HEART DISEASE PREDICTION USING DIFFERENT MACHINE LEARNING TECHNIQUES

1Srijan Kumar Singh, 2Siddharth Dhawan, 3Tanmay Chordia, 4Bindu Garg
1Student, 2Student, 3Student, 4Professor
1Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
2Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
3Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India,
4Bharati Vidyapeeth Deemed to be University College of Engineering, Pune India

Abstract: Heart-related diseases or cardiovascular diseases (CVDs) have been the main cause of a large number of deaths in the world over the last few decades and have emerged as the most life-threatening illness, not just in India but throughout the world. Therefore, a robust, effective and viable system is required to diagnose these diseases in time for proper medical care. Machine Learning algorithms and methods have been applied to different clinical datasets to mechanize the examination of enormous and complex information. In recent years, several researchers have used many techniques of machine learning to help the health care industry and the experts identify heart-related diseases. This paper presents an overview of different models dependent on such calculations and procedures and analyze their performance. Models based on supervised learning algorithms such as, Gaussian Naïve Bayes, Decision Trees (DT), Support Vector Machines (SVM), Linear SVC, Random Forest (RF) are quite popular among the researchers.

I. INTRODUCTION
Heart is a central organ of the human body. Blood is pumped to every part of our anatomy. If it does not function properly, then the brain and various other organs will cease to operate, and the individual will die within a few minutes. Transition in lifestyle, work-related stress, and poor eating habits lead to elevated levels of multiple heart-related illnesses. Cardiac disorders have emerged as one of the most common causes of death worldwide. Heart-related illnesses are responsible for claiming 17.7 million lives per year, 31 percent of all global deaths, according to the World Health Organization. Heart related diseases have also been the leading cause of mortality in India. According to the 2016 Global Burden of Disease Survey, published September 15, 2017, heart disease has killed 1.7 million Indians in 2016. Heart-related diseases increase health-care costs and can reduce an individual's productivity. The World Health Organization (WHO) figures indicate that from 2005 to 2015, India lost up to $237 billion due to infectious or cardiovascular diseases [2]. Therefore, it is very important to predict heart-related diseases correctly and practically.

Health organizations around the globe are collecting data on various health-related issues. Using various machine learning techniques these data can be used to obtain useful insights. But the data collected is very large, and the data can be very noisy at times. Such datasets, which are too complex for the comprehension of human minds, can be easily examined using various techniques of machine-learning. Therefore, in recent times, these algorithms have become very useful for reliably predicting the existence or absence of heart-related diseases.

II. DIMENSIONALITY REDUCTION
The method of reducing the number of random variables under consideration by acquiring a collection of key variables is in math, machine learning and knowledge theory. Approaches can be broken down into collection of features and extraction of features. The data required for a function or a problem that consist of many attributes or parameters, but not all of these attributes may influence performance equally. A large number of attributes, or features, that affect the complexity of the computation and may even result in over-fitting that results in poor results.

Feature Extraction
In machine learning and analytics, selection of features, also known as selection of component, selection of attributes or selection of component subsets, is the method of choosing a subset of specific features (variables, predictors) for use in model construction. Why is it necessary: -
- Simplifying the templates to make them easy for researchers/users to view.
- To reduce the complexity or training time.
- To avoid the curse of dimensionality phrase.
- Better generalization by minimized overfitting.
Feature Selection

In machine learning, pattern recognition and image processing, the extraction of features starts with an initial set of measured data and constructs derived values (faces) that are intended and non-redundant, enabling the corresponding learning and generalization phases, and in certain cases contributing to better human interpretations. Extraction of the function corresponds to reduction of the dimensionality. Extraction of features means to reduce the amount of resources used to define a large range of details.

III. ALGORITHMS AND TECHNIQUES USED

A. Decision tree

Decision tree is the most effective and common classification and prediction method. A Decision tree is a graph-like flowchart in which each internal node signifies a check on an element, each branch reflects a check outcome, and each leaf node (terminal node) carries a class name. This algorithm splits the population into two or more identical sets depending on the most important predictors. The Decision Tree algorithm measures each and every attribute’s entropy first. The data set is then broken with the help of variables or predictors for maximum gain of knowledge or minimal entropy. Such two measures are achieved recursively for the rest of the attributes.

\[
\text{Entropy}(S) = \sum_{i=1}^{c} -p_i \log_2 p_i
\]

\[
\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)
\]

Using this technique to classify our problem we got the accuracy of 77.05% using all the features in Heart Disease UCI dataset.

B. Random Forest

Random Forest is a popular supervised machine learning algorithm. This method may be used for regression and classification tasks but usually works best in classification tasks. Random forest classifier generates a collection of decision trees from randomly identified training group sub-set. It then aggregates the votes from various decision trees for the final class of the test item to be determined. Basic parameters for the Random Forest Classifier may be total number of trees to produce and parameters relevant to the decision tree such as minimum split, split criterion etc.

Using this technique to classify our problem we got the accuracy of 80.33% using all the features in Heart Disease UCI dataset.

C. Support Vector Machine

"Support Vector Machine" (SVM) is a supervised algorithm for machine learning that can be used both for classification and regression problems. It is primarily used in classification issues, though. In the SVM algorithm, each data object is plotted as a point in n-dimensional space (where n is the number of features you have) with each field being the value of a particular coordinate.
1. **Gaussian radial basis function (RBF)**

It is a general-purpose kernel, used when there is no prior knowledge about the data.

\[ k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \]

Using this technique to classify our problem we got the accuracy of 68.85% using all the features in Heart Disease UCI dataset.

2. **Linear SVM**

Linear SVM is the latest incredibly quick machine learning (data mining) algorithm to solve multiclass classification problems from ultra-large data sets that implements an original patented variant of a cutting plane algorithm to construct a linear vector support system.

Using this technique to classify our problem we got the accuracy of 81.97% using all the features in Heart Disease UCI dataset.

D. **Gaussian Naive Bayes**

Gaussian Naive Bayes methods are a series of supervised learning algorithms focused on the implementation of the theorem of Bayes with the "naive" presumption of conditional independence for each pair of characteristics given the class variable value. The theorem of Bayes notes the following relation, provided the variable class and dependent function vector x1 through x2.

\[ P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(\frac{-(x_i - \mu_y)^2}{2\sigma_y^2}\right) \]

Using this technique to classify our problem we got the accuracy of 85.25% using all the features in Heart Disease UCI dataset.

E. **Logistic Regression**

Logistic Regression is used where the target variable is categorical. Logistic regression name comes from the function used at computational level to compute probabilities; the logistic function also called sigmoid function.

\[ P = \frac{1}{1 + e^{-(a+bX)}} \]

Statisticians have been established to explain the features of population growth in biodiversity, increasing increasingly and optimizing environmental efficiency. This is an S-shaped curve that can take any number evaluated and map it to a value between 0 and 1.

Using this technique to classify our problem we got the accuracy of 85.25% using all the features in Heart Disease UCI dataset.
IV. CONCLUSION

Based on the above analysis, it can be concluded that predicting cardiovascular diseases or heart-related diseases offer substantial scope for machine learning algorithms. Gaussian Naïve Bayes and Logistic Regression, performed exceptionally well with 85.25% of accuracy whereas Decision trees performed very poorly with mere accuracy of 77.07%. Random Forest (80.33%) and SVM (81.97%) models have performed moderately well, as they overcome the problem of overfitting by using multiple algorithms (Random Forest multiple decision trees). Models based on the Naïve Bayes classifier were very efficient in computational terms. Systems based on machine learning algorithms and techniques were very accurate in heart-related predictions.

REFERENCES