Application Of Various Data Mining Techniques To Classify Heart Diseases

MSc Research Project
Data Analytics

Saswata Ghosh
x16104170

School of Computing
National College of Ireland

Supervisor: Mr.Jorge Basilio

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Saswata Ghosh
Student ID:	x16104170
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Mr.Jorge Basilio
Submission Due Date:	11/12/2017
Project Title:	Application Of Various Data Mining Techniques To Classify Heart Diseases
Word Count:	4413

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

1 Introduction 3

2 Related Work 5

2.1 Heart Disease 5

2.2 Literature review 5

3 Methodology 7

3.1 Project Development 7

3.2 Data Mining Algorithm To Detect Heart Disease For My Research . . . 9

4 Implementation 10

4.1 Data Preparation and Feature engineering 10

4.2 Design of the Model 12

4.3 Implementing the machine learning algorithm 12

5 Evaluation 13

5.1 Exploratory data analysis result 13

5.2 Evaluation of the machine learning Algorithm 14

5.3 Graphical Representation 16

6 Conclusion and Future Work 16

7 Acknowledgement 17

Application Of Various Data Mining Techniques To Classify Heart Diseases

Saswata Ghosh
x16104170
MSc Research Project in Data Analytics

10th December 2017

Abstract

Diseases are quite common in human beings but the most common disease that affects human beings are heart diseases which are the main cause of death in developed and developing countries.Detecting heart disease is a great challenge for medical practitioners with high level of accuracy so that a patients life is saved.Large amount of medical data are being generated which can be useful for the physicians,various data mining algorithm helps us to find meaningful insights from data which would otherwise remain undiscovered. My research aims to deploy a model that gives us the highest performance metric(accuracy,sensitivity,specificity and kappa) among the three algorithm Random Forest,Logistic regression and Artificial Neural Network so that a patient life is saved in places where there is lack of medical practitioners .I would be following the CRISPDM data mining methodology .Supervised classification algorithm is used to classify the heart disease in the UCI-ML data set which consists of 14 attribute and 303 observations.The effect of detecting heart disease is to render in the different data mining algorithm in terms of accuracy,sensitivity, specificity and kappa with the sole motive of obtaining more superior results than any other research method.The accuracy,sensitivity,specificity and kappa of ANN is 79%,85%,73% and 59%,whereas that of Random forest is 80%,83%,78% and 61% percent and Logistic regression is 85%,89%,80% and 70% percent,which clearly shows that Logistic regression outperforms the other two algorithm in terms of all the performance measures.Thus this research will enable the medical industry as a whole in timely treatment of heart disease without the intervention of medical practitioners and also this will help the general practitioners to decide the next step for treatment without the intervention of trained cardiologists in remote area.

Keywords-Heart Disease,ANN,Random Forest,Logistic Regression,Classification

1 Introduction

The goal of this research is to classify whether there is heart disease or not in patients by looking at the patients medical history in the data set and by using machine learning algorithm to aid in early treatment in places where cardiologists are not easily available.The risk of having heart disease is invariably present in the elderly persons but are quite prevalent in young and middle aged persons due to their fast life style.There are drugs that can cure heart disease but it needs to be detected at an early stage.A heart disease can be classified into present or absent.

The research examines the machine learning approach that can be applied on the heart disease patient which detect the inception of heart disease by exploring the UCI dataset that was not much being explored.The challenge is to identify parameters that separate heart disease form other kind of disease. Detecting heart disease at an early stages is important so that a patient live is saved. Machine learning algorithm helps cardiologists in the time consuming diagnosis of heart disease. In developing countries where there is shortage of qualified cardiologists or over-worked doctors this kind of machine learning algorithm will go a long way in helping all the stakeholders.This technology will not only help the doctors to make near perfect diagnoses but will also help in cost-reduction so that a large portion of the population which are poor in the developing world can take advantage of it.This research has huge potential and can be considered as a yardstick before performing real life experiments which in turn will save lot of time and resource.

Heart disease in developed countries are more than the developing countries,17million people are affected due to heart attack and strokes.In rural India people at the age of 25-69 years people suffers from heart problem(Son et al.; 2012; Mahajan et al.; 2017).According to WHO figures it shows that heart diseases are the sole reason of human morbidity in developed and developing countries(Desai et al.; 2017).The most common and complex health problems around the globe is heart failure which usually means that the heart fails to pump sufficient blood to meets the needs of the body(Samuel et al.; 2017) Some of the common causes of heart disease are alcohol intake,smoking,family history,diabetes,hypertension, irregular food diet(Krishnaiah et al.; 2014).Some of the major symptoms of heart diseases are shortness of breath, pain in the heart, pain or discomfort in both arms or neck or stomach,sweating and pressure in the upper back.There are different type of heart disease some common types are coronary heart disease,congenital heart disease ,angina pectoris(Tarun and Tyagi; n.d.). Most common way to detect heart disease is echocardiography, this is a painless test which gives the shape of the heart and how well the valves and arteries are functioning,this test can also diagnose part of the heart muscle that are not contracting due to poor blood flow. Another test that can be performed to detect heart problem is doppler ultrasound which detects how well the blood is flowing through all the valve and arteries(Anbarasi et al.; 2010)This medical techniques has certain flaws and at sometimes fails to detect heart problem.Effective and efficient heart disease detection system will not only be beneficial but also a path breaking phenomenon in the field of healthcare sector that can save lakhs of people lives.Data mining techniques combines statistical analysis,database technology and machine learning to extract undiscovered data and relationships from large database(Chaurasia and Pal; 2013).Hence with the help of this machine learning algorithm detecting heart disease would not only save time but also reduce the number of test that is taken by the patients.

My motivation in doing this research is to save human lives as there is lack of trained cardiologists in many areas hence due to this patients are dying without any treatment.Providing quality services in the healthcare sector at a optimum cost is a challenge ,hence if we can use data Mining techniques to classify heart disease it would not only provide quality healthcare services but also reduce excessive dependence on human being thus by reducing cost,time and effort which will ultimately add to increased life expectancy and promotion of national health.

In this research I will follow the cross industry standard process for data mining technique(CRISP-DM) is used.This research deals with the following sections .Section 1 gives us the introduction. Section 2 explores the related work in the area of data mining paradigm to classify heart disease.Section 3 gives us the methodology used.Section 4 gives us the implementation of the research. Section 5 gives us Result/Evaluation and section 6 gives us Conclusion and Future work.Section 7 gives us acknowledgement.

2 Related Work

2.1 Heart Disease

Heart is a vital organ of our body, daily working of a human being depends on the proper functioning of heart, which is used to pump blood throughout the body(Deekshatulu et al.; 2013) There are various kind of heart disease some of them are

- Unstable angina
- Myocardial Infarction
- Complete heart block
- Lipid levels(Mohan et al.; 2014; Gandhi and Singh; 2015)

Symptoms of heart attack are-

- Vomiting or nausea
- Irregular heart beat
- Use of tobacco
- Family history(Deekshatulu et al.; 2013)

Cardiovascular disease is most commonly defined as abnormal functioning of heart causing cardiac attacks.As there are no particular test for detecting heart disease,clinical diagnosis is mainly dependent on patients history and physical examination which is backed up by ECG and chest cardiography(Son et al.; 2012)Estimated 5 million people died due to heart disease in the year 2008 itself which depicts 30 percent of the death worldwide.Out of these 2 million died due to stroke and 3 million were due to heart attack(Methaila et al.; 2014)

2.2 Literature review

In this section I am going to discuss the previous works of various researchers and their techniques to classify heart disease using various data mining algorithm and tools.Also I will critically review the works of each researches and how these related works helped me in my research

Hnin Wint Khaing [2011] presented a research paper to predict heart disease by extracting data relevant to heart disease from the database using the k-means clustering algorithm.Then the patterns are mined from the extracted data that contain heart disease using the MAFIA algorithm .The machine learning algorithm is trained with the extracted pattern for efficient prediction of heart disease.Decision tree was used as a training algorithm to show the level of heart attack.The method showed better prediction capability and very efficiently.In this method there was no use of any performance metric like AUC,Accuracy,sensitivity and specificity but it had predicted heart disease very efficiently.(Chaurasia and Pal; 2013)

The research paper by Chaurasia and Pal to predict heart disease from the database of heart disease with using just 11 vital attribute using data mining algorithm like naive bayes,decision tree and bagging approaches like random forest. The result of the research is that bagging algorithm outperformed naive bayes and decision tree algorithm.This research method shows that bagging algorithm performs better than the other two algorithm but the accuracy or any performance measure has not been mentioned in the paper(Chaurasia and Pal; 2013)

M.A Jabbar et al proposed a technique to predict heart disease using neural network as a classification algorithm and PCA ,chi-square as a feature subset for classification of heart disease.Then they compared it with decision tree,naive bayes, PART and Neural network which helped in predicting heart disease very efficiently.This techniques gives us clear and effective picture of predicting heart disease than the other methods as proposed by the other two researchers above.(Jabbar et al.; 2014)

Latha et al(2007) devised a technique called coactive neuro fuzzy interface system and genetic algorithm that was used in detecting heart disease.Two techniques were combined neural networks and genetic algorithm ,they together formed a hybrid method.Sole motive of this method is to use in heart disease detection and also to reduce the risk of heart disease in patients.In this technique no performance metrics were used like accuracy,sensitivity,specificity.(Banu and Swamy; 2016).

John Gennari et al devised a technique to predict heart disease using the logistic regression data mining algorithm with a accuracy of 77 percent.In this method the accuracy of the model is quite high using logistic regression but i will try to improve further the accuracy of the model and also there is no other performance metric like sensitivity,specificity.(Ahmed and Hannan; 2012).

Guru et al devised a technique to detect heart disease on a sample database of patient records.Neural network was used to test and train on the 13 attribute.Training of the attribute was done with the help of back propagation algorithm.The results of the unknown dataset was successful .In this method no performance measure was considered for comparison.Only one algorithm was used to detect heart which I think is not a good comparison criteria to classify heart diseases.(Anbarasi et al.; 2010)

Sellappan Palaniappan et al devised a technique called IHDPS (intelligent heart disease prediction system) using data mining technique like decision tree,naive bayes and neural network .The hidden pattern and relationship were used to develop the model.This model is very useful in detecting heart disease as it is user-friendly,scalable,reliable and web-based.This kind of model is different from the usual data mining algorithm where accuracy is a metric (Krishnaiah et al.; 2014)

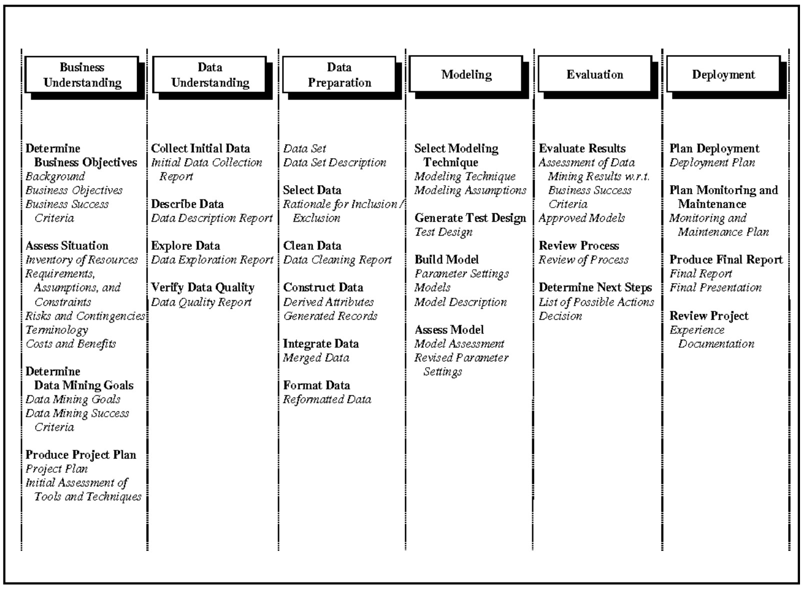
Sitar-taul et al devised a mechanism to detect heart disease using different data mining algorithm like naive-bayes,KNN and Random Forest.The accuracy of naive-bayes was 52.33%,KNN is 45.67% and Random Forest is 52% which helped to detect heart disease .This kind of model is not such an effective process to detect heart disease because the accuracy of all the three algorithm is very less.(Krishnaiah et al.; 2014)

Resul das et al used SAS enterprise miner 5.2 to develop an ensemble based neural network to classify heart disease.To build the ensemble three independent based neural network model was developed.There was no performance improvement when increasing the neural network in the ensemble.(El Bialy et al.; 2016)

3 Methodology

This research is carried out based on the CRISP-DM methodology.This is the most popular data mining technique that is followed in the industry as well as research works.It consists of six different phases which thereby indicates the design of the entire research project

3.1 Project Development



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 1: Stages of Data mining

The figure shown above depicts the six main stages of CRISP-DM data mining technique.Hence the application of machine learning to heart disease prediction would help the cardiologists around the world to save thousand of lifes.In healthcare industry there are lots of data that are undiscovered which some way or the other might have been useful in drawing meaningful insights from the data that could have led to early detection of heart disease. Data mining algorithm helps the healthcare industry in finding the insights and meaningful information from the data. This process is beneficial for both the doctors and the patients as early detection of disease saves a lot of time for the doctors which they can eventually use it to save the life of a patient.On the other hand it is beneficial to the patients as it reduces the cost of healthcare system which otherwise would have aggravated in due course of time.Data mining also helps us reduce the number of resource as there is a shortage of skilled resource in the developing countries.

Six main stages of CRISP-DM that correlate to my research are:-

- Business Understanding-In my research this step is used to define the research question i.e. application of various data mining algorithm to classify heart diseases and thus converting the business question into data mining paradigm.

- Data Understanding-The data set has been obtained from the UCI machine repository.Cleaning of the data set was done in R.Exploratory data analysis was done in R to get a insight of my data set by depicting it through graphs to understand correlation between the variables,outliers and class imbalance.

- Data preparation-After the raw data was obtained from the UCI machine repository feature engineering,PCA and correlation matrix was done to prepare the data set for the best result of the model and to select the best attribute for the model.As the accuracy of the model depends on the data set so feature engineering was done to the data set.

- Modelling-In my Research various data mining algorithm was implemented(ANN,LR,RF) to my data set to check the performance metrics(accuracy,sensitivity,specificity,kappa) of the three algorithm.

- Evaluation-The results of the exploratory data analysis(correlation matrix,graphs) and the data mining algorithm with all the metrics such as accuracy,sensitivity,specificity and kappa was presented in this section.

- Deployment-As it is a research project it can be deployed as a medical software in some hospitals so that a doctor can use it as a yardstick before treating any heart disease patients.

All the models will be evaluated on the basis of training data(the model is used to learn about the data in this phase)and from this learning the model will be tested on the testing data.In my research project performance metrics like accuracy,sensitivity, specificity and kappa is used.

- Accuracy- It gives us the fraction of the classifiers prediction that is correct.

- Sensitivity- It is also known as recall which is the amount of actual positives which are correctly identified by the classifier(percentage of sick patients who are correctly classified as having the disease).

- Specificity-It gives us the amount of actual negatives which is correctly identified by the classifier(Healthy person who are correctly classified as not having the disease).

- Kappa- It is used to compare the observed accuracy with the expected accuracy.

The performance of the all the models was evaluated and compared with the sole motive to look as to which algorithm provides the best performance.One should keep in mind that different methods are more favourable over the other in different business cases.In this research I would be focusing on accuracy,sensitivity,specificity and kappa metric as discussed with all the stakeholders.

3.2 Data Mining Algorithm To Detect Heart Disease For My Research

There are various kind of data mining algorithm to classify heart disease but I am going to use the following for my research because as I would be classifying the presence of heart disease or not on the categorical binary data i.e. num(target variable) in my data set

Random Forest-This is an ensemble based supervised classification algorithm that is used to build multiple decision trees. Hence it is less prone to over-fitting.This is one of the most widely used classification algorithm. It consists of two parameters ntree(number of trees) and mtry(number of attribute to select the best subset) (Kumar and Shaikh; 2017). Classification algorithm has direct impact on the accuracy of the model and in recent years researches has found out that the accuracy of ensemble learning is more than that of a single classifier(Xu et al.; 2017)

Advantages

- New data can be added
- Easy to understand
- Does not suffers from over-fitting

Disadvantage

- Time consuming
- Non-Numeric data is difficult to understand(Gandhi and Singh; 2015)

Artificial Neural network-It is similar to biological model,which is used generally for a computing system that is made of number of a simple,interconnected elements that is used to process information by their dynamic state response to the external inputs.(Methaila et al.; 2014). Neural networks are known to generate highly accurate results in practical applications. Artificial neural networks is a powerful algorithm that helps the doctors to analyze and find meaningful insights from the model which otherwise would have been difficult to find. Hence it is used for disease classification(Rani; 2011)

Advantages

- It can be used in large data sets
- Once the trained data is entered we dodnt need to reprogram.

- It can handle missing or noisy data.

Disadvantage

- Time consuming
- It needs to be properly trained to work well
- If any modifications are done on the data then it cant be added to the existing neural nets.(Gandhi and Singh; 2015)

Logistic regression-This is considered as a special case of linear regression model but the binary response variable violates the normality assumptions of linear regression model.It works best when there is binary classification of data. This is the most commonly used machine learning algorithm incase of medical data.(Khemphila and Boonjing; 2010)

Advantage

- It can produce simple probabilistic formula of classification
- It works the best incase of binary classification.

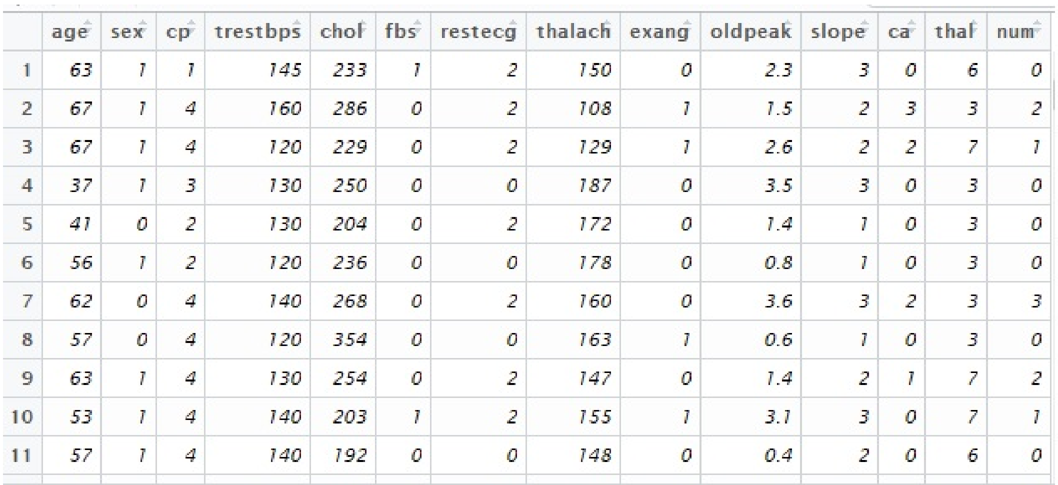
Disadvantage

- It generally requires large data sets to implement this algorithm.(Khemphila and Boonjing; 2010).

4 Implementation

4.1 Data Preparation and Feature engineering

The data set consists of 303 observation with 14 attribute including sex, age ,cp of the patient.http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data. Most common problem is with binary categorical variable which may affect the performance of the model.Here age,trestbp,chol,thalach and oldpeak is a continuous variable.Whereas sex,cp,fbs,restecg,exang,slope,thal and num are categorical variable.Only Ca is a discrete variable.Encoding has been done to the data set. The raw data set and the description of the data set is given below.



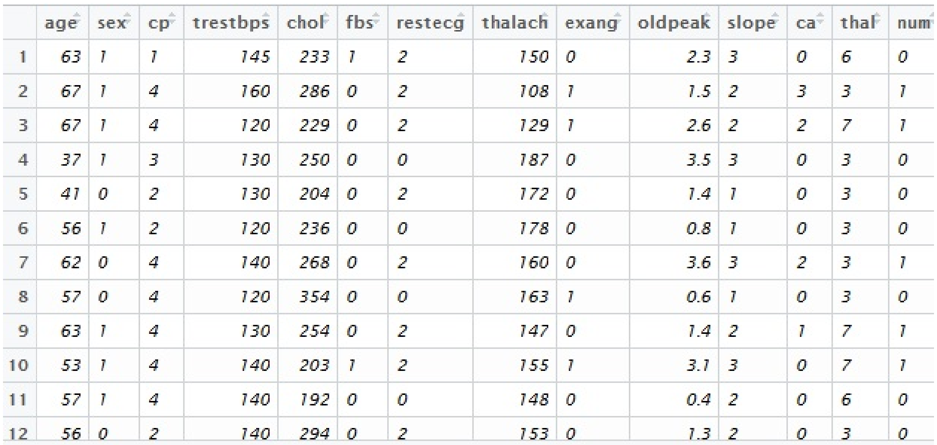
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
1	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
6	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
7	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
8	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
9	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
10	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
11	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0

Figure 2: UCI Raw Data

Table 1: Description of Attributes

Name	**Description**
Age	Age in years
Sex	0=female 1=male
Cp	chest pain type 1: typical angina, 2: atypical angina, 3: non-angina, 4: asymptomatic angina
Trestbps	resting blood pressure on admission to hospital in mmHg
Chol	serum cholesterol level in mg/dl
Fbs	fasting blood sugar 0: <= 120 mg/dl, 1: >120 mg/dl
Restecg	resting electrocardiography 0: normal, 1: ST-T wave abnormality, 2: left ventricular hypertrophy
Thalach	maximum heart rate achieved
Exang	exercise induced angina 0: no, 1: yes
Oldpeak	ST depression induced by exercise relative to rest
Slope	slope of peak exercise ST segment 1: upsloping, 2: flat, 3: downsloping
Ca	number of major vessels colored by fluoroscopy 0,1,2,3
Thal	heart status 3: normal, 6: fixed defect, 7: reversible defect
Num	Class label 0:No heart disease 1:Heart disease

In explanatory data analysis I have implemented boxplot and mosaic plot to depict the characteristics of data .Cleaning of the data set was done in R programming language which includes eliminating null values, missing values etc. In my data set only 6 rows were present that consists of missing values and null values.After the initial data preparation part I tried to do PCA but as my data set is small and the correlation between the attribute is not so strong hence it did not provide any improvement in the performance of the model hence I am considering all the attributes.PCA is mainly done for dimensionality reduction .R was used for implementing different data mining algorithm in my research project.Correlation matrix was shown to depict the relationship between the variables(strong or weak).Now feature engineering was applied to the data set to enhance the performance of the model.Binary categorical variables are either encoded as 0 or 1.So all the categorical variables are converted into numeric value.In the research Thal was coded as normal(3),fixed defect(6) and reversible defect(7).Chest pain were coded as typical angina(1),atypical angina(2),non-anginal pain(3),asymptomatic angina(4).Feature Engineering was done in the last column num where all the variables are encoded as 0 or 1.



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
1	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
2	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
3	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
4	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
5	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
6	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
7	62	0	4	140	268	0	2	160	0	3.6	3	2	3	1
8	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
9	63	1	4	130	254	0	2	147	0	1.4	2	1	7	1
10	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
11	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
12	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0

Figure 3: Feature Engineering

In this research project classification of heart disease is done as present or not and our target variable is num which we are going to predict by using the various machine learning algorithm.

4.2 Design of the Model

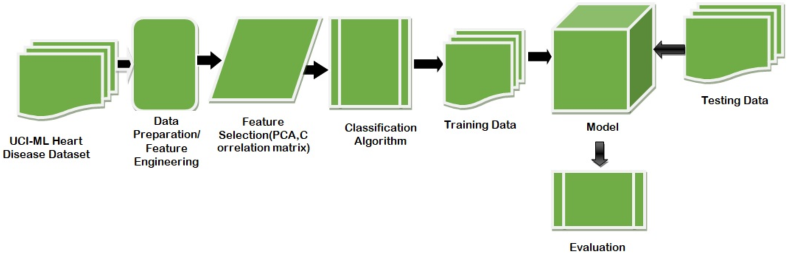

UCI-ML Heart Disease Dataset
Data Preparation/ Feature Engineering
Feature Selection(PCA,C orrelation matrix)
Classification Algorithm
Training Data
Model
Testing Data
Evaluation

Figure 4: Workflow of the model

The data set is obtained from the UCI-ML machine repository,preparation of the data set was done in R.Feature selection was done by using PCA and Correlation matrix.The classification algorithm was applied on the model.After that the data was splitted in 70:30 ratio of training and testing and applied on the model to find the performance of the model and compare all the metrics which among the three is the best.

4.3 Implementing the machine learning algorithm

Random Forest-It is mostly used when there is a categorical binary data classification hence it is used.It works best with categorical binary data. It is implemented in R by the random forest and caret package(Kuhn et al.; 2014) with mtry=10 and ntree=200 for the initial tests.It was run in R to check the performance metrics of the model.

Artificial Neural Network-In this algorithm we use the neural net and the caret package(Kuhn et al.; 2014).In our model we take neurons as 8 for our model which means the second layer which is known as the hidden layer is the weighted sum of the values in the first layer.This second layer then projects to the third layer(Rani; 2011; Sohn and Dagli; 2004) i.e. the target variable which is num in my model.

Logistic Regression-It is used because for binary classification problem.It can handle missing data.(Khemphila and Boonjing; 2010)It uses the caret library package(Kuhn et al.; 2014) .The method used here is glm(generalized linear model) and the family used here is binomial as it a binary classification problem.

5 Evaluation

5.1 Exploratory data analysis result

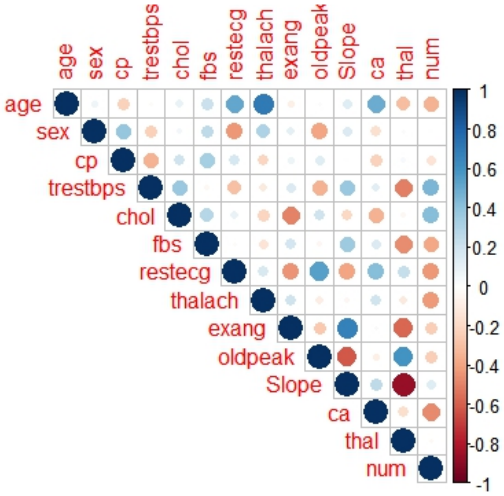


Figure 5: Correlation Matrix

The correlation matrix is used to depict the correlation between the variables.Corrplot package was used in R(Wei and Simko; 2016).Blue colour is used to depict positive correlation and red colour is used to depict negative correlation.Intensity of colour and size of the circle is directly related to correlation coefficients.

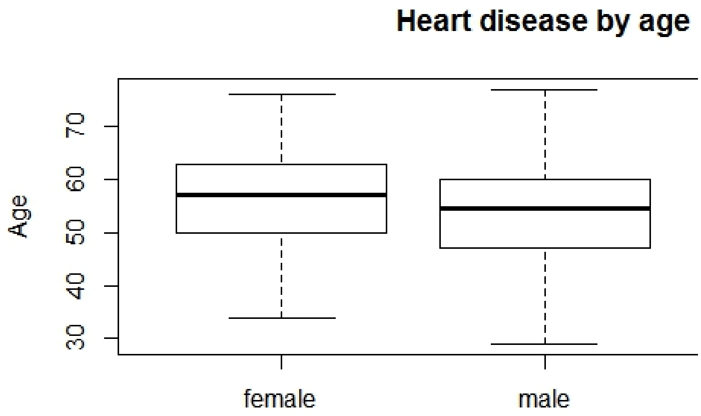


Figure 6: Box Plot

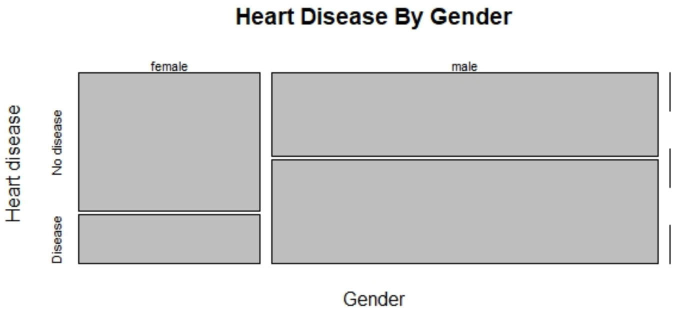
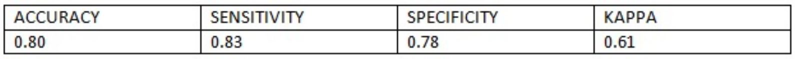


Figure 7: Mosaic Plot

In the above explanatory data analysis of the data set various graphs were shown to depict the characteristics of the data set. In the boxplot we observe that in females heart disease is detected at the age of 35 upto to the age of 70 and more but it is more frequent at the age of 50 to 60.Whereas in males are affected by heart disease at the age of 30 to the age of 70 but more affected in the age of 45 to 60.Mosiac plot is used to depict heart disease by gender.We can see that males are more affected from heart disease than the females.

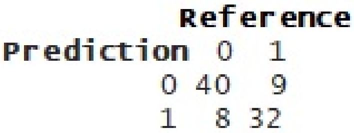
5.2 Evaluation of the machine learning Algorithm

Random Forest-It was implemented using the caret and the random forest library.(Kuhn et al.; 2014)Null and Missing values were removed from the data set.According to the confusion matrix given the number of true positive cases i.e. number of correctly classified positive instances is 40(TP) and the number of true negative cases i.e.number of correctly classified negative prediction is 32(TN).Whereas the number of false negative cases i.e. number of incorrect negative prediction is(TYPE II error) 8(FN). and the number of false positive cases is i.e. number of incorrect positive prediction (TYPE 1 error) is 9(FP).Performance measures like accuracy,specificity,sensitivity and kappa of the algorithm was found out to be 80.9%,78.05%,83.33% and 61.49% with the 70:30 splitting of the training and testing data set.



ACCURACY	SENSITIVITY	SPECIFICITY	KAPPA
0.80	0.83	0.78	0.61

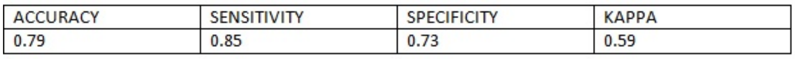
Figure 8: Performance Metrics in Random Forest



```
          Reference
Prediction  0  1
         0 40  9
         1  8 32
```

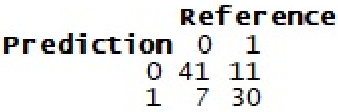
Figure 9: Confusion Matrix

Artificial Neural Network-It was implemented using the nnet,caret,e1071 and devtools package(Bergmeir and Benítez Sánchez; 2012; Kuhn et al.; 2014; Hothorn; 2017; Wickham and Chang; n.d.).According to the confusion matrix given the number of true positive cases i.e. number of correctly classified positive instances is 41(TP) and the number of true negative cases i.e.number of correctly classified negative prediction is 30(TN).Whereas the number of false negative cases i.e. number of incorrect negative prediction is(TYPE II error) is 7(FN) and the number of false positive cases i.e. number of incorrect positive prediction (TYPE 1 error) is 11(FP)Performance metrics like accuracy,sensitivity,specificity and kappa of the algorithm was found out to be 79.78%,85.42%,73.17% and 59.01% percent with the 70:30 splitting of the data set into training and testing data.



ACCURACY	SENSITIVITY	SPECIFICITY	KAPPA
0.79	0.85	0.73	0.59

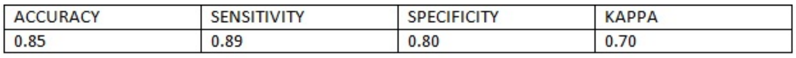
Figure 10: Performance Metrics in Artificial Neural Network



```
          Reference
Prediction  0  1
         0 41 11
         1  7 30
```

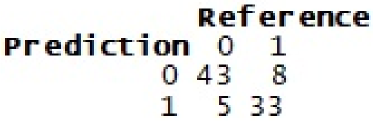
Figure 11: Confusion Matrix

Logistic Regression-It was implemented using the caret package.(Kuhn et al.; 2014)According to the confusion matrix given the number of true positive cases i.e. number of correctly classified positive instances is 43(TP) and the number of true negative cases i.e.number of correctly classified negative prediction is 33(TN).Whereas the number of false negative cases i.e. number of incorrect negative prediction is(TYPE II error) is 5(FN) and the number of false positive cases i.e. number of incorrect positive prediction (TYPE 1 error) is 8(FP)Performance metrics like accuracy,sensitivity,specificity and kappa of the algorithm was found out to be 85.39%,89.58%,80.49% and 70.45% with the 70:30 splitting of the data set into training and testing data.



ACCURACY	SENSITIVITY	SPECIFICITY	KAPPA
0.85	0.89	0.80	0.70

Figure 12: Performance Metrics in Logistic Regression



```
          Reference
Prediction  0  1
         0 43  8
         1  5 33
```

Figure 13: Confusion Matrix

5.3 Graphical Representation

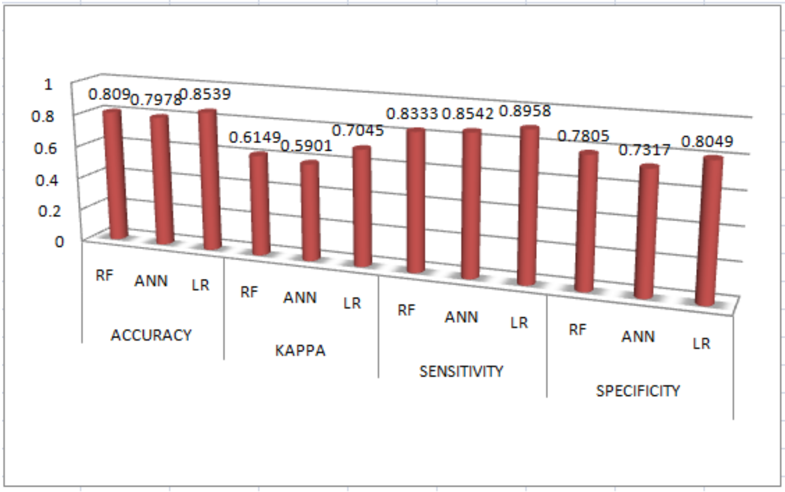


Figure 14: Comparison of the three algorithm

6 Conclusion and Future Work

The research project is use to evaluate the application of various data mining algorithm to classify heart disease and check which algorithm provides the best performance metric(accuracy,sensitivity,specificity and kappa). Some of the Research Findings are-

- We can see from the Figure 14 that logistic regression is the best performing algorithm in terms of all the performance metrics as compared to the other two algorithm.
- We can see that logistic regression is mainly used for binary classification and also if the dependent variable is categorical(num is the target variable/dependent variable which is encoded as 0 and 1) then it provides the best classification result than any other classification algorithm.

- We can see that number of false positive(patients having the disease but classified as having no disease) is 8 which is less than the other two algorithm.
- If we look at the kappa value which is 70.45% which falls in the good agreement category which is higher than the other two algorithm.

There are several gaps that can be taken care of in the future-

- One thing that can be done for future is applying Deep learning technique and check whether they can classify heart disease more accurately than the above mentioned machine learning algorithm but for that we need a large data set as deep learning works on large data set.
- We can use unsupervised algorithm to classify the heart disease

7 Acknowledgement

This research work though an individual piece of work but benefited from the able guidance and motivation of my thesis supervisor Mr.Jorge Basilio who has guided me not only in technical matters of the thesis but also throughout the whole process of the thesis.I am grateful to him for his invaluable advice,knowledge and encouragement throughout the entire thesis.Next I would like to thank Dr.Bidyut Basu for his valuable insights into the clinical data set that I have obtained from UCI.I would also like to thanks UCI for making the data set public so that everybody can use it, as clinical data is hard to find due to their non-disclosure agreement. Finally I would like to thank my friend Mr Prashanth Avverahalli Ramesha for his patient proof reading of my thesis and also for his valuable inputs.

References

Ahmed, A. and Hannan, S. A. (2012). Data mining techniques to find out heart diseases: An overview, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* **1**(4): 18–23.

Anbarasi, M., Anupriya, E. and Iyengar, N. (2010). Enhanced prediction of heart disease with feature subset selection using genetic algorithm, *International Journal of Engineering Science and Technology* **2**(10): 5370–5376.

Banu, N. S. and Swamy, S. (2016). Prediction of heart disease at early stage using data mining and big data analytics: A survey, pp. 256–261.

Bergmeir, C. N. and Benítez Sánchez, J. M. (2012). Neural networks in r using the stuttgart neural network simulator: Rsnns.

Chaurasia, V. and Pal, S. (2013). Early prediction of heart diseases using data mining techniques, *Caribbean Journal of Science and Technology* **1**: 208–217.

Deekshatulu, B., Chandra, P. et al. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm, *Procedia Technology* **10**: 85–94.

Desai, U., Nayak, C. G., Seshikala, G. and Martis, R. J. (2017). Automated diagnosis of coronary artery disease using pattern recognition approach, pp. 434–437.

El Bialy, R., Salama, M. A. and Karam, O. (2016). An ensemble model for heart disease data sets: a generalized model, pp. 191–196.

Gandhi, M. and Singh, S. N. (2015). Predictions in heart disease using techniques of data mining, pp. 520–525.

Hothorn, T. (2017). Cran task view: Machine learning & statistical learning.

Jabbar, M., Deekshatulu, B. and Chndra, P. (2014). Alternating decision trees for early diagnosis of heart disease, pp. 322–328.

Khemphila, A. and Boonjing, V. (2010). Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients, pp. 193–198.

Krishnaiah, V., Srinivas, M., Narsimha, G. and Chandra, N. S. (2014). Diagnosis of heart disease patients using fuzzy classification technique, pp. 1–7.

Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., Team, R. et al. (2014). caret: Classification and regression training. r package version 6.0–21, *CRAN: Wien, Austria* .

Kumar, S. S. and Shaikh, T. (2017). Empirical evaluation of the performance of feature selection approaches on random forest, pp. 227–231.

Mahajan, R., Viangteeravat, T. and Akbilgic, O. (2017). Improved detection of congestive heart failure via probabilistic symbolic pattern recognition and heart rate variability metrics, *International Journal of Medical Informatics* **108**: 55–63.

Methaila, A., Kansal, P., Arya, H. and Kumar, P. (2014). Early heart disease prediction using data mining techniques, *Computer Science & Information Technology Journal* pp. 53–59.

Mohan, K. R., Paramasivam, I. and Narayan, S. S. (2014). Prediction and diagnosis of cardio vascular disease–a critical survey, pp. 246–251.

Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach, *arXiv preprint arXiv:1110.2626* .

Samuel, O. W., Asogbon, G. M., Sangaiah, A. K., Fang, P. and Li, G. (2017). An integrated decision support system based on ann and fuzzy_ahp for heart failure risk prediction, *Expert Systems with Applications* **68**: 163–172.

Sohn, S. and Dagli, C. H. (2004). Ensemble of evolving neural networks in classification, *Neural Processing Letters* **19**(3): 191–203.

Son, C.-S., Kim, Y.-N., Kim, H.-S., Park, H.-S. and Kim, M.-S. (2012). Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches, *Journal of biomedical informatics* **45**(5): 999–1008.

Tarun, A. S. and Tyagi, D. (n.d.). Comparative analysis of machine learning techniques in heart disease prediction by r language.

Wei, T. and Simko, V. (2016). corrplot: Visualization of a correlation matrix. r package version 0.77, *CRAN, Vienna, Austria* .

Wickham, H. and Chang, W. (n.d.). devtools: Tools to make developing r packages easier, 2015, *URL http://CRAN. R-project. org/package= devtools. R package version* **1**(0): 185.

Xu, S., Zhang, Z., Wang, D., Hu, J., Duan, X. and Zhu, T. (2017). Cardiovascular risk prediction method based on cfs subset evaluation and random forest classification framework, pp. 228–232.