HEART DISEASE PREDICTION USING MACHINE LEARNING

November - 2022

Submitted in partial fulfilment of the

Degree of Bachelor of Technology

in

Computer Science Engineering

SUBMITTED BY:		SUBMITTED TO:
Rohan Vats	(20103034)	Dr. Sulabh Tyagi
Sapan Sharma	(20103042)	
Anshuman Singh Jaswal	(20103057)	

DECLARATION

I/we hereby declare that this submission in my/ our own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Signature:

Date: Name:

Enrolment No:

CEERTIFICATE

This is to certify that the work titled **Heart- Disease Prediction using Machine Learning** submitted by **Rohan Vats, Sapan Sharma, and Anshuman Singh Jaswal** in partial fulfilment for the award of degree of Bachelor of Technology of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor ……………………….

Name of Supervisor ………………………..

Designation ………………………..

Date ………………………..

ACKNOWLEDGEMENT

We would like to convey my heartfelt gratitude to Dr. Sulabh Tyagi for his tremendous support and assistance in the completion of our project and for providing us with this wonderful opportunity to work on a project with the topic Ethics in AI. Your useful advice and suggestions were really helpful to us during the project’s completion. In this aspect, we are eternally grateful to you. The completion of the project would not have been possible without their help and insights. It was a great learning experience. Also, we would like to acknowledge that this project was completed entirely by us and not by someone else.

Signature of Supervisor ……………………….
Name of Supervisor ………………………..
Designation ………………………..
Date ………………………..

ABSTRACT

The health care industries collect huge amounts of data that contain some hidden information, which is useful for making effective decisions. For providing appropriate results and making effective decisions on data, some advanced data mining techniques are used. In this study, a Heart Disease Prediction System (is developed using several algorithms like KNN Algorithm, Decision Tree Algorithm, Random Forest Algorithm, Logistic Regression Algorithm, Naive Bayes, XG Boost, Support Vector Machine (SVM) for predicting the risk level of heart disease. The system uses 13 medical parameters such as age, sex, blood pressure, cholesterol, and obesity for prediction. The Heart Disease Prediction System predicts the likelihood of patients getting heart disease. It enables significant knowledge. E.g., Relationships between medical factors related to heart disease and patterns, to be established. The obtained results have illustrated that the designed diagnostic system can effectively predict the risk level of heart diseases.

INTRODUCTION

The highest mortality of both India and abroad is mainly because of heart disease. According to World Health Organization (WHO), heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths . Hence, this is vital time to check this death rate by identifying the disease correctly in the initial stage. We can use data mining technologies to discover knowledge from the datasets. The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by medical practitioners to reduce the number of adverse drug effect, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient’s future behavior on the given history is one of the important applications of data mining techniques that can be used in healthcare management.

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Healthcare data is massive [8]. It includes patient data, resource management data, and transformed data. Healthcare organizations must have the ability to analyze data. Treatment records of millions of patients can be stored, and computerized and data mining techniques may help in answering several important and critical questions related to health care. Clinical decisions are often made based on doctors’ intuition and experience rather than on the knowledgerich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

PROBLEM STATEMENT

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

OBJECTIVES

The main objective of developing this project are:

1. To develop machine learning model to predict future possibility of heart disease by implementing different Algorithms.

2. To determine significant risk factors based on medical dataset which may lead to heart disease.

3. To analyze feature selection methods and understand their working principle.

DATASET

The dataset is publicly available on the Kaggle Website at which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 304 records and 13 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

The dataset that was used for this project is a subset of a much larger dataset and has the following feature vectors:

1. age (Age of the patient in years)
2. sex (Male/Female)
3. cp chest pain type ([typical angina, atypical angina, non-anginal, asymptomatic])
4. trestbps resting blood pressure (resting blood pressure (in mm Hg on admission to the hospital))
5. chol (serum cholesterol in mg/dl)
6. fbs (if fasting blood sugar > 120 mg/dl)
7. restecg (resting electrocardiographic results)
   -- Values: [normal, stt abnormality, lv hypertrophy]
8. thalach: maximum heart rate achieved

9. exang: exercise-induced angina (True/ False)
10. 10. oldpeak: ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
12. ca: number of major vessels (0-3) colored by fluoroscopy
13. thal: [normal; fixed defect; reversible defect]
14. Target: 1 means person hi heart disease patient and 0 means person is not heart disease patient

In the dataset, there are 303 example vectors. Expert Systems have been used in the field of medical science to assist the doctors in making certain diagnoses, and this can help save lives. Coronary Heart Disease is a disease where a waxy substance builds up inside the coronary arteries, and hence this may lead to heart attack, and even death.

When diagnosed and treated, the treatment can go a long way in helping the patient. This classification task is important because the expert system, when correctly generalized, can tell the doctor which patient may have the disease, and the doctor can take a look at that case in more detail. Moreover, if the doctor makes a slip, i.e.

misdiagnoses someone, the expert system can help rectify his mistake. It results in two doctors, one of them virtual, instead of one doctor diagnosing every case which has a greater chance of accuracy and precision.

We are going to use a variety of Machine Learning algorithms, implemented in Python, to predict the presence of heart disease in a patient. This is a classification problem, with input features as a variety of parameters, and the target variable as a binary variable, predicting whether heart disease is present or not.

SUPERVISED LEARNING

Supervised learning, or classification is the machine learning task of inferring a function from a labelled data. In Supervised learning, we have a training set, and a test set. The training and test set consists of a set of examples consisting of input and output vectors, and the goal of the supervised learning algorithm is to infer a function that maps the input vector to the output vector with minimal error. In an optimal scenario, a model trained on a set of examples will classify an unseen example in a correct fashion, which requires the model to generalize from the training set in a reasonable way. In layman's terms, supervised learning can be termed as the process of concept learning, where a brain is exposed to a set of inputs and result vectors and the brain learns the concept that relates said inputs to outputs. A wide array of supervised machine learning algorithms are available to the machine learning enthusiast, for example Neural Networks, Decision Trees, Support Vector Machines, Random Forest, Naïve Bayes Classifier, Bayes Net, Majority Classifier etc., and they each have their own merits and demerits. There is no single algorithm that works for all cases, as merited by the No free lunch theorem.

Before we get started, we must know about how to pick a good machine learning algorithm for the given dataset. To intelligently pick an algorithm to use for a supervised learning task, we must consider the following factors:

1. Heterogeneity of Data:

Many algorithms like neural networks and support vector machines like their feature vectors to be homogeneous numeric and normalized. The algorithms that employ distance metrics are very sensitive to this, and hence if the data is heterogeneous, these methods should be the afterthought. Decision Trees can handle heterogeneous data very easily.

2. Redundancy of Data:

   If the data contains redundant information, i.e. contain highly correlated values, then it’s useless to use distance based methods because of numerical instability. In this case, some sort of Regularization can be employed to the data to prevent this situation.

3. Dependent Features:

   If there is some dependence between the feature vectors, then algorithms that monitor complex interactions like Neural Networks and Decision Trees fare better than other algorithms.

4. Bias-Variance Trade-off:

   A learning algorithm is biased for a particular input x if, when trained on each of these data sets, it is systematically incorrect when predicting the correct output for x, whereas a learning algorithm has high variance for a particular input x if it predicts different output values when trained on different training sets. The prediction error of a learned classifier can be related to the sum of bias and variance of the learning algorithm, and neither can be high as they will make the prediction error to be high. A key feature of machine learning algorithms is that they are able to tune the balance between bias and variance automatically, or by manual tuning using bias parameters, and using such algorithms will resolve this situation.

5. Curse of Dimensionality:

   If the problem has an input space that has a large number of dimensions, and the problem only depends on a subspace of the input space with small dimensions, the machine learning algorithm can be confused by the huge number of dimensions and hence the variance of the algorithm can be high.

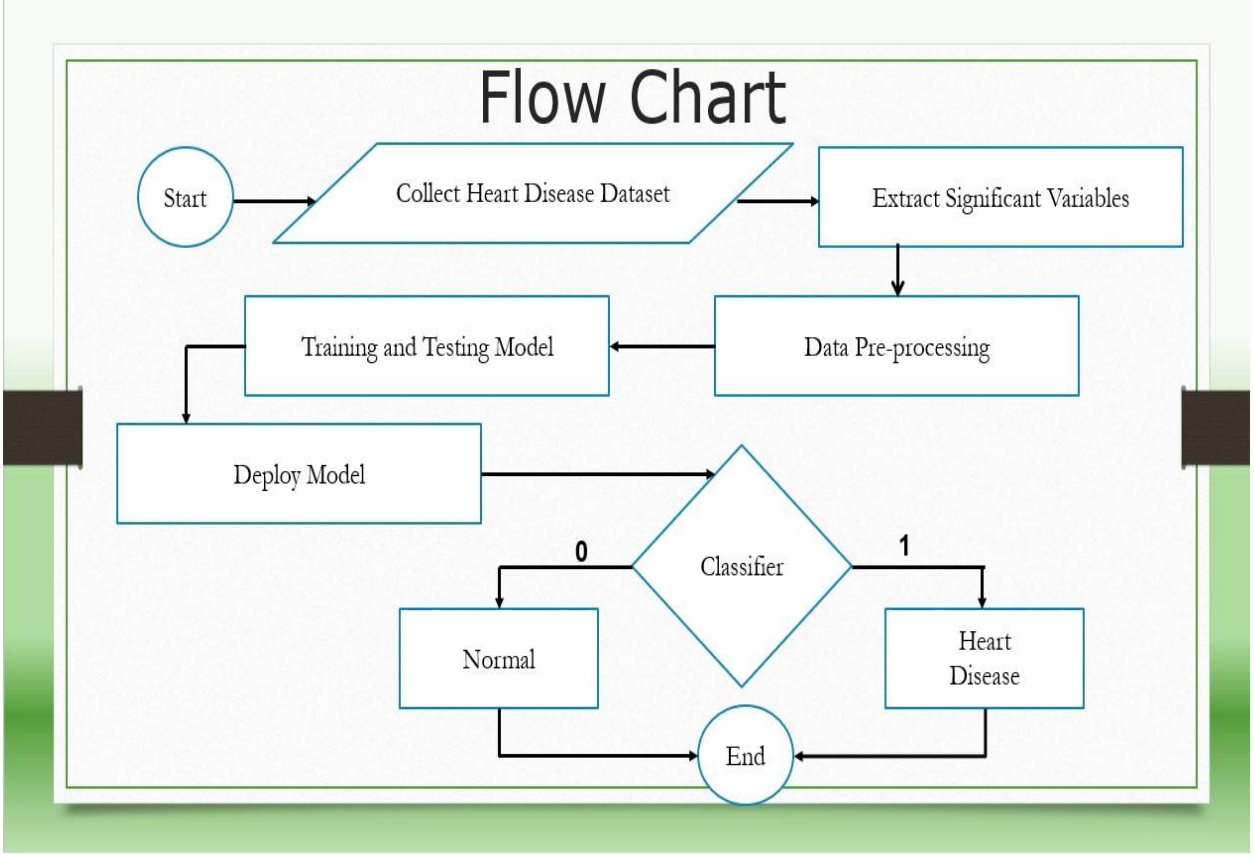
In practice, if the data scientist can manually remove irrelevant features from the input data, this is likely to improve the accuracy of the learned function. In addition, there are many algorithms for feature selection that seek to identify the relevant features and discard the irrelevant ones, for instance Principle Component Analysis for unsupervised learning. This reduces the dimensionality.

6. Overfitting:

The programmer should know that there is a possibility that the output values may constitute an inherent noise which is the result of human or sensor errors. In this case, the algorithm must not attempt to infer the function that exactly matches all the data. Being too careful in fitting the data can cause overfitting, after which the model will answer perfectly for all training examples but will have a very high error for unseen samples. A practical way of preventing this is stopping the learning process prematurely, as well as applying filters to the data in the pre-learning phase to remove noises.

Only after considering all these factors can we pick a supervised learning algorithm that works for the dataset we are working on. For example, if we were working with a dataset consisting of heterogeneous data, then decision trees would fare better than other algorithms. If the input space of the dataset we were working on had 1000 dimensions, then it’s better to first perform PCA on the data before using a supervised learning algorithm on it.

FLOW CHART



Flow Chart

Start

Collect Heart Disease Dataset

Extract Significant Variables

Data Pre-processing

Training and Testing Model

Deploy Model

Classifier

0

1

Normal

Heart Disease

End

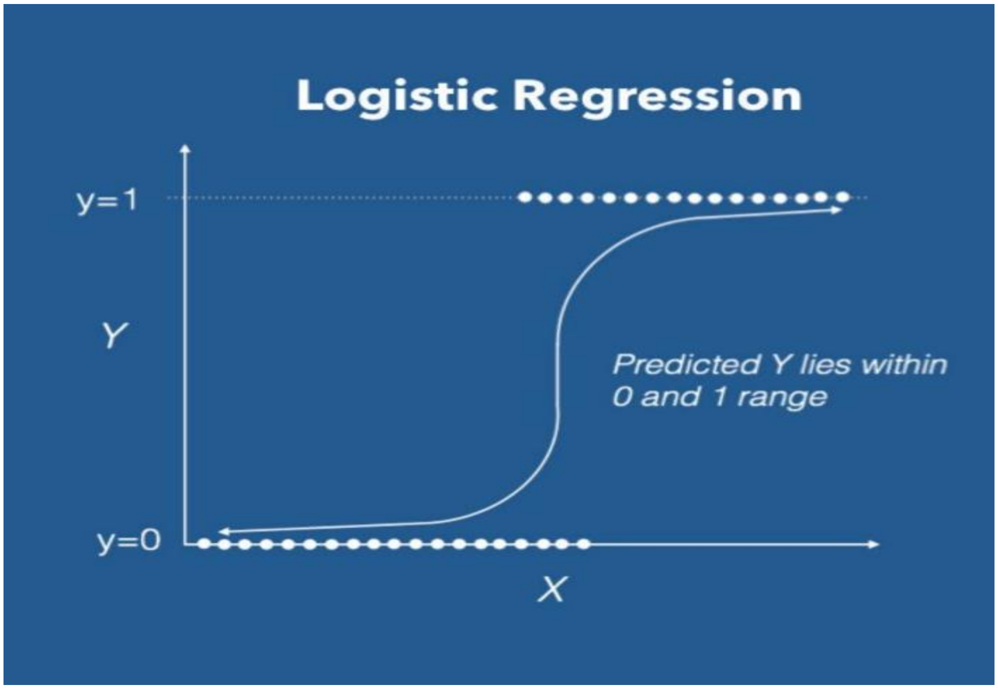
METHODS AND ALGORITHMS USED

To find out the accurate results we have applied several algorithms on our dataset and which ever gives us the best result, will be choose to predict the heart disease in the patient.

Logistic Regression –

In statistics, the logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression[1] (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling;[2] the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names.

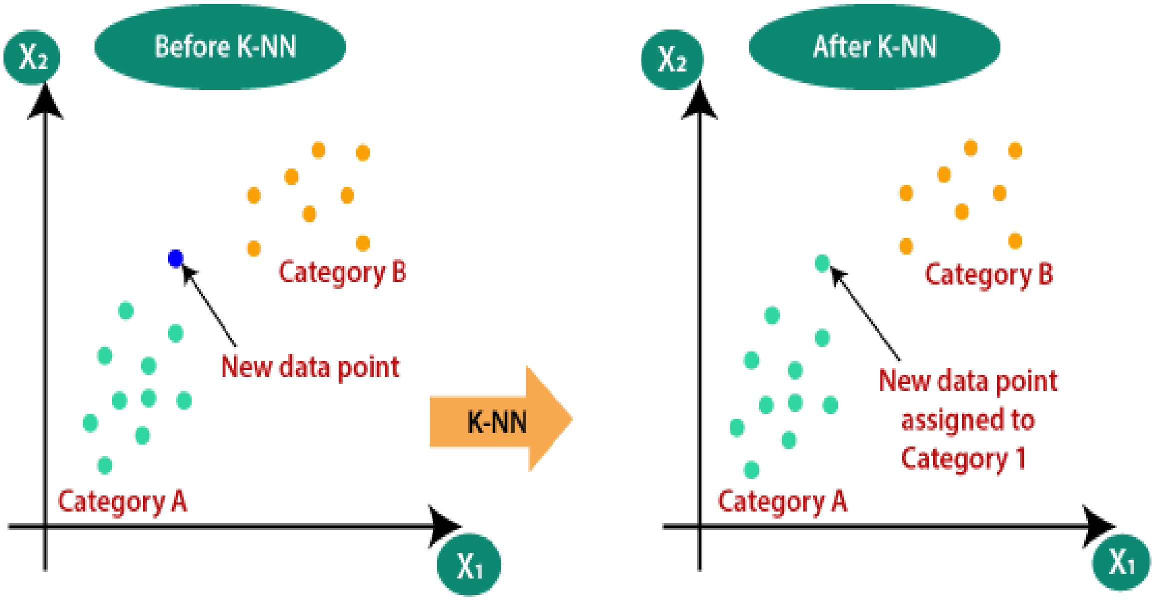
An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a sigmoid function, which takes any real input {\displaystyle t}t, and outputs a value between zero and one.[2] For the logit, this is interpreted as taking input log-odds and having output probability.



K- Nearest Neighbor (KNN) –

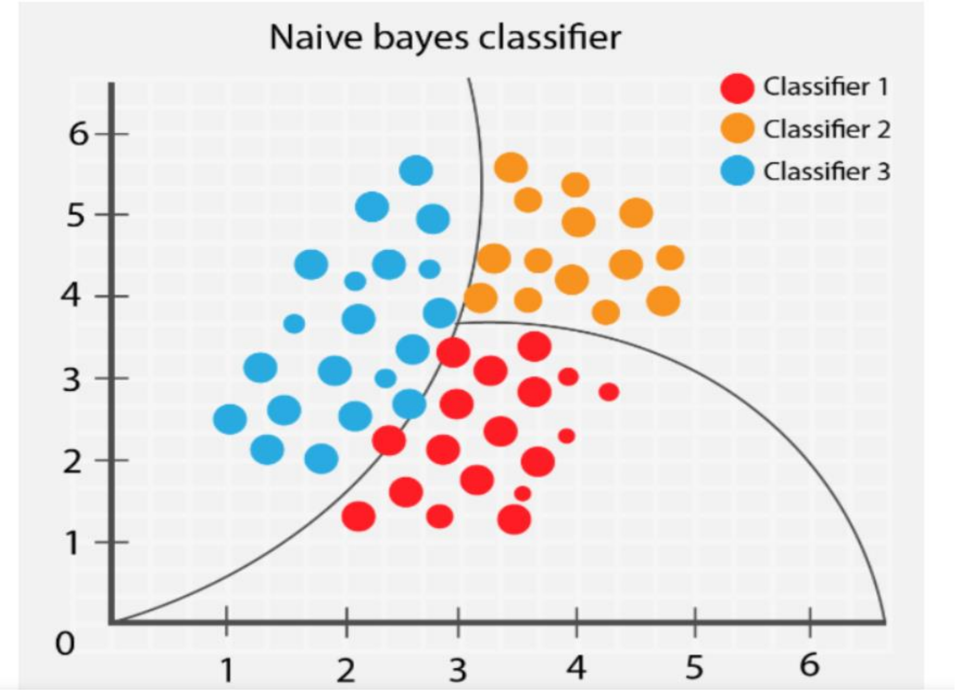
K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

It stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. It can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.



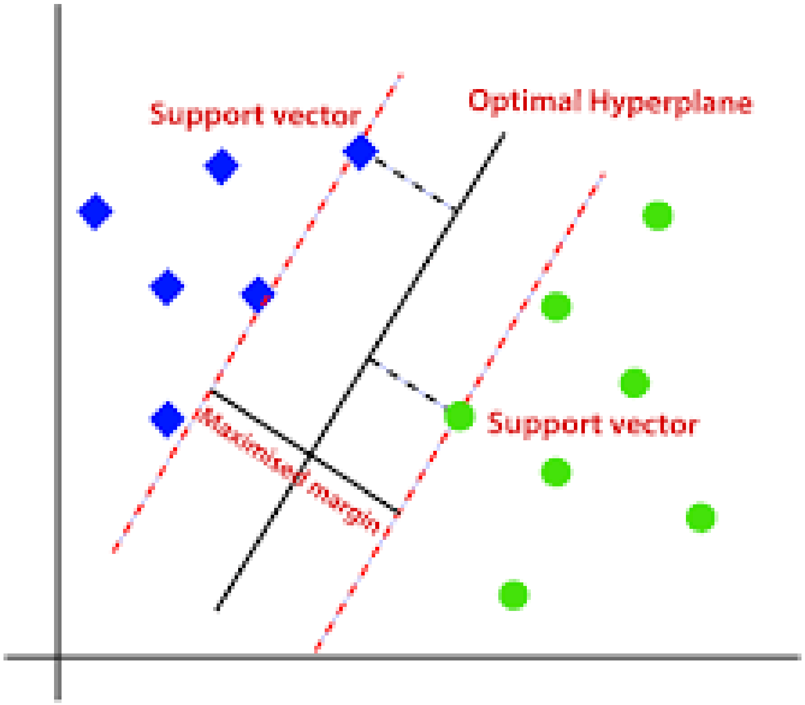
Naïve Bayes –

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done by evaluating a closed-form expression, which takes linear time, rather than by expensive iterative approximation as used for many other types of classifiers.



Support Vector Machine (SVM) –

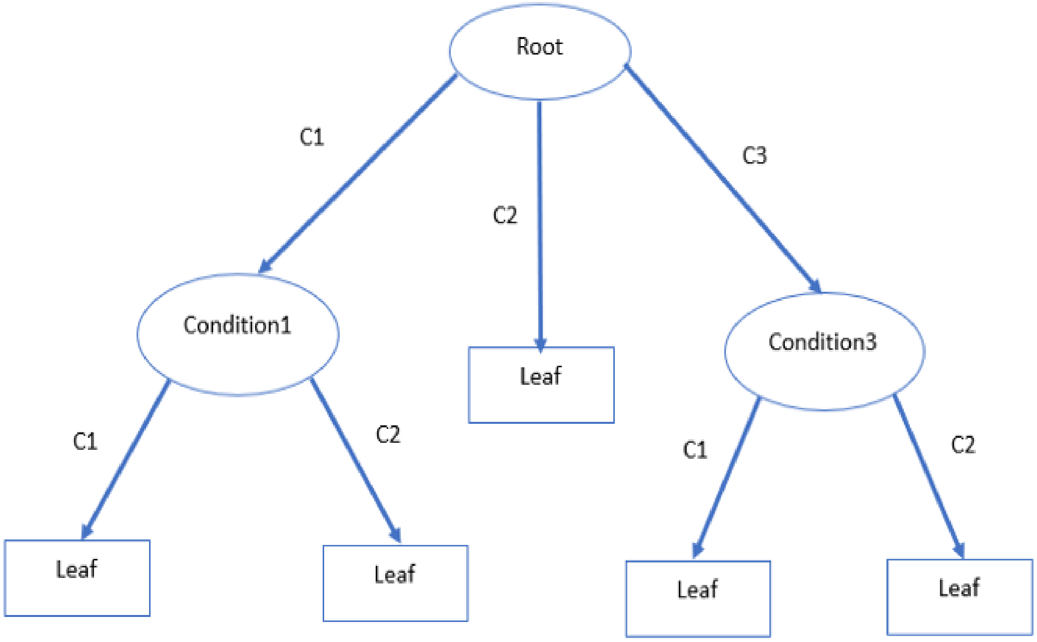
In machine learning, support vector machines (SVMs, also support vector networks[1]) are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by Vladimir Vapnik with colleagues (Boser et al., 1992, Guyon et al., 1993, Cortes and Vapnik, 1995, Vapnik et al., 1997[citation needed]) SVMs are one of the most robust prediction methods, being based on statistical learning frameworks or VC theory proposed by Vapnik (1982, 1995) and Chervonenkis (1974). Given a set of training examples, each marked as belonging to one of two categories, 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 so as to maximise 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.



Decision Tree –

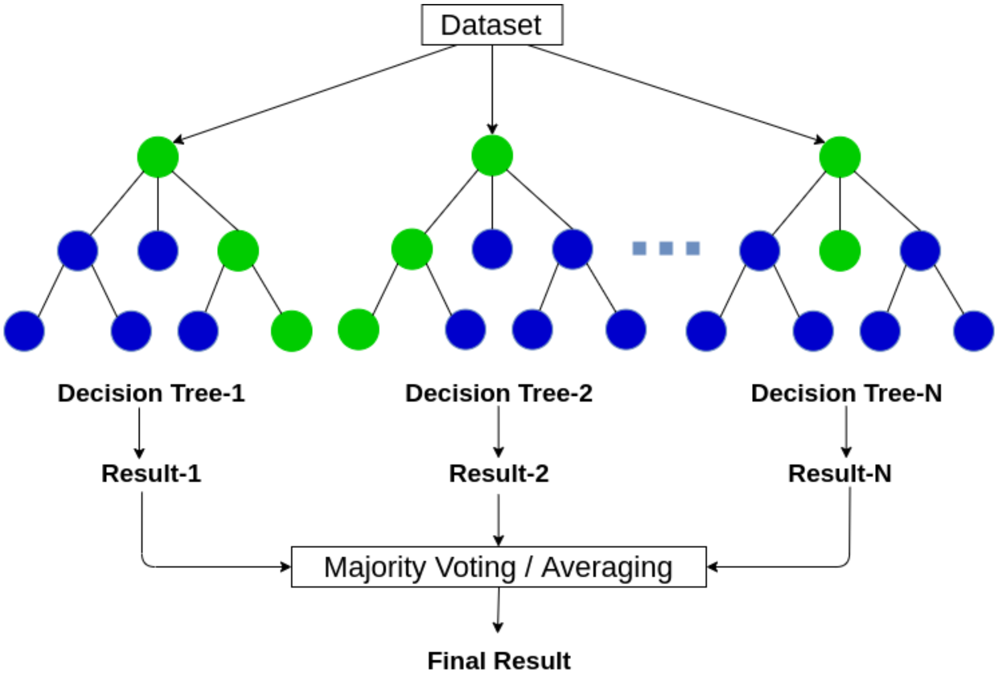
A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.



Random Forest –

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees.[citation needed] However, data characteristics can affect their performance.



After applying all the algorithms have been implemented, we will finalize the one which will give us the best accuracy out of all.

ETHICS in AI

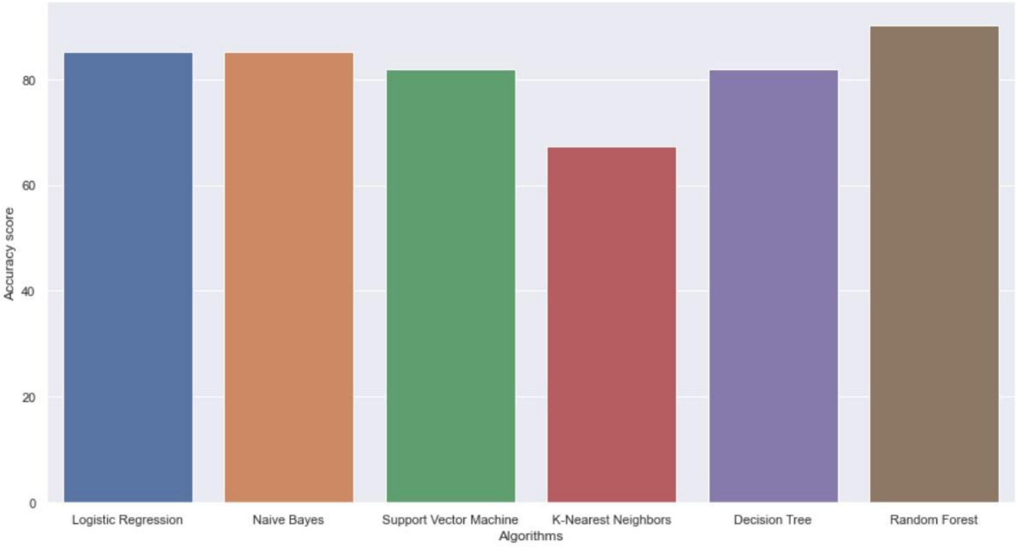
For implementing ethics in our A.I project we will deploy it on the web so that everyone can access our heart disease prediction model. Our project will not use someone else data and will maintain the privacy of the user's data.

We have ensured and took appropriate steps so that our dataset does not have any null values and any wrong values (data cleaning).

RESULTS

After applying all these algorithms, we got to know that Random Forest gives us the best accuracy.

So, we use this to predict the persons heart disease.



REFERENCES

1. "Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Datasets:Coronary Heart Disease Dataset." Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Accessed April 27, 2016.
   - http://statweb.stanford.edu/~tibs/ElemStatLearn/

2. Datasets https://www.kaggle.com/datasets/redwankarimsony/heart-disease-dat a
3. https://towardsdatascience.com/machine-learning-target-feature-label -imbalance-problem-and-solutions-98c5ae89ad0
4. https://www.researchgate.net/publication/331589020_Heart_Disease _Prediction_System
5. Research paper -

A. https://www.ijresm.com/Vol.2_2019/Vol2_Iss2_February19/IJRESM_V2_I2_89.pdf

B. heart-disease-prediction-using-machine-learning-IJERTV9IS040614202005l0-84272-1pfgh18-with-cover-page-v2.pdf (d1wqtxts1xzle7.cloudfront.net)