Predictive Modeling

Ma ChengYuan

detasets intro

datasets :

Based on
epiz_inform_stationary_risks_10_events.csv

1.final_obj_hospitalization.txt
- personal info

2.from all_epizodes_risks_strat.pkl –
- operation code
- diagnosis

3.all_analisis_risk_stratif.txt
- test result

target disease :

1) Желудочковая тахикардия
2) Острый коронарный синдром
3) Медиастинит
4) ОНМК

feature info

feature info :

Клинический_диагноз_рубрика : 723

Код_МЭС : 611

Код_теста : 2469

patients number by target :

желудочковая_тахикардия : 1723

острый_коронарный_синдром :712

медиастинит : 46

онмк : 437

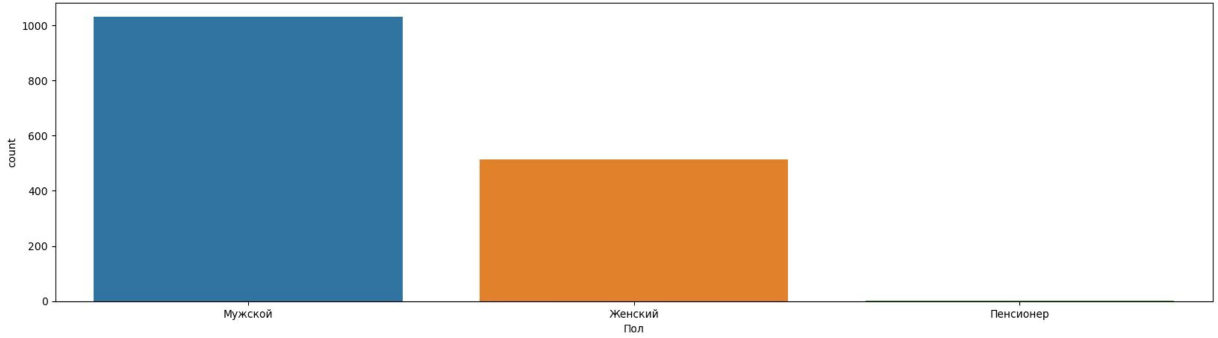
basic data distribution

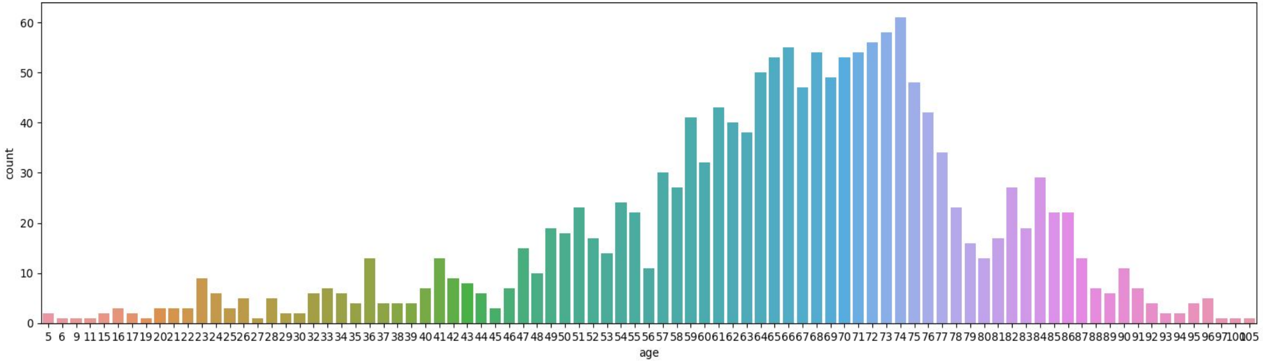
order by target as below :

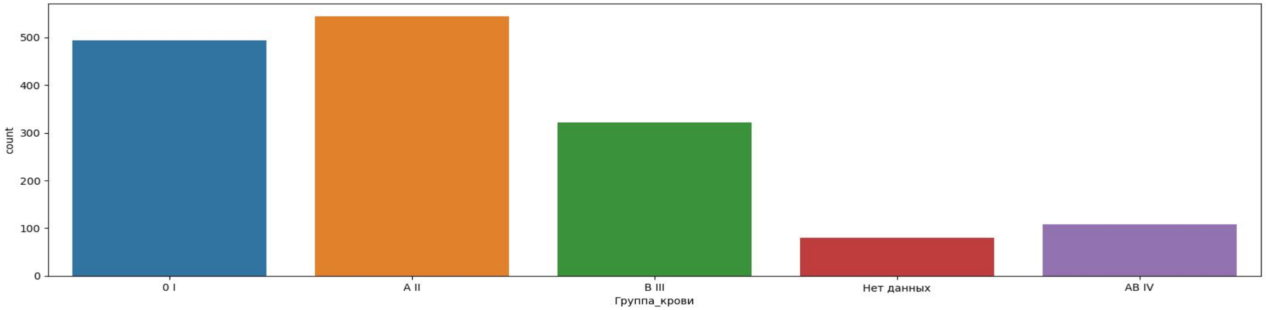
- желудочковая_тахикардия
- острый_коронарный_синдром
- медиастинит
- онмк

graph showed from top to bottom:

- gender
- age
- blood type



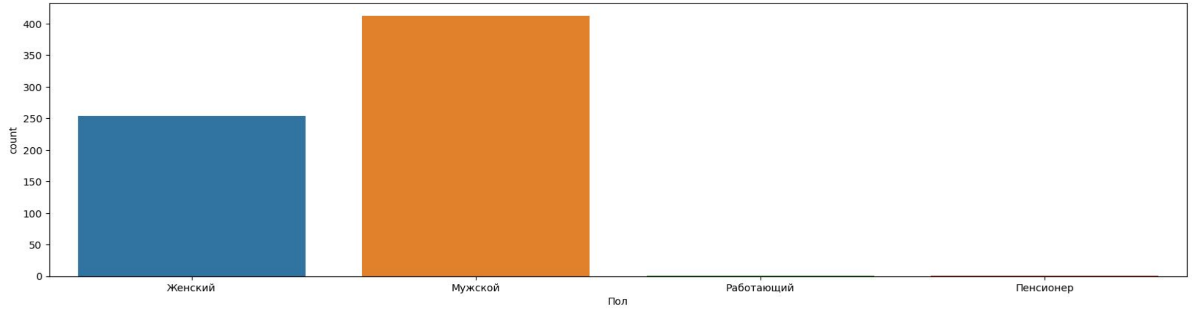


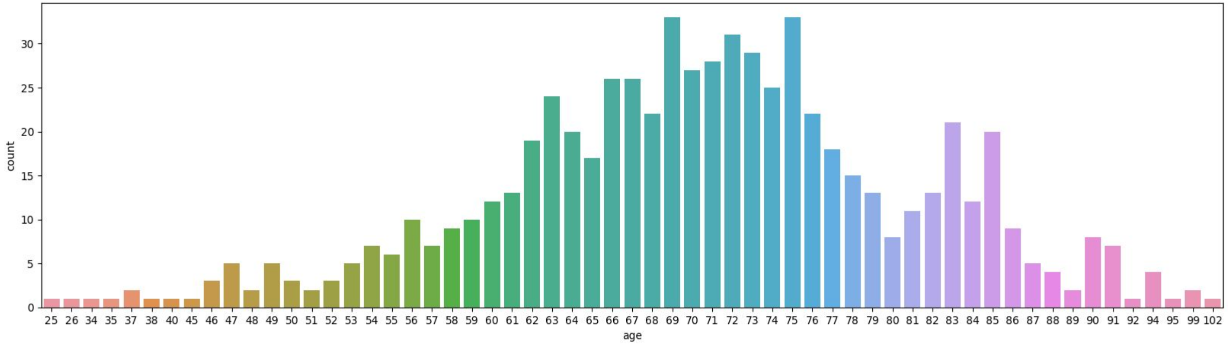


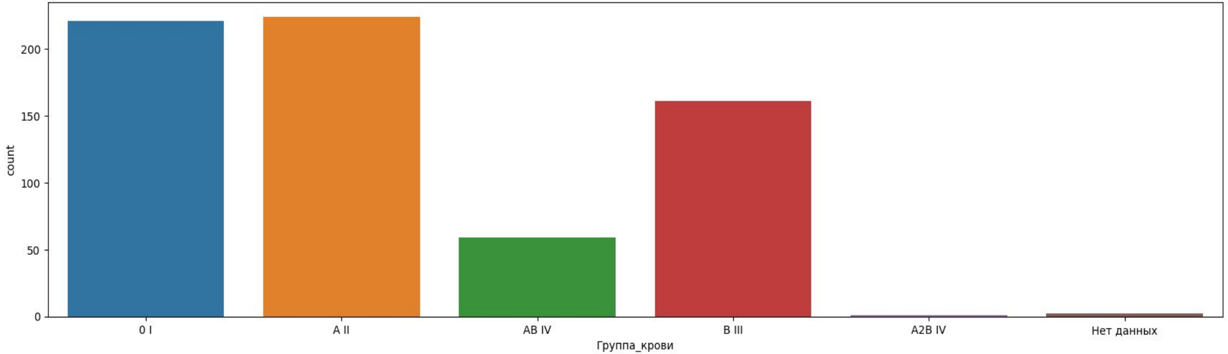
here we can see that obviously this disease concentrates on group of people whose age is around 70 years old .

Male shows nearly 2 times of possibility to get this disease

From the blood type , we can conclude that from higher proportion to lower , it is A2 , O1 , B3



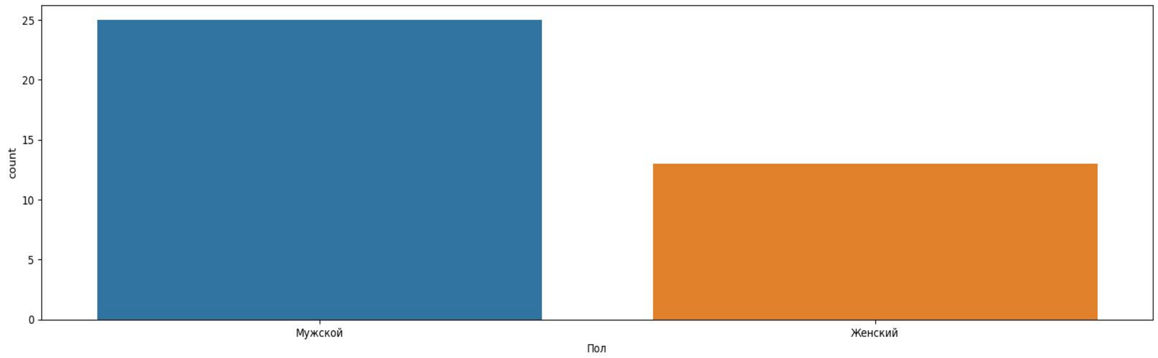


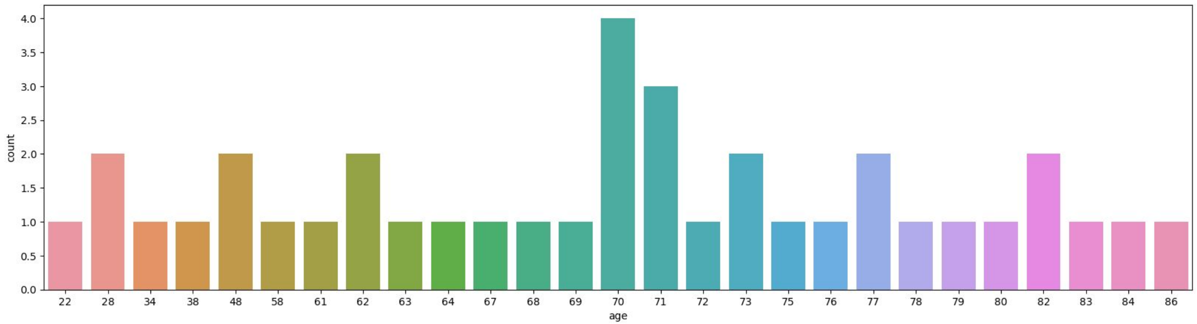


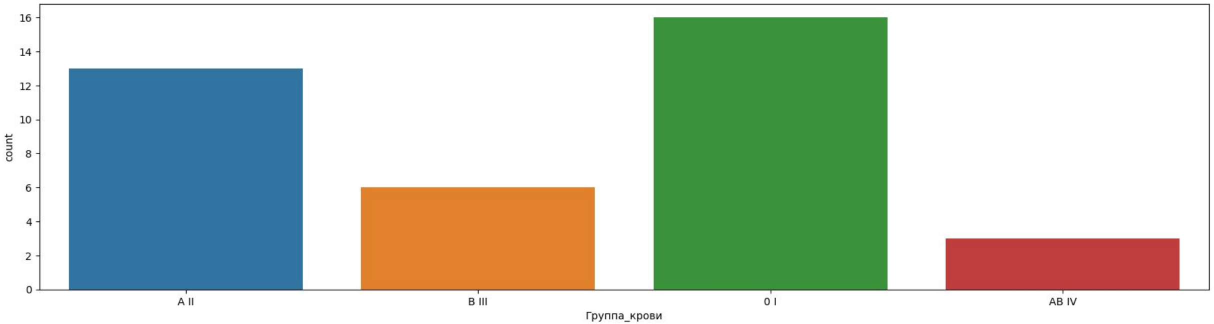
here we can see that obviously this disease concentrates on group of people whose age is around 70 years old .

Male shows nearly 2 times of possibility to get this disease

From the blood type , we can conclude that O1 , B3 , A2 shows higher proportion



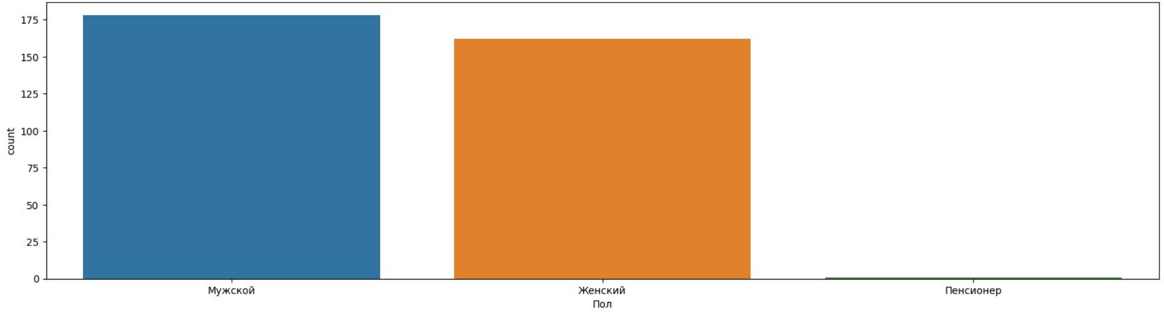


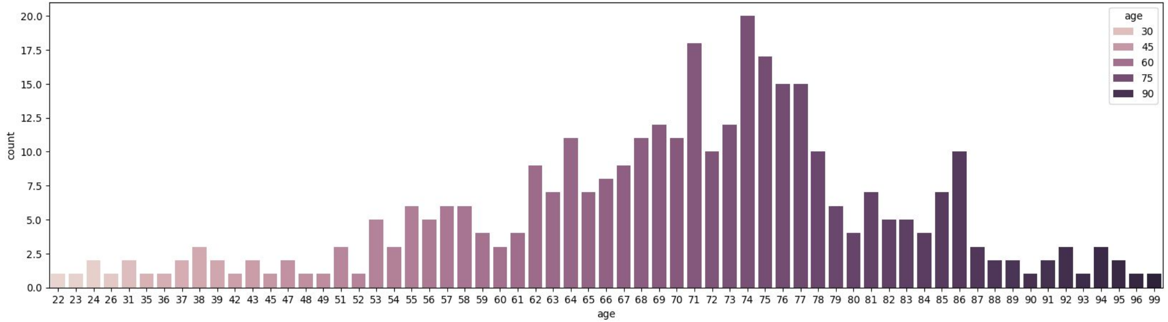


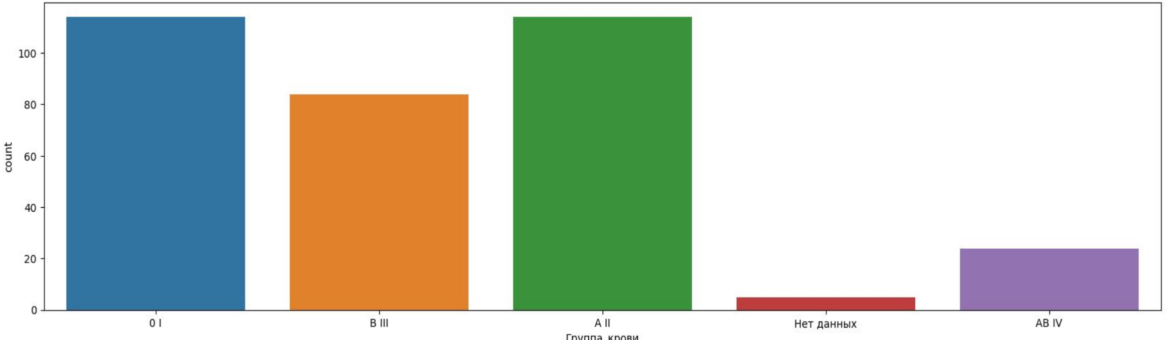
here we can see that obviously this disease concentrates on group of senior people

Male shows nearly 2 times of possibility to get this disease

From the blood type , we can conclude that O1 , A2 shows higher proportion



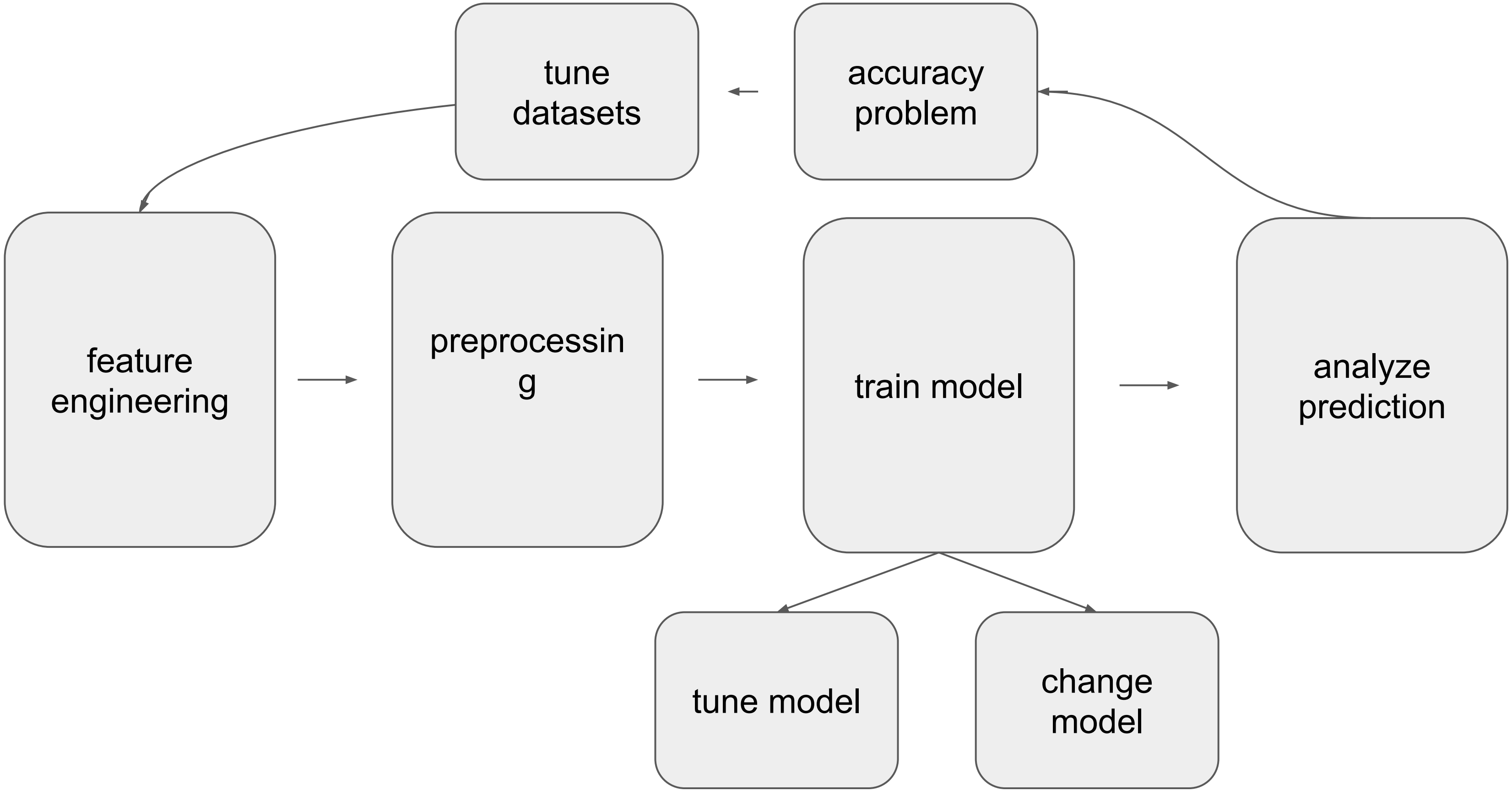




here we can see that obviously this disease concentrates on group of people whose age is around 70 years old .

From the blood type , we can conclude that O1 , B3 , A2 shows higher proportion

pipeline


tune
datasets
accuracy
problem
feature
engineering
preprocessin
g
train model
analyze
prediction
tune model
change
model

preprocessing

1. select potential columns from each df (**feature engineering**)
2. convert categorization value to numerical value using one-hot encoding (**preprocessing**)

*** main datasets ***
all_analisis_risk_stratif:

- Код_теста

1) fill null cell in col(‘Значение_число’) with average value from criteria ((lower + upper) / 2)

clinical_diag_293_strat_risk:

- Код_МЭС
- Клинический_диагноз_рубрика

model training

1. all_analisis_risk_stratif:
- Код_теста

test sets :

1) numerical value using col(‘Значение_число’) with average value in null cell
2) classified into lower & higher of criteria
3) without value in null cell
4) numerical columns of echo data included , without value in null cell + test sets (1)

2. clinical_diag_293_strat_risk:
- Код_МЭС
- Клинический_диагноз_рубрика

test result(1)

1)

row x column :
4909x3121

1) original

желудочковая_тахикардия :
Neural Net : f1 : 0.567878
xgb : f1 : 0.693428

острый_коронарный_синдром :
Neural Net : f1 : 0.592496
xgb : f1 : 0.759434

медиастинит :
Neural Net : f1 : 0.498126
xgb : f1 : 0.498126

онмк :
Neural Net : f1 : 0.510228
xgb : f1 : 0.599690

1) oversample

желудочковая_тахикардия :
Neural Net : f1 : 0.566822
xgb : f1 : 0.679563

острый_коронарный_синдром :
Neural Net : f1 : 0.574827
xgb : f1 : 0.754864

медиастинит :
Neural Net : f1 : 0.483158
xgb : f1 : 0.498126

онмк :
Neural Net : f1 : 0.420461
xgb : f1 : 0.618609

1) cross validation

желудочковая_тахикардия :
xgb : f1 : 0.684981

острый_коронарный_синдром :

xgb : f1 : 0.742531

медиастинит :

xgb : f1 : 0.52202

онмк :

xgb : f1 : 0.59033

test result(2)

2)

row x column :
4909x3435

2) original

желудочковая_тахикардия :
Neural Net : f1 :0.671587
xgb : f1 : 0.688900

острый_коронарный_синдром :
Neural Net : f1 : 0.725524
xgb : f1 : 0.691100

медиастинит :
Neural Net : f1 : 0.497954
xgb : f1 : 0.497954

онмк :
Neural Net : f1 : 0.589569
xgb : f1 : 0.602336

2) oversample

желудочковая_тахикардия :
Neural Net : f1 : 0.636534
xgb : f1 : 0.694334

острый_коронарный_синдром :
Neural Net : f1 : 0.702459
xgb : f1 : 0.684404

медиастинит :
Neural Net : f1 :0.544369
xgb : f1 : 0.569381

онмк :
Neural Net : f1 : 0.602206
xgb : f1 : 0.589569

2) cross validation

желудочковая_тахикардия :
xgb : f1 : 0.693104

острый_коронарный_синдром :

xgb : f1 : 0.69856

медиастинит :

xgb : f1 : 0.498106

онмк :

xgb : f1 : 0.610675

test result(3)

желудочковая_тахикардия :
xgb :
f1 : 0.693428
decsiontree :
f1 : 0.640268
LGB :
f1 : 0.686466
catboost :
f1 : 0.667132

original from (1):
желудочковая_тахикардия
:

xgb : f1 : 0.693428

острый_коронарный_синдром :
xgb :
f1 : 0.754944
decsiontree :
f1 : 0.652398
LGB :
f1 : 0.768578
catboost :
f1 : 0.743238

острый_коронарный_синдром :

xgb : f1 : 0.759434

медиастинит :
Neural Net :
f1 : 0.498126
xgb :
f1 : 0.498297
decsiontree :
f1 : 0.497784
LGB :
f1 : 0.498297
catboost :
f1 : 0.498297

медиастинит :

xgb : f1 : 0.498126

онмк :
xgb :
f1 : 0.619066
decsiontree :
f1 : 0.659900
LGB :
f1 : 0.614116
catboost :
f1 : 0.536594

онмк :

xgb : f1 : 0.599690

test result(4)

желудочковая_тахикардия :
xgb :
f1 : 0.708828

острый_коронарный_синдром :
xgb :
f1 : 0.752493

медиастинит :
Neural Net :
f1 : 0.498297

онмк :
xgb :
f1 : 0.626451

original from (1):
желудочковая_тахикардия
:

xgb : f1 : 0.693428

острый_коронарный_синдром :

xgb : f1 : 0.759434

медиастинит :

xgb : f1 : 0.498126

онмк :

xgb : f1 : 0.599690

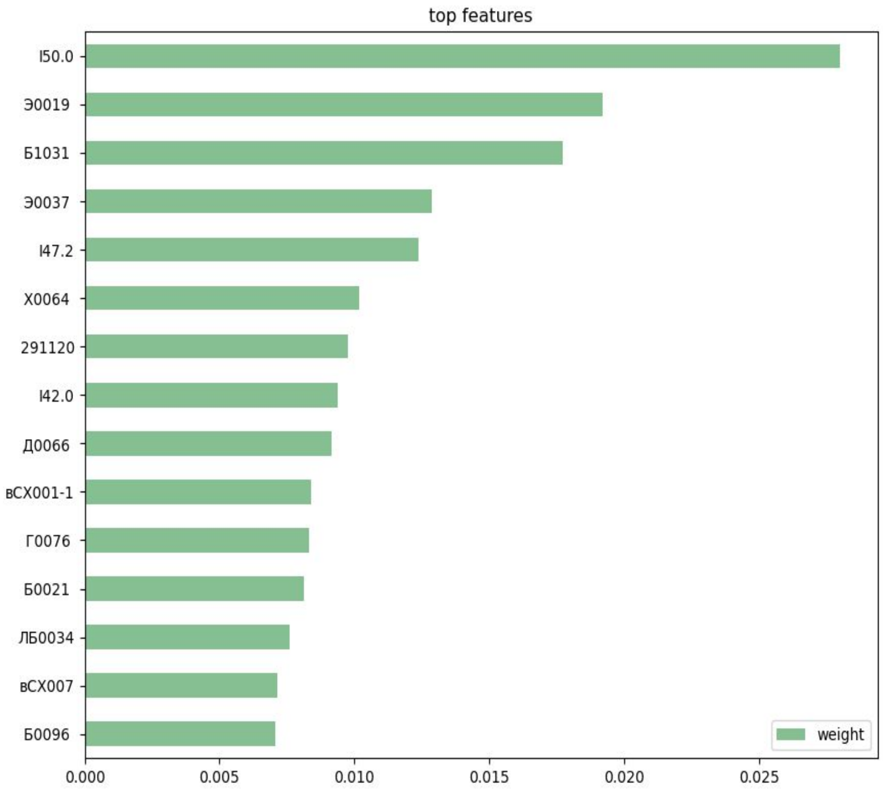
conclusion

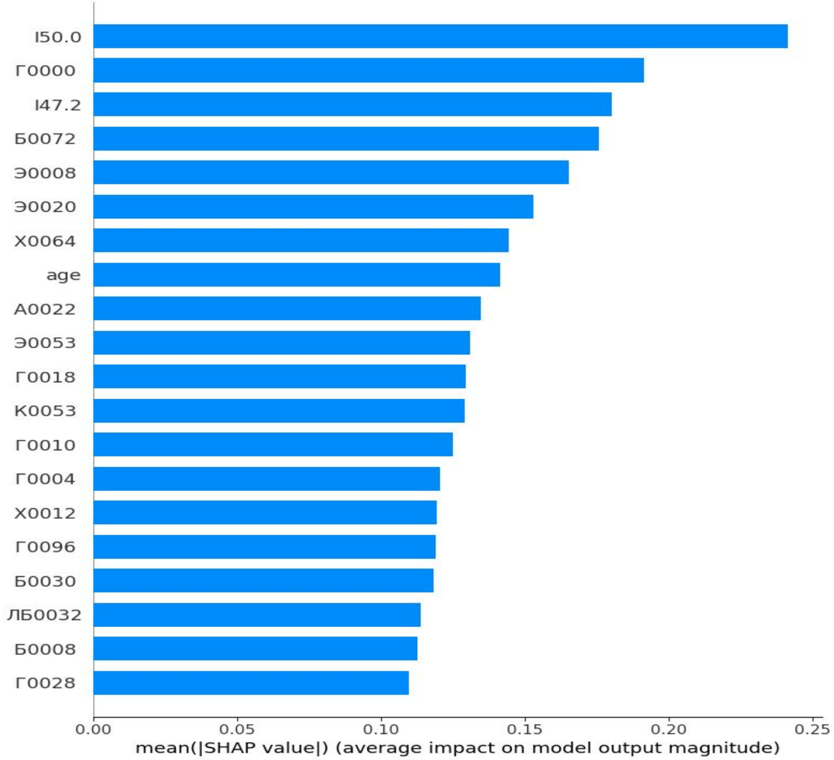
after tests , test set 1 is considered as baseline to compare with other test sets , test set 3 and 4 shows better accuracy on **онмк**

therefore , it can be considered that numerical value in echo data is clarifying the classification of **онмк**

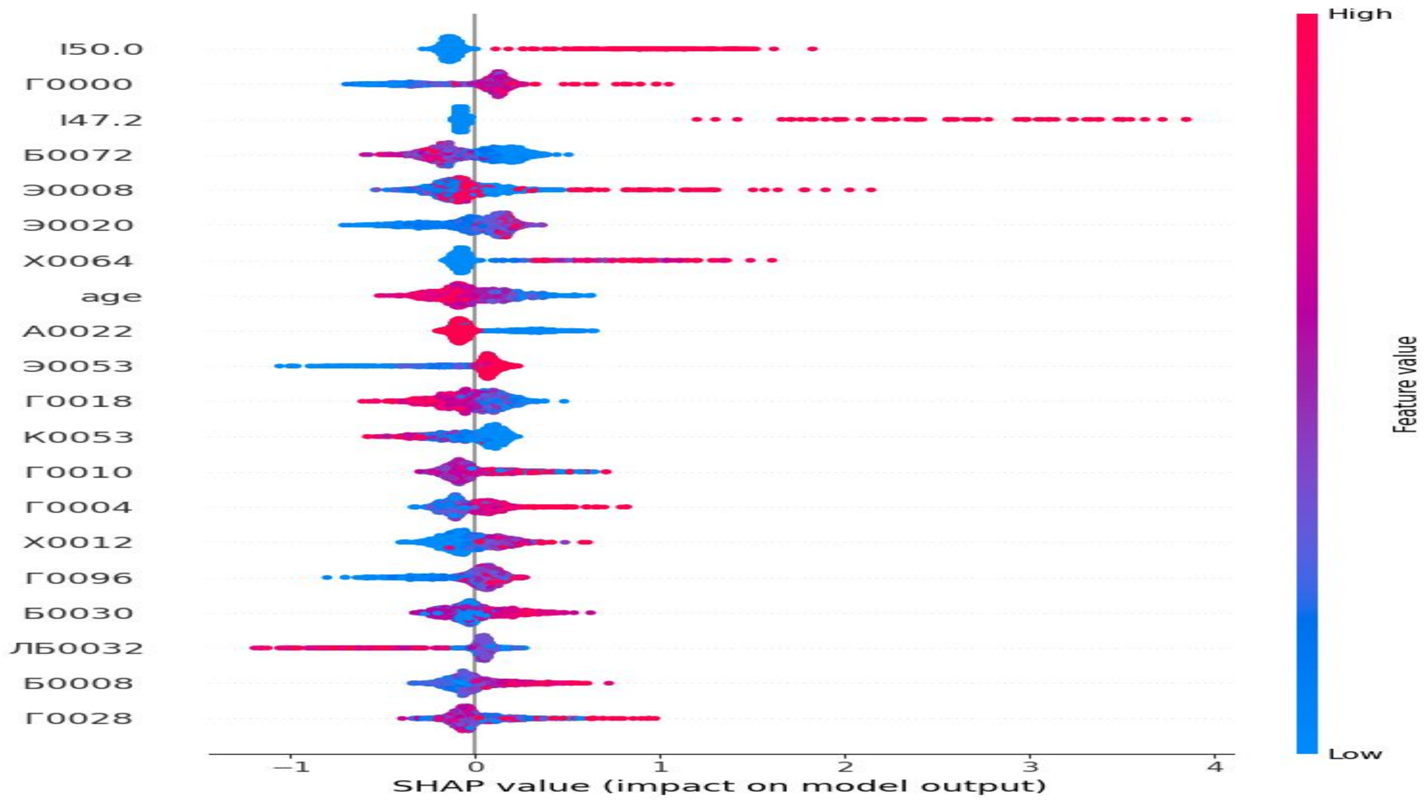
further investigation was processed with test set 1

feature importance (left xgb , right shap)

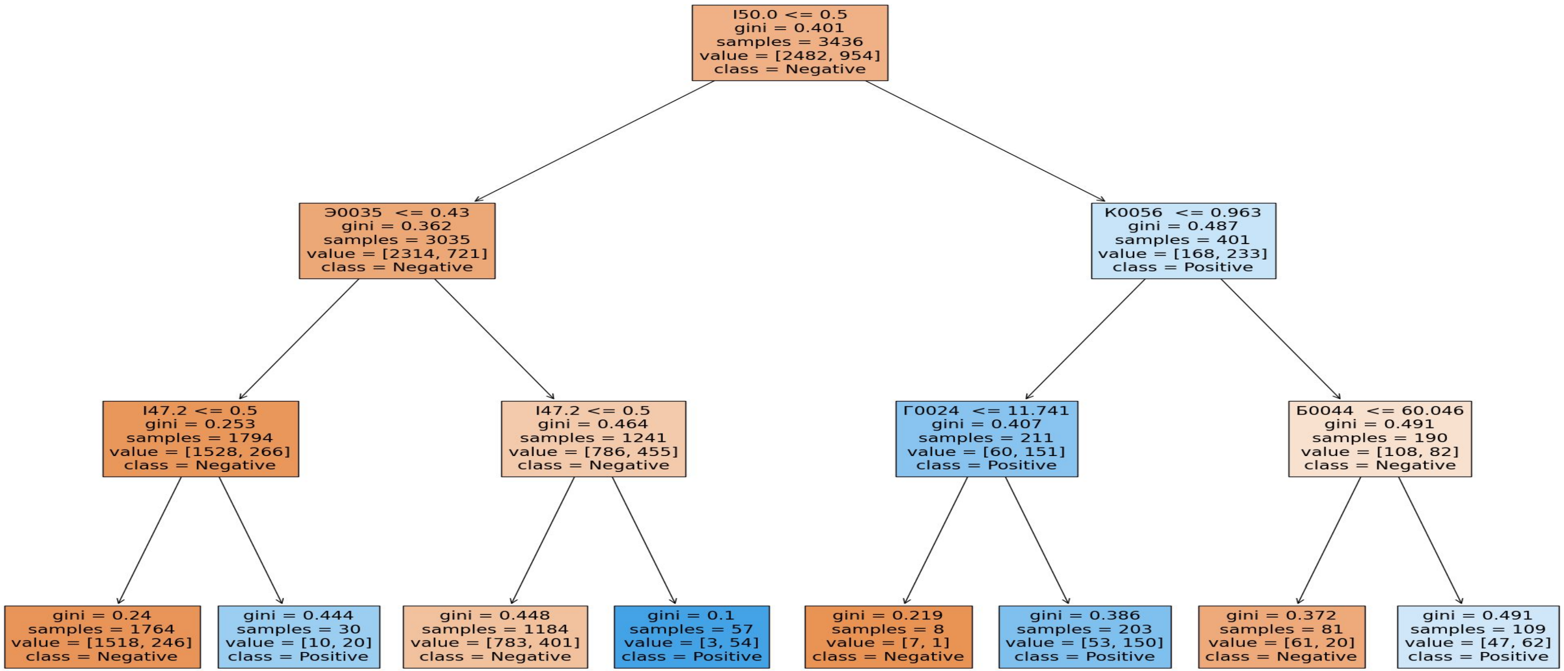




feature importance (shap)(black box model)



feature importance (decision tree)



feature selection(pearson)

X : correlation rate used to remove features
(higher rates means less feature removed)

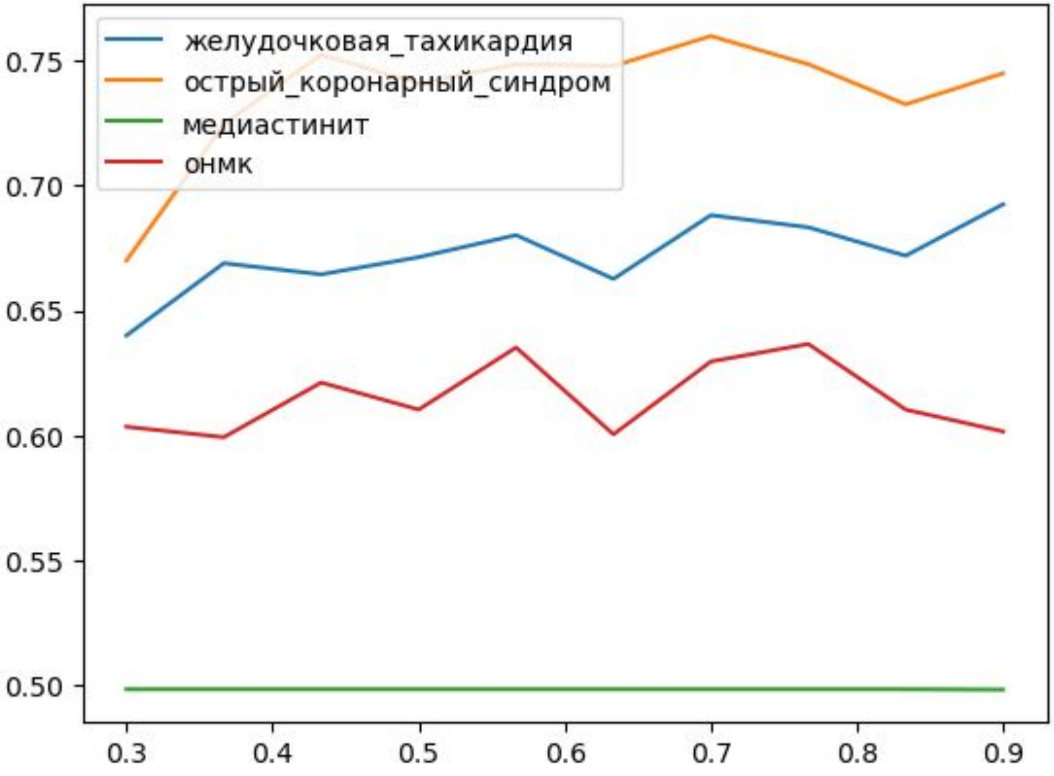
Y : F1 score

желудочковая_тахикардия best at : 0.9

острый_коронарный_синдром best at 0.7

медиастинит best at 0.3

онмк best at 0.75



feature selection(L2 regularization after removal correlated columns)

желудочковая_тахикардия :
total features: 2415

selected features: 776

f1 : 0.678326

острый_коронарный_синдром :
total features: 1870

selected features: 580

f1 : 0.739063

медиастинит :
total features: 802

selected features: 226

f1 : 0.498126

онмк :
total features: 2148

selected features: 671

f1 : 0.599242

original from (1):
желудочковая_тахикардия :

xgb : f1 : 0.693428

острый_коронарный_синдром :

xgb : f1 : 0.759434

медиастинит :

xgb : f1 : 0.498126

онмк :

xgb : f1 : 0.599690

feature selection(L2 regularization with original data)

желудочковая_тахикардия :
total features: 3117

selected features: 958

f1 : 0.678892

test feature is 606
operations feature is 153
diagnosis feature is 196
['Пол', 'age', '0 I']

compared with original from test(1):
желудочковая_тахикардия :

xgb : f1 : 0.693428

острый_коронарный_синдром :
total features: 3117

selected features: 813

f1 : 0.737259

test feature is 530
operations feature is 125
diagnosis feature is 154
['age', 'A II', 'AB IV', 'B III']

острый_коронарный_синдром :

xgb : f1 : 0.759434

онмк :
total features: 3117

selected features: 853

f1 : 0.617764

test feature is 495
operations feature is 98
diagnosis feature is 99
['Пол', '0 I', 'A II', 'B III']

онмк :

xgb : f1 : 0.599690

shapley extraction

желудочковая_тахикардия :
total features: 958

selected features: 236

f1 : 0.689613

['age', 'Пол', '0 I']

test feature is 210

operations feature is 8

diagnosis feature is 15

original from previous result
желудочковая_тахикардия :

xgb : f1 : 0.678892
roc auc score : 0.7758469115345669

острый_коронарный_синдром :
total features: 814

selected features: 207

f1 : 0.737874

['B III', 'age', 'A II']

test feature is 189
operations feature is 9
diagnosis feature is 6

острый_коронарный_синдром :

xgb : f1 : 0.737259

онмк :
total features: 853

selected features: 218

f1 : 0.633716

['Пол', 'age', '0 I', 'A II']

test feature is 204

operations feature is 7

diagnosis feature is 3

онмк :

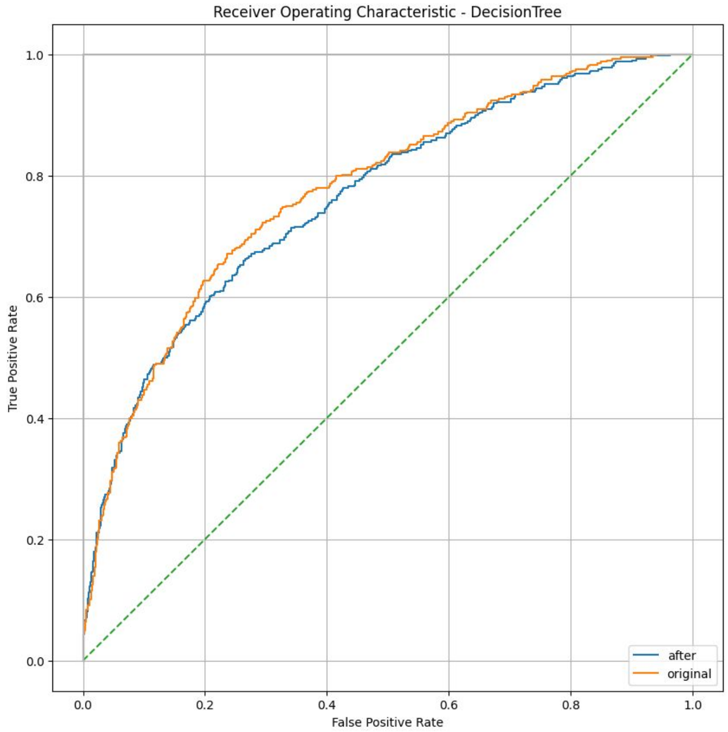
xgb : f1 : 0.617764

metrics (желудочковая_тахикардия)

f1 : 0.689613

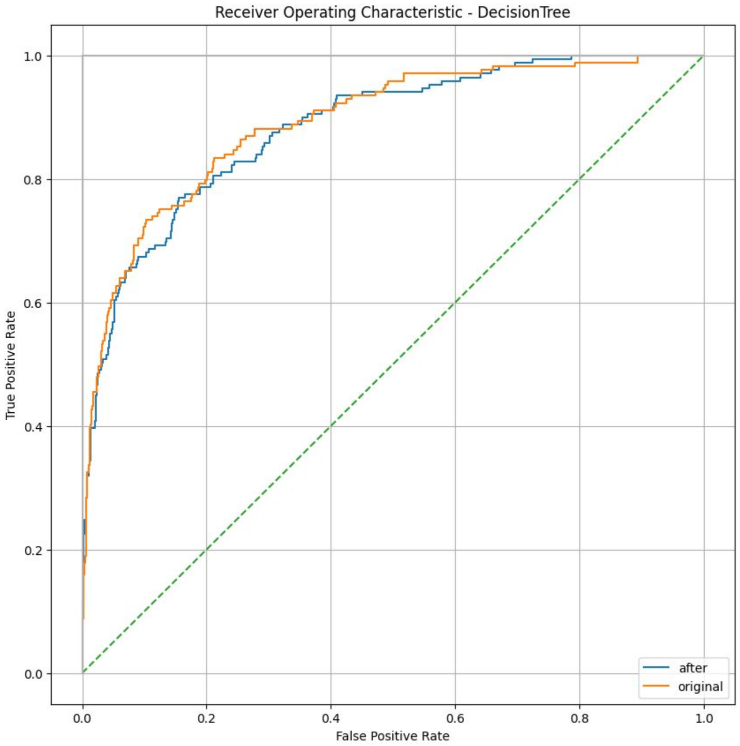
	precision	recall	f1-score	support
0	0.81	0.90	0.85	1065
1	0.64	0.45	0.53	408
accuracy			0.78	1473
macro avg	0.72	0.67	0.69	1473
weighted avg	0.76	0.78	0.76	1473

roc auc score is : 0.7625172604252969



metrics (острый_коронарный_синдром)

roc auc score is : 0.8834628090173159




```
f1 : 0.737874
              precision    recall  f1-score   support

           0       0.93      0.98      0.95      1304
           1       0.73      0.41      0.52       169

    accuracy                           0.91      1473
   macro avg       0.83      0.69      0.74      1473
weighted avg       0.90      0.91      0.90      1473
```

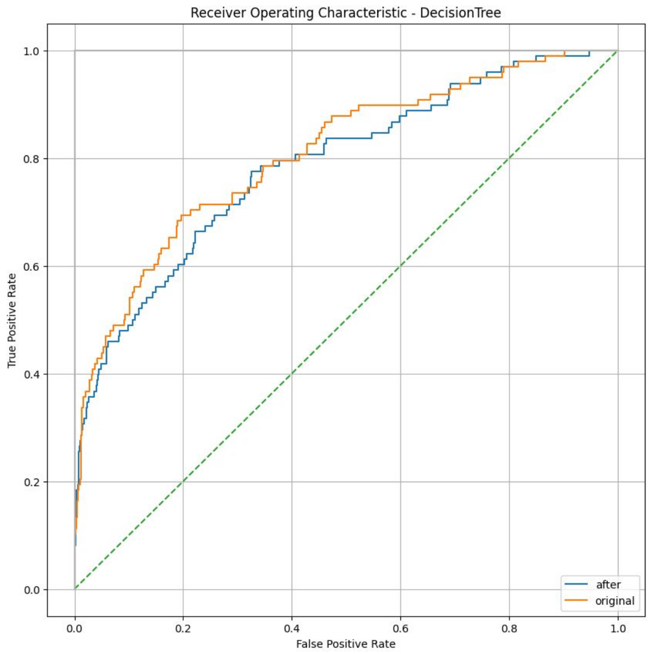
metrics (онмк)

roc auc score is : 0.7870278293135435

```
f1 : 0.633716
              precision    recall  f1-score   support

           0       0.94      1.00      0.97      1375
           1       0.78      0.18      0.30        98

    accuracy                           0.94      1473
   macro avg       0.86      0.59      0.63      1473
weighted avg       0.93      0.94      0.93      1473
```



feedback

After feature extraction , features were trimmed to around 200 , and f1 score still can maintain with good quality .

It is obvious to see that **test feature** can influence the result of prediction more than other features

related previous work

желудочковая_тахикардия :

A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy

link : https://www.sciencedirect.com/science/article/pii/S001048252100442X

острый_коронарный_синдром :

A Machine Learning-Based Approach for the Prediction of Acute Coronary Syndrome Requiring Revascularization

link : https://link.springer.com/article/10.1007/s10916-019-1359-5

медиастинит :

Performance of a Machine Learning Algorithm in Predicting Outcomes of Aortic Valve Replacement

link : https://www.sciencedirect.com/science/article/abs/pii/S0003497520311565

онмк :

Performance Analysis of Machine Learning Approaches in Stroke Prediction

link: https://ieeexplore.ieee.org/abstract/document/9297525?casa_token=TfM_OTIj2BEAAAAA:vV39yNcKMpzQc9jI_oopWu0eggmUj9CRoMETefwiKE7d3W07qChFVgS8HmEnqhtRvggkcX0FChDokA

related previous work

желудочковая_тахикардия :

A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy

link : https://www.sciencedirect.com/science/article/pii/S001048252100442X

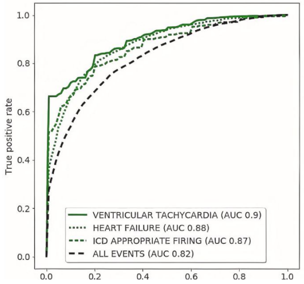
Table 1
Patient demographic, physical and clinical characteristics. The upper part of th table shows patient overview characteristics and the lower part shows the sta tistics in terms of baseline and follow-up measurements.

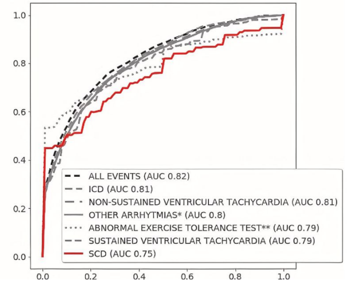
Patients	Total (N = 2302)	
Characteristics	no. (%)	
Sex		
Male	1448 (62.9%)	
Female	854 (37.1%)	
Family history of HCM	983 (42.7%)	
Family history of SCD	426 (18.5%)	
Family history of CAD	104 (4.5%)	
Diabetes	82 (3.6%)	
Type 2 diabetes	73 (3.2%)	
Hypertension	214 (9.3%)	
Hypercholesterolemia	478 (20.8%)	
Genetic mutations	**Total tests performed (N = 1321)**	
MYBPC3	455 (34.4%)	
MYH7	254 (19.2%)	
MYL2	13 (1.0%)	
MYL3	7 (0.5%)	
TNNI3	42 (3.2%)	
TNNT2	45 (3.4%)	
TPM1	8 (0.6%)	
TTN	3 (0.2%)	

Table 2
Performance of the machine learning algorithms on the task of risk stratification of HCM patients. The results of the 10-fold cross-validation for predicting high-risk patients five years ahead are shown. The reported values are mean values and standard deviation between cross-validation folds. The best results for each metric are in bold.

Model	Accuracy	AUC	Specificity	Sensitivity	Precision	F_1 score
Random forest	0.72 ± 0.03	0.79 ± 0.03	0.81 ± 0.05	0.62 ± 0.03	0.74 ± 0.05	0.68 ± 0.03
SVM (linear)	0.69 ± 0.05	0.74 ± 0.04	0.69 ± 0.05	**0.69 ± 0.08**	0.59 ± 0.08	0.63 ± 0.07
SVM (RBF)	0.67 ± 0.02	0.73 ± 0.03	0.68 ± 0.03	0.64 ± 0.05	0.62 ± 0.04	0.63 ± 0.04
Boosted trees	**0.75 ± 0.02**	**0.82 ± 0.02**	0.81 ± 0.03	0.67 ± 0.04	0.78 ± 0.02	**0.72 ± 0.02**
Neural-Networks	0.74 ± 0.03	0.80 ± 0.04	**0.86 ± 0.05**	0.61 ± 0.07	**0.79 ± 0.05**	0.68 ± 0.05

AUC – Area Under Curve, SVM – support vector machine, RBF – radial basis kernel.





comparison(желудочковая_тахикардия)

my	(желудочковая_тахикардия)
removed rows when there is missing value	copying past/known values of the last result (in the range of five years)
filled missing value with average value from range of criteria	missing numerical values were replaced by random samples from the normal distributions
	combining all possible pairs of patient measurements as features
4909 (1723:3186)	2302 (undefined)
XGB	XGB
AUC 0.76 , F1 0.69	AUC 0.82 , F1 0.71

related previous work

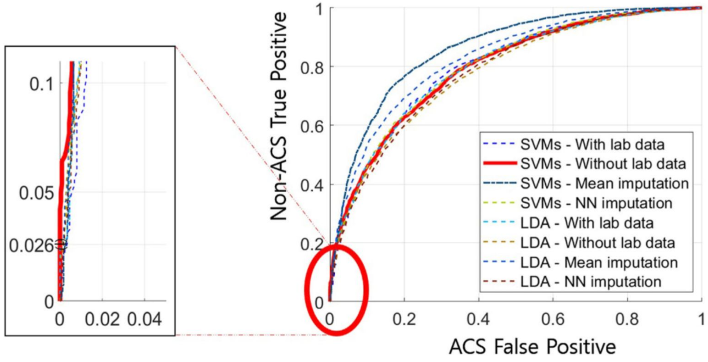
острый_коронарный_синдром :

A Machine Learning-Based Approach for the Prediction of Acute Coronary Syndrome Requiring Revascularization

link : https://link.springer.com/article/10.1007/s10916-019-1359-5

Table 1 Features for analysis

		Features
Epidemiological data	1	Gender
	2	Age
Clinical data at emergency department or outpatient clinic	3	Systolic BP
	4	Diastolic BP
	5	HR
Past medical history before presenting chest pain	6	CAD
	7	MI
	8	CABG
	9	PCI
	10	Hypertension
	11	DM
	12	Hyperlipidemia
	13	CVA
	14	PCI
	15	History of Smoking
	16	Current smoking
Laboratory data before presenting chest pain	17	TC
	18	LDL- cholesterol
	19	HDL-cholesterol
	20	TG



comparison(острый_коронарный_синдром)

my	(острый_коронарный_синдром)
filled missing value with average value from range of criteria	Nearest neighbor imputation to fill missing value
4909 (712:4197)	5838 (2311 : 3527)
XGB	SVM
AUC 0.88 , F1 0.74	AUC 0.86

CCDS

interface : telegram

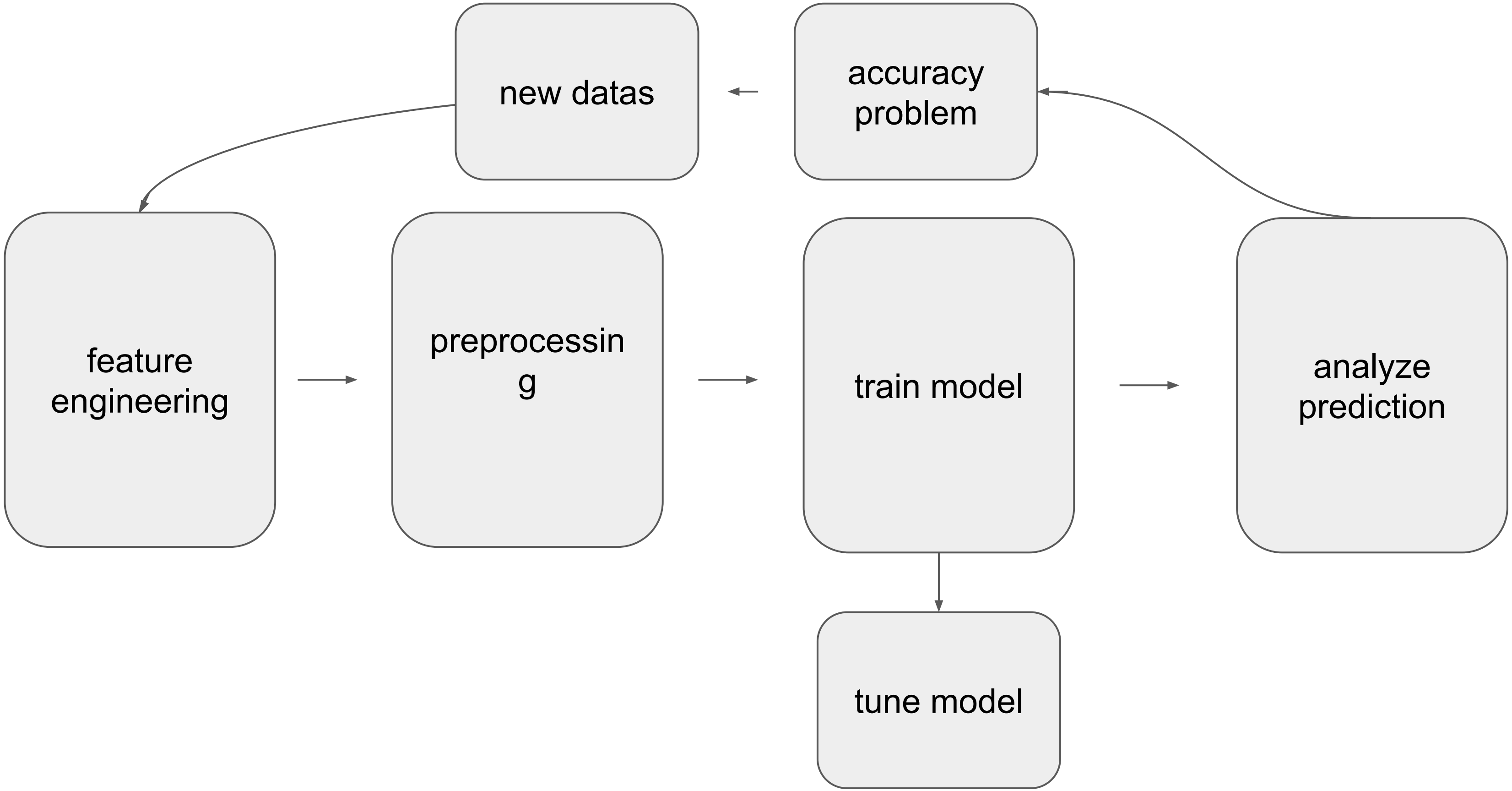
language : python

database : mongodb

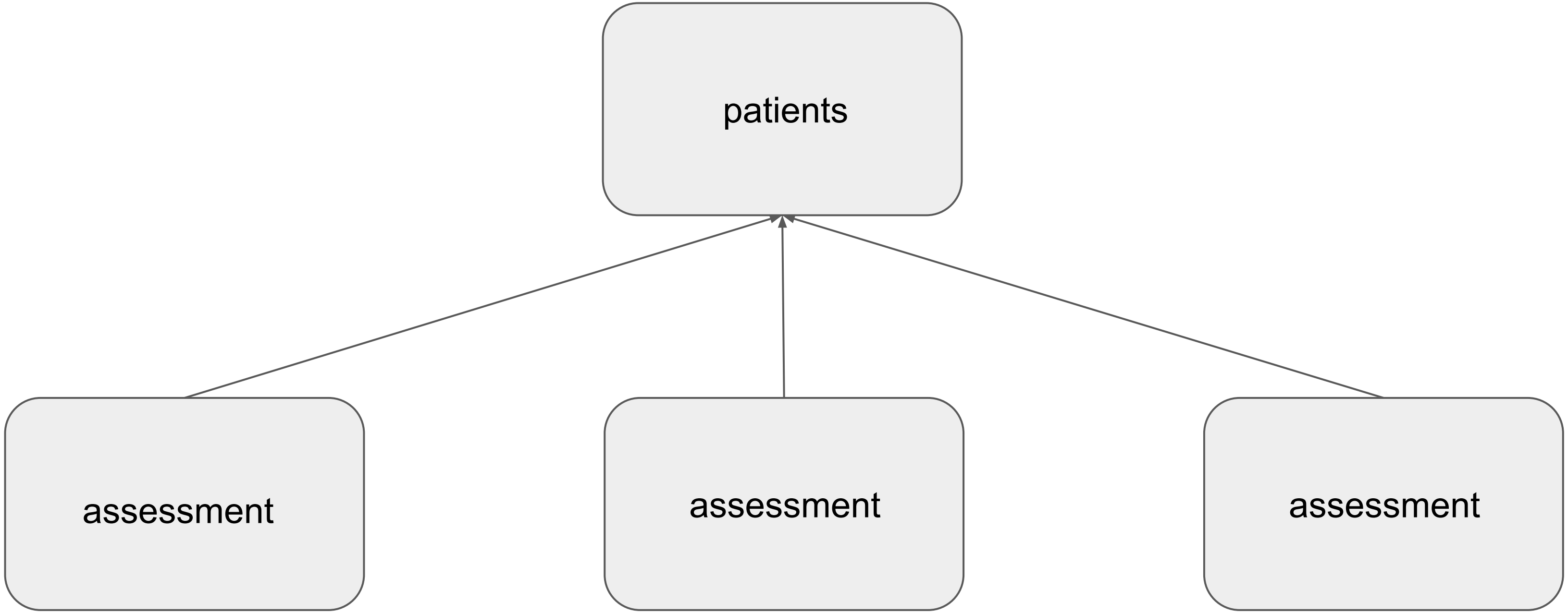
motivation : To support doctor by prediction and possible factors

goal : To reduce the diagnosis time in order to offer faster and more accurate treatment and relieve labourious workload for medical staffs

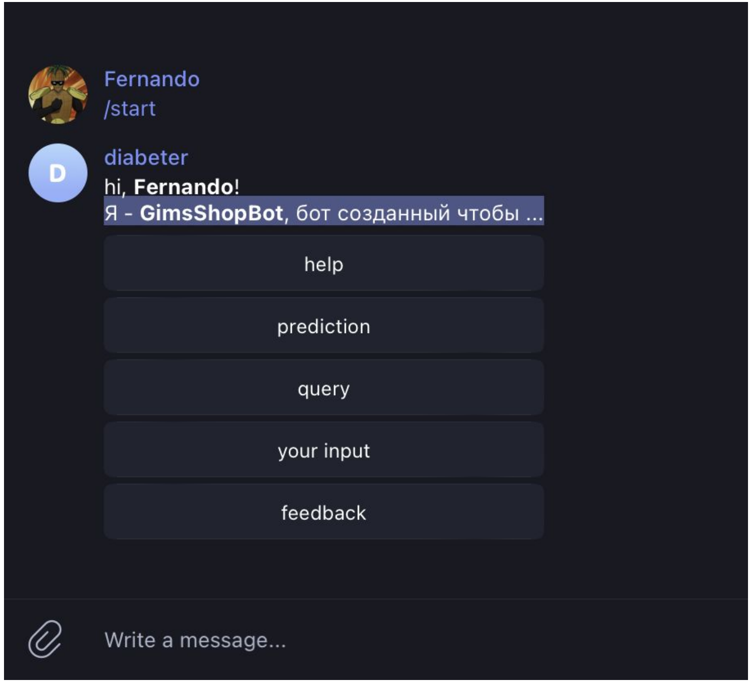
pipeline


new datas
accuracy problem
feature engineering
preprocessin g
train model
analyze prediction
tune model

database design



demo image



/help - to offer instruction

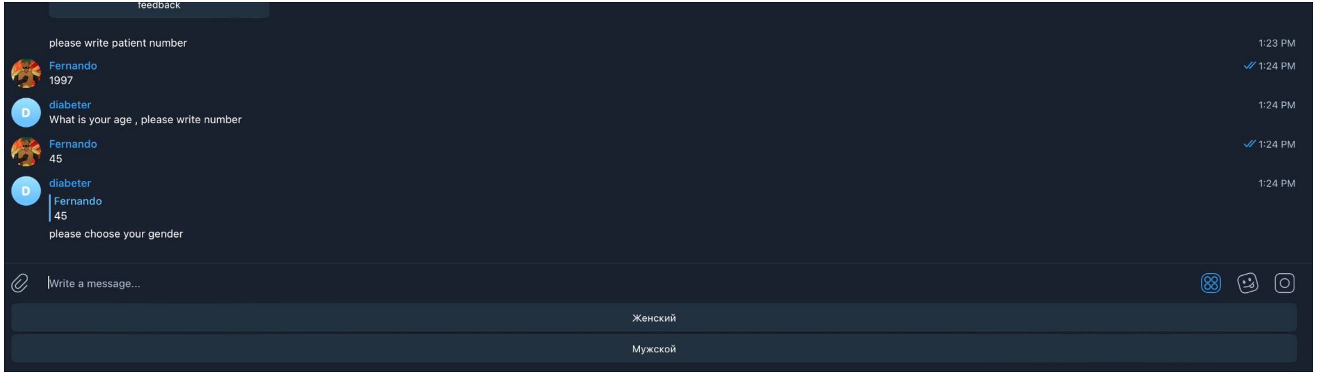
/prediction - main functionality to predict

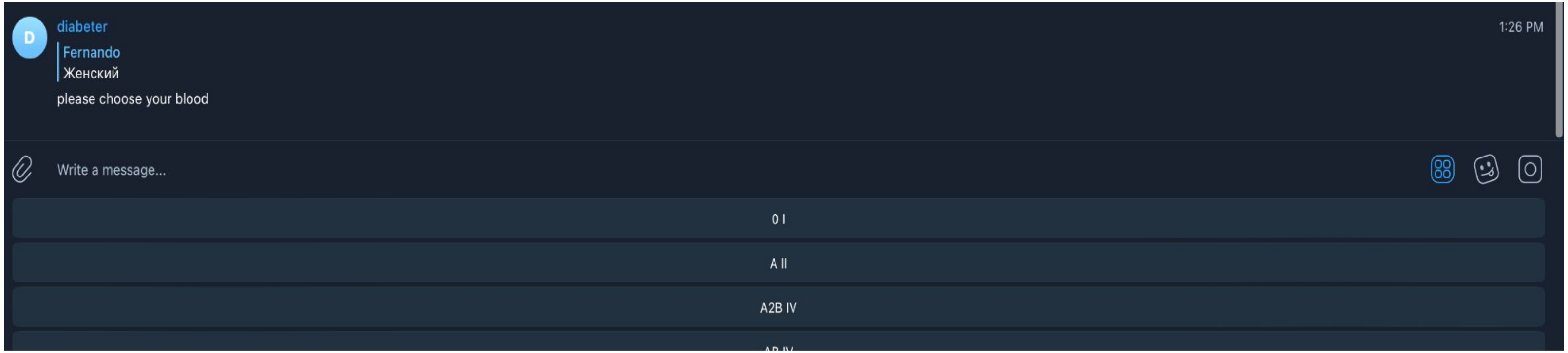
/query - to offer the name of code

/your input - to showcase the current input

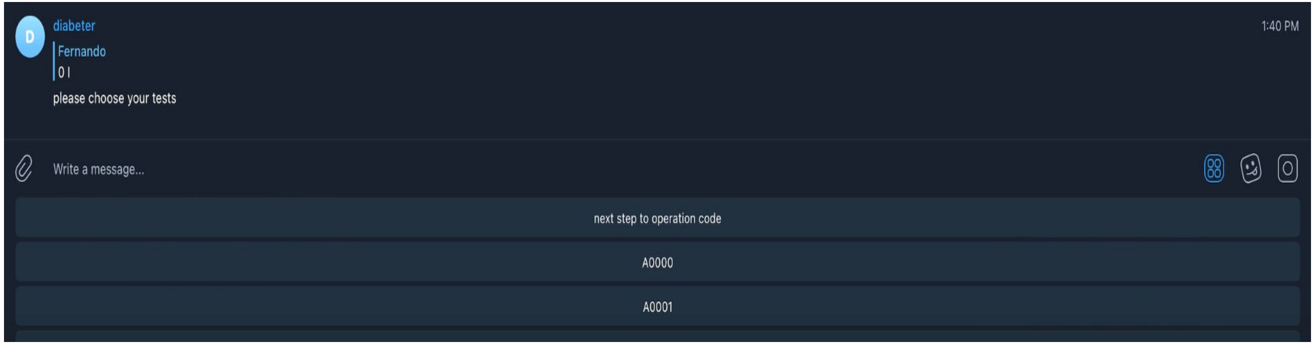
/feedback - to assess the prediction in order to trace the accuracy

demo image


feedback
please write patient number
1:23 PM
Fernando
1997
1:24 PM
diabeter
What is your age , please write number
1:24 PM
Fernando
45
1:24 PM
diabeter
Fernando
45
please choose your gender
1:24 PM
Write a message...
Женский
Мужской


diabeter
Fernando
Женский
please choose your blood
1:26 PM
Write a message...
0 I
A II
A2B IV

demo image

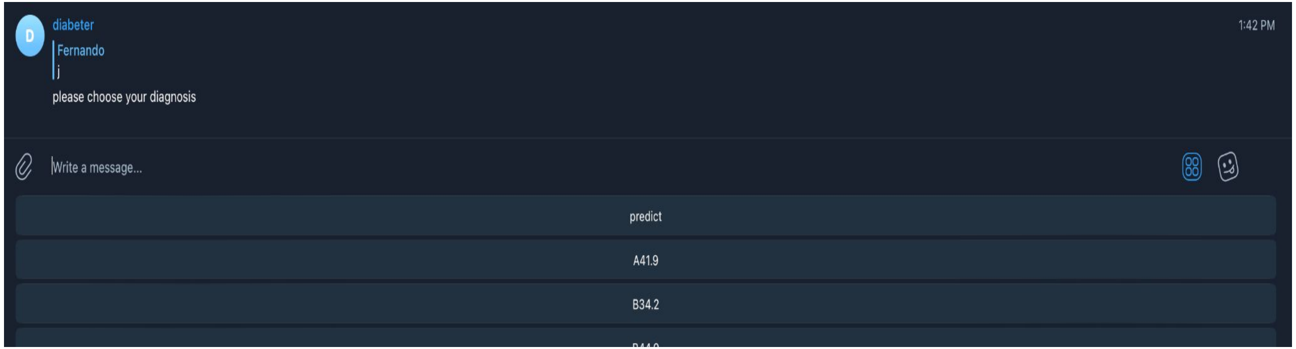


diabeter
1:40 PM
Fernando
0 I
please choose your tests
Write a message...
next step to operation code
A0000
A0001

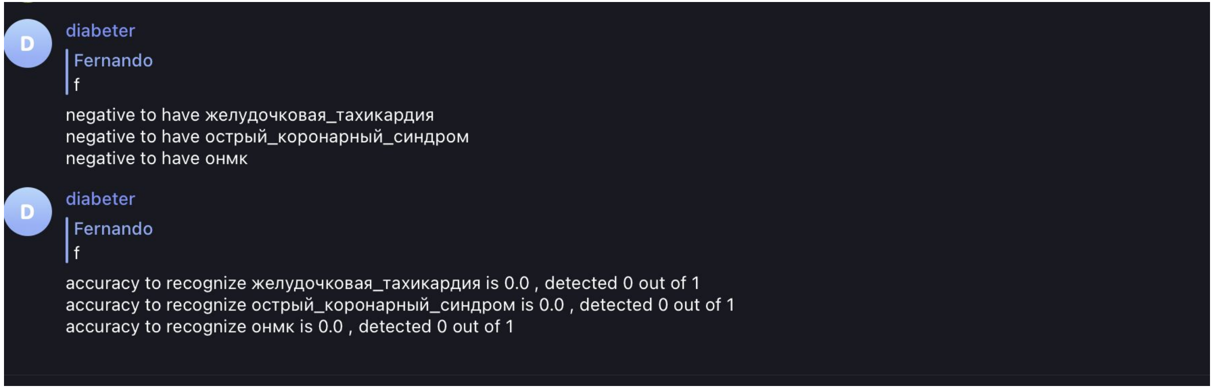
diabeter
1:40 PM
Fernando
A0000
please input the number of your test A0000
Fernando
1:40 PM
24

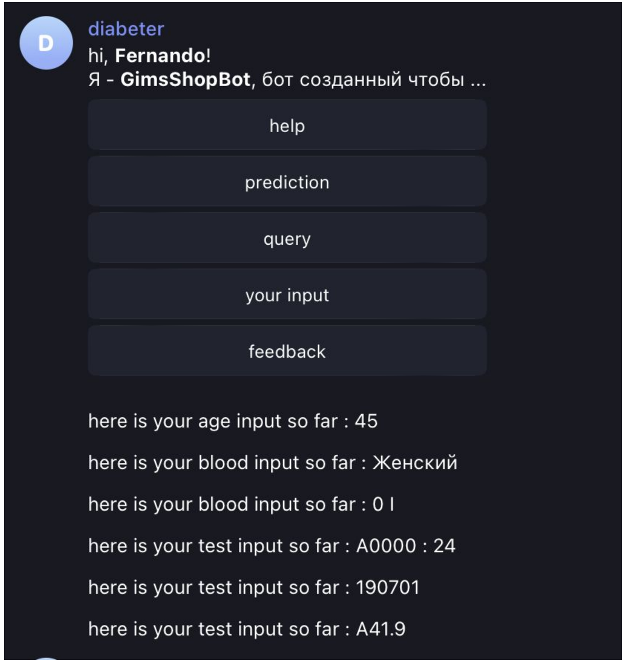
demo image


diabeter
Fernando
f
please choose your operations
1:41 PM
Write a message...
next step to diagnosis code
190701
190705

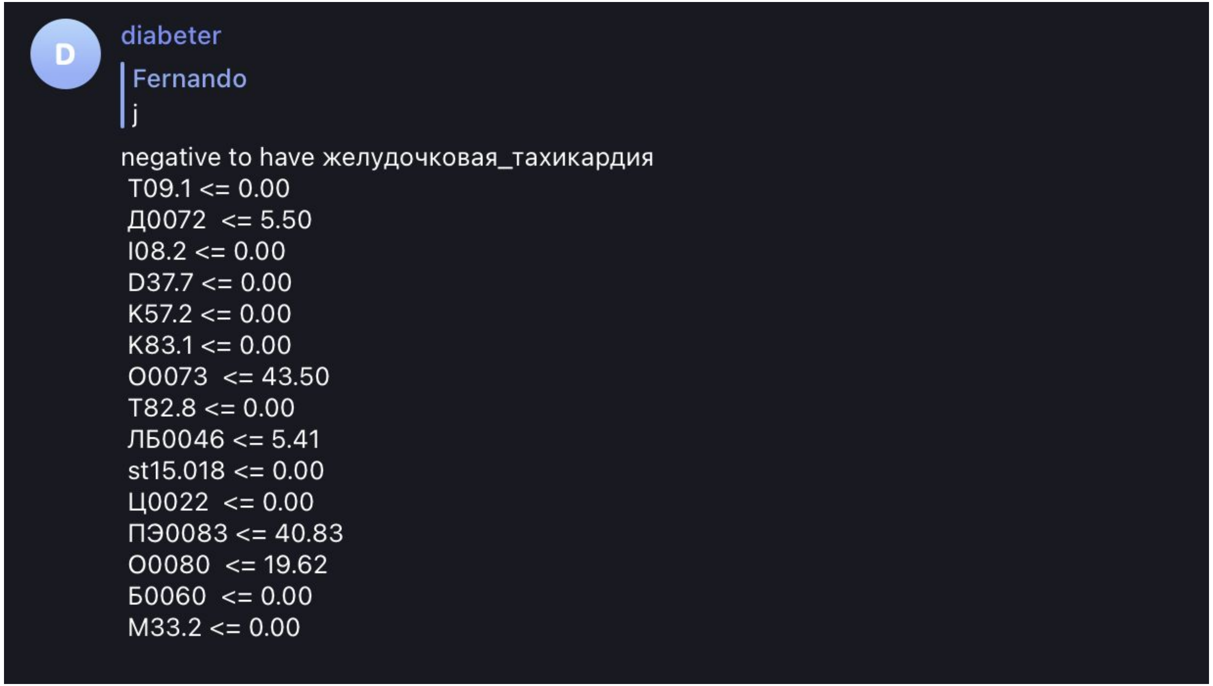

diabeter
Fernando
j
please choose your diagnosis
1:42 PM
Write a message...
predict
A41.9
B34.2

demo image


diabeter
Fernando
f
negative to have желудочковая_тахикардия
negative to have острый_коронарный_синдром
negative to have онмк
diabeter
Fernando
f
accuracy to recognize желудочковая_тахикардия is 0.0 , detected 0 out of 1
accuracy to recognize острый_коронарный_синдром is 0.0 , detected 0 out of 1
accuracy to recognize онмк is 0.0 , detected 0 out of 1


diabeter
hi, Fernando!
Я - GimsShopBot, бот созданный чтобы ...
help
prediction
query
your input
feedback
here is your age input so far : 45
here is your blood input so far : Женский
here is your blood input so far : 0 l
here is your test input so far : A0000 : 24
here is your test input so far : 190701
here is your test input so far : A41.9

demo image


D
diabeter
Fernando
j
negative to have желудочковая_тахикардия
T09.1 <= 0.00
Д0072 <= 5.50
I08.2 <= 0.00
D37.7 <= 0.00
K57.2 <= 0.00
K83.1 <= 0.00
O0073 <= 43.50
T82.8 <= 0.00
ЛБ0046 <= 5.41
st15.018 <= 0.00
Ц0022 <= 0.00
ПЭ0083 <= 40.83
O0080 <= 19.62
Б0060 <= 0.00
M33.2 <= 0.00

furthre improvement

1) advice of treatment
2) implement CDS hooks
3) implement censorship to input in order to prevent abnormal input
4) implement more clear explanation instead of code

End

(thank you very much)

Ma ChengYuan