Volume 6, Issue 7, July 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Suggesting Precautions by Finding Causes with Dataset Analysis on Heart Disease

A. Hareesh[*]
M-Tech Student, Department of CSE, JNTUA College of Engineering, Ananthapuramu, India

Dr. A. P. Siva Kumar
Assistant Professor, Department of CSE, JNTUA College of Engineering, Ananthapuramu, India

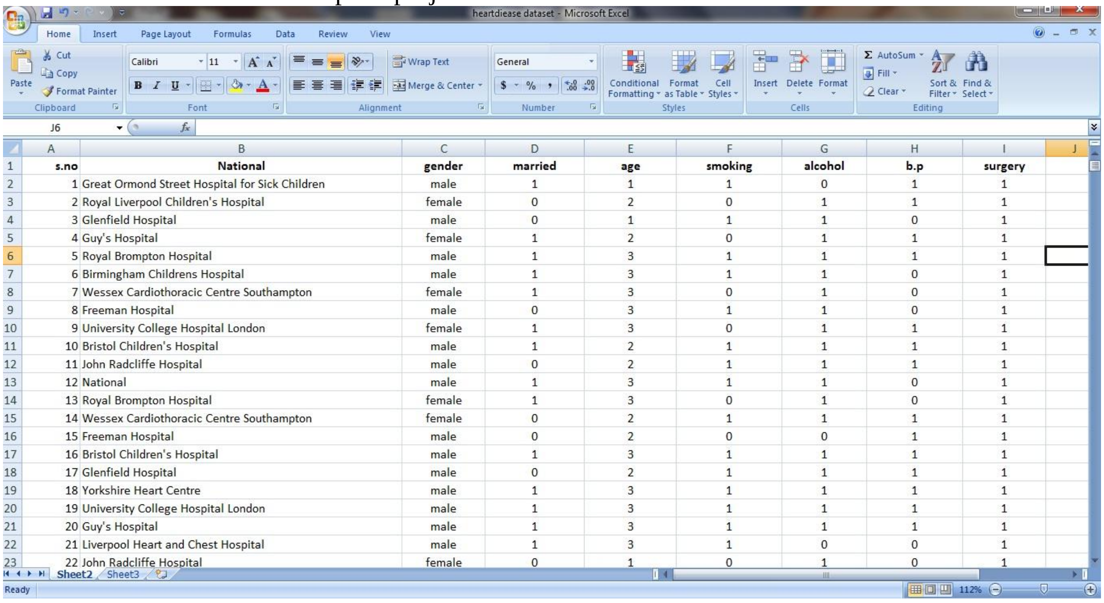
Abstract— This paper gives suggestion to the people who are suffering from heart disease. Now-a- days lot of the people are affecting by heart disease. In India heart disease become a major health issue as majority of the people are unaware of this heart disease. Because of this many people are leaving their lives. In this project we are taken a predefined dataset of heart disease which constitutes data items of 736 patients with 9 attributes. Based on this dataset we are able to perform statistical computations on that dataset by using R programming. Finally, we design a tree which gives probability of the patients who are suffering from heart disease. Based on this data we generate a report on that dataset. This report contains the complete information based on the attributes and gives us an idea to know which attributes are worst affecting the patients. Now we can give precautions and suggestions to the people who are suffering from the heart disease.

Keywords— Heart Disease, Dataset, Venn-Diagrams, Probability, Tree Structure, Set theory, Data analytics , R-Programming.

I. INTRODUCTION

In India Rural area people have a lack of awareness on the diseases. So, this reason may reflect the major effect on the people life. The Government has conducted a lot of awareness programs to get awareness for the people. A Lot of organizations are work in this Area and they are doing research on newly threatening disease to find causes for that disease due to take awareness for the people to overcome those diseases. In the research purpose, Datasets of the disease plays a major role to find out the solution for that problem. The Dataset is having the information about the data corresponding to some real time matter.

We want to perform operations on the dataset using several approaches among that R-programming technique is the best approach to perform statistical operation on Dataset . R programming is used to perform the statistical operations on Dataset. Now R programming is widely used in many platforms in this world. Majorly R programming is using in the automobile industry, weather forecasting, company fiscal budget analyzing purposes. Even in Bio-medical category also R programming plays a crucial role. By using this R programming, we can perform different operations on the dataset with the help of mathematical operations .There are a lot of mathematical operations to perform analysis on the dataset. By using R-Studio we can perform R programming operations. This R-studio is used to perform the statistical operation on the datasets also we can save the complete project for future reference.



	A	B	C	D	E	F	G	H	I	J
1	s.no	National	gender	married	age	smoking	alcohol	b.p	surgery	
2	1	Great Ormond Street Hospital for Sick Children	male	1	1	1	0	1	1	
3	2	Royal Liverpool Children's Hospital	female	0	2	0	1	1	1	
4	3	Glenfield Hospital	male	0	1	1	1	0	1	
5	4	Guy's Hospital	female	1	2	0	1	1	1	
6	5	Royal Brompton Hospital	male	1	3	1	1	1	1	
7	6	Birmingham Childrens Hospital	male	1	3	1	1	0	1	
8	7	Wessex Cardiothoracic Centre Southampton	female	1	3	0	1	0	1	
9	8	Freeman Hospital	male	0	3	1	1	0	1	
10	9	University College Hospital London	female	1	3	0	1	1	1	
11	10	Bristol Children's Hospital	male	1	2	1	1	1	1	
12	11	John Radcliffe Hospital	male	0	2	1	1	1	1	
13	12	National	male	1	3	1	1	0	1	
14	13	Royal Brompton Hospital	female	1	3	0	1	0	1	
15	14	Wessex Cardiothoracic Centre Southampton	female	0	2	1	1	1	1	
16	15	Freeman Hospital	male	0	2	0	0	1	1	
17	16	Bristol Children's Hospital	male	1	3	1	1	1	1	
18	17	Glenfield Hospital	male	0	2	1	1	1	1	
19	18	Yorkshire Heart Centre	male	1	3	1	1	1	1	
20	19	University College Hospital London	male	1	3	1	1	1	1	
21	20	Guy's Hospital	male	1	3	1	1	1	1	
22	21	Liverpool Heart and Chest Hospital	male	1	3	1	0	0	1	
23	22	John Radcliffe Hospital	female	0	1	0	1	0	1	

Fig. 1 Dataset of Heart Disease patients

We are gathering the information related to heart disease patients Dataset. In this dataset patients, all are suffering from heart disease and they are doing treatment from that disease. In this data set, each row may represent the information related to every patient behavior and each column represents the heart disease attributes. This project we follow the tree structure approach because basically, it is very help full for this kind of projects. Tree approach helps us to easily recognize the data than the linear approach.

We are using basic probability and set theory concepts to get the results from the dataset. Probability concepts are used to finding out probabilities of each node representing in the tree-structure. This project we maintain a tree structure in that tree structure is used calculate the conditional probability of different kind of data. Conditional probability majorly used in bio-medical engineering for detecting and preventing the disease. And Set operations are very helpful for us to recognize the data in an easier manner.

II. DATA AND METHODOLOGY

A. Dataset

The dataset is having the data identified with coronary illness patients among a few healing facilities. We gather dataset on various fields like doctor's facilities, the Internet or National therapeutic exploration focus. we arrange the dataset into our required fashion. We have collected the information among 736 patients records related to 30 Hospitals. A Dataset of Heart disease patients is shown in Fig. 1 .We perform Mathematical operations on dataset row wise and column wise based on the requirement by using R-programming language. It shows our dataset into graphical manner. Based on this operation we identify the root node for further proceeding. In the graphical representation of the Dataset is shown in Fig. 2., Fig. 3 and Fig. 4 as given below

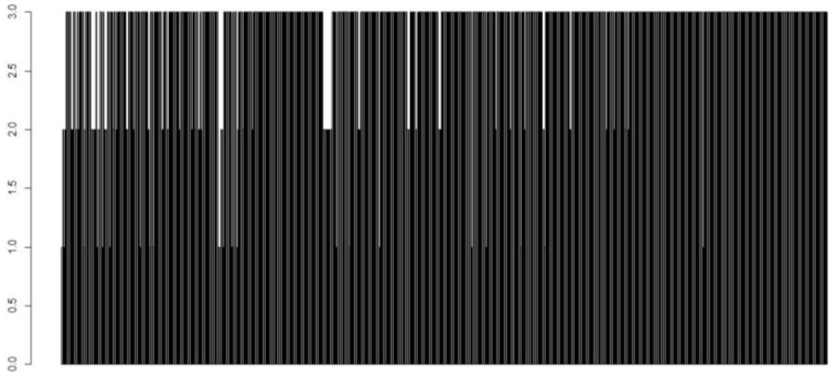


Fig. 2 Age group representation for different patients in the Dataset

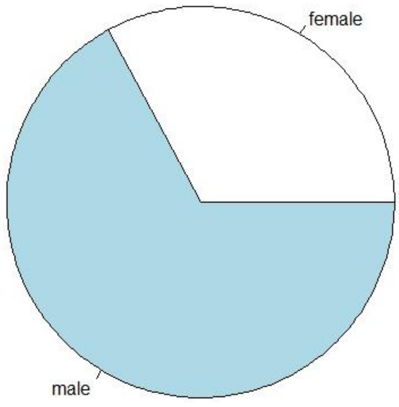


Fig. 3 Gender representation of Dataset

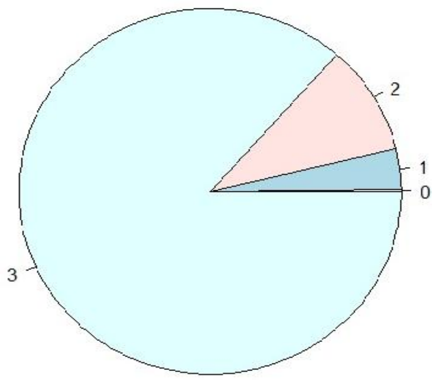


Fig. 4 Age representation of Dataset

B. Tree Structure:

After identifying the root of the dataset we start constructing a tree for easier identification of the Dataset. Our dataset id having the major attributes like Gender, Smoking , Drinking, Blood Pressure And Age. We have arranged attributes into tree manner and this tree plays a crucial role in the project. Because all the further proceeding is to be based on this tree. In the dataset, gender is categorized into male and female. The major attributes of the dataset Smoking, Alcohol, Blood Pressure is arranged in results like 0(No) or 1(Yes). Another property Age is classified into three classifications Age1 (Up to 25 years),Age2 (25years to 50 years) and Age3(above 50 years).The Tree structure of the given Dataset is shown in Fig. 5

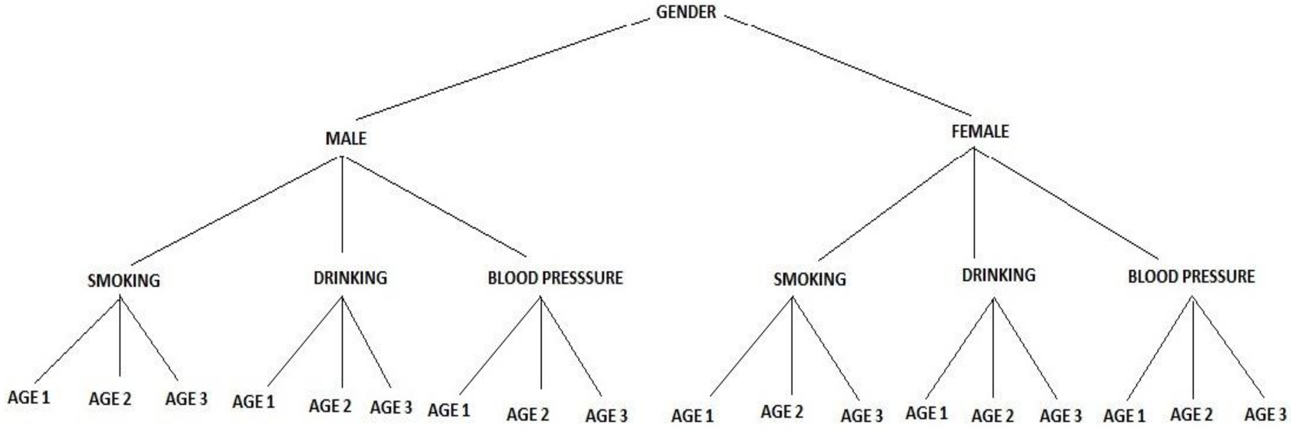


Fig. 5 Tree structure representation of Dataset

C. Statistical Operations

To arrange dataset into the above tree structure then we perform statistical operations on that dataset. We divide entire dataset according to the tree structure. After that, we calculate the probabilities of each node mentioned in the dataset. R-programming language helps us to perform operations on the heart disease dataset. It split the dataset into small datasets. It named it as F,M, FS,FA,FB,MS,MA, MB etc. And then we calculate the probabilities of each sub-datasets and every variable named it as Fp,Mp,FSp,FAp,FBp,MSp,MAp,MBp etc..Basic probability operation are used to detect the results for which kind of persons are most effecting heart disease. We calculate the probabilities of each variable based on the given formula

Probability =[No. of possible outcomes/Total no. of outcomes]

For an example

Fp = No. of female patients/Total No. of Patients in a Dataset

After getting the Results of subsets then we further divide base Dataset based on the attributes Smoking, Drinking and Blood pressure. To get all the probabilities with their combinations.

D. Graphical Representation of Data

We represent every subset of a dataset into Venn-diagram form because it is very useful to identified each and every combination of a sub-datasets. Based on this process we calculate every variable probability and finding out their combinations. Venn-Diagrams are plays crucial role for this project to identify all the variables and finding out all the variable. We consider Total Dataset is treated as sample space. Representation of Venn-Diagrams for this project is shown in Fig. 6, Fig. 7 and Fig. 8. We are using basic union and intersection formulas to differentiate the Dataset into Sub datasets.

For example

$$n(S \cup A)=n(S)+n(A)-n(S \cap A)$$

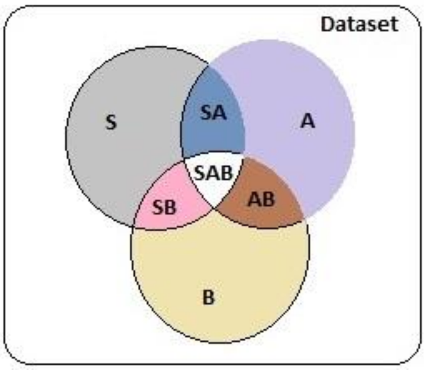


Fig. 6 Venn-diagram of smoking, Alcohol and Blood Pressure

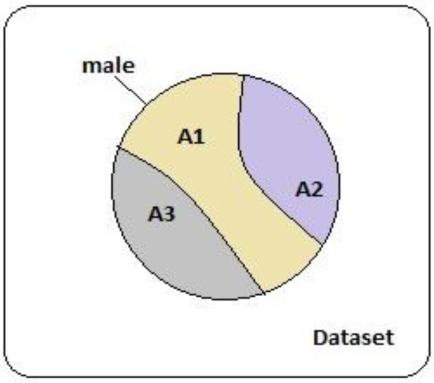


Fig. 7 Venn –diagram representation for male Patients with respect to Age

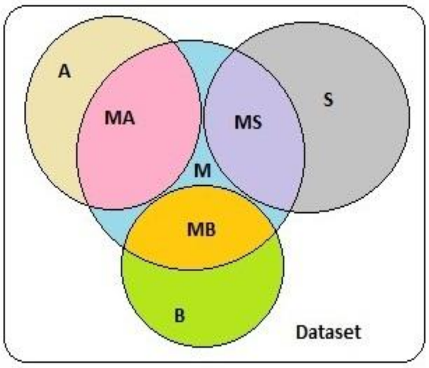
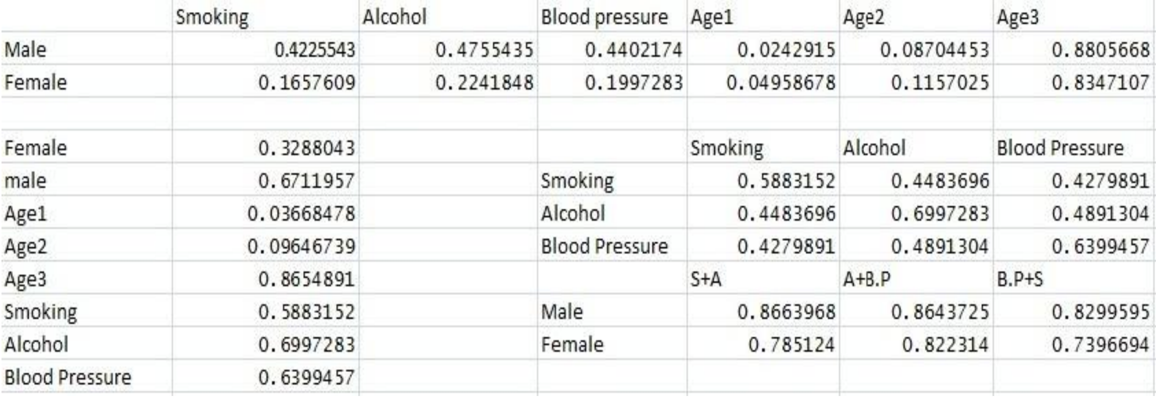


Fig. 8 Representation of Male patients with combination of Smoking, Alcohol, Blood Pressure

III. EXPERIMENTAL RESULTS

We perform statistical operations on Dataset by using R-Programming Language. We are divide Dataset into several Sub Datasets by applying R-Programming commands on that dataset. And also we maintain comma separate value files of sub Datasets. We prepared a report of having probability values of each node in Tree Structure. We calculate the probability values of union and intersections of attributes mentioned in above procedure. This report I having the probabilities of patients suffer from heart disease. In this way, That we discover which sort of patients are more impact by Heart Disease. In view of the procedure we assess the reasons for Heart Disease . To that causes, we give Suggestions for the general population. At the end of this project we give awareness for the people who are suffering from Heart Disease. There are a lot of organizations doing research on new kind diseases are suffered by the people. To completion of the research, they are announced precautions for the people who suffers to those diseases.

Finally, we prepare a report is having total probabilities of different combinations which are mentioned in our analysis phase. The Report helps us to understand which category people are more suffering from the disease. Final Report of this project is shown in Fig. 9 . This report is prepared in excel sheet format.



	Smoking	Alcohol	Blood pressure	Age1	Age2	Age3
Male	0.4225543	0.4755435	0.4402174	0.0242915	0.08704453	0.8805668
Female	0.1657609	0.2241848	0.1997283	0.04958678	0.1157025	0.8347107
Female	0.3288043			Smoking	Alcohol	Blood Pressure
male	0.6711957		Smoking	0.5883152	0.4483696	0.4279891
Age1	0.03668478		Alcohol	0.4483696	0.6997283	0.4891304
Age2	0.09646739		Blood Pressure	0.4279891	0.4891304	0.6399457
Age3	0.8654891			S+A	A+B.P	B.P+S
Smoking	0.5883152		Male	0.8663968	0.8643725	0.8299595
Alcohol	0.6997283		Female	0.785124	0.822314	0.7396694
Blood Pressure	0.6399457					

Fig. 9 Final Report of Heart Disease Patients

IV. CONCLUSION AND FUTURE ENHANCEMENT

This project is helping us to give awareness for the rural people who are suffered from Heart Disease. To apply basic Probability and set theory concepts on Dataset regarding Heart Disease patients. R-Programming language is used to perform the operation on the Datasets. To apply this procedure into any kind of disease Dataset. R-Programming language plays a major role in this project to obtain probabilities. Dataset is storing information regarding particular topic.

It helps to understand the problem of the topic. R-Programming is majorly used in Data analytics side. To combine R-programming concept into Computer Networks concepts to reduce the traffic in metropolitan cities. R-Programming is having special features to perform operations on any kind of dataset irrespective of the information, data etc. In future work, we apply this kind of tree structure procedure to any kind of disease dataset to find causes therefore to create awareness for the society. Reduce the traffic signal in metropolitan cities we use traveling salesman problem topic with the combination of R-Programming. We collect information about daily traffic areas and every hour vehicles travel to that junctions. Based on that information we apply computer network problem to move some vehicles into another junction. R-Programming is used to find out company fiscal results and weather forecasting.

REFERENCES

[1] “Google Search engine”.Available:http://www.google.com.

[2] “R Programming tutorial site” Available:http://www.tutorialspoint.com/r/ .

[3] “Git hub website” Available:http://github.com/R datasets.

[4] “UCI Machine learning website” Available:http://archive.ics.uci/edu/ml/datasets/Heart+Disease.

[5] K. Sudhakar and M. Manimekalai, *Study of Heart Disease Prediction using Data Mining*, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 4, no. 1, 2014.

[6] Rupali R. Patil., *The Heart Disease Prediction System using Naive Bayes and Jelinek-mercer smoothing*, International Journal of Advanced Research in Computer and Communication Engineering, vol. 3, no. 5, 2014.

[7] Jianting Sheng, Fuhai Li, and Stephen T. C. Wong “Optimal Drug Prediction From Personal Genomics Profiles” IEEE journal of biomedical and health informatics, vol. 19, no. 4, July 2015.

[8] Deepali Chandna, "Diagnosis of Heart Disease Using Data Mining Algorithm", (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 5, no. 2, pp. 1678-1680, 2014.

[9] Aqueel Ahmed and Shaikh Abdul Hannan, "Data Mining Techniques to Find Out Heart Diseases: An Overview", International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2012.

AUTHOR’S PROFILE

A.Hareesh obtained B.Tech degree in Computer Science Engineering from Priyadarshini College of Engineering and Technology, Nellore Affiliated to Jawaharlal Nehru Technological University, Anantapur,A.P,India. Currently pursuing M.Tech in Software Engineering from Jawaharlal Nehru Technological University Anantapur College of Engineering,JNT University, Anantapur, A.P, India,during 2014 to 2016. His research interests include Data Analytics and Data Mining.

Dr A.P.Siva Kumar is currently working as an Assistant Professor in Computer Science at JNTUA College of Engineering, JNT University, Anantapur, A.P, India. He received his Ph.D degree in Computer Science And Engineering from JNT University, anantapur,A.P, India. He received the bachelor’s degree in 2002 and the Master’s degree in 2004, both from JNTU Hyderabad, A.P, India. He has around 10 years of experience as a Lecturer/Research and Development with strong analytical background in the education sector. His research interests are cross Lingual Information retrieval and Natural Language Processing.