WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Survey Study about Heart Disease Prediction using Machine learning

Krinal Thakkar[1], Dr Gayatri S Pandi[2]

[1]Post Graduate Scholar, Post Graduate Department LJ University, Ahmedabad
[2]Head of Department, Post Graduate Department, LJ University

ABSTRACT

According to World Health Organization 32% of death is due to the heart disease and 85% of death is due to heart attacks and stroke.This is mostly in the low and middle-income countries , In US someone is dying due to heart attacks every 40 seconds which has been a huge concern for heart disease prediction in it. Every year 805,000 people have heart attacks in the US. Many other diseases can also be responsible for heart disease and which can put people's health at a high risk. In 2016 17.6 million people died due to cardiovascular disease which is 31% of global mortality. Each heart condition has different symptoms and other different warnings which can help a person to know about it. Most CVDs can stem interference with normal heart function Conditions like obesity , blood pressure , high cholesterol can contribute to such heart disease. All this can be prevented by proper diet , good lifestyle ,and proper sleep.The treatment of such CVD is based on its severity and several are easy to treat this disease. For patients with high BP , high cholesterol should be provided with certain medicine and drugs that can lower the condition of bad cholesterol like Lipitor and crestor. Many people are given beta blockers for lowering the high blood pressure conditions.The study focuses on developing a good prediction for heart disease using machine learning models. This should be a proper predictive model to know the hidden factors and data from the database which will be compared with the values in the trained and test dataset. This model should answer complex answers with proper data which should be useful for the people , doctors and healthcare practitioners can make this as cost effective and should help in providing effective treatments. Machine learning is a very good way for the prediction of such diseases which will be very much helpful in effective treatments. It provides the facility to improve the computer programs explicitly performed on it.

Keywords: Treatments, Machine learning models, Diagnosing heart disease, effective treatments

Introduction:

As per the modern world we are very much busy in this running race of earning money , living a luxurious and comfortable life but due to this we are not taking care of yourself and due to which our health is suffered a lot In this tension and pressure we are not having proper food habits and not proper lifestyle , people nowadays do not have a 8 hours of proper sleep and exercise due to which the body is affected a lot and we need to change our lifestyle otherwise we will suffer a lot. The quality of weather and food has also been changed nowadays due to which our heart and lungs do not get a pure air and pure food and we start suffering from heart and lungs diseases

We know that if our heart gets affected then that can cause issues in other organs too. It is a world known fact that the heart is the most essential body part and it plays a very important role in our life and so it is mandatory to take care of our heart in a proper manner.

Machine learning is a type of Artificial Intelligence where that provides the facility to improve the computer programs and also the inputs and outputs given by the programs are based on the model and depending on the quality of the data processing made on the dataset Prediction means the relationship study between the response variable and predicted variable

There is a technique which should be performed on a response variable to predict the predictor variable and find different predictions of the same response. When a patient is having a heart disease there are many factors for the same, factors can be Unhealthy food, hypertension, hypercholesterolemia, diabetes, tobacco, any major surgeries, etc. Heart attacks should be predicted months before with all this risk factors

Related Work:

The main aim for this study is to develop a prototype where we want to help the people and doctors by developing a healthcare prediction system which can bring the knowledge from hidden data in data storages,by providing such effective treatment it helps the people to reduce the costs of the treatment and we can visualize the data in tabular form and PDF forms

This Prediction means the output predicted for the input variables given to the algorithm, this process starts with the datasets when the historical or a huge data is taken, that should be pre-processed first - Data cleaning, Removing the noisy data, data transformation data reduction should be performed on the given dataset, Then the algorithm used for the model should be fit for the model and should identify the data precisely. The word "Prediction "can give a very good output in the health care, financial and stock market sectors. Machine learning model predictions allow business and private sectors to make highly accurate outcomes in the future and can also predict the fraud, hacking and such crime activities.

We need to predict the likelihood of patients getting the heart diseases, there are a lot of types of heart disease as CVD,CHD, Heart attacks, we need to find the risk factors and the relationship between those risk factors and the treatments found in them, the risk factors of the patient, which include hypercholesterolemia, hypertension, diabetes and tobacco use, along with obesity, less exercise, and elevated inflammatory markers such as CRP.

Naïve Bayes was found to be the best algorithm and with some other algorithms Artificial neural networks are also employed for the prediction of diseases.

There is maximum 14 attributes taken by the researchers for heart disease prediction and more than 300 records for it, they have used supervised learning techniques such as Naive Bayes, decision tree, random forest,KNN,etc. for predicting the accuracy of it. The dataset contains different types of attributes with their unit features. There are many types of datasets like UCI dataset, Cleveland dataset This dataset contains of 303 instances and 76 features, only 14 features are taken for testing dataset and compared with different models like KNN,Random Forest, Naive bayes, the important performance of is of different algorithms

Classification of Machine learning techniques

The classification task is used for prediction and with the different comparison of all algorithms used. Many machine learning models are used like Naive bayes,KNN, Genetic Algorithm, Random Forest with Linear model, - This model should provide a good performance and accuracy for the prediction of dataset, we need to give a improvement in health care for better results, We can use Weka tools for dataset and for more accuracy, preprocessing of data and graphs, There is a usage of different attributes taken from Weka tool and then the pre-processing is made on it. Only 16 attributes are taken from 83 attributes for testing. The comparison and analysis is made using different algorithms usingWeka tool, heart disease can be predicted and treated early and previously.

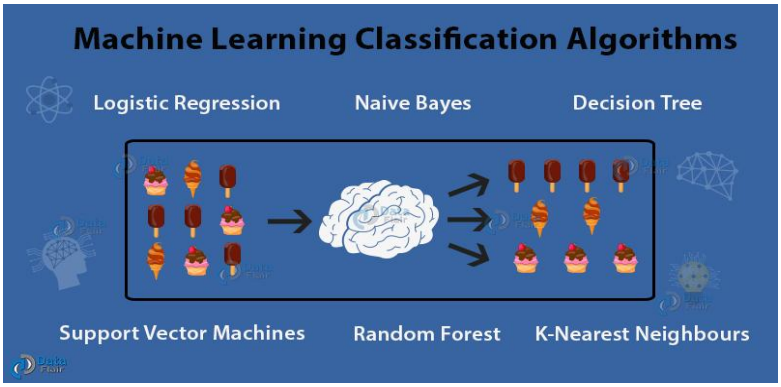

Figure 1 - Machine learning algorithms [1]

Literature Study

The paper shows that the data samples are categorized into different steps Step 1, the KNHANES-VI dataset was examined and data was selected. In step 2, statistical analysis was performed to identify features related to CHD risk. In step 3, predictors of CHD risk were selected using feature sensitivity-based feature selection in step 4, NN-based CHD risk predictors were trained using feature correlation analysis of features. In step 5, performance measurements were made to validate NN-based CHD risk predictions using feature correlation analysis.[2]

In this paper, we propose a novel method that aims at ending Signiant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. [3] The prediction model is introduced with different combinations of features and several known class cation techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFM).[3]

In this study, I consider only 14 essential attributes. I applied four data mining classification techniques, K-nearest neighbor, Naive Bayes, decision tree and random forest. The data were pre-processed and then used in the model. K-nearest neighbor, Naïve Bayes, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in K-nearest neighbors (k = 7). We can further expand this research incorporating other machine learning techniques such as time series, clustering and association rules, support vector machines, and genetic algorithms. [4]

Considering the limitations of this study, there is a need to implement more complex and combination of models to get higher accuracy for early prediction of heart disease.[4]

Cardiac surgery patients develop some form of AKI post-surgery 1–3and 1–5% develop severe kidney injury necessitating dialysis AKID.2,4–5 The mortality following AKI-D has been reported to be very high in the range of 50–80%.6,7 Even milder forms of AKI can have an impact on short term and long-term morbidity and mortality. AKI associated with cardiac surgery increases infectious risk, extends the length of stay in the intensive care unit thereby increasing the utilization of health care resources and independently predicts death.8 Recent advancements have led to less invasive surgical techniques and off–pump coronary artery bypass procedures have reduced mortality;however, the incidence of renal dysfunction has more or less remained the same.[4]

This paper defines that the number of patients with heart failure increases every day. There is a need for a system that can identify heart failure illness. For this purpose, our approach consists of developing a system that resolves the missing data problem in the preprocessing step by using the MICE model that we prove is the best algorithm to fill inexistent values.[5]

We used Boost, Ad boost, gradient boosting, extra trees, light gradient boosting Light gbm, SGDC, Nu SVM and the stacking algorithm in the classification step, the score accuracy of 95.83% was obtained by using the stacking algorithm. [5]

Cardiovascular diseases had been for a long time one of the essential medical problems. As indicated by the World Health Association, heart ailments are at the highest point of ten leading reasons for death. Correct and early identification is a vital step in rehabilitation and treatment. To diagnose heart defects, it would be necessary to implement a system able to predict the existence of heart diseases. [5]

The perspective of said project is to make out a modernistic predictive model concerning the particulars of the eudaimonia for heart sickness based on predictive modeling. Numerous supervised machine learning techniques such as classification have been applied to existing therapeutic material. Different classification techniques will be implemented and compared upon standard performance metric such as accuracy.[6]

A predictive system based on hybrid intelligent machine learning technique was adopted for treatment of heart sickness, which uses seven classifier algorithms namely LR, K-NN, DT, NB, ANN, SVM and RF with three different feature selection methods, tested on Cleveland heart sickness dataset. In terms of accuracy, from all seven algorithms, the correctness result of logistic regression with the relief method provides an improved predictive outcome for heart sickness used a naïve bayes classifier with a system for more enhanced predictions that was further developed by ensemble-based classifiers for more improvisation of system performance.[6]

Algorithms Studied

1 Naive Bayes Theorem:

Naive bayes is a supervised algorithm which helps in classifying the model efficiently , bayes theorem is based on conditional probability theory where there is condition based on variables The predicted variables are not related to each other in correlation with one another All the features used independently contribute to the data given and work with the Bayesian methods , Many complex combinations , situations use this theorem with posterior probability also in this P(X) is the prior probability and P(Y) is the predictor probability Naive bayes is very easy and simple to implement and efficient in supervised learning with non-linear data , there can be a loss of accuracy as it is based on conditional probability and which can affect the accuracy and performance of the model to be low , Average 84% accuracy can be achieved with this and not more than that with the important predictors.

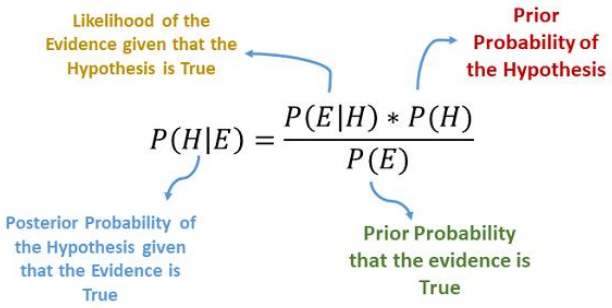

Figure: 2 Naive Bayes Theorem [7]

2 K Nearest Neighbor:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on the Supervised Learning technique. The algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories'- 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 good suite category by using K- NN algorithm. The K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for Classification problems.[8]

K-NN is a non-parametric algorithm, which means it does not make any assumptions on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.[8]

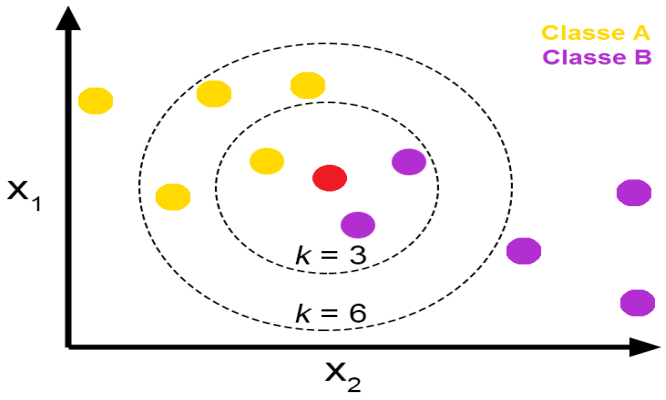

Figure 3 K nearest neighbor - calculating the difference between classes [9]

3 Support Vector Machine:

Support Vector Machine is a supervised learning technique which can be used for classification and regression problems. It is mostly used for classification problems only;in this we have a hyperplane to plot the points in it. Each point plotting with data item as point is with n dimensional space, we can perform classification with hyperplane that is different from other classes given. They are separated into different classes of hyperplane.
The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.[10]

Support Vector Machine is a classification technique of Machine learning, which is used to analyze data and discover patterns in classification and regression analysis. SVM is typically mull over when data is characterized as a two-class problem. In this strategy, data is characterized by finding the best hyperplane that isolates all data points of one class to the other class. The higher separation or edge between the two classes is, the better is the model, considered. The data points lying on the limit of the margin are called support vectors. The actual basis of SVM is mathematical methods used to design complex real-world problems. We have chosen SVM for this experiment because our dataset - Cleveland Heart Disease [11]
Dataset CHDD has multi classes to predict based on various parameters. In SVM, the mapping of training data is to be done with a function called kernel [11]

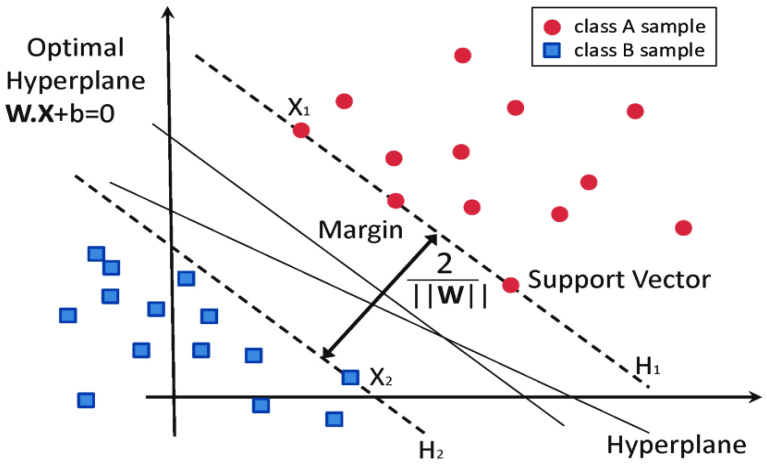

Figure: 4 Support vector machine example [12]

4 Random Forest Classification:

It is one of the supervised machine learning algorithms which is used for both classification and regression also. However, it is mainly used for classification purposes. The name itself suggests that it is a forest, a forest is a group of trees similarly in a random forest algorithm we will have trees these trees are the decision trees. If we have a higher number of decision trees prediction results will be more accurate. Random forest algorithm works this way; first it will collect random samples from the dataset and then it will create decision trees for each sample from those available trees. We will select the tree which will produce the best prediction results.[13]

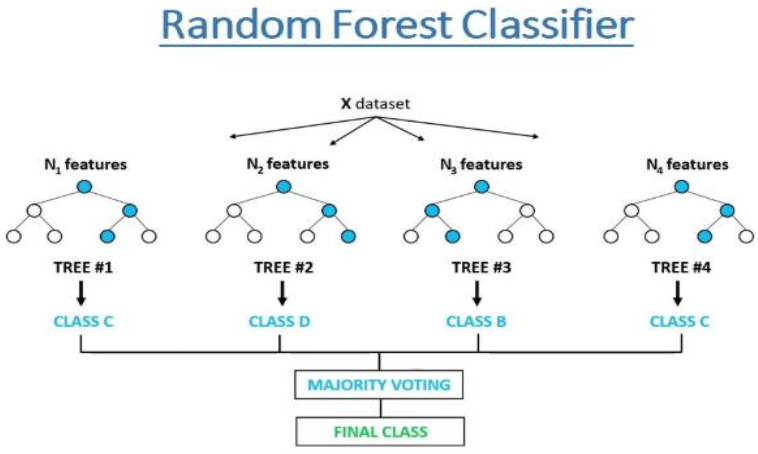

Figure:5 Random Forest Classifier [14]

5 Multilayer Perceptron Classifier:

The multilayer perceptron (MLP) is a widely used supervised neural network used in medical decision support systems as a result of its comprehensible architecture as well as its comparably easy algorithm. The nodes of the network organized in layers form the multilayer perceptron. A standard multilayer perceptron network entails three or more processing nodes levels, that is, an input layer whereby inputs are obtained externally, several hidden layers, and, finally, an output layer that generates the outcomes of the classification process. The analysis of these classifiers is based on the overall accuracy of instance classification and the use of the confusion matrix. [15]
The accuracy is the proportion of instances that are correctly classified. The confusion matrix is used to display the number of accurate and inaccurate projections that the model formulated concerning the true classifications in the experiment data. [15]
MLP is a relatively simple form of neural network because the information travels in one direction only. It enters through the input nodes and exits through output nodes. This is referred to as front propagation only. Interestingly, backpropagation is a training algorithm where you feed forward the values, calculate the error and propagate it back to the earlier layers. In other words, forward-propagation is part of the backpropagation algorithm but comes before back-propagating the signals from the nodes [16].

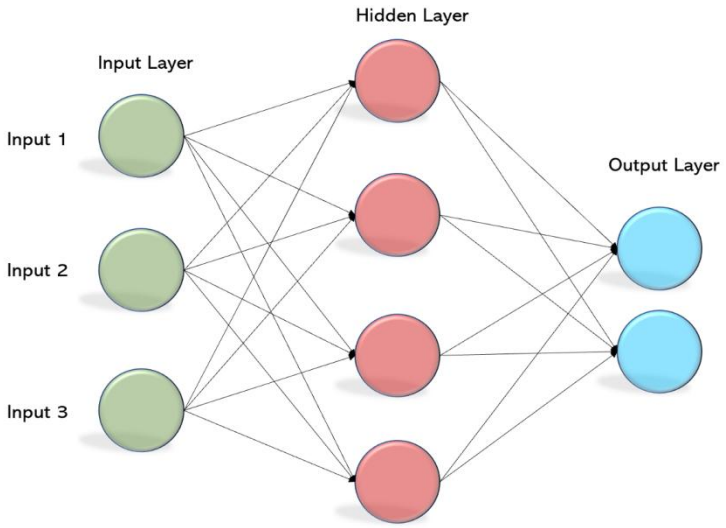

Figure:6 Multilayer Perceptron Classifier [17]

Proposed System:

GA is search heuristic approach , process of natural selection where the fittest individuals are selected for reduction in order to produce the offspring of the next generation ,We need to make a proper sequential process of GA and start with the process The natural selection which is the process of selecting the fitness of the features selected and then process it and create the next offspring, how better are this features will be as the offspring as how are parents as the children becomes , This is an iterative process at end generating the offspring and the most fittest individuals should be found

The main phases are
1 Initial Population
2 Fitness function
3 Selection
4 Crossover
5 Mutation

1 Initial Population: The process where the individuals are called population,each individual is a set of variables and are called Genes, they join to form a chromosome

2 Fitness function: It displays how to fit the individuals and gives the scores to each individual, the probability that one should be selected is dependent on the fitness score

3 Selection: This is the phase where to select the fittest individual and then send them for the next generation. Two pairs should be selected and should be sent as per selected fitness scores and individuals with good fitness score should have more chance to be selected.

4 Crossover: This is the most important phase in GA, where each pair of individuals are to be mated, a point is taken for random with the genes. Offspring are created by giving the genes of the fittest parents to each other until crossover is reached to it point

5 Mutation: Here, new offspring are formed and some genes are given for the mutation with low probability, some of the offspring can be swapped if needed This is mandatory to maintain the diversity with the individuals and all coming offspring.

Pseudocode

START

Generate the initial population

Compute fitness

REPEAT

Selection

Crossover

Mutation

Compute fitness

UNTIL population has converged

STOP

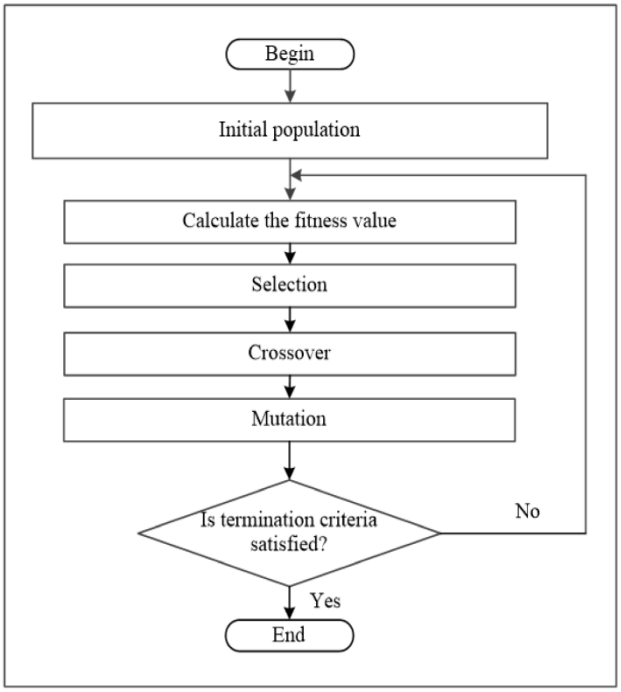

Figure:7 Flow of Genetic Algorithm [18]

Conclusion:

This study will help us to know the strength and weaknesses of the other algorithms and models used and so we will be able to improve the prediction of heart disease by increasing the accuracy and performance of the model by getting the best dataset and the most efficient attributes in it which will provide the best accuracy, in model fitting the most important thing is feature selection and data pre-processing. The existing systems are focused on the single attributes and single models, we can use more complex combinations of models and analyze and compare our results with other models , we should used more than 16 attributes to and make a model based on the fitness functions given by genetic algorithm and send them for crossover process , we can also implement the number of training and testing data for more accurate results by the whole process of genetic algorithm we can make the probability of the fitness function used by using Bayes theorem and make better results In this way it should be more accurate and can have good performance of the prediction methods used

References:

[1] Malini Shukla Machine Learning Classification – 8 Algorithms for Data Science Aspirants Access from https://data-flair.training/blogs/machine-learning-classification-algorithms/ Access on 29/01/22

[2] Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis Jae Kwon Kim and Sanggil Kang

[3] Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava

[4] Heart Disease Prediction using Machine Learning Techniques Devansh Shah, Samir Patel& Santosh Kumar Bharti Article number: 345 (2020)

[5] Machine learning-based identification of patients with a cardiovascular defect Nabaouia Louridi, Samira Douzi &Bouabid El Ouahidi
[6] Prediction of Heart Disease using Machine Learning Algorithms Garima Choudhary Dr. Shailendra Narayan Singh Computer Science & Engineering
[7] LaptrihnX LaptrinhX Blog about Programming, Open Source, Technology, Software and IT Jobs.Access from - https://laptrinhx.com/ Access on : 29/01/22
[8] Sonoo Jaiswal Javatpoint, Access from - www.javatpoint.com, Access on 29/01/22
[9] Vincent Granville Towards Data Science -Access from - https://towardsdatascience.com/ Access on 29/01/22
[10] Sandeep Jain GeeksforGeeks Access from https://www.geeksforgeeks.org/ Access on 29/01/22
[11] Heart Disease Prediction using Machine Learning Techniques by VijetaSharma,ShrinkhalaYadav,Manjari Gupta
[12] Virologist Dr. Ijad Madisch Discover scientific knowledge and stay connected to the world of scienceAccess from www.researchgate.net/ Access on 29/01/22
[13] Predicting Heart Disease at Early Stages using Machine Learning: A Survey by Rahul Katarya,Polipireddy Srinivas
[14] Quincy Larson Learn to code — for free. Build projects. Earn certifications.Access from www.freecodecamp.org , Access on 29/01/22
[15] Heart Disease Risk Level Prediction: Knitting Machine Learning Classifiers by Mrs. Kelibone Eva Mamabolo Dr. MoeketsiMosia
[16] Michael Heinze DataSklr Access from www.datasklr.com Access on 29/01/22
[17] Max Tegmark Becoming Human-Exploring Artificial Intelligence and what it makes to human Access from https://becominghuman.ai/ Access on 29/01/22
[18] Dr. Shu-Kun Lin Advancing Open Science for more than 25 years Access from /www.mdpi.com Access on 29/02/22