Heart Disease Prediction using Data Mining Techniques

Anand Sharma
dept of Computer Science and Engineering
Indraprashtra Institute of Information Technology Delhi
Delhi, India
anand19059@iiitd.ac.in

***Abstract*—In this Project we have gathered database from 3 different countries Hungarian,Long beach Va,Switzerland hospitals. We have Preprocessed data cleaned it ,remove dummy and not usesful data . 5 Data Mining Techniques is applied on the preprocessed data Logistics Regressions , Decision Tree Clasifier , KNN, Support Vector Machines,Naive Bayes and obtained the results . I have obtained highest accuracy on Decision Tree Classifier . By this Project we can predict the heart disease in future**

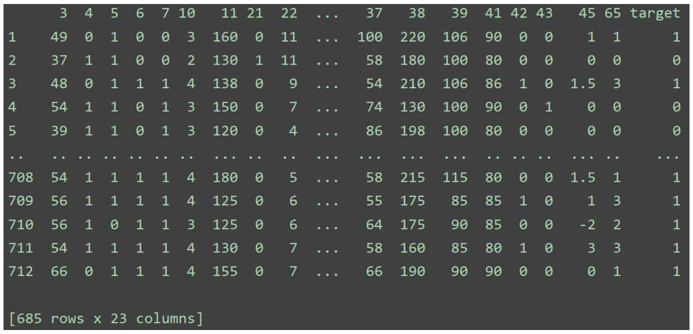
***Index Terms*—Heart Disease , Data Mining , Classification,Prediction**

I. Introduction

According to World Health Organisation more than 12 million death occour due to cardiovascular disease .Data Mining is a technique which is performed on large and wide complex database to obtain some meaningful contexts from it . Several Statistical ,graphical, and machine learning , database technique are used [6]. Data Mining is extremely useful in Medical Fields . So I have used the power of data mining to improve life of people by identifying the possiblity of Heart Attack before the occurrence . In this project I have applied several data mining techniques such as Logistics Regressions , Decision Tree Classifier , KNN, Support Vector Machines,Naive Bayes and obtained the accuracy [5]. Initally i have taken dataset of 3 differnt country and combined it .Then a proper ananlysis is done by checking the attribute deependence and relations of attributes with class . Finally i have obtained the maximum accuracy of approx 94

II. Dataset Description

I have taken the Dataset of Heart Disease from website https://archive.ics.uci.edu/ml/index.php . It was originally distributed data on basis of country .Originally data is from 1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D. 2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. 3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D . 4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation I have Merged datset of 3 countries . It contain total 714 rows and 75 columns . Missing Value are represented as '-9' and name as dummy name in dataset .



```
      3  4  5  6  7 10   11 21  22  ...   37   38   39  41 42 43   45 65  target
1    49  0  1  0  0  3  160  0  11  ...  100  220  106  90  0  0    1  1       1
2    37  1  1  0  0  2  130  1  11  ...   58  180  100  80  0  0    0  0       0
3    48  0  1  1  1  4  138  0   9  ...   54  210  106  86  1  0  1.5  3       1
4    54  1  1  0  1  3  150  0   7  ...   74  130  100  90  0  1    0  0       0
5    39  1  1  0  1  3  120  0   4  ...   86  198  100  80  0  0    0  0       0
..   .. .. .. .. .. ..  ... ..  ..  ...  ...  ...  ...  .. .. ..  ... ..     ...
708  54  1  1  1  1  4  180  0   5  ...   58  215  115  80  0  0  1.5  1       1
709  56  1  1  1  1  4  125  0   6  ...   55  175   85  85  1  0    1  3       1
710  56  1  0  1  1  3  125  0   6  ...   64  175   90  85  0  0   -2  2       1
711  54  1  1  1  1  4  130  0   7  ...   58  160   85  80  1  0    3  3       1
712  66  0  1  1  1  4  155  0   7  ...   66  190   90  90  0  0    0  1       1

[685 rows x 23 columns]
```

Fig. 1. Example of a figure caption.

III. Data Preprocessing

Originally data was in .data file . I have read the 3 .data file of each country separately . Convereted that data file to string . Removed the spaces in string and divide whole string data to seperate rows and columns according to info. dropped the spaces in between and created a dataframe for 3 seperate datset . Finally combined the 3 dataset to make one dataframe for further processing. Rows that don't contain any value are removed ,columns with more than 50% null values are removed . From remaining dataset Replaced nan ,?,None,-9.0,-9,NULL with the mean value of that attribute column . Removed some column which was added during preprocessing from data file dropped that column . Removed the noise values from the remaining dataset by checking any out of range values. converted the class from 0,1,2,3,4 to 0,1 .simply we have made it a classification problem .Plotted various graphs between different attribute to visualise data .I have plotted the correlation matrix and checked whether any attribute has a high correlation . In our data i do not found major correlated attribute . Finally after all preprocessing steps i am left with 685 rows and 23 columns .I have created a heart dataset.csv file of the obtained preprocessed data .

IV. Classification

Data has been splitted by using train test split in 80 is to 20 ratio for training and testing . Now we will use 80% for training and 20% for testing There are several Data Mining classifiers availabe to us by using sklearn library.I have used Logistics Regressions , Decision Tree Classifier ,

Identify applicable funding agency here. If none, delete this.

KNN, Support Vector Machines,Naive Bayes classifiers and compared the accuracy obtained from it.

A. Logistics Regression

Logistics Regression [4] is statistical method for predicting binary classes . It is an extension of Linear Regression which is used on classification problem . It follows Bernoulli Distribution.Estimation is done through maximum likelihood of the attributes . By using Logistsics Regression Classifiers We have obtained the accuracy of 83.96%

B. Decision Tree

Decision Tree (DT) [1] [3] is a simple classifier. Decision tree builds classification or regression models in the structure of a tree based on different depths making it simple to handle the classifiction problem. Decisions trees can be used for both categorical and numerical data. The algorithm works by finding the information gain of the attributes and taking out the attributes for splitting the branches in tree. I have applied Decision Tree Classifiers by taking different depths and checked thee accuracy score and plotted it . After the depht of 14 i can clearly see no change in accuracy . The maximum accuracy obtained from it was 95%

C. Naive Bayes

Naïve Bayes (NB) [2] is a statistical classifier which assumes no dependence among various attributes.it will treat each attribute as independent and apply Bayes Theorem for classification problem Using Naive Bayes maximum acccuracy of 83.21% is obtained.

D. KNN

K Nearest Neighbour [3] is classifier which predict the class into 0 or 1 by the choice of the k nearest neighbour. It chooses k neighbour points and based on that it gives the prediction of the classes . In this project we have applied the KNN with different values of Neighbour and plotted the accuracy . We have obtained the maximum accuracy of 66% . We are not getting good accuracy with KNN

E. Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning model which is used for classification .This algorithm create a optimal hyperplane which is used to categorize new data. It is used to draw hyperplane for multi dimensional space .I have applied the SVM classifier with kernel =linear and obtained the accuracy of 86.13%

V. Results

The results of accuracy of various classifiers are plotted on the preprocessed dataset . As we can see Decision Tree classifier has given us the maximum accuracy among all the classifiers ie 95%. Logistics Regressions and SVM has performed same with respect to accuracy ie 86% . KNN has failed to fit the dataset as acuracy is too low ie 64%. as we are increasing the neighbour values still accuracy is not increasing . it doesn't work that well on our problem. Accuracy of

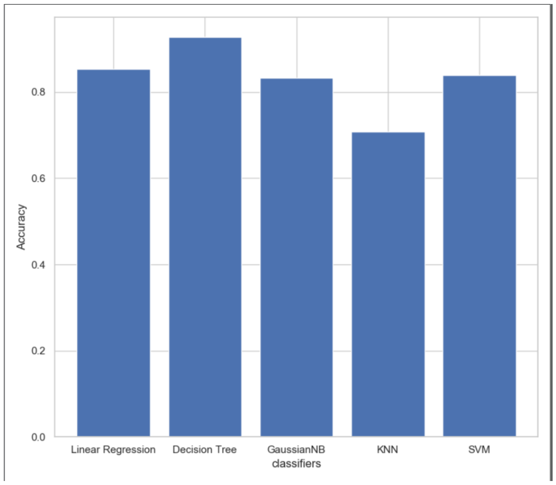


Fig. 2. Accuracy of Different Classifiers.

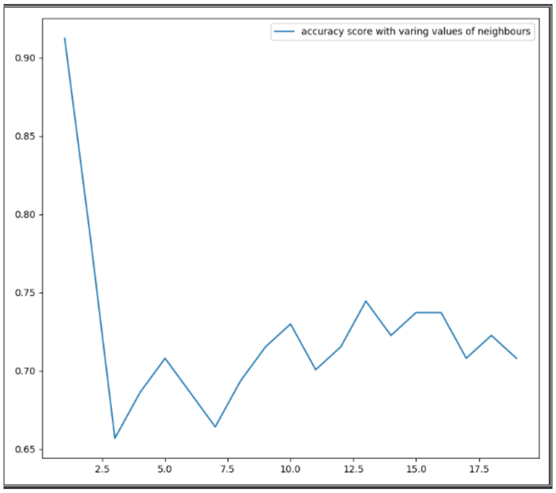


Fig. 3. Accuracy of KNN with different values of Neighbour.

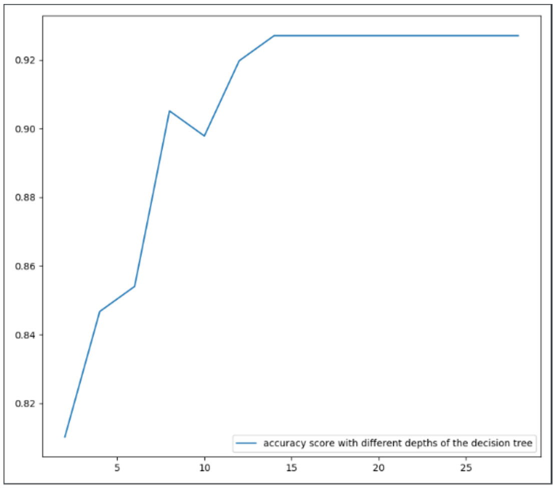


Fig. 4. Accuracy of Decision Tree Classifier with different depths.

Gaussian NB was moderate as it gives more than 80% As we can see In Decision Tree classifier after increasing the depth above 15 accuracy has been constant no fluctations. Initally with increase in depth our accuracy is increasing In KNN we are getting approx 64% accuracy with different neighbour values

VI. References

References

[1] J. Ross Quinlan, “Induction of Decision Trees”, Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986.

[2] K. Ming Leung, “Naive Bayesian Classifier”, Master Thesis,Department of Computer Science and Engineering,Polytechnic University, 2007.

[3] Sayali D. Jadhav, H. P. Channe, “Comparative Study of KNN, Naive Bayes and Decision Tree Classification Techniques”, International Journal of Science and Research (IJSR) , Volume 5, Issue 1 , Paper ID: NOV153131,2016.

[4] Johan Holmgren, Sebastian Aspegren, Jonas Dahlström, “Prediction of bicycle counter data using regression”, Procedia Computer Science 113,pp. 502–507,2017

[5] SellappanPalaniappan, RafiahAwang “Intelligent Heart Disease Prediction System Using Data Mining Techniques” IEEE, pp.978-1-4244-1968,2008 .

[6] Mustafa A. Al-Fayoumi, “Enhanced Associative classification based on incremental mining Algorithm (E-ACIM)”, IJCSI International Journal of Computer Science Issues, Volume 12, Issue 1, 2015.