IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HEART DISEASE PREDICTION USING DATA MINING

[1]U. Sairam [2]Santhosh Voruganti,

[1, 2] Assistant Professor, Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, India-500075

Abstract: Medical services provide gigantic information on every day ground having diverse structures like printed ,images, numbers pool and so forth. However, there is absence of devices accessible in healthcare to process this data. Data mining frame works are utilized to extricate information from this data which can be utilized by media proficient individual to figure future procedures. Heart illness is the primary driver of death in the masses. Early recognizing and hazard expectations are essential for patient's medicines and specialists' analysis.

Data mining algorithms like Decision trees (J48), Bayesian classifiers, Multilayer preceptor, Simple logistic and Ensemble techniques are utilized to determine the heart ailments. In this work, different data mining classification procedures are analyzed for testing their precision and execution on preparing medicinal informational index. The classification results will be envisioned by various representation procedures like 2D diagrams, pie graphs, and different techniques. The beforehand mentioned calculations are analyzed and assessed based on their exactness, time utilization factor, territory under ROC and so on.

Index Terms **- Data Mining, Hear Illness, Classifier.**

I. INTRODUCTION

A. Overview

Heart ailments are one of the significant reasons of death and disability on the planet effecting 17.5 million individuals every year and more than twenty-three million anticipated passing from cardiovascular sickness by 2030. Coronary illness incorporates different sorts of conditions that can influence center reason. The heart is an important organ of human body. On the off chance that the blood dissemination to the body is lacking, the organs of the body that is cerebrum and heart quit working and passing happens in couple of minutes. The peril factors related are distinguished as age, family history, diabetes, hypertension, elevated cholesterol, tobacco, smoke, liquor inward breath, heftiness, physical idleness, chest torment write and less than stellar eating routine . Medical industry is data rich yet learning poor. There is requirement for a wise emotionally supportive network for ailment forecast. Data mining strategies like Classification, regression is utilized to anticipate the infection[15]. With the advancements of computing facility gave by software engineering innovation, it is currently conceivable to anticipate many states of infirmities more accurately. The objective of this work is to analyze the potential utilization of classification[14] based data mining techniques like naive bayes, decision tree(j48), ensemble algorithms and simple logistic and so forth.

B. Methodologies

Data mining is a cognitive procedure of discovering the hidden approach patterns from large data set. It is generally utilized for applications, for example, financial data ,analytic thinking, retail, media transmission industry, genome data analysis, logical applications and health mind frameworks and so on. Data mining holds extraordinary potential to improve heath frameworks by utilizing data and analytics to recognize the accepted procedures that enhance care and reduce cost. WEKA is a effective tool as it contains both supervised and unsupervised learning techniques[14]. We utilize WEKA because it causes us to evaluate and compare data mining techniques (like Classification, Clustering, and Regression etc.) conveniently on real data.

C. Problem Statement

II. The rate of heart diseases is increasing at an exponential rate. The busy lifestyle of people in this era with all the fast food in the lunch break and getting back to sitting and working has pushed as over the edge. Along with this people today have a lack of exercise and are less active. For most of them recreation is just another movie in bed or anything technology based. Physical activities have reduced drastically. These factors boosted the rate of heart diseases to an unfortunately high percentage.

II. LITERATURE SURVEY

Various work has been improved the situation disease forecast concentrating on heart illness utilizing different data mining systems. Authors have connected distinctive data mining techniques like decision trees, KNN, support vector machine, neural network that contrast in their accuracy, execution time.

Mr. Chintan Shah , clarifies dialog of different classification algorithms in view of specific parameters like time taken to build the model, accurately and inaccurately classified instances and so on. Theresa Princy. R. [2] proposed a framework to precisely foresee heart disease utilizing ID3 and KNN classifiers and accuracy level also provided for different number of attributes.

Finding of coronary illness with the assistance of Bayesian Network calculation has been characterized by Xue et al [3]. Abraham proposed a methodology so as to increase classification accuracy of medical data based on Naive Bayes classifier algorithm [4]. Palaniappan & Awang [5] recommended a model of IHDPS (Intelligent Heart Disease Prediction System) actualizing data mining calculations, like Naive-Bayes, Decision Trees and Neural Network. The last yield of these algorithm depicts that every strategy has its distinctive capacities in the reason for the portrayed mining objectives.

Jagdeep Singh implemented different association and classification methods on the heart datasets to foresee the heart illness. The association algorithm like Apriori and FPGrowth are used to discover association rules of heart dataset attributes[6].

In [7], diverse machine learning systems including Decision Tree (DT), Naive Bayes (NB), Multilayer Perceptron (MLP), K-Nearest Neighbor (K-NN), Single Conjunctive Rule Learner (SCRL), Radial Basis Function (RBF) and Support Vector Machine (SVM) have been applied, individually and in combination, using ensemble machine learning approaches on the Cleveland Heart Disease data set keeping in mind the end goal to analyze the execution of every strategy. Gudadhe et al. [8] realized a design base with both the MLP network and the SVM approach. This design accomplished an accuracy of 80.41% in terms of the classification between two classes (the presence or absence of heart disease, respectively).

Author in [9] assesses the disease categorization using three different machine learning calculations by WEKA Tool. We compare the results in terms of time taken to build the model and its accuracy. This work demonstrate the Random Forest is best classifier for disease categorization of WEKA tool because it runs efficiently on large datasets.

In paper [10], author applied HNB classifier for analysis of coronary illness tested execution for heart stalog data collection. Experimental result demonstrates that HNB model exhibits a predominant execution compared with other Approaches. Proposed approach applies discretization and IQR filters to enhance the efficiency of Hidden naïve bayes.

Authors in [12] executed the framework that extracts hidden knowledge from a historical heart disease database. Mamta Sharma[13] uncovers that the Neural Networks with 15 attributes shows significant results over all other data mining techniques. Decision Tree methods has proven excellent precision with C4.5, ID3, CART and J48.

III. PROPOSED METHODOLOGY

A.Existing Solutions

The Artificial Neural Network (ANN), also a supervised learning strategy, contains three layers: input, hidden and output. The connection between the input units and the hidden and the output units are based on relevance of the assigned weight of that specific input unit. Usually, if the weight is higher, then it is considered more important. ANN may use linear and sigmoid transfer (activation) functions. Also, the ANNs are suitable for the training of large amounts of data with limited inputs. For multi-layer feed forward ANN, the mostly used learning algorithm see Fig 1 is the Backpropagation learning tool [4], [5]. In ANN, the input data records should be separated into three sub-datasets for the purpose of training, validation and testing.

B. Flow chart of the proposed methodology

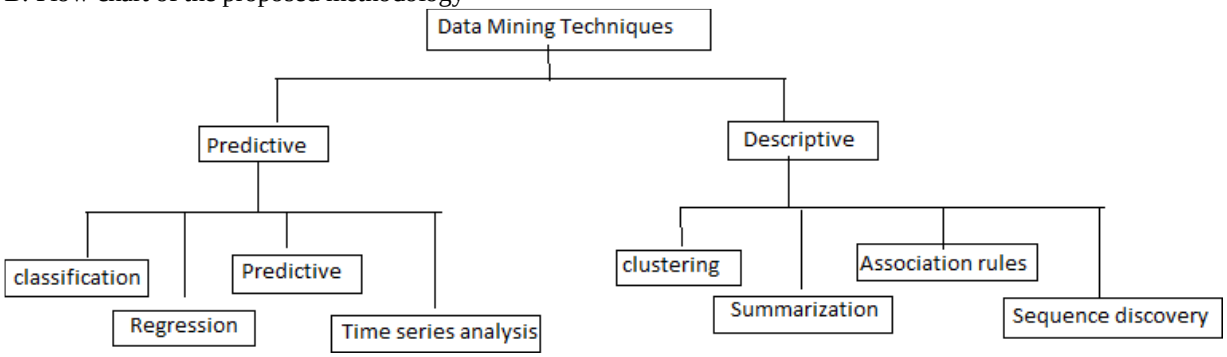
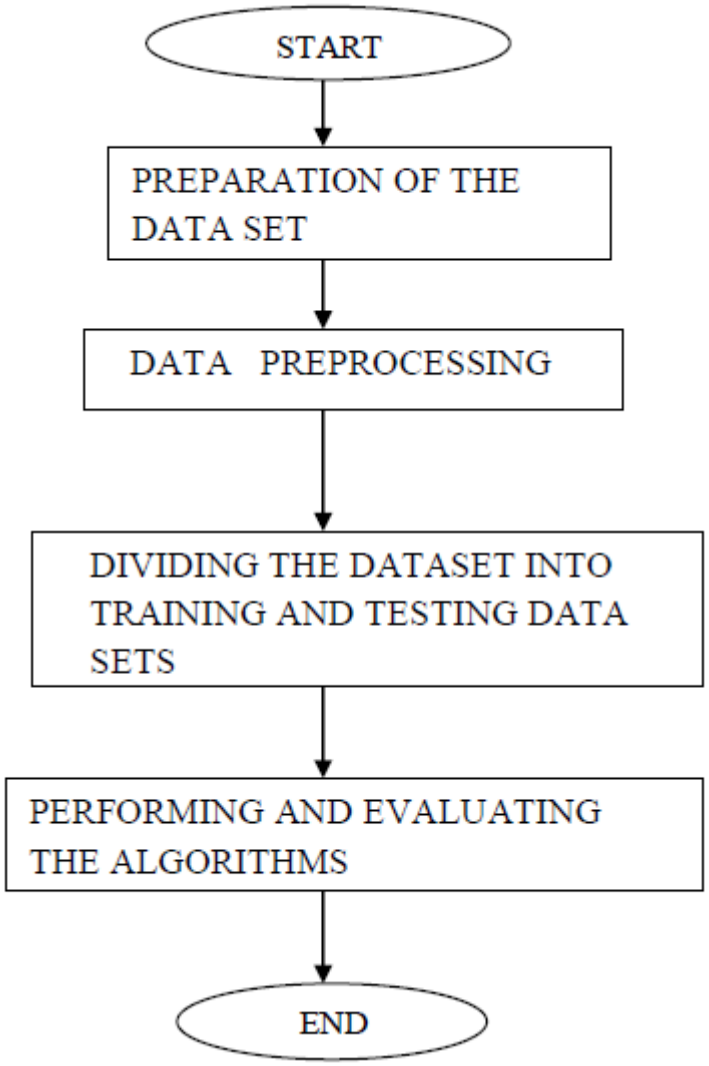


Fig 2 Taxonomy of data mining techniques



The above figure 1 indicates the work flow of the project where operations are performed in a step-by-step manner starting from dataset loading followed by pre-processing, partitioning and evaluation of algorithms.

IV. IMPLEMENTATION

The various modules in this project are as follows:
1) Importing required packages , algorithms and functions
2) Data set loading and preprocessing
3) Implementation of algorithm

A. Data Set Loading and Preprocessing

```
names = ["age", "sex", "cp", "trestbps", "chol", "fbs", "restecg", "thalach", "exang", “Oldpeak,” “Slope”, “ca”, “thal”, “target”
df = pd.read_csv('C:/Users/K SHIVARAM TEJA/Desktop/h.csv', names=names, header=None, na_values='')
```

In the above lines of code it is observed that all the fourteen attributes in the dataset are given different names and in the next step the dataset is loaded using the pandas package read_csv is the function required .

```
from sklearn.preprocessing import StandardScaler
x_std = StandardScaler().fit_transform(x)
```

The first line imports the standardization scaler which is used for data pre processing and then the imported function is used for preprocessing

B. Implementation of Algorithms

Adaboost:

```
x = df.iloc[:,:-1]
y = df.iloc[:,-1]
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.8,random_state=0)
clf = AdaBoostClassifier()
clf.fit(x_train,y_train)
prediction = clf.predict(x_test)
accuracy = accuracy_score(prediction,y_test)
cm = confusion_matrix(prediction,y_test)
prfs = precision_recall_fscore_support(prediction,y_test)
print('Accuracy: ',accuracy)
```

```
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall: ', prfs[1])
print('Fscore: ', prfs[2])
print('Support: ', prfs[3]
```

firstly the data set is partitioned as 80% training data and 20% testing data. The above lines of code are for finding the testing data accuracy along with precision, recall, fscore and support. A confusion matrix is also obtained.

```
x = df.iloc[:,:-1]
y = df.iloc[:,-1]
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.80,random_state=0)
clf = AdaBoostClassifier()
clf.fit(x_train,y_train)
prediction = clf.predict(x_train)
accuracy = accuracy_score(prediction,y_train)
cm = confusion_matrix(prediction,y_train)
prfs = precision_recall_fscore_support(prediction,y_train)
print('Accuracy: ',accuracy)
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall: ', prfs[1])
print('Fscore: ', prfs[2])
print('Support: ', prfs[3])
```

```
clf = AdaBoostClassifier()
clf.fit(x_train,y_train)
prediction = clf.predict(x_test)
accuracy = accuracy_score(prediction,y_test)
cm = confusion_matrix(prediction,y_test)
prfs = precision_recall_fscore_support(prediction,y_test)
print('Accuracy: ',accuracy)
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall:    ', prfs[1])
print('Fscore:    ', prfs[2])
print('Support:   ', prfs[3])
```

```
Accuracy:  0.75

Confusion Matrix:  [[25 10]
 [ 5 20]]

Precision:  [0.83333333 0.66666667]
Recall:     [0.71428571 0.8        ]
Fscore:     [0.76923077 0.72727273]
Support:    [35 25]
```

Fig 3 Testing accuracy of adaboost algorithm

By using testing data, we have achieved accuracy of 75% and remaining parameters precision, recall, fscore, support are also calculated and displayed. On contrary to testing accuracy training accuracy is much higher with 93% and other 23 parameters precision, fscore, recall, support are calculated . Training data achieved higher accuracy because we have 80% of training data which has more continuous and centered values than 20 % of testing data.

The above figure 5 shows the testing accuracy for logistic regression which is far better than testing accuracy of adaboost and the other parameters precision, recall, fscore and support are also calculated a confusion matrix is also obtained.
Testing accuracy of logistic regression is 81.6%
The above figure 6 is the training accuracy for logistic regression which is better than tesing accuracy of logistic regression but less than training accuracy of adaboost algorithm. Training accuracy is 87.34% and the other parameters precision, recall, fscore and support also calculated.

```
x = df.iloc[:,:-1]
y = df.iloc[:,-1]
x_train,x_test,y_train,y_test = train_test_split(x,y,train_size=0.80,random_state=0)
clf = AdaBoostClassifier()
clf.fit(x_train,y_train)
prediction = clf.predict(x_train)
accuracy = accuracy_score(prediction,y_train)
cm = confusion_matrix(prediction,y_train)
prfs = precision_recall_fscore_support(prediction,y_train)
print('Accuracy: ',accuracy)
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall:    ', prfs[1])
print('Fscore:    ', prfs[2])
print('Support:   ', prfs[3])
```

```
Accuracy:  0.9451476793248945


Confusion Matrix:  [[125   8]
 [  5  99]]


Precision:  [0.96153846 0.92523364]
Recall:     [0.93984962 0.95192308]
Fscore:     [0.95057034 0.93838863]
Support:    [133 104]
```

Fig 4 training accuracy for adaboost algorithm

```
clf = LogisticRegression()
clf.fit(x_train,y_train)
prediction = clf.predict(x_test)
accuracy = accuracy_score(prediction,y_test)
cm = confusion_matrix(prediction,y_test)
prfs = precision_recall_fscore_support(prediction,y_test)
print('Accuracy: ',accuracy)
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall:    ', prfs[1])
print('Fscore:    ', prfs[2])
print('Support:   ', prfs[3])
```

```
Accuracy:  0.8166666666666667


Confusion Matrix:  [[27  8]
 [ 3 22]]


Precision:  [0.9        0.73333333]
Recall:     [0.77142857 0.88      ]
Fscore:     [0.83076923 0.8       ]
Support:    [35 25]
```

Fig 5 testing accuracy for logistic regression

```
clf = LogisticRegression()
clf.fit(x_train,y_train)
prediction = clf.predict(x_train)
accuracy = accuracy_score(prediction,y_train)
cm = confusion_matrix(prediction,y_train)
prfs = precision_recall_fscore_support(prediction,y_train)
print('Accuracy: ',accuracy)
print('\n')
print('Confusion Matrix: ',cm)
print('\n')
print('Precision: ', prfs[0])
print('Recall:    ', prfs[1])
print('Fscore:    ', prfs[2])
print('Support:   ', prfs[3])
```

```
Accuracy:  0.8734177215189873


Confusion Matrix:  [[122  22]
 [  8  85]]


Precision:  [0.93846154 0.79439252]
Recall:     [0.84722222 0.91397849]
Fscore:     [0.89051095 0.85      ]
Support:    [144  93]
```

Fig 6 training accuracy of logistic regression

The above figure 6 is the training accuracy for logistic regression which is better than tesing accuracy of logistic regression but less than training accuracy of adaboost algorithm. Training accuracy is 87.34% and the other parameters precision, recall, fscore and support also calculated.

```
X = array[:,0:8]
Y = array[:,8]
seed = 7
num_trees = 100
kfold = model_selection.KFold(n_splits=10, random_state=seed)
model = GradientBoostingClassifier(n_estimators=num_trees, random_state=seed)
results = model_selection.cross_val_score(model, X, Y, cv=kfold)
print(results.mean())
```

```
0.6973563218390806
```

Fig 7 Boosting ensemble

ALGORITHM	ACCURACY
Adaboost	93.44
Logistic Regression	87.48
Bagging Ensemble	81.33
Boosting Ensemble	69.73
Voting Ensemble	71.06

Table 1 Results showing accuracy of each algorithm

C. Data visualisation

Describing the dataset

```
[13]: df.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	299.000000	301.(
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.672241	4.7
std	9.038662	0.467299	0.960126	17.599748	51.776918	0.356198	0.994971	22.875003	0.469794	1.161075	0.616226	0.937438	1.9
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.(
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	3.(
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000	3.(
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.(
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.(

Fig 8 describing data set

The above figure describes the dataset and various parameters like count, standard deviation, variance and so on.

Heat map of null values

Fig 9 heat map of null values

The x-axis is the attributes in the data set and the y-axis is the count. There are a total of 14 attributes including categorical and numerical attributes

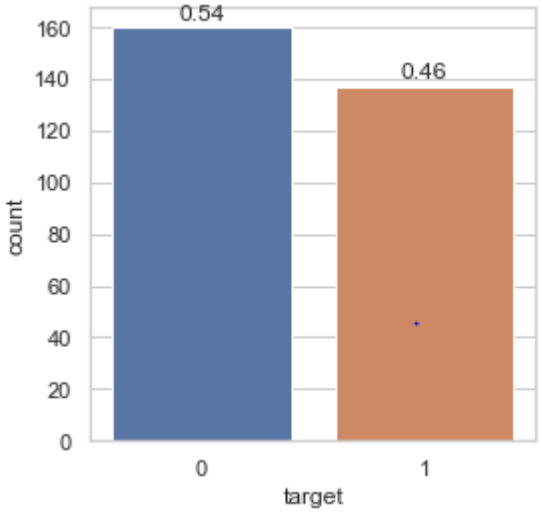


Fig 10 target plot

Target plot gives the count of number of male and female in the data . It can be interpreted that female are more than male with blue line indicating female and orange line indicating male.

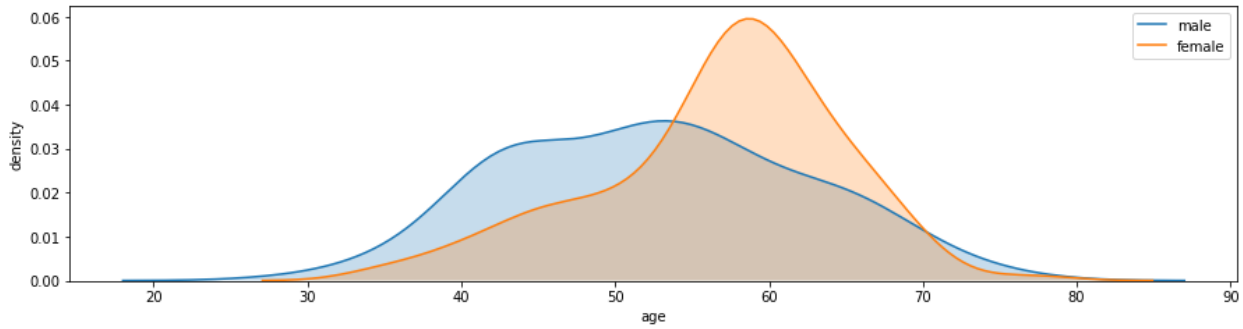


Fig 11 Plot of Age comparing male and female

The above plot shows the probability of male and female getting heart disease within different ages. The density is high from 50 to 70 years age.

In this heart disease data set, there are total 14 attributes out of which 8 are categorical and 6 are numerical so there are different plots for numerical as well as categorical. The below plots are for categorical attributes.

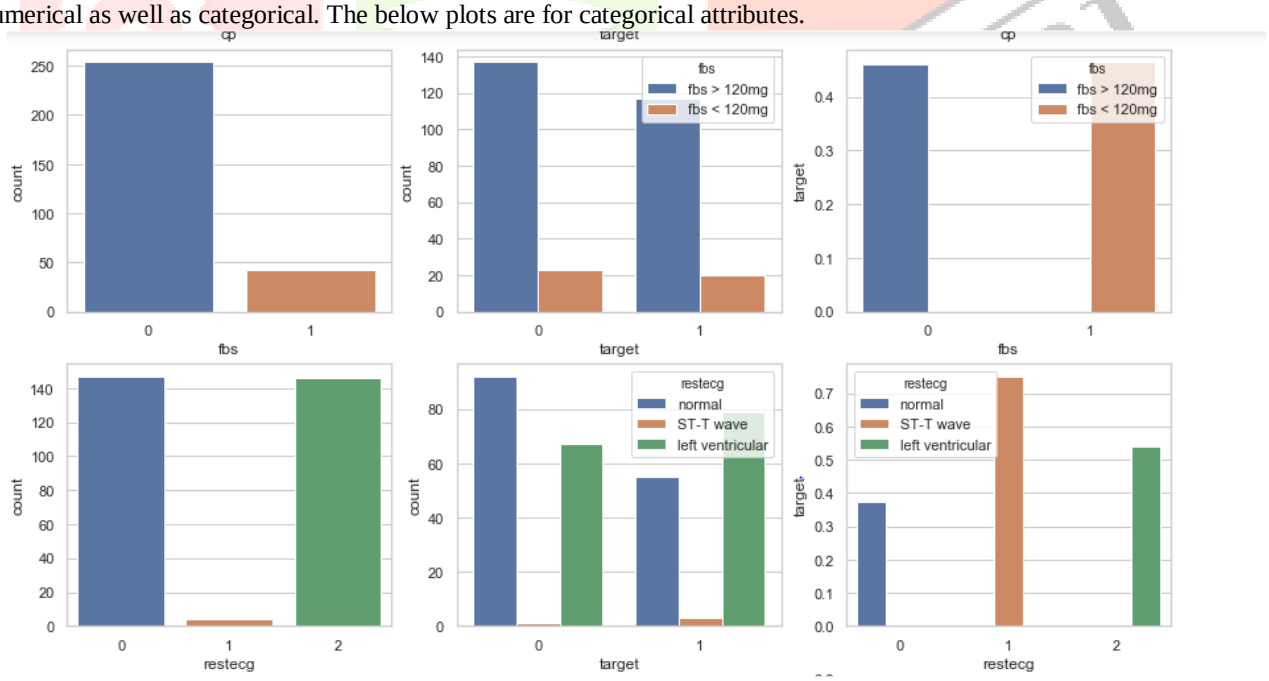


Fig 12 Plot of chest pain type and fasting blood sugar

The above plot is the categorical attributes plot and the categorical attributes are chest pain type and fasting blood sugar. The plot describes the count of male and female in three plots but the range is different for each plot x- axis is target and y-axis count.

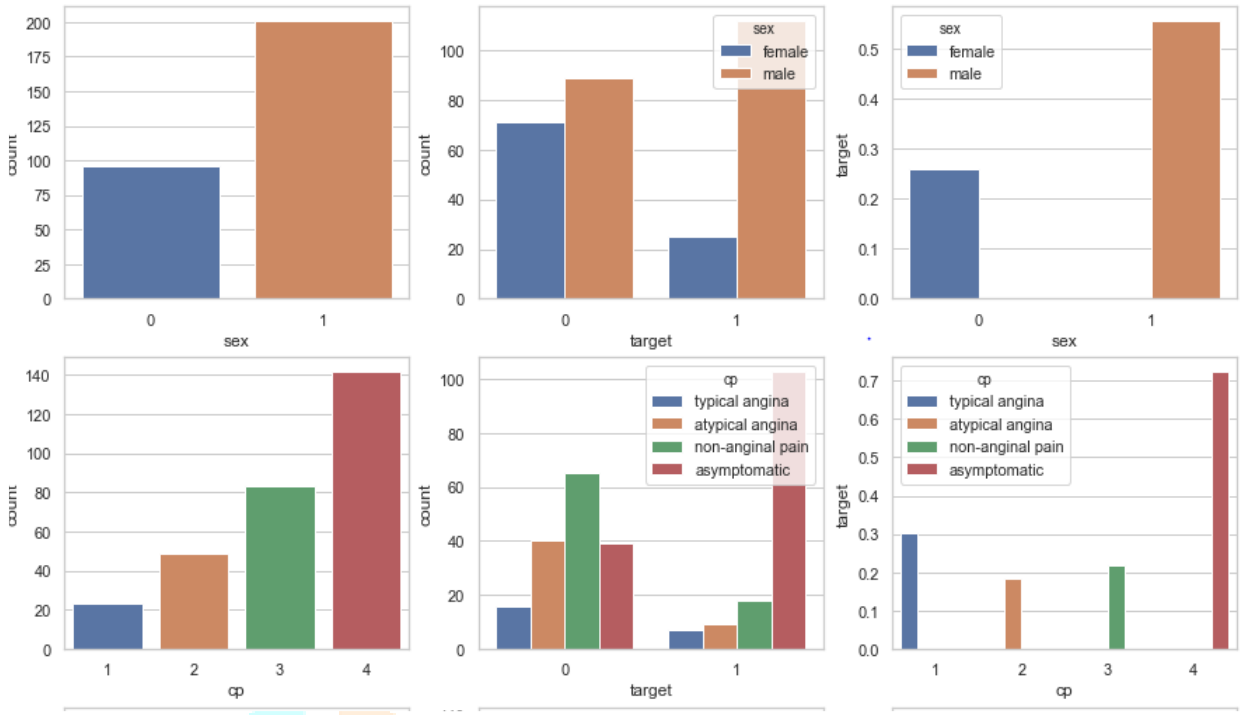


Fig 13 plot of sex and Exercise induced

The above plot is the categorical attributes plot and the categorical attributes are sex and Exercise induced. . The plot describes the count of male and female in three plots but the range is different for each plot x- axis is target and y-axis count.

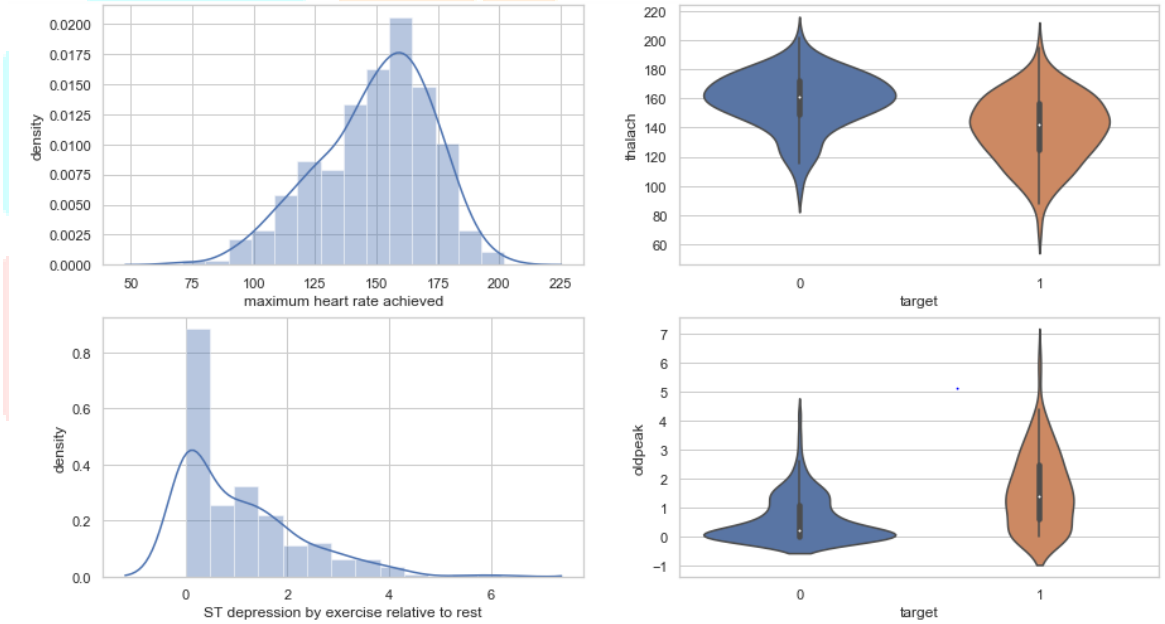


Fig 14 Plot of Maximum heart rate and depression

V. CONCLUSIONS

Heart Disease is a fatal disease by its nature. This disease makes a life-threatening complexity, for example heart attack and death. The significance of Data Mining in the Medical Domain is acknowledged and steps are taken to apply relevant techniques in the Disease Prediction. The various research works with some effective techniques done by different people were studied.

A New technique of predicting heart disease is developed which resulted in better accuracy than the existing works. In this work adaboost algorithm has the greater accuracy than logistic regression for training data and ensemble have comparatively less accuracy than both adaboost and logistic regression. This project can be developed further by using the concept of Internet of things where sensors can be arranged in the vehicles of people who are probable of facing a heart disease and alerts can be sent faster to the nearest medical facilities.

REFERENCES

[1] Parisa Naraei, Abdolreza Abhari and Alireza Sadeghian,"Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data", IEEE, 2016.

[2] DeepaliChandna "Diagnosis of Heart Disease Using Data Mining Algorithm", IEEE Conf. on International Journal of Computer Science and Information Technologies, 2015, pp 1678-1680

[3] Asghar, S. "Automated Data Mining Techniques: A Critical Literature Review" 978-0- 7695-3595-1, 75 – 79, IEEE, 2009.
[4] M.Akhiljabbara"Heart Disease Prediction System using Associative Classification and Genetic Algorithm"IEEE, 2012.
[5] Niti Guru, Anil Dahiya, Navin Rajpal, "Decision Support System for Heart Disease Diagnosis Using Neural Network", Delhi Business Review, Vol.8, No.1, 2007
[6] K. Srinivas, B.Kavitha Rani and Dr. A. Govrdhan "Application of Data Mining Techniques in Healthcare and Predication of Heart Attacks", International Journal on Computer Science and Engineering, Vol. 02, No. 02, pp.250-255,2011.
[7] N. Deepika and K... Chandrashekar, "Association rule for classification of Heart Attack Patients", International Journal of Advanced Engineering Science and Technologies, Vol.11, No.2, pp253-257, 2011.
[8] D. Shanthi, G.Sahoo and Dr. N. Saravanan, "Designing an Artificial Neural Network Model for the Prediction of Thrombo-embolic Stroke", International Journal of Biometric and Bioinformatics, Vol. 3, No.1,pp250-255,2008
[9] Chaitrali S. Dangare and Sulabha S.Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications , Vol.47, No. 10, pp.0975-888, 2012

[10] Ashish Kumar Sen1 "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach TwoLevel" ISSN 2319-7242Volume 2, 2663-2671, IEEE, 2013.
[11] M V Bhanu Prakash U Sairam, Feature Prospect of the VAST Applications of Machine Learning, Research Review international Journal of Multidisciplinary, volume 4 and issue 4 in April 2019.
[12] Y. Adepu, V. R. Boga and S. U, "Interviewee Performance Analyzer Using Facial Emotion Recognition and Speech Fluency Recognition," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-5, doi: 10.1109/INOCON50539.2020.9298427.
[13] V. Kunta, C. Tuniki and U. Sairam, "Multi-Functional Blind Stick for Visually Impaired People," 2020 5th International Conference on Communication and Electronics Systems (ICCES), 2020, pp. 895-899, doi: 10.1109/ICCES48766.2020.9137870.
[14] Santhosh Voruganti Enhanced Rating Prediction Based On Location And Friend Set published in JETIR May 2019 volume 6 issue 5 ISSN-2349-5162.
[15] Santhosh Voruganti Local Security Enhancement and Intrusion Prevention in Android Devices published in International Research Journal of Engineering and Technology Volume: 07 Issue: 01 January 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072.
[16]. Mr U Sairam, Mr V Santhosh published a paper titled "Breast Cancer Prediction using CNN and Machine Learning Algorithms with Comparative Analysis" in International Journal for Research in Applied Science & Engineering Technology Volume 9, Issue VI, June 2021
[17]. Mr. Santhosh Voruganti and Mr U Sairam published a paper titled "Digital Image Watermarking using Chaotic Encryption and Arnold Transform" in International Journal for Research in Applied Science & Engineering Technology Volume 9, Issue VI, June 2021.
[18]. Mr U Sairam, Mr V Santhosh, Mr MV Bhanu Prakash, Mr R Govardhan Reddy published a paper titled A Study on IoT Applications Towards Impact of Loss of Data in IEEE on 21 June 2021 ISBN 978-1-6654-1571-2 DOI: 10.1109/ICOEI51242.2021.9452935.
[19]. Mr U Sairam published a paper title " Multi-Functional Blind Stick for Visually Impaired People" in IEEE Explore on 11-July-2020 with ISBN:978-1-7281-5371-1.
[20]. Mr U Sairam Published a paper title Interview Performance Analyzer Using Facial Emotion Recognition and speech Fluency Recognition in IEEE Explore in November 2020 with ISBN: 978-1-7281-9744-9/20.
[21]. Mr Santhosh V and U Sairam Published a paper title "Visual Question Answering with External Knowledge" in Elsevier (SSRN Journal) ISSN 1556-5068 and May 25, 2021. http://dx.doi.org/10.2139/ssrn.3853031.
[22]. Mr. U Sairam, Mrs. Surya Samantha (2020) "Applications and Enabling Technologies for IoT "Alochana Chakra Journal ISSN NO: 2231-3990 pp. 9115-9125.
[23]. Mr. U Sairam, Mr. M.V Bhanu prakash (2020) "Dl And Ml Approaches Along With Block chain Towards IoT Security" International Journal of Advanced Science and Technology (Scopus indexed journal) Vol. 29 No. 4s pp. 826-832.
[24]. U. Sai Ram, B. Surya Samantha (2019)"Technology Fundamentals of Block chain and Consideration for Block chain Security International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue- 1C2 pp.415-420.
[25]. Mr. U Sairam, Mr. M.V Bhanu prakash (2019) "Feature Prospect of the VAST Applications of Machine Learning" RESEARCH REVIEW International Journal of Multidisciplinary volume no 4 issue no 04 issn 2455-3085 pp.1266-1271
[26]. Mr. U Sairam, Mr. M.V Bhanu prakash (2019) A Review on Block chain Technology RESEARCH REVIEW International Journal of Multidisciplinary volume no 4 issue any 01 ISSN 2455-3085 pp. 498-501.
[27]. Mr. U Sairam, Mrs. Surya Samantha (2018) "A Survey on Challenges and Benefits towards the Adoption of DevOps Approach" International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653 Volume 6 Issue III pp.1004-1009.

[28]. B. Surya Samantha , M. Truth, U. Sairam (2018) "A Review on Using Crow Search Algorithm in Solving the Problems of Constrained Optimization " Volume 4 Issue 2 Print ISSN: 2395-6011 pp.1004-1009.

[29]. Santhosh Voruganti Map reduce A programming model for cloud computing based on hadoop ecosystem published in International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3794-3799.

[30] Santhosh Voruganti Survey on Data-intensive Applications, Tools and Techniques for Mining Unstructured Data. International Journal of Computer Applications (0975 – 8887), volume 146-No.12, July 2016.

[31] Santhosh Voruganti Comparative Analysis of Dimensionality Reduction Techniques for Machine Learning IJSRST Volume 4 Issue 8 Print ISSN: 2395-6011 Online ISSN: 2395-602X Themed Section: Science and Technology June 2018.

[32].Santhosh Voruganti EFFECTIVE IOT TECHNIQUES TO MONITOR THE LEVELSOF GARBAGE IN SMART DUSTBINS published in International Research Journal of Engineering and Technology Volume: 07 Issue: 06 June 2020 e-ISSN: 2395-0056 p-ISSN: 2395-0072.

[33]. U.Sairam,Santhosh Voruganti ,Mental Health Prediction Using Deep Learning, International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653,Volume 10 Issue II Feb 2022