# At the Heart of the Matter

*Using modeling and machine learning techniques to classify heart disease in patients*

Annika Cleven, Yajie He, Tyler Humpherys, Dinelka Nanayakkra, Matthew Sutcliffe

## Introduction and Objective:

Heart disease is a leading cause of death in the United States, accounting for over 700,000 deaths each year and posing a significant healthcare burden.[1] Accurate detection of heart disease is essential for improving patient outcomes, with implications for treatment plans, medication strategies, and lifestyle modifications. While traditional diagnostic methods, such as stress testing and electrocardiography, provide valuable insights, integrating this data can enhance heart disease classification accuracy. The objective of our project is to develop a data-driven approach to predict the presence or absence of heart disease.

## Dataset:

We will analyze a publicly available dataset collected by the Cleveland Clinic between 1981 and 1984. This dataset includes 303 patients with or without heart disease, as well as a range of demographic (e.g. age, sex, smoking status), clinical (e.g. resting blood pressure, serum cholesterol), and diagnostic (e.g. chest pain location and response to exercise) features on which to train a predictive model. Specifically, the heart disease outcome of interest is a categorical value with 0 being no heart disease and 1 to 4 being increasing heart disease severity. For our analysis, we will simplify this to a binary outcome of 0 (no heart disease) and 1 (heart disease, any severity).

Dataset repository: https://doi.org/10.24432/C52P4X

Introductory paper: https://doi.org/10.1016/0002-9149%2889%2990524-9

## Aims:

- Implement the Newton-Raphson optimization algorithm from scratch to estimate parameters for a logistic regression model of heart disease
- Train a random forest model to improve heart disease classification performance and to examine feature importance

[1] Martin, SS., … American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee (2024). 2024 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. Circulation, 149(8), e347–e913. https://doi.org/10.1161/CIR.0000000000001209

- Develop a robust R package that implements the prior aims and provide suitable documentation and unit tests

## Models:

The goal of this project is to classify patients into two categories–presence and absence of heart disease–using the other covariates. To achieve this we will begin by applying a logistic model. This model allows us to delineate the relationship between the outcome variable (presence of heart disease) and the predictor covariates, while also providing the log odds of heart disease presence given a set of covariates. We selected the logistic model for our analysis plan because of its interpretability and as a strong baseline to compare against more complex machine learning techniques.

To fit the logistic model, we will use the Newton-Raphson approach, coded from scratch. Depending on the time and bandwidth of this team, we may also investigate fitting the model with Broyden-Fletcher-Goldfarb-Shanno (BFGS) and Stochastic Gradient Descent (SGD) methods. Each of these methods pose their own strengths and challenges. The Newton Raphson method requires us to find a second derivative, which we anticipate to be feasible. If we encounter challenges with the second derivative, we can pivot to use BFGS, which only requires a first derivative.

The second model we will use is random forest. We chose random forest for this project due to its ability to handle complex, high-dimensional data with many predictor covariates. One key advantage of this model is its robustness to noise and outliers, leading to more stable and accurate results. Additionally, random forest provides insights into variable importance, allowing us to identify the most influential predictors of heart disease. This method also does not require assumptions about the underlying distribution of the data, making it a flexible and reliable approach for this classification task.

After fitting the random forest model, if time and resources allow, we would like to extend our investigation to compare our random forest model performance to other machine learning methods. Support Vector Machine (SVM), a popular choice in machine learning literature for classification problems, is of particular interest for application to our heart disease classification problem.

## Procedure to fit the model:

### Logistic Regression

As mentioned in the prior section, we aim to use logistic regression (coded from scratch) to predict the presence of heart disease. The logistic model can be defined as follows:

$$y_i \mid x_i \sim Binomial(1\,,\, p_i)$$

$$logit(p_i) = x_i^T \beta$$

where,

$y_i$- binary outcome variable indicating presence/absence of heart disease of individual **i**

$p_i$- probability of the presence of heart disease in individual **i**

The log-likelihood of our logistic regression model that we aim to maximize can be written down as follows:

$$l_n(\beta) = \sum_{i=1}^{n} \{y_i x_i \beta - log[1 + exp(x_i^T \beta)]\}$$

We can take the derivative of this log-likelihood, set it equal to zero, and try to find an optimal solution analytically. However, we choose not to, given that it does not have a closed form solution. We keep the log-likelihood as it is, and apply numerical optimization techniques learned in module 2 to solve this problem.

The first technique we foresee implementation from scratch is the Newton Raphson algorithm.

**Newton-Raphson optimization**

let,

$l_n^{'}(\beta^{(t)})$: first derivative of the log-likelihood at the t$^{th}$ iteration.

$l_n^{''}(\beta^{(t)})$: second derivative of the log-likelihood at the t$^{th}$ iteration (hessian matrix).

$h^{(t)} = \; - l_n^{''}(\beta^{(t)})^{-1} l_n^{'}(\beta^{(t)})$

Now, the beta values at the (t+1)$^{th}$ iteration can be estimated as:

$$\beta^{(t+1)} = \beta^{(t)} + h^{(t)}$$

For the logistic framework we are working with, we can moreover see that in matrix form:

$$h^{(t)} = (X^{'} W^{(t)} X)^{-1} X^{'} (Y - p^{(t)})$$

where,

$p = (p_1, p_2, \dots , p_n)$

If time allows, we look to use the BFGS algorithm directly from the ‘optimx’ package from R to see if we can achieve better computational efficiency compared to Newton-Raphson.

### Random Forest

To fit the random forest model in R, we will use the Rtemis library’s “s_Ranger” method. By specifying “probability = True”, we can classify having heart disease for individuals whose probability is p>= 0.5. If there is a heavy class imbalance we may use the class weighting techniques accompanying this package. A grid search will be employed for tuning hyperparameters like number of trees, node size, and so on. We will evaluate this model using train-test split and then referencing the confusion matrix to get the testing accuracy, precision, recall, and specificity of our fitted random forest model.

As mentioned earlier, if time and resources allow, we may also compare performance with SVM. In this case we will fit SVM with the “e1071” R package. If this heart data is not linearly separable, then an appropriate kernel will be identified and employed to transform the data into a higher-dimensional space where it is linearly separable and allow for more complex decision boundaries.

Rtemis: https://github.com/egenn/rtemis

e1071: https://cran.r-project.org/web/packages/e1071/index.html

### R Package

During the course of building our R package we will implement thorough documentation for our methods using roxygen, write appropriate package vignettes, and include package tests to ensure the proper use of and functioning of our package. Included in this package will be capabilities to fit the logistic regression with Newton Raphson, fit the random forest model, evaluate them, and compare them with appropriate plots.

This heart disease dataset does not have memory-intensive demands (~60 KB) and so advanced data loading and manipulation strategies (e.g. datatable, sqlite, hdf5, sparse matrices, etc) will not be as important for this project.