DataMining Features

Relevance of Features

['age', 'sex', 'painloc', 'painexer', 'relrest', 'cp', 'trestbps', 'htn', 'chol', 'smoke', 'cigs', 'years', 'fbs', 'dm', 'famhist', 'restecg', 'dig', 'prop', 'nitr', 'pro', 'diuretic', 'proto', 'thaldur', 'thaltime', 'met', 'thalach', 'thalrest', 'tpeakbps', 'tpeakbpd', 'trestbpd', 'exang', 'xhypo', 'oldpeak', 'slope', 'rldv5', 'rldv5e', 'ca', 'restef', 'restwm', 'exeref', 'exerwm', 'thal', 'num', 'lmt', 'ladprox', 'laddist', 'diag', 'cxmain', 'ramus', 'om1', 'om2', 'rcaprox', 'rcadist', 'cathef', 'junk', 'dataset']

Todos:

- Features to OneHotEncode:
  - cp
  - slope
  - dataset
- Features to maybe use binning on
  - Age
  - trestbps
  - chol
  - thaldur
  - thaltime
  - thalach
  - thalrest
  - tpeakbps
  - tpeakbpd
- Features to think about:
  - chol - too many zeros/to high values
  - prop - one value wrong
  - thaldur
  - met - i dont know what that is
  - thalrest - some values under 50 are unhealthy
  - tpeakbpd - some values a bit low
  - trestbpd - some persons are dead - 0 puls
  - ca - 67% missing, 1 value wrong
  - restef - 97% missing
  - restwm - 97% missing
  - exeref - 99.8% missing
  - exerwm - 99.8% missing

    - thal - values out of bounds
    - num - should be binary but have 2,3,4
    - lmt
    - 'ladprox', 'laddist', 'diag', 'cxmain', 'ramus', 'om1', 'om2', 'rcaprox', 'rcadist'
    - cathef
    - junk

Feature Analysis

Age

- Values normal
- TODO: maybe try binning in 5 or so bins

Sex

- 1 = male
- 0 = female
- 0.2% missing -> try at the end with different drop rates
- values normal

painloc

- chest pain location (1 = substernal; 0 = otherwise)
- Values normal
- 30% missing

painexer

- painexer (1 = provoked by exertion; 0 = otherwise)
- Values normal
- 31.5% missing

relrest

- relrest (1 = relieved after rest; 0 = otherwise)
- Values normal
- 32% missing

cp

- cp: chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- TODO: OneHotEncoding
- Values normal

trestbps

- resting blood pressure (in mm Hg on admission to the hospital)
- values normal
- TODO: maybe binning

htn

- Hypertension
- Values normal
- 0 = false
- 1 = true
- 4% missing

chol

- Too many zeros!
- Maybe binning because many discrete values
- Probably many values to high. Chol values over 240 are dangerous
- https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean
- Median: 224
- Maybe drop feature

smoke

- Values normal
- 75% missing

cigs

- values pretty normal
- median: 20/day
- probably drop

years

- values normal
- must look if years - age is negative

fbs

- (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- Values normal

dm

- dm (1 = history of diabetes; 0 = no such history)
- Too many missing
- 89.5% missing
- values normal

famhist

- family history of coronary artery disease (1 = yes; 0 = no)
- values normal
- 47% missing

restecg

- Values normal

dig

- (digitalis used furing exercise ECG: 1 = yes; 0 = no)
- Values normal
- 8% missing

prop

- (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
- one value wrong (22)

nitr

- (nitrates used during exercise ECG: 1 = yes; 0 = no)

- Values normal
- 7% missing

pro

- (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
- Values normal
- 7% missing

diuretic

- (diuretic used used during exercise ECG: 1 = yes; 0 = no)
- Values normal
- 9% missing

proto

TODO:!

thaldur

- duration of exercise test in minutes
- TODO: Maybe binning due to 86 discrete values
- 6% missing
- values normal

thaltime

- time when ST measure depression was noted
- 50% missing
- 64 distinct, maybe binning
- Lots of zeros, but maybe due to not taking the test

met

- mets achieved?
- Dont really know that this tells us :D

thalach

- maximum heart rate achieved

- values normal
- TODO: maybe binning

thalrest

- resting heart rate
- some values under 50 maybe look into it
- TODO: maybe binning

tpeakbps

- peak exercise blood pressure (first of 2 parts)
- values normal
- TODO: maybe binning

tpeakbpd

- peak exercise blood pressure (second of 2 parts)
- some values are a bit low
- TODO: maybe binning

trestbpd

- resting blood pressure
- some values are a bit low
- TODO: maybe binning

exang

- exercise induced angina (1 = yes; 0 = no)
- values normal

xhypo

- xhypo: (1 = yes; 0 = no)
- values normal
- pretty unbalanced - 90% - 2%

oldpeak

- ST depression induced by exercise relative to rest

- values normal

slope

- the slope of the peak exercise ST segment
  - Value 1: upsloping
  - Value 2: flat
  - Value 3: downsloping
- TODO: OneHotEncoding
- one zero value

rldv5

- height at rest
- values normal?
- Missing 47%

rldv5e

- height at peak exercise
- values normal

ca

- number of major vessels (0-3) colored by flourosopy
- Missing 67%!!!
- one value out of bounds (9)

restef

- rest raidonuclid (sp?) ejection fraction
- MISSING 97%

restwm

- rest wall (sp?) motion abnormality
- 0 = none
- 1 = mild or moderate
- 2 = moderate or severe
- 3 = akinesis or dyskmem (sp?)
- values normal
- 97% missing

exeref

- 99,8% missing

exerwm

- 99.8% missing

thal

- 3 = normal; 6 = fixed defect; 7 = reversable defect
- normal values 3,6,7 - but also 1,5,4,2
- look into them

num

- diagnosis of heart disease (angiographic disease status)
  - Value 0: < 50% diameter narrowing
  - Value 1: > 50% diameter narrowing
  - (in any major vessel: attributes 59 through 68 are vessels)
- look into it should be values 0,1 but have 3,4,5 also

lmt

- 30% missing
- should be binary i think but only 1x 0, 580x 1, 42x 2 and 1x 162?

ladprox

- 26% missing
- should be binary but is 1 or 2

-> same with the next features 'laddist', 'diag', 'cxmain', 'ramus', 'om1', 'om2', 'rcaprox', 'rcadist'

cathef

- "not used"
- 65% missing
- look into

junk

- "not used" not described
- 87% missing

dataset

- describes from what dataset the data is
- if we keep this one hot encoding