IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

HEART DISEASE PREDICTION USING MACHINE LEARNING

U D Prasan[1*], M. Narsinga Rao[2]

[1]Department of Computer Science and Engineering,
Aditya Institute of Technology and Management, Tekkali - 532201, India.
[2]Dept of CSE, Gandhi Institute of Science and Technology
* Corresponding author: udprasanna@gmail.com

Abstract

In this recent era, cardiovascular disease (CVD) propagation rate has been intensifying the cause of death worldwide among the non-communicable disease. In particular the south Asian countries have a tremendous risk of cardiovascular disease at an early age than any other ethnic group. Most often it's challenging for medical practitioners to predict cardiovascular disease as it requires experience and knowledge which is a complex task to accomplish. This health industry has enormous amounts of data which is useful for making effective conclusions using their hidden information. So, using appropriate results and making effective decisions on data, some superior data analysis techniques are used, for example Naive Bayes, Decision Tree. By using some properties like (age, gender, BP, stress, etc) it can be predicted the chances of cardiovascular disease. Logistic regression, Decision tree, and Naive bayes.

Keywords: cardiovascular disease, Naive Bayes, Decision Tree, Logistic regression

Introduction:

Heart disease predictor is an offline platform designed and developed to explore the path of machine learning. The goal is to predict the health of a patient from collective data, so as to be able to detect configurations at risk for the patient, and therefore, in cases requiring emergency medical assistance, alert the appropriate medical staff of the situation of the latter. We initially have dataset collecting information of many patients with which we are able to conclude the results into a complete form and can predict data precisely. The main challenge in today's healthcare is provision of best quality services and effective accurate diagnosis. Even if heart diseases are found as the prime source of death in the world in recent years, they are also the ones that can be controlled and managed effectively. The whole accuracy in management of a disease lies on the proper time of detection of that disease. The proposed work makes an attempt to detect these heart diseases at early stage to avoid disastrous consequences. Records of large set of medical data created by medical experts are available for analyzing and extracting valuable knowledge from it. Data mining techniques are the means of extracting valuable and hidden information from the large amount of data available.

Hence, decision making using discrete data becomes complex and tough task. Machine Learning (ML) which is subfield of data mining handles large scale well-formatted dataset efficiently. In the medical field, machine learning can be used for diagnosis, detection and prediction of various diseases. The main goal of this paper is to provide a tool for doctors to detect heart disease as early stage. This in turn will help to provide effective treatment to patients and avoid severe consequences. ML plays a very important role to detect the hidden discrete patterns and thereby analyze the given data. After analysis of data ML techniques help in heart disease prediction and early

Research paper © 2012 IJFANS. All Rights Reserved,

diagnosis. This paper presents performance analysis of various ML techniques such as Decision Tree, KNN algorithms for predicting heart disease at an early stage.

Problem Definition

Machine learning allows building models to quickly analyze data and deliver results, leveraging the historical and real-time data, with machine learning that will help healthcare service providers to make better decisions on patient's disease diagnosis. By analyzing the data we can predict the occurrence of the disease in our project. This intelligent system for disease prediction plays a major role in controlling the disease and maintaining the good health status of people by predicting accurate disease risk. Machine learning algorithms can also be helpful in providing vital statistics, real-time data and advanced analytics in terms of the patient's disease, lab test results, blood pressure, family history, clinical trial data etc.

Literature Survey

Tom Mitchell states machine learning as "A computer program is said to learn from experience and from some tasks and some performance on, as measured by, improves with experience". Machine Learning is combination of correlations and relationships, most machine learning algorithms in existence are concerned with finding and/or exploiting relationship between datasets. Once Machine Learning Algorithms can pinpoint on certain correlations, the model can either use these relationships to predict future observations or generalize the data to reveal interesting patterns. In Machine Learning there are various types of algorithms such as Regression, Linear Regression, Logistic Regression, Naive Bayes Classifier, Bayes theorem, KNN (K-Nearest Neighbor Classifier), Decision Tress, Entropy, ID3, SVM (Support Vector Machines), K-means Algorithm, Random Forest and etc.,

The name machine learning was coined in 1959 by Arthur Samuel. Machine learning I explores the study and construction of algorithms that can learn from and make predictions on data Machine learning is closely related to (and often overlaps with) computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. These analytical models allow researchers, data scientists, engineers, and analysts to "produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical relationships and trends in the data.

Existing system

Prediction using traditional methods and models involves Various risk factors and it consists of various measures of algorithms such as datasets, programs and much more to add on. High-risk and Low-risk patient classification is done on the basis of the tests that are done in group. But these models are only valuable in clinical situations and not in big industry sector. So, to include the disease predictions in various health related industries, we have used the concepts of machine learning and supervised learning methods to build the predictions system. After doing the research and comparison of all the algorithms and theorems of machine learning we have come to conclusion that all those algorithms such as Decision Tree, KNN, Naïve Bayes, Regression and Random Forest Algorithm all are important in building a disease prediction system which predicts the disease of the patients from which he/she is suffering from and to do this we have used some performance measures like ROC, KAPPA Statistics, RMSE, MEA and various other tools.

After using various techniques such as neural networks to make predictions of the diseases and after doing that we come to conclusion that it can predicts up to 90% accuracy rate after doing the experimentation and verifying the

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

results. The information of patient statistics, results, disease history in recorded in EHR, which enables to identify the potential data centric solution, which reduces the cost of medical case studies. Existing system can predict the disease but not the sub type of the disease and it fails to predict the condition of the people, the predictions of disease have been indefinite and non-specific.

Proposed System

Heart disease is the leading cause of death among all other diseases, even cancers. The number of men & women facing heart disease is on a raise each year. This prompts for its early diagnosis data has been collected from Kaggle. Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand & treatment. Due to lack of resources in the medical field, the prediction of heart disease occasionally may be a problem. Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity & patients. This issue can be resolved by adopting Machine Learning techniques. This paper intends to adopt DECISION TREE CLASSIFIER & KNN (K-Nearest Neighbour) - two techniques for the effective prediction of Heart disease. Here we have combined the overall structure and unstructured form structure of data for the overall risk analysis that is required for doing the prediction of the disease. Using the structured analysis, we can identify the chronic types of disease in a particular region and particular community.

In unstructured analysis we select the features automatically with the help of algorithms and techniques. This system takes symptoms from the user and predicts the disease accordingly based on the symptoms that it takes and also from the previous datasets, it also helps in continuous evaluation of viral diseases, heart rate, blood pressure, sugar level and much more which is in the system and along with other external symptoms its predicts the appropriate and accurate disease.

Requirements And Technical Description

Python is a high-level, general-purpose and a very popular programming language. Python programming language (latest Python 3) is being used in web development used in this research Colab is used extensively in the machine learning community with applications including: Getting started with TensorFlow, Developing and training neural networks , Experimenting with TPUs , Disseminating AI research, Creating tutorials

Scikit-Learn-It is a Python library associated with NumPy and SciPy. It is considered as one of the best libraries for working with complex data. There are a lot of changes being made in this library. One modification is the cross validation feature, providing the ability to use more than one metric. Lots of training methods like logistics regression and nearest neighbors' have received some little improvements. It contains a numerous number of algorithms for implementing standard machine learning and data mining tasks like reducing dimensionality, classification, regression, clustering, and model selection.

Pandas-pandas are mainly used for data analysis. A panda allows importing data from various file formats such as comma-separated values, JSON, SQL, and Microsoft Excel. Pandas allow various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

Numpy- Numpy is a python library used for working with arrays. It also has functions for working in domain of linear algebra, flourier transform, and matrices. Numpy was created in 2005 by Travis Oliphant. It is an open source project.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

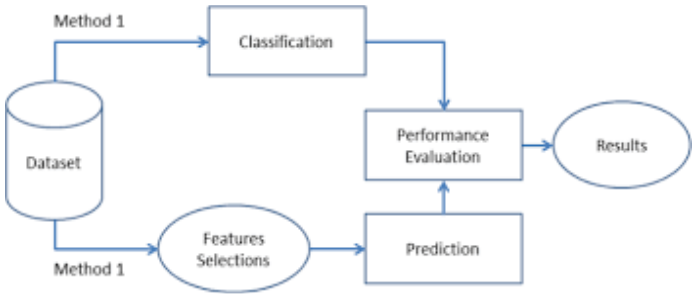
Matplotlib- matplotlib. Pyplot is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

Seaborn- Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics

Methodology

Disease prediction using machine learning predicts the presence of the disease for the user based on various symptoms and the information the user gives such as sugar level, haemoglobin level and many more such general information through the symptoms. The architecture of the system disease prediction using machine learning consist of various datasets through which we will compare the symptoms of the user and predicts it, then the datasets are transformed into the smaller sets and from there it gets classified based on the classification algorithms later on the classified data is then processed into the machine learning technologies through which the data gets processed and goes in to the disease prediction model using all the inputs from the user that is mentioned above.

Then after user entering the above information and overall processed data combines and compares in the prediction model of the system and finally predicts the disease. An architecture diagram is a graphical representation of a set of concepts, that are part of architecture, including their principles, elements and components. The diagram explains about the system software in perception of overview of the system.



System architecture

Datasets

Table 1: Summary of the heart disease dataset

Sno	Attributes	Description
1	Age	Age in years
2	Sex	Male or Female
3	Cp	Chest pain type
4	Thestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	Restecg	Resting electrographic results
7	Fbs	Fasting blood sugar
8	Thalach	Max. heart rate achieved
9	exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of the peak exercise ST segment
12	Ca	No. of major vessels colored
13	Thal	Defect type

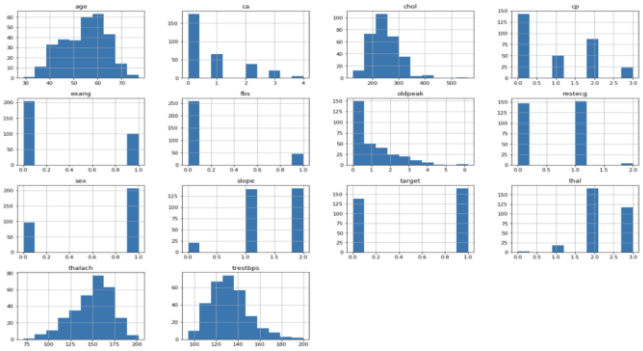
ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

Heart Disease Dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
29	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1
261	38	1	3	120	231	0	1	182	1	3.8	1	0	3	0
262	66	0	0	178	228	1	1	165	1	1	1	2	3	0
263	52	1	0	112	230	0	1	160	0	0	2	1	2	0
264	53	1	0	123	282	0	1	95	1	2	1	2	3	0
265	63	0	0	108	269	0	1	169	1	1.8	1	2	2	0
266	54	1	0	110	206	0	0	108	1	0	1	1	2	0
267	66	1	0	112	212	0	0	132	1	0.1	2	1	2	0
268	55	0	0	180	327	0	2	117	1	3.4	1	0	2	0
269	49	1	2	118	149	0	0	126	0	0.8	2	3	2	0
270	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
271	56	1	0	130	283	1	0	103	1	1.6	0	0	3	0
272	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
273	61	1	3	134	234	0	1	145	0	2.6	1	2	2	0
274	67	1	0	120	237	0	1	71	0	1	1	0	2	0
275	58	1	0	100	234	0	1	156	0	0.1	2	1	3	0
276	47	1	0	110	275	0	0	118	1	1	1	1	2	0
277	52	1	0	125	212	0	1	168	0	1	2	2	3	0
278	58	1	0	146	218	0	1	105	0	2	1	1	3	0
279	57	1	1	124	261	0	1	141	0	0.3	2	0	3	0
280	58	0	1	136	319	1	0	152	0	0	2	2	2	0
281	61	1	0	138	166	0	0	125	1	3.6	1	1	2	0
282	42	1	0	136	315	0	1	125	1	1.8	1	0	1	0
283	52	1	0	128	204	1	1	156	1	1	1	0	0	0
284	59	1	2	126	218	1	1	134	0	2.2	1	1	1	0
285	40	1	0	152	223	0	1	181	0	0	2	0	3	0
286	61	1	0	140	207	0	0	138	1	1.9	2	1	3	0
287	46	1	0	140	311	0	1	120	1	1.8	1	2	3	0
288	59	1	3	134	204	0	1	162	0	0.8	2	2	2	0
289	57	1	1	154	232	0	0	164	0	0	2	1	2	0

Results and Discussions



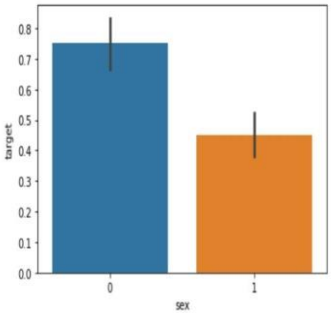
Histogram of the Attributes

ISSN PRINT 2319 1775 Online 2320 7876

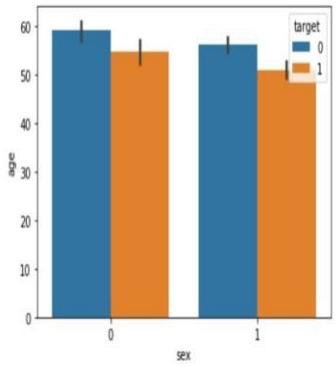
Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

```
In [53]: import warnings
         warnings.filterwarnings('ignore')
```

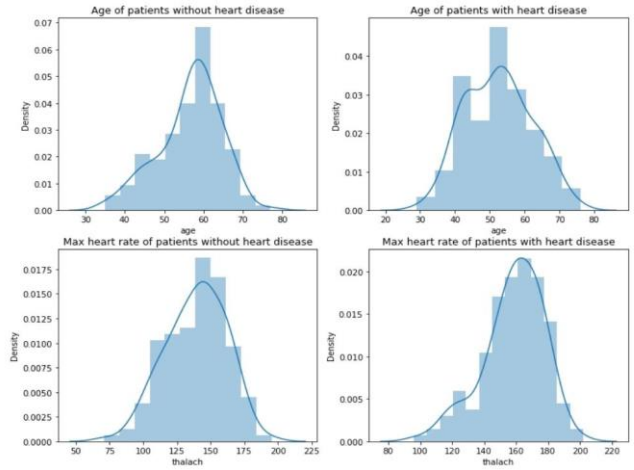
```
In [54]: #seaborn
         sns.barplot(df['sex'],df['target'])
         plt.show()
```




```
In [55]: sns.barplot(df['sex'],df['age'],hue=df['target'])
         plt.show()
```




```
plt.figure(figsize=(12,10))
plt.subplot(221)
sns.distplot(df[df['target']==0].age)
plt.title('Age of patients without heart disease')
plt.subplot(222)
sns.distplot(df[df['target']==1].age)
plt.title('Age of patients with heart disease')
plt.subplot(223)
sns.distplot(df[df['target']==0].thalach )
plt.title('Max heart rate of patients without heart disease')
plt.subplot(224)
sns.distplot(df[df['target']==1].thalach )
plt.title('Max heart rate of patients with heart disease')
plt.show()
```



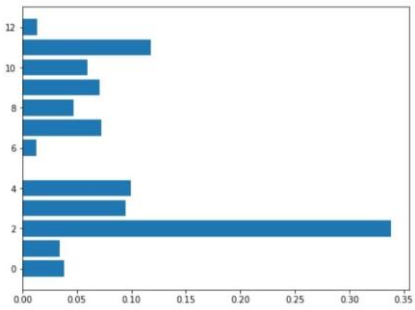
ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

Decision tree algorthim

```
In [132]: from sklearn.tree import DecisionTreeClassifier

          dt=DecisionTreeClassifier()
          dt.fit(x_train,y_train)

Out[132]: DecisionTreeClassifier()
```

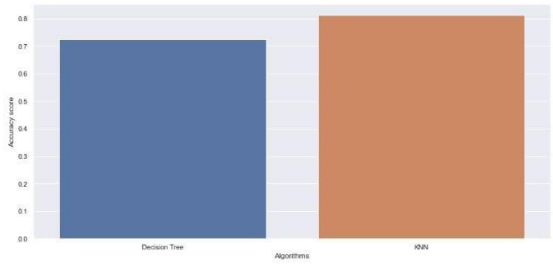
```
In [140]: def plot_f(model):
              plt.figure(figsize=(8,6))
              n=13
              plt.barh(range(n),model.feature_importances_,align='center')

          plot_f(dt)
```



KNN algorithm

```
In [108]: from sklearn.preprocessing import StandardScaler
          std=StandardScaler().fit(x)
          x_std=std.transform(x)
```

```
Out[63]: <AxesSubplot:xlabel='Algorithms', ylabel='Accuracy score'>
```



Conclusion and Future Scope

Heart Disease is one of the major concerns for society today. It is difficult to manually determine the odds of getting heart disease based on risk factors. However, machine learning techniques are useful to predict the output from existing data. Heart disease is the leading cause of death among all other diseases, even cancers. The number of men & women facing heart disease is on a raise each year. This prompts for its early diagnosis data has been collected from Kaggle. Data collection is the process of gathering and measuring information from countless different sources. In order to use the data we collect to develop practical artificial intelligence (AI) and machine learning solutions, it must be collected and stored in a way that makes sense for the business problem at hand & treatment. Due to lack of resources in the medical field, the prediction of heart disease occasionally may be a

ISSN PRINT 2319 1775 Online 2320 7876

Research paper © 2012 IJFANS. All Rights Reserved, UGC CARE Listed (Group -I) Journal Volume 11, Iss 7, July 2022

problem. Utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity & patients. This issue can be resolved by adopting Machine Learning techniques. This intends to adopt DECISION TREE CLASSIFIER & KNN (K-Nearest Neighbour) - two techniques for the effective prediction of Heart disease.

References

1. Disease Prediction And Doctor Recommendation System By Www.Iriet.Net
2. Disease Prediction Based On Prior Us. Ahrg.Gov/Nisoverview.Jsp Knowledge By Www.Hcup
3. Gdps - General Disease Prediction System By Www.Iriet.Net
4. Disease Prediction Using Machine Learning By International Research Journal Of Engineering And Technology (Irjet).