

## Code book

1. Name of the data file: 中風

2. Data preprocessing:

Number	Event
1	將「序列」欄位刪除
2	刪除含缺失值的資料，共 202 筆
3	將「吸煙狀況」欄位內容為「Unknow」的資料行數刪除，共 1544 筆資料

3. Data overview:

Total sample size	3425
Independent variables (X)	
Total of categorical variables	7
Total of numeric variables	3
Sum	10
Dependent variables (Y)	
Total of categorical variables	1
Total of numeric variables	

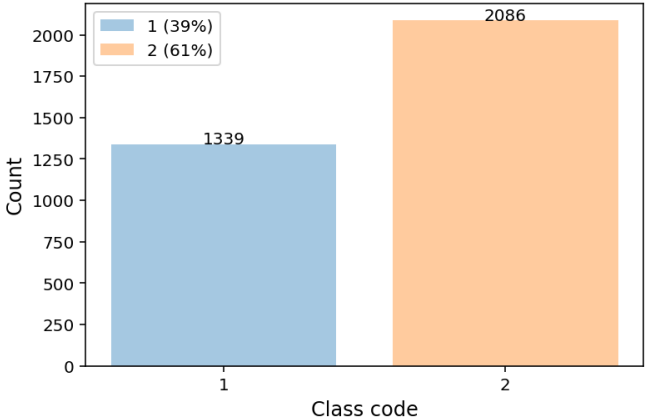
4. Variables overview:

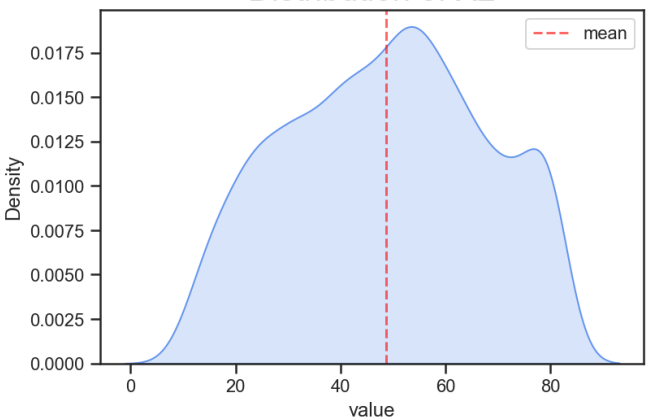
Index	Variable	Description / Unit	Miss
X1	性別	1:Male 2:Female	1
X2	年齡	Continuous	0
X3	高血壓	0:No 1:Yes	0
X4	心臟病	0:No 1:Yes	0
X5	結過婚	0:No 1:Yes	0
X6	工作類型	1:children 2:Never_worked 3:Private 4:Self-employed 5:Govt_job	0
X7	居住類型	1:Rural 2:Urban	0

X8	平均血糖水平	Continuous	0
X9	BMI	Continuous	201
X10	吸煙狀況	1:never smoked 2:formerly smoked 3:smokes	0
Y	中風	0:No 1:Yes	0

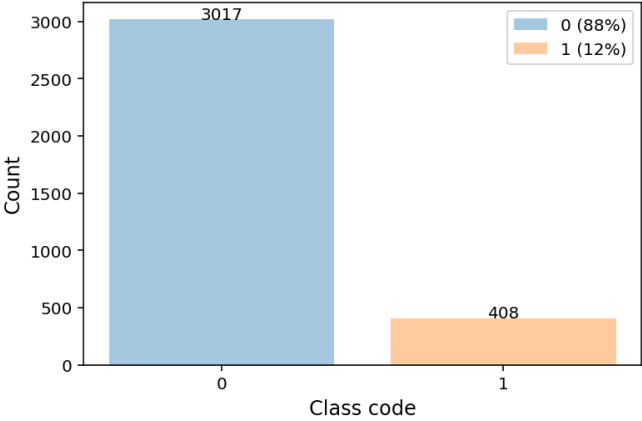
## 5. Variable description:

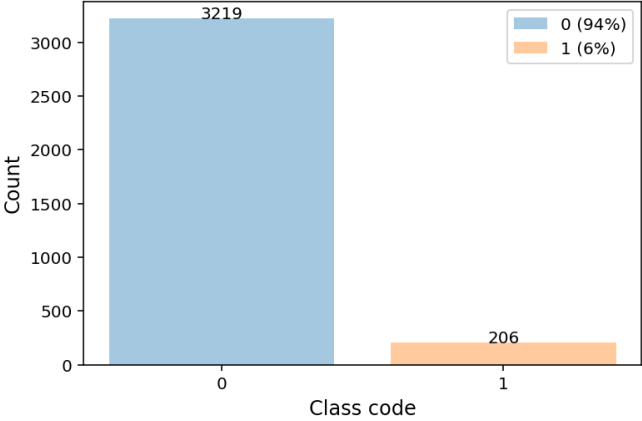
Independent variable (known as “X” or input)

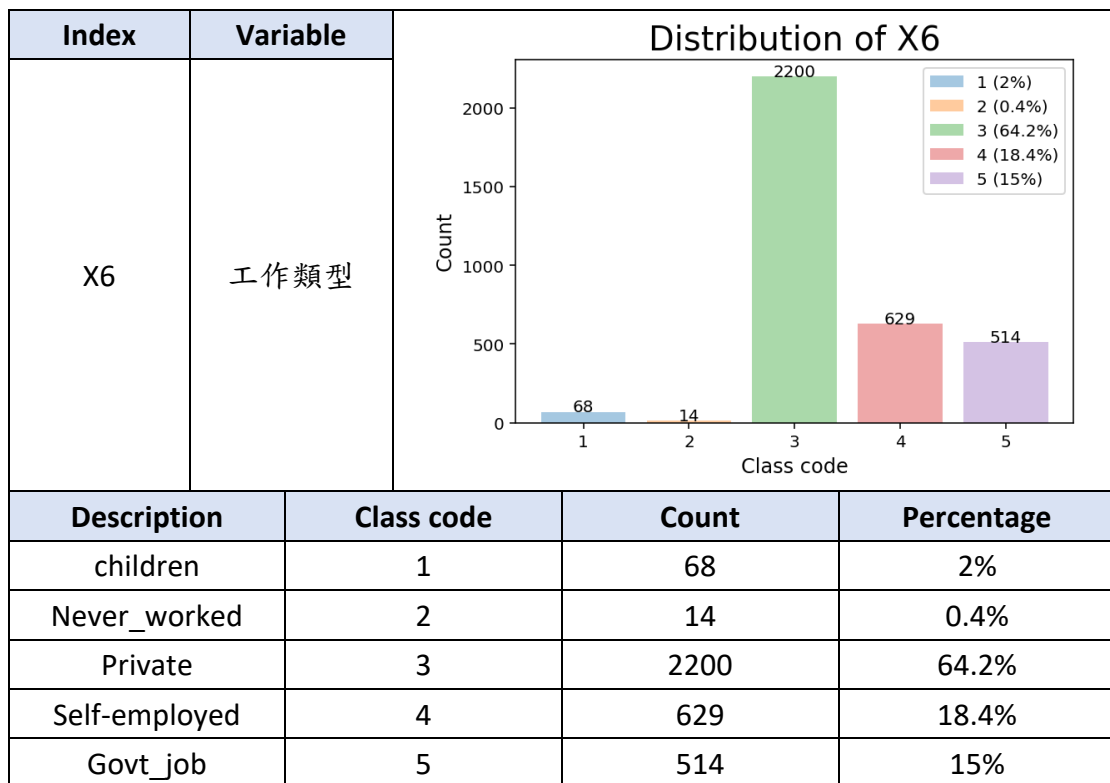
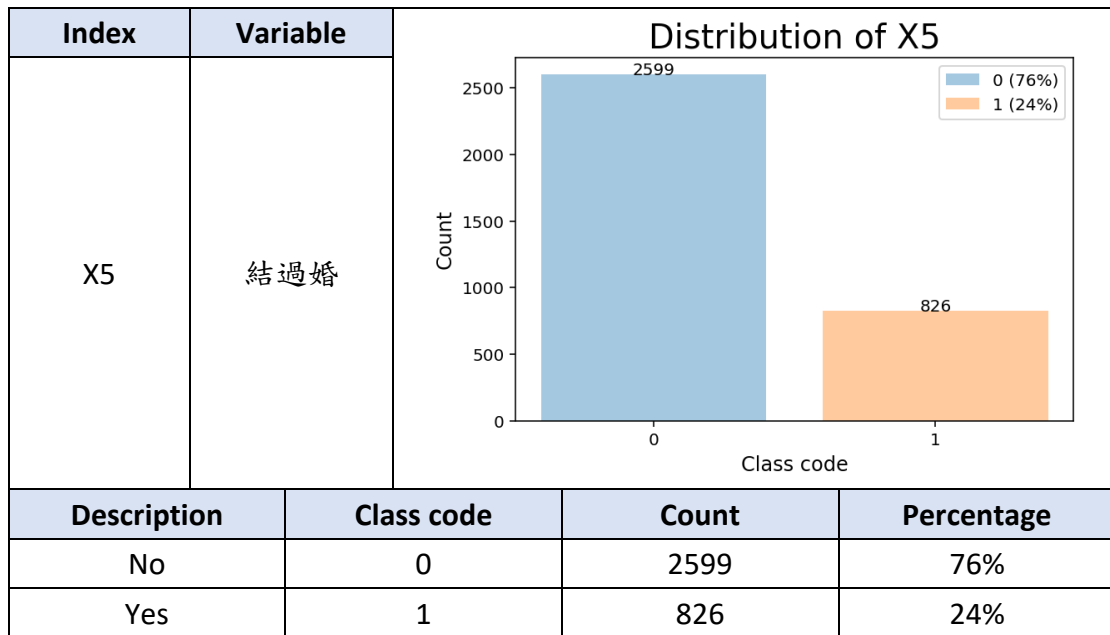
Index	Variable	Distribution of X1		
X1	性別			
Description		Class code	Count	Percentage
Male		1	1339	39%
Female		2	2086	61%

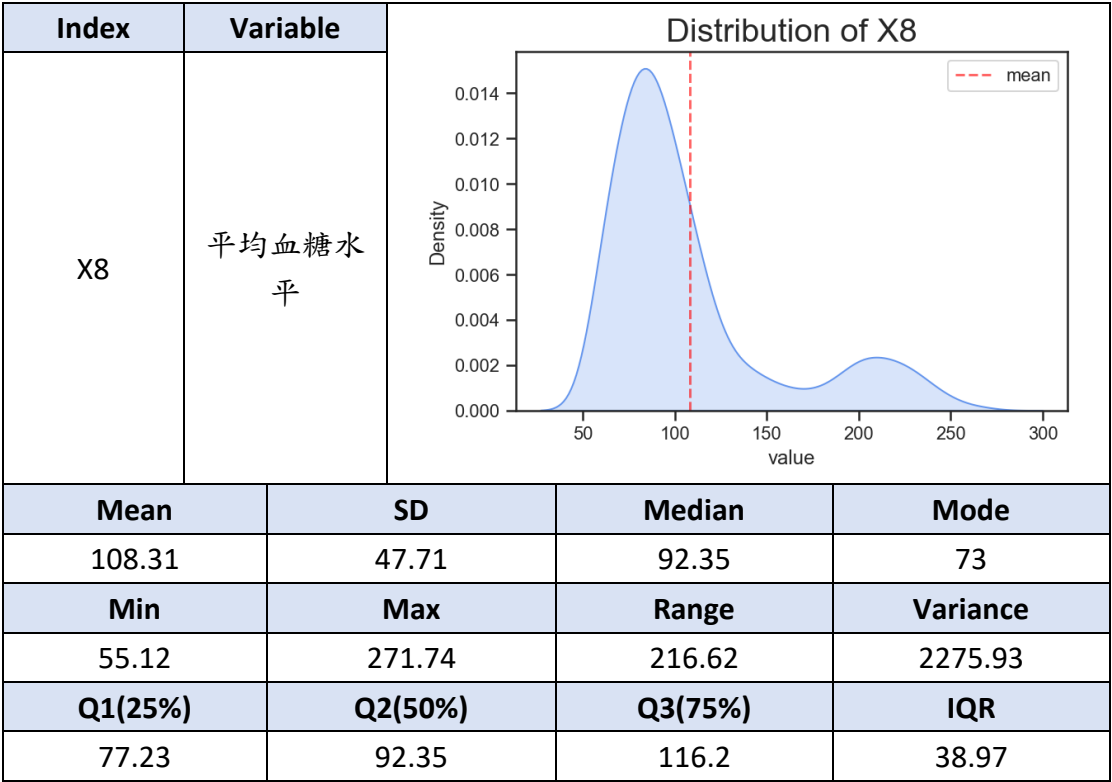
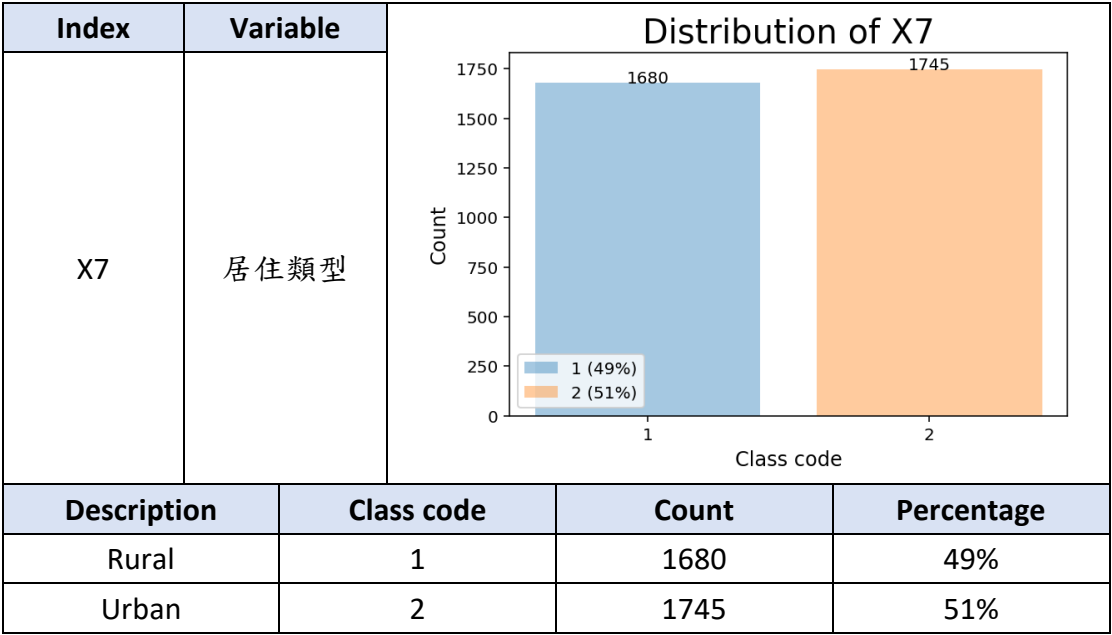
Index	Variable	Distribution of X2			
X2	年齡				
		Mean	SD	Median	Mode
		48.65	18.85	50	54
		Min	Max	Range	Variance
		10	82	72	355.32

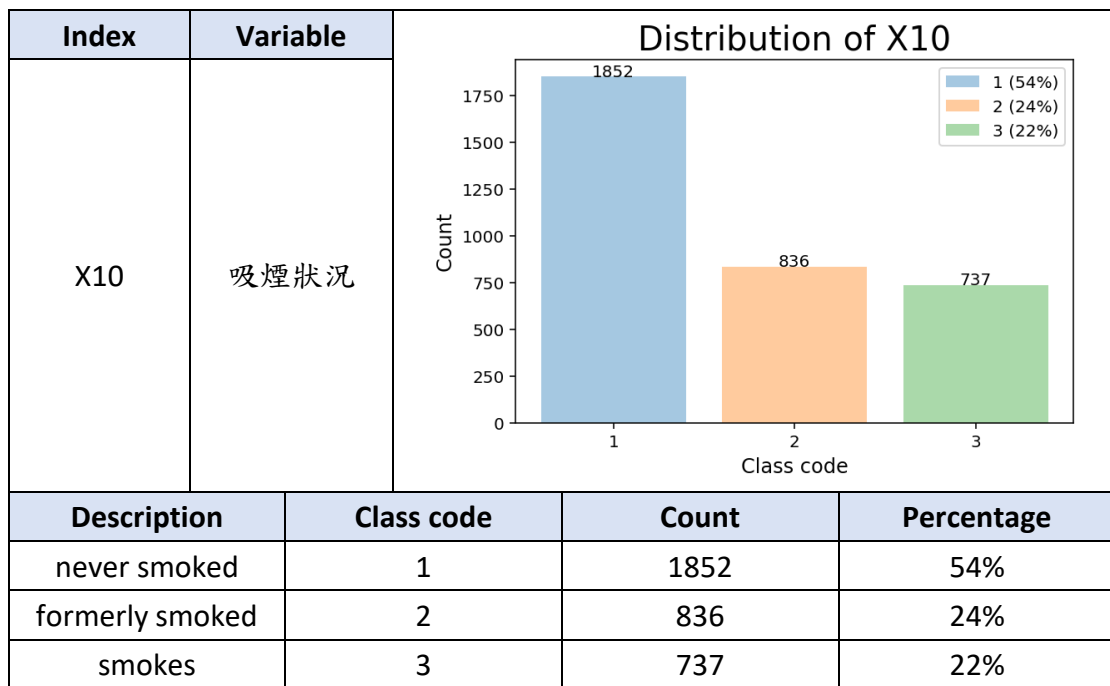
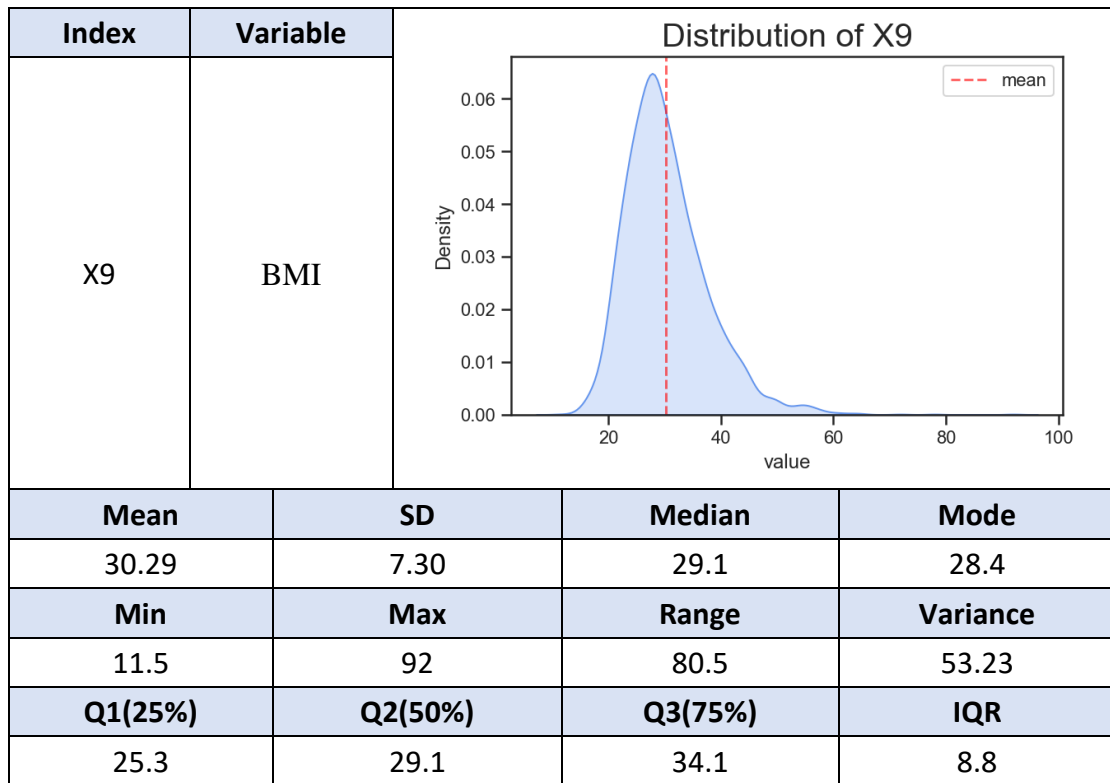
Q1(25%)	Q2(50%)	Q3(75%)	IQR
34	50	63	29

Index	Variable	<div>Distribution of X3</div> 	
X3	高血壓		
Description	Class code	Count	Percentage
No	0	3017	88%
Yes	1	408	12%

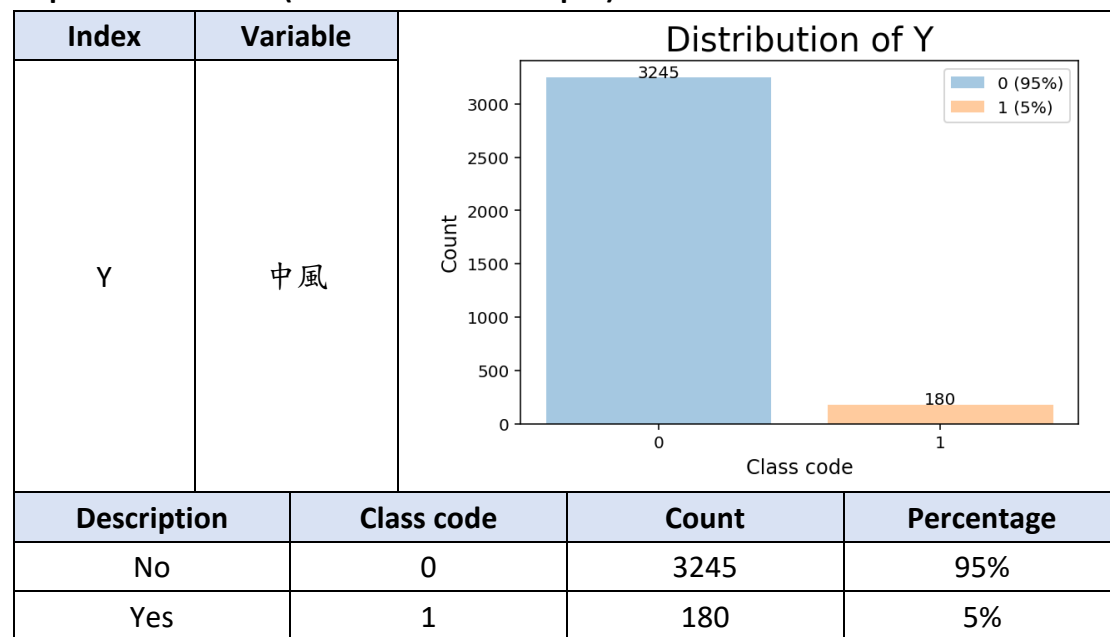
Index	Variable	<div>Distribution of X4</div> 	
X4	心臟病		
Description	Class code	Count	Percentage
No	0	3219	94%
Yes	1	206	6%







### Dependent variable (known as “Y” or output)



## 6. Cross Tabulations, Odds Ratio, Chi-Square Test, T-test:

### Categorical variables

Variable	Y,0	Y,1	Odds Ratio	P value
X1_1	1264(39%)	75(42%)	1	0.517
X1_2	1981(61%)	105(58%)	1.119462025	
X3_0	2894(89%)	123(68%)	1	1.159
X3_1	351(11%)	57(32%)	0.261721165	
X4_0	3075(95%)	144(80%)	1	1.915
X4_1	170(5%)	36(20%)	0.221138211	
X5_0	806(25%)	20(11%)	1	4.115
X5_1	2439(75%)	160(89%)	0.378256824	
X6_1	68(2.1%)	0(0%)	0	0.013
X6_2	14(0.4%)	0(0%)	0	
X6_3	2091(64.4%)	109(60%)	1	
X6_4	581(18%)	48(27%)	0.630968038	
X6_5	491(15.1%)	23(13%)	1.112823072	
X7_1	1594(49%)	86(48%)	1	0.784
X7_2	1651(51%)	94(52%)	0.947609386	
X10_1	1768(54%)	84(46%)	1	0.049
X10_2	779(24%)	57(32%)	0.649321267	
X10_3	698(22%)	39(42%)	0.850330665	

### Numeric variables

Variable	Y,0 (mean ± SD)	Y,1 (mean ± SD)	P value
X2	47.6±18.6	68±12	1.3106964821728848e-57 (p<0.05)
X8	106.7±46.2	136.8±63.1	2.244359093511492e-09 (p<0.05)
X9	30.3±7.3	30.7±6.3	0.438

### 7. Hyper parameter tuning: Logistic Regression

Round	Hyperparameter	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	confusion matrix	
								TN	FP
								FN	TP
Seed=1	C=1	0.724	0.149	0.971	0.711	0.259	0.899	463 1	188 33
Seed=2	C=1	0.771	0.188	0.773	0.771	0.302	0.833	494 10	147 34
Seed=3	C=1	0.711	0.12	0.788	0.707	0.208	0.79	461 7	191 26
Seed=4	C=1	0.616	0.103	0.909	0.601	0.186	0.791	392 3	260 30
Seed=5	C=1	0.74	0.141	0.778	0.738	0.239	0.792	479 8	170 28
Seed=6	C=1	0.769	0.183	0.75	0.771	0.295	0.814	494 11	147 33
Seed=7	C=1	<b>0.82</b>	0.152	0.778	0.822	0.255	0.843	541 6	117 21
Seed=8	C=1	0.712	0.136	0.833	0.706	0.233	0.804	458 6	191 30





Mean	0.716	0.142	0.836	0.71	0.241	0.818	
SD	0.06	0.02	0.07	0.07	0.03	0.03	

CART

Round	Hyperparameter	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	confusion matrix	
								TN	FP
								FN	TP
Seed=1	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.835	0.198	0.765	0.839	0.315	0.853	546 8	105 26
Seed=2	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.803	0.163	0.5	0.824	0.246	0.665	528 22	113 22
Seed=3	max_depth=2, max_leaf_nodes=4, min_samples_leaf=5,	0.826	0.169	0.667	0.834	0.27	0.786	544 11	108 22
Seed=4	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.52	0.087	0.939	0.498	0.159	0.758	325 2	327 31
Seed=5	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.6	0.098	0.806	0.589	0.175	0.721	382 7	267 29
Seed=6	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.556	0.118	0.909	0.909	0.208	0.777	341 4	300 40
Seed=7	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.803	0.125	0.667	0.809	0.211	0.783	532 9	126 18
Seed=8	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.803	0.163	0.667	0.81	0.262	0.773	526 12	123 24
Seed=9	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1,	0.819	0.17	0.657	0.828	0.271	0.741	538 12	112 23

Seed=10	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1,	0.797	0.159	0.575	0.811	0.249	0.752	523	122
								17	23
Mean		0.736	0.145	0.715	0.737	0.237	0.761		
SD		0.12	0.03	0.13	0.13	0.05	0.05		

KNN

Round	Hyperparameter	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	confusion matrix	
								TN	FP
								FN	TP
Seed=1	n_neighbors=19	0.547	0.082	0.794	0.535	0.148	0.704	348	303
								7	27
Seed=2	n_neighbors=20	0.79	0.162	0.545	0.807	0.25	0.709	517	124
								20	24
Seed=3	n_neighbors=20	0.526	0.078	0.818	0.511	0.142	0.682	333	319
								6	27
Seed=4	n_neighbors=20	0.556	0.078	0.758	0.546	0.141	0.658	356	296
								8	25
Seed=5	n_neighbors=20	0.774	0.116	0.5	0.789	0.188	0.686	512	137
								18	18
Seed=6	n_neighbors=17	0.619	0.114	0.727	0.612	0.197	0.687	392	249
								12	32
Seed=7	n_neighbors=20	0.687	0.098	0.667	0.748	0.171	0.762	492	166
								9	18
Seed=8	n_neighbors=20	0.75	0.101	0.472	0.766	0.166	0.603	497	152
								19	17
Seed=9	n_neighbors=20	0.549	0.08	0.743	0.538	0.144	0.668	350	300
								9	26
Seed=10	n_neighbors=20	0.745	0.119	0.525	0.758	0.194	0.667	489	156
								19	21
Mean		0.66	0.103	0.655	0.661	0.174	0.683		
SD		0.1	0.03	0.12	0.12	0.03	0.04		

# RF

Round	Hyperparameter	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	confusion matrix	
								TN	FP
								FN	TP
Seed=1	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=50	0.793	0.179	0.882	0.788	0.297	0.897	513 4	138 30
Seed=2	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=50	0.752	0.182	0.818	0.747	0.298	0.832	479 8	162 36
Seed=3	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=100	0.682	0.11	0.788	0.676	0.193	0.751	441 7	211 26
Seed=4	max_depth=3, max_leaf_nodes=4, min_samples_leaf=3, n_estimators=100	0.638	0.106	0.879	0.626	0.19	0.779	408 4	244 29
Seed=5	max_depth=3, max_leaf_nodes=4, min_samples_leaf=5, n_estimators=50	0.717	0.127	0.75	0.715	0.218	0.754	464 9	185 27
Seed=6	max_depth=3, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=100	0.749	0.17	0.75	0.749	0.277	0.787	480 11	161 33
Seed=7	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=100	<b>0.826</b>	0.157	0.778	0.828	0.261	0.847	545 6	113 21
Seed=8	max_depth=2, max_leaf_nodes=4, min_samples_leaf=1, n_estimators=50	0.699	0.127	0.806	0.693	0.22	0.784	450 7	199 29

Seed=9	max_depth=3, max_leaf_nodes=4, min_samples_leaf=3, n_estimators=50	0.615	0.107	0.886	0.6	0.19	0.803	390260	
								431	
Seed=10	max_depth=3, max_leaf_nodes=4, min_samples_leaf=3, n_estimators=100	0.677	0.131	0.8	0.67	0.225	0.785	432213	
								832	
Mean		0.715	0.14	0.814	0.709	0.237	0.802		
SD		0.06	0.03	0.05	0.07	0.04	0.04		

SVM

Round	Hyperparameter	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC	confusion matrix	
								TN	FP
								FN	TP
Seed=1	C=9	0.747	0.158	0.941	0.737	0.27	0.893	480 2	171 32
Seed=2	C=1	0.685	0.159	0.909	0.669	0.27	0.832	429 4	212 40
Seed=3	C=8	0.65	0.109	0.879	0.638	0.195	0.779	416 4	236 29
Seed=4	C=4	0.632	0.108	0.909	0.618	0.192	0.79	403 3	249 30
Seed=5	C=6	0.674	0.121	0.833	0.666	0.212	0.788	432 6	217 30
Seed=6	C=8	0.718	0.163	0.818	0.711	0.272	0.816	456 8	185 36
Seed=7	C=2	<b>0.806</b>	0.142	0.778	0.807	0.24	0.84	531 6	127 21
Seed=8	C=1	0.667	0.122	0.861	0.656	0.214	0.797	426 5	223 31

Seed=9	C=3	0.657	0.118	0.886	0.645	0.645	0.814	419	231
								4	31
Seed=10	C=1	0.804	0.178	0.178	0.814	0.28	0.28	525	120
								14	26
Mean		0.704	0.138	0.846	0.696	0.235	0.814		
SD		0.06	0.02	0.08	0.07	0.03	0.03		

#### Summary of results

	Accuracy	Precision	Sensitivity	Specificity	F1 score	AUC
<b>Logistic Regression</b>	0.722(±0.06)	0.143(±0.03)	0.824(±0.07)	0.717(±0.06)	0.242(±0.04)	0.817(±0.03)
<b>LR+Lasso</b>	0.716(±0.06)	0.142(±0.02)	0.836(±0.07)	0.71(±0.07)	0.241(±0.03)	0.818(±0.03)
<b>CART</b>	<b>0.736(±0.12)</b>	0.145(±0.03)	0.715(±0.13)	0.737(±0.13)	0.237(±0.05)	0.761(±0.05)
<b>KNN</b>	0.66(±0.1)	0.10(±0.03)	0.655(±0.12)	0.661(±0.12)	0.174(±0.03)	0.683(±0.04)
<b>RF</b>	0.715(±0.06)	0.14(±0.03)	0.814(±0.05)	0.709(±0.07)	0.237(±0.04)	0.802(±0.04)
<b>SVM</b>	0.704(±0.06)	0.138(±0.02)	0.846(±0.08)	0.696(±0.07)	0.235(±0.03)	0.814(±0.03)

## 8. Variables importance and ranking:

### Feature importance

Variable	Logistic Regression	LR+Lasso	CART	RF
X1_1	1.53783	0.15472	0	0.00106
X1_2	1.88058	0	0	0.00504
X2	5.28520	4.85768	0.83233	0.36580
X3_0	1.88058	0.51649	0	0.10126
X3_1	1.37924	0	0	0.09793
X4_0	1.85489	0.51339	0	0.11453
X4_1	1.40493	0	0	0.04112
X5_0	1.54407	0	0	0.02895
X5_1	1.71575	0.11714	0	0.01286
X6_1	16.79456	0	0	0.00327
X6_2	5.33805	0	0	0.00004
X6_3	6.47546	0.14132	0	0.00555
X6_4	6.12449	0.12770	0	0.00407
X6_5	6.27283	0	0	0.00326
X7_1	1.61549	0	0	0.00326

X7_2	1.64433	0.00515	0	0.00145
X8	0.88051	0.72156	0.16767	0.14672
X9	0.54402	0	0	0.05350
X10_1	1.23718	0.31461	0	0.00482
X10_2	1.18531	0.25167	0	0.00390
X10_3	0.83733	0	0	0.00437

### Feature ranking

Variable	Logistic Regression	LR+Lasso	CART	RF	Average
X1_1	14	7	21	20	15.5
X1_2	7	21	21	11	15
X2	6	1	1	1	2.3
X3_0	7	3	21	4	8.8
X3_1	16	21	21	5	15.8
X4_0	9	4	21	3	9.3
X4_1	15	21	21	7	16
X5_0	13	21	21	8	15.8
X5_1	10	10	21	9	12.5
X6_1	1	21	21	16	14.8
X6_2	5	21	21	21	17
X6_3	2	8	21	10	10.3
X6_4	4	9	21	14	12
X6_5	3	21	21	17	15.5
X7_1	12	21	21	17	17.8
X7_2	11	11	21	19	15.5
X8	19	2	2	2	6.3
X9	21	21	21	6	17.3
X10_1	17	5	21	12	13.8
X10_2	18	6	21	15	15
X10_3	20	21	21	13	18.8

### Final Feature ranking

Rank	Variable
1	X2
2	X8
3	X3_0
4	X4_0
5	X6_3
6	X6_4
7	X5_1
8	X10_1
9	X6_1
10	X1_2
10	X10_2
12	X1_1
12	X6_5
12	X7_2
15	X3_1
15	X5_0
17	X4_1
18	X6_2
19	X9
20	X7_1
21	X10_3