**Analysis of using different Machine Learning algorithms for Heart Disease identification**

Logo

Description automatically generated

**Design and analysis of algorithm: project submission**

|  |  |  |  |
| --- | --- | --- | --- |
| *Student Name* | *Raghu priya pasham* | *Student ID:* | *F2018A7PS0241U* |

1. **ABSTRACT:**

Cardiovascular disease is the number one cause of death globally, it is fatal and often goes undetected. The death count from cardiovascular disease is 17.9 million annually. Heart disease is a condition which can be defined as an obstruction to the artery due to formation of clots. The American Heart association predicts that nearly half of the adult population suffer from heart conditions which can be fatal. Nearly half the patients diagnosed with heart disease die within a short period of time. In developing or underdeveloped countries, where the number of doctors per 1000 of the population is lesser as compared to a developed country, therefore, detecting heart illness is a challenge as it requires surfeit amount of tests and medical expertise. Most hospitals keep a huge database containing the information about their patients. Early diagnosis of heart related diseases can assuage heart disease and the patient can seek treatment options which further increases their chances of recovery. Machine Learning has shown to produce effective results when given plethora of data for training in prediction analysis. Deep learning algorithms use a lot of hidden layers and vast amount of inputs to give the correct output. Machine learning has the power to detect the patterns and delate against abnormalities. A lot of factors can lead to heart disease, choosing the factors that are most important to give the correct results is one the challenges. For an efficient algorithm we need a prodigious data cleaning algorithm, classifiers, and accurate prediction model. What factors you choose for the inputs to your model affects the accuracy of your model too. As cardiovascular diseases affect in men than women especially older men, factors like these affects the prediction model. Developing a good prediction model for heart disease detection can improve the mortality of the country. In this paper we imbibe the knowledge of different Machine Learning algorithms to predict cardiovascular diseases and analyse the efficacy of these methods.

**Key Words: Machine Learning, Cardiovascular Disease, prediction model.**

1. **Problem Statement:**

It is a challenge to predict heart disease as many factors contribute towards a person’s susceptivity towards heart disease like their medical history, cholesterol levels, diet etc. In this paper we study look at different methods of data cleaning, feature selection methods and machine learning classifiers and analyse their performance.

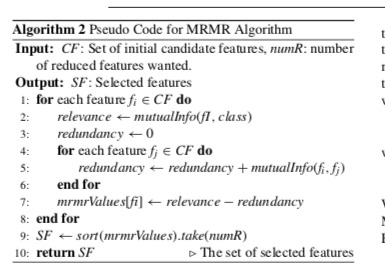
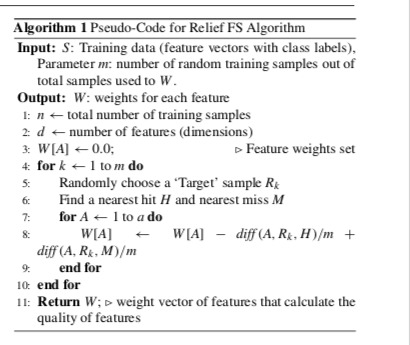
1. **Objectives:**

1. To analyse and understand the 15 different published papers and the algorithm used in them for pre-processing of data and the classifiers used.

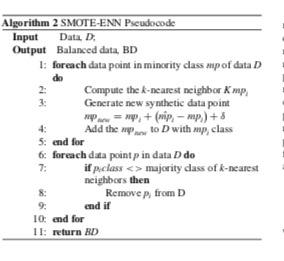
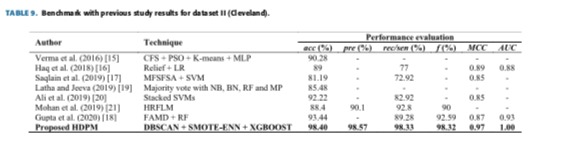
2. To discuss the disadvantages and advantages of the models in the given papers

3. To look at different feature extraction methods that accurately gives more weight to the features which has more impact on the health of the person and analyse the efficacy of the different feature selection with classical and hybrid model of classifiers.

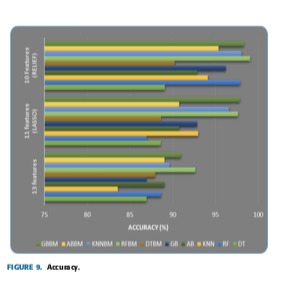
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| REFERENCE | OBJECTIVES | PROBLEM STATEMENT | METHODOLOGY | DATASET | Algorithm | Advantage | disadvantage | Performance measure value |
| 1. | 1.Heart Disease prediction using hybrid machine learning algorithm  2. To analyse the efficacy of different classification models for heart disease prediction | The clinical records of 600 patients are taken from UCI machine learning repository we need to classify each patient on the susceptivity to heart disease. | They make a hybrid model called HRF M (hybrid Random forest with Linear model), which uses Data Mining with ANN back propagation neural network. Classification modelling is done using Decision tree, Support vector Machine, Naïve bayes Neural network and K nearest neighbour and results obtained from each are compared | Cleveland dataset from UCI Machine learning repository | The code consists of 6 stages first pre-processing in which data cleaning is done and entries with missing clinical details are removed from the data sets. The 7 classification models are applied, and accuracy is computed based on | The introduced model has achieved the highest accuracy than the previous traditional method | They do not use variety of datasets and only focus on Cleveland  Datasets.  In pre-processing we omit the data with missing entries, so this algorithm only works on structured data. | HRFLM achieved 88.4% accuracy, classification error 11.6%, precision of 90.1% , f measure of 90 and sensitivity of 92.8. which is much better than the traditional naïve bayes model that has an accuracy 75.8%, classification error of 24.2%, precision 90.5, f measure of 84.5%, sensitivity of 79.8. The other classification method like generalized Linear model of 85.1, Logistic Regression 82.9, deep learning 87.4%, Decision tree 85%, Random forest 86.1%, Gradient Boosted tress of 78.3% SVM 86.1%, VOTE 87.41%. |
| 2. | 1. Heart disease diagnosis using Machine learning  2. To improve the predictive capability of machine learning algorithm | To develop a non-invasive algorithm using various machine learning classifiers which is computationally efficient in diagnosis of a heart disease given various parameter as input | For feature selection various state of the art techniques are used like relief, LASSO, MRMR and FCMIM . Various classifiers are used and their efficacy analysed. For validation of the model they have used LOSO(Leave one subject out), in this we use data of all the subjects to train the model and leave one out which is used for testing the data. For performance evaluation they have used confusion matrix | Cleveland dataset from UCI Machine learning repository. This dataset has 75 attributed | Pre-processing of dataset: In this paper they have applied minmax scalar and standard scalar . Feature selection: They have introduced 4 feature selection methods.  1.Relief: In this they have assigned weights to each feature and updates the weights automatically, the algorithm for relief works like knn  2.Minimal redundancy maximal relevance:  This algorithm removes redundant features .  Least Absolute Shrinkage Selection Operator Algorithm(Lasso) and  Fast Conditional mutual information which uses conditional mutual information for efficient feature selection to compute the value of feature relevance and redundancy .  Classifiers: They have proposed the following classifiers:  Logistic regression which classifies the output as 0 or 1  Support vector machine, Naïve Bayes which calculates the conditional probability and uses posterior probability, ANN Decision tree and KNN | The MCC of Logistic regression model using the FCMIM feature selection model is 91% when used FCMIM proposed mechanism for feature selection  The processing time of logistic regression with relief, FCMIM, LASSO, and LLBFS algorithm reduced as compared to MRMR algorithm and other classifiers  The proposed feature selection chooses features that are more effective in detecting heart disease | They have considered only Cleveland dataset with 300 data entries, they could have taken much more number of entries. | They have calculated the various classifiers performance based on the accuracy, specificity and sensitivity. Logistic regression showed 84% accuracy, 93 % specificity and 75 % sensitivity KNN showed excellent performance when k=7. ANN with 10 hidden layers gave the best performance with 60 percent accuracy 100 specificity and 0 percent sensitivity. The naïve bayes classifier gave |
| 3. | To make a clinical decision support system for prediction of heart disease  Create a hybrid model using DBSCAN to eliminate outliners, Synthetic minority over-sampling technique edited nearest neighbour(SMOTE-ENN) to balance out the training and testing datasets and XGBoost to predict heart disease  Performance analysis of the proposed model with state of the art models. | Prediction of heart disease by a hybrid model that uses DBSCAN, SMOTE-ENN and XGBoost | First collect dataset, then they used pre-processing of data to improve the accuracy of their model. Next DBSCAN is applied for the optimal parameter selection. The data is balanced based on SMOTENN method and XGBoost is used train the model and generate HDPM. They have also applied cross fitting to maximise the results | They have used two datasets the Stat log and Cleveland dataset from UCI machine learning repository. The dataset have 270 subject with 13 attributes and one output class for Stat log data set and for Cleveland the original dataset has 303 subjects and 79 attributes, they have only used 13 of them, additionally they have removed 6 subjects due to missing values | DBSCAN: The aim of using DBSCAN is to eliminate outliners, we need two parameters epsilon and minimum points the epsilon is described as the radius of the neighbourhood and the Minpts is described as the minimum number of points that needs to be in a neighbourhood. In this algorithm we go through every point in the dataset and count the number of neighbours of that point and form clusters if the point is not in any of the cluster we mark it as an outliner.  SMOTE-ENN(edited nearest neighbour), SMOTENN is a method that uses under sampling and oversampling to balance the data, it generates data in the minority class until data is balanced and then deletes the overlapping or redundant data while balancing the data simultaneously.  XGBoost(extreme gradient boosting) | Improved accuracy of heart prediction model.  Using hybrid model for balancing of data SMOTE-ENN as it performs better than either alone. | Analysis of different methods of outliner detection can be studied.  Datasets with more number of subjects can be taken which will give better prediction about the efficiency of the proposed model. | The proposed model outperformed the existing model by achieving the accuracy of 95% ,precision of 97.14%, sensitivity of 94.67%, f measure of 95.35%for dataset 1 and for dataset 2 it achieved 98.40%, 98.57%, 98..33%, 98.32% . MCC value was 0.92 and 0.97 for dataset 1 and 2 respectively. |
| 4. | To make efficient machine learning algorithm using LASSO and relief for prediction of cardiovascular diseases.  Develop new hybrid classifiers using traditional classifiers that has improved efficacy. | To classify a patients vulnerability to heart disease using Relief and LASSO, and hybrid approach Bagging and Boosting to improve the testing rate and reduce the time complexity and to compare performance of different models | First, they took 1190 subject with 14 attributed to generate ‘num’ as output where num can be a value between 0 to 4 , 0 representing healthy person , 1-4 giving various stages of illness. They have filled the missing values using KNN and standardised the data. For feature selection they have used Relief and LASSO (Least absolute shrinkage and selection operator). They have used Bagging and boosting method to solve the problem of overfitting of data. For classification purpose they have used decision tree ,random forest algorithm, KNN Adaboost and gradient boosting | They have used combined dataset of Cleveland , Long Beach, VA, Switzerland, Hungarian and Stat log | For feature selection they have used the following algorithms:  Relief: In this they have assigned weights to each feature and updates the weights automatically, the algorithm for relief works like KNN  Lasso: It removes features with negative coefficient values from the subset of features  Bagging: This algorithm is used to decrease the variance in the chosen collection of subset data. It resolves the problem of overfitting data.  Boosting: Boosting is a repetitive process that uses the last prediction to better understand the flow and it updates the weights accordingly | Using Bagging resolves the problem of overfitting data and also higher dimensionality data problem  They have used KNN algorithm to fill in the missing values | They can include more feature extraction algorithm and test the accuracy of the model | The most accurate prediction was done by AB classifier as it obtained 89.07%, KNN 83.61%, the accuracy of DT and GB are similar at 86.97%. The accuracy of RFBM is 92.65%. When used relief to select features by RFBM we get error rate of 0.95%. But when used LASSO with KNN we get error rate of 2.2%. |
| 5. | To develop a new signalling model for ECG machine to improve the accuracy of the prediction model  To introduce a fast corelation based selection model to improve the accuracy by eliminating the redundant features. | To evaluate accuracy of a hybrid model named RFRF-ILM which is a mixture of recursion enhance random forest with an improved linear model  To design an ANN with backpropagation learning that implements a feature selection | They have used decision tree which divides the datasets into smaller datasets on the basis of leaf node equation. Then they have applied several classifiers like KNN, SVM, Random forest etc. | Cleveland dataset from UCI Machine learning repository | Decision tree classier for RFRF-IFLM first import the dataset and create a matrix of the data, then define the target variable and map it to variable that you want ad output, then call decision tree classifier to classify the data | The proposed method has high sensitivity as compared to GC,DLT,EHPS ETC.  The proposed method has high specificity. | They do not use variety of datasets and only focus on Cleveland | The proposed method gives higher accuracy than other models for various sizes of datasets which are used |
| 6. | Prediction of heart disease using random search algorithm  Using optimized random forest | An intelligent system which is based on random search algorithm and optimized random forest model for detection of heart diseases | Two types of experiments are performed in which the first experiment has only about development of random forest model and the second one is the proposed random search algorithm based on random forest algorithm is developed. In the first iteration the random search algorithm generates a Boolean mask vector with only true or false values where the Boolean mask vector is anded with the feature vector where only that feature is selected whose Boolean vector is true. In the second iteration another Boolean mask vector is anded with feature vector and the same process is repeated for N-1 subset of features. Each of these produced subset features by RSA is applied to Random forest algorithm for classification | It is an online heart failure database from Cleveland. The dataset consist of 303 instances in which 297 have complete attributes with 6 missing and there are 13 features which are age, sex, chest pain type, resting blood pressure, serum cholesterol etc.… | In the algorithm RSA-RF after the first steps i.e., input NF that are full features set which are hyperparameters of RF and Fsub that are optimal subset of features. First we initialize counter variable to 1 and best\_accuracy variable to 0 then, a random mask vector of size NF-1 is generated and its total number of true values are equal to counter. Then constructing a subset of features where only those features which are present at location of Boolean mask vector. The a nested for loop is initialized from 1 to maximum values where inside those loops accuracy of each combination of random forest is evaluated. | The proposed RSA-RF system shows better performance than eleven proposed methods and the proposed learning system can help to improve quality of heart failure detection | The time complexity of the models is reduced only when the features are reduced | The proposed RSA-RF improves performance of random forest by 3.3% |
| 7. | To find an optimized XGBoost based system for prediction of heart disease  To find the efficiency of the model using 5 metrics: accuracy, sensitivity, F1-score, and area under the curve | An optimized XGBOOST for effective prediction of heart disease | Bayesian optimization is used for optimizing the hyperparameters of the XGBoost and One-Hot encoding technique to encode categorial features in the dataset. The proposed technique efficiency is compared with Random Forest (RF) and extra tree (ET) classifiers. The four common methods for optimization of hyperparameters are Manual search, grid search, random search, Bayesian optimization | It is an online heart failure database from Cleveland. The dataset consist of 303 instances in which 297 have complete attributes with 6 missing and there are 13 features which are age, sex, chest pain type, resting blood pressure, serum cholesterol etc.… | First is to preprocessing the dataset i.e., splitting the dataset into train and test by 80% and 20% where we train the model using training dataset and then Xgboost training and hyper-parameter optimization where we apply XGBoost classifier on the pre-processed data for binary classification of the target variable with the Bayesian Optimization on training data and finally we evaluate the efficiency of the proposed method using metrics as accuracy, sensitivity, specificity, F1-score, AUC. | The tree-based proposed method performs better than the other three proposed methods for accuracy. Bayesian optimization method as a hyper-parameter optimization technique is proved to be very efficient to get best hyper-parameters | This system can only be applied for similar datasets to evaluate its ability for producing similar accuracy | This model outperforms both models by 3.28% in terms of accuracy |
| 8. | To build a Naïve Bayes, decision tree, K-nearest neighbor and Random forest model for predicting heart diseases  To find the accuracy of the above models | Detection of Heart diseases using machine learning techniques | First is to pre-processing the dataset i.e., splitting the dataset into train and test by 80% and 20% where we train the model using training dataset and then cleaning the data as collected data usually has noise and missing values and then the format of data is transformed from one form to another integration must be done from various sources before processing and then reduction of data to achieve effective results | It is an online heart failure database from Cleveland. The dataset consist of 303 instances in which 297 have complete attributes with 6 missing and there are 13 features which are age, sex, chest pain type, resting blood pressure, serum cholesterol etc.… | Algorithms used are Naïve Bayes classifier, Decision tree, K-nearest neighbour, Random forest algorithm | Forest RI, Forest RC and combination of forest both are used for regression as well as classification models | Accurate and effective predictions are done only for lesser number of attributes. More complex and combinations of models can be applied to get higher accuracy | The accuracy scores for proposed model are 88.157% for Naïve Bayes, 90.789% for K-NN, 80.263% for Decision Tree, 86.84% for Random forest |
| 9. | The aim is to use the medical data to predict coronary heart disease using machine learning technology  Naïve Bayes, Support vector machine and decision tree are used to discover correlations in the coronary heart disease data to improve the prediction rate | An experimental analysis of the prediction of heart disease by machine learning techniques | To conduct the experiment WEKA, an open source data analysis software version 3.8.3 is used to conduct the experiments where the WEKA supports various ML tasks like pre-processing, classification, regression, feature selection, clustering and visualization. The methodology used for testing ML technique is 10-fold cross validation in which the entire dataset is split into 10 folds and at each iteration 9 folds are used for training the ML models and the 1-fold is used for evaluation and this process is repeated 10 times and at every iteration error rate is noted down. The average of all the errors during 10 runs is the final error | It is obtained from South African heart disease which is a subset of a larger dataset and it contains 462 instances and 10 attributes and 9 of them are independent factors and 1 variable is that is CHD is dependent variable or labelled class. The attributes are systolic blood pressure(sbp), cumulative tobacco in kg, low density lipoprotein cholesterol (ldl), family history of heart disease(famhist), type-A behavior (typea), obesity, current alcohol consumption(alcohol), and age at onset(age) | Decision Tree  Naïve Bayes Algorithm  Support Vector Machine | This research highlights few of the available techniques like SVM, DT and NB and their their performances | The research was strictly on SVM, DT and NB models. NB algorithm didn’t reach the threshold of 80%. Specificity and Sensitivity rate turned out as not the best classifiers. There were not enough instances and the CHD class was unbalanced | NB outperforms SVM and DT with respect to accuracy. The three models show 70% accuracy |
| 10. | Modernize the machine learning techniques for effective prediction of heart disease  To find a strong and effective machine learning algorithm | Statistical solution for machine learning to analyze heart disease data | In this research the three different machine learning algorithms are used that are logistic regression, ANN, SVM to predict the classification’s accuracy. The logistic model is a statistical model for binary development variables, ANN is neural network of simple elements that get inputs and this KNN is based on Keras framework for analyzing the heart disease. SVC model is used to classify the data into multiclass, K-fold cross-validation is used by setting K=10. In this paper both the datasets are emerged into a single dataset with 573 samples. The attributes are chest pain(cp), resting blood pressure(trestbps), cholesterine (chol), fasting blood sugar(fbs), resting electrocardiographic(restecg), exercise induced angina(exang), oldpeak, slope, number of major vessels(ca), thallasemia(thal), After preprocessing the data the desired data is obtained. Four tables are obtained after pretreatment and the person table contains age and gender of 271 subjects and the heart table contains heart attributes of these 271 users, symptoms table contain information such as symptomID, chestpaintype. The heartfact table consists of final diagnosis result number, personID, heartID, symptomID, and bloodID of the four tables as primary keys | It contains two different datasets of heart disease, one from the UCI website and another is Statlog Heart dataset. The first dataset consists of 270 person information and the second one consists of 76 attributes. | Logistic Regression  Artificial Neural Network  Support vector classifier | It is analysed that SVC machine learning algorithm is better compared to LR and ANN. SVC could be the best risk prediction model | The dataset is unstructured and has to go through the process of imposing structure | The implementation of these show that the SVC has highest accuracy of 92%, LR has 85% and ANN of 82% |
| 11. | Heart disease prediction using Ensemble machine learning technique. Compare the accuracy of different classification techniques. | Find accuracy of datamining algorithms. Compare and investigate the accuracy of different data mining algorithms. | They used concepts such as data portion models and data mining techniques. As there is limited data. K-Fold Cross Validation is used for data portioning. This technique is chosen as it has less variance compared to other estimators. The classifiers used are decision tree, Naïve Bayes, Multilayer Perceptron, K-nearest Neighbors, Single Conjunctive Rule Learner, Radial Basis Function, Support Vector Machine and combining these classifiers using ensemble methods such as bagging, boasting, and stacking. Performance, accuracy, precision, recall and F-measure are calculated. Receiver Operation Characteristic (ROC) curve area has been employed to compare the performance of each classifier. In first experiment- Apply the DT, NB, MLP, SVM, K-NN, RBF and SCRL classifiers dataset. while use the 10-fold cross validation technique for estimating the performance of each classifiers. In second Experiment, Bagging, Boosting, and stacking is done. | The Cleveland dataset for heart diseases. It has 303 records and 76 attributes. | Decision tree, Naïve Bayes, Multilayer Perceptron, K-nearest Neighbors, Single Conjunctive Rule Learner, Radial Basis Function, Support Vector Machine. | The efficiency of each individual classifier, and such classifiers in combination, by employing the bagging, boosting, and stacking techniques, has been evaluated. | The number of instances are limited in this dataset. | Highest accuracy is given by SVM with 84.15% and least is given by SCRL with 69.96%. After bagging applied, SVM with highest accuracy of 84.15%. and least is given by decision tree with 78.54%. After boasting applied, SVM gives highest accuracy with 84.81%. |
| 12. | Heart disease prediction using classification techniques.  To review all different classifiers and find accuracy with different reduction methods. | Experiment to find accuracy using different classifiers and different number of parameters taken. | Different classifiers used for predicting Heart disease. The classifiers used on the dataset are k-nearest neighbour, Decision tree, Naïve Bayes, and Bagging. Multilayer perceptron (NN) is used as baseline and SMOTE is used for balancing the class distribution for finding accuracy. In case 1, different classifiers are applied on dataset to find accuracy. In case 2, by reducing number of variables accuracy is found. Reduction of variables is done using specificity like specific to type of heart disease. In case 3, Accuracy is calculated using Re-sample option in Weka tool by taking 7 attributes. In case 4, accuracy is found using all 14 attributes applied on Re-sample filter. In case 5, SMOTE option in Weka is used for calculating accuracy. | Heart disease Dataset is taken from UCI Machine Learning Repository. The dataset consists of total 303 records and 14 attributes. | KDD (knowledge discovery database) data mining process is used for finding knowledge of data stored in database. KDD process consists of Data selection, Data pre-processing, Data transformation and then Data mining techniques. Algorithms used are for KNN, Decision tree, Naïve Bayes, and bagging. | Pre-processing of the dataset is done which helps in giving better accuracy. | There are many other effective hybrid machine learning algorithms which can give more accuracy. | The highest accuracy of 78.20% is given by Re-sampling Weka option with KNN classifier. |
| 13. | Prediction of heart disease using Machine learning. To find Performance using decision tree classifier. | Evaluate different classification techniques in heart diseases. | Heart attack prediction mechanism which first learns deep features and trains the learns features. When trained with all the attributes one of the classifiers gives better performance than other. The classifiers used on the data set are decision tree. Performance of these classifiers are found. | Cleveland heart diseases dataset from UCI machine learning repository. | Upload dataset. Symptoms are given as input attributes. Heart disease is set of output attributes. Decision tree is retuned after training data on function iterative dichotomise. 1. Creating a root node for the tree. 2. If(all inputs positive, return leaf node positive) if else(if all input are negative, return leaf node as negative) else(Some inputs are positive and some inputs are negative, check condition, then return result) 3. Calculate the entropy of current state H(S). 4. For each of the attribute, calculate the entropy with respect to the attribute ‘x’ denotes by H(S,X). Select the attribute which gives max value IG(S,X). 6. Remove the attributes which gives highest value from the set of attributes. 7. Repeat until no attributes are left or decision tree has all leaf nodes. Output: retrieving dataset value. | Decision trees ease interpretations, boasts stability and accuracy. | Prediction using low population and high dimensional dataset is challenging because of insufficient samples to get accurate mapping among features. | Results achieved using Precision is 78.70, F-measure is 74.31, Accuracy is 87.26. |
| 14. | Prediction of heart disease using machine learning techniques. Finding which classifier gives better accuracy to detect the heart-disease. | Predicting accuracy for future problems that heart disease may cause and which algorithm gives better results. | Dataset taken is pre-processed. 4- machine learning algorithms SVM, Decision tree, Random forest, and k-nearest neighbor are compared to get better accuracy. Confusion matrix decides accuracy perfectly. | Cardiovascular Disease Dataset, 2019 taken from Kaggle website. Dataset consists of 70000 records of patients and 11 attributes. | Dataset is pre-processed. Algorithms used in this paper are, 1. K-nearest neighbor algorithm. 2. Random Forest Classifier- It gives higher accuracy and training should be less. At first, model is built by splitting variables to training and test set. After splitting, training of dependent variables and predicting the response is done. 3. Decision tree classifier- initially data is split into training and test set. Normalizing the values is done before prediction. Training of both dependent and independent variables is done. Prediction of accuracy is done using Gini-index criterion for the test data. 4. Support Vector Machine (SVM). | SVM is effective in high dimensional spaces where number of dimensions are greater than number of attributes. | KNN when applied on large dataset takes long time to process. | Accuracy achieved by KNN is 63.4%, Random forest classifier is 71.4%, decision tree gives 68.4% and SVM gives 72.5% accuracy using linear SVM and 86.2% using Gaussian SVM kernel. Highest accuracy is given SVM. |
| 15. | Prediction of heart disease using Neural Networks. Analyse performance of Machine learning technique. | Optimize network and to get best performance using ANN. | While predicting heart disease, the input variables are disease risk factors obtained from dataset and output variables are like disease absence and disease presence. Prediction of heart disease is supervised learning problem. Backpropagation Algorithm is one of the most used ANN (Artificial Neural Network) learning technique, which is used for heart disease prediction. Backpropagation Algorithm is classification technique as it uses nonlinear relationships. Performance is found with and without PCA (Principal Component Analysis) | Cleveland Database is used for prediction of heart disease. The dataset consists of 303 records and 76 attributes. | Backpropagation algorithm is used. The prediction system is designed as a multilayer perception (NN). The ANN which is designed has three layers: input layer, hidden layer, and output layer. Input layer has 13 neurons which is equal to number of attributes in dataset. Hidden layer is designed to have 3 neurons. Number of neurons in hidden layer are mean of number of neurons in input and output layers. Output is designed to have 2 neurons (like “disease absence” or “disease presence”). | It needs less formal statistical training. It can implicitly detect complex nonlinear relationships between independent and dependent variables. | Only backpropagation algorithm is used for prediction of heart disease. The method that has been proposed can be improved for more accurate diagnosis. | Accuracy using ANN is nearly 95%. |
|  |  |  |  |  |  |  |  |  |

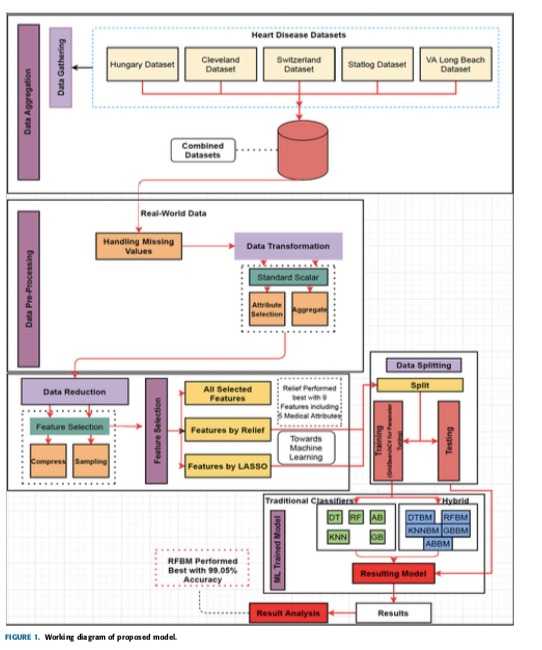
In paper 2 the algorithms used are

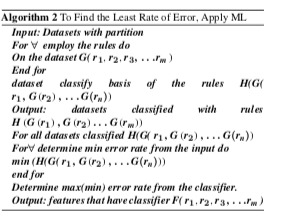
In paper 3, the algorithms used are

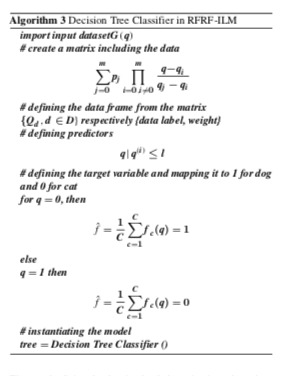


For paper 4 & 5,

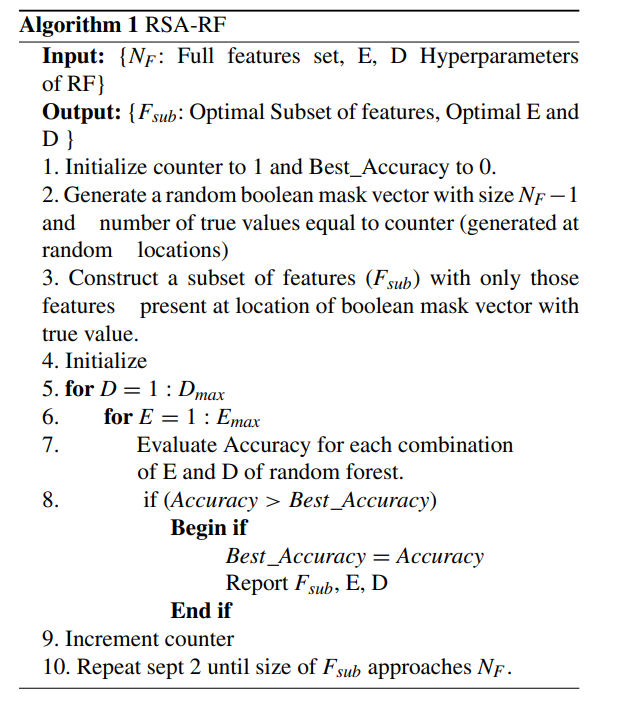








In paper 6,



**Detailed analysis of problem and algorithm used:**

In this problem statement we are applying various machine learning algorithms for detection of heart disease The following algoritms are applied in this project:

1. Logistic Regression (Scikit-learn)
2. Naive Bayes (Scikit-learn)
3. Support Vector Machine (Linear) (Scikit-learn)
4. K-Nearest Neighbours (Scikit-learn)
5. Decision Tree (Scikit-learn)
6. Random Forest (Scikit-learn)
7. XGBoost (Scikit-learn)
8. Artificial Neural Network with 1 Hidden layer (Keras)

Accuracy achieved: 95% (Random Forest)

Dataset used:  <https://www.kaggle.com/ronitf/heart-disease-uci>

In the data set we used the following columns are there :  
1. age: age sex: 1: male, 0: female

2. cp: chest pain type, 1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic

3.trestbps: resting blood pressure

4.chol: serum cholesterol in mg/dl

5.fbs: fasting blood sugar > 120 mg/dl

6.restecg: resting electrocardiographic results (values 0,1,2)

7. thalach: maximum heart rate achieved

8.exang: exercise induced angina

9.oldpeak: old peak = ST depression induced by exercise relative to rest

10.slope: the slope of the peak exercise ST segment

11.ca: number of major vessels (0-3) coloured by fluoroscopy

12.thal: thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

The dataset we have also have target value as 0 and 1, as it’s a classification problem.

In our algorithm we first find the correlation between the various data columns given in the dataset and analysed the impact of each variable on the target

**Training and testing split ratio:**

We have split our dataset in the ration of 80:20, 80 percent of dataset has been used for training and the 20 percent has been used for testing.

We have then applied the following algorithms:

1. **Naïve Bayes**: Naive Bayes is a supervised machine Learning algorithm. This classifier is based on conditional probability as given by Bayes Theorem based on the assumption of independence among predictors. By Bayes Theorem calculates we can calculate posterior probability:

=

X= data of unknown class

C= Hypothesis that data is of specific class

Here p(x|c) is the likelihood

P(c) prior probability

P(x)marginal likelihood

**Advantages of naïve bayes:**

* Very simple and easy to implement
* Needs less training data so great for datasets with lesser size
* Handles both continuous and discrete data
* Highly scalable with number of predictors and data points
* It can be used for real time data analysis
* Not sensitive to irrelevant features

**Implementation of naïve Bayes:**

In our implementation of the model we have imported naïve bayes from sklearn and implemented it using gaussianNB() function. We Achieved an accuracy score of 85.25%

Disadvantages of Naïve bayes:

* + It assumes all predictors are independent which doesn’t happen in real life.
  + It assigns zero probability to categories whose data wasn’t available in training data set

1. **K-Nearest Neighbours:** This is a supervised machine learning algorithm based on feature similarity, it is used to classify the data, in this method we start with data sets in known category. If we take the value of K too low we may not get accurate answer and if we take value of K too high we can get processing issues. Usually we take K value to be square root of n where n is the number of data points in our sample. It is preferable to take odd number to get accurate results To use KNN. we need small dataset and noise free dataset. We initialise a value of K and for each sample in the training we do the following steps:
2. Find the Euclidean distance between the query and the all the data point
3. pick the first k entries and classify the data based on the most common classification of these entries

Chart, line chart

Description automatically generated

Formula to find Euclidean distance

Chart, scatter chart

Description automatically generated

Take the distance of the current point from all the known points

**Implementation of KNN in our model:**

In our KNN classifier the number of neighbours is set to 7 , the accuracy we got is 67.21 %

**Artificial Neural Network with Hidden Layers:**

Neural networks is used for problems for which its algorithm is expensive or does not exist. Artificial neural network is an information processing system that has certain performance characteristics in common with biological neural network.

It is based on certain assumption that:

1. Information processing occurs at many simple elements called neurons.
2. Signals are passed between neurons over connection links
3. Each connection has an associated weight which in a typical neural network multiplies the signal transmitted
4. Each neuron applies an activation function(usually non-linear) to its net input(sum of weighted input signal) to determine its output signal.

Neural network is characterised by:

1. Its architecture (Number of hidden layers) and it gives information about its connections
2. Its activation function.

A picture containing diagram

Description automatically generated

**Activation function:**

The 3 main activation function that we use are:

Step:x={

Sigmoid :

A picture containing text, clock, gauge

Description automatically generated

Sign={-1 if x<0

{0 if x=0

{ x=1 if x>0

**Graphical user interface, text, application, chat or text message

Description automatically generated**

**Implementation of ANN:**

We have used keras library in python to implement Artificial neural network. A sequential model is a stack of layers where each layer has on input tensor and output tensor we have added two layers in this with 154 params and 12 params respectively.

Our first layer will take in 13 dimension and give 11 dimension output and use relu as activation function, while the second layer will use sigmoid as activation function. We can then run the model. This model achieved an accuracy of 83%.

**Logistic Regression**:

Logistic Regression is a supervised learning and it comes under classification technique where we classify as 0 or 1. Logistic Regression is an algorithm for performing binary classification. The logistic regression model is a transformation of the linear regression model that helps one to model binary variables probabilistically. A generalized linear model with a logit-link is another name for it. Some basic examples are to find out if it rains or not or classification of objects, survival rate of a patient

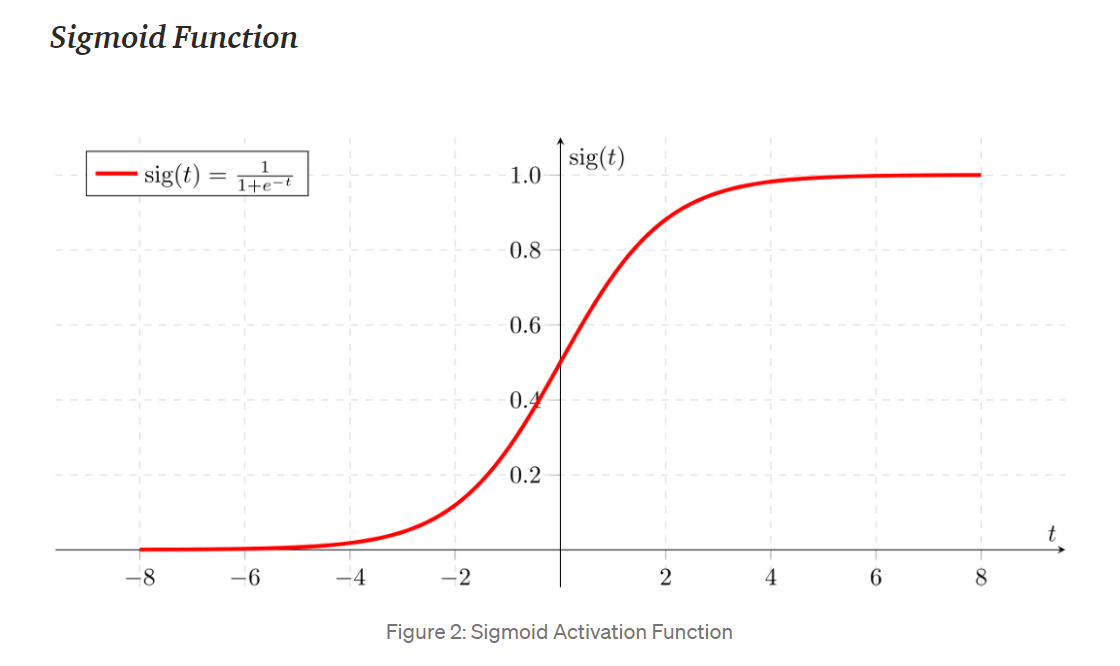
The three types of logistic Regression are 1) Binary Logistic Regression where it has two possible outcomes like spam or not 2) Multinominal Logistic Regression where the outcome would be three or more categories without ordering 3) Ordinal Logistic Regression which has three or more outcomes with ordering like rating a movie from 1 to 10

It helps in calculating the possibility of a particular event taking place. When plotted on the graph it gives an S shaped curve.

Output = 0 or 1

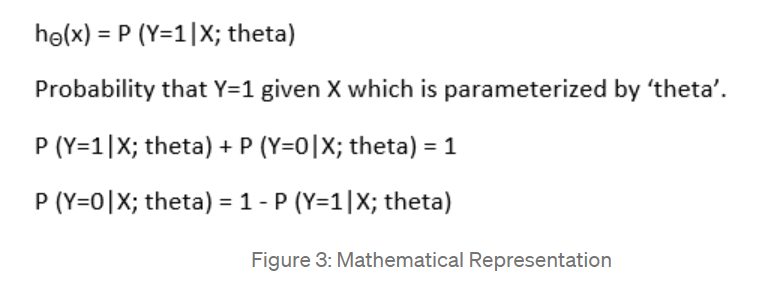
Hypothesis => Z=WX+B

h(x) = sigmoid(Z)

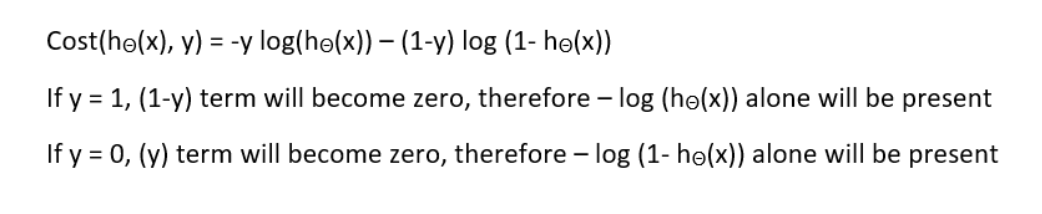


When Z goes to infinity YPredicted will be 1 and when Z goes to negative infinity YPredicted will be 0

Mathematically it can be written as,



Cost function is



**Advantages of Logistic Regression:**

* Logistic regression is more straightforward to apply, analyse, and train.
* Very quick at classifying unknown records
* It doesn't make any predictions about class distributions in function space.
* It performs well when the dataset is linearly separable and has good consistency for several basic data sets.
* It's simple to apply to several classes (multinomial regression) and a probabilistic interpretation of class predictions.

**Implementation of Logistic Regression:**

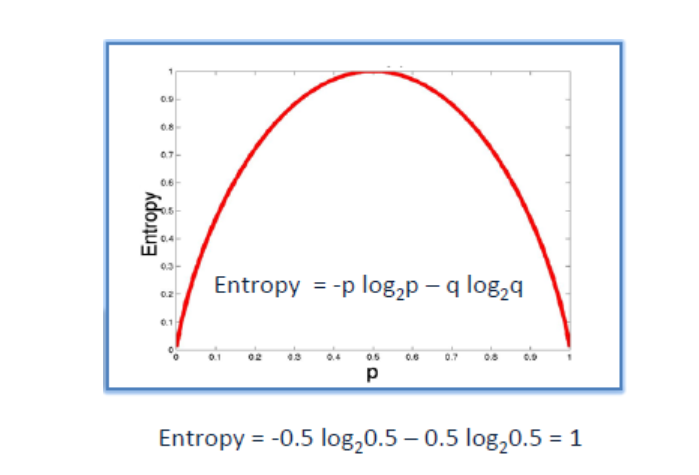
In our implementation we imported LogisticRegression from sklearn and then trained our model by fitting the X\_train and Y\_train an then predicting the Y\_pred\_lr by using the trained model and therefore an accuracy of 85.25% has been achieved

**Decision Tree:**

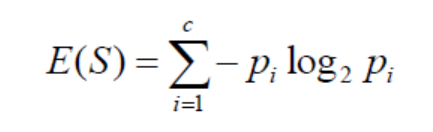
Decision tree is a tree shaped diagram that is used to determine a course of action. Each branch of the tree represents a possible decision, occurrence of reaction. It is an expansion of the decision model that helps you to see the connections between decisions and their outcomes. There are two types of decision trees 1) Categorical Variable Decision tree which has categorical target variable belongs to this type 2) Continuous Variable Decision Tree which has a continuous target variable belongs to this type. The decision to make strategic splits has a significant impact on a tree's accuracy. To determine whether to divide a node into two or more sub-nodes, decision trees employ a variety of algorithms. The existence of sub-nodes improves the homogeneity of the sub-nodes that follow.

In order to solve the attribute selection researchers devised some solutions and that include

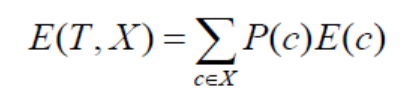
* Entropy: To measure the randomness from the information that is being processed. If the entropy is higher then it is hard to draw any conclusions



The formula for calculating entropy using frequency table of one attribute is:



The formula for calculating entropy using frequency table of two attribute is:

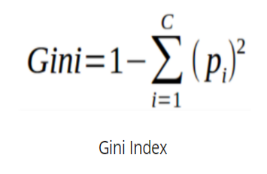


* Information Gain: Information benefit, or IG, is a mathematical property that calculates how well a given feature distinguishes training examples based on their classification objective. IG is a decrease in entropy.

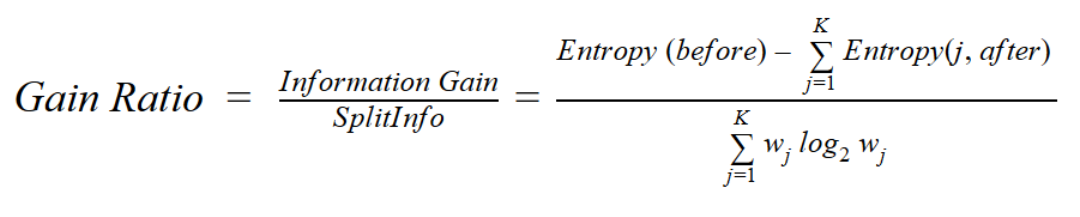
Mathematically represented as:

Information Gain(T,X) = Entropy(T) – Entropy(T,X)

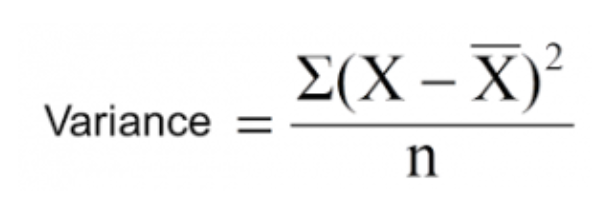
* Gini Index: It is cost function that is used to evaluate splits in the dataset



* Gain Ratio: It is calculated using Information gain and split info



* Reduction in Variance: It is an algorithm used for continuous target variables.

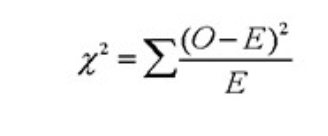


X is actual

N is number of values

X-bar is the mean of values

* Chi-square: It is represented as

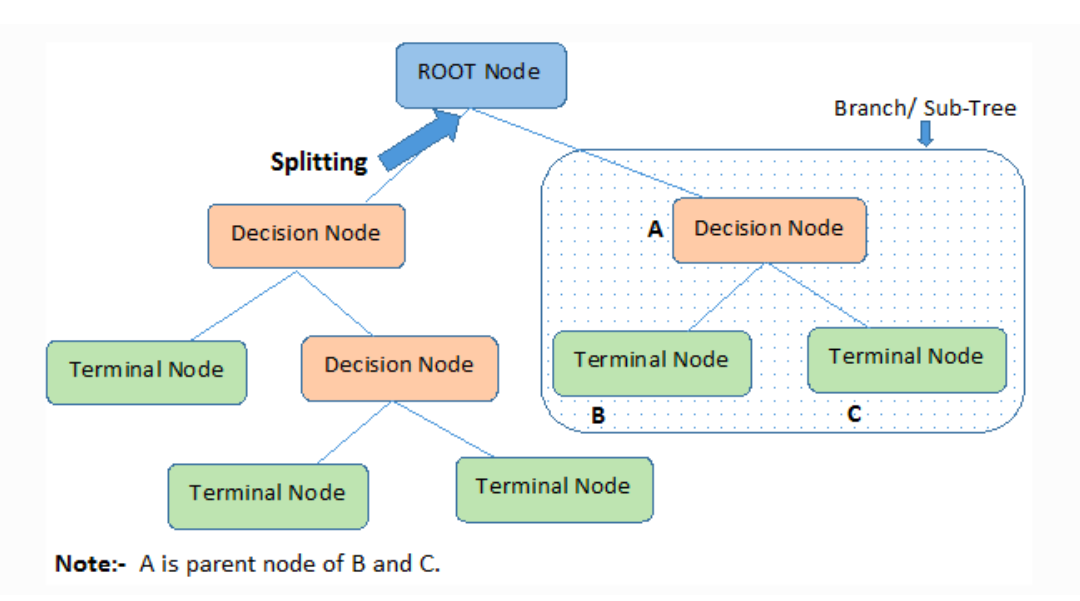


χ2 = chi square

*O* = Observed score

*E* = Expected score

A decision tree would look like the below picture



**Advantages of Decision Tree:**

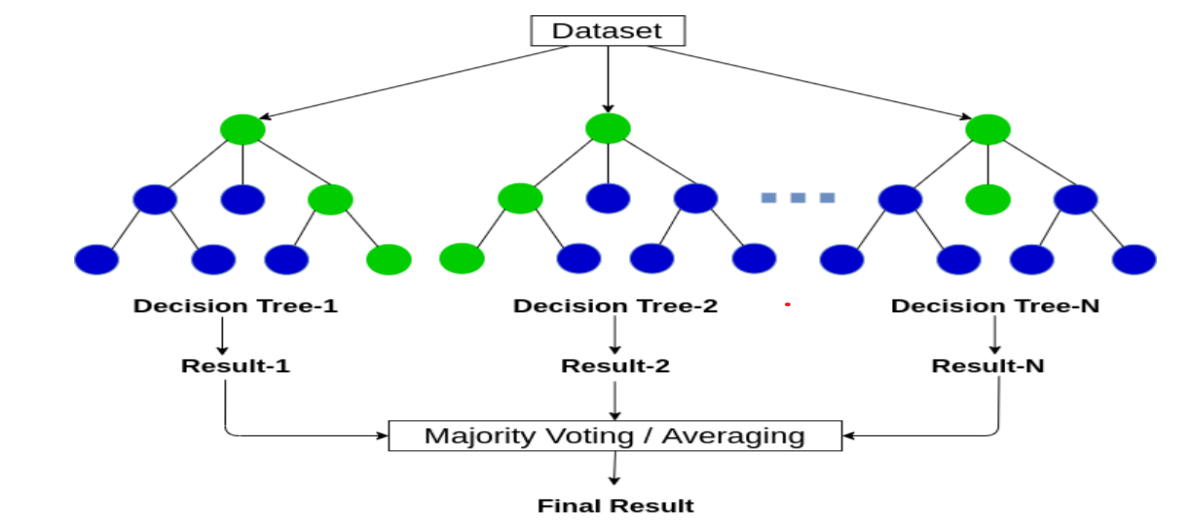
* It is simple to understand, interpret and visualize
* It handles both numerical and categorical data
* Less data cleaning is required once the variables are created
* Less effort for preparation of data

**Implementation of Decision Tree:**

In our implementation we imported DecisionTreeClassifier from sklearn and initially its max accuracy is zero then trained our model by fitting the X\_train and Y\_train an then predicting the Y\_pred\_lr by using the trained model and therefore an accuracy of 81.97% has been achieved

**Random Forest**:

Random forest or Random decision forest is a method that is operated by construction of multiple decision trees during the training phase. Random forest is a versatile, easy-to-use machine learning algorithm that, in most cases, produces excellent results even without hyper-parameter tuning. Like its name random forest consists of large number of individual decision trees which operates as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model.



**Advantages of Random Forest:**

* There is no overfitting that is use of multiple trees will reduce the risk of overfitting
* Training time is less
* Larger the data, higher the accuracy

**Implementation of Random Forest**:

In our implementation we imported RandomForestClassifier from sklearn then trained our model by fitting the X\_train and Y\_train an then predicting the Y\_pred\_lr by using the trained model and therefore an accuracy of 95.08% has been achieved

**Support Vector Machine**:

SVM is a supervised learning method, used for regression and classification problems. SVM algorithm aims to create a decision boundary or a line that separates n-dimensional space to classes. The best decision boundary is known as hyperplane. In simple words, it is representation of different categories which are separated as points in space by a gap as wide as possible. Support vector machine work is to segregate the given data in the best possible way so when the segregation is done, the distance between the nearest points is known as margin, the approach is to select a hyperplane with the maximum possible margin between the support vectors in the given data set. For selecting the maximum hyperplane in the given set, the support vector machine follows the following

1. It generates a hyperplane which segregates the classes in the best possible way.
2. It selects the right hyperplane with the maximum segregation from either nearest data point.
3. inseparable and nonlinear planes - In cases where hyperplanes are not very efficient and in those cases the support vector machine uses a kernel trick to transform the input into higher dimensional speeds, so it becomes easier to segregate the points.

Let the training samples have dataset Data = {yi, xi} where i = 1, 2, . . . , n and yi ∈ R^n represents the target item and xi ∈ R^n represents the ith vector. The linear SVM finds the optimal hyperplane of the form f (x) = wT\*x + b where w is dimensional coefficient vector and b is a offset. This is done by solving the subsequent optimization problem:

Min w,b,ξi 1/2\*w2+ C ∑i=1n ξi

s.t. yi ( wT\*xi + b) ≥ 1 − ξi, ξi ≥ 0, ∀i ∈ {1, 2, . . . ,m}

Chart, scatter chart

Description automatically generated

**Advantages of Support vector machine:**

* It is effective in high dimensional space and it is still effective in spaces well number of dimensions are greater than number of samples.
* It uses a subset of training points in the decision functions that makes it memory sufficient.
* Different functions can be specified for the decision functions which makes it versatile.

**Advantages of Support vector machine:**

* If the number of the features is much larger than the number of samples we have to avoid overfitting in choosing the kernel functions and regularization term is actually crucial.
* SVM do not directly provide probability estimates, these are calculated using 5-fold cross validation.

**Implementation of Support vector machine:**

In our implementation we imported Support Vector Machine Classifier from sklearn then trained our model by fitting the X\_train and Y\_train an then predicting the Y\_pred\_lr by using the trained model and therefore an accuracy of 81.97% has been achieved.

**XGBoost**:

**XGBoost** is an optimized gradient boosting library which is made to be highly **efficient**, **portable,** and **flexible**. It is supervised learning. XGBoost is use for implementing machine learning algorithms under the [Gradient Boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. XGBoost provides a parallel tree boosting which is used for solving data science problems faster and accurately.

**Features of XGBoost:**

* It supports parallelization by creating decision trees parallelly, there is no sequential modeling.
* Implements distributed computing methods for evaluating any large or complex models.
* It also uses out of core computing to analyze huge and varied datasets. It implements cache optimization to make best use of its hardware.

**Implementation of XGBoost:**

In our implementation we imported Support Vector Machine Classifier from sklearn then trained our model by fitting the X\_train and Y\_train and then predicting the Y\_pred\_lr by using the trained model and therefore an accuracy of 85.25% has been achieved.

**Results & Discussion**

|  |  |
| --- | --- |
| Model | Accuracy |
| Naïve Bayes | 85.25% |
| KNN | 67.21% |
| Logistic regression | 85.25% |
| Decision Tree | 81.97% |
| Random Forest | 90.16% |
| SVM | 81.97% |
| XGBoost | 85.25% |
| ANN | 83.61% |

**Conclusion**

Detecting the processing of different heart related diseases which will help humans in long run. Dataset from Kaggle with 76 attributes has been used for prediction. Machine learning techniques are used for processing raw data. For data, first correlation between the various data columns is found and analysed the impact of each variable on the target. The techniques used are Naïve Bayes, KNN, logistic regression, decision tree, random forest, SVM, XGBoost, and ANN and their respective accuracies are 85.25%, 67.21%, 85.25%,81.97%, 90.16%, 81.97%, 85.25%, 83.61%. Among this Random forest has achieved highest accuracy.

**Discussion:**

* For future the random forest algorithm can be integrated with machine to provide services to remote patients.
* A larger dataset can be used to give better understanding about the accuracy of the model.
* New factors that emerge from various studies can be included to see if it improves the efficacy of the model.

References:

1. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
2. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
3. N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
4. P. Ghosh et al., "Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques," in IEEE Access, vol. 9, pp. 19304-19326, 2021, doi: 10.1109/ACCESS.2021.3053759.
5. C. Guo, J. Zhang, Y. Liu, Y. Xie, Z. Han and J. Yu, "Recursion Enhanced Random Forest With an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform," in IEEE Access, vol. 8, pp. 59247-59256, 2020, doi: 10.1109/ACCESS.2020.2981159.
6. A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor and R. Nour, "An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection," in IEEE Access, vol. 7, pp. 180235-180243, 2019, doi: 10.1109/ACCESS.2019.2952107.
7. Kartik Budholiya, Shailendra Kumar Shrivastava, Vivek Sharma,An optimized XGBoost based diagnostic system for effective prediction of heart disease,

Journal of King Saud University - Computer and Information Sciences,

2020,ISSN 1319-1578,https://doi.org/10.1016/j.jksuci.2020.10.013.

(<https://www.sciencedirect.com/science/article/pii/S1319157820304936>)

1. Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* **1,**345 (2020). <https://doi.org/10.1007/s42979-020-00365-y>
2. Amanda H. Gonsalves, Fadi Thabtah, Rami Mustafa A. Mohammad, and Gurpreet Singh. 2019. Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies (ICDLT 2019). Association for Computing Machinery, New York, NY, USA, 51–56. DOI:https://doi.org/10.1145/3342999.3343015
3. Abdur Rasool, Ran Tao, Kaleem Kashif, Waqas Khan, Promise Agbedanu, and Neeta Choudhry. 2020. Statistic Solution for Machine Learning to Analyze Heart Disease Data. In <i>Proceedings of the 2020 12th International Conference on Machine Learning and Computing</i> (<i>ICMLC 2020</i>). Association for Computing Machinery, New York, NY, USA, 134–139. DOI:https://doi.org/10.1145/3383972.3384061
4. S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia and J. Gutierrez, "A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.
5. Nida Khateeb and Muhammad Usman. 2017. Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. In <i>Proceedings of the International Conference on Big Data and Internet of Thing</i> (<i>BDIOT2017</i>). Association for Computing Machinery, New York, NY, USA, 21–26. DOI:https://doi.org/10.1145/3175684.3175703
6. T, Chithambaram & N, Logesh & M, Gowsalya. (2020). Heart Disease Detection Using Machine Learning. 10.21203/rs.3.rs-97004/v1.
7. M. A. Khan and F. Algarni, "A Healthcare Monitoring System for the Diagnosis of Heart Disease in the IoMT Cloud Environment Using MSSO-ANFIS," in IEEE Access, vol. 8, pp. 122259-122269, 2020, doi: 10.1109/ACCESS.2020.3006424.
8. T. Karayılan and Ö. Kılıç, "Prediction of heart disease using neural network," 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 2017, pp. 719-723, doi: 10.1109/UBMK.2017.8093512.