HUMAN DISEASE PREDICTION – Python & GUI

1. Abstract:

The Healthcare industry is a big business. It produces large amounts of data daily that can be used to extract information for predicting diseases, which can be used for early diagnosis of a patient while also referring to their health history. This field can improve a lot with the help of AI & Machine learning. In today's day and age, people face various diseases due to environmental conditions, lifestyle, transmittable bacterial and viral infections, etc. Prediction of such diseases at an early stage can prove to be extremely beneficial and ultimately aid in saving lives. Accuracy in disease prediction becomes challenging for medical professionals with the various mutants, types of diseases, and various symptoms that come specific to each type. The inevitable growth in medical data each year needs to be analyzed with data mining to find patterns. For disease prediction, we have used the Decision tree algorithm and Naïve Bayes model, both of which produce an accuracy rate of approximately 80- 90%. This prediction model is done using Python and GUI Tkinter. Using this system, the client is able to narrow down a set of symptoms to two diseases efficiently and quickly while also eliminating human error, the predicted diseases can further be analyzed by the accuracy rate of each machine learning model used.

2. Introduction

- Healthcare is one of the most urgent matters in human society. It is widely distributed and fragmented.
- It is collaborative in nature and consists of a large number of physicians of different specialties, nurses, pathologists, laboratory technicians, etc.
- These stakeholders generate data from various heterogeneous sources like physical examinations, clinical notes, patient history, patient observation, laboratory results, etc.
- Thus, the amount of medical data being digitally collected and stored is vast and expanding quickly, the science of data management and analysis is also advancing to convert this vast resource into information and knowledge that helps the respective healthcare workers gain insights.
- Thus, there is a clear need for an effective and robust methodology that allows for early disease detection and it can be used by doctors in decision-making.
- Machine learning provides several classifiers for intelligent data analysis, by using these classifiers we will be able to analyze the symptoms and come to a close conclusion as to what the disease might be, we will be able to train the model and eventually predict the disease.

3. Problem Statement

- Present day Health industry is loaded with gadgets that give out off-base or unaccepted outcomes, to prevent such errors in the medical field we have come up with a disease prediction model which will provide precise information and predictions based on data given out by clients.
- Most of the time there is a need to be physically present at the clinic, or the availability of a specialist to analyze the symptoms and predict the cause for the same.
- Self-researching on the internet is very time-consuming and can sometimes be an incorrect diagnosis.
- High dependency on medical professionals can increase the chance of human error, as they would not remember the various symptoms for different diseases and slight variations in different diseases leading to different symptoms.
- It enables clients to get acquainted with the possible diseases at an early stage and treat it accordingly.

4. Existing System

The working of the existing system is as follows:

1. Traditional methods involve various complex algorithms, datasets, and rigid programming.
2. High dependency on manual updating and rectification of code.
3. Takes more development time.
4. They are handy only in clinical situations and do not suit big industry sectors.
5. Diagnosis of the condition solely depends upon the medical profession's intuition and patient's records.
6. Due to this early detection is not possible at an earlier stage

5. Proposed System

The proposed system tests the hypothesis that supervised ML algorithms can improve health care by the accurate and early detection of diseases. In this system, we utilize more than one supervised ML model for each disease recognition problem. This approach renders more comprehensiveness and precision because the evaluation of the performance of a single algorithm over various studies indicates a certain level of induced bias which generates imprecise results. The analysis of ML models will be conducted on a few diseases with various symptoms. The detection of the disease is done through two classification models Decision Tree and Naïve Bayes respectively. At the end, the best performing ML models with respect to the specific disease will be concluded.

Hardware & Software requirements

- Hardware Requirements
    - Processor - Intel Core i3 and above
    - Speed - 2.5 GHz
    - RAM - 4 GB (min)
    - Hard Disk - 200 GB
- Operating System - Windows 7 and above
    - Programming Language - Python 3.7

6. Implementation

<u>Decision Tree Classifier</u>

- It is a Supervised Machine Learning where the data is continuously split according to a certain parameter.

The decision Tree consists of:

- Nodes: Test for the value of a certain attribute.
- Edges/ Branch: Correspond to the outcome of a test and connect to the next node or leaf.
- Leaf nodes: Terminal nodes that predict the outcome (represent class labels or class distribution).
- It has two main types, Classification Trees and Regression Trees.

- Classification Trees are built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions and then splitting it up further on each of the branches.
- Regression Trees are decision trees where the target variable can take continuous values (typically real numbers, e.g. the price of a house, or a patient's length of stay in a hospital).
- They are built using the divide and conquer method because it splits the data into subsets, which are then split repeatedly into even smaller subsets, and so on and so forth until the process stops when the algorithm determines the data within the subsets are sufficiently homogenous, or another stopping criterion has been met.

Advantages of Classification with Decision Trees:

- Inexpensive to construct. Extremely fast at classifying unknown records.
- Easy to interpret for small-sized trees and excludes unimportant features.

Disadvantages of Classification with Decision Trees:

- Easy to overfit and models are often biased toward splits on features having a large number of levels.
- Small changes in the training data can result in large changes to decision logic.
- Large trees can be difficult to interpret and the decisions they make may seem to counter-intuitive.

Naïve Bayes Classifier

- Naive Bayes classifier is a probabilistic machine learning model based on Bayes Theorem. It is called naïve as the presence of one particular feature does not affect the other.
- We can find the probability of **A** happening, given that **B** has occurred. Here, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent.
- Types of Naive Bayes Classifier:
- Multinomial Naive Bayes-This is mostly used for document classification problems, i.e. whether a document belongs to the category of sports, politics, technology, etc. The features/predictors used by the classifier are the frequency of the words present in the document.
- Bernoulli Naive Bayes-This is similar to the multinomial naive Bayes but the predictors are Boolean variables. The parameters that we use to predict the class variable take up only values of yes or no, for example, if a word occurs in the text or not.
- Gaussian Naive Bayes-When the predictors take up a continuous value and are not discrete, we assume that these values are sampled from a gaussian distribution. Gaussian Naïve Bayes-
  It is easiest to work with the distribution of data. It only needs the estimation of the mean & standard deviation from the training dataset. It follows normal distribution & supports continuous data.

Working of GNB -
Data preprocessing: It includes data preparation so that it can be effectively used in code. It includes Importing the libraries, Importing the dataset, Splitting of the dataset into training and testing, and Feature scaling. Fitting NB to the training set. Predicting test results. Visualizing the training set result.

Advantages

- Simple, fast & effective predicting.
- Can be used for multiple class prediction problems.

- Performs well in the case of text analytics problems.

Disadvantages

- Relies more on independent features, less likely to occur in real life.
- Not ideal for data sets with many numerical attributes.
- Model assigns 0 probability for Unknown categories which is incorrect.

7. Comparison of two models used

Name	Decision Tree Accuracy	Naïve Bayes Accuracy	Higher Accuracy
Monisha	85 %	82%	Decision Tree
Roshni	88 %	80 %	Decision Tree
Nivedha	87 %	83 %	Decision Tree
Rushika	82 %	81 %	Decision Tree
Monisha	81 %	81 %	None

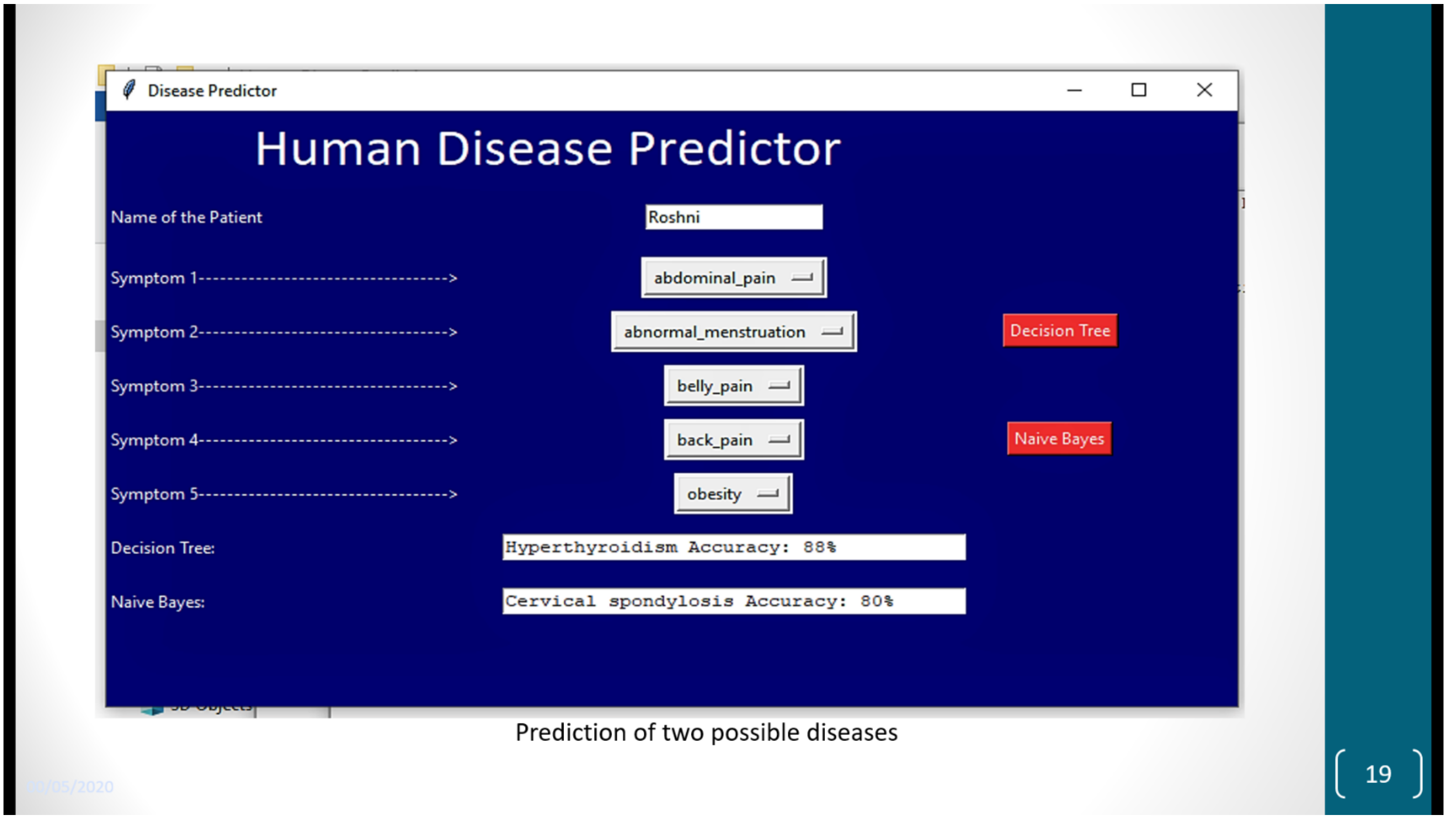
8. Conclusion

- This project predicts disease using machine learning which is very useful in today's age. It is extremely beneficial for the healthcare sector as they would need to use these systems to predict the diseases of the patients based on their general information and symptoms and come up with an immediate and accurate diagnosis.
- Many chronic diseases can be prevented or cured if they are dealt with in the early stages. If the health industry adopts this project then the work of the doctors can be reduced and they can easily start treatment of the disease. Disease prediction is to provide a prediction for the various and generally occurring diseases that when unchecked and sometimes ignored can turns into fatal disease and cause lot of problem to the patient and as well as their family members.

9. References

1. 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.
2. B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.
3. IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, " Wearable 2.0: Enable human-cloud integration in next generation healthcare system," IEEE Commun. , vol. 55, no. 1, pp. 54–61, Jan. 2017.
4. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "HealthCPS: Healthcare cyberphysical system assisted by cloud and big data," IEEE Syst. J., vol. 11, no. 1, pp. 88–95, Mar. 2017.

10. Output Snapshots



Disease Predictor

Human Disease Predictor

Name of the Patient: Roshni

Symptom 1---------------------------------> abdominal_pain

Symptom 2---------------------------------> abnormal_menstruation

Symptom 3---------------------------------> belly_pain

Symptom 4---------------------------------> back_pain

Symptom 5---------------------------------> obesity

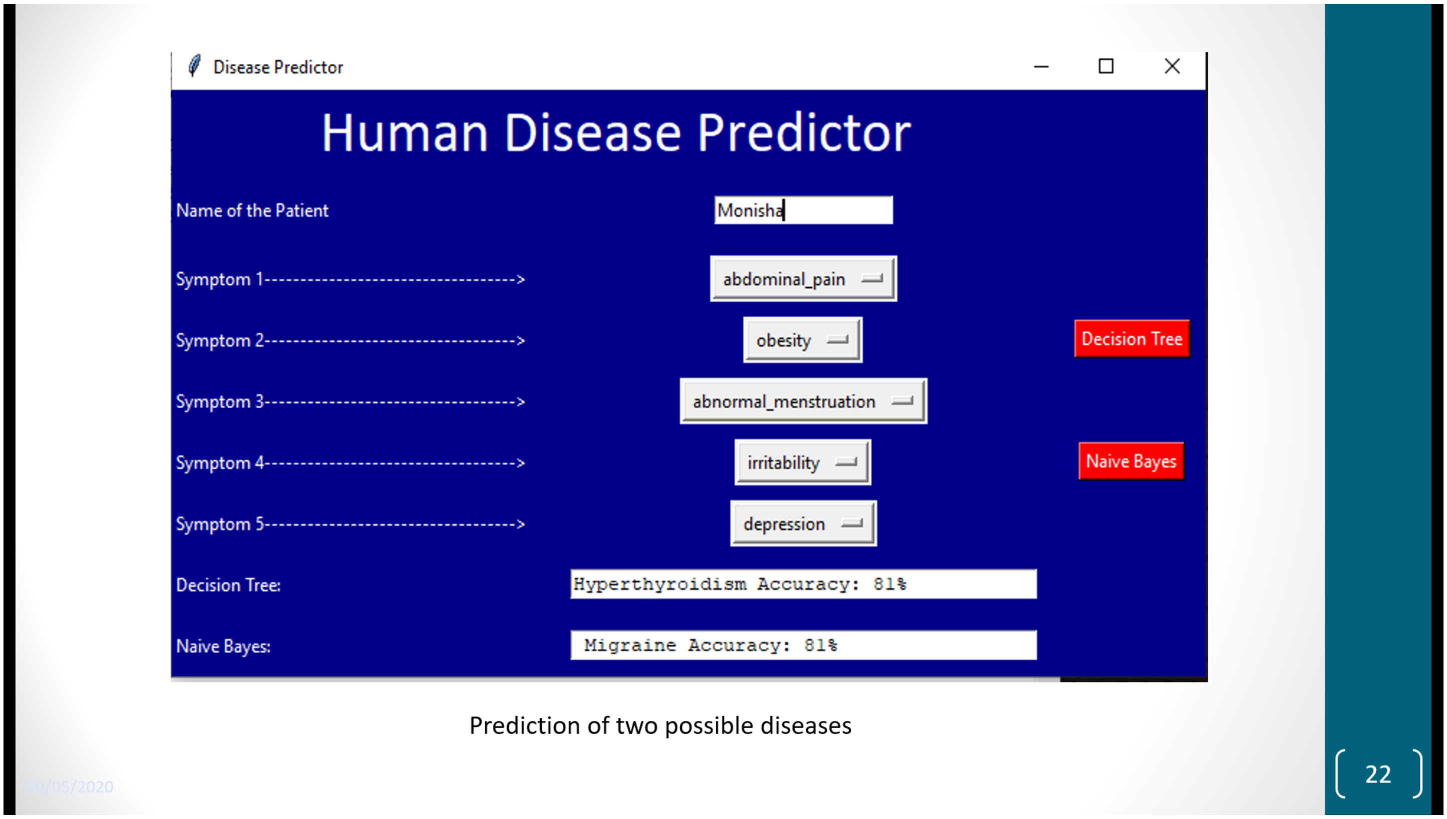
Decision Tree: Hyperthyroidism Accuracy: 88%

Naive Bayes: Cervical spondylosis Accuracy: 80%

Prediction of two possible diseases

20/05/2020

19



Disease Predictor

Human Disease Predictor

Name of the Patient: Monisha

Symptom 1---------------------------------> abdominal_pain

Symptom 2---------------------------------> obesity

Symptom 3---------------------------------> abnormal_menstruation

Symptom 4---------------------------------> irritability

Symptom 5---------------------------------> depression

Decision Tree: Hyperthyroidism Accuracy: 81%

Naive Bayes: Migraine Accuracy: 81%

Prediction of two possible diseases

20/05/2020

22