Team Project Proposal

Raghda Kailany , Cameron Tinney, Fabliha Bushra and **Md Omar Faruque**

1. **Problem Statement**:

Heart disease remains one of the leading causes of morbidity and mortality worldwide. Accurate prediction and understanding of factors contributing to heart disease can lead to better preventive measures, early detection, and management strategies. Given a dataset with 14 key attributes related to patient health and heart function, we aim to employ data mining techniques to understand and predict the presence of heart disease.

The primary objective of our data mining project is twofold:

Predictive Analysis: To build a model that can predict the likelihood of a patient having heart disease based on the provided attributes. This will entail classifying patients into two categories - those with heart disease and those without.

Descriptive Analysis: To extract meaningful insights and patterns from the dataset which can help in understanding the key factors or combinations of factors that are most indicative of heart disease. This might include finding relationships or correlations between various attributes like age, cholesterol levels, type of chest pain, etc., and the presence or absence of heart disease.

1.1. **Motivation:**

Medical Importance: Early detection and prevention of heart disease can save lives. An accurate predictive model can act as an auxiliary tool for medical practitioners in diagnosing heart disease, ensuring timely and appropriate treatment.

Comprehensive Understanding: With the variety of attributes available, ranging from demographic details to specific heart performance indicators, the dataset offers an opportunity to gain a holistic understanding of heart health and the factors influencing it.

Addressing Knowledge Gaps: Previous studies have focused only on a subset of attributes. By analyzing all 14 attributes, we aim to derive a more comprehensive model and potentially uncover lesser-known relationships or factors contributing to heart disease.

In conclusion, our data mining project is not just about predicting heart disease but also aims to contribute to the broader understanding of factors influencing heart health, enabling better preventive and diagnostic strategies in the future.

2. **Data mining task:**

Our team proposes a comprehensive analytical approach to develop an accurate machine-learning model for predicting heart disease using the UCI dataset. We would like to perform the following data mining tasks in our project.

Exploratory Data Analysis: We will first thoroughly explore and understand the data through visualizations and statistical analysis to identify patterns, relationships, and potential issues.

Data Cleaning: We will then employ data cleaning techniques like handling missing values, anomaly detection, and outlier removal to prepare high-quality data.

Feature Engineering: Transforming and combining features through engineering will help us derive new perspectives from the data.

Feature Selection: Feature selection techniques will derive the most informative subset of variables to improve our model performance.

Preprocessing: We will Standardize and normalize data so that features are on the same scale and compatible with models.

Model Training and Selection: For modeling, we will train and rigorously evaluate supervised classification algorithms like decision trees, random forest, XGBoost, etc. on the engineered features. Hyperparameter tuning using cross-validation will optimize model complexity. Comparing performance across accuracy metrics on train and validation sets will help select the best model.

Evaluation: We will leverage learning curves to detect overfitting and confusion matrices to analyze prediction errors. The evaluation will be iterative, guiding the feature selection and model optimization.

Our team will try to bring together skillsets in statistics, machine learning, and data visualization. Combined with rigorous evaluation at each stage, we believe this holistic approach will lead to deep insights and an accurate model.

3. Dataset:

The UCI Heart Disease dataset, often referred to as the "Heart Disease Cleveland dataset," was collected primarily from the Cleveland Clinic Foundation in Cleveland, Ohio, USA. The dataset contains data related to heart disease diagnoses and is sourced from patients who were either diagnosed with or suspected to have heart disease.

The provided dataset is related to heart disease and contains information about various clinical, demographic, and diagnostic aspects for approximately 920 participants.

The description of the dataset, including the mentioned categories:

1. **id:**
A unique identifier for each participant.

2. **age:**
The age of the participant.

3. **sex:**
The gender of the participant (e.g., Male, Female).

4. **dataset:**
The source or dataset from which this data is derived (e.g., Cleveland).

5. **cp (chest pain type):**
The type of chest pain reported by the participant (e.g., typical angina, asymptomatic, non-anginal).

6. **trestbps (resting blood pressure):**
The participant's resting blood pressure.

7. **chol (serum cholesterol):**
The serum cholesterol levels of the participant.

8. **fbs (fasting blood sugar):**
Indicates if the fasting blood sugar of the participant is above a certain threshold (TRUE or FALSE).

9. **restecg (resting electrocardiographic results):**
Results of the resting electrocardiogram (e.g., lv hypertrophy, normal).

10. **thalch (maximum heart rate achieved during exercise):**
The maximum heart rate achieved by the participant during exercise.

11. **exang (exercise-induced angina):**
Indicates if the participant experienced exercise-induced angina (TRUE or FALSE).

12. **oldpeak (depression induced by exercise relative to rest):**
ST depression induced by exercise relative to rest.

13. **slope:**
The slope of the peak exercise ST segment (e.g., downsloping, flat).

14. **ca (number of major vessels colored by fluoroscopy):**
The number of major vessels colored by fluoroscopy.

15. **thal (thalassemia):**
Thalassemia type (e.g., normal, fixed defect, reversible defect).

16. **num (diagnosis of heart disease):**
Indicates the presence or absence of heart disease (0: No heart disease, 1-4: Presence of heart disease).

3.1 Data Collection Methodologies:

The dataset likely originated from patients who were undergoing diagnostic tests related to heart conditions. Information such as age, sex, clinical symptoms (chest pain type), blood pressure, cholesterol levels, and electrocardiogram results would typically be collected during routine clinical assessments and diagnostic tests.

3.2 Data Anonymization:

To ensure privacy and confidentiality, the dataset has been anonymized by replacing direct identifiers (e.g., names) with unique numerical IDs. Additionally, sensitive information might have been generalized or removed to prevent re-identification of individuals. This practice aligns with ethical guidelines to protect the privacy of the participants in the study.

4. Tentative Schedule:

- Week 1: Exploratory Data Analysis & Data Cleaning
- Weeks 2-3: Feature Engineering
- Week 4: Feature Selection
- Week 5: Preprocessing
- Weeks 6-8: Model Training and Selection
- Weeks 9-10: Evaluation

Total Timeline: 2 months