The dataset used in the R Shiny app is called The Framingham Heart Study Dataset taken from BioLINCC website. The data is continuously collected every two years through sampled participants from Massachusetts, USA. The objective is to study the incidence and prevalence of coronary heart disease (CHD) and its risk factors over time – whether a patient develops CHD in ten years and what risk factors can be monitored for medical intervention. The initial data included 4240 observations. However, to fit the project requirement, stratified sampling was performed of 500 observations based on 15.23% split of the dependent variable 'TenYearCHD.' Additionally, categorical variables were converted into factors from integers, so that they can be used for statistical modelling correctly with fixed and known values.

The purpose of the Shiny app is to create a learning-based platform that allows the user to explore how machine learning works through the two classification models. The app's structure consists of a landing/home page and four other tabs. The landing/home page gives user a background of the dataset used for the prediction, provides a lexicon to simplify user's understanding of the dataset and briefly informs the user on the workings of the two models - setting the user for an enriching experience on the app. Following that, is a tab for predicting whether the user is likely to have CHD in 10 years based on the decision tree model and the inputs on the independent variables given by the user. The information inputted by the user is read as a row of the dataset, the model is applied, and the prediction is presented along with suggested references for additional information. The next tab visualizes the decision tree and briefly explains the process of the model, as well as provides an interactive feature of understanding the impact of the tuning parameter i.e. tuneLength - to understand how the tree changes when different numbers of complexity parameters are tested. Similarly for the random forest model, the app visualizes the Cross-Validation approach and how the changes in 'mtry' parameter reflect on the accuracy of the model whilst explaining why visualizing random forest is not possible - adding yet another learning experience for the user. Lastly there is a dashboard tab that compares the two models along the accuracy metric for test and training sets of both the models, percentage of test set observations predicted to have CHD is shown for both the models, and variable importance which highlights the most influential risk factors in classifying a patient as Yes or No.

To build and test the accuracy of both the models, the dataset was separated into training and testing data with a 60-40 split. Though the models' predictions are similar, random forest does tend to perform better with a higher accuracy. However, the exact measures change every time despite setting seeds perhaps due to the dynamic sampling occurring in random forests. Interestingly, while running both the models several times, the predicted classifications were not consistent with the actual classifications. For example, the actual number of patients who are diagnosed with CHD in the test set is roughly 15%, however, both the models predicted much fewer patients to be diagnosed with CHD, implying importance on the True Negative Rate.