



Chapter 8

Heart Disease Prediction Using ML Algorithm

Atharva Deshmukh

 <https://orcid.org/0000-0002-8039-3523>
Terna Engineering College, India

Amit Kumar Tyagi

 <https://orcid.org/0000-0003-2657-8700>
Vellore Institute of Technology, India

Sangita Krishnaram Toppo

Terna Engineering College, India

ABSTRACT

Many patients don't get proper treatment due to a shortage of doctors. Thus, predicting a disease using the patient's symptoms has become an important task these days. To solve this there must be a predicting system for predicting diseases. In this chapter, a model is proposed for predicting the disease suffered by a person by knowing the symptoms. The model uses the logistic regression algorithm, which assigns observations to a discrete set of classes and provides a good level of accuracy. It collects the data of a person's symptoms and suggests a suitable disease accordingly. To showcase the accuracy of the proposed model, it has been implemented on a heart disease dataset to predict the occurrence of heart disease. The implementation will illustrate the effectiveness of the proposed model, which can help in the development of an intelligent healthcare system and reduce the cost of treatment.

INTRODUCTION

Overall Description

Numerous patients go untreated or are not treated precisely by the specialists, legitimate treatment is important for the prosperity of the individuals. Consequently, foreseeing a sickness utilizing the patient's manifestations has become a significant undertaking nowadays. There is an absence of specialists in

DOI: 10.4018/978-1-6684-3533-5.ch008

Heart Disease Prediction Using ML Algorithm

India, there is 1 specialist for every 10,198 people in India (WHO suggests the proportion of 1:100). To address this intense deficiency of specialists there should be a foreseeing framework for anticipating the overall sicknesses which would help in the legitimate use of the assets.

The initial phase in treating a patient is the right location of the wellbeing of a person by utilizing the given side effects. The expectation of the illness has become an indispensable errand of late anyway the right forecast of sicknesses has gotten excessively extreme for a specialist. The framework proposed in this paper is intended to build up a sickness forecast framework by using ML. The order inside the forecasting framework is finished with the assistance of the logistic regression algorithm. This may encourage the right expectation of health and furthermore encourage the right treatment of illness.

The fundamental spotlight is on utilizing ML in medical services to enhance understanding consideration for better outcomes. ML has made it simpler to recognize various sicknesses and finding accurately. Prescient investigation with the assistance of effective numerous ML calculations assists with foreseeing the infection all the more accurately and helps treat patients. The medical care industry creates a lot of medical services information day by day that can be utilized to separate data for anticipating sickness that can happen to a patient in the future while utilizing the therapy history and wellbeing information. This shrouded data in the medical care information will be later utilized for full of feeling dynamic for patient's wellbeing. Likewise, this zone needs improvement by utilizing the educational information in medical care.

Information volume is a huge test in any industry however especially in medical care where information will in general sit inert in information bases oversaw by dated EHR(Electronic Health Record) frameworks. Numerous organizations assemble their business on getting enormous volumes of information from these frameworks to make them accessible and significant as they power prescient investigation, choice help, imaging, activity streamlining, and different applications. Different associations utilize protection claims information that has as of late become accessible through state governments. In any case, rules for obtaining entrance for business objects are incipient using complex cycles, so fruitful applications have been rare.

ML in medical services has as of late stood out as truly newsworthy. Google has built up an ML calculation to help distinguish malignant tumors on mammograms. Stanford is utilizing a profound learning calculation to distinguish skin malignant growth. A new JAMA article announced the consequences of a profound ML calculation that had the option to analyze diabetic retinopathy in retinal pictures. Obviously, ML places another bolt in the bunch of clinical dynamics.

All things considered, ML fits a few cycles in a way that is better than others. Calculations can furnish prompt advantage to disciplines with measures that are reproducible or normalized. Likewise, those with enormous picture datasets, for example, radiology, cardiology, and pathology are solid competitors. ML can be prepared to see pictures recognize anomalies, and highlight territories that need consideration, accordingly improving the precision of every one of these cycles. Long haul,

Machine Learning will profit the family expert or internist at the bedside. ML can offer a target assessment to improve effectiveness, unwavering quality, and precision. Human insight can scarcely be contrasted with some other marvel. ML knowledge in medical services has a great deal of potential outcomes to improve the savvy choices made by people. The particular advantages of including ML into medication incorporate precise information can educate experts about normal examples, ML can proceed just as a human does and invalidates pressure and fatigue factors, informational collections can prepare ML calculations and models to address key medication creation issues which would help in relieving more individuals under lower cost and saving the customized approach.

Heart Disease Prediction Using ML Algorithm

Computer-based intelligence utilizes modern calculations to remove, learn, anticipate, and predict from gigantic measures of clinical information while additionally offering proficient help and help. Concerning sicknesses, disease, cardiovascular, and sensory system issues are the most much of the time explored including ML instruments. Self-prepared frameworks can follow administered and unaided getting the hang of, encouraging early discovery and analysis significantly. To perform well, self-prepared frameworks ought to cooperate continually with the clinical examinations information, so clearly, human movement is interconnected with machine learning.

The medical care industry has become an enormous business. The medical services industry delivers a lot of medical care information every day that can be utilized to extricate data for foreseeing illness that can happen to a patient in the future while utilizing the therapy history and wellbeing information. This shrouded data in the medical care information will be later utilized for full of feeling dynamic for patient's wellbeing. Additionally, this zone needs improvement by utilizing the educational information in medical care. The significant test is the means by which to remove the data from this information on the grounds that the sum is huge so some information mining and machine learning strategies can be utilized. Likewise, the normal result and extent of this venture are that if sickness can be anticipated then early therapy can be given to the patients which can decrease the danger of life and save the life of patients, and the cost to get therapy of illnesses can be diminished up somewhat by early acknowledgment. The fast selection of electronic wellbeing records has made an abundance of new information about patients, which is a goldmine for improving the comprehension of human wellbeing.

At the point when machine learning is utilized in guide to enhancing dealing with patients, high outcomes are accomplished. It has made it simpler to spot different sorts of infections and perform analysis precisely. Performing prescient examination with the help of numerous productive ML calculations may encourage foreseeing any infection with extraordinary precision and help to treat patients. The gigantic measure of clinical data containing treatment history and wellbeing information is regularly used to extricate information for anticipating infections that may happen to a patient inside what's to come. The concealed information inside the clinical data is frequently later utilized for a viable dynamic cycle for the patient's wellbeing.

One of the chief crucial utilization of ML is inside the field of medical care. The medical care offices must be constrained to be progressed with the goal that better choices for quiet therapy are regularly made. When ML is utilized in medical care, it causes people to handle immense and complex illness datasets to examine them into accommodating clinical bits of knowledge. At that point, this will be additionally utilized by clinical specialists to give exact treatment to patients. Thus, ML, when upheld in medical care will bring about high patient fulfillment. In this paper, the calculated relapse calculation will be utilized to foresee sicknesses utilizing the patient's therapy history and wellbeing information.

Logistic Regression Algorithm

The logistic regression is likewise alluded to as the sigmoid function that helps the simple portrayal of charts. It also gives high precision. In this calculation, the information ought to be first imported and afterward, it ought to be prepared. It is a kind of regression examination calculation, which is utilized for an expectation of the result of a categorical dependent variable from a bunch of independent or predictor variables.

The vital portrayal in logistic regression is the coefficients, much the same as linear regression. The coefficients in logistic regression are assessed utilizing a cycle called maximum-likelihood estimation.

Heart Disease Prediction Using ML Algorithm

In logistical regression, the variable amount is typically binary. It is chiefly utilized for prediction and furthermore ascertains the probability. Logistic regression is utilized in light of the fact that it is an effective regression predictive analysis algorithm that doesn't dispose of any information and uses all information productively.

Naïve Bayes

Naive Bayes algorithm based on Bayes' theorem with the assumption of independence between every pair of features. Naive Bayes classifiers work well in many real-world situations such as document classification and spam filtering.

Decision Tree

Given a data of attributes together with its classes, a decision tree produces a sequence of rules that can be used to classify the data.

Stochastic Gradient Descent

Stochastic gradient descent is a simple and very efficient approach to fit linear models. It is particularly useful when the number of samples is very large. It supports different loss functions and penalties for classification.

K-Nearest Neighbours

Neighbors-based classification is a type of lazy learning as it does not attempt to construct a general internal model but simply stores instances of the training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point.

Random Forest

Random forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses an average to improve the predictive accuracy of the model and controls over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement.

Support Vector Machine

Support vector machine is a representation of the training data as points in space separated into categories by a clear gap that is as wide as possible. 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.

Heart Disease Prediction Using ML Algorithm

LITERATURE SURVEY

Researchers worked on deep learning and data mining to predict the disease (Bharti and Singh, 2015; Gandhi and Singh, 2015). Their approach was to check the medical history of the patient and observe the new symptoms in the patient and depending on both of them predict the possibility of the patient developing heart disease. Using this model it was believed that the accuracy to predict the heart attack would increase. The model proposed (Babu et al., 2017) uses deep learning algorithms but even after using these models, the accuracy was still very less.

A system was proposed (Chen et al., 2016) for health monitoring using smart clothing. He was able to achieve cost minimization for heterogeneous systems. The disease history of the patient is recorded which helps to give solutions that minimize the cost of medical treatment. But the drawback of the system is that it could predict the disease but not the sub-type of disease. The system proposed (Jabbar, Deekshatulu, and Chandra, 2013) uses ANN(Artificial Neural Network) to classify heart disease. Their model has an accuracy level of 92.8%. The model proposed (Abdullah, 2012) uses Data mining to detect coronary heart disease (CHD). Using feature selection, the number of attributes was reduced. This model has an accuracy of 60.74%.

According to the model proposed (DhfarHamed, Ibrahim, and Naeem, 2017), a Support Vector Machine (SVM) is used to practically implement the theories which were theoretically laid. Shouman, Turner, and Stocker (2012) aims at finding out the death rate by using the ML tool and implementing it on the data set. Kanchan and Kishor (2016) proposed a system that uses a lesser number of characteristics to predict coronary illness by using advanced machine learning calculations. (Jamgade and Zade, 2019) One such application of machine learning algorithms is in the area of healthcare. Machine learning in healthcare aids humans to process big and complex medical datasets and then examining them into clinical intuitions. This can further be used by physicians in giving medical care. (Vinitha et al., 2018) In the proposed structure, provide machine learning algorithms for effectual prediction of various disease occurrences in disease-frequent societies. It experiments with the altered estimate models over real-life hospital data collected.

(Patel and Patel, 2016) This article many types of Data Mining techniques such as classification, clustering, association and also highlights related work to examine and foresee human disease. (Chen et al., 2017), they organize machine learning algorithms for efficacious divination of chronic disease outbreaks in disease-frequent communities. (Sriram et al., 2013) Diagnosis of Parkinson's disease with the help of a machine learning approach provides a better understanding of the PD dataset in the present decennium.

(Jabbar and Samreen, 2016) Hidden Naïve Bayes is a data mining model that lessens the traditional Naïve Bayes conditional independence supposition. This model gives that the Hidden Naïve Bayes (HNB) can be applied to heart disease classification (prediction) (Prasad et al., 2019). In this paper, the logistic regression algorithms are used, and the health care data gives the details of the patients whatever they are having heart diseases or not in accordance with the information in the database record. Also, they will try to use this data as a model which predicts the patient if they are having heart disease or not (Gawande and Barhatte, 2017). In this ECG (electrocardiography) signals are used. ECG is likely to the show condition of the patient and for the diagnosis and treatment of all the types of cardiac diseases.

Wiharto, Kusnanto, and Herianto (2016) presented a paper titled *Intelligence System for Coronary Heart Disease Diagnosis Level using K-Star Algorithm*. They present an expectation framework for heart infection using Learning Vector Quantization neural system computation in this research. This

Heart Disease Prediction Using ML Algorithm

framework's neural system recognizes 13 clinical features as input and predicts the presence or absence of coronary disease in the patient, as well as numerous execution measures.

Prediction for illness similarities by using ID3 algorithm in television and mobile phone was presented by Kumar and Padmapriya (2012) in a paper titled Prediction for disease similarities by using ID3 algorithm in television and mobile phone. This study explains how to deal with identifying designs that are hidden by cardiac sickness in a planned and hidden manner. The offered framework makes use of information mining techniques such as the ID3 algorithm. This proposed strategy not only informs people about illnesses but also has the potential to lower the death rate and the number of disease victims.

Disease Predicting System Using Data Mining Techniques was presented by Banu and Gomathy (2014). MAFIA (Maximal Frequent Itemset Algorithm) and K-Means clustering is discussed in this study. Because categorization is crucial for illness prognosis. The accuracy of the categorization based on MAFIA and K-Means is achieved.

Golande and Pavan Kumar (2019) investigated a variety of machine learning techniques that can be used to classify cardiac disease. A study was conducted to examine the accuracy of Decision Tree, KNN, and K-Means algorithms that may be utilized for classification. This study found that the Decision Tree had the best accuracy and that it may be made more efficient by combining several methodologies and fine-tuning parameters.

Heart disease is a life-threatening condition that should not be taken lightly. Males are more likely than females to get heart disease, according to Harvard Health Publishing (2016), men were rough twice as likely as women to suffer a heart attack over their lives, according to researchers. Even after controlling for established heart disease risk factors such as high cholesterol, high blood pressure, diabetes, body mass index, and physical activity, the greater risk persisted. The researchers are working on this dataset since it contains essential features such as dates from 1998 and is regarded as one of the benchmark datasets for heart disease prediction.

Much research has been conducted, and several machine learning models are employed for the categorization and prediction of heart disease diagnosis. Melillo et al. (2013) developed an automatic classifier for detecting congestive heart failure that distinguishes between patients at high and low risk; they used a machine-learning algorithm known as CART, which stands for Classification and Regression, and achieved a sensitivity of 93.3 percent and specificity of 63.5 percent. Al Rahhal et al. (2016) then propose an electrocardiogram (ECG) strategy for enhancing performance, in which deep neural networks are utilized to choose the best characteristics and then employ them. (Guidi et al., 2014) then add a clinical decision support system for recognizing cardiac failures and avoiding them at an early stage. They attempted to examine several machine learning and deep learning models, particularly neural networks such as the support vector machine, random forest, and CART algorithms. Random forest and CART surpassed everyone else in the categorization with an accuracy of 87.6 percent. Zhang et al. combined natural language processing with a rule-based method. When the NYHA HF class was determined from unstructured clinical records, (Zhang et al., 2017) obtained 93.37 percent accuracy. SVM approaches employed by (Parthiban and Srivatsa, 2012) for recognizing individuals who already have diabetes and subsequently forecasting heart disease produced a 94.60 percent accuracy rate, and the characteristics utilized were common, such as blood sugar level, patient age, and blood pressure data.

Beyene and Kamat (2018) suggested using data mining techniques to predict and analyze the occurrence of heart disease. The major goal is to forecast the emergence of cardiac disease in order to provide an early automated diagnostic of the condition with a quick result. The proposed technique is equally important in a healthcare organization with specialists who lack knowledge and expertise. It analyses

Heart Disease Prediction Using ML Algorithm

many physiological variables such as blood sugar and heart rate, as well as age and gender, to determine whether or not a person has heart disease. WEKA software is used to compute dataset analyses.

Zhang et al. (2012) proposed a Support Vector Machine (SVM) algorithm-based heart attack, prediction model. The imperative characteristics and distinct kernel functions were retrieved using Principal Component Analysis. Radial Basis Function provided the most accurate results. Grid search in SVM was used to find the optimum parameter values, and optimum values were discovered. The highest level of categorization accuracy was around 88.64 percent.

One of the leading causes of mortality worldwide is heart disease. In both men and women, the number of persons afflicted by heart disease rises with age. Other variables that contribute to heart disease include gender, diabetes, and BMI. We attempted to forecast and analyze heart disease in this work by taking into account variables such as age, gender, blood pressure, heart rate, diabetes, and so on. Because there are so many other factors that have a role in heart disease, predicting it will be difficult.

MOTIVATIONS AND SCOPE

At present to stay sound, normal body finding is important. Today, there are different sources accessible as individual forecast or proposal frameworks yet the need of great importance is to have a coordinated model involving both. Likewise, it would be more proper and helpful if individuals could get fundamental determination online 24x7 instead of visiting medical clinics and centers often. Subsequently, lessening cost and saving time. On the off chance that specific abnormalities were found in the analysis, at that point suggestion of close-by subject matter experts and medical clinics as per the client's inclination would encourage fast and suitable therapy. Medical care is an area advancing persistently and producing an enormous measure of information builds up a need to utilize the information for valuable information which pulls in huge associations to put vigorously in this field.

Here the extent of the venture is that coordination of clinical choice help with PC-based patient records could diminish clinical blunders, upgrade quiet security, decline undesirable practice variety, and improve the understanding result. The application grants client to share their heart-associated issues. It at that point measures client explicit subtleties to determine for a differed ailment that may be identified with it. Here we will in general utilize some astute information mining methods to figure the preeminent right disease that may be identified with the patient's subtleties. In view of the result, the framework consequently shows the outcome explicit specialists for greater treatment. The framework grants client to see the specialist's subtleties and can likewise be utilized if there should be an occurrence of crisis.

BACKGROUND

Millions of individuals are affected by heart disease, and it is still the leading cause of mortality worldwide. To lower the effective cost of diagnostic testing, a medical diagnosis should be efficient, reliable, and supported by computer technology. Data mining is a type of software that aids computers in constructing and classifying numerous features. To forecast cardiac disease, this study report uses classification approaches. This part provides an overview of relevant topics such as machine learning and its methodologies, data pre-processing, evaluation metrics, and a description of the dataset utilized in this study.

Heart Disease Prediction Using ML Algorithm

Machine learning is a new branch of artificial intelligence that is gaining popularity. Its main goal is to create systems that can learn and make predictions based on previous experiences. It creates a model by training machine learning algorithms with a training dataset. The model predicts heart disease using the new input data. It builds models by detecting hidden patterns in the input dataset using machine learning. For fresh datasets, it makes accurate predictions. The dataset has been cleaned, and any missing values have been filled in. The model is then evaluated for accuracy using the new input data to forecast heart disease.

PURPOSE

Illness forecast utilizing understanding treatment history and wellbeing information by applying data mining and ML procedures is progressing battle for as long as many years. Numerous works have been applied data mining strategies to obsessive information or clinical profiles for an expectation of explicit illnesses. These methodologies attempted to anticipate the reoccurrence of the illness. Additionally, a few methodologies attempt to do expectation on control and movement of sickness. The new achievement of profound learning in unique territories of ML has driven a move towards ML models that can learn rich, various leveled portrayals of crude information with minimal pre-preparing and produce more precise outcomes. Quantities of papers have been distributed on a few information-digging procedures for analysis of coronary illness, for example, Decision Tree, Naive Bayes, neural organization, part thickness, consequently characterized gatherings, sacking calculation, and backing vector machine indicating various degrees of exactness in infections forecast.

EXISTING SYSTEM

In the current framework, useful utilization of different gathered information is tedious, machine can anticipate illnesses yet can't foresee the sub sorts of infections brought about by the event of one sickness. It neglects to foresee all potential states of the individuals. The existing framework handles just organized information. A machine can distinguish an illness yet can't expect the sub sorts of the infections and sicknesses brought about by the presence of one bug. The expectations of illnesses have been vague and inconclusive. For the event, if a gathering of individuals are predicted with Diabetes, without a doubt some of them may have complex danger for Heart infections because of the reality of Diabetes. Determination of the condition exclusively relies on the specialist's instinct and the patient's records. Discovery is absurd at a previous stage that may later possibly hurt the patient.

PROPOSED SYSTEM

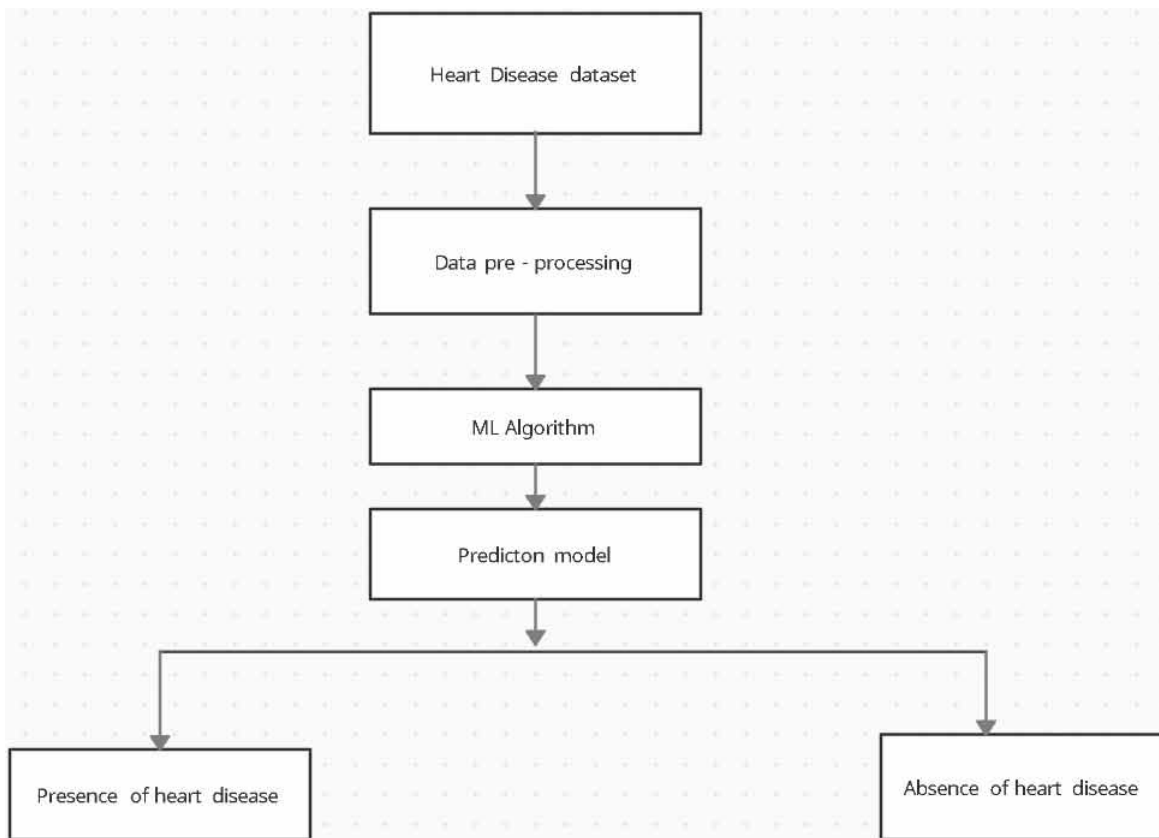
The proposed framework has been created to order individuals, who are blasted by illness and solid individuals. The presentation of the prescient model with chosen highlights is tried to anticipate the probabilities of experiencing coronary illness. Highlight determination calculation was utilized to choose significant highlights, and on these chosen highlights, the exhibition of the classifiers was tried. The Framingham heart condition dataset is taken from Kaggle and has been utilized in our examination.

Heart Disease Prediction Using ML Algorithm

The mainstream ML classifier logistic regression is utilized inside the framework. The model's approval and execution assessment measurements are processed. It is adaptable and can be generally utilized for different sicknesses with high paces of accomplishment. The strategy of the proposed framework is organized into five phases which include:

1. pre-processing of a dataset
2. feature selection
3. cross-validation method
4. machine learning classifiers
5. classifiers' performance evaluation methods.

Figure 1. Disease predicting model framework for predicting heart disease



IMPLEMENTATION

The proposed framework is actualized on a coronary illness dataset, which is accessible on the Kaggle site and it contains a cardiovascular report on inhabitants of the town of Framingham, Massachusetts. The grouping objective is to anticipate whether the patient has a danger of future heart disease(HD) in

Heart Disease Prediction Using ML Algorithm

10 years or not. The dataset gives the patients' data like segment, social and clinical danger factors. The dataset contains more than 4,000 records and around 15 credits.

Figure 2. An imported dataset from Kaggle

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	male	age	education	CurrentSm	CigsPerDay	BPMeds	PrevalentS	PrevalentD	Diabetes	TotChol	SysBP	DiaBP	BMI	HeartRate	Glucose	TenYearHD	
2	1	39	4	0	0	0	0	0	0	195	106	70	26.97	80	77	0	
3	0	46	2	0	0	0	0	0	0	250	121	81	28.73	95	76	0	
4	1	48	1	1	20	0	0	0	0	245	127.5	80	25.34	75	70	0	
5	0	61	3	1	30	0	0	1	0	225	150	95	28.58	65	103	1	
6	0	46	3	1	23	0	0	0	0	285	130	84	23.1	85	85	0	
7	0	43	2	0	0	0	0	1	0	228	180	110	30.3	77	99	0	
8	0	63	1	0	0	0	0	0	0	205	138	71	33.11	60	85	1	
9	0	45	2	1	20	0	0	0	0	313	100	71	21.68	79	78	0	
10	1	52	1	0	0	0	0	1	0	260	141.5	89	26.36	76	79	0	
11	1	43	1	1	30	0	0	1	0	225	162	107	23.61	93	88	0	
12	0	50	1	0	0	0	0	0	0	254	133	76	22.91	75	76	0	
13	0	43	2	0	0	0	0	0	0	247	131	88	27.64	72	61	0	
14	1	46	1	1	15	0	0	1	0	294	142	94	26.31	98	64	0	
15	0	41	3	0	0	1	0	1	0	332	124	88	31.31	65	84	0	
16	0	39	2	1	9	0	0	0	0	226	114	64	22.35	85	NA	0	
17	0	38	2	1	20	0	0	1	0	221	140	90	21.35	95	70	1	
18	1	48	3	1	10	0	0	1	0	232	138	90	22.37	64	72	0	
19	0	46	2	1	20	0	0	0	0	291	112	78	23.38	80	89	1	
20	0	38	2	1	5	0	0	0	0	195	122	84.5	23.24	75	78	0	
21	1	41	2	0	0	0	0	0	0	195	139	88	26.88	85	65	0	
22	0	42	2	1	30	0	0	0	0	190	108	70.5	21.59	72	85	0	
23	0	43	1	0	0	0	0	0	0	185	123.5	77.5	29.89	70	NA	0	
24	0	52	1	0	0	0	0	0	0	234	148	78	34.17	70	113	0	
25	0	52	3	1	20	0	0	0	0	215	132	82	25.11	71	75	0	
26	1	44	2	1	30	0	0	1	0	270	137.5	90	21.96	75	83	0	
27	1	47	4	1	20	0	0	0	0	294	102	68	24.18	62	66	1	
28	0	60	1	0	0	0	0	0	0	260	110	72.5	26.59	65	NA	0	
29	1	35	2	1	20	0	0	1	0	225	132	91	26.09	73	83	0	
30	0	61	3	0	0	0	0	1	0	272	182	121	32.8	85	65	1	
31	0	60	1	0	0	0	0	0	0	247	130	88	30.36	72	74	0	
32	1	36	4	1	35	0	0	0	0	295	102	68	28.15	60	63	0	
33	1	43	4	1	43	0	0	0	0	226	115	85.5	27.57	75	75	0	
34	0	59	1	0	0	0	0	1	0	209	150	85	20.77	90	88	1	

At that point, we settle on a decision for picking the variable amount and variable amount, where each quality is considered as a potential danger factor. There are a few segment, social and clinical danger factors included.

Demographic

sex: male or female(Nominal)

age: age of the patient(Continuous)

Behavioral

CurrentSmoker: whether or not the patient may be a current smoker (Nominal)

CigsPerDay: the number of cigarettes that the person smoked on average in one day(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

Heart Disease Prediction Using ML Algorithm

Medical (History)

BPMeds: whether or not the patient was on vital sign medication (Nominal)

PrevalentStroke: whether or not the patient had a stroke before (Nominal)

PrevalentHyp: whether the patient was hypertensive or not (Nominal)

Diabetes: whether the patient had diabetes or not (Nominal)

Medical (Current)

TotChol: Total Cholesterol level (Continuous)

SysBP: Systolic Blood Pressure (Continuous)

DiaBP: Diastolic Blood Pressure (Continuous)

BMI: Body Mass Index (Continuous)

HeartRate: pulse rate (Continuous - In medical research, variables like pulse rate though after all discrete, yet are considered continuous thanks to an outsized number of possible values.)

Glucose: Glucose level (Continuous)

Predict variable (desired target): risk of future heart disease(HD) in 10 years (binary: “1” means “Yes” and “0” means “No”)

Figure 3. Number of people with a history of heart disease

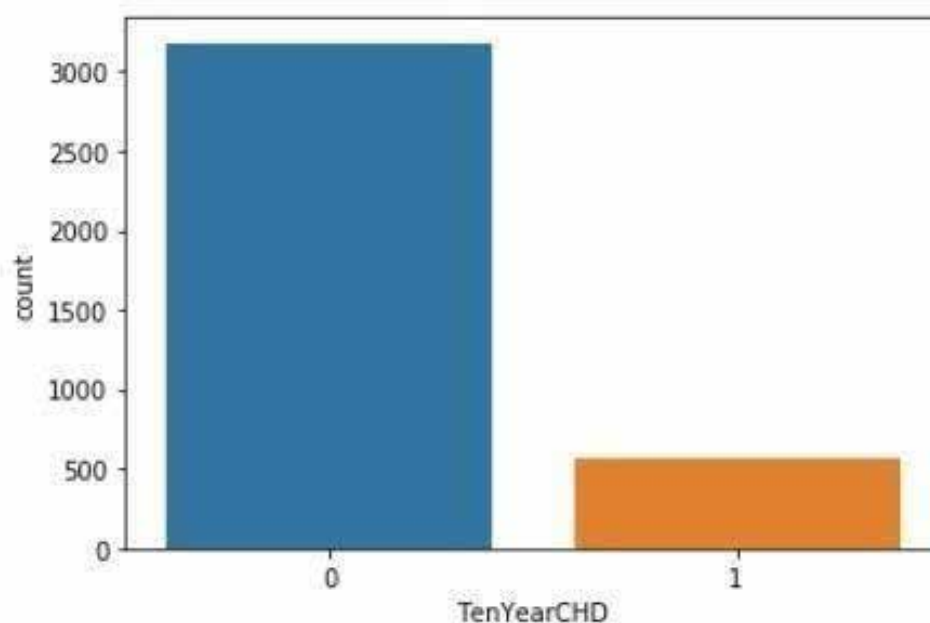


Figure 3 shows the clinical record of 4,000 individuals out of which around 3,500 individuals have not experienced cardiovascular sickness before though 5,000 individuals have experienced cardiovascular infection.

Heart Disease Prediction Using ML Algorithm

Figure 4. independent and dependent variables part-1

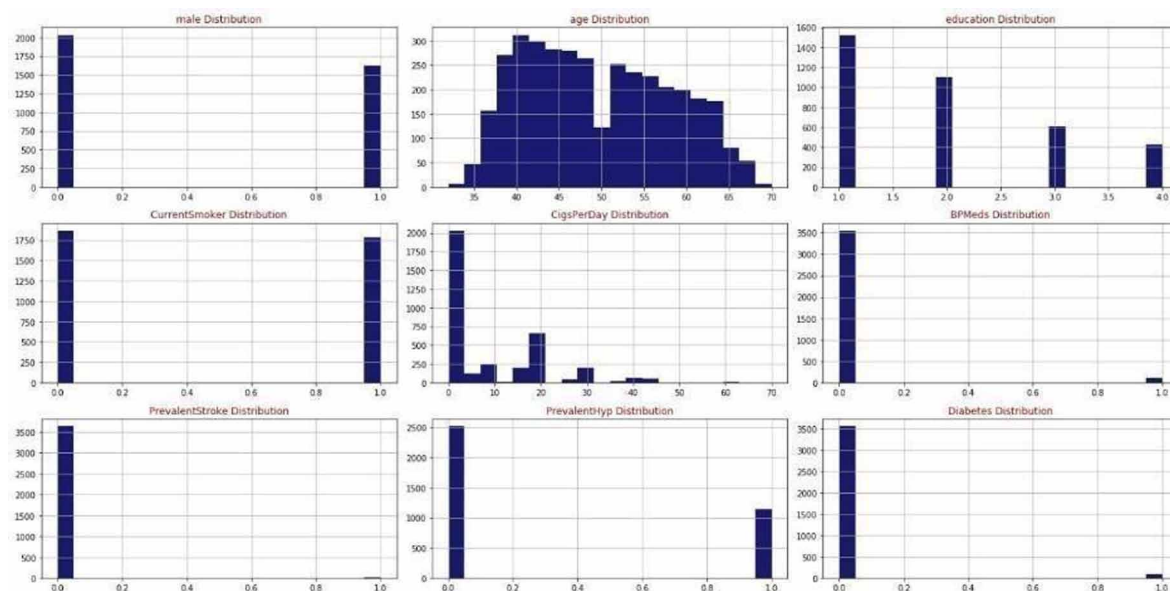
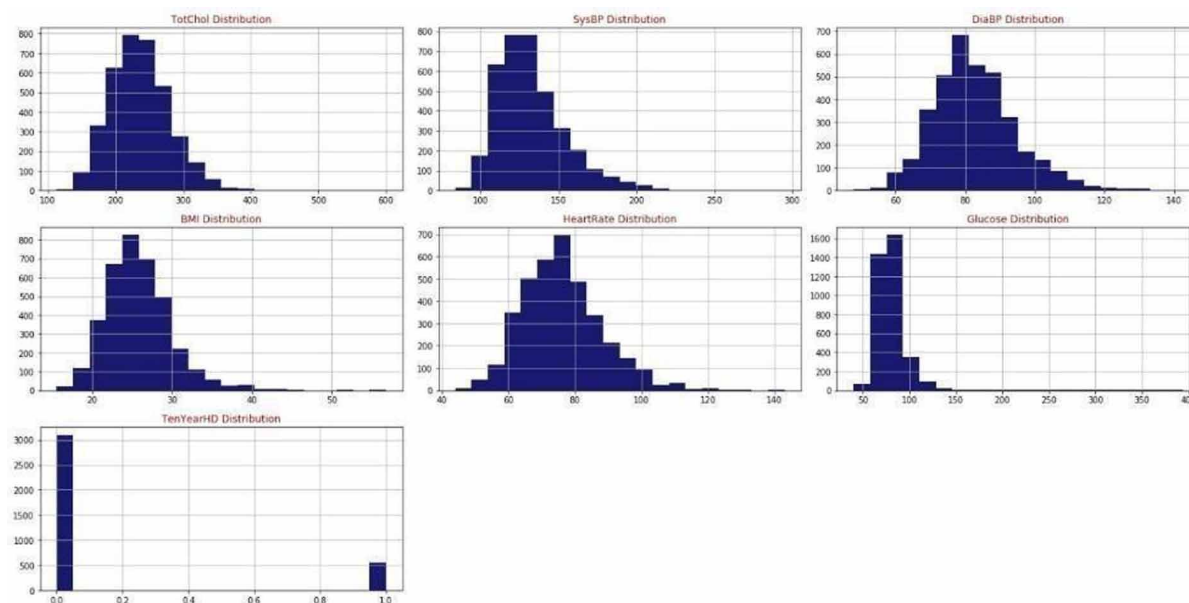


Figure 5. independent and dependent variables part-2



The coronary illness dataset is then part into two subsets, for example, preparing information and testing information and that we fit our model on train information to frame expectations on the test information After that two things can wind up occurring, we may overfit our model or we may underfit

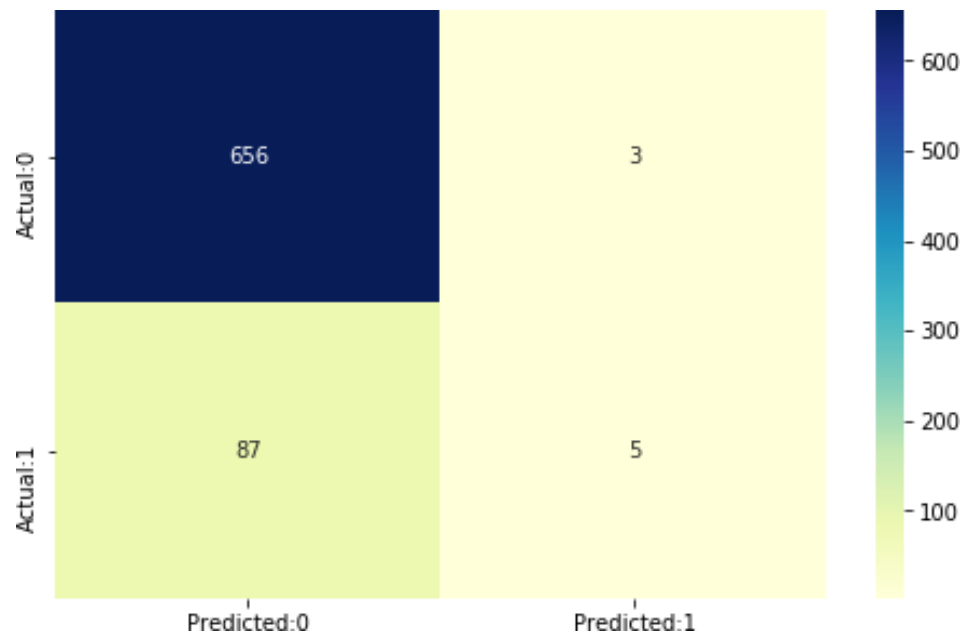
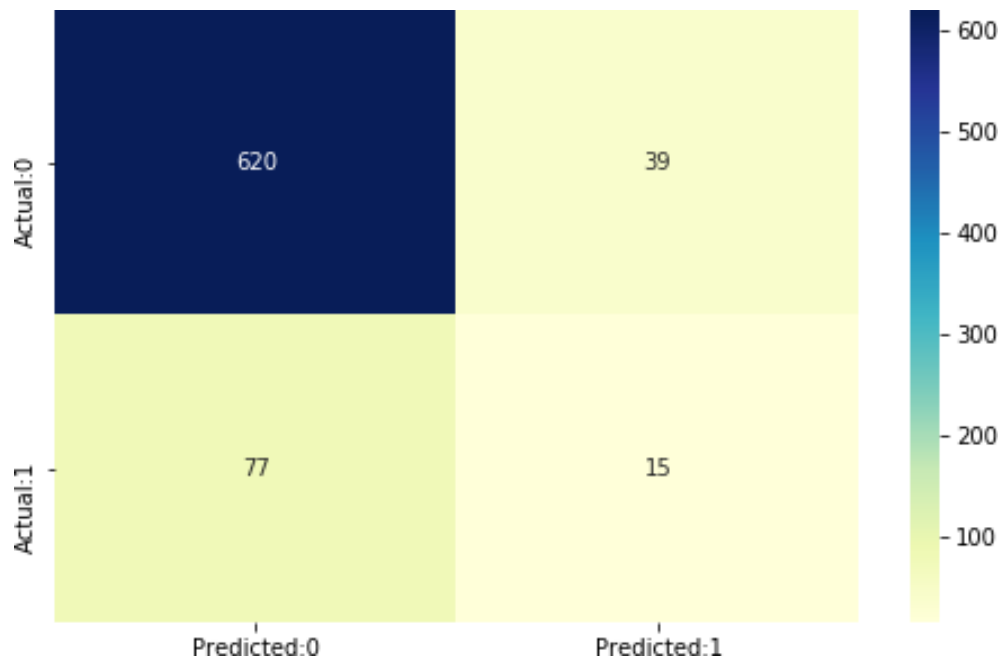
Heart Disease Prediction Using ML Algorithm

our model. Any of those things happening would influence the consistency of our model, so we may end up utilizing a model with lower exactness. For the coronary illness dataset, the training set is taken as 80% of the genuine information and the test set as 20% of the information.

Figure 6. Logistic regression results using backward elimination (P-value approach)

Logit Regression Results						
Dep. Variable:	TenYearHD	No. Observations:	3751			
Model:	Logit	Df Residuals:	3744			
Method:	MLE	Df Model:	6			
Date:	Sat, 30 May 2020	Pseudo R-squ.:	0.1149			
Time:	23:18:09	Log-Likelihood:	-1417.7			
converged:	True	LL-Null:	-1601.7			
Covariance Type:	nonrobust	LLR p-value:	2.127e-76			
	coef	std err	z	P> z	[0.025	0.975]
const	-9.1264	0.468	-19.504	0.000	-10.043	-8.209
sex_male	0.5815	0.105	5.524	0.000	0.375	0.788
age	0.0655	0.006	10.343	0.000	0.053	0.078
CigsPerDay	0.0197	0.004	4.805	0.000	0.012	0.028
TotChol	0.0023	0.001	2.106	0.035	0.000	0.004
SysBP	0.0174	0.002	8.162	0.000	0.013	0.022
Glucose	0.0076	0.002	4.574	0.000	0.004	0.011

The results show that there are some attributes with a P-value higher than the favored alpha (5%) and subsequently indicate a low genuinely critical relationship with the likelihood of coronary illness. We use a backward elimination way to deal with and eliminate those ascribes with the highest P-value each in turn, at that point the regression is run over and again until all attributes have P Values under 0.05. A property having P-value under 0.05 shows that the adjustment in its value will cause a change in the chances of having a coronary illness.

Heart Disease Prediction Using ML Algorithm*Figure 7. Confusion matrix for model evaluation using Logistic Regression**Figure 8. Confusion matrix for model evaluation using Naïve Bayes*

Heart Disease Prediction Using ML Algorithm

Figure 9. Confusion matrix for model evaluation using Decision Tree Classifier

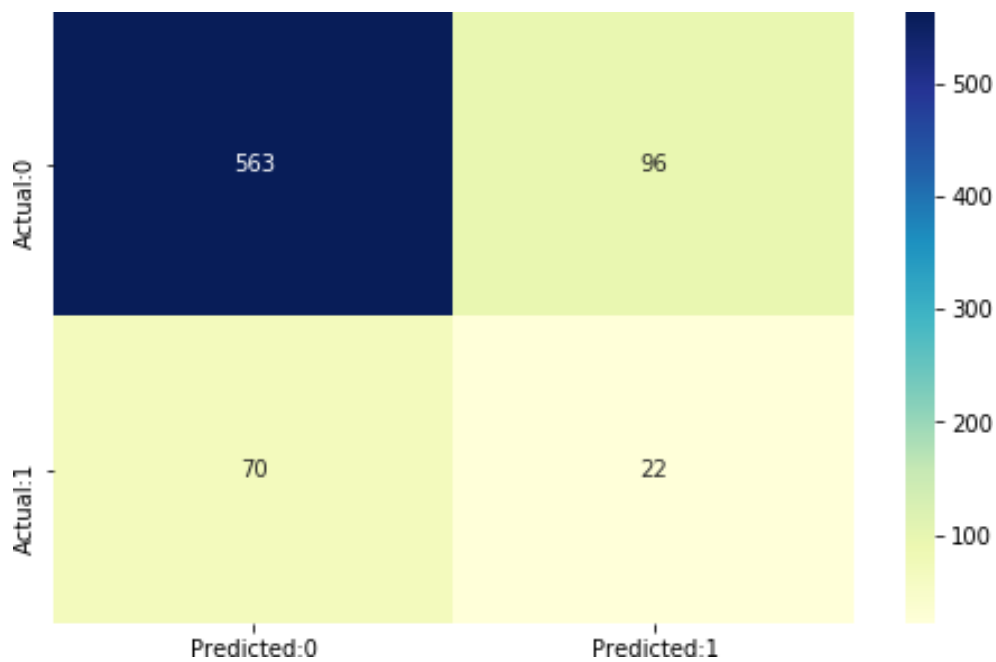
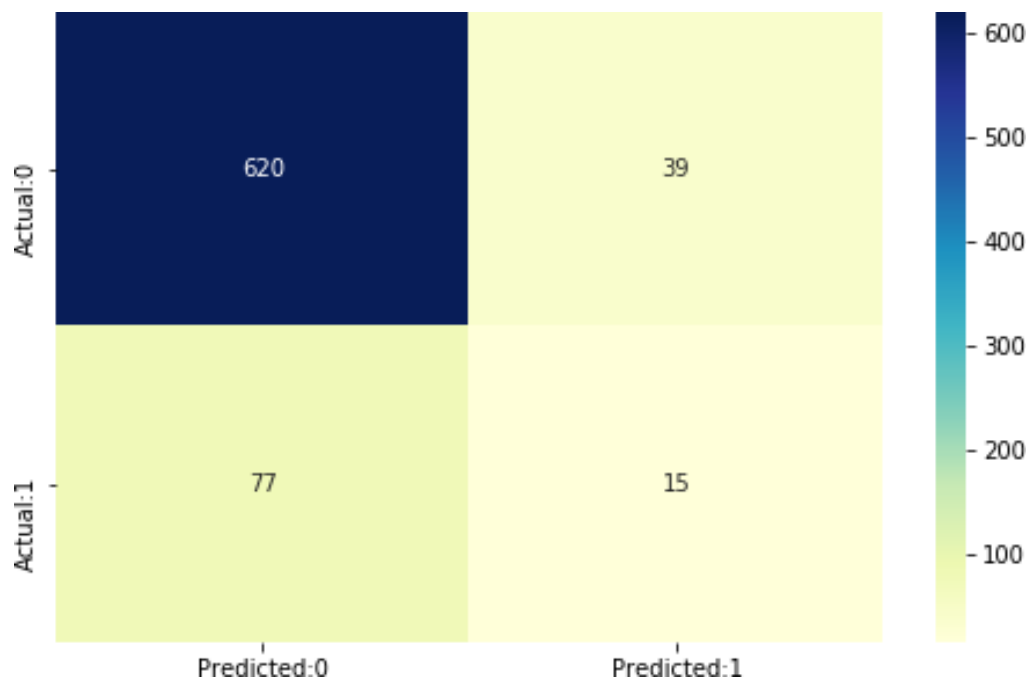
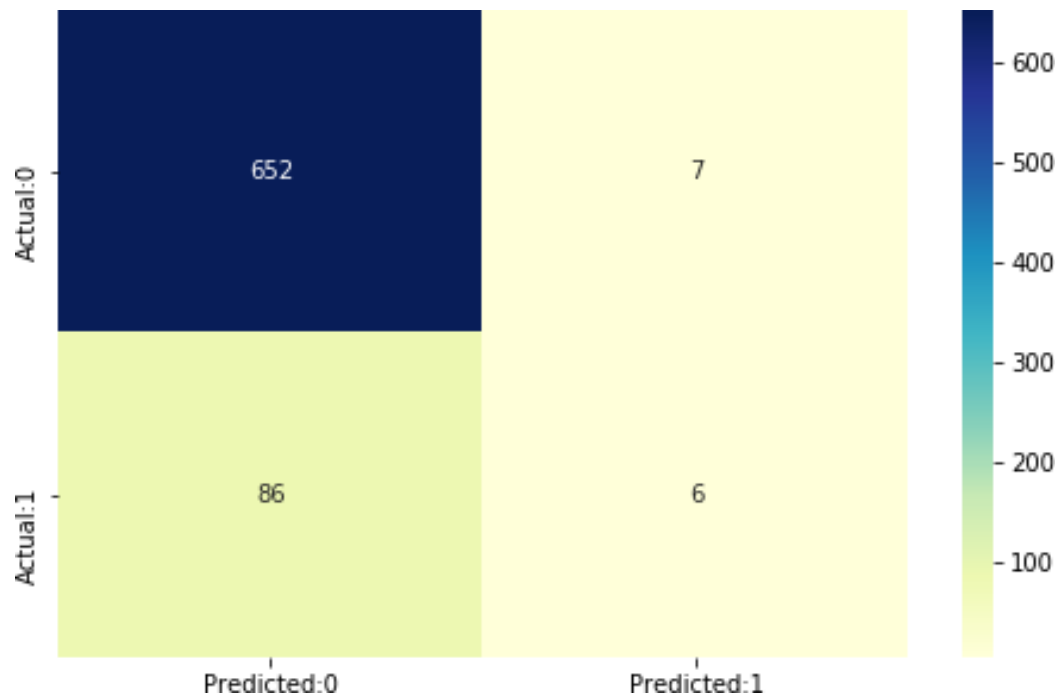
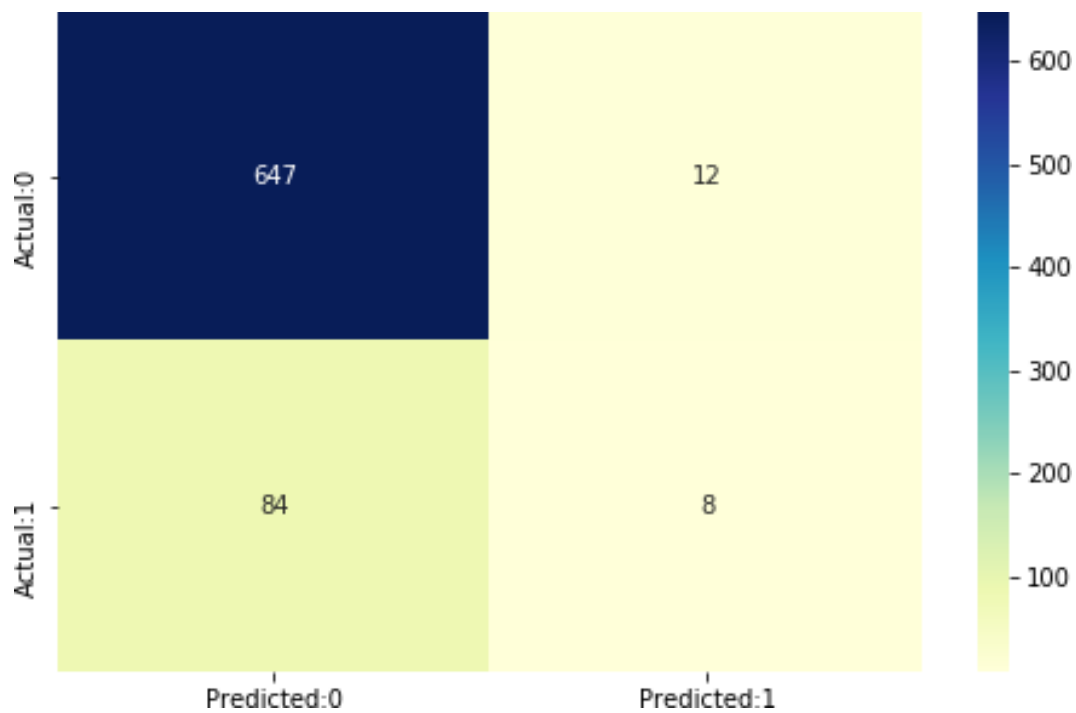


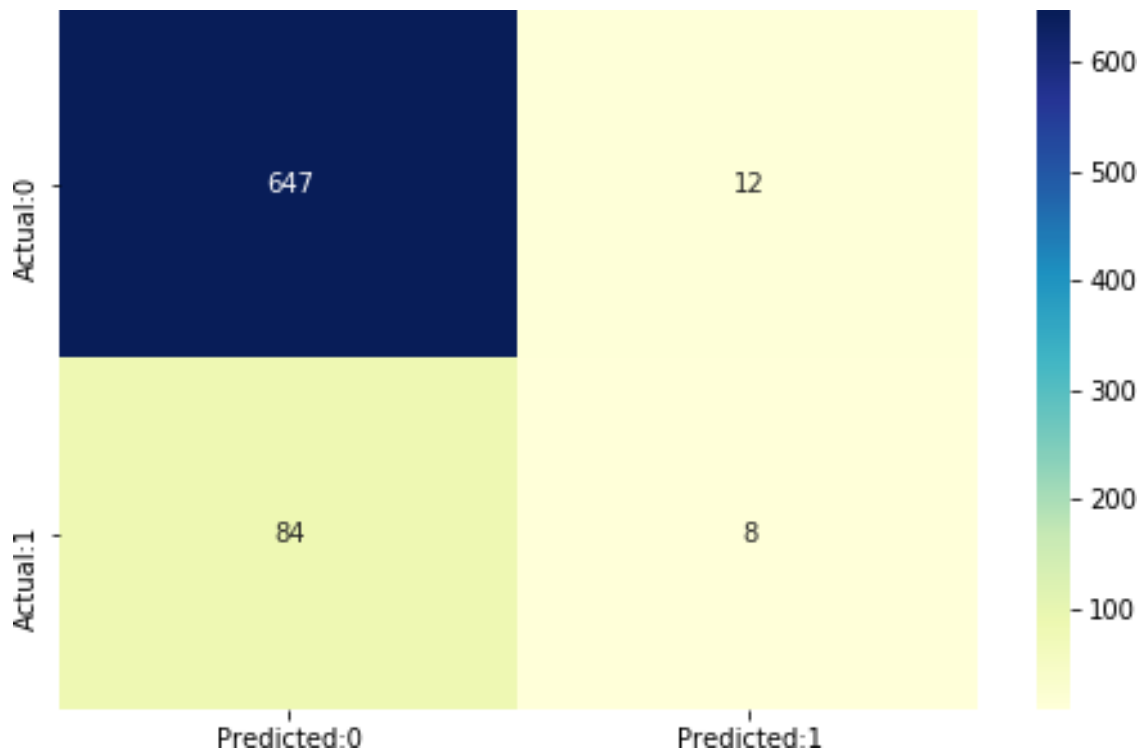
Figure 10. Confusion matrix for model evaluation using Stochastic Gradient Descent Classifier



Heart Disease Prediction Using ML Algorithm*Figure 11. Confusion matrix for model evaluation using KNN**Figure 12. Confusion matrix for model evaluation using Random Forest Classifier*

Heart Disease Prediction Using ML Algorithm

Figure 13. Confusion matrix for model evaluation using Support Vector Machine



Algorithm:

- Step 1: Import all the important and relevant libraries
- Step 2: Import the dataset
- Step 3: Remove the rows and columns that are irrelevant
- Step 4: Remove all the null values from the dataset
- Step 5: Perform data exploratory analysis of clean data
- Step 6: Make TenYear CHD dependent variable and others independent variable
- Step 7: Eliminate the features that are irrelevant using backward elimination (P value) approach
- Step 8: Split data into 80:20 and train the model using

- 8.1: Logistic Regression
- 8.2: Naïve Bayes
- 8.3: Decision Tree Classifier
- 8.4: Stochastic Gradient Descent
- 8.5: K-nearest neighbor
- 8.6: Random Forest classifier
- 8.7: Support vector machine

Step 9: Evaluate the correctness of the model and using confusion matrix calculate the amount of correct and incorrect prediction Step10: Stop

Heart Disease Prediction Using ML Algorithm

The confusion matrix shows that the model evaluation using Logistic Regression made 661 right expectations which are the highest of all the algorithms tested and 90 inaccurate ones. From the above insights, it very well clarified that the model is profoundly specific than sensitive. The negative values are anticipated more precisely than the positives. Likewise, the model accomplishes a precision pace of 88%.

Table 1. Accuracy of different algorithms

Algorithms	Accuracy
Logistic Regression	0.8801597869507324
Naive Bayes	0.8455392809587217
Decision Tree Classifier	0.7789613848202397
Stochastic Gradient Descent Classifier	0.8455392809587217
KNN	0.8761651131824234
Random Forest Classifier	0.8721704394141145
Support Vector Machine	0.8721704394141145

OUTPUT

Figure 14. Effect in odds of heart disease

	CI 95%(2.5%)	CI 95%(97.5%)	Odds Ratio	pvalue
const	0.000043	0.000272	0.000109	0.000
sex_male	1.455242	2.198536	1.788687	0.000
age	1.054483	1.080969	1.067644	0.000
CigsPerDay	1.011733	1.028128	1.019897	0.000
TotChol	1.000158	1.004394	1.002273	0.035
SysBP	1.013292	1.021784	1.017529	0.000
Glucose	1.004346	1.010898	1.007617	0.000

The fitted model shows that holding any remaining highlights steady, the chances of experiencing cardiovascular illness for guys are 78.8% higher than the chances for females. The coefficient for age says that holding all others consistent, we will see a 6.76% expansion in the chances of experiencing cardiovascular sickness with one year increment in age. Additionally, with each additional cigarette one smokes there is a 2% expansion in the chances of getting the cardiovascular illness. For Total cholesterol level and glucose level, there is no critical change. Likewise, there is a 1.7% expansion in chances for each unit increment in systolic Blood Pressure.

Heart Disease Prediction Using ML Algorithm

FUTURE SCOPE

In the future, an intelligent system that can guide the selection of appropriate treatment approaches for a patient diagnosed with heart disease may be developed. There has already been a lot of effort put into developing models that can predict whether a patient is going to acquire heart disease or not. Once a patient is diagnosed with a certain type of heart disease, he or she can choose from a number of therapy options. By extracting knowledge from such appropriate databases, data mining may be of great assistance in determining the course of therapy to be taken.

Additionally, the computing time was lowered, which is beneficial for deploying a model. It was also discovered that the dataset should be normalized; otherwise, the training model becomes overfitted and the accuracy gained is insufficient when a model is assessed for real-world data issues that differ much from the dataset on which the model was trained. It was also discovered that statistical analysis is crucial when analyzing a dataset, and it should have a Gaussian distribution, and then outlier detection is vital, and a technique called Isolation Forest is used to handle this. The challenge that arose here is that the dataset's sample size is not large (Pramod et al., 2021). When a huge dataset is available, the outcomes of deep learning and machine learning can improve significantly. When compared to other studies, the algorithm we used in ANN architecture boosted accuracy. The dataset size may be raised, and deep learning with many additional improvements can be employed to produce more promising outcomes. Machine learning and numerous other optimization approaches can also be applied to improve the evaluation findings. Different methods for normalizing data may be utilized, and the results can be compared. More techniques to integrate heart-disease-trained ML and DL models with specific multimedia for the benefit of patients and clinicians should be discovered.

In today's society, the majority of data is digital, dispersed, and underutilized. We may also use the given data to look for unknown patterns by analyzing them. The fundamental goal of this research is to predict cardiac problems with great accuracy. In machine learning, we may utilize logistic regression, naive Bayes, and sklearn to predict cardiac disease. The paper's future scope is the prediction of cardiac illnesses utilizing new approaches and algorithms with reduced time complexity.

The machine learning model will, at some time in the future, employ a bigger training dataset, maybe more than a million individual data points stored in an electronic health record system. Although it would be a big jump in terms of computer power and software sophistication, an artificial intelligence-based system might allow a medical practitioner to choose the appropriate therapy for a patient as quickly as feasible. A software API can be created to allow health websites and applications to give free access to patients. The probability forecast would be made with no or very little processing time.

CONCLUSION

It has been presumed that men are more inclined to coronary illness than ladies. An expansion in age, alongside the number of cigarettes, smoked every day and systolic pulse likewise show expanded chances of getting a cardiovascular infection. Absolute cholesterol shows no huge change in the chances of Heart Disease (HD), this could be because of the presence of good cholesterol in the cholesterol perusing. Essentially, glucose also causes a truly immaterial change in chances (0.2%). The logistic regression approach was used to create a machine learning model and build a heart disease risk prediction for people at risk of future heart disease. The ratio of a total number of correct predictions to a total number of expected

Heart Disease Prediction Using ML Algorithm

outputs was used to calculate accuracy. It can be written as $(TP+FN)/(TP+TN+FP+FN)$ Where, TP = True positive, TN = True negative, FN = False negative, FP = False positive. The future improvement of the proposed framework will bring about forecast illnesses by utilizing progressed procedures and calculations in less time unpredictability. An insightful framework might be created utilizing the proposed model that can prompt the choice of legitimate treatment techniques. Information investigation and Machine learning can be of awesome assistance in choosing the line of treatment to be trailed by removing information from such appropriate information bases.

REFERENCES

- Abdullah, A. S. (2012). A data mining model to predict and analyze the events related to coronary heart disease using decision trees with particle swarm optimization for feature selection. *International Journal of Computers and Applications*, 55(8).
- Al Rahhal, M. M., Bazi, Y., AlHichri, H., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345, 340–354.
- Babu, S., Vivek, E. M., Famina, K. P., Fida, K., Aswathi, P., Shanid, M., & Hena, M. (2017, April). Heart disease diagnosis using data mining technique. In 2017 international conference of electronics, communication and aerospace technology (ICECA) (Vol. 1, pp. 750-753). IEEE. doi:10.1109/ICECA.2017.8203643
- Banu, M. N., & Gomathy, B. (2014, March). Disease forecasting system using data mining methods. In *2014 International conference on intelligent computing applications* (pp. 130-133). IEEE.
- Beyene, C., & Kamat, P. (2018). Survey on prediction and analysis the occurrence of heart disease using data mining techniques. *International Journal of Pure and Applied Mathematics*, 118(8), 165–174.
- Bharti, S., & Singh, S. N. (2015, May). Analytical study of heart disease prediction comparing with different algorithms. In *International Conference on Computing, Communication & Automation* (pp. 78-82). IEEE. 10.1109/CCAA.2015.7148347
- Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *IEEE Access: Practical Innovations, Open Solutions*, 5, 8869–8879.
- DhafarHamed, J. K. A., Ibrahim, M., & Naeem, M. B. (2017, March). The Utilisation of Machine Learning Approaches for Medical Data Classification. *Annual conference on new trends in information & communications technology applications*.
- Gandhi, M., & Singh, S. N. (2015, February). Predictions in heart disease using techniques of data mining. In *2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE)* (pp. 520-525). IEEE. 10.1109/ABLAZE.2015.7154917
- Gawande, N., & Barhatte, A. (2017, October). Heart diseases classification using convolutional neural network. In *2017 2nd International Conference on Communication and Electronics Systems (ICCES)* (pp. 17-20). IEEE.

Heart Disease Prediction Using ML Algorithm

Golande, A., & Pavan Kumar, T. (2019). Heart disease prediction using effective machine learning techniques. *International Journal of Recent Technology and Engineering*, 8(1), 944–950.

Guidi, G., Pettenati, M. C., Melillo, P., & Iadanza, E. (2014). A machine learning system to improve heart failure patient assistance. *IEEE Journal of Biomedical and Health Informatics*, 18(6), 1750–1756.

Harvard Health Publishing. (2016, November 8). *Throughout life, heart attacks are twice as common in men than women*. Harvard Medical School. Retrieved from <https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women#:~:text=Researchers%20found%20that%20throughout%20life,mass%20index%2C%20and%20physical%20activity>

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using artificial neural network and feature subset selection. *Global Journal of Computer Science and Technology Neural & Artificial Intelligence*, 13(3), 4–8.

Jabbar, M. A., & Samreen, S. (2016, October). Heart disease prediction system based on hidden naïve bayes classifier. In *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)* (pp. 1-5). IEEE.

Jamgade, A.C., & Zade, S.D. (2019). Disease Prediction Using Machine Learning. *International Research Journal of Engineering and Technology*, 5(6).

Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In *2016 international conference on global trends in signal processing, information computing and communication (ICGTSPICC)* (pp. 5-10). IEEE.

Kumar, L. S., & Padmapriya, A. (2012). Prediction for Common Disease using ID3 Algorithm in Mobile Phone and Television. *International Journal of Computers and Applications*, 975, 8887.

Melillo, P., De Luca, N., Bracale, M., & Pecchia, L. (2013). Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability. *IEEE Journal of Biomedical and Health Informatics*, 17(3), 727–733.

Mishra, S., & Tyagi, A. K. (2022). The Role of Machine Learning Techniques in Internet of Things-Based Cloud Applications. In S. Pal, D. De, & R. Buyya (Eds.), *Artificial Intelligence-based Internet of Things Systems. Internet of Things (Technology, Communications and Computing)*. Springer., doi:10.1007/978-3-030-87059-1_4

Parthiban, G., & Srivatsa, S. K. (2012). Applying machine learning methods in diagnosing heart disease for diabetic patients. *International Journal of Applied Information Systems*, 3(7), 25–30.

Pramod, A., Naicker, H. S., & Tyagi, A. K. (2021). Machine learning and deep learning: Open issues and future research directions for the next 10 years. *Computational analysis and deep learning for medical care: Principles, methods, and applications*, 463-490.

Prasad, R., Anjali, P., Adil, S., & Deepa, N. (2019). Heart disease prediction using logistic regression algorithm using machine learning. *International Journal of Engineering and Advanced Technology*, 8(3S), 659–662.

Heart Disease Prediction Using ML Algorithm

Shouman, M., Turner, T., & Stocker, R. (2012). Applying k-nearest neighbour in diagnosing heart disease patients. *International Journal of Information and Education Technology (IJJET)*, 2(3), 220–223.

Sriram, T. V., Rao, M. V., Narayana, G. S., Kaladhar, D. S. V. G. K., & Vital, T. P. R. (2013). Intelligent Parkinson disease prediction using machine learning algorithms. *Int. J. Eng. Innov. Technol*, 3, 212–215.

Tyagi, A. K., Aswathy, S. U., Aghila, G., & Sreenath, N. (2021, October). AARIN: Affordable, Accurate, Reliable and INnovative Mechanism to Protect a Medical Cyber-Physical System using Blockchain Technology. *IJIN*, 2, 175–183.

Tyagi, A. K., & Nair, M. M. (2021, July). Deep Learning for Clinical and Health Informatics. In *Computational Analysis and Deep Learning for Medical Care: Principles, Methods, and Applications*. DOI: <https://doi.org/doi:10.1002/9781119785750.ch5>

Wiharto, W., Kusnanto, H., & Herianto, H. (2016). Intelligence system for diagnosis level of coronary heart disease with K-star algorithm. *Healthcare Informatics Research*, 22(1), 30–38.

Zhang, R., Ma, S., Shanahan, L., Munroe, J., Horn, S., & Speedie, S. (2017, November). *Automatic methods to extract New York heart association classification from clinical notes*. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE.

Zhang, Y., Liu, F., Zhao, Z., Li, D., Zhou, X., & Wang, J. (2012, June). Studies on application of Support Vector Machine in diagnose of coronary heart disease. In *2012 Sixth International Conference on Electromagnetic Field Problems and Applications* (pp. 1-4). IEEE.