Heart Disease Prediction

Project Report

Prof. Shujing Sun

BUAN 6356:502 Business Analytics with R

Group 15

Group Members

Akash Gupta (AXG200040)

Priyal Gupta (PXG200016)

Rahul Gaikwad (RAG200000)

Sanjana Kale (SXK200056)

Shamik Pendse (SAP200015)

Siddhant Amidwar (SXA00004)

Setting

Business Context

Taking decisions and making predictions using data science is rapidly gaining traction in the healthcare industry. Data Science has been shown to be useful, accurate, and cost-effective across a variety of fields in articles and research publications. Three machine learning models were used on datasets collected from the UCI machine learning repository for this study. Cleveland, Hungarian Institute Of Cardiology, and Switzerland Heart Disease datasets were chosen since they were all connected to predicting the presence or absence of heart disorders in patients. On these datasets, logistic regression, decision trees, and Random forest were used as the models. Variable parameters for several models were used to generate the results. The project's purpose was to check model consistency on datasets and determine the most appropriate algorithm for that dataset.

Problem

Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict possible heart disease. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia, or already established disease) need early detection and management wherein a machine learning model can be of great help.

The following are the project's key goals:

- Using area codes, combine datasets from multiple locations into a single dataset.
- Data pre-processing to replace null and out-of-bounds values.
- Data training with the chosen machine learning models.
- Choosing the most accurate model and putting it into practice for this dataset.

Data Description

There are 3 available databases concerning heart disease diagnosis. All attributes are numeric valued. The data was collected from the 3 following locations:

• Cleveland Clinic Foundation

• Hungarian Institute Of Cardiology

• University Hospital, Switzerland

A brief description of the dataset used: The dataset has 14 attributes (categorical, Integer and Real) and 720 instances. The 'num' attribute is the response variable, used for the prediction.

(Link: https://archive.ics.uci.edu/ml/datasets/heart+Disease)

There are a total of 14 attributes in the dataset we have used.

Problems Faced

1) Cleaning the data was one of the challenges faced for our project.
2) Selecting which algorithms gives the highest accuracy for this project.

Field	Description	Range and Values
Age	Age of the patient	0-100 in years
Sex	Gender of the patient	0-1 (1:Male 0:Female)
Chest Pain	Type of chest pain	1-4 (1: Typical Angina, 2: Atypical Angina, 3: Non-angina, 4: Asymptotic)
Resting Blood Pressure	Blood pressure during rest	mm Hg
Cholesterol	Serum Cholesterol	mg / dl
Fasting Blood Sugar	Blood sugar content before food intake if >120 mg/dl	0-1 (0: False, 1: True)
ECG	Resting Electrocardiographic results	0-1 (0: Normal, 1: Having ST-T wave)
Max Heart Rate	Maximum heartbeat rate.	Beats/min
Exercise Induced Angina	Has pain been induced by exercise	0-1 (0: No, 1: Yes)
Old Peak	ST depression induced by exercise relative to rest	0-4
Slope of Peak Exercise	Slope of the peak exercise ST segment	1-3 (1: Up sloping, 2: Flat, 3: Down sloping)
Ca	Number of vessels colored by fluoroscopy	0-3
Thal		3- normal 6- Fixed Defect 7- Reversible Defect
Num	Diagnostics of Heart Disease	0-1 (0: <50% Narrowing 1: >50% Narrowing)

Data Preprocessing:

The preprocessing of the data was done by cleaning the data by removing the null values and replacing them with NA in order to maintain consistency in the data and retain the statistical data. We have further replaced NA values with Mean and Mode to maintain the constant mean and mode of the dataset. The conversion of character values to integer values is one of the steps we followed in the data processing.

Algorithms Used:

1. Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents an attribute "test," each branch reflects the test's result, and each leaf node represents the class label (decision taken after computing all attributes). A decision tree and its closely related influence diagram are used as a visual and analytical decision support tool in decision analysis to calculate the expected values (or expected utility) of competing alternatives.

2. Binomial Logistic Regression

When the dependent variable is dichotomous, logistic regression is the best regression strategy to use (binary). The logistic regression, like all regression studies, is a predictive analysis. The logistic function, also known as the sigmoid function, was created by statisticians to characterize the properties of population increase in ecology, such as how it rises swiftly and eventually reaches the environment's carrying capacity. It's an S-shaped curve that can map every real-valued number to a value between 0 and 1, but never exactly.

1 / (1 + e-value), where e is the natural logarithms' base and value is the actual numerical value to be transformed.

3. Random Forest

Being a flexible to use model, Random Forest is used for both classifications as well as regression tasks. In this algorithm, each data point is assessed based on whether it passes a logical test based on one or more variables. The test's outcome defines which branch of the tree the data point will descend, and hence which logical test it will face next. Random forest is a bootstrapped decision tree in which a huge number of decision trees are generated, and the mode result is taken, generally yielding a more accurate result.

Analysis:

On comparing the resultant accuracies of all the algorithms, the observation was as follows:

- Decision Tree: 81.82% (Without Pruning)
  82.72% (With Pruning)

- Logistic Regression: 82.72%
  83.18% (with significant features)

- Random Forest: 83.18%

Based on the above results, we can conclude that the best results were given by Logistic regression and Random Forest as compared to the decision tree algorithm for this dataset.

The most useful variables which proved to be efficient in the analysis were:

- In the decision tree the significant variable in deciding the Heart Disease in this data set is CP, Chol, Oldpeak, CA after pruning.
- In Logistic regression, the important variables which increase or decrease the odds of having heart disease in this dataset are: CP, Chol, Thalach, Exang, Oldpeak,Slope, CA, and Thal.
- In Random forest, the high importance variables in deciding the presence of heart disease for this dataset is CP, Chol, Thalach, Oldpeak, Exang

References:

[1] http://csjournals.com/IJCSC/PDF7-1/18.%20Tejpal.pdf
[2] https://archive.ics.uci.edu/ml/datasets/heart+Disease