Methodology for Generating a Decision Tree for Cardiovascular Disease (CVD) Prediction

This report outlines the theoretical and practical steps required to generate a **Decision Tree (DT)** model for Cardiovascular Disease (CVD) prediction, using the types of variables and data characteristics identified in the reviewed academic literature on Bayesian Networks [1] [2] [3].

Decision Trees are a type of supervised machine learning algorithm that are highly valued in clinical settings for their **interpretability**. Unlike complex "black-box" models, a DT provides a clear, rule-based flow chart that mimics human decision-making, making it easy for clinicians to understand the path leading to a prediction.

1. Data Preparation and Feature Engineering

The quality of the final decision tree is highly dependent on the preparation of the input data. The variables identified in the Bayesian Network studies provide an excellent starting point for feature selection.

1.1. Feature Selection and Target Variable

The input features (or nodes in the BN) would include the key risk factors identified in the literature, such as:

- **Demographic:** Age, Sex, Education level, Socioeconomic status [2] [3].
- **Lifestyle/Behavioral:** Physical activity, Sleep duration, Smoker profile, Diet [2] [3].
- **Clinical/Biochemical:** Hypertension, Hypercholesterolemia, Diabetes, Blood pressure, Lipid levels [2] [3].

The **Target Variable** (the outcome to be predicted) would be a binary classification:

- **CVD Outcome:** Coronary Heart Disease (CHD) / Coronary Artery Disease (CAD) (Yes/No) [1] [3].

1.2. Handling Missing Data

The studies, particularly Suo et al. [1], highlight the challenge of **missing data** in Electronic Health Records (EHRs). Before training a DT, missing values must be addressed:

- **Imputation:** Missing values for continuous variables (e.g., blood pressure) can be replaced with the mean or median. For categorical variables (e.g., smoking status), the mode or a separate "Unknown" category can be used.

- **Advanced Techniques:** The Bayesian Network approach used by Suo et al. [1] (Weighted Survival Bayesian Network) is specifically designed to handle missing and censored data, which suggests that a simple DT model might require more robust imputation or the use of algorithms that can handle missing data intrinsically.

1.3. Discretization of Continuous Variables

Decision trees perform optimally with categorical or discrete data. The Ordovas et al. study [2] provides a clear example of how continuous variables are converted into discrete levels:

- **Age:** Grouped into ranges, e.g., (24,34], (34,44], (44,54], etc.
- **Body Mass Index (BMI):** Categorized into discrete levels: *underweight, normal, overweight, obese*.
- **Sleep Duration:** Categorized into *short, normal, excessive*.

This discretization process is crucial as it transforms complex continuous data into simple, clinically meaningful rules (e.g., "If Age is > 54 and BMI is Obese...").

2. Decision Tree Algorithm and Training

The core of generating a decision tree is the recursive partitioning process, which involves repeatedly splitting the dataset based on the feature that provides the most information gain about the target variable.

2.1. Algorithm Selection

Two primary algorithms are commonly used for classification:

- **Classification and Regression Tree (CART):** Uses the **Gini Index** as the measure of impurity. It tends to produce binary trees (each node splits into two branches).
- **C4.5 / C5.0:** Uses **Entropy** and **Information Gain** to select the best split. It can produce multi-way splits.

In the context of CVD prediction, the goal is to find the most effective split that separates patients with CVD from those without.

2.2. The Splitting Criterion (Impurity Measure)

The algorithm evaluates every possible split for every feature and selects the one that results in the lowest impurity in the resulting child nodes.

Impurity Measure	Formula (Conceptual)	Goal

Gini Index (CART)	$1 - \sum_{i=1}^{C} (p_i)^2$	Measures how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. A lower Gini Index is better.
Information Gain (C4.5)	$Entropy(Parent) - \sum_{j=1}^{k} \frac{N_j}{N} Entropy(Child_j)$	Measures the reduction in entropy (or uncertainty) achieved by the split. A higher Information Gain is better.

For example, the model might first test a split on **Age**. If splitting at "Age > 60" results in a much purer set of "CVD Positive" cases in one branch and "CVD Negative" cases in the other, that split is chosen as the root of the tree.

2.3. Stopping Criteria and Pruning

The recursive splitting continues until a **stopping criterion** is met:

1. All instances in a node belong to the same class (perfect purity).
2. The number of instances in a node falls below a pre-defined minimum (e.g., minimum 20 patients per leaf node).
3. The Information Gain from any further split is below a threshold.

After the tree is fully grown, **pruning** is often applied. This involves removing branches that have little power to generalize (i.e., they only fit the training data too closely, a phenomenon known as **overfitting**). Pruning is essential for creating a model that performs well on new, unseen patient data.

3. Validation and Evaluation

To ensure the model is reliable, its performance must be rigorously evaluated using a separate **validation dataset**, as demonstrated by the split used in the Suo et al. study (training set $n=110,325$, validation set $n=59,367$) [1].

3.1. Key Performance Metrics

The model's predictive power is assessed using metrics similar to those reported in the BN literature:

Metric	Description	Relevance to CVD Prediction
Accuracy	The proportion of total predictions that were correct (e.g., 85.34% for CAD in Kong et al. [3]).	A general measure of model correctness.
Area Under the ROC Curve (AUC)	Measures the model's ability to distinguish between the positive class (CVD) and the negative class (No CVD). An AUC of 1.0 is perfect; 0.5 is random chance (e.g., 0.852 for CAD in Kong et al. [3]).	The gold standard for evaluating diagnostic and predictive models.
Sensitivity (Recall)	The proportion of actual CVD cases that were correctly identified.	Crucial for ensuring that high-risk patients are not missed (minimizing False Negatives).
Specificity	The proportion of actual No-CVD cases that were correctly identified.	Important for avoiding unnecessary follow-up or treatment for healthy patients (minimizing False Positives).

4. Interpretation and Clinical Application

The final decision tree provides a set of clinical rules for risk stratification.

4.1. Reading the Tree

A decision tree is read from the **Root Node** (the top) down to the **Leaf Nodes** (the bottom). The feature at the root node is the single most important predictor of the outcome.

Example Rule (Conceptual):

1. **Root Node:** Is **Age** > 60?
    - **If YES:** Go to next node. Is **Hypertension** present?
        - **If YES:** Go to next node. Is **Physical Activity** insufficient?
            - **If YES (Leaf Node):** High Risk of CVD (e.g., 90% probability).
            - **If NO (Leaf Node):** Moderate Risk of CVD (e.g., 65% probability).
    - **If NO:** Go to next node. Is **Smoker Profile** 'Smoker'?
        - ...and so on.

4.2. Comparison to Bayesian Networks

While both Decision Trees and Bayesian Networks are highly interpretable models, they serve slightly different purposes:

- **Decision Trees:** Provide a clear, sequential set of rules for **classification** (e.g., High Risk vs. Low Risk). They are excellent for simple, actionable clinical protocols.
- **Bayesian Networks:** Provide a probabilistic model of **interdependencies** between all variables. They are superior for understanding the *causal structure* of risk factors and for performing complex *probabilistic inference* (e.g., "If we observe a patient has high cholesterol, what is the updated probability of them being a smoker?").

In summary, generating a decision tree for CVD prediction requires meticulous data preparation, careful selection of a splitting algorithm, and rigorous validation, all guided by the rich feature set and data characteristics highlighted in the Bayesian Network literature.

References

[1] Suo, X., Huang, X., Zhong, L., Luo, Q., Ding, L., & Li, H. (2024). Development and Validation of a Bayesian Network ‐ Based Model for Predicting Coronary Heart Disease Risk From Electronic Health Records. Journal of the American Heart Association, 13(10), e029400.

[2] Ordovas, J. M., Rios-Insua, D., Santos-Lozano, A., Lucia, A., Torres, A., Korre, M., & Camacho, J. M. (2023). A Bayesian network model for predicting cardiovascular risk. Computer Methods and Programs in Biomedicine, 231, 107405.

[3] Kong, D., Chen, R., Chen, Y., Zhao, L., Huang, R., Luo, L., ... & Wu, K. (2024). Bayesian network analysis of factors influencing type 2 diabetes, coronary heart disease, and their comorbidities. BMC Public Health, 24(1), 1267.