IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comparison of Several Ml Algorithms for Effective Heart Disease Prediction

SANDHYA GANDI [#1],VADAMODULA VIJAY KUMAR [#2]

[#1] Assistant Professor, Department of Computer Science and Engineering, Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh.

[#2] MCA Student, Department of Computer Science and Application, Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh.

Abstract

Heart disease is becoming one of the most significant reasons for mortality and almost a lot of human beings are suffering from this problem. As we all know it is not so easy to predict the heart disease prior without having very good clinical knowledge. In current days all the predictions are done manually with error rate to find out the abnormalities. In general the manual prediction always makes a lot of errors and a lot of effort is required to process the manual records and hence this motivated me to propose this article in which heart disease prediction can be done by using several Machine Learning algorithms. In general, machine learning is a domain which greatly increases its capabilities in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. In this proposed work, we propose an ensemble model by collecting several ML classification models in one location and then test which model gives more accuracy in the prediction of heart diseases. This proposed work is trained by using several ML algorithms and then checking the following factors such as accuracy, precision, recall and F1-Score.By conducting various experiments on several ML Algorithms by taking UCI dataset, we finally check which algorithm fits best for efficient heart disease prediction.

KEYWORDS:

Machine Learning, UCI Dataset, Heart Disease, Ensemble Model, Disease Prediction.

1. INTRODUCTION

Clinical analysis (CA) is one kind of investigation in the field of examination for AI, incompletely on the grounds that the information is moderately organized and marked, and all things considered, this will be where patients initially cooperate with working, handy computerized reasoning frameworks. This is noteworthy for two reasons. Initially, as far as genuine patient measurements, clinical picture investigation is a litmus test regarding whether man-made brainpower frameworks will really improve understanding results and endurance.

Besides, it gives a tried to human-AI collaboration, of how responsive patients will be towards wellbeing modifying decisions being made, or helped by a non-human entertainer. From the figure 1, we can clearly identify that some of the major causes of death as per the analysis report of DEC 2020.If we look in that figure almost 16587 deaths are occurred in the month of DEC 2020 worldwide due to coronary heart diseases and next to that there are several other causes for deaths. From the figure 1,we can clearly see the difference between men and women who are severely affected with heart diseases and one main cause for all these deaths is no pro-active mechanism to identify the heart disease in the early stages. In general, lack of medical knowledge is one of the root causes for these many deaths.

Hence this motivated me to propose this article in which heart disease prediction can be done by using several machine learning classification algorithms .Here we try to classify the heart disease prediction using several ML algorithms and then check which algorithm is having more accuracy by comparing with several factors like precision,recall,F1-score and so on.

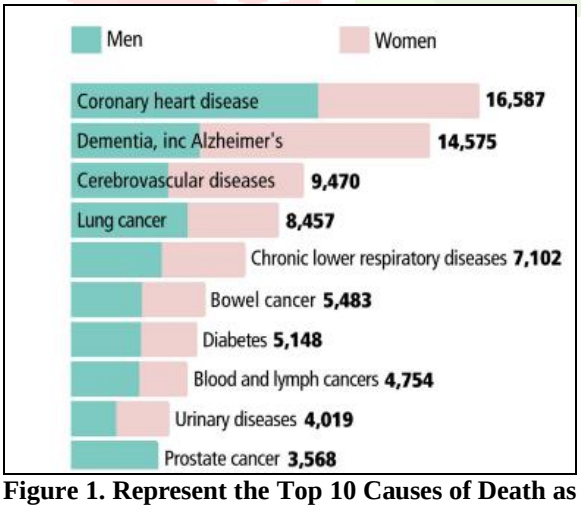


Figure 1. Represent the Top 10 Causes of Death as per the Dec 2020 World Wide Report

In this proposed article we try to compare nearly 8 classification algorithms and then check each and every algorithm with factors such as accuracy,precion,recall and F1-Score and then check which one is best among all these 8 algorithms.

2. LITERATURE SURVEY

Literature survey is that the most vital step in software development process. Before developing the new application or model, it's necessary to work out the time factor, economy and company strength. Once all these factors are confirmed and got an approval then we can start building the application. The literature survey is one which is mainly deal with all the previous work which is done by several users and what are the advantages and limitations in those previous models. This literature survey is mainly used for identifying the list of resources to construct this proposed application.

MOTIVATION

Mohammed Abdul Khaleel[2] has given paper in the Survey of Techniques for mining of data on Medical Data for Finding Frequent Diseases locally. In this proposed work,the author mainly focus on the distinct information about several data mining procedures and also techniques which are required for medical knowledge processing. For example, heart infirmities, lung malignancy and lot more disease information's. The information mining or data mining is the process of extracting the valuable or useful information from hidden data and then try to analyze the data based on user inputs. Here the author mainly concentrated on the importance of Naïve Bayes Algorithm and its importance in order to classify the medical data. The used data-set is obtained from diabetic research institutes of Chennai, Tamilnadu which is leading institute. There are more than 500 patients in the dataset. The tool used is Weka and classification is executed by using 70% of Percentage Split. The accuracy offered by Naive Bayes is 86.419%.

Haik Kalantarian and Mohammad Pourhomayoun [3] has given a paper named Remote Health Monitoring Outcome Success prediction using First Month and Baseline Intervention Data. RHS systems are effective

in saving costs and reducing illness. In this proposed work the authors mainly concentrated on the RHM framework, which is almost cell or mobile based for instructing the remote users and help to connect several users who wish to find the information related to medical. This RHM will help each and every individual to navigate the flow from one location to another and then gather the suitable information in very less time without any data loss.

L.Sathish Kumar and A. Padmapriya [4] has given a paper named Prediction for similarities of disease by using ID3 algorithm in television and mobile phone. In this proposed work the authors mainly concentrated on the major impact of coronary illness and how that is affecting the others.

The given framework utilizes information mining methods, for example, ID3 algorithm. This proposed method helps the people not only to know about the diseases but it can also help's to reduce the death rate and count of disease affected people.

M.A.NisharaBanu and B.Gomathy [5] has given a paper named Disease Predicting system using data mining techniques. In this work the authors mainly concentrated on MAFIA (Maximal Frequent Item set algorithm) and K-Means clustering. As we all know that classification is major process for prediction of any disease present in humans, the authors concentrated on these two classification techniques and found which one is giving best accuracy and which one gives more efficient results.

3. EXISTING SYSTEM AND ITS LIMITATIONS

In the existing system there was no proper method to identify the heart disease prediction using machine learning algorithms. All the existing systems try to use manual approach for predicting the heart disease by collecting symptoms and medical record values and hence the existing system contains several limitations such as :

LIMITATION OF PRIMITIVE SYSTEM

1. All the existing systems use manual approach for heart disease prediction.
2. There is less accuracy in heart disease prediction.
3. More Time Delay in finding the root cause of heart diseases
4. There is no prevention technique due to late prediction.
5. There is no early prediction of heart disease.
6. There should be enough medical knowledge to predict the heart disease based on symptoms.

4. PROPOSED SYSTEM AND ITS ADVANTAGES

In the proposed system used several ML Classification algorithms for efficient heart disease prediction. The proposed system use ML-Approach for classifying the each and every factor very accurately for diagnosis of heart disease and then check which one is best for identifying the disease in very accurate manner and less time complexity to retrieve the data.

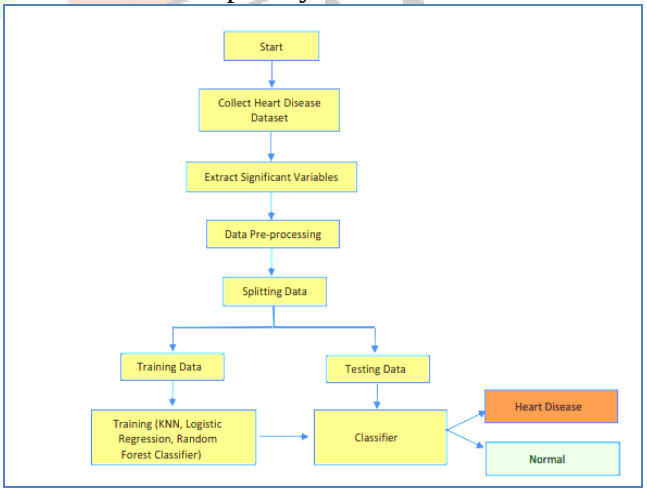


Figure 2. Represent the Working of Proposed Model

ADVANTAGES OF THE PROPOSED SYSTEM

1) By using the proposed ml algorithms we can able to classify the outcome very easily.

2) The proposed mechanism is best for early prediction of heart disease.

3) The proposed system is very efficient in comparing multiple algorithms and picks one among the several algorithms.

4) The proposed system can provide reliable performance for diagnosis of heart disease in early stages.

5) By applying several Ml algorithms of heart patients dataset ,we can get several factors which can show the performance of Ml algorithm accurately.

6) The proposed application is best in preventing heart diseases at the early stage.

5. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes Google Collaboratory and as a Back-End Data base we took UCI Heart Patients Records as dataset. Here we are using Python as Programming Language to implement the current application. The application is divided mainly into following 5 modules. They are as follows:

1. Import Necessary Libraries
2. Load Dataset Module
3. Data Pre-Processing
4. Train the Model Using Several ML Algorithms
5. Find the Performance of ML Algorithms

Now let us discuss about each and every module in detail as follows:

5. 1 IMPORT NECESSARY LIBRARIES

In this module initially we need to import all the necessary libraries which are required for building the model. Here we try to use all the libraries which are used to convert the data into meaningful manner.Here the data is divided into numerical values which are easily identified by the system,hence we try to import numpy module and for plotting the data in graphs and charts we used matplot library.

I. Importing essential libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

%matplotlib inline

import os
print(os.listdir())

import warnings
warnings.filterwarnings('ignore')
```

```
['.ipynb_checkpoints', 'heart.csv', 'Heart_disease_prediction.ipynb', 'README.md']
```

5. 2 LOAD DATASET MODULE

In this module the we try to load the dataset which is downloaded or collected from UCI repository. Here we store the dataset names as ‘Heart.csv ‘file and this dataset contains the following information such as :

```
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

Each and every attribute contains some information which are tested and collected based on individual patient id.

5.3 DATA PRE-PROCESSING MODULE

Here in this section we try to pre-process the input dataset and find out if there are any missing values or in-complete data

present in the dataset. If there is any such data present in the dataset, the application will ignore those values and load only valid rows which have all the valid inputs.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

5.4 TRAIN THE MODEL USING SEVERAL ML ALGORITHMS

Here we try to train the current model on given dataset using several ML classification algorithms and then try to find out which algorithms suits best in order to identify and classify the input dataset accurately and efficiently. Here we try to use following algorithms on input dataset such as:

1. Logistic Regression
2. Naïve Bayes
3. Support Vector Machine
4. K-Nearest Neighbors
5. Decision Tree
6. Random Forest
7. XGBoost
8. Neural Networks

5.5 PERFORMANCE ANALYSIS MODULE

Here in this module we try to compare each and every classification algorithm on given input dataset and then try to find out which one suits best for finding the accurate results. Finally we will identify the best algorithm which gives accurate results in very less time. Here we can see **Random Forest** gives more accurate result compared with other ML Algorithms.

6. EXPERIMENTAL RESULTS

In this section we try to design our current model using PYTHON as programming language and taking Heart Disease Dataset from UCI Machine Learning Repository as storage database. Here we try to construct the application by using several ML classification algorithms to predict heart disease present or not based on set of features which are recorded for every patient.

OUTPUT FINAL SCORE

In: scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn]

algorithms = ["Logistic Regression","NaiveBayes","Support Vector Machine","K-Nearest Neighbors","DecisionTree","Random Forest","XGBoost","Neural Network"]

for i in range(len(algorithms)):

print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")

The accuracy score achieved using Logistic Regression is: 85.25 %

The accuracy score achieved using Naive Bayes is: 85.25 %

The accuracy score achieved using Support Vector Machine is: 81.97 %

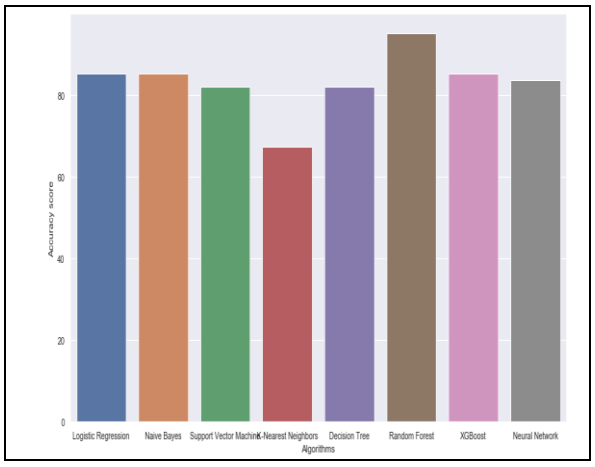
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %

The accuracy score achieved using Decision Tree is: 81.97 %

The accuracy score achieved using Random Forest is: 95.08 %

The accuracy score achieved using XGBoost is: 85.25 %

The accuracy score achieved using Neural Network is: 83.61 %



From the above Graph we can conclude that Random Forest Gives Best Accuracy when compared with many other ML Algorithms for Efficient Prediction of Heart Disease Patients Dataset.

7. CONCLUSION

In this current work we for the first time proposed an ensemble model which can combine multiple machine learning algorithms and propose a new model which can give more accuracy. Here we gathered several machine learning classification algorithms and try to compare each and every individual algorithm on heart disease patients dataset and finally conclude which algorithm is best for predicting heart disease very effectively. For this we gathered some heart patient's data from KAGGLE or UCI repository and then trained the system with this dataset. The dataset is trained using all the individual ML algorithms and then try to compare which algorithm has advantages and which algorithm has limitation in order to find out the heart disease prediction.

By conducting various experiments on our proposed dataset using several ML classification algorithms, our comparative results clearly state that out of 100% accuracy the random forest achieved 95.08 % accuracy and effectively extracted the heart patient's data. In future we want to extend the same application by taking deep learning models and then find out which one is best suited for heart disease prediction on large datasets.

REFERENCES

[1] LathaParthiban and R.Subramanian, Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm, International Journal of Biological and Medical Sciences, 2008.

[2] JesminNahar, TasadduqImama, Kevin S. Tickle, Yi-Ping Phoebe Chen, Association rule mining to detect factors which contribute to heart disease in males and females, Elsevier, 2013.

[3] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, HaikKalantarian, MajidSarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics.

[4] PonrathiAthilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Medical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march2017.

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[7] ShantakumarB.Patil, Dr.Y.S. Kumaraswamy, Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction, (IJCSNS) International Journal of Computer Science and Network 228 Security ,2009.

[8] Abhishektaneja, Heart Disease Prediction System Using Data Mining Techniques, Oriental Scientific Publishing Co., India, 2013.

[9] M. Anbarasi, E. Anupriya, N.ch.s.n.Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology,2010.

[10] Miss. Chaitrali S. Dangare, Dr. Mrs.Sulabha S. Apte, A data mining approach for prediction of heart disease using neural networks, international journal of computer engineering and technology, 2012.

[11] Shadab Adam Pattekari and AsmaParveen, prediction system for heart disease using naïve

bayes, International Journal of Advanced Computer and Mathematical Sciences, 2012.

[12] N. AdityaSundar, P. PushpaLatha, M. Rama Chandra,performance analysis of classification data mining techniques over heart diseases data base, international journal of engineering science and advanced technology, 2012.

About the Authors

SANDHYA GANDI is currently working as an Assistant Professor in Department of Computer Science and Engineering at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh. She has more than 10 years of teaching experience. Her research interest includes Java, Python, .Net, HTML.

VADAMODULA VIJAY KUMAR is currently pursuing his 2 years MCA in Department of Computer Science and Applications at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh.His area of interest includes C, C++, Java and Python.