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.

5. Exploratory Data Analysis Simon Tao Laura Carbaugh

An exploratory analysis was conducted to better understand the structure and relationships of variables within the dataset before model building. Summary statistics and visualizations were used to examine distributions, detect potential outliers, and identify correlations between key variables. Exploring the relationship between these variables helps us determine which model techniques will be effective.

Preliminary exploration of the merged heart disease dataset revealed several important trends. The summary statistics show that the average patient age is approximately 54 years, and most participants have resting blood pressure values around 132 mmHg and serum cholesterol near 200 mg/dL. The standard deviations suggest moderate variability across most numeric features, with cholesterol and maximum heart rate (`thalach`) showing the greatest spread.

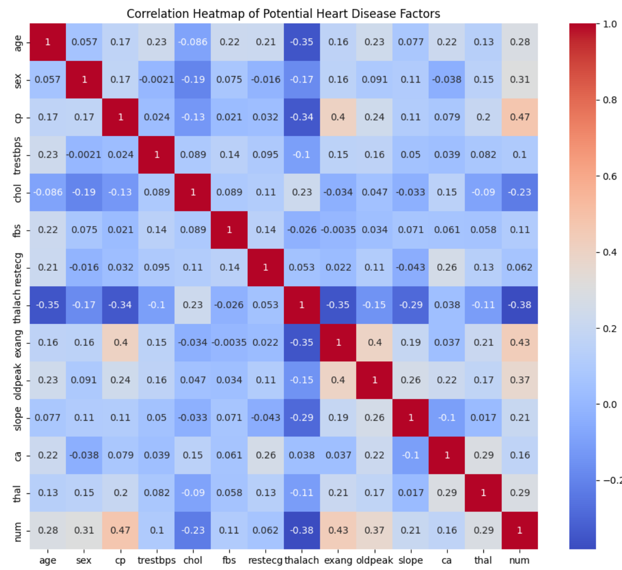

Figure 1. Correlation heatmap of continuous predictors and the target `num`.

The correlation heatmap highlights notable relationships between variables. Chest pain type (`cp`) and maximum heart rate (`thalach`) both display moderate positive correlations with the presence of heart disease, while exercise-induced angina (`exang`) and ST depression (`oldpeak`) are negatively correlated with heart health. These relationships suggest that exercise response variables and pain characteristics may serve as meaningful predictors in the classification models.

Density plots of numeric variables classified by disease status show that patients with heart disease tend to have slightly lower maximum heart rates and higher `oldpeak` values, indicating greater ST depression after exercise. These visualizations and summary tables help us identify which variables could be key factors in determining the presence of heart disease.

6. Methods Simon Tao

The goal of this project is to build a model that will be able to classify whether or not an individual in the dataset has heart disease. This process will involve building and comparing multiple different models and evaluating their performance. The prediction of heart disease presence is based on the predictors in the dataset that were listed in section 2. In order to find a strong classification model for this data, we are going to explore a variety of machine learning models, such as different types of regression, k-nearest neighbors, and decision trees. These models are

Table 1. Summary statistics for continuous variables.

Variable	Count	Mean	SD	Min	P25	P50	P75	Max
age	920	53.51	9.42	28.00	47.00	54.00	60.00	77.00
trestbps	920	132.00	18.45	0.00	120.00	130.00	140.00	200.00
chol	920	199.91	109.04	0.00	177.75	223.00	267.00	603.00
thalach	920	137.69	25.15	60.00	120.00	140.00	156.00	202.00
oldpeak	920	0.85	1.06	-2.60	0.00	0.50	1.50	6.20
ca	920	0.23	0.63	0.00	0.00	0.00	0.00	3.00

Table 2. Counts for categorical variables.

Variable	Category	Count
sex	1	726
	0	194
fbs	0	782
	1	138
exang	0	583
	1	337
num	1	509
	0	411
cp	4	496
	3	204
	2	174
	1	46
restecg	0	553
	2	188
	1	179
slope	2	654
	1	203
	3	63
thal	3	682
	7	192
	6	46

appropriate because of their interpretability and predictive accuracy. The modeling process will involve integrating and cleaning the data, exploratory data analysis, model training, and validation and performance comparison of the different models.

6.1. Models Used and Justification

Laura Carbaugh

There are many different methods that we could use to predict the presence of heart disease based on clinical and demographic factors. Logistic regression is a potential option due to its simplicity and interpretability. This method is appropriate because it estimates the probability that an observation belongs to the "disease" or "no disease" class, making it ideal for binary classification problems. The coefficients from the model are easily understandable, which is important in a healthcare context where outcomes must be transparent and explainable. Additionally, logistic regression allows for the identification of statistically significant predictors, making it useful for understanding which variables have the strongest association with heart disease risk. Another option that we would like to explore is k-nearest neighbors (kNN), in which observations are classified based on their nearest neighbors in the feature space. This specific method does not assume linearity, which makes it flexible for finding complex or irregular patterns within the data. KNN is also advantageous because it does not require a predefined model structure or assumptions about the data distribution. However, it can be sensitive to the choice of k and to differences in variable scales, so appropriate feature scaling and model tuning will be important steps during implementation.

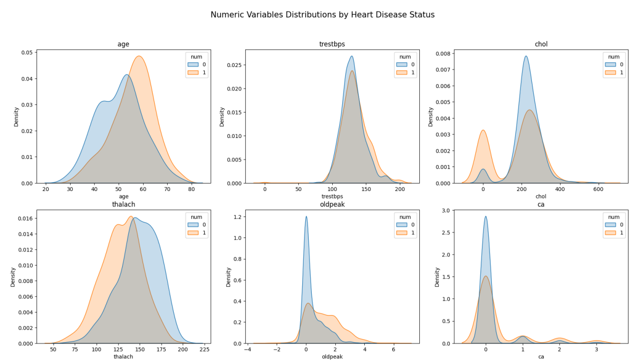

Figure 2. Distributions of key numeric variables classified by heart disease status.

The third technique we plan to explore is decision trees, which will allow us to examine non-linear relationships between the predictors and target variables. Decision trees split the data based on certain attributes, forming a series of decision rules that can be visualized in a hierarchical manner. This approach is particularly effective for identifying key variables that contribute the most to predicting heart disease, and it can handle both numeric and categorical data without requiring feature scaling. Decision trees also provide a level of interpretability that is valuable in medical analysis, as the decision paths can be clearly traced and explained.

By exploring these three models, we aim to evaluate the trade-offs between interpretability, efficiency, and accuracy. Comparing their results will allow us to determine which method performs best on our dataset while still offering meaningful insights into the underlying factors that contribute to heart disease.

6.2. Model Training Procedure Atty Ehui

Before beginning model training, the four datasets from Cleveland, Hungary, Switzerland, and VA Long Beach will be combined into a single dataset and cleaned for analysis. The combined dataset will then be divided into a training set and a testing set. We will use an 80/20 split so that the majority of the data is used to train the models while a smaller portion is used for evaluating performance on unseen data. This separation helps assess how well the models generalize beyond the data they were trained on. Additionally, to prevent data leakage, any transformations applied to the training data will be applied to the testing data as well.

During model training each of the described methods will be fitted on the training data. For logistic regression, regularization may be considered to prevent overfitting, which is useful given the potential multicollinearity among similar variables. For k-nearest neighbors, the optimal value of k number of neighbors will be determined using cross-validation to balance bias and variance. Decision trees will be trained with various splitting criteria to identify in order to determine key patterns without becoming too complex.

6.3. Model Validation Plan Laura Carbaugh

When assessing to see if the models are valid, we plan to look at the accuracy score, confusion matrix, and the AUC value. The accuracy score will assess the overall success of the classification models. The confusion matrix, which includes the sensitivity and specificity values, will assess how good the models are at predicting the different classes in the data (whether or not an individual has heart disease). The Area Under the Curve (AUC) will be used evaluate the trade-off between sensitivity and specificity across various classification thresholds. Together, these metrics will provide a well-rounded understanding of model performance and reliability.

6.4. Model Implementation Atty Ehui

After cleaning and merging the data, the models will be implemented in Python using libraries such as scikit-learn, NumPy, and pandas. Each model will follow the same preprocessing steps to ensure fair comparison. For logistic regression, we will test different predictors and regularization strengths to prevent overfitting and identify the most influential variables. For kNN, we will experiment with different k values to ensure that all features are properly scaled. For decision trees, model depth and splitting criteria will be tuned to balance accuracy and interpretability while avoiding overfitting. All models will be evaluated using cross-validation and the same metrics to compare performance.