A Novel Approach to Predict Cardio Disease Using Naive Bayes Algorithm

P Praveen, Department of CS & AI, SR University, Warangal, Telangana State, India, prawin1731@gmail.com

Virinchi bethanamudi 18K41A0542@sru.edu.in
Bhavitha Ramadugu 18K41A0546@sru.edu.in
Gursimran kour sardar 18K41A0548@sru.edu.in
Nimisha rebelly 18K41A0552@sru.edu.in

Abstract

The project aims to generate a model to predict the likelihood of an individual being affected by heart disease in the next 10 years. Many hospitals generate data about patients that can identify early symptoms by identifying the patterns and facts present in them. This huge data makes it difficult for the Doctors to identify patterns. As the old saying goes "Prevention is better than Cure", early detection and continuous management by clinicians will bring down the mortality rate. All that said, it's impractical to observe patients daily to accurately predict and provide 24x7 consultations, since it requires heaps of knowledge and time. During this paper we have developed and researched models to predict heart disease through multifarious heart attributes of patients and identified machine learning techniques just round the corner to effectively predict heart diseases. Data Science, with its vast applied capabilities, has a key role in processing mammoth amounts of data to extract meaningful insights from it in every walk of life, with specific mention of healthcare services.

In this paper the Naive Bayes Algorithm is used and it is built on the Framingham dataset. And this data is divided as training of data and testing of data in ratio of 80:20. The training set is used to train the classifier and test data is used to generate a confusion matrix to calculate accuracy. Several techniques can be used to solve this problem. The proposed Naive Bayes algorithm has achieved an accuracy of approximately 81.5% on the testing data, which is greater than Decision Tree accuracy.

Keywords: Classification, Clustering, Naive-Bayes, Decision-tree, Random Forest Classifier, Machine Learning

Introduction

Heart disease has become a prevalent disease known to everyone these days as it accounts for 31% of all deaths globally. Annually there are more deaths caused by heart diseases than deaths caused by any other disease. Every year around 18 million people die because of heart diseases. Of these 18 million deaths, 3/4 of deaths occur in countries whose income falls in the low and middle-income group.

Why should heart disease be considered a major concern? Because even young people in the age group 20-30 years are succumbing to heart infections. The expansion of coronary illness among youngsters can be credited to countless reasons like dietary patterns, lack of sleep, sadness, stoutness, unequal and unfavourable eating routine, genetic, high BP, cholesterol, smoking, liquor utilization other than other way of life habits. Heart disease is even considered a silent killer that results in the death of a person with no obvious symptoms.

Heart disease was the key explanation for casualties across the globe including India, which kills one person every 34 seconds. A study shows that from 1990 to 2016 the deaths because of heart problems has risen by approximately 34% from 155.7 to 209.1 deaths per one lakh population in India. In India, deaths because of heart diseases are increasing at an alarming rate. A total of 18,309 casualties were recorded in India for 2014. These numbers skyrocketed and accounted for 28,005 deaths in 2019, which is approximately 53% increase from 2014.

Finding a way to prevent heart diseases has become imperative, given the exponential increase in heart ailments and their associated mortality rate. Early scientific prediction and detection seconded by medical data of heart condition play an essential role in creating choices on lifestyle changes in high-risk patients and successively cutting back the complications. Hence, improving the overall data analysis, to ensure individual well-being is the need of the hour, this is where Machine Learning pops up as the way forward. AI helps in foreseeing heart infections, and furthermore the expectations made are a unit very right.

The venture expects to consider and investigate whether a patient is inclined to be determined to have any heart sicknesses given their clinical qualities like sex, age, torment, fasting sugar level, and so forth A dataset is browsed the UCI vault with the patient's clinical record and characteristics. This dataset is utilized to foresee whether a patient will have a heart condition. 15 clinical traits of a patient will go about as the reason for us to anticipate and arrange him/her, to track down a patient's likelihood to get influenced by a heart condition. These clinical properties are prepared under three calculations:

- Naive-Bayes
- Decision-tree
- Random Forest Classifier.

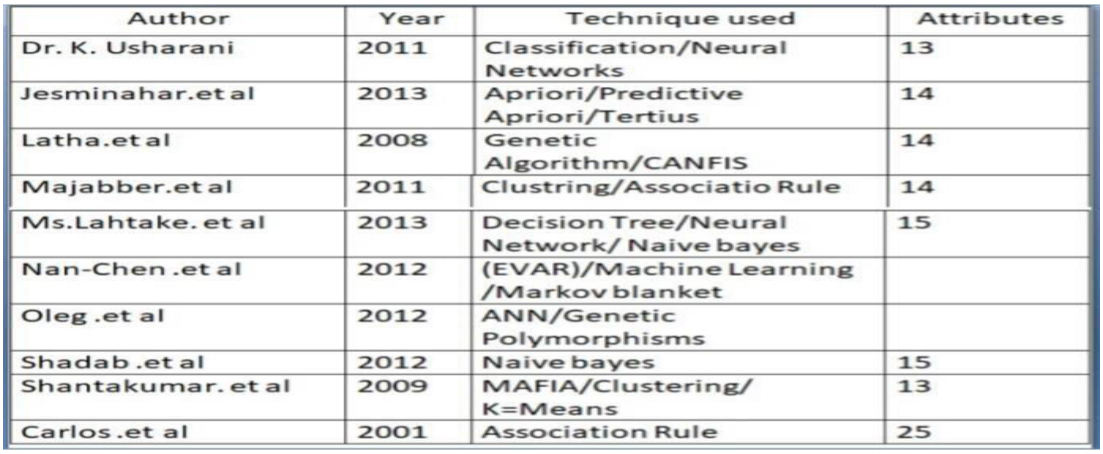
The classification algorithm built should be trained and tested to predict whether an individual will be affected by heart condition or not.

Related Work

Heart disease describes a range of medical conditions that affect the heart. These conditions are a reflection of the abnormal health and behaviour of the heart, with varied symptoms. Maintaining a healthy heart should be the primary priority. Hence, various data mining techniques have been devised and used in recent years for disease prediction, as stepping in realize the well-being of the heart.

The table shows diverse information mining strategies utilized in the forecast of coronary illness over the distinctive coronary illness datasets. In certain papers, only one technique is used to predict heart problem on the other hand in other papers used more than one data mining technique.

TABLE 1: The table below shows data from different Publications used for the prediction of heart disease by different authors with details of published year, technique, and the number of attributes used.



Author	Year	Technique used	Attributes
Dr. K. Usharani	2011	Classification/Neural Networks	13
Jesminahar.et al	2013	Apriori/Predictive Apriori/Tertius	14
Latha.et al	2008	Genetic Algorithm/CANFIS	14
Majabber.et al	2011	Clustring/Associatio Rule	14
Ms.Lahtake. et al	2013	Decision Tree/Neural Network/ Naive bayes	15
Nan-Chen .et al	2012	(EVAR)/Machine Learning /Markov blanket	
Oleg .et al	2012	ANN/Genetic Polymorphisms	
Shadab .et al	2012	Naive bayes	15
Shantakumar. et al	2009	MAFIA/Clustering/ K=Means	13
Carlos .et al	2001	Association Rule	25

Existing Methods

Classification using Random Forest

These are a blend of trees that foresee an exploitation call tree any place each tree relies upon the upsides of examined information severally and with a steady circulation for a woods. There ought to be no connection between the choice tree utilized in an arbitrary backwoods. Irregular Forest might be a regulated algorithmic standard utilized for the expectation and it is picked because of the improved exactness that is a direct result of the presence of an assortment of trees. These trees square measure a unit prepared individually thus the expectations of these trees square measure a unit joined thus a definitive yield is picked by abuse of the common cost. Irregular Forest calculation can be utilized for the order and the relapse issues dependent on our prerequisite. The calculation for an irregular backwoods is given underneath:

Stage 1: From m highlights haphazardly select k highlights, fulfilling the condition k << m.

Stage 2: For the k highlights, utilizing the best parted point ascertain the hub "d".

Stage 3: utilizing the best Split we need to divide the Nodes.

Stage 4: Repeat 1 to 3 stages until the necessary number of hubs has been reached.

Stage 5: Construct backwoods by rehashing stages 1 to 4 for n number occasions to make n number of trees.

To start with, the k alternatives, the square measure, are removed from the complete m choices. In the following stage, in each tree, arbitrarily select k highlights to discover the root hub by utilizing the best parted methodology. The following stage includes ascertaining the girl hubs utilizing a similar best split methodology for the coronary illness dataset. Essentially, the tree is produced using the premise hub and till all the leaf hubs square measure created from the properties. This arbitrarily made tree shapes the irregular timberland that is utilized for making the sickness forecast in patients.

Classification using Decision Tree:

A Decision Tree is a basic and simplest calculation to execute. It initially chooses a trait as a root hub, at that point computes either the data acquire or a gin file to recognize powerful hubs in the tree structure. A choice tree assembles different models as a tree structure simplifying it to comprehend and examine the information. The calculation works by figuring the data acquired for each characteristic present in the dataset and dividing the qualities by utilizing acquired values. The calculation for the choice tree is:

Step1: An information gain is calculated for all attributes present in the dataset.
Step2: An information gain calculated for the attributes should be sorted in descending order.
Step3: Assign the attribute with the high information gain to the root node of the tree.
Step4: In the same way gain should be calculated.
Step5: Split the nodes by comparing gain values.
Step6: Repeat this process until every attribute present in dataset becomes a leaf node in the tree.

Problem Statement

The effects of heart disease can be reduced effectively by following a healthy lifestyle, using effective medicine and if required surgeries. The symptoms of heart disease can be found early then the above steps should be taken to avoid future complications. To find the symptoms we use data and predict, so that heart disease can be prevented.

The predominant objective of our work is to build a model that will predict the likelihood of an individual being affected by heart disease in the next 10 years. Data mining is changing the healthcare industry by enabling health systems to use data and analytics to identify the best practices improving care and reduce costs by identifying symptoms of particular disease before it becomes severe. According to scholars by using Machine learning techniques in healthcare about 31% of healthcare spending can be reduced.

4.1 Algorithm with Description

- **Naive Bayes classification:**

Naive Bayes Algorithm works on the principle Bayesian theorem, which calculates the probability of an event given the other event as occurred. Naive Bayes assumes that there is no relationship between the attributes.

The method of naive Bayes algorithm is:

- **Training Step:** It expects credits to be autonomous restrictively and assesses likelihood appropriation "Contingent likelihood" from the preparation information.
- **Prediction Step**: For test information, the strategy ascertains the likelihood of the dataset which has a place with each class. The strategy at long last characterizes the test information dependent on the likelihood determined.

Why Naive Bayes algorithm?

- This algorithm can be implemented if the data is very large.
- If there's no relationship between the attributes (independent of each other).
- if we want best output.

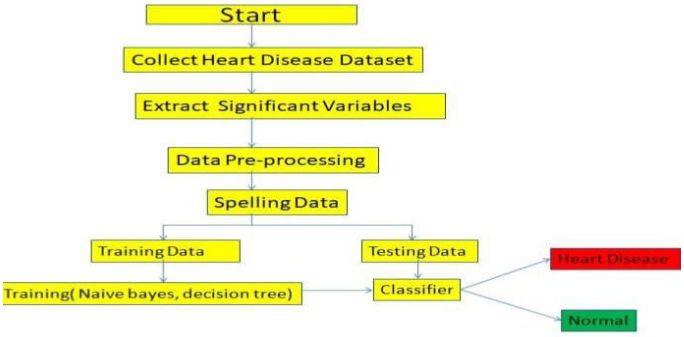
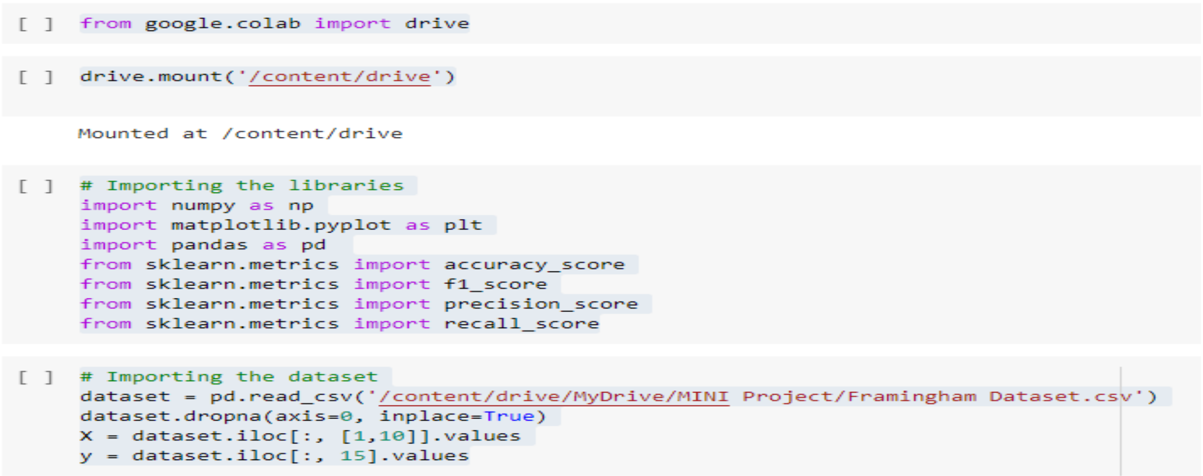
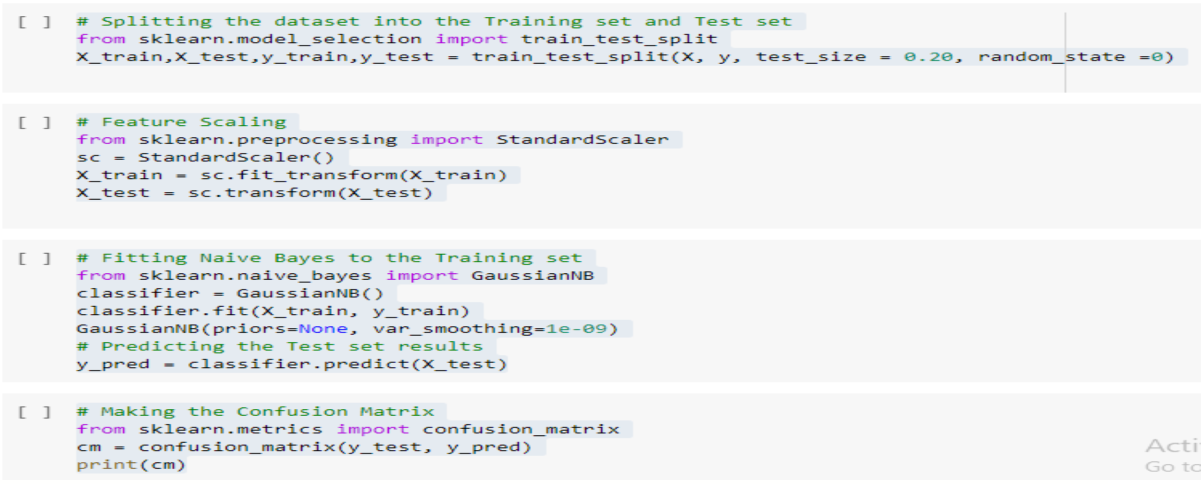

Start
Collect Heart Disease Dataset
Extract Significant Variables
Data Pre-processing
Spelling Data
Training Data
Testing Data
Training(Naive bayes, decision tree)
Classifier
Heart Disease
Normal

Figure 1: Process of Naive Bayes Algorithm

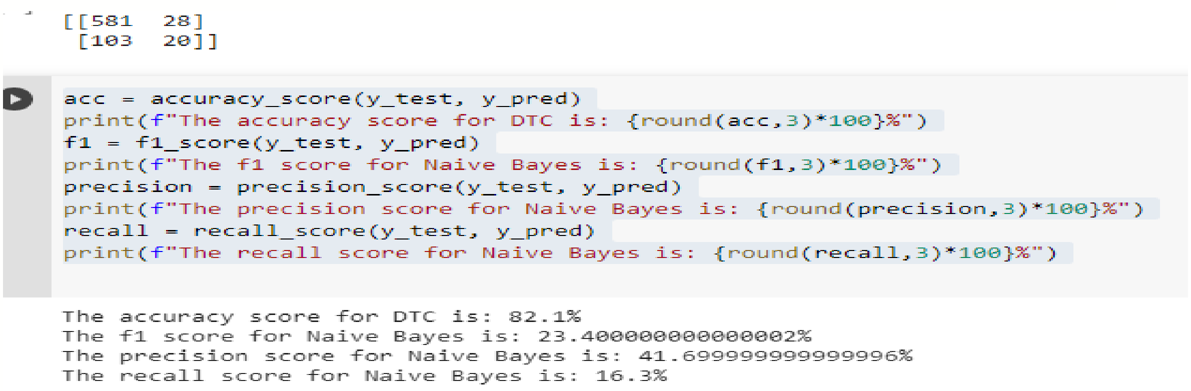
4.2 Execution



```
[ ]  from google.colab import drive

[ ]  drive.mount('/content/drive')

     Mounted at /content/drive

[ ]  # Importing the libraries
     import numpy as np
     import matplotlib.pyplot as plt
     import pandas as pd
     from sklearn.metrics import accuracy_score
     from sklearn.metrics import f1_score
     from sklearn.metrics import precision_score
     from sklearn.metrics import recall_score

[ ]  # Importing the dataset
     dataset = pd.read_csv('/content/drive/MyDrive/MINI Project/Framingham Dataset.csv')
     dataset.dropna(axis=0, inplace=True)
     X = dataset.iloc[:, [1,10]].values
     y = dataset.iloc[:, 15].values
```



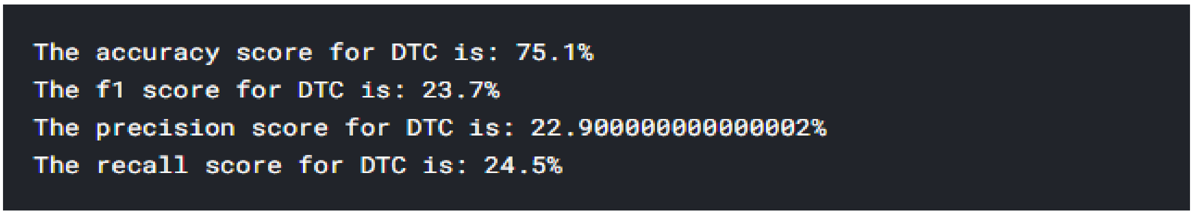
```
[ ]  # Splitting the dataset into the Training set and Test set
     from sklearn.model_selection import train_test_split
     X_train,X_test,y_train,y_test = train_test_split(X, y, test_size = 0.20, random_state =0)

[ ]  # Feature Scaling
     from sklearn.preprocessing import StandardScaler
     sc = StandardScaler()
     X_train = sc.fit_transform(X_train)
     X_test = sc.transform(X_test)

[ ]  # Fitting Naive Bayes to the Training set
     from sklearn.naive_bayes import GaussianNB
     classifier = GaussianNB()
     classifier.fit(X_train, y_train)
     GaussianNB(priors=None, var_smoothing=1e-09)
     # Predicting the Test set results
     y_pred = classifier.predict(X_test)

[ ]  # Making the Confusion Matrix
     from sklearn.metrics import confusion_matrix
     cm = confusion_matrix(y_test, y_pred)
     print(cm)
```



```
[[581  28]
 [103  20]]
```

```
acc = accuracy_score(y_test, y_pred)
print(f"The accuracy score for DTC is: {round(acc,3)*100}%")
f1 = f1_score(y_test, y_pred)
print(f"The f1 score for Naive Bayes is: {round(f1,3)*100}%")
precision = precision_score(y_test, y_pred)
print(f"The precision score for Naive Bayes is: {round(precision,3)*100}%")
recall = recall_score(y_test, y_pred)
print(f"The recall score for Naive Bayes is: {round(recall,3)*100}%")
```

```
The accuracy score for DTC is: 82.1%
The f1 score for Naive Bayes is: 23.400000000000002%
The precision score for Naive Bayes is: 41.699999999999996%
The recall score for Naive Bayes is: 16.3%
```



```
The accuracy score for DTC is: 75.1%
The f1 score for DTC is: 23.7%
The precision score for DTC is: 22.900000000000002%
The recall score for DTC is: 24.5%
```

The utilization of initial two lines of the code is to offer admittance to collab to get to the dataset, which is put away in google drive. Furthermore, the third piece of the code is to import python libraries.

⇨ In the fourth part dataset is perused utilizing read _csv utilizing it's area and put away in a variable called 'dataset'. Also, dropna work is utilized to erase the columns which contains "NA" values. At that point independent factors are put away in "X" and target variable is put away in "y".

⇨ In the fifth part, train _test _split is utilized to part the information into 80% of preparing and 20% test data. Next include scaling is done to standardize credits, with the end goal that their reach is 0 to 1. In seventh part, Gaussian NB strategy, which is imported from s-k learn.

⇨ Naïve-Bayes is utilized to prepare the classifier utilizing classifier.fit() technique and it is utilized to foresee the y esteems and store them in "y _pred".

⇨ At that point y _test and y _pred are utilized to build disarray lattice. Which is utilized to ascertain exactness, f1_score, accuracy and review. For these estimations, in-fabricated strategies are utilized.

5 Results

1. The built model predicts whether the individual is affected by heart disease or not(i.e., target variable). Once the model classifies an individual as probable for heart disease then care can be taken like early treatment and a good diet. So that deaths due to heart disease can be controlled or reduced. The accuracy of the Naive Bayes algorithm which was built using the Framingham dataset is 82%.
2. The built model output binary values 0 & 1.
   - 1 - Represents “Risk”.
   - 0 - Represents “no Risk”.
3. The accuracy for the model built by Naive Bayes algorithm is 82% whereas accuracy for model built using the Decision tree algorithm is 75%. In this case, Naive Bayes outperforms the Decision Tree Algorithm.
4. But the accuracy as of Random forest is about 88%. Naive Bayes Algorithm cannot outperform Random Forest because it uses the output from many decision trees and calculates the majority votes of prediction to generate the output of the random forest.

6 Conclusion

In this project Naive Bayes algorithm is implemented for the prediction of heart disease. But it is analyzed that only marginal accuracy is achieved for the predictive model of heart disease and hence more complex models are needed to increase the accuracy of predicting the early heart disease. The accuracy achieved for Naive Bayes algorithm is 82% which is greater than the accuracy of Decision tree. But when it comes to comparison with Random forest, it has low accuracy because Random forest is a more sophisticated than Naive Bayes algorithm.

In the near future, we may have different methods proposed for the early prediction of heart disease with high accuracy.

References

1. Senthilkumar Mohan, ChandrasegarThirumalai, Gautam Srivastava Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Digital Object Identifier 10.1109/ACCESS.2019.2923707, IEEE Access, VOLUME 7, 2019 S.P. Bingulac, On the Compatibility of Adaptive Controllers, Proc. Fourth Ann. Allerton Conf. Circuits and Systems Theory, pp. 8-16, 1994. (Conference proceedings)
2. SonamNikhar, A.M. Karandikar Prediction of Heart Disease Using Machine Learning Algorithms International Journal of Advanced Engineering, Management and Science (IJAEMS) Infogain Publication, [Vol-2, Issue-6, June - 2016]. I.s. Jacobs and C.P. Bean, Fine particles, thin films and exchange anisotropy, in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
3. Aditi Gavhane, GouthamiKokkula, Isha Pandya, Prof. Kailas Devadkar (PhD), Prediction of Heart Disease Using Machine Learning, Proceedings of the 2nd International

conference on Electronics, Communication and Aerospace Technology (ICECA 2018). IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1- 5386-0965-1

4. Abhay Kishore1, Ajay Kumar2, Karan Singp, Maninder Punia4, Yogita Hambir5, Heart Attack Prediction Using Deep Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04 | Apr-2018.
5. A.Lakshmanarao, Y.Swathi, P.Sri Sai Sundareswar, Machine Learning Techniques For Heart Disease Prediction, International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019
6. B. Rama, P. Praveen, H. Sinha and T. Choudhury, "A study on causal rule discovery with PC algorithm," 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017, pp. 616-621.doi: 10.1109/ICTUS.2017.8286083.
7. AvinashGolande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of
8. Recent Technology and Engineering (IJRTE) ISN: 2277-3878, Volume-8, Issue-1S4, June 2019.
9. V.V. Ramalingam, AyantanDandapath, M Karthik Raja, Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology, 7 (2.8) (2018) 684-687.
10. Praveen., P and Ch. Jayanth Babu. “Big Data Clustering: Applying Conventional Data Mining Techniques in Big Data Environment.” (2019). Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 74, ISSN 2367-3370, https://doi.org/10.1007/978-981-13-7082-3_58 Springer Singapore.
11. M. S. Amin, Y. K. Chiam, K. D. Varathan,Identication of signicant features and data mining techniques in predicting heart disease, Telematics Inform., vol. 36, pp. 8293, Mar.2019.
12. R Ravi Kumar M Babu Reddy P Praveen, “An Evaluation Of Feature Selection Algorithms In Machine Learning” International Journal Of Scientific & Technology Research Volume 8, Issue 12, December 2019 ISSN 2277-8616,PP. 2071-2074
13. N. Al-milli, Backpropagation neural network for prediction of heart disease, J. Theor. Appl.Inf. Technol., vol. 56, no. 1, pp.131135, 2013.
14. A. S. Abdullah and R. R. Rajalaxmi, A data mining model for predicting the coronary heart disease using random forest classier, in Proc. Int. Conf. Recent Trends Comput. Methods, Commun. Controls, Apr. 2012, pp. 2225.
15. P. Praveen, C. J. Babu and B. Rama, "Big data environment for geospatial data analysis," 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2016, pp. 1-6. doi: 10.1109/CESYS.2016.7889816