Current Known Dataset(clean): https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

Useful similar link projects: https://github.com/caravanuden/cardio/tree/master

Note on Revisions:

This document serves as our team's agreed-upon plan for the 7-week prototype. To ensure we stay aligned and meet client expectations, please discuss any potential changes to major sections (especially Section 3: Timeline and Section 4: Deliverables) as a group. We can set a brief agenda for these discussions to keep them focused. All approved changes will be logged in Trello.

Cardiovascular Risk Bayesian Network Analysis (Prototype Phase)

Project Goal & Objectives

The primary goal of this research project is to develop a predictive Bayesian network model to understand the complex inter-dependencies between various cardiovascular risk factors.

Key Objectives:

- To identify the key variables that are probabilistically linked to cardiovascular events.
- To quantify the strength of these relationships.
- To build an interactive prototype model capable of performing "what-if" scenario analysis (inference).
- To provide actionable insights that can aid in risk assessment and decision-making.
- To lay the groundwork for a comparative analysis of this model against other predictive methods (e.g., LLM-based approaches).

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 Roles & Responsibilities

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

COMP2003 project plan guidelines

This “Prototype Project Plan” completes:

- 1, 2,3, 4, 6, 7, 9, 13, 15, 17,

Project Plan Structure

1. **Project Title:** Clearly state the name of the project.
2. **Project Overview:** Provide a brief description of the project, including its purpose, goals, and objectives.
3. **Project Scope:** Define the boundaries of the project, specifying what is included and excluded.
4. **Project Objectives:** Clearly state the measurable and achievable outcomes the project aims to accomplish.
5. **Stakeholders:** Identify and list all stakeholders involved in the project, including their roles and responsibilities.
6. **Project Team:** Outline the members of the project team, their roles, and reporting relationships.
7. **Timeline:** Create a detailed timeline with milestones and deadlines for key deliverables. Use a Gantt chart to illustrate this clearly.
8. **Research:** present findings from existing work that address the problem statement, scope and objectives of the project. Main part of this section is to address existing competition and solutions, and how your project is unique in its approach. You may pull material from your design document for this.
9. **Proposed Solution:** an overview of your project's solution and methodologies. This should be in line with your overview, scope and, objectives, timeline and further elaborated on under the Work Breakdown Structure next.
10. **Work Breakdown Structure (WBS):** Break down the proposed solution into smaller, manageable tasks and subtasks. Create a hierarchical structure showing the relationship between different tasks. You may pull material from your design document for this.
11. **Resource Plan:** Identify the resources (human, financial, equipment) required for each task and allocate them accordingly.
12. **Risk Management:** Identify potential risks and develop strategies for risk mitigation. Include contingency plans for addressing unforeseen issues.
13. **Communication Plan:** Define how communication will be handled throughout the project, including regular meetings, reporting mechanisms, and channels of communication.
14. **Quality Management:** Specify the quality standards and processes that will be used to ensure the project's deliverables meet the required criteria.
15. **Monitoring and Evaluation:** Outline how the project's progress will be monitored and evaluated, including key performance indicators (KPIs).
16. **Budget:** Provide a detailed budget outlining the costs associated with the project, including resources, materials, and any other relevant expenses. Cost savings using open-source solutions that are free are a big plus point to be mentioned here.
17. **Approval Process:** Clearly define the process for obtaining approvals at different stages of the project. For this module, you will require the client's approval/signoff for this.
18. **Change Management:** Describe how changes to the project scope, schedule, or resources will be identified, evaluated, and implemented. This is in case the client wants to request changes in semester 2, what's your process for doing so?
19. **Closure and Evaluation:** Outline the steps for closing out the project, including a post-implementation review and lessons learned. This should be in line with the final submission checklist provided in the Handbook.
20. **Appendices:** Include any additional documentation or reference materials that support the project plan.

BUT, it currently does not achieve:

- 5, 8, 10, 11, 12, 14, 16, 18, 19, 20