Predicting Heart Disease
Using Machine Learning Classification Methods

Laura Carbaugh [*1] **Simon Tao** [*] **Atty Ehui** [*]

1. Introduction Laura Carbaugh Atty Ehui

Heart disease is one of the leading causes of death worldwide, making the early and accurate prediction of cardiac risk a critical public health goal to prevent fatalities. Using a data set from the UC Irvine Machine Learning Repository, we plan to explore factors that are involved in predicting whether an individual has heart disease or not. The UCI data set combines patient records from four different sources. Cleveland, Hungary, Switzerland, and VA Long Beach. Although this database contains 76 variables, previous research focuses on a subset of 14 variables that capture the most relevant information. By building and evaluating numerous models on this data, we aim to improve our understanding of various cardiovascular risk factors and illustrate how early intervention can provide better patient outcomes. Our study offers a chance to compare different modeling strategies in a real-world health context, highlighting how careful algorithm selection contributes to meaningful insights.

2. Variable Description Simon Tao

For this project, we are using a dataset from the UC Irvine Machine Learning Repository. The data has 303 observations across 14 different variables regarding the presence or absence of heart disease in a patient. Below is a description of each of the variables that will be used.

age Patient age in years.

sex Biological sex of the patient (1 = male, 0 = female).

cp Chest pain type: 1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic.

trestbps Resting blood pressure (mm Hg).

chol Serum cholesterol (mg/dL).

fbs Indicator of fasting blood sugar > 120 mg/dL (1 = true, 0 = false).

restecg Resting electrocardiographic results: 0 = normal; 1 = ST–T wave abnormality; 2 = left ventricular hypertrophy.

thalach Maximum heart rate achieved during exercise.

exang Exercise-induced angina (1 = yes, 0 = no).

oldpeak ST depression induced by exercise relative to rest.

slope Slope of the peak exercise ST segment: 1 = upsloping; 2 = flat; 3 = downsloping.

ca Number of major coronary vessels (0–3) colored by fluoroscopy.

thal Thalassemia status: 3 = normal; 6 = fixed defect; 7 = reversible defect.

num Target variable: diagnosis of heart disease, originally coded 0–4 to reflect severity (often binarized to 0 vs. 1 for classification).

2.1. Key Variables Simon Tao

The variables include demographic factors, laboratory measurements, and clinical assessments which allows for a robust modeling of heart disease risk. While all of these variables are important for modeling, “cp,” “trestbps,” and “ca” are powerful predictors of heart disease. Other continuous variables such as “age,” “chol,” and “thal” are risk factors that can influence long-term disease development. “Sex” has also shown predictive value because men have a higher incidence of heart disease at earlier ages. Exercise-related measurements also play a key role. “Oldpeak,” which measures ST depression induced by exercise relative to rest, and “slope,” describing the shape of the ST segment during peak exercise, have strong diagnostic relevance.

3. Data Cleaning Laura Carbaugh

In order to build a reliable classification model, the data must be cleaned so that it is ready for testing and analysis. This data set does include missing values, particularly in variables such as “ca” which is the number of major vessels colored by fluoroscopy, and “thal,” which indicates thalassemia status. Missing values must either be handled through imputation methods, such as replacing them with the median for numerical features or the mode for categorical ones, or by removal. Because this dataset contains

303 observations, removal must be handled carefully as this can affect model performance if the sample becomes much smaller.

This database contains both numeric and categorical variables, which require different processing techniques. All numeric variables must be converted to integers to ensure that the model interprets these values correctly. Continuous variables such as "trestbps" (resting blood pressure) and "chol" (serum cholesterol) may need normalization to ensure that features measured on different scales do not dominate the algorithm. Categorical variables such as "cp" (chest pain type) and "restecg" (resting electrocardiographic results) must be encoded either using one-hot or ordinal encoding so that it is presented in a readable format for the model. The target variable "num" records the severity of heart disease on a scale from 0 to 4, but this variable may need to be transformed into a binary outcome by grouping all nonzero values as "disease present." After all necessary changes, the dataset must be inspected to confirm that no null values remain and that every feature is in a readable format for classification.

4. Preparation for Analysis Atty Ehui

After cleaning and refining the data, the next step is to prepare it for model building and evaluation. The processed data will be randomly divided into training and testing subsets to allow for unbiased assessment of the model's capabilities. Any transformations done to the data will be fit on the training set and then applied to the test set in order to prevent data leakage that could harm the model. These measures ensure that the final classification models are trained on consistent, well-formatted data.