iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** III **Month of publication:** March 2021

DOI: https://doi.org/10.22214/ijraset.2021.33228

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

Comparative Study of Naive Bayes, Gaussian Naive Bayes Classifier and Decision Tree Algorithms for Prediction of Heart Diseases

Sushma S A[1], Keerthan Kumar T G[2]
[1, 2]*Assistant Professor Department of Information Science, Siddaganga Institute of Technology, Tumakuru, Karnataka.*

Abstract: ***Nowadays death due to heart disease has been common in the world. It has become a hard task for the medical practitioners to diagnose in the initial stage and requires more expertise and demand in the medical field for prediction. Designing an automated system by using machine learning algorithm will improve the medical efficiency and also reduce the cost. In this paper we are planning to design an automated system that can be used for efficiently predicting the results which give information about the risks need to be faced by the patients with respect to heart diseases by using the parameter available in the dataset. We are extracting the hidden patterns from the parameters by applying data mining techniques. Since the heart data is too massive and complex for analysis using traditional techniques, we are using machine learning algorithm for computation using the parameters available in the dataset and produce accurate prediction of heart disease. Machine Learning Prediction techniques like Naive Bayes Classifier, Gaussian Naïve Bayes Classifier and Decision tree can be used to analyze and predict the heart diseases.***
Keywords: Naive Bayes Classifier, Gaussian Naive Bayes Classifier, Decision tree, Prediction Engine.

I. INTRODUCTION

Data Mining is an important technique used for extracting meaningful information from given dataset which can be used for analysis and provide preventive measure to individual with the help of prediction techniques to cure that disease in the very beginning stage. Over last 10 years health organizations are maintaining huge amount of patient data in digitized form so it is available for researchers to do analysis and predictions than keeping it in hard copy which is very difficult to manage. Due to advancement in the technology big data can be used for biomedical research and healthcare communities to manage and store complex data and at the same time processing is also fast by using different Map Reduce techniques. But one thing we should keep in mind is that the medical data should be complete otherwise there can be weakness or inefficiency in predicting the risk of the disease relevant to the research study we do. In this paper we are collecting real time data from the hospitals for analysis purpose. To avoid any deviations in the analysis we can use a latent factor to reform the missing data. The data might be in the form of structured, unstructured and semi structured data. So, we can use Map reduce algorithm to process these types of different data. The objective of this paper is to give awareness to the people in the very beginning stage to identify the risk by inputting basic health parameters related to patient like blood pressure, weight, height, body mass index etc. so that it is easy for us to predict heart diseases.

Nowadays irrespective of people are living either in village or cities death due to heart disease has been common from age of 15 to 40. Due to change in lifestyle and food habits of an individual there has been a major effect in people suffering from heart diseases. It has been found that people of age range between 30 to 69 years around 1.3 million people have died because of cardiovascular diseases,0.9 million have died because of coronary heart disease and 0.4 million by stroke.

It is also found that people born before 1970 are the victims of these heart diseases and majority of them are from urban cities. This has motivated us to make analysis and predict the heart disease at the very initial stage and help them in undergoing proper medical diagnosis and decrease the death of the people.

II. LITERATURE SURVEY

The impediment in recognizing the heart illnesses just as issues because of different components like cholesterol, resting ECG, hypertension, diabetes, unusual heartbeat rate and numerous different elements. The procedures and techniques like information mining and neural systems have been used to discover the seriousness just as reality of heart illnesses among patients. The reality of the illness is classified and recognized based on techniques like K-Nearest Neighbor Algorithm (KNN), Genetic calculation (GA), Decision Trees (DT) just as Naive Bayes (NB) [1].

The natural highlights of coronary illness are mind boggling just as intense and henceforth, the ailment must be maneuvered carefully. The viewpoint and origination of clinical science just as information digging are utilized for finding different sorts of conditions identifying with digestion. Information mining and examination with classification has a critical and crucial job in foreseeing heart sicknesses just as exploring information.

To deliver an expectation model at least two than two methods have been utilized together regularly called as half-breed model. The pulse time arrangement have been utilized to present Neural system. Neural system calculation consolidates back probabilities and anticipated qualities from various forerunner procedures. This model accomplishes a precision of 89.01% which is a superior outcome contrasted with past works.

This technique utilizes different and numerous other clinical records for forecast, for example, Left pack branch square (LBBB), Right group branch square (RBBB), Normal Sinus Rhythm (NSR), Sinus bradycardia (SBR) Second degree square (BII) to find out the exact and precise state of the patient relating to coronary illness. An outspread premise work arranges (RBFN) is available in dataset that has been utilized for classification, where 70% of the information is placed being used for preparing and the staying 30% is for classification. In the field of clinical and exploration, Computer Aided Decision Support System (CADSS) is likewise presented. The usage of information mining strategies in the medicinal services industry and clinical field has appeared to set aside a lot lesser effort for forecast and assurance of heart illnesses with more exact and exact outcomes. The proposed technique utilizes 15 boundaries for the coronary illness forecast and examination. The yield results show an expanded degree of execution contrasting it with the current ways just as techniques.

There is sufficient and enormous number of works in this field legitimately identified with this venture. Counterfeit Neural Network has been acquainted with give the most elevated exactness and accuracy in clinical field. The outcomes that are acquired are contrasted and the aftereffects of existing models inside a similar space and those supposedly was improved. The information of patients experiencing heart ailments gathered and collected from the University of California (UCI) research center and were utilized to find designs with Neural Networks (NN), DT, Support Vector Machines (SVM) [3] and Naive Bayes. The outcomes are thought about for execution just as precision with these calculations. A gigantic measure of information produced and gathered by the clinical business has not been utilized adequately whenever beforehand. The new methodologies and techniques introduced following limits the expense just as improve the forecast and assurance of coronary illness in a simple and powerful manner. Numerous studies have been done that have focus on diagnosis as well as analysis of heart disease. There have been applied different data mining techniques for diagnosis and for achieving different probabilities for different methods. Smart Heart Disease Prediction System (IHDPS) has been placed being developed by utilizing information mining methods, for example, Naive Bayes, Neural Network, and Decision Trees has been proposed by Sellappan Palaniappan. Every technique has its own quality and ability to reach to proper outcomes. For building this framework concealed examples and connection between them is utilized too. It is electronic, easy to understand and furthermore expandable.

1) To build up the multi-parametric element with direct and nonlinear attributes of HRV (Heart Rate Variability) a novel method was proposed by Heon Gyu Lee etal. To accomplish this, they have utilized a few classifiers for example CMAR, Bayesian Classifiers (Classification based on Multiple Association Rules), (Decision Tree) just as SVM (Support Vector Machine).

2) The trouble in distinguishing compelled affiliation rules for coronary illness expectation just as investigation was concentrated via Carlos Ordonez. The subsequent dataset got contains records of patients experiencing coronary illness. Three imperatives were acquainted with decline the sum. The prediction and determination of Heart disease, Blood Pressure as well as Sugar with the aid of neural networks has been proposed by Niti Guru. The dataset containing records with 13 attributes in each record. The supervised networks i.e. Neural Network with back propagation algorithm is used for training as well as testing of data of patterns [6].

They are as follows:

Separate the attributes into groups. i.e. uninteresting groups.

a) In a rule, there should not be unlimited number of attributes, but limited. The result of this is divided into two section of rules i.e. either the existence or absence of heart disease.

b) Franck Le Duff has built a decision tree along with database of patient for a medical problem.

c) Latha Parthiban likewise anticipated an effective methodology on premise of coactive neuro-fluffy induction framework (CANFIS) for forecast and breaking down of coronary illness. The CANFIS model uses neural system abilities with the fluffy rationale just as hereditary calculation.

d) Kiyong Noh utilized an arrangement calculation for the extraction of highlights that multiparametric in nature by surveying HRV (Heart Rate Variability) from ECG, information pre-preparing and coronary illness design. The dataset including 670 people groups, circulated and partitioned into two gatherings, in particular customary ordinary individuals and patients enduring with coronary illness, were utilized to complete the investigation for the affiliated classifier. ANN has been acquainted with produce the one of the most noteworthy precision forecasts in the clinical field [6]. The back-spread multilayer observation (MLP) of ANN has been utilized to foresee heart illnesses. The acquired yield results are then contrasted and the aftereffects of existing models inside a similar area and saw as improved. The information of coronary illness patients gathered from the UCI research center is utilized to find designs with DT, Support Vector machines SVM, NN and Naive Bayes. The yield results are contrasted for execution and precision and these calculations. This proposed half and half strategy portrays aftereffects of 86.8% for F-measure, contending with the other existing techniques. The arrangement without division of Convolutional Neural Networks (CNN) is presented here. This strategy considers the heart cycles with numerous sorts of introductory situations from the Electrocardiogram (ECG) signals in the preparation stage. CNN can produce highlights with different situations in the testing phase of the patient. An enormous measure of information produced by the clinical business has not been utilized adequately already. The new methodologies introduced here reduction the expense just as improve the forecast and investigation of coronary illness in a simple and successful way.

III. PROPOSED SYSTEM

Patients suffering from Cardiovascular diseases are around 80% in India's total population. and primary reason for death are symptoms of panic heart attack and stroke. Due to expensive medical costs people are not affordable to undergo treatments at medical hospitals. Quality service indicates diagnosing patients correctly and administering treatments that are effective. Clinical decisions are often made based on doctors' perception and practice rather than on the knowledge-rich data hidden in the database. This practice points to uninvited biases, mistakes and extreme medical expenses which affects the quality of facility delivered to patients [2]. Supervised learning trains a model on known input and output data so that it can predict future outputs, and unsupervised learning, which finds hidden patterns or intrinsic structures in input data. This research work is intended to use supervised machine learning algorithms to predict the heart diseases. Supervised methods are an effort to determine the association between input attributes and a target attribute. The relationship revealed is represented in a structure referred to as a model. Classification model and regression model are the two main models in supervised learning. Here this work concentrates on classification model. Classification deals with allocating observations into distinct classes, rather than appraising continuous quantities. This research work uses some of the classification algorithms like Naïve Bayes, Gaussian Naïve bayes and Decision tree to predict the heart diseases and compare their performance.

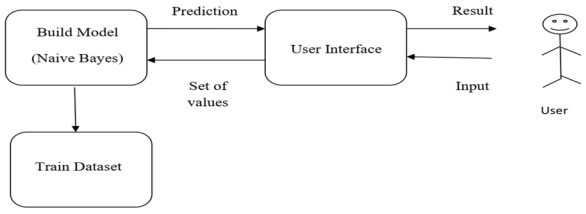


Fig 3 Proposed System for Heart Disease Prediction

The proposed system consists of a user interface where an individual can enter his health details along with some parameters related to heart data done after testing through ECG. This parameter set is passed to the data mining model where we apply different algorithm like Naive Bayes, Gaussian Naive Bayes and Decision tree algorithms which will enable us to predict the heart diseases and tell how healthy the heart is.

A. *Dataset*

The dataset consists of 920 individuals' data. There are 15 columns within the dataset from age to diagnosis of cardiopathy.

1) *Age:* Represents the age of the individual.
2) *Sex:* Represents the gender of the individual using the subsequent format: 1 =male 0 = female.
3) *Chest-pain Type:* This displays the sort of chest-pain experienced by the individual using the subsequent format: 1 = typical angina 2 = atypical angina 3 = non - angina pain 4 = asymptotic

4) *Resting Blood Pressure:* This contains the resting pressure level value of a person in mmHg (unit).
5) *Serum Cholesterol:* This contains the amount serum cholesterol in mg/dl(unit).
6) *Fasting Blood Sugar:* In this we are comparing the fasting blood glucose value of a private with 120mg/dl. If fasting blood glucose > 120mg/dl, then: 1 (true) else: 0 (false).
7) *Resting ECG:* This is described as 0 for normal 1for having ST-T wave abnormality and 2 for left ventricular hypertrophy.
8) *Max Rate Achieved:* This describes the max rate achieved by a person.
9) *Exercise Induced Angina:* This describes as 1 for yes and 0 for no.
10) *ST Depression:* It induced by exercise related to rest and displays the worth which is an integer or can be float too.
11) *Peak Exercise ST Segment:* This described as 1 for upsloping, 2 for flat and 3 for down sloping.
12) *The Number of Major Vessels Ranging From 0 To 3 Colored By Fluoroscopy:* It describes the worth as integer or float.
13) *Thal:* It displays the thalassemia: 3 for normal, 6 for fixed defect and 7 for reversible defect
14) *Diagnosis of Cardiopathy:* It describes whether the individual is affected by heart disease or not: 0 forabsence and 1,2,3,4 for present

B. *Need for These Dataset Parameters*

1) *Age:* Age is that the most fundamental hazard thinks about creating cardiovascular or heart ailments, with around a significantly increasing of hazard with every time of life. Coronary greasy streaks can start to make in youth. it's assessed that 82 percent of people who pass on of coronary cardiopathy are 65 and more established. At the same time, the possibility of stroke pairs each decade after age 55.
2) *Sex:* Men are at more danger of cardiopathy than pre-menopausal ladies. Once past menopause, it's been contended that a lady's hazard is practically identical to a man's albeit more present-day information from the WHO and UN questions this. On the off chance that a female has diabetes, she is bound to create cardiopathy than a male with diabetes.
3) *Angina (Chest Pain):* Angina is agony or uneasiness caused when your solid tissue doesn't get enough oxygen-rich blood. it will want weight or crushing in your chest. The anxiety can likewise happen in shoulders, arms, neck, jaw, or might be toward the rear. Angina torment may even want acid reflux.
4) *Resting Blood Pressure:* After some time, the high-pressure level can harm conduits that feed your heart. The high-pressure level that occurs with different conditions, similar to weight, elevated cholesterol or diabetes, expands your hazard significantly more
5) *Serum Cholesterol:* A significant level of beta-lipoprotein (LDL) cholesterol (the"terrible" cholesterol) is conceivable to limit courses. An elevated level of fatty oils, such a blood fat related with your eating regimen, likewise ups your danger of coronary disappointment. Be that as it may, an elevated level of lipoprotein (HDL) cholesterol (the "great" cholesterol) brings down your danger of coronary disappointment
6) *Fasting Blood Sugar:* Not creating a sufficient hormone discharged by your pancreas (insulin) or not reacting to insulin appropriately causes your body's blood glucose levels to rise, expanding your danger of coronary disappointment.
7) *Resting ECG:* For individuals at generally safe of upset, the USPSTF closes with moderate assurance that the possible damages of screening with resting or exercise ECG rise to or surpass the likely advantages. For individuals at middle of the road to high hazard, current proof is inadequate to evaluate the equalization of focal points and damages of screening.
8) *Max Rate Achieved:* the ascent inside the cardiovascular hazard, identified with the increasing speed of rate, was, for example, the ascent in chance saw with high-pressure level. it's been indicated that an ascent in rate by 10 beats for every moment was identified with an ascent inside the danger of cardiovascular passing by at least 20%, and this expansion inside the hazard is practically identical to the one saw with an ascent in systolic weight level by 10 weight unit.
9) *Exercise-induced Angina:* The agony or distress identified with angina for the most part feels tight, grasping or crushing, and may differ from gentle to extreme. Angina is here and there felt inside the focal point of your chest yet may spread to either or both of your shoulders, or your back, neck, jaw or arm. It can even be felt in your hands sorts of Angina

a) Stable Angina/heart disease
b) Unstable Angina
c) Variant (Prinzmetal) Angina
d) Microvascular Angina
e) ST depression induced by exercise relative to rest

Peak exercise ST segment: A treadmill ECG check is considered abnormal when there's a horizontal or down-sloping ST-segment depression ≥ 1 mm at 60– 80 ms after the J point. By and large, the event of level or downs lanting ST-fragment gloom at a lower outstanding task at hand (determined in METs) or rate demonstrates a more awful visualization and better probability of multi-vessel malady. The term of ST-fragment sadness is furthermore significant, as drawn-out recuperation after pinnacle pressure is as per a positive treadmill ECG check. Another finding that is exceptionally demonstrative of genuine CAD is that the event of ST-portion rise > 1 mm (regularly proposing transmural ischemia); these patients are every now and again alluded critically for coronary angiography.

IV. MODEL TRAINING AND PREDICTION

The proposed expectation model will be prepared by investigating existing information since we definitely know whether every patient has cardiopathy or not. This technique is moreover referenced to as management and learning. The prepared model is then wont to anticipate and decide whether clients endure cardiopathy. The preparation just as expectation strategy is portrayed as follows

A. Splitting

Initially, information is separated into two division utilizing part parting. In this, information is part in a proportion of 75:25 for the preparation set just as the forecast set. The information of preparing set is utilized in part of the calculated relapse for preparing of the dataset, while the expectation set information is utilized in the segment of forecast.

B. Prediction

The two contributions of the part of forecast are the model too the expectation set. The forecast outcome shows the anticipated, decided information, genuine information, just as the likelihood of various and different outcomes in each gathering.

V. ALGORITHMS USED FOR PREDICTION OF HEART DISEASES.

A. Naive Bayes Classifier

Credulous Bayes classifiers are a lot of arrangement calculations dependent on Bayes Theorem. Bayes Theorem: We can find that with Bayes hypothesis A will all the more most likely occur if B occurs. Here, the proof is B, and the thought is A. It is expected that the proof and the thought are commonly free. This nearness doesn't influence different qualities. It's called credulous, in this way. These capacities as per the possibility that all sets of highlights are arranged. It is a kind of probabilistic AI model, which is utilized for arrangement undertakings. The classifiers in Naïve Bayes are exceptionally recursive and permit various boundaries to be reliable with the quantity of usefulness/indicators for a learning issue. By utilizing it related to an articulation that takes direct time as opposed to emphasizing it on the whole informational collection, the classifier can be prepared on the most likely result. There are two varieties of classifiers for credulous bayes, as:

1) Bernoulli Naive Bayes

2) Gaussian Naive Bayes

Innocent Bayes calculations are for the most part utilized in separating and suggestion frameworks.

a) Naive Bayes Algorithm: Bayesian rational is useful to decision making. The representation for Naive Bayes is probabilities. It works on Bayes theorem of probability to predict the class of unknown data set. A list of probabilities is stored to file for a learned naive Bayes model. This includes:

- Class Likelihoods: The likelihoods of each class in the training dataset.
- Conditional Likelihoods: The conditional likelihoods of each input value given each class value.

b) Pseudo Code

- *Learning Phase:* Learning a naive Bayes model from your training data is fast.
  Given a training set S and F features and L classes,
  For each target value of ci(ci=c1,....,cL)(ci)⇓ estimate P(ci) with examples in S;
  For all feature value xjk of each feature xj(j=1,...,F;k=1,...,Nj)(xj=xjk| ci) ⇓
  estimate P(xjk|ci) with examples in S;
  Output: F* L conditional probabilistic models
- *Testing Phase:* Training is fast because only the probability of each class and the probability of each class given different input (x) values need to be calculated. Given an unknown instance x'=(a'1 ,...,a'n)Look up tables to assign the label c* to X' if[(a'1|c*)... (a'n|c*)] (c*)>[(a'1|ci)... (a'n|ci)] (ci),ci ≠c*,ci=c1,...,cL

B. Gaussian Naive Bayes Classifier

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be-

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Sometimes assume variance

1) is independent of Y (i.e., σi),

2) or independent of Xi (i.e., σk)

3) or both (i.e., σ)

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution. An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed o define such a distribution.

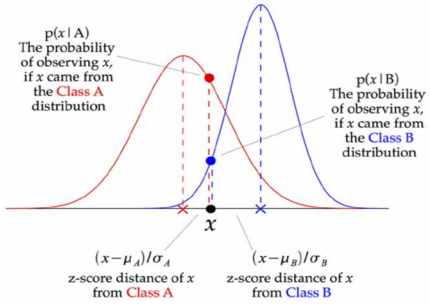


Fig 5.2 Gaussian Naive Bayes Classifier

The above illustration indicates how a Gaussian Naive Bayes (GNB) classifier works. At every data point, the z-score distance between that point and each class-mean is calculated, namely the distance from the class mean divided by the standard deviation of that class.Thus, we see that the Gaussian Naive Bayes has a slightly different approach and can be used efficiently.

C. Decision Tree Algorithm

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

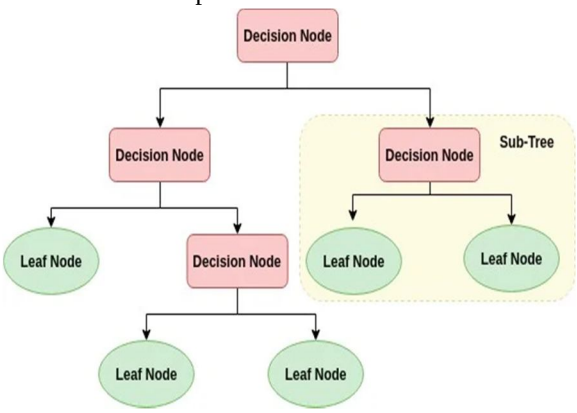


Fig 5.3 Decision Tree Classifier

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

VI. IMPLEMENTATION AND RESULTS

We have a dataset containing 303 rows and 14 columns. The columns corresponds to the different attributes such as age, sex, cp, threstbps, chol, fbs, restecg, thalach, exang, old peak, ca, thal and target. Target is the output variable which is stored in the set Y whereas all the other variables are stored in set X.

```
In [8]: heart = pd.read_csv("dataset.csv")
        heart
```

Out[8]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

Fig 6.1: Dataset

As we can see in the figure 5.1, the dataset is stored as a dataframe in variable heart.Dataframe is created by using the “dataset.csv” file containing the dataset for positive and negative cases of individuals having heart diseases. If the target value is 1 we can say the person has a heart condition and for the value 0 we can say that the person does not have a heart condition.

A. Describing The Dataset

After loading the dataset we need to understand the nature of the dataset and the we need to treat the dataset accordingly for null values, outliers etc. Null values and outliers effect the efficiency of the model a lot because for null values the model won’t understand what to do with those values. Also, for outliers the dataset will have sudden change in the values different than the natural trend in the values which will lead to some unwanted results.

```
5]: heart.describe()
    # print(heart["chol"].max())
```

5]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.00
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.31
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.61
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.00
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.00
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.00
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.00
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.00

Fig: 6.2: datsaset description

In the figure 6.2, we can see that the heart.describe() chunk of code calculate the total count, mean,standard deviation, minimum value, 25 percentile, 50 percentile, 75 percentile and maximum value of each column and each column constitute of one attribute.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

B. Understanding The Dataset

Once the we have a little knowledge of the dataset the main task before fitting the model is to understand the dataset and process it. Knowing about the relation between different attributes is one of the most important tasks to build a model with good fit.

```
heart.target.value_counts()
```

```
1.0    165
0.0    138
Name: target, dtype: int64
```

Fig 6.3: Number of 1's and 0's in target variable

The figure 6.3 shows that number of 1's and 0's in target variable. We should check this in order to check if the dataset is balanced or not. The dataset should not contain a lot of 1s compared to 0s or vice-versa. The same thing is represented graphically in the figure 5.4, i.e, visualizing the dataset. Target variable contains 165 1s, i.e., 54.45% of the total data and 138 0s,i.e., 45.55% of the total data.

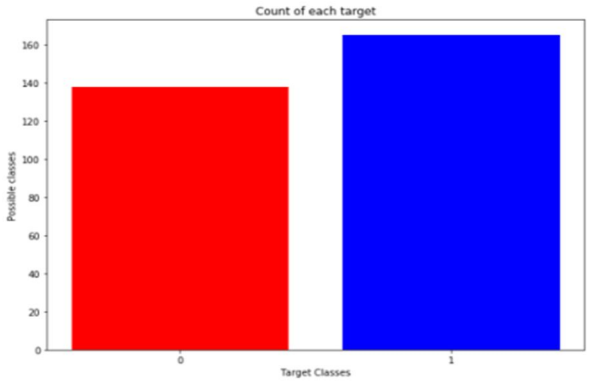


Fig 6.4: Graphical representation of number of 1's and 0's in target variable

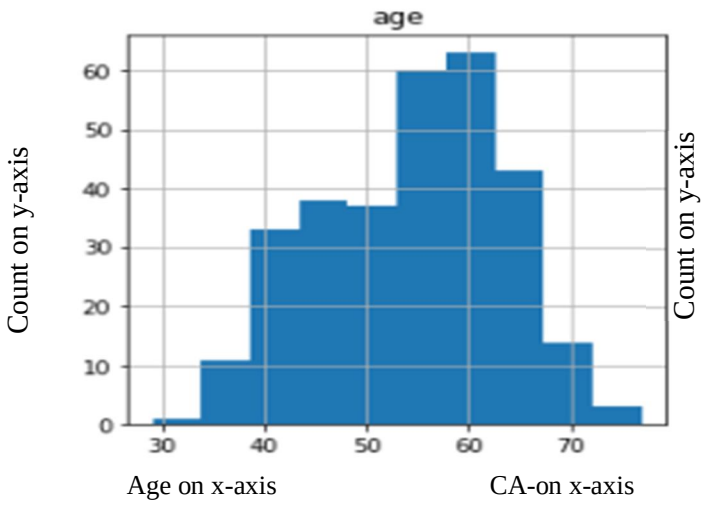


Fig 6.5: Visualizing "age" attribute

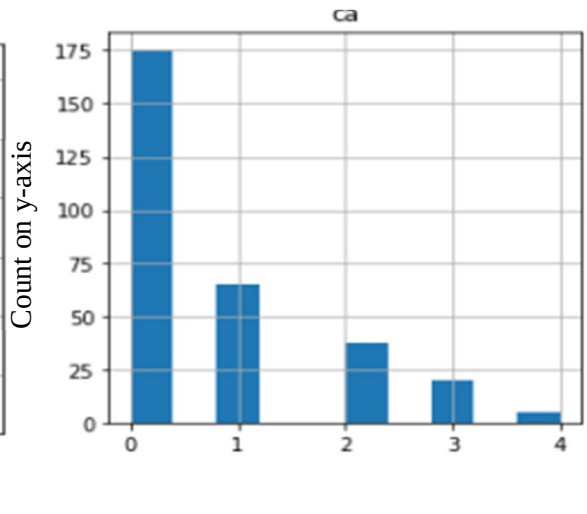


Fig 6.6: Visualizing "ca" attribute

In the figure 6.5, we have plotted "age" at the X-axis and count on the Y-axis. It is good to visualize each attribute to have a good understanding of each feature. We did this for each attribute by plotting the different values against their counts. In figure 6.6, we have plotted "ca" attribute against their count. In the similar way other attributes are also measured before building the model. It is important to understand the degree of associations between the features or attributes. For that purpose, we generate a correlation matrix to check the correlation between the features.

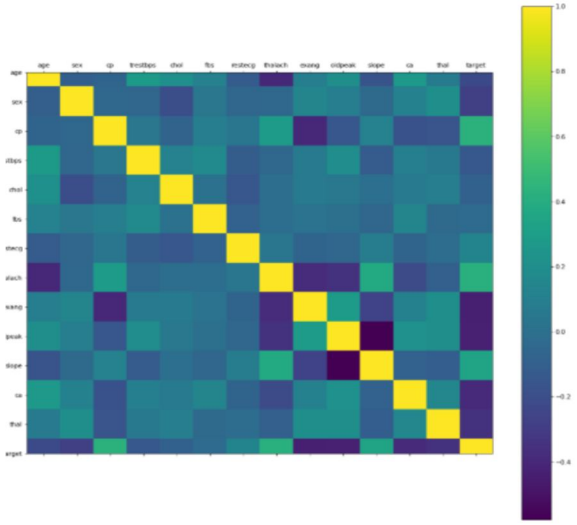


Fig 6.7: Correlation graph

Figure 6.7 shows the correlation graph which shows that there is almost no feature that has significant correlation to the target variable. Also, there are few features that even have negative correlation, and some have lower positive correlation.

C. Processing the Dataset

1) Outlier Treatment: Outlier treatment is an important feature while treating the dataset. Outliers are extreme values that deviate from other observations on data, which effects the efficiency of the model. The model fails to understand on how to comprehend those values. There are multiple ways to treat outliers such as Z-Score or Extreme Value Analysis, Probabilistic and Statistical Modeling, etc. Here we have included Z-Score analysis and removed the values which lie above 75 percentile and those below 25 percentile score.

```
def outlier_treatment(datacolumn):
 sorted(datacolumn)
 Q1,Q3 = np.percentile(datacolumn , [25,75])
 IQR = Q3 - Q1
 lower_range = Q1 - (1.5 * IQR)
 upper_range = Q3 + (1.5 * IQR)
 return lower_range,upper_range
```

Fig 6.8: Outlier Treatment

The fig 6.8 shows the chunk of code which is used to treat the outliers present in the dataset.

D. Scaling The Dataset

Once the outliers are removed from the dataset it is important to scale the dataset within a common range so that we don't get vague results while training the model. So we scale our dataset. There are two ways to scale the dataset in python, one is the MinMax Scalar and the other is Standard Scalar. The one used in this project is the MinMax Scalar. It scales the values in the dataset between 0 and 1. The fig 5.19 shows the chunk of code for the MinMax Scalar.

```
min_max = MinMaxScaler()
columns_to_scale = ['age','cp', 'trestbps', 'chol', 'thalach', 'oldpeak','thal','ca','slope']
heart[columns_to_scale ] = min_max.fit_transform(heart[columns_to_scale ])
```

Fig 6.9: MinMax Scalar

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue III Mar 2021- Available at www.ijraset.com

E. Building the Model

Once the dataset is processed, we are ready to build the model, for the purpose oftraining and testing the fit of the model we need to have a training dataset and a testing dataset. Therefore, we use a built-in function in the sk learn library in python train_test_split. The train_test_split function splits the dataset in such a way that there is not uneven distribution of the target values and maintains the same ratio that was present in the dataset initially. After splitting the dataset into test set and train set, we build the model and check the fit by adding each feature, i.e., first we train the dataset on the one feature and check its first then we keep adding other features one by one until we are done with all the features and keep checking the fit while we add the features. We do it for two classifiers, one is the Decision Tree and the other is the Naïve Bayes Classifier and then compare the results obtained by both the classifiers.

```
y = heart['target']
X = heart.drop(['target'], axis = 1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
dt_scores = []
for i in range(1, len(X_train.columns) + 1):
    dt_classifier = DecisionTreeClassifier(max_features = i, random_state = 0)
    dt_classifier.fit(X_train, y_train)
    dt_scores.append(dt_classifier.score(X_test, y_test))
```

Fig 6.10: Decision tree classifier

```
gnb_scores = []
#l=len(X.columns)
for i in range(1, len(X_train.columns) + 1):
    gaussian=GaussianNB()
    gaussian.fit(X_train.iloc[:,0:i],y_train)
    bayes_pred=gaussian.predict(X_test.iloc[:,0:i])
    #bayes_cm=confusion_matrix(y_test,bayes_pred)
    gnb_scores.append(accuracy_score(bayes_pred,y_test))
```

Fig 6.11: Gaussian Naïve Bayes Classifier

The fits are compared in a tabular form in the Table 6.1.

Number of features	Decision Tree fit	Gaussian Naïve Bayes fit
1	0.676056338028169	0.6619718309859155
2	0.8028169014084507	0.6197183098591549
3	0.676056338028169	0.7464788732394366
4	0.7887323943661971	0.7323943661971831
5	0.7746478873239436	0.7183098591549296
6	0.647887323943662	0.6619718309859155
7	0.8450704225352113	0.704225352112676
8	0.8169014084507042	0.7323943661971831
9	0.8309859154929577	0.7183098591549296
10	0.8309859154929577	0.7464788732394366
11	0.8028169014084507	0.7464788732394366
12	0.7605633802816901	0.7323943661971831
13	0.8169014084507042	0.7746478873239436

Table 6.1: Comparison of Decision Tree and Gaussian Naïve Bayes classifiers

From the table 6.1, we have easily observe that the decision tree classifier almost every time irrespective of the number of features giving the best fit with 0.8450 with seven features (age, sex, cp, threstbps, chol, fbs, restecg, thalach) whereas Gaussian Naïve Bayes give the fit of 0.7042 with the same number of features.

F. Applying Lasso Regression

After building the model and obtaining the above fit, it is not always necessary that these combinations of features are important and they give more accuracy. It is possible that we make use of features which are not as significant and somehow fit the model and obtain a better accuracy. So to make feature selection more significant we apply the Lasso Regression. Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point. The acronym "LASSO" stands for Least Absolute Shrinkage and Selection Operator. Lasso regression performs L1 regularization, which adds a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models With few coefficients, some coefficients can become zero and eliminated from the model. Larger penalties result in coefficient values closer to zero, which is the ideal for producing simpler models

```
lasso=Lasso()
lasso = Lasso(alpha=0.01, max_iter=10e5)
lasso.fit(X_train,y_train)
coef_dict = {}
for coef, feat in zip(lasso.coef_,X_train):
    coef_dict[feat] = coef
coeff_used00001 = np.sum(lasso.coef_!=0)
```

Fig 6.11: Lasso Regression

Fig 6.11 shows the code for applying Lasso Regression and Table 6.2 shows the coefficient obtained for different features.

Features	Co-efficients
age	-0.0
sex	-0.19973138942523366
cp	0.2678074655988135
trestbps	-0.0
chol	-0.0
fbs	-0.0
restecg	0.028909572471217244
thalach	0.0
exang	-0.2016491380437256
oldpeak	-0.31345950953438256
slope	0.12297024165416119
ca	-0.4141596080891296
thal	-0.09888311043791062

Table 6.2: Outcome of Lasso Regression

The co-efficients of Lasso Regression as shown in Table 6.2 clearly shows that the features which have co-efficients 0 are not significant. So now we again train and test the model by removing the insignificant features.

G. Building the Model Again

After applying the Lasso Regression, we build the model again. The model is built similarly as build previously but with the following 8 ('sex', 'cp','restecg', 'exang','oldpeak', 'slope', 'ca', 'thal') features and plotted. The results are compared in the Table 6.3.

Number of features	Decision Tree fit	Gaussian Naïve Bayes fit
1	0.8169014084507042	0.5352112676056338
2	0.7464788732394366	0.4507042253521127
3	0.8450704225352113	0.4507042253521127
4	0.8591549295774648	0.4647887323943662
5	0.8591549295774648	0.5915492957746479
6	0.8450704225352113	0.6338028169014085
7	0.8450704225352113	0.49295774647887325
8	0.802816901408450	0.5070422535211268

Table 6.3: Comparison of Decision Tree and Gaussian Naïve Bayes classifiers fit after Lasso Regression

From the Table 6.3, we can easily observe that the Decision Tree Classifier gives better fit compared to the Gaussian Naïve Bayes Classifier. With all the significant features the decision tree classifier gives a fit of 0.8028 whereas the Gaussian Naïve Bayes gives a lot less, i.e., 0.5070. Therefore, we can easily conclude that the decision tree classifier is a better method to model.

VII. CONCLUSION AND FUTURE ENHANCEMENTS

The scope of this project will help us find multiple opportunities in the future regarding the current medical application scenario. We are continuing to tweak the project with added functionality and modifications to make it useful for people working in the medical field. The main motive would be to improve the data set and machine learning model in order to increase the project 's efficiency. We are currently using the Naïve Bayes classifier, but we look forward to implementing Particle swarm optimization (PSO) in the future which will be a more robust solution to the problem at hand. We also look forward to implementing the following features:

1) Creating an easy to use User Interface for patients to enter their health details and get the result in real-time.

2) Simulate the project using neural networks to get an upper hand in efficiency and complexity.

In the results of the simulation, it was evaluated that this method could only change the set of entries with a limited number of features and improve the efficiency of the classification than all the features used. We want to develop a system for recommending early onset heart disease in the future. In addition, the use of Naïve Bayes for selection of features in data sets with a large number of features can also be studied to realize the different features of naïve Bayes in feature selection. It may also incorporate other techniques of data mining to create accurate and computationally effective classifications for medical applications.

REFERENCES

[1] Dr. Kanak Saxena, Purushottam, Richa Sharma, "Efficient Heart Disease Prediction System", Artificial Intelligence and Signal Processing Conference(AISP), 2016, pp 962-969.

[2] Ashok Kumar Dwivedi, "Evaluate the performance of different machine learning techniques for prediction of heart disease using ten-fold cross-validation" ,Springer, 17 September 2016.

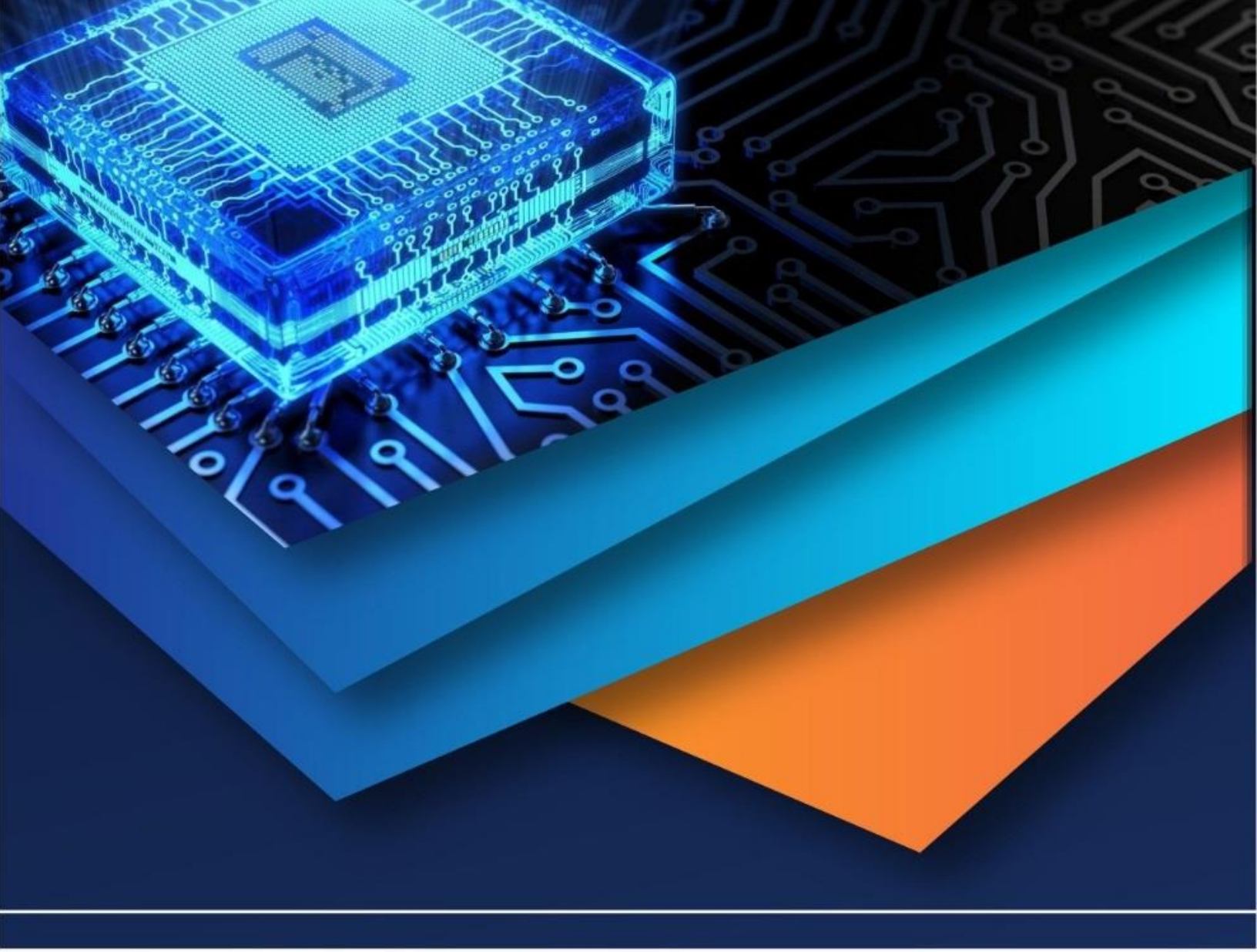
[3] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018.

[4] Amir Al, M.Zain Amin, "An Intuitive Guide of Naïve Bayes Classifier with Practical Implementation in Scikit Learn", Wavy AI Research Foundation.

[5] Ali Haghpanah Jahromi, Mohammad Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features", Artificial Intelligence and Signal Processing Conference (AISP), 2017.

[6] Arundhati Navada, Aamir Nizam Ansari, Siddharth Patil, Balwant A. Sonkamble, "Overview of use of decision tree algorithms in machine learning", IEEE Control and System Graduate Research Colloquium, 2011.

[7] Raj Bhatnagar, Lalit Kumar, "An efficient map-reduce algorithm for computing formal concepts from binary data", IEEE International Conference on Big Data (Big Data), 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY