Prediction of Heart Disease Using Machine Learning

Akash Chandra Patel¹, Anash Shameem², Sunil Chaurasiya³, Manish Mishra⁴, Prof. Abhishek Saxena⁵

¹,²,³,B.Tech (CSE) Final Year Students, ⁵Assistant Professor
Department of CSE.
School of Engineering & Technology,
Noida International University, (U.P.)

Abstract: In today’s environment Heart disease is a major life-threatening disease that can cause either death or a serious long term disability [1] since its diagnosis in most cases depends upon a complex combination of clinical and pathological data. Since medical diagnosis is a complicated task and plays a vital role in saving human lives so it needs to be executed accurately and efficiently but there is lack of effective tools to discover hidden relationships and trends in e-health data. Due to this complexity, an appropriate and accurate computer-based automated decision support system is required to reduce the cost for achieving clinical tests. In this paper, we proposed a Heart Disease prediction system that can assist medical professionals in predicting the status of heart disease based on the clinical data of patients. The main objective of this study is to build a model that can predict the occurrence of heart disease, based on a combination of features (risk-factors). Different machine learning classification techniques will be implemented and compared upon standard performance metric such as accuracy for comparison between different machine learning algorithms.

Keywords: Diagnosis, Heart Disease, Machine Learning, Classification.

I. INTRODUCTION

The heart is one of the most important organs in the human body. It is the center of the circulatory system. The functions heart include pumping the blood to different parts of the human body through a network of blood vessels, supplying a constant supply of oxygen as well as other vital nutritional components [2]. If the heart ever stops functioning and ceases to pump blood, the body will shut down and within a fraction of time a person will lose his life. The usage of information technology in the health care industry is increasing day by day to aid doctors in decision-making activities. It helps medical practitioners in disease management, medications, and discovery of patterns and relationships among diagnosis data. Current approaches to predicting cardiovascular risk fail to identify many people who would benefit from preventive treatment, while others receive unnecessary intervention [3]. Machine-learning offers the opportunity to improve accuracy by exploiting complex interactions between risk factors.

II. METHODOLOGY

Below steps shows the method through which the Heart disease prediction model has developed.

A. Data Collection

Data has been collected from the UCI machine learning repository named as “Heart Disease Data Set”[4].

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>chest_pain</th>
<th>bloodPressure</th>
<th>serum_cholesterol</th>
<th>fasting_blood_sugar</th>
<th>electrocardiographic</th>
<th>max_heart_rate</th>
<th>induced_angina</th>
<th>ST_depression</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>145.0</td>
<td>233.0</td>
<td>1.0</td>
<td>2.0</td>
<td>150.0</td>
<td>0.0</td>
<td>2.3</td>
</tr>
<tr>
<td>1</td>
<td>1.0</td>
<td>4.0</td>
<td>160.0</td>
<td>286.0</td>
<td>0.0</td>
<td>2.0</td>
<td>108.0</td>
<td>1.0</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>1.0</td>
<td>4.0</td>
<td>120.0</td>
<td>229.0</td>
<td>0.0</td>
<td>2.0</td>
<td>120.0</td>
<td>1.0</td>
<td>2.6</td>
</tr>
<tr>
<td>3</td>
<td>1.0</td>
<td>3.0</td>
<td>130.0</td>
<td>250.0</td>
<td>0.0</td>
<td>0.0</td>
<td>187.0</td>
<td>0.0</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>0.0</td>
<td>2.0</td>
<td>130.0</td>
<td>204.0</td>
<td>0.0</td>
<td>2.0</td>
<td>172.0</td>
<td>0.0</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table 2.1 Actual Dataset

B. Data Preprocessing

The dataset used in this project contains 14 variables. The independent variable that needs to be predicted, ‘diagnosis’, determines whether a person is healthy or suffer from heart disease. Experiments with the Cleveland database have concentrated on endeavors to distinguish disease presence (values 1, 2, 3, 4) from absence (value 0). There are several missing attribute values, distinguished with the symbol '?'. The header row is missing in this dataset, so the column names have to be inserted manually. The missing values are numerical entries so we have filled the values by taking Mode of the columns.
C. Building Model

The predictor model has been built using Jupyter Notebook, an interactive tool for data analysis and machine learning. It is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations. It is used for data cleaning, transformation, numerical simulation, statistical modeling, data visualization, machine learning[5]. The data used in the model is normalized for better accuracy of the model.

D. Various Techniques

**Support Vector Machine (SVM):** Support Vector Machine is a supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier.

Given some training data \( D \), a set of \( n \) points of the form

\[
D = \{(X_i, y_i) \mid X_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^n
\]

Where the \( y_i \) is either 1 or -1, indicating the class to which the point \( X_i \) belongs. Each \( X_i \) is a \( p \)-dimensional real vector. We want to find the maximum-margin hyperplane that divides the points having \( y_i = 1 \) from those having \( y_i = -1 \). Any hyperplane can be written as the set of points \( x \) satisfying, \( w \cdot x - b = 0 \), where \( \cdot \) denotes the dot product and \( w \) the normal vector to the hyperplane. The parameter \( b/\|w\| \) determines the offset of the hyperplane from the origin along the normal vector \( w \).

If the training data are linearly separable, we can select two hyperplanes in a way that they separate the data and there are no points between them, and then try to maximize their distance. The region bounded by them is called "the margin". These hyperplanes can be described by the equations [6]

\[
w \cdot x - b = 1, \text{ and } w \cdot x - b = -1
\]

**Decision Tree:** Decision tree are a supervised method used for the prediction of categorical as well as the numerical value they represent the data instances along with their class label in the form of a tree. A set of rules can be deduced from the tree which can be used to classify the unknown data record to its output value. A test on an attribute is performed on the internal node. The result of the test is depicted by the branch of tree and class label are present at the leaf node. In this technique, the whole data set or the
whole collection of sample points in split into two or more homogenous classes. The split is established from the parameter or the factor which is determined to be the best splitter or differentiator [7].

**Naïve Bayes:** Rather than a single classifier it actually is a combination of multiple classifiers all working on the basic Naïve Bayes principle of independent features. Hence each feature is assumed to be independent and autonomous contributing individually to the training data point’s probability of belonging to a particular class. As per the Bayes theorem,

\[ P(c|x) = \frac{P(x|c)P(c)}{P(x)} \]

Here \( P(c|x) \) is the posterior probability of the class given predictor \( P(c) \) is the prior probability of class \( P(x|c) \) is the likelihood probability of predictor given class \( P(x) \) is the prior probability of predictor.

**KNN:** \( K \)-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure \( k \)-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the \( k \) closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its \( k \) nearest neighbors (\( k \) is a positive integer, typically small). If \( k = 1 \), then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of \( k \) nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of \( 1/d \), where \( d \) is the distance to the neighbor [8].

**Random decision forests:** This is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

**E. Choosing best predictor**

The Similarity score has been used to compare performance parameters of different machine learning model considered for the comparison of accuracy. The model build by Random Forest has an accuracy of 0.89 SVM has an accuracy of 0.86 for a given set of test data. The Random Forest model is giving a higher number of correct YES or NO value as Heart Disease prediction output against certain values of input parameters as compared to other machine learning algorithms.

**III. RESULT**

Analysis of data shows that the performance of Random Forest Machine is more accurate than other Machine learning Algorithms (Table 3.1) for the specified testing data set and sample training data set used in this study. This study was based on the idea that values used for various parameters affecting heart disease. Random forests model which is having a low error rate has proven to be useful in Hearth disease prediction among other prediction techniques.

<table>
<thead>
<tr>
<th>Models</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN</td>
<td>0.868132</td>
</tr>
<tr>
<td>Decision Trees</td>
<td>0.791209</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.857143</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.868132</td>
</tr>
<tr>
<td>SVM</td>
<td>0.879121</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.890110</td>
</tr>
</tbody>
</table>

**Table 3.1 Accuracy of Models**
Different observation has also been recorded that is illustrated as follows

Table 3.2 Diseases Based on Sex

IV. CONCLUSION

It has been observed by the above observation that men are much more prone to get heart disease than women. The higher number of vessels detected through fluoroscopy confirms the higher risk of diseases. While soft chest pain may be a bad symptom of approaching problems with heart (especially in the case of men), strong pain is a serious warning. The flat slope (value = 2) and downslope (value = 3) of the peak exercise indicates a high risk of getting the disease.

REFERENCES

[3] https://www.ahajournals.org/doi/full/10.1161/01.cir.97.18.1837