- 預測「10年內是否患有心臟病」

戴偉哲(105830555)

內容

- 主題選擇與預測項目
- 資料集欄位敘述
- 第一次用KNN建模型
- 第二次用KNN建模型-對數值型資料做相關係數的加權
- 使用Naïve Bayes建模型-Gauss, Bernoulli
- 使用Decision Tree建模型
- 結果統整

主題選擇與預測項目

- 從Koggle找與醫療照護產業相關的資料集，原因為組員背景皆有健康相關。

- 選擇心臟病為主題原因為觀察112年十大死因當中，前三名依序為惡性腫瘤、心臟病、肺炎，與111年相較，癌症及心臟病續居前兩名。

- 選擇預測項目「10年內是否患有心臟病」原因為觀察各心臟病相關資料集，發現有些臨床生理數據異常會成為心臟病的潛在因子，因此選擇未來是否會患有心臟病資料的資料集為主要資料集並且以此為預測項目

- 觀察①病患生活背景與習慣②臨床的生理監測數據資料集，可以發現有許多資料欄位是心臟病的風險因子，可用來預測「10年內是否患有心臟病」

資料集欄位敘述

- 背景資料(Background)

1. Sex: Gender of the patient, male or female ("M" or "F")

1. Age: Age of the patient (Number ，Discrete)

1. Education: The level of education of the patient
   (categorical values - 1,2,3,4)

資料集欄位敘述

- 行為資料(Behavior):

4. is_smoking(是否吸菸): Whether or not the patient is a current smoker ("YES" or "NO")

4. CigsPerDay(每天吸菸數): The number of cigarettes that the person smoked on average in one day. (Number，Discrete)

資料集欄位敘述

- 病史(History):

6. BPMeds(Blood Pressure Medication，是否正在服用降血壓藥物): Whether or not the patient was on blood pressure medication ("NO":0,"YES":1)
7. Prevalent Stroke(有無中風病史): Whether or not the patient had previously had a stroke ("NO":0,"YES":1)
8. Prevalent Hyp(有無高血壓病史): Whether or not the patient was hypertensive ("NO":0,"YES":1)
9. Diabetes(有無糖尿病): Whether or not the patient had diabetes ("NO":0,"YES":1)

資料集欄位敘述

- 目前臨床生理數據 (Current):

10. TotChol(總膽固醇，mg/dL): Total cholesterol level (Continuous)
11. Sys BP(收縮壓): Systolic blood pressure (Continuous)
12. Dia BP(舒張壓): Diastolic blood pressure (Continuous)
13. BMI(身體質量指數): Body Mass Index (Continuous)
14. HR(心率): Heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
15. Glucose(血糖): Glucose level，Blood sugar (Continuous)

資料集欄位敘述

- 預測項目 (Desired Target):

16. TenYearCHD(10年冠狀血管心臟病風險):10-year risk of coronary heart disease ("NO":0,"YES":1)

第一次用KNN建模型

原始資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	F	YES	3.0	0.0	0	0	0	221.0	148.0	85.0	NaN	90.0	80.0	1
1	1	36	4.0	M	NO	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.0	0
2	2	46	1.0	F	YES	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.0	0
3	3	50	1.0	M	YES	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.0	1
4	4	64	1.0	F	YES	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.0	0
5	5	61	3.0	F	NO	0.0	0.0	0	1	0	272.0	182.0	121.0	32.80	85.0	65.0	1
6	6	61	1.0	M	NO	0.0	0.0	0	1	0	238.0	232.0	136.0	24.83	75.0	79.0	0
7	7	36	4.0	M	YES	35.0	0.0	0	0	0	295.0	102.0	68.0	28.15	60.0	63.0	0
8	8	41	2.0	F	YES	20.0	NaN	0	0	0	220.0	126.0	78.0	20.70	86.0	79.0	0
9	9	55	2.0	F	NO	0.0	0.0	0	1	0	326.0	144.0	81.0	25.71	85.0	NaN	0

建立KNN模型資料前處理(1/2)

	age	education	sex	is_smoking	cigsPerDay	BPMeds	pre_stroke	pre_hyp
空值處理	無	平均值	無	無	平均值	眾數	無	無
正規化	Z-Score	X	X	X	極值 正規化	X	X	X
資料離散	X	X	X	X	X	X	X	X
型態轉換	X	X	"M":1, "F":2	"NO":0, "YES":1	X	X	X	X

建立KNN模型資料前處理(2/2)

	diabetes	totChol (mg/dL)	sysBP (mmHg)	diaBP (mmHg)	BMI (kg/(m**2)	HR (bpm)	Glucose (mg/dL)	TenYearCHD
空值處理	無	平均值	無	無	平均值	平均值	平均值	無
正規化	X	X	X	X	X	X	X	X
資料離散	X	<150過低 150<=正常<200 >=200過高	<90過低 90<=正常<130 >=130過高	<60過低 60<=正常<80 >=80過高	<18.5過輕 18.5<=正常<24 >=24過重	<60過慢 60<=正常<100 >=100過快	<70過低 70<=正常<100 >=100過高	X
型態轉換	X	"過低":1, "正常":2, "過高":3	"過低":1, "正常":2, "過高":3	"過低":1, "正常":2, "過高":3	"過低":1, "正常":2, "過高":3	"過低":1, "正常":2, "過高":3	"過低":1, "正常":2, "過高":3	X

處理後資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	1.682535	2.0	2	1	0.042857	0.0	0	0	0	3	3	3	3	2	2	1
1	1	-1.575978	4.0	1	0	0.000000	0.0	0	1	0	3	3	3	3	2	2	0
2	2	-0.412223	1.0	2	1	0.142857	0.0	0	0	0	3	2	2	2	2	2	0
3	3	0.053279	1.0	1	1	0.285714	0.0	0	1	0	3	3	3	3	2	2	1
4	4	1.682535	1.0	2	1	0.428571	0.0	0	0	0	3	3	3	3	2	2	0
5	5	1.333409	3.0	2	0	0.000000	0.0	0	1	0	3	3	3	3	2	1	1
6	6	1.333409	1.0	1	0	0.000000	0.0	0	1	0	3	3	3	3	2	2	0
7	7	-1.575978	4.0	1	1	0.500000	0.0	0	0	0	3	2	2	3	2	1	0
8	8	-0.994100	2.0	2	1	0.285714	0.0	0	0	0	3	2	2	2	2	2	0
9	9	0.635156	2.0	2	0	0.000000	0.0	0	1	0	3	3	3	3	2	2	0

建立模型與結果數據

- 資料前處理後，將20%資料集設定為test_data
- GridSearchCV搜尋K value最佳解
- KNN使用最佳的K value建立KnnModel模型
- 得到KnnModel模型的Accuracy與F1_score
- Knn_model.score十次平均：0.84
- F1_score十次平均：0.03

第二次用KNN建模型

對數值型資料做相關係數的加權

二次建立KNN模型資料前處理(1/2)

	age	education	sex	is_smoking	cigsPerDay	BPMeds	pre_stroke	pre_hyp
空值處理	無	平均值	無	無	平均值	眾數	無	無
正規化	Z-Score	X	X	X	極值 正規化	X	X	X
型態轉換	X	X	"M":1, "F":2	"NO":0, "YES":1	X	X	X	X

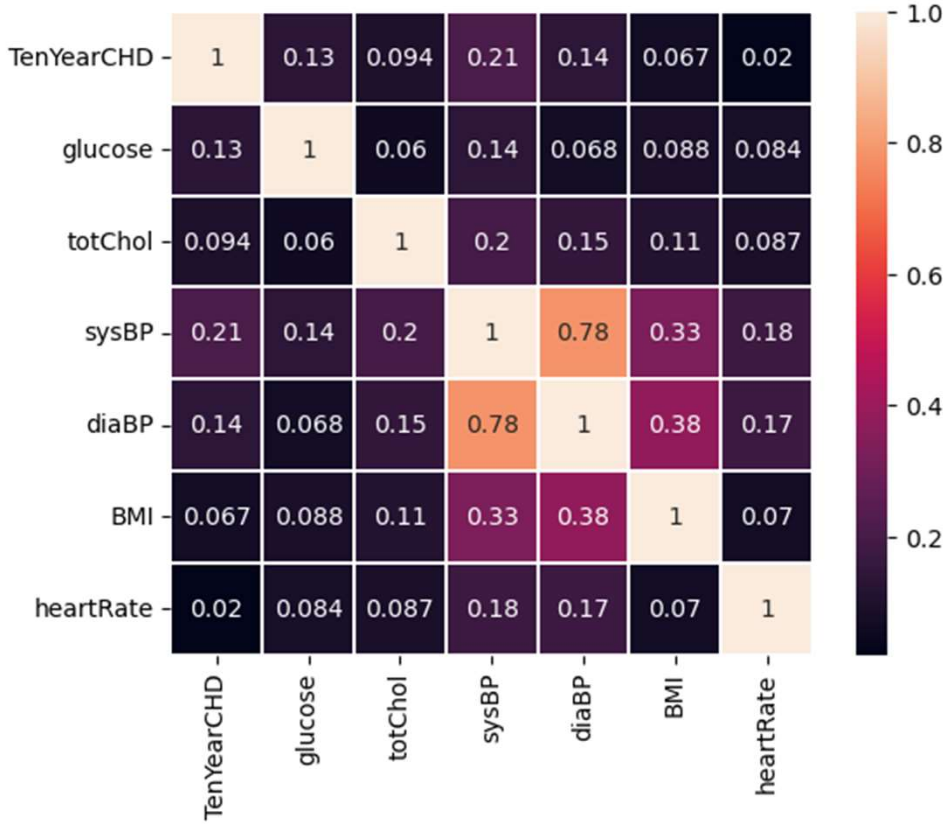
二次建立KNN模型資料前處理(2/2)

	diabetes	totChol (mg/dL)	sysBP (mmHg)	diaBP (mmHg)	BMI (kg/(m**2)	HR (bpm)	Glucose (mg/dL)	TenYearCHD
空值處理	無	平均值	無	無	平均值	平均值	平均值	無
正規化	X	X	X	X	X	X	X	X
型態轉換	X	X	X	X	X	X	X	X

處理後資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	1.682535	2.0	2	1	0.042857	0.0	0	0	0	221.0	148.0	85.0	25.79	90.0	80.00	1
1	1	-1.575978	4.0	1	0	0.000000	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.00	0
2	2	-0.412223	1.0	2	1	0.142857	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.00	0
3	3	0.053279	1.0	1	1	0.285714	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.00	1
4	4	1.682535	1.0	2	1	0.428571	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.00	0
5	5	1.333409	3.0	2	0	0.000000	0.0	0	1	0	272.0	182.0	121.0	32.80	85.0	65.00	1
6	6	1.333409	1.0	1	0	0.000000	0.0	0	1	0	238.0	232.0	136.0	24.83	75.0	79.00	0
7	7	-1.575978	4.0	1	1	0.500000	0.0	0	0	0	295.0	102.0	68.0	28.15	60.0	63.00	0
8	8	-0.994100	2.0	2	1	0.285714	0.0	0	0	0	220.0	126.0	78.0	20.70	86.0	79.00	0
9	9	0.635156	2.0	2	0	0.000000	0.0	0	1	0	326.0	144.0	81.0	25.71	85.0	82.09	0

相關係數




```
Axes(0.1675,0.11;0.5775x0.77)
id                0.009866
age               0.224927
education        -0.052074
sex              -0.084647
is_smoking        0.034143
cigsPerDay        0.066686
BPMeds            0.087349
prevalentStroke   0.068627
prevalentHyp      0.166544
diabetes          0.103681
totChol           0.093679
sysBP             0.212703
diaBP             0.135979
BMI               0.066538
heartRate         0.020224
glucose           0.132648
TenYearCHD        1.000000
Name: TenYearCHD, dtype: float64
```

加權後數據

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	1.682535	2.0	2	1	0.042857	0.0	0	0	0	241.774	179.080	96.90	27.51793	91.80	90.4000	1
1	1	-1.575978	4.0	1	0	0.000000	0.0	0	1	0	231.928	203.280	111.72	31.76459	73.44	84.7500	0
2	2	-0.412223	1.0	2	1	0.142857	0.0	0	0	0	273.500	140.360	80.94	21.71345	89.76	106.2200	0
3	3	0.053279	1.0	1	1	0.285714	0.0	0	1	0	254.902	191.180	100.32	30.15342	69.36	106.2200	1
4	4	1.682535	1.0	2	1	0.428571	0.0	0	0	0	263.654	165.165	96.90	28.19014	71.40	87.0100	0
5	5	1.333409	3.0	2	0	0.000000	0.0	0	1	0	297.568	220.220	137.94	34.99760	86.70	73.4500	1
6	6	1.333409	1.0	1	0	0.000000	0.0	0	1	0	260.372	280.720	155.04	26.49361	76.50	89.2700	0
7	7	-1.575978	4.0	1	1	0.500000	0.0	0	0	0	322.730	123.420	77.52	30.03605	61.20	71.1900	0
8	8	-0.994100	2.0	2	1	0.285714	0.0	0	0	0	240.680	152.460	88.92	22.08690	87.72	89.2700	0
9	9	0.635156	2.0	2	0	0.000000	0.0	0	1	0	356.644	174.240	92.34	27.43257	86.70	92.7617	0

預測方法

- 資料前處理後，加入相關係數加權
- 將20%資料集設定為test_data
- GridSearchCV搜尋K value最佳解
- KNN使用最佳的K value建立KnnModel模型
- 得到KnnModel模型的Accuracy與F1_score

建立模型與結果數據

- 資料前處理後，將20%資料集設定為test_data，加入相關係數
- GridSearchCV搜尋K value最佳解
- KNN使用最佳的K value建立KnnModel模型
- 得到KnnModel模型的Accuracy與F1_score
- Knn_model.score十次平均：加權前0.85，加權後0.86
- F1_score十次平均：加權前0.04，加權後0.06

使用Naïve Bayes建模型

Gauss, Bernoulli

建立Gauss模型資料前處理(1/2)

	age	education	sex	is_smoking	cigsPerDay	BPMeds	pre_stroke	pre_hyp
空值處理	無	平均值	無	無	平均值	眾數	無	無
正規化	Z-Score	X	X	X	極值 正規化	X	X	X
型態轉換	X	X	"M":1, "F":2	"NO":0, "YES":1	X	X	X	X

建立Gauss模型資料前處理(2/2)

	diabetes	totChol (mg/dL)	sysBP (mmHg)	diaBP (mmHg)	BMI (kg/(m**2)	HR (bpm)	Glucose (mg/dL)	TenYear CHD
空值處理	無	平均值	無	無	平均值	平均值	平均值	無
正規化	x	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	Z-Score	x
型態轉換	x	x	x	x	x	x	x	x

處理後資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	1.682535	2.0	2	1	0.042857	0.0	0	0	0	-0.357260	0.690777	0.176067	-0.001204	1.171479	-0.090215	1
1	1	-1.575978	4.0	1	0	0.000000	0.0	0	1	0	-0.557291	1.587959	1.257276	0.967888	-0.332268	-0.306367	0
2	2	-0.412223	1.0	2	1	0.142857	0.0	0	0	0	0.287284	-0.744714	-0.988311	-1.325791	1.004396	0.515012	0
3	3	0.053279	1.0	1	1	0.285714	0.0	0	1	0	-0.090553	1.139368	0.425577	0.600217	-0.666434	0.515012	1
4	4	1.682535	1.0	2	1	0.428571	0.0	0	0	0	0.087253	0.174897	0.176067	0.152195	-0.499351	-0.219906	0
5	5	1.333409	3.0	2	0	0.000000	0.0	0	1	0	0.776248	2.215986	3.170184	1.705663	0.753771	-0.738673	1
6	6	1.333409	1.0	1	0	0.000000	0.0	0	1	0	0.020576	4.458940	4.417732	-0.234955	-0.081643	-0.133445	0
7	7	-1.575978	4.0	1	1	0.500000	0.0	0	0	0	1.287438	-1.372741	-1.237821	0.573433	-1.334766	-0.825134	0
8	8	-0.994100	2.0	2	1	0.285714	0.0	0	0	0	-0.379486	-0.296123	-0.406122	-1.240570	0.837313	-0.133445	0
9	9	0.635156	2.0	2	0	0.000000	0.0	0	1	0	1.976434	0.511341	-0.156612	-0.020683	0.753771	0.000137	0

建立模型與結果數據

- 資料前處理後，將20%資料集設定為test_data
- 建立Gauss模型
- 得到Gauss模型的Accuracy與F1_score
- Gauss_model.score十次平均：0.82
- F1_score十次平均：0.24

建立Bernoulli模型資料前處理(1/2)

	age	education	sex	is_smoking	cigsPerDay	BPMeds	pre_stroke	pre_hyp
空值處理	無	平均值	無	無	平均值	眾數	無	無
正規化	Z-Score	X	X	X	極值 正規化	X	X	X
資料離散	X	X	X	X	X	X	X	X
型態轉換	X	X	"M":1, "F":2	"NO":0, "YES":1	X	X	X	X

建立Bernoulli模型資料前處理(2/2)

	diabetes	totChol (mg/dL)	sysBP (mmHg)	diaBP (mmHg)	BMI (kg/(m**2)	HR (bpm)	Glucose (mg/dL)	TenYear CHD
空值處理	無	平均值	無	無	平均值	平均值	平均值	無
正規化	X	X	X	X	X	X	X	X
資料離散	X	正常<200 >=200過高	正常<130 >=130過高	正常<80 >=80過高	正常<24 >=24過重	正常<100 >=100過快	正常<100 >=100過高	X
型態轉換	X	"正常":0, "過高":1	"正常":0, "過高":1	"正常":0, "過高":1	"正常":0, “過重":1	"正常":0, “過快":1	"正常":0, "過高":1	X

處理後資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	1.682535	2.0	2	1	0.042857	0.0	0	0	0	1	1	1	1	0	0	1
1	1	-1.575978	4.0	1	0	0.000000	0.0	0	1	0	1	1	1	1	0	0	0
2	2	-0.412223	1.0	2	1	0.142857	0.0	0	0	0	1	0	0	0	0	0	0
3	3	0.053279	1.0	1	1	0.285714	0.0	0	1	0	1	1	1	1	0	0	1
4	4	1.682535	1.0	2	1	0.428571	0.0	0	0	0	1	1	1	1	0	0	0
5	5	1.333409	3.0	2	0	0.000000	0.0	0	1	0	1	1	1	1	0	0	1
6	6	1.333409	1.0	1	0	0.000000	0.0	0	1	0	1	1	1	1	0	0	0
7	7	-1.575978	4.0	1	1	0.500000	0.0	0	0	0	1	0	0	1	0	0	0
8	8	-0.994100	2.0	2	1	0.285714	0.0	0	0	0	1	0	0	0	0	0	0
9	9	0.635156	2.0	2	0	0.000000	0.0	0	1	0	1	1	1	1	0	0	0

建立模型與結果數據

- 資料前處理後，將20%資料集設定為test_data
- 建立Bernoulli模型
- 得到Bernoulli模型的Accuracy與F1_score
- Bernoulli_model.score十次平均：0.82
- F1_score十次平均：0.23

使用Decision Tree建模型

建立Tree模型資料前處理(1/2)

	age	education	sex	is_smoking	cigsPerDay	BPMeds	pre_stroke	pre_hyp
空值處理	無	平均值	無	無	平均值	眾數	無	無
型態轉換	X	X	"M":1, "F":2	"NO":0, "YES":1	X	X	X	X

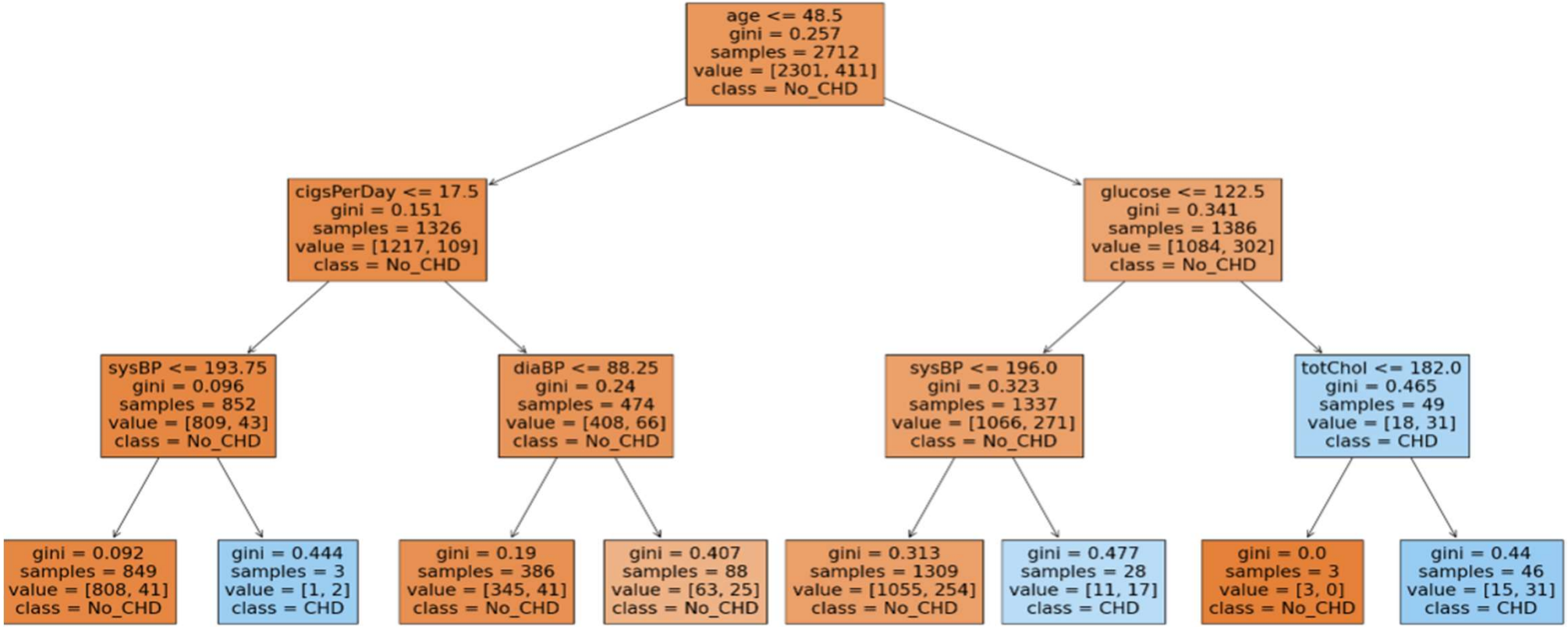
建立Tree模型資料前處理(2/2)

	diabetes	totChol (mg/dL)	sysBP (mmHg)	diaBP (mmHg)	BMI (kg/(m**2)	HR (bpm)	Glucose (mg/dL)	TenYear CHD
空值處理	無	平均值	無	無	平均值	平均值	平均值	無
型態轉換	X	X	X	X	X	X	X	X

處理後資料

	id	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
0	0	64	2.0	2	1	3.0	0.0	0	0	0	221.0	148.0	85.0	25.79	90.0	80.00	1
1	1	36	4.0	1	0	0.0	0.0	0	1	0	212.0	168.0	98.0	29.77	72.0	75.00	0
2	2	46	1.0	2	1	10.0	0.0	0	0	0	250.0	116.0	71.0	20.35	88.0	94.00	0
3	3	50	1.0	1	1	20.0	0.0	0	1	0	233.0	158.0	88.0	28.26	68.0	94.00	1
4	4	64	1.0	2	1	30.0	0.0	0	0	0	241.0	136.5	85.0	26.42	70.0	77.00	0
5	5	61	3.0	2	0	0.0	0.0	0	1	0	272.0	182.0	121.0	32.80	85.0	65.00	1
6	6	61	1.0	1	0	0.0	0.0	0	1	0	238.0	232.0	136.0	24.83	75.0	79.00	0
7	7	36	4.0	1	1	35.0	0.0	0	0	0	295.0	102.0	68.0	28.15	60.0	63.00	0
8	8	41	2.0	2	1	20.0	0.0	0	0	0	220.0	126.0	78.0	20.70	86.0	79.00	0
9	9	55	2.0	2	0	0.0	0.0	0	1	0	326.0	144.0	81.0	25.71	85.0	82.09	0

DecisionTree



建立模型與結果數據

- 資料前處理後，將20%資料集設定為test_data
- 建立DecisionTree模型
- 得到DecisionTree模型的Accuracy與F1_score
- Bernoulli_model.score十次平均：0.85
- F1_score十次平均：0.06

結果統整

	第一次KNN	第二次KNN	KNN(加權)	NB Gaussian	NB Bernoulli	Decision Tree	KNN-最終目標變數數相同
Accuracy	0.84	0.85	0.86	0.82	0.82	0.85	0.59
f1_score	0.03	0.04	0.06	0.24	0.23	0.06	0.57
	生理數據不做離散化，所建模的精準率較高			Precision與Recall結果相近(約0.20~0.25)		各項精準率等等結果 與KNN數據較為相近	目標數據數相同後所得結果， "Accuracy"與"f1_score"相近
		將相關係數較高之數據加權後建模，精準率再次提高					
	precision_score約0.3~0.5 recall_score < 0.1						
	Precision結果較高，而Recall結果極低導致"f1_score"結果低						

報告到此結束謝謝大家