ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

Comparative Analysis of Heart Disease Dataset using KNN and Decision Tree Classification

[1] Bhavini Bhatia, [2] Vamika Razdan
[1][2] UG Scholar Department of B.Tech(I.T)
[1][2] Bharati Vidyapeeth's College of Engineering,New Delhi,India

***Abstract*— There are huge amount of data in the medical industry which requires prediction and analysis so that right decisions can be made and help in proper analysis. As the data is large and the decision made by the doctor may not be accurate which may result in failure in some cases and can sometimes put someone's life at stake. Data mining in healthcare is an intelligent diagnostic tool. Heart disease is the leading cause of death in the world over the past 10 years.29.2% of deaths are due to Cardiovascular Diseases(CVD). Researchers have been using several data mining techniques to help healthcare professionals in the diagnosis of heart disease. Decision Tree is one of the successful data mining techniques used. However, most researchers have applied KNN . Number of experiments has been conducted to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms other predictive methods like KNN, Neural Networks.This research paper intends to provide the knowledge that which data mining technique gives better accuracy and should be used in Heart Disease analysis.**

***Index Terms*— KNN,Decision Tree,heart diseases**

1. INTRODUCTION

Data analysis comprises of vast potential in the field of healthcare that enables health systems to methodologically use data mining and analytics to classify the inefficiencies and henceforth improve care and reduce costs.

The healthcare industry has immensely benefitted from data mining and also finds future applications such as detection of analysis of health care centers for betterment in health policy-making and preventing the errors occurring in hospitals, prevention from various diseases, detecting fraudulent insurance claims etc.. The enormous data being added each and every day motivates the researchers to extract useful knowledge and use data mining techniques to help the healthcare professionals in the diagnosis of heart disease.The enormous sea of data is not unmethodical.It has a hidden pattern ,significant relationships and knowledge which are difficult to detect with traditional statistical methods. Thus data mining helps to obtain such information and helps to predict substantial results. Data mining in healthcare is of great importance and is an emerging field for providing accurate prognostication which is very difficult and a deeper understanding of medical data. Several data mining techniques are used in the diagnosis of heart diseases and many other health related problems such as Naïve Bayes, Decision Tree, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies.

This paper presents a model that gives light on the fact that the Decision Tree accuracy is far better than KNN in identifying heart disease patients. The rest of the paper is divided as follows: the methodology section explains the proposed methodology for using KNN and the Decision Tree accuracy in diagnosing heart disease.

II.RELATED WORK

The role of psychosocial work stress as a risk factor for chronic disease has been the subject of considerable debate. Many researchers argue in support of a causal connection while others remain skeptical and have argued that the effect on specific health conditions is either negligible or confounded[1].

Some researchers have used data mining, K-means clustering, MAFIA (Maximal Frequent Itemset Algorithm and C4.5 algorithm. The outcome shows that the prediction model designed is capable of predicting the

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

heart attack with good accuracy.The heart disease prediction system was developed using clustering and classification algorithms to predict the effective risk level and accuracy of the patients. In the coming future they have planned to propose an effective disease prediction system to predict the heart disease with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms[2].

Various data mining algorithms such as Aprior, FP-Growth, Naive bayes, ZeroR, OneR, J48 and k-nearest neighbor are applied to study the prediction of heart diseases. On basis of best outcomes the development of heart disease prediction system is done by using hybrid technique that is used for classification associative rules (CARs).The prediction accuracy achieved is 99.19%[3].

III.PROBLEM

In this problem we apply a predictive analysis using the two techniques of classification(KNN and Decision Tree) and the problem lies in deciding which technique gives us the most accurate result. We want to predict the occurrence of heart diseases. We have 12 predictors and 270 instances.

We are provided with the dataset "heart.txt" where the class is taken on the basis of 'coeur' and henceforth the values are predicted. On this basis we further apply the classification techniques:-

a. KNN
b. Decision Tree.

The problem lies in deciding which technique is beneficial and which classification should be used in order to find the accurate result.

Dataset Used

The dataset used for the above problem is heart.txt which was available from the following link:
https://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/heart.txt

age	sexe	type_douleur	pression	cholester	sucre	electro	taux_max	angine	depression	pic	vaisseau	coeur
70	masculin	D	130	322	A	C	109	non	24	2	D	presence
67	feminin	C	115	564	A	C	160	non	16	2	A	absence
57	masculin	B	124	261	A	A	141	non	3	1	A	presence
64	masculin	D	128	263	A	A	105	oui	2	2	B	absence
74	feminin	B	120	269	A	C	121	oui	2	1	B	absence
65	masculin	D	120	177	A	A	140	non	4	1	A	absence
56	masculin	C	130	256	B	C	142	oui	6	2	B	presence
59	masculin	D	110	239	A	C	142	oui	12	2	B	presence
60	masculin	D	140	293	A	C	170	non	12	2	C	presence
63	feminin	D	150	407	A	C	154	non	40	2	D	presence
59	masculin	D	135	234	A	A	161	non	5	2	A	absence
53	masculin	D	142	226	A	C	111	oui	0	1	A	absence
44	masculin	C	140	235	A	C	180	non	0	1	A	absence
61	masculin	A	134	234	A	A	145	non	26	2	C	presence
57	feminin	D	128	303	A	C	159	non	0	1	B	absence
71	feminin	D	112	149	A	A	125	non	16	2	A	absence
46	masculin	D	140	311	A	A	120	oui	18	2	C	presence
53	masculin	D	140	203	B	C	155	oui	31	3	A	presence

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

64	masculin		A	110	211	A	C	144	oui	18	2	A	absence
40	masculin		A	140	199	A	A	178	oui	14	1	A	absence
67	masculin		D	120	229	A	C	129	oui	26	2	C	presence
48	masculin		B	130	245	A	C	180	non	2	2	A	absence
43	masculin		D	115	303	A	A	181	non	12	2	A	absence
47	masculin		D	112	204	A	A	143	non	1	1	A	absence

IV.METHODOLOGY

The two main methods KNN and Decision Tree are used in this paper to predict which gives better results and accuracy.

METHOD 1: KNN

K-NEAREST NEIGHBORS METHOD:

The k-nearest neighbors algorithm (k-NN) is a method of classification of objects. It is easy to understand and here we train the data based on closest training examples in the feature space. KNN is a type of non-parametric method. In machine learning the algorithm KNN is the simplest.Due to irrelevant featuresthe accuracy of this algorithm can be immensely reduced.The advantage of KNN is that it can be used for both classification and regression predictive problems. However, it is popularly used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Output must be interpreted easily
2. Time for calculation should be less
3. Power of prediction should be high.

It is commonly used for its ease of understanding and low calculation time.The experiment is performed on "heart.txt" dataset, where the class is taken on the basis of 'coeur' and henceforth the values are predicted. "HEART" dataset has various columns :sexe,angine,age,pression,cholesterol,depression,pic,taux_max,sucre,electro,vaisseau and type_douleur. The dataset is divided into two parts that is 80% of data is used for training and 20% is used for testing.

Methods for normalization used are:

1. Z score
2. Min-Max
3. Decimal Scaling

METHOD 2: DECISION TREE

Decision tree is a visual representation in form of a graph to represent various choices and their output in form of a tree. The nodes in the graph represent an event or choice and the edges of the graph represent the decision rules or conditions. It is majorly used in the field of Machine Learning and Data Mining applications using R.The important aspects of decision tree approach are:

- The classification of data is done by partitioning attribute space
- It helps to find axis-parallel decision boundaries for specific ideal criteria.
- The classification decisions are represented by the leaf nodes which contain the class labels.
- Keeps splitting nodes based on split criterion, such as GINI index, information gain or entropy
- To avoid overfitting pruning is necessary

The package used to create decision trees in R is "party". We applied the DECISION TREE on "heart.txt" dataset, where the class is taken on the basis of 'coeur' and henceforth the values are predicted. "HEART" dataset has various columns sexe, angine, age, pression, cholesterol, depression, pic, taux_max, sucre, electro, vaisseau and type_douleur.

V.EXPERIMENTAL WORK

PseudoCode:

KNN

Step 1: Data is taken as input.

Step 2: The data is divided into training and testing data which is 80% and 20 % respectively.

Normalization Technique used :Z-score

Step 3: **Z-Score** method is applied for normalization.

Step 4:The column names "sexe","type_douleur","sucre","angine","vaisseau","electro" are made NULL.

Step 5: Install package "class" to apply KNN.

Step 6:Prediction table is made using the command table.

Step 7: To predict the accuracy the sum of the diagonal elements from the table is divided by the sum of all the elements from the predicted table.

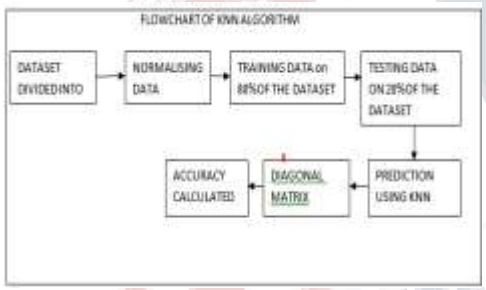
ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

Min-Max
Step 8: **Z-Score** method is applied for normalization.
Step 9: Repeat the steps from 4-7.
Step 10:**Decimal Scaling** method is applied for normalization.
Step 11: Repeat the steps from 4-7.

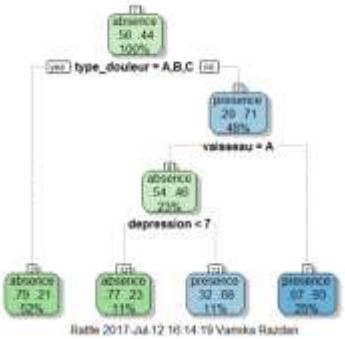
Decision Tree
Step 1: Extracting the data from the given dataset.
Step 2: Use rpart for decision tree and specify the method=”class”.
Step 3: Plotting the decision tree.
Step 4: For a better visualization use fancyRpartPlot available in the package rattle .
Step 4: Use printcp to allocate the points and further use plotcp to specify the minimum line for deciding the pruning factor.
Step 5:UsefancyRpartPlot
Step 6:Mutate is used to give values to “presence” and “absence” as ‘1’ and ‘0’.
Step 7: ROC curve analysis is used for prediction of accuracy.

KNN:



Decision Tree:

DECISION TREE:



VI. RESULT ANALYSIS

KNN Method

We have taken 80% of data as training data and 20% as test data.We see that the accuracy changes with the change in the value of k,as shown in the table below:

Z-Score Approach

S.NO.	Value Of K	Accuracy
1	1	57.40741
2	3	70.37037
3	4	72.22222
4	5	62.96296

Max -Min Approach

S.NO.	Value Of K	Accuracy
1	1	57.40741
2	3	72.22222
3	4	68.51852
4	5	75.92593

Decimal Scaling

S.NO.	Value Of K	Accuracy
1	1	59.25926
2	3	57.40741
3	4	62.96296
4	5	68.51852

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

Decision Tree Method:
Area under the curve:

- The accuracy of the decision tree is 0.877840112201964
- The percentage of accuracy of the decision tree is 87.784011220%.

Results obtained from the experiment

Method Used	Accuracy
KNN-Min-Max	75.9%
Decision Tree	87.7%
KNN-Z-Score	72.2%
KNN -Decimal Scaling	68.5%

VII.CONCLUSION

After applying both the methods we find that accuracy is more in decision tree rather than KNN. We conclude that to find accuracy on a dataset decision tree is more accurate than KNN.Further in KNN when we used Z-Score , Min-Max and Decimal Scaling for normalization .The accuracy in KNN is most with Min-Max approach and least with Min-Max approach.
Data Mining has gained popularity in almost all applications of real world and is expected to become more popular due to its applications and various advantages it offers. One of the data mining techniques i.e., classification is an interesting as well as an important topic to the researchers as it is highly efficient and accurately classifies the data for knowledge discovery. Decision trees have gained popularity because they produce human readable classification rules and are easy to interpret as compared to other classification methods.The most commonly used decision tree classifiers are examined and the experiments are conducted to find the most accurate classifier for Medical Diagnosis.

REFERENCES

[1]MikaKivimaki and Ichiro Kawachi,Work Stress as a Risk Factor for Cardiovascular Disease.

[2] Heart Disease Analysis System Using Data Mining Techniques by G.Karthiga1, C.Preethi1 and R.DelshiHowsalya Devi.

[3]JagdeepSingh,AmitKamra,HarbhagSingh"Prediction of heart diseases using associative classification".Published in 2016 5th International Conference on Wireless Networks and Embedded Systems.

[4]Pignone M, Phillips C, Mulrow C,Use of lipid lowering drugs for primary prevention of coronary heart disease: meta-analysis of randomised trials. BMJ 2000; 321: 1–5.

[5]SudlowCL,Warlow CP. Comparable studies of the incidence of stroke and its pathological types: results from an international collaboration. International Stroke Incidence Collaboration. Stroke 1997; 28: 491–9.

[6]Brotman DJ, Golden SH, Wittstein IS. The cardiovascular toll of stress. Lancet. 2007;370:1089–100. doi: 10.1016/S0140-6736(07)61305-1.

[7]Steptoe A, Kivimäki M. Stress and cardiovascular disease. Nat Rev Cardiol. 2012;9:360–70. doi: 10.1038/nrcardio.2012.45.

[8]Landsbergis PA, Dobson M, Koutsouras G, Schnall P. Job strain and ambulatory blood pressure: a meta-analysis and systematic review. Am J Public Health. 2013;103:e61–71. doi: 10.2105/AJPH.2012.301153.

[9] Kivimäki M, IPD-Work. Long working hours and risk of coronary heart disease and stroke: a meta-analysis of 603,838 men and women. Lancet. 2015 [in press].

[10]Hemingway H, Marmot M. Evidence based cardiology: psychosocial factors in the aetiology and prognosis of coronary heart disease. Systematic review of prospective cohort studies. BMJ. 1999;318:1460–7. doi:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

10.1136/bmj.318.7196.1460.

[11]Ferrie JE. Is job insecurity harmful to health? J R Soc Med. 2001;94:71–6

[12]Law MR, Wald NH, Wu T, HackshawA,BaileyA.Systematic underestimation of association between serum cholesterol concentration and ischaemic heart disease in observational studies: data from the BUPA study. BMJ 1994; 308: 363–6.

APPENDIX:

```
#Decision Tree
donees<-read.table(file="heart.txt",dec="
",header=TRUE)
summary(donees)
arbre.full<-rpart(coeur~ .,data=donees,method="class")
print(arbre.full)

library(rpart)
library(rpart.plot)
plot(arbre.full,margin=0.1,main="Occurence of heart
diseases")
text(arbre.full, use.n=TRUE, all=TRUE, cex=.7)
library(rattle)
fancyRpartPlot(arbre.full)
printcp(arbre.full)
plotcp(arbre.full, minline = TRUE)
arbre.full1<-prune(arbre.full,cp= 0.047)
fancyRpartPlot(arbre.full1)
arbre.full1
#Predicting accuracy
donees%>%mutate(Target=ifelse(coeur=='presence',1,0))
->donees
donees%>%select(-coeur)->donees
#Confusion Matrix
actual<-donees$Target
predicted<-predict(arbre.full1,type = "class")

head(predicted)
head(as.numeric(predicted))
predicted<-as.numeric(predicted)
predicted<-ifelse(predicted==2,1,0)
confusionMatrix(predicted,actual,positive="1")

#kappa metric
kappa2(data.frame(actual,predicted))

#ROC curve analysis
library(ROCR)
pred<-prediction(actual,predicted)
perf<-performance(pred,"tpr","fpr")
plot(perf,col="red")
abline(0,1, lty = 8, col = "grey")

auc<-performance(pred,"auc")
unlist(auc@y.values)
paste("The accuracy of the decision tree is", a)


#KNN
r1<-read.table(file="heart.txt",dec=" ",header=TRUE)
summary(r1)
View(r1)
r1$sexe<-NULL
r1$type_douleur<-NULL
r1$sucre<-NULL
r1$angine<-NULL
r1$vaisseau<-NULL
r1$electro<-NULL
#Z-Score
r1NormZ <- as.data.frame( scale(r1[1:6] ))
r1NormZ
library("class")
View(r1)
row_index=sample(1:nrow(r1),0.8*nrow(r1))
train_data_i=r1NormZ[row_index,]
test_data_i=r1NormZ[-row_index,]
pred_model=knn(train_data_i,test_data_i,r1[row_index,7]
,k=3)
pred_model
predicted_table=table(pred_model,r1[-row_index,7])
predicted_table
a=0.2*nrow(r1)
diag_element=diag(predicted_table)
diag_element
accuracy_r1=(sum(diag_element)/a)*100
accuracy_r1
#Min-Max
row_index=sample(1:nrow(r1),0.8*nrow(r1))
d=function(x){((x-min(x))/(max(x)-min(x)))}
norm_data=lapply(r1[,-7],d)
class(norm_data)
norm_data=as.data.frame(norm_data)
train_data_i=norm_data[row_index,]
test_data_i=norm_data[-row_index,]
pred_model=knn(train_data_i,test_data_i,r1[row_index,7]
,k=5)
pred_model
predicted_table=table(pred_model,r1[-row_index,7])
predicted_table
```

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 8, August 2017

```
a=0.2*nrow(r1)
diag_element=diag(predicted_table)
diag_element
accuracy_r1=(sum(diag_element)/a)*100
accuracy_r1
#   Decimal Scaling
install.packages("dprep")
library(dprep)
row_index=sample(1:nrow(r1),0.8*nrow(r1))
decimal<-as.data.frame( decscale(r1[1:6] ))
train_data_i=decimal[row_index,]
test_data_i=decimal[-row_index,]
pred_model=knn(train_data_i,test_data_i,r1[row_index,7]
,k=1)
pred_model
predicted_table=table(pred_model,r1[-row_index,7])
predicted_table
a=0.2*nrow(r1)
diag_element=diag(predicted_table)
diag_element
accuracy_r1=(sum(diag_element)/a)*100
accuracy_r1
```