CCDS

(prediction-oriented)

Ma ChengYuan

CCDS

interface : telegram

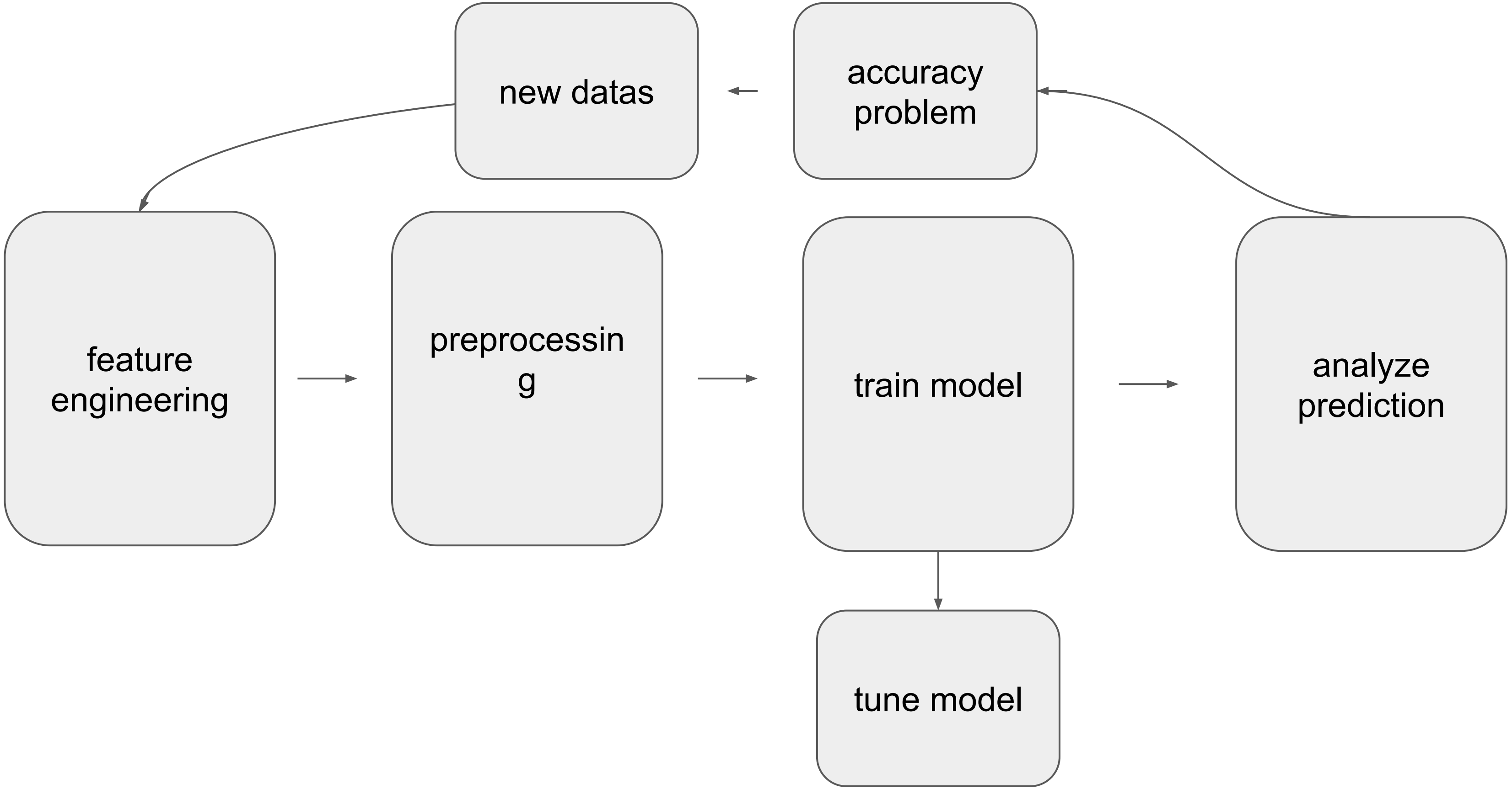
language : python

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

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

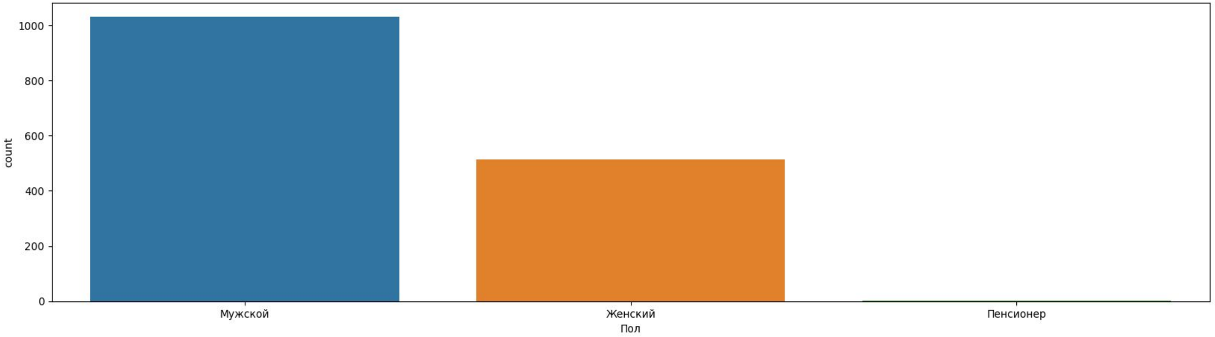
basic data distribution

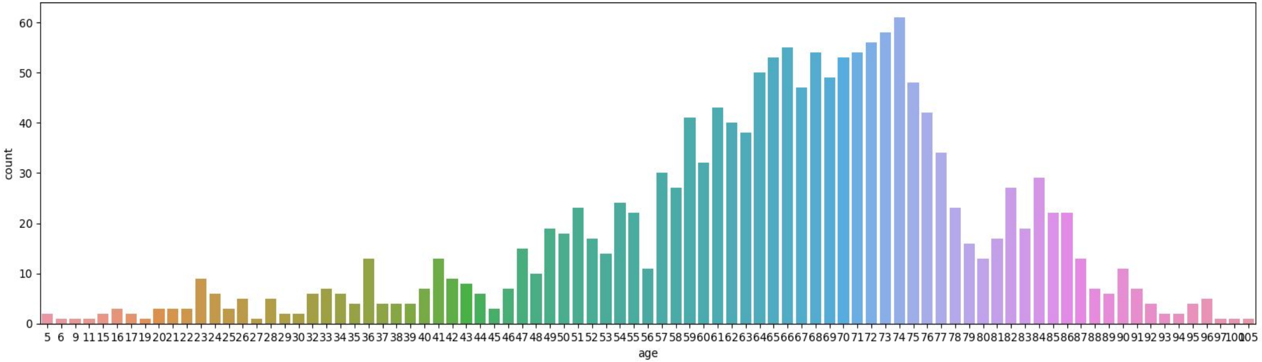
order by target as below :

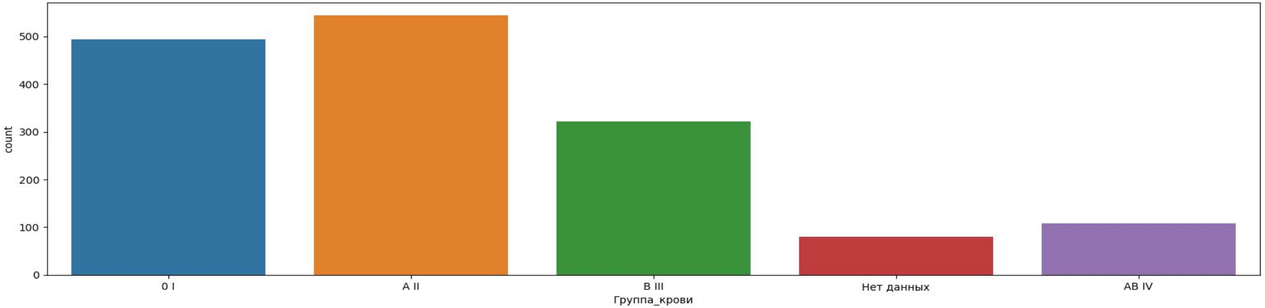
- желудочковая_тахикардия
- острый_коронарный_синдром
- медиастинит
- онмк

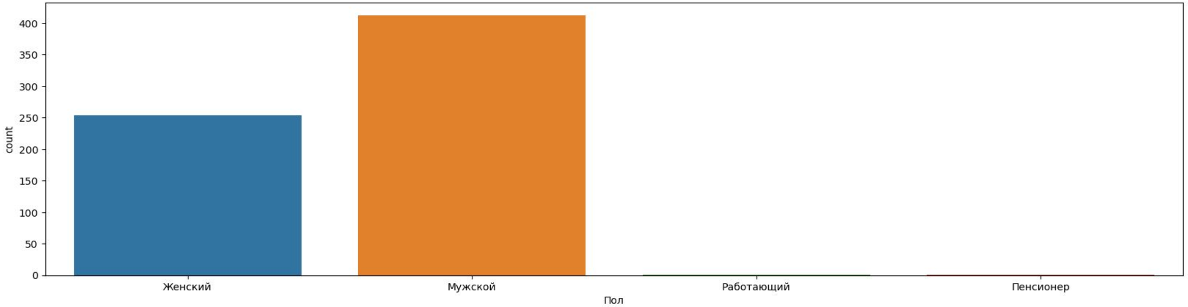
graph showed from top to bottom:

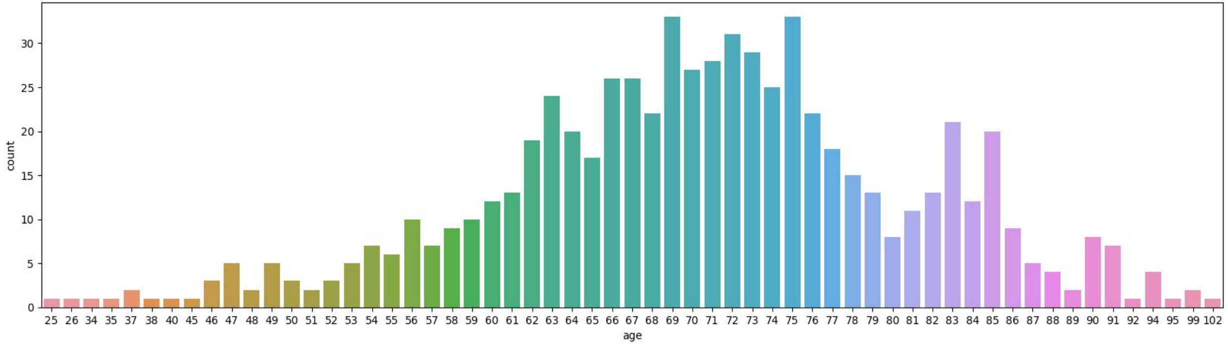
- gender
- age
- blood type

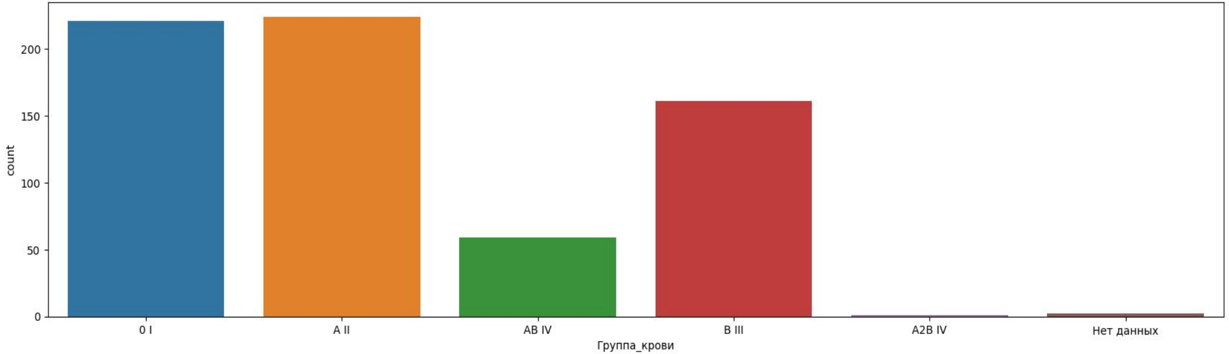

count
0
200
400
600
800
1000
Мужской
Женский
Пенсионер
Пол

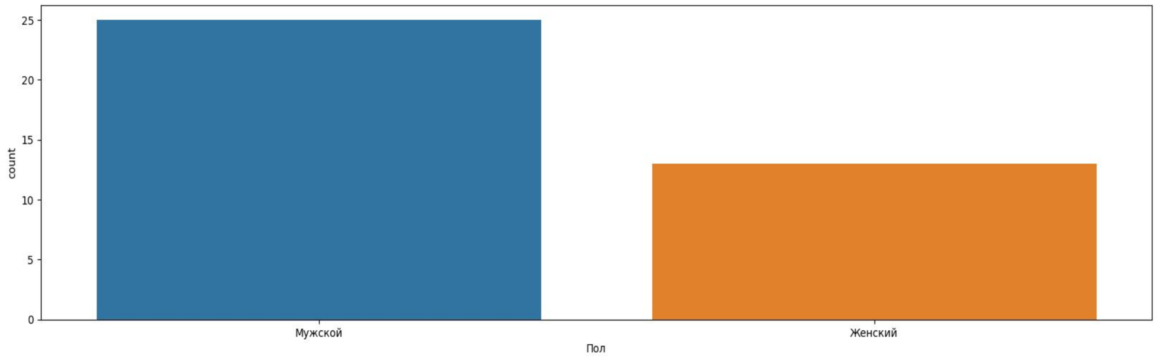

count
0
10
20
30
40
50
60
5 6 9 11 15 16 17 19 20 21 22 23 24 25 26 27 28 29 30 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 100 105
age

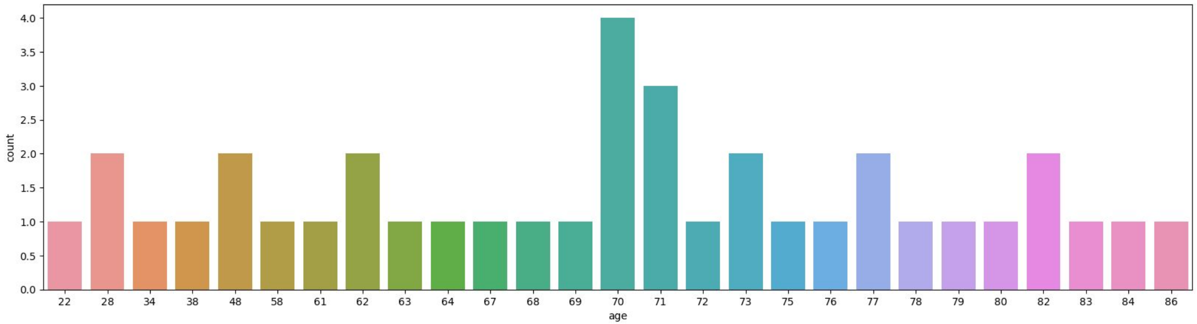

count
0
100
200
300
400
500
0 I
A II
B III
Нет данных
AB IV
Группа_крови

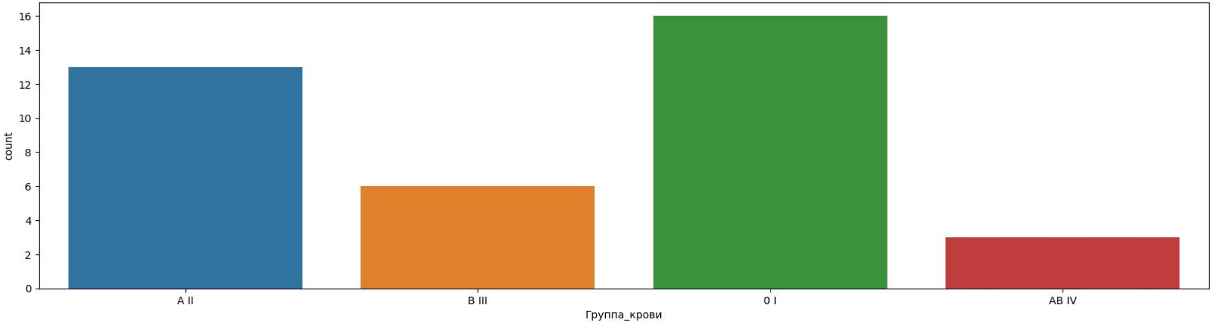

count
0
50
100
150
200
250
300
350
400
Женский
Мужской
Работающий
Пенсионер
Пол

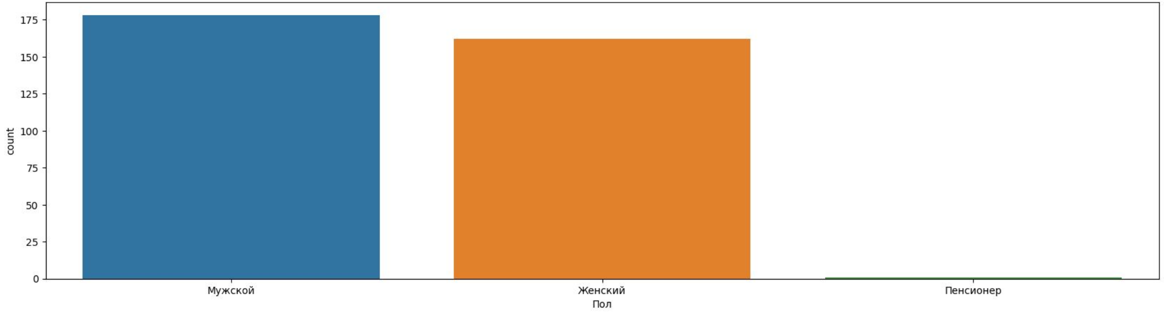

count
0
5
10
15
20
25
30
25 26 34 35 37 38 40 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 94 95 99 102
age


count
0
50
100
150
200
0 I
A II
AB IV
B III
A2B IV
Нет данных
Группа_крови

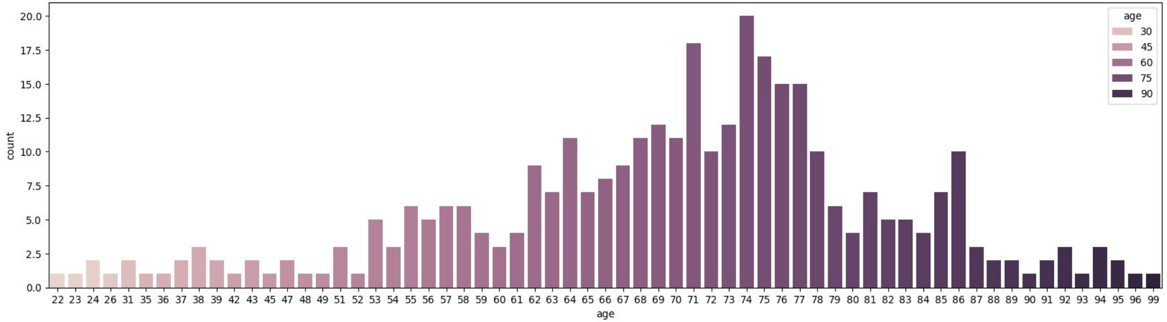

25
20
15
10
5
0
count
Мужской
Женский
Пол


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
count
22
28
34
38
48
58
61
62
63
64
67
68
69
70
71
72
73
75
76
77
78
79
80
82
83
84
86
age

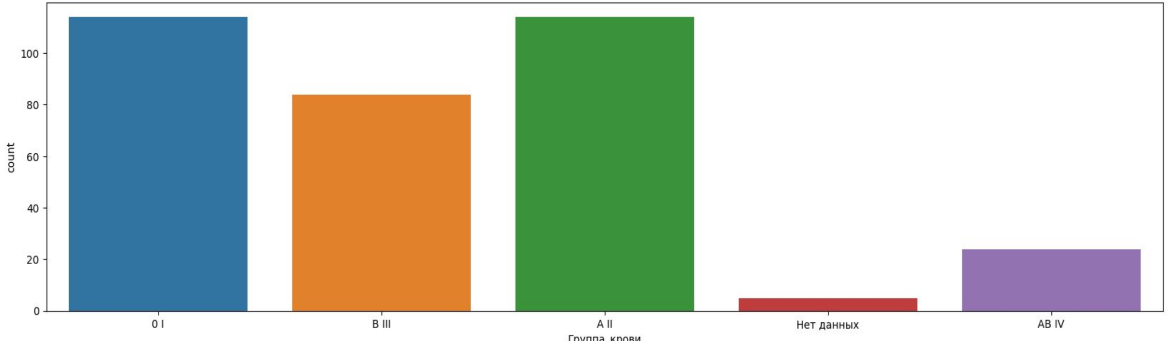

16
14
12
10
8
6
4
2
0
count
A II
B III
0 I
AB IV
Группа_крови



175
150
125
100
75
50
25
0
count
Мужской
Женский
Пенсионер
Пол



20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
count
age
30
45
60
75
90
22 23 24 26 31 35 36 37 38 39 42 43 45 47 48 49 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 99
age



100
80
60
40
20
0
count
0 I
B III
A II
Нет данных
AB IV
Группа крови

preprocessing

1. select potential columns from each df
2. find target disease from epiz_inform_stationary_risks_10_events , and insert extra column to classify patient's disease
3. convert categorization value to numerical value using one-hot encoding
4. merge dfs

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

2. clinical_diag_293_strat_risk:
  - Код_МЭС
  - Клинический_диагноз_рубрика

training result(1)

1)

row x column :
4909x3121

желудочковая_тахикардия
: 1359

острый_коронарный_синдром
: 595

медиастинит
: 36

онмк
: 340

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

training result(2)

2)

row x column :
4909x3435

желудочковая_тахикардия
: 1359

острый_коронарный_синдром
: 595

медиастинит
: 36

онмк
: 340

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

training 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

training 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

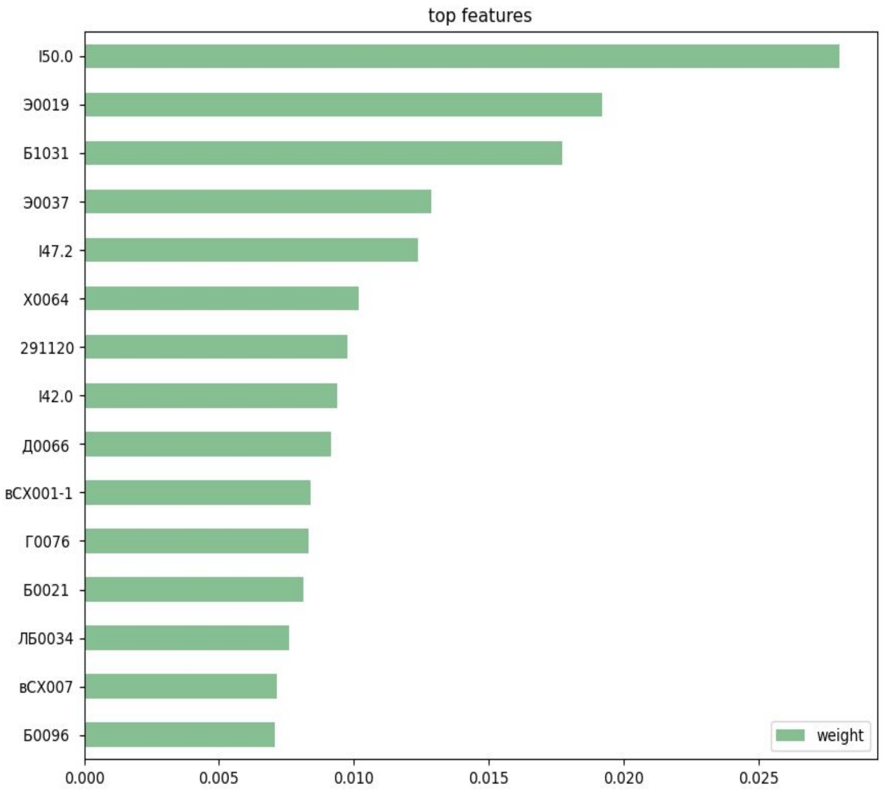
медиастинит :

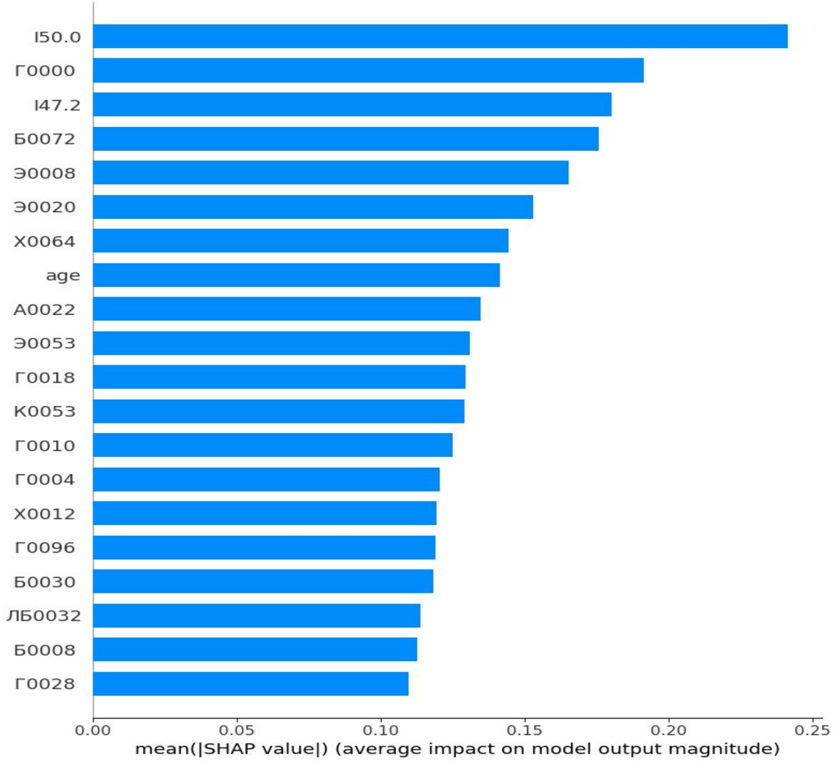
xgb : f1 : 0.498126

онмк :

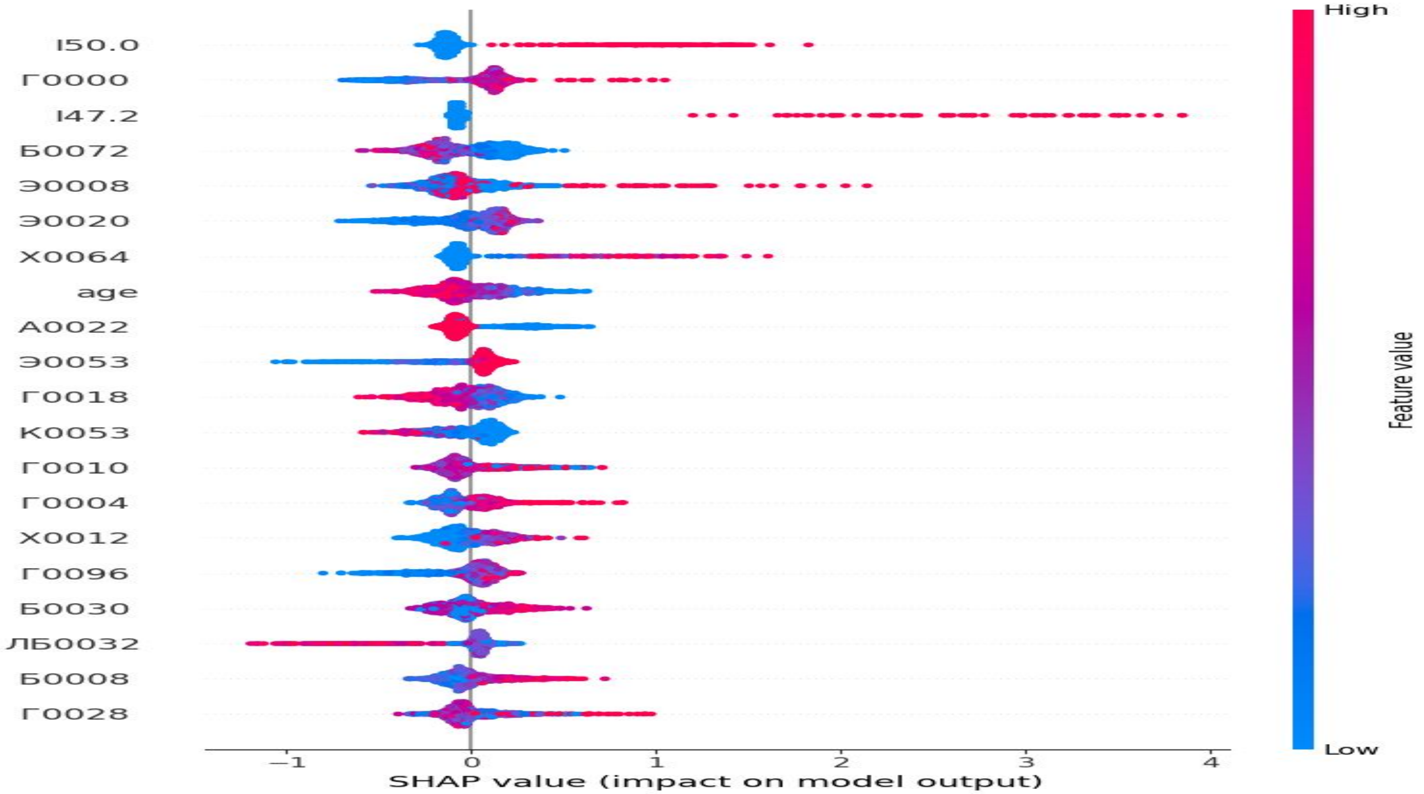
xgb : f1 : 0.599690

feature importance (left xgb , right shap)

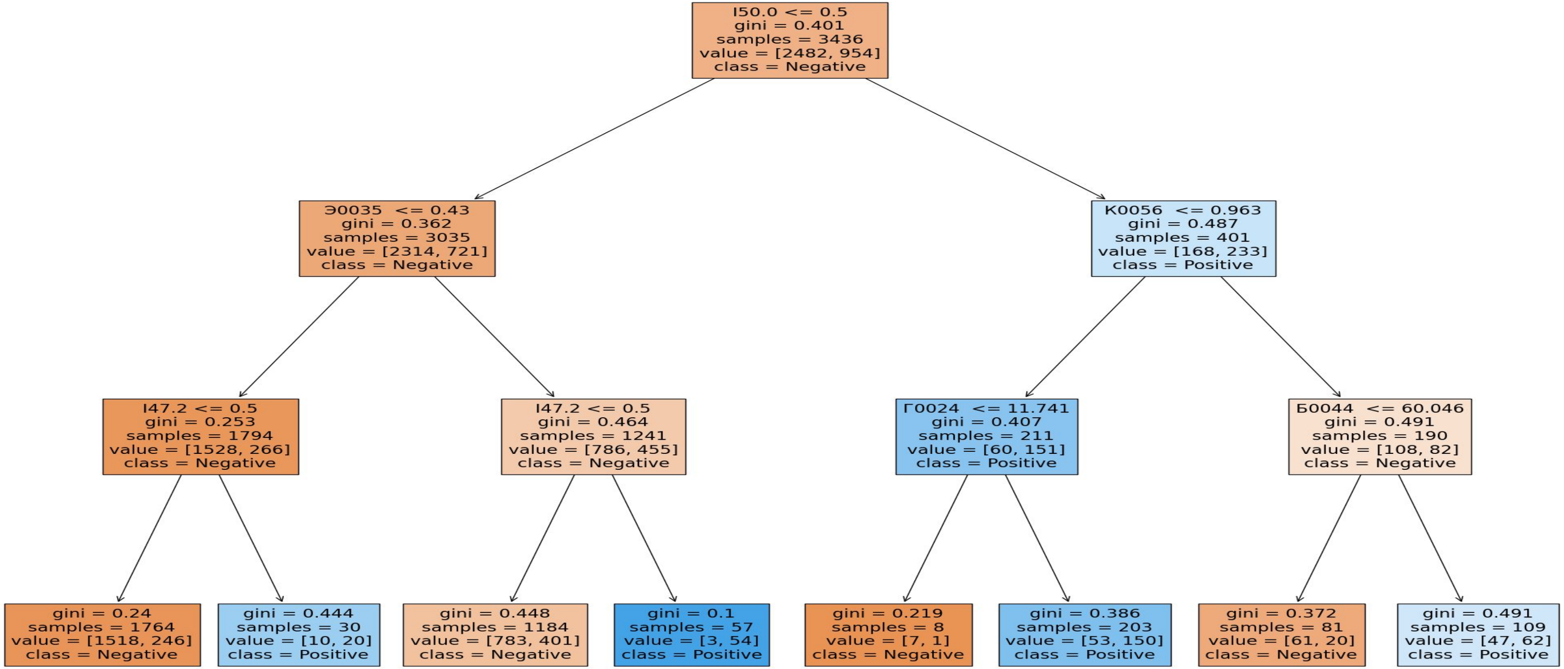




feature importance (shap)(black box model)



feature importance (decision tree)

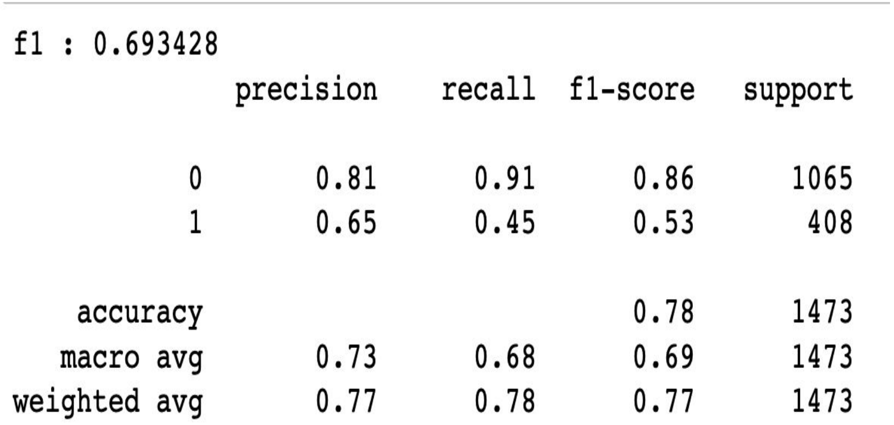


model training with feature importances(shap)

extracting columns with more thna 0 shap value

original :

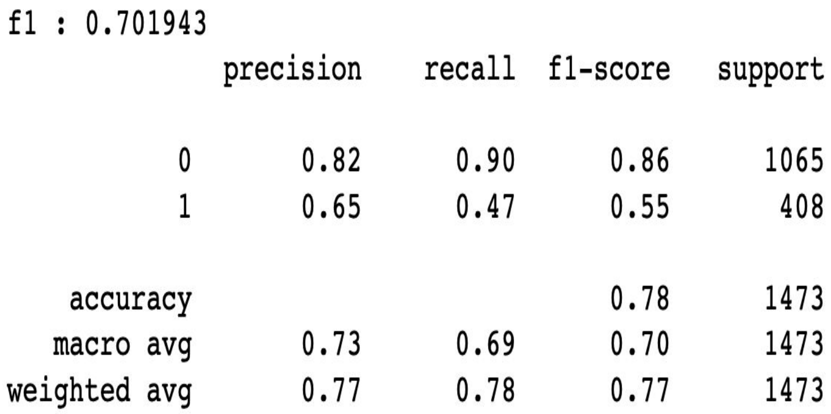
after :



```
f1 : 0.693428
              precision    recall  f1-score   support

           0       0.81      0.91      0.86      1065
           1       0.65      0.45      0.53       408

    accuracy                           0.78      1473
   macro avg       0.73      0.68      0.69      1473
weighted avg       0.77      0.78      0.77      1473
```



```
f1 : 0.701943
              precision    recall  f1-score   support

           0       0.82      0.90      0.86      1065
           1       0.65      0.47      0.55       408

    accuracy                           0.78      1473
   macro avg       0.73      0.69      0.70      1473
weighted avg       0.77      0.78      0.77      1473
```

feature selection(L2 regularization)

желудочковая_тахикардия :
total features: 3117

selected features: 963

f1 : 0.704869

острый_коронарный_синдром :
total features: 3117

selected features: 811

f1 : 0.743238

медиастинит :
total features: 3117

selected features: 696

f1 : 0.498297

онмк :
total features: 3117

selected features: 852

f1 : 0.593904

compared with original from test(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(pearson)

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

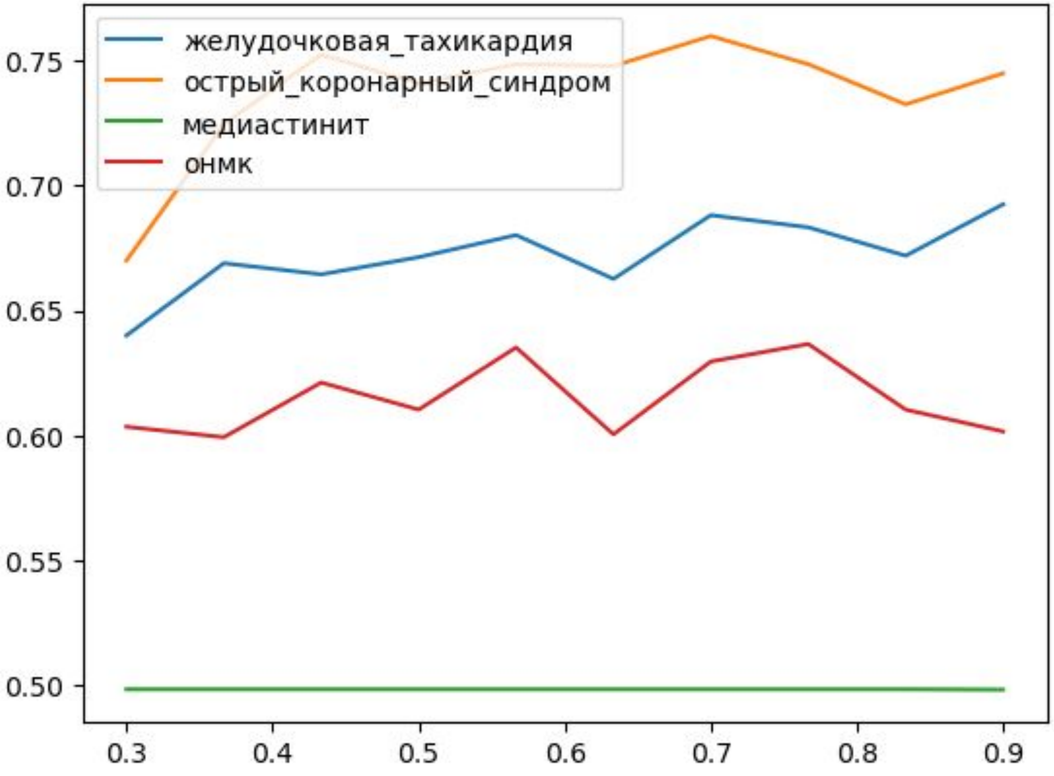
Y : F1 score

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

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

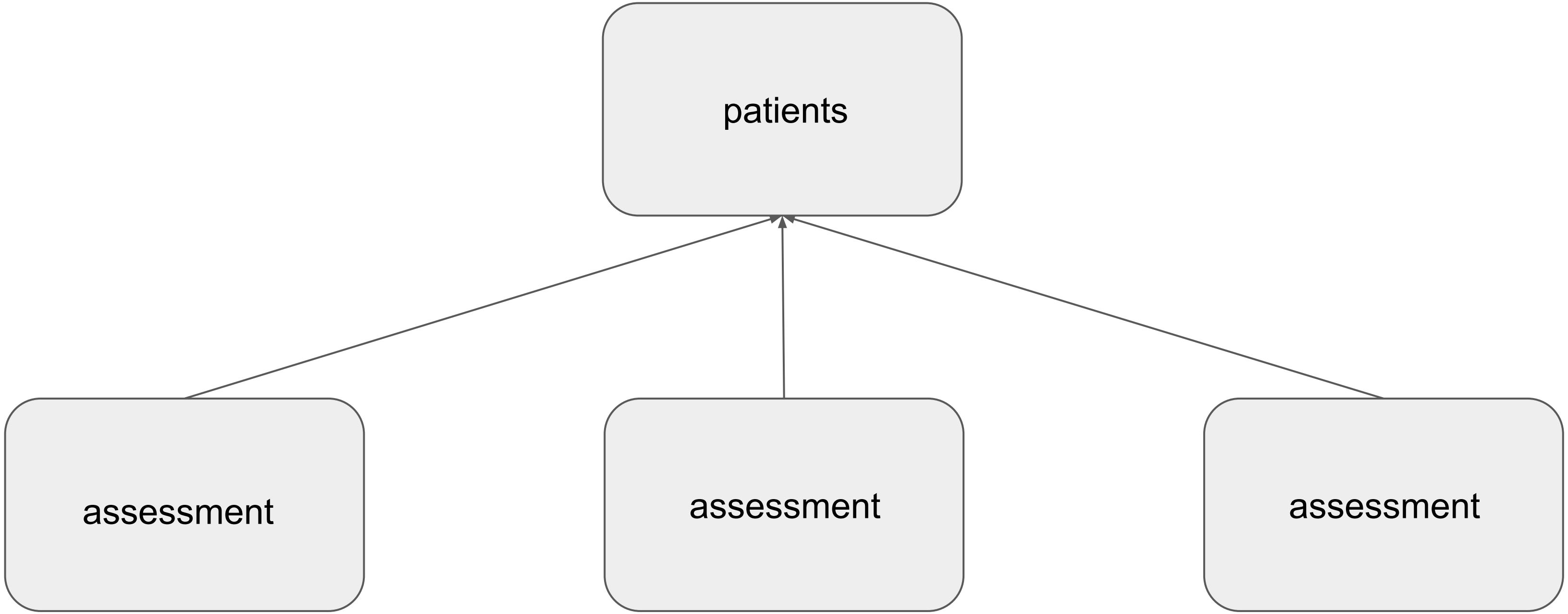
link : https://www.sciencedirect.com/science/article/abs/pii/S0003497520311565

онмк :

Performance Analysis of Machine Learning Approaches in Stroke Prediction

link: https://ieeexplore.ieee.org/abstract/document/9297525?casa_token=TfM_OTlj2BEAAAAA:vV39yNcKMpzQc9jI_oopWu0eggmUj9CRoMETefwiKE7d3W07qChFVgS8HmEnqhtRvggkcX0FChDokA

database design



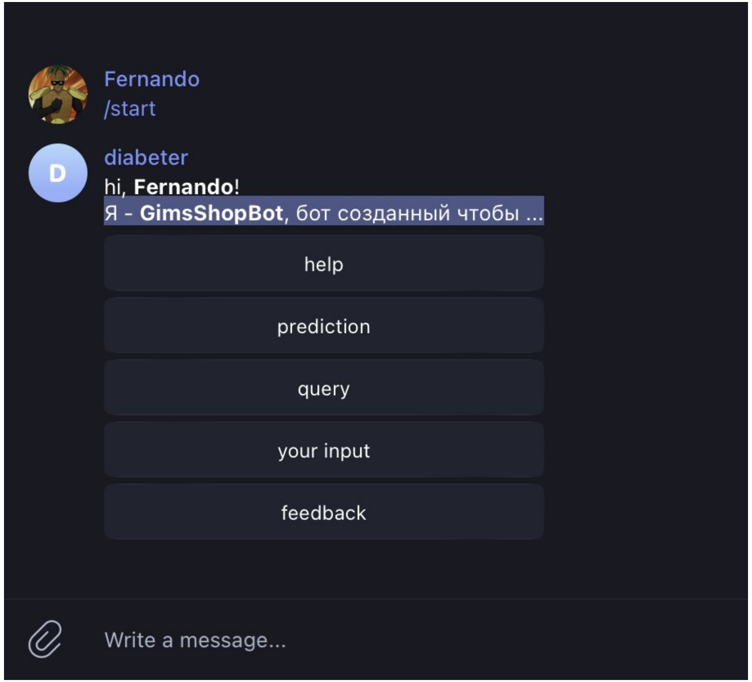
patients

assessment

assessment

assessment

demo image



Fernando
/start

diabeter
hi, **Fernando**!
Я - **GimsShopBot**, бот созданный чтобы ...

/help - to offer instruction

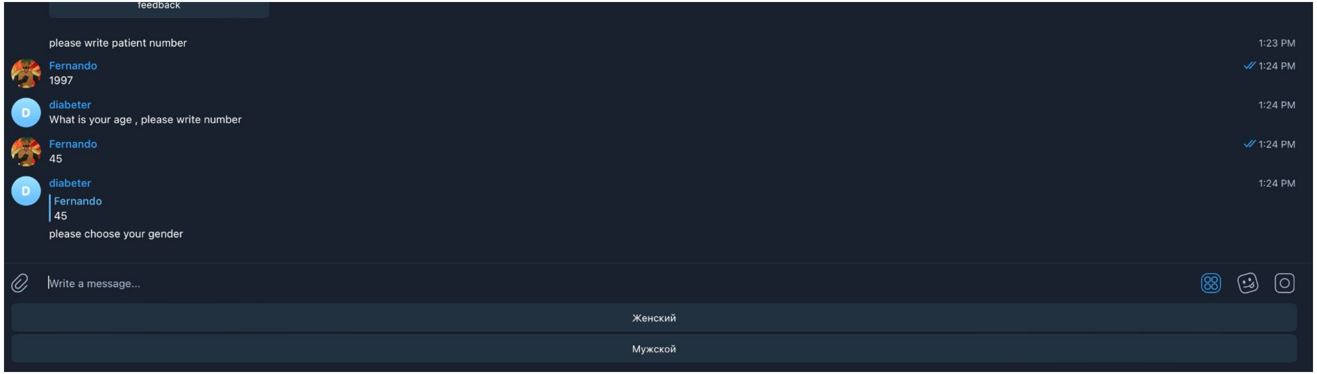
/prediction - main functionality to predict

/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



please write patient number 1:23 PM

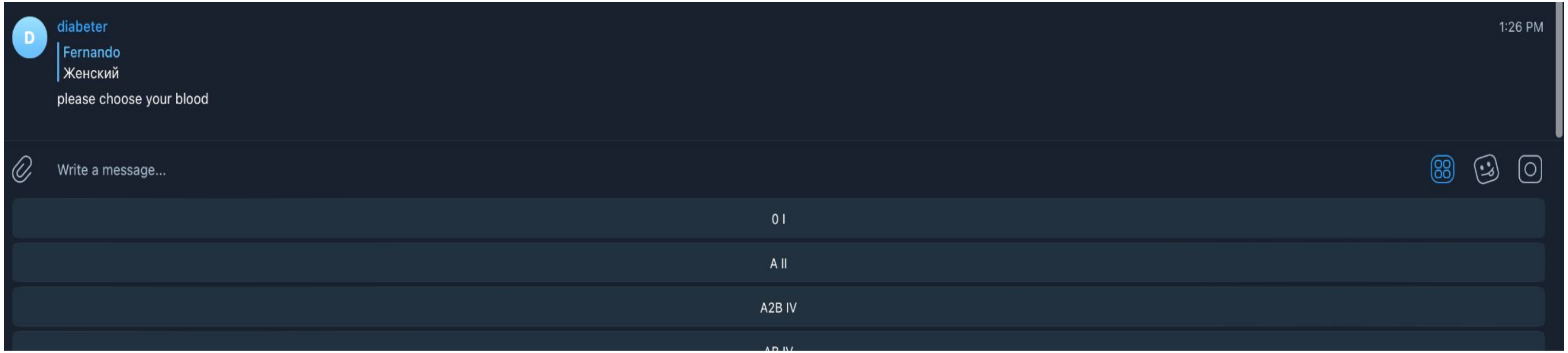
Fernando 1:24 PM
1997

diabeter 1:24 PM
What is your age , please write number

Fernando 1:24 PM
45

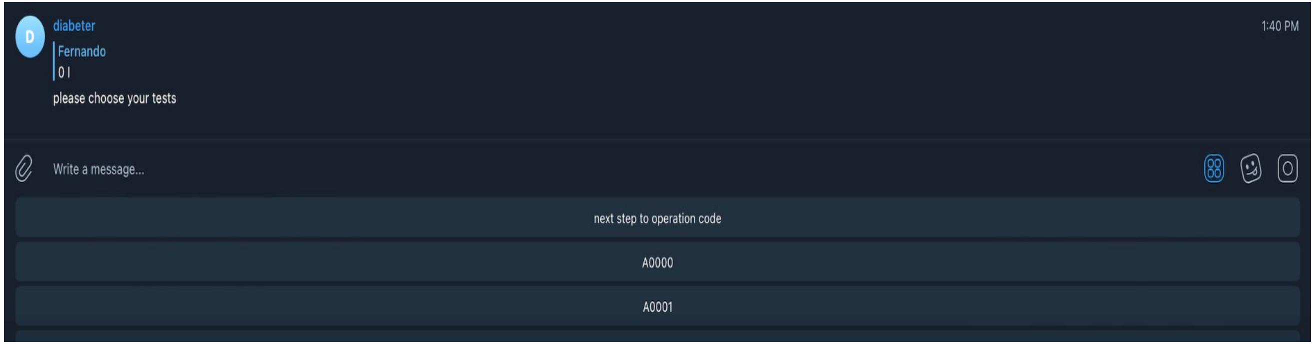
diabeter 1:24 PM
> Fernando
> 45

please choose your gender


diabeter 1:26 PM
> Fernando
> Женский

please choose your blood

demo image



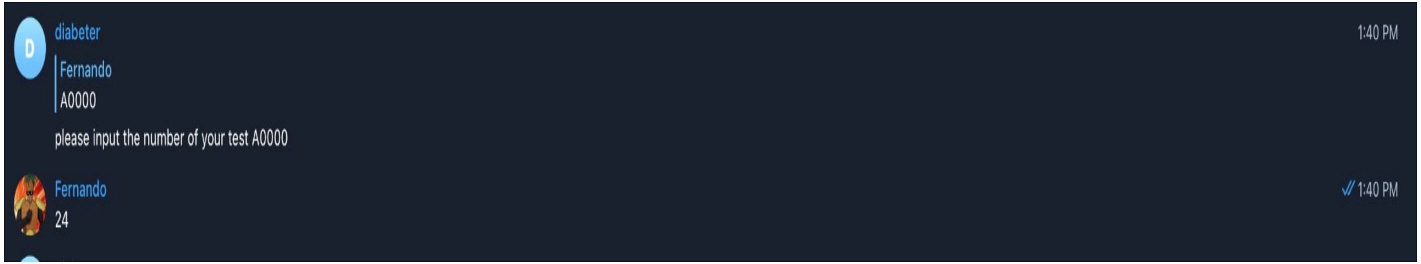
diabeter

Fernando

0 I

please choose your tests

1:40 PM



diabeter

Fernando

A0000

please input the number of your test A0000

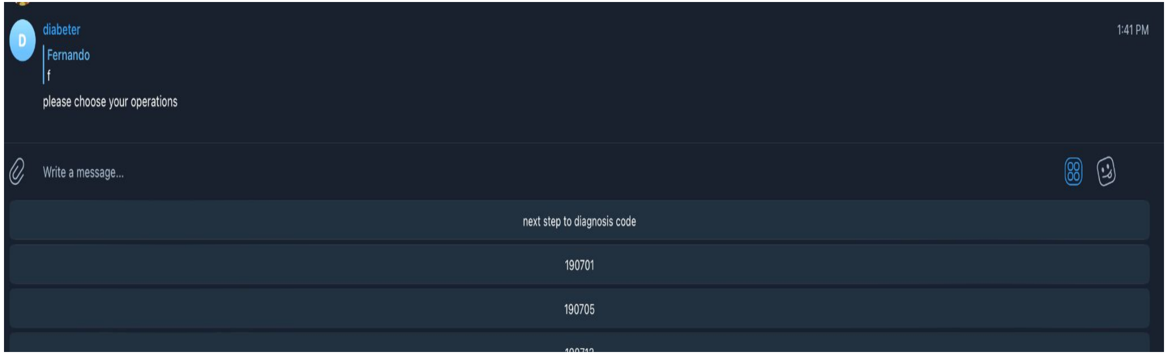
1:40 PM

Fernando

24

1:40 PM

demo image



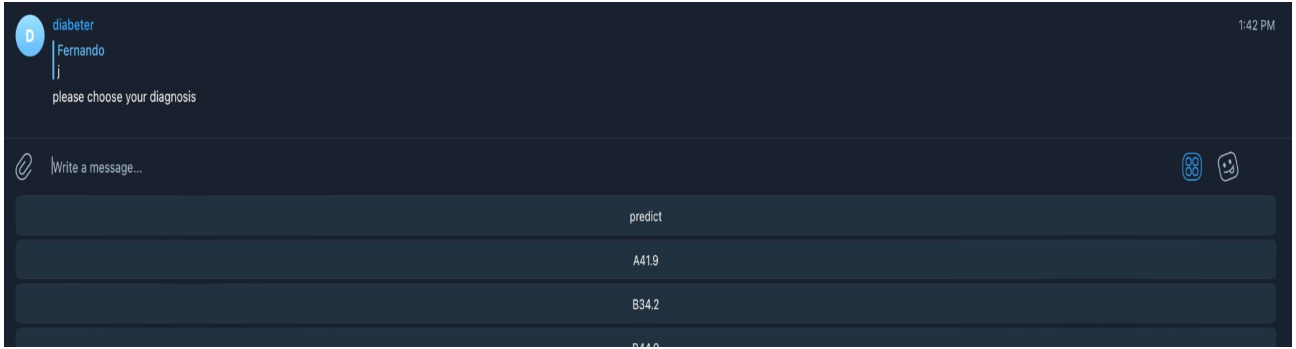
diabeter

Fernando

f

please choose your operations

1:41 PM



diabeter

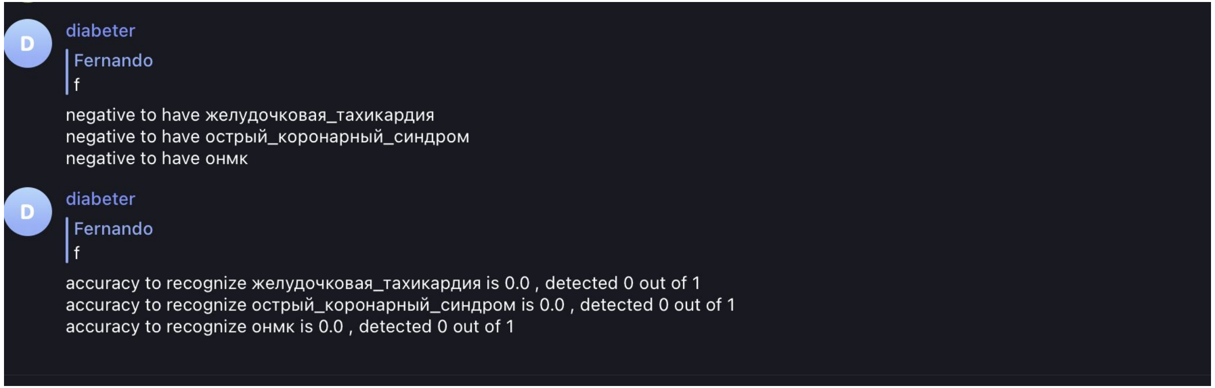
Fernando

j

please choose your diagnosis

1:42 PM

demo image



diabeter

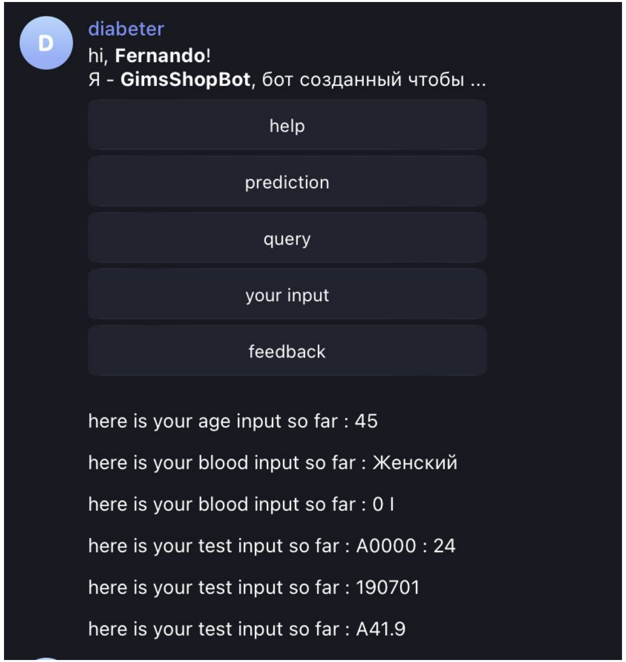
Fernando
f

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 ...

here is your age input so far : 45

here is your blood input so far : Женский

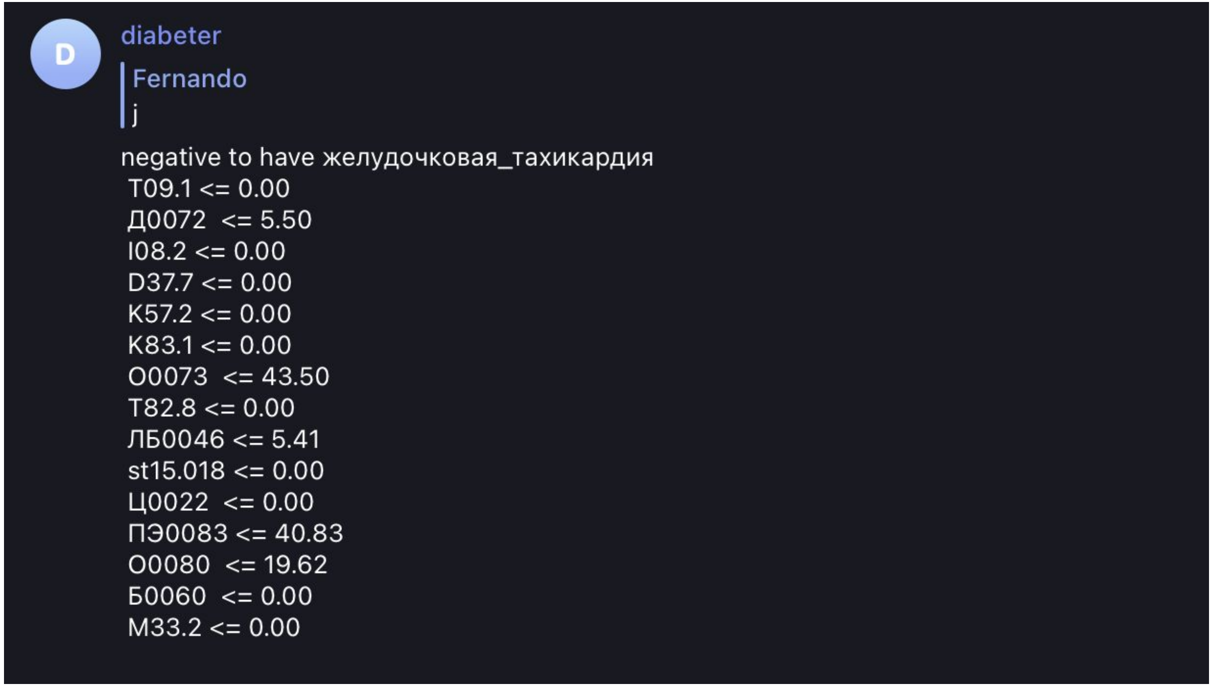
here is your blood input so far : 0 l

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



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
 О0073  <= 43.50
 T82.8 <= 0.00
 ЛБ0046 <= 5.41
 st15.018 <= 0.00
 Ц0022  <= 0.00
 ПЭ0083 <= 40.83
 О0080  <= 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
5) feature selection , remove correlated columns

End

(thank you very much)

Ma ChengYuan