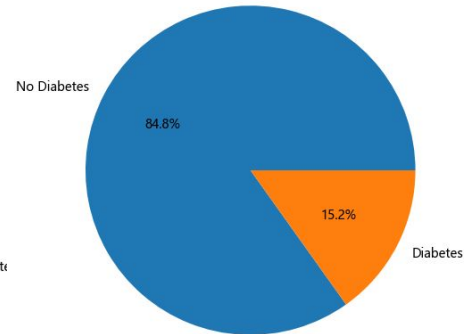
Diabetes Distribution for HighChol = 1



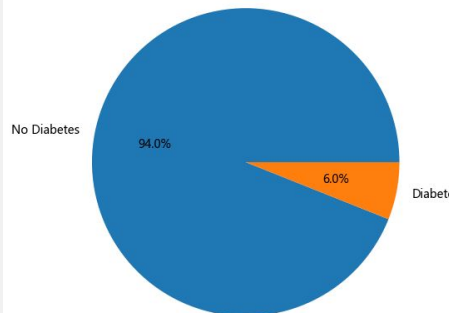
Diabetes Distribution for Smoker = 1



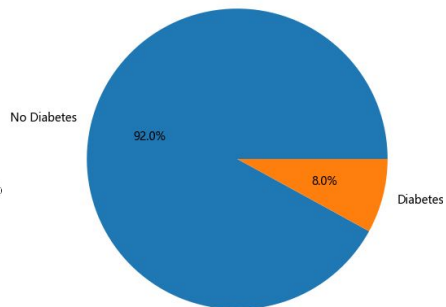
Diabetes Distribution for Sex = 1



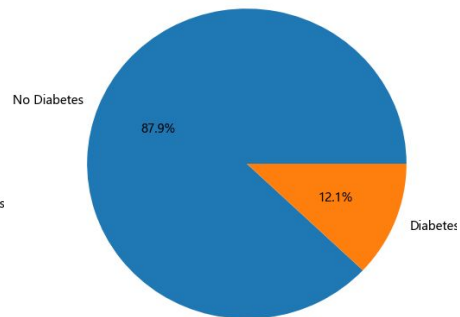
Diabetes Distribution for HighBP = 0



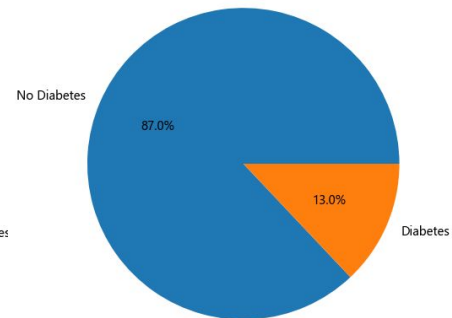
Diabetes Distribution for HighChol = 0



Diabetes Distribution for Smoker = 0



Diabetes Distribution for Sex = 0

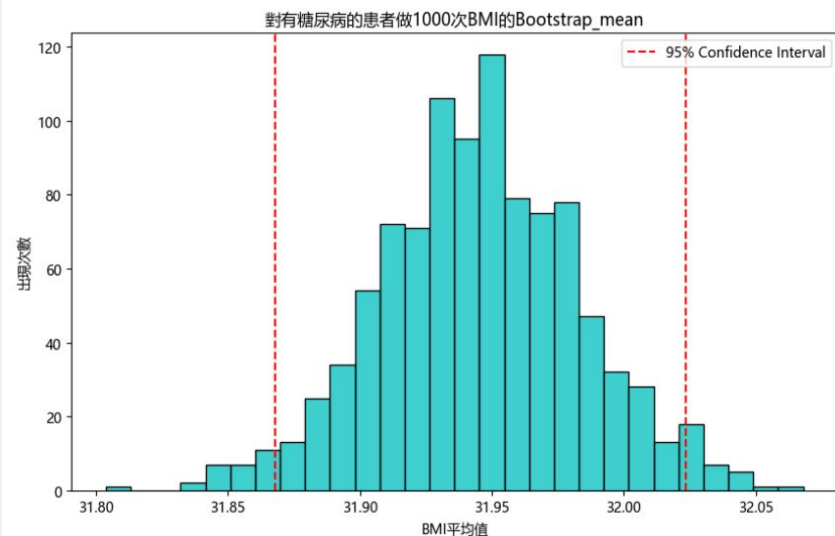


95%
CI

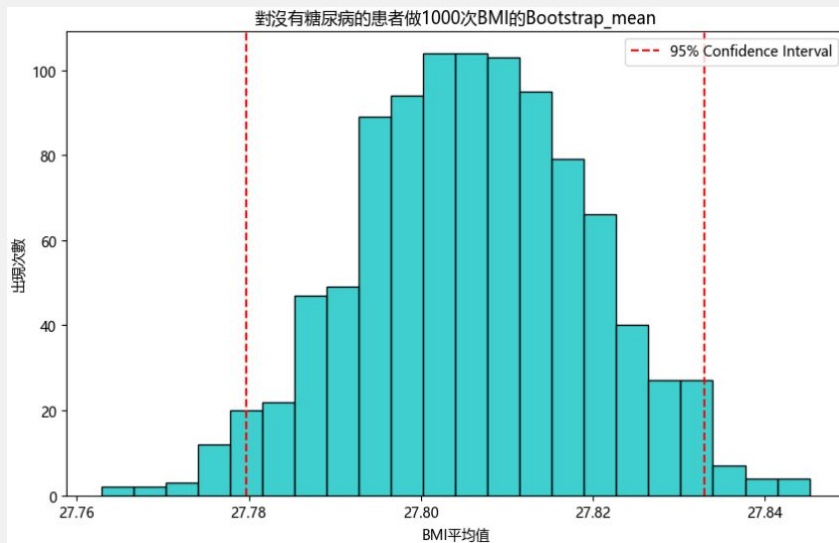
整數型特徵bootstrap

正常範圍: $18.5 \leq \text{BMI} < 24$
過重: $24 \leq \text{BMI} < 27$
輕度肥胖: $27 \leq \text{BMI} < 30$
中度肥胖: $30 \leq \text{BMI} < 35$
重度肥胖: $\text{BMI} \geq 35$

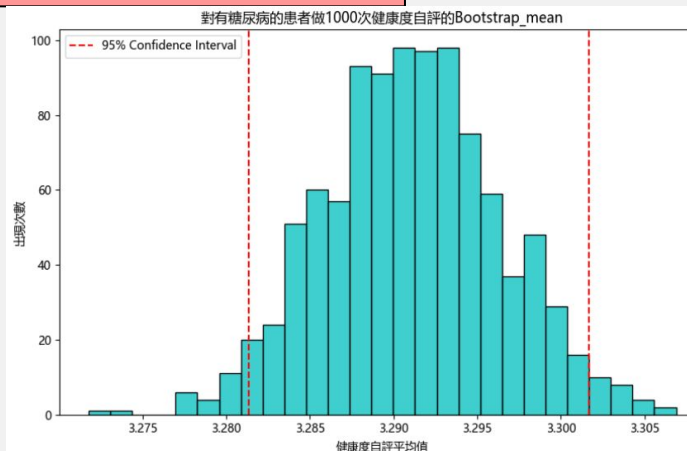
有糖尿病之BMI [31.87, 32.02]



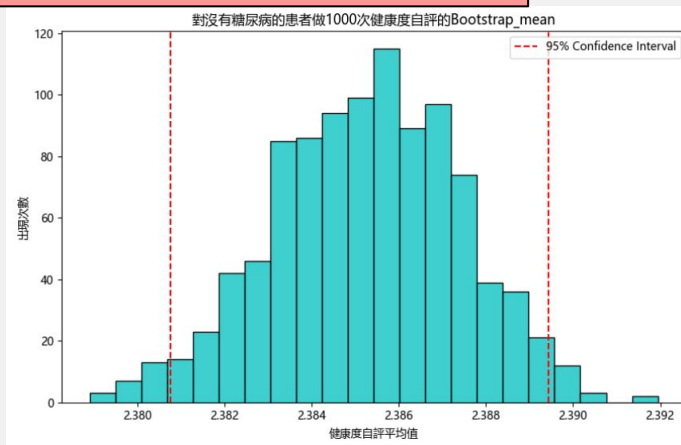
無糖尿病之BMI [[27.78, 27.83]



有糖尿病之健康度自評[3.28, 3.30]
中位數皆為3

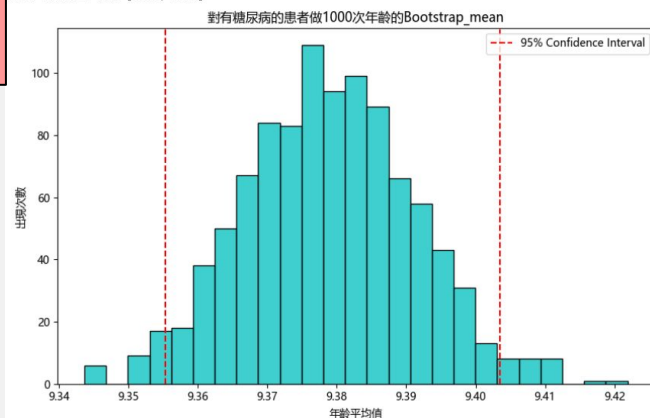


無糖尿病之健康度自評[2.38, 2.39]
中位數皆為2

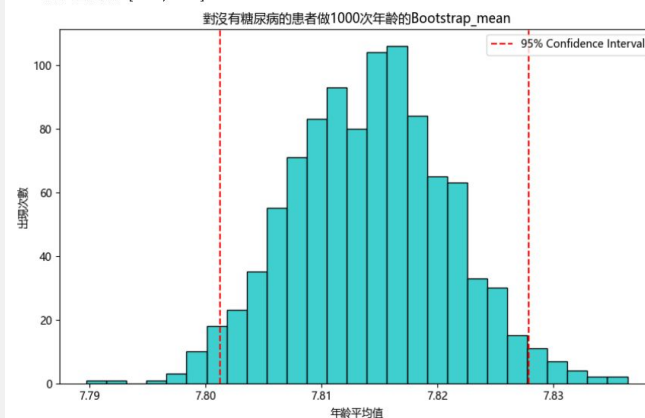


年齡
bootstrap

95% 信賴區間為: [9.36, 9.40]



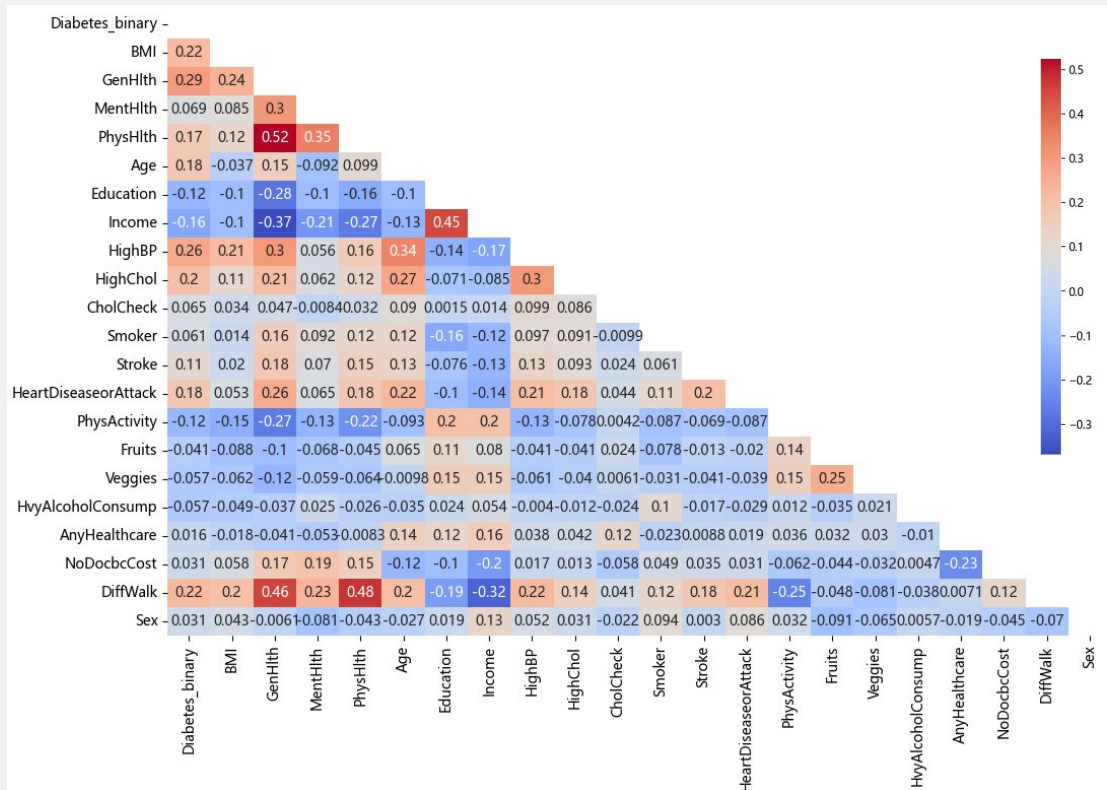
95% 信賴區間為: [7.80, 7.83]



各特徵相關性分析 (Pearson correlation)

相關性較高之組合(>0.3)

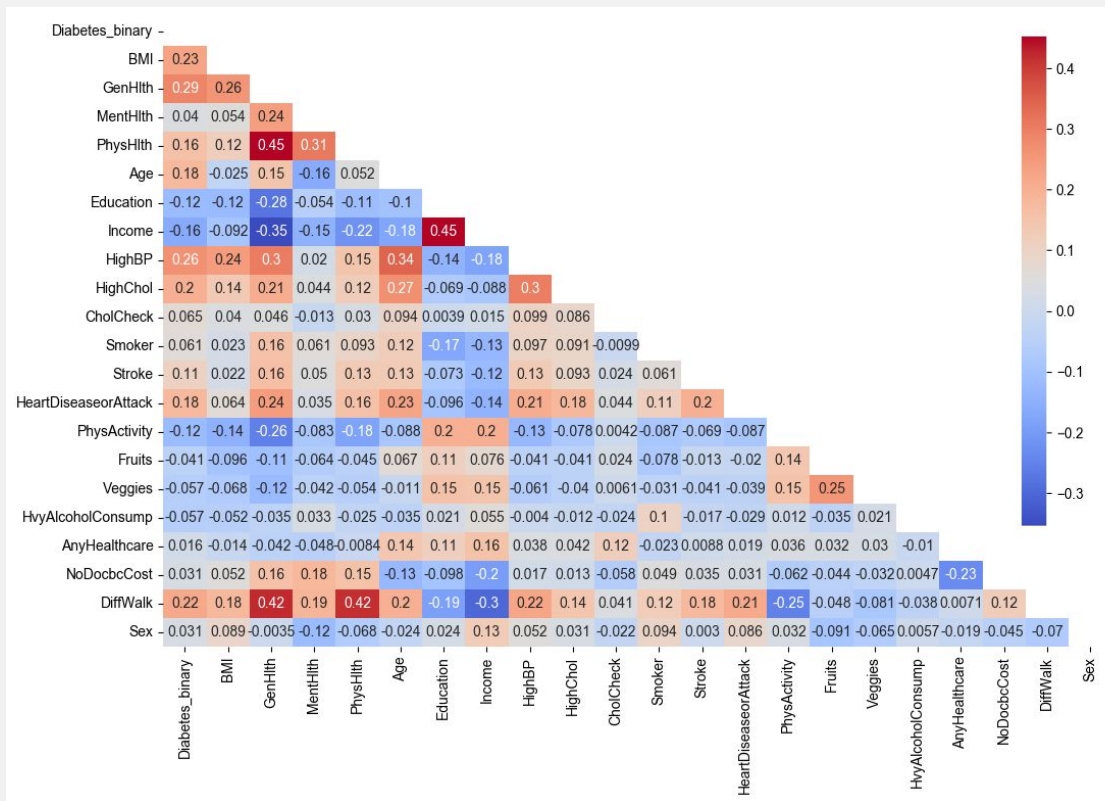
- GenHlth(自我報告健康狀):
 - MenHlth(0.3),
 - PhysHlth(0.52)
 - Income(-0.37)
 - HighBP(0.3)
- MentHlth(心理健康)
 - PhysHlth(0.35)
- PhysHlth(身體健康)
 - DiffWalk(0.48)
 - Age: HighBP(0.34)
- Education
 - Income(0.45)
- Income:
 - DiffWalk(-0.32)
- HighBP(高血壓)
 - HighCol(0.3)



各特徵相關性分析 (spearman correlation)

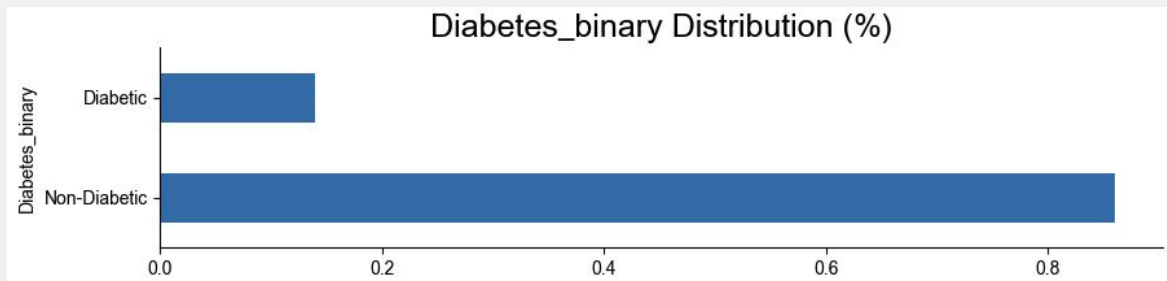
相關性較高之組合(>0.3)

- **PhysHlth(身體健康)**
 - GenHlth(0.45)
- **Income**
 - GenHlth(-0.35)
 - Education(0.45)
- **DiffWalk**
 - GenHlth(0.42)
 - PhysHlth(0.42)
 - Income(-0.3)
- **HighBP(高血壓)**
 - GenHlth(0.3)
 - Age(0.34)
 - HighChol(0.3)



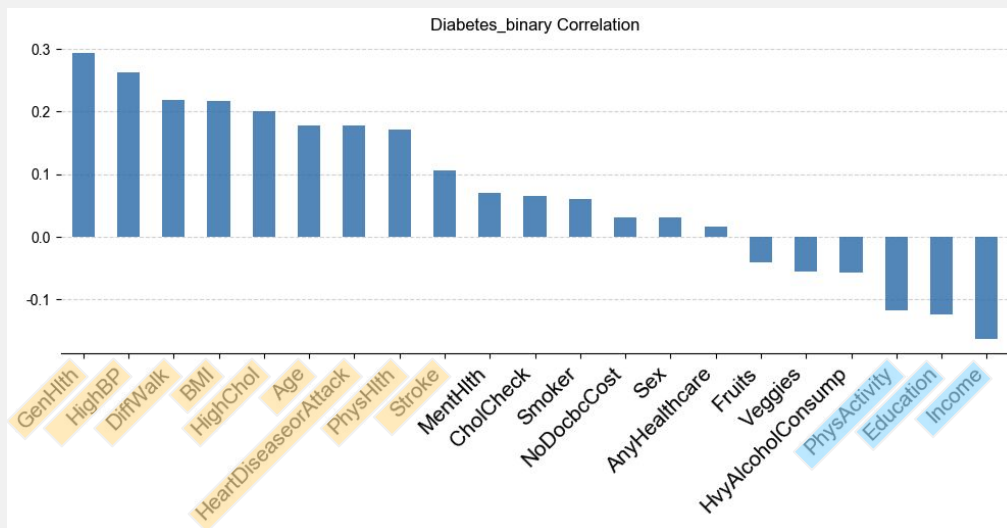
EDA

資料集糖尿病比例
14 : 86



Spearman Correlation
糖尿病vs 各變數

設定 $\alpha = 0.1$ →



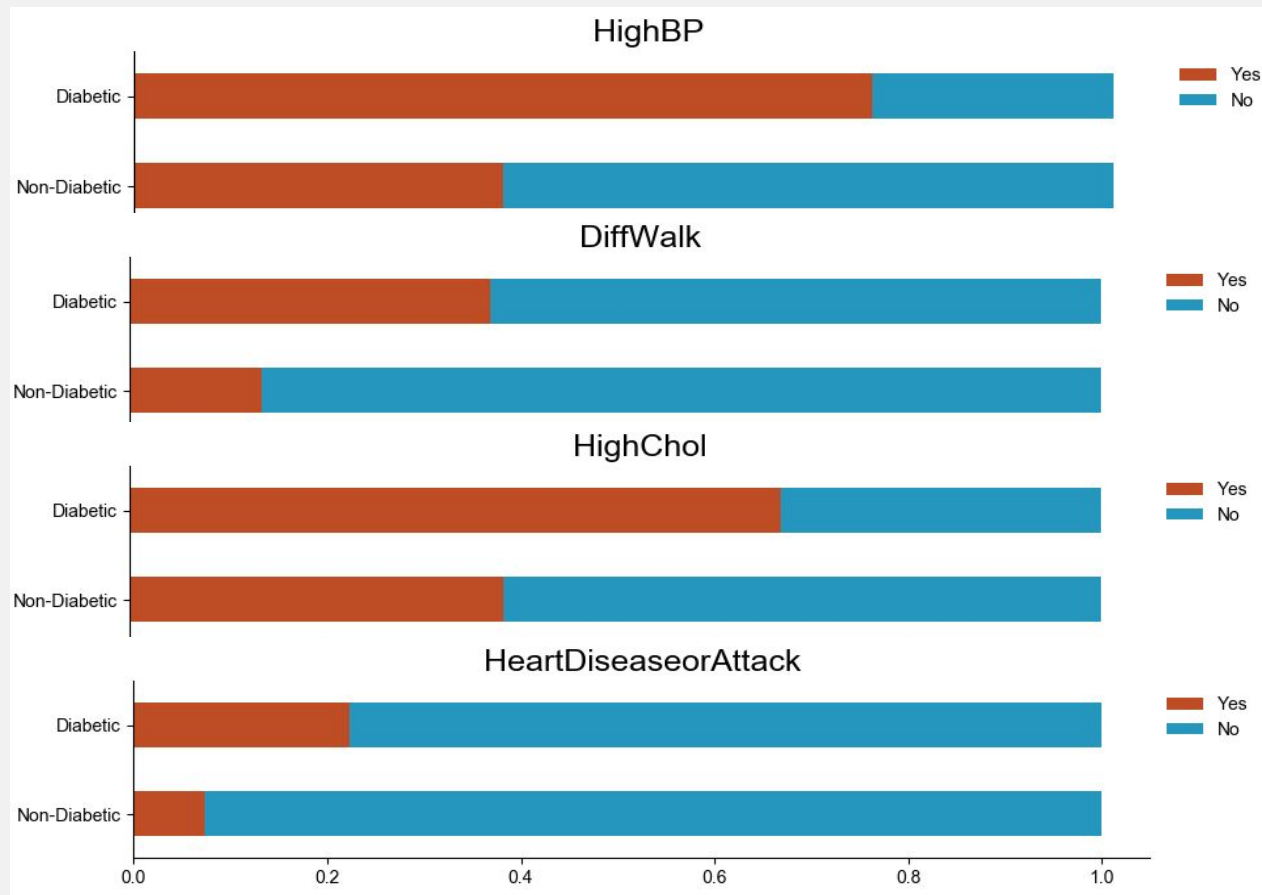
正相關

➤ 0.26312

➤ 0.21834

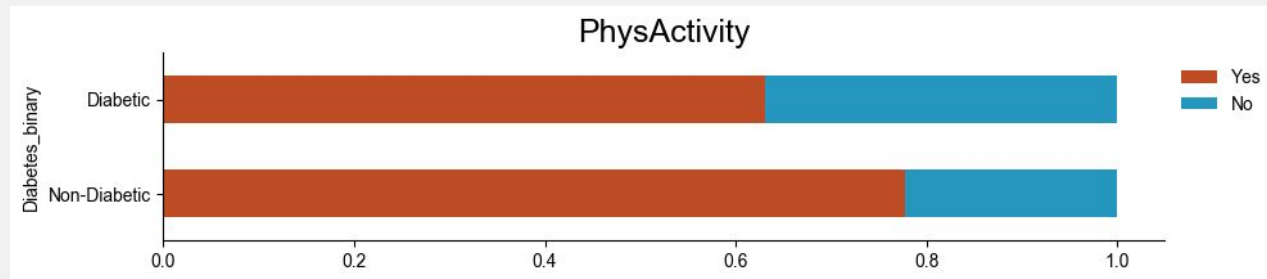
➤ 0.20027

➤ 0.17728



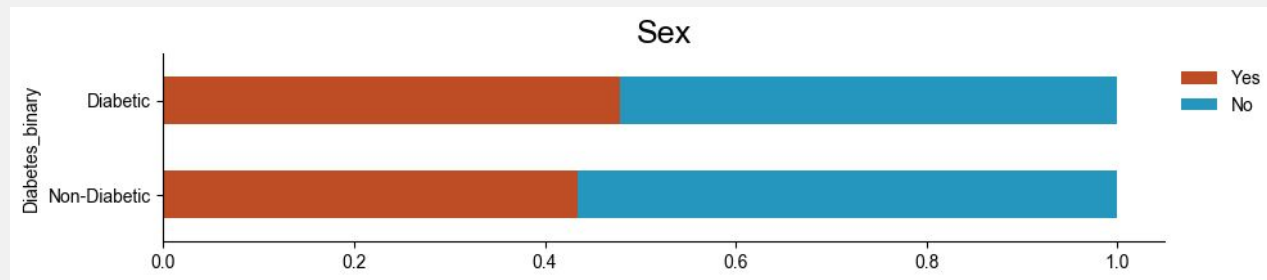
負相關

➤ -0.26312



無相關

➤ 0.03143



卡方檢定

顯著的特徵:

DiffWalk
HighBP
HeartDiseaseorAttack
HighChol
Stroke
PhysActivity
HvyAlcoholConsump
Smoker
NoDocbcCost
Fruits
Veggies
Sex
CholCheck

非顯著的特徵:

AnyHealthcare

	ChiSqr_Score	ChiSqr_pValue
DiffWalk	10059.51	0.00e+00
HighBP	10029.01	0.00e+00
HeartDiseaseorAttack	7221.98	0.00e+00
HighChol	5859.71	0.00e+00
Stroke	2725.23	0.00e+00
PhysActivity	861.89	1.89e-189
HvyAlcoholConsump	779.42	1.61e-171
Smoker	521.98	1.57e-115
NoDocbcCost	229.54	7.50e-52
Fruits	154.29	2.00e-35
Veggies	153.17	3.52e-35
Sex	140.25	2.35e-32
CholCheck	39.72	2.94e-10
AnyHealthcare	3.28	0.07

卡方檢定

顯著的特徵 ($p < 0.05$):

HeartDiseaseorAttack

DiffWalk

HighBP

HighChol

非顯著的特徵:

Fruits

Smoker

HvyAlcoholConsump

NoDocbcCost

Sex

Veggies

PhysActivity

CholCheck

AnyHealthcare

Stroke

Random sample 100

	ChiSqr_Score	ChiSqr_pValue
HeartDiseaseorAttack	3.07e+00	0.08
DiffWalk	2.43e+00	0.12
HighBP	1.98e+00	0.16
HighChol	1.98e+00	0.16
Fruits	7.13e-01	0.40
Smoker	3.73e-01	0.54
HvyAlcoholConsump	3.03e-01	0.58
NoDocbcCost	3.03e-01	0.58
Sex	4.74e-02	0.83
Veggies	2.11e-02	0.88
PhysActivity	1.57e-02	0.90
CholCheck	5.57e-03	0.94
AnyHealthcare	0.00e+00	1.00
Stroke	NAN	NAN

卡方檢定

顯著的特徵 ($p < 0.05$):

Stroke
HighChol
HighBP
DiffWalk
NoDocbcCost
HeartDiseaseorAttack

非顯著的特

徵: HvyAlcoholConsump
Fruits
Smoker
PhysActivity
Veggies
Sex
CholCheck
AnyHealthcare

Random sample 300

	ChiSqr_Score	ChiSqr_pValue
Stroke	1.25e+01	4.17e-04
HighChol	1.01e+01	1.46e-03
HighBP	9.76e+00	1.78e-03
DiffWalk	6.93e+00	8.46e-03
NoDocbcCost	5.45e+00	1.95e-02
HeartDiseaseorAttack	4.50e+00	3.39e-02
HvyAlcoholConsump	8.51e-01	3.56e-01
Fruits	5.66e-01	4.52e-01
Smoker	3.15e-01	5.75e-01
PhysActivity	8.25e-02	7.74e-01
Veggies	4.14e-02	8.39e-01
Sex	3.75e-02	8.46e-01
CholCheck	2.80e-02	8.67e-01
AnyHealthcare	5.05e-03	9.43e-01

邏輯回歸

BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income

Feature: BMI

Logit Regression Results

```

=====
Dep. Variable:    Diabetes_binary    No. Observations:    253680
Model:            Logit              Df Residuals:        253678
Method:           MLE                Df Model:            1
Date:            Mon, 10 Jun 2024    Pseudo R-squ.:      0.04904
Time:            17: 35: 04          Log-Likelihood:     -97401.
converged:        True               LL-Null:            -1.0242e+05
Covariance Type: nonrobust          LLR p-value:         0.000
=====

```

```

=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const         -4.1406      0.025    -162.688      0.000     -4.190     -4.091
BMI             0.0786      0.001     96.505      0.000      0.077      0.080
=====

```

Feature	COEF	PVALUE
BMI	0.0786	0
GenHlth	0.7875	0
MentHlth	0.0231	0
PhysHlth	0.0441	0
Age	0.1887	0
Education	-0.3393	0
Income	-0.2071	0

- BMI, GenHlth, PhysHlth, Age, Education, Income
 - 影響非常顯著
- MentHlth
 - 有一定的預測作用, 但相較於其他變數效應較小

邏輯回歸

BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income

Feature: BMI

Logit Regression Results

```

=====
Dep. Variable:    Diabetes_binary    No. Observations:    100
Model:            Logit              Df Residuals:         98
Method:            MLE                Df Model:              1
Date:             Tue, 11 Jun 2024    Pseudo R-squ.:       0.06337
Time:             15:34:42            Log-Likelihood:      -34.367
converged:        True                LL-Null:              -36.692
Covariance Type:  nonrobust           LLR p-value:         0.03104
=====

```

```

=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const         -5.2585        1.592      -3.303      0.001      -8.378      -2.139
BMI             0.1128         0.052       2.182      0.029       0.011       0.214
=====

```

- BMI, GenHlth, MentHlth
 - 顯著差異
- PhysHlth, Age, Education, Income
 - 沒有顯著差異

>0.1 <0.05

Feature	COEF	PVALUE
BMI	0.1128	0.029
GenHlth	1.0771	0.006
MentHlth	0.0840	0.026
PhysHlth	0.0076	0.878
Age	0.2357	0.062
Education	-0.0899	0.787
Income	-0.1211	0.423

Random sample 100

邏輯回歸

BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income

<0.05

Logit Regression Results

```

=====
Dep. Variable:    Diabetes_binary    No. Observations:    300
Model:            Logit              Df Residuals:         298
Method:           MLE                Df Model:             1
Date:             Tue, 11 Jun 2024    Pseudo R-squ.:       0.01609
Time:             16: 07: 22          Log-Likelihood:      -121.31
converged:        True               LL-Null:             -123.29
Covariance Type:  nonrobust          LLR p-value:         0.04641
=====
               coef    std err          z      P>|z|      [0.025    0.975]
-----
const        -2.9608     0.607    -4.881     0.000    -4.150    -1.772
BMI           0.0403     0.020     2.054     0.040     0.002     0.079
=====

```

Feature	COEF	PVALUE
BMI	0.0403	0.04
GenHlth	0.8144	0
MentHlth	0.0354	0.033
PhysHlth	0.0635	0
Age	0.2032	0.002
Education	-0.1249	0.464
Income	-0.2087	0.005

Random sample 300

- BMI, GenHlth, MentHlth, PhysHlth, Age, Income
 - 顯著差異
- Education
 - 沒有顯著差異

邏輯回歸—預測糖尿病

Optimization terminated successfully.

Current function value: 0.355211

Iterations 7

Logit Regression Results

```
=====
Dep. Variable:    Diabetes_binary    No. Observations:    253680
Model:            Logit              Df Residuals:        253675
Method:           MLE                Df Model:           4
Date:             Sat, 08 Jun 2024    Pseudo R-squ.:       0.1202
Time:             17:28:27            Log-Likelihood:       -90110.
converged:        True                LL-Null:             -1.0242e+05
Covariance Type:  nonrobust           LLR p-value:          0.000
=====
```

	coef	std err	z	P> z	[0.025	0.975]
const	-4.9879	0.044	-113.165	0.000	-5.074	-4.901
BMI	0.0832	0.001	95.691	0.000	0.082	0.085
Age	0.2052	0.002	87.138	0.000	0.201	0.210
PhysActivity	-0.3094	0.013	-23.421	0.000	-0.335	-0.284
Income	-0.1495	0.003	-52.909	0.000	-0.155	-0.144

```
=====
```

Confusion Matrix:

```
[[216158  2176]
 [ 33622  1724]]
```

Classification Report:

	precision	recall	f1-score	support
0	0.87	0.99	0.92	218334
1	0.44	0.05	0.09	35346
accuracy			0.86	253680
macro avg	0.65	0.52	0.51	253680
weighted avg	0.81	0.86	0.81	253680

feature VIF

0	const	45.772267
1	BMI	1.031343
2	Age	1.025130
3	PhysActivity	1.065332
4	Income	1.060973

Cochran-Armitage趨勢檢定

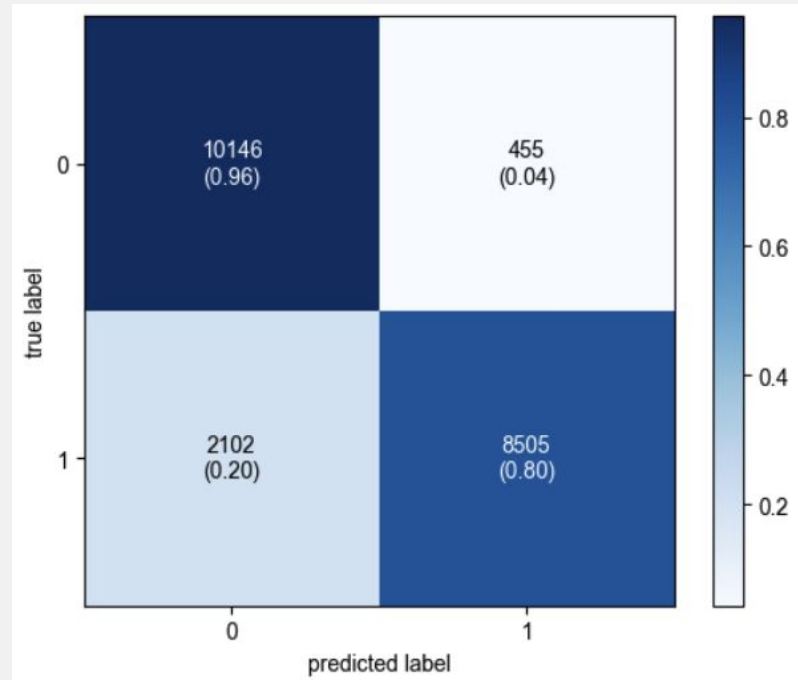
	GenHlth	MentHlth	PhysHlth	Age	Education	Income
Null_mean (H0 expected)	53421.67	112568.96	149940.58	248557.29	143166.63	178634.26
null_sd (H0 expected)	186.36	1292.92	1520.56	532.71	171.94	361.24
p-value	0.0	4.95e-267	0.0	0.0	0.0	0.0
Statistic (actual value)	80977.0	157707.0	281159.0	296166.0	132389.0	148810.0
z-score (correlation direction)	147.86	34.91	86.30	89.37	-62.68	-82.56

Data Mining Techniques

Classification

- Decision tree model
- Train-test: 0.7-0.3
- Depth: 12
- 特徴取用: BMI, GenHlth, MentHlth, PhysHlth, Age, Education, Income, HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, PhysActivity, Veggies, HvyAlcoholConsump, DiffWalk

Confusion matrix



Association rules

有糖尿病

min_sup = 0.7, min_conf = 0.7

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(CholCheck)	(AnyHealthcare)	0.993182	0.959769	0.954705	0.961259	1.001552	0.001480	1.038459	0.227327
1	(AnyHealthcare)	(CholCheck)	0.959769	0.993182	0.954705	0.994723	1.001552	0.001480	1.292203	0.038527
2	(AnyHealthcare)	(HighBP)	0.959769	0.752674	0.723844	0.754186	1.002009	0.001451	1.006152	0.049841
3	(HighBP)	(AnyHealthcare)	0.752674	0.959769	0.723844	0.961697	1.002009	0.001451	1.050345	0.008107
4	(AnyHealthcare)	(Veggies)	0.959769	0.756408	0.728060	0.758578	1.002869	0.002083	1.008988	0.071102
5	(Veggies)	(AnyHealthcare)	0.756408	0.959769	0.728060	0.962522	1.002869	0.002083	1.073465	0.011743
6	(CholCheck)	(HighBP)	0.993182	0.752674	0.748628	0.753767	1.001453	0.001086	1.004442	0.212806
7	(HighBP)	(CholCheck)	0.752674	0.993182	0.748628	0.994625	1.001453	0.001086	1.268492	0.005867
8	(CholCheck)	(Veggies)	0.993182	0.756408	0.751514	0.756673	1.000350	0.000263	1.001088	0.051317
9	(Veggies)	(CholCheck)	0.756408	0.993182	0.751514	0.993529	1.000350	0.000263	1.053725	0.001436
10	(CholCheck, AnyHealthcare)	(HighBP)	0.954705	0.752674	0.720647	0.754838	1.002875	0.002066	1.008828	0.063298
11	(CholCheck, HighBP)	(AnyHealthcare)	0.748628	0.959769	0.720647	0.962624	1.002975	0.002137	1.076389	0.011799
12	(AnyHealthcare, HighBP)	(CholCheck)	0.723844	0.993182	0.720647	0.995583	1.002418	0.001738	1.543774	0.008735
13	(CholCheck)	(AnyHealthcare, HighBP)	0.993182	0.723844	0.720647	0.725595	1.002418	0.001738	1.006379	0.353799
14	(AnyHealthcare)	(CholCheck, HighBP)	0.959769	0.748628	0.720647	0.750855	1.002975	0.002137	1.008939	0.073723
15	(HighBP)	(CholCheck, AnyHealthcare)	0.752674	0.954705	0.720647	0.957450	1.002875	0.002066	1.064514	0.011592
16	(CholCheck, AnyHealthcare)	(Veggies)	0.954705	0.756408	0.724382	0.758749	1.003095	0.002235	1.009705	0.068127
17	(CholCheck, Veggies)	(AnyHealthcare)	0.751514	0.959769	0.724382	0.963897	1.004301	0.003102	1.114340	0.017235
18	(AnyHealthcare, Veggies)	(CholCheck)	0.728060	0.993182	0.724382	0.994948	1.001779	0.001286	1.349711	0.006529
19	(CholCheck)	(AnyHealthcare, Veggies)	0.993182	0.728060	0.724382	0.729355	1.001779	0.001286	1.004785	0.260416
20	(AnyHealthcare)	(CholCheck, Veggies)	0.959769	0.751514	0.724382	0.754746	1.004301	0.003102	1.013179	0.106451
21	(Veggies)	(CholCheck, AnyHealthcare)	0.756408	0.954705	0.724382	0.957660	1.003095	0.002235	1.069796	0.012668

antecedents_len > 1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	antecedents_len
10	(CholCheck, AnyHealthcare)	(HighBP)	0.954705	0.752674	0.720647	0.754838	1.002875	0.002066	1.008828	0.063298	2
11	(CholCheck, HighBP)	(AnyHealthcare)	0.748628	0.959769	0.720647	0.962624	1.002975	0.002137	1.076389	0.011799	2
12	(AnyHealthcare, HighBP)	(CholCheck)	0.723844	0.993182	0.720647	0.995583	1.002418	0.001738	1.543774	0.008735	2
16	(CholCheck, AnyHealthcare)	(Veggies)	0.954705	0.756408	0.724382	0.758749	1.003095	0.002235	1.009705	0.068127	2
17	(CholCheck, Veggies)	(AnyHealthcare)	0.751514	0.959769	0.724382	0.963897	1.004301	0.003102	1.114340	0.017235	2
18	(AnyHealthcare, Veggies)	(CholCheck)	0.728060	0.993182	0.724382	0.994948	1.001779	0.001286	1.349711	0.006529	2

min_sup = 0.7, min_conf = 0.7

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(CholCheck)	(AnyHealthcare)	0.957730	0.949641	0.914800	0.955176	1.005828	0.005300	1.123465	0.137070
1	(AnyHealthcare)	(CholCheck)	0.949641	0.957730	0.914800	0.963311	1.005828	0.005300	1.152128	0.115054
2	(AnyHealthcare)	(PhysActivity)	0.949641	0.776943	0.741474	0.780794	1.004957	0.003657	1.017568	0.097943
3	(PhysActivity)	(AnyHealthcare)	0.776943	0.949641	0.741474	0.954349	1.004957	0.003657	1.103110	0.022112
4	(AnyHealthcare)	(Veggies)	0.949641	0.820326	0.781669	0.823121	1.003407	0.002654	1.015801	0.067425
5	(Veggies)	(AnyHealthcare)	0.820326	0.949641	0.781669	0.952877	1.003407	0.002654	1.068659	0.018898
6	(CholCheck)	(PhysActivity)	0.957730	0.776943	0.745124	0.778011	1.001375	0.001023	1.004813	0.032487
7	(PhysActivity)	(CholCheck)	0.776943	0.957730	0.745124	0.959047	1.001375	0.001023	1.032158	0.006156
8	(CholCheck)	(Veggies)	0.957730	0.820326	0.786451	0.821162	1.001019	0.000801	1.004674	0.024081
9	(Veggies)	(CholCheck)	0.820326	0.957730	0.786451	0.958706	1.001019	0.000801	1.023633	0.005665
10	(CholCheck, AnyHealthcare)	(PhysActivity)	0.914800	0.776943	0.714790	0.781362	1.005688	0.004043	1.020213	0.066385
11	(CholCheck, PhysActivity)	(AnyHealthcare)	0.745124	0.949641	0.714790	0.959290	1.010160	0.007189	1.236999	0.039461
12	(AnyHealthcare, PhysActivity)	(CholCheck)	0.741474	0.957730	0.714790	0.964012	1.006560	0.004658	1.174573	0.025208
13	(CholCheck)	(AnyHealthcare, PhysActivity)	0.957730	0.741474	0.714790	0.746338	1.006560	0.004658	1.019175	0.154175
14	(AnyHealthcare)	(CholCheck, PhysActivity)	0.949641	0.745124	0.714790	0.752695	1.010160	0.007189	1.030612	0.199723
15	(PhysActivity)	(CholCheck, AnyHealthcare)	0.776943	0.914800	0.714790	0.920004	1.005688	0.004043	1.065047	0.025357
16	(CholCheck, AnyHealthcare)	(Veggies)	0.914800	0.820326	0.753561	0.823744	1.004167	0.003127	1.019393	0.048703
17	(CholCheck, Veggies)	(AnyHealthcare)	0.786451	0.949641	0.753561	0.958179	1.008991	0.006715	1.204154	0.041726
18	(AnyHealthcare, Veggies)	(CholCheck)	0.781669	0.957730	0.753561	0.964041	1.006589	0.004933	1.175497	0.029983
19	(CholCheck)	(AnyHealthcare, Veggies)	0.957730	0.781669	0.753561	0.786820	1.006589	0.004933	1.024161	0.154865
20	(AnyHealthcare)	(CholCheck, Veggies)	0.949641	0.786451	0.753561	0.793522	1.008991	0.006715	1.034244	0.176941
21	(Veggies)	(CholCheck, AnyHealthcare)	0.820326	0.914800	0.753561	0.918612	1.004167	0.003127	1.046834	0.023094

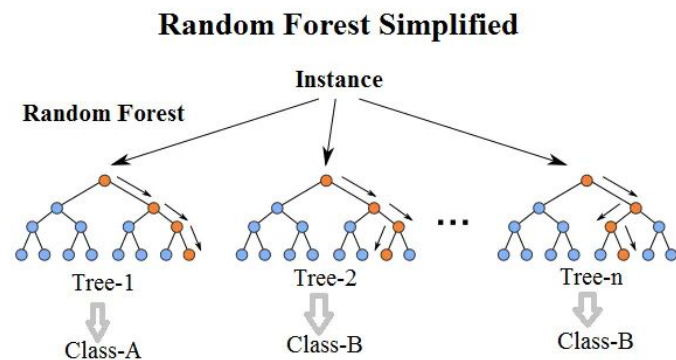
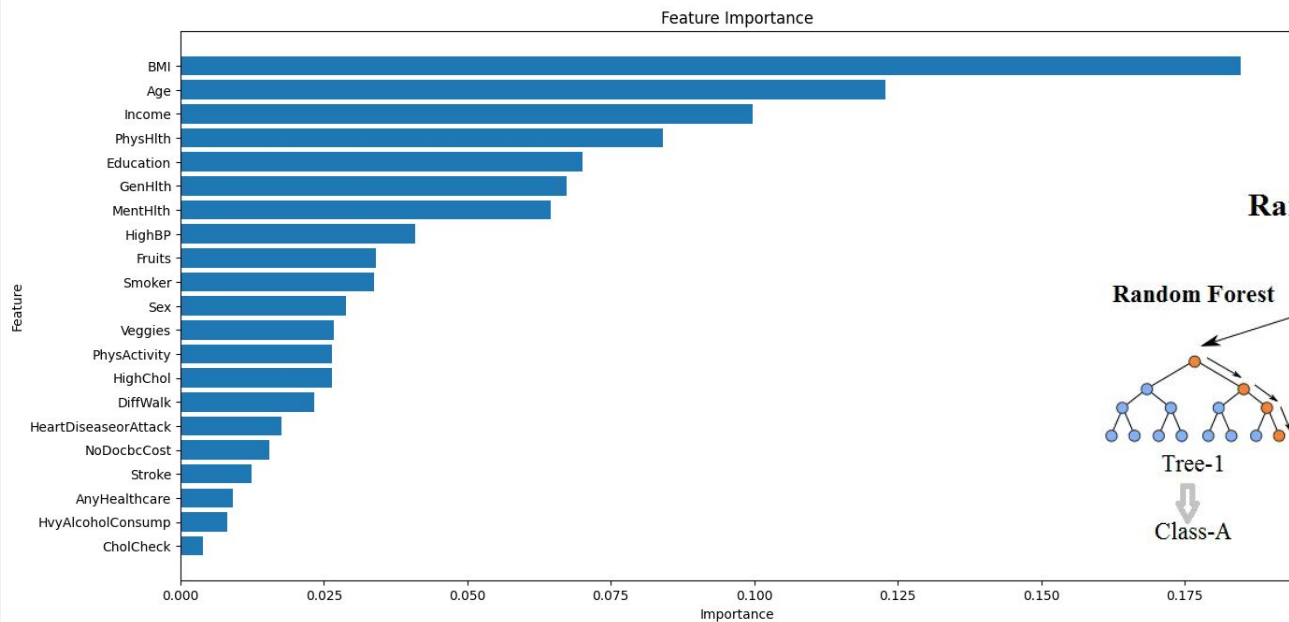
antecedents_len > 1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric	antecedents_len
10	(CholCheck, AnyHealthcare)	(PhysActivity)	0.914800	0.776943	0.714790	0.781362	1.005688	0.004043	1.020213	0.066385	2
11	(CholCheck, PhysActivity)	(AnyHealthcare)	0.745124	0.949641	0.714790	0.959290	1.010160	0.007189	1.236999	0.039461	2
12	(AnyHealthcare, PhysActivity)	(CholCheck)	0.741474	0.957730	0.714790	0.964012	1.006560	0.004658	1.174573	0.025208	2
16	(CholCheck, AnyHealthcare)	(Veggies)	0.914800	0.820326	0.753561	0.823744	1.004167	0.003127	1.019393	0.048703	2
17	(CholCheck, Veggies)	(AnyHealthcare)	0.786451	0.949641	0.753561	0.958179	1.008991	0.006715	1.204154	0.041726	2
18	(AnyHealthcare, Veggies)	(CholCheck)	0.781669	0.957730	0.753561	0.964041	1.006589	0.004933	1.175497	0.029983	2

特徵選擇

原始Random Forest模型

➤ 模型特徵重要性排序



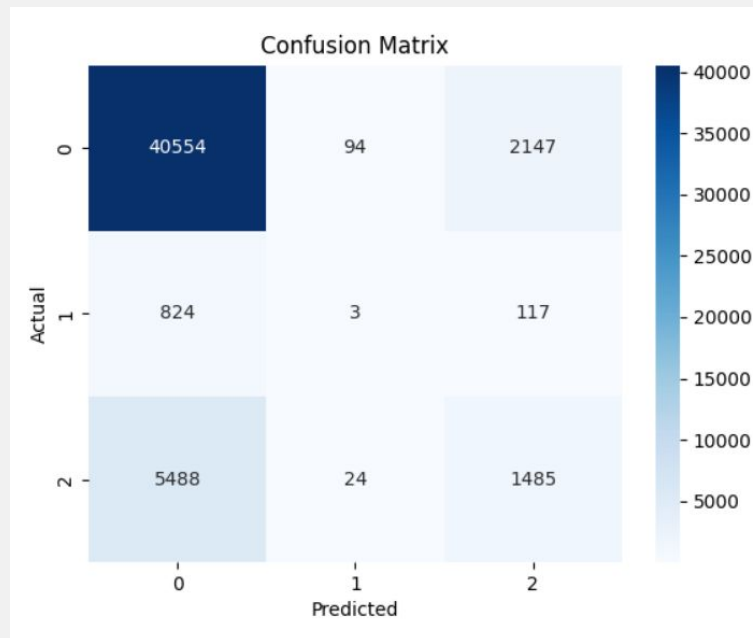
選擇前十個特徵並評估

Confusion matrix

Accuracy: 0.8286423841059603

F1 Score: 0.8012423971371091

	precision	recall	f1-score	support
0	0.87	0.95	0.90	42795
1	0.02	0.00	0.01	944
2	0.40	0.21	0.28	6997
macro avg	0.43	0.39	0.40	50736
weighted avg	0.78	0.83	0.80	50736



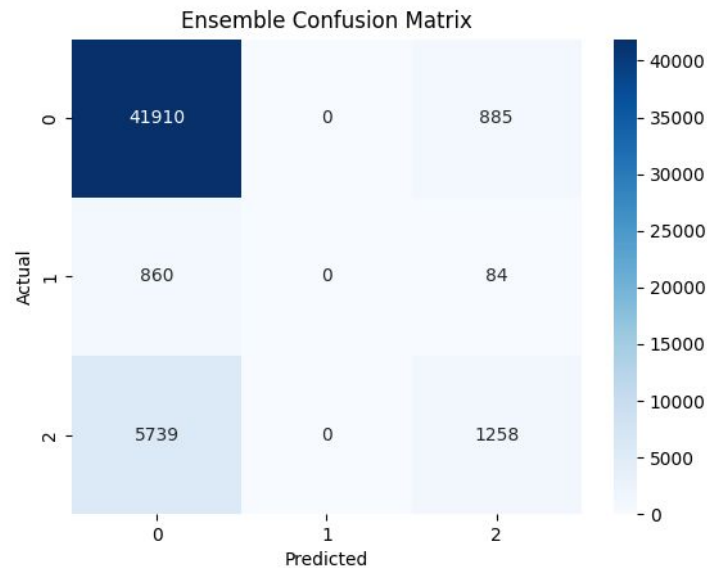
嘗試模型集成改良

Random Forest / Logistic Regression / Gradient Boosting

Ensemble Accuracy: 0.8508356985178177

Ensemble F1 Score: 0.8119625305245382

	precision	recall	f1-score	support
0	0.86	0.98	0.92	42795
1	0.00	0.00	0.00	944
2	0.56	0.18	0.27	6997
macro avg	0.48	0.39	0.40	50736
weighted avg	0.81	0.85	0.81	50736





Reference

- Dataset
 - CDC Diabetes Health Indicators
 - <https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators>
- Tools
 - chatGPT
 - Python
 - fetch_ucirepo
 - sklearn
 - scipy
 - statsmodels
 - mlxtend

Q&A