

Predictive Modeling

Ma ChengYuan

feature info

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

patients number by target (ratio for class) :

желудочковая_тахикардия : 1723 (1:0.38)

острый_коронарный_синдром : 712 (1:0.14)

медиастинит : 46

онмк : 437 (1:0.07)

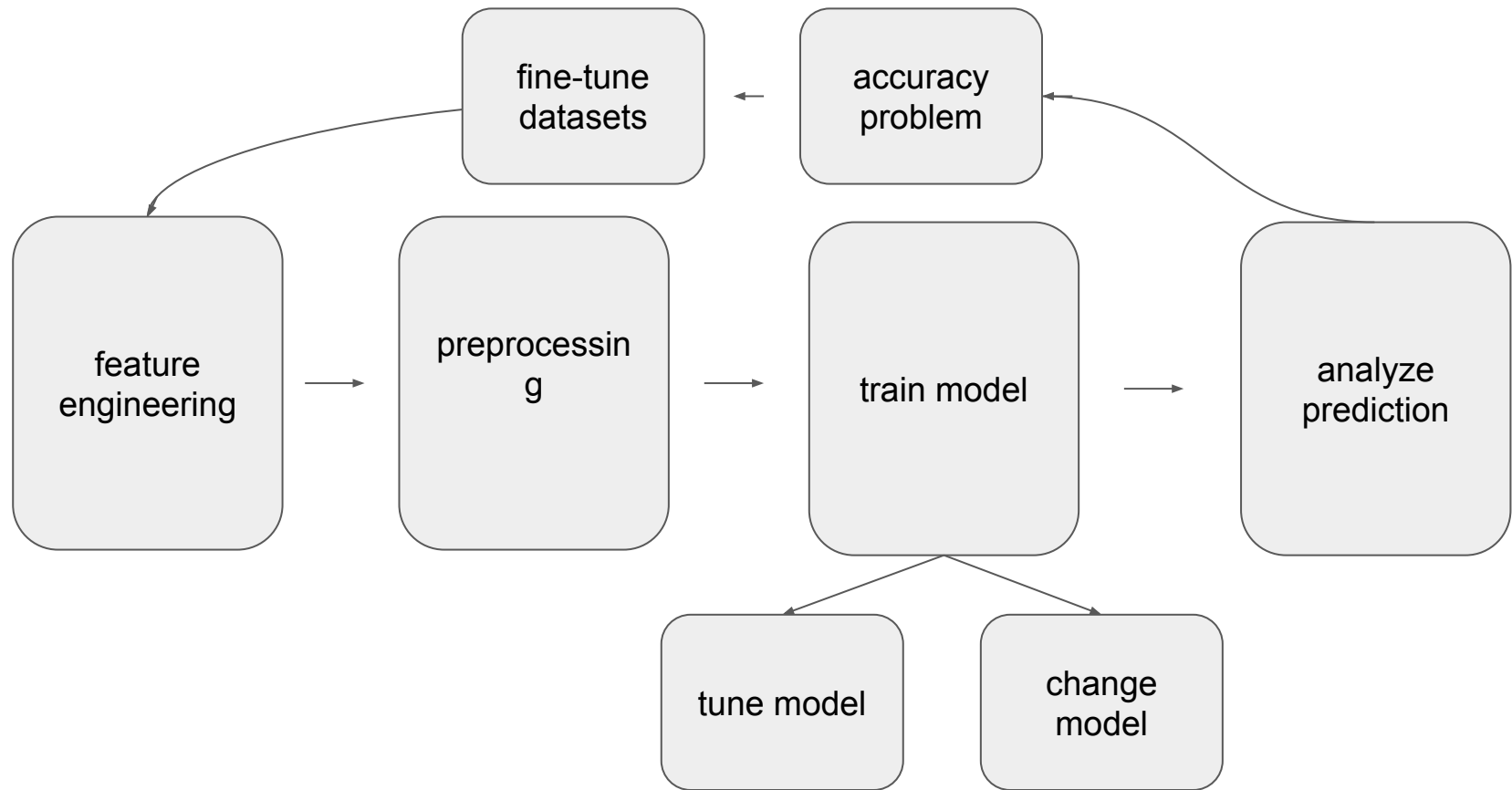
feature info :

Клинический_диагноз_рубрика : 723

Код_МЭС : 611

Код_теста : 2469

pipeline





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_analysis_risk_stratif:

- Код_теста 
- 1) fill null cell in col('Значение_число') with average value from criteria $((\text{lower} + \text{upper}) / 2)$

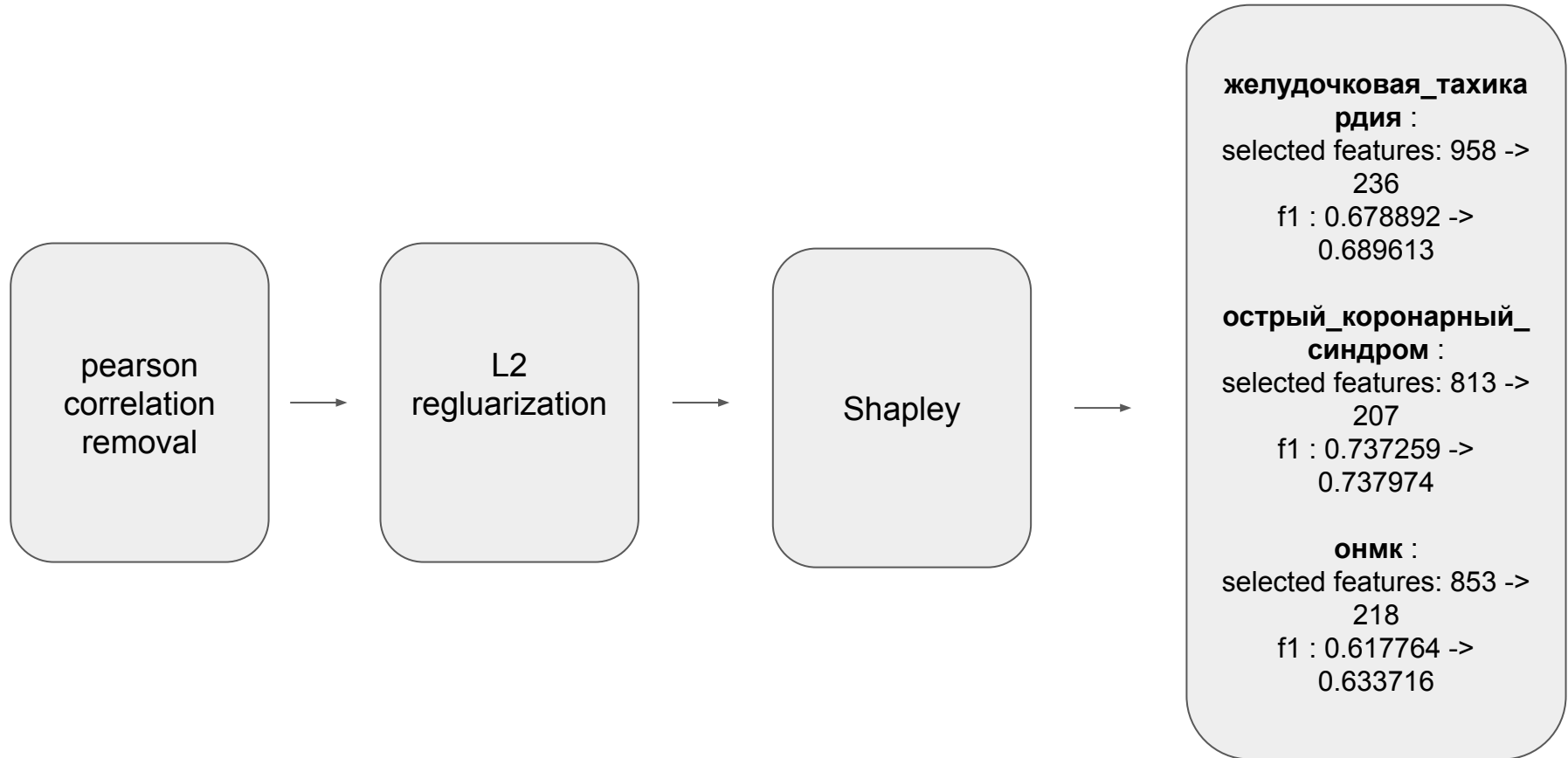
clinical_diag_293_strat_risk: 

- Код_МЭС
- Клинический_диагноз_рубрика

test result (f1 metrics)

numerical value using col('Значение_число') with average value i	original	original	oversampling	oversampling
	xgb	neural network	xgb	neural network
желудочковая_тахикардия	0.693428	0.567878	0.679563	0.566822
острый_коронарный_синдром	0.759434	0.592496	0.754864	0.574827
онмк	0.59969	0.510228	0.618609	0.420461
classified into lower & higher of criteria				
	original	original	oversampling	oversampling
желудочковая_тахикардия	0.6889	0.671587	0.694334	0.636534
острый_коронарный_синдром	0.6911	0.725524	0.684404	0.702459
онмк	0.602336	0.589569	0.589569	0.602206
without value in null cell				
	xgb	lgb	decision tree	catboost
желудочковая_тахикардия	0.589569	0.640268	0.686466	0.667132
острый_коронарный_синдром	0.754944	0.652398	0.768578	0.743238
онмк	0.619066	0.6599	0.614116	0.536594

Feature Extraction Method



Final extration with detail (XGB)

Disease	желудочковая_тахикардия	острый_коронарный_синдром	ОНМК
Original features	958	813	853
Current features	236	207	218
Personal info	age', 'Пол', 'O I'	B III', 'age', 'A II'	Пол', 'age', 'O I', 'A II'
Test feature	210	189	204
Operation feature	8	9	7
Diagnosis feature	15	6	3
F1	0.689613	0.737874	0.633716
Before F1	0.678892	0.737259	0.61776

metrics before and after oversampling(training class ratio)

желудочковая тахикардия (1:0.39) :

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

острый коронарный синдром (1:0.14) :

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

ОНМК (1:0.08):

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

after oversampling : (1:1)

Counter({0: 2482, 1: 2482})				
f1 : 0.687290				
	precision	recall	f1-score	support
0	0.82	0.86	0.84	1065
1	0.58	0.50	0.54	408
accuracy			0.76	1473
macro avg	0.70	0.68	0.69	1473
weighted avg	0.75	0.76	0.75	1473

roc auc score is : 0.7515626438368775

(1:1)

Counter({0: 3010, 1: 3010})				
f1 : 0.739423				
	precision	recall	f1-score	support
0	0.93	0.97	0.95	1304
1	0.64	0.45	0.53	169
accuracy			0.91	1473
macro avg	0.79	0.71	0.74	1473
weighted avg	0.90	0.91	0.90	1473

roc auc score is : 0.8710022870003993

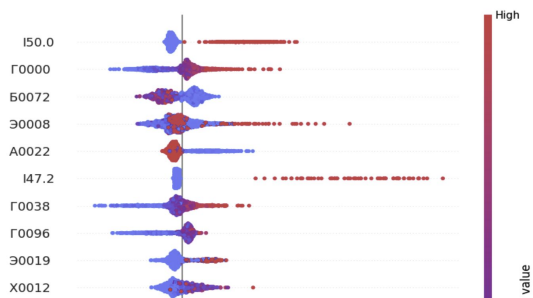
(1:1)

Counter({1: 3194, 0: 3194})				
f1 : 0.606814				
	precision	recall	f1-score	support
0	0.94	0.99	0.97	1375
1	0.52	0.16	0.25	98
accuracy			0.93	1473
macro avg	0.73	0.58	0.61	1473
weighted avg	0.91	0.93	0.92	1473

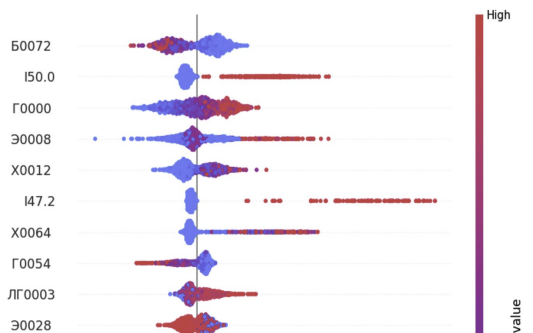
roc auc score is : 0.7438070500927643

shapley metrics

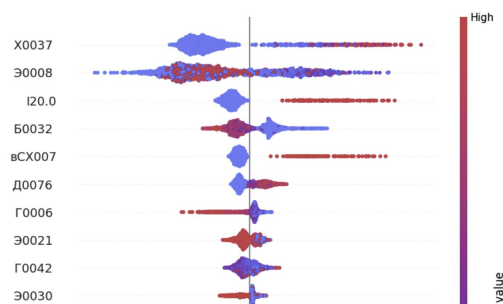
желудочковая тахикардия :



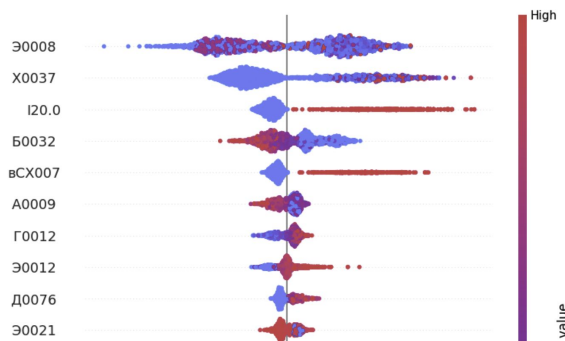
after oversampling : **match(7 / 10)**



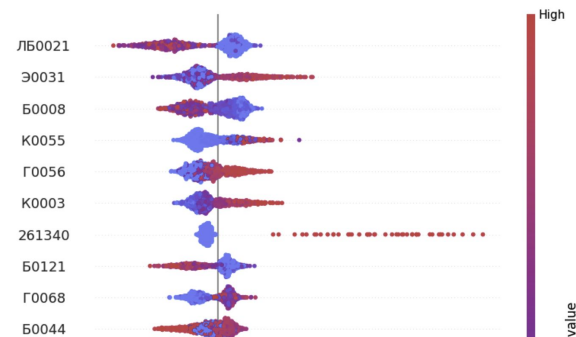
острый коронарный синдром :



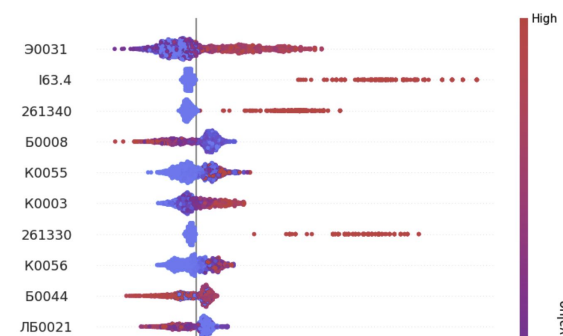
(6 / 10)



ОНМК :



(7 / 10)



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_oopWu0eggmUj9CRoMETefwiKE7d3W07qChFVqS8HmEnqhtRvggkcX0FChDokA

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

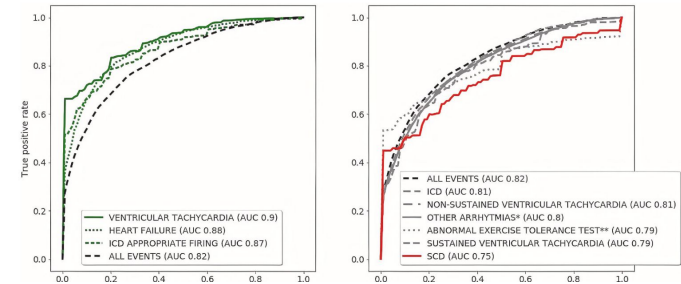
Patient demographic, physical and clinical characteristics. The upper part of the table shows patient overview characteristics and the lower part shows the statistics in terms of baseline and follow-up measurements.

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%)

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 :

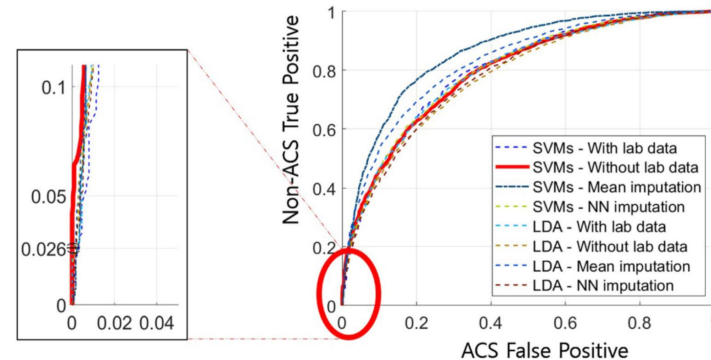
A Machine Learning-Based Approach for the Prediction of Acute Coronary Syndrome Requiring Revascularization

link :

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

Table 1 Features for analysis

		Features
Epidemiological data	1	Gender
	2	Age
	3	Systolic BP
Clinical data at emergency department or outpatient clinic	4	Diastolic BP
	5	HR
	6	CAD
Past medical history before presenting chest pain	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

End
(thank you very much)

Ma ChengYuan