EARLY HEART DISEASE PREDICTION USING MACHINE LEARNING

Prateek Pal (prateekpal641@gmail.com)

ABSTRACT

Heart failure is sometimes known as congestive heart failure occurs when the heart muscle does not pump blood as well as it should. When this happens, blood often backs up and fluid can build up in the lungs, causing shortness of breath. This activity relies heavily on classic methods such as ECG, occultation, and blood pressure, blood sugar, and cholesterol measurements. So, many predictive algorithms can be fitted to data like age gender cholesterol, and status of different physical and chemical states of body, and an outcome that shows the risk of heart disease in an individual can be obtained.

INTRODUCTION

In many circumstances, preserving a patient's life is dependent on the amount of time between seeing a doctor and receiving the necessary hospitalization, so providing physicians with constant updates on their patients' health conditions will significantly reduce the number of deaths.

Cardiopathy can be caused by a variety of factors, including dynamic lifestyle changes, smoking, eating habits, physical activity, obesity, diabetes, and biochemical factors such as blood pressure or glycemia, with pain in the arms and chest being the most common symptom of cardiovascular disease.

To reduce the danger of this illness, it is critical to track key heart behavior for all types of CVD and create a system that assists clinicians in making accurate and efficient judgments.

The doctor uses a range of criteria to try to distinguish the heart abnormality in a medical diagnostic.

This activity relies heavily on classic methods such as ECG, occultation, and blood pressure, blood sugar, and cholesterol measurements.

These procedures, on the other hand, are costly and time-consuming, and they may result in human errors.

Machine-learning algorithms, on the other hand, enable for a faster and more accurate identification of cardiovascular illness.

Deep learning allows the healthcare business to evaluate data quickly without sacrificing accuracy.

It is neither machine learning nor artificial intelligence; rather, it is a sophisticated hybrid of the two that sifts through data at breakneck speed using a layered algorithmic architecture.

Through a programming interface, deep learning frameworks provide building blocks for developing, training, and evaluating deep neural networks.

GPU-accelerated libraries like cuDNN, NCCL, and DALI are utilized by popular deep learning frameworks like MXNet, Py Torch, TensorFlow, and others to enable high-performance, multi-GPU accelerated training.

In numerous studies, a range of machine learning algorithms have been utilized to aid in the diagnosis of heart disorders, as well as the classification of cardiovascular diseases.

In this analysis, ten different machine learning classification algorithms are used to select the best model for the early detection of heart disease. These algorithms include Logistic Regression, Decision Tree, Nave Bayes, Random Forest, Artificial Neural Network, and others.

For selecting the vital and more correlated features that truly reflect the motif of the desired target, four feature selection algorithms were used: Fast Correlation-Based Filter Solution (FCBF), minimal redundancy maximal relevance (mRMR), Least Absolute Shrinkage and Selection Operator (LASSO), and Relief.

Anaconda IDE was used to accomplish all the processing and calculations. All the classifiers were created using Python. Pandas, NumPy, Matplotlib, Sci-kit Learn (sklearn), and Seaborn are the most used packages and libraries.

LITERATURE REVIEW

Preexisting data is being used in every field to predict and present better outcomes strategies for future use. Here we have used data of individuals which relates to their health conditions and use their data to build a model which based on given data would predict probabilities of heart diseases in near future. This would play a vital role in health industry as the heart disease are not the one which are taken lightly or are the one which are easily curable. Having a knowledge of how likely a person is to have a heart disease in future will significantly reduce the number of deaths occurring due to cardiovascular diseases.

Mamatha Alex P et.al: - we are living in a postmodern era and there are tremendous changes happening to our daily routines which make an impact on our health positively and negatively. As a result of these changes, various kinds of diseases are enormously increasing. Especially heart disease has become more common these days. The life of people is at a risk. Variation in Blood pressure, sugar, pulse rate etc. can lead to cardiovascular diseases that include narrowed or blocked blood vessels. It may cause Heart failure, Aneurysm, Peripheral artery disease, Heart attack, stroke, and even sudden cardiac arrest. Many forms of heart disease can be detected or diagnosed with different medical tests by considering the family's medical history and other factors. t the prediction of heart diseases without doing any medical tests is quite difficult. The aim of this project is to diagnose different heart diseases and to take all precautions to prevent them at an early stage at an affordable rate. We follow ‘Data mining’ technique in which attributes are fed into SVM, Random Forest, KNN, and ANN Classification Algorithms for the prediction of heart diseases. The preliminary readings and studies obtained from this technique are used to know the possibility of detecting heart diseases at an early age and can be completely cured by proper diagnosis.

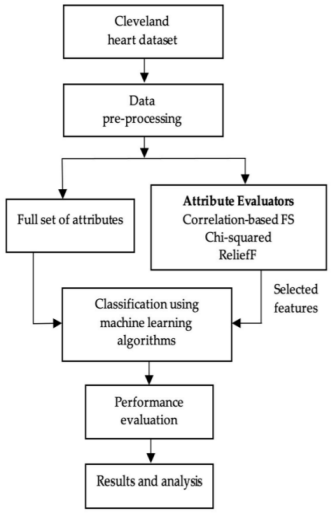
Hu Yuliang et.al: - to analyze heart valve disease accurately and effectively, a new quantized diagnosis method was proposed to analyze four clinical heart valve sounds, namely cardiac sound characteristic waveform (CSCW). BIOPAC acquiring system was used to collect signals. The recorded data is transmitted to a computer by ethernet for storage, analysis, and display in real-time. Analytical model of

single-degree-of-freedom (SDOF) was established to extract characteristic waveform. Furthermore, diagnosis parameters were calculated to discriminate heart sound of normal and heart valve disease by easy-understanding graphical representation, so that, even for an inexperienced user, he or her can monitor his or her pathology progress easily. Finally, a case study on a heart valve disease patient before and after surgery is demonstrated to validate the usefulness and efficiency of the proposed method.

M.A.Jabbar et.al :- coronary heart disease is a major cause of death worldwide. The diagnosis of heart disease is a tedious task. There is a need for an intelligent decision support system for disease prediction. Data mining techniques are often used to classify whether a patient is normal or having heart disease. Hidden Naïve Bayes is a data mining model that relaxes the traditional Naïve Bayes conditional independence assumption. Our proposed model claims that the Hidden Naïve Bayes (HNB) can be applied to heart disease classification (prediction). Our experimental results on heart disease data set show that the HNB records 100% in terms of accuracy and out performs Naïve bayes.

G Krstacic et.al: The article emphasizes clinical and prognostic significance of non-linear measures of the heart rate variability, applied on the group of patients with coronary heart disease (CHD) and age-matched healthy control group. Three different methods were applied: Hurst exponent (H), Detrended Fluctuation Analysis (DFA) and approximate entropy (ApEn). Hurst exponent of the R-R series was determined by the range rescaled analysis technique. DFA was used to quantify fractal long-range-correlation properties of heart rate variability. Approximate entropy measures the unpredictability of fluctuations in a time series. It was found that the short-term fractal scaling exponent (/spl alpha//sub 1/) is significantly lower in patients with CHD (0.93 /splplusmn/ 0.07 vs. 1.09 /splplusmn/ 0.04; $p < 0.001$). The patients with CHD had lower Hurst exponent in each program of exercise test separately, as well as approximate entropy than healthy control group ($P < 0.001$).

METHODOLOGY



Dataset Collection

We use a data set of people who have performed analyses and tests to detect heart disease. The data set is a matrix where the rows represent the patients and the columns represent the factors or attributes (features) to be tested.

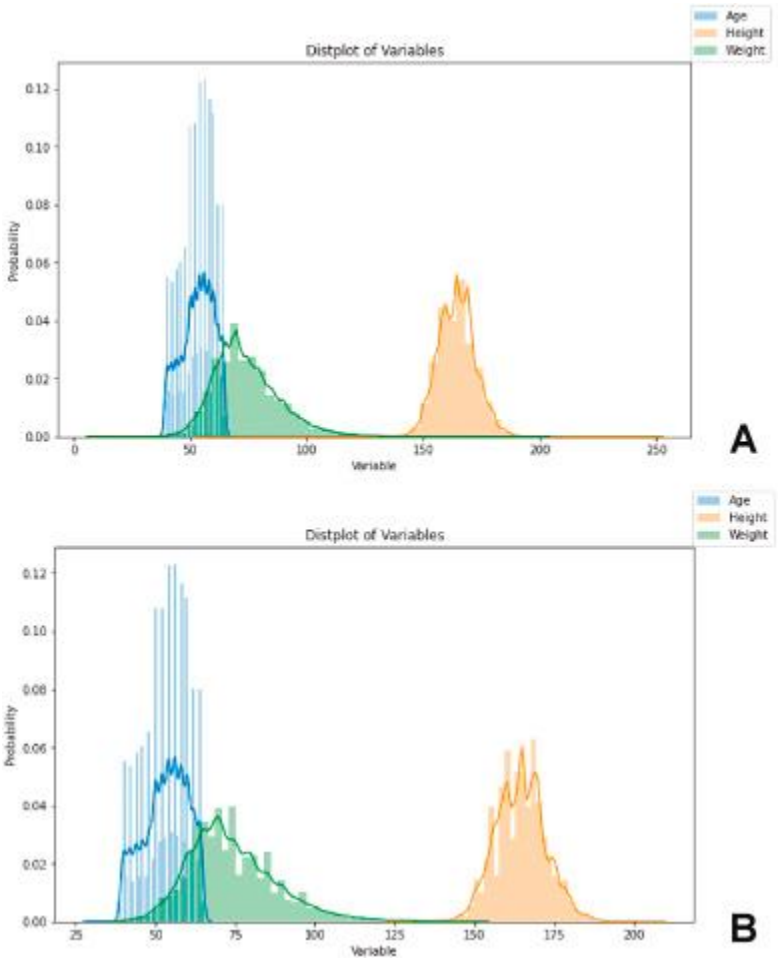
Data Sets Used -

1. https://www.kaggle.com/ronitf/heart-disease-uci
2. https://www.kaggle.com/johnsmith88/heart-disease-dataset.
3. Framingham Dataset.
4. Stat Log dataset in UCI repository.
5. Online repository of University of California, Irvine (UCI) for machine learning.

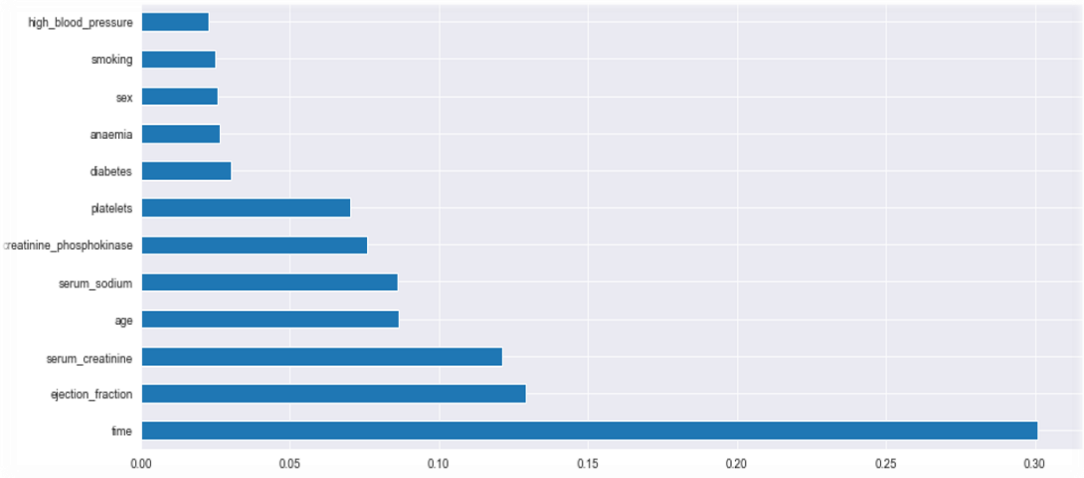
Manual Exploration

Here we try to prepare our data for the model building and analyze our data to choose our features which will give good accuracy for our model.

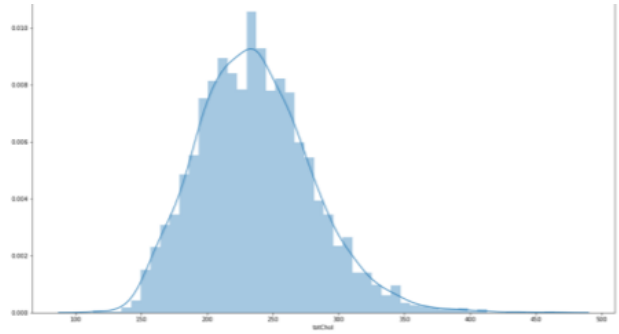
The figure below shows the distribution of three physical features which are age, height, and weight before and after preprocessing of the data.



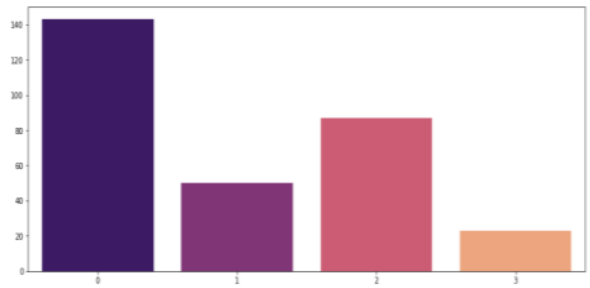
Following table has been drawn to analyze our data using extra tree classifier and obtaining the value of importance which each feature hold in predicting heart disease.



Let us try to deduce a relation between cholesterol level and the heart attack chances by plotting a distplot.

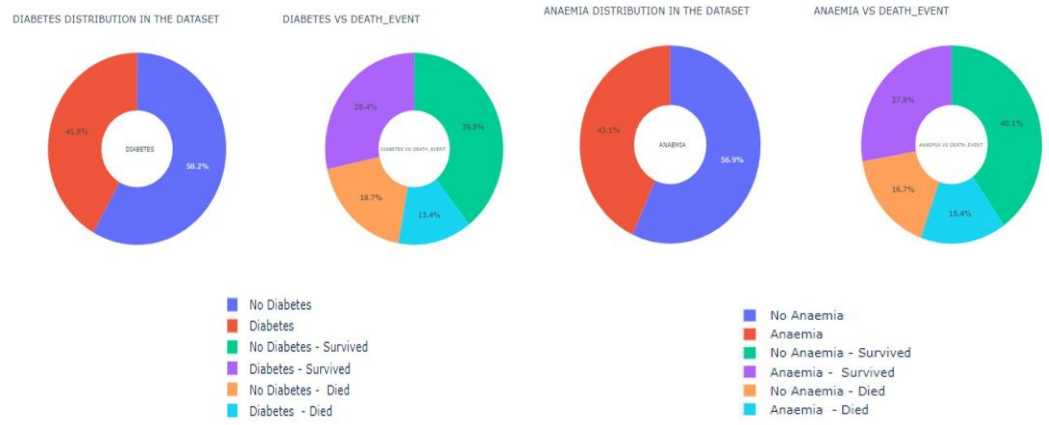


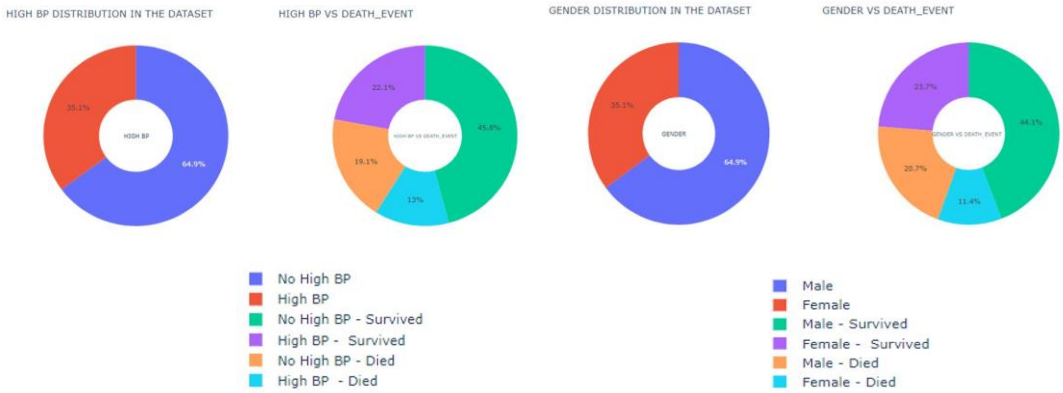
The following bar graph shows the relation between different chest pains and their proneness to a heart attack.



Here, 0, 1, 2, 3 are categorized into four types from 0 to 3 defining: 0 = typical angina, 1 = atypical angina, 2 = non-anginal pain, and 3 = asymptotic. As per this graph-plot, it is found that people having chest pain of type-0 i.e., typical angina is more prone to heart attacks with respect to others whereas people having chest pain of type-2 i.e., non-anginal pain is mild prone to heart attacks.

Below we analyze some important traits and see how they relate to deaths by plotting their pie charts.

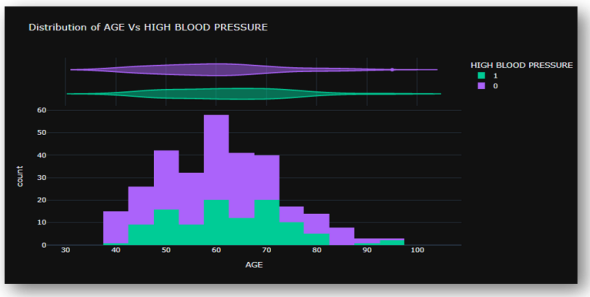


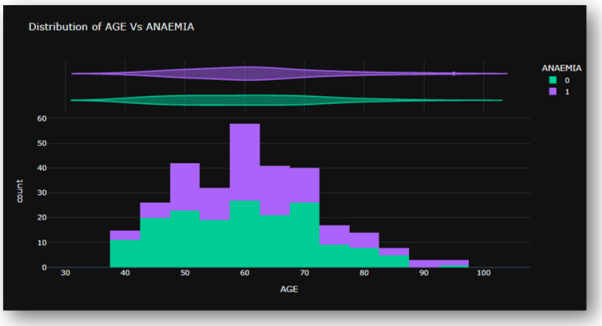


From the above inference, we analyze that the proportion of death and living is almost equal in people, so gender does not prove to be an important feature in our model

Similarly, we see for some other features too that hoe they correlates with the death column.

As there are several diseases which appear with increasing age here we see that how two of the diseases- high bp and anemia show themselves with increasing age.





High BP is a widespread problem when a person reaches the age of 70.

While the anemia does not have any correlation with age.

Many diseases occur with increasing age in our dataset there are two diseases so let's see how those diseases correlates with the data.

Data Pre-processing

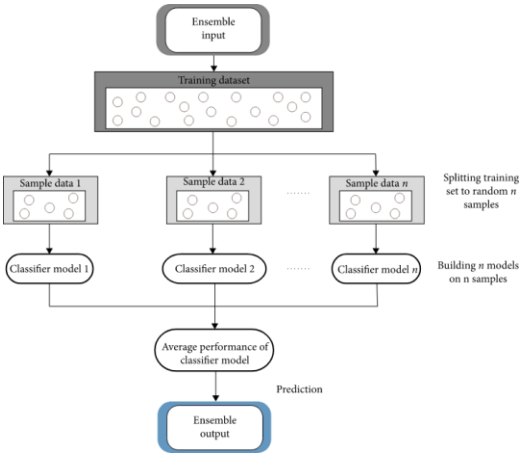
In this step we take our data fill in the missing values or remove the columns or rows which have many numbers of values missing, find outliers and fix them, normalize our data so that the algorithms would take less time to reach their optimal minima.

We also do feature selection where we choose only those features which correlates with the result and remove others.

The last step we do is to split our data for training validating and testing.

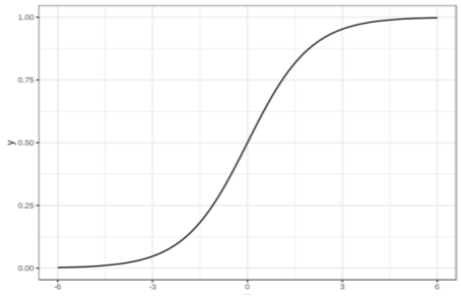
MACHINE LEARNING MODELS

Random Forest- They are used for classification as well as regression. It creates a tree for the data and makes prediction based on that. Random Forest algorithm can be used on large datasets and can produce the same result even when large sets record values are missing. The generated samples from the decision tree can be saved so that it can be used on other data. In random forest there are two stages, firstly create a random forest then make a prediction using a random forest classifier created in the first stage.



Decision Tree - This algorithm is in the form of a flowchart where the inner node represents the dataset attributes and the outer branches are the outcome. Decision Tree is chosen because they are fast, reliable, easy to interpret, and truly little data preparation is required. In Decision Tree, the prediction of class label originates from root of the tree. The value of the root attribute is compared to record's attribute. On the result of comparison, the corresponding branch is followed to that value and jump is made to the next node.

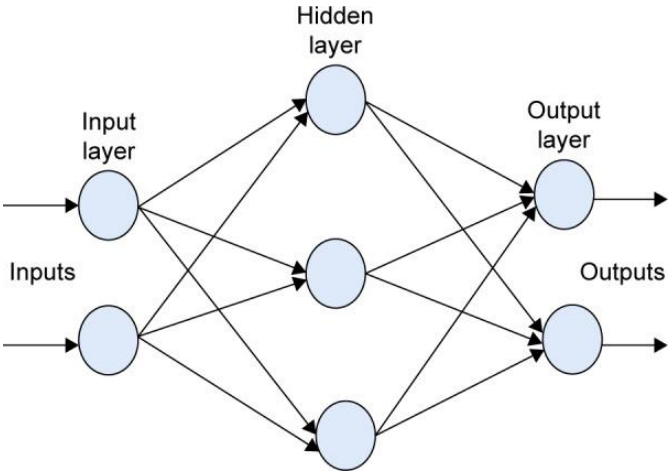
Logistic Regression - it is a classification algorithm mostly used for binary classification problems. In logistic regression instead of fitting a straight line or hyper plane, the logistic regression algorithm uses the logistic function to squeeze the output of a linear equation between 0 and 1.



Naive Bayes- This algorithm is based on the Bayes rule. The independence between the attributes of the dataset is the main assumption and the most important in making a classification. It is easy and fast to predict and holds best when the assumption of independence holds. Bayes' theorem calculates the posterior probability of an event (A) given some prior probability of event B represented by P(A/B) [10] as shown in equation $P(A|B) = (P(B|A)\ P(A)) / P(B)$.

Neural Networks - Deep neural network represents the type of machine learning when the system uses many layers of nodes to derive high-level functions from input information. It means transforming the data into a more creative and abstract component. All layer computes result for next layer and at last loss is calculated and weights are reassigned with the help of loss function this continues till a model does not reach a minimal.

$Yj = f\ (\Sigma\ wji\ xi)$



Support Vector Machine - an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to points in space to maximize the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Extra Tree Classifier - is a type of ensemble learning technique that aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is remarkably like a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Apriori Algorithm - Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. A frequent itemset can be a defined as a subset of frequent itemset i.e., if {PQ} is a frequent itemset, both {P} and {Q} should be a frequent itemset.

Frequent Pattern Mining using MAFIA - Mining frequent itemset is an active area in data mining that aims at searching interesting relationships between items in databases. It can be used to address a wide variety of problems such as discovering association rules, sequential patterns, correlations, and much more. The proposed approach utilizes an efficient algorithm called MAFIA (Maximal Frequent Itemset Algorithm) which combines diverse old and new algorithmic ideas to form a practical algorithm. The proposed algorithm is employed for the extraction of association rules from the clustered dataset besides performing efficiently when the database consists of exceptionally long itemset specifically.

ID3 Algorithm - The ID3 algorithm (Quinlan86) is a Decision tree building algorithm that determines the classification of objects by testing the values of the properties. It builds the tree in a top-down fashion, starting from a set of objects and the specification of properties. At each node of the tree, a property is tested, and the results used to partition the object at that point are set. This process is recursively

continued till the set-in a given subtree is homogeneous with respect to the classification criteria. Then it becomes a leaf node. At each node, information gain is maximized, and entropy is minimized. In simpler words, that property is tested which divides the candidate set in the most homogeneous subsets.

CART - CART stands for Classification and Regression Trees methodology. In classification trees the target variable is categorical, and the tree is used to identify the "class" within which a target variable would fall into. In regression trees, the target variable is continuous, and a tree is used to predict its value. The CART algorithm is structured as a sequence of questions, the answers to which determine what will be the next question if there should be any questions. The result of these questions looks like a tree structure where the ends are terminal nodes which represent that there are no more queries.

EVALUATING MODELS

In classification problems, we do not calculate accuracies just by calculating true predictions/total predictions as it can give a particularly good accuracy, but your model was not good. This happens in classification problems where classes are highly imbalanced so for classification problems, we calculate the functions below to check how well different models perform.

Accuracy = (TP + TN)/ (TP + FP + TN + FN) × 100%

Recall = TP/(TP+FN)

Precision = TP/(TP+FP)

F-measure = (2*precision*recall)/ (precision + recall)

TP - classes marked as positive which were positive.

TN – classes marked as negative which were negative.

FP – classes which were not positive but were marked positive.

TN – classes which were positive but were marked negative.

A confusion matrix

	A (patients with heart disease)	B (patients with no heart disease)
A (patients with heart disease)	TP	FN
B (patients with no heart disease)	FP	TN

Abbreviations: TP, true positive; FN, false negative; FP, false positive; TN, true negative.

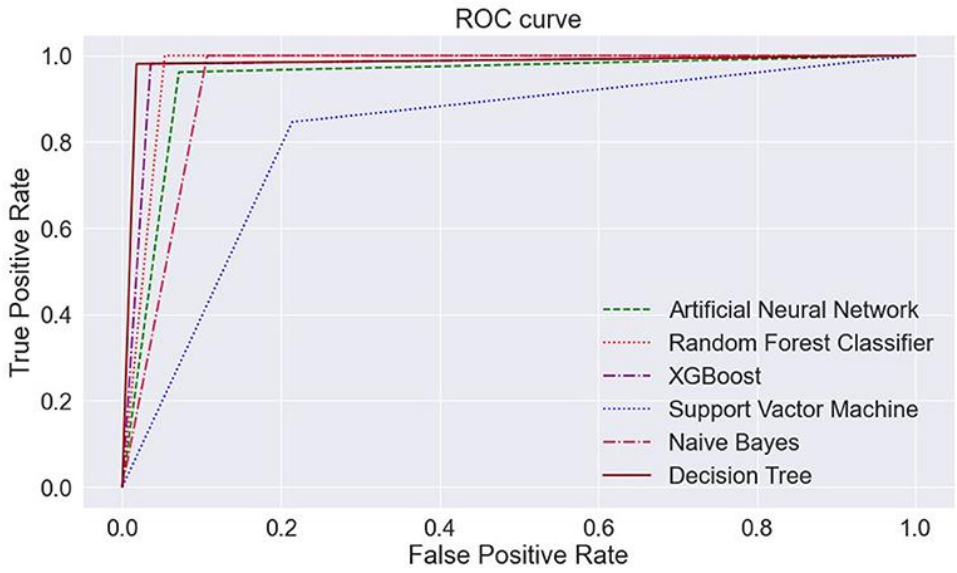
The calculated accuracies for above classifiers were obtained as following -

Algorithm	Recall	Precision	F-Score	AUC	Accuracy

KNN	99	97	98	97.9	99%
SVM	88	83	85	91	84.7%
Decision Trees	98.7	98	98	98	98.3%
Random Forest	98	97.5	98	98	97.9%
Naïve Bayes	88	81.9	85	92	83.7%
Gradient Boost	90.32	92.14	92	96.87	91.34%
Logistic Regression	89.92	81.24	85	92.28	84.08%
ANN	84.35	83.19	82	92.54	85.07%
Extra Tree	91.82	92.84	92	97.92	92.09%

CONCLUSION

Heart diseases have become increasingly frequent among people including our country (Algeria). Therefore, predicting the disease before becoming infected decreases the risk of death. This prediction is an area that is widely researched. Our paper is part of the research on the detection and prediction of heart disease. It is based on the application of Machine Learning algorithms, of which we have chosen the most used algorithms (Neural Network, Random Forest, Decision Trees, Logistic Regression, Naïve Bayes), on a real data set of people, where we had particularly satisfactory results, we arrived at 99% of accuracy with KNN. Though these algorithms show accuracies for different datasets too therefore at last we can say some models prove to be exceptionally better than the others which are KNN, Extra Tree, Gradient Boost, and Decision Trees. The strong point of our study, we tested the stability of the algorithm on varied sizes of our data set, we noticed at the end that Neural Network gives the best results. Also, we made a study on the features selection, or we used the correlation matrix to detect the dependencies between the attributes. This approach can be improved in several aspects, for example applying deep Learning algorithms, using other methods for attribute selection, and even increasing the size of the data set.



REFRENCES

1. Marimuthu Muthuvel, M Abinaya, K S Hariesh, K Madhankumar, V Pavithra: A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach- September 2018.

https://www.researchgate.net/publication/327722009_A_Review_on_Heart_Disease_Prediction_using_Machine_Learning_and_Data_Analytics_Approach

2. Xiao-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar: Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method. 19 Dec 2020.

https://www.hindawi.com/journals/complexity/2021/6663455

3. Armin Yazdani, Kasturi Dewi Varathan, Yin Kia Chiam, Asad Waqar Malik & Wan Azman Wan Ahmad: A novel approach for heart disease prediction using strength scores with significant predictors

21 June 2021

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01527-5

4. Yar Muhammad, Muhammad Tahir, Maqsood Hayat & Kil To Chong: Early and accurate detection and diagnosis of heart disease using intelligent computational model. 12 November 2020.

https://www.nature.com/articles/s41598-020-76635-9

5. Vardhan Shorewala: Early detection of coronary heart disease using ensemble techniques

11 July 2021.

https://www.sciencedirect.com/science/article/pii/S235291482100143X

6. Suraj Kumar Gupta, Aditya Shrivastava, S. P. Upadhyay, Pawan Kumar Chaurasia: A Machine Learning Approach for Heart Attack Prediction. 6th august 2021.

https://www.ijeat.org/wp-content/uploads/papers/v10i6/F30430810621.pdf

7. Sundas Naqeeb Khan, Nazri Mohd Nawi, Asim Shahzad, Arif Ullah, Muhammad Faheem Mushtaq, Jamaluddin Mir, Muhammad Aamir: Comparative Analysis for Heart Disease Prediction.

https://joiv.org/index.php/joiv/article/view/66

8. Apeksha Shah, Swati Ahirrao, Sharnil Pandya, Ketan Kotecha and Suresh Rathod: Smart Cardiac Framework for an Early Detection of Cardiac Arrest Condition and Risk. 22 October 2021.

https://www.frontiersin.org/articles/10.3389/fpubh.2021.762303/full

9. Jagdeep Singh, Amit Kamra, Harbhag Singh: Prediction of heart diseases using associative classification. 14-16 Oct. 2016.

https://ieeexplore.ieee.org/document/7993480

10. Poornima Singh, Sanjay Singh, and Gayatri S Pandi-Jain: Effective heart disease prediction system using data mining techniques 2018 Mar 15.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5863635/

11. Aditya Methaila, Prince Kansal, Himanshu Arya, Pankaj Kumar: EARLY HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES 2014.

https://airccj.org/CSCP/vol4/csit42607.pdf

12. H. Benjamin Fredrick David and S. Antony Belcy: HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES

http://ictactjournals.in/paper/IJSC_Vol_9_Iss_1_Paper_6_1817_1823.pdf

13. Animesh Hazra, Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee and Asmita Mukherjee: Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review November 7 2017.

https://www.ripublication.com/acst17/acstv10n7_13.pdf

14. Santhosh Gupta Dogiparthi, Dr. Jayanthi K, Dr. Ajith Ananthakrishna Pillai: A Comprehensive survey on Heart Disease Prediction using Machine Intelligence July 6th 2021.

https://assets.researchsquare.com/files/rs-680505/v1/7af5df3f-05eb-47a9-8f8d-0b859505321d.pdf?c=1631885784