HEART DISEASE PREDICTION USING MACHINE LEARNING ALGORITHM

***Abstract:***

**The human heart plays a vital role in maintaining the circulation of blood throughout our bodies. Heart-related ailments have become a significant global health concern, leading to numerous fatalities. Symptoms like chest discomfort and irregular heartbeats are commonly reported. The healthcare sector has amassed extensive data, enabling researchers to explore predictive methods for heart disease using machine learning algorithms.In this paper, we propose a comprehensive approach to predict heart disease, employing a range of machine learning algorithms, Random Forest, K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, and Logistic Regression. These algorithms are applied based on various features to accurately forecast the occurrence of heart disease. Our study aims to assess and compare the predictive accuracy of these diverse machine learning techniques.By focusing on these aspects, we aim to contribute valuable insights into the effectiveness of different machine learning approaches in diagnosing heart disease, potentially leading to improved diagnostic and treatment strategies.**

# Introduction

# Heart disease poses a significant threat to our overall well-being, causing a detrimental impact on health. It claims a substantial number of lives annually. Heart disease can manifest as a result of the deterioration of the heart muscle, leading to its impaired function. Essentially, heart disease can be described as a condition where the heart struggles to efficiently pump blood throughout the body. Another term commonly used to refer to this condition is "Coronary artery disease" or "Coronary heart disease" (CAD), which occurs when there is an inadequate blood supply to the arteries.

Mostcommonindications ofheartattackare:

* Chestpain.
* Shortnessofbreath.
* SweatingandFatigue.
* Nausea,Indigestion,Heartburn,orStomachpain.
* Pressureintheupperbackpainthatspreadstoanarm.

1. **Prevalence and Significance:**
   * Heart disease is a pervasive health issue, affecting millions of people across the globe. It encompasses a variety of conditions, including coronary artery disease, heart failure, arrhythmias, and valvular heart diseases.
2. **Risk Factors:**
   * Several factors contribute to the development of heart disease. These include:
     + **Modifiable Risk Factors:** such as high blood pressure, high cholesterol, smoking, obesity, lack of physical activity, and poor diet.
     + **Non-Modifiable Risk Factors:** including age, family history, gender, and certain genetic predispositions.
3. **Pathophysiology:**
   * Heart disease often involves a process called atherosclerosis, where fatty deposits (plaque) build up in the walls of the arteries, narrowing them and reducing blood flow to the heart muscle. This can lead to angina (chest pain) or heart attacks (myocardial infarctions).
4. **Types of Heart Disease:**
   * *Coronary Artery Disease (CAD):* This is the most common form of heart disease and occurs when the blood vessels that supply the heart with oxygen and nutrients become narrowed or blocked.
   * *Heart Failure:* This condition results from the heart's inability to pump blood effectively to meet the body's needs.
   * *Arrhythmias:* These are abnormal heart rhythms, which can be too fast, too slow, or irregular.
   * *Valvular Heart Disease:* This involves problems with the heart valves, which control the flow of blood into and out of the heart's chambers.
5. **Diagnostic Methods:**
   * Various diagnostic tests are used to evaluate heart health, including electrocardiograms (ECGs), echocardiograms, stress tests, and cardiac catheterization.
6. **Treatment and Management:**
   * Treatment strategies for heart disease depend on the specific condition and its severity. They may include lifestyle modifications (e.g., diet, exercise), medications, surgical interventions (e.g., angioplasty, bypass surgery), and cardiac rehabilitation.
7. **Prevention:**
   * Prevention plays a crucial role in reducing the risk of heart disease. This involves adopting a heart-healthy lifestyle, which includes a balanced diet, regular exercise, avoiding tobacco, and managing stress.
8. **Ongoing Research and Innovation:**
   * The field of cardiology is continually advancing, with ongoing research focused on understanding the underlying mechanisms of heart disease and developing innovative treatments and interventions.
9. **Public Health Impact:**
   * Heart disease has a substantial impact on public health, leading to significant healthcare costs and societal burdens. Efforts in prevention, early detection, and effective management are critical in addressing this global health challenge.

# In light of the aforementioned considerations, this study aims to provide a comprehensive effort in predicting the likelihood of heart disease. The domain of heart disease prediction has seen extensive research using machine learning algorithms from various authors. The primary objective of this research is to attain superior predictive accuracy for determining the probability of a heart attack occurrence. Utilizing data mining techniques, specifically K-Nearest Neighbor, Decision Tree, Random Forest, Support Vector Classifier, Logistic Regression, and Naïve Bayes, this paper classifies patient risk levels. Several critical risk factors, such as age, gender, blood pressure, cholesterol levels, chest pain, heart rate, among others, are taken into account.

# This study employs the supervised machine learning paradigm to facilitate these predictions. It employs various machine learning algorithms, including K-Nearest Neighbor, Random Forest Classifier, and Logistic Regression, in conjunction with a heart disease dataset, to make accurate predictions.

# LiteratureReview

Amandeep Kaur [1] compared various algorithms such asartificial neural network, K-nearest neighbor, Naïve Bayes,Supportvectormachineon heartdiseaseprediction.

JThomas,RTheresaPrincy[2]madeuseofKnearestneighboralgorithm,neuralnetwork,naïveBayesanddecision tree for heart disease prediction. They used dataminingtechniquesto detecttheheartdiseaseriskrate.

Monika Gandhi et. Al, [3] used Naïve Bayes, Decision Treeandneuralnetworkalgorithmsandanalyzedthemedicaldataset. There are a huge number of features involved. So,there is need to reduce the number of features. This can bedonebyfeatureselection.Ondoingthis,theysaythattimeisreduced.

Ramandeep Kaur, Er. Prabhsharn Kaur [4] have showed thattheheartdiseasedatacontainsunnecessary,duplicateinformation.Thishastobepre-processed.

Sonam Nikhar et. Al, [5] has built up the paper titled asPredictionofHeartDiseaseUsingMachineLearningAlgorithmsusingdecisiontreeclassifierandnaïvebayes.

Mr. Santhana Krishnan. J and Dr. Geetha. S, [6] has writtenpaperthatpredictsheartdiseaseformalepatientusingclassificationtechniques.

# Methodology

### In this research paper, we've leveraged our dataset to implement various machine learning algorithms for the purpose of discerning whether an individual has heart disease. Our approach involves addressing missing data within the dataset, visually representing the data, and assessing the performance of distinct machine learning algorithms in terms of accuracy. The specific machine learning algorithms employed are detailed below.

### Data Collection

In this paper, the datasetis obtainedfrom the ClevelandHeartDiseasedatabaseatUCIRepository.Thereare14attributesinthe dataset.

Thedescriptionof datasetisgivenasfollows:

1. Age:describestheageofaperson.
2. Sex: describes the sex of a person; 1 for male, 0 forfemale.
3. A diagram of a data processing process

   Description automatically generatedCp: describes the chest pain type in a person ( 1 forangina, 2 for a typical angina, 3 for non-angina, 4 forasymptomatic).
4. Trestbps:describesthe restingbloodpressure.
5. Chol:describesthe serumcholesterol.
6. FBS:describes the Fasting Blood Sugar ( 1 for true & 0forfalse).
7. Restecg: describes the resting electro-graphic results( 0for normal, 1for ST-Twave abnormality, 2 for leftventricular hypertrophy).
8. Thalach:describesthemaximumheartrate.
9. Exang:describestheexerciseinducedangina
10. Oldpeak:describesthedepressionraisedbyexerciserelativetorest.
11. Slope:describestheslopeofthepeakexerciseSTsegment(1forupsloping,2forflat,3fordownsloping).
12. Ca:describesthenumberofbloodvessels.
13. Thal: describes thal feature (3 for normal, 6 for fixeddefect,7forreversibleeffect).
14. Target: describes the target class (0 for no heart disease,1 234forhaving heartdisease).

***FlowDiagram***

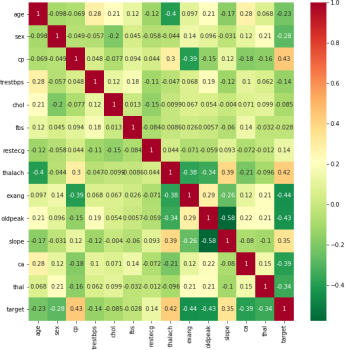
A diagram of a data processing process

Description automatically generated

# ResultsandDiscussion

### CorrelationMatrix

Let’s see the correlation matrix of features. From this graph,we can observe that some features are highly correlated andsomeare not.



**Figure1:**Thisfigureshowsthe correlationmatrix

### .Histogram

The histogram is best and easy way to visualize the databecause it only takes a single line of code to make the plots.Let’s take a look at the plots. Before applying any machinelearningalgorithmswewillhavetolookforcategoricalvariables. The target class is used for describing whether apersonishavingheartdiseaseor not.

A screenshot of a graph

Description automatically generated

**Figure2:**Thisfigureshowsthehistogram.

### ExploratoryDataAnalysis:

Exploratory Data Analysis (EDA) is an approach to analyzethe data sets to describe their main highlights using visualmethods. There are many different methods to conductingexploratory data analysis out there, so it can be hard to knowwhat analysis to perform and how to do it properly. EDA,featureselection,andfeatureengineeringareoftentiedtogetherandare importantsteps inthemachinelearningjourney.

### Barplotfortargetclasswithdifferentfeatures:

It is very important that the dataset we are using should bepre-processed and cleaned. This graph shows the count ofeachtarget class.

A blue and orange rectangles

Description automatically generated

**Figure3:**TargetversusCount

Feature

Theabovegraphshowsthedistributionoftargetversuscount class that is used to predict the total number of heartdisease whether someone has heart disease or not (0 = noheartdisease,1 = havingheartdisease).

A graph of different colored squares

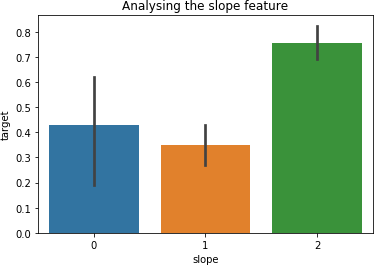
Description automatically generated with medium confidence

**Figure4:**TargetversusThal Feature.

A graph of different colored squares

Description automatically generated

**Figure5:**Targetversus CaFeature.



**Figure6:**TargetversusSlopeFeature.

### 

**Figure9:**Targetversus FbsFeature.

**Figure8:**Targetversus RestecgFeature.

**Figure7:**TargetversusExangFeature.

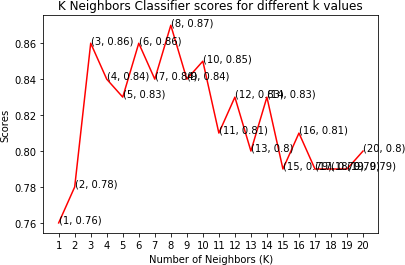
### MachineLearningAlgorithms

***LogisticRegression:***

Logistic regression is asupervised learning algorithm usedtopredictthebinaryformofatargetvariable.Itistheeasiest and simplest algorithm used in machine learning thatcan be used for various problems such as disease prediction,cancer detection and so on. In this paper, we achieved theaccuracyof84%byusingthismodel.

### KNearestNeighborsClassifier:

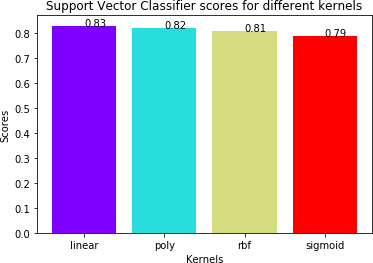
K Nearest Neighbors is a non parametric method used forclassification.Itislazylearningalgorithmwhereallcomputationisdeferreduntilclassification.Itisalsoaninstancebasedlearningalgorithm,wherethefunctionisapproximatedlocally.Thisalgorithmisusedwhentheamount of data is large and there are non-linear decisionboundariesbetweenclasses.KNNexplainsacategoricalvalue using the majority votes of nearest neighbors. Not onlyforclassification,KNNcanbeusedforfunctionapproximationproblem.



**Figure 10:** This figure shows

This graph shows that the maximum accuracy achieved by Kneighborsclassifieris87%.

### SupportVectorClassifier:

SVM (Support Vector Machine)isa supervisedmachinelearning algorithm which can be used for classification andregression problems as support vector classification (SVC)andsupportvectorregression(SVR).Thisclassifierseparates data points using a hyper plane with the largestamount of margin. Support vectors

are the data points whichare closest to the hyper plane. There are several kernels onwhich the hyper plane can be decided. This paper mainlyfocuses on four kernels namely linear, polynomial (poly),radialbasisfunction(rbf)andsigmoid.Thistypeofclassifier uses less memory because they use a subset oftrainingpointsinthe decisionphase.

**Figure 12:** This figure shows the Decision Tree Classifierscores

This graph shows the linegraph from which weobservedthatthemaximumaccuracyis79%andisobtainedbynumber ofmaximumfeatures(2,4,18).

### RandomForestClassifier:

Random forest is a supervised learning algorithm. It can beusedforclassificationandregression.Itissimpleandeasyto implement. A forest is comprised of trees. This classifiercreates decision trees on randomly selected data samples,getspredictionfromeachtreeandselectsthebestsolutionbymeansofvoting.Therandomforestcomposedofmultipledecisiontrees.Itcreatesaforestoftrees.

A colorful rectangular bars with numbers

Description automatically generated

**Figure14:**Thisfigureshows theRandomForest Classifierscores.

This graph shows that the maximum accuracy is 84% andwasobtainedfor both100&500tree.

**Table1:**AccuracyValues

|  |  |
| --- | --- |
| Algorithms | Accuracy |
| LogisticRegression | 78.688525 |
| NaïveBayesClassifier | 80.327869 |
| KNearestNeighborsClassifier | 73.770492 |
| DecisionTreeClassifier | 73.770492 |
| SupportVectorClassifier | 80.327869 |
| RandomForestClassifier | 85.245902 |

Table 1 shows that random forest classifier with 85%incomparisonwiththeothermachinelearningalgorithmsusedinthispaper.BecauseKNN algorithm is based on feature similarity and is one ofthe most famous classification algorithms as of now in theindustry simply due to its simplicity and accuracy.

**Figure11:**ThisfigureshowstheSupport

VectorClassifierscores.

Thisgraphshowsthatthelinearkernelishavingthehighestaccuracyof83%byusingthisdataset.

### DecisionTree Classifier

Thisclassifierfallsunderthecategoryofsupervisedlearning. It can be used to solve regression and classificationproblems. We can use this algorithm for issues where wehavecontinuousbutalsocategoricalinputandtargetfeatures. It is the most effective machine learning algorithmused fordescribingthetreesina graphicalmanner.

K nearestneighbors is a simple algorithm that stores all the accessiblecasesandclassifies newcasesbasedonasimilaritymeasure.

# ConclusionandFutureWork

This paper involves prediction of the heart disease datasetwith proper data processing and implementation of machinelearningalgorithms.Inthispaper,weusesfive machinelearningalgorithmsfor prediction.

Amongallthemachinelearningalgorithmsusedinthispaper,thehighestaccuracyisachievedbyrandom forest classifier with 85%. This paper shows that themachine learning algorithms can be used to predict the heartdiseaseeasilywithdifferentparametersandmodels.Machinelearningisveryusefulinprediction,solvingproblems and other areas. Machine learning is an effectivewaytosolvethe problemsindifferentareastoo.

# Acknowledgement

Without their active guidance, help, cooperation &encouragement, I would not have been able to write thispaper. I am very thankful for their guidance and help oncompletionofthispaper.

I would like to express my gratitude to “IMS EngineeringCollege” for giving me this great opportunity. I would alsoliketo express my specialthanks tomy parents andmyfamily members, who has always supported me morally aswellaseconomically.

ThankingYou.

# References

1. Avinash Golande, Pavan Kumar T, (June 2019): HeartDisease Prediction Using Effective Machine LearningTechniques,InternationalJournalofRecentTechnology and Engineering (IRTE), ISN: 2277-3878,Volume-8,Issue-1S4.
2. A. SahayaArthy, G.Murugeshwari, (April 2018):Asurvey on heart disease prediction using data miningtechniques.
3. AmitaMalav,KalyaniKadam,(2018):“AHybridApproach for Heart Disease Prediction Using ArtificialNeural Network and K – Means”, International JournalofPureand AppliedMathematics.
4. Benjamin EJ et.al, (2018):Heart Diseaseand StrokeStatisticsAt-a-Glance.
5. DhafarHamed,JwanK.Alwan,MohamedIbrahim,MohammadB.Naeem,(march–2017):“TheUtilizationofMachineLearningApproachesforMedical Data Classification” in Annual Conference onNewTrendsinInformation&CommunicationsTechnologyApplications.
6. HimanshuSharma,MARizvi,(August2017):Prediction of Heart Disease Using Machine LearningAlgorithms:ASurvey.
7. IKetutAgungEnriko,MuhammadSuryanegara,DadangGunawanal,(June2018):“HeartDiseaseDiagnosis System with k – Nearest Neighbors MethodUsingRealClinicalMedicalRecords”,4thInternationalConference
8. Monika Gandhi, Shailendra Narayanan Singh, (2015):Predictions in heart diseases using techniques of datamining.
9. M.S.Amin,Y.K.Chiam,K.D.Varathan,(Mar.2019):Identication of significant features anddataminingtechniquesinpredictingheartdisease,TelematicsInform., vol. 36,pp. 8293.
10. SenthilkumarMohan,ChandrasegarThirumalai,Gautam Srivastava, (2019):Effective Heart DiseasePredictionUsingHybridMachineLearningTechniques, Digital Object Identifier10.1109/ACCESS.2019.2923707,IEEEAccess,VOLUME 7.
11. V.V.Ramalingam,AyantanDandapath,MKarthikRaja,(2018):heartdiseasepredictionusingmachine

# Lakshmanarao,Y.Swathi,P.SriSaiSundareswar,(November 2019): Machine Learning Techniques ForHeartDiseasePrediction,InternationalJournalOfScientific

learningtechniques:asurvey,InternationalJournalofEngineering&Technology(IJET),7(2.8)684-687.

# Author Profile

**Mahak Tayal** currently studying in 2nd year in the department of Information Technology at Indira Gandhi Delhi Technical University. She is deeply interested in python and machine learning.

**Yuvika Mittal** currently studying in 2nd year in the department of Mechanical engineering at Indira Gandhi Delhi Technical University. She is deeply interested in python and machine learning.

**Jiya Yadav** currently studying in 2nd year in the department of Electronics and communication engineering at Indira Gandhi Delhi Technical University. She is deeply interested in python and machine learning.