TITLE : HEART DISEASE STATLOG DATASET USING MACHINE LEARNING ALGORITHMS

Project Title : Heart disease statlog dataset using machine learning algorithms
Mail-id : moezzema10@gmail.com
GitHub:https://github.com/sayyedamoezzema

1. ABSTRACT : Heart Disease prediction is one of the most complicated task in medical field.In the modern era.approximately one person dies per minute due to heart disease.Data science plays crucial role in processing huge amount of data in the field of health care.As heart disease prediction is a complex task, there is a need to automate the prediction process to avoid the risk associated with it and alert the patient well in advance.

The purposed work predicts the chances of the heart disease and classifies patient's risk level by implementing different data mining techniques such as Logistic Regression,Decision Tree, Naïve Bayes, SVM Classifier, KNN Classiefier.These algorithms work by learning from historical data to predict the probability.

In this case we have analyzed a dataset of different patient have different information from each other.The patient information collect from various source to predict the occurrence of heart disease.In this dataset contains patient various attributes such as age,sex,cp,chol,ca and sloop etc.

In this dataset we have applied multiple classification algorithms on the dataset and compared their performance based on various metrics such as accuracy,recall,precision,sensitivity and specificity.

Keywords – Logistic Regression, Decision Tree,SVM Classifier, KNN Classifier, Naïve Bayes, Heart Disease Prediction.

After applied different models they shows different result, the logistic regression outperforms other algorithms in predicting the heart disease,achieving an acuuracy of 83.33%.We have also identified the most important features that contribute to the prediction of heart disease,which can help healthcare providers to focus on this risk factors and preventive measures.

In conclusion, Machine learning (ML) has shown great potential in improving the diagnosis, prognosis, and treatment of heart disease. ML algorithms can analyze large datasets and identify patterns and associations that may not be easily identifiable by humans. This can lead to more accurate predictions of disease outcomes and better personalized treatment plans.

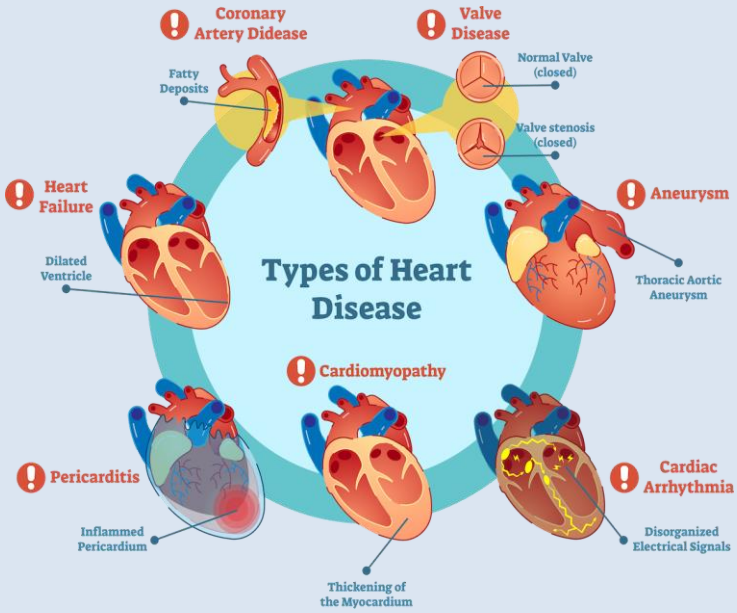
However, there are also limitations and challenges in using ML for heart disease. One major concern is the potential for bias and errors in the algorithms, which can lead to incorrect diagnoses and treatment recommendations. Additionally, there are issues around data privacy and security when using sensitive medical data.

Despite these challenges, ML is likely to continue playing an important role in heart disease research and clinical practice. Ongoing research and development in ML algorithms,as well as efforts to address ethical and technical concerns, will be critical in maximizing the potential benefits of ML for heart disease. Ultimately, the integration of ML with traditional medical approaches may lead to more effective and efficient heart disease care, ultimately improving patient outcomes and quality of life.

2. INTRODUCTION : The work proposed of this the data set to prediction of heart disease.Human heart is the principal part of human body.Basically ,it regulates blood flow throughout our body any irregularity to heart can cause distress in other part of body.Any sort of disturbance to normal functioning of the heart can be classified as a heart disease.In today's contemporary world,heart disease is one of the primary reasons for occurrence of most deaths.Heart Disease may occur due to unhealthy life style,smoking,alcohol and high intake of fat which may cause hypertension.According to the world health organization more than ten million die due to heart disease every single year around the world.A healthy life style and earliest are only ways to prevent the heart related disease.The main goal of a machine learning project may vary depending on the specific problem being addressed.It could involve predicting customer behavior,diagnosing a disease.

In this project the patient if he or she is having disease or not based on the variouse columns such as below.

1.Age: Patients Age in years (Numeric)
2.Sex: Gender (Male : 1; Female : 0) (Nominal)
3.cp: Type of chest pain experienced by patient. This term categorized into 4 category.
4.trestbps: patient's level of blood pressure at resting mode in mm/HG (Numerical)
5.chol: Serum cholesterol in mg/dl (Numeric)
6.fbs: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)
7.restecg: Result of electrocardiogram while at rest are represented in 3 distinct values
8.thalach: Maximum heart rate achieved (Numeric)
9.exang: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)
10.oldpeak: Exercise induced ST-depression in relative with the state of rest (Numeric)
11.slope: ST segment measured in terms of slope during peak exercise.
12.ca: The number of major vessels (0–3)(nominal)
13.thal: A blood disorder called thalassemia
14.target: It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.



The main challenge in today's healthcare is provision of best quality service and effective accurate diagnosis,Even if heart disease are found as the prime source of death in the world recent years,they are also the ones that can be controlled and manage effectively.The whole accuracy in management of disease lies on the proper time of detection of that disease.The purposed work makes an attempt to detect these heart disease at early stage to avoid disastrous consequence.

The main goal of study case is to provide a tool for doctors to detect heart disease as early stage.This in turnwill help to provide effective treatment to patients and avoid severe consequences.ML plays a very important role to detect the hidden discrete patterns and there by analyse the given data.After analysis of data ML techniques help in heart disease prediction and diagnosis.This study case presents performance analysis of various ML techniques such as Logisitc Regression, Naïve Bayes, SVM Classifier, Decision Tree, KNN Classifier.

3. BRIEF LITERATURE REVIEW : With growing development in the field of medical science along side machine learning various experiments and researches has been carried out in these releasing the relevant significant papers.

Santhana Krishnan. J ,et ,al proposed a paper "Prediction of Heart Disease Using Machine Learning Algorithms" using decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables.But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Clevel and data set.Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

Sonam Nikhar et al proposed paper " Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease.Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier.

Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.

Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an opensource data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

Overall, it is a important to be aware of the biases present in the heart disease dataset take step to patient them in order to ensure accurate analysis and prediction.

4. METHODOLOGY :

4.1.1 DATA COLLECTION : Initially, we collect a dataset for our heart disease prediction from kaggle. The data set include information of columns in that columns multiple attributes like, age,sex, ca, chol,fbs, and cp etc. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 80% of training data is used and 20% of data is used for testing.

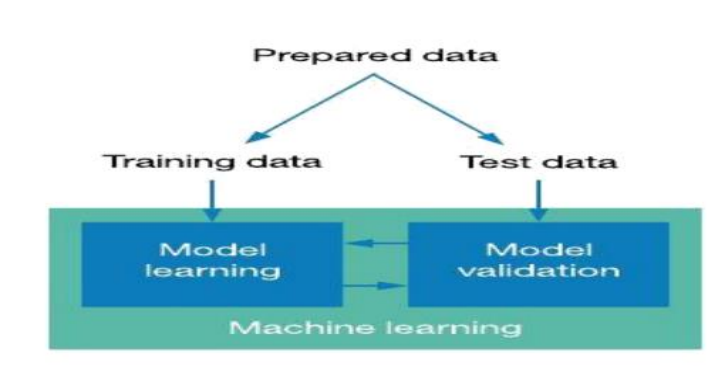


Fig : Data Collection

4.1.2 SELECTION OF ATTRIBUTES : Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

4.1.3 DATA PREPROCESSING : Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset.

Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

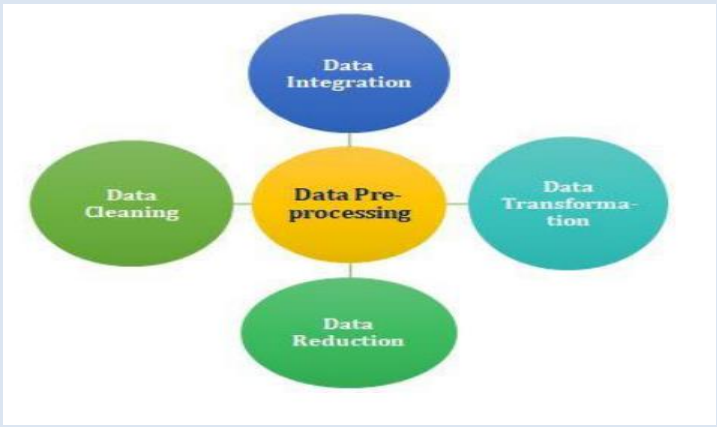


Fig : Data Preprocessing

4.1.4 PREDICTION OF DISEASE : Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, KNN are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

4.1.5 CLASSIFICATION MODEL DEVELOPMENT:

Logistic regression is a widely used statistical method for analyzing and modeling binary or categorical data. In machine learning, logistic regression is a popular algorithm for developing classification models.

To develop a classification model using logistic regression, the following steps can be taken:

4.2 MODEL SELECTION : Logistic Regression model selected for classification model based on its simplicity,interpretability, and ability to handel binary classification problem.

4.2.1FEARUTRE SELECTION : Identify the most relevant features that can be used to predict the outcome variable.

4.2.2MODEL BUILDING : Build the logistic regression model by training it on the pre-processed data with the selected features.

4.2.3MODEL EVALUATION : Evaluate the performance of the model by using metrics such as accuracy, precision, recall, and F1 score.

4.2.4MODEL TUNING : If the model performance is not satisfactory, tune the hyperparameters of the model and repeat the evaluation process until the desired performance is achieved.

4.2.5MODEL TRAINING : The logistic regression model was trained on the preprocessed training data.

4.2.6PERFORMANCE METRICS : Performance metrics such as accuracy, precision, recall, F1 score, ROC curve, and AUC score are commonly used to evaluate the effectiveness and accuracy of a logistic regression model. These metrics help to assess the model's ability to correctly predict the outcome variable and provide a way to compare the performance of different models. It is important to choose the appropriate metric(s) based on the specific problem and the goals of the analysis.

4.2.7BIAS CRITERIA : Bias criteria should be taken into consideration when developing heart disease prediction models to ensure that the model is accurate, reliable, and generalizable. Some important bias criteria to consider when developing a heart disease prediction model include:

Bias can occur if there are differences in other factors (e.g., age, gender, lifestyle habits,cp,chol) between the groups being studied that can influence the outcome. To address this, the model should identify and control for potential confounding variables.

Bias can occur if the evaluation metrics used to assess model performance are biased, leading to inaccurate conclusions. Evaluation metrics should be chosen carefully and should be sensitive to the population being studied.By addressing these sources of bias, heart disease prediction models can be developed that are accurate, reliable, and can improve healthcare outcomes for all patients.

DATA VISUALIZATION :

Box Plot : Box plot we created for some attribute like age,trestbps,chol and thalach. In this box we show that the some attribute of in this outlier is present .

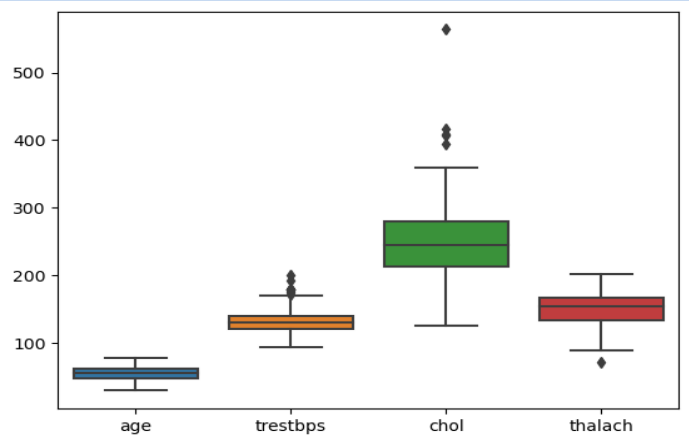


Fig : BoxPlot on age,trestbps,chol,thalach

Histogram : Histogram were created for the continuous variable such as age .Histogram shows the distribution of the variable and can help to identity any outlier or unusual patterns.

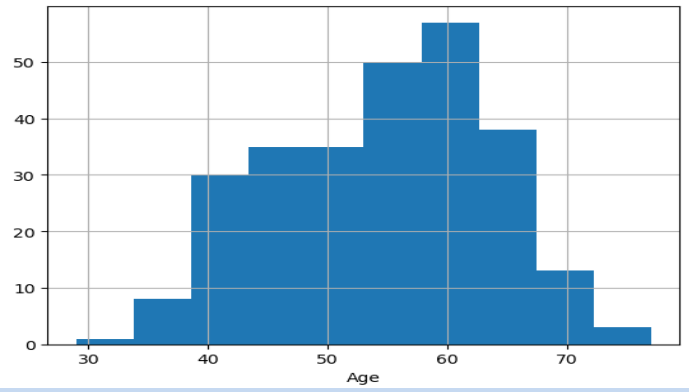


Fig : Histogram on Age

Regression Chart : In a regression chart for heart disease, we typically plot the relevant independent variables such as age of patient on the x-axis, and the dependent variable, and the presence or absence of heart disease trestpbs on the y-axis.

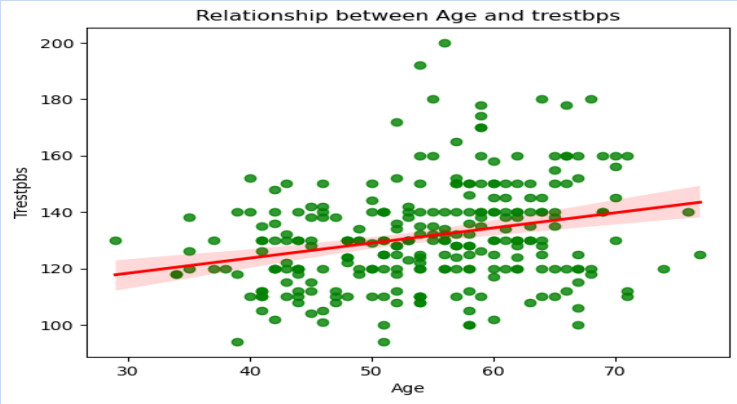


Fig : Regression chart age Vs trestpbs

Heatmaps : Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

Correlation matrix heatmap

5. FINDINGS AND DISCUSSSION :

After performing data analysis and modeling on the heart disease dataset, the following finding were observed.

Heart disease is a major cause of mortality worldwide, and there have been several efforts to use machine learning to predict the risk of heart disease and improve early detection.

Machine learning algorithms have been trained on datasets containing a range of risk factors, such as age, sex, blood pressure, cholesterol levels, and cp.

Trained on these datasets to predict the likelihood of heart disease based on patient characteristics and risk factors. These algorithms can also be used to identify patterns in the data that may not be apparent to human analysts.

Ethical and Practical Implications : The result of analysis have ethical and practical implication, The use of machine learning for heart disease risk prediction has several practical implications. It may lead to earlier detection of heart disease, which could improve patient outcomes and reduce healthcare costs. However, it is also essential to consider the practical limitations of these models, such as their reliability, accuracy, and cost-effectiveness.

Overall, heart disease can have significant ethical and practical implications for individuals, families, and society as a whole. Addressing these implications requires a comprehensive approach that involves healthcare professionals, policy makers,and the broader community.

Possible Unintended Consequences : Heart disease can result in increased healthcare costs, both for the individual and the healthcare system. These costs can be associated with medical treatment, hospitalizations, and rehabilitation, placing a significant burden on the healthcare system and individuals.

Heart disease datasets can be used to inform public health interventions, such as policies, programs, and campaigns. However, there is a risk that these interventions may have unintended consequences,such as stigmatization, reduced access to healthcare, or adverse outcomes for certain population groups.

Possible unintended consequences of heart disease include emotional distress, disability, increased healthcare costs, reduced productivity, social disparities, privacy concerns, bias, medical errors,misinterpretation, unintended consequences of interventions, and data breaches.

Overall, the use of heart disease datasets can have unintended consequences that require careful consideration and management. Addressing these unintended consequences requires a comprehensive approach that addresses privacy concerns, bias, data accuracy, expertise, and potential unintended consequences of interventions.

6.CONCLUSION :

1.Our machine learning algorithm can now classify patients with heart disease,Now we can properly diagnose patients, and get them the help they need to recover.By diagnosing detecting these feature carly,we may prevent worse symptoms from arising later.

2.Our Logistic Regression Classifier yield the highest accuracy 83.33% any accuracy 70% is considered good,but be careful because if your accuracy is extremely high,it may be too good to be true(an example of over fitting).Thus 83.33% is the ideal accuracy!

7.REFERENCES :

1.Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982 92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

2.Patel S & Chauhan Y (2014). Heart attack detection and medical attention using motion sensing device - kinect. International Journal of Scientific and Research Publications, 4(1), 1-4.

3.Maas, A.H.; Appelman, Y.E. Gender differences in coronary heart disease. Neth. Heart J. 2010, 18, 598–602.

4.A. Aldallal and A. A. A. Al-Moosa, "Using Data Mining Techniques to Predict Diabetes and Heart Diseases", 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), pp. 150-154, 2018, September.

5.Ankita Dewan and Meghna Sharma, "Prediction of heart disease using a hybrid technique in data mining classification", 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)