Understanding Data-Binning

Currently i am researching into what a BN is, i have discovered that the data must be discrete continous variables, after asking an ai engine what would be the best means of cleaning my dataset, there are 4 ways:

- **Equal-Width Binning:** Create bins of the same size
- **Equal-Frequency Binning (Quantiles):** Create bins with the same number of data points in each
- **Clustering:** Use an algorithm like K-Means to find natural "clusters" in the data to define the bins.
- **Domain Knowledge:** Use expert knowledge (e.g., a doctor defining "high" blood pressure as > 130)

Some of the attributes require “domain knowledge” for absolute precision, some would be best to use a clustering algorithm and some will require equal binning, i will have to make a python script that sorts the csv file of all 12 attributes using with predetermined sorting means.

New dataset

New dataset: https://archive.ics.uci.edu/dataset/45/heart+disease

Reference: “Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.”

Catagories

When creating a BN it is known best to catagorise what attributes are known:

Only 14 attributes used:

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

Tier Sorting system

I have Separated it into 5 Tiers:

Tier 1: Root Cause

- Age
- Sex

Tier 2: Clinical Factors

- Trestbps (resting bp)
- Chol
- Fbs (fasting blood sugar)
- Restecg (resting ecg)
- Thal (thalassemia)

Tier 3: Exercise test results

- Thalach (max heart rate)
- Oldpeak
- Slope
- Ca (vessels)

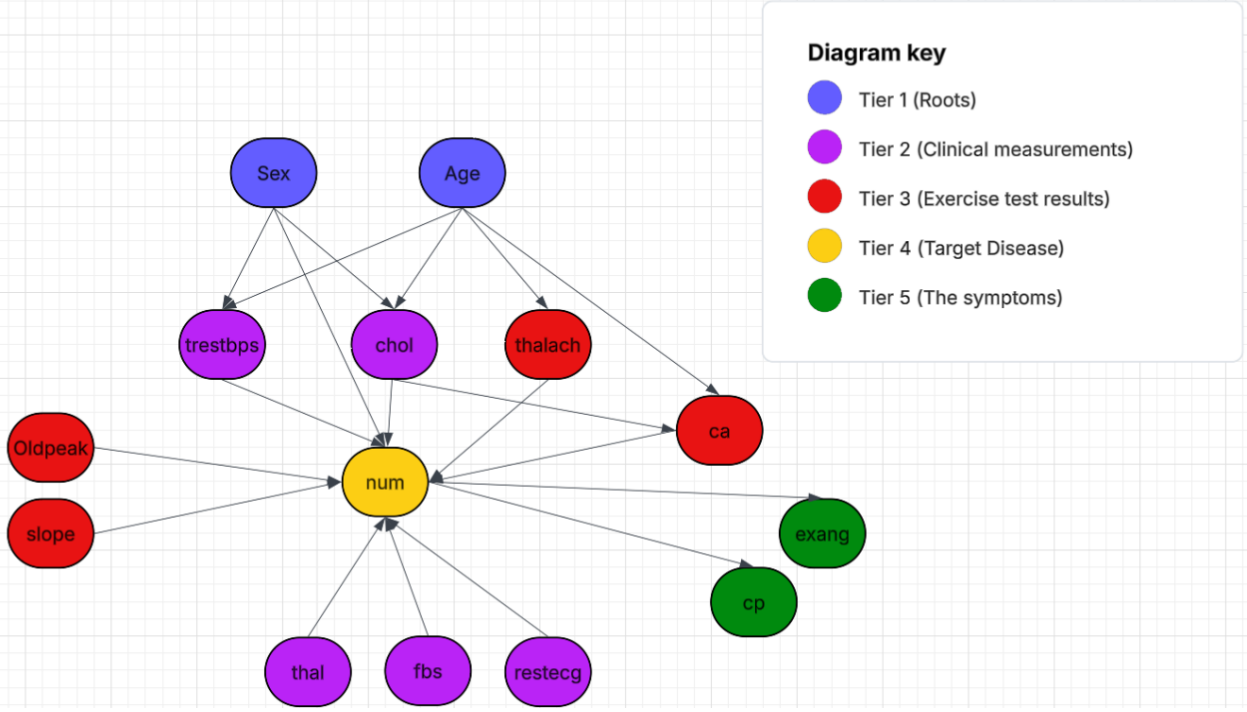
Tier4: the disease (target)

- Num (central node, presence of heart disease)

Tier 5: The symptoms

- Cp (chest pains)
- Exang (exercise induced angina)

BN Mock draw up



“compare diagram from paper”
“how ai version that they made the bayesian network, and how they achieved these relationships?, vs research paper bayesian network designs”

“maxiumum estimator function to determine the cpt, is i appropiate compared to what the research papers functions”

BN Relationship Validation

In the EDA File on github

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/SourceCode/Jorjit/jupyter/eda.ipynb

Heatmap Correlation Analysis

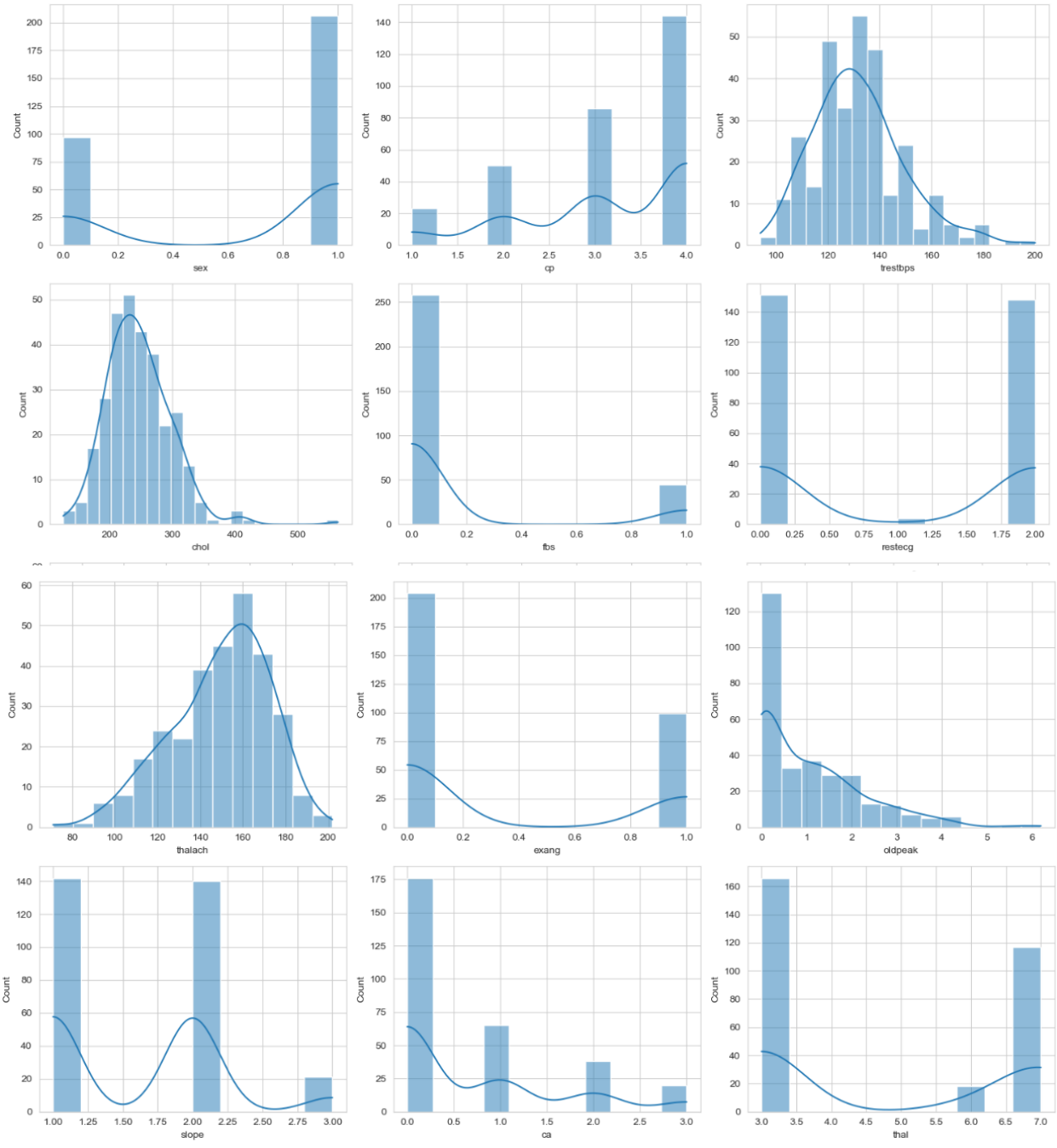
- **Target Variable (`num`):** The strongest positive predictors for heart disease are `ca` (0.52), `thal` (0.51), and `oldpeak` (0.50).
- **Inverse Relationship:** `thalach` (Max Heart Rate) has a strong negative correlation (-0.42) with the target, indicating that patients with heart disease often achieve lower max heart rates.
- **Multicollinearity:** `oldpeak` and `slope` have a notably high correlation (0.58), suggesting they provide overlapping information regarding ST depression.
- **Biological Trend:** `age` and `thalach` are negatively correlated (-0.39), confirming the natural trend that max heart rate decreases as age increases.

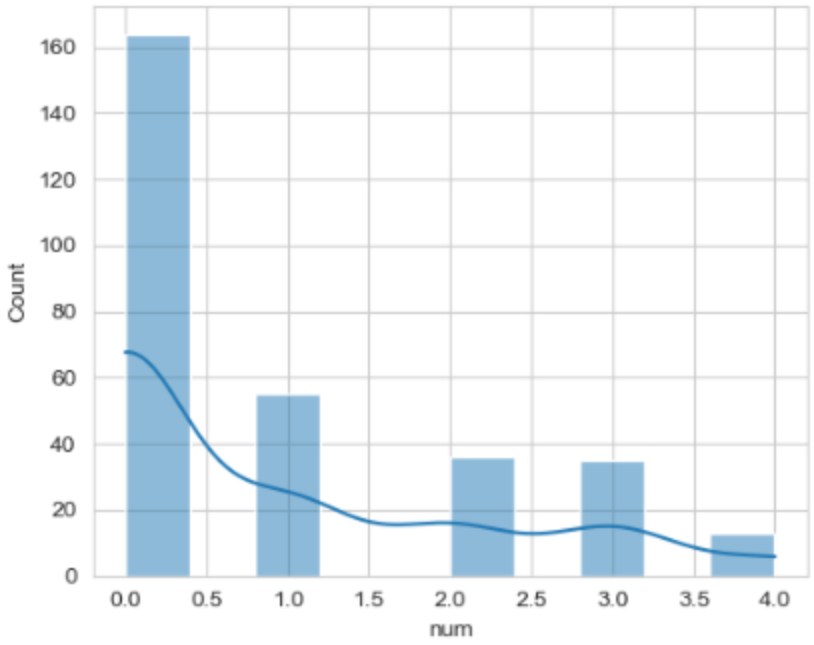
Here is the document of “Medical Justification: Complete Analysis of Heart Disease Attributes”

All of the Necessary analysis based off the EDA.IPYNB, of relationships are explained in this document.

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/Medical%20Justification_%20Inverse%20vs.%20Direct%20Relationships%20in%20Heart%20Disease%20Attributes%20(1).pdf

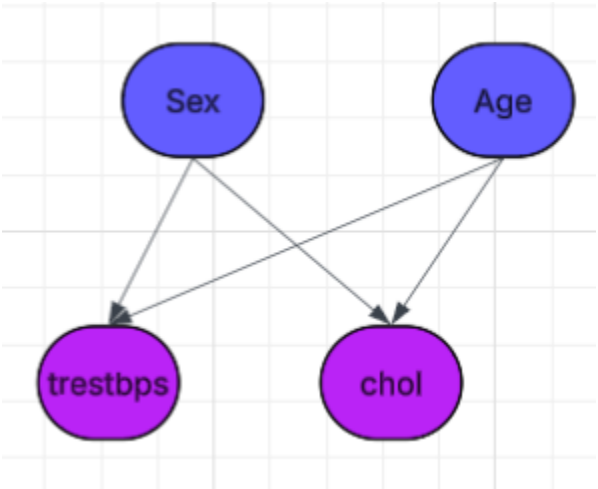
Here are images of the data histplot using “Seaborn” Library


Count
sex
cp
trestbps
chol
fbs
restecg
thalach
exang
oldpeak
slope
ca
thal



BN Conditional Probability tables

Tier 1 --> Tier 2



--- Connection: Sex_Label -> BP_Bin ---

	Sex_Label	Prob(Elevated)	Prob(High_BP)	Prob(Normal)
0	Female	0.4830	0.3101	0.2069
1	Male	0.4809	0.3211	0.1980

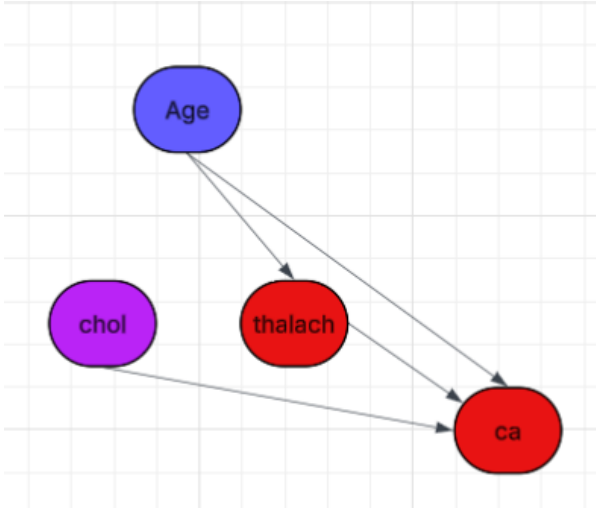
--- Connection: Age_Bin -> BP_Bin ---

	Age_Bin	Prob(Elevated)	Prob(High_BP)	Prob(Normal)
0	Middle	0.5037	0.3155	0.1807
1	Old	0.3987	0.4703	0.1310
2	Young	0.5296	0.1328	0.3376

--- Connection: Sex_Label -> Chol_Bin ---

	Sex_Label	Prob(Borderline)	Prob(Desirable)	Prob(High_Chol)
0	Female	0.2554	0.1521	0.5925
1	Male	0.3571	0.1711	0.4717

Tier 1&2 --> 3



--- Connection: Age_Bin -> HR_Bin ---

	Age_Bin	Prob(High_Rate)	Prob(Low_Rate)	Prob(Normal_Rate)
0	Middle	0.5544	0.0615	0.3840
1	Old	0.3414	0.1228	0.5358
2	Young	0.7253	0.0168	0.2580

--- Connection: Age_Bin -> CA_Label ---

	Age_Bin	Prob(0.0_Vessels)	Prob(1.0_Vessels)	Prob(2.0_Vessels)	Prob(3.0_Vessels)	Prob(nan_Vessels)
0	Middle	0.5588	0.2490	0.1213	0.0549	0.0160
1	Old	0.3814	0.2482	0.2242	0.1335	0.0128
2	Young	0.8195	0.0900	0.0266	0.0232	0.0407

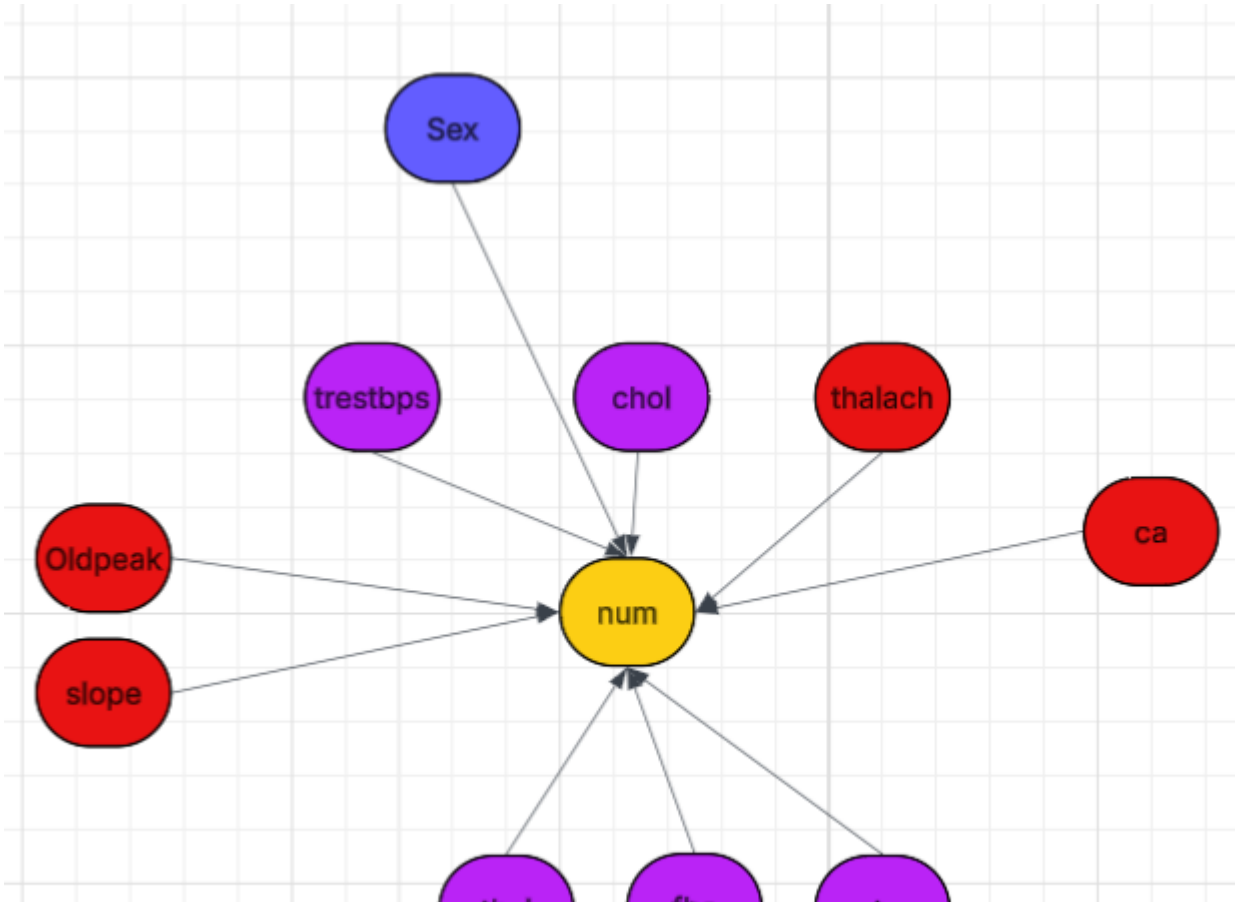
--- Connection: Chol_Bin -> CA_Label ---

	Chol_Bin	Prob(0.0_Vessels)	Prob(1.0_Vessels)	Prob(2.0_Vessels)	Prob(3.0_Vessels)	Prob(nan_Vessels)
0	Borderline	0.5842	0.2167	0.1265	0.0448	0.0278
1	Desirable	0.6157	0.1675	0.1015	0.0753	0.0399
2	High_Chol	0.5411	0.2295	0.1380	0.0820	0.0094

--- Connection: HR_Bin -> CA_Label ---

	HR_Bin	Prob(0.0_Vessels)	Prob(1.0_Vessels)	Prob(2.0_Vessels)	Prob(3.0_Vessels)	Prob(nan_Vessels)
0	High_Rate	0.6788	0.1621	0.1216	0.0203	0.0172
1	Low_Rate	0.2229	0.4389	0.1539	0.1410	0.0433
2	Normal_Rate	0.4766	0.2480	0.1328	0.1218	0.0208

Direct parents --> Disease



--- Connection: Sex_Label -> Disease_Target ---

	Sex_Label	Prob(Negative)	Prob(Positive)
0	Female	0.5385	0.4615
1	Male	0.5132	0.4868

--- Connection: BP_Bin -> Disease_Target ---

	BP_Bin	Prob(Negative)	Prob(Positive)
0	Elevated	0.5324	0.4676
1	High_BP	0.5056	0.4944
2	Normal	0.5202	0.4798

--- Connection: Chol_Bin -> Disease_Target ---

	Chol_Bin	Prob(Negative)	Prob(Positive)
0	Borderline	0.5329	0.4671
1	Desirable	0.5141	0.4859
2	High_Chol	0.5166	0.4834

--- Connection: HR_Bin -> Disease_Target ---

	HR_Bin	Prob(Negative)	Prob(Positive)
0	High_Rate	0.5445	0.4555
1	Low_Rate	0.4938	0.5062
2	Normal_Rate	0.4952	0.5048

--- Connection: Oldpeak_Bin -> Disease_Target ---

	Oldpeak_Bin	Prob(Negative)	Prob(Positive)
0	Ischemia	0.5255	0.4745
1	No_Depression	0.5307	0.4693
2	Severe_Ischemia	0.4919	0.5081

--- Connection: Slope_Label -> Disease_Target ---

	Slope_Label	Prob(Negative)	Prob(Positive)
0	Downsloping	0.5033	0.4967
1	Flat	0.5058	0.4942
2	Upsloping	0.5400	0.4600

--- Connection: CA_Label -> Disease_Target ---

	CA_Label	Prob(Negative)	Prob(Positive)
0	0.0_Vessels	0.5429	0.4571
1	1.0_Vessels	0.4925	0.5075
2	2.0_Vessels	0.4931	0.5069
3	3.0_Vessels	0.4936	0.5064
4	nan_Vessels	0.5037	0.4963

--- Connection: Thal_Label -> Disease_Target ---

	Thal_Label	Prob(Negative)	Prob(Positive)
0	Fixed_Defect	0.4993	0.5007
1	Normal	0.5486	0.4514
2	Reversible_Defect	0.4872	0.5128

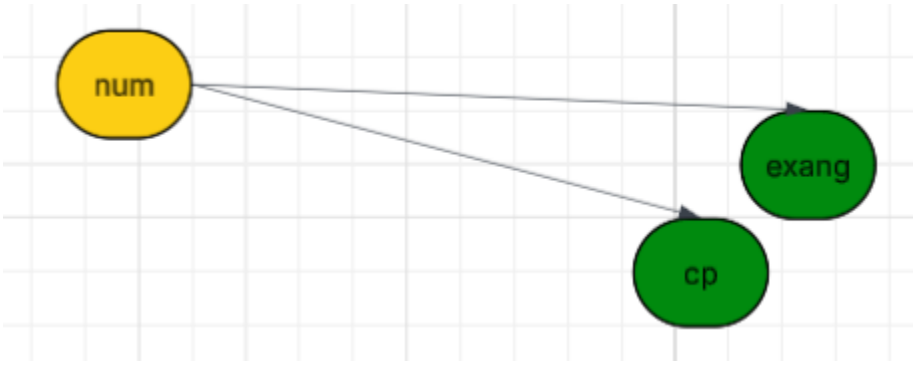
--- Connection: FBS_Label -> Disease_Target ---

	FBS_Label	Prob(Negative)	Prob(Positive)
0	High_Sugar	0.5049	0.4951
1	Normal_Sugar	0.5246	0.4754

--- Connection: ECG_Label -> Disease_Target ---

	ECG_Label	Prob(Negative)	Prob(Positive)
0	LVH	0.5169	0.4831
1	Normal	0.5270	0.4730
2	ST_Abnorm	0.4998	0.5002

Disease --> Symptoms



--- Connection: Disease_Target -> Exang_Label ---

	Disease_Target	Prob(No)	Prob(Yes)
0	Negative	0.8491	0.1509
1	Positive	0.4549	0.5451

--- Connection: Disease_Target -> CP_Label ---

	Disease_Target	Prob(Asymptomatic)	Prob(Atypical_Angina)	Prob(Non_Anginal)	Prob(Typical_Angina)
0	Negative	0.2382	0.2500	0.4098	0.1021
1	Positive	0.7378	0.0712	0.1337	0.0573

How my CPT was found?

Bayesian Parameter Estimation was used, originally we talked about maximum liklihood estimator as alot of other Research articles we have seen use; looking into it, it wont work, if a specific combination never appears in our data the probability becomes 0. This would crash our Bayesian network.

Bayesian Parameter Estimation Math:

$$\theta_{ijk} = \frac{N_{ijk} + \alpha_{ijk}}{N_{ij} + \alpha_{ij}}$$

θ_{ijk} Represents the estimated probability that node “i” is in state “k” given its parents are in configuration “j”

-- e.g node “i” is the “Effect” we are studying like Heart Disease

-- State “k” is the outcome of the specific answer we are looking for like ‘Positive’ they have the disease

-- Parents in config “j” is the setup of risk factors thats being observed like patient is ‘Old’

N_{ijk} Represents the actual count oberved in the data e.g., number of patients who are Male AND High BP P(Male¦High_BP)

N_{ij} Is the total actual count for that parent config e.g total Male patients

α_{ijk} is known as the "Pseudo-count" derived from BDeu prior

The "BDeu Prior": ESS (Equivalent Sample Size) = 10 is distributed uniformly across all possible combinations in the table

$$\alpha_{ijk} = \frac{\text{Equivalent Sample Size}}{\text{Number of Child States} \times \text{Number of Parent Configurations}}\}$$

Example Calculation SEX--> BP_Bin:

- Sex has 2 states Male and Female
- BP_BIN has 3 states (Nrmal, Elevated, High)
- Total cells in this CPT = 2 x 3 = 6
- ESS = 10
- Therefore the added pseudo-count for each cell is **α = 10/6 == 1.67**

If you observe 0 real cases of female with elevated BP it wouldnt say 0% ,instead it calculates:

P(Elevated ¦ Female)= (0 + 1.67) / (Total Females + Total Prior for Females)

What Research articles did i take Inspiration from?

https://link.springer.com/article/10.1186/s12888-025-07189-1

https://publichealth.jmir.org/2022/3/e25658/citations

Where should we plan our next steps in the network building now that we have all of the CPT estimations calculated and confirmed?

This Link will demonstrate where we have idea formed the next steps.

https://www.bayesserver.com/docs/introduction/bayesian-networks/

Under the "Graphical"

Bayesian networks can be depicted graphically as shown in *Figure 2*, which shows the well known *Asia network*. Although visualizing the structure of a Bayesian network is optional, it is a great way to understand a model.

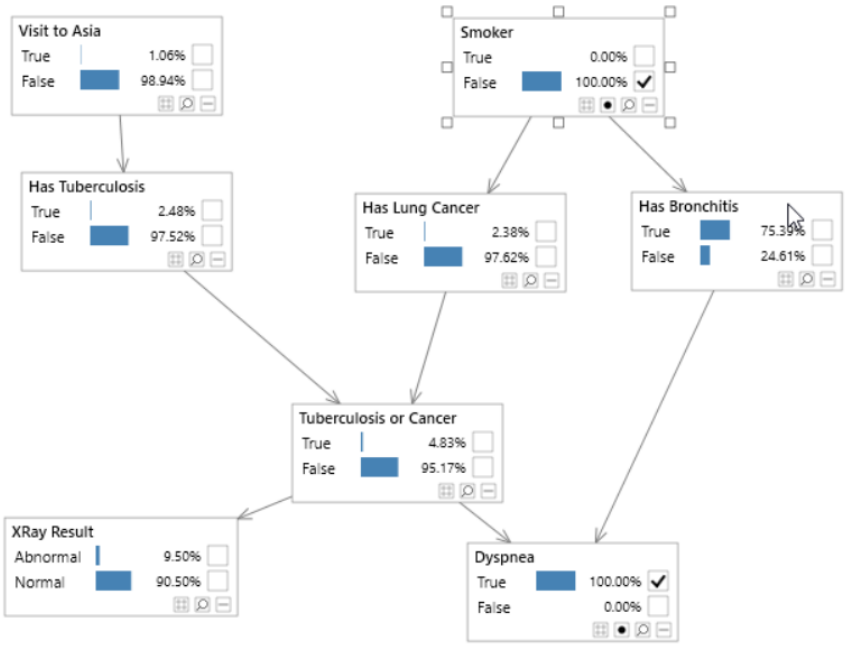


Figure 2 - A simple Bayesian network, known as the Asia network.

The next steps would be to create a visulisation table and combine it with our BN CPT's

Referencing the PROTOYPE project plan on the github

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/PROTOTYPE%20project%20plan%20(1).pdf

We will be heading into Week 5 Stages.