HEART DISEASE PREDICTION USING MACHINE LEARNING

1st GOKKUL E
2019503517
BATCH 1

2nd RAHUL B
2019503545
BATCH 1

3rd BHARATH M
2019503509
BATCH 1

***Abstract*—Heart disease is a broad term used for diseases and conditions affecting the heart and circulatory system. They are also referred to as cardiovascular diseases. It is a major cause of disability all around the world. Since the heart is amongst the most vital organs of the body, its diseases affect other organs and part of the body as well. There are several different types and forms of heart diseases. The most common ones cause narrowing or blockage of the coronary arteries, malfunctioning in the valves of the heart, enlargement in the size of heart and several others leading to heart failure and heart attack.Key facts according to WHO (World Health Organizations) :**

- **Cardiovascular diseases (CVDs) are the leading cause of death globally.**
- **An estimated 17.9 million people died from CVDs in 2019, representing 32% of all global deaths. Of these deaths, 85% were due to heart attack and stroke.**
- **Over three quarters of CVD deaths take place in low- and middle-income countries.**
- **Out of the 17 million premature deaths (under the age of 70) due to noncommunicable diseases in 2019, 38% were caused by CVDs.**
- **Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol.**
- **It is important to detect cardiovascular disease as early as possible so that management with counseling and medicines can begin.**

Cardiovascular disease or heart disease describes a range of conditions that affect your heart. Diseases under the heart disease umbrella include blood vessel diseases, such as coronary artery disease. From WHO statistics every year 17.9 million die from heart disease. The medical study says that human lifestyle is the main reason behind this heart problem. Apart from this there are many key factors which warns that the person may/may not be getting chance of heart disease.Using a dataset containing information on patients who had health checkups, if we create suitable machine learning techniques which predicts the heart disease more accurately, it is very helpful to the health organization as well as patients.

I. INTRODUCTION

In this project, we will be building a heart disease prediction website which gives result for the data entered.We use an online dataset which contains some medical information of patients which tells whether that person getting a heart attack chance is less or more. Using the information explore the dataset and classify the target variable using different Machine Learning models and find out which algorithm suitable for this dataset and find the most accurate model.We use that model in our webpage to predict whether the given new data has a heart disease or not.

II. LITERATURE SURVEY.

A. Machine Learning Techniques For Heart Disease Prediction by A.Lakshmanarao,Y.Swathi,P.Sri Sai Sundareswar,2019.

"Lakshmana Rao et al,Machine Learning Techniques for Heart Disease Prediction in which the contributing elements for heart disease are more (circulatory strain, diabetes, current smoker, high cholesterol, etc..). So, it is difficult to distinguish heart disease. Different systems in data mining and neural systems have been utilized to discover the seriousness of heart disease among people. The idea of CHD ailment is bewildering, in addition, in this manner, the disease must be dealt with warily. Not doing early identification, may impact the heart or cause sudden passing. The perspective of therapeutic science furthermore, data burrowing is used for finding various sorts of metabolic machine learning a procedure that causes the framework to gain from past information tests, models without being expressly customized Machine learning makes rationale dependent on chronicled information."

B. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques by Senthilkumar Mohan , Chandrasegar Thirumalai,Gautam Srivastava , 2019

"The authors have proposed Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques in which strategy that objective is to finding critical includes by applying Machine Learning bringing about improving the exactness in the expectation of cardiovascular malady. The expectation model is created with various blends of highlights and a few known arrangement strategies. We produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with a linear model (HRFLM) they likewise educated about Diverse data mining approaches and expectation techniques, Such as, KNN, LR, SVM, NN, and Vote have been fairly famous of late to distinguish and predict heart disease."

C. Prediction of Heart Disease using Multiple Linear Regression Model by K. Polaraju, D. Durga Prasad, 2017

"K. Polaraju et al, proposed Prediction of Heart Disease using Multiple Regression Model and it proves that Multiple Linear Regression is appropriate for predicting heart disease

chance. The work is performed using training data set consists of 3000 instances with 13 different attributes which has mentioned earlier. The data set is divided into two parts that is 70% of the data are used for training and 30% used for testing. Based on the results, it is clear that the classification accuracy of Regression algorithm is better compared to other algorithms."

D. Prediction of Heart Disease Using Machine Learning Algorithms by Mr.Santhana Krishnan.J, Dr.Geetha.S, 2019

"Mr. Santhana Krishnan.J and Dr. Geetha.S, Prediction of heart disease using machine learning algorithm This Paper predicts heart disease for Male Patient using Classification Techniques. The detailed information about Coronary Heart diseases such as its Facts, Common Types, and Risk Factors has been explained in this paper. The Data Mining tool used is WEKA (Waikato Environment for Knowledge Analysis), a good Data Mining Tool for Bioinformatics Fields. The all three available Interface in WEKA is used here; Naive Bayes, Artificial Neural Networks and Decision Tree are Main Data Mining Techniques and through this techniques heart disease is predicted in this System. The main Methodology used for prediction is Decision Trees like CART, C4.5, CHAID, J48, ID3 Algorithms, and Naive Bayes Techniques."

E. Heart Disease Prediction Using Effective Machine Learning Techniques by Avinash Golande, Pavan Kumar T, 2004

"Avinash Golande et al,proposed Heart Disease Prediction Using Effective Machine Learning Techniques in which Specialists utilize a few data mining strategies that are available to support the authorities or doctors distinguish the heart disease. Usually utilized methodology utilized are decision tree, k- closest and Naïve Bayes. Other unique characterization-based strategies utilized are packing calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel self arranging guide and SVM (Bolster Vector Machine). The following area obviously gives subtleties of systems that were utilized in the examination. ".

F. Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques by Mr. Chala Beyene, Prof. Pooja Kamat, 2018

"Chala Beyene et al, recommended Prediction and Analysis the occurrence of Heart Disease Using Data Mining Techniques. The main objective is to predict the occurrence of heart disease for early automatic diagnosis of the disease within result in short time. The proposed methodology is also critical in healthcare organisation with experts that have no more knowledge and skill. It uses different medical attributes such as blood sugar and heart rate, age, sex are some of the attributes are included to identify if the person has heart disease or not. Analyses of dataset are computed using WEKA software.".

G. Machine Learning Technology-Based Heart Disease Detection Models by Umarani Nagavelli, Debabrata Samanta, Partha Chakraborty, 2020

"The authors propose a hybrid approach for diagnosing cardiac disease.Using the WEKA and KEEL tools, performance of these algorithms is evaluated.PCA is the widely used dimensionality reduction technique for Feature extraction under WEKA tool.Wrapper method is used for feature selection under KEEL tool.Selected model is tested only when the model is completely trained.".

H. Heart Disease Prediction Based on the Embedded Feature Selection Method and Deep Neural Network by Dengqing Zhang, Yunyi Chen, Yuxuan Chen, Shengyi Ye, Wenyu Cai, Junxue Jiang, Yechuan Xu, Gongfeng Zheng, Ming Chen, 2019

"The goal of this study was to examine machine learning algorithms using several performance criteria in order to enhance accuracy. It requires using the mean value to replace missing data in the pre-processing step.A score accuracy of 86.8% was achieved using SVM with a linear kernel.".

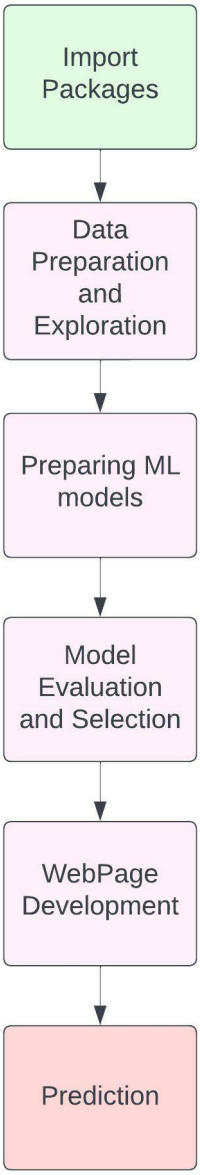
I. HDPF: Heart Disease Prediction Framework Based on Hybrid Classifiers and Genetic Algorithm by S. E. A. Ashri, M. M. El-Gayar and E. M. El-Daydamony, 2021

"Inorder to improve the prediction accuracy, here the authors have used hybrid approach using ensemble model.To enhance the prediction performance and time consumption they have proposed Pre-processing and Feature Selection based on a genetic algorithm.To overcome the Overfitting problem 10-folds cross-validation technique used.The proposed ensemble classifier model achieved a classification accuracy of 98.18%.".

J. Heart disease identification from patients' social posts, machine learning solution on Spark by Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, 2020

"The authors have proposed the best machine learning algorithm for predicting heart disease with high accuracy.Two types of feature selection techniques are used: univariate feature selection and Relief . Hyperparameter optimization and cross-validation were used to improve the accuracy of four types of machine learning algorithms. Twitter data streams containing patient data can be efficiently handled by merging Apache Kafka with Apache Spark as the system's underlying architecture.Random forest classifier outperforms the other models by achieving the highest accuracy at 94.9%.".

III. ARCHITECTURE DIAGRAM



IV. PROPOSED WORK.

A. Import Packages

- Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.
- NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.
- Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support-vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.
- The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.
- Python pickle module is used for serializing and de-serializing python object structures. The process to converts any kind of python objects (list, dict, etc.) into byte streams (0s and 1s) is called pickling or serialization or flattening or marshalling.
- These are the packages imported to use in our modeling

B. Data Preparation and Exploration

The dataset has 14 attributes:

- age: age in years.
- sex: sex (1 = male; 0 = female).
- cp: chest pain type (Value 0: typical angina; Value 1: atypical angina; Value 2: non-anginal pain; Value 3: asymptomatic).
- trestbps: resting blood pressure in mm Hg on admission to the hospital.
- chol: serum cholestoral in mg/dl.
- fbs: fasting blood sugar ¿ 120 mg/dl (1 = true; 0 = false).
- restecg: resting electrocardiographic results (Value 0: normal; Value 1: having ST-T wave abnormality; Value 2: probable or definite left ventricular hypertrophy).
- thalach: maximum heart rate achieved.
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak: ST depression induced by exercise relative to rest.
- slope: the slope of the peak exercise ST segment (Value 0: upsloping; Value 1: flat; Value 2: downsloping).
- ca: number of major vessels (0-3) colored by flourosopy.
- thal: thalassemia (3 = normal; 6 = fixed defect; 7 = reversable defect).
- target: heart disease (1 = no, 2 = yes)

C. Preparing ML models

1) K-Nearest Neighbor:

- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.
- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

2) Random Forest Classifier:

- 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.
- Random Forest can be used for both Classification and Regression problems in ML.
- Random Forest 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.

3) Support Vector Machine:

- 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.
- Support Vector machine is used for Classification as well as Regression problems.
- SVM algorithm can be used for Face detection, image classification, text categorization.

4) Decision Tree:

- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
- For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

5) Logistic Regression:

- 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.
- Logistic regression is used for solving the classification problems.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.

6) Naive Bayes Classifier:

- Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- 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.

7) Extreme Gradient Boost:

- Gradient boosting refers to a class of ensemble machine learning algorithms that can be used for classification or regression predictive modeling problems.
- Extreme Gradient Boosting, or XGBoost for short is an efficient open-source implementation of the gradient boosting algorithm. As such, XGBoost is an algorithm, an open-source project, and a Python library

D. Models evaluation and selection

- Evaluating accuracy of each model and then the best model is found.
- Save the model as serialized object pickle.

E. WebPage Development

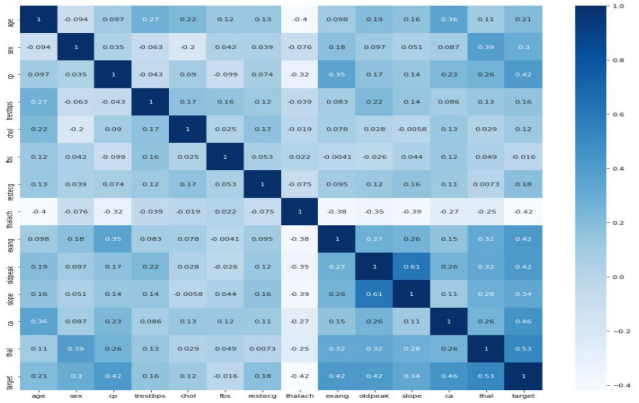
- A Webpage containing a form is created, the form contains details to be entered by the user.
- The webpage uses the best model which we obtained after comparing all the models for evaluating the user entered data.

F. Prediction

- The user data is used to predict whether there is a heart disease for the given user detail.
- It uses the best model which gives the most accuracy to evaluate for the given data.
- It then predicts the result.

V. IMPLEMENTATION

A. Attributes Correlation



From the above correlation plot, the chest pain type (cp), exercise induced angina (exang), ST depression induced by exercise relative to rest (oldpeak), the slope of the peak

exercise ST segment (slope), number of major vessels (0-3) colored by flourosopy (ca) and thalassemia (thal) are correlated with the heart disease (target) directly. We see also that there is an inverse proportion between the heart disease and maximum heart rate (thalch). We can see also, there are a relation between the following attributes:

- The number of major vessels (0-3) colored by flourosopy (ca) and the age.
- ST depression induced by exercise relative to rest (old-peak) and the slope of the peak exercise ST segment (slope).
- The chest pain type (cp), exercise induced angina (exang).
- maximum heart rate (thalch) and the age.

B. Logistics Regression

```
confussion matrix
[[26  4]
 [ 5 19]]


Accuracy of Logistic Regression: 83.33333333333334
```

	precision	recall	f1-score	support
1	0.84	0.87	0.85	30
2	0.83	0.79	0.81	24
accuracy			0.83	54
macro avg	0.83	0.83	0.83	54
weighted avg	0.83	0.83	0.83	54

For the given dataset, Logistic Regression shows an accuracy of 83

C. Naive Bayes

```
confussion matrix
[[23  7]
 [ 7 17]]


Accuracy of Naive Bayes model: 74.07407407407408
```

	precision	recall	f1-score	support
1	0.77	0.77	0.77	30
2	0.71	0.71	0.71	24
accuracy			0.74	54
macro avg	0.74	0.74	0.74	54
weighted avg	0.74	0.74	0.74	54

For the given dataset, Naive Bayes shows an accuracy of 74

D. Random Forest

```
confussion matrix
[[28  2]
 [ 6 18]]


Accuracy of Random Forest: 85.18518518518519
```

	precision	recall	f1-score	support
1	0.82	0.93	0.87	30
2	0.90	0.75	0.82	24
accuracy			0.85	54
macro avg	0.86	0.84	0.85	54
weighted avg	0.86	0.85	0.85	54

For the given dataset, Random Forest shows an accuracy of 85

E. Extreme Gradient Boost

```
confussion matrix
[[25  5]
 [ 7 17]]


Accuracy of Extreme Gradient Boost: 77.77777777777779
```

	precision	recall	f1-score	support
1	0.78	0.83	0.81	30
2	0.77	0.71	0.74	24
accuracy			0.78	54
macro avg	0.78	0.77	0.77	54
weighted avg	0.78	0.78	0.78	54

For the given dataset, Extreme Gradient Boost shows an accuracy of 77

F. K-Neighbors Classifier

```
confussion matrix
[[23  7]
 [ 9 15]]


Accuracy of K-NeighborsClassifier: 70.37037037037037
```

	precision	recall	f1-score	support
1	0.72	0.77	0.74	30
2	0.68	0.62	0.65	24
accuracy			0.70	54
macro avg	0.70	0.70	0.70	54
weighted avg	0.70	0.70	0.70	54

For the given dataset, K-Neighbors Classifier shows an accuracy of 70

G. Decision Tree Classifier

```
confussion matrix
[[23  7]
 [ 7 17]]


Accuracy of DecisionTreeClassifier: 74.07407407407408

              precision    recall  f1-score   support

           1       0.77      0.77      0.77        30
           2       0.71      0.71      0.71        24

    accuracy                           0.74        54
   macro avg       0.74      0.74      0.74        54
weighted avg       0.74      0.74      0.74        54
```

For the given dataset, Decision Tree Classifier shows an accuracy of 74

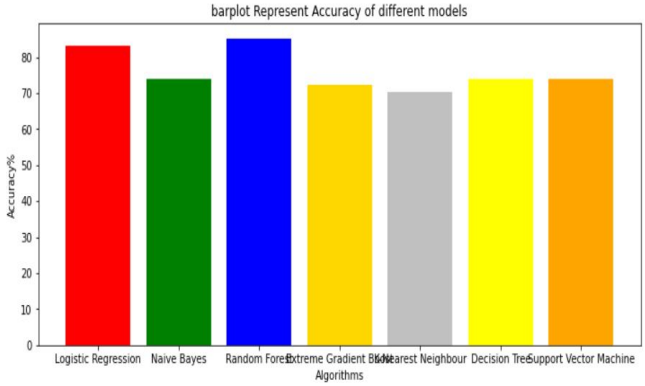
H. Support Vector Classifier

```
confussion matrix
[[27  3]
 [11 13]]


Accuracy of Support Vector Classifier: 74.07407407407408

              precision    recall  f1-score   support

           1       0.71      0.90      0.79        30
           2       0.81      0.54      0.65        24

    accuracy                           0.74        54
   macro avg       0.76      0.72      0.72        54
weighted avg       0.76      0.74      0.73        54
```

For the given dataset, Support Vector Classifier shows an accuracy of 74

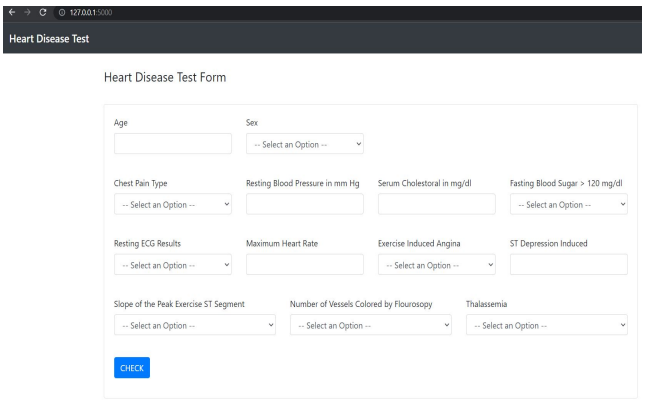
I. Accuracy Comparison of different models

	Model	Accuracy
0	Logistic Regression	83.333333
1	Naive Bayes	74.074074
2	Random Forest	85.185185
3	Extreme Gradient Boost	72.222222
4	K-Nearest Neighbour	70.370370
5	Decision Tree	74.074074
6	Support Vector Machine	74.074074



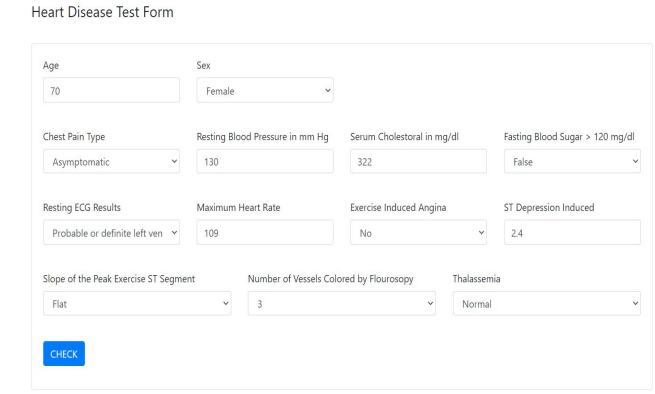
Comparing all models and the accuracy of Random Forest Classifier seems to be the highest so we will select Random Forest as the primary classifier.

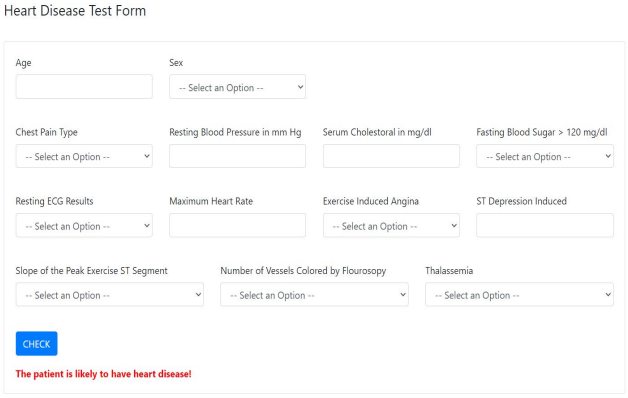
J. Webpage Outlook



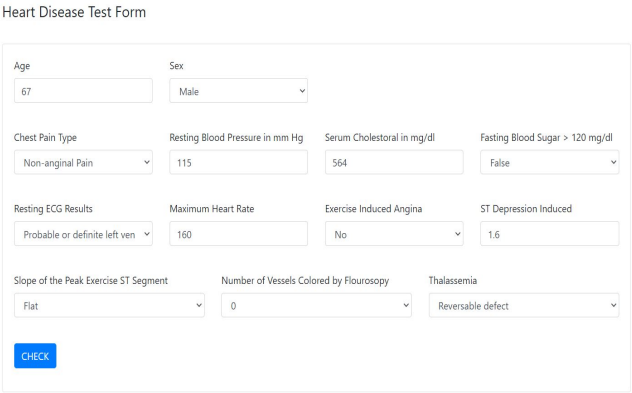
Using HTML, we have created several forms for getting inputs from users and then compared with our model and the result is displayed.

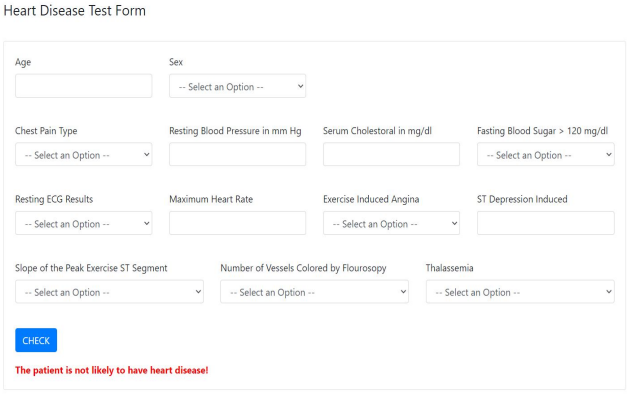
K. Prediction





From the above inputs , we have predicted that the person is having an heart disease.





From the above inputs , we have predicted that the person is not having an heart disease.

VI. CONCLUSION

Heart is one of the essential and vital organ of human body and prediction about heart diseases is also important concern for the human beings so that the accuracy for algorithm is one of parameter for analysis of performance of algorithms. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose. When we perform the analysis of algorithms on the basis of dataset and on the basis of confusion matrix, we find Random Forest as the best one. For the Future Scope more machine learning approach will be used for best analysis of the heart diseases and for earlier prediction of diseases so that the rate of the death cases can be minimized by the awareness about the diseases.

REFERENCES

[1] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, "Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning", Computational Intelligence and Neuroscience, vol. 2021, Article ID 8387680, 11 pages, 2021.

[2] N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020.

[3] Abdeldjouad, F.Z., Brahami, M., Matta, N. (2020). A Hybrid Approach for Heart Disease Diagnosis and Prediction Using Machine Learning Techniques. In: Jmaiel, M., Mokhtari, M., Abdulrazak, B., Aloulou, H., Kallel, S. (eds) The Impact of Digital Technologies on Public Health in Developed and Developing Countries. ICOST 2020. Lecture Notes in Computer Science(), vol 12157.

[4] Swain, D., Ballal, P., Dolase, V., Dash, B., Santhappan, J. (2020). An Efficient Heart Disease Prediction System Using Machine Learning. In: Swain, D., Pattnaik, P., Gupta, P. (eds) Machine Learning and Information Processing. Advances in Intelligent Systems and Computing, vol 1101.

[5] N. Louridi, M. Amar and B. E. Ouahidi, "Identification of Cardiovascular Diseases Using Machine Learning," 2019 7th Mediterranean Congress of Telecommunications (CMT), 2019, pp. 1-6.

[6] Anna Karen Gárate-Escamila, Amir Hajjam El Hassani, Emmanuel Andrès, Classification models for heart disease prediction using feature selection and PCA, Informatics in Medicine Unlocked, Volume 19, 2020,100330, ISSN 2352-9148.

[7] G. S. Sajja, M. Mustafa, K. Phasinam, K. Kaliyaperumal, R. J. M. Ventayen and T. Kassanuk, "Towards Application of Machine Learning in Classification and Prediction of Heart Disease," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1664-1669.

[8] Amen, Khalid and Zohdy, Mohamed and Mahmoud, Mohammed. (2020). Machine Learning for Multiple Stage Heart Disease Prediction. 10.5121/csit.2020.101118.

[9] P. Ramprakash, R. Sarumathi, R. Mowriya and S. Nithyavishnupriya, "Heart Disease Prediction Using Deep Neural Network," 2020 International Conference on Inventive Computation Technologies (ICICT), 2020, pp. 666-670.

[10] Hager Ahmed, Eman M.G. Younis, Abdeltawab Hendawi, Abdelmgeid A. Ali, Heart disease identification from patients' social posts, machine learning solution on Spark, Future Generation Computer Systems, Volume 111, 2020, Pages 714-722, ISSN 0167-739X