# DISEASE PREDICTION USING NAIVE BAYES CLASSIFIER

**Roop Chandrika Mallela**[1]
[1]Undergraduate Student, CSE Department, Sathyabama Institute of Science and Technology, Chennai. India.
[1]mrchandrika2000@gmail.com,

**Reddy Lakshmi Bhavani**[2]
[2] Undergraduate Student, CSE Department, Sathyabama Institute of Science and Technology, Chennai. India.
[2]reddybhavani4985@gmail.com,

**Dr. B. Ankayarkanni**[3]
[3]Assistant Professor, CSE Department, Sathyabama Institute of Science and Technology,Chennai. India.
[3]ankayarkanni.s@gmail.com

**Abstract- Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy. Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one is willing to go to hospital for health check-up due to the fear of spreading virus. In this situation, technology plays and important role. The domain we used here is Machine Learning, it is the technique by which machines can learn from past experiences like a human being and make it efficient in future. ML is the domain which is widely used nowadays and it is the most efficient domain in health care. We will develop a GUI to get the symptoms from the user. The models used in this paper are Naive Bayes and Decision Tree. The output is the disease, the accuracy of model, its definition and the treatment of the particular disease based on the symptoms given by the individual. As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease". This paper shows detailed explanation of how to find the diseases from symptoms, so that the individual can contact the respective doctor and stay healthy at an early stage.**

***Keywords:*** **Disease Prediction, Machine Learning, Tkinter, Chatbot, Symptoms.**

## I. INTRODUCTION

A disease is a condition that affects the individual functioning of body totally. Diseases if neglected will lead to the death of an individual. Diseases can be identified by the symptoms of the body of an individual.

Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy. Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus.

As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease*"*.

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

Some approaches tried to do predict the disease but they are restricted to particular disease like liver disease, heart disease, diabetes and so on. Machine Learning is the domain used in this paper. Machine Learning is a technique by which machines will be capable of learning and improving from past experiences using the algorithms.

The dataset which is used in this paper contains symptoms and the particular disease which was processed in Gaussian Naive Bayes and Decision Tree, which are the most easy models for disease prediction. While processing the data, symptoms are given as input and the disease was received as an output.

Machine learning technique makes easier to predict the disease and get treatment based on the disease. Various algorithms are implemented on the dataset and the accuracy of the same is also calculated. By this, we will consider only the algorithms which has the highest accuracy and it is used for accurate disease prediction.

Developing a project based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method is the main objective of the project.

We have designed a disease prediction system using ML algorithm (Naive Bayes), find the most accurate algorithm, and used it to find the disease and Tkinter for GUI.

And also, we have created a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

This project helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor.

A disease predictor can also be called as a virtual doctor, which can predict the disease based on symptoms. Recently due to covid-19 no one are willing to go outside. This disease predictor system can be a most useful as it identifies the disease without even contacting the individual.

In this way machine learning when implemented in healthcare can help in satisfying the individual and also take care of their particular disease easily.

Naïve Bayes, which is the most easy model helps to get the idea about the disease of an individual based on the symptoms he/she have, and get the treatment easily by contacting the concern doctor. This system not only reduces expense of the individual but also gives the accurate prediction.

## II.LITERATURE SURVEY

- Tarigoppula V.S Sriram et al. in [1] collects the voice dataset from UCI Machine Learning repository and train four algorithms on that dataset. The result is the prediction of Parkinson Disease by considering the most accurate algorithm.
- Shubham Bind et al. in [2] studies about all the available researches in literature to predict the Parkinson diseases.
- K. Gomathi, D. Shanmuga Priya in [3] used different data mining techniques to predict Heart disease, Breast Cancer, Diabetes. The models used are Decision Tree and Naive Bayes Classifier. Performance of both the models was compared and the best classifier is used to predict the above diseases.
- Isha Pandya et al. in [4] used two supervised machine learning algorithms Decision Tree, accuracy 91% and Naïve Bayes classifier, accuracy 87%. Here, they used the combination of both to get the best accuracy. Naïve Bayes Classifier accuracy should be improved.
- Akash C. Jamgade, Prof. S. D. Zade in [5] paper determined the most danger diseases which occur in a person in a locality and community. But, the data collection is difficult.
- Siddhika Arunachalam in [6] six classification algorithms are used after analyzing 14 attributes in the dataset. But, we may get confused which algorithm to use.
- Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE methodologies in [7] used are SVM, ANN, Logistic Regression so on. The predicted diseases are breast cancer, lung cancer, heart diseases, diabetes, thyroid or kidney diseases.
- H BENJAMIN FREDRICK DAVID in [8] predicted the occurrence of heart disease using ensemble learning algorithms. Accuracy of both Decision Tree and Naïve Bayes and analysed.
- Harshit Anand et al. in [9], domains of Machine learning and Data Science are used and models are built using numpy, pandas, sklearn, and so on and the model are deployed using Django.
- Goutam Chakraboty et al. in [10], takes into account six features from 23 features in the dataset and predict the risk of chronic kidney disease. It is used to reduce the impact of Chronic Kidney disease (CKD), where creatinine test is not available for all.
- A. Durga Praveen et al. in [11], applied 5 models namely KNN, SVM, Random Forest, Naïve Bayes and Adaboost and found that KNN, Adaboost has the highest accuracy of all the models. So, any of these two are used for prediction and prevention of the liver disease.
- Sejin Park et al. in [12], early prediction of disease using the previous real-time stroke symptoms. It is implemented at a low cost. Random Forest algorithm is used for validating clinical significance.
- Ahan Chatterjee et al. in [13], they have used machine learning models like Decision tree, SVM, Random Forest, and so on, find the best classifier using the accuracy and use it to predict cancer disease risk in the early stage and prevent it. Simulation model is also designed to manage the patient flow in OPDs.
- Sergio Grueso et al. in [14], they have used dataset taken from ADNI database and selected 47 out of 159 studies for analysis. Deep learning combined with multimodal and multidimensional data is used to achieve the best performance.
- Upendra Kumar in [15], have used a technique which is a computer screening tool. It checks for disease of an individual and gives the efficient and correct treatment in a fraction of seconds.

### *(A) Existing System*

From the above literature survey, We have inferred that all the systems existing predict only particular diseases namely lung disease, breast cancer, heart disease, diabetes by implementing various algorithms on the particular datasets.

After implementing various algorithms, the most accurate one is selected and it is used for prediction of disease. Sometimes, we may get confused of what algorithm to use. Also, all the systems find only the

particular disease and not the disease based on the symptoms.

### *(B) Proposed System*

We are proposing a system, which uses tkinter for GUI interface. It is a simple user Interface and also time efficient. Our aim with is to get the disease based on symptoms given by the user.

The domain we will use is the machine learning, in that we will be using Naïve Bayes Classifier, which will help us in getting the most accurate predictions easily and also the accuracy is given as output.

To reduce time consuming, we will ask only less questions namely the name of the individual and the symptoms the individual is facing. In this way, our system will be less time consuming and give accurate predictions. And also, we are creating a chatbot using Decision Tree, which will help us in getting accurate predictions by taking into account the symptoms faced by an individual.

### *Algorithm of Proposed System*

- ✓ Collect the dataset.
- ✓ Import the necessary libraries.
- ✓ Visualise the dataset.
- ✓ Train the dataset using Naive Bayes classifier and Decision Tree.
- ✓ Test the model and find the accuracies of both.
- ✓ Deploy the model- a) Naïve Bayes as GUI Interface using Tkinter

  b) Chatbot using Decision Tree
- ✓ Predict the disease based on the symptoms given by user.

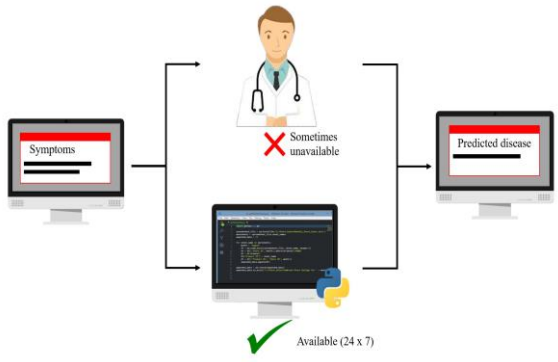


*Fig 1 Proposed system for disease prediction. The doctor may not be available always when needed. By using this system we can predict the disease based on our symptoms anytime.*

## III ARCHITECTURE OF THE SYSTEM

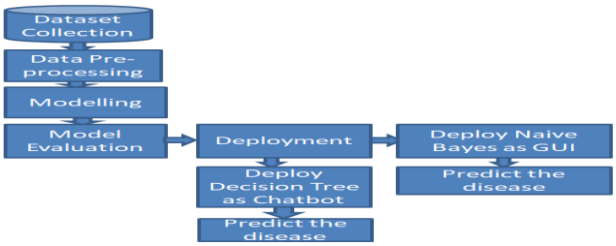


*Fig 2 Architecture of the system*

## IV METHODOLOGY

The modules used in this project are-

- ✓ **Tkinter**- Tkinter is library of python used oftenly by everyone. It is a library which is useed to create GUI based applications easily. It contains so many widgets like radiobutton, textfiled and so on. We have used this for creating account registration screen, login or register screen, prediction interface which is a GUI based application
- ✓ **Sklearn**- Scikit Learn also known as sklearn is a open source library for python programming used for implementing machine learning algorithms. It features various classification, clustering, regression machine learning algorithms. In this it is used for importing machine learning models, get accuracy, get confusion matrix.
- ✓ **Pandas**- Library of python which can be used easily. It gives speed results and also easily understandable. It is a library which can be used without any cost. We have used it for data analysis and to read the dataset.
- ✓ **Matplotlib**- Library of python used for visualising the data using graphs, scatterplots and so on. Here, we have used it for data visualisation.
- ✓ **Numpy**- Library of python used for arrays computation. It has so many functions. We have used this module to change 2-dimensional array into contiguous flattened array by using ravel function.
- ✓ **Pandas Profiling**-This is library of python which can be used by anyone free of cost. It is used for data analysis. We have used this for getting the report of the dataset.

### Naive Bayes Classifier

Naïve Bayes, a supervised machine learning algorithm used for classifying problems which uses Bayes theorem.

Naïve Bayes is mainly used for doing the text classification. The Bayes theorem can be defined by:

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)}$$

P (y|x) is the chance of assumption y for given data x This is also called chance or possibilty of updation.

P(x|y) is the chances of data x given the assumption y was true.

P(y) is the possibility of assumption y being true. This is called the initial possibility or chance of y .

P(x) is the possibility and evidence of data and is called marginal possibility or chance.

**Gaussian Naive Bayes**

It follows the same procedure as the Naive Bayes. But for Naive Bayes we need a categorical dataset and for Gaussian Naive Bayes we need a dataset that has all the continuous features.

**Decision Tree**

Decision Tree an algorithm whose input and output or known. The information is divided repeatedly using the particular parameter. Decision tree consists of two main parts namely decision node and leaf nodes. The decision nodes specify the decision at which the parameter should be spilt. The leaf nodes are the output bought by the decisions. A decision tree asks for either true or false to divide the data.

Project contains three parts:

1. DATASET COLLECTION.
2. TRAIN AND TEST THE MODEL.
3. DEPLOY THE MODEL USING TKINTER.

- ✓ **Dataset Collection-** We had collected dataset from kaggle notebooks. The dataset contains the symptoms and the corresponding disease. It contains 4920 rows and 133 columns.
- ✓ **Train and Test the model-** We had used the Naïve Bayes Classifier as a model to train the dataset. After training, we had tested the model and found its accuracy.
- ✓ **Deploy the model using Tkinter-** Deployed Naive Bayes by creating interface to get the name, symptoms of an individual. By this, we will get the disease and accuracy of model as the output. We have also created a chatbot using Decision Tree which helps an individual to get the corresponding disease by checking whether he/she is being faced by the symptoms.

## V RESULTS

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where essential resources are unavailable and people are also not willing to go outside in fear of spreading virus.

Developing a project based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method is the main objective of the project.

Our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease.This also helps in reduction of the cost and give the correct and fast result.

In addition to the above, to reduce time consuming, our system asks only less questions namely the name of the individual and the symptoms the individual is facing. In this way, our system will be less time consuming and give accurate predictions.

The chatbot which we have created using decision tree will be very helpful in finding the disease because its asks a symptom which individual is facing and also it asks from how many days he/she is facing with that symptoms.

After which, the chatbot will ask the symptoms related to the user given symptoms, combination which causes disease. At last it gives the disease, how much it is affecting an individual, briefing about the disease and its precautions as an output.

| **Algorithm** | **Accuracy** |
|---|---|
| Decision Tree | 97% |

| | |
|---|---|
| Naive Bayes | 100% |

Accuracies Table

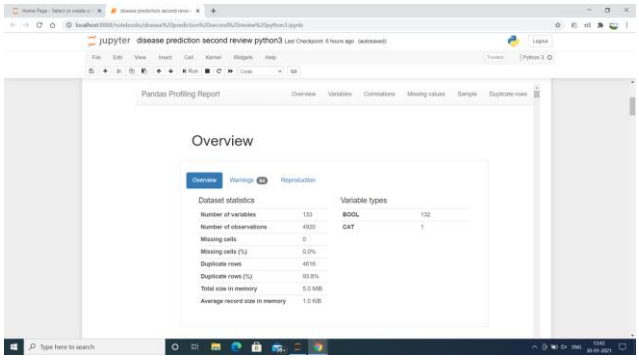


**Fig 3 Report of dataset** which contains all the details of the dataset like how many attributes are present, their correlation and so on.

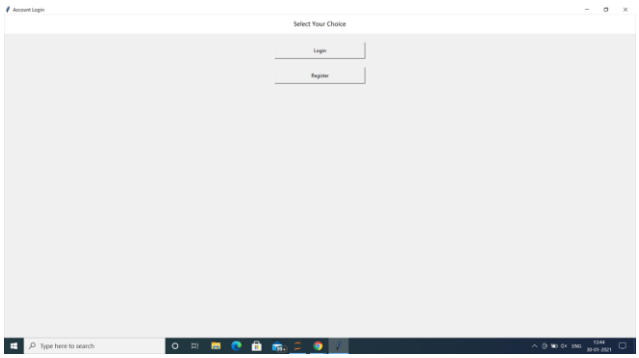


**Fig 4 Account Registration Screen** contains two options which an individual can choose. The two options are login and register. If an individual has already registered, he/she can login. If an individual has not registered, they need to register to get the interface.

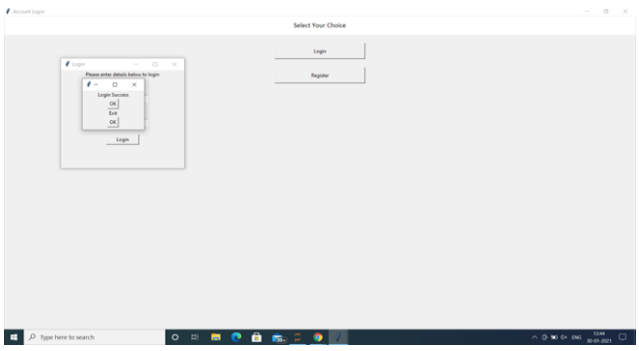


**Fig 5 Login Successful Screen.** If an individual gives the correct details, login successful dialog box appears which asks for two option whether to login or exit.

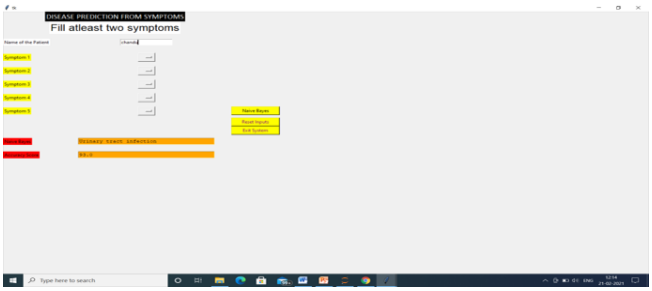


**Fig 6 GUI Interface,** this is the interface which appears after an individual gives the correct details. It asks for name of an individual, 5 symptoms faced. After giving the symptoms, when we click Naïve Bayes, the disease faced by individual and also the accuracy of model is given as an output and also individual can reset the inputs and exit system based on their choice.

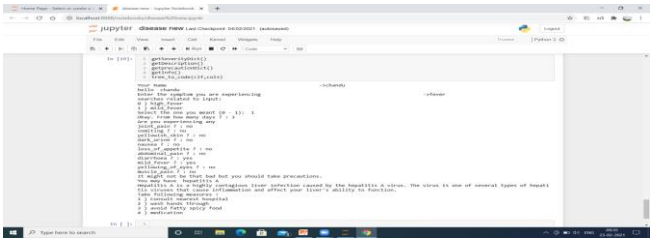

**Fig 7 Chatbot** represents the chatbot of an individual with the machine. The disease, its severity, its definition, and precautions are given as the output.

## VI CONCLUSION

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease.

The project presented the technique of predicting the disease based on the symptoms of an individual patient. Once the disease is predicted, we could easily manage the medicine resources required for the treatment.

Our prediction system will be very helpful for finding the disease based on the symptoms in the early stage and get the correct diagnosis of a disease.This also helps in reduction of the cost and give the correct and fast result.

The chatbot which we have created using decision tree will be very helpful in finding the disease because its asks a symptom which individual is facing and also it asks from how many days he/she is facing with that symptoms.

In addition to the above, to reduce time consuming, our system asks only less questions namely the name of the individual and the symptoms the individual is facing. In this way, our system will be less time consuming and give accurate predictions.

**FUTURE WORK**

A web page which gets symptoms from the user and give the disease as an output can be implemented. And also Naïve Bayes accuracy depends on the symptoms given by the user in tkinter. If we normally fit the Naive Bayes model into the dataset, it is showing 100 percent accurate. The reason for this should be found and solve it.

**REFERENCES**


[1] Tarigoppula V.S Sriram et al, "Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms" International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 3, September 2013

[2]Shubham Bind et al, "A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction" International Journal of Computer Science and Information Technologies, Vol. 6 (2) , 2015, 1648-1655.

[3]K. Gomathi, D. Shanmuga Priya, "Multi Disease Prediction using Data Mining Techniques", 2016.

[4]Isha Pandya et al, "Prediction of Heart Disease Using Machine Learning Algorithms", 2018.

[5] Akash C. Jamgade, Prof. S. D. Zade,"Disease Prediction using Machine Learning", International Research Journal of Engineering and Technology Volume: 06 Issue: 05 May 2019.

[6] Siddhika Arunachalam," Cardiovascular Disease Prediction Model using Machine Learning Algorithms", International Journal for Research in Applied Science & Engineering Technology ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 8 Issue VI June 2020.

[7] Ionela-Catalina ZAMFIR, Ana-Maria Mihaela IORDACHE, "A REVIEW OF DATA MINING TECHNIQUES IN MEDICINE", JOURNAL OF INFORMATION SYSTEMS & OPERATIONS MANAGEMENT, Vol. 14.1, May 2020.

[8] H BENJAMIN FREDRICK DAVID, "IMPACT OF ENSEMBLE LEARNING ALGORITHMS TOWARDS ACCURATE HEART DISEASE PREDICTION", DOI: 10.21917/ijsc.2020.0296.

[9] Harshit Anand et al. "Hridaya Kalp: A Prototype for Second Generation Chronic Heart Disease Detection and Classification",31 July,2020.

[10] Goutam Chakraboty et al. "Predicting the Risk of Chronic Kidney Disease using Machine Learning Algorithm", 11(1), 28 December, 2020, 202.

[11] A. Durga Praveen et al. "Intelligent Liver Disease Prediction system using Machine Learning Models", vol 702.Springer,Singapore. 5 Jan, 2021.

[12] Sejin Park et al. "Machine-Learning-Based Elderly Stroke Monitoring System Using Electroencephalography Vital Signals", 2021, *11*(4), 1761.

[13] Ahan Chatterjee et al. "A Machine Learning Approach to prevent cancer", DOI: 10.4018/978-1-7998-2742-9.ch007, 2021.

[14] Sergio Grueso et al. "Machine Learning methods for predicting conversion from mild cognitive impairment to Alzheimer's disease", 2021.

[15] Upendra Kumar, "Applications of Machine Learning In Disease Pre-Screening", 10.4018/978-1-7998-7705-9.ch049, 2