SC End-Sem Project

Predictive model using Bayesian networks in R

Madhav Thakker | 2016159

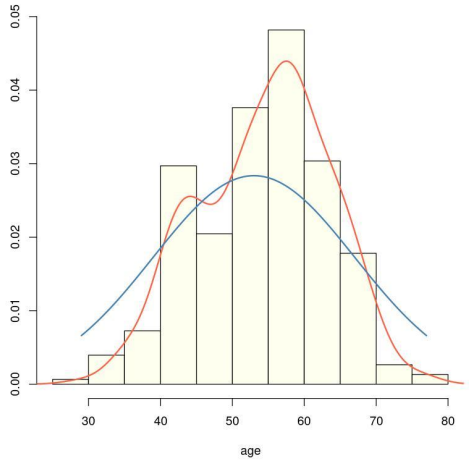
INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

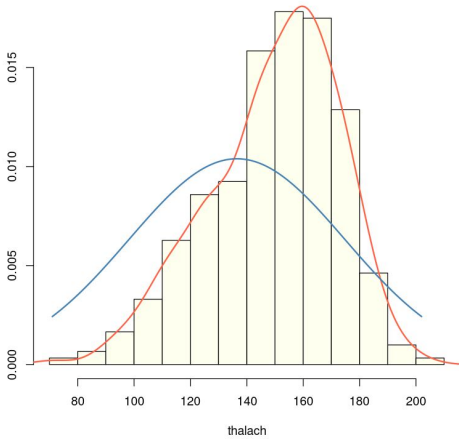
Data-Overview

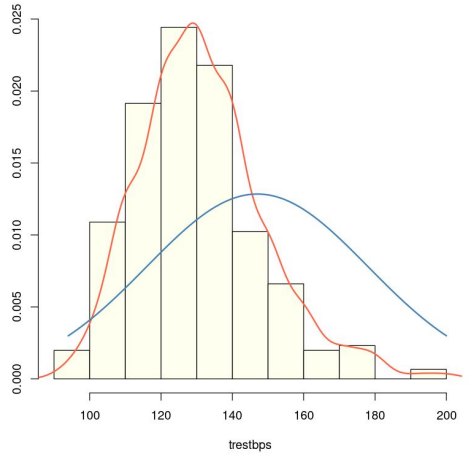
Features used -

- age, trestbps, chol, thalach, oldpeak
- cp, exang, ca, thal, num

Converted the target feature to binary column.







Structure Learning

exang: exercise induced angina, chol: serum cholesterol, trestbps: resting blood pressure, thalach: maximum heart rate

Adding whitelists in the graph -

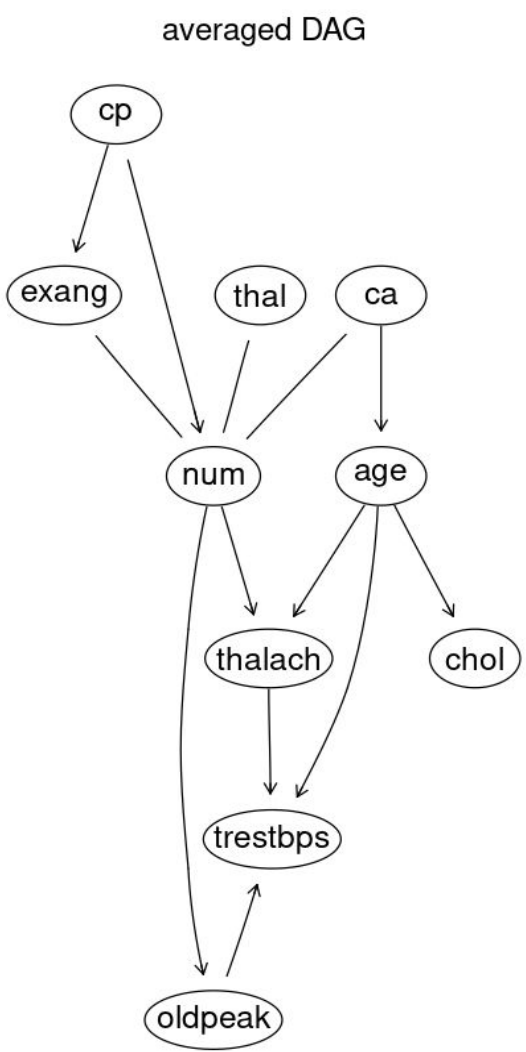
- exang ➜ chol & chol ➜ exang
- exang ➜ trestbps & trestbps ➜ exang

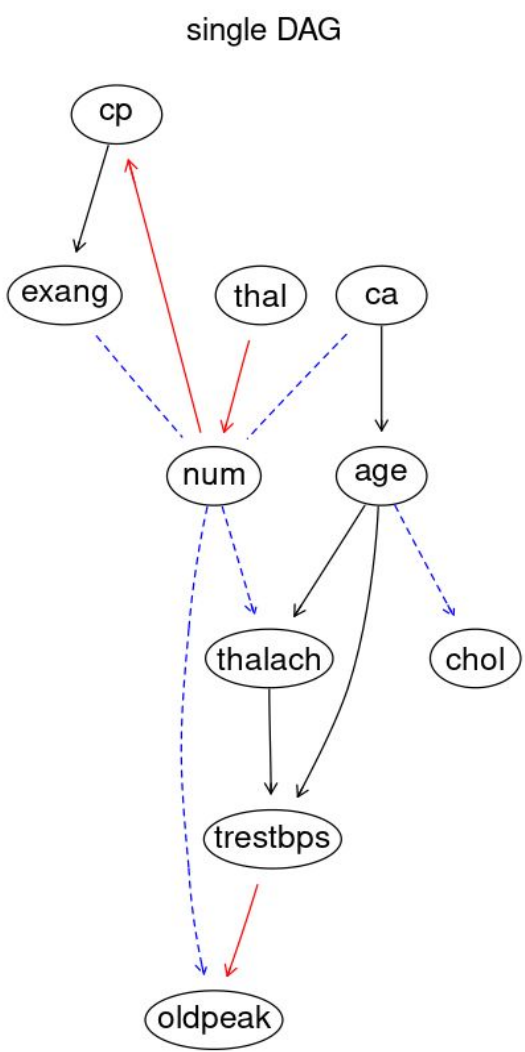
Adding blacklists in the graph -

- thalach ➜ trestbps & trestbps ➜ thalach

Structure Learning

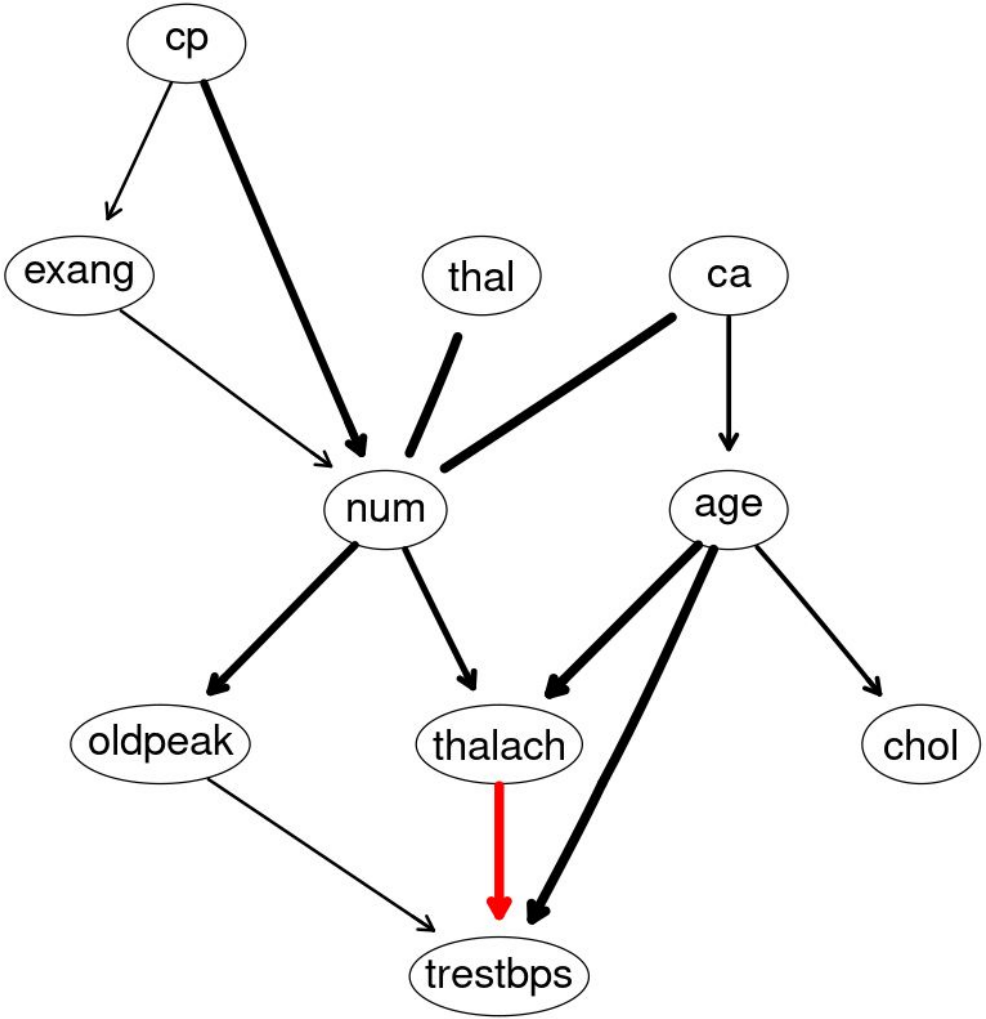
- Single-DAG: find the network structure with the best goodness-of-fit on the whole data.
- Averaged DAG:
  - resample the data using bootstrap
  - learn a separate network from each bootstrap sample
  - check how often each possible arc appears in the networks
  - construct a consensus network with the arcs that appear more often





Final DAG Structure

The final network is learnt by taking the threshold over the strengths of the averaged network and combining it with the network learned using the original data.



Missing Data

These are the steps taken to handle missing data:

- removed the rows with at-least 1 missing feature-value.
- A DAG is learnt on the data with all features available.
- Using the trained DAG, the missing values are imputed using bnlearn's *impute* function.

Results

The data is divided into train and test sets randomly.

Learn the network on the train-set and predict the values of node “num” for the test-set.

Data-set	Accuracy
Train	85%
Test	84%