ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4

Journal of Coastal Life Medicine

Heart Disease Prediction Using Logistic Regression

Received: 24 October 2022, **Revised:** 28 November 2022, **Accepted:** 30 December 2022

Kavya S M, PrathanyaSree C, Deepasindhu M, Nowshika B, Shijitha R,
Department of Biomedical Instrumentation Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamilnadu

Keywords
Machine Learning, Logistic regression, Framingham dataset, heart diseases.

Abstract
Myocardial Infarction and Brain attacks are responsible for the fatalities of individuals from cardiovascular diseases (CVDs), and especially the deaths occur before age 70. 17.9 million people are thought to pass away from CVDs annually. Accurate monitoring for each patient individually is not always possible, and clinicians cannot consult with patients every 24 hours due to the additional time and knowledge required. Using the patient's various cardiac characteristics and the machine learning approach of logistic regression on a publicly accessible dataset from Kaggle, we developed and examined models for predicting heart disease in this research. The main objective is to ascertain of acquiring coronary heart disease (CHD) upto 10 years of health risk. More than 4,000 records, 15 attributes, and patient data are included in the collection. To forecast outcomes, it makes predictions about a dependent variable based on one or more sets of independent variables. Both binary classification and multi-class classification can use it. This study aims to establish the most significant heart disease risk factors and estimate the overall risk using logistic regression.

1. Introduction

In the modern world, cardiovascular disease, often known as heart disease, is a severe sickness spreading globally. For the past several years, the prevalence of heart disease has been rising quickly in our daily lives. Smoking, having high blood pressure, and having high cholesterol are three major risk factors for heart disease. “Women” typically exhibit distinct heart disease symptoms than men do, particularly when it comes to coronary artery disease (CAD) and other cardiovascular issues. A set of illnesses that impact the circulatory system are referred to as heart disease. There are a unique set of causes for each form of heart disease. The main behavioral risk factors for myocardial infarction and brainattacksinclude poor eating habits, inactivity, tobacco use, and alcohol dependence. People may develop obesity, high blood lipids, high blood sugar, and high blood pressure as a result of behavioral risk factors. According to the primary care clinics, measurements of these "intermediate risk variables" include a higher probability of experiencing a myocardial infarction and “brain attacks”, congestive “heart failure”, or other problems.

We have to improve the performance of previous work architecture in this project using a preprocessing method that includes a normalization phase that fills in missing data with the mean value of each feature. A critical step in the machine learning process, data preparation improves the quality of the input data and extracts useful information. The evaluation of various ML algorithms for the identification of heart disease and the prediction of CVD are the main objectives of that work. Machine learning is a technique that enables a machine to learn without having to train it to do so explicitly. Artificial intelligence is a subfield that uses clever software to allow devices to perform tasks expertly.

To overcome difficulties with classification, we employ logistic regression. Instead of using linear regression, which denotes continuous progress, it does this by forecasting categorical outcomes. An example of a binomial, which has two possible results in the most straightforward instance, is the prediction of heart disease, which has been steadily rising in prevalence worldwide. When compared to the current method, using logistic regression

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

increases Accuracy. Many academics are working to create various IoT-based medical gadgets as a result. Below are some of the researcher's findings.

It has been demonstrated that less sophisticated systems, like logistic regression and support vector machines with linear kernels, produce results that are more accurate than those of more complex systems. ROC curves and F1 ratings have been utilized as evaluation methods [1]. This study's work analyses heart illness by applying machine learning methodologies to gauge the levels of disease severity. On the UCI heart disease dataset, experiments are conducted [2]. The intention of this work is to help people better understand their conditions and to encourage them to seek professional care early when necessary [3]. The study's foundation is publicly accessible medical data on heart disease. There are 208 entries in this dataset, and each contains eight details on the patient, including their age, type of chest pain, blood sugar level, blood pressure, heart rate, ECG, and more [4]. This paper proposes a system to uses a logistic regression classification algorithm to classify the risk level [5].This paper will cover some of the most recent studies employing data mining approaches to forecast cardiac diseases to ascertain whether data mining methodologies are relevant and practical. Additionally, it will evaluate the various mining algorithm combinations used. The standard data set for heart disease include the “UCI Machine Learning Repository”. There are 270 records total, some of which pertain to patients without cardiac disease and others to those who do. There are a total of 13 features in full, including “age”, “gender”, “chest pain”, “resting blood pressure level”, “cholesterol”, “fasting blood sugar”, resting “ECG” results, maximal “heart rate”, exercise-induced angina, old peak ST depression brought on by exercise relative to sleep, the slope of the peak exercise ST segment, and the number of significant vessels colored by fluoroscopy [6].

In the current healthcare industry, machine learning is frequently used to identify diseases and forecast their incidence using data models. For investigations including the risk assessment of complicated settings, the machine learning algorithm of logistic regression is comparatively popular. The goal of the study is to use logistic regression to predict overall risk and identify the most important determinants of cardiovascular disease [7]. In light of cardiovascular data, machine learning looks at how computers may learn (or enhance their performance) [8]. This section displays the risk level drawn from the heart disease database. The cardiovascular disease database's patient clinical treatment data has undergone pretreatment to improve the mining process [9].

Such proactive measures can avoid both the onset of sickness and the progression of the disease into a severe stage. As a result of the collection of various risk factors as a set of data, the data were then grouped into a number of risk factors that people are known to experience in their daily lives. “These risk factors are listed in Table 1”:

2. Methodology

The primary goal of developing this approach was to forecast the likelihood of developing heart disease ten years from now. To train our system, we used a variety of feature selection strategies, including backward elimination and logistic regression, as a machine learning approach. The details of these algorithms are covered below.

The Kaggle dataset, which has 4240 observations, forecasts the likelihood that a patient in a specific location will develop heart disease. In this study, we used SK Learn software to predict heart disease using the patient data provided. Over the next ten years, with the aid of this patient data, heart disease development can be anticipated. With the collected data, pre-processing and loading were carried out. Pre-processing and data loading was carried out using the obtained data. The preparation procedure comprises deleting the main error and any superfluous data from the database. This technique is also used to find missing data in a database. Then, utilizing feature selection, the data pertinent to the prognosis of cardiac illnesses is extracted.

Table 1: List of risk factors

Sex:	male or female
Age	age of the patient
Education	no further information provided
Current Smoker	whether or not the patient is a current smoker
Cigs Per Day	the number of cigarettes that the person smoked on average in one day

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

BP Medication	whether or not the patient was on blood pressure medication
Prevalent Stroke	whether or not the patient had previously had a stroke
Prevalent Hypertension	whether or not the patient was hypertensive
Diabetes	whether or not the patient had diabetes
TotalCholesterol	total cholesterol level
Systolic BP	systolic blood pressure
Diastolic BP	diastolic blood pressure
BMI	Body Mass Index
Heart Rate	heart rate values
Glucose	glucose level
10-year risk of coronary heart disease (CHD)	(Binary: "1", means "Yes", "0" means "No")

The processed data are analyzed using Exploratory Data Analysis. This step determines whether a predictive model is a feasible analytical tool to achieve a specific task with Accuracy. Then the data are refined to ensure that they are relevant and categorized, which enables the user to get an accurate result. The data which are required for the prediction have been separated. Before being tested with test data, the segregated data is trained using logistic regression. The precise results are obtained using logistic regression. It operates using a categorical dependent variable that can categorize the outcomes into discrete or binary categorical variables, 0 or 1, representing the absence or presence of cardiac disease. The data was validated when the logistic regression models had been trained and tested on the data. The association between the features has been established through analysis of the verified data. Finally, the computer displays the expected outcomes to the user "as shown in figure 1".

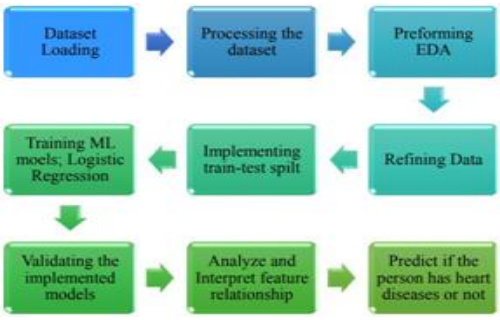

Figure 1: Block diagram of proposed model

3. Software Architecture

Logistic Regression

The model of the logistic regression result is shown in Figure 2. An algorithm for supervised classification is logistic regression. This algorithm for predictive analysis is built on the idea of probability. By calculating probabilities using the underlying logistic function, it assesses the relationship between the dependent variable (Ten-year CHD) and one or more independent variables (risk factors) (sigmoid function). As a cost function, the sigmoid function is used as a cost function to limit the logistic regression hypothesis between 0 and 1 (squashing), that is, 0 h (x) 1. In logistic regression, the cost function is referred to as:

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \ldots . (1)$$

The accurate presentation of data is crucial to the success of logistic regression. Essential elements from the available data set are thus chosen utilizing backward elimination and recursive elimination strategies to increase the model's potency. "In statistics, the outcome of a categorical dependent variable is forecast from a set of independent or predictor factors using a type of regression analysis called logistic regression. In logistic regression, the dependent variable is always binary". Prediction and success probability estimateare the two main uses of logistic regression.

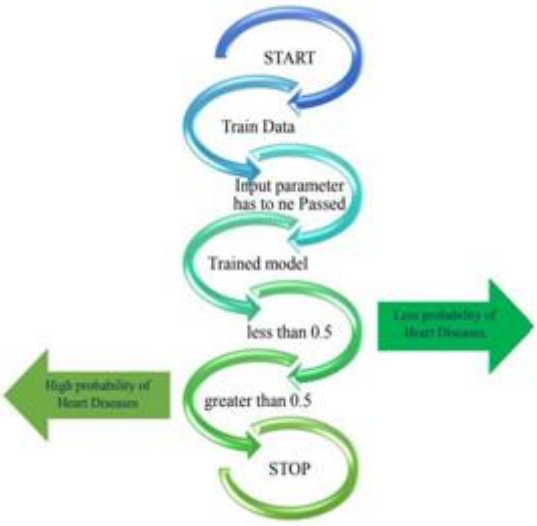

Figure 2: Software architecture of remote health monitoring system

Some of the qualities have P values that are greater than the preferred alpha (5%) in the results

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

(table 2), which indicates a weak statistically significant link between them and the likelihood of developing "heart disease". Here, the regression is performed repeatedly until all of the attributes have P values less than 0.05. The backward elimination strategy is used to remove the attributes with the highest P values one at a time (table 2).

Table 2:" Logistic Regression Results"

Coef	Coef	Std err	z	P>|z|	[0.025	0.975]
const	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
Sex_male	0.5742	0.107	5.345	0.000	0.364	0.785
age	0.0641	0.007	9.799	0.000	0.051	0.077
Current Smoker	0.0739	0.155	0.478	0.633	-0.229	0.377
Cigs Per Day	0.0184	0.006	3.000	0.003	0.006	0.030
BPMeds	0.1448	0.232	0.623	0.533	-0.310	0.600
Prevalent Stroke	0.7193	0.7193	1.471	0.141	-0.239	1.678
Prevalent Hyp	0.2142	0.136	1.571	0.116	-0.053	0.481
diabetes	0.0022	0.312	0.007	0.994	-0.610	0.614
Tot Chol	0.0023	0.001	2.081	0.037	0.000	0.004
sysBP	0.0154	0.004	4.082	0.000	0.008	0.023
diaBP	0.0040	0.006	-0.623	0.533	-0.016	0.009
BMI	0.0103	0.013	0.827	0.408	-0.014	0.035
Heart Rate	-0.0023	0.004	-0.549	0.583	-0.010	0.006
glucose	0.0076	0.002	3.409	0.001	0.003	0.012

A statistical analysis technology called logistical prediction uses previous data from a dataset to forecast a binary outcome, such as true or false. By examining the link between one or more earlier independent factors, an arithmetic regression model predicts a dependent data variable. The logistic regression approach is used to forecast the kind of people based on one or more predictor factors (x). It is used to simulate a variable with a binary conclusion that can only have two feasible values: 0 or 1, yes or no, or diseased or not.

4. Data Preparation

Due to the dataset's 4240 observations, 388 missing values, and 644 observations at risk for heart disease, two separate experiments were carried out to prepare the data (shown in figure 3). First, we made sure that the missing data were removed, leaving only 3751 data and 572 observations at risk for heart disease.

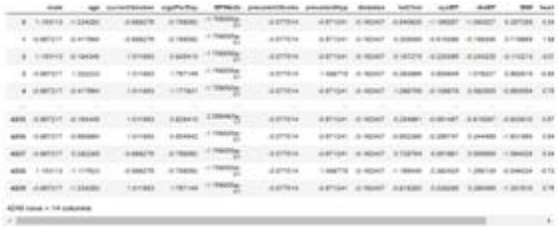

Figure 3: Data Preparation

As a result, our model receives less training from irrelevant observations. So, using the Simple Imputer and Standard Scaler modules of S_k learn, we moved forward with the imputation of data using the mean value of the observations and scaling them.

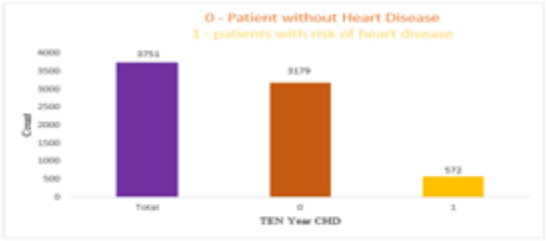

Figure 4: Data after Spanning and Attributing

In this figure 6, we explore four methods of feature scaling techniques that are implemented in scikit-learn. We will apply those techniques to the following three features sex, cp, and fbs. Then, we will plot how using those scalers affects the feature distributions. Except for the cp values, which fall within the [1–4] range, it is clear that the values of the majority of features fall within the [0, 1] range. According to sickit-learn documentation, the normalizer acts row-wise on the data. It adjusts the entire row to the unit norm rather than removing the mean and scaling by deviation (which is shown in figure 4). This is seen by the second plot produced using the dataset's Normalizer.

Exploratory Analysis:

There are 572 cases when a heart complaint is threatened and 3179 cases where there are no heart complaints. The graph (shown in Figure 5) explains the different threat factors that may help with the vaticination of the heart complaint of a set of people. The brace plot of the exploratory analysis enables us to about fantasize both distributions of a single variable as well as the relation between dyads of variables which is used to prognosticate the heart complaint (as shown in figure 5).

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

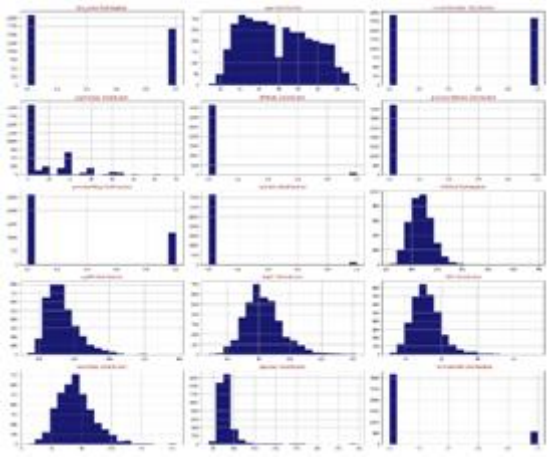

Figure 5: Exploratory Analysis

Confusion Matrix visualization:

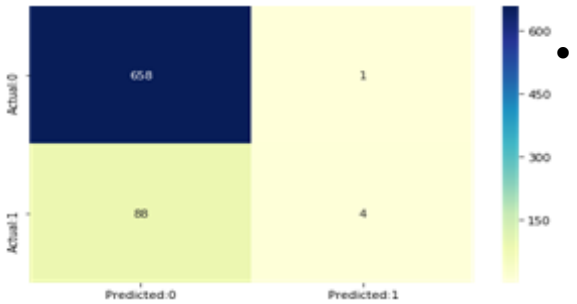

Figure 6: Confusion Matrix Visualization

The confusion matrix (shown in the figure 6) shows 658+4=662 correct predictions and 88+1= 89 in correct ones.

- True Positives: 4
- True Negatives: 658
- False Positives: 1
- False Negatives: 88

Feature Selection

Feature Selection using Backward Elimination (P-value) algorithm:

The data was further put through the backward elimination process to identify the most relevant features, which produced the results below:

Table 3: Result from Feature Selection using Back ward Elimination Method

Dep. Variable:	Ten Year CHD	**Time**	21:52:58
Model:	Logit	**Log-Likelihood**	-1414.3
No. Observations	3751	**converged**	True
Method	MLE	**LL-Null**	-1601.7
Df Model	14	**LLR p-value**	2.439e-71
Date	Fri, 18 May 2018	**Coef**	std err z
Pseudo R- sq u.:	0.1170	**P>|z|**	[0.025 0.975]

Based on the result above, the columns (table 1) were dropped. According to the fitted model (shown in table 4), the chances of being diagnosed with heart disease for males (gender male=1) over exp 0.5815) = 1.788687 for females (gender male = 0) when all other variables are held constant. The odds for men are 78.8% higher than the odds for women, based on of percentage difference.

- The coefficient for age states that, when all oth er factors are held constant, a one-year increase in age will result in a 7% increase in the likelihood of receiving a CDH diagnosis s ince exp (0.0655) =1.067644.
- Similarly, with every extra cigarette one smokes, there is a 2% increase in the odds of CDH.

There is no apparent change in the total cholesterol and glucose levels. There is a 1.7% increase in odds for every unit increase in systolic Blood Pressure.

Table 5: Dataset after Dropping Columns after Feature Selection

Coef	Coef	Std err	z	P>|z|	[0.025	0.975]
const	-8.6532	0.687	-12.589	0.000	-10.000	-7.306
Sex_male	0.5742	0.107	5.345	0.000	0.364	0.785
age	0.0641	0.007	9.799	0.000	0.051	0.077
Current Smoker	0.0739	0.155	0.478	0.633	-0.229	0.377
Cigs Per Day	0.0184	0.006	3.000	0.003	0.006	0.030
BPMeds	0.1448	0.232	0.623	0.533	-0.310	0.600
Prevalent Stroke	0.7193	0.7193	1.471	0.141	-0.239	1.678
Prevalent Hyp	0.2142	0.136	1.571	0.116	-0.053	0.481
diabetes	0.0022	0.312	0.007	0.994	-0.610	0.614
Tot Chol	0.0023	0.001	2.081	0.037	0.000	0.004
sysBP	0.0154	0.004	4.082	0.000	0.008	0.023
diaBP	0.0040	0.006	-0.623	0.533	-0.016	0.009
BMI	0.0103	0.013	0.827	0.408	-0.014	0.035
Heart Rate	-0.0023	0.004	-0.549	0.583	-0.010	0.006
glucose	0.0076	0.002	3.409	0.001	0.003	0.012

ROC Curve:

The trade-off between specificity and sensitivity (as well as TPR) is demonstrated by the ROC curve (1 – FPR). Classifiers that give curves that are closer to the top-left corner exhibit better performance. Points that are diagonal are expected to be provided by a random classifier by default. The ROC curve is generated by computing and plotting the true positive and false positive rates for a single classifier at different thresholds. For

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

instance, the predicted likelihood of positive class membership for observation would serve as the threshold in logistic regression.

The area under the curve (AUC):

$$\text{Accuracy} \quad \frac{TP + TN}{(TP + TN + FP + FN)} \ldots (2)$$

AUC is preferred over Accuracy as it is a much better indicator of model performance (shown in figure 7). This is because AUC uses the True Positive Rate and, It is advisable to use other measures in addition to the accuracy metric if you are using the False Positive Rate of the model across various cut-off values. A classifier is useless if it produces a score of less than 0.5 because it merely indicates that it performs worse than a random classifier. It might be between 0.5 and 1, and the higher the number, the better.

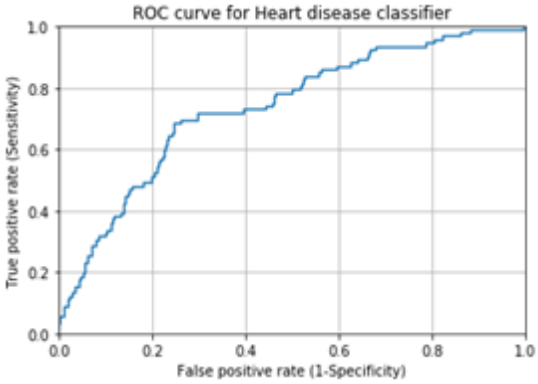

Figure 7: ROC curve for heart disease

5. Evaluation Metrics

For the evaluation of our output from our training the data, the accuracy was analyzed "Confusion matrix".

Confusion Matrix:

A confusion matrix is used to analyze the achievement of a classification model on a set of test data, the true values of which are known. It makes it possible to visualize how well an algorithm performs. It enables quick diagnosis of class labeling confusion, such as when one class is frequently mislabeled as the other. The crucial to the confusion matrix is the number of correct and incorrect prognostications or epitomized with count values and broken down by each class, not just a number of errors made.

Accuracy

The accuracy is calculated as follows:

- "True Positive (TP)" = Positive observation and is positively predicted.
- "False Negative (FN)" = Positive observation but negatively predicted.
- "True Negative (TN)" = Negative observation and negatively predicted.
- "False Positive (FP)" =Negative observation, but positively predicted.

Here,

- True Positives: 4
- True Negatives: 658
- False Positives: 1
- False Negatives: 88

Precision:

To determine the value of completeness, we divide the total number of correctly diagnosed positive consequently makes by the entire number of predicted positive exemplifications. A positive illustration is confirmed to be positive by High

"True positive (TP)"	"A Positive observation is predicted and is actually positive".
"True Negative (TN)"	"The output "TN" stands for True Negative and represents the number of correctly classified negative examples"
"False Positive (FP)"	"A Positive observation is predicted and is actually negative".
"False Negative (FN)"	"The predicted value is negative, but the actual value is positive".

Precision (a small number of FP). Precision is calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \ldots (3)$$

They attained perfection while training the data after point selection using backward elimination.

6. Results

"The confusion matrix shows 658+4=662 correct predictions and 88+1=89 incorrect ones".

Here,

- True Positives: 4

ISSN: 2309-5288 (Print)
ISSN: 2309-6152 (Online)
CODEN: JCLMC4
Journal of Coastal Life Medicine

- True Negatives: 658
- False Positives: 1
- False Negatives: 88

The coding was done to put together the data, to create it, pre-process it, constitute the model and then assess it. Jupyter Notebook as IDE is used to write the code in Python programming. Python libraries are used to do the experiments and to create the models (as in table 7.1).

Table7.1: Major modules and classes used from SK learn

Modules used	**Imported class from Respective modules**
Sklearn input	Simple Imputer
Sklearn.pre-processing	Standard Scaler
Sklearn. pipeline	Pipeline
Sklearn. Feature_ selection	RFECV (Recursive Feature Elimination)
Sklearn. ensemble	Random Forest Classifier
Sklearn.model_selection	Train_test_split, Stratified KFold
Sklearn.linear_model	Logistic Regression,
Sklearn. utils	Shuffle
Sklearn. metrics	Accuracy_score, confusion_ matrix

7. Conclusion

In this work, all the features chosen suggest a P value of less than 5%, indicating its important part in heart disease prediction after the elimination process. Heart disease is known to strike men more frequently than it does women. Ageing, daily cigarette smoking, and fluctuating blood pressure all of these factors raise the chance of acquiring heart disease. The chance of congenital heart abnormalities has not changed much as a result of total cholesterol. High-density lipoproteins may be too responsible for this. Glucose also has no significant change in the odds of congenital heart defects (0.2%). This model presents 88% accuracy, and the model has the 74% in the area under the ROC curve.

Future Scope:

Tool to determine a patient's personal illness risk. The framework can be modified to work with additional models, including neural networks, ensemble techniques, etc. Use various machine learning techniques to predict cardiac diseases, such "DT, NB, and SVM".

References

[1] Divyansh Khanna, Rohan Sahu, Veeky Baths, and Bharat Deshpande Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease, International Journal of Machine Learning and Computing, Vol. 5, No. 5, October 2015.

[2] N Satyanandam, Dr. Ch Satyanarayana,Heart Disease Prediction using predictive optimization techniques, International Journal of image, graphics and signal processing, Vol. 11, No. 9, September 2019.

[3] Paria Soleimani, ArezooNeshati, Applying the regression technique for the prediction of acute heart attack, World Academy of Science Engineering and Technology, International Journal of Biomedical Biological Engineering, Vol. 9, No. 11, 2015.

[4] Dr T Lalitha, future prediction of heart disease through exploratory analysis of data, smart green connected societies, vol. 1 no. 01, 2021

[5] Abhijna Bhat, Pragathi, Pranamya M S, Smitha,Prediction of Heart Disease Using Logistic Regression, International Research Journal of Engineering and Technology, Vol 07, Issue: 06June, 2020.

[6] AnimeshHazra,Subrata Kumar Mandal, Amit Gupta, Arkomita Mukherjee, Asmita Mukherjee, Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review, Advances in Computational Sciences and Technology Vol 10, Number 7, 2017.

[7] A, S ThanujaNishadipapers, Predicting Heart Diseases in Logistic Regression of Machine Learning Algorithms by PythinJupyterlab, International Journal of Advanced Research and Publications,Vol3, Number 8, 2019.

[8] Dinesh Kumar G, Prediction of cardiovascular disease using machine learning algorithms, preceding of 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore,India, 2018.

[9] Purushottam, Efficient heart disease prediction system using decision tree, 06 July 2015.