JOURNAL OF OPTOELECTRONICS LASER ISSN: 1005-0086

Volume 41 Issue 8, 2022

Analysis Of Heart Disease Prediction Using Machine Learning Classification Algorithms

Dr. V. Krishna*[1], Mrs. Ch. Sumalatha[2], Mr. Y. David Solomon Raju[3], Dr. K V Murali Mohan[4]

[1]Professor, Dept. of CSE, TKR College of Engineering & Technology, Hyderabad.
[2]Asst. Prof., Dept of ECE, Shadan Women's College of Engineering and Technology, Hyderabad.
[3]Assoc. Prof., Dept of ECE, Holy Mary Institute of Technology & Science, Hyderabad.
[4]Professor of ECE, Teegala Krishna Reddy Engineering College, Hyderabad Hyderabad.

ABSTRACT

Heart disease prediction is an important task in the health care domain now a day. Because for every minute, the number of people passing away from a heart attack. It is difficult to heart disease prediction by physicians with huge health records. Due to this problem, the healthcare industry wants to implement an automatic system technique that will deliver productive decisions from a huge dataset. So, the machine learning techniques are effective in resolving these kinds of issues very well. In the medical industry, important data can be gathered from various patients' manifestations and clinical reports for analysis by physicians. These days, at this stage of their lifetime lot of people are getting heart failure symptoms. But comparing old people and young people, senior citizens are facing this type of problems. However, the machine learning techniques can find correlations between different features for the prediction of heart disease status from the training dataset. To overcome this complexity, we need to implement the automatic heard disease prediction system to notify the patient and get to recover from the disease. Here to gain the automatic system we are using machine learning techniques to easily perform heart disease prediction with huge data. The machine learning techniques can be split into multiple types like unsupervised and supervised learning classifiers. The unsupervised learning techniques are used for prediction with unstructured data. But the supervised learning techniques working with structured data which is recommended to implement these classifiers. So, in this system, we are using supervised machine learning techniques such as KNN, RF, LR, DT, NB, and SVM classifiers. For the heart disease prediction, this system is using a training dataset, accessed from the Kaggle repository. As well as this system is comparing accuracy performance between various machine learning algorithms and shows the accuracy results with a graphical presentation.

Keywords: Machine Learning, Classifiers, Dataset, Prediction.

I. INTRODUCTION

Day by day data of huge health records are raised in the healthcare medical industry.So, it is recommended to manage a huge data and make them useful information for favorable decision making. Due to this problem, the healthcare industry wants to implement anautomatic system technique that will deliver productive decisions from a huge dataset. So, the machine learning techniques are effective inresolving these kinds of issues very well. In the medical industry, important data can be gathered from various patients' manifestations and clinical reports for analysis by physicians. These days, at this stage of their lifetime lot of people are getting heart failure symptoms. But comparing old people and young people, senior citizens are facing this type ofproblems. However, the machine learning techniques can find correlations between different features for the prediction of heart disease status from the training dataset. By using this kind of training model, it can detect heart disease patients without the help of medical practitioners. Then it can pretend as an automatic system to categorize positive heart disease patients and negative heart disease patients accurately, so then it reduces the diagnosis time and cost of treatment.

In the health care domain, providing quality services and predicting the diagnosis status accurately is the main challenge task. According to a survey a lot of people who passed away with heart disease were even managed and controlled effectively by an automatic system. So, in this system, the proposed system can predict the heart disease status at an advanced stage to notify the patients and help them to recovery from that disease. A huge of medical records are generated by medical experts for analyze and retrieve the useful information from that database. The health care database contains mostly unattached information which is tedious task for prediction of heart disease. Therefore, this system proposed a automatic system to physicians for prediction of heart disease at advance stage then they can provide treatment to patient and save them from ruggedseriousness. So, the machine learning techniques have an important role in heart disease prediction with supervised classifiers at advance stage to diagnoses the patients.

II. PROBLEM STATEMENT

The detection or prediction of heart disease is toughest task in the health care domain. The physicians can detect heart disease with some symptoms such as smoking, high in taking of fat and consuming alcohol etc. But with these symptoms the physicians can detect disease may or may not be accurately. Due to this reasons doctors cannot do treatment to patients at early stages then there is a chance to face the harmful outcomes by patients. So, to detect or prediction of heart disease we need to develop the tool for detection of any disease at early stages then the physicians will do the treatment to patients to preventharmful consequences. Here the prediction tool can be implementing by the supervised machine learning techniques to heart disease prediction. Many clinical or hospitals generate huge medical records which is unstructured format. So, by using this prediction tool can easily fetch the useful information to make the training dataset for further references. The machine leaning algorithms will take this training dataset as input and predict the heart disease status with current patient details as testing dataset.

DOI: 10050086.2022.08.68

2.1. Contribution

In this system we proposed heart disease prediction with six machine learning classification algorithms and shows the accuracy results between these six algorithms. Here the main task of this system is to predict accurately when patient feels pain with heart disease. Regarding implementation, with help of heart disease training dataset and machine learning classifiers we build the train model file and then the physicians can enter the input values from patients' health records and giving to the training model as input for heart disease prediction. Here the heart disease training dataset is downloaded from the Kaggle repository which is uploaded by the medical department experts. To build this system we had chosen python language which has pre-trained libraries or packages to access the machine learning classifiers. Here all six algorithms providing best accuracies for prediction of heart disease.

2.2 Existing System

The physicians can detect heart disease with some symptoms such as smoking, high in taking of fat and consuming alcohol etc. But with these symptoms the physicians can detect disease may or may not be accurately. Due to this reasons doctors cannot do treatment to patients at early stages then there is a chance to face the harmful outcomes by patients. So, to detect or prediction of heart disease we need to develop the tool for detection of any disease at early stages then the physicians will do the treatment to patients to prevent harmful consequences.

2.3 Proposed System

The proposed systems aim to predict heart diseases more accurately than other systems. The treatments for heart diseases are lagging due to several silent symptoms.   Monitoring the health condition using the data collected from various resources supports to predict the health condition of the patients and to take appropriate measures. Health care is facing the issue of predicting and diagnosing the disease. The major issue is information overloading. Machine Learning techniques are adapted to predict heart disease and help doctors to take appropriate decisions and treat the disease. Predicting heart disease by applying machine learning algorithms such as Random Forest, Naïve Bayes, Logistic Regression, Decision Tree, Support Vector Machine, and KNN classification are widely used. The gained information in the dataset is used to predict the chance of heart disease. As a result, machine learning helps to achieve natural evolution in the medical field for the diagnosis of heart diseases.

III. DESCRIPTION TO MACHINE LEARNING

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: The ability to learn. Machine learning is actively being used today, perhaps in many more places than one would expect.

Basic Difference in ML and Traditional Programming

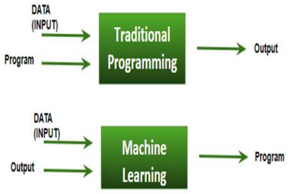


Fig.1 Block Diagram of ML and Traditional Programming

Traditional Programming: We feed in DATA (Input) + PROGRAM (logic), run it on machine and get output.

Machine Learning: We feed in DATA (Input) + Output, run it on machine during training and the machine creates its own program (logic), which can be evaluated while testing.

3.1. Block Diagram

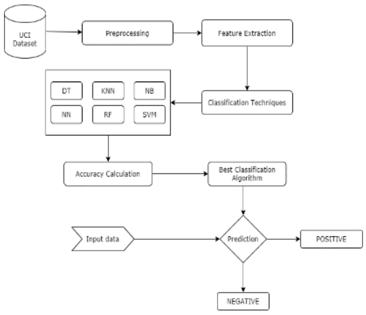


Fig.2 Block Diagram

Fig.2 depicts about our proposed system model. In this system model they used heart disease training dataset which is downloaded from the Kaggle repository. Later by preprocessing, it can read the training dataset and split the independent and dependent attributes by feature extraction and then build the training model with classification algorithm for heart disease prediction by giving input data, finally calculate the accuracy between six machine learning classifiers.

A. Dataset Collection

In this system, we are using the "Framingham" heart disease dataset shown in figure. 2 which is accessed from the Kaggle web repository, the following link can providethe dataset (https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset).This dataset includes over 4,240 records among them 3596 records belong to NEGATIVE and 139 records belong to POSITIVE classes and it contains 15 attributes which are defined in Table.1. The goal of the dataset is to predict whether the patient has the risk of future (CHD) coronary heart disease.

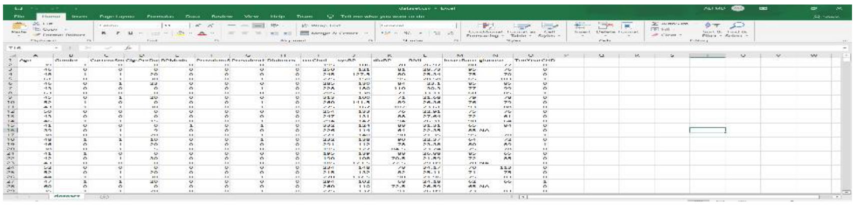


Fig.3 Heart Disease Dataset

B. Preprocessing

In the preprocessing, we need to load or read the training dataset with help of the panda's library and by importing the panda's library we can invoke the read_csv () method for reading the entire dataset and store in a variable and using the dropna (), it will remove records, if the dataset contains the empty or NaN values.

C. Feature Extraction

After completion of preprocessing such as loading the training dataset, we need to get the features of the given dataset. By using the feature extraction method Chi-Square Test, this system will get the top 10 selected features to train the classification models. UsingSelectKBest () function, we can get the top 10 best features based on the chi2 methodology which is imported from sklearn. feature_selection package.

3.2 Classification Techniques

Decision Tree (DT): The DT classifier is supervised machine learning technique, in this classifier it has one root and multiple nodes and finished with leaf nodes. Here the DT classifier can prepare the dataset in the tree form. Later when the user enters the testing dataset for heart disease prediction then it follows the IF and THEN rules which means it compares with every node, until it reaches to leaf node. The leaf nodes contain the target column values such as POSITIVE or NEGATIVE.

Table 1: **Attributes** Of Heart Disease Dataset

S. No	Attribute	Description
1.	Age	Age at the time of medical examination in years.
2.	Gender	The gender of the observations.
3.	Current Smoker	Current cigarette smoking at the time of examinations.
4.	CigsPerDay	The number of cigarettes smoked each day.
5.	BPmeds	Use of Anti-hypertensive medication at exam
6.	PrevalentStroke	Prevalent Stroke (0 = free of disease
7.	PrevalentHyp	Prevalent Hypertensive. The subject was defined as hypertensive if treated.
8.	Diabetes	Diabetic according to criteria of the first exam treated.
9.	TotChol	Total cholesterol (mg/dL).
10.	SysBP	Systolic Blood Pressure (mmHg).
11.	DiaBP	Diastolic blood pressure (mmHg).
12.	BMI	Body Mass Index, weight (kg)/height (m)^2
13.	Heartrate	Heart rate (beats/minute)
14.	Glucose	Blood glucose level (mg/dL).
15.	TenYearCHD	risk of coronary heart disease (CHD).

Finally, the testing dataset value matches the final leaf node value that can be the system predictable values. In this system, we are using sklearn. tree package to import the Decision Tree Classifier to build training model for heart disease prediction.

The branches/edges represent the result of the node and the nodes have either:
1. Conditions [Decision Nodes]
2. Result [End Nodes]

The branches/edges represent the truth/falsity of the statement and take makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:

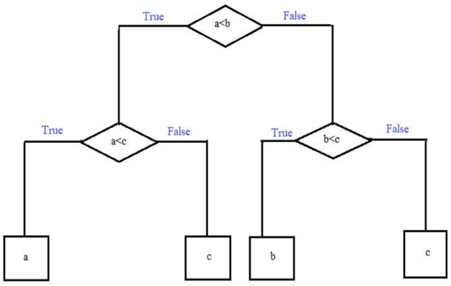


Fig4 Example of Decision Tree Classifier

Strengths and Weakness of Decision Tree approach
The strengths of decision tree methods are:
Decision tree are able to generate understandable rules.
Decision tree performs classification without requiring much computation.
The weakness of decision tree methods:
Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
Decision trees are prone to errors in classification problems with many classes and relatively small number of training examples.

The below snippet will show the building of DT classifier model.

```
from sklearn. tree importDecisionTreeClassifier
rf = DecisionTreeClassifier()
rf.fit (x_train, y_train)
pre_cls = rf.predict(x_test)
```

Random Forest (RF):
The RF classifier is a collection of decision tress. It is also belonging to supervised machine learning algorithm. Here the RF classifier will gather number of decision tress randomly to taking the decision. While prediction of disease it takes the all-output values of decision trees randomly and which class is voted more than other class then that become the system predictable output status. In this system RF classifier providing 98% best accuracy compare with remaining algorithms. This system will use sklearn.ensemblepackage to import the RandomForestClassifier to build training model for heart disease prediction.

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data, and hence the output doesn't depend on one decision tree but on multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. This part is called Aggregation.

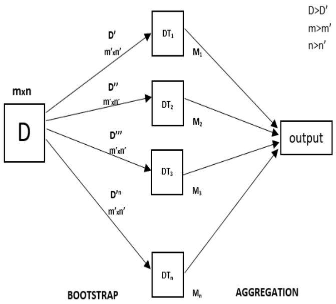


Fig.5 Example of Random Forest Classifier

Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

The below syntax will show the preparation of build model.

```
from sklearn.ensembleimportRandomForestClassifier
rf = RandomForestClassifier()
rf.fit(x_train,y_train)
pre_cls = rf.predict(x_test)
```

Logistic Regression (LR):

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.This classifier also can import *sklearn.linear_model*package Logistic Regression for heart disease prediction.

Logistic regression is named for the function used at the core of the method, the logistic function. The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.
1 / (1 + e^-value)
Where e is the base of the natural logarithms' (Euler's number or the EXP () function in your spreadsheet) and value is the actual numerical value that you want to transform. Below is a plot of the numbers between -5 and 5 transformed into the range 0 and 1 using the logistic function.

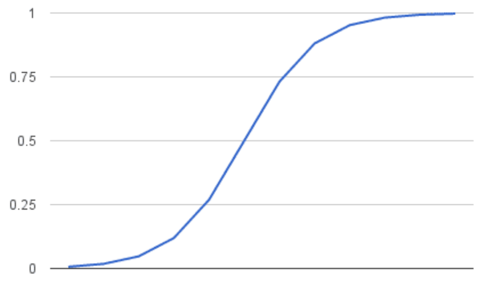


Fig 6 Example of Logistic Regression

Follow the below snippet code:

```
fromsklearn.linear_modelimport LogisticRegression
lr_clf = LogisticRegression()
lr_clf.fit(x_train,y_train)
predicted=lr_clf.predict(x_test)
```

Naïve Bayes (NB):

In this system, the naïve Bayes algorithm can be used for the prediction of heart disease. This algorithm follows the Bayes rule for heart disease prediction. It is the fastest and most easily predictable classifier and it calculates posterior probability events with other events this algorithm uses mostly for text classifications. This classifier Multinomial NB is imported from sklearn. naive_bayes package.
Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.Bayes' Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes' theorem is stated mathematically as the following equation:

NB= P(A/B) =P(B/A). P(A)/P(B)
where A and B are events and P(B) $\neq$ 0.
Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.
P(A) is the priori of A (the prior probability, i.e., Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance (here, it is event B).
P(A|B) is a posteriori probability of B, i.e., probability of event after evidence is seen.
The classifier follows the below snippet.

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinominalNB()
nb.fit(x_train,y_train)
pre_cls = nb.predict(x_test)
```

Support Vector Machine (SVM):

The support vector machine classifier is an important classifier because of their classification advantages. The SVM classifier while the classification of features, first can draw the margins between different classes, and the hyper plane line can separate with support vectors which means the nearest classes to that hyper plane line. Here this system can separate hyper plane with POSITIVE and NEGATIVE features and select the nearest support vectors and build the training model to predict the heart disease status.

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

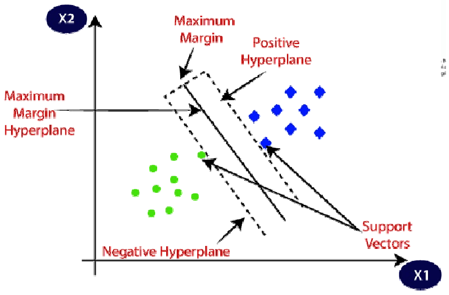


Fig.7 Example of SVM

The below syntax is following the code of heart disease prediction

```
from sklearnimportsvm
svm = svm.SVC()
svm.fit(x_train,y_train)
pre_cls = svm.predict(x_test)
```

K-Nearest Neighbour (K-NN):

The K-nearest neighbor classifier is a different learning classifier compare with another machine learning classifier, because it follows the Euclidian distance formula to calculate the distance. This classifier while prediction it calculates the distance between each record then it returns the distance and store it like this follow the last record and it can return the predictable output value which distance is less to compare with all distances and that one become our heart disease predictable output like POSITIVE or NEGATIVE. It is also import the KNeighbour'sClassifier module from this sklearn. neighbors.

How Does K-NN Work?

The K-NN working can be explained on the basis of the below algorithm:

Step-1: Select the number K of the neighbours
Step-2: Calculate the Euclidean distance of **K number of neighbours**
Step-3: Take the K nearest neighbours as per the calculated Euclidean distance.
Step-4: Among these k neighbours, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbour is maximum.
Step-6: Our model is ready.

Suppose we have a new data point and we need to put it in the required category.
Consider the below image:

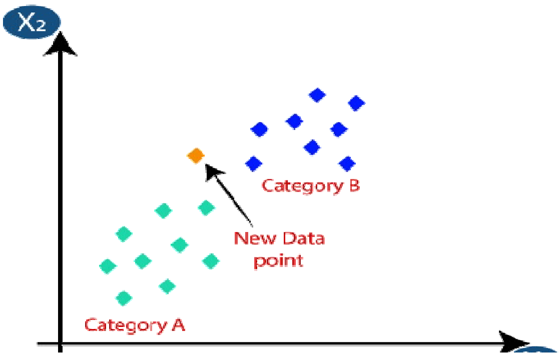


Fig.8 Example of K-NN Algorithm

Firstly, we will choose the number of neighbours, so we will choose the k=5.

Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:

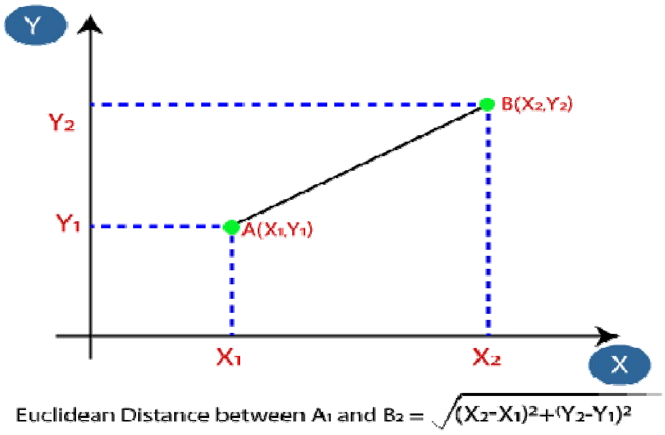


Fig.9 Calculation of Euclidean Distance

By calculating the Euclidean distance, we got the nearest neighbours, as three nearest neighbours in category A and two nearest neighbours in category B. Consider the below image:

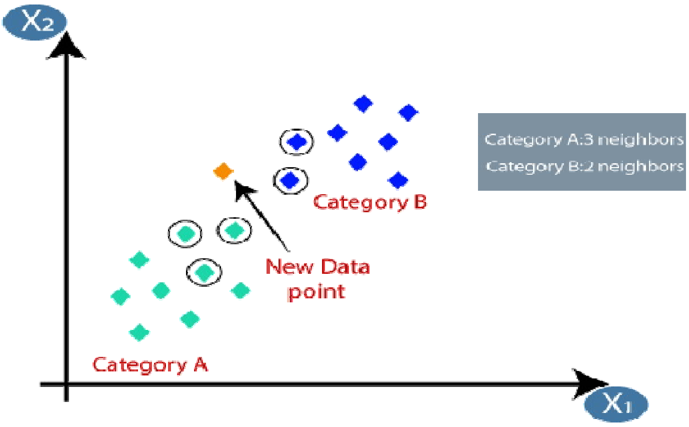


Fig.10 Example of Euclidean Distance

As we can see the 3 nearest neighbours are from category A, hence this new data point must belong to category A.Here we took *K* value is to pick the nearest distance value as output.

```
from sklearn.neighborsimport KNeighborsClassifier
knn=KNeighborsClassifier()
knn.fit(x_train,y_train)
pre_cls = knn.predict(x_test)
```

3.3 Calculation of Accuracy

Here the system can calculate accuracy between six supervised machine learning algorithms. For this we need to split the heart disease dataset with 70% as training dataset and remaining 30% as testing dataset can be done with help of *train_test_split()* method which is importing from *sklearn.model_selection*package. Therefore it can return the *x_train, x_test, y_train, y_test*parameters then by taking the inputs as *x_train*,*y_train*and build the training model with respective classifiers and get the predicted classes *pre_cls* by invoke the prediction function by taking input as *x_test*then finally calling *metrics.accuracy_score ()*withinput*y_test*and*pre_cls* then it returns the accuracy of each respective classifier.

Accuracy = (TP+TN) / (TP+FP+TN+FN)

True Positive (TP) signifies how many positive class samples your model predicted correctly.

True Negative(TN) signifies how many negative class samples your model predicted correctly.

False Positive (FP) signifies how many negative class samples your model predicted incorrectly. This factor represents Type-I error in statistical nomenclature. This error positioning in the confusion matrix depends on the choice of the null hypothesis.

False Negative (FN) Signifies how many positive class samples your model predicted incorrectly. This factor represents Type-II error in statistical nomenclature. This error positioning in the confusion matrix also depends on the choice of the null hypothesis.

3.4 Prediction

This module will be executed after building the training model with the respective best classifier. For heart disease prediction we need to invoke *predict ()*method with the testing dataset as input. This method will be available in every classifier. By calling this function it can start to compare with training dataset with the given testing dataset with the respective classifier and it returns the target column as output which matches to near with training dataset. This method is used for the prediction of heart disease as POSITIVE or NEGATIVE.

After executing the code, we will get a link, when we click on that link cardio care page will be opened, after main page we'll get our admin page in which we need to enter the username and password after giving the credentials we are logged in to our page and it will display welcome admin. First step is uploading the dataset. After that, the preprocessing stage will be started, in this we need to choose file from our system which is dataset.csv after uploading the dataset, it will give us the preprocessed dataset i.e. Not applicable (NA) values and null values are dropped using dropna () function, and itdisplays the original records and filtered records, then it displays 10 records from the entire dataset.After preprocessing we need to click on the Feature selection which is present on the top of the admin page.

In this stage, the system will automatically chooses the preprocessed dataset and it will select best attributesfrom 15 attributes and displays on the screen, they are sysbp, glucose, diabp, age, gender, Bpmeds, Diabetes, totchol, prevalenthyp, cigsperday, tenyearCHD. After getting selected attributes we need to click on the Evaluation which is on the top of the admin page, the system will evaluate all six machine learning algorithmsand it will display the evaluated table which consists of accuracy,precision,recall,F1score and it will plot a graph on this evaluation process which is accuracy of various techniques.

After evaluating, the system will select the best classification algorithm and that algorithm is used in prediction stage.Here we got Random Forest Classifier has best algorithm with highest accuracy and F1score, so the system will select RFC in prediction stage. After that we need to click on the Prediction which is on the top of the page.In this Prediction Stage, we need to enter the details of the patient, the details are the selected attributes. After entering the attributes, the RFC will process and predict the result as POSITIVE or NEGATIVE.

IV. SOFTWARE SPECIFICATIONS

Hardware requirements:

Processor	: Any Update Processor
Ram	: Min 4 GB
Hard Disk	: Min 100 GB

Software requirements:

Operating System	: Windows family
Technology	: Python 3.6
Front-end Technology	:HTML, CSS, JS

IDE	: PyCharm
Web framework	: Flask

V. SYSTEM TESTING AND UML DIAGRAMS

5.1 System Testing

The system testing can be classified into multiple parts such as unit testing and validation testing. The following steps given brief descriptions about two testing categories.

Unit testing

The unit testing is an important or initial testing tool in software testing. In this testing the shortest function or modules will be tested independently. In this testing developer and tester both can involve for testing the functionality and debugging the errors if any found and try to resolve those bugs at advance stage of SDLC.

Validation Testing

In the validation testing, tests the functionalities which compiles for getting expected results or not. These types of testing will be done at client side. The examples for validating with registration form, login form and any form filling scenario it can be used. When the validation testing is satisfied then only the user can go for another process of application functionalities. In this system we are using at dataset upload validation like admin should select the dataset for storing in database.

Integration testing

The testing is very useful and important of the software testing model. Here it can test the interaction between interfaces of various functionalities. As well as it can test with hardware, OS and with various software system for interaction of the system.

Transparent testing

The transparent testing is known also white box testing and it is also type of unit testing. This testing will be performing for components behavior of the software and operated by the programming developers to validate their functionality working flow.

Black-box Testing

In this testing, the tester does not aware of the internal process of functionalities. So that it can allow only testing the structure of the functionalities like the user interface.

5.2. UML diagrams

The UML diagrams are categorized into structural diagrams, behavioral diagrams, and also interaction overview diagrams. The diagrams are hierarchically classified in the following figure:

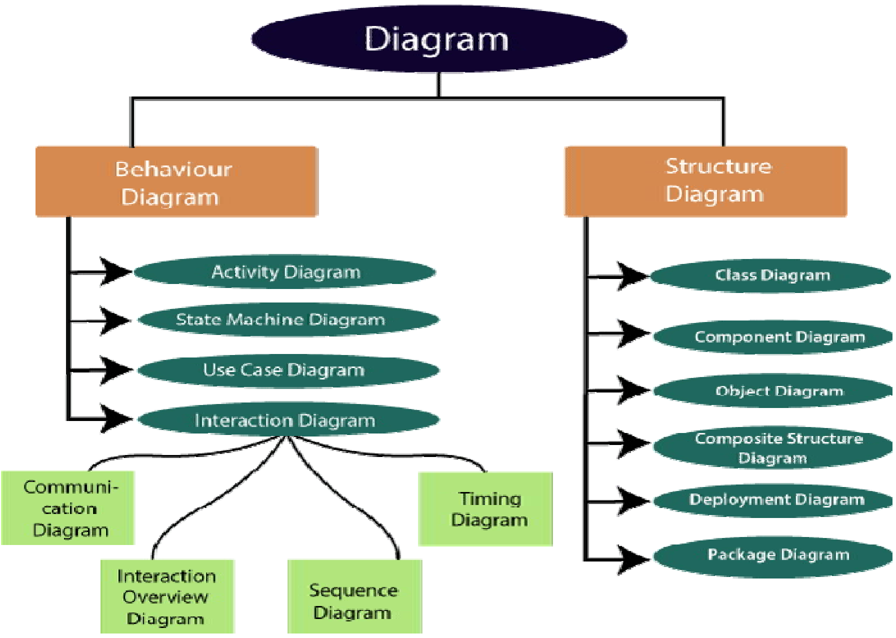


Fig 10 Flowchart of UML Diagrams

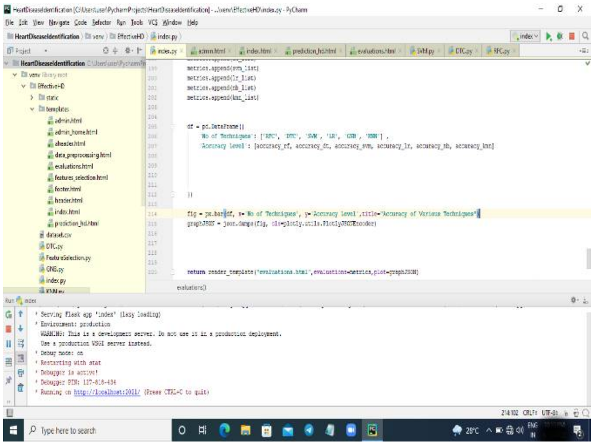
DOI: 10050086.2022.08.68

Use case:

Use case diagrams are a set of use cases, actors, and their relationships. They represent the use case view of a system.A use case represents a particular functionality of a system. Hence, use case diagram is used to describe the relationships among the functionalities and their internal/external controllers. These controllers are known as actors.

VI. RESULTS

When we execute our code and click on the click it gives us the following output.



```
metrics.append(svm_list)
metrics.append(lr_list)
metrics.append(nb_list)
metrics.append(knn_list)


df = pd.DataFrame({
    'No of Techniques': ['RFC', 'DTC', 'SVM', 'LR', 'GNB', 'KNN'] ,
    'Accuracy Level': [accuracy_rf, accuracy_dt, accuracy_svm, accuracy_lr, accuracy_nb, accuracy_knn]




})

fig = px.bar(df, x='No of Techniques', y='Accuracy Level',title="Accuracy of Various Techniques")
graphJSON = json.dumps(fig, cls=plotly.utils.PlotlyJSONEncoder)




return render_template("evaluations.html",evaluations=metrics,plot=graphJSON)
```

```
 * Serving Flask app 'index' (lazy loading)
 * Environment: production
   WARNING: This is a development server. Do not use it in a production deployment.
   Use a production WSGI server instead.
 * Debug mode: on
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 127-818-434
 * Running on http://localhost:2021/ (Press CTRL+C to quit)
```

Fig11 URL link for the Output

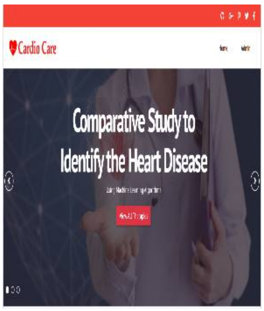


Fig. 12 Home Page

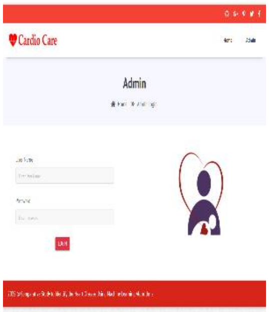
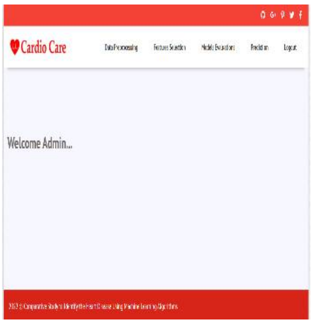


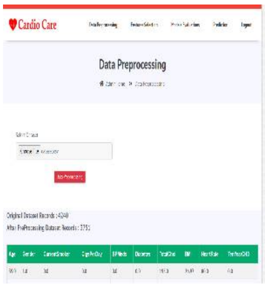
Fig 13 Admin login page



Welcome Admin...

Fig 14 Admin Home Page

Fig 12 shows the home page of the cardio care. After executing the code, we will get a link, when we click on that link cardio care page will be opened. Fig 13 shows the admin page, after main page we'll get our admin page in which we need to enter the admin details such as, username and password. Fig 14 after entering all details or credinals we are logged in to the main page, which shows welcome admin...

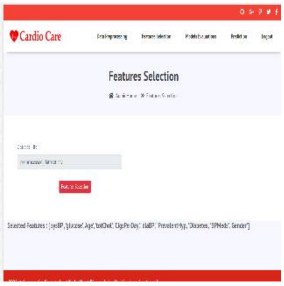


Original Dataset Records : 4240

After PreProcessing Dataset Records : 3751

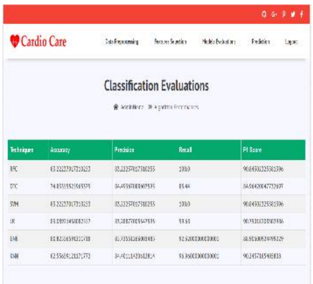
Age	Gender	CurrentSmoker	CigsPerDay	BPMeds	Diabetes	TotalChol	BMI	HeartRate	TenYearCHD
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Fig 15 Data Preprocessing



Selected Features : ['sysBP','glucose','Age','totChol','CigsPerDay','diaBP','PrevalentHyp','Diabetes','BPMeds','Gender']

Fig 16 Features Selection



Techniques	Accuracy	Precision	Recall	F1 Score
RFC	83.22237017310253	83.22237017310255	100.0	90.84502225581596
DTC	74.83555525565575	84.45556708807595	85.44	84.96420047732697
SVM	83.22237017310253	83.22237017310255	100.0	90.84502225581596
LR	83.08995458082557	83.28877005547595	99.68	90.75018338503386
ENB	80.82386591231718	83.73551326900485	92.62000000000001	88.50960624495229
KNN	82.55659121171772	84.40111420612814	96.36000000000001	90.1457165408833

Fig 17 Models Evaluations

Fig 15 describes the data preprocessing in this we have to choose the data file from our system.First step is uploading the dataset. After that, the preprocessing stage will be started, in this we need to choose file from our system which is dataset.csv after uploading the dataset, it will give us the preprocessed dataset i.e. Not applicable (NA) values and null values are dropped using dropna () function, and itdisplays the original records and filtered records, then it displays 10 records from the entire dataset. Fig 16 shows the 10 selected features among 15 features. After preprocessing we need to click on the Feature selection which is present on the top of the admin page. In this stage, the system will automatically choose the preprocessed dataset and it will select best attributesfrom 15 attributes and displays on the screen, they are sysbp, glucose, diabp,age, gender, Bpmeds, Diabetes, totchol,prevalenthyp, cigsperday, tenyearCHD. Fig 17 describes the evaluation of all algorithms that is accuracy, precision, recall and F1 score.

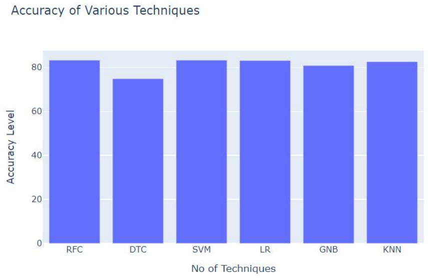
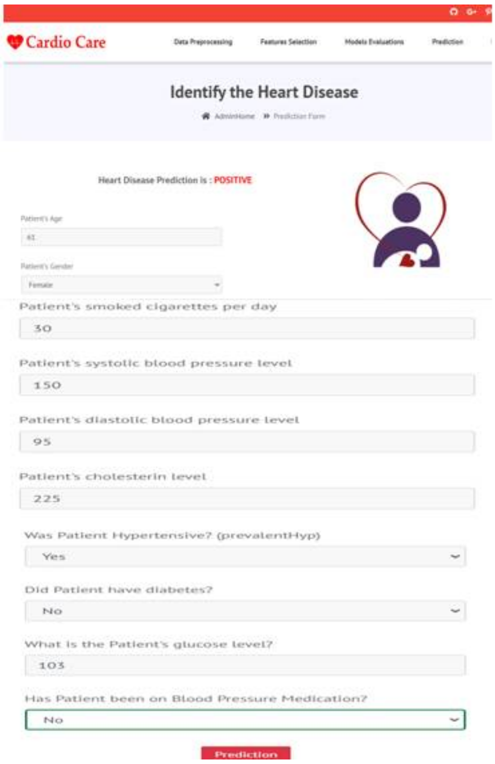


Fig 18 Classifiers accuracies

Fig 18 displays the accuracy graph of various techniques. After getting selected attributes we need to click on the Evaluation which is on the top of the admin page, the system will evaluate all six machine learning algorithms and it will display the evaluated table which consists of accuracy, precision, recall, F1 score and it will plot a graph on this evaluation process which is accuracy of various techniques.



Identify the Heart Disease

Heart Disease Prediction is : POSITIVE

Patient's Age
61

Patient's Gender
Female

Patient's smoked cigarettes per day
30

Patient's systolic blood pressure level
150

Patient's diastolic blood pressure level
95

Patient's cholesterin level
225

Was Patient Hypertensive? (prevalentHyp)
Yes

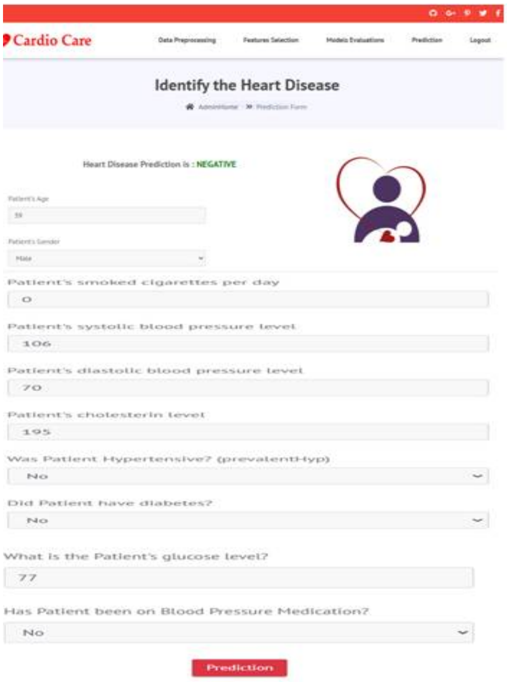
Did Patient have diabetes?
No

What is the Patient's glucose level?
103

Has Patient been on Blood Pressure Medication?
No

Fig 19 Output of Heart Disease Identification as POSITIVE

Fig 19 shows the prediction part, here we need to enter the details of the patient and it shows the result of the patient as POSITIVE.



Identify the Heart Disease

AdminHome » Prediction Form

Heart Disease Prediction is : NEGATIVE

Patient's Age
39

Patient's Gender
Male

Patient's smoked cigarettes per day
0

Patient's systolic blood pressure level
106

Patient's diastolic blood pressure level
70

Patient's cholesterin level
195

Was Patient Hypertensive? (prevalentHyp)
No

Did Patient have diabetes?
No

What is the Patient's glucose level?
77

Has Patient been on Blood Pressure Medication?
No

Fig 20 Output of Heart Disease Identification as NEGATIVE

Fig 20 shows the prediction part, here we need to enter the details of the patient and it shows the result of the patient as NEGATIVE.

After evaluating, the system will select the best classification algorithm and that algorithm is used in prediction stage.Here we got Random Forest Classifier has best algorithm with highest accuracy and F1score, so the system will select RFC in prediction stage. After that we need to click on the Prediction which is on the top of the page.In this Prediction Stage, we need to enter the details of the patient, the details are the selected attributes. After entering the attributes, the RFC will process and predict the result as POSITIVE or NEGATIVE.

VII. CONCLUSION

By taking the advantages of online world we have lot of medical history data is available. Extracting and analysis medical history data is become very necessary for the prediction of the diseases. Especially in heart diseases, the rate of deaths due to heart attacks is increasing day by day. This rate we can decrease by predicting disease by analyzing the heart patient’s medical history data. In this paper we propose a comparative analysis of heart disease prediction using popular classification algorithms. We classify and compare the results in terms of Accuracy calculation. Here we have used KNN, SVM, NB, LR, DT and Random Forest for classifying the heart attack medical data and calculate the accuracy score. In these algorithms we got 83% of accuracy for Random Forest algorithm. We deployed Random Forest algorithm for user heart disease predictions.

REFERENCES

[1] Prabakaran, G, Ni, R & Ramesh, M 2013, ‘A robust QR-code video watermarking scheme based on SVD and DWT composite domain’, Proceedings of the international conference on pattern recognition, informatics and mobile engineering, pp. 251-257.

[1] T.Nagamani, S.Logeswari, B.Gomathy,” Heart Disease Predictionusing Data Mining with Mapreduce Algorithm”, International Journalof Innovative Technology and Exploring Engineering (IJITEE) ISSN:2278-3075, Volume-8 Issue-3, January 2019.

[2] Fahd Saleh Alotaibi,” Implementation of Machine Learning Model toPredict Heart Failure Disease”, (IJACSA) International Journal ofAdvanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[3] Avinash Golande, Pavan Kumar T, ”Heart Disease Prediction UsingEffective Machine Learning Techniques”, International Journal ofRecent Technology and Engineering, Vol 8, pp.944-950,2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin,”Design And Implementation Heart Disease Prediction Using NaivesBayesian”,International Conference on Trends in Electronics andInformation(ICOEI 2019).

[5] Nagaraj M Lutimath,ChethanC,Basavaraj S Pol.,’Prediction Of HeartDisease using Machine Learning’, International journal Of RecentTechnology and Engineering,8,(2S10), pp 474-477, 2019.

[6] Theresa Princy R,J.Thomas,’Human heart Disease Prediction Systemusing Data Mining Techniques’, International Conference on CircuitPower and Computing Technologies,Bangalore,2016.

[7] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, ―MedicalData Mining for Heart Diseases and the Future of Sequential Mining inMedical Field,‖ in Machine Learning Paradigms, 2019, pp. 71–99.

[8] Puneet Bansal and Ridhi Saini et al. "Classification of heartdiseases from ECG signals using wavelet transform and kNNclassifier", International Conference on Computing,Communication and Automation (ICCCA2015).

[9] V. Krishnaiah, G. Narsimha, and N. Subhash, ''Heart disease prediction system using data mining techniques and intelligent fuzzyapproach: A review,'' Int. J. Comput. Appl., vol. 136, no. 2, pp. 43–51,2016.

[10] S. Radhimeenakshi, ''Classification and prediction of heart disease riskusing data mining techniques of support vector machine and artificialneural network,'' in Proc. 3rd Int. Conf. Comput. Sustain. Global Develop.(INDIACom), New Delhi, India, Mar. 2016, pp. 3107–3111.

[11] T. Vivekanandan and N. C. S. N. Iyengar, ''Optimal feature selectionusing a modified differential evolution algorithm and its effectiveness forprediction of heart disease,'' Comput. Biol. Med., vol. 90, pp. 125–136,Nov. 2017.

[12] J. Nahar, T. Imam, K. S. Tickle, and Y.-P. P. Chen, ''Association rulemining to detect factors which contribute to heart disease in malesand females,'' Expert Syst. Appl., vol. 40, no. 4, pp. 1086–1093, 2013.

[13] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ''Prediction ofheart disease using machine learning,'' in Proc. 2nd Int. Conf. Electron.,Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[14] R. Das, I. Turkoglu, and A. Sengur, ''Effective diagnosis of heart diseasethrough neural networks ensembles,'' Expert Syst. Appl., vol. 36, no. 4,pp. 7675–7680, May 2009.

[15] J. S. Sonawane and D. R. Patil, ''Prediction of heart disease using multilayer perceptron neural network,'' in Proc. Int. Conf. Inf. Commun. Embedded Syst., Feb. 2014, pp. 1–6