**Heart Disease Prediction Using Machine Learning**

Samhita Ghosh (11915175), Student of Computer Science and Engineering at Lovely Professional University.

**Abstract:**

Heart-related diseases, also known as Cardiovascular Diseases (CVDs), have been the leading cause of death in the world during the last several decades, and have risen to become the most life-threatening disease not only in India but worldwide. As a result, a reliable, accurate, and practical approach to diagnosis such disorders in time for adequate treatment is required. It can be prevented with accurate prediction whether a person could get a heart disease or not by analyzing different test result, but it can also be fatal if the prediction is erroneous. Several machine learning algorithms have recently been used by many researchers to aid the health care industry and experts in the detection of heart-related disorders. The results and analyses of the machine learning dataset are compared in this research using various machine learning methods. The dataset contains 14 key attributes that were used in the investigation. The accuracy and confusion matrix are used to validate a number of promising outcomes. The major goal of this research effort is to use machine learning algorithms to anticipate a patient's cardiac condition. The data is also standardized for better results. 88.5% percent accuracy was achieved using Logistic Regression method.

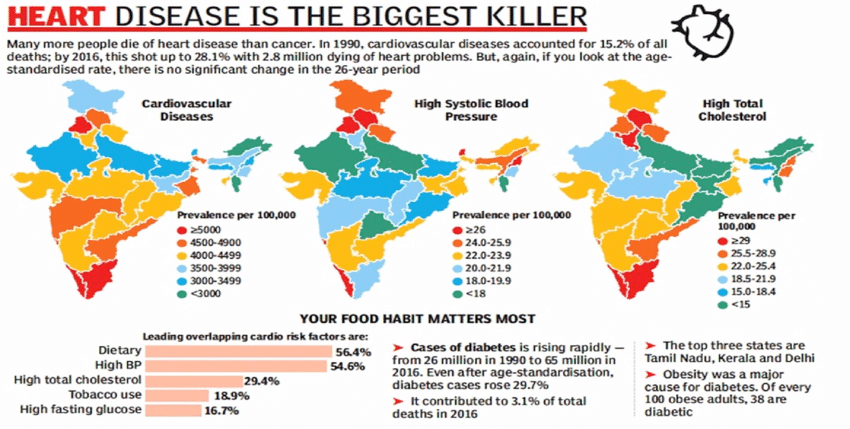
**Introduction:**

The heart is a vital organ in the human body. It delivers blood to every region of our bodies. If it fails to function properly, the brain and other organs will stop operating, and the individual will die within minutes.

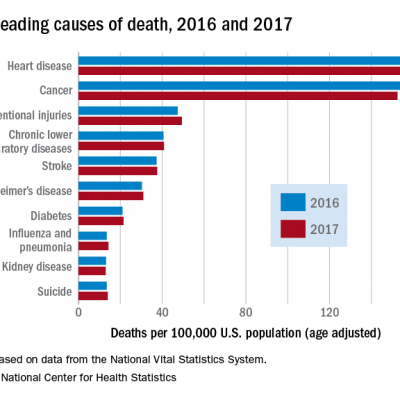
Changes in lifestyle, work-related stress, and poor eating habits all contribute to a rise in the incidence of heart disease. Heart disease has risen to become one of the leading causes of death all over the world. According to the World Health Organization, cardiac illnesses claim the lives of 17.7 million people each year, accounting for 31% of all fatalities worldwide. Heart disease has become the top cause of death in India as well. According to the 2016 Global Burden of Disease Report, which was issued on September 15, 2017, heart disorders killed 1.7 million Indians in 2016. Heart disease raises health-care costs and reduces an individual's productivity.

According to World Health Organization (WHO) estimates, India lost up to $237 billion owing to heart-related or cardiovascular disorders between 2005 and 2015. As a result, it is critical to be able to forecast heart-related disorders in a reliable and precise manner. The term "heart disease" encompasses a wide range of heart conditions. Coronary artery disease is the most prevalent form, which can lead to a heart attack. Other types of cardiac illness may affect the heart's valves, or the heart may fail to pump properly, resulting in heart failure.

Heart disease is a condition that some people are born with. Heart disease can affect anyone, including youngsters. It occurs when a material known as plaque accumulates in your arteries. Heart disease is increased by smoking, poor dietary habits, and a lack of exercise. Heart disease can also be exacerbated by excessive cholesterol, high blood pressure, or diabetes. There are various natural techniques for preventing this condition, such as quitting smoking, maintaining a healthy weight, eating a nutritious diet, and participating in sports on a regular basis. We have scientific methods as well, such as medications and surgery. Data on numerous health-related concerns is collected by medical institutions all over the world. These data can be used to gain meaningful insights utilizing a variety of machine learning techniques. However, the amount of data collected is enormous, and it is frequently noisy. Machine learning approaches can quickly investigate these datasets, which are too huge for human minds to understand. As a result, these algorithms have recently proven to be extremely useful in accurately predicting the presence or absence of heart-related diseases. The use of information technology in the health-care profession is growing by the day to help doctors make better decisions. It aids doctors and physicians in disease management, pharmaceutical development, and the identification of patterns and linkages among diagnosis data. Many people who would benefit from preventative care are missed by current methods for predicting cardiovascular risk, while others receive unneeded treatment. By utilizing intricate connections between risk factors, machine learning provides an opportunity to increase accuracy. We wanted to see if machine learning could help with cardiovascular risk prediction.



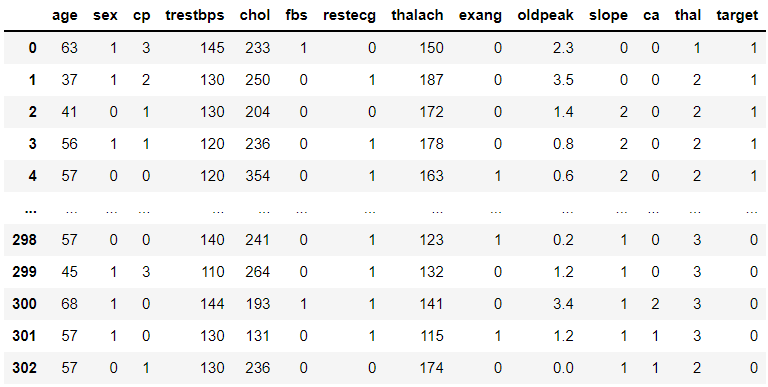
The above figure is the analysis of the 2016 heart disease in India. As I said, there is a 12.9% of increase in the heart disease by 2016 from 1990.



The above is the analysis of the death of people due to heart disease in the year 2016 & 2017 in US. In this also, we see that the number of deaths are due heart disease.

**DATASET:**

On one dataset, I ran a computer simulation. A Heart dataset is a collection of data. Machine Learning Repository hosts the dataset. There are 303 samples in this dataset, including 14 input features and 1 output feature. The characteristics of loan applicants are described in terms of their financial, personal, and social characteristics.



The above picture shows the first 5 and last 5 records of the dataset, that has been used in the research. We see that the main column is the ‘target’ column in the whole data set.

Link for the dataset - https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

**PLANNING:**

Firstly, we take the dataset and store it into the database. Then analyze the dataset by keeping the required data and delete the rest for more accuracy. After that clean the data so that the accuracy to predict the heart disease will not be affected. Then select different machine learning models and check which models provides higher accuracy. Then try some methods in the selected machine learning model to check whether the accuracy could be increased. Test the model with the testing dataset for the prediction.

There is a pictorial representation of the flowchart is been given in the next page.

Analyse and understand the data

Store the dataset into the database for usage.

Dataset

Accuracy Measurement

Flowchart of the complete experiment step by step.

Machine learning models such as Logistic regression, KNeighbour Classifier and Random Forest Classifier

Testing the model for prediction with most accuracy

Training the model for prediction

Splitting the data into train and test

Dataset for training

Dataset for testing

Clean the dataset.

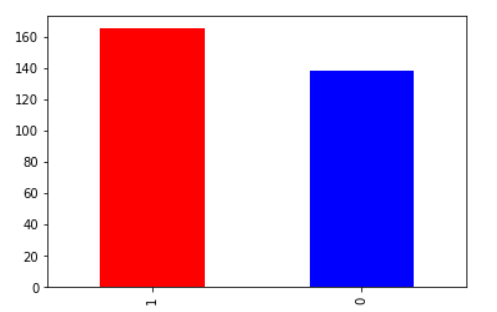
**Analysis:**

The components of the dataset:

1. age: Age of the person is taken in years.
2. sex: 1 => Male ; 0 => Female
3. cp: Chest Pain Type

* 0 => Typical Angina : chest pain caused due to decrease in blood supply to the heart.
* 1 => Atypical Angina : chest pain not related to heart.
* 2 => Non-anginal pain : typically esophageal spasms (not related to heart)
* 3 => Asymptomatic : chest pain not showing signs of disease.

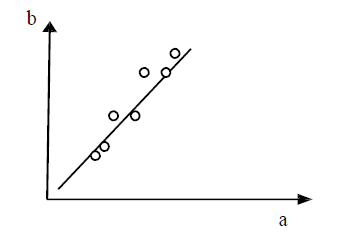
1. trestbps: Resting blood pressure (in mmHg on admission to the hospital) anything above 130-140 is typically cause for concern.
2. chol : Serum cholesterol in mg/dl
   * 1. serum = LDL(Low Density Lipids) + HDL(High Density Lipids) +.2 \* triglycerides
     2. above 200 is a cause for concern
3. fbs: Fasting blood sugar > 120mg/dl
   * 1. 1 => true , 2 -> false
     2. Above 126 mg/dl signals diabetes
4. restecg: Resting electrocardiographic results
   * 1. 0 => Nothing to note
     2. 1 => ST –T wave abnormality ; can range from mild symptoms to severe problems ; signals abnormal heart beats.
     3. 2 => Possible or definite left ventricular hypertrophy ; enlarged heart’s main pumping chamber
5. exang: Exercise induced angina.
   * 1. 0 => no , 1 => yes
6. oldpeak: ST depression induced by exercise relative to rest looks at stress of heart during exercise ; unhealthy heart will stress more
7. slope : the slope of the peak exercise ST segment
   * 1. 0 => Upslopping: better heart rate with exercise (uncommon)
     2. 1 => Flatslopping : minimal change (typical healthy heart)
     3. 2 => Downsloping : signs of unhealthy heart
8. ca: number of major vessels (0 – 3) colored by fluoroscopy colored vessel means the doctor can see the blood passing through ; more blood movement means more better (no clots)
9. thal: thalium stress result
   * 1. 1 – 3: normal
     2. 6: fixed defect
     3. 7: reversible defect



The categorical representation of the ‘target’ column after the data preprocessing.

**Logistic Regression:**

One of the supervised machine learning methods is known as Logistic Regression. It is based on the relationship between dependent and independent variables. The dependent and independent variables are "a" and "b," respectively, and the relationship between them is shown by an equation of line, which is linear in nature, which is why this method is called linear regression.



The relation between a and b by logistic regression.

As shown in Fig.5, it provides a relation equation to forecast the value of a dependent variable "b" based on the value of an independent variable "a." As a result, it can be argued that linear regression technique provides a linear relationship between a and b.

**K-Nearest Neighbor:**

K-Nearest Neighbor (KNN) is a supervise learning-based classification method. It classifies entities that are dependent on their closest neighbors. KNN is a widely used classifier and regression methodology in a variety of fields, including image processing, data processing, pattern recognition, and other applications. The algorithmic program's output is determined by the K-nearest neighbor class, which is enforced by locating K-number of coaching points that are closest to the requested character and considering the votes among the K objects. The algorithmic program is quite simple to use. On the other hand, capable of learning non-linear call boundaries and regression functions that are quite complex. Similar instances should have similar category labels (in classification) or target values, according to KNN's intuition (regression). On the negative side, the algorithmic program is computationally expensive and prone to overfitting.



Pictorial representation of the K-nearest Neighbor method function.

**Random Forest:**

Random Forest is a supervised machine learning technique that performs better in classification tasks than other algorithms. Before producing an output, the Random Forest approach evaluates numerous decision trees, as the name implies. As a result, it's essentially a collection of decision trees.

This method is based on the idea that a larger number of trees would eventually lead to the correct selection. It employs a voting method for classification and then determines the class, whereas it uses the mean of all the decision tree outputs for regression. It works well with huge datasets that have a lot of dimensions.

One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. When it comes to categorization difficulties, it outperforms the competition.



**ACCURACY SCORES:**

|  |  |
| --- | --- |
| Logistic Regression | 0.8852459016393442 |
| K-Nearest Neighbor | 0.6885245901639344 |
| Random Forest | 0.8360655737704918 |

After performing hyperparameter tuning, the accuracy score is:

|  |  |
| --- | --- |
| Logistic Regression | 0.8852459016393442 |
| K-Nearest Neighbor | 0.7540983606557377 |
| Random Forest | 0.8688524590163934 |

Even though the accuracy of K-Nearest Neighbor and Random Forest increased than the previous score but they are still less than the Logistic Regression accuracy score. So, for the prediction of heart disease, we could prefer Logistic Regression for most accuracy.

**CONFUSION MATRIX:**

The confusion matrix is created by various classifier and it includes expected and real classifications information. The confusion matrix provides analysis to judge the effectiveness of proposed methodology.

Where,

The number of actual negative cases in the data = Condition Negative (N)

Condition Negative (N) = Total number of negative cases

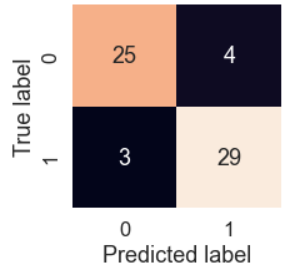
Condition Positive (P) = Total number of positive cases

True Positive (TP) = number of correct positive prediction

True Negative (TN) = number of correct negative prediction

False Positive (FP) = Type I Error, No. of incorrect positive prediction

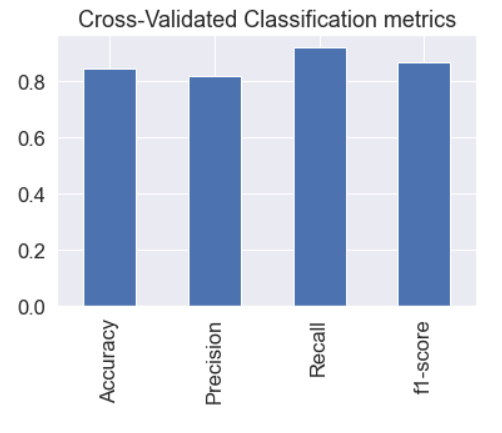
False Negative (FN) = Type II Error, No. of incorrect negative prediction.

****

**CROSS VALIDATION M ATRIX:**

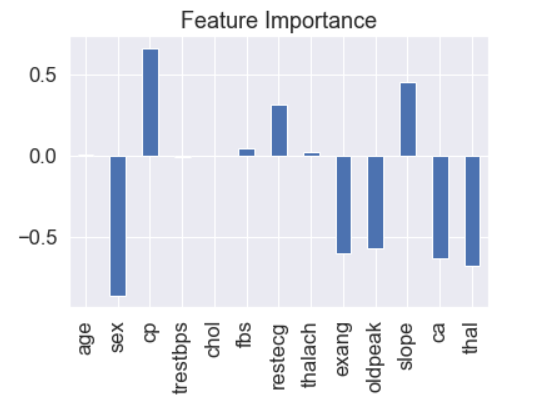
When it comes to developing machine learning models, cross-validation is the first and most important stage. You can develop a solid machine learning model that is highly generalizable if you use a good cross-validation method in which validation data is representative of training and real-world data.

So, we performed the cross-validation matrix for the Logistic Regression model (as we discussed that it is the one with most accuracy compared to the other two machine learning model).



**FEATURE IMPORTANCE:**

Feature Importance refers to methods for calculating a score for each of a model's input features; the scores simply describe the "importance" of each feature. A higher score indicates that a certain feature will have a greater impact on the model used to forecast a given variable.



In our research, the chest pain i.e., cp is the most important feature which is the primary attribute for the prediction of heart disease.

**CONCLUSION:**

In the physical organism, the heart plays a crucial role. Heart disease necessitates greater precision and accuracy in diagnosis and analysis. Heart disorders may not be detectable in their early stages in real time. This has to be looked at more. Using a data set of heart disorders, the suggested work presents a reliable and early heart disease prediction. The provided methodology necessitates the use of a variety of machine learning methods. The analysis is based on the Confusion matrix, and the accuracy of the algorithms is compared, and Logistic Regression is shown to be the best. As a result, the efficacy of the work given has been confirmed. This method could be utilized to aid in the early and precise detection of cardiac disease. There are numerous machine learning techniques that can be employed for finer exploration and early prediction of heart illnesses in the future. This has to be looked at more.

**REFERENCES:**

[1] www.analyticsvidhya.com

[2] www.towardsdatascience.com

[3] w3schools

[4] www.kaggle.com

[5] https://[www.cancer.gov/publications/dictionaries/cancer-terms/def/heart-](http://www.cancer.gov/publications/dictionaries/cancer-terms/def/heart-) disease

[6] https://[www.w3schools.com/python/numpy/numpy\_intro.asp](http://www.w3schools.com/python/numpy/numpy_intro.asp)

[7] https://[www.mygreatlearning.com/blog/python-numpy-tutorial/](http://www.mygreatlearning.com/blog/python-numpy-tutorial/)

[8] https://[www.w3schools.com/python/numpy/numpy\_array\_indexing.asp](http://www.w3schools.com/python/numpy/numpy_array_indexing.asp)

[9] https://pandas.pydata.org/about/index.html

[10] https://[www.geeksforgeeks.org/introduction-to-seaborn-python/](http://www.geeksforgeeks.org/introduction-to-seaborn-python/)