Turkish Journal of Computer and Mathematics Education Vol.12 No.9 (2021), 2023 - 2037

Research Article

A Survey on Heart Disease Early Prediction Methodologies

Saiyed Faiayaz Waris [a], S. Koteeswaran [b]

[a]Research Scholar , [b]Department of Computer Science and Engineering ,VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India
[b]Associate Professor , Department of Computer Science and Engineering ,VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India.

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract: In today health trend, the deaths are increasing due to the heart diseases. The lives to be saved and deaths need to be minimized by determining the heart disease on the samples of health patient sources derived from medical clinics. The appropriate treatments are to be guided and prescribed as the follow up. For predicting the heart diseases in advance, one of the factors significantly assumed is accuracy. Based on this factor, there were many methodologies are taken as study and compared with few factors. The review over those methodologies suggest that new methods are more sophisticated and are more reliable in determining the heart disease with more accuracy. The approaches that would be described in terms of their working theme and their accuracies are noted. The domains from which the techniques, the tools, the datasets are taken are from data mining, machine learning, deep learning and other type, python environment and other relevant type, Cleveland or Kaggle and other specified kind of data respectively. The accuracy of the described approaches are represented in a pictorial form.

Keywords: Methodologies, accuracy, heart disease, attributes and performance

1. Introduction

There are many methodologies existed from one domain to other domain such as from data mining to machine learning. They are all useful to analyze over data manually or over a dataset taken from medical or hospital dataset. The deaths are increasing from the world population because of many diseases in which heart disease is threatening in many ways. In order to save the lives before getting heart problem, prediction to be made using available methodologies based on accuracy. The comparison among them is to be done based on accuracy.

There are the approaches that need to be described in order to know their working procedure. These approaches are demonstrated in terms of short pseudo codes, and accuracy along with performance.

Table 1: The analysis requirement on the approaches

Name of the Methodology
Support Vector Machines
Naïve Bayes
Logistic Regression
Decision Tree
KNN
Random Forest
LightGBM
XGBoost
Hybrid frameworks, if any
Specific tool Environment over specific methodology for HDP

The above methods are explained and their accuracy to be increased using any efficient tool or application available in the market.

The attributes considered from a Cleveland dataset is provided as follows. There are certain cases where one or few extra attributes are included based on their impact on heart.

Table 2: List of attributes that have relationship with the heart

Factor	**Meaning**	**To be in specific Range**
Age	Person's age	29 - 79
Sex	Person's gender	0,1
Cp	Chest Pain Type	1,2,3,4
Tresbps	Resting blood pressure in mm Hg	94 to 200
Chol	Cholesterol in mg/dl	126 to 564
Fbs	Fasting Blood Sugar in mm/dl	0,1
Restecg	Resting Electro cardio-graphic results	0,1,2
Thalach	Maximum heart rate achieved	71 to 202
Exang	Exercise Induced Angina	0,1
Oldpeak	ST depression by exercise relative to rest	1 to 3
Slope	Slope of the peak exercise ST segment	1,2,3
Ca	Number of major vessels colored by fluoroscopy	0 to 3
Thal	Kind of defect	3,6,7
Target or Num	Class attribute	0 or 1

Each value for an attribute that listed in Table addresses the seriousness of the coronary illness. Basically, an ensemble method like bagging and boosting aimed to increase the accuracy of the classifiers. They consider a combination of the classifiers for further process of evaluation.

2. Literature Review

There are many methodologies are taken as study. The following are the works that demonstrate the heart disease prediction. As per the source of (S.Raguvaran,2016), there are four best methodologies are applied such as Logit Model, Neural Network, KNN, Random Forest Classifier in determining the early prediction of coronary illness. The point of this investigation is to provide more accuracy with minimum error rate. Among 4 studies, Logistic Regression is proved having best accuracy over others. With the respect of (K. Srinivas,2020), the machine learning method called HLRM is used in 2 phases where the first phase comprises of data pre-processing, KNN, and Principal Component Analysis are applied to determine factors that caused the heart disease extrapolation and the second phase consists of stochastic gradient descent linear regression is used to quantify the correlation between the predictor and dependent outcome variables. Finally, the accuracy achieved is around 89% and this would helpful for medical and research analysis. This would help saving many lives in our world. In the view of work demonstrated in (Kusuma.S,2018), the benefit of using machine learning and deep learning techniques and their simulation tools are used not only for heart disease prediction but also for bioinformatics in future too. The aim of this study is survey on various approaches over HD in terms of cardiovascular type in order to obtain accuracy by comparing the existing methodologies. In the view of (Purushottama C,2016), the

framework is built that would use certain possible rules which would generate the prediction of the heart disease for the tenets. Even the non-specialized doctors would tell the status of the disease. It is proved as well said potential in judging the illness of the heart. The variations of the rules generated such as original, pruned, without duplicates, classified and polish. As per view of (Kennedy Ngure Ngare,2019), the effective method to predict heart disease can be achieve using naiveBayes and decision tree algorithms. In this, multilayer discernment neural organization with back engendering as preparing calculation, and the particular symptomatic framework is utilized to precisely foresee the heart illness. In the regard to (V. Krishnaiah,2013), the decision trees is proposed to predict the heart disease issues using CRISP-ADM kind of management that collects medical data and manage the large sized information databases provided from day to day storage. It uses decision tree growth algorithm and four phases of CRISP-ADM in which interaction happens between reference model and user guide. It works on pages of a simulated tool, and produces the status as PASS or FAIL though the query. In the source mentioned w.r.to (Mohamed Djerioui,2020), the two techniques such as MLP and LSTM are applied in which LSTM is proved to provide best accurate results in the prediction of heart disease. The LSTM is proven in giving the accurate results for issues of heart disease. In the view of source mentioned in (C. Beulah Christalin Latha,2019), there is a need to apply ensemble methods over classification algorithms in order to increase the accuracy more than 7% than usual accuracy of them. The ensemble method applied in terms of boosting and bagging is not only for enhancing the accuracy but also the predict the coronary infection in the inception stage. In the regard of resource (Subhankar Rawat,2019), the sample data recommended such as age, sex, chest pain type, resting blood pressure, cholesterol serum, fasting blood sugar, resting electrocardiography, maximum heart rate exercise induced angina, old peak, slope, ca, thal, and numare taken from Cleveland database available from UCI repository. There were certain methodologies are taken from machine learning domain and are demonstrated in detail and are compared based on accuracy. As per the source given in (Keshav Srivastava,2020) KNN which is a data mining technique is applied using frameworks such as Flask and Piggle packages over the web app and is processed over certain attributes in order to predict the heart disease with the better accuracy. In the regard of source given in (Honey Pandey,2021), the particular machine learning strategy called Support Vector Machine is applied over the clinical information and the subtleties are put away in a google sheet for planning and testing. With the view of (Devansh Shah,2020), there were data mining and KNN proven with more accuracy compare to other ML techniques such as random forest, Decision tree and Naïve Bayes. The referencesprovided in (Santhana Krishnan J,2019), the specific ascribes are considered for assessment of HD utilizing the information mining procedures, for example, Naive Bayes and Decision Trees with great exactness. In the information provided in (Archana Singh,2020), many ML methods, for example, KNN, DT, logic model and Support vector machines are applied for training and testing over a UCI repository and are checked in python anaconda environment. As per many approaches provided in (Khaled Mohamad Almustafa,2020), all those approaches are compared over Cleveland and other country data sources with respect to minimal attributes. Many machine learning techniques are analyzed in which KNN is proved with best accuracy more than 86% in the phases of training and testing. As per the source mentioned in (Amit Chauhan,2020), the huge arrangement of libraries upheld by the python, the hardware and planned space of ML and information mining classifiers are utilized to anticipate heart disease. With respect to studies specified in (Latchoumi, T. P.2013), the KNN with 10-fold mechanism is applied over the dataset to reduce noisy points in the training set and KNN is defined for two examples in which one is pizza on which pineapple flavor and other is generating recommendations relative to post movie for MovieDB website. As per the source mentioned in (K-Nearest Neighbours, https://www.geeksforgeeks.org/k-nearest-neighbours/), few environments are defined for KNN in order to predict heart disease. With respect to, the validation and training error to be minimized by picking the k value from the KNN in the environments such as R and Python for the prediction of HD. In the view of (Ranjeeth, S,2019), the attributes such as age and loan are considered for credit assessment over a set of samples for determining the optimal k based on Testing and justification of dataset..

As per, KNN is to be demonstrated for the application intended. In the opinion of (S.Raguvaran,2020)**,** the approaches such as ANN, KNN, random forests and logistic regression are applied for predicting heart diseasein which latter KNN is found having more accuracy. As per the data mentioned in (Galla Siva Sai Bindhika,2020), the best accuracy is obtained without any equipment by using combination of statistical model and random forests. With respect to (Shankar, G,2020; Megha Kamboj,2018) first study deals with the various ML algorithmare used to predict heart condition beforehandto avoid mishaps and suggest the better treatment, and second study deals with comparing them based on accuracy. As per data provided in , the machine leaning and deep learning methods are applied in order to extract the significant details for making the decisions using available clinical and EHR results. In the view of , the first study deals with increasing the accuracy by increasing the k value, the process is stopped when saturation is met, and second study deals with the combination of KNN and genetic algorithms are used for diagnosing the heart disease with more accuracy. As per the data of [31], one of the stages of classification called PSO is used to move noise along with KNN for improving the accuracy among several classification methods. As per the sources mentioned in, the primary study deals with KNN in the python

environment with best k value which in turn decrease the inaccuracy andrise the precision, and second study deals with few platforms on which python with certain libraries are how they are to be installed in order to predict the outcome based on the intended methodologies. In the view of source indicated by [34], the infrared light absorption detector is applied over the patient finger tip, to predict the CVD risk. The wave form is generated from the extracted features and SVM is applied to predict accuracy as directed by PWV. Hence, this approach helps as a tool to avoid CVD and guarantees accuracy > 85%. In the view of , the first study focus on the ways of carrying out the approach, also description of how to predict the class for given dataset., and second is focus on the steps that take up from loading the data, transforming using specific fitted function, and outputs based on kind of logistic regression. In the source of, the first study deals with the decision tree is used for predicting the class for each example in the problem using heuristics such as entropy with information gain. In this, when the tree becoming larger and to avoid such scenario, set minimum number of coaching inputs on each leaf or length of longest path from root to leaf, and second study demonstrates on the scikit-learn package is used for building optimal Decision Trees in order to predict the classes using decision tree classification as well as attribute selection measure concepts. In the view of , it demonstrates the advantages as well as disadvantages of the random forest, also its features are discussed along with differences between decision trees and random forests. In the view of , the accuracy for catBoost is proved 88% and XGB is proved 85% for the considered dataset. In regard to source is about SVM mentioned in, the formulae are provided in order to compute accuracy, misclassification, precision, sensitivity, and specificity and is a two-fold process for a database of pregnant women. In regard of , the methodology of SVM is described in python environment and is applied over non-linear dataset. In the view of (Naive Bayes Classifiers, https://www.geeksforgeeks.org/naive-bayes-classifiers/) the working methodology of naïve bayes is demonstrated with iris, and other datasets and provides how to compute the accuracy also. In the view of, the theme of logistic regression is demonstrated through an example database and given how to compute accuracy also. In the sources specified in the accuracy and theme of the approach is described with an example. In the view of, the two measures are taken into consideration for the categorical dataset and is applied over heterogeneous datasets. The working of KNN is also demonstrated in this. As per the sources specified in, the working flow of random forest approach is provided along with an example.

All the above methods which are all discussed demonstrate directly or indirectly about the prediction of class label with certain amount of accuracy and certain time of executing the procedure.

3. Working of the methodologies: There are certain methodologies are taken in to account, for prediction of heart disease. The core concept of each methodology is defined using their pseudo procedures.

The pseudo procedures of existing methodologies are defined in the following:

Pseudo_Procedure Support_Vector_Machines(Dataset[][])
Aim: Converts lower data set into higher dimensional data set, No. of epochs are reduced with the regularization parameter λ. Input: Dataset Output: Class label for each tuple • Classify the proper hyper plane w.r.to numeroussituations. • Categorize the two modules • Find the right hyper plane to isolate into classeswhere margin is used between nearest points and hyperplane. • Additional feature is added such as $z=x^2+y^2$ and plot on x and z now. • the Kernel trick and decision function in non-linear scenario is defined as $g(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \right)$. Where n indicate support vectors , yi are target marks to x, and b is determined.

The following is an example where SVM is applied in order to predict the class label for the given image.

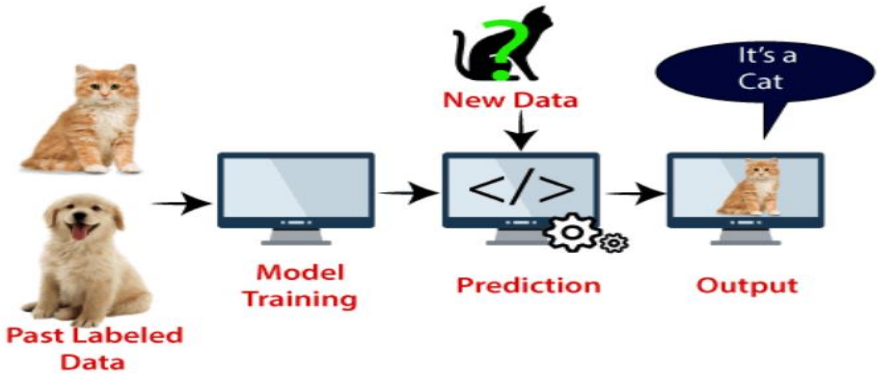


Fig.1: Detecting new sample based on trained objects using SVM

In order to detect the accuracy based on confusion matrix, the following formulae are used:

Exactness = (SV+ SN)/ (SV + VS+ VN + SV

SV =True positive

SN= True Negative

VS=False Positive

VN= False Negative.

Mismeasure (all incorrect / all) = VS + VN / SV + SN + VS + VN

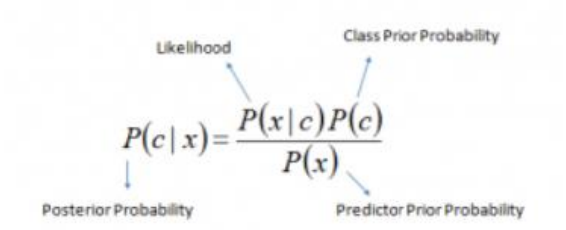
Precision = SV / SV + VS

aka Sensitivity = SV / SV+ VN

Selectivity =SV / SN +VN

Nextmethodology is to be discussed is Naives Bayes Theorem, which is based on Bayes Theorem:

The posteriori probability is obtained from the following formula:



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Above,

- $P(c|x)$ is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

Fig.2: Formulae for computing posteriori probability based on Bayes Theorem

Pseudo_ProcedureNaïve Bayes (Dataset[][])

Aim: Predicts class for a tuple where one attribute is independent of other attribute.

Input: Dataset

Output: Class label for each tuple

- Change data into frequency table
- The result with most elevated posteriori likelihood is taken for arrangement and formulae are. It predicts the tuple X has a place with class Ci if and just if

$$P(C_i|X) > P(C_j|X) \quad for \;\; 1 \leq j \leq m, \; j \neq i$$

- For maximizing $\boxed{P(C_i|X)}$ for class C_i, the bayes theorem defines the following

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

- If the attributes are conditionally independent, the following to be hold

$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$ where Xk alludes the estimation of characteristic Ak for tuple X

- It predicts the class mark for X is Ci if and if $P(X|C_i)\,P(C_i) > P(X|C_j)\,P(C_j) \qquad for \qquad 1 \leq j \leq m, \; j \neq i$

The accuracy of this approach is computed based on

Accuracy= No. of tuples Correctly_predicted / Total number of actural tuples.

Pseudo_ProcedureLogistic Regression(Dataset[][])

Aim: Predict the output which either 0 or 1 for our case. It also have multinomial and ordinal which output any one from more than 3 possible dependent variables.

Input: Dataset

Output: Class label for each tuple

- The hypothesis is defined for prediction using the following where β kinds are regression coefficients.

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{bmatrix}$$

Then, in a more compact form,

- $h(x_i) = \beta^T x_i$
- Some modifications are incurred over the hypothesis for further classification.

$$h(x_i) = g(\beta^T x_i) = \frac{1}{1 + e^{-\beta^T x_i}}$$

where,

$$g(z) = \frac{1}{1 + e^{-z}}$$

is called **logistic function** or the **sigmoid function**.

The accuracy of this approach is computed based on

Accuracy= No. of tuples Correctly_predicted / Total number of actural tuples.

Pseudo_Procedure Decision Tree(Dataset[][])
Aim: It is initialized by Leo Breiman, University of california. It is based on Entropy and Information gain in order to determine the class labels in a visualization tree. Input: Dataset Output: Class label for each tuple • Create a root first. • Calculate entropy for current state i.e. H(S) through an example for the given dataset. Yes No Total 9 5 14 $Entropy(S) = \sum_{x \epsilon X} p(x) \log_2 \frac{1}{p(x)}$ $Entropy(S) = -\left(\frac{9}{14}\right) \log_2 \left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2 \left(\frac{5}{14}\right)$ $= 0.940$ • Calculate entropy for the attribute 'x' denoted by H(S, x) for each attribute in order to split that node into branches. • Calculate the Information gain for the attribute x for its values using attribute determination estimates, for example, Information gain $IG(S, Wind) = H(S) - \sum_{i=0}^{n} P(x) * H(x)$ • For all the examples given, the labels are determined using above measures.

The following specifies general structure to be provided by decision tree induction:

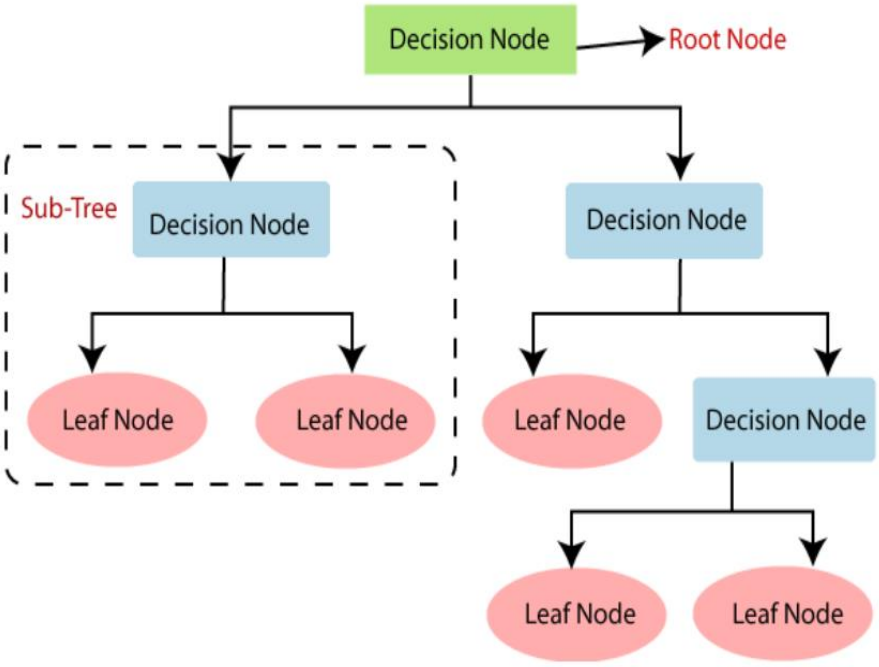


Fig.2: Structure of Tree provided by Decision Tree Induction classification

The accuracy is determined based on number of tuples correctly classified and total number of tuples.

Accuracy= No. of tuples Correctly_predicted / Total number of actural tuples.

Pseudo_Procedure KNN(Dataset[][])
Aim: It is determined initially by Evelyn Fix and Joseph Hodges . It determines class label as most frequent class from results for the new sample. Input: Dataset Output: Class label for each tuple • Feed the dataset • Fix the value for k • Iterate till the last sample in the training set for the correct prediction – - Compute distance between training and test sample using Euclidian or manhattan or hamming distance, Here how to compute Euclidian distance is as follows: $$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$ - Assign the class label for each attribute based on most frequent class defined for the given samples.

Pseudo_Procedure Random Forest(Dataset[][])
Aim: It is invented by Tin Kam Ho and it considers the multiple decision trees, average the outcome of such trees and output the higher accuracy in determining the class label for the attribute. Input: Dataset Output: Class label for each tuple • Selection of random k information focuses in the dataset. • Build up Decision trees for related chosen information focuses. • Pick the number N for the Decision trees to develop. • Recurrence first and second points • For each new data point, assign the class that obtained from majority of voting.

The operational of random forest algorithm is representedin the following diagram:

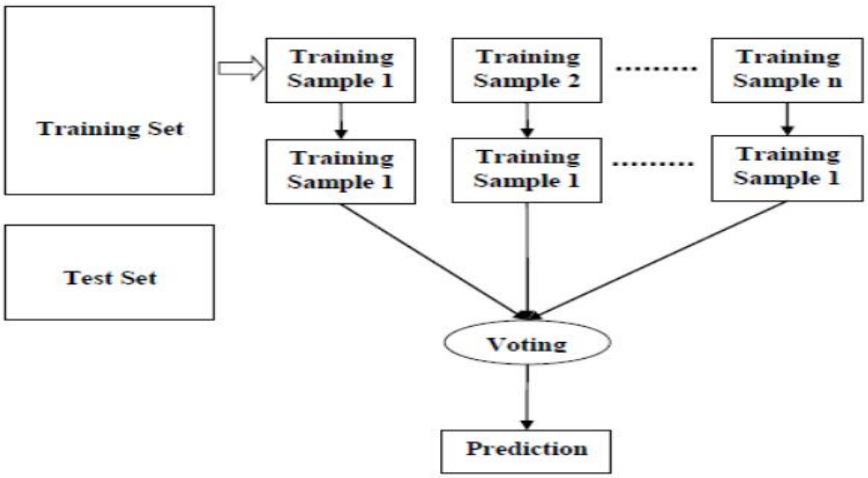


Fig.3: Theme of Random Forest approach

Pseudo_Procedure LightGBM(Dataset[][])
Aim: It is de-facto algorithm invented by Guolin Ke that wins competitions over kaggle site analytics and is extremely powerful. Input: Dataset Output: Class label for each tuple • Declare feature vector and target variable • Splitting dataset into training and test set • Model development and training where loaded dataset converted into LGBM format with parameters and their values, accuracy is modeled based on these values by performing the many number of iterations. • Predict the class and compute the accuracy.

Pseudo_Procedure XGBoost(Dataset[][])

Aim: It is coined by Taingi Chen. It is based on gradient boosting and other variations and trains the model by efficiently making use of available resources.

Input: Dataset

Output: Class label for each tuple

- The target variable y to be prophesied from early model F0.
- A new model h1 is used to fit residuals from the precedingstage.
- New model F1 is formed by combining F0 and h1, which is improvedtype of F0. The mean square error from F1 is inferior than F0 is defined as $F_1(x) \leftarrow F_0(x) + h_1(x)$
- New model F2 to be formed to progress the enactment of F1 and after residuals of F1 is defined as $F_2(x) \leftarrow F_1(x) + h_2(x)$.
- This process is repeated for n iterations until residuals are minimized and generalized as $F_m(x) \leftarrow F_{m-1}(x) + h_m(x)$.
- Additional learners are used to bring down the errors without affecting the functions formed in previous steps.
- For an example, F0(x) for given dataset is defined to limit misfortune capacity or MSE in this circumstance as $F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$ $argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} (y_i - \gamma)^2$.
- The boosting capacity is characterized by taking differential w.r.to γ as $F_0(x) = \frac{\sum_{i=1}^{n} y_i}{n}$.
- In this process, the gradient of loss function is defined iteratively as $r_{im} = -\alpha \left[\frac{\partial (L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}$, *where α is the learning rate* .

The boosted model Fm(x) for each hm(x) on each step using multiplicative factor γm is defined as $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$.

Pseudo_Procedure CatBoost(Dataset[][])

Aim: It is developed by Yandex. It is useful for gradient boosting framework that attempts for categorical features. It uses efficient encoding similar to mean encoding and reduces over-fitting.

- Multiple random permutations are applied over a set of input observations.
- Perform conversion from categorical or floating to integer using

$$avg\_target = \frac{countInClass + prior}{totalCount + 1}$$

Where countInClass denotes how many times label value became 1 for object s with current categorical value, prior is preliminary value determined by preliminary parameters, and totalcount denote objects whose categorical values matching the current one.

4. Results: In this, the various methodologies which have discussed with their procedures in previous chapter would be demonstrated based on certain factors such as accuracy and performance metrics.

The following table represents the name of the methodology, purpose, advantages of each of it.Table 3: Role and the advantages of the existing methodologies

Name of the Methodology	**Purpose**	**Advantages**
SVM	It is a non-linear separable problem and uses a kernel trick which transforms from low dimensional space into high dimensional space.	• Powerful in high dimensional space. • It is memory proficient as a result of subset of preparing focuses utilized in choice capacity (SV).
Naïve Bayes	It is probabilistic classifier that determines class for the tuple irrespective of occurrence of one attribute with other attribute.	• It is applicable to high dimensional data. • Make quick predictions.
Logistic Regression	It classifies based on sigmoid function, and rounds based on the condition to nearest binary value 0 or 1.	• Need not pick up learning rate. • Run faster occasionally.
Decision Tree	It generates a tree where conditions are the nodes which cause splitting of the node, and leaves denote the class labels for the given problem.	• Its performance won't be affected by on-linear relationship. • It performs feature selection implicitly. • It performs human activity as a automation process.
KNN	It is also a classification which determines class label for new sample based on most frequent classes for the top rows.	• Decreases the error rate as k value increases
Random Forest	It is based on kind of ensemble approach called bagging and improves the accuracy by combining the multiple classifiers.	• Takes less time for training. • Produce higher accuracy even for large datasets and avoids over-fitting. • Produce higher accuracy even large portion of data is missed.
LightGBM	It is 6 times faster than XGBoost and is extremely good for large scale datasets. It increases actual accuracy of any classifier into best accuracy by combining more number of classifiers.	• It grows the tree vertically and is leaf-wise grow the tree, which leads to reduce more loss. • Handles large dataset, faster training speed, less memory usage, and supports GPU learning.
XGBoost	It develops a predictive model with accuracy on the unseen data.	• To get execution speed • Model the prediction • Fast learning through distributed and parallel computing.

Cat Boost	It is useful for prediction over the categorical features. It works on indices of categorical attributes and also without indices.	• More accuracy with prediction than previous two boosters LGBM and XGB. • Execution takes very less time compared to other boosters.

The following table gives the accuracies to be guaranteed by the various machine learning approaches that taken into the study.

Table 4: Accuracies of described methodologies for the heart prediction

Name of the Methodology	**Accuracy (%), X**
Support Vector Machines	88<=X
Naïve Bayes	95<=X
Logistic Regression	92<=X
Decision Tree	79<=X
KNN	88<=X
Random Forest	97<=X
LightGBM	92<=X
XGBoost	85<=X
CatBoost	88<=X

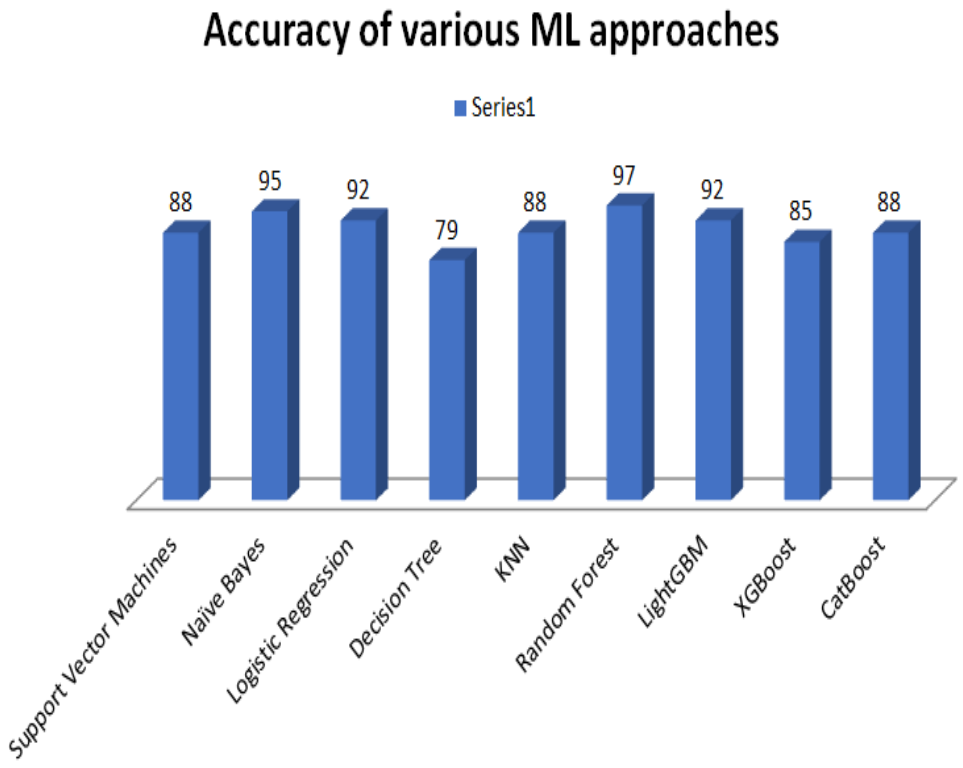


Fig. 4: Accuracy of reviewed methodologies

The two parameters such as sensivity and specificity observed as a test to prove correctly the person has a disease and prove the person has not correctly a disease.

The Figure 1 represent a graph on these existing approaches and gives visual appearance of their accuracies for the end user to understand easily which method is having what accuracy. From this, which approach is best applied and which approach is having least significant in the prediction of class labels based on accuracy.

5. Conclusion

In account of saving many lives, the early prediction of cardiovascular disease might be helpful in suggesting the further steps to adapt. The accuracy of the described methodologies for the prediction of HD is observed with specific accuracies. The order of methodologies if consider based on accuracy in descending manner are Random Forest, Linear Regression, LightGBM, KNN, CatBoost,XGBoost, SVM, Naïve Bayes, and Decision Trees. Among these, certain methodologies accuracy is to be increased by 7% using ensemble methods. Still, among these, the accuracy of few methods are increased using certain intermediate boosters. Hence, the review over many existing approaches over HD would helpful to recommend next treatment to the patients to save their lives.

References

1. Allison Ragan, Taking the Confusion Out of Confusion Matrices, October, 2018, https://towardsdatascience.com/taking-the-confusion-out-of-confusion-matrices-c1ce054b3d3e.
2. Amit Chauhan, Heart Disease Prediction using Machine Learning with Python, October 2020, https://towardsai.net/p/machine-learning/heart-disease-prediction-using-machine-learning-with-python
3. Archana Singh, Rakesh Kumar,Heart Disease Prediction Using Machine Learning Algorithms, ICE3,2020, https://ieeexplore.ieee.org/abstract/document/9122958
4. Aroulanandam, V.V., Latchoumi, T.P., Balamurugan, K., Yookesh, T.L. (2020). Improving the energy efficiency in mobile Ad-Hoc network using learning-based routing. Revue d'Intelligence Artificielle, Vol. 34, No. 3, pp. 337-343. https://doi.org/10.18280/ria.340312
5. Avinash Navlani, Decision Tree Classification in Python,https:// www.datacamp.com/community/tutorials/decision-tree-classification-python.
6. Balamurugan, K., Uthayakumar, M., Gowthaman, S. and Pandurangan, R., 2018. A study on the compressive residual stress due to waterjet cavitation peening. Engineering Failure Analysis, 92, pp.268-277.
7. Balamurugan, K., Uthayakumar, M., Ramakrishna, M., & Pillai, U. T. S. (2020). Air jet Erosion studies on mg/SiC composite. Silicon, 12(2), 413-423.
8. Balamurugan, K., Uthayakumar, M., Sankar, S., Hareesh, U. S., & Warrier, K. G. K. (2019). Predicting correlations in abrasive waterjet cutting parameters of Lanthanum phosphate/Yttria composite by response surface methodology. Measurement, 131, 309-318.
9. Balamurugan, K., 2020. Metrological changes in surface profile, chip, and temperature on end milling of M2HSS die steel. International Journal of Machining and Machinability of Materials, 22(6), pp.443-453.
10. Bhasha, A. C., & Balamurugan, K. (2019). Fabrication and property evaluation of Al 6061+ x%(RHA+ TiC) hybrid metal matrix composite. SN Applied Sciences, 1(9), 1-9.
11. C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Volume 16, 2019, 100203, https://doi.org/10.1016/j.imu.2019.100203.
12. Catboost and other class.algos with 88% accuracy, https://www.kaggle.com/ kanav0183/catboost-and-other-class-algos-with-88-accuracy.
13. Classification Algorithms - Random Forest, https://www.tutorialspoint.com/ machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
14. Decision Tree Classification Algorithm, https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
15. Decision Tree Introduction with example, https://www.geeksforgeeks.org/decision-tree-introduction-example/
16. Devansh Shah, Samir Patel & Santosh Kumar Bharti, Heart Disease Prediction using Machine Learning Techniques, October, 2020, SN Computer Science volume 1, Article number: 345(2020),https://link.springer.com/article /10.1007/s42979-020-00365-y

17. Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—A Smart Web Application to Manage Network Environments. In Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 97-108). Springer, Singapore.
18. Galla Siva Sai Bindhika, Munaga Meghana, Manchuri Sathvika Reddy, Rajalakshmi, Heart Disease Prediction Using Machine Learning Techniques, April, 2020, https://www.irjet.net/archives/V7/i4/IRJET-V7I4993.pdf
19. Gowthaman, S., Balamurugan, K., Kumar, P. M., Ali, S. A., Kumar, K. M., & Gopal, N. V. R. (2018). Electrical discharge machining studies on monel-super alloy. Procedia Manufacturing, 20, 386-391.
20. Honey Pandey,S.Prabha, Smart Health Monitoring System using IOT and Machine Learning Techniques, January 02,2021, IEEE Explore.
21. Jason Brownlee, How to Setup Your Python Environment for Machine Learning with Anaconda, September, 2020, https://machinelearningmastery.com/setup-python-environment-machine-learning-deep-learning-anaconda/
22. Jason Brownlee, Naive Bayes Classifier From Scratch in Python, October, 2019, https://machinelearningmastery.com/naive-bayes-classifier-scratch-python/
23. K. Srinivas,B. Kavitha Rani, M. Vara Prasad Rao, Raj Kumar Patra, G.Madhukar, A.Mahendar, Prediction Of Heart Disease Using Hybrid Linear Regression, Volume 07, Issue 05, 2020, ISSN 2515-8260.
24. Kennedy Ngure Ngare, Heart disease prediction, March, 2019, https://www. researchgate.net/publication/331589020_Heart_Disease_Prediction_System.
25. Keshav Srivastava, Dilip Kumar Choubey, Heart Disease Prediction using Machine Learning and Data Mining, May,2020, DOI: 10.35940/ijrte.F9199.059120
26. Khaled Mohamad Almustafa, Prediction of heart disease and classifiers' sensitivity analysis, 02 July 2020, BMC Bioinformatics volume 21, Article number: 278 (2020), https://bmcbioinformatics.biomedcentral.com /articles/10.1186/s12859-020-03626-y
27. K-Nearest Neighbours, https://www.geeksforgeeks.org/k-nearest-neighbours/
28. Kusuma.S, Divya Udayan.J, Machine Learning and Deep Learning Methods in Heart Disease (HD) Research, International Journal of Pure and Applied Mathematics, ISSN: 1314-3395 (on-line version), Volume 119 No. 18 2018, 1483-1496.
29. Latchoumi, T. P., & Kannan, V. V. (2013). Synthetic Identity of Crime Detection. International Journal, 3(7), 124-129.
30. Latchoumi, T.P., Ezhilarasi, T.P. and Balamurugan, K., 2019. Bio-inspired weighed quantum particle swarm optimization and smooth support vector machine ensembles for identification of abnormalities in medical data. SN Applied Sciences, 1(10), pp.1-10.
31. Loganathan, J., Janakiraman, S. and Latchoumi, T.P., 2017. A Novel Architecture for Next Generation Cellular Network Using Opportunistic Spectrum Access Scheme. Journal of Advanced Research in Dynamical and Control Systems,(12), pp.1388-1400.
32. *Logistic Regression in Machine Learning, https://www.javatpoint.com/logistic-regression-in-machine-learning.
33. Logistic Regression in Machine Learning, https://www.javatpoint.com/logistic-regression-in-machine-learning
34. M.W. Kenyhercz, N.V. Passalacqua, Missing Data Imputation Methods and Their Performance With Biodistance Analyses, Biological Distance Analysis, 2016, https://www.sciencedirect.com/topics/immunology-and-microbiology/k-nearest-neighbor
35. Megha Kamboj, Heart Disease Prediction with Machine Learning Approaches, International Journal of Science and Research (IJSR), ISSN: 2319-7064, 2018, https://www.ijsr.net/archive/v9i7/SR20724113128.pdf
36. Mohamed Djerioui, Youcef Brik, Mohamed Ladjal,Bilal Attallah, Heart Disease prediction using MLP and LSTM models, IEEE Xplore, NOVEMBER, 2020, 10.1109/ICEE49691.2020.9249935.
37. Nagesh Singh Chauhan, Building Heart disease classifier using K-NN algorithm, https://www.kdnuggets.com/2019/07/classifying-heart-disease-using-k-nearest-neighbors. html/2
38. Naive Bayes Classifiers, https://www.geeksforgeeks.org/naive-bayes-classifiers/
39. Najat Ali, Daniel Neagu & Paul Trundle, Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets, https://link.springer.com/article/10.1007/s42452-019-1356-9.
40. Niklas Donges, A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM, September, 2010, https://builtin.com/data-science/random-forest-algorithm.

41. Onel Harrison, Machine Learning Basics with the K-Nearest Neighbors Algorithm, September, 2018, https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
42. Prashant Gupta, Decision Trees in Machine Learning, May, 2017, https:// towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052.
43. Predicting heart disease using machine learning, Python notebook using data from Heart Disease UCI, https://www. kaggle. com /faressayah/predicting-heart-disease-using-machine-learning
44. Purushottama C , Kanak Saxena C, Richa Sharma C, Efficient Heart Disease Prediction System, Procedia Computer Science 85 (2016), 962 – 969, 2016, https:// doi.org/10.1016/j.procs.2016.05.288.
45. Random Forest Algorithm, https://www.javatpoint.com/machine-learning-random-forest-algorithm.
46. Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2020). Role of gender on academic performance based on different parameters: Data from secondary school education. Data in brief, 29, 105257.
47. Ranjeeth, S., Latchoumi, T. P., Sivaram, M., Jayanthiladevi, A., & Kumar, T. S. (2019, December). Predicting Student Performance with ANNQ3H: A Case Study in Secondary Education. In 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE) (pp. 603-607). IEEE.
48. Ranjeeth, S., Latchoumi, T.P. and Victer Paul, P., 2019. Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. *Recent Advances in Computer Science and Communications. https://doi. org/10.2174/2666255813666191116150319*.
49. S.Raguvaran , R.Anandhi , A.Anbarasi, T.Megala, Heart Disease Prediction Using Hybrid Machine Learning Algorithms, Vol. 13 No. 01 (2020): Vol 13 No 1 (2020), http://sersc.org/ journals/index.php/IJGDC/article/view/26283
50. S.Raguvaran, R.Anandhi, A.Anbarasi, T.Megala, Heart Disease Prediction Using Hybrid Machine Learning Algorithms, Vol. 13 No. 01 (2020), IJGRC, Web of Science, June,2016.
51. Santhana Krishnan J., Geetha S., Prediction of Heart Disease Using Machine Learning Algorithms, 2019, ICIICT, https://ieeexplore.ieee.org/document/8741465
52. Shankar, G., Latchoumi, T. P., Chithambarathanu, M., Balayesu, N., & Shanmugapriya, C. (2020). An Efficient Survey on Energy Conservation System with Video Surveillance. Journal of Xian University of Architecture and Technology, 12(7), 100-106.
53. Subhankar Rawat, Heart Disease Prediction, Cleveland Heart Disease (UCI Repository) dataset-classification with various models, August, 2019, https:// towardsdatascience.com/heart-disease-prediction-73468d630cfc.
54. Sunil ray, 6 Easy Steps to Learn Naive Bayes Algorithm with codes in Python and R, September, 2017.
55. Support Vector Machine Algorithm, https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
56. TAVISH SRIVASTAVA, Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R), https://www. analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/
57. Tharuka Sewwandi, Predicting Cardiovascular Disease Using K Nearest Neighbors Algorithm, September,2020, https://towardsdatascience.com/predicting-cardiovascular-disease-using-k-nearest-neighbors-algorithm-614b0ecbf122
58. V. Krishnaiah ,Dr.G.Narsimha, Dr.N.Subhash Chandra, Heart Disease Prediction System Using CRISPADM and Decision Trees, Vol 5 (2013): CVR Journal of Science and Technology, 2013, http://cvr.ac.in/ojs/index.php/cvracin/article/view/297.