JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

# HEART DISEASE PREDICTION

**I.Amrutha[1], A.V.S.S. Deekshitha[1], G. Nagendra[1], T. Kasturi[1], Dr. E. Laxmi Lydia[2]**

[1] Department of Computer Science and Engineering, Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh 530049, India, email: amruthaittamsetty@gmail.com

[1] Department of Computer Science and Engineering, Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh 530049, India, email:deekshitha.agraharapu@gmail.com

[1] Department of Computer Science and Engineering, Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh 530049, India, email: gonuguntan@gmail.com

[2] Department of Computer Science and Engineering. Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh 530049, India, email: thotakasturi111@gmail.com

[2]Profressor, Department of Computer Science and Engineering, Vignan's Institute of Information Technology (Autonomous), Visakhapatnam, Andhra Pradesh 530049, India, email:elaxmi2002@yahoo.com

**Abstract:**

Heart diseases are one the serious problem that is one of the causes of the increased mortality rate. To overcome this problem heart disease prediction is needed. Machine Learning, a subfield of Artificial Intelligence is used for any kind of prediction where the model is trained from natural events. In this paper, we will predict heart disease from machine learning algorithms by calculating the accuracy of each algorithm. The algorithms are Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree, and Random Forest. The Kaggle dataset is used for training and testing. The model is implemented using python programming in the Anaconda (Jupyter notebook) tool which consists of various libraries that help in predictions.

**Keywords:** Supervised Learning, Unsupervised Learning, Reinforcement Learning, Logistic Regression, SVM, KNN, Naïve Bayes, Decision Tree, Random Forest.

## I. Introduction :

The heart is one of the most essential and vital organ of the body. The heart is the engine of the body such that its functionality is essential for life. Nowadays most diseases are related to the heart so the prediction of heart diseases is mandatory. Hence, various machine algorithms are used to predict heart diseases. This early prediction of heart disease using efficient machine learning algorithms helps to decrease the mortality rate and turns people towards their health care.

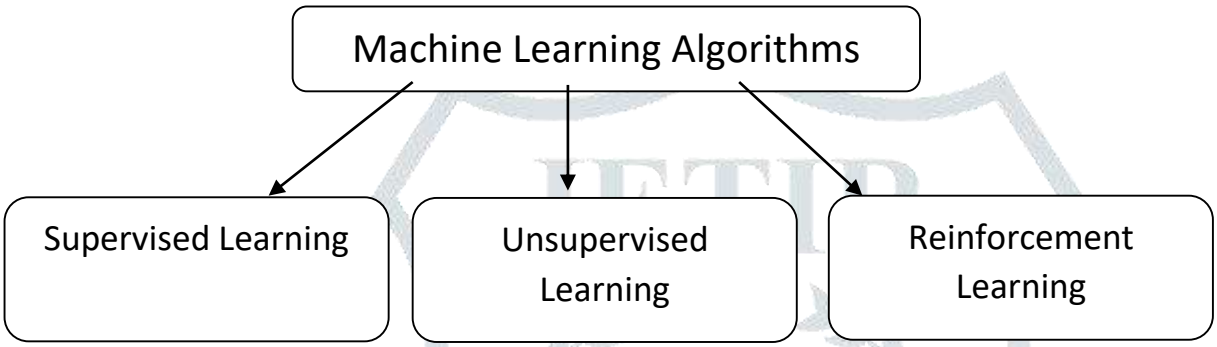
Machine Learning is the technology which is based on training and testing the model using a dataset and predicts new data. The model takes various parameters like age, gender, blood pressure, cholesterol, etc. as input and predicts the condition of the heart and the chance of heart disease.

This paper consists of the following sections: Section II is about the machine learning concept. Section III described the various researcher's studies. Section IV discussed the proposed system. Section V illustrates about methodology followed for predicting the system. Section VI tells about classification algorithms used in predicting heart diseases. Section VII represents the results of each algorithm. Section VIII gives the conclusion of this paper. And finally, Section IX is about Acknowledgement.

## II. MACHINE LEARNING

Machine learning is a technique that makes a machine that predicts things like humans without being programmed and also improves performance from past predictions or learning from the data.

Machine learning is purely based on training and testing datasets. The training dataset is used to prepare the model whereas the testing dataset is used to check the model output and whether the model built by the training dataset is giving the correct output or wrong output.




### A. Supervised Learning

Supervised learning is preparing a machine learning model based on a dataset (training dataset + testing dataset) under proper guidance where the model is built by training dataset consisting of input and output acts as guide and model is evaluated by testing dataset consists of input and output. Supervised learning is based on training. It has 2 types 1) Classification 2) Regression [1].

Below are some of the supervised learning algorithms:

- Linear Regression
- Logistic Regression
- Support Vector Machine (SVM)
- Decision tree

### B. Unsupervised Learning

Unsupervised learning is preparing a machine learning model without any proper guidance. If a dataset is given to unsupervised learning, it works on data present in the dataset and finds hidden patterns and relationships between the attributes present in the dataset then it builds the model. Unsupervised learning is based on the self-study concept. It has 2 types 1) Clustering 2) Association [1].

Following are some unsupervised learning algorithms

- K-means
- KNN (k-nearest neighbors)
- Neural network
- PCA (principal component analysis)

### C. Reinforcement Learning

Reinforcement learning is a hit and trial technique where the machine gets a reward or penalty points for each activity it performs. For every correct decision machine makes it acquires an award point or acquires penalty

point for every incorrect decision made by the machine model. It is an interaction between the learning agent and the environment where the learning agent depends on exploration and exploitation [1].

## III. LITERATURE SURVEY

Various researches have been done which focused on predicting heart diseases. Many researchers applied various data mining techniques for analysis and accomplished unique probabilities for various techniques

Chaitrali S.Dangare and Sulabha S. Apte stated that chances of suffering from heart diseases will also depend on family history, obesity, and smoking. In their research they used Cleveland and Statlog Heart Disease databases along with that they also included 2 more parameters like obesity and smoking as inputs. On these databases, they performed data mining techniques like Naïve Bayes, decision trees, and neural networks which resulted in an accuracy of 94.44%,96.66%, 99.25 with 13 attributes, and 90.74%,99.62%,100% with 15 attributes respectively [2].

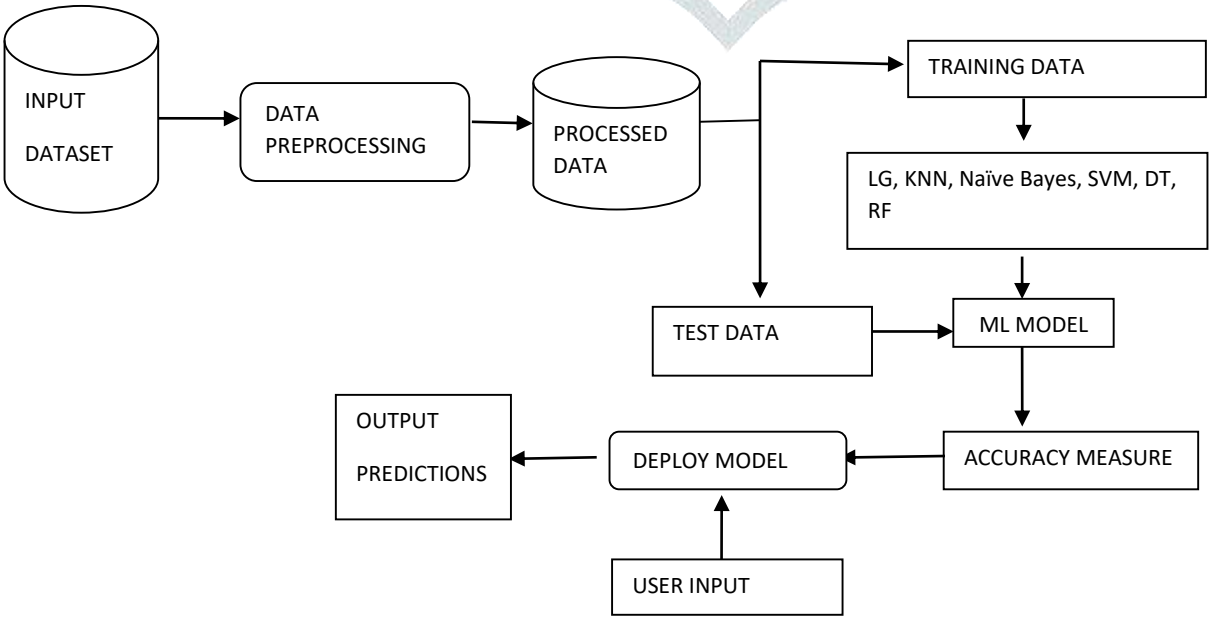
Avinash Golande and Pavan Kumar T investigated various classification algorithms that were used with dissimilar attributes and obtained different accuracies depending upon tools designed for execution, and examined accuracy can be improved by combining different data mining techniques and by parameter tuning [3].

The Intelligent Heart Disease Prediction System (IHDPS) was proposed by Sellappan Palaniappan and used data mining techniques like Decision trees, Naïve Bayes, and Neural networks. Each technique has its unique results. This model is built with hidden patterns and associated with their relationship. IHDPS is web-based, user-friendly, mountable, and expandable [4].

## IV. PROPOSED SYSTEM

In this present system, only registered users can use the heart disease prediction model in order to provide security. If the new user wants to use the prediction model, they need to register else the user can simply log in. Once the user logged in, the user is directed to the prediction model where input data is obtained from users. Using Machine learning classification algorithms like logistic regression, knn, naïve bayes, SVM, decision tree, and random forest user data are analysed. Now, results are obtained based on how we train the model.

## V. METHODOLOGY FOR PREDICTING SYSTEM



### A. Data Collection and Data Pre-processing

Dataset used in this paper is collected for Kaggle which is the data set collected from 1988 from four databases: Cleveland (303 observations), Hungary (294 observations), Switzerland (123 observations), Long Beach V (200 observations), and Stalog (123 observations) [5]. The heart disease dataset is a collection of a total of 918 observations that combined over 11 features that can be used to predict heart disease. Data pre-processing is a process of cleaning raw data into clean data that can be used for training models which helps in building machine learning models more accurately. We have used 80% as training data and 20% as testing data in this project.

**Table 1 Attributes along with description and values**

| S. No | Attributes | Description of attributes | Values |
|---|---|---|---|
| 1 | Age | Age of patient | Continuous |
| 2 | Sex | Gender of patient | 1- male<br>0- female |
| 3 | Cheat Pain Type | Discomfort is caused when the heart muscle doesn't get enough oxygen | 0- Typical angina<br>1- Atypical angina<br>2- Non-anginal pain<br>3- asymptomatic |
| 4 | Resting BP | High Blood pressure can damage arteries that circulate blood in heart | Continuous value in mm hg |
| 5 | Cholestrol | Bad cholesterol (blood fat related to diet) narrow arteries | Continuous value in mm/dl |
| 6 | FastingBS | When pancreas not producing enough hormones increase blood sugar | 0 < 120 mg/dl<br>1 >= 120 md/dl |
| 7 | RestingECG | Electrocardiography | 0 – Normal<br>1 – ST-T wave abnormality<br>2 – Showing left ventricular hypertrophy |
| 8 | MaxHR | Increase in heart rate increase the risk of cardiac death | A numeric value between 60 and 202 |
| 9 | ExerciseAngina | Pain or discomfort<br>Feels like a tight or gripping felt in the center of the chest | 1 – Yes<br>0 – No |
| 10 | Oldpeak | ST value measured in terms of depression | Numeric value |
| 11 | ST_Slope | After doing the exercise slope of the curve | 0- Upsloping<br>1- Flat<br>2- Downsloping |
| 12 | HeartDisease | The output of the prediction | 1- Heart disease<br>0- Normal person |

## VI. CLASSIFICATION ALGORITHMS

### A. LOGISTIC REGRESSION

This is a supervised machine learning technique that is used for classification purposes. It produces the results in categorical or discrete values. Instead of giving an exact 0 or 1, it gives probabilistic values that lie from 0 to 1.

## B. K-NEAREST NEIGHBORS (KNN)

K nearest neighbors algorithm is a simple unsupervised algorithm that can be used for solving classification or regression problems. Each data point is classified into a different category based on the distance between the points. Euclidean distance is mostly used to find out the distance between the data point and the standard point. This algorithm follows the principle of Birds of a feather flocking together. The steps followed by the knn algorithm are as follows

- Load the data from the dataset.
- Initialize the value for K.
- Find out the distance of K points.
- Among these K points, take note of data points in each category.
- Categorize the new data points to that category for which the number of the neighbor is maximum.

## C. NAÏVE BAYES

Naive Bayes is a machine learning algorithm that's used for classification purposes and is powered by probabilistic applications. The base of this classifier is based on the Bayes theorem. Bayes Theorem states that:

$$P(M/N) = (P(N/M) * P(M))/P(N)$$

Using this above formula the probability of M happening, given that N has occurred is found. This algorithm is commonly used in sentiment analysis, spam filtering, recommendations, etc. The biggest downside of this algorithm is that the predicting variables should be independent.

## D. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine produces significant accuracy with less computational power this can be used for regression or classification, it’s clear that we've used it for classification. This algorithm finds a hyperplane that makes the job of classification simpler.

Considering the training samples having the dataset Data as {yi,xi} (i=1,2,...,n) where xi ∈ R, n represents the ith vector and yi ∈ R n represents the target item. This algorithm finds the optimal hyperplane of the form

f(x)= wTx + b ; w is a dimensional coefficient vector and b is an offset.

This is performed by solving the subsequent optimization problem:

Minw, b, ξi 1 2 w 2 + C Xn i=1 ξi s.t. yi , w T

xi + b ≥ 1 − ξi, ξi ≥ 0, ∀i ∈ {1, 2, . . . ,}[7]

## E. DECISION TREE (DT)

In decision trees for training, a data set following path is performed. The construction of trees is based on entropy inputs and is drawn upside down with its root at the top. Recursive Binary splitting is performed to choose the nodes based on the cost function, we should make sure when to stop splitting hence reducing the complexity of the tree. Pruning is carried out to improve the performance of the tree.

Input: Features with a target class in the D dataset

for ∀features do

for Each sample

do Execute the Decision Tree algorithm

end for

Feature space are identified f1, f2, . . . , fx of dataset UCI.

end for

Obtain the total number of leaf nodes l1,l2,l3,. . . ,ln with its limitations split the dataset D into d1, d2, d3,. . . , dn based on the limitations of the leaf nodes.

Output: Partition datasets d1, d2, d3,…,dn.[7]

## F. RANDOM FOREST (RF)

Random Forest solves the problem of variance that arises in the case of decision trees. In this case, we'll be training a forest of bagged decision trees hence the results produced will be more accurate.

This algorithm follows the below steps

Input: original dataset

create N bagged samples of size n; where n is smaller than the original dataset

Decision Tree is trained with input as every N bagged dataset. while performing a node split, we should randomly select a smaller number with M features from all available features in the training set. The best split is selected based on impurity measures like Gini impurity or Entropy.

Combine the outputs of all the individual decision trees into a single output

find out the majority vote across all trees for each observation.

## VII. RESULTS

To calculate the classification accuracy, a confusion matrix is used. The number of attributes belonging to a particular class can be seen in a confusion matrix. In our research, we have a 2 x 2 confusion matrix because there are 2 values in the output heart disease attribute.

Class P represents heart disease predicted

Class Q represents no heart disease

**Table 2 Confusion matrix**

| | Positive | Negative |
|---|---|---|
| P | TP | TN |
| Q | FP | FN |

The term TP (True Positive) refers to the number of records that were classed as true while really being true. The term FN (False Negative) refers to the number of records that were classed as false while really being true. The term FP (False Positive) refers to the number of records that were classed as true while really being false. The term TN (True Negative) refers to the number of records that were classed as false while really being false.[1]

Accuracy = (TP+TN) / (TP+FN+FP+TN) [1]

**Confusion matrix for Logistic Regression:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 58 | 19 |
| Q (no heart disease) | 12 | 95 |

Accuracy for Logistic Regression = 85.33%

**Confusion matrix for Gaussian Naïve Bayes:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 61 | 16 |
| Q (no heart disease) | 15 | 92 |

Accuracy for Gaussian Naïve Bayes = 80.98%

**Confusion matrix for SVM:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 61 | 16 |
| Q (no heart disease) | 13 | 94 |

Accuracy for SVM = 86.41%

**Confusion matrix for K-nearest neighbors:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 51 | 26 |
| Q (no heart disease) | 27 | 80 |

Accuracy for KNN = 85.33%

**Confusion matrix for Decision Tree:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 60 | 17 |
| Q (no heart disease) | 23 | 84 |

Accuracy for Decision Tree = 77.17%

**Confusion matrix for Random Forest:**

| | Positive | Negative |
|---|---|---|
| P (heart disease predicted) | 64 | 13 |
| Q (no heart disease) | 9 | 98 |

Accuracy for Random Forest = 88.59%

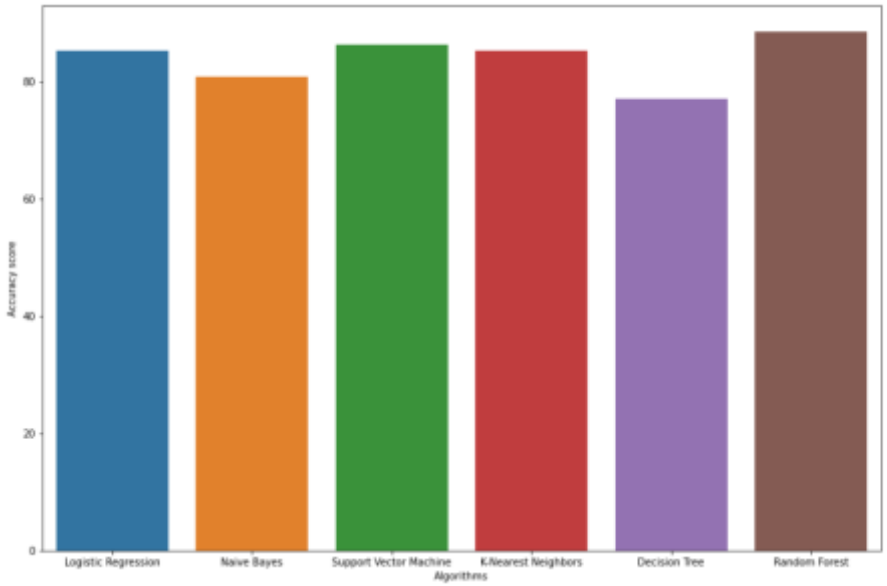



Fig 1 **Bar graph of classification algorithms**

This is the webpage of predict page when the user is logged in successfully where input is taken from the users. Once the user enters input parameters, they click Predict button to get results.

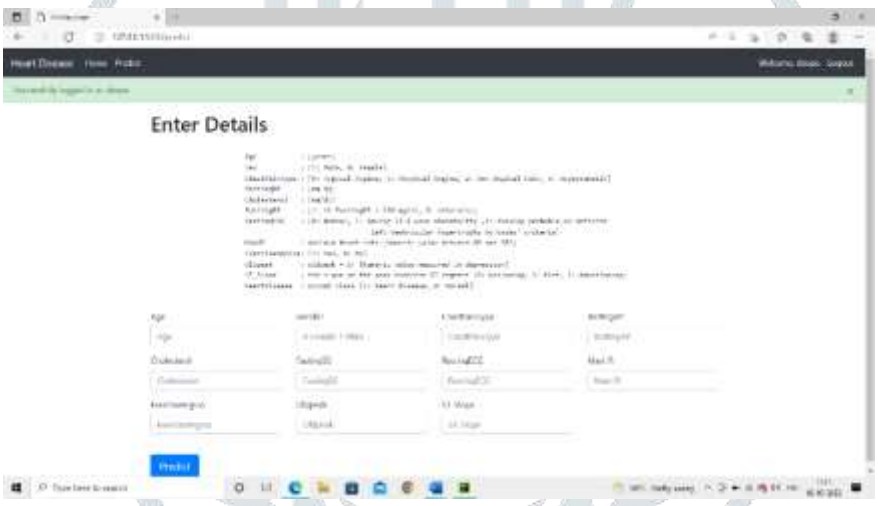



Fig 2 Successful login page

These are the web pages displayed when the user entered details have chances of getting heart disease and do not have a chance of getting heart disease respectively.

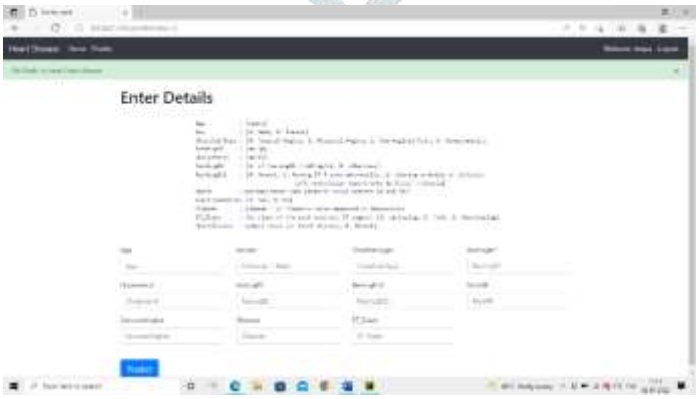



Fig 3 Patient does not have heart disease

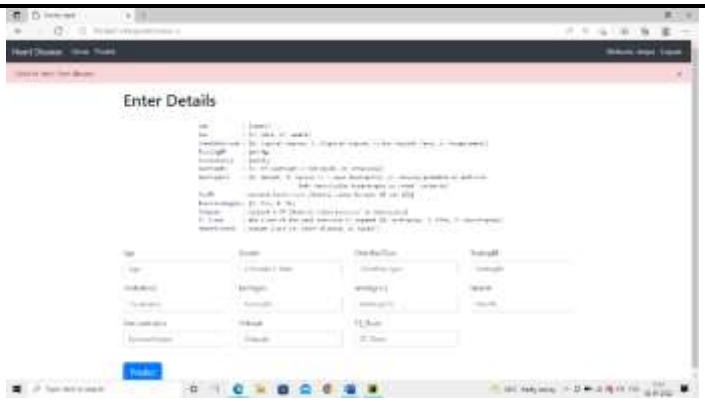



Fig 4 Patient likely to have heart disease

## VIII. Conclusion

In this paper, we put forward an approach to predicting Heart disease using the Random Forest algorithm which gives the highest accuracy than other machine learning algorithms. From this prediction model, we can estimate the condition of the heart using the medical record of the patient. We can help patients to take necessary precautions by predicting whether they're prone to heart diseases.

## IX. Acknowledgement:

This paper is a part of B. Tech final year major project. The completion of this undertaking could not have been possible without the assistance of our college - Vignan's Institute of Information Technology, Duvvada, Visakhapatnam, Andhra Pradesh 530049, India. Their contributions and guidance are gratefully acknowledged. Their support and contribution ensured the completion of the project. We are highly indebted to our Principal, our Training and Placement Dean, and our training and placement cell, who have granted us with necessary data and made it possible for us to work on it. We express our gratitude to our project review panel for gracing us with their advice. We are highly grateful to our guide Dr. E. Laxmi Lydia, who has put us on the right path and has guided us through the project from the start to the end with her abundant experience and has ensured the completion of the paper.

## References

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

[2] Chaitrali S.Dangare, Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", International Journal of Computer Applications, Volume-47, No.10, pp.44-448,2012.

[3] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction using Effective Machine Learning Techniques", International Research Journal of Engineering and Technology(IRJET), Volume-08, Issue-1S4, 2019.

[4] Sellappan Palaniappan, RafiahAwang, "Intelligent Heart Disease Prediction System Using Data Mining Techniques", (IJCSNS), Vol.8 No.8, August 2008.

[5] https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

[6] https://towardsdatascience.com/heart-disease-prediction-73468d630cfc.

[7] Galla Siva Bindhika, Munaga Meghana, Manchuri sathvika Reddy, and Rajalakshmi, "Heart Disease Prediction Using Machine Learning Techniques", Volume-01, Issue-04, 2020.