Bayesian Network for Cardiovascular Disease Prediction

Project Meeting Summary

November 17th, 2025

OVERVIEW

The team presented significant progress on their Bayesian network project, including dataset selection, exploratory data analysis, and critical technical decisions about network structure. The meeting focused on validating the team's approach against published research and determining methods for conditional probability table estimation.

DATA AND DATASET PROGRESS

Dataset Selection Completed: Converted original data from .data file to CSV using Python. Dataset cleaned and ready for analysis with 303 patient records and 14 selected attributes from credible cardiovascular disease literature.

Selected Attributes (14 total): Age, gender, smoking, alcohol intake, resting blood pressure (systolic and diastolic), cholesterol, glucose levels, chest pain types (typical angina, atypical angina, asymptomatic), maximum heart rate achieved, exercise-induced chest pain, ST depression, and target: cardiovascular disease severity (0-4 scale).

Scale Explanation: 0 = no cardiovascular disease, 1 = mild disease presence, 2-3 = moderate disease, 4 = extremely severe disease.

CRITICAL TECHNICAL CLARIFICATION

The team confirmed they are building a Bayesian network, not a decision tree. This distinction matters significantly for the project's architecture and implementation approach.

Decision Trees: Use sequential, linear rules where one path leads to one outcome. Less suitable for complex interdependencies typical in medical diagnosis.

Bayesian Networks: Model probabilistic relationships with multiple variables influencing outcomes. Variables link bidirectionally and can represent complex causality, making them better suited to medical diagnosis where multiple factors interact.

Why This Matters for CPTs: One connection line in a Bayesian network does not equal one probability value. If multiple nodes connect to a single outcome node, the connection must account for all combinations, creating a table rather than a simple probability.

NETWORK STRUCTURE DEVELOPED

5-Tier Hierarchical Model: Tier 1 (Root Causes): Age, gender, lifestyle factors. Tier 2 (Intermediate): Blood pressure, cholesterol, glucose. Tier 3 (Clinical Factors): Symptom types, exercise tolerance. Tier 4 (Exercise Response): Exercise-induced symptoms, ST depression. Tier 5 (Target): Cardiovascular disease outcome.

Relationship Example: Age increases with higher chance of elevated cholesterol, which increases chest pain likelihood, which raises cardiovascular disease probability. All connections validated through medical literature.

EXPLORATORY DATA ANALYSIS COMPLETED

Correlation Heatmap Analysis: Using pandas and seaborn libraries, the team generated a correlation matrix showing all variable relationships. Chest pain shows 0.41 positive correlation with disease presence. Purpose: verify that hypothesized relationships exist in actual data.

Medical Plausibility Analysis: Created separate documentation explaining why each variable was selected, medical reasoning for each connection, supporting research references, and addressing why certain attributes matter for the diagnosis model.

CONDITIONAL PROBABILITY TABLE METHODS

Method 1: Maximum Likelihood Estimation: Available in PGMPY Python library. Formula: (count of specific case) / (total population). Mathematical, straightforward, and automatically calculated from data.

Method 2: Value Binning with Ranges: Break continuous variables into bins (e.g., age: <40, 40-60, >60). Calculate separate probabilities for each bin. Create discrete lookup tables. Used in published research papers.

Method 3: Greedy Hill-Climbing Algorithm: Statistical learning approach to find optimal network structure using conditional dependence measures. Identifies important relationships. More complex but potentially more accurate.

RESEARCH PAPERS AND COMPARISONS

Papers Reviewed: Three major papers on Bayesian networks for cardiovascular disease prediction. Papers used 13 attributes (team using 14, close match). Expert-validated networks showed refinements from initial versions. All papers documented probability estimation methodology clearly.

Advisor Guidance: Consider adopting network structures from published papers to avoid reinventing relationships already established medically. If team's network differs from papers, must justify with strong medical evidence.

PROBABILITY DEFENSIBILITY CONCERNS

Concern 1: Mathematical vs. Medical Meaning: Can maximum likelihood estimation accurately capture medical relationships? Mathematical correlation does not guarantee meaningful causality. Weak correlation might be included mathematically but not medically relevant.

Concern 2: Assumption of Relationships: Need proof that relationships are significant and directional. Cannot assume all correlations represent true causality. Every connection requires medical justification.

Concern 3: Arbitrary Probability Assignment: Cannot assign probabilities based on intuition. Every CPT value must have documented source. Must explain reasoning for each probability value.

WORK BREAKDOWN

Completed (Weeks 1-2): Data cleaning and CSV conversion. Dataset and attribute selection. EDA with correlation analysis. Network structure design. Medical plausibility documentation. Research paper review. CPT method identification.

Still Required: Select best CPT estimation method. Get AI-generated network for comparison. Document AI usage methodology. Implement CPT calculation programmatically. Finalize network structure with medical justification. Create final probability tables.

NEXT WEEK DELIVERABLES

Task 1: Medical Justification Document: Explain every connection in the network. Provide medical literature support. Defend why each variable is included. Clarify inverse vs. direct relationships.

Task 2: CPT Estimation Method Selection: Compare maximum likelihood vs. other methods. Determine best approach for team's data. Understand statistical theory behind chosen method.

Task 3: Dual-Track Implementation: Track A: Use AI to generate decision tree from attributes and compare with published research. Track B: Implement chosen estimation method programmatically and calculate probability values.

Task 4: AI Usage Documentation: Explain how AI was used in network creation. Show validation against research papers. Document what was AI-generated vs. team-determined.

KEY TECHNICAL INSIGHTS

Network vs. Tree Understanding: A network means a single line has multiple variations of values. A decision tree means one line represents one probabilistic value. This distinction drove all technical decisions.

Circular Networks Acceptable: Papers reviewed showed networks that were not strictly hierarchical, validating the team's complex structure approach.

One Line, Multiple Probabilities: Connections cannot have single probabilities. This explains why CPT tables exist separate from connection lines in the diagram.

ADVISOR'S FINAL RECOMMENDATIONS

Focus primarily on CPT estimation methodology. Ensure every probability has documented source. Use published research as validation framework. Maintain clear documentation for eventual publication. Test probability calculation programmatically before finalizing.

TIMELINE ASSESSMENT

Progress Rate: Good with two weeks of substantial groundwork completed. Remaining Time: 5 weeks, adequate for implementation. CPT estimation is main remaining blocker. Once resolved, visualization and documentation will follow.