Repository
Data

# 1 Dataset description:

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease.
People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.
Attributes:

1. **Age:** age of the patient [years]
2. **Sex:** gender of the patient [M: Male, F: Female]
3. **ChestPainType:** chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP:** resting blood pressure [mm Hg]
5. **Cholesterol:** serum cholesterol [mm/dl]
6. **FastingBS:** fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
7. **RestingECG:** resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR:** maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina:** exercise-induced angina [Y: Yes, N: No]
10. **Oldpeak:** oldpeak = ST [Numeric value measured in depression]
11. **ST_Slope:** the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease:** target [1: heart disease, 0: Normal]

## 1.1 Binarization

### 1.1.1 Dichotomic Scale

The columns Sex, FastingBS and ExerciseAngina contain only two possible values, so a dichotomous scale is best suited for their analysis. This will allow you to create binary attributes indicating the presence or absence of each of these attributes for objects in the dataset.

### 1.1.2 Nominal Scale

While the ChestPainType, RestingECG, and ST_Slope columns have more than two possible values, a nominal scale will be used to binarize them. This will create separate binary attributes for each unique value in these columns.

### 1.1.3 Inter-Ordinal Scale

To save as much information as possible due to the importance of the task, an inter-ordinal scale is proposed.
Age: Age is one of the most significant risk factors for cardiovascular diseases, including strokes, heart attacks and other related conditions. A high relationship with the target variable requires better binarization of this trait (narrow boundaries), while the boundaries may be wider at the beginning of the scale, since young people have a lower risk of having diseases of the cardiovascular system, according to statistics
Boundaries: 20, 30, 40, 50, 55, 60, 65, 70, 75, 80.
RestingBP: Normal blood pressure is considered to be in the range of 90-120 mmHg. Values above 130 may indicate prehypertension or hypertension, which is associated with an increased risk of cardiovascular diseases.
Boundaries: 70, 90, 120, 135, 150, 160, 170, 180.
Cholesterol: Cholesterol levels below 200 mg/dl are considered normal, 200-239 mg/dl is borderline, and $\geq$240 mg/dl is high. High cholesterol is a significant risk factor for cardiovascular disease.
Boundaries: 100, 150, 200, 240, 300, 400.
MaxHR: Normal values of the maximum heart rate vary depending on age and level of physical fitness. A decrease in the maximum heart rate may indicate cardiovascular problems. Logical reasoning suggests the importance of this feature, therefore, a fairly frequent division over the entire length of the values is proposed.
Boundaries: 60, 70, 80, 90, 100, 110, 120.
Oldpeak: Oldpeak is an important indicator that is used to assess myocardial ischemia. Values close to 0 indicate the absence of ischemia, while higher values may indicate serious heart problems.
Boundaries: 1, 2, 3.

## 1.2 Another binarization

As part of the assignment, another way of binarization was required. Binary and categorical will be encoded in the same way, but numeric only ordinal $\geq$.

# 2 Base models

7 standard classifiers were used to create baseline: kNN, GaussianNB, logistic regression, decision tree, random forest, CatBoost and XGB. Each of the models was configured using a selection of hyperparameters during cross-validation.

| | classifier | accuracy | precision | recall | f1_score |
|---|---|---|---|---|---|
| 2 | LogReg | 0.880435 | 0.917197 | 0.878049 | 0.897196 |
| 0 | NB | 0.844203 | 0.890323 | 0.841463 | 0.865204 |
| 3 | NB | 0.844203 | 0.890323 | 0.841463 | 0.865204 |
| 4 | NB | 0.844203 | 0.890323 | 0.841463 | 0.865204 |
| 5 | NB | 0.844203 | 0.890323 | 0.841463 | 0.865204 |
| 6 | NB | 0.844203 | 0.890323 | 0.841463 | 0.865204 |
| 1 | NB | 0.612319 | 0.870130 | 0.408537 | 0.556017 |

Figure 1: Binarization №1

| | classifier | accuracy | precision | recall | f1_score |
|---|---|---|---|---|---|
| **2** | LogReg | 0.884058 | 0.923077 | 0.878049 | 0.900000 |
| **0** | NB | 0.869565 | 0.905063 | 0.871951 | 0.888199 |
| **3** | NB | 0.869565 | 0.905063 | 0.871951 | 0.888199 |
| **4** | NB | 0.869565 | 0.905063 | 0.871951 | 0.888199 |
| **5** | NB | 0.869565 | 0.905063 | 0.871951 | 0.888199 |
| **6** | NB | 0.869565 | 0.905063 | 0.871951 | 0.888199 |
| **1** | NB | 0.721014 | 0.914286 | 0.585366 | 0.713755 |

Figure 2: Binarization №2

The best quality was achieved by the logistic regression model using the second binarization method: f1=0.9

# 3 FCA

## 3.1 Concept Lattice

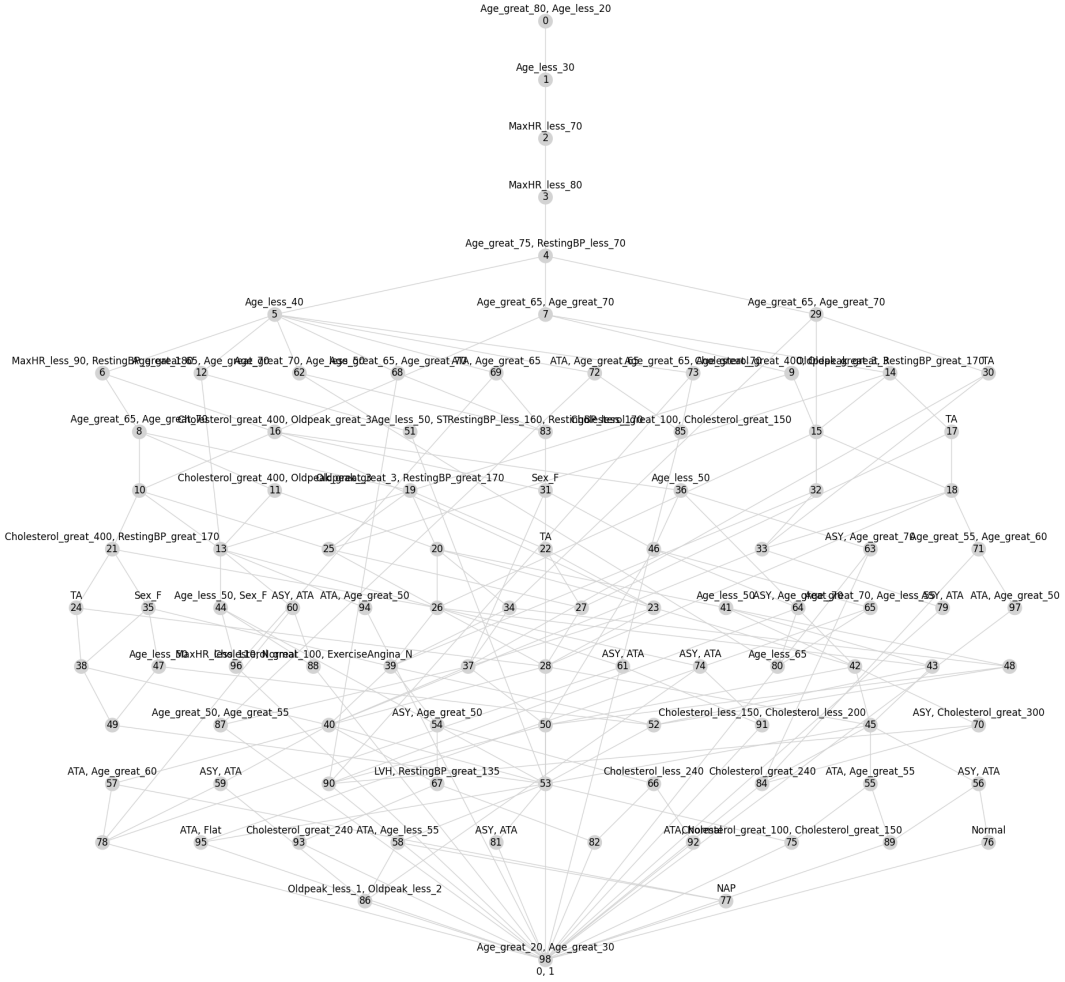


Figure 3: Concept Lattice №3

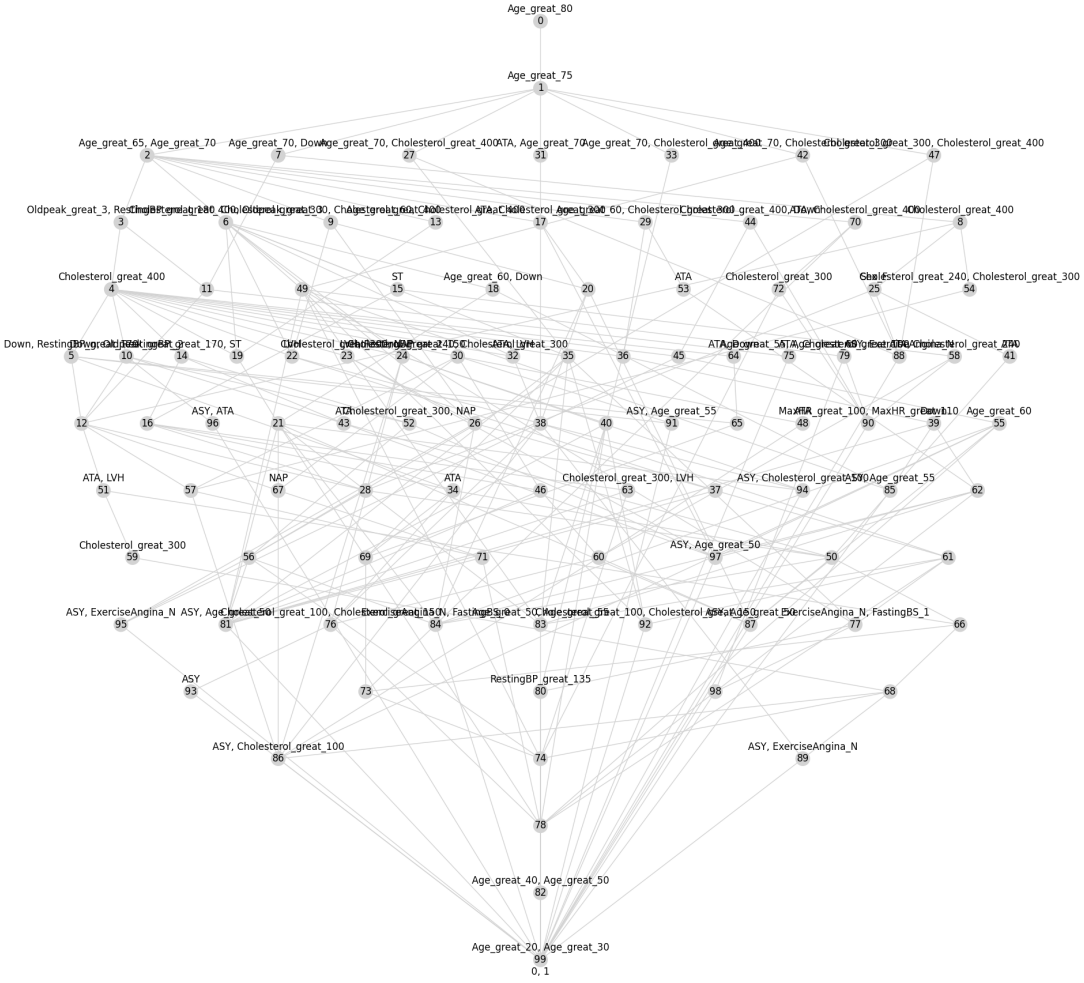


Figure 4: Concept Lattice №3

## 3.2 Concept Network

**Best concepts:** The best concepts are selected based on two metrics: f1 and recall.
**Nonlinearities:** ReLU, Tanh, Sigmoid.

### 3.2.1 Best model

Best model parameters: metric for selecting concepts = recall, nonlinearity function = Sigmoid.
Best metric: f1 = 0.91, recall = 0.91.

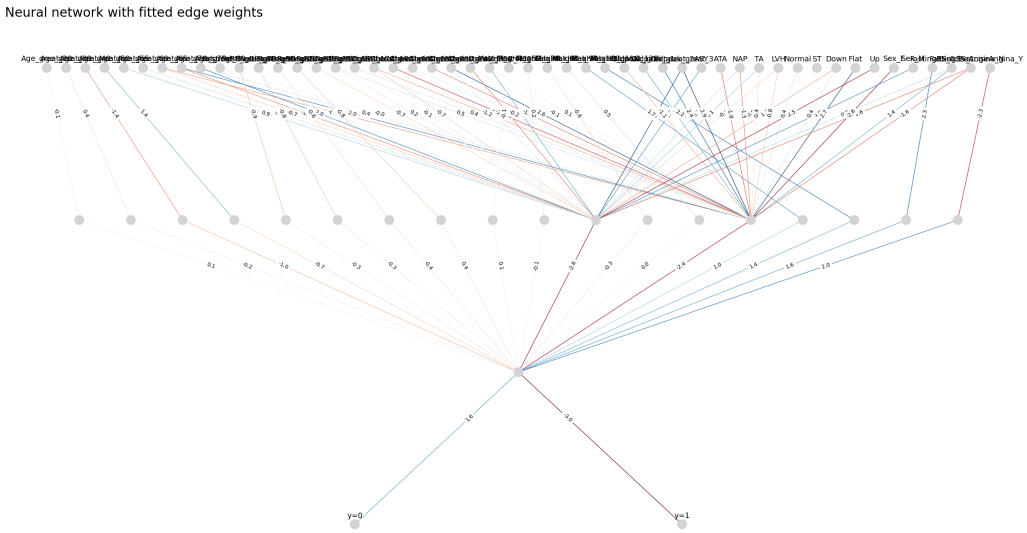


Figure 5: Best Concept Network

Age≥40, oldpeak≥2, flat, fasting_BS = 0, cholesterol≥200 contribute the most to the likelihood of having heart disease. Signs such as young age, ATA, Female (surprisingly), in contrast, reduce this probability more than others.

# 4 Conclusion

In conclusion, the following points can be highlighted:

1. The CbO algorithm requires considerable time to work on large sample sizes.
2. Installing the correct dependencies is often a difficult task.
3. F1 score is a more appropriate indicator for choosing the best concepts.
4. Fewer features lead to better quality, which is probably due to a decrease in multicollinearity in the feature matrix.
5. It even turned out to be a little baseline!