Heart Disease Prediction Using Effective Machine Learning Techniques

**ABSTRACT:**

In today’s era deaths due to heart disease has become a major issue approximately one person dies per minute due to heart disease. This is considering both male and female category and this ratio may vary according to the region also this ratio is considered for the people of age group 25-69. This does not indicate that the people with other age group will not be affected by heart diseases. This problem may start in early age group also and predict the cause and disease is a major challenge nowadays. Here in this paper, we have discussed various algorithms used for prediction of heart diseases. And create a GUI based system for predicting heart disease.

**Keywords:** dataset, Machine learning-Classification method

**Introduction:**

The major challenge that the Healthcare industry faces now-a-days is superiority of facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted.

Records or data of medical historyis very large, but these are from many dissimilarfoundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database.

Even if cardiovascular diseases is found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on on the well-timed judgment of that disease. An correct and methodical tool for recognizing high-risk patients and mining data for timely analysis of heart infection looks a serious want.

Even though heart disease is acknowledged as the supreme chronic sort of diseasein the world, it can be most avoidable one also at the same time. A healthy way of life (main prevention) and timely analysis (inferior prevention) are the two major origins of heart disease director. Conducting steady check-ups (inferior prevention) showsoutstanding role in the judgment and early prevention of heart disease difficulties. Several tests comprising of angiography, chest X-rays, echocardiography and exercise tolerance test support to this significant issue. Nevertheless, these tests are expensive and involve availability of accurate medical equipment.

Heart expert’s create a good and huge record of patient’s database and store them. It also delivers a great prospect for mining a valued knowledge from such sort of datasets.

Researchers make use of several data mining techniques that are accessible to help the specialists or physicians identify the heart disease. Commonly used procedures used are decision tree, k-nearest and Naïve Bayes. Other different classification based techniques used are bagging algorithm, kernel density, sequential minimal optimization and neural networks, straight Kernel self-organizing map and SVM (Support Vector Machine).

**Domain overview:**

Machine learning is to predict the future from past data. Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of Computer Programs that can change when exposed to new data and the basics of Machine Learning, implementation of a simple machine learning algorithm using python. Process of training and prediction involves use of specialized algorithms. It feed the training data to an algorithm, and the algorithm uses this training data to give predictions on a new test data. Machine learning can be roughly separated in to three categories. There are supervised learning, unsupervised learning and reinforcement learning. Supervised learning program is both given the input data and the corresponding labeling to learn data has to be labeled by a human being beforehand. Unsupervised learning is no labels. It provided to the learning algorithm. This algorithm has to figure out the clustering of the input data. Finally, Reinforcement learning dynamically interacts with its environment and it receives positive or negative feedback to improve its performance.

Data scientists use many different kinds of machine learning algorithms to discover patterns in python that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning. Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function from input variables(X) to discrete output variables(y). In machine learning and statistics, classification is a supervised learning approach in which the computer program learns from the data input given to it and then uses this learning to classify new observation. This data set may simply be bi-class (like identifying whether the person is male or female or that the mail is spam or non-spam) or it may be multi-class too. Some examples of classification problems are: speech recognition, handwriting recognition, bio metric identification, document classification etc.

Analyses Predicts

Machine Learning

Past Dataset

Trains

Fig: Process of Machine learning

[Supervised Machine Learning](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) **is the** majority of practical machine learning uses supervised learning. Supervised learning is where have input variables (X) and an output variable (y) and use an algorithm to learn the mapping function from the input to the output**is y = f(X).** The goal is to approximate the mapping function so well that when you have new input data (X) that you can predict the output variables (y) for that data. Techniques of Supervised Machine Learning algorithms include **logistic regression**, **multi-class classification**, **Decision Trees** and **support vector machines etc**. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers. Supervised learning problems can be further grouped into **Classification** problems. This problem has as goal the construction of a succinct model that can predict the value of the dependent attribute from the attribute variables. The difference between the two tasks is the fact that the dependent attribute is numerical for categorical for classification. A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes. A classification problem is when the output variable is a category, such as “red” or “blue”.

Agriculture is one of the most important occupations practiced in our country. It is the broadest economic sector and plays an important role in overall development of the country. About 60 % of the land in the country is used for agriculture in order to suffice the needs of 1.2 billion people. Thus, modernization of agriculture is very important and thus will lead the farmers of our country towards profit. Data analytic (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software. Earlier yield prediction was performed by considering the farmer's experience on a particular field and crop. However, as the conditions change day by day very rapidly, farmers are forced to cultivate more and more crops. Being this as the current situation, many of them don’t have enough knowledge about the new crops and are not completely aware of the benefits they get while farming them. Also, the farm productivity can be increased by understanding and forecasting crop performance in a variety of environmental conditions. Thus, the proposed system takes the location of the user as an input. From the location, the nutrients of the soil such as Nitrogen, Phosphorous, Potassium is obtained. This static data is the crop production and data related to demands of various crops obtained from various websites. It applies machine learning and prediction algorithm to identify the pattern among data and then process it as per input conditions.

**Preparing the Dataset:**

The demo dataset is now supplied to machine learning model on the basis of this data set the model is trained. Every new detail filled at the time of application form acts as a test data set. After the operation of testing, model prediction based upon the inference it concludes on the basis of the training data sets. Satellite Imagery (Remote Sensing Data), has been widely used for predicting crop yield. This dataset is collected using the sensors mounted on satellites or planes, which detect the energy (electromagnetic waves), reflected or diffracted from surface of the earth. Remote sensing data has a lot of energy bands to offer, but mainly only few of them have been used for crop yield prediction. Yet, there are some people who have tried generating relevant features using the bands which are typically ignored, and they have been successful with improving results with that. In case of this dataset, most people rarely explore the high-order moments of the features. Based on these datasets people have used algorithms like Regression models, Random Forest and Nearest Neighbor etc.

Table shows details of the datasets:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Age | Age of person |
| Sex | Male or Female |
| cp | Chest pain type |
| trestbps | Resting blood pressure |
| chol | Cholesterol level |
| fbs | Fasting blood sugar |
| Rest ecg | Resting electrocardiogram |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | the slope of the peak exercise ST segment |
| ca | number of major vessels (0-3) colored by flourosopy |
| thal | 3 = normal; 6 = fixed defect; 7 = reversable defect |

**Problem Description:**

India has witnessed an alarming rise in the occurrence of heart disease, stroke, diabetes and cancers in the past 25 years, a series of new studies published on Wednesday in [The Lancet](https://www.thelancet.com/journals/langlo/article/PIIS2214-109X(18)30407-8/fulltext) and its associated journals have revealed.

Detailed estimates of cardiovascular diseases, diabetes, chronic respiratory diseases, cancer, and suicide show that their prevalence has gone up in every Indian state between 1990 and 2016, but there is vast variation among states.

The prevalence of heart disease and stroke has increased by over 50% from 1990 to 2016 in India, with an increase observed in every state. The contribution of these diseases to total deaths and disease burden in the country has almost doubled in the past 25 years. Heart disease now is the leading individual cause of disease burden in India, and stroke is the fifth leading cause.

**Scope:**

The scope of this project is to investigate a dataset of heart disease records for patients using machine learning technique. To identifying heart disease is more difficult. We try to reduce this problem by applying machine learning techniques.

**Objectives:**

* Data validation
* Data Cleaning/ Preparing
* Data Visualization
* Using more algorithm with comparing to predict more accuracy (Like random forest, Decision tree Logistic classification algorithm)

**LITERATURE SURVEY**

**General:**

A literature review is a body of text that aims to review the critical points of current knowledge on and/or methodological approaches to a particular topic. It is secondary sources and discuss published information in a particular subject area and sometimes information in a particular subject area within a certain time period. Its ultimate goal is to bring the reader up to date with current literature on a topic and forms the basis for another goal, such as future research that may be needed in the area and precedes a research proposal and may be just a simple summary of sources. Usually, it has an organizational pattern and combines both summary and synthesis.

A summary is a recap of important information about the source, but a synthesis is a re-organization, reshuffling of information. It might give a new interpretation of old material or combine new with old interpretations or it might trace the intellectual progression of the field, including major debates. Depending on the situation, the literature review may evaluate the sources and advise the reader on the most pertinent or relevant of them.

**Review of Literature Survey:**

1. **PROBLEMS WITH MINING MEDICAL DATA**

In year 2000, research conducted by ShusakuTsumoto says that as we human beings are unable to arrange data if it is huge in size we should use the data mining techniques that are available for finding different patterns from the available huge database and can be used again for clinical research and perform various operations on it.

2. **EVIDENCE COMBINATION IN MEDICAL DATA MINING**

Y. Alp Aslandogan, et. al. (2004), worked on three different classifiers called K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and used Dempsters’ rule for this three viewpoint to appear as one concluding decision. This classification based on the combined idea show increased accuracy.

**3. IMPROVING HEART DISEASE PREDICTION USING CONSTRAINED ASSOCIATION RULES**

Carlos Ordonez (2004), Assessed the problematic to recognize and forecast the rule of relationship for the heart disease. Adataset involving medical history of the patients having heart disease with the aspects of risk factors was accessed by him, measurements of narrowed artery and heart perfusion. All these restrictions were announced to shrink the digit of designs, these are as follows:

1) The features should seem on a single side of the rule.

2) The rule should distinct variousfeatures into the different groups.

3) The count of featuresavailable from the rule is organized by medical history of people having heart disease only. The occurrence or the nonappearance of heart disease was predicted by the author in four heart veins with the two clusters of rules.

**4. PREDICTING SURVIVAL CAUSES AFTER OUT OF HOSPITAL CARDIAC ARREST USING DATA MINING METHOD**

Franck Le Duff (2004), worked on creating Decision tree quickly with clinical data of the physician or service. He suggested few data mining techniques which can help cardiologists in the predication survival of patients. The main drawback of the system was that the user needs to have knowledge of the techniques and we should collect sufficient data for creating an suitable model.

**5. USING EFFICIENT SUPANOVA KERNEL FOR HEART DISEASE DIAGNOSIS**

Boleslaw Szymanski, et. al. (2006), operated on a novel experiential to check the aptitude of calculation of scarce kernel in SUPANOVA. The author used this technique on a standard boston housing market dataset for discovering heart diseases, measurement of heart activities and prediction of heart diseases were found 83.7% correct which were measured with the help of support vector machine and kernel equivalent to it. A quality result is gained by spline kernel with the help of standard boston housing market database.

**6. ASSOCIATIVE CLASSIFICATION APPROACH FOR DIAGNOSING CARDIOVASCULAR DISEASE**

Kiyong Noh, et. al. (2006) made use of a classification technique for removal of multi-parametric structures by accessing HRV and ECG signals. Kiyong used the FP-growth algorithm as the foundation of this technique that is associative. A rule consistency degree was gained which allows a robust press on trimming designs in the method of producing designs.

**7. MININGBIOSIGNAL DATA: CORONARY ARTERY DISEASE DIAGNOSIS USING LINEAR AND NONLINEAR FEATURES OF HRV**

HeonGyu Lee, et. al. (2007), operated for the operation systems of Arithmetical and cataloguing for the addition chief of the multi-parametric feature through direct and nonlinear features of Heart Rate Variability (HRV). The dissimilar classifiers existing are cataloguing grounded on Decision Tree (C4.5), Multiple Association Rules (CMAR) and Bayesian classifiers, and Support Vector Machine (SVM) that are investigated for the valuation of the linear and nonlinear features of the HRV tables.

**8. DECISION SUPPORT SYSTEM FOR HEART DISEASE DIAGNOSIS USING NEURAL NETWORK**

Niti Guru, et. al. (2007), functioned for forecasting of heart disease, Blood Stress and Sugar by the aid of neural systems. Hearings were accepted out on example best ever of patients. The neural system is verified with 13 types, as blood pressure,period, angiography etc. [12].

**Existing System:**

In hospitals doctors use many procedures which are costly to find heart diseases.

These techniques include stress test, chest x-ray, computed tomography (CT), nuclear scanning, echocardiogram (**heart** ultrasound), or magnetic resonance (MR). ... A wide range of modalities may also be used to **monitor** patients with **heart disease**/damage during and after treatment, including: Electrocardiogram (ECG or EKG)

**Drawbacks:**

* Many patients cant afford huge amount for tests.
* Some times even after verifying testing reports its difficult to say that a patient has heart disease or not..

**Proposed System:**

## Exploratory Data Analysis:

## In this section of the report, you will load in the data, check for cleanliness, and then trim and clean your dataset for analysis. Make sure that you document your steps carefully and justify your cleaning decisions.

Training the Dataset:

* The first line imports iris data set which is already predefined in sklearn module. Iris data set is basically a table which contains information about various varieties of iris flowers.
* For example, to import any algorithm and train\_test\_split class from sklearn and numpy module for use in this program.
* Then we encapsulate load\_data() method in data\_dataset variable. Further we divide the dataset into training data and test data using train\_test\_split method. The X prefix in variable denotes the feature values and y prefix denotes target values.
* This method divides dataset into training and test data randomly in ratio of 67:33. Then we encapsulate any algorithm.
* In the next line, we fit our training data into this algorithm so that computer can get trained using this data. Now the training part is complete.

Testing the Dataset:

* Now we have dimensions of a new flower in a numpy array called ‘n’ and we want to predict the species of this flower. We do this using the predict method which takes this array as input and spits out predicted target value as output.
* So the predicted target value comes out to be 0. Finally we find the test score which is the ratio of no. of predictions found correct and total predictions made. We do this using the score method which basically compares the actual values of the test set with the predicted values.

Advantages:

* Our goal is push for assisting doctors and patients using our predictions. All these publications state they have done better than their competitors but there is no article or public mention of their work being used practically to assist the doctors. If there are some genuine problems in rolling out that work to next stage, then identify those problems and try solving them.

Application:

* medical sector to automate to identify the heart disease (real time world) and predicting by desktop application / web application.

Overview of the system:

This helps all others department to carried out other formalities. It have to find Accuracy of the training dataset, Accuracy of the testing dataset, Specification, False Positive rate, precision and recall by comparing algorithm using python code. The following Involvement steps are,

* Define a problem
* Preparing data
* Evaluating algorithms
* Improving results
* Predicting results

The steps involved in Building the data model is depicted below.

**Prediction (**Heart Disease prediction**)**

**Building classification Model**

**Pre Processing** (Outlier Detection, Outlier Removal, Imputation Removal)

**Data collection** (Splitting Training set & Test set)

Fig: data flow diagram for Machine learning model

Project Goals:

# Exploration data analysis of variable identification

* Loading the given dataset
* Import required libraries packages
* Analyze the general properties
* Find duplicate and missing values
* Checking unique and count values

# Uni-variate data analysis

* Rename, add data and drop the data
* To specify data type

# Exploration data analysis of bi-variate and multi-variate

* Plot diagram of pairplot, heatmap, bar chart and Histogram

# Method of Outlier detection with feature engineering

* Pre-processing the given dataset
* Splitting the test and training dataset
* Comparing the Decision tree and Logistic regression model and random forest

# Comparing algorithm to predict the result

* Based on the accuracy

Data collection:

The data set collected for predicting past farmer list of yield is split into Training set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Random Forest , logistic , Decision tree algorithms are applied on the Training set and based on the test result accuracy, Test set prediction is done.

Data Gathering

Data Pre-Processing

Choose model

Train model

Test model

Tune model

Prediction

Fig: Process of dataflow diagram

General Properties:

Create cells freely to explore your data and you should not perform too many operations in each cell. One option that you can take with this project is to do a lot of explorations in an initial notebook. These don't have to be organized, but make sure you use enough comments to understand the purpose of each code cell. Then, after you're done with your analysis, create a duplicate notebook where you will trim the excess and organize your steps so that you have a flowing, cohesive report and make sure that you keep your reader informed on the steps that you are taking in your investigation. Follow every code cell, or every set of related code cells, with a markdown cell to describe to the reader what was found in the preceding cell. Try to make it so that the reader can then understand what they will be seeing in the following cell.

Patient details

Test dataset

Data Processing

Model

Classification ML Algorithm

Training dataset

Fig: Architecture of Proposed model

**Project Requirements**

**General:**

Requirements are the basic constrains that are required to develop a system. Requirements are collected while designing the system. The following are the requirements that are to be discussed.

1. Functional requirements

2. Non-Functional requirements

3. Environment requirements

A. Hardware requirements

B. software requirements

**Functional requirements:**

The software requirements specification is a technical specification of requirements for the software product. It is the first step in the requirements analysis process. It lists requirements of a particular software system. The following details to follow the special libraries like sk-learn, pandas, numpy, matplotlib and seaborn.

**Non-Functional Requirements:**

Process of functional steps,

1. Problem define
2. Preparing data
3. Evaluating algorithms
4. Improving results
5. Prediction the result

**Environmental Requirements:**

1. Software Requirements:

Operating System : Windows

Tool : Anaconda with Jupyter Notebook

2. Hardware requirements:

Processor : Pentium IV/III

Hard disk : minimum 80 GB

RAM : minimum 2 GB

**Performance of Machine Learning parameters:**

Comparison of accuracy results of crop yield prediction:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameters** | **LR** | **DT** | **RF** | **SVC** |
| **Precision** | 0.96 | 1 | 1 | 0.62 |
| **Recall** | 0.96 | 1 | 1 | 1 |
| **F1-Score** | 0.96 | 1 | 1 | 0.77 |
| **Sensitivity** | 0.95 | 1 | 1 | 1 |
| **Specificity** | 0.92 | 1 | 1 | 0 |
| **Accuracy (%)** | 94.73 | 100 | 100 | 62.06 |

Comparison of accuracy results of crop yield cost prediction:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameters** | **LR** | **DT** | **RF** | **SVC** |
| **Precision** | 0.70 | 1 | 1 | 0.69 |
| **Recall** | 0.88 | 1 | 1 | 1 |
| **F1-Score** | 0.78 | 1 | 1 | 0.82 |
| **Sensitivity** | 0.87 | 1 | 1 | 1 |
| **Specificity** | 0.13 | 1 | 1 | 0 |
| **Accuracy (%)** | 65.17 | 100 | 100 | 69.36 |

**Conclusion:**

The analytical process started from data cleaning and processing, missing value, exploratory analysis and finally model building and evaluation. Finally we predict the heart disease using machine learning algorithm with different results. This brings some of the following insights about heart disease prediction. As maximum types of dataset will be covered under this system, doctor may get to know about the disease exactly using ML algorithms, it helps the doctor in decision making weather patient has heart disease or not.

**Future Work:**

* Remaining SMLT algorithms will be involve to finding the best accuracy with applying to predict the crop yield and cost.
* Hospitals wants to automate the detecting the heart disease from eligibility process (real time).
* To automate this process by show the prediction result in web application or desktop application.
* To optimize the work to implement in Artificial Intelligence environment.