DISEASE PREDICTION SYSTEM – SMART HEALTHCARE TOOL using Machine Learning Algorithm

Manisha Mohan Pancholi ¹, Prachi Priya², Shubham Raina ³, Chakradhar Panchal⁴
¹,²,³, BE Student, Department of Computer Engineering, Sinhgad Institute Of Technology, Lonavala, Maharashtra, India

Abstract:

The world is moving at a fast speed and to keep up ourselves with the whole world, we tend to ignore the symptoms of disease which can affect our health to a large extent. Many working professionals get a heart attack, bad cholesterol, eye disease and they are unable to treat it at the right time as they are busy coping up with the progressive world. God has granted every individual a beautiful gift called life, so it is our responsibility to live our life to the fullest and try to stay safe from the dangers of the world. So we have developed a logistic regression model for Covid-19 with the machine learning algorithms like decision tree, a random forest which takes into account the symptoms felt by a person, and according to those symptoms it predicts the disease which the person, can be suffering from. It saves time as well as makes it easy to get a warning about your health before it’s too late.

Key Words: Machine Learning, Kaggle, UCI-REPOSITORY, Decision tree, logistic regression, CKD, COVID-19.

1. INTRODUCTION

Today, the healthcare industry has become a big money-making business. The healthcare industry uses and produces quite a large amount of data that can be used to extract information about a particular disease for a patient. This information of healthcare will further be used for effective and best possible treatment for patient's health. This area also needs some improvement by using informative data in healthcare sciences. But a major challenge is to extract the information from the data because the data is present in a huge amount so some data mining and machine learning techniques are used. The expected result and of this project is to predict the disease so that the risk of life can be prevented at an early stage and save the life of people and the cost of treatment can be reduced to a particular extent. In India also we should adopt the non-manual system of medical treatment which is the best for improving and understanding human health. The main motive is to use the concept of machine learning in healthcare to improvis the treatment of patients. Machine learning has already made it much easier to identify and predict various diseases. Predictive analysis of the disease with the help of many machine learning algorithms helps us to predict the disease and helps in effectively treating the patients. Disease prediction using machine learning also uses the patient history and health data by applying various concepts like data mining and machine learning techniques and also some algorithm. Many works have also applied data mining techniques to the pathological data for the prediction of some particular diseases. These approaches were intended to beforehand predict the reoccurrence of certain diseases. Also, some approaches tried to do prediction while controlling the disease. The recent work of deep learning was in disparate areas of machine learning which have driven a shift to machine learning models that can learn and understand the hierarchical representations of raw data with some pre-processing. With the development of this concept called big data technology, more attention is paid to disease prediction.

2. OBJECTIVES

Nowadays, people face various diseases due to environmental condition and their living habits. So the prediction of disease at an earlier stage becomes an important task. But the accurate prediction based on symptoms becomes too difficult for the doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role to predict the disease. Medical science has a large amount of data growth per year. Due to the increasing amount of data growth in the medical and healthcare field the accurate analysis on medical data has been benefiting from early patient care. With the help of disease data, data mining finds hidden pattern information in a huge amount of medical data. We proposed general disease prediction based on the symptoms of the patient. For the disease prediction, we use Decision tree Classifier for Chronic Kidney Disease and general disease Prediction, Logistic regression for Covid-19 prediction machine learning algorithm for accurate prediction of disease, Random forest for Cardiovascular disease prediction. For disease prediction required disease symptoms dataset. In this general disease prediction, the living habits of a person and checkup information consider for the accurate prediction. The accuracy of general disease prediction by using a Decision tree is 97% which is more than the
Logistic regression algorithm. Similarly, accuracy is the measure for each disease using a confusion matrix.

2. OVERVIEW

The dataset we have considered consists of 132 symptoms, the combination or permutations of which leads to 41 diseases. Based on the 4920 records of patients, we aim to develop a prediction model that takes in the symptoms from the user and predicts the disease he is more likely to have. We created 4 models in our system:

- Chronic Kidney Disease Predictor
- Corona Virus (COVID-19) Predictor
- Cardiovascular Disease (CVD) Predictor
- General Disease Predictor

The considered symptoms are:

Fig. 1. Attributes Used in General Disease Prediction System.

2.1 General Disease Predictor:
Dataset has taken for general disease prediction system from UCI-Repository which contain records of more than 5000 patients. The accuracy of 97% is archived through the Decision Tree algorithm.

Fig. 2. Dataset from UCI-Repository for General Disease Prediction

2.2 Chronic Kidney Disease (CKD) Predictor:
Dataset used for kidney disease prediction from kaggle.com which contains records of more than 400 patients. The Algorithm used to predict kidney disease is the Decision tree it provides 100% accuracy. Attributes choose for this model are sugar christening, Albumin, Hypertension, Packed volume cell, Serum christening, Specific gravity, etc.

Fig. 3. Dataset from kaggle for prediction of Kidney Disease

2.3 Cardiovascular Disease (CVD) Predictor:
Heart Disease is the most common disease in today's era. There are lots of people who die every year due to these diseases. To overcome the death rate we predicted this system with nearly 98% accuracy result, we use here a Random Forest Machine learning algorithm. Dataset chooses from Kaggle.

If age=<30 and Overweight=no and Alcohol Intake=never. Then
Heart Disease level is Low.
(OR)
If Age=> and Blood Pressure=High and smoking=current
Then
Heart Disease level is High.

Fig. 4. Heart Disease Prediction Using Random Forest.

2.4 Corona Virus (COVID-19) Predictor:
As we all know in the current pandemic situation it's too risky to go outside. So, we created this model with 96% accuracy to predict COVID-19 diseases at one touch. The attribute used in these models are breathing, Cough, Cold, Fever, Oxygen level, travelled history etc. Logistic regression algorithm gives the best result for this model.

Fig. 5. Covid-19 Predictor

3. METHODOLOGY

Disease prediction system follows following steps:

Fig. 6. PREDICTION MODEL

A. Input (Symptoms):
While designing the model we have assumed that the user has a clear idea about the symptoms he is experiencing. The Prediction developed considers 95 symptoms amidst which the user can give the symptoms his processing as the input.

B. Data preprocessing:
The data mining technique that transforms the raw data or encodes the data to a form which can be easily interpreted by the algorithm is called data preprocessing. The preprocessing techniques used in the presented work are:

• Data Cleaning: Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data.
• Data Reduction: The analysis becomes hard when dealing with huge database. Hence, we eliminate those independent variables (symptoms) which might have less or no impact on the target variable (disease). In the present work, 95 of 132 symptoms closely related to the diseases are selected.

C. Models selected:

The system is trained to predict the diseases using three algorithms: Disease Tree Classifier x Random forest Classifier x Naïve Bayes Classifier. A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database.

D. Output (diseases):

Once the system is trained with the training set using the mentioned algorithms a rule set is formed and when the user the symptoms are given as an input to the model, those symptoms are processed according to the rule set developed.

3.1. ALGORITHMS USED:

A. Decision Tree Classifier:

The classification models built by decision tree resemble the structure of tree. By learning the series of explicit if-then rules on feature values (symptoms in our case), it breaks down the dataset into smaller and smaller subsets that results in predicting a target value (disease). A decision tree consists of the decision nodes and leaf nodes.

• Decision node: Has two or more branches. In our work presented, all the symptoms are considered as decision nodes.
• Leaf node: Represents the classification that is, the Decision of any branch. Here the Diseases correspond to the leaf nodes.

B. Logistic Regression:

It is a machine learning approach, which can be used for classification problem. In a paper published in Proceedings of Belgian Royal Academy, Pierre Francais Verhulst described the logistic function and its properties by defining three parameters and the curve passing through these. [5]. It is very simple and one of the most used machine learning algorithms. Logistic regression is a statistical model for predicting binary classes. The maximum likelihood Estimation model is used in logistic regression. The dependent variable here follows the Bernoulli distribution. Logistic function or sigmoid function is an ‘S’ shaped curve that takes a value between 0 and 1. 1 will be predicted if the curve goes to positive infinity and 0 if it goes to negative infinity. Figure 1 is a representation of a logistic function.

C. Random Forest Algorithm:

Random forest is a flexible, easy to use machine learning algorithm that provides exceptional results most of the time even without hyper-tuning. As mentioned in the Decision tree, the major limitation of the decision tree algorithm is overfitting. It appears as if the tree has memorized the data. Random Forest prevents this problem: It is a version of ensemble learning. Ensemble learning refers to using multiple algorithms or the same algorithm multiple times. Random forest is a team of Decision trees. And greater the number of these decision trees in Random forest, the better the generalization. More precisely, Random forest works as follows:

1. Selects k symptoms from dataset (medical record) with a total of m symptoms randomly (where k<<m) Then, it builds a decision tree from those k symptoms.
2. Repeats n times so that we have n decision trees built from different random combinations of k symptoms (or a different random sample of the data, called bootstrap sample).

Fig. 7. DECISION TREE FOR GENERAL DISEASE PREDICTION

Fig. 8. LOGISTIC REGRESSION MODEL FOR COVID-19
3. Takes each of the n-built decision trees and passes a random variable to predict the Disease. Stores the predicted Disease, so that we have a total of n Diseases predicted from n Decision trees.

4. Calculates the votes for each predicted Disease and takes the mode (most frequent Disease predicted) as the final prediction from the random forest algorithm.

4. RESOURCES USED

4.1 Hardware:
1. Processor: Minimum 2.0GHz requires.
2. Ram: 2 GB.
3. Hard Disk: 100 GB.
5. Output device: VGA and High-Resolution Monitor

4.2 Software:
2. Language: Python
3. Algorithm Used: Decision Tree, Random Forest, Logistic regression

5. RESULT AND ANALYSIS

When person having chances to particular diseases it display the following result:

Fig. 9. Person causes Disease

When person is safe for does not causes any disease it display the following result:

Fig. 10. Person does not causes Disease or person is safe

The results obtained by applying Random Forest, Decision Tree and Logistic Regression are shown in this section. The metrics used to carry out performance analysis of the algorithm are Accuracy score, Precision (P), Recall (R) and F-measure. Precision (mentioned in equation (2)) metric provides the measure of positive analysis that is correct. Recall [mentioned in equation (3)] defines the measure of actual positives that are correct. F-measure accuracy.

(1)Precision = (TP) / (TP + FP )
(2) Recall = (TP) / (TP+FN)
(3) F- Measure =(2 * Precision * Recall) / (Precision +Recall)
(4) Accuracy=TP+TN/TP+FP+FN+TN

• TP (True Positive): The patient has the disease and the test is positive.
• FP (False Positive): The patient does not have the disease but the test is positive.
• TN (True Negative): The patient does not have the disease and the test is negative.
• FN (False Negative): The patient has the disease but the test is negative.

<table>
<thead>
<tr>
<th>Diseases</th>
<th>Heart Disease Prediction</th>
<th>General Disease Prediction</th>
<th>Corona Virus Disease Prediction</th>
<th>Kidney Disease Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>0.696533</td>
<td>-</td>
<td>0.921111</td>
<td>0.94657</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.99971</td>
<td>0.98154</td>
<td>-</td>
<td>1.00000</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.97568</td>
<td>-</td>
<td>-</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Fig. 11. Accuracy Measure using confusion Matrix

7. CONCLUSION AND FUTURE WORK

In the proposed system, hidden knowledge will be extracted from the historical data by preparing datasets by applying the Naïve Bayes algorithm. Predicting smart health can be done only the system responds that way. These datasets will be compared with the incoming queries and the final report will be generated using Association Rule
Mining. Since this proposed methodology will work on real historical data, it will provide accurate and efficient results, which will help patients get diagnosis instantly. This system will also guide the users on how to remain healthy and fit using tips provided here. The further enhancements that can be done would be integrating this web application into an Android app. This will be available to users on a mobile basis and its use can be further increased. Also feature like getting the doctor online on chat so that patients can directly talk to the concerned doctors. The modules doing cancer analysis can be integrated to find how close the person associated with cancer is. This will make this web application predictable in a true sense.

7. REFERENCES


