BigHW, LazyFCA

Mikhail Sambuev

December 2023

1 Description of first dataset

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

I have chosen heart failure clinical records dataset. About this dataset:

1. Age: displays the person's age
2. Anaemia: displays whether there is anemia or not

   1 = there is anemia

   0 = there is not anemia
3. Creatinine phosphokinase: displays the level of the CPK enzyme in the blood (mcg/L)
4. Diabetes: displays whether the person has diabetes
5. Ejection fraction: displays the percentage of blood leaving the heart at each contraction
6. High Blood pressure: displays whether the person has hypertension

1 = hypertension

0 = no hypertension

7. Platelets: displays the platelets in the blood

8. Serum creatinine: displays the level of serum creatinine in the blood

9. Serum sodium: displays the level of serum sodium in the blood

10. Sex: displays the person's gender

    1 = male

    0 = female

11. Smoking: displays whether the person has smokes or not

    1 = smoking

    0 = no smoking

12. Time: displays the follow-up period

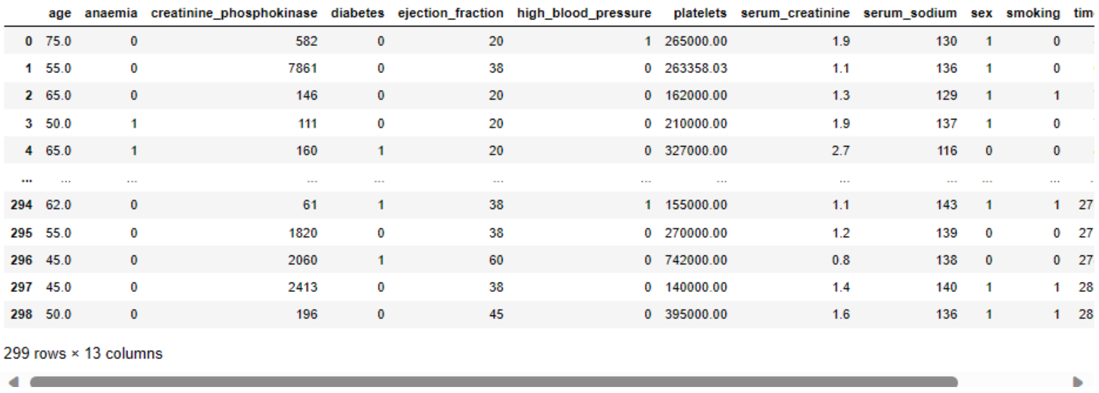
13. Death event: displays whether the person died during the follow-up period

    1 = died

    0 = didn't die

2 Data Pre-Processing

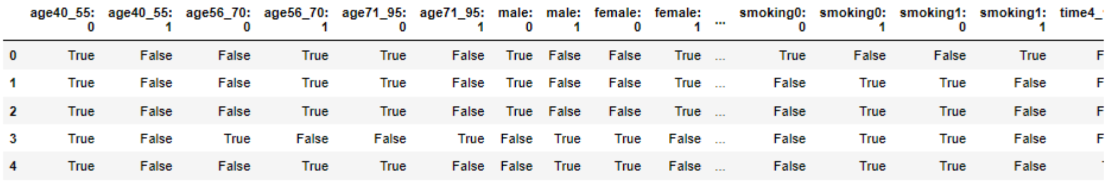
The dataset is shown below



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	tim
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	.
294	62.0	0	61	1	38	1	155000.00	1.1	143	1	1	27
295	55.0	0	1820	0	38	0	270000.00	1.2	139	0	0	27
296	45.0	0	2060	1	60	0	742000.00	0.8	138	0	0	27
297	45.0	0	2413	0	38	0	140000.00	1.4	140	1	1	28
298	50.0	0	196	0	45	0	395000.00	1.6	136	1	1	28

299 rows × 13 columns

First of all, we should binarize dataset. Details in code.



	age40_55: 0	age40_55: 1	age56_70: 0	age56_70: 1	age71_95: 0	age71_95: 1	male: 0	male: 1	female: 0	female: 1	...	smoking0: 0	smoking0: 1	smoking1: 0	smoking1: 1	time4_
0	True	False	False	True	True	False	True	False	False	True	...	True	False	False	True	F
1	True	False	False	True	True	False	True	False	False	True	...	False	True	True	False	F
2	True	False	False	True	True	False	True	False	False	True	...	False	True	True	False	F
3	True	False	True	False	False	True	False	True	True	False	...	False	True	True	False	F
4	True	False	False	True	True	False	False	True	True	False	...	False	True	True	False	

3 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. Random Forest Classifier
2. Decision Tree
3. XGB Classifier
4. Naive Bayes Classifier

Classifier	Accuracy
LazyFCA	71.5
F1score	54.5
Random Forest	63.4
Decision Tree	53.4
XGB	60
Naive Bayes	76.7

4 Description of first dataset

`https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-predictio`

I have chosen heart failure clinical records dataset. About this dataset:

1. Age: displays the person's age

2. Sex: displays the person's gender

   1 = male

   0 = female

3. Chest-pain type("cp"): displays the type of chest-pain experienced by the individual

   0 = typical angina

   1 = atypical angina

   2 = non — anginal pain

   3 = asymptotic

4. Resting Blood Pressure("trestbps"): displays the resting blood pressure value of an individual in mmHg

5. Serum Cholestrol("chol"): displays the serum cholesterol in mg/dl

6. Fasting Blood Sugar("fbs"): compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar $>$ 120mg/dl then :

   1 = true

   0 = false

7. Resting ECG("restecg") : displays resting electrocardiographic results

   0 = normal

   1 = having ST-T wave abnormality

   2 = left ventricular hyperthrophy

8. Max heart rate achieved : displays the max heart rate achieved by an individual

9. Exercise induced angina :

   1 = yes

   0 = no

10. ST depression induced by exercise relative to rest: displays the value which is an integer or float

11. Peak exercise ST segment :

    0 = upsloping

    1 = flat

    2 = downsloping

12. Number of major vessels (0–3) colored by flourosopy

13. Thal : displays the thalassemia

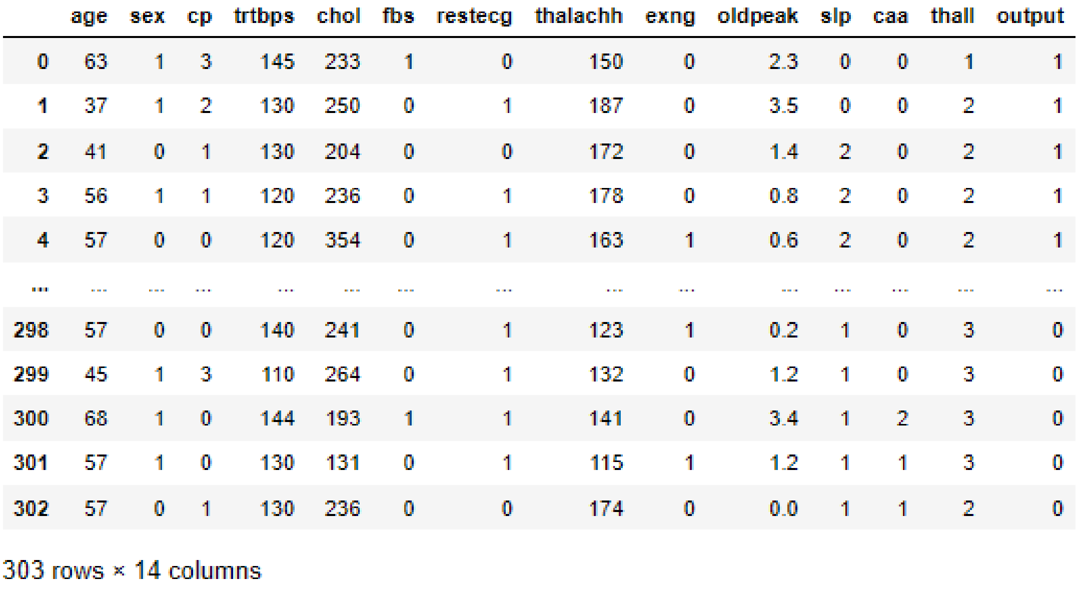
14. Diagnosis of heart disease : Displays whether the individual is suffering from heart disease or not:

    0 = absence

    1 = present

5 Data Pre-Processing

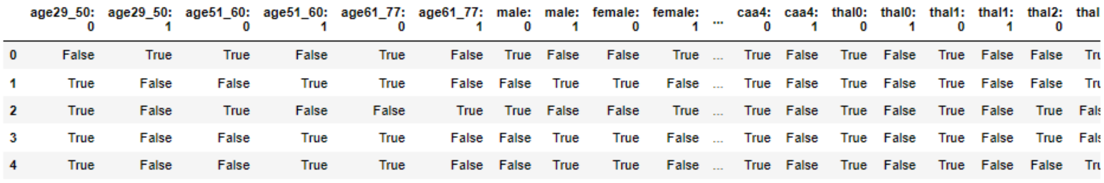
The dataset is shown below



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

303 rows × 14 columns

First of all, we should binarize dataset. Details in code.



	age29_50: 0	age29_50: 1	age51_60: 0	age51_60: 1	age61_77: 0	age61_77: 1	male: 0	male: 1	female: 0	female: 1	...	caa4: 0	caa4: 1	thal0: 0	thal0: 1	thal1: 0	thal1: 1	thal2: 0	thal
0	False	True	True	False	True	False	True	False	False	True	...	True	False	True	False	True	False	False	Tru
1	True	False	False	True	True	False	False	True	True	False	...	True	False	True	False	True	False	False	Tru
2	True	False	True	False	False	True	True	False	False	True	...	True	False	True	False	True	False	True	Fals
3	True	False	False	True	True	False	False	True	True	False	...	True	False	True	False	True	False	True	Fals
4	True	False	False	True	True	False	False	True	True	False	...	True	False	True	False	True	False	False	Tru

6 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. Random Forest Classifier
2. Decision Tree
3. XGB Classifier
4. Naive Bayes Classifier

Classifier	Accuracy
LazyFCA	78.8
F1score	79
Random Forest	71
Decision Tree	77.4
XGB	77.4
Naive Bayes	77.4

7 Description of first dataset

`https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset`

I have chosen heart failure clinical records dataset. About this dataset:

1. Age: displays the person's age
2. Pregnancies: displays to express the Number of pregnancies
3. Glucose: displays the Glucose level in blood
4. BloodPressure: displays the Blood pressure measurement

5. SkinThickness: displays the thickness of the skin

6. Insulin: displays the Insulin level in blood

7. BMI: displays the Body mass index

8. DiabetesPedigreeFunction: displays the Diabetes percentage

9. Serum sodium: displays the level of serum sodium in the blood

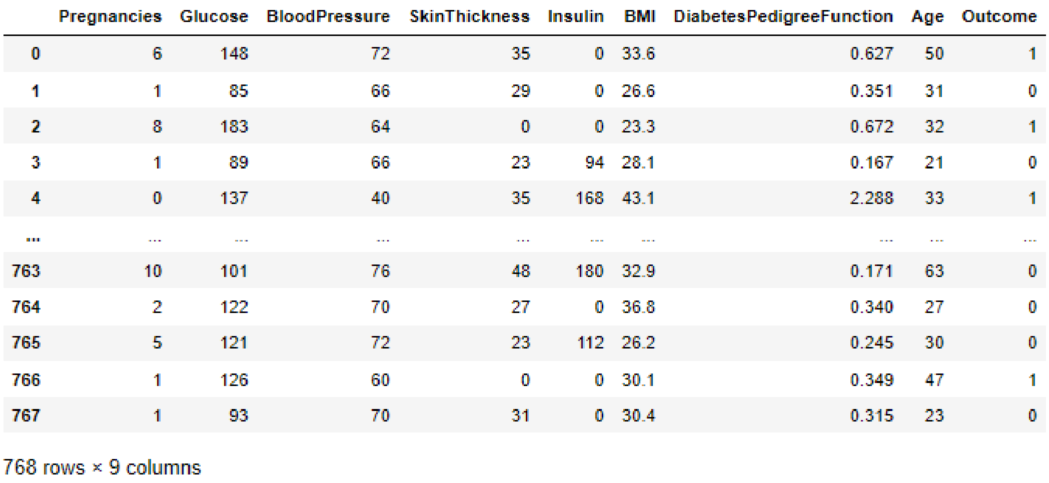
10. Outcome: displays the final result whether a person has diabetes or not

    1 = yes

    0 = no

8 Data Pre-Processing

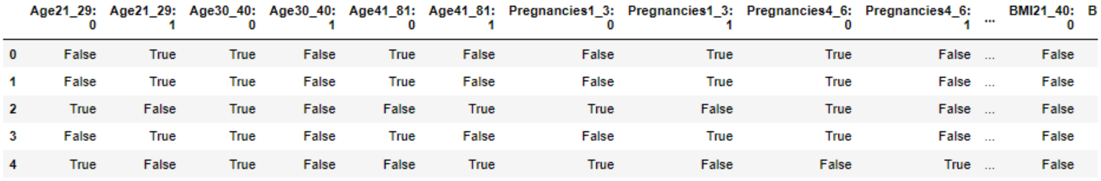
The dataset is shown below



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

First of all, we should binarize dataset. Details in code.



	Age21_29: 0	Age21_29: 1	Age30_40: 0	Age30_40: 1	Age41_81: 0	Age41_81: 1	Pregnancies1_3: 0	Pregnancies1_3: 1	Pregnancies4_6: 0	Pregnancies4_6: 1	...	BMI21_40: 0	B
0	False	True	True	False	True	False	False	True	True	False	...	False	
1	False	True	True	False	True	False	False	True	True	False	...	False	
2	True	False	True	False	False	True	True	False	True	False	...	False	
3	False	True	True	False	True	False	False	True	True	False	...	False	
4	True	False	True	False	False	True	True	False	False	True	...	False	

9 Comparison with classical classification algorithms

Here are the comparative results of classification algorithms. There were used next classifiers:

1. Random Forest Classifier
2. Decision Tree
3. XGB Classifier
4. Naive Bayes Classifier

Classifier	Accuracy
LazyFCA	68.2
F1score	34.5
Random Forest	67.5
Decision Tree	63.6
XGB	67.5
Naive Bayes	67.5