

Analyzing the trends in Heart Disease using historical data

Team Technocratz

Abhishek Shakwala
avs2368@g.rit.edu

Rinkesh Shah
rps2227@g.rit.edu

Shristika Yadav
sy2109@g.rit.edu

1. Overview

It might have happened so many times that a group of doctors needs to examine methodically and in detail, the trends seen in the heart disease, typically for explanation and interpretation based on some historical records of the patients. Sometimes it is challenging for the doctors to gather, manage, and discover or reveal some facts based on these historical data. Along with these discoveries, there are users (possible patient) who might just want to know whether they have any heart disease. The Heart Disease analysis application that we plan to build will help both these users to get the significant results.

We propose to use Heart disease data sets, available to us via UCI Machine Learning Dataset Repository. These are the data sets that contains information regarding patient's personal history which includes chest pain location, chest pain type, whether the patient smoke and various other attributes. The application will feed various personal and heart diseases related details and will help the users by sharing the heart-related issues. We will use this data in conjunction with some data mining techniques to guess the accurate heart-related illness that the patients must experience.

Our first phase will involve building the relational database and parsing CSV data. This will include building a robust system of querying information and building relational database models.

Our second phase will involve cleaning datasets, developing the algorithm and dummy models for the data sets and correlate it with past results present in the training data. This will help us in getting the significant results requested by the user.

Our third phase will contribute towards consolidating all the information we have gathered in the form of observations and visualizations that will interest the user.

2. Current Work And Challenges

2.1. Data Acquisition

2.1.1 Gathering Data

In the first phase of the project, our major focus has been on gathering the data. We have gathered most of the heart disease data sets from UCI Machine Learning Dataset Repository. The datasets that were available to us were in .data format which we extracted into .csv format using Python script. It was necessary to convert these files to a format so that it becomes feasible to parse these data for cleaning and to create a relational database.

We have used Java Database Connectivity(JDBC), which is a standard API for establishing the connection to the relational database. Initially, we created a Java project, then we imported the packages containing JDBC classes required for database programming, registered the JDBC driver and open a connection which represents the physical connection to the database. Once these steps were performed, we executed a query using an object of type statement for building and submitting the SQL query to the database. Finally, we used the appropriate result set method to retrieve the data from the result set and explicitly close all the database resources once the task of accessing database is completed.

Various challenges have been encountered while transforming these data files into relational databases. The fields in the data files available are in comma separated form and not in the new line separated form. This problem was eventually solved by parsing the .data extension file using python script and extracting the data and writing it to .csv extension file. This extracted .csv file was then given as the input to the load CSV file SQL query which stored the data inside the CSV file into appropriate attributes in the relation database tables.

2.1.2 Understanding the scope of data at hand

Initially, all the data that was present in CSV file was loaded into a single relational table (catch all in one table schema). This helped us to understand the scope of the data and get an intuition regarding the feature selection and what discoveries can be made from the available datasets. Based on this understanding regarding the scope of data we have revised the list of attributes that will be used to perform analysis that help us in getting insight and how these insights help us in better decision-making. Once the scope of available datasets was aligned with the analysis to be performed we transformed the single relational table into multiple entities.

2.2. Data Cleaning

The raw data which we have gathered has some corrupted observations as well. One such corrupted observation that we found in our data is, some rows contained “?” or some negative value instead of the value allowed by the domain of that attribute. So, in this phase we have just started with the data cleaning process by replacing these “?” values with null values but in future we will be more concrete about such corrupted data by carefully studying various methods which can help us get clean data (e.g. One way which we found is by reviewing all the values of the attribute which holds such corrupted data and analysing the distribution of the data in that column).

2.3. Exploratory Data analysis

Exploratory Data Analysis (EDA) is an approach to analyse datasets that outlines the main characteristics of data sets and is often done with visual methods. We have performed the heart disease data analysis using multivariate type of data analysis. With the help of Rattle and R we defined the scope of our attributes to be used for performing heart disease analysis by using various parameters like kurtosis and skewness. We got to know the noise attributes which were not so effective while designing our model. We performed correlation analysis on our data using the attributes defined in our scope which helped us to filter the attributes to get the appropriate prediction result. The result of correlation analysis performed on the attributes defined for performing analysis is shown in the Figure 1.

In the initial phases we had performed analysis based on the heart diagnosis test conducted by user such as: Cp, Chol, Trestbps, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca and Thal, and based on these parameters we built the decision tree shown in figure 3. We were successful in predicting whether the user has heart disease or not based on these parameters. Finally, we delve deeper into analyzing the precautionary measures to be taken to prevent the

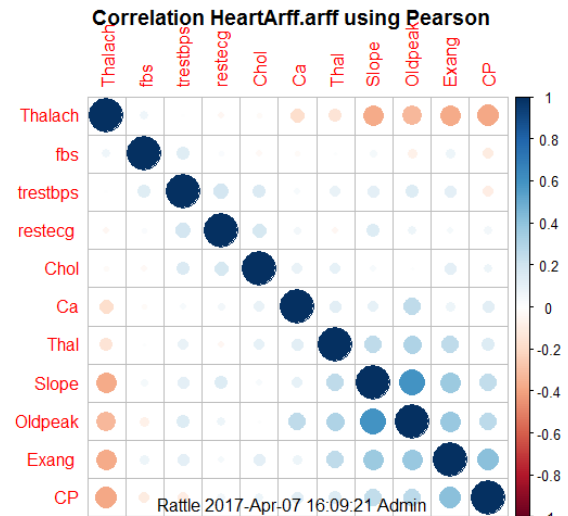


Figure 1. Correlation analysis

heart disease. To conduct this analysis we have used person history data set which includes cigarettes per day, years as smoker, duration of exercise and month of exercise. We have combined the 11 parameters used previously for predicting whether the person suffers from heart disease or not along with the person history data parameters and based on this we built the decision tree shown in figure 2. As shown in figure 2 the decision tree is quite identical to the one shown in figure 3 but in figure 2 most of the leaf nodes of the decision tree are the person history data parameters which are contributing towards predicting the precautionary measures. Thus, we are successful in predicting whether the person has heart disease or not and suggesting the precautionary measures that should be taken by the user if they suffer from heart disease.

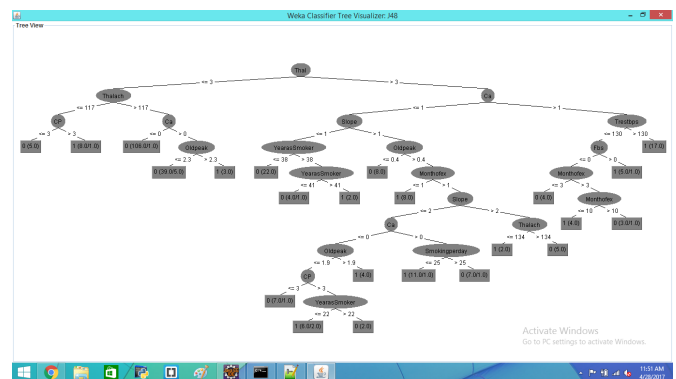


Figure 2. Precautionary measures decision tree

2.4. Modelling

In this phase we have performed various forms of analysis that eventually helped us in building the descriptive and

predictive model. We have performed the following modelling steps using Rattle and R initially:

- Building machine learning models
- Validating Models
- Testing Models
- Model Evaluation, Interpretation, and Comparisons

Once we created our application we have used weka to build our models which we will discuss further under section 2.5.

Using rattle we tested various models such as SVM, Tree, Forest and Linear, with different parameters which helped us to select the most effective model for our data set. The models on which we have performed analysis using our datasets are as follows:

- J48 (Decision Tree)
- Random Forest
- Bagging

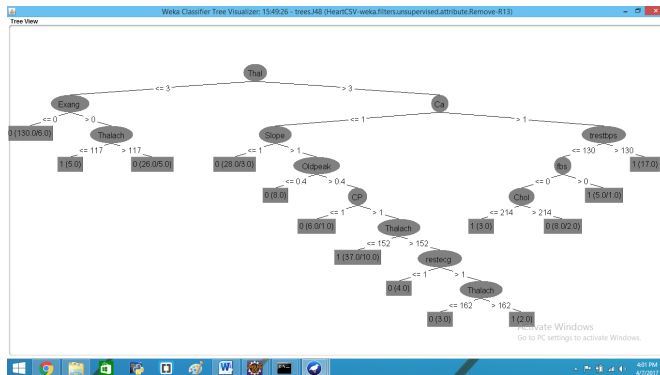


Figure 3. Decision Tree

The confusion matrix for each of the models are as follows:

- J48 (Decision Tree):

Correctly Classified Instances 73 85.8824%
 === Confusion Matrix ===

a b ← classified as
 61 2 | a = 0
 10 12 | b = 1

- Random forest:

Correctly Classified Instances 69 81.1765%
 === Confusion Matrix ===

a b ← classified as
 61 2 | a = 0

14 8 | b = 1

- Bagging:

Correctly Classified Instances 78 91.7647%
 === Confusion Matrix ===

a b ← classified as
 62 1 | a = 0
 6 16 | b = 1

The percentage specified above in each of the models indicate the accuracy of the model on test data. Using the above three models we initially provided the input training data to all the three models to build the model and thereafter we performed the mean of all the three models output to get the result of prediction of the sample input.

2.5. Application Architecture and Implementation

We have developed the application using J2EE for handling the business logic of our application and we have used MySQL database for the backend support for our application. The relational database created from the csv file which are under section 4. represent the following:

The Personal details table consists of various information regarding the person who wishes to analyse his personal history and draw some conclusion to make a healthier and risk free life from heart disease. The Personal history table consists of various information regarding the persons personal life history such as family history of heart disease, cigarettes smoked per day, years of smoking and various information regarding the exercise carried out by them. The Person test table consists of information for the medical test or examination performed on the person to analyse the risk of heart disease.

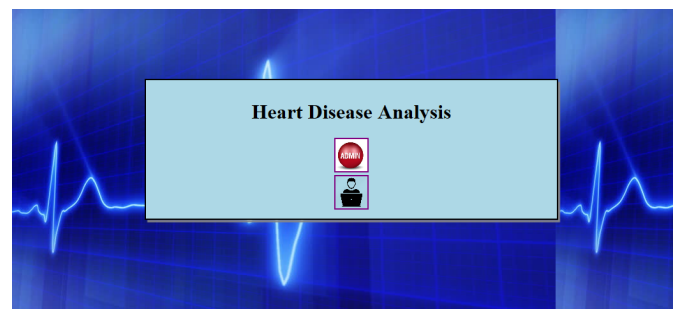
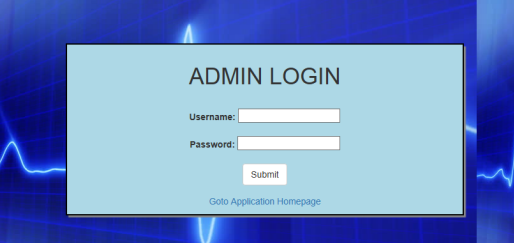


Figure 4. HomeScreen

Initially various statistical analysis was being done using R and Rattle on these tables to get useful insights which would help in driving the decision of the prediction model.



ADMIN LOGIN

Username:

Password:

[Go to Application Homepage](#)

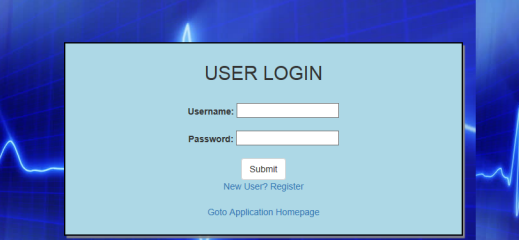
[illegible]

Add Smoking History Data

UPLOAD THE FILE

Choose the file To Upload: Browse...

Keep Enjoying!! You don't have Heart Disease.



The screenshot shows a web application interface for user login. The background is a solid blue color. In the center, there is a white rectangular box with a thin black border. Inside this box, the text "USER LOGIN" is displayed in a large, bold, black font. Below this title, there are two input fields: "Username:" and "Password:", each followed by a white rectangular input box. Below the password field is a "Submit" button, which is a white rectangle with a black border and the word "Submit" in black text. At the bottom of the white box, there are two links: "New User? Register" and "Goto Application Homepage", both in a smaller black font. The entire page is framed by a dark blue border with a subtle grid pattern and a glowing blue ECG line running vertically on both sides.

USER LOGIN

Username:

Password:

[New User? Register](#)

[Goto Application Homepage](#)

New User Register

Enter Username:

Enter Password:

Enter Age:

Gender:

☐ Male

☐ Female

Welcome User

Previous Heart Status:

You had Heart Disease on your previous visit.

Test Details:

Chest Pain Type:

typical angina

Resting Blood Pressure:

Serum Cholesterol:

Fasting Blood Sugar:

FBS > 120 mg/dl

But as we started constructing our application we replaced R and Rattle with Weka for modelling and for performing various analysis on the input data. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access. As our application was being constructed in Java it became feasible for us to use Weka to

build models and perform analysis over using Rattle and R.

3. Future Scope

Our application is currently predicting two things: first, whether the person has heart disease or not and second, pre-

cautionary measures to be taken if the person suffers from heart disease. Though we are successful in predicting both the things, still there is a good scope for predicting the precautionary measures more effectively. In future to enhance our application development we would like to be more specific about the preventive measures, for example, we would like to see prediction such as "If you increase the duration of exercise time by 50 minutes than you can prevent your heart disease in next four months". By giving such intuitive results, our application can be more useful and efficient for the users who are diagnosed with heart conditions.

4. Tables and Figures

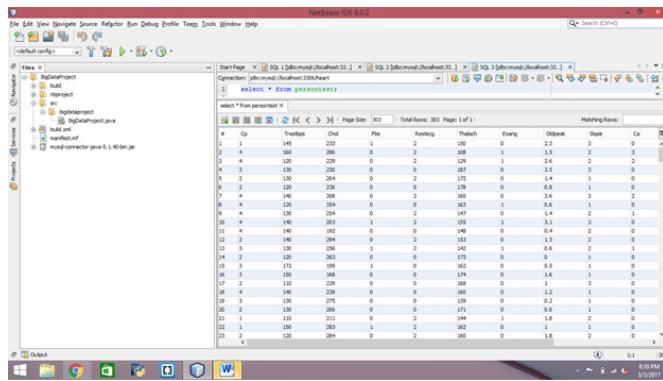


Figure 13. personstest

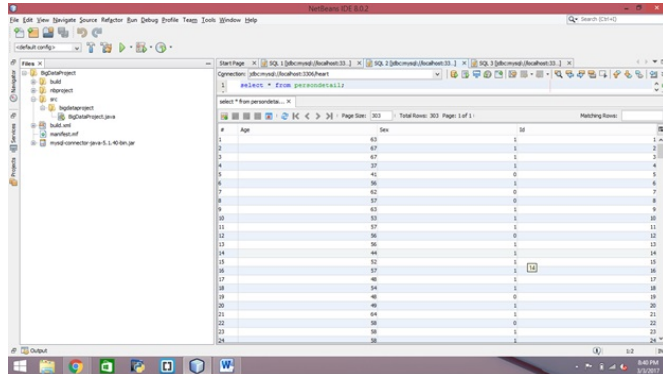


Figure 14. persondetails

Attribute	Type	Null	Default
Id	int	No	
Username	varchar	No	
Age	int	Yes	Null
Sex	int	Yes	Null

Table 1. Person Details

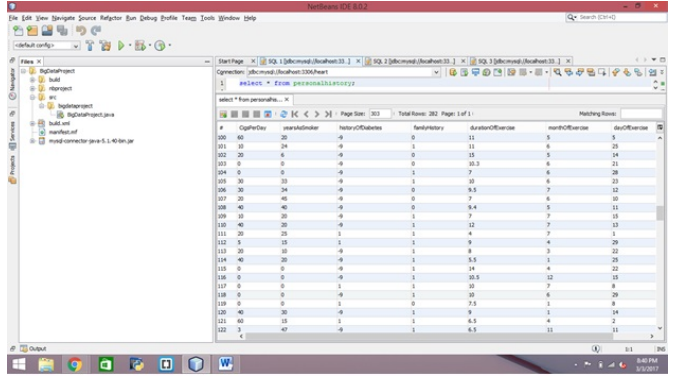


Figure 15. personhistory

Attribute	Type	Null	Default
Id	int	No	
Cp	varchar(10)	Yes	Null
Trestbps	varchar(10)	Yes	Null
Chol	varchar(10)	Yes	Null
Fbs	varchar(10)	Yes	Null
Restecg	varchar(10)	Yes	Null
Thalach	varchar(10)	Yes	Null
Exang	varchar(10)	Yes	Null
Oldpeak	varchar(10)	Yes	Null
Slope	varchar(10)	Yes	Null
Ca	varchar(10)	Yes	Null
Thal	varchar(10)	Yes	Null
Num(Output)	varchar(10)	Yes	Null

Table 2. Person Test

Attribute	Type	Null	Default
Id	int	No	
CigsPerDay	varchar(10)	Yes	Null
YearAsSmoker	varchar(10)	Yes	Null
HistoryOfDiabetes	varchar(10)	Yes	Null
FamilyHistory	varchar(10)	Yes	Null
DurationOfExercise	varchar(10)	Yes	Null
MonthOfExercise	varchar(10)	Yes	Null
DayOfExercise	varchar(10)	Yes	Null

Table 3. Person History

5. Final Remarks

After thoroughly studying various approach for data cleaning, modeling and analyzing using Rattle and R we have successfully created an application which predicts whether the person suffers from heart disease or not and if the person suffers from heart disease than what are the preventive measures taken against it. Throughout the application development, we faced some challenges in gathering, cleaning, modeling and analyzing the data

sets. It was hard to gather different datasets with similar parameters which can be used to perform the analysis. The data sets which we gathered before developing the application was not holding clean records, and for that, we used different techniques to get clean data for analysis.

It was absorbing to observe that the data set which we gathered for analysis was highly skewed because of which our prediction for preventive measures was not useful as we expected it would have been. However, it was a great learning in understanding such nature of the data and being more careful about data gathering and cleaning in future to built effective data models and decision trees.