International Journal of Future Generation Communication and Networking
Vol.14, No. 1, (2021), pp. 4086 – 4096

Web Based Machine Learning Application for Heart Disease Prediction

Lokesh M. Giripunje[1], Tejas P. Sonar[2], Rohit S. Mali[3], Jayant C. Modhave[4], Mahesh B. Gaikwad[5]

[1] Assistant Professor, [2, 3, 4, 5]Students.

Department of Electronics and Telecommunication, Savitribai Phule Pune University, Pune, Maharashtra, India.

[1]lokeshgiripunje@gmail.com

[2]tejassonar24@gmail.com

[3]rohitmali72667@gmail.com

[4]jayantmodhave2@gmail.com

[5]maheshgaikwad193@gmail.com

Abstract

In the healthcare sector, doctors are facing many issues for predicting diseases. Coronary heart disease is one of them and it should be managed precisely and efficiently. Hospitals & clinics are offering costly therapies and operations to treat heart disease. Therefore, prediction of heart diseases at an early stage will give insight to doctors. Optimized solution for this problem can be solved by machine learning techniques. Machine learning in healthcare helps to analyze the tons of data & provide precise outcomes. In this work we have referred to some previous papers and improved their accuracy. And then compared the accuracies of machine learning algorithms like Decision tree, Random forest, KNN and SVM.

Heart disease prediction is considered a classification-based problem. So, the Random Forest classifier is the most accurate algorithm which gives accuracy of 97% and it is a suitable model for further processes. In addition, deployment of machine learning models using web applications is done with the help of flask framework, HTML, GitHub and Heroku server. Under deployment, the webpage will take input attributes from the user and give the output regarding the patient's condition with probability in percentage for occurrence of coronary heart disease in upcoming ten years.

Keywords— *Random Forest, Deployment, flask, Heroku, Web Application.*

I. INTRODUCTION

Many people are facing heart related problems due to Coronary heart disease. CHD happens in your coronary arteries that delivers a constant blood supply to your heart, these arteries are set on top of the heart. To your heart in order to work properly those muscles of the heart have to have constant fresh oxygenated blood going to it. In the arteries when it starts to develop the fatty plaques which cause blockages or restrict blood flow to the heart and these fatty plaques are caused by a condition called atherosclerosis and this occurs in the artery wall. Heart disease has become a serious issue which is major cause of death in humans [1]

Here comes machine learning in action. We can predict the CHD very early with the help of history and current condition of a person and we can save that person from a heart attack or stroke. In machine learning it learns from the natural phenomenon or events as in this project we are using the biological parameters such as cholesterol, blood pressure, gender, age, etc. for testing the data and on the basis of these, comparison and prediction is done in the terms of accuracy of algorithms such as in this project we are using Random Forest for prediction [2]. Testing and prediction of heart diseases is important in the restorative field and machine learning systems are used for processing the gathered

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

information from healthcare industries and gives novel and better understanding towards heart disease [3]. Many people are addicted with chronic habitual behaviors, like consumption of Cigars, Alcohol. Those peoples are suffering from chronic diseases like heart diseases, cancer, liver problems, kidney failures etc. To cure such persons with chronic disease is a hard task for doctors. For such hard challenges, professionals are providing hand to hand support to predict such disease early and cure as well as recover the patients from the chronic disease [5], likewise this paper is proposing the same task but in a different style we are proposing a web application for the early prediction of heart disease which would be very easy to handle, this will help many people to save their money and time from costly heart tests and as according to the statistical data from WHO, one-third population worldwide died from heart disease[6] so this project will save many lives.

Heart disease risk factor include:

High Cholesterol
High blood pressure
Diabetics
Smoking
Consuming too much alcohol
Being overweight

Symptoms of Heart attack:

Shortness of breath
Pain and discomfort in chest
Pain may spread to left or right arm or to neck, jaw, back or stomach

ITERATURE REVIEW

	Using Machine Learning Algorithms [2]			done using data balancing technique. 2. KNN is most efficient algorithm with accuracy of 87%.
2	Improving the Accuracy in Prediction of Heart Disease Using Machine Learning Algorithms [3]	2020	B. Keerthi Samhitha, Sarika Priya. M. R, Sanjana.C, Suja Cherukullapurath Mana and Jithina Jose	1. Cleveland heart dataset is collected from UCI for preparing and testing purposes. 2. KNN is most efficient algorithm with accuracy of 88.7%.
3	Prediction of Heart Disease Using Machine Learning [4]	2018	Aditi Gavhane, Isha Pandya, Gouthami Kokkula, Prof. Kailas Devadkar	1. Cleveland heart dataset is collected from UCI library. 2. Algorithm used for heart disease prediction is multi-layer perceptron (MLP)
4	Early Detection of Heart Syndrome Using Machine Learning Technique [5]	2019	Noor Basha, Gopal Krishna C., Ashok Kumar P. S. and Venkatesh P.	1. Dataset is collected from Kaggle web to analyse and implement the data on different algorithm to check the accuracy score. 2. Out of all machine learning algorithm KNN gives more

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

				accuracy score of 85%
5	Prediction of Heart Disease Using Machine Learning Algorithms [6]	2019	Mr. Santhana Krishana.j, Dr. Geetha.S	1. Dataset used here for predicting heart disease is taken from UCI Machine learning repository. 2. Algorithms used for prediction of heart disease is Naïve Bayes and Decision Tree out of these Decision Tree gives more accuracy of 91%.

TABLE I: LITERATURE REVIEW

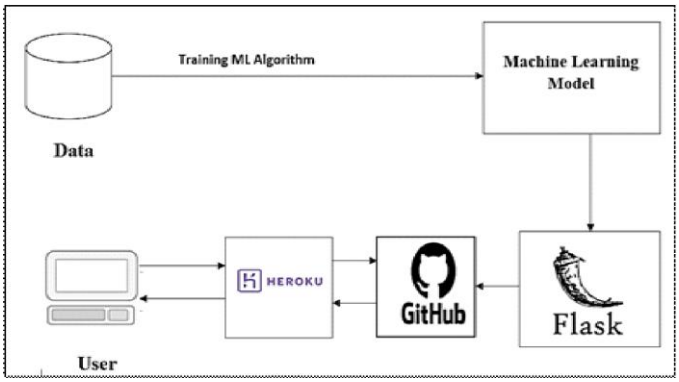


FIGURE 1: BLOCK DIAGRAM

II. ALGORITHM SELECTION

In this paper we are Working the various machine learning algorithms such as Decision Tree, Random Forest, K- Nearest Neighbor and Support Vector Machine. Out of these four algorithms Random Forest gives the highest accuracy of 97% hence Random Forest algorithm used for further process of deployment.

Algorithm	**Accuracy**
Decision Tree	91.31%
Random Forest	97.35%
K- Nearest Neighbor	85.26%
Support Vector Machine	66.01%

TABLE III: ACCURACY SCORE OF MACHINE LEARNING ALGORITHMS.

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

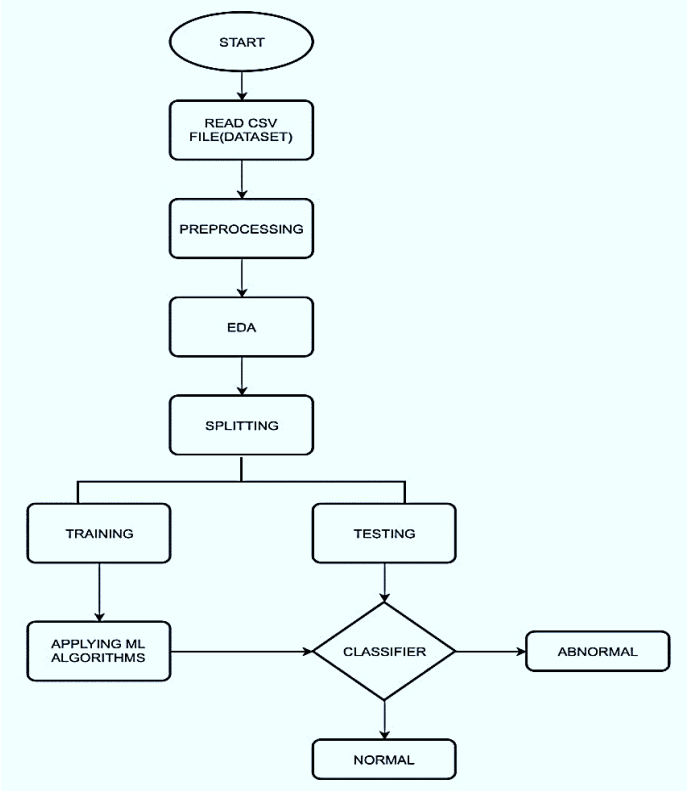


FIGURE 2: FLOWCHART OF MACHINE LEARNING MODEL

III. METHODOLOGY

A. Collection of Dataset

- Dataset is collected from Kaggle.[7]
- Framingham Heart Study dataset which contains about 4240 records and 15 attributes.
- Import necessary libraries which are required for the prediction like Pandas, NumPy, Sklearn, Pickle etc.

Attributes:

Demographic:

- gender: male(1) or female(0)
- Age: Age of the patient
- Education: no further information provided

Behavioral:

- Current Smoker: whether or not the patient is a current smoker
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day.

Information on medical history:

- BP Meds: whether or not the patient was on blood pressure medication
- Prevalent Stroke: whether or not the patient had previously had a stroke
- Prevalent Hyp: whether or not the patient was hypertensive
- Diabetes: whether or not the patient had diabetes

Information on current medical condition:

- Tot Chol: total cholesterol level
- Sys BP: systolic blood pressure
- Dia BP: diastolic blood pressure
- BMI: Body Mass Index
- Heart Rate: heart rate
- Glucose: glucose level

Target variable to predict:

- 10 year risk of heart disease (CHD) - (binary: "1", means "Yes", "0" means "No")

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

FIGURE 3I: ATTRIBUTES[7]

B. Data Pre-processing

- Dataset may include missing values or irrelevant data hence data pre-processing is important and it also effect on accuracy score.
- Firstly, check if there is any null/ missing value present in dataset.
- If data contain missing values, then to remove these missing values There are several methods for removing null/missing value in dataset some of them are:
  i. Replace null/missing values by the mean value.
  ii. Replace null/missing values by the median value.
  iii. Drop the records that contains null/missing value.
- Also check for outliers If any attributes contain outliers, then remove outliers using IOR(Inter Quartile Range) method.

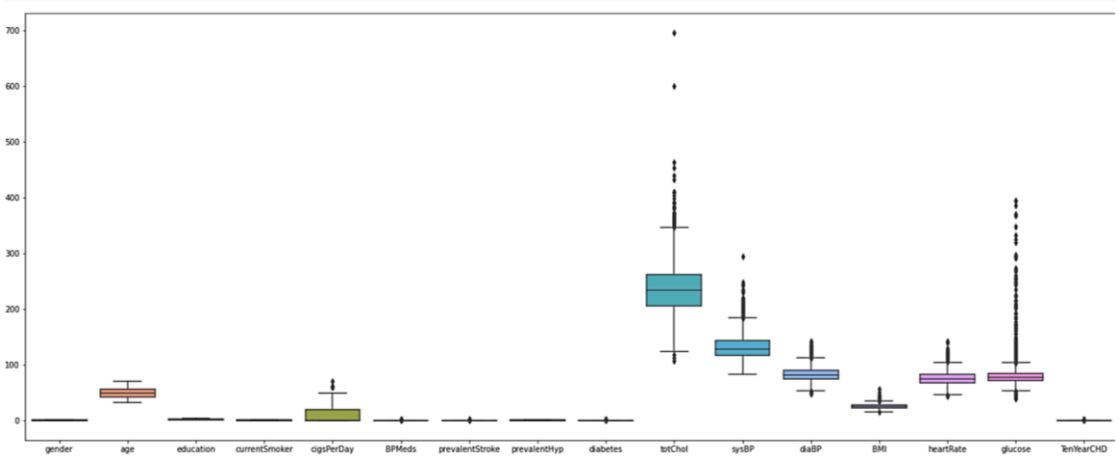


FIGURE 4: BOXPLOT

C. Resampling

- Resampling technique used to deal with highly imbalanced dataset.

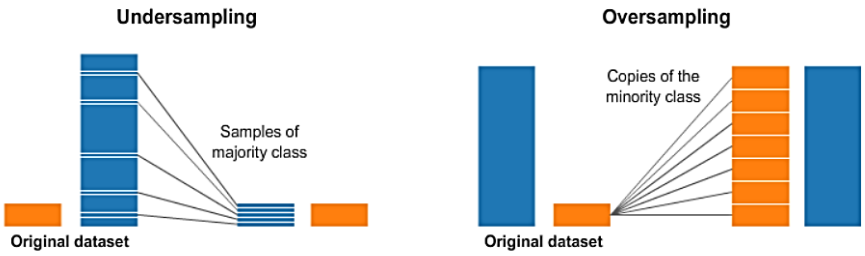
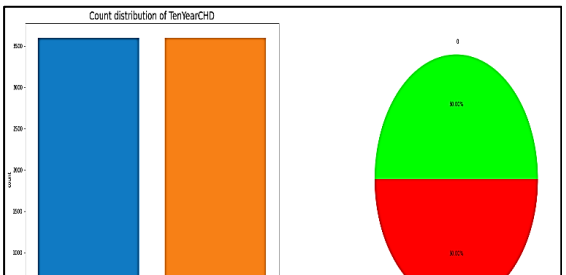


FIGURE 5: RESAMPLING TECHNIQUES[8]

- As shown in above figure v resampling consists of removing samples from majority class which called as undersampling and adding more samples to minority class called as oversampling.[8]
- In this paper we are using resample method to balance our data.



ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

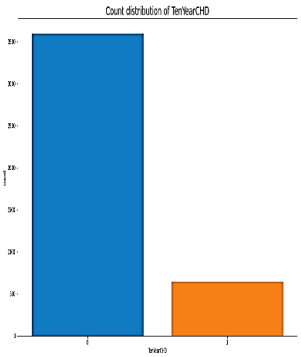


FIGURE 6: BEFORE RESAMPLING

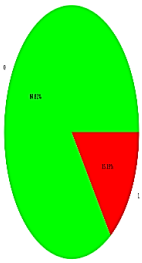


FIGURE VII: AFTER RESAMPLING

D. Feature Selection

- As our dataset contains 15 attributes not all attributes are that much important for prediction.
- Feature selection is a technique to choose the attributes which contribute most to the target variable.
- Feature importance gives the score for each attribute in the dataset, the higher score indicates the more important attributes towards the target attribute.
- In this paper we are using Extra Tree Classifier for selecting top 9 features and it is inbuilt class in tree-based classification algorithm.

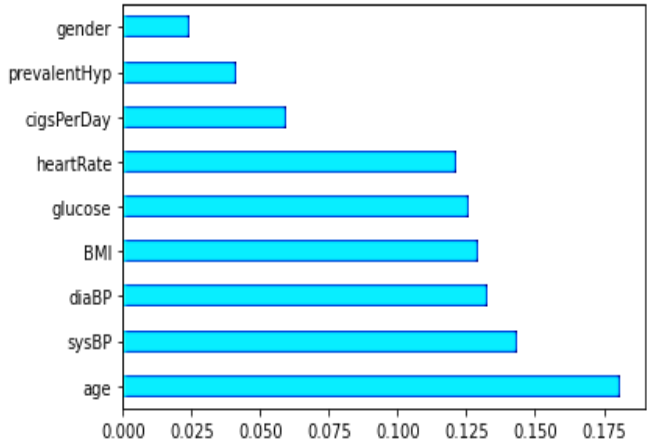


FIGURE 7: FEATURE IMPORTANCE

E. Model Selection

- In this we are using the Random Forest algorithm as we discussed earlier in paper it gives a maximum accuracy of 97%.
- Random Forest classifier is basically a bagging technique.
- In bagging, there are many base learner models. In the random forest this model is basically called as decision tree.
- From dataset we are picking some sample rows and columns give it to the 1st decision tree and it will get train on this particular data.
- Similarly remaining all decision tree will get train on various samples of data and using majority voting random forest will gives the output.

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

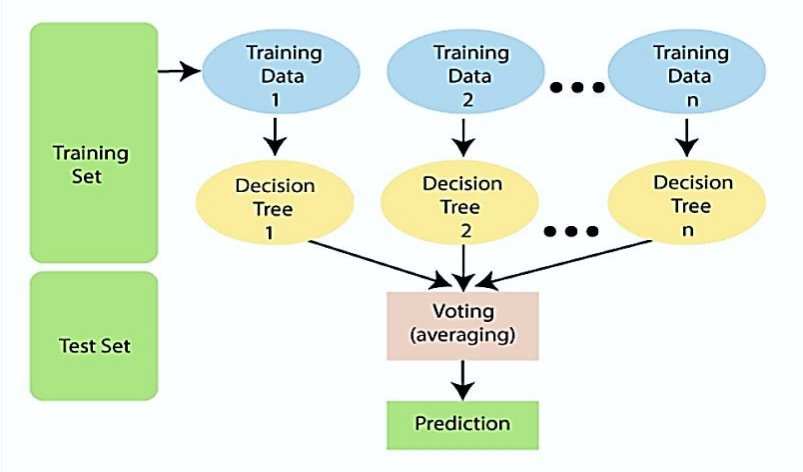


FIGURE 8: RANDOM FOREST

F. Train & Test model

- The goal of training is to make a final prediction correctly as often as possible.
- Split data into the training and testing part. Generally, 80% of data is took for train the model and 20% data are for the testing part.
- Our machine learning model gives accuracy of 97% and below figure shows the sample outputs.

EXAMPLE 1

```
[ ]  pred_new=my_model.predict([[55,1,20.0,1,75.0,32.2,180.0,90.0,85.0]])
     pred_new

     array([1])
```

"1", means Risk of Heart Disease

EXAMPLE 2

```
[ ]  pred_new=my_model.predict([[35,1,0,1,75.0,32.2,120.0,90.0,85.0]])
     pred_new

     array([0])
```

"0", means No Risk of Heart Disease

FIGURE 9: EXAMPLES

- Confusion Matrix is a matrix used to summarize the performance of a machine learning algorithm, it is showing how many predictions are right and how many are wrong.
- There are mainly four Terms associated with confusion matrices.

i. True Positive (TP): If the input is positive and the prediction made by the system is true.

ii. True Negative (TN): If the input is negative and the prediction made by the system is true.

iii. False Positive (FP): If the input is positive and the prediction made by the system is false.

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

iv. False Negative (FN): If the input is negative and the prediction made by the system is false.

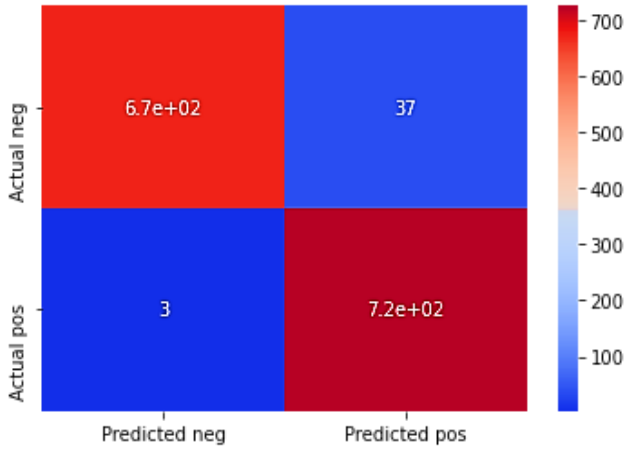


FIGURE 10: CONFUSION MATRIX

- For the good predicted model TP rate and TN rate is always greater that FP and FN rate

True positive (TP) = 725
True negative (TN) = 673
False Positive (FP) = 3
False negative (FN) = 37

G. *Deployment*

- Deployment of machine learning model is done with the help of flask, Heroku and GitHub.
- Flask is lightweight micro web application framework [9]
- GitHub is used as version controller. The version control is useful for if we have made any changes in our code so this system keep track of it.

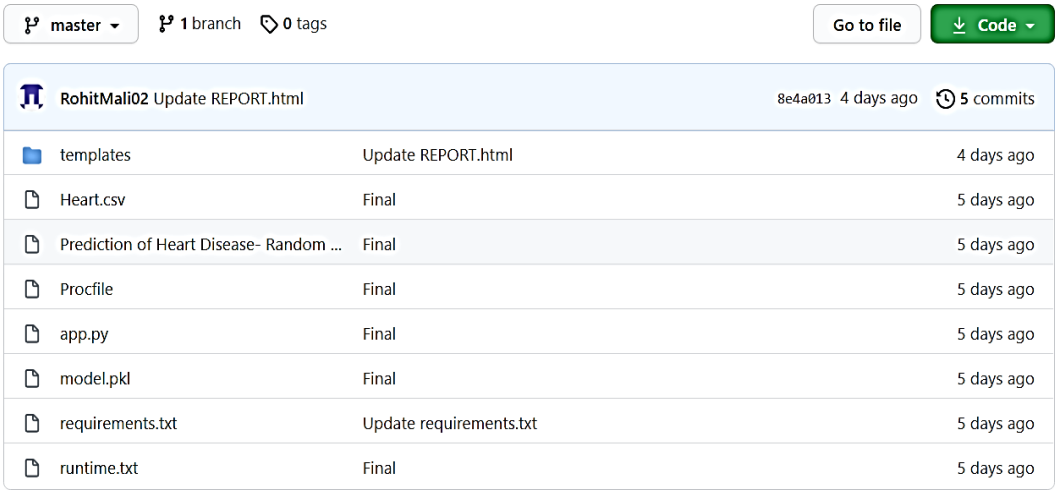


FIGURE 11: GITHUB REPOSITORY

- Heroku is a platform as service (PaaS) which is useful for operating and running your model

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

in the cloud.

FIGURE 12: INPUT ATTRIBUTES

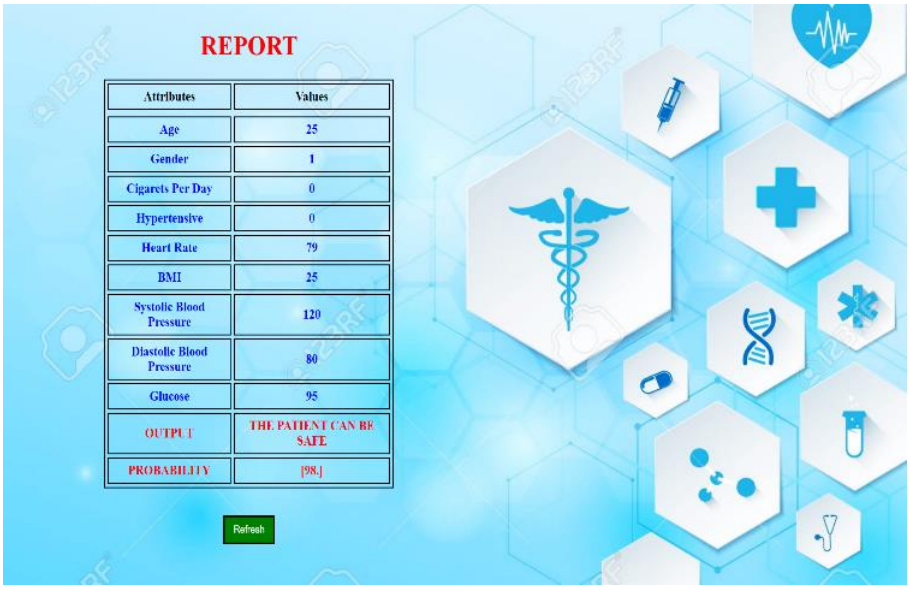


FIGURE 13: REPORT

IV. CONCLUSIONS

In this work, we have implemented a real-time web application for heart disease prediction. This application allows users to input different attributes like Gender, Age, Cigarettes Per Day, Hypertensive, BMI, Heart Rate, Systolic Blood Pressure, Diastolic Blood Pressure, Glucose Level as input parameters so that the user can get quick predictions depending upon inputs and previous data.

Hence, we can predict the Coronary Heart Disease at an early stage with the help of medical history and current condition of a person and we can save that person from a heart attack or stroke.

V. FUTURE WORK

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

For the future consideration, different types of disease can be predicted within the website only by providing input attributes for particular disease.

REFERENCES

[1] Lokesh M. Giripunje, Tejas P. Sonar, Rohit S. Mali, Jayant C. Modhave, Mahesh B. Gaikwad "Review on Machine Learning in Healthcare Industry -Heart Disease Prediction". LINO Journal ISSN – 0211-2574 April 2021 Volume: 11, Issue: 3.

[2] Archana Singh, Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms". 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020)

[3] B. Keerthi Samhitha, Sarika Priya M. R, Sanjana.C, Suja Cherukullapurath Mana and Jithina Jose. "Improving the Accuracy in Prediction of Heart Disease using Machine Learning Algorithms". International Conference on Communication and Signal Processing, July 28 - 30, 2020, India

[4] Aditi Gavhane, Isha Pandya, Prof. Kailas Devadkar. "Prediction of Heart Disease Using Machine Learning". Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

[5] Noor Basha, Gopal Krishna, Ashok Kumar, Venkatesh P. "Early Detection of Heart Syndrome Using Machine Learning Technique". 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).

[6] Mr. Santhana Krishana.j, Dr. Geetha.S. "Prediction of Heart Disease Using Machine Learning Algorithms". IEEE 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT).

[7] https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset?select=framingham.csv

[8] https://medium.com/analytics-vidhya/undersampling-and-oversampling-an-old-and-a-new-approach-4f984a0e8392

[9] https://palletsprojects.com/p/flask/

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC

ISSN: 2233-7853 IJFGCN
Copyright ⓒ 2020 SERSC