HEART DISEASE PREDICTION USING MACHINE LEARNING

[1]Rishabh Magar, [2]Rohan Memane, [3]Suraj Raut
[1]Prof. V. S. Rupnar
[1]Computer Department,
[1]MMCOE, Pune, India.

Abstract : Heart disease is one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines Statistical analysis machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases data set from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques.

***IndexTerms* - Machine learning (ML), support vector machines (SVM), supervised learning.**

I. INTRODUCTION

A. Basics and backgrounds

Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases dataset from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques.

B. Literature Survey

Senthilkumar Mohan have suggested mine to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced ML techniques can help remedy this situation. This research concludes with various models that can be used for prediction.

Anjan N. Repaka stated the performance of prediction for two classification models, which is analyzed and compared to previous work. Experimental results show the improved accuracy percentage of risk prediction of our proposed method compared to other works**.**

Aditi Gavhane addresses the issue of prediction of heart disease according to input attributes on the basis of various data mining techniques and represented them with their accuracy in tabular format. It proposes to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

Santhana Krishnan predicts the arising possibilities of Heart Disease. The outcomes of this system the chances of occurring heart disease in terms of percentage. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using four main Machine Learning Algorithm Namely Decision Tree, Logistic Regression, Support

Vector Machine and Naive Bayes Algorithm which shows the best algorithm among these two in terms of accuracy level of heart disease.

II. PROPOSED SYSTEM

It is a web-based machine learning application which is trained by a UCI dataset. The user inputs its specific medical details to get the prediction of heart disease for that user. The algorithm will calculate the probability of presence of heart disease. The result will be displayed on the webpage itself. Thus, minimizing the cost and time required to predict the disease.

Format of data plays crucial part in this application. At the time of uploading the user data application will check its proper file format and if it not as per need then ERROR dialog box will be prompted.

There will be the following four algorithms implemented:

- Support Vector Machine (SVM)
- Decision Tree
- Naïve Bayes Algorithm
- Logistic Regression

The working of these algorithms has been explained in the sections ahead.

The algorithms have been trained using the data set obtained from University of California, Irvine.75% of the entries in the data set have been used for training and the remaining 25% for testing the accuracy of the algorithm. Furthermore, some steps have been taken for optimizing the algorithms thereby improving the accuracy. These steps include cleaning the dataset and data preprocessing.

The algorithms were judged based on their accuracy and it was observed that the SVM was the most accurate out of the three with 64.4% efficiency. Hence, it was selected for the main application.

The main application is a web application which accepts the various parameters from the user as input and computes the result. The result is displayed along with the accuracy of prediction.

Inputs: Data set, User Data

Outputs: Prediction of disease successfully.

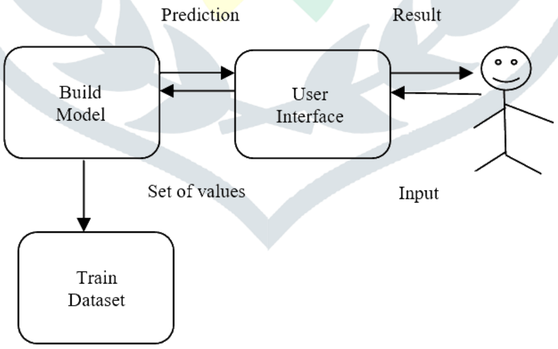


Fig (I). Architecture Diagram

III. PROCESS FLOW

1. Start
2. Input the details
    a. Check the format of details
    b. Process the details
    c. Ignore Spaces as Delimiters
3. Train Dataset
4. Prediction of the result using:
    - Support Vector Machine algorithm
    - Decision Tree

   - Naïve Bayes Algorithm
   - Logistic Regression
5. Display result
6. END

IV. METHODOLOGY

A. Data Preprocessing

The dataset we obtained was not completely accurate and error free. Hence, we first carried out the following operations on it:
Data Cleaning

- NA values in the dataset were the major setback for us as it was reducing the accuracy of the prediction profoundly so, we removed the fields which had NA value. We substituted it with the mean value of the column. This way, we removed all the NA values in the data set.
- Feature Scaling
- Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without feature scaling. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be scaled so that each feature contributes approximately proportionately to the final distance. So, we scaled the various fields in order to get them closer in terms of values.
- Factorization
- In this section, we assigned a meaning to the values so that the algorithm doesn't confuse between them. For example, assigning meaning to 0 and 1 in the age section so that the algorithm doesn't consider 1 as greater than 0 in that section.

B. Support Vector Machine

Support vector machine (SVM) are supervised learning method that analyze data used for classification and regression analysis. It is given a set of training data, marked as belonging to either one of two categories, an SVM training algorithm then builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier . An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. The points are separated based on hyper plane that separate them.

C. Decision Tree

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data (training data).In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

D. Naïve Bayes

Naive Bayes is a family of probabilistic algorithms that take advantage of probability theory and Bayes' Theorem to predict the tag of a text (like a piece of news or a customer review). They are probabilistic, which means that they calculate the probability of each tag for a given text, and then output the tag with the highest one. The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

Bayes' Theorem is useful when working with conditional probabilities, for this we use formula:

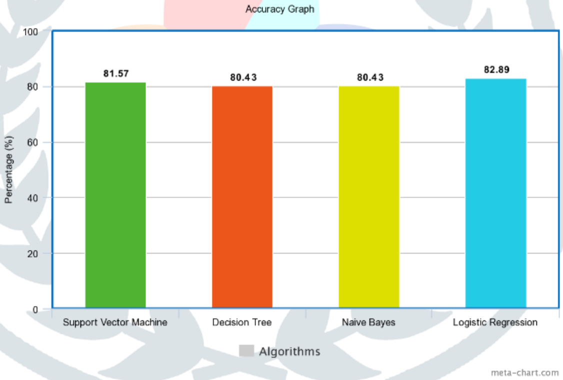
$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

E. Logistic Regression

Logistic regression algorithm also uses a linear equation with independent predictors to predict a value. The predicted value can be anywhere between negative infinity to positive infinity. We need the output of the algorithm to be class variable, i.e. 0-no, 1-yes. Therefore, we are squashing the output of the linear equation into a range of [0, 1]. To squash the predicted value between 0 and 1, we use the sigmoid function.

IV. RESULT

The heart disease prediction system evaluates the risk of disease for patient. HDPS gives result on the basis of accuracy in percentage factor. This percentage shows the accuracy for a particular trained model on user given data values. The following chart describes the accurate behaviour of each algorithm. Green, Red, Yellow, Blue pillars in the following diagram represents result for SVM, Decision Tree, Naïve Bayes, Logistic Regression respectively. As predicted result which is given by HDPS is in bivalued in nature i.e. (Yes/No).



V. COCLUSION

At first, the four algorithms were implemented. Datasets were trained for all the algorithms individually. After this, all of them were tested. The most efficient algorithm was to be selected based on various criteria. We found out that Logistic Regression algorithm has the most efficient out of the four with an accuracy of 82.89%. Decision tree and Naïve Bayes had accuracy of 80.43% and 80.43% respectively, and SVM was having 81.57% Thus these four algorithms were further implemented using a better user interface. In this, HTML, CSS and Django framework of Python were used to build interactive web application. And this web application with machine learning algorithms forms a robust model to predict a heart disease. This would help the end users get a preliminary prediction about the condition of their heart. Since heart diseases are a major killer in India and throughout the world, application of a promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on the society.

VI. FUTURE WORK

Using the machine learning concept newly trained dataset can be used for an even more accurate prediction system. Accounts can be created for each user and then by referring the past choice history of user's heart condition can be monitored to tell if there is any improvement or if the condition has deteriorated.

REFERENCES

[1] A. N. Repaka, S. D. Ravikanti and R. G. Franklin, "Design and Implementing Heart Disease Prediction Using Naive Bayesian," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 292-297.

[2] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019.

[3] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), CHENNAI, India, 2019, pp. 1-5.

[4] A. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1275-1278.

[5] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4.

[6] C. T. and A. Choudhary, "Heart Disease Diagnosis using a Machine Learning Algorithm," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vellore, India, 2019, pp. 1-4.