Disease Prediction from Various Symptoms Using Machine Learning

Mrs. M.S.Sandhya Rani[1], Guda Ankitha[2], Polasani Harini[2], G Ravi[2]

[1]Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India

[2]Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India

ARTICLEINFO

ABSTRACT

The main reason to develop this system is that we need to be aware of diseases that may start with small symptoms and end up with dangerous health conditions. So, to avoid this we need to have regular health check-up's but in our modern lifestyle there is no much time left for health check-up or any kind of diagnosis. That's why with this human disease prediction from various symptoms we can easily check the underlying diseases just buy selecting the symptoms that people may have in the start and check the problem and all this process will happen in a span of few seconds and mainly at your convenient place. In this system we have used naive bayes algorithm to predict the disease based on the your inputs and we have used the dataset collected by the Columbia University and this system uses this dataset and algorithm to predict the most likely disease according to the entered or selected symptoms.

Keywords: Disease Prediction Symptoms Machine Learning

I. INTRODUCTION

Now-a-days, many people having various diseases due to environmental conditions and their living habits. So the prediction of disease at earlier stage becomes an important task. The normal medical check-up is a process where the patient has to visit a doctor and undergo many medical tests, and then come to a conclusion. This process is very time consuming. To save time required for the initial process of diagnosis symptoms, this project proposes an automated disease prediction system that relies on user input. This Disease Prediction From Various Symptoms Using Machine Learning is completely done with the help of Machine Learning and Python Programming language with Tkinter Interface for it and also using the dataset which is taken from a study conducted at Columbia University which is readily available, using that we will predict the disease. Today's doctors use a variety of scientific technology and methodologies for both the detection and diagnosis of numerous fatal diseases as well as ordinary illnesses.

Correct and thorough diagnosis is always responsible for a successful course of treatment. Doctors may sometimes be able to make accurate decisions while diagnosing the disease of a patient, therefore disease prediction systems which use machine learning algorithms assist in such cases to get accurate results. Because of the competitive environment of economic development and the prevalence of general disease, the project disease prediction using machine learning was created to combat general disease in its earlier stages involved so much that he/she is not concerned about health according to research there are 40% peoples how ignores about general disease which leads to harmful disease later. The main reason of ignorance is laziness to consult a doctor and time concern the peoples have involved themselves so much that they have no time to take an appointment and go to the doctor which later results into fatal disease. According to research there are 70% peoples in India suffers from general disease and 25% of peoples face death due to early ignorance of the symptoms.

II. LITERATURE SURVEY

In the paper called "Disease Prediction System using data mining techniques" the author has discussed about the data mining techniques like association rule mining, classification, clustering to analyse the different kinds of heart disease based problems. The database used contains collection of records, each with a single class label. A classifier performs a simple and clear definition for each class that can be used to classify successive records. The data classification depends on the algorithm called MAFIA which causes accuracy, the info is calculable exploitation entropy primarily based cross validations and partition techniques and also the results are compared with other outputs. C4.5 algorithmic rule is employed because the coaching algorithmic rule to indicate rank of attack with the choice tree. The heart disease problem information is clustered mistreatment which

is the K-means clump algorithmic rule, which will remove the data applicable to heart attack from the database. Some limitations are measured by the measurements like, time complexity is more due to DFS traversal, C4.5- Time complexity increases while searching for insignificant branches and lastly no precautions are defined.

In the paper "A study on data mining prediction techniques in healthcare sector" 21 fields that mentioned are, information Discovery method (KDD) is that the method of adjusting the low level data into high-level knowledge. The repetition method consists of the subsequent steps information cleansing, information integration, information choice, information transformation, data processing, Pattern analysis, Knowledge is the process.

Healthcare data processing prediction will support data processing techniques which are as follows: Neural network, Bayesian Classifiers, call tree, Support Vector Machine. The paper states that comparative study of various aid predictions, Study of information mining techniques and tools for prediction of cardiovascular disease, numerous cancers, and diabetes, disease and medicine conditions. Some limitations are if attributes are not related then Decision tree prediction is less accurate and ANN is computationally intensive to train also it will not lead to specific conclusion. So, we can use Bayes algorithm.

The paper "Predicting Disease by Using Data Mining Based on Healthcare information System" published in the year 2016 applies the information mining process to predict high blood pressure from patient medical records with eight alternative diseases. The data was taken from a true world health care system information containing medical records. Under sampling technique has been applied to come up with coaching knowledge sets, and data processing tool

will not generate the Naive Bayesian and J-48 classifiers created to improve the prediction performance, and rough set tools were wanted to scale back the ensemble supported the concept of second order approximation.

Experimental outputs showed a bit improvement of the ensemble approach over pure Naive Bayesian and J-48 in accuracy, sensitivity and F-measure. Initially they used classification called ensemble the classifier and so the reduction of Ensemble Classifiers is employed. But the choice trees generated by J-48 is typically lacking within the levelling therefore the overall improvement of victimization ensemble approach is a small amount.

The paper "An approach to create an Interactive software solution for smart health prediction using data mining" published in the year 2018 mainly aims on developing a computerized system to check and maintain your health by knowing the symptoms. It has a symptom checking module which actually defines our body structure and gives us liability to select the affected area and checkout the symptoms. Technologies which are implemented in this paper are: The front end is developed with using HTML, Java Script and CSS. The back end is developed using MySQL which is used to design the database queries. This paper also contains the information of testing like Alpha testing which is performed at server side or we can say at the developer's end, this is an actual testing, tested with potential users or as an independent testing process at server side. And the Beta testing is done after performing alpha testing, versions of a system or software known as beta versions are given to a specific audience outside the programming team to check. Only the limitation of this paper is that it suggests that only the award winning doctors and not the nearby doctors to the patient who is using this algorithm.

III. METHODOLOGY

3.1 PRIMARY STEPS IN DESIGNING A MACHINE LEARNING MODEL:

Data Collection: This is the first thing we are going to deal with. In this we will be gathering all the data that is required for our model.

Pre-Processing: In this step, the symptoms are pre-processed to extract the useful features from the given image and it also does various functions like image adjustment, pixel changing.

Training: Once the cleaning is completed, it has to be trained. We have predefined dataset which contains the symptoms and diseases. This particular dataset is used to train the model.

Detection: In this step the model predicts the disease of a user. If the user enters minimum of two symptoms or maximum of five symptoms then it returns the disease as an output and if the user enters one symptom or if user doesn't enter then it will display a dialog box.

Evaluate the Model: After we done with training of our model it is now evaluated and tested in before head so that it is ready for further process.

3.2 INTRODUCTION TO TECHNOLOGIES USED

3.2.1 Python

It is a multi-paradigm programming language. Object oriented programming and structured programming is fully supported, and many of its features support functional programming and object oriented programming. Many other paradigms are also supported via extensions, including design by contract and logic programming. Python uses dynamic typing method and a combination of reference counting and a cycle detecting garbage collector for memory management. It also features dynamic name

resolution which is also called as late binding, which binds method and variable names during program execution i.e., when program is running. Python's developers strive to avoid premature optimization, and reject patches to non critical parts of Python that would offer marginal increases in speed at the cost of clarity. When speed is primary concern then a Python programmer can move time-critical functions to extension modules written in languages such as C, or use PyPy, a just-in-time compiler which can be used. Cython or CPython is also available, which translates a Python script into C and makes direct C level API calls into the Python interpreter. A primary goal of Python's developers is that keeping the language easy and fun to use. Python's design offers some support for functional programming in the Lisp like tradition. It has filter, map, and reduce functions or methods, list comprehensions, dictionaries, sets and generator expressions. The standard library of python has two modules they are itertools and functools that implement functional tools borrowed from Haskell and Standard ML languages.

3.2.2 Tkinter Interface

Tkinter is a Python's version in binding to the Tk GUI toolkit. Tkinter is the standard Python interface to the Tk implementation of the GUI toolkit and is Python's de facto standard GUI. It is pre-loaded with standard Linux, Microsoft Windows and Mac OS X installers of Python. The name Tkinter comes from Tk interface. Tkinter was written by Fredrik Lundh. Tkinter is free and open software.

3.2.3 Machine Learning

ML is the study of computer algorithms that improve automatically through experience and no of times it runs. It is seen as a part of artificial intelligence (AI). Machine learning algorithms build a model based on the sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed by a programmer. Machine learning algorithms are used in many applications, such as email filtering and computer vision, where it is difficult to develop conventional algorithms to perform the needed tasks.

3.2.4 Scikit

Scikit-learn(sklearn) is a free software, machine learning library for the Python programming language. It is bundled with various classification, regression and clustering algorithms including support-vector machines. And this is developed by a person named David Cournapeau.

3.3 NAIVE BAYES ALGORITHM

Naive Bayes algorithm is a supervised machine learning algorithm, which is based totally on Bayes theorem and used for solving classification problems. It is generally used in text classification that includes a variety of high dimensional training dataset. Naive Bayes is one of the simple and most effective Classification algorithms which help in building the fast machine learning models that can make very quick predictions. This particular system accepts the input from the user and predicts the most probable disease. This is achieved with the help of the dataset and the machine learning algorithm. The algorithm here is Naive Bayesian which works on a probabilistic approach. We will import Scikit to learn the library for its implementation. For this, we have used multinomial Navie Bayes (NB) since multiple variants i.e., multiple symptoms are taken. Bayes theorem is also known as Bayes Rule or Bayes law, which is used to determine the probability of a hypothesis with previous knowledge. It depends on the conditional probability. The general formula for Bayes theorem is given as

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

IV. RESULTS AND DISCUSSION

This system can predict disease based on symptoms. And these above output is the combination of tkinter and python programming.
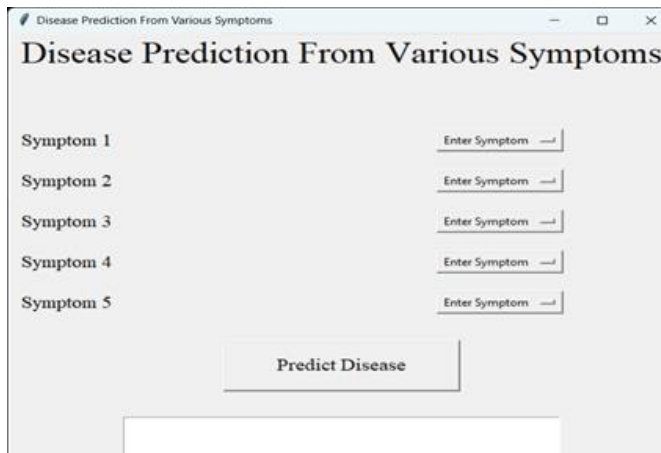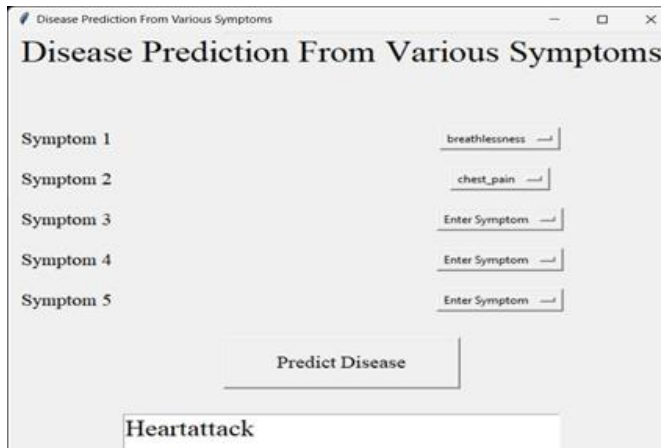

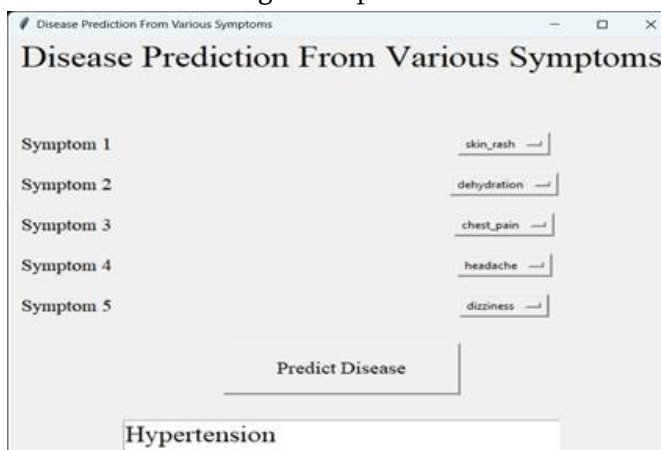
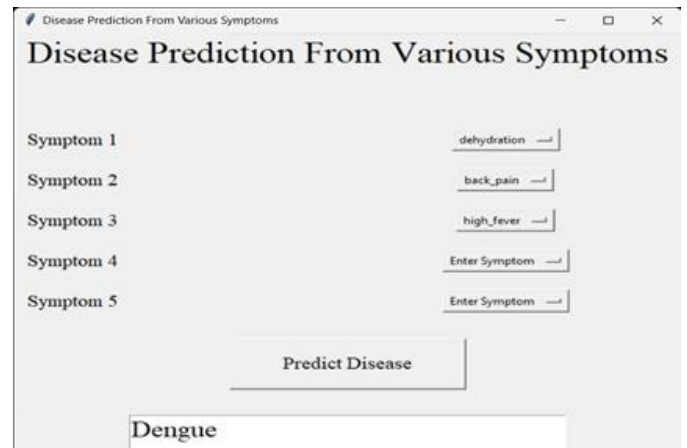Fig 1: System When Started

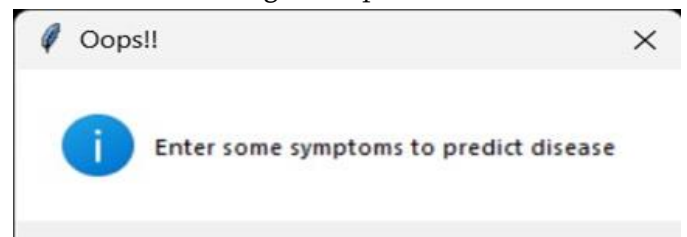
Fig 2: Output 1

Fig 3: Output 2

Fig 4: Output 3

Fig 5: Output 4 (When no symptom is entered)

V. CONCLUSION

So, finally I conclude by describing that, this project disease prediction from various symptoms using machine learning (ML) is very useful in everyone's everyday life and it is mainly more important for the healthcare sector, because they are the one that daily uses these systems to predict the diseases of the patients based on their general information and their symptoms that they have been through. Now-a-day's health industry plays major role in treating the diseases of the patients so this is also some kind of help for the health sector to tell the user and also it is much useful for the user in case he/she doesn't want to go to the hospital then they can just select the symptoms and the user can get to know the disease that they are suffering from and the health industry can also get benefited from this system by just asking the symptoms from the user and entering in the system and then in just a few seconds they can tell the exact and up to some extent the accurate diseases. If health industry takes up this project then the work of the doctors can be reduced drastically and they can easily predict the disease of the patient very quickly. The Disease Prediction From Various Symptoms project is to provide prediction for the various generally occurring diseases that when unchecked

and sometimes ignored can turns into dangerous disease and cause a lot of problem to the patient and as well as the people in their family. So, many people who do not want to go to hospitals and clinics they can use this project to know their health and disease condition very conveniently and quickly sitting in their homes.

VI. FUTURE SCOPE

As now-a-days we can clearly witness the increase in use of computers and technology to consider a huge amount of data, computers are being used to perform various complex tasks with commendable accuracy rates. Machine learning (ML) is a collection of multiple techniques and algorithms which permit computers to execute such complex tasks in a simplified manner. It is also used in both academics which is for students or learners and also in industry to make accurate predictions and use these diverse sources of dataset and information. Till this date we can say we got developed in the fields of Big Data, Machine learning, and Data Sciences etc and have been a part of one of those big industries which were able to collect such data and the staff to transform their goods and services in a desired manner. The learning methods developed for these industries and researches offer excellent potential to further improve in medical research and clinical care for the patients in the best possible manner. Machine learning uses mathematical algorithms and procedures which are used to describe the relationship between variables used in the model and the others. Our paper will explain the process of training the model and learning a suitable algorithm to predict the presence of a particular disease from the sample of the tissue based on its features. Though these algorithms work in different and unique manners depending on the way in which they are developed and used by the researchers. One method is to consider their main goals. The goal of our project is to reach to a conclusion about the data which are collected from a wide variety of samples from our population. Though many techniques, like bayes theorem can be able to predict the diseases. For example, consider a case where, if we can create a model which described and understood the relationship between clinical variables and their transience then we can follow the organ transplant surgery i.e. we would need the factors and features which differentiate low mortality rate from high if we can develop such outcomes and reduce mortality rate to a desired rate in the future and also nothing can be said to be better than such situations.

VII.REFERENCES

[1]. D. W. Bates, S . Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," Health Affairs.

[2]. K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, "Performance comparison of three data mining techniques for predicting kidney disease survivability", International Journal of Advances in Engineering Technology.

[3]. Mr. Chala Beyene and Prof. Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics.

[4]. Boshra Brahmi, Mirsaeid Hosseini Shirvani, "Prediction and Diagnosis of Heart Disease by Data Mining Techniques", Journals of Multidisciplinary Engineering Science and Technology.

Cite this article as :