Big Homework Neural FCA Report

Ayub Makhamaev

19.12.2023

1 Heart Disease.

First dataset that we use is "Heart Disease"dataset from `https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset/data`. It has 14 columns and 1025 rows, but after dropping duplicates it has only 302 rows.

Each row represents a person and contains general and medical information about him. And the target variable is the rate of heart disease of this person (0- patient is healthy).

Description of the dataset

column	description	unique values
age	Age in years	41
sex	Male/Female	2
cp	Chest pain type	4
trestbps	Resting blood pressure	49
chol	Serum cholestoral in mg/dl	152
fbs	fasting blood sugar $>$ 120 mg/dl	2
restecg	resting electrocardiographic results	3
thalach	maximum heart rate achieved	91
exang	exercise-induced angina (True/ False)	2
oldpeak	ST depression induced by exercise relative to rest	40
slope	the slope of the peak exercise ST segment	3
ca	number of major vessels (0-3) colored by fluoroscopy	5
thal	0=normal; 1=fixed defect; 2=reversible defect	4
target	heart disease presence	2

Plot the corellation heatmap of features .

Metrics For disease prediction recall is more important than accuracy, because is person gets false-negative classification it is very dangerous for him. So we choose "f1 score"as main metric, but we will also consider recall.

Standard ML baseline First of all let us build baseline with usual ML methods . We will use XGBoost classifier.

We get the confusion matrix on Рис 2.

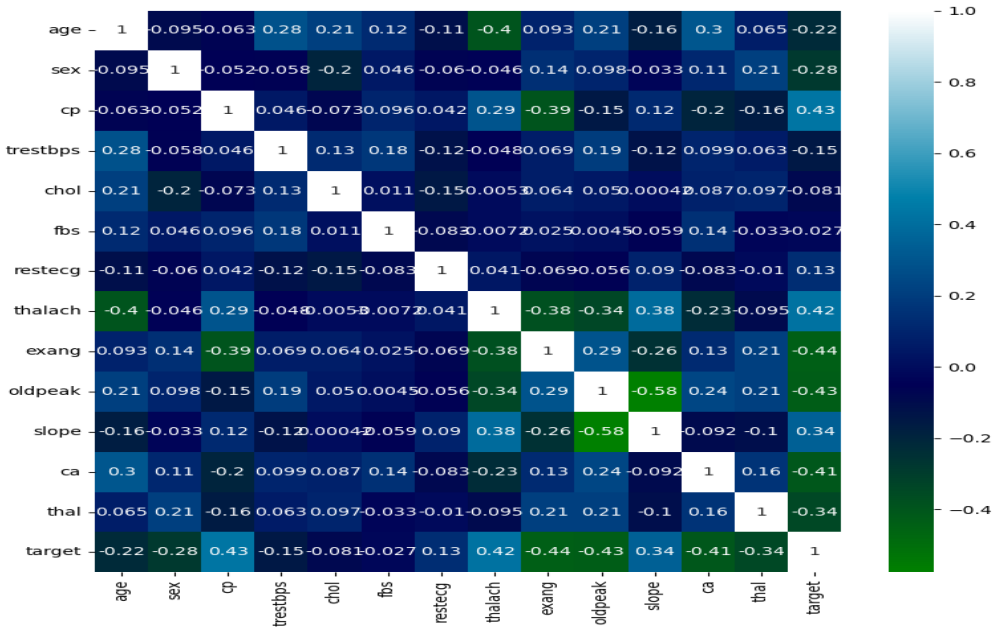

Рис. 1: Corellation heatmap

Some scores for XGBoost Classifier

	f1-score	recall	roc-auc
XGBoost	0.782	0.86	0.723

1.1 Neural FCA

Baseline 1. Lets build first baseline based only on categorical features. We rename categories of categorical columns with their real values (for example for sex 0-"female 1-"male") and one-hot encode them.

Categorical columns: ['sex', 'cp', 'fbs', 'restecg', 'exang', 'thal'].

Data after one-hot encoding and binarizing is on the Рис 3.

Now we build neural network. Choose 4 best concepts based on "f1metric.

Concept lattice and fitted network with edge weights are in Рис 5. and Рис 6.

Prediction scores:

	f1-score	recall	roc-auc
1 baseline	0.8	0.833	0.792

Baseline 2. Now we use all the attributes and scale numerical features using different strategies.

To binarize numerical attributes we will use KBins Discretizer from sklearn.preprocessing. It divides range of attribute in n intervals. We will consider number of intervals 2 and 3, and 2 strategies - "uniform"and "quantile".

We pick "uniform "n = 2". Scaled numerical part of dataset is on Рис 4.

At this step we also use different methods to find best concepts: based on f1-score and recall. The metrics of the prediction given in the next table:

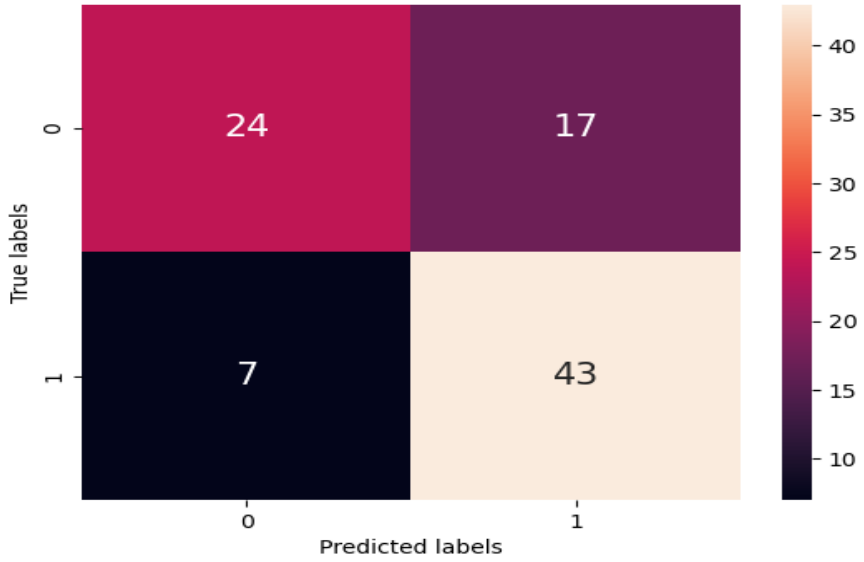

Рис. 2: Confusion matrix

best concept	num of concepts	f1-score	recall
f1	7	0.778	0.712
recall	8	0.792	0.737

Now use some non-linearities: Leaky Relu and hyperbolic tangent. And combine it with methods of selecting best concepts

best concept	non-lin	num of concepts	f1-score	recall
f1	Leaky Relu	9	0.792	0.737
f1	Tahn	10	0.804	0.774
recall	Leaky Relu	10	0.78	0.764
recall	Tahn	9	0.804	0.774

We also use "quantile"strategy for scaling numerical attributes. The difference between uniform and quantile is that in uniform cases all intervals are the same width, and in quantile case- all intervals have (approximately) the same number of objects.

Table for scaling numerical part of the data with n-bins=3 and strategy="quantile".

	1 bin	2 bin	3 bin
age	[29:51]	[51:59]	[59:77]
trestbps	[94:124]	[124:138]	[138:200]
chol	[126:222]	[222:260]	[260:564]
thalach	[71:142]	[142:161]	[161:202]
oldpeak	[0.0:0.2]	[0.2:1.4]	[1.4:6.2]
ca	[0:1]	[1:4]	

Tables of metrics for this variant:

	fbs	restecg	exang	female	male	cp 0	cp 1	cp 2	cp 3	thal fixed defect	thal normal	thal rev. defect
87	False	True	True	True	False	True	False	False	False	True	False	False
139	False	True	False	False	True	False	True	False	False	True	False	False
473	False	False	False	False	True	False	False	True	False	True	False	False
0	False	True	False	False	True	True	False	False	False	False	False	True
216	False	False	False	False	True	True	False	False	False	False	False	True
...	...	...	...	...	...	...	...	...	...	...	...	...
49	True	True	True	False	True	True	False	False	False	False	False	True
135	True	False	True	True	False	True	False	False	False	False	True	False
163	False	False	False	False	True	True	False	False	False	False	False	True
190	False	True	False	False	True	False	False	True	False	True	False	False
271	False	True	False	False	True	False	True	False	False	False	False	True

211 rows × 12 columns

Рис. 3: Binarized data in 1 baseline

	age|29.0:53.0|	age|53.0:77.0|	sex|0.0:0.5|	sex|0.5:1.0|	cp|0.0:1.5|	cp|1.5:3.0|	trestbps|94.0:147.0|	trestbps|147.0:200.0|	chol|126.0:345.0|	chol|345.0:564.0|	...
189	False	True	False	True	False	True	True	False	True	False	...
167	False	True	True	False	True	False	True	False	False	True	...
343	True	False	False	True	False	True	False	True	True	False	...
240	False	True	True	False	True	False	True	False	True	False	...
110	True	False	False	True	True	False	True	False	True	False	...
...	...	...	...	...	...	...	...	...	...	...	...
124	False	True	False	True	True	False	True	False	True	False	...
287	False	True	True	False	True	False	False	True	True	False	...
151	False	True	False	True	True	False	False	True	True	False	...
323	False	True	False	True	True	False	False	True	True	False	...
73	False	True	False	True	True	False	True	False	True	False	...

Рис. 4: KBinsDiscretizer(n=2, uniform)

best concept	non-lin	num of concepts	f1-score	recall
f1	Leaky Relu	10	0.758	0.722
f1	Tahn	7	0.811	0.734
recall	Leaky Relu	9	0.789	0.745
recall	Tahn	8	0.78	0.786

Networks

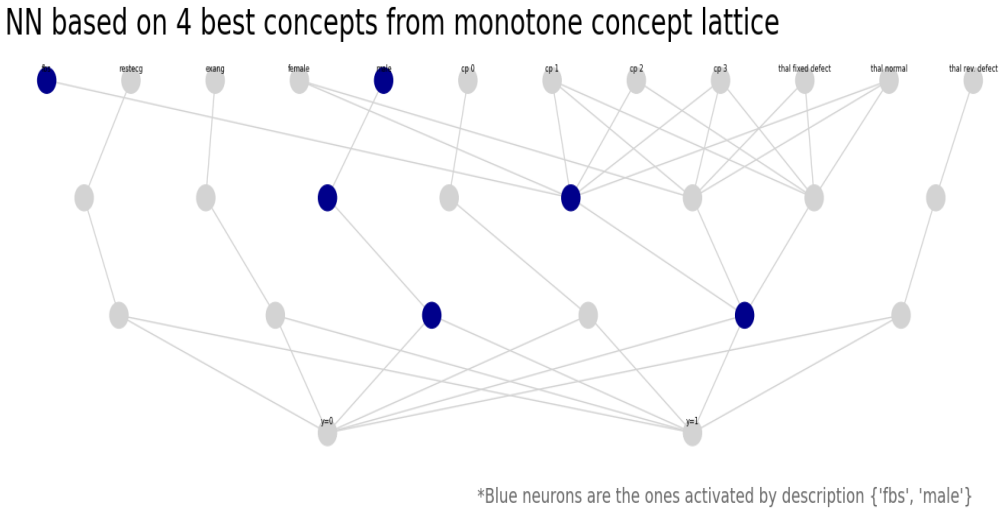

Рис. 5: Concept lattice in 1 baseline

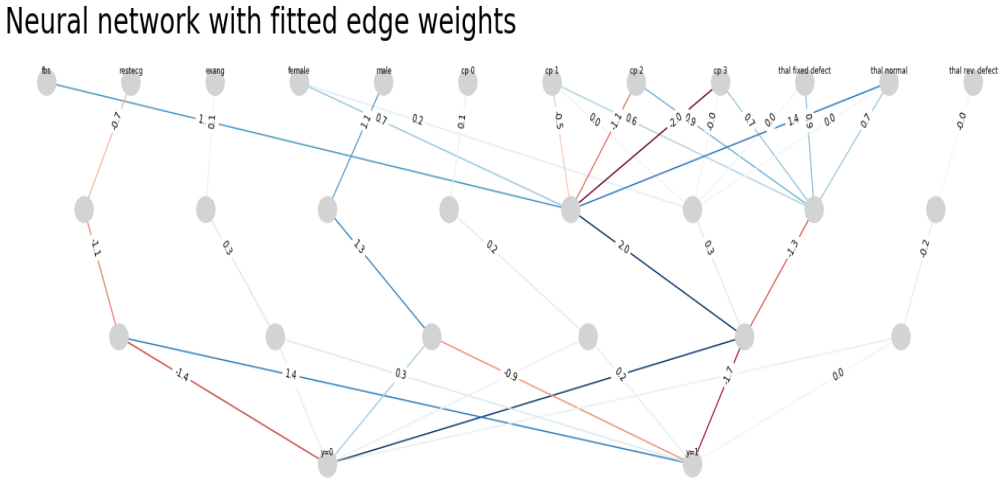

Рис. 6: Fitted network in 1 baseline

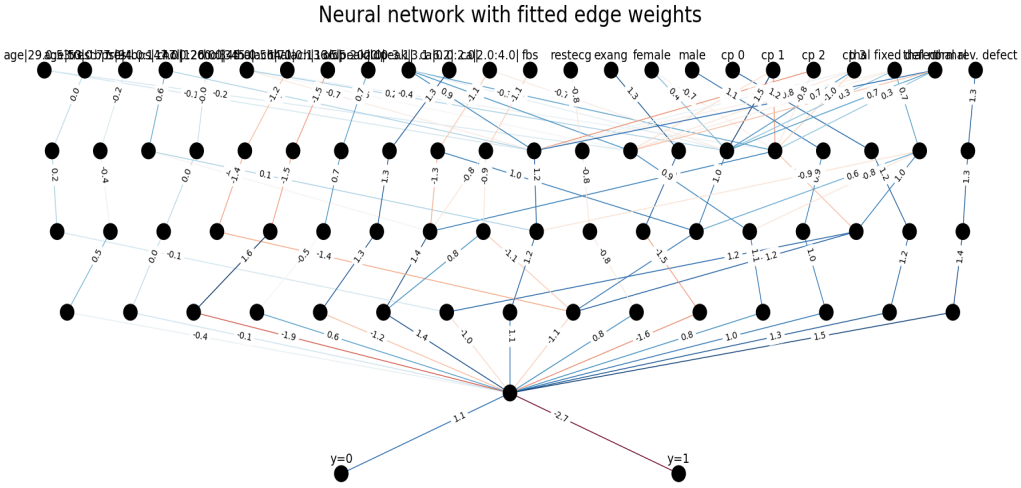

Рис. 7: Fitted network for n=2, uniform, nonlinearity Tahn

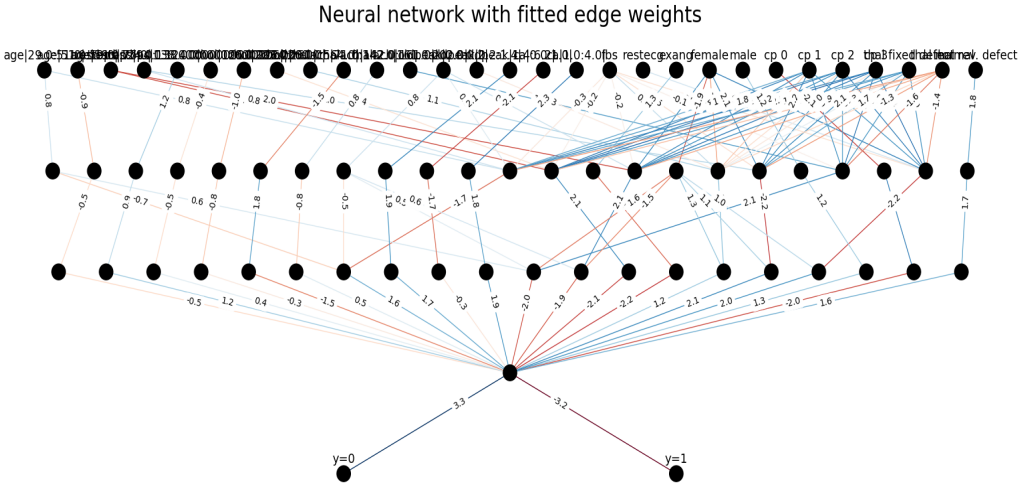

Рис. 8: Fitted network for n=3, quantile, nonlinearity LeakyReLU