Region-wise Disease Prediction

[1]Geetha. R, [2]J Sri Sai Sangeetha, [3]Manikeshwari, [4]Umamaheshwari. G, [5]Bhavana. P

[1]Geetha. R CSE, Cambridge Institute of Technology
[2]J Sri Sai Sangeetha CSE, Cambridge Institute of Technology
[3]Manikeshwari CSE, Cambridge Institute of Technology
[4]Umamaheshwari. G CSE, Cambridge Institute of Technology
[5]Bhavana. p CSE, Cambridge Institute of Technology

Abstract - Due to Machine Learning there is progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care, and community services. When the quality of medical data is incomplete the accuracy of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may result in weakening the prediction of disease outbreaks [2]. This paper mainly focuses on using a machine learning algorithm such as Naïve Bayes, Decision Tree, and Random Forest. The prediction of a particular disease in a given region based on the symptoms of a patient is done for structured data. The Comparative analysis of predicting a disease based on the symptoms using the above-mentioned algorithms is also a part of the proposed model.

Key Words: Machine Learning, Supervised Learning, Neural Network.

1. INTRODUCTION

The concept of Machine Learning is constantly changing in the healthcare sector. Machine-learning lends itself better to some processes than others. Algorithms provide immediate benefit to disciplines with processes that are reproducible or standardized. Also, those with large image datasets, such as radiology, cardiology, and pathology, are strong candidates. In Machine learning a system can be trained to look at images, data, identify abnormalities and patterns, and point to areas that need attention, thus improving the accuracy of all these processes. On a long term, machine learning will benefit the family practitioner or internist at the bedside. Machine Learning can improve efficiency, reliability and accuracy.

2. Body of Paper

The accuracy of the analysis is reduced when the quality of medical data is incomplete. Moreover, different regions have unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. The main motive in this paper is to predict the disease of a patient in a given region using algorithms such as Naïve Bayes, Decision Tree, and Random Forest. Comparative analysis of prediction of a disease using the above-mentioned algorithms.

At the first occurrence of an acronym, spell it out followed by the acronym in parentheses, e.g., charge-coupled diode (CCD).

2.1 PROPOSED SYSTEM

The drawback of the existing system is that it does not deal with region wise data because for a particular disease, the number of cases in various regions will vary, some will have more number of cases and some will have none causing in over fitting and biasing leading to extra preprocessing steps and this can be time-consuming when you have a lot of data to process.

The proposed system is to do Region-wise disease prediction by not considering particular disease but to consider all the diseases in the region perform preprocessing such as label encoding, removing null and missing values. Only these preprocessing tasks are done since the dataset is binary-valued multi-class text.The dataset is then divided into testing and training data. Algorithms implemented are Naïve Bayes, decision tree, random forest, and a neural network is also implemented Using the sklearn package methods.

The front end is developed using Tkinter package. Through the front end a user will enter the patient name and five of major symptoms patient has. After entering the choice from a drop-down menu, they can click on any of the four buttons where each indicates one of the algorithms. On clicking the button, in the backend, the corresponding algorithm gets executed and the result is displayed. The result is the disease predicted by that algorithm. Along with this, accuracy and precision for all the implementation are displayed. Another output is a graph that shows the number of cases for each of the diseases in that region. This helps in real-time statistical analysis.

The proposed system is very easy to operate. Main advantages of the proposed system are speed and accuracy. There is no redundancy of data. The data is stored in the computer’s secondary memories like hard disk, etc. it can be easily retrieved and used at any time. In proposed systems drawbacks of the existing system are eliminated to a great extent and it provides tight security to data.

2.2 SYSTEM DESIGN

As shown in Fig-1**,** Front end is developed using the Tkinter package in python. The application can be accessed by anybody. A person has to enter the patient's name and five of the major symptoms that the patient has. Based on these symptoms, when the user clicks on any of the four buttons corresponding to the machine learning concepts we have implemented, in the backend it will be executed and the disease predicted will be displayed in the front end.

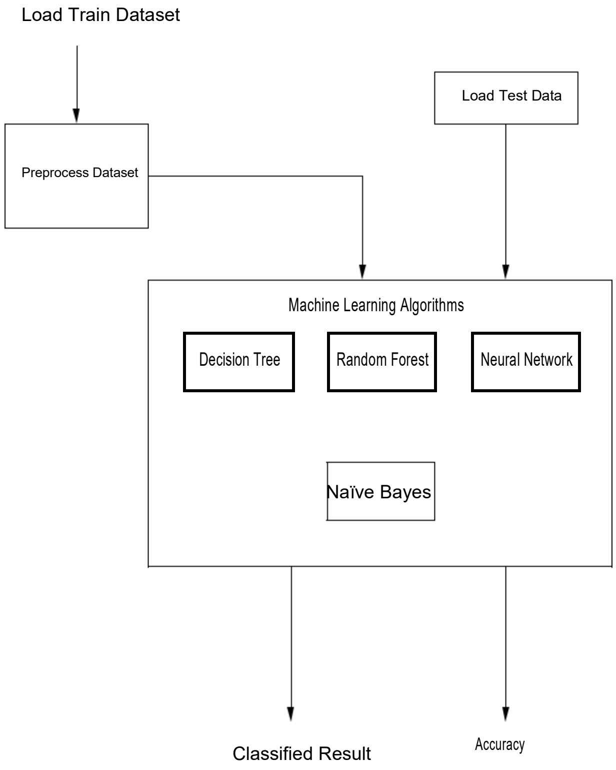


Fig -1: Design of our Proposed System

Datasets from different hospitals in different regions are gathered in a tabulated format. These datasets are both numeric and textual. The datasets are binary-valued multi-class datasets. Since these data are real-time data, preprocessing is a must to understand the context of the datasets. Missing data and null data are replaced with a replacement, text data is cleaned and converted to machine-understandable form, numbers. Once preprocessing is accomplished, the dataset is divided into training and testing data.

Once the data is ready, three supervised machine learning algorithms, Naïve Bayes, decision tree, random forest, and a neural network is also implemented. Fig -2 shows the use-case diagram of the system.

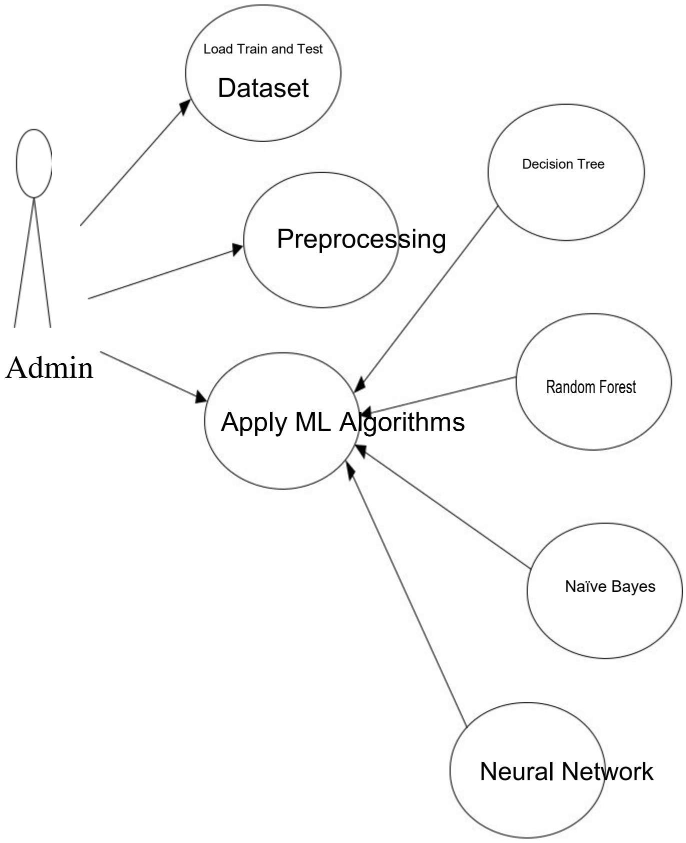


Fig -2: Use case diagram

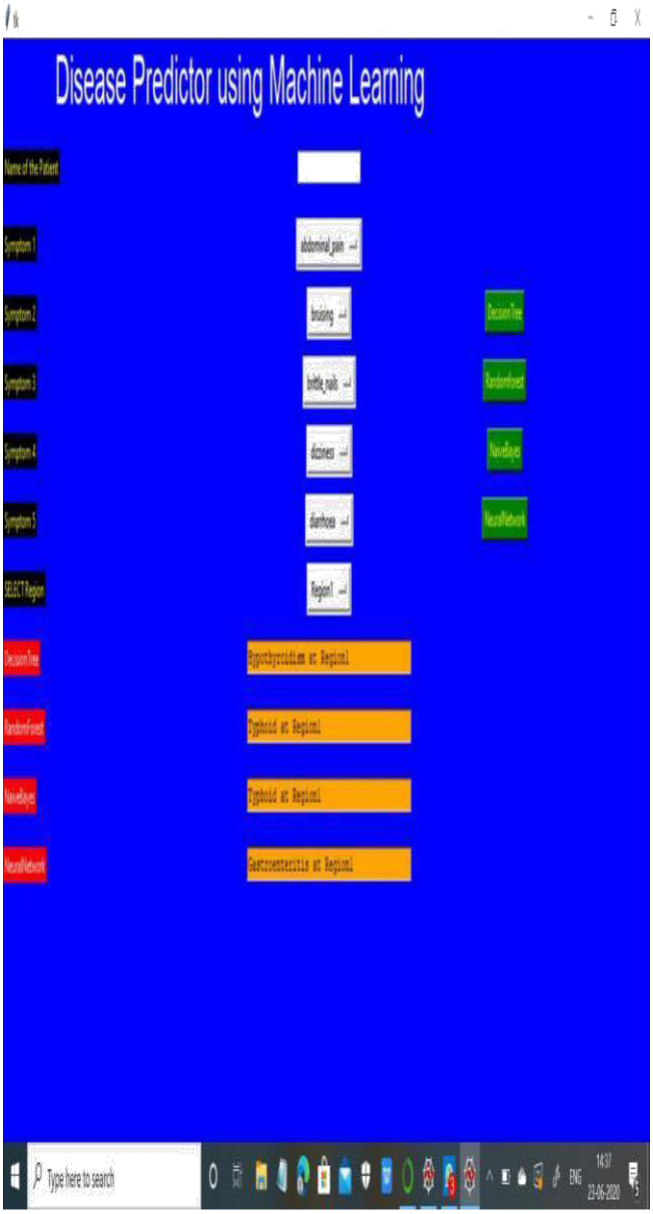
2.3 IMPLEMENTATION

The project has been implemented entirely in python version 3.x. The platform used is in Spyder available in Anaconda Navigator. The dataset that we have used for training has 132 columns where each column denotes a symptom. Each of the entries in a column has either a 0 or 1 denoting the absence or presence of the symptom.

There are 4921 rows and around 40 diseases and each of the diseases has a uniform number of entries so there is no biasing or overfitting or underfitting of data. In short, an explanation of the training data is that it a binary-valued, multi-class dataset.

The front end as shown in Fig 3, is developed using the Tkinter package, where any user can access it. When a user enters the patient name, and five of the major symptoms they have and click on any of the four buttons that denote any of the implementations, in the backend, the symptoms are taken and based on the button clicked, and the training that chosen algorithm has gone through will be implemented and predicted output along with accuracy and precision will be displayed.

Along with this when a region is chosen, a graph is displayed as shown in Fig -4, which indicates the diseases versus the number of cases that can be used to analyze which disease is more in a region.



Hypothyroidism at Region1

Typhoid at Region1

Typhoid at Region1

Gastroenteritis at Region1

Fig -3: Front-end

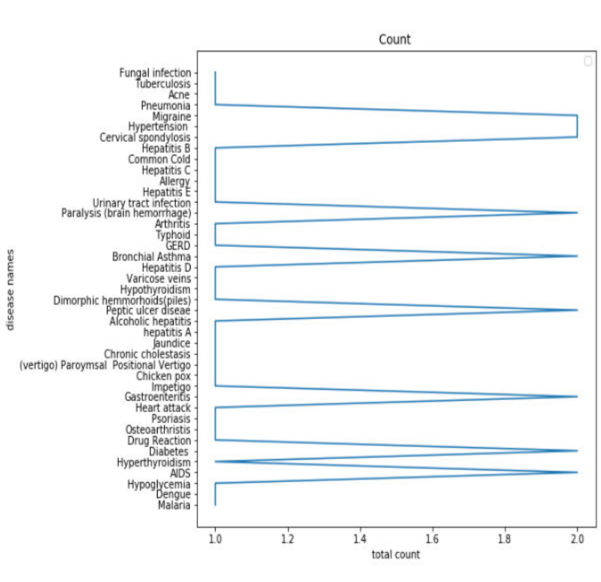


Fig -4: Graph representing total cases in a region

Overall, this project gives a comparative analysis on which algorithm performs better for binary-valued multi-class data and can be used for studying the different supervised algorithms and neural network implemented.

2.4 RESULTS

The drawback of the existing system is that it does not deal with region wise data because for a particular disease, the number of cases in various regions will vary, some will have more number of cases and some will have none causing in overfitting and biasing leading to extra preprocessing steps and this can be time-consuming when you have a lot of data to process.

Implementation of the system in python makes it secure, easier to understand, easier to use and it can be used for learning purposes. The structured data is cleaned so the system cannot show false results. Running the project, showed that the accuracy and precision of all the implementations are similar but the neural network takes more time compared to other algorithms.

3. CONCLUSIONS

The proposed system is to do Region-wise disease prediction by not considering particular disease but to consider all the diseases in the region perform preprocessing such as label encoding, removing null and missing values. The proposed system is very easy to operate. Speed and accuracy are the main advantages of this system. There is no redundancy of data. The data is stored in the hard disk, etc. it can be easily retrieved and used at any time. It can eliminate the drawbacks of the existing system to a great extent and it provides tight security to data. It can also be used for learning purposes.

ACKNOWLEDGEMENT

We are very much thankful for our project guide, Prof. Bhavana P who supported us during the project completion and also helped us to improve manuscripts.

REFERENCES

[1] "Mohd Usama, Belal Ahmad, Jiafu Wan, M. Shamim Hossain, Mohammed F. Alhamid and M. Anwar Hossain: Deep Feature Learning for disease risk assessment based on Convolutional Neural Network With Intra-Layer Recurrent Connection by using Hospital Big Data, Dec 3, 2018, IEEE"

[2] Shraddha Subhash Shirsath, Prof. Shubhangi Patil Disease Prediction Using Machine Learn.Over BigData". I international Journal of Innovative Research in Science, Engineering and Technology, [2018].ISSN (Online) : 2319-8753, ISSN (Print) : 2347-6710.

[3] Dr S. Vijayarani, Mr. S. Dhayanand, "Liver Disease Prediction using SVM and Naive Bayes Algorithms."

[4] In Chen, Yixue Hao, Kai Hwang, Fellow, IEEE, Lu Wang, and Lin Wang, Disease Prediction by Machine Learning over Big Data from Healthcare Communities, 2169-3536 (c) 2016 IEEE.

[5] hahab Tayeb*, Matin Pirouz*, Johann Sun1, Kaylee Hall1, Andrew Chang1, Jessica Li1, Connor Song1, Apoorva Chauhan2, Michael Ferra3, Theresa Sager3, Justin Zhan*, Shahram Latifi, Toward Predicting Med-ical Conditions Using k-Nearest Neighbours, 2017 IEEE International Conference on Big Data.

[6] hen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee, Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database, 2017 IEEE.

[7] Reekanth Rallapalli Faculty of computing Botho University Gaborone, Botswana rallapalli.sreekanth@bothouniversity.ac.bw,Predicting the Risk of Diabetes in Big Data Electronic Health Records by using Scalable Random Forest Classification Algorithm, 2016 IEEE.

[8] rof. Dhomse Kanchan B. Assistant Professor of IT department METS BKC IOE, Nasik Nasik, India kdhomse@gmail.com , Mr. Mahale Kishor M. Technical Assistant of IT department METS BKC IOE, Nasik, India kishu2006.kishor@gmail.com, Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analy-sis,2016 IEEE.