Cardiovascular Risk Bayesian Network Analysis (Prototype Phase)

Github: https://github.com/Plymouth-University/comp2003-2025-2026-team-32

Project Made by:

Jorjit Singh Dasoria

Hussain Ali Raza

Kapeesh Dussain

Project Goal & Objectives

Project Goal "To develop an Interventional Decision Support System based on a Weighted Survival Bayesian Network (WSBN). While our interim baseline models the 14 core UCI Heart Disease variables, the final project scope will integrate Treatment Nodes (e.g., medication for cholesterol). This will evolve the software from a static risk predictor into a dynamic simulator, allowing users to quantify how specific medical interventions reduce cardiovascular risk."

SMART Objectives

1)Validate Baseline Network (Current Status):

We will finalize the current 14-variable, 5-Tier Hierarchical Network using Bayesian Parameter Estimation with BDeu Priors (ESS=10). This establishes a mathematically stable foundation that prevents "zero-probability" errors before introducing complex treatment layers.

2)Integrate Treatment Nodes (Next MVP Phase):

For the upcoming software version, we will research and map Treatment Nodes as binary parents to modifiable risk factors (e.g., Statin Therapy ---> Cholesterol). This objective explicitly expands the network topology to model the causal impact of medical interventions on patient outcomes.

3)Deliver Interactive Simulation Software:

We will develop a functional GUI featuring a 'Treatment Simulator' dashboard. This final deliverable will visualize the expanded network and enable users to toggle treatment nodes, instantly displaying the probabilistic reduction in heart disease risk.

Project Scope

This prototype phase is strictly defined by the ingestion, cleaning, and modeling of the UCI Heart Disease dataset[6]. The data processing scope involves applying four specific binning strategies—Equal-Width, Equal-Frequency, Clustering, and Domain Knowledge—to convert continuous variables into the discrete categories required for Bayesian analysis.

How does the model handle uncertainty without crashing on unseen data?

The modeling scope focuses on applying the Bayesian Estimator formula (θ_{ijk}) rather than simple Maximum Likelihood Estimation. This ensures that unobserved combinations, such as a young patient with high blood pressure, are assigned non-zero probabilities through "pseudo-counts," preventing the model from failing during inference. The deliverables for this phase include the Python source code, the reproducible Jupyter Notebook, and the functional network model.

What lies beyond this current prototype?

To maintain a focused 7-week timeline, we have explicitly excluded the development of the Large Language Model (LLM) wrapper and real-time clinical deployment from this phase. These elements are reserved for Phase 2, where they will be used for a comparative analysis against the baseline established by this Bayesian model.

Software Description for Project Plan

Overview

The team is developing a web-based clinical decision-support prototype for cardiovascular/heart disease (CHD) risk assessment. The software enables a user to enter a patient evidence profile and receive a probabilistic estimate of heart disease risk. The system is designed to prioritise interpretability and traceability of assumptions by exposing both (i) the factors contributing to the computed risk and (ii) the underlying Bayesian network structure that links patient attributes to the disease outcome.

Core User Workflow

The intended workflow is evidence-driven and interactive:

1. The user selects categorical values for patient attributes (demographics, clinical factors, and exercise test results) through a structured input panel.

2. The software computes a heart disease probability estimate from the selected evidence states.
3. The user is presented with a numerical probability, a qualitative risk band (e.g., low/moderate/high), and an at-a-glance positive/negative comparison bar.

4. The system provides explainability through ranked influential factors and a visualisation of the Bayesian network structure.

5. The user can reset the evidence profile to support repeatable evaluation and scenario comparison.

User Interface Components

The interface is organised into three main functional areas:

Patient Attributes (Evidence Input Panel).
A left-side panel collects user-selected patient evidence using predefined categorical states (e.g., age group, biological sex, blood pressure category, cholesterol category, fasting blood sugar category, resting ECG state, thalassaemia state, and exercise test variables such as ST depression/oldpeak, ST slope, and major vessels coloured). These categories reflect the discretised representation used by the model and ensure consistent inputs for probability estimation.

Risk Output (Decision-Support Result).
The central output area presents the computed heart disease risk as a probability value and an associated

qualitative risk label. A positive-versus-negative bar provides an immediate visual interpretation of the estimated likelihood of disease presence versus absence for the entered evidence profile.

Explainability Layer (Rationale and Transparency).
The software provides interpretability through:

- A "Top Influential Factors" panel ranking the factors that most strongly impact the computed risk for the current evidence profile, including quantitative indicators of influence.

- A "Bayesian Network Structure" view that visualises the network relationships between variables and the disease node, grouped into clinically meaningful categories (e.g., demographics/root causes, clinical factors, exercise test variables, symptoms, and disease). Highlighting supports inspection of how specific evidence relates to the outcome.

Probabilistic Computation Approach

The risk estimate is generated using a Bayesian network–style formulation over discretised evidence states. The implemented approach uses conditional probability tables (CPTs) estimated via Bayesian estimation methods, producing a probability of disease for a given evidence profile. This supports an interpretable, probabilistic output aligned with decision support under uncertainty. Where appropriate, the approach can incorporate weighting to reflect differing predictive strengths of variables in the model.

Treatment Support (Decision Support Extension)

The prototype includes treatment concepts as part of the decision-support model. Treatments are represented alongside clinical states to support interpretation and scenario exploration. The intent is not automated prescribing but rather contextual decision support: treatments are associated with modifiable factors and risk-related states so that the interface can communicate how interventions relate to the evidence profile and the computed risk estimate. This supports user understanding of potential management considerations in relation to the model structure and current risk output.

Key Outputs and Current Capabilities

At its current stage, the software demonstrably provides:

- Structured capture of discretised patient evidence.
- Computation of a heart disease probability estimate and risk categorisation
- Transparent explanation through influential factor ranking and Bayesian network visualisation

- Treatment-linked decision-support concepts presented in relation to patient risk context.

Implementation Notes (Deployment)

The software is delivered as a browser-based web application, enabling rapid demonstration and evaluation without specialist installation. The interface supports repeatable testing through evidence reset and consistent categorical input states, aligning with the project's emphasis on reproducibility and traceability.

Research Workflow & Methodology

This project will follow a structured 7-step process. The initial 7-week prototype will focus on rapidly executing this workflow to deliver a functional model, which will be expanded upon in the full project.

Stages:

1. **Data Collection**
    a. Ingest the defined cardiovascular dataset. We will validate the 12 core features and confirm data types.
    b. Pre-defined CSV, `pandas`
2. **Data Preprocessing**
    a. Rigorously clean and prepare the data. This involves managing missing values, correcting inconsistencies
    b. `pandas`, `numpy`
3. **Feature Understanding**
    a. Focusing on the relationship with the `cardio` target variable
    b. Exploratory Data Analysis (EDA) on the three defined feature categories:
        i. **Objective:** age, `height`, `weight`, `gender`
        ii. **Examination:** `ap_hi`, `ap_lo`, `cholesterol`, `gluc`

iii. **Subjective:** smoke, alco, active

4. **Structure Learning**
    a. Mathematically discover the dependency structure between the 12 features
    b. pgmpy, bnlearn
5. **Parameter Learning**
    a. We will calculate the Conditional Probability Tables (CPTs) for each variable, defining *how strongly* it depends on its "parent" variables
    b. pgmpy (e.g., BayesianEstimator)
6. **Inference & Evaluation**
    a. Use the trained model to make predictions and answer queries (e.g., "What is the probability of cardio=1 given cholesterol=3 and smoke=1?")
    b. scikit-learn (Accuracy, Precision, Recall)
7. **Interpretation & Visualization**
    a. visualize the learned network structure
    b. networkx, matplotlib

Weekly plan

Week1: (DATA ACQUIRED & VALIDATED)

Begin Data preprocessing, Understand where the data is required and how accurate it will be by validating it. (Find comparisions with other open source projects)

Week2: (DATA CLEANED & PREPARED)

Understand what kind of data is considered ready for the Bayesian Network.

Converting continuous features like “age” and “weight” into meaningful catagories, personal research shows that bayesian networks work best with catagorical data (Essential for probability tables later)

Removing any impossible readings, Create a script to determine what limits will be deemed unrealistic to ensure they dont skew the model.

Week3: (iNITIAL INSIGHTS REPORT)

Start and complete EDA, Creating visualisations (Graphs and charts) to understand the dataset. “E.g. What is the distribution of cholesterol levels?” “How many individuals in the dataset smoke?” Each Visualisation of different attributes must be shown whats best to visualise any form of spread or relationship.

Start looking into initial relationships between these features and target variables “cardio”.

Week 4: (NETWORK STRUCTURE DEFINED “skeleton”)

During this week the core skeleton will be built. Using certain algorithms HillClimbSearch is an example of one, this is to automatically generate a graph that shows the probabilistic dependencies between all 12 attributes.

Cant trust this algorithm blindly, review the connections made if it seems bizarre investigate by looking into proof on articles that they depend on each other.

Ensuring the graphs are medically plausible.

Then after the SKELETON is done CPTs must be made for the variables (Conditional Probability Tables)

Week5 (VERSION 0.1):

The model must be brought to life.

Complete the Parameter Learning (CPTs), the model will know all relevant and must know probabilities.

Begin on the interface, start testing the models on “What-if" questions. E.g. “what is the risk for a 45 year old non-smoker?, what changes to that risk if they start smoking?”.

Week6: (MODEL PERFORMANCE VALIDATED)

Test the models accuracy, by using a part of our data that the model has never seen also known as a “test set” and see how well it predicts the desired outcome. The hard given metrics will prove accuracy “85%....”.

Then we think deeper:

what are the strongest predictors of heart disease?

What relationships are suprising?

Create an easy-to-read visualisation of the network graph for the client (Yvonne).

Start by creating a presentation showcasing accuracy scores and look into different ways to validate accuracy scores (how many tests?, different data?, will skewed variables work?)

Week7 (PROTOTYPE COMPLETE)

Demonstrate interface capabilites to stakeholders.

Live-demo of what-if questions in realtime, to demonstrate its value as an interactive tool.

Walk through accuracy scores, and the key risk factors the model identified.

Use the success of this prototype to push further completing additional complexities that have been encountered across the first prototype build. Is it visual enough?, Adding a LLM wrapper (e.g. chatgpt) predicting its own answer to the what-ifs and showcasing the comparative analysis agaisnt the bayes network model.

Key Prototype Deliverables

1. Cleaned & processed Dataset
   - Final, analysis ready dataset as a CSV file (or any format thats optimal, tbd...)
   - Data dictionary explaining changed data types changes and description

2. Prototype Bayesian Network Model
   - complete model file (e.g. `.json` format) that can be loaded for interaction
   - A demonstration of its "what-if" query capabilities

3. Source Code & Notebooks
   - A well-commented Jupyter Notebook that allows for a complete, reproducible run of the analysis

4. Initial Findings Report & Presentation
    - A brief report (e.g., Google Doc) summarizing the process, key insights, and prototype performance.
    - A slide deck highlighting the most important findings for all audiences.

5. Phase 2 Project Plan
    - A detailed draft proposal for the full project for the rest of the weeks, more focus on front end display visuals
    - This will include a clear roadmap for the comparative analysis (Bayesian Network vs. LLM) and a plan for developing a publishable research document

Team Role Responsibilities & Stakeholders

Team

This project will be executed by a dedicated team of three, with clear responsibilities aligned with the project workflow.

1. **Project Lead & ML Developer (Jorjit)**
    a. **Primary Focus:** Overall project management, client communication, and core model development.
    b. Key Responsibilities:
        i. Manages the project timeline and ensures milestones are met (Weeks 1-7).
        ii. Leads the technical execution of Structure Learning (Step 4), Parameter Learning (Step 5), and Inference (Step 6).
        iii. Integrates all components into the final prototype model.
        iv. Serves as the primary point of contact for the client.
2. **Data & Domain Research Analyst (Rush)**

a. **Primary Focus:** Data integrity, domain-specific validation, and inference design.
b. **Key Responsibilities:**
    i. Leads deep Exploratory Data Analysis (Step 3).
    ii. Researches and defines "unrealistic" data limits (e.g., for `ap_hi`, `ap_lo`) to guide data cleaning (Step 2).
    iii. Validates the learned network structure for medical plausibility (Step 4).
    iv. Leads research on the "what-if" query interface, defining the user interaction for the prototype demo (Step 6 & 7).

3. **Data Processing & Evaluation Engineer (Hussain)**
    a. **Primary Focus:** Building the data pipeline and rigorously evaluating the final model.
    b. **Key Responsibilities:**
        i. Executes the Data Collection and Preprocessing pipeline (Step 1 & 2), including discretization and outlier handling.
        ii. Develops the model evaluation framework (Step 6), including train/test splits and accuracy metrics (Accuracy, Precision, Recall).
        iii. Leads the final Interpretation & Visualization (Step 7), creating the network graphs and performance charts for the final report

Stakeholders

Stakeholders:

- **Client (Yvonne):** Primary approver; reviews the network visualization and final prototype functionality.
- **Teaching Staff/Tutors:** Evaluate the project against the "Interim Threshold" rubric, specifically looking for detailed AI descriptions and professional execution.
- **Project Team:** Responsible for execution, documentation, and delivery.

Research

Enhanced Bayesian Network Prototype Documentation This technical documentation details the architectural enhancements made to the Bayesian Network prototype, specifically the integration of "Tier 3" treatment nodes for lifestyle changes, blood pressure

medications, and cholesterol medications to address supervisor feedback. It introduces a temporal separation between "baseline" and "current" clinical factors to model how interventions modify disease probabilities, and outlines the logic for dynamic risk calculation based on protective and risk factors. Additionally, the document defines the visual standards for the interface and validates the probability values against medical literature, such as the specific risk reductions associated with statins and ACE inhibitors.

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/treatment_prototype_documentation..pdf

Methodology for Generating a Decision Tree for CVD Prediction This report outlines the methodology for generating a Decision Tree (DT) for cardiovascular disease prediction, utilizing feature selection and data preparation techniques derived from Bayesian Network literature to ensure high data quality. It details the critical steps of handling missing data via imputation, discretizing continuous variables like age and BMI into clinically meaningful categories, and selecting appropriate splitting algorithms such as CART or C4.5 to maximize information gain. Furthermore, it establishes a validation framework using metrics like Accuracy, Sensitivity, and AUC to systematically compare the interpretability of rule-based Decision Trees against probabilistic Bayesian models

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/Methodology_for_Generating_a_Decision_Tree_for_Cardiovascular_Disease_(CVD)_Prediction.pdf

Medical Plausibility Analysis: UCI Heart Disease Dataset This analysis provides a comprehensive medical validation of the 14 core attributes within the UCI Heart Disease dataset, confirming their biological plausibility and relevance for predicting coronary artery disease. It systematically reviews each variable—ranging from physiological markers like resting blood pressure and cholesterol to diagnostic test results like ST depression and fluoroscopy—to ensure the data units and categorical encodings align with 1980s clinical standards. This document serves as the foundation for the network's domain knowledge, confirming strong expected dependencies such as the inverse relationship between age and maximum heart rate.

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/Medical%20Plausibility%20Analysis_%20UCI%20Heart%20Disease%20Dataset%20.pdf

Medical Justification: Inverse vs. Direct Relationships This document establishes the "Ground Truth" for the Bayesian Network's structure by defining the specific causal, inverse, and direct relationships between the 14 dataset attributes based on medical science. It details the biological logic dictating network connections, such as the "inverse" relationship where aging reduces maximum heart rate, and the "causal" link where hypertension leads to left ventricular hypertrophy on ECGs. The report concludes with a summary table categorizing each node as a root cause, symptom, or test result, ensuring the model's probability calculations reflect physiological reality rather than random correlation.

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

Analysis of Probability Determination in Bayesian Networks This research summary contrasts the parameter learning methodologies used in three key cardiovascular risk studies (Suo et al., Ordovas et al., and Kong et al.) to determine Conditional Probability Tables (CPTs). It analyzes distinct approaches including Weighted Maximum Likelihood Estimation using survival analysis to handle censored data, Bayesian Parameter Estimation utilizing Dirichlet priors for expert-informed models, and standard Maximum Likelihood Estimation for complete datasets. This comparison informs the project's choice of estimation techniques by highlighting how different methods address data completeness and uncertainty.

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/Analysis%20of%20Probability%20Determination%20in%20Bayesian%20Networks%20for%20Cardiovascular%20Risk%20Prediction.pdf

Bayesian Network Treatment Integration Analysis This document proposes a strategic extension to the Bayesian Network to address supervisor feedback by integrating "Tier 4A" treatment intervention nodes for lifestyle changes and medications. It outlines a transition from a purely observational model to an interventional one, using literature-based estimation (e.g., meta-analyses of antihypertensive efficacy) to construct CPTs where dataset values are missing. The plan details the technical implementation of "do-calculus" for query functionality and establishes a validation strategy using sensitivity analysis to model the probabilistic impact of medical interventions on disease risk.

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/bayesian_network_treatment_analysis.pdf

Solution

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)

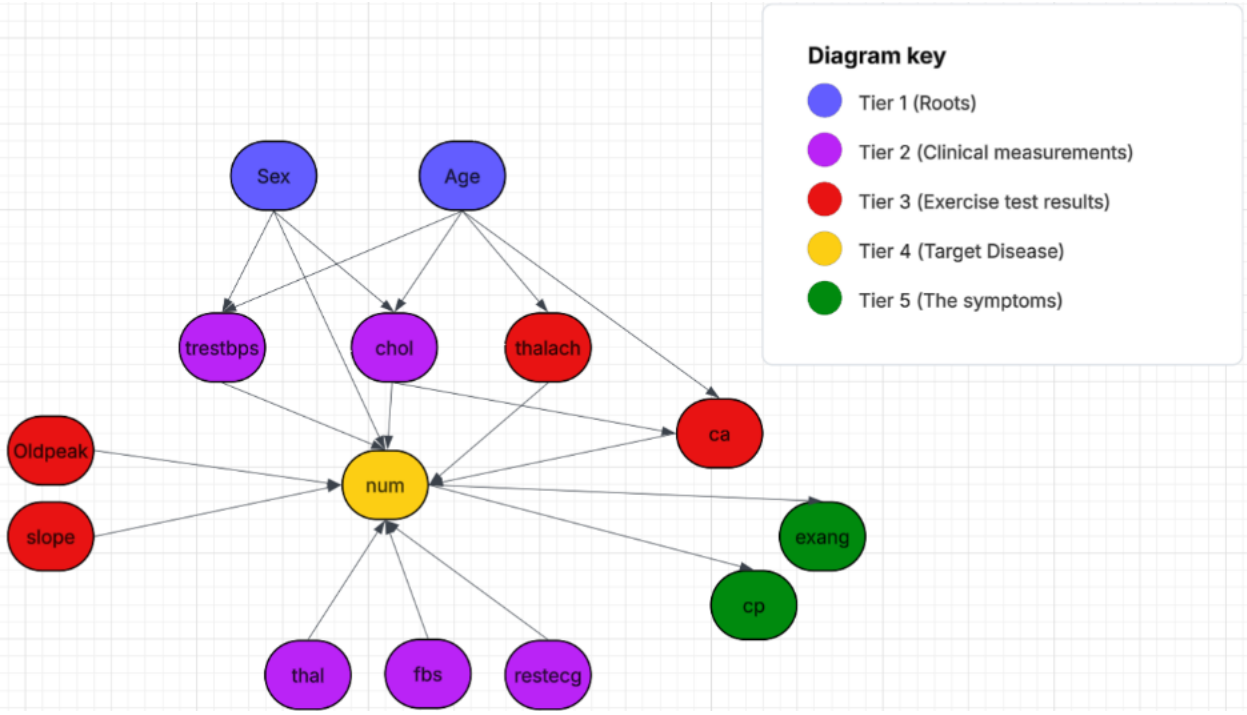
Tier4: the disease (target)

- 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”

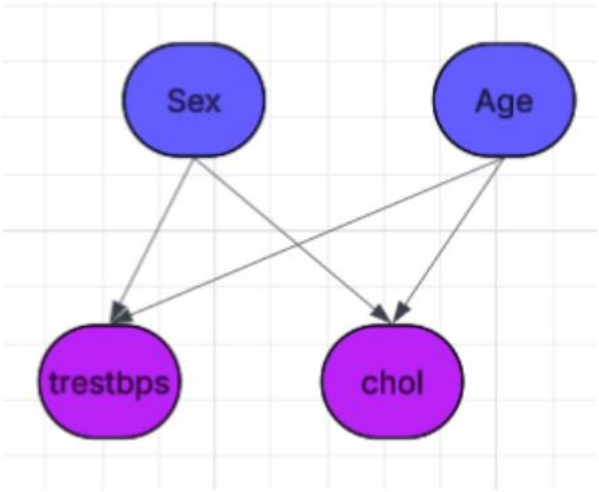
All of the Necessary analysis based off the EDA.IPYNB, of relationships are explained in this document.

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

BN Conditional Probability tables

Tier 1 --> Tier 2


Sex
Age
trestbps
chol

--- 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

--- Connection: Age_Bin -> Chol_Bin ---

	Age_Bin	Prob(Borderline)	Prob(Desirable)	Prob(High_Chol)
0	Middle	0.3220	0.1381	0.5399
1	Old	0.2937	0.1310	0.5752
2	Young	0.3665	0.2734	0.3601

--- 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

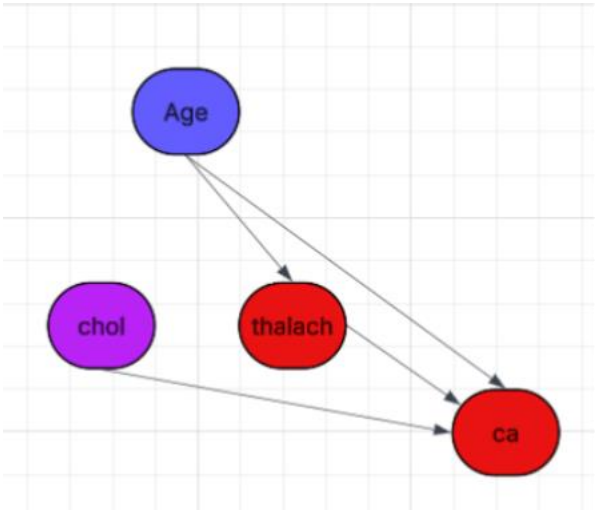
--- Connection: Chol_Bin -> CA_Label ---

	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

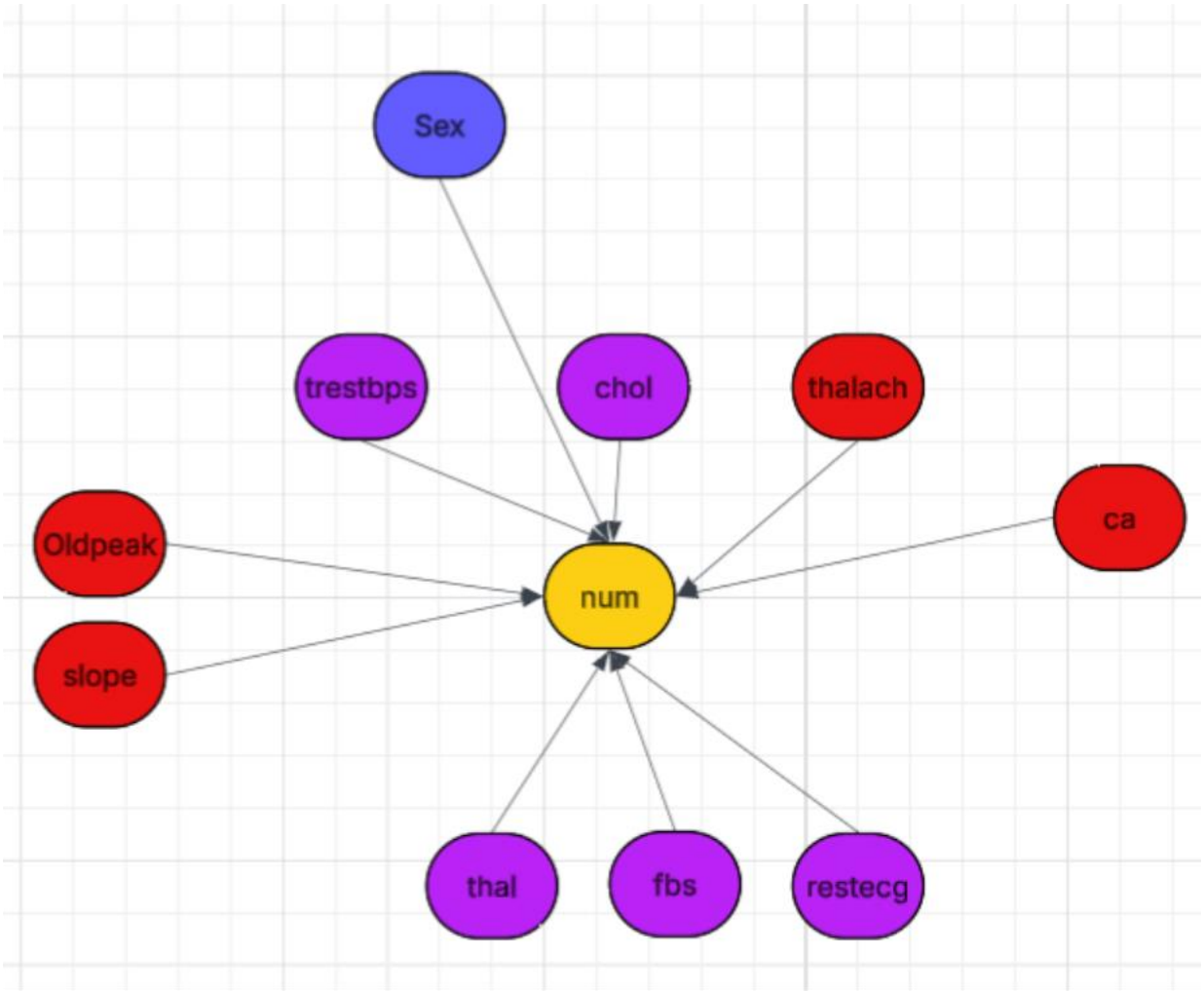
--- Connection: HR_Bin -> CA_Label ---

	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

Tier 1&2 --> 3



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

Wrong way links

--- 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.

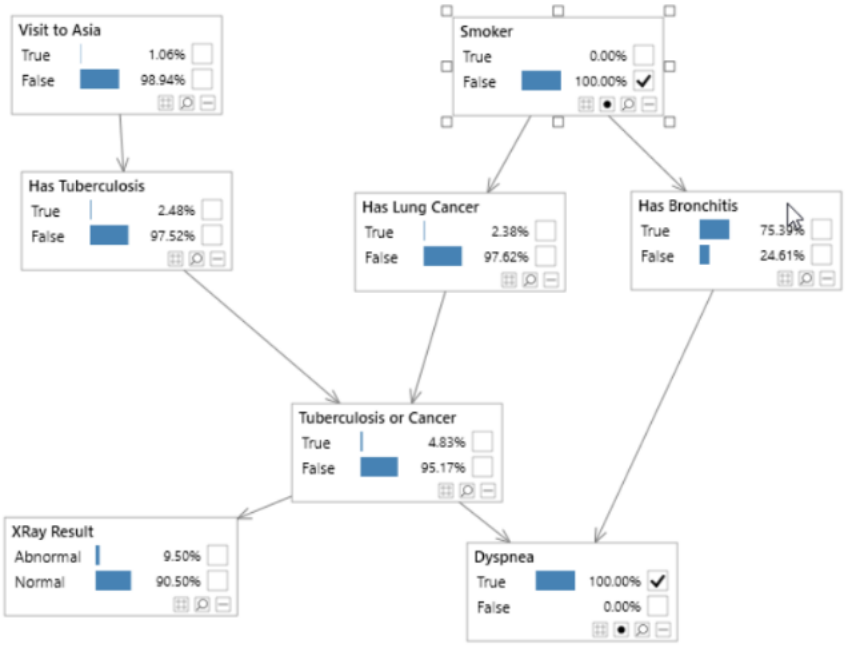


Figure 2 - A simple Bayesian network, known as the Asia network.

The next steps would be to create a Visualisation table and combine it with our BN CPT’s.

Resource Plan

Human Resources: Roles defined earlier in the document (Jorjit, Rushil, Hussain)

Hardware: High-performance Laptops capable of running jupyter Notebooks and local server environments

Software & Development Architectures

This Prototype utilises dual-prototype strategy to maximise both AI-assisted rapid development and robust backend scalability

Main interactive Prototype

- Tech Stack: Vite, TypeScript, React, shadcn-ui, Tailwind CSS (built via Lovable).

- Purpose: This version represents our "extensive AI use" deliverable. It uses the Lovable AI engine to synthesize the frontend structure (created by Rush/Hussain) with the probabilistic logic derived from the data_binning.ipynb notebook (developed by Jorjit). This allows for rapid visualization of the CPT variables in a modern, responsive interface.

By combining the:

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

With

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/SourceCode/bayesian_network_with_treatments.html

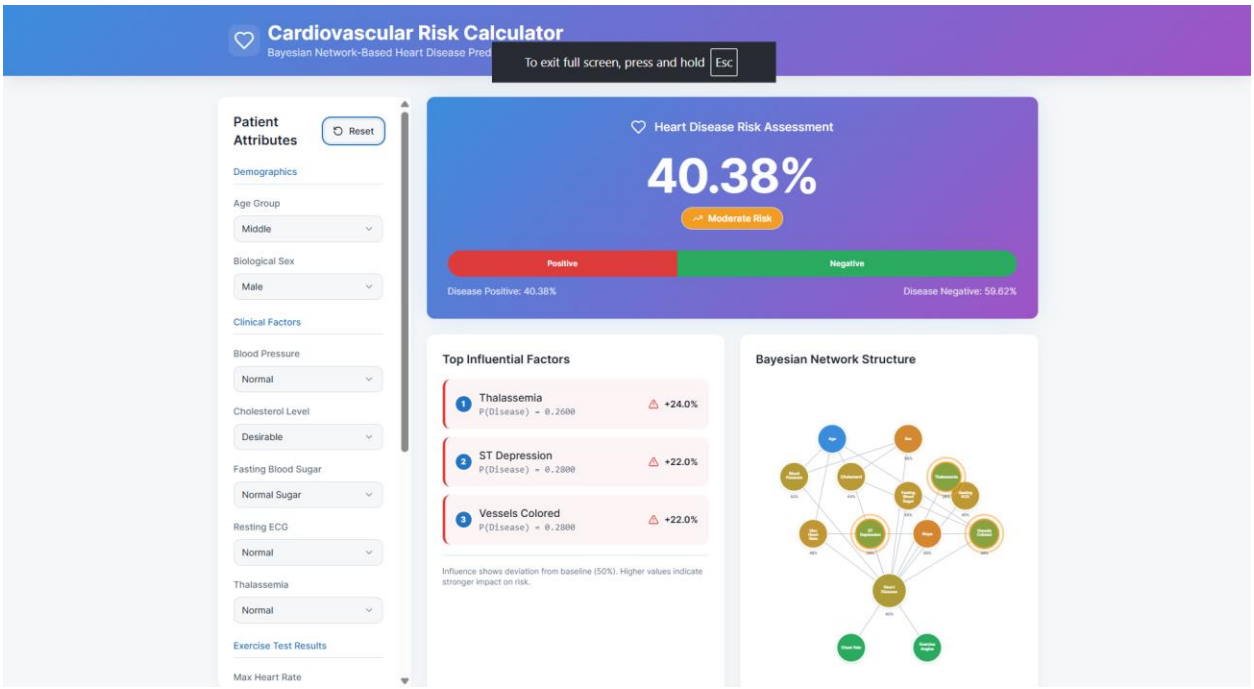
To Reference the prompts used:

https://github.com/Plymouth-University/comp2003-2025-2026-team-32/blob/main/Design%20Documents/Prompts/lovable_ai_deployment_instructions.md

Prototype Interactive URL

CLICK THIS URL BELOW TO INTERACT WITH OUR PROTOTYPE

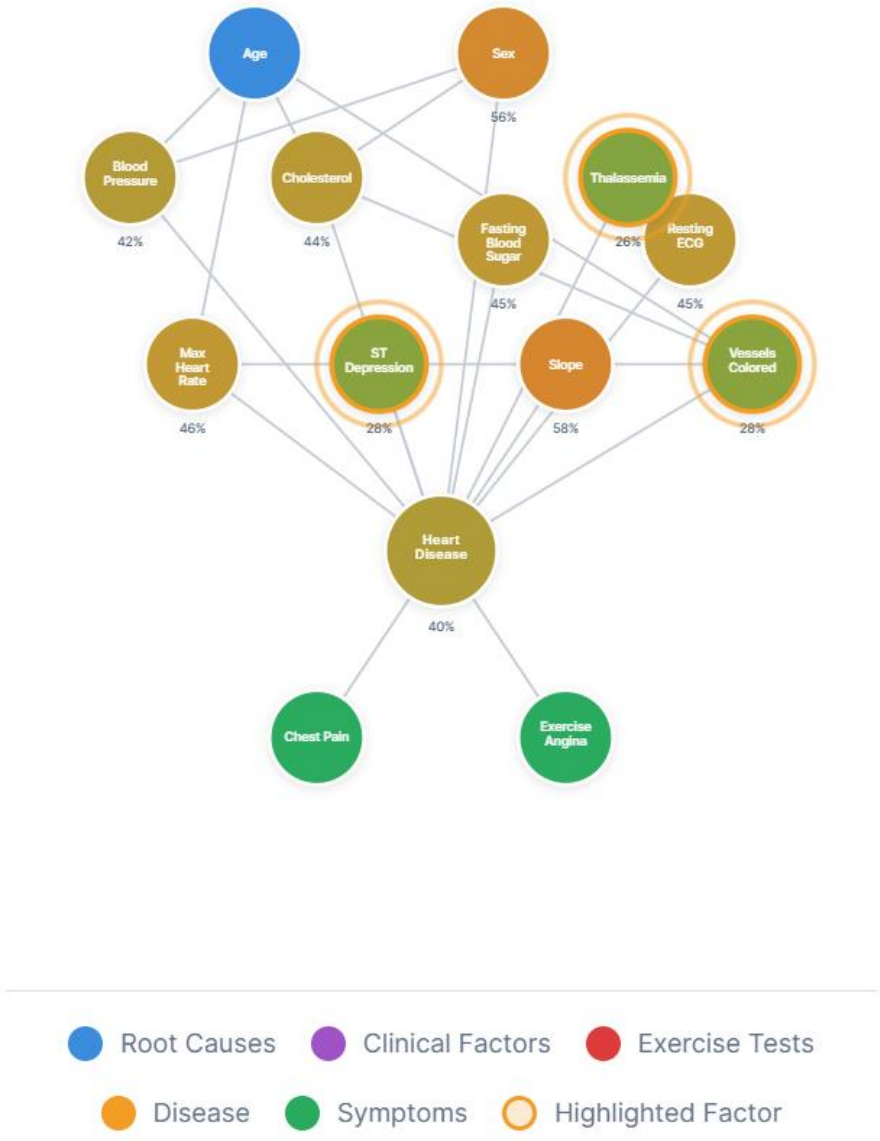
https://cardiovascular-risk-calculator.lovable.app/



How does lovable work?

- **Synthesis of Components:** It combines the structural frontend elements (the index.html created by **Hussain and Rush**) with the mathematical logic (the Conditional Probability Variables/CPTs) derived from **Jorjit's** data_binning.ipynb.
- **Modern Tech Stack:** The Lovable engine automates the construction of a modern web application stack, building the project with:
    - **Vite:** For fast build tooling and development.
    - **TypeScript & React:** For a robust, component-based user interface.
    - **Shadcn-ui & Tailwind CSS:** For rapid, responsive, and professional styling.
- **Function:** It acts as the "Working Interactive Prototype," serving as the visual front-end that allows stakeholders to interact with the Bayesian probabilities immediately, distinct from the Python-based backend reserved for the final submission.

Network Visualization & UI Enhancements



To align with the project's shift from theoretical analysis to a deliverable software product, we have significantly upgraded the network visualization. Using the **Lovable** interface design platform, we transitioned from the initial schematic (Figure 2) to a user-centric, interactive dashboard (Figure 1).

Key improvements include:

Semantic Labeling: We replaced cryptic variable codes (e.g., *trestbps*, *num*, *ca*) with clear, natural language labels (e.g., *Blood Pressure*, *Heart Disease*, *Vessels Colored*). This

ensures the model is immediately interpretable by clinicians and non-technical stakeholders without requiring a reference manual.

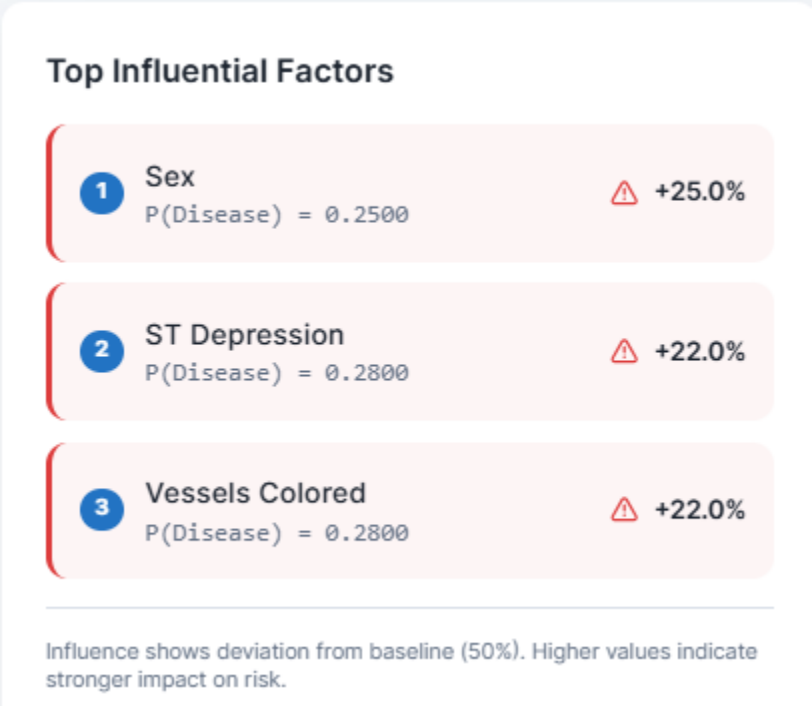
Integrated Probability Indicators: Unlike the static previous version, the new interface displays real-time probability percentages directly below each node (e.g., "Max Heart Rate 46%"). This reflects the "Weighted Survival" nature of our model, giving users instant quantitative feedback on the network's state.

Refined Hierarchy & Key: We implemented a streamlined legend at the bottom of the interface to clearly distinguish our **5-Tier Hierarchy**. The color-coding is now softer and more distinct:

- **Blue (Root Causes):** Fixed attributes like Age.
- **Purple/Brown (Clinical Factors):** Modifiable biomarkers like Cholesterol.
- **Red/Orange (Exercise Tests):** Stress-test results.
- **Green (Symptoms):** Patient-reported outcomes.

Modern UI Polish: The new radial layout reduces visual clutter, using clearer connection lines and "Highlighted Factor" rings (orange outlines) to draw attention to critical nodes. This cleaner aesthetic improves the usability of the Decision Support System, making complex dependencies easier to trace during "what-if" simulations.

Top Influential Factors Dashboard



To enhance diagnostic clarity, we implemented an algorithmic ranking system (Figure 3) that isolates the primary drivers of patient risk. Instead of requiring users to manually trace network paths, this feature explicitly quantifies how much specific attributes—such as *ST Depression*—deviate from the baseline probability (e.g., +22.0%). This acts as an immediate "triage list," filtering complex dependencies into actionable insights that will dynamically reshuffle as users simulate treatments in the final software.

Risk Management

To ensure project stability, we have identified three critical risks and their corresponding mitigation strategies:

- **Zero-Probability Crash:** There is a risk that specific patient profiles (e.g., "Young" + "High BP") may not exist in the training data, causing standard Maximum Likelihood Estimation models to crash.
    - *Mitigation:* We will implement **Bayesian Parameter Estimation** using **BDeu Priors** to add "pseudo-counts" ($\alpha = 1.67$), ensuring the probability distribution is smoothed and valid for all combinations.

- **Unrealistic Data Outliers:** Impossible physiological readings (e.g., Blood Pressure > 250) could skew the model's predictions.
    - *Mitigation:* We will apply **Domain Knowledge limits** during preprocessing to rigorously filter data against established medical thresholds.
- **Unequal Team Contribution:** Risks regarding unbalanced workloads.
    - *Mitigation:* We will enforce strict **Git commit tracking** and conduct sprint reviews as per the "Individual Process" rubric, ensuring every member meets the contribution threshold.

Communication Plan

- **Internal:** The team will maintain momentum through weekly sprint meetings and daily asynchronous updates via Trello.
- **External:** Progress will be validated through scheduled feedback sessions with the Client (Yvonne) and Sprint Tutors.
- **Documentation:** To ensure transparency and reproducibility, all major architectural decisions will be logged in Trello, while technical logic will be documented via detailed comments within the Jupyter Notebooks.

Quality Management

- **Medical Plausibility:** The network structure is strictly validated against the **"Medical Justification"** document to ensure that causal links—such as the inverse relationship between Age and Max Heart Rate—align with 1980s clinical standards and biological trends.
- **Code Quality:** All deliverables will adhere to the **"70+ (First Class)"** standards, characterized by extensive design detail, structured commits, and clear AI use descriptions.

Monitoring and Evaluation

We will track progress using the following Key Performance Indicators (KPIs):

- **Model Accuracy:** Achieving a target predictive accuracy of >85% on the test set[11].

- **Inference Capability:** The successful execution of "What-if" scenarios (e.g., verifying that increasing Cholesterol levels correctly propagates an increased probability to the Disease node).
- **Contribution Metrics:** Maintaining a Git commit frequency above the threshold of 10 meaningful commits per member per semester.

Budget

- **Financial Cost:** £0.00.
- **Strategic Savings:** The project leverages the **Python Open Source ecosystem** (pandas, pgmpy) and free public data from the **UCI Machine Learning Repository**, effectively eliminating costs associated with data licensing and proprietary software fees.
- **Using free** lovable credits to utilise the server running the interim prototype

Approval Process

- **Milestones:** Structure "Skeleton" approval is scheduled for Week 4, with the Final Prototype approval set for Week 7.
- **Signoff:** Formal signoff is required from the Client (Yvonne) for both the finalized Project Plan and the Prototype delivery.

Change Management

- **Protocol:** Any changes to major sections, such as the Timeline or Deliverables, must be discussed and agreed upon as a group.
- **Logging:** All approved changes will be logged in Trello to maintain a clear audit trail, satisfying the documentation requirements of the "Process" rubric.
- **Semester 2 Expansion:** This plan is designed to be extensible, allowing for future scope expansion (e.g., LLM integration) via a formal change request in the following semester.

Closure and Evaluation

- **Final Deliverables:** The project will conclude with the delivery of the Cleaned Dataset, the functional Prototype Model (.json/python), Reproducible Notebooks, and the Initial Findings Report.

- **Post-Implementation Review:** A final review will be conducted to assess whether the project met the "First Class" criteria, specifically evaluating the depth of the plan, the clarity of AI descriptions, and the polish of the final prototype.

Appendices