Modelling cardiovascular risk using Bayesian Networks

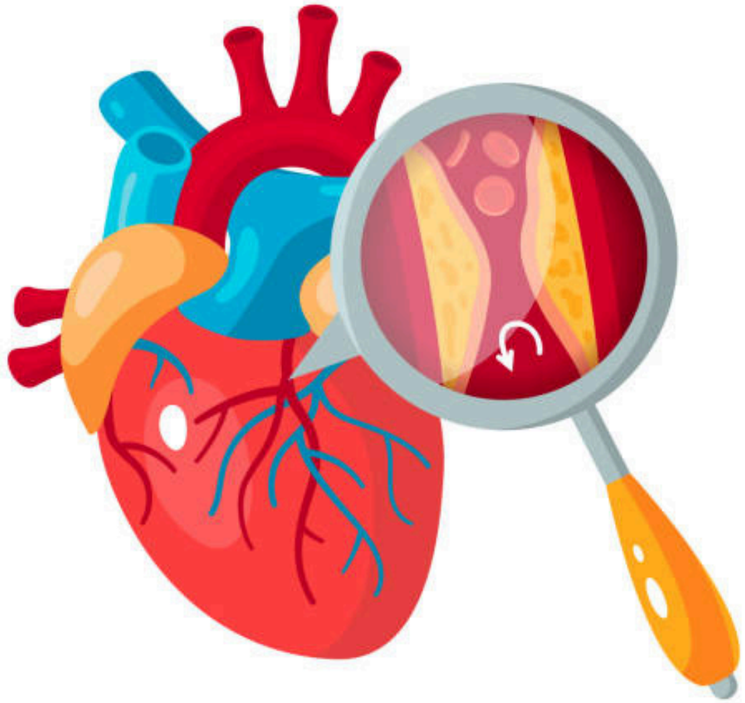
Matteo Tarenzi

BAYESIAN ANALYSIS PROJECT

https://github.com/TarMatt/Coronary_heart_disease.git

A growing problem

Coronary heart disease occurs when the arteries of the heart cannot deliver enough oxygen to the heart muscle due to narrowing from the buildup of fatty deposits called plaque .



The causes

The risk is influenced by a combination of genetic, lifestyle and environmental factors.

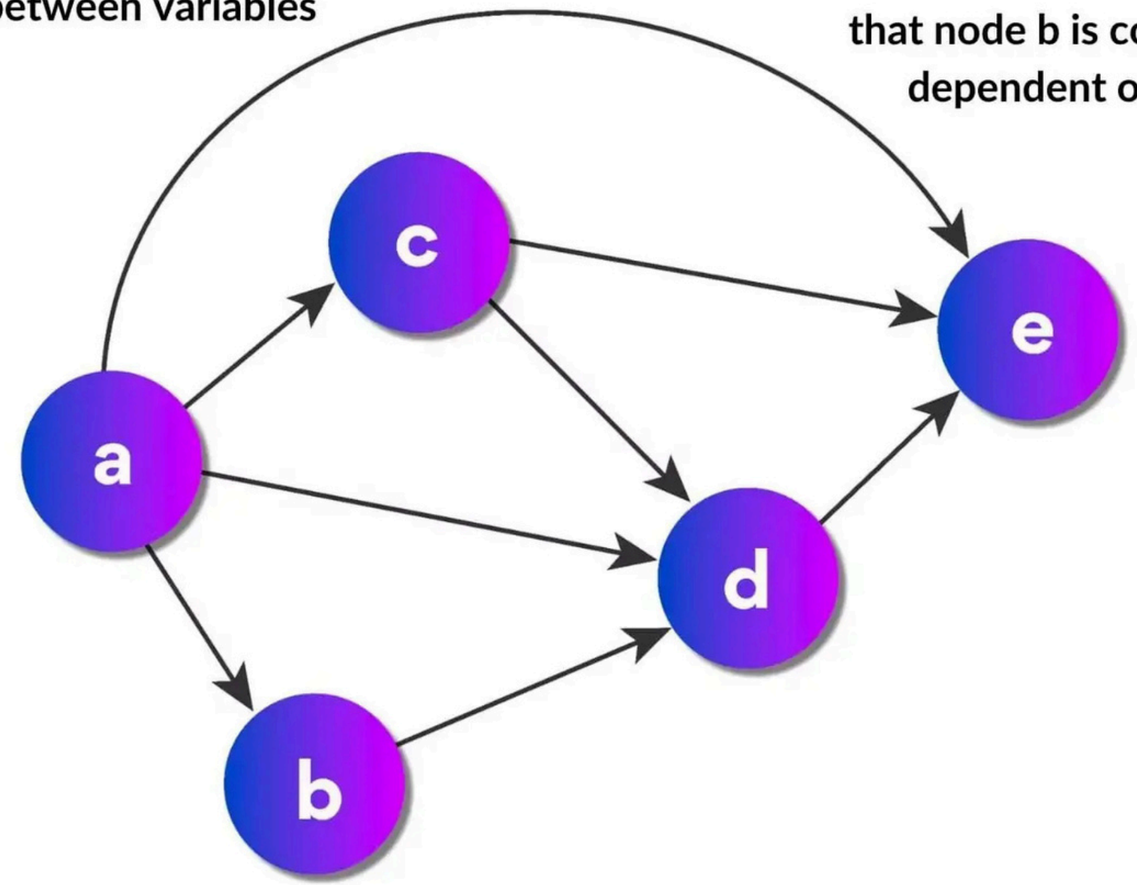
A worrying record

CHD is the single largest cause of death in the developed countries.

Bayesian Networks

Directed edges represent probabilistic dependencies between variables

A directed edge from node a to node b indicates that node b is conditionally dependent on node a



Nodes represent random variables

Modularity
Causal Reasoning
Uncertainty Handling
Dynamic Learning

Project phases

- **Data Exploration & Engineering**
- **Definition of the Bayesian Network**
- **Variable Elimination**
- **Classification**

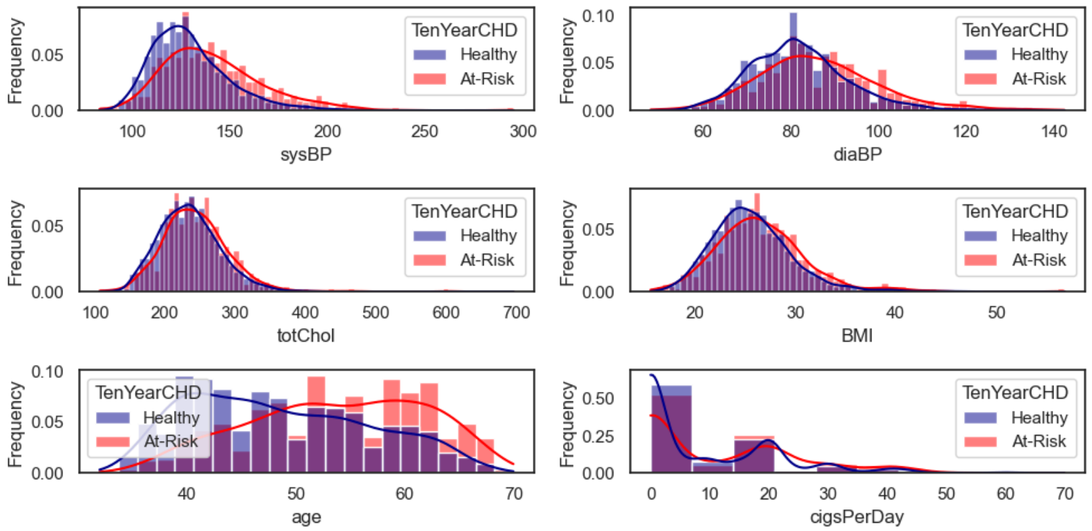
Data Exploration & Engineering

The Dataset

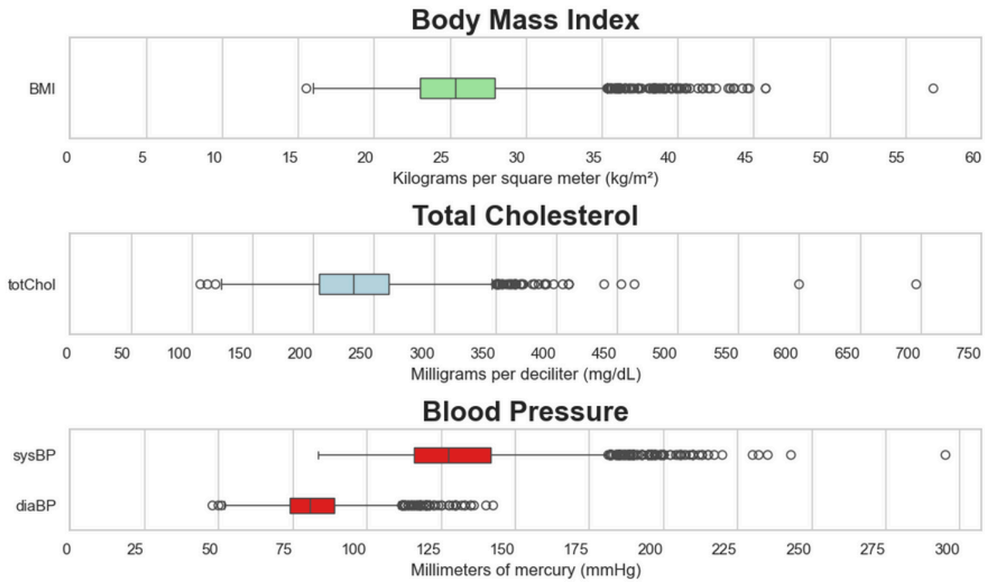
The dataset includes 4 categorical and 5 continuous features.

- Age
- TotatCholesterol
- SystolicBP
- DiastolicBP
- BMI
- CigsxDay
- Gender
- Diabetes
- TenYearCHD

Frequencies



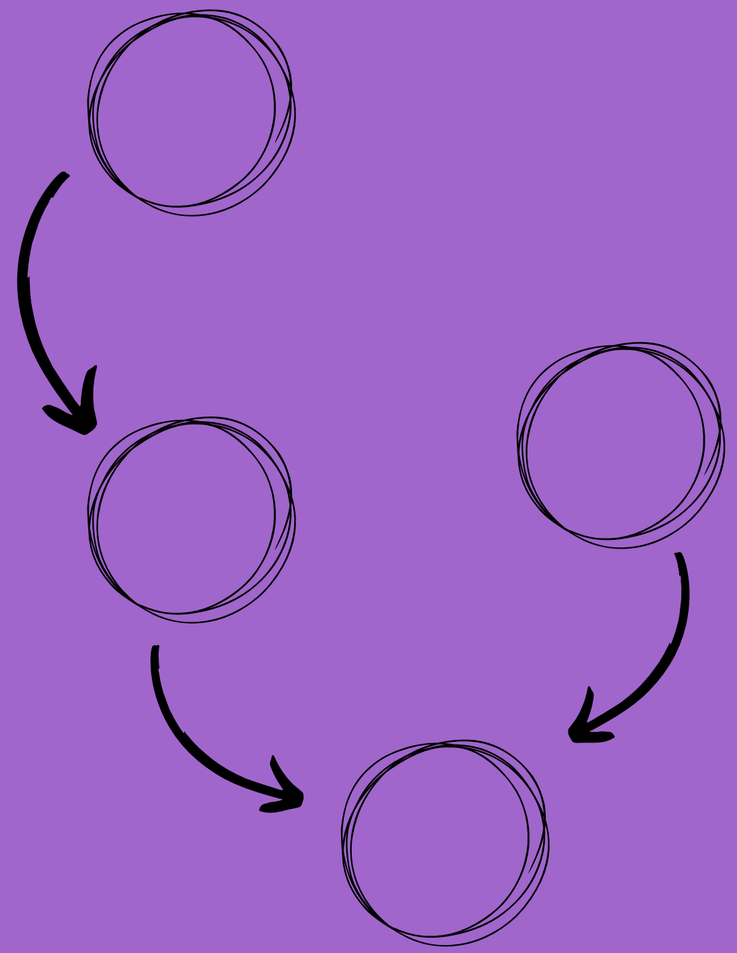
Outliers



The binning process

Continuos variables have been converted into categorical through binning.

Age	→	30-44 45-54 55-64 65-75
TotCholesterol	→	Optimal Elevated High
BMI	→	Under Normal Over Obese
CigsxDay	→	NonSmoker Light Moderate Severe
SystolicBP **DiastolicBP**	→	Optimal Elevated Hypertension Crisis



Definition of the Bayesian Network

Estimation

Three different strategies have been tested to define the best structure.

Hill Climb

BIC score: -26518,28

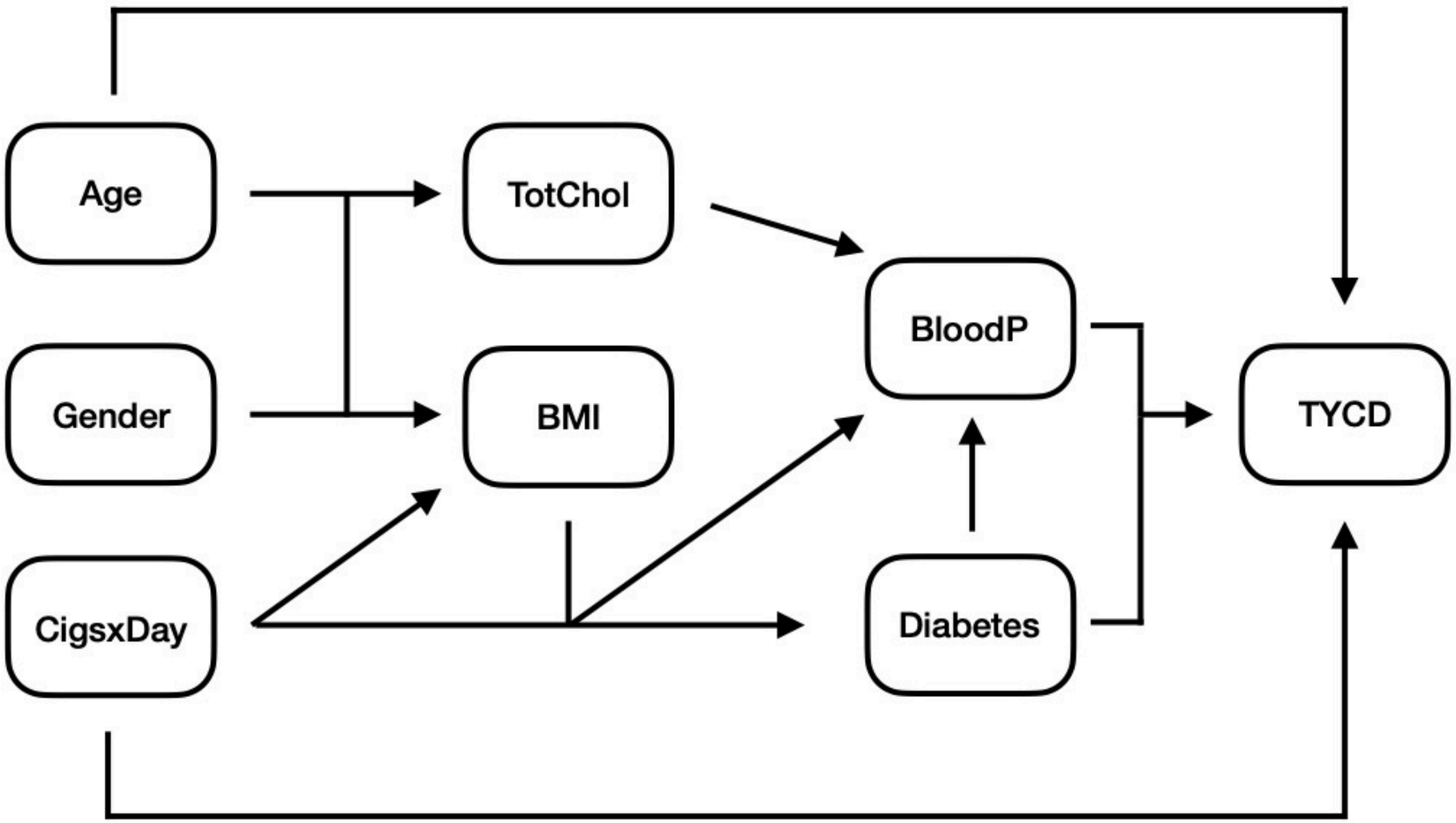
Custom definition

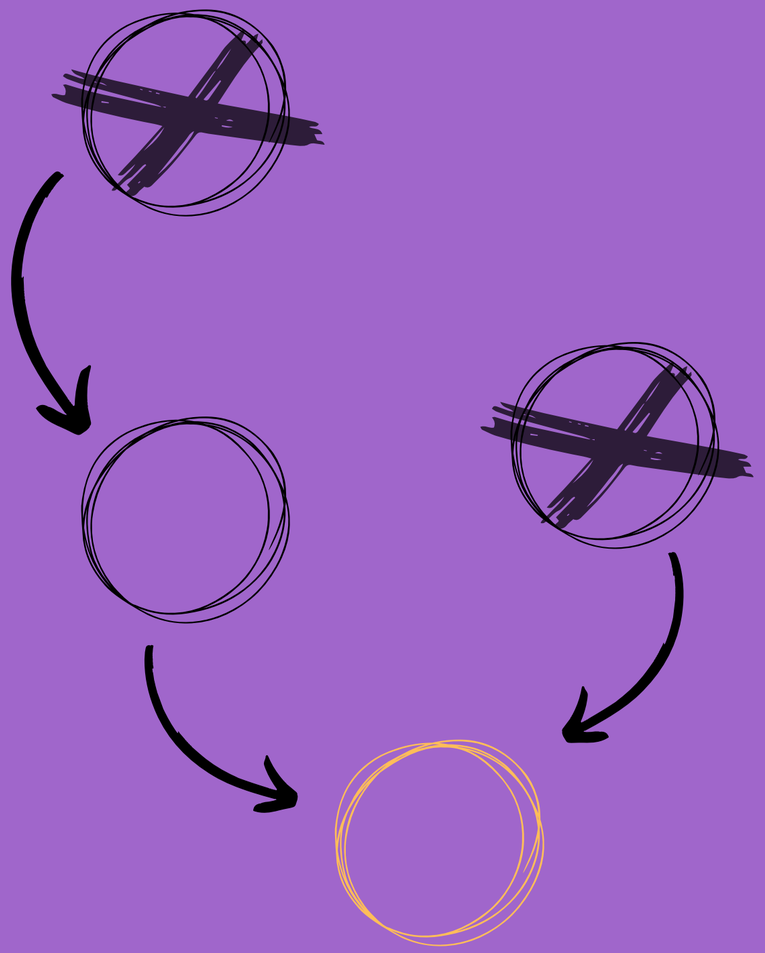
Bic score: -28221,91

Hybrid learning

BIC score: -26248,75

Custom definition





Variable Elimination

Healty

{BMI:'Normal', **BloodP**:'Optimal', **CigsxDay**:'NonSmoker'**}**

TYCD	phi(TYCD)
At-Risk	0,0603
Healthy	0,9397

Diseased

{BMI:'Obese', **BloodP**:'Crisis', **CigsxDay**:'SevereSmoker'**}**

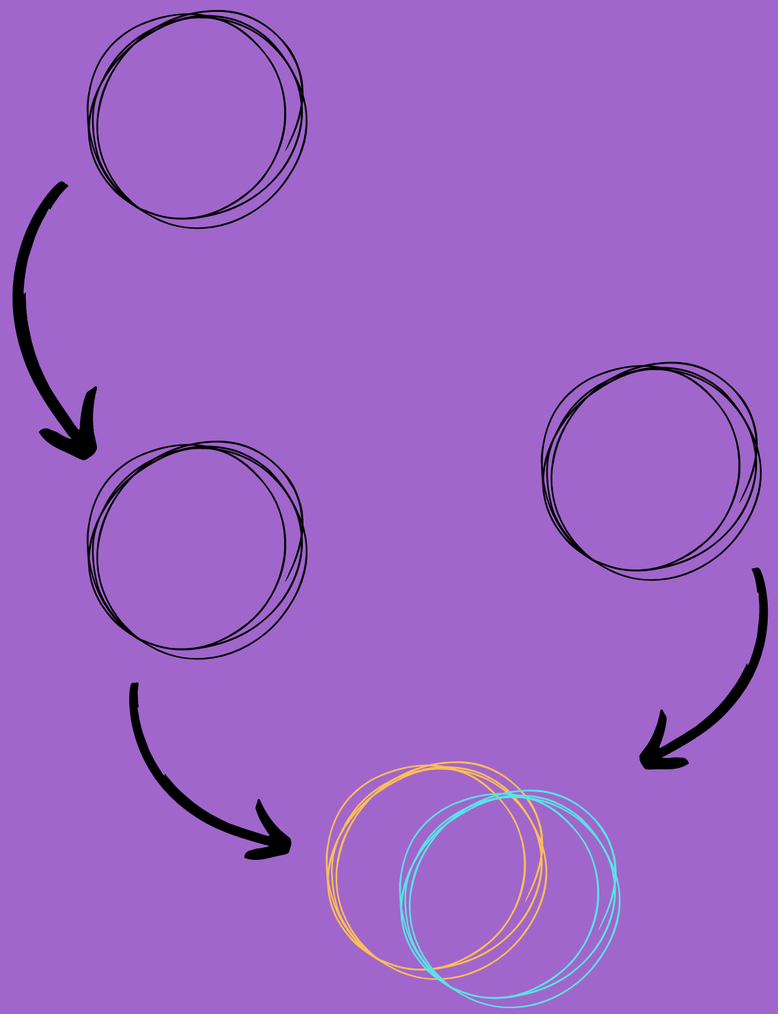
TYCD	phi(TYCD)
At-Risk	0,5668
Healthy	0,4332

Blood Pressure

TYCD \ BP	Optimal	Elevated	Hypertention	Crisis
At Risk	0,103	0,123	0,198	0,285
Healty	0,897	0,877	0,804	0,715

Body Mass Index

BP \ BMI	Underweight	Normal	Overweight	Obese
Optimal	0,511	0,397	0,215	0,119
Elevated	0,332	0,397	0,412	0,320
Hypertension	0,119	0,189	0,343	0,491
Crisis	0,038	0,015	0,030	0,070



Classification

Classification Workflow

TRAINING SET → LEARN the NETWORK

VALIDATION SET → SELECT the THRESHOLD

TEST SET → EVALUATE the PERFORMANCE

Threshold: 0,17

Classification Report

	Precision	Recall	F1-score	Support
Healthy	0,93	0,71	0,81	538
At-Risk	0,29	0,68	0,41	94
Accuracy			0,71	632
Macro avg	0,61	0,70	0,61	632
Weighted avg	0,83	0,71	0,75	632