Multiple Disease Prediction using Streamlit

Jadala Srilipi [1], Kesa Sruthi[2], Vodyati Vyshnavi[3], Dr Sreedhar Bhukya[4]
[1,2,3] *Sreenidhi Institute of Science and Technology, Hyderabad*
[4]*Dr Sreedhar Bhukya, Professor, Sreenidhi Institute of Science and Technology, Hyderabad*

***Abstract*—Wellbeing is one of the significant variables to be considered by an individual. Healthcare falls under the essential conveniences to be given to the society. Many of the current AI models for medical services examination are focusing on one disease prediction for each analysis. Our point is to anticipate the various sorts of illness in single stage by utilizing inbuilt python module Streamlit. In this task we are utilizing Naïve Bayes algorithm, random forest, decision tree and svm classifier are utilized for prediction of a particular disease .The calculation which gives more accuracy is used to train the data set before implementation. To implement multiple disease analysis used machine learning algorithms, Streamlit and python pickling is utilized to save the model behaviour. In this article we analyse Diabetes analysis, Heart disease and parkinson's disease by using some of the basic parameters such as Pulse Rate, Cholesterol, Blood Pressure, Heart Rate, etc., and also the risk factors associated with the disease can be found using prediction model with good accuracy and Precision. Further we can include other kind of chronic diseases, skin diseases and many other. In this work, demonstrated that using only core health parameters many diseases can be predicted. The significance of this analysis to analyse the maximum diseases to screen the patient's condition and caution the patients ahead of time to diminish mortality proportion.**

Keywords*—Prediction, Random forest, Decision Tree, SVM Classifier, Exploratory Data Analysis, Machine Learning.

I. INTRODUCTION

The medical services industry can go with a successful choice by "mining" the huge data set they have for example by extracting the hidden relationships and connections in the data set. Data mining algorithms like Decision Tree, Random Forest and Naïve Bayes calculations can give a solution for this present circumstance. Thus, we have developed a computerized framework that can discover and extract hidden knowledge associated with the diseases from a historical(diseases-side effects) data set by the standard arrangement of the particular algorithm.

The medical care and clinical area are more in need of data mining today. At the point when certain information mining strategies are utilized in a correct manner, significant data can be removed from enormous data sets and that can assist the clinical specialist with taking early choice and further develop healthcare administrations. The spirit is to use the classification in order to assist the physician.

During a ton of examinations over existing frameworks in medical services, examination thought about just a single sickness at a time. Most extreme articles center around a specific sickness. At the point when any association needs to break down their patient's well being reports then they need to send many models. The methodology in the current framework is helpful to dissect just specific illnesses. These days mortality has expanded because of not distinguishing the specific infection. Indeed, even the patient who got restored from one sickness might be experiencing another infection. Inside experiencing heart issues which are not distinguished. Like this many occasions are seen in many individuals' life stories.In numerous sickness expectation frameworks a client can break down more than one illness on a solitary site. The client doesn't have to cross better places to foresee whether he/she has a specific infection or not. In this, the client needs to choose the name of the specific illness, enter its boundaries and simply click on submit. The comparing AI model will be summoned and it will anticipate the result and show it on the screen.

II. LITERATURE SURVEY

There have been various examinations done connected with predicting the disease using different Techniques and algorithms which can be used by Healthcare centers. This paper reviews on the strategies and results used by the research papers:

These are some of the techniques used.

1) Sateesh Ambesange [1] detected the health parameters by various sensors,The Arduino boards processed the data received from the sensors and demonstrated the prediction of Diabetes, using only core health parameters and compared the results with the complete PIDD data set ,resulted in 81.91% precision for KNN algorithm 81.81%
2) Akkem Yaganteeswarudu [2] conducted comparative study on the effectiveness of Decision Tree, Random Forest and logistic regression algorithms in predicting multi Disease which resulted in logistic regression results 92% accuracy, for heart disease classification Randomforest yield 95% accuracy and for cancer detection SVM yield 96 % accuracy.
3) Chetan Sagarnal [3] in this the algorithms are selected, the symptoms are processed, and the disease is predicted which is resulted with 95.12%
4) Nuzhat F.Shaikh [4] In the visualization of the modules by different techniques for understanding and algorithm selected for comparison basis of accuracy and time taken for the class labels with the best accuracy 98.12 by J48 algorithm.
5) Rashmi G Saboji et al, [5] tried to find a scalable solution that can predict heart disease utilizing classification mining and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifiers but Random Forest gives more accurate results with accuracy 98%.
6) Pahulpreet Singh Kohli et al, [6] suggested disease prediction by using applications and methods of machine learning and used techniques like Logistic Regression, Decision Tree, Support Vector Machine, Random Forest and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for Heart disease.
7) Lambodar Jena et al, [7] focused on risk prediction for chronic diseases by taking advantage of distributed machine learning classifiers and used techniques like Naive Bayes and Multilayer Perceptron. This paper tries to predict Chronic-Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.
8) Naganna Chetty et al, [8] developed a system that gives improved results for disease prediction and used a fuzzy approach. And used techniques like KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. In this paper diabetes disease and liver disorder prediction is done and the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.
9) Sayali Ambekar et al, [9] recommended Disease Risk Prediction and used a convolution neural network to perform the task. In this paper machine learning techniques like CNN-UDRP algorithm, Naive Bayes, and KNN algorithm are used. The system uses structured data to be trained and its accuracy reaches 82% and is achieved by using Naïve Bayes.
10) MinChen et al, [10] proposed a disease prediction system in his paper where he used machine learning algorithms. In the prediction of disease, he used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, and Decision Tree. This proposed system had an accuracy of 94.8%.

III. PROPOSED METHODOLOGY

Proposed methodology has 4 major steps –Data Pre-processing, Model Selection, Random Forest and Model Building

A. Data Pre-processing

For a system to predict proper results, first it should be trained properly with existing data. Pre-Processing the data is important so that good quality data is used for training the model. Data cleaning and removal of noise are some of the processes involved in Pre-Processing. We used the Diabetes Dataset (PIDD) of UCI Machine learning Repository, For heart disease analysis Cleveland, Hungarian and Switzerland heart disease patient's data sets are used. And for Parkinsons Data Set which is available in machine learning repository. Data from various sources has been collected and aggregated. Now by using the preprocessing techniques like:

Data Cleaning: Data is cleansed through processes such as filling in missing value, thus resolving the inconsistencies in the data. x

Data Reduction: The analysis becomes hard when dealing with a 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.

B. Model Selected

The system is trained to predict the diseases using three algorithms
Decision Tree Classifier
Random forest Classifier
Naïve Bayes Classifier
SVM Classifier
A comparative study is presented at the end of work, thus analyzing the performance of each algorithm of the considered database. Algorithm with the best accuracy and precision will be considered.

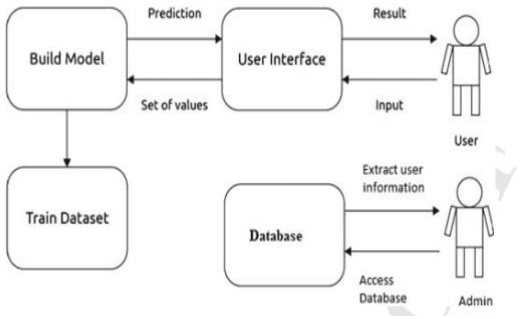
C. Random forest Classifier

Random forest is an adaptable, simple to use machine learning algorithm that provides exceptional outcomes more often than not even without hyper-tuning. The major limitation of decision tree algorithms 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 group 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).
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.

D. Model Building

Multiple disease prediction is a classification problem. So, we have implemented various classification algorithms like Decision Tree Classifier, Random Forest Classifier, SVM, Naïve Bias Classifier to choose the algorithm with best results. Next step is building the prediction version. First we need training and testing of the data by using the classification algorithm then we are fitting the model and load the model by using a pickle module.. Second, by using the python module Streamlit we are developing the user interface where the user is able to see the prediction of disease. Algorithm is chosen with the present stage of version accuracy.



Data Flow in the Model

The Data Flow Diagram shows the progression of the model. Here the information will be put away in the data set and which is gotten to by the Admin. At the point when the client has visited the site and enters the legitimate sources of info , the data sources will be shipped off the model and assess the data sources confronted to the prepared dataset and the forecast will be sent back to the UI by the model.

IV. EXPERIMENTATION

We have used a large dataset which consists of 70% training data and 30% testing data. The algorithms used for comparison were Naive Bayes, Decision Tree, SVM and Random Forest. The algorithms selected for comparison were based on the accuracy and time taken for prediction of class label. The accuracy analysis of algorithms on the dataset can be seen in Table 1.

TABLE I Accuracy analysis of algorithm on datasets

Algorithm	Accuracy on Diabetes	Accuracy on Heart Disease	Accuracy on Parkinson's Disease
Decision Tree Classifier	0.89	0.93	0.81
Random Forest Classifier	0.95	0.94	0.92
Support Vector Classifier	0.76	0.83	0.71
Naïve Bias Classifier	0.76	0.75	0.77

V. FUTURE WORK AND CONCLUSION

Because of this project the user doesn't need to traverse different websites which saves time as well. Diseases if predicted early can increase your life expectancy as well as save you from financial troubles.Multi disease prediction model is used to predict multiple diseases at a time. Here based on the user input disease will be predicted. The choice will be given to the user. If the user wants to predict a particular disease or if the user doesn't enter any disease type then based on user entered inputs corresponding disease model will be invoked and predicted. The advantage of a multi disease prediction model in advance can predict the probability of occurrence of various diseases and also can reduce mortality ratio.

In the future we can add more diseases in the existing API. We can try to improve the accuracy of prediction in order to decrease the mortality rate. Try to make the system user-friendly and provide a chatbot for normal queries.

REFERENCES

[1] Vijayalaxmi A, Sridevi S, Dr.Sridhar and Sateesh Ambesange "Multi-Disease Prediction with Artificial Intelligence from Core Health Parameters Measured through Non-invasive Technique" IEEE 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)

[2] Akkem Yaganteeswarudu "Multi Disease Prediction Model by using Machine Learning and Flask API.

[3] Chetan Sagarnal,Sneha Grampurohit "Disease Prediction using Machine Learning Algorithms" IEEE 2020 International Conference for Emerging Technology (INCET)

[4] Ajinkya Kunjir,Harshal Sawant and Nuzhat F. Shaikh "Data Mining and Visualization for Prediction of Multiple Diseases in Healthcare" 2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).

[5] Rashmi G Saboji and Prem Kumar Ramesh, "A Scalable Solution for Heart Disease Prediction using International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume- 8, Issue- 4, July- 2020 https://doi.org/10.21276/ijircst. 2020.8.4.14 www.ijircst.org Copyright © 2020. Innovative Research Publication. All Rights Reserve 330 Classification Mining Technique" IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.

[6] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[7] Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.

[8] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach" IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.

[9] Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.