

A Synopsis On
Medical Diagnosis
For The Heart Using Prediction System

Submitted By
Rutuja Joshi – SYAIB18
Sagar Waghmare – SYAIB29
Parth Tilay – SYAIB58
Himanshu Dhere – SYAIB61

Guided By
Mr. Prof. Sharad Adsure

Department of Artificial Intelligence
G.H.Raisoni College of Engineering and Managment
Wagholi, Pune – 412207
2021-22

Contents

Sr. No.	Topics	Page No.
1	Abstract & Technical Keywords	
2	Introduction	
3	Literature Review/Related work	
4	Proposed Work and Objectives	
5	Methodology	
6	Desired Implications	
7	Conclusion	
8	References: (as per IEEE format)	

Mr. Sharad Adsure
Guide

Prof. Rachna Sable
H.O.D

ABSTRACT –

Heart diseases these days are main reason for deaths over the decade not only in India all over the world. Over the years researchers have developed several machine learning models for accurate and feasible system to diagnose such disease and could give treatment in time. Many researches in recent times have been using several machine learning techniques to help the health care industry and many heart surgeons in the diagnosis of heart related diseases. Our model acts like a predictive model for heart diagnosis diseases. So a quiet significant amount of pressure has been lift off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict the patients with heart disease. It is implemented on the.pynb format. Machine learning model gives a quite good approach as it learns as per input given to it, in our project we have used Regression a method of machine learning. The proposed model is quite satisfying and was able to predict evidence of having a heart disease by using Logistic Regression.

As we have used multiple algorithms such as Logistic Regression, Naïve Bayes, SVM & Random Forest Algorithms. Analysing data with different algorithms gives us a perfect prediction & accuracy.

INTRODUCTION –

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die. Change in lifestyle, work related stress and bad food habits contribute to the increase in rate of several heart related diseases.

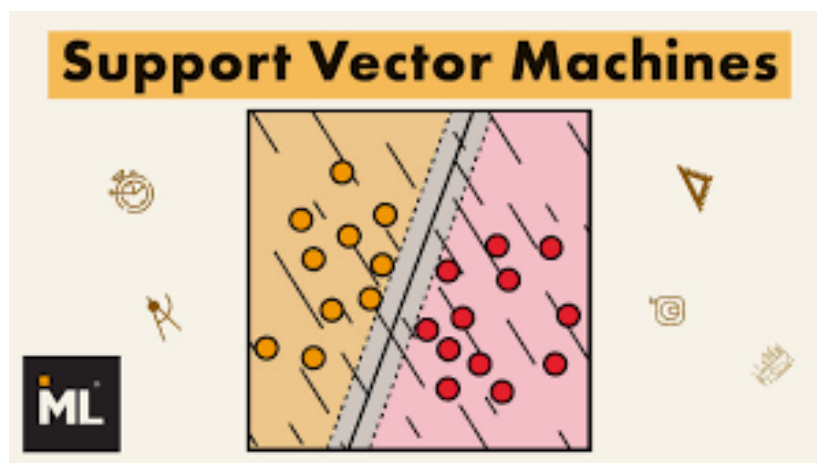
Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately.

A major challenge faced by health care organizations, such as hospitals and medical centres, is the provision of quality services at affordable costs. The quality service implies diagnosing patients properly and administering effective treatments. The available heart disease database consists of both numerical and categorical data. Before further processing, cleaning and filtering are applied on these records in order to filter the irrelevant data from the database. The proposed system can determine an exact hidden knowledge, i.e., patterns and relationships associated with heart disease from a historical heart disease database. It can also answer the complex queries for diagnosing heart disease; therefore, it can be helpful to health care practitioners to make intelligent clinical decisions. This machine learning model could help in estimating the probability of people having good hearts and people who have problems of heart, it helps taking important features from the dataset and making predictions based on these features. As these

machine learning model features are based on different algorithms which are used for analysing actual value from predicted value. These models are –

Support Vector Machine (SVM) –

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. There are specific types of SVMs you can use for particular machine learning problems, like support vector regression (SVR) which is an extension of support vector classification (SVC). The main thing to keep in mind here is that these are just math equations tuned to give you the most accurate answer possible as quickly as possible. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximizes the distance from the nearest data points of all the classes. The decision boundary created by SVMs is called the maximum margin classifier or the maximum margin hyper plane.



Logistic Regression –

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. Multinomial logistic regression can model scenarios where there are more than two possible discrete outcomes. Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. As aspects of cyber security are classification problems, such as attack detection, logistic regression is a useful analytic technique.



Random Forest Regression –

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and

Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.




Naïve Bayes –

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of

the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

Naïve Bayes Classifier

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$


Thomas Bayes
1702 - 1761

LITERATURE REVIEW –

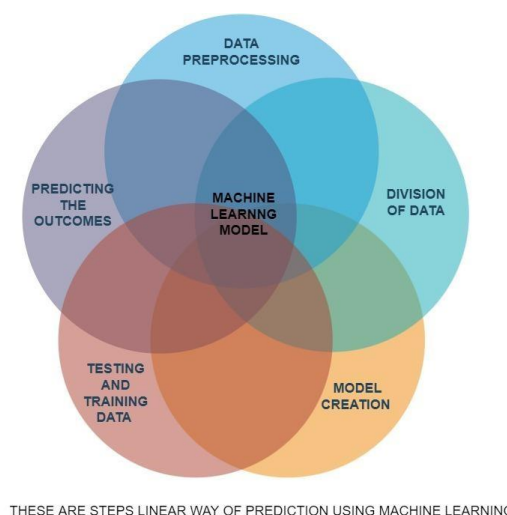
Human heart being the electro-mechanical pump supplies blood via a cardiovascular network. Its rhythmic beating gives rise to a pattern which when recorded can be used to find out the functionality of a heart. The diagnostic tool is called as Electrocardiogram (ECG) and its tracing contains a lot of attributes whose proper analysis may detect any cardiac peculiarity. Among them, is an entity called as the beat-to-beat interval (R-R interval). The analysis of beat to beat fluctuations of heart rate is known as heart rate variability (HRV) which is a concise marker to study the health of the heart along with a lot of measures clinically. This report talks about how the heart functioning is proper or whether is the heart healthy or any disease is measured inside it.

As technology is growing slowly & gradually it plays an important role in medical field, as artificial intelligence is used in surgeries and diagnosis treatment for better and quick outcome as patient suffering from “heart diseases” can be cured by exact medicines and treatment.

Machine Learning models are used in heart disease prediction, to make those prediction successfully accurate it uses different machine learning based algorithms for prediction on the given whole lot datasets, as it analysis the dataset categorize it for easy further implementations. While studying this project, we studied about the different heart diseases and how our technology is not that accurate or advance for these types of prediction, as different articles, IEEE papers were still incomplete, because these medical field models are critical to design as the prediction made by the model are different as the working algorithm changes. We tried implementing different algorithms on our same model and datasets & we got actual desired outcome.

PROPOSED WORK AND OBJECTIVES –

Medical has came so far in terms of capitalization and mastering in varied diagnosis. Thus however this has been only possible because of Machine Learning and Artificial Intelligence. However the objective of this system is to predict the heart Disease with as many data as possible from the datasets.



This model will work properly if established through a website and published online. However this model has very short datasets to work on, while making higher datasets in size, the model can be eligible to work for the real-time users and can simply display the data outcomes based on the previous features

Methodology

There are several methods or linear ways in which machine learning model predicts the outcome, as per the given input to model. There are linear ways in which we have to import data and commands the model –

- **Importing Datasets –**

The dataset consists of 1026 individual data and there are 14 columns in dataset

1. ***Age***: displays the age of the individual.
2. ***Sex***: displays the gender of the individual using the following format :
1 = male
0 = female
3. ***Chest-pain type***: displays the type of chest-pain experienced by the individual using the following format :
1 = typical angina
2 = atypical angina
3 = non — anginal pain
4 = asymptotic
4. ***Resting Blood Pressure***: displays the resting blood pressure value of an individual in mmHg (unit)
5. ***Serum Cholestrol***: displays the serum cholesterol in mg/dl (unit)

6. ***Fasting Blood Sugar***: compares the fasting blood sugar value of an individual with 120mg/dl.
If fasting blood sugar > 120mg/dl then : 1 (true)
else : 0 (false)
7. ***Resting ECG*** : displays resting electrocardiographic results
0 = normal
1 = having ST-T wave abnormality
2 = left ventricular hypertrophy
8. ***Max heart rate achieved*** : displays the max heart rate achieved by an individual.
9. ***Exercise induced angina*** :
1 = yes
0 = no
10. ***ST depression induced by exercise relative to rest***: displays the value which is an integer or float.
11. ***Peak exercise ST segment*** :
1 = upsloping
2 = flat
3 = downsloping
12. ***Number of major vessels (0–3) colored by flourosopy*** : displays the value as integer or float.
13. ***Thal*** : displays the thalassemia :
3 = normal
6 = fixed defect
7 = reversible defect
14. ***Diagnosis of heart disease*** : Displays whether the individual is suffering from heart disease or not :

0 = absence

1, 2, 3, 4 = present.

Why these parameters:

In the actual dataset, we had 76 features but for our study, we chose only the above 14 because :

1. Age: Age is the most important risk factor in developing cardiovascular or heart diseases, with approximately a tripling of risk with each decade of life. Coronary fatty streaks can begin to form in adolescence. It is estimated that 82 percent of people who die of coronary heart disease are 65 and older. Simultaneously, the risk of stroke doubles every decade after age 55.
2. Sex: Men are at greater risk of heart disease than pre-menopausal women. Once past menopause, it has been argued that a woman's risk is similar to a man's although more recent data from the WHO and UN disputes this. If a female has diabetes, she is more likely to develop heart disease than a male with diabetes.
3. Angina (Chest Pain): Angina is chest pain or discomfort caused when your heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in your chest. The discomfort also can occur in your shoulders, arms, neck, jaw, or back. Angina pain may even feel like indigestion.
4. Resting Blood Pressure: Over time, high blood pressure can damage arteries that feed your heart. High blood pressure that occurs with other conditions, such as obesity, high cholesterol or diabetes, increases your risk even more.

5. Serum Cholesterol: A high level of low-density lipoprotein (LDL) cholesterol (the “bad” cholesterol) is most likely to narrow arteries. A high level of triglycerides, a type of blood fat related to your diet, also ups your risk of a heart attack. However, a high level of high-density lipoprotein (HDL) cholesterol (the “good” cholesterol) lowers your risk of a heart attack.
6. Fasting Blood Sugar: Not producing enough of a hormone secreted by your pancreas (insulin) or not responding to insulin properly causes your body’s blood sugar levels to rise, increasing your risk of a heart attack.
7. Resting ECG: For people at low risk of cardiovascular disease, the USPSTF concludes with moderate certainty that the potential harms of screening with resting or exercise ECG equal or exceed the potential benefits. For people at intermediate to high risk, current evidence is insufficient to assess the balance of benefits and harms of screening.
8. Max heart rate achieved: The increase in cardiovascular risk, associated with the acceleration of heart rate, was comparable to the increase in risk observed with high blood pressure. It has been shown that an increase in heart rate by 10 beats per minute was associated with an increase in the risk of cardiac death by at least 20%, and this increase in the risk is similar to the one observed with an increase in systolic blood pressure by 10 mm Hg.
9. Exercise induced angina: The pain or discomfort associated with angina usually feels tight, gripping or squeezing, and can vary from mild to severe. Angina is usually felt in the center of your chest but may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands.
 - o Types of Angina
 - a. Stable Angina / Angina Pectoris
 - b. Unstable Angina
 - c. Variant (Prinzmetal) Angina
 - d. Microvascular Angina.
10. Peak exercise ST segment: A treadmill ECG stress test is considered abnormal when there is a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60–80 ms after the J point. Exercise ECGs with up-sloping ST-segment depressions are typically reported as an ‘equivocal’

test. In general, the occurrence of horizontal or down-sloping ST-segment depression at a lower workload (calculated in METs) or heart rate indicates a worse prognosis and higher likelihood of multi-vessel disease. The duration of ST-segment depression is also important, as prolonged recovery after peak stress is consistent with a positive treadmill ECG stress test. Another finding that is highly indicative of significant CAD is the occurrence of ST-segment elevation > 1 mm (often suggesting transmural ischemia); these patients are frequently referred urgently for coronary angiography.

Data Analysis

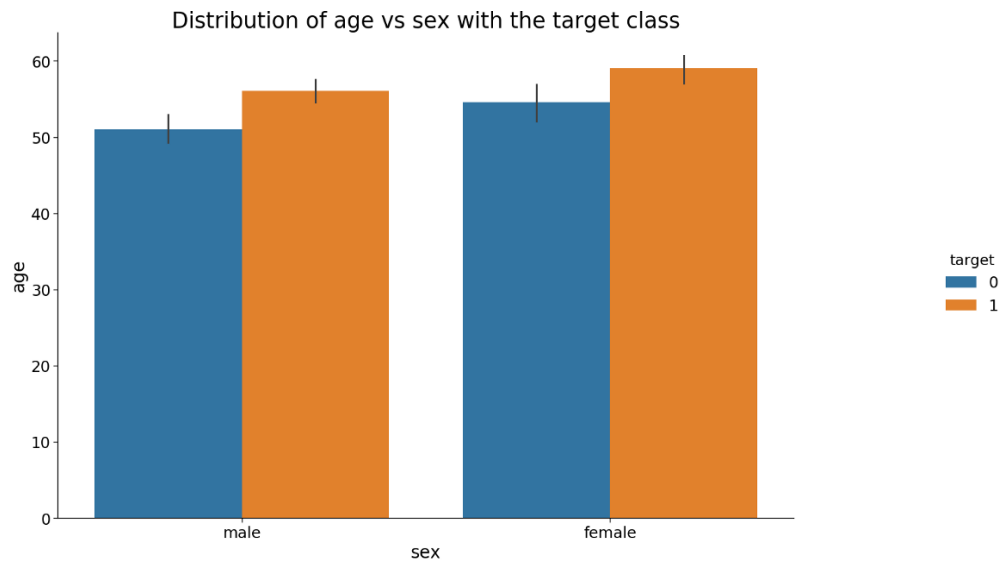
Let us look at the people's age who are suffering from the disease or not.

Here, target = 1 implies that the person is suffering from heart disease and target = 0 implies the person is not suffering.



We see that most people who are suffering are of the age of 58, followed by 57. Majorly, people belonging to the age group 50+ are suffering from the disease.

Next, let us look at the distribution of age and gender for each target class.



We see that for females who are suffering from the disease are older than males.

From this step we use different machine learning algorithms for predicting output:

Support Vector Machine (SVM) –

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. And we test accuracy of the model after using **Support Vector Machine (SVM)**

```
[ ] from sklearn.svm import SVC
    classifier = SVC(kernel = 'rbf')
    classifier.fit(X_train, y_train)

SVC()

[ ] # Predicting the Test set results
    y_pred = classifier.predict(X_test)

[ ] from sklearn.metrics import confusion_matrix
    cm_test = confusion_matrix(y_pred, y_test)

[ ]
    y_pred_train = classifier.predict(X_train)
    cm_train = confusion_matrix(y_pred_train, y_train)

    print()
    print('Accuracy for training set for svm = {}'.format((cm_train[0][0] + cm_train[1][1])/len(y_train)))
    print('Accuracy for test set for svm = {}'.format((cm_test[0][0] + cm_test[1][1])/len(y_test)))

Accuracy for training set for svm = 0.948780487804878
Accuracy for test set for svm = 0.9658536585365853
```

Naïve Bayes Algorithm –

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. And we test accuracy of the model after using Naïve Bayes Algorithm. The fundamental Naive Bayes assumption is that each feature makes an:

1. independent
2. equal

contribution to the outcome. With relation to our dataset, this concept can be understood as:

We assume that no pair of features are dependent. For example, the temperature being ‘Hot’ has nothing to do with the humidity or the outlook being ‘Rainy’ has no effect on the winds. Hence, the features are assumed to be independent.

Secondly, each feature is given the same weight(or importance). For example, knowing only temperature and humidity alone can’t predict the outcome accurately. None of the attributes is irrelevant and assumed to be contributing equally to the outcome.


```
[ ] X = df.iloc[:, :-1].values
    y = df.iloc[:, -1].values

[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

[ ] from sklearn.naive_bayes import GaussianNB
    classifier = GaussianNB()
    classifier.fit(X_train, y_train)

GaussianNB()

[ ] # Predicting the Test set results
    y_pred = classifier.predict(X_test)

[ ] from sklearn.metrics import confusion_matrix
    cm_test = confusion_matrix(y_pred, y_test)

[ ] y_pred_train = classifier.predict(X_train)
    cm_train = confusion_matrix(y_pred_train, y_train)
    print()
    print('Accuracy for training set for Naive Bayes = {}'.format((cm_train[0][0] + cm_train[1][1])/len(y_train)))
    print('Accuracy for test set for Naive Bayes = {}'.format((cm_test[0][0] + cm_test[1][1])/len(y_test)))

Accuracy for training set for Naive Bayes = 0.8207317073170731
Accuracy for test set for Naive Bayes = 0.8536585365853658
```

Logistic Regression –

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

And we test accuracy of the model after using **Logistic regression**.

```
[ ] X = df.iloc[:, :-1].values
    y = df.iloc[:, -1].values

[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

[ ] from sklearn.linear_model import LogisticRegression
    classifier = LogisticRegression()
    classifier.fit(X_train, y_train)

/usr/local/lib/python3.7/dist-packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
https://scikit-learn.org/stable/modules/linear\_model.html#logistic-regression
extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
LogisticRegression()

▶ # Predicting the Test set results
y_pred = classifier.predict(X_test)

from sklearn.metrics import confusion_matrix
cm_test = confusion_matrix(y_pred, y_test)

[ ] y_pred_train = classifier.predict(X_train)
    cm_train = confusion_matrix(y_pred_train, y_train)

print()
print('Accuracy for training set for Logistic Regression = {}'.format((cm_train[0][0] + cm_train[1][1])/len(y_train)))
print('Accuracy for test set for Logistic Regression = {}'.format((cm_test[0][0] + cm_test[1][1])/len(y_test)))

Accuracy for training set for Logistic Regression = 0.8621951219512195
Accuracy for test set for Logistic Regression = 0.8634146341463415
```

Random Forest Regression –

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. And we test accuracy of the model after using **Random Forest**.

```
[ ] X = df.iloc[:, :-1].values
    y = df.iloc[:, -1].values

[ ] from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 0)

[ ] from sklearn.ensemble import RandomForestClassifier
    classifier = RandomForestClassifier(n_estimators = 10)
    classifier.fit(X_train, y_train)

RandomForestClassifier(n_estimators=10)

[ ] y_pred = classifier.predict(X_test)

from sklearn.metrics import confusion_matrix
cm_test = confusion_matrix(y_pred, y_test)

[ ] y_pred_train = classifier.predict(X_train)
    cm_train = confusion_matrix(y_pred_train, y_train)

[ ] print()
print('Accuracy for training set for Random Forest = {}'.format((cm_train[0][0] + cm_train[1][1])/len(y_train)))
print('Accuracy for test set for Random Forest = {}'.format((cm_test[0][0] + cm_test[1][1])/len(y_test)))

Accuracy for training set for Random Forest = 1.0
Accuracy for test set for Random Forest = 1.0
```

DESIRED IMPLICATIONS –

Based on the above project, it can be concluded that there is a huge scope for machine learning algorithms in predicting cardiovascular diseases or heart related diseases. Each of the abovementioned algorithms have performed extremely well in some cases but poorly in some other cases. Systems based on machine learning algorithms and techniques have been very accurate in predicting the heart related diseases but still there is a lot scope of research to be done on how to handle high dimensional data and overfitting. A lot of research can also be done on the correct ensemble of algorithms to use for a particular type of data.

Conclusion –

Thus, by differentiating a model into 4 distinct machine learning models help us predict the heart disease for a person. It helps us to detect the heart disease for the given person by comparing the datasets data and the given values. Hence this is a system created only on the model. Thus, this model can be further implemented on the system using and adding web development models into it.

REFERENCES –

1. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. *Int J Comput Sci Net Secur*. 2008;**8**:343–350. [[Google Scholar](#)]
2. Sayad AT, Halkarnikar PP. Diagnosis of heart disease using neural network approach. *Int J Adv Sci Eng Technol*. 2014;**2**:88–92. [[Google Scholar](#)]
3. Gudadhe M, Wankhade K, Dongre S. Decision support system for heart disease based on support vector machine and Artificial Neural Network. Computer and

Communication Technology (ICCCT), 2010 International Conference on; 2010. pp. 741–745. [[Google Scholar](#)]

4. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating error. *Nature*. 1986;**323**:533–536. [[Google Scholar](#)]
5. [1] Ramadoss and Shah B et al. “A. Responding to the threat of chronic diseases in India”. *Lancet*. 2005; 366:1744–1749. doi: 10.1016/S0140-6736(05)67343-6.
6. [2] Global Atlas on Cardiovascular Disease Prevention and Control. Geneva, Switzerland: World Health Organization, 2011
7. [3] Dhomse Kanchan B and Mahale Kishor M. et al. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis”, 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication.
8. [4] R.Kavitha and E.Kannan et al. “An Efficient Framework for Heart Disease Classification using Feature Extraction and Feature Selection Technique in Data Mining “, 2016
9. [5] Shan Xu ,Tiangang Zhu, Zhen Zang, Daoxian Wang, Junfeng Hu and Xiaohui Duan et al. “Cardiovascular Risk Prediction Method Based on CFS Subset Evaluation and Random Forest Classification Framework”, 2017 IEEE 2nd International Conference on Big Data Analysis.
10. [6] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis and Vijay K. Mago et al. “ Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map”, 978-1-5090-0626-7/16/\$31.00 c 2016 IEEE.
11. [7] Kanika Pahwa and Ravinder Kumar et al. “Prediction of Heart Disease Using Hybrid Technique For Selecting Features”, 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON).