A Survey on Heart Disease Early Prediction Methodologies

Saiyed Faiayaz Waris\(^a\), and S. Koteeswaran\(^b\)

\(^a\) Research Scholar, 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.

**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>
<thead>
<tr>
<th>Name of the Methodology</th>
<th>Specific tool Environment over specific methodology for HDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machines</td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
</tr>
<tr>
<td>LightGBM</td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td></td>
</tr>
<tr>
<td>Hybrid frameworks, if any</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: The analysis requirement on the approaches

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>
<thead>
<tr>
<th>Factor</th>
<th>Meaning</th>
<th>To be in specific Range</th>
</tr>
</thead>
</table>

Table 2: List of attributes that have relationship with the heart
There are many methodologies are taken as study. The following are the works that demonstrate the heart disease prediction. As per the source of [1], 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 [2], 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 [3], 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 [4], 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 [5], the effective method to predict heart disease can be achieved 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 [6], the decision trees are 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 [7], 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 [8], 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 [9], 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 [10] 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 [11], 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 [12], 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 references provided in [13], the specific ascribes are considered for assessment of HD utilizing the information mining procedures, for example, Naïve Bayes and Decision Trees with great exactness. In the information provided in [14], 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 [15], all those approaches are compared over Cleveland and other country data sources with respect to minimal attributes. In the view of [16], 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 [17], 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 [18],[19], 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 [20], few environments are defined for KNN in order to predict heart disease. With respect to [21], 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 [22], 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 [23], KNN is to be demonstrated for the application intended. In the opinion of [24], the approaches such as ANN, KNN, random forests and logistic regression are applied for predicting heart disease in which latter KNN is found having more accuracy. As per the data mentioned in [25], the best accuracy is obtained without any equipment by using combination of statistical model and random forests. With respect to [26],[27], first study deals with the various ML algorithm used to predict heart condition beforehand to avoid mishaps and suggest the better treatment, and second study deals with comparing them based on accuracy. As per data provided in [28], 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 [29],[30], 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 [32],[33], the primary study deals with KNN in the python environment with best k value which in turn decrease the inaccuracy and rise 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 [35],[36], 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 [37],[38], 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 [39], 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 [40], 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 [41], 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 [42], the methodology of SVM is described in python environment and is applied over non-linear dataset. In the view of [43], [44], [45], 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 [46], [47], the theme of logistic regression is demonstrated through an example database and given how to compute accuracy also. In the sources specified in [48],[49], the accuracy and theme of the approach is described with an example. In the view of [50], 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 [51], [52], 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:

<table>
<thead>
<tr>
<th>Pseudo Procedure</th>
<th>Support_Vector_Machines(Dataset[])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim:</td>
<td>Converts lower data set into higher dimensional data set. No. of epochs are reduced with the regularization parameter λ.</td>
</tr>
<tr>
<td>Input:</td>
<td>Dataset</td>
</tr>
<tr>
<td>Output:</td>
<td>Class label for each tuple</td>
</tr>
<tr>
<td>• Classify the proper hyper plane w.r.to numerous situations.</td>
<td></td>
</tr>
<tr>
<td>• Categorize the two modules</td>
<td></td>
</tr>
<tr>
<td>• Find the right hyper plane to isolate into classes where margin is used between nearest points and hyperplane.</td>
<td></td>
</tr>
<tr>
<td>• Additional feature is added such as z=x²+y² and plot on x and z now.</td>
<td></td>
</tr>
<tr>
<td>• 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, $y_i$ are target marks to x, and b is determined.</td>
<td></td>
</tr>
</tbody>
</table>

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

![Image of SVM process]

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

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

The posteriori probability is obtained from the following formula:
Fig.2: Formulae for computing posterior probability based on Bayes Theorem

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.

Pseudo Procedure Naïve 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
  \[ P(C_i|X) > P(C_j|X) \quad \text{for} \quad 1 \leq j \leq m, \ j \neq i \]
- For maximizing $P(C_i|X)$ for class Ci, 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
  \[ P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad \text{where Xk alludes the estimation of characteristic Ak for tuple X} \]
- It predicts the class mark for X is Ci if and if
  \[ P(X|C_i) P(C_i) > P(X|C_j) P(C_j) \quad \text{for} \quad 1 \leq j \leq m, \ j \neq i \]

The accuracy of this approach is computed based on

Accuracy = No. of tuples Correctly predicted / Total number of actual tuples.

Pseudo Procedure Logistic 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 $\beta$ kinds are regression coefficients.
Some modifications are incurred over the hypothesis for further classification.

The accuracy of this approach is computed based on

\[
\text{Accuracy} = \frac{\text{No. of tuples Correctly.predicted}}{\text{Total number of actual tuples}}.
\]

### 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.

\[
\text{Entropy}(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}
\]

\[
\text{Entropy}(S) = \log_2 (\frac{9}{14}) \log_2 (\frac{9}{14}) = \log_2 (\frac{5}{14}) \log_2 (\frac{5}{14}) = 0.940
\]

- 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

\[
\text{IG}(S, W\text{ind}) = H(S) - \sum_{i=0}^{n} P(x) \times H(x)
\]

- For all the examples given, the labels are determined using above measures.

The following specifies general structure to be provided by decision tree induction:
The accuracy is determined based on number of tuples correctly classified and total number of tuples. 

\[ \text{Accuracy} = \frac{\text{No. of tuples Correctly predicted}}{\text{Total number of actual tuples}}. \]

---

**PseudoProcedure 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:

\[
\text{d}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}
\]

  - Assign the class label for each attribute based on most frequent class defined for the given samples.

---

**PseudoProcedure 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 represented in the following diagram:

![Diagram of Random Forest approach](image)

**Fig.3: Theme of Random Forest approach**

<table>
<thead>
<tr>
<th>Pseudo_Procedure LightGBM(Dataset[][])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aim:</strong> It is de-facto algorithm invented by Guolin Ke that wins competitions over kaggle site analytics and is extremely powerful.</td>
</tr>
<tr>
<td><strong>Input:</strong> Dataset</td>
</tr>
<tr>
<td><strong>Output:</strong> Class label for each tuple</td>
</tr>
<tr>
<td>• Declare feature vector and target variable</td>
</tr>
<tr>
<td>• Splitting dataset into training and test set</td>
</tr>
<tr>
<td>• 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.</td>
</tr>
<tr>
<td>• Predict the class and compute the accuracy.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pseudo_Procedure XGBoost(Dataset[][])</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Aim:</strong> 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.</td>
</tr>
<tr>
<td><strong>Input:</strong> Dataset</td>
</tr>
<tr>
<td><strong>Output:</strong> Class label for each tuple</td>
</tr>
<tr>
<td>• The target variable ( y ) to be prophesied from early model ( F_0 ).</td>
</tr>
<tr>
<td>• A new model ( h_1 ) is used to fit residuals from the preceding stage.</td>
</tr>
<tr>
<td>• New model ( F_1 ) is formed by combining ( F_0 ) and ( h_1 ), which is improved type of ( F_0 ). The mean square error from ( F_1 ) is inferior than ( F_0 ) is defined as</td>
</tr>
</tbody>
</table>
| \[
F_1(x) < -F_0(x) + h_1(x)
\]
| • New model \( F_2 \) to be formed to progress the enactment of \( F_1 \) and after residuals of \( F_1 \) is defined as |
| \[
F_2(x) < -F_1(x) + h_2(x)
\]
| • This process is repeated for \( n \) iterations until residuals are minimized and generalized as |
| \[
F_m(x) < -F_{m-1}(x) + h_m(x)
\]
| • 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

\[ F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \]

\[ \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) = \arg\min_{\gamma} \left( \sum_{i=1}^{n} (y_i - \gamma)^2 \right) \]

The boosting capacity is characterized by taking differential w.r.t. \( \gamma \) as

\[ F_{j+1}(x) = \frac{\sum y_i}{n} \]

\[ r_m = -a \left[ \frac{\partial L(y_i, F_m(x))}{\partial F_m(x)} \right]_{F_m(x)=F_m(x)} , \text{where } a \text{ is the learning rate} \]

The boosted model \( F_m(x) \) for each \( h_m(x) \) on each step using multiplicative factor \( \gamma_m \) 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

\[ \text{avg\_target} = \frac{\text{countInClass} + \text{prior}}{\text{totalCount} + 1} \]

Where \( \text{countInClass} \) denotes how many times label value became 1 for object \( s \) with current categorical value, \( \text{prior} \) is preliminary value determined by preliminary parameters, and \( \text{totalCount} \) 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>
<thead>
<tr>
<th>Name of the Methodology</th>
<th>Purpose</th>
<th>Advantages</th>
</tr>
</thead>
</table>
| 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-line 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>
<thead>
<tr>
<th>Name of the Methodology</th>
<th>Accuracy (%), X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support Vector Machines</td>
<td>88&lt;=X</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>95&lt;=X</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>92&lt;=X</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>79&lt;=X</td>
</tr>
<tr>
<td>KNN</td>
<td>88&lt;=X</td>
</tr>
<tr>
<td>Random Forest</td>
<td>97&lt;=X</td>
</tr>
<tr>
<td>LightGBM</td>
<td>92&lt;=X</td>
</tr>
<tr>
<td>XGBoost</td>
<td>85&lt;=X</td>
</tr>
<tr>
<td>CatBoost</td>
<td>88&lt;=X</td>
</tr>
</tbody>
</table>
Fig. 4: Accuracy of reviewed methodologies

The two parameters such as sensitivity 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:

7. Mohamed Djerioui, Youcef Brik, Mohamed Ladjal, Bilal Attaallah, Heart Disease prediction using MLP and LSTM models, IEEE Xplore, NOVEMBER, 2020, 10.1109/ICEE49691.2020.9249935.


49. Decision Tree Introduction with example, https://www.geeksforgeeks.org/decision-tree-introduction-example/