JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

GUI based Heart Disease Prediction System Using Machine Learning

[1]Manasa H R, [2]Dr.Anil Kumar K M
[1]Student, [2]Associate Professor
[1]Department of Computer Science and Engineering,
[1] Sri Jayachamarajendra college of Engineering,
[1]JSS science and Technological University, Mysore, India

Abstract: Heart disease is one of the major causes of death worldwide and the early prediction of heart disease is important. The computer-based heart disease prediction system helps the physician as a tool for heart disease diagnosis. Hence, the early diagnosis of heart diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. Machine learning techniques used to develop an appropriate computer-based system and decision support that can aid to early detection or prediction of heart disease, in this project we have developed a model which classifies if patient will have heart disease or not based on various features. The data collection, data preprocessing, feature selection and classification algorithms placed a vital role in heart disease prediction. In this proposed work data include 70000 records of patients with 13 attributes such as age, gender, weight, height, blood pressure types, glucose, smoke, alcohol, etc. In addition, many classification algorithms are used such as KNN, Random Forest, Logistic Regression, SVM, Decision Tree. Feature selection is of three types: Filter Method, Wrapper Method, and Embedded Method. Here, we used Filter based method for feature selection, Filter Based feature selection techniques such as chi-square, mutual information, correlation it will select relevant feature for model creation among all 13 attributes only 10 features are selected. It will reduce the computational cost of modeling and, in some cases, used to improve the performance of the model. Finally, we will compare all algorithms for accuracy, and conclude which algorithm gives accurate results. By using this GUI application, user can sit at their convenience and have a health check-up. The GUI is designed in such a way that anyone can easily operate it and have a check-up that will enable to predicts a patient will have heart disease or not.

***Index Terms* - KNN, SVM, Logistic Regression, Random Forest, Decision Tree, Filter-Based Feature selection, chi-square, Mutual Information, Correlation.**

I. INTRODUCTION

According to WHO (world Health Organization), heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. Among all human diseases. Many research have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is highlighted as silent killer which leads to the death of the person without obvious symptoms. The early detection of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This work aims to predict future heart disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning. Here, the major risk factors associated with the heart disease includes excessive alcohol consumption, high blood pressure, and the sex and age of a patient, cholesterol, BMI, etc.

Evolution of modern technologies like machine learning and Data science has opened the path for healthcare industries and medical institutions. To detect or predict disease earliest as possible and it helps to provide better patient care. It is a type of system that is made using machine learning algorithms for guessing the possible disease based on the risk factor and symptoms. In this proposed system is designed by using the supervised learning techniques such as K-Nearest neighbor, Logistic Regression, SVM, Random Forest, Decision Tree. Machine learning algorithm plays a significant role in disease prediction. It will predict whether the patient has a particular disease or not based on an efficient learning technique. Data set is collected from the Kaggle data repository that contains 70000 records of patients and 13 attributes such as age, gender, weight, height, ap_lo, ap_hi, cholesterol, alco, smoke, active, cardio, etc.

In addition, Feature Selection technique can improve model performance, reduces computational cost, and building better systematic models and decrease the required storage. The feature selection technique can be broadly classified into three categories: Filter Method, Wrapper Method, and Embedded Method. This proposed system Filter-based features selection algorithm used to select best features such as chi-square, mutual information, correlation. Feature selection placed a very important role to select relevant feature, this approach calculates the score, or rank based on their dependency on the class label for each feature. The filter technique can employ various feature selection criteria, including correlation coefficient (Pearson), Chi-square, Mutual information gain. it will select relevant feature for model creation among all 13 attributes.

A. *AIM AND OBJECTIVES*

The main and aim and objectives of this thesis:

- To analyse feature selection methods and understand their working principle.
- To study and research on various machine learning algorithms such as Random Forest, SVM, K-NN, etc.
- To compare numerous machine learning algorithms for accuracy and is visualized.
- To design a GUI by using Machine Learning model to predict the heart disease.

II. RELATED WORK

Jian Ping Li, et.al, proposed an efficient machine learning based diagnosis system has been developed for the diagnosis of heart disease. Different Machine learning classifiers are used LR, K-NN, ANN, SVM, NB, and DT to designing the system. Four standard feature selection algorithms are used Relief, LLBFS, LASSO, MRMR, and proposed a novel feature selection algorithm FCMIM used to solve feature selection problem. And cross-validation method is used in the system for hyper parameters selection. The Cleveland heart disease dataset has been used for testing the system and performance evaluation metrics are used to check the performance of the system. ANN classifier is best with Relief FS algorithm as compared to the specificity of MRMR, LASSO, LLBFS, and FCMIM feature selection algorithms. ANN with relief is the best predictive system to detect the healthy people [1]. The sensitivity value of the NB classifier is selected set of features by using LASSO FS algorithm and it is giving the best result as compared to the sensitivity values of the Relief FS algorithm with SVM (linear) classifier. The Logistic Regression classifier MCC with 91% on selected features selected by FCMIM FS algorithm. The processing time of the Logistic Regression with Relief, LASSO, FCMIM and LLBFS FS algorithm best as compared to the MRMR FS algorithms, and other classifiers. This result shows that the proposed features selection algorithm selects features that are more effective and with high classification accuracy than the standard feature selection algorithms. [1] The accuracy of SVM classifier with proposed FS algorithm (FCMIM) with 92.37% that is very accurate as compared to other methods.[1]

N. Komal Kumar et.al, proposed a different machine learning classifiers such as Random Forest, Decision Tree, Logistic Regression, Support vector machine (SVM), K-nearest neighbors (KNN) are used to predict cardiovascular disease (CVD). The proposed method using a random forest machine learning classifier has achieved a greater accuracy of 85.71% with a ROC AUC score of 0.8675 which is outperformed compared to all the classifiers are analyzed and classified patients with cardiovascular disease.[2]

KarenGárate-Escamila et al. In this research work chi-square and principal component analysis techniques has been used to predict heart disease. This study was conducted using 3 different datasets collected from UCI machine learning repository, The dataset contains 74 features and are validated by six ML classifiers used multilayer perceptron, and logistic regression, random forests, gradient-boosted tree, decision tree. The Chi-square and principal component analysis (CHI-PCA) with random forests (RF). From the analysis, ChiSqSelector derived features of physiological relevance, such as cholesterol, highest heart rate, chest pain, features related to ST depression, and heart vessels. The results proved that the combination of chi-square with PCA gives greater performance in most of the classifiers. Random forest classifier showed highest accuracy, with 98.7% for Cleveland, 99.0% for Hungarian, and 99.4% for Cleveland-Hungarian (CH) datasets. Respectively [3].

Sharma Purushottam et.al, Proposed an Efficient Heart Disease Prediction System using data mining techniques. In this exploration, the heart disease database contains screening clinical information of the heart patients like Age, Sex, Chest Pain, resting blood pressure, Serum cholesterol, Fasting blood sugar, resting electrocardiographic results, Maximum heart rate achieved, Exercise induced angina, ST depression, Number of major vessels colored by fluoroscopy and thal, Slope of the peak exercise ST segment. This study was conducted by using classification decision rule, this system can help medical practitioner in efficient decision making based on the given parameter. Training and testing phase using 10-fold method to find out the accuracy 86.3 % in testing phase and 87.3 % in training phase.[4]

Praneetha M et.al, proposed novel system used AI methods to improve the accuracy of the estimated cardiovascular infirmity. This forecast of cardiovascular disease is precisely done by utilizing diverse AI methods like SVC, KNN, DT, RFC. This approach used web applications to permit anybody to check the likelihood of having the disease all alone too. Decision Trees Classifier with 79% efficiency. Support Vector Classifier with 83% efficiency. Random Forest Classifier with 84% efficiency. K-Neighbors Classifier with efficiency of 87%. Since K-NN gives high efficiency, this algorithm is integrated with the web page to predict cardiovascular disease accurately.[5]

Tannishtha Mandal et.al, MLP performs very well in this domain and is to be used for prediction of CVD risk. In this approach discussion based on the importance of training model, training model is very important because it is a continuous process for prediction capability of the model changes with the composition of data set. Used classification algorithms SVM, KNN, LR, MLP. MLP gives good accuracy result.[6]

Senthil Kumar Mohan et.al**,** In this Identification of raw healthcare data of heart information will used in early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data towards the heart disease. And introduced a prediction model with different combinations of features, and several known classification techniques. It produced an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with Hybrid Random Forest with Linear Model (HRFM) [11].

Devansh Shah, Samir Patel et.al, presents various attributes related to heart disease, and the model on basis of supervised machine learning algorithms such as Naïve Bayes, decision tree, K-nearest neighbor, and random forest algorithm. It uses the existing dataset from the Cleveland data of UCI repository of heart disease patients. The dataset comprises of 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to the performance of different algorithms. This research aims to envision the probability of developing heart disease in the patients. The result K-nearest neighbor algorithm with highest accuracy.[8]

Muhammad Salman Pathan et.al, Filter-based feature selection technique has been used in this research namely the ANOVA-F test. The ANOVA-F test can be implemented in python using the f_classif () function that is provided by scikit-learn library. The f_classif () function is used in selecting the important features based on the values. via the SelectKBest class. SelectKBest is a method is available in the sklearn library. which takes a scoring function and ranks the features by these scores.[7] Here The scoring function is f_classif () i.e., ANOVA-F test and they have defined SelectKBest class to identify most important features from datasets. ANOVA -F test feature selection techniques is used for reduced feature subset. In this study they have used 2 different datasets, one is CVD and Framingham dataset. CVD dataset contains 29072 patient records with 11 features and Framingham datasets contains 4240 patient records with 15 features. The main impact of this research is how feature selection is impacting the performance of ML models. Here they used classification algorithms to predict the disease and accuracy of CVD dataset is 0.75 and Framingham 0.71.[7]

Rachael Hagan, Charles J et.al, the application of three machine learning methods to predict cardiovascular disease and search the space of hyper parameters for each method. This study used SVM, multi-layer perceptron neural networks and decision trees. Highlighted the importance of applying 10-fold cross validation, for comparison of the training models using two datasets with different characteristics, this dataset contains features such as the presence or absence of hypertension, cholesterol levels etc. Both datasets records age and gender among other features. SVM method applied to the UCI dataset produced an accuracy of 92% for Kaggle dataset reach only 72% accuracy. The three-layer MLP model trained on the UCI dataset with 74% accuracy was achieved. In the case of the UCI dataset, all ensemble methods achieved 90% accuracy, and Extra Trees models with 96% accuracy was obtained.[9]

K. Arul Jothi, S. Subburam et.al, proposed heart disease prediction Model using Machine Learning. There are two Data Mining Techniques have been used namely Decision Tree and K-Nearest Neighbor algorithm used for predicting the heart diseases. Decision Tree algorithm can be used on the dataset, to predict the chances of getting heart diseases of a patient in advance with an accuracy 81% and K-Nearest Neighbor algorithm can be used on the same dataset, to predict the chances of getting heart disease of patient in advance with an accuracy 67%. In this system, both algorithms are the best algorithms for managing and carrying medical data set and minimizes the medical tests as well as the death rate of the patients.[10]

III. METHODOLOGY

This section describes the various methods and materials used for this project: The Kaggle repository dataset have been used in this project work to carry out data analysis for heart disease prediction and the detailed information about the dataset, research design, data preprocessing techniques used for this study.

A. System Design and Methods:

In this section, we are going to discuss how we prepared or designed the whole system and various methods used for prediction of heart disease. As shown in below Figure 1.

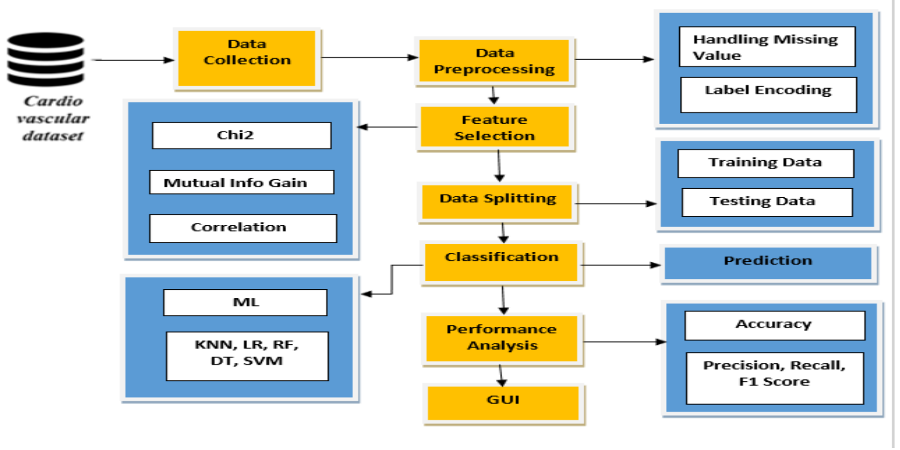


*Figure 1: Heart Disease Prediction System Design***.**

The Heart Disease Prediction System describes the workflow of the entire system. The working of the system is starts with the collection of datasets and selecting the important attributes or features. Then we preprocessed the required data into the required format. The data is splitting into two parts training and testing data. The different classification algorithms are applied, and the model is trained by using the training data. The accuracy of the system is obtained by using the testing data. This system is implemented by using the following modules.

1) Data collection

The dataset used for this project purpose was the Public Health Dataset. The Dataset is collected from UCI, Kaggle or GitHub data repository. In our thesis, the cardiovascular disease dataset is used. The dataset is used for predicting the heart disease. The dataset which contains the information about the patient details such as age, sex, glucose, cholesterol and so on. We read the dataset by using python panda's packages. Our dataset is in the form of '.csv' file format.

2) Data Description

The dataset contains 70000 records of patient's data in 14 features, such as age, gender, systolic blood pressure, diastolic blood pressure, etc. The target class "cardio" equals to 0, when patient is healthy and it's 1 if patient has cardiovascular disease. The task is to predict the absence or presence or of cardiovascular disease (CVD) using the patient examination results. Data description as shown in Table 1. There are 3 types of input features: Objective Feature: factual information. Examination Feature: results of medical examination. Subjective Feature: information given by the patient. Now the attributes which are used in this thesis research purpose are described as follows and what they are used or resemble as follows:

- age (in days)
- gender (0: female; 1: male)
- height (cm)
- weight (kg)
- ap_hi (systolic blood pressure)
- ap_lo (diastolic blood pressure)
- cholesterol (0: normal; 1: above normal; 2: well above normal)
- gluc (0: normal; 1: above normal; 2: well above normal)
- smoke (whether the patient is smokes or not)
- alco (alcohol intake)
- active (physical activity)
- cardio (target)

3) Pre-processing of Data

Before we start let us give a brief information about what data preprocessing is. Data pre-processing is very important step for the machine learning model creation. Initially, data may not be clean it contains some noise or unwanted information. Which can cause misleading outcomes by using the data preprocessing we can transform data into our required specific format. It is used to deal with duplicates, noises, Label Encoding, and Handling missing values of the dataset. Pre-processing of data has the activities like importing datasets, splitting datasets, attribute scaling, etc. Data Preprocessing techniques is required for improving the accuracy of the model. In our thesis we are using standard scaler from the sklearn library for preprocess our data. We choose this one over the other ones because it suits very well with our prediction system. In our thesis preprocess involves Handling /checking Missing Values, Label encoding.

a) Handling/Checking Missing Values*:* The Real-world data contains lot of missing values. the cause of missing value can be data corruption or failure to record the data. The handling of missing value is very important step during the preprocessing of the dataset many machines learning algorithm does not support for missing values. In this process, the Nan values such as missing values and Null values are replaced by 0. duplicate and Missing values are removed, and data was cleaned of any abnormalities.

b) Label Encoding**:** The label Encoding is very simple, and it involves converting each value in a column into a number. Consider the dataset having many more columns, to understand label-encoding, we will focus on one categorical column. In our research dataset contains some categorical (string) values that is gender, glucose, cholesterol. This attribute is containing values like Gender (female, male) and glucose (Normal, above normal, well above normal) cholesterol (normal, above normal, well above normal). Here, we can convert the strings into integer numeric value (0,1,2).

Table 1. Dataset Description

	Features	Description		
		Value Type	Variable	Variable Type
Categorical Data	Age (in days)	Objective Feature	age	Int(days)
	Age (in years)	Objective Feature	age	Int (years)
	Height	Objective Feature	height	Int(cm)
	Weight	Objective Feature	weight	float (Kg)
	Systolic blood pressure	Examination Feature	ap_hi	Int
	Diastolic blood pressure	Examination Feature	ap_lo	Int
Numerical Data	Gender	Objective Feature	gender	String (0: female; 1: male)
	Cholesterol	Examination Feature	Cholesterol	String (0: normal; 1: above normal; 2: well above normal)
	Glucose	Examination Feature	gluc	string (0: normal; 1: above normal; 2: well above normal)
	Smoking	subjective features	smoke	Binary (0,1)
	Alcohol intake	subjective features	alco	binary
	Physical activity	subjective features	active	binary
	Presence or absence of cardio disease	Target variable	cardio	binary
	id	Objective Feature	id	Int

4) Data Splitting

After the collection and preprocessing of the dataset, we split the dataset into training data and testing data. The training data is used for prediction model and testing data is used for evaluating the prediction model. In our process, we considered 80% of the dataset to be the training data and the remaining 20% to be the testing data. Then the Separated training and testing data is an important part of evaluating data models. Typically, when we are separating a data set into a training and testing set, most of the data is used for training, and a small portion of the data is used for testing.

5) Feature Selection

Features are the input variables that we are used in our machine learning models. Each column in our dataset contains a feature. To train an optimal model, we need to make sure that we use only the important features. If we have so many features, the model can capture the unimportant features and learn from noise. The method of choosing the important features of our data is called Feature Selection. Feature selection is the method of reducing the number of features to develop a predictive model. It is desirable to reduce the number of input features to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. There are mainly two types of feature selection:

- supervised Feature selection
- unsupervised Feature selection

The main difference has to do with whether features are selected based on the target variable or not. Unsupervised feature selection method is ignoring the target variable, these methods are used to remove the redundant variables by using correlation. Supervised feature selection method is using target variable, these methods that are used to remove the irrelevant variables. Again, the supervised methods are divided into three types, that is shown in the below Figure 2, illustrates different Types of Feature Selection Methods.

1) Wrapper Based Method
2) Filter Based Method
3) Intrinsic Based Method

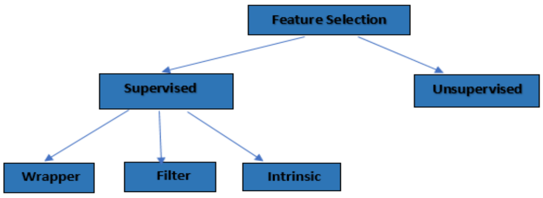


Figure 2: Types of Feature Selection Methods

In our proposed work mainly uses three filter-based feature selection algorithms, namely, Chi-square, mutual information, Correlation. Figure 3 shows the different Filter based feature selection methods. The filter-based feature selection method chooses the best subset immediately before passing it to the learning model. The remaining two approaches, wrapper and embedded, creates the optimal subset in combination with the learning model.

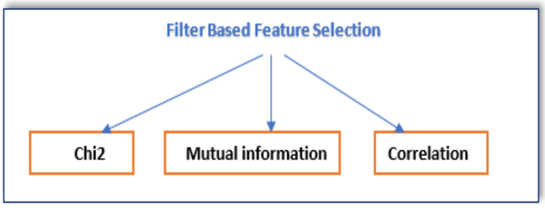


Figure 3: Filter Based Feature Selection

A) Feature Selection Using Filter Methods: Filter-based selection techniques is a statistical method used to determine the dependence or association among input attributes and the target attribute. In filter-based method, without depending on any learning models, features are assessed based on the general properties like feature importance and ranking (scores) of the feature. And characteristics are chosen based on the output results or scores generated by the various statistical methods used to validate them. It primarily qualifies the properties of features using distinct types of measurement criteria, including information, distance, dependence, consistency, similarity, and statistical measures.

a) Chi-Square: A chi-squared test, also called as $\chi 2$ test. The Chi-Square test is useful in feature selection and is an important problem in machine learning, where we have several features in our dataset and must select the best features to build the model by using SelectKBest method. The chi-square test helps us to solve the problem in feature selection by testing the relationship between the features. A chi-square test is used test the independence of two events. Given the data of two variables, we can get observed count is O and expected count is E. Chi-square test follows the univariate feature selection method to identify the dependence of each feature with the target feature. Here, we remove the features that are independent of the target feature and other dependent features is highly important to predict the target feature. The process steps are given below.

Step 1: First select all the features from the dataset.
Step 2: calculate the Chi-square feature scores by using chi2() function that is in the scikit-learn machine library.
Step 3: the feature with a large Chi-square value relies more on the target feature and is chosen for model construction. SelectKBest () was used to select the ten features with the high Chi-square score value. K denotes the number of features selected in the final dataset.
Step 4: based on the top ten ranking of the features the next step is needed to select the number of features that are used for model creation.

b) Mutual information: The mutual information (MI) is calculated between two variables and measures the reduction in uncertainty for one feature given a known value of the other feature. The process is listed below.
Step 1: First select all the features from the dataset.
Step 2: calculate the mutual information feature scores and rank to determine the relationship. Between the features and the target feature using the mutual_info_classif () function.
Step 3: The higher score indicates the additional dependency of the feature on target variable. SelectKBest() was used to select the ten features with the high mutual information scores used for model construction.

c) Correlation based feature selection (CFS): The correlation-based feature selection (CFS) method is a filter-based method and therefore independent of the final classification model. we can evaluate the relationship between each feature and target using a correlation and selecting those features that have strongest relationship with the target feature.

6) Classification

In machine learning, Classification is referring to a predictive modeling problem where a class label is predicted for a given example of input data. Classification is a method for labeling a particular set of data values with different classes, which is carried out on structured and unstructured data. The aim of classification is to create predictive model to approximate mapping between input variables and discrete output target variables. There are different classification algorithms available. This relies on the nature of the dataset available and the domain of the problem. In this analysis, we used five different supervised learning algorithms, which are decision tree, random forest, and support vector machine, K- nearest neighbor, logistic regression.

1) K-Nearest Neighbor Algorithm

K-closest neighbors (KNN) algorithm is a supervised machine learning algorithms which can be used for both classification and regression problems. KNN is a lazy classifier since it doesn't have a specific training phase and uses all the data for training while classification. KNN is additionally a non-parametric learning algorithm calculation since it doesn't expect anything about the fundamental data.
In this KNN approach First we need to select the number K neighbors and calculating the distance between each neighbor, select the nearest neighbor and then counting number of data points in each category and assign new datapoints to category for which number of the neighbor is maximum.

2) Random Forest

Random Forest is an easy and flexible supervised machine learning algorithm it is used for both regression and classification problems. It will build trees by using the ensemble learning techniques. It contains various subsets of the given datasets that is used to predict the accuracy of the dataset. RF algorithm takes less training time compared to other algorithm and predict output with high accuracy. In this approach first it will select the random k points from the training set and build the decision tree associate with selected subset data points and then choose N number of decision tree that we want to build again the new process is repeats for new datapoints. Finally, it will predict the category of the datapoint.

3) Logistic Regression

Logistic Regression is a classifier used to predict the probability of the target features or attributes. It can be either 0 or 1, True or False, and Yes or No. In Logistic regression the dependent variable is must be categorical value in nature and independent should not have multicollinearity. To Implement Logistic Regression algorithm first we need to import our dataset and required libraries and then perform data preprocessing, data splitting, fitting model to the test and train sets and predict the result, test accuracy, and create confusion matrix to visualizing test result.

4) Decision Tree

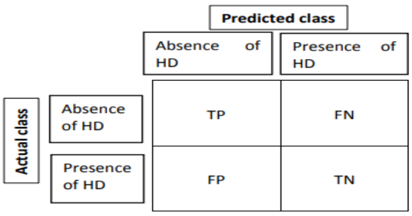
Decision Tree is one of the popular and easiest classification algorithms to understand and interpret is belongs to the supervised learning algorithms family and used for both classification & regression problem solving. In this algorithm the training data model is used to predict the class and value of the target variables by learning decision rules from training data. Class prediction is start from root of the tree and compare the values of the root feature with target attribute based on the value it will jump to the next node in tree.

5) SVM algorithm

The SVM is a supervised machine learning algorithm use for both regression and classification, SVM is mainly used for finding hyperplane in an N dimensional space it classifies the data points it depends on the number of features. If the feature input is 2 the hyperplane is just a line. Feature input is 3, hyperplane becomes 2D plane its mainly depends on the number of features that we used in our algorithm.

7) Performance Analysis: Performance metrics are used to evaluate how different algorithms perform based on various criteria such as accuracy, precision, recall, f1-scoore etc. Different performance metrics are discussed below.

i) Confusion Matrix: The confusion matrix is used to analyze the performance of the algorithm. The rows indicate the actual instance of the class label while the columns indicate the predicted class instances. It gives us a matrix as output and gives the total performance of the system. The Figure 4 shows a confusion matrix data.



		Predicted class	
		Absence of HD	Presence of HD
Actual class	Absence of HD	TP	FN
	Presence of HD	FP	TN

Figure 4: Confusion matrix

True Positive value means the positive value is correctly predicted, false positive means the positive value is falsely classified, false negative means the negative value is falsely predicted while the negative value is correctly classified the true negative values. Confusion matrix table is used to calculate different performance evaluation metrics as discussed below.
ii) Accuracy: Accuracy is the ratio of the number of correctly classified samples to all the cases. It is equal to the sum of TP and TN divided by the total number of instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

iii) Precision: Precision is defining the number of positive class prediction that belong to the positive class.

$$Precision = \frac{TP}{TP + FP} \qquad (2)$$

iv) Recall: Recall is calculating ratio between the number of positive instances correctly classified as positive to the total number of positive instances. Recall is also often called sensitivity.

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

v) F1 Score: F1 score is a measure both precision and recall and tries to find a balance between both.

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall} \qquad (4)$$

IV. IMPLEMENTATION

First, we start process with data collection, in our thesis work we are collected CVD dataset from Kaggle Data Repository. That contains 70000 records of patient's data in 14 features, such as age, gender, systolic blood pressure, diastolic blood pressure, etc. The target class "cardio" equals to 0, when patient is healthy and it's 1 if patient has cardiovascular disease. The task is to predict the absence or presence or of cardiovascular disease (CVD) using the patient examination results. second step is importing all necessary python libraries, by using the Jupyter or Spyder IDE. Next step is Data preprocessing checking or handling missing values and perform label encoding for string values to numeric values as we discussed already in the above section III. Here, we considered 80% of the dataset to be the training data and the remaining 20% to be the testing data. Then the Separated training and testing data is an important part of evaluating data models.
Then apply feature selection Methods such as chi2, mutual information, CFS. Correlation heat map each square shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no relationship between two variables. Correlation is close to 1 is more positively correlated. The diagonals are all white because those squares are correlating each variable to itself (so it's a perfect correlation). The Figure 5 shows the correlation heat map for all 13 attributes and one target cell. The selected features are age, weight, cholesterol, ap_lo, ap_hi, glucose, active, alco, smoke for the further analysis. Then we will build a model by using Machine learning algorithms.

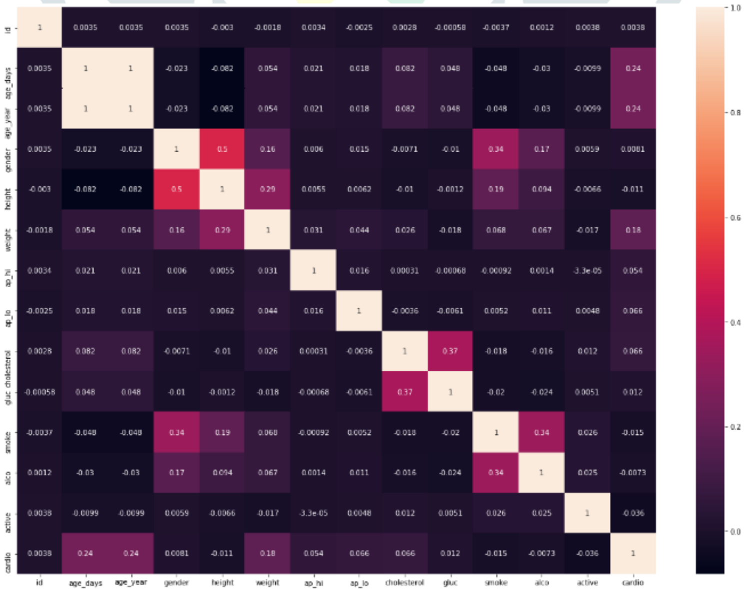


Figure 5: Correlation heat map

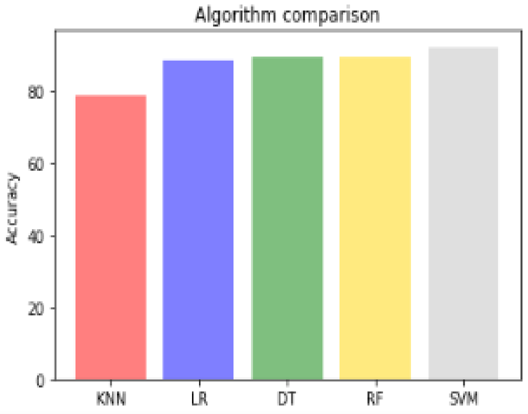


Figure 6: Algorithm Comparison

Finally, we will compare all algorithms for accuracy. And conclude that SVM algorithm gives accurate results as shown in Figure 6. Then deploy the models into GUI application. The GUI is created by using Tkinter python package that we discussed already. The GUI application is user friendly. After entering the symptoms details the predicted output will be displayed on the screen. Users can frequently use this application. That is shown in the Figure 7 and Figure 8. we have created heart disease system prediction system. The user can enter these details like name patient, age, height, weight, blood pressure types, whether the patient is smokes, alcohol intaking, and physically active or not, cholesterol, glucose etc, and press submit button. By using these details system will predict that whether the patient is having a heart disease or not. Finally, result will be displayed on the screen and with accuracy of all algorithms.

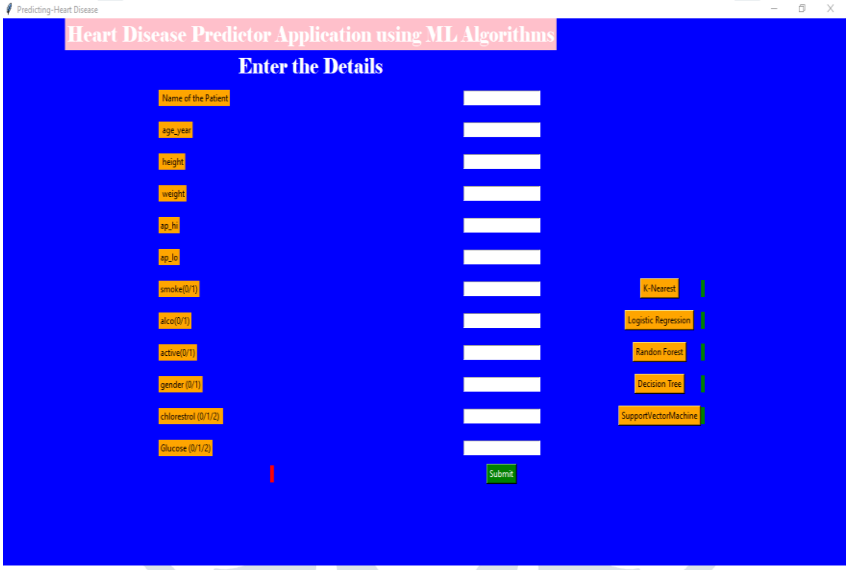
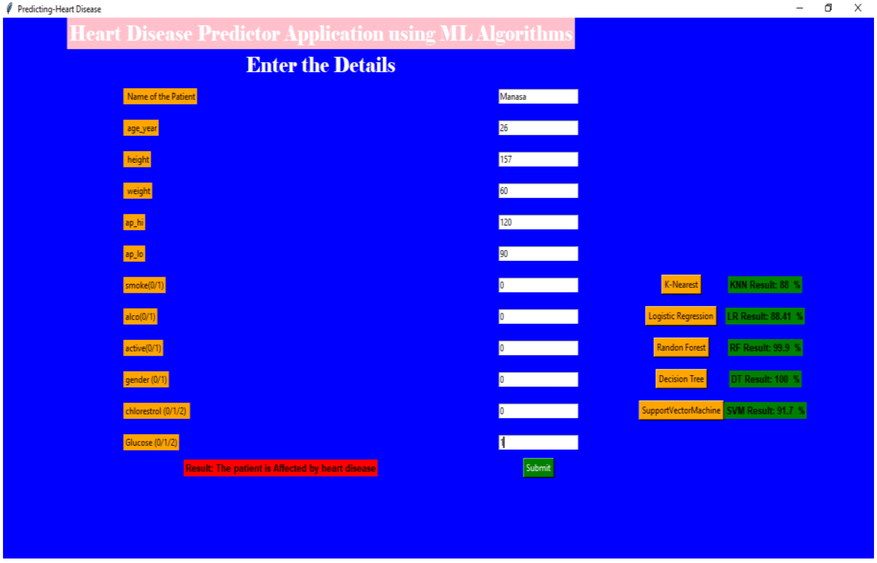


Figure 7 Graphical User Interface for Heart disease prediction System



Name of the Patient Manasa
age_year 26
height 157
weight 60
ap_hi 120
ap_lo 90
smoke(0/1) 0
alco(0/1) 0
active(0/1) 0
gender (0/1) 0
chlorestrol (0/1/2) 0
Glucose (0/1/2) 1

KNN Result: 88 %
LR Result: 88.41 %
RF Result: 99.9 %
DT Result: 100 %
SVM Result: 91.7 %

Result: The patient is Affected by heart disease

Figure 8 Graphical User Interface with Result

V. RESULTS

The CVD heart disease dataset gathered from the Kaggle Repository contains 70000 samples: 50% samples indicate the absence of heart disease, and 50% samples indicate the presence of heart disease. The optimal feature subsets for classification were selected by using feature selection algorithms, namely, Chi-square, mutual information, correlation. The result analysis was conducted on five classification algorithms: decision tree, random forest, support vector machine, K-nearest neighbor, logistic regression, and comparing all models for accuracy with different number of features is shown in the Table 3.

Table 2. Comparison of all Algorithms Results

Feature selection method	Model	Features	Dataset		Accuracy	Precision	Recall	F1-Score	Support	
			Train	Test					0	1
Chi2	KNN	10	56000	14000	78.87%	55%	55%	55%	7086	6914
	RF	10	56000	14000	89.40%	67%	71%	69%	7487	6513
	LR	10	56000	14000	88.62%	67%	71%	69%	7487	6513
	DT	10	56000	14000	89.50%	65%	73%	69%	7815	6185
	SVM	10	200	200	92.34%	48%	76%	59%	138	62

Table 2 summarizes the best accuracy performance of chi2 feature selection technique. The feature subset selectedby the chi2 feature selection technique has achieved thehighest classification accuracy of 92.34%, precision 48%, recall 76%, F1 score is 59%, with the SVM classifier. The selected features are age_days, weight, gender, ap_lo, ap_hi, cholesterol, gluc, active, smoke, alco. By using the chi2 feature selection method we have selected 10 best features among all 13 features. The experimental findings suggest that using feature selection algorithm with machine learning models is capable of classifying the disease well with a smaller number of features.

Table 3. Comparison of Different Number of Features

FS Method	Model	Features	Dataset		Accuracy	Precision	Recall	F1-Score	Support	
			Train	Test					0	1
chi2	KNN	5	56000	14000	66%	65%	65%	65%	7086	6914
		6	56000	14000	70%	77%	55%	64%		
		7	56000	14000	69%	76%	56%	65%		
		8	56000	14000	69%	75%	55%	63%		
		9	56000	14000	66%	70%	55%	62%		
	RF	5	56000	14000	67%	66%	67%	66%	7069	6931
		6	56000	14000	72%	75%	65%	70%		
		7	56000	14000	72%	75%	65%	70%		
		8	56000	14000	71%	75%	63%	69%		
		9	56000	14000	69%	71%	61%	66%		
	LR	5	56000	14000	71%	72%	67%	69%	7069	6931
		6	56000	14000	71%	76%	60%	67%		
		7	56000	14000	71%	76%	60%	67%		
		8	56000	14000	71%	76%	60%	67%		
		9	56000	14000	71%	75%	62%	68%		
	DT	5	56000	14000	62%	62%	61%	61%	7069	6931
		6	56000	14000	72%	75%	64%	69%		
		7	56000	14000	72%	75%	65%	69%		
		8	56000	14000	71%	75%	63%	68%		
		9	56000	14000	66%	71%	55%	62%		
	SVM	5	56000	14000	89%	80%	64.70%	63.82%	7069	6931
		6	56000	14000	89%	80%	61.53%	63.82%		
		7	56000	14000	85%	77%	68.18%	70.83%		
		8	56000	14000	85%	84.21%	82.60%	71.11%		
		9	56000	14000	89%	80.00%	64.70%	63.82%		

V. CONCLUSION AND FUTURE WORK

Machine learning techniques used to develop an appropriate computer-based system and decision support that can aid to early detection or prediction of heart disease, in this project we have developed a model which classifies if patient will have heart disease or not based on various features. The computer-based heart disease prediction system helps the physician as a tool for heart disease diagnosis. Hence, the early diagnosis of heart diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. Moreover, if we increase the attributes, maybe we can find more accurate result, but it will take more time to process, and the system will be slower. So, considering these possible things we took a decision which is better for us to work with.

Limitation and Future work: The dataset that is used in our thesis is very small and old, in our future work we are collecting different attribute and datasets are taken as input and are processed by the machine learning algorithm integrated with the web framework, we are unable to explore other feature and different type of heart disease prediction we are just predicting whether a patient having heart having disease or not. Future enhancements of heart disease prediction system are to find specific type of heart disease such as CAD, CVD, heart attack, etc.

REFERENCES

[1] Jian Ping Li, Amin ul Haqq, Salah Uddin, Jamaluddin Khan, Asif Khan, and Abdus Sa boor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare", IEEE Access, Digital Object Identifier 10.1109/ACCESS.2020.3001149

[2] N. Komal Kumar, G. Sarika Sindhu, D. Krishna Prasanth, A. Shaheed Sultana, "Analysis and Prediction of Cardiovascular Disease using Machine Learning Classifiers". 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS), 978-1-7281-5197-7/20/$31.00 ©2020 IEEE.

[3] A. KarenGárate-Escamila, A. E. Hassani, and E. Andrés, "Classification models for heart disease prediction using feature selection and PCA," Elsevier, Informatics in Medicine Unlocked, vol. 19, Article ID 100330, 2020. https://doi.org/10.1016/j.imu.2020.100330

[4]. Sharma Purushottam, Dr Kanak Saxena, Richa Sharma" Efficient Heart Disease Prediction System using Decision Tree" in IEEE International Conference on Computing Communication and Automation (ICCCA-2015), May 2015.

[5] Praneetha M, Sri Varsha M, Jesudoss A, Albert Mayan, "Cardiovascular Disorder Prediction using Machine Learning", 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432199

[6] Tanisha Mandal, Sougata Bera, Debashis Saha"A Comparative Study of AI-based Predictive Models for Cardiovascular Disease (CVD) Prevention in Next Generation Primary Healthcare Services", 2020 IEEE International Conference for Innovation in Technology (INOCON) Bengaluru, India. Nov 6-8, 2020

[7] Muhammad Salman Pathan, Avishek Nag, Muhammad Mohsin Pathan, Soumyabrata Dev "Analyzing the impact of feature selection on the accuracy of heart disease prediction", Volume 2, November 2022, 100060 Healthcare Analytics journal www.elsevier.com/locate/health

[8] Devansh Shah, Samir Patel, Santosh Kumar Bharti "Heart Disease Prediction using Machine Learning Techniques", Received: 27 September 2020 / Accepted: 2 October 2020 © Springer Nature Singapore Pte Ltd 2020

[9] Rachael Hagan, Charles J. Gillen, Fiona Mallett, "Comparison of machine learning methods for the classification of cardiovascular disease", Available online 20 May 2021 2352-9148/© 2021 The Authors. Published by Elsevier Ltd. https://doi.org/10.1016/j.imu.2021.100606

[10] K. Arul Jothi, S. Subbu ram, V. Umadevi, K. Hemavathy, "heart disease prediction system using machine learning" https://doi.org/10.1016/j.matpr.2020.12.901 2214-7853/ 2021 Elsevier Ltd. All rights reserved.

[11]. Mohan, Senthil Kumar, Chandrasekar Thirumal Ai, and Gautam Srivastava, "Effective heart disease prediction using hybrid machine learning techniques" IEEE Access 7 (2019): 81542-81554

[12] [World Health Organization, Cardiovascular Diseases, WHO, Geneva,Switzerland,2020, https://www.who.int/healthtopics/cardiovascular-diseases.

[13] **Xiao**-Yan Gao, Abdelmegeid Amin Ali, Hassan Shaban Hassan, and Eman M. Anwar, “Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method”. Hindawi Complexity Volume 2021, Article ID 6663455, 10 pages https://doi.org/10.1155/2021/6663455

[14] Anuradha P, Dr. Vasantha Kalyani David, “Feature Selection and Prediction of Heart diseases using Gradient Boosting Algorithms”, 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) | 978-1-7281-9537-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICAIS50930.2021.9395819

[15] C. BeulahChristianLatha, S. Carolina Sanjeev, “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”. Informatics in Medicine Unlocked 16 (2019) 100203 Available online 02 July 2019 2352-9148/ © 2019 Published by Elsevier Ltd. https://doi.org/10.1016/j.imu.2019.100203

[16] Juan-Jose Beunzaa, Enrique Puerta’s, Ester García-Ovejero, Gemma Villella, Emilia Condesa, Gergana Koleva, Cristian Hurtado, Manuel. Landechoa, “Comparison of machine learning algorithms for clinical event prediction” (risk of coronary heart disease) 1532-0464/ © 2019 Elsevier Inc. All rights reserved. https://doi.org/10.1016/j.jbi.2019.103257

[17] J. Jeyaranjani, T. Dhiliphan Rajkumar, T. Ananth Kumar “coronary heart disease diagnosis using the efficient ANN model”, https://doi.org/10.1016/j.matpr.2021.01.257 2214-7853/ 2021 Elsevier Ltd. All rights reserved.

[18] Farzana Tasnim, Sultana Umme Habiba,” A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection”, 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) | 978-0-7381-3042-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICREST51555.2021.9331158

[19] VardhanShorewala “Early detection of coronary heart disease using ensemble techniques”,. https://doi.org/10.1016/j.imu.2021.100655, Available online 11 July 2021 2352-9148/© 2021 Published by Elsevier Ltd.

[20] Atharva Nikam, Saket Bhandari, Aditya Mhaske, Shamla Mantri“Cardiovascular Disease Prediction Using Machine Learning Models”,2020 IEEE Pune Section International Conference (PuneCon) | 978-1-7281-9600-8/20/$31.00 ©2020 IEEE | DOI: 10.1109/PuneCon50868.2020.9362367