SMART HEALTH PREDICTION FOR AVOIDING FUTURE HEALTH RISKS

A Capstone project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ELECTRONICS & COMMUNICATION ENGINEERING

by

P. SHIVAPRIYA	**(19K41A04A9)**
S. SUPRAJA LAKSHMI DEVI	**(20K45A0409)**
R. NAGRAJ	**(19K41A0483)**
N. BHARGAVI	**(19K41A0479)**

Under the guidance of

Mr. Rajendra Prasad

Assistant Professor, Department of ECE.

SR ENGINEERING COLLEGE

Ananthasagar, Warangal-506371.

CERTIFICATE

This is to certify that this project entitled **"SMART HEALTH PREDICTION FOR AVOIDING FUTURE HEALTH RISKS"** is the bona-fide work carried out by **P. SHIVAPRIYA , S. SUPRAJA LAKSHMI DEVI, N. BHARGAVI and R. NAGRAJ** as a Capstone project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ELECTRONICS & COMMUNICATION ENGINEERING** the academic year 2022-2023 under our guidance and Supervision.

Mr. Rajendra Prasad   **Dr. Sandip Bhattacharya**

Assoc. Prof.   (ECE),   Assoc. Prof. & HOD (ECE),

SR Engineering College,   SR Engineering College,

Ananthasagar, Warangal.   Ananthasagar, Warangal

ACKNOWLEDGEMENT

ABSTRACT

Now-a-days, Health care industries are literally playing a major role in curing the diseases that are suffering the people. And this will be one kind of help to health care industries. In, present days, people are facing lot of issues related to their health due to their life style and their livelihood. Due to their busy schedule in their lives, people are not at all taking care of their health. They, are not having the time to consult doctors and know what they are going through and it may lead to severe risk for them. So, People should be aware of what they are going through at early stage will reduce the high risk.

Disease Prediction is a supervised model that is used for prediction of diseases from the symptoms or the information provides by the user. This system will process the symptoms entered by the user and provide the predicted disease as an output. In our proposed system we used naive Bayes Classifier, Random Forest Classifier and Decision tree classifier in prediction of the disease.

CONTENTS

5   CONCLUSION & FUTUTRE SCOPE

LIST OF FIGURES

LIST OF TABLES

LIST OF GRAPHS

LIST OF ACRONYMS

ACRONYM	ABBREVIATION
CNN	Convolution Neural Networks
KNN	K-Nearest Neighbors
SVM	Support Vector Machine
NN	Neural Network.
CNN MDRP	Convolution Neural Networks Based multi modal infection chance prediction

CHAPTER-1

INTRODUCTION

1.1 OVERVIEW OF THE PROJECT

Machine learning is used across the globe. They play a Crucial role in the health care Industries. It is the scientific study of various algorithms and statistical models which computer uses to perform a specific task. Machine learning model learns from the data and experience. The machine learning algorithm has two phases: 1) Training & 2) Testing. This model will predict the disease from information provided by the user.

Machine Learning is also used by the healthcare industry to bring improvement in their techniques so that they can provide better quality of service to their patients. The Disease prediction system predicts diseases by analyzing the symptoms of user provided to the system. The system will process the symptoms which are given by the user and gives the predicted disease as an output. Naive Bayes classifier, which is a supervised machine learning algorithm is used in Disease prediction. With the help of this algorithm, we can calculate the probability of the disease.

The purpose of introducing this project called "Smart Health Prediction for Avoiding Future Health Risks" is to predict the disease which disease of a patient by completely analyzing and processing of the symptoms provided by the user. In our project we perform the comparative analysis among the classifiers we used, Decision Tree, Random Forest and Naive Bayes. As we have many samples of data for training as well as Validating data, we perform the analysis of existing as well as proposed classifiers like Random Forest, Decision Tree and Naïve Bayes which we can get the better accuracy and the analysis done through prediction.

1.2 EXISTING METHODS

The big data is applied in medical field both side developing the better growth in both fields, that is big data is applied in medical fields develops the medical fields at the same time increase the growth in big data field. health care sectors. It additionally, provides the more merits gives, (i) medical data analysis with accuracy, (ii) early prediction for disease, (iii) patient-oriented data with accuracy, (iv) The medical data, is securely stored and used in many places, (v) incomplete regional data are reduced and give the accuracy result. Goal of the concept is choosers the region and collects the hospital data or medical data of particular selected region, this process is using the machine learning algorithm. This term based on the data mining technique is used for disease prediction with accuracy. Then, finding the missing data based on latent factor get the incomplete data and it is reduced. The CNN-MDRP is uses the hospital data, that is structured and unstructured data. The CNN-MDRP algorithm- based prediction is produce more accurate, this accuracy is compared with previous system. The advantages of the concept will better feature description and better accuracy, and the disadvantages of this system is, this feature is only applicable for the structured data so it is not good in disease description. Authors, in this paper has proposed in to the concept is machine

learning based disease prediction using the big data for overcome the machine learning drawbacks. The smooth progress of big data is moves in the biomedical and healthcare communities in hospital for accurate results in any experiment result. This concept is (a) reduces the incomplete data and (b) effective disease prediction.

Existing method -1:

"To predict diseases by drugs" and Problem solution-disease prediction, & use of drugs like in fig.1.1 are very harmful, the best medicine prescribed by doctor is suggested to the patients, this is considered as the solution for this model and the methods used are - Random Forest, KNN etc. And the Year of publish Jan-2019 which and successful paper in the internet, providing a wide knowledge to readers. And the main advantage is that - This is the major problem of the youth, which is resolved by the use of ML model. Disadvantages-Accurate drug may not be predicted that easily.

Fig .1.1 drug disease prediction

Existing method -2:

Problem statement of this model is found out to be "Heart disease prediction" and Problem solution- Prediction using input symptoms givenshown in fig 1.2. This is considered as the solution for this model and the methods used are - Linear Regression, K-Nearest Neighbors. And the Year of publish - Aug- 2022which became a famous and successful paper in the internet, providing a wide knowledge to readers. And the main advantage is that, Any kind of heart disease can be easily predicted, such as high heart rate, heart attack. this could be highly beneficial to the patients. And the major Disadvantages- If there is small break in the model this total project goes waste.

Fig .1.2 Heart Disease Prediction

Existing method -3:

Problem statement of this model is found out to be "Diabetes disease prediction". Problem solution-Using Pina Indians diabetes – whose age is 21yrs + in fig 1.3 . This is considered as the solution for this model and the methods used are Decision Tree, random forest, Naive Bayes, SVM. The Authors of this paper and publication are Ms. Mitushi Soni, Dr. Ms. Sunita Varma. And the Year of publish is September-2022. And the main advantage is that, Having Diabetes or not is known prior. This is a huge advantage for the patients. This could be highly beneficial to the patients. Disadvantages- If there is small break in the model this total project goes waste.

Fig .1.3 Diabetes disease prediction

Existing method -4:

Problem statement of this model is found out to be "Malaria disease prediction using image processing", shown in fig 1.4.

Problem solution- Heterogeneous data set That works for only particular photos. This is considered as the solution for this model and the methods used are Robust - homogeneity. And the Year of publish, Aug-2022. And the main advantage is that- people can medicate themselves and be in proper hygiene This could be highly beneficial to the patients. Disadvantages-Sometimes image processing may understand the wrong information.
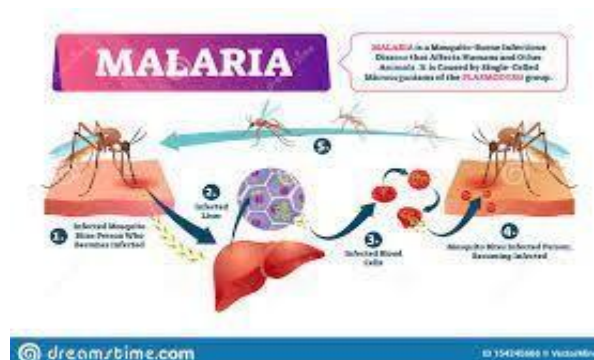

Fig 1.4 Malaria disease prediction

Existing method -5

Problem statement of this model is found out to be "heart disease prediction using hybrid ML", in fig 1.5. Problem solution is Exactness in cardiovascular problems. This is considered as the solution for this model and the methods used are KNN, SVM- NN- Neutral Network. The And the Year of publish, Oct- 2016. And the main advantage is that this model is used to detect heart-attack, high heart rate. Disadvantages- classifiers may take more time to predict the output.
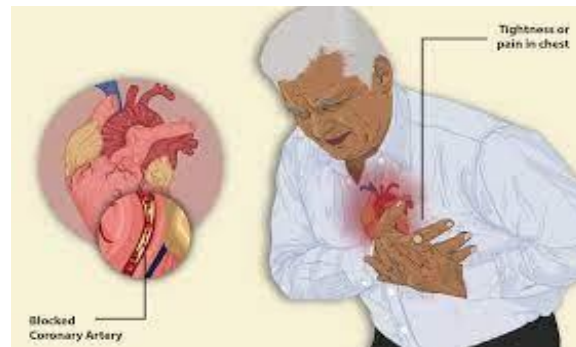
Fig .1.5 Heart Disease Prediction.

Existing method -6:

Problem statement of this model is found out to be "Prediction of muscular paralysis disease based on hybrid feature". shown in fig.1.6 Problem solution is Focused on muscular paralysis using Electromyography –EMG signals. This is considered as the solution for this model and the methods used are Relief-F Selection, SVM- Support vector machine, NN- Neutral Network. And the Year of publish, Feb 2013. And the main advantage is that this model used the Hybrid features like muscle and bone testing is highly beneficial for the patients Disadvantages-Sometimes this model may not be accurate.

.

Fig 1.6 Prediction of muscular paralysis disease

4

Existing method -7:

Problem statement of this model is found out to be "Brain Tumour disease prediction system". Problem solution is multi-modal disease risk prediction model with higher accuracy. This is considered as the solution for this model and the methods used are, Naive Bayes, K-Nearest Neighbors, CNN. The Authors of this paper and publication are Mr. M. Al-Ayyoub, Mr. Ghaith Husari, Mr. Omar Darwish. And the year of publish - April-2012. And the main advantage is that this model used the new technology like neural network which has the perfect output. Disadvantages- only regarding brain is predicted.

Fig 1.7 Brain tumor disease

1.3 MOTIVATION FOR THE WORK:

The main motive of doing this project is to introduce a Disease Prediction Model. Now-a-days there is a significant increase in the number of people who are affected with numerous diseases. Detecting it in the early stage and following some dietary regulations can greatly reduce the risk without even using medication. Further in future, this work will be aimed towards identifying the better classifier for disease prediction. Our main moto is to develop a machine learning model that predicts the disease in the patients.

1.4 PROBLEM STATEMENT:

Managing the health conditions of people is still a challenging problem for middle class person because of Lack of capital, they step back to consult a doctor. This is considered as the major problem of the people. Knowing the disease in prior is far better than, curing and caring.

CHAPTER-2

2.1 LITERATURE SURVEY

In [1] In 2016, The authors associates proposed creating a cardiovascular disease prediction system based on machine learning (CVD). Its goal was to increase the Framing-ham risk score's precision and consistency. The system was put to the test by having its capabilities evaluated using information gathered from 689 people who had symptoms of CVD and a set of verified information from the Framing-ham research. The accuracy of the proposed system, which is substantially higher than that of the existing approaches, was found to be 98.57% by the researchers.

In [2] The authors of this study investigate the various classification algorithms applied in the WEKA platform to identify and forecast the possibility of cardiac disease in individuals. They want to increase the precision of their diagnosis and find hidden patterns using data mining techniques. The most accurate algorithm, J48, outperformed other algorithms when limited error trimming was utilized, according to the study's findings. When UCI data were used, it was also more accurate. Sadly, a clinical model has developed that can forecast a person's risk of developing heart disease. The creation of a more precise and thorough model is still required due to the complexity of the data mining process. Although still in its early stages of development, this technique and will need to be thoroughly checked to ensure that it is not affected by other models. In addition to this, the researchers also need to implement a Multivariate Decision Tree technique to improve the accuracy and efficiency of their model.

In [3] A model created by the authors can be used to accurately forecast the likelihood of diabetes in a patient. To identify diabetes early on, researchers employed SVM, Naive Bayes, and Decision Tree techniques, three machine learning classification algorithms. The research make use of the UCI machine learning repository's Pima Indians Diabetes Database (PIDD). Nonetheless, the development of a prediction model for diabetic sickness did achieve an accuracy of 76%; as a result, they emphasized the need for more complicated models to improve these models' predictability. They asserted that future study may expand and improve the automation of diabetes analysis to include more machine learning methods.

In [4] 14 unique characteristics of CKD patients were examined in [4]. According to the study's findings, decision tree techniques are more accurate than SVM algorithms (91.75% vs. 96.75%). The prediction process advances more swiftly with this technology, which is advantageous. It will facilitate quicker patient diagnosis and earlier CKD patient treatment for medical professionals. A reduced degree of data strength due to the volume of data collection and the missing attribute values is one of this study's flaws. Millions of records with no missing data are necessary for a machine learning model with a target accuracy of 99.99%.

In [4] Machine learning (ML) techniques such as SVM, Random Forest, Decision Trees, Artificial Intelligence, and Naive Bayes were utilized by Auxilia (2018) in [5] to accurately predict liver illness. 11 features and 583 occurrences from the Indian Liver Patient Records data collection were used in this investigation. SVM (77%), Random

Forest (77%), Decision Trees (81%), Artificial Intelligence (71%), and Nave Bayes (37%), with Decision Trees having the highest and Nave Bayes having the lowest accuracy, all produced accurate results.

In [6] The machine learning model developed by the author in [6] analyses five alternative strategies. Compared to Matlab and Weka, the Rapid Miner tool performed more accurately. This study assessed the classification accuracy of the methods Decision Tree, Logistic Regression, Random Forest, Naive Bayes, and SVM. The most accurate algorithm was the decision tree algorithm.

In [7] The Breast Cancer Wisconsin Diagnostic data set has been subjected to five supervised machine learning algorithms: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Decision tree (C4.5), and K-Nearest Neighbors (KNN). When the models were precisely compared, Support Vector Machine outperformed all other techniques, with an efficiency of 97.2%. The WBCD database was the only one for which the results were applicable, and there is no proof that this is true for other datasets. They suggested applying both previously used machine learning methods and more recent machine learning algorithms using brand-new parameters on bigger data sets with more types of illnesses.

In [8] Knowledge discovery was employed as a key data mining application in [8] by the authors who also included a variety of processing phases. Based on the CKD datasets, the research makes an effort to offer a data mining paradigm for knowledge discovery. Data preparation and pre-processing are done using conventional data mining techniques. Decision trees, random forests, and support vector machines are three machine learning methods that are utilised to forecast the early onset of CKD. The Random Forest technique has excellent accuracy (99.16%). As an extension of this research, future work might compare the feature set selection and execution time.

In [9] Author's project sought to create a model for the prediction of cardiovascular illness using machine learning methods. The data for this study were generated from the Cleveland heart disease data collection from the UCI machine learning repository, which has 303 cases and 17 characteristics. The authors used a range of supervised classification techniques, such as k-nearest neighbour (KNN), naive Bayes, decision trees, and random forests. According to the study's findings, the KKN model here demonstrated the highest level of precision, with an accuracy rate of 90.8%. This highlights the need of using the right models and methodologies in order to get the best results while demonstrating the potential value of machine learning techniques for predicting cardiovascular disease.

In [10] Author's selection of three machine learning methods for detecting breast cancer. They include Bayesian networks, RF, and SVM. The Wisconsin dataset, which had been gathered from the UCI Repository, was used by the researchers to assess these learning models. They might contrast the models in terms of recall and accuracy. K-fold validation was used to assess the three models. The simulation results demonstrated that SVM performed better in terms of precision, recall, and accuracy. Yet, RF had a higher probability of successfully identifying the tumour. Yao tested a number of methods and the most accurate when it came to identifying breast cancer.

CHAPTER-3

PROJECT IMPLEMENTATION

3.1 HARDWARE / SOFTWARE TOOLS

Hardware:

System	: Pentium 4, Intel Core i3, i5, i7 and 2GHz
Minimum	
RAM	: 4GB / above
Hard Disk	: 10GB / above
Input	: Keyboard and Mouse
Output	: Monitor or PC

Software:

OS	: Windows 8 / Higher Versions
Platform	: Visual Studio Code / Jupyter Notebook
Program Language	: Python

3.2 USE CASE DIAGRAM

In fig 3.1 we can see that firstly we need to gather the data from the source data set, here, we have included the data set regarding the hospital, and then selecting the algorithms for the model, after the data is introduced, we need to process the data. Also, this processed data is passed through the training and testing datasets model, then it calculates the accuracy of these two models, based on its performance towards the work.

The first five steps of the work is totally done by the developer, to develop this model in a efficient way. And the last two steps include the part done by the user to test the model, which are giving the inputs as symptoms and retrieving the output as the predicted disease.
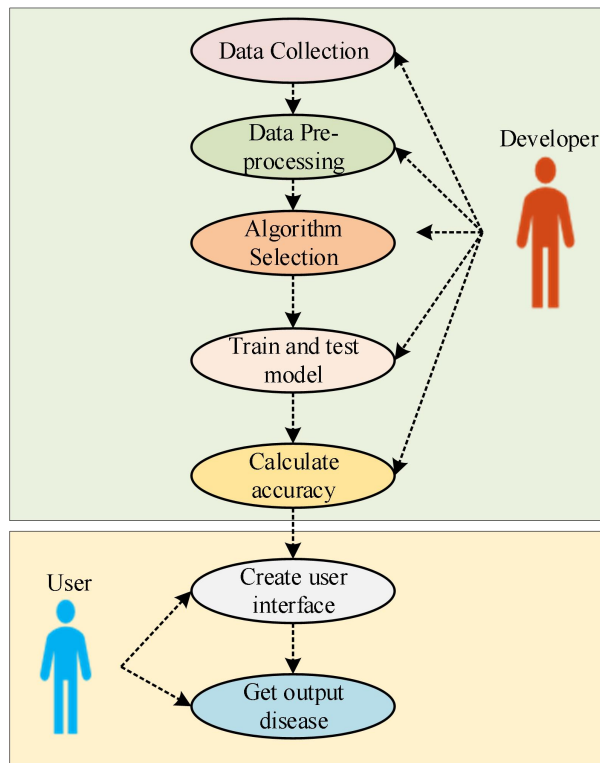
Fig 3.1 Use case diagram of Disease Prediction

3.3 BLOCK DIAGRAM

The data regarding the patient health details are collected from the Kurimotla general hospital database.

Processing the data: After collecting the data, this is given as input to the model for processing the data, processsesing involves the various methods and algorithms to evaluate the data into refined form. Three algorithms are used predict the forthcoming diseases in prior. they are:

Train data:

Training data is also known as training dataset, learning set, and training set. It's an essential component of every machine learning model and helps them make accurate predictions or perform a desired task. Training data builds the machine learning model. It teaches what the expected output looks like. The model analyzes the dataset repeatedly to deeply understand its characteristics and adjust itself for better performance. As it is observed in the fig 2, and fig 6 we see that training data set is identified to be 80 percent.

Test data:

Once we train the model with the training dataset, it's time to test the model with the test dataset. This dataset evaluates the performance of the model and ensures that the model can generalize well with the new or unseen dataset. The test dataset is another

subset of original data, which is independent of the training dataset. As it is observed in the fig 2, and fig 6 we see that testing data set is identified to be 20 percent.

Proposed model:

The model we have used here is the K-fold cross-validation model. K-fold cross-validation improves the model by validating the data. This technique ensures that the model's score does not relate to the technique we use to choose the test or training dataset. K-fold cross-validation method divides the data set into subsets as K number. Therefore, it repeats the holdout method k number of times.

Checking Accuracy:

Accuracy is the degree of how close a calculated or measured value is to the actual value. It measures the statistical error, which is the difference between the measured value and the actual value. The range in those values indicates the accuracy of the measurement. The accuracies of training and testing data are shown in the table I, table II , table III , table IV, table V. In table V we can see the accuracy scores of all the three algorithms in detail (Naive bayes, Decision Tree, Random Forest.
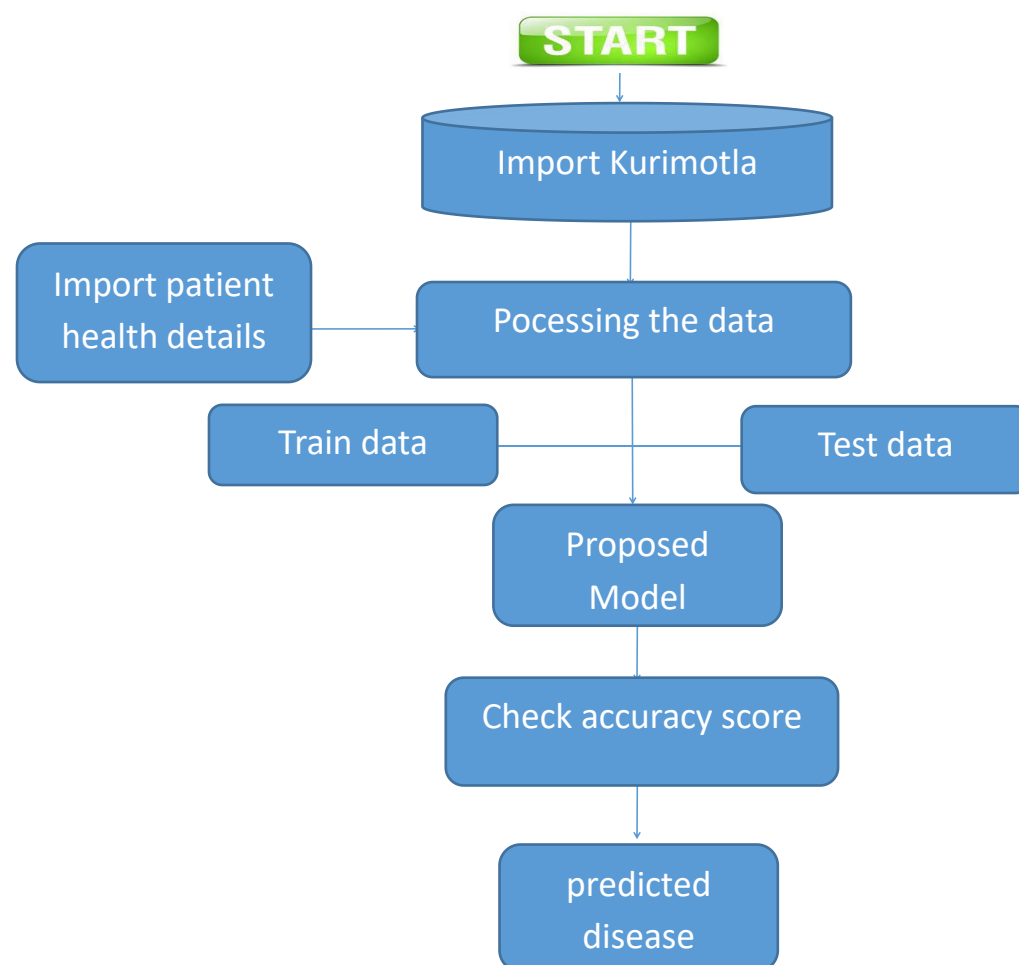

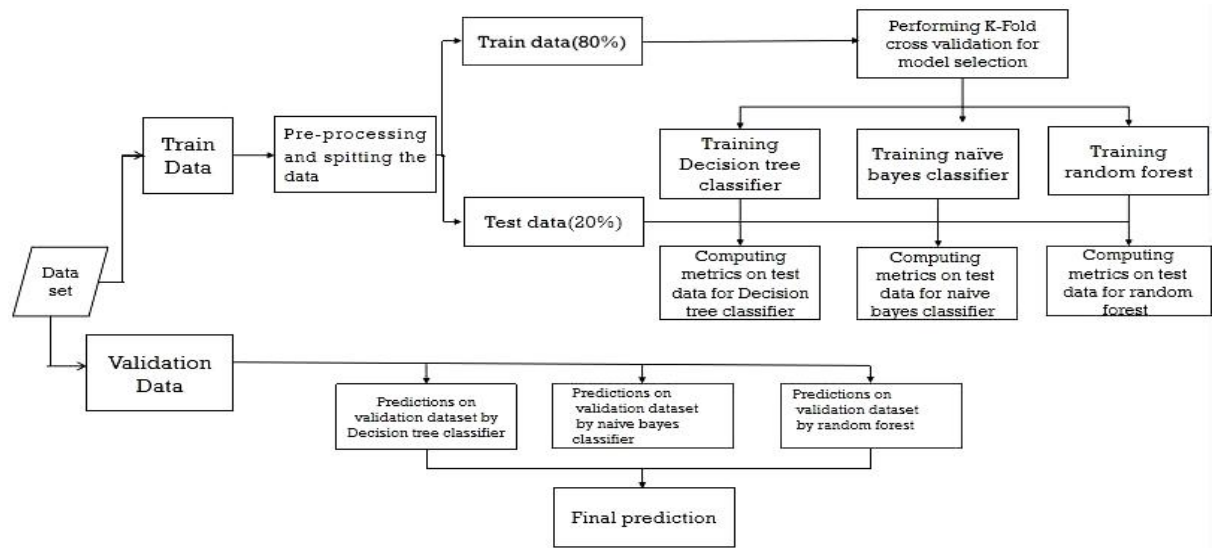Fig. 3.2 Block diagram of Disease Prediction

10

3.4 FLOW CHART

Fig. 3.3 Flow chart of Disease Prediction

3.5 Classifier Naive Bayes Algorithm

These are the classifiers which belong to the family of "Probabilistic Classifiers". These are also called as "Simple Bayes" or "Independence Bayes" is completely based on applying Bayes theorem along with independent assumptions between the events. They are mostly used for machine learning because they are very simple and easy to implement. These are majorly used in solving the classification related problems. It helps in the fast building of machine learning models that gives us fast predicted outcomes. The main utility of this algorithm is to find the probability of an object in a particular event.

$$P(E|F) = P(F|E) *P(E)|P(F)$$

Whereas, P(E|F) is a conditional probability which is the probability of occurrence of any event E, when another event F in relation to E has already occurred.

This also means the probability of event E depends on another event F.

P(B|A) is the probability of event E depends on another event F.

P(A) is the probability of event E.

 P(B) is the probability of event F.

Working of Naive Bayes algorithm:

First it will convert the given source data set into the frequency tabular format.

After finding the probabilities of given data a likelihood table will be generated.

Now we need to apply the Bayes theorem and calculate the posterior probability.
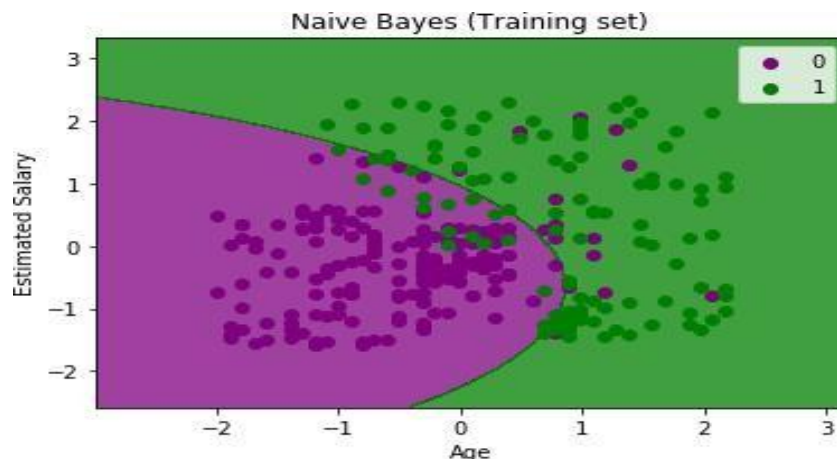

Fig.3.4 Naive bayes algorithm

Advantages:

- ✓ This algorithm is most suitable for mainly solving the multi-class prediction problems.

- ✓ Naive Bayes algorithm is also suitable for the categorical input variables than numerical value.

- ✓ This algorithm performance is very fast and it mainly saves a lot of time.

- ✓ Naive Bayes algorithm assumes all the features or predators are true. So, it performs far better than other models and they require very little amount of training data.

Disadvantages:

- ✓ This algorithm will face the trouble with "Zero-frequency problem" where it can assign zero probability to categorical variable which is in the test data set was not available in the training data set. Hence to overcome this, we need to use smoothing techniques.

- ✓ This algorithm will assume that all its features are independent. So, this algorithm is not used mostly in real world use.

✓ This algorithms estimation can be wrong in fewer cases.

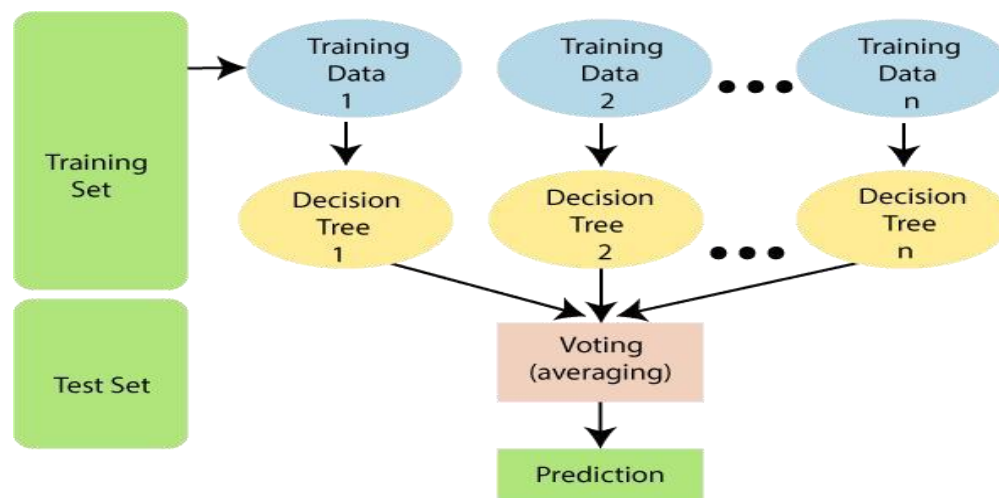## 3.5 Classifier Random Forest Algorithm

The Random Forest Algorithm consists of various types of decision trees, each with the same nodes, but by using different data that can leads to different leaves. It combines the decisions of multiple decision trees to find a single answer that which is the average of all these decision trees.

Random Forest is a machine learning algorithm which belongs to the supervised learning technique. It is used for solving both Classification and Regression problems in Machine Learning, this is based on the ensemble learning, it is a process of joining multiple classifiers to solve a complex problem and to improve the model performance.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given data set and takes the average to improve the predictive accuracy of that data set." Instead of depending on single decision tree, the random forest takes the prediction from each and every tree and it will check the majority votes for particular disease and it predicts the final output disease.

The higher the number of trees in the forest, it leads to high accuracy and it prevents the problem of over-fitting.

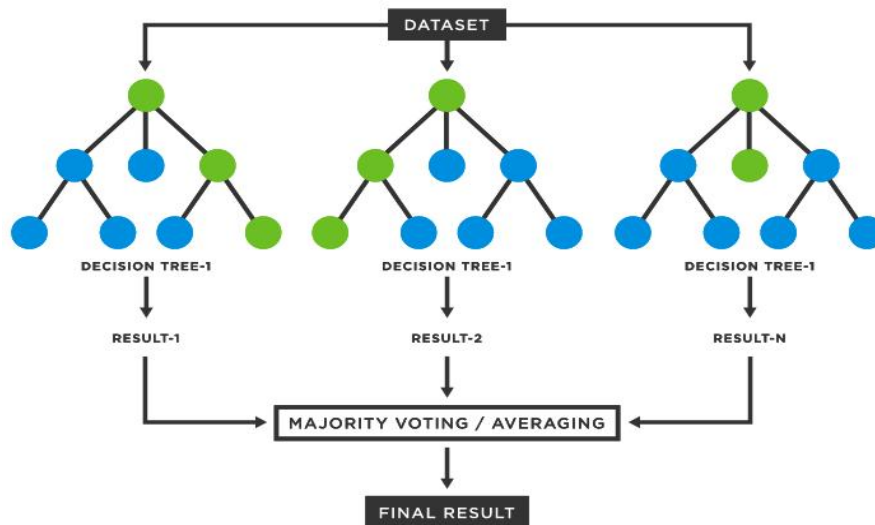The diagram which is shown below explains the working of the Random Forest algorithm:

Fig 3.5 Random Forest algorithm

Below are a few points which explains why do we prefer to use Random Forest Algorithm:

✓ It consumes very less training time as compared to the other classifier algorithms.

✓ It runs very efficiently for larger data set also.

✓ It provides the output with high accuracy.

✓ Even though large proportion of data is missing there will be no change in the accuracy.

Advantages of Random Forest:

✓ Random Forest algorithm will improve the accuracy of the model and prevents the over fitting issues.

✓ It has the capability to perform all the Classification as well as the Regression Tasks.

✓ This is capable to handle large number of datasets with high dimensionality.

Disadvantages of Random Forest

✓ Although it is capable to perform Classification tasks and the Regression tasks, these are not majorly suitable for the Regression tasks.

3.6 Classifier Decision Tree Algorithm

This Algorithm is belonged to the family known as "Supervised Learning Algorithm". Unlike this particular algorithm, this is used for solving the Regression as well as the classification problems too. It is a tree like structured classifier, whereas internal nodes will represent the features of a data set. The branches will represent the decision rules and the each, leaf node represents the outcome.

- ✓ In a Decision tree, there are two nodes, they are Decision Node and Leaf Node.

- ✓ Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

- ✓ The decisions or the test are performed on the basis of features of the given data set.

- ✓ The main aim of using this algorithm is to create a training model which can be used for predicting the class or the value of the target.

- ✓ For building a tree, we need to use CART [Classification and Regression Tree algorithm].

- ✓ A decision tree is going to simply ask a question, and based on the answer whether it is yes or no, further, the tree will split into the sub-trees.
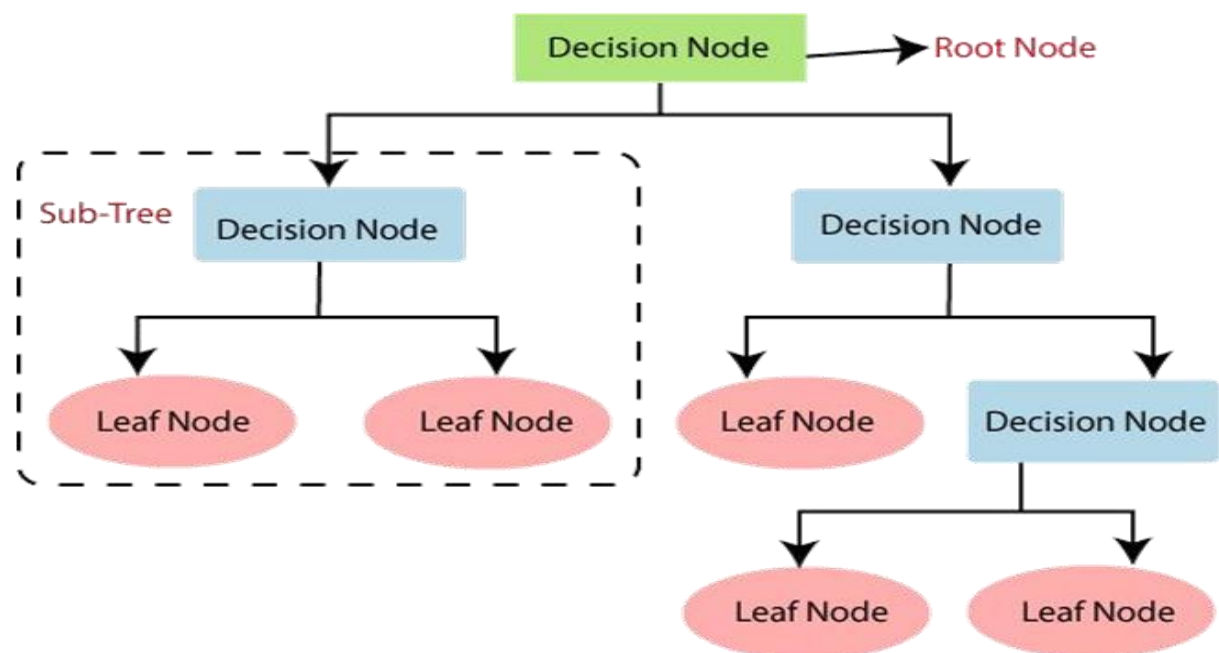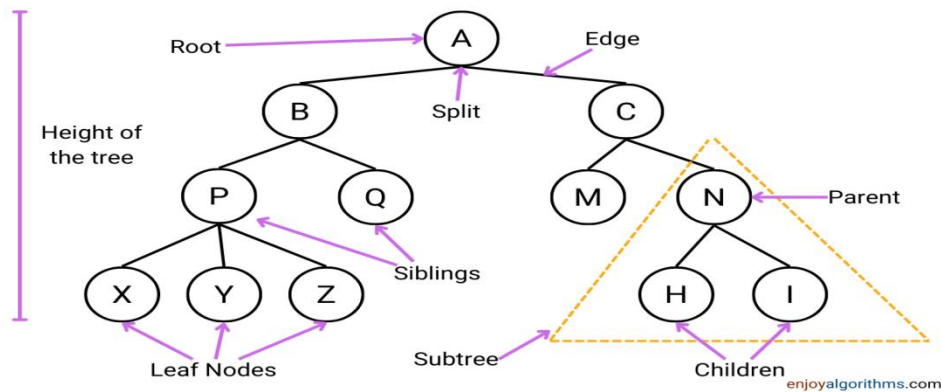
Fig.3.6 Decision Tree algorithm.

Advantages:

- ✓ It is used in solving the classification problems and regression problems.

- ✓ Decision trees are the one of the fastest Way to identify all the relations and variables between two or more variables.

- ✓ These are very simple and easily understandable.

- ✓ Decision trees does not require any kind of transformation of features if we are dealing with non-linear data.

- ✓ This algorithm is scale-invariant.

- ✓ The de4cision tree will not suffer from any kind of missing data values. Hence, it requires very fewer data.

- ✓ These are non-Parametric.

Disadvantages:

- ✓ This algorithm is not used in the Big Data.

- ✓ These are re-usable.

- ✓ In this algorithm over fitting comes into play.

- ✓ It consumes more time for training time complexity.

- ✓ If you want to overcome the limitations of the decision tree, then you should use the random forest method, because it does not depend on a single tree. It creates a forest with multiple trees and takes the decision based on the number of majority of votes.

CHAPTER-4

SIMULATION RESULTS & ANALYSIS

4.1 INTERFACE:

The interface which is displayed in fig 4.1. for disease prediction. Interface consists of Five Slots for symptoms selection. After the selection of symptoms, model uses three algorithms Decision tree, Random Forest and Naive bayes which we installed in it to find out the disease the patient is facing and displays the resultant disease.

The accurate disease is shown, we can observe the accurate disease by the majority of the algorithms detected. For example, if two algorithm's calculations choose the same disease as the output, then the majority disease is considered to be the final output. And that is considered as the predicted disease. And if the remaining one algorithm shows any different diseased then it is not at all considered. Only the majority of the disease is considered.
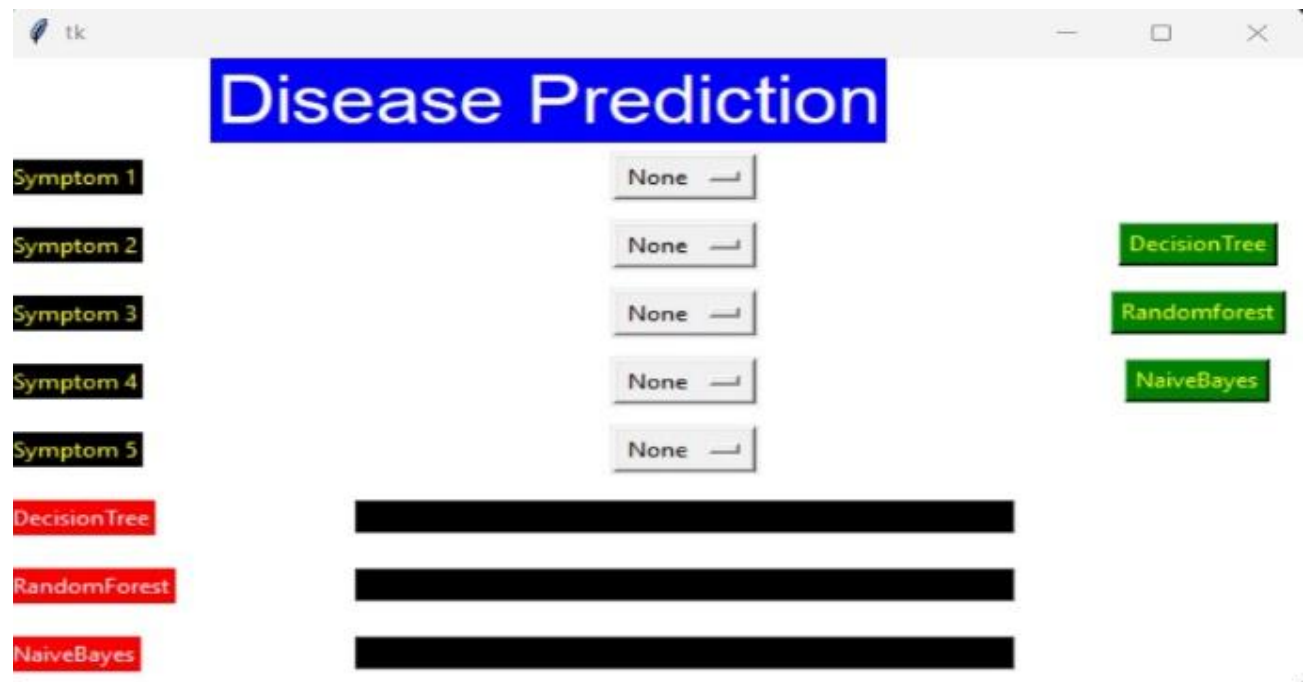
Fig. 4.1 Interface of the user with the model

4.2 RESULT:

This is how disease is predicted. Patient's symptoms should be entered.

Using this algorithms they decide from which disease a patient is suffering and display the diseases as shown in fig 4.2.

Fig 4.2. Predicted disease

4.3 COMPARATIVE RESULTS ANALYSIS:

4.3.1 MODEL ACCURACIES:

Model	Training Accuracy	Testing Accuracy
Logistic Regression	73.4%	73%
Decision Tree	95%	72.4%
Random Forest	95%	70.4%

4.3.2 TRAINING ACCURACIES:
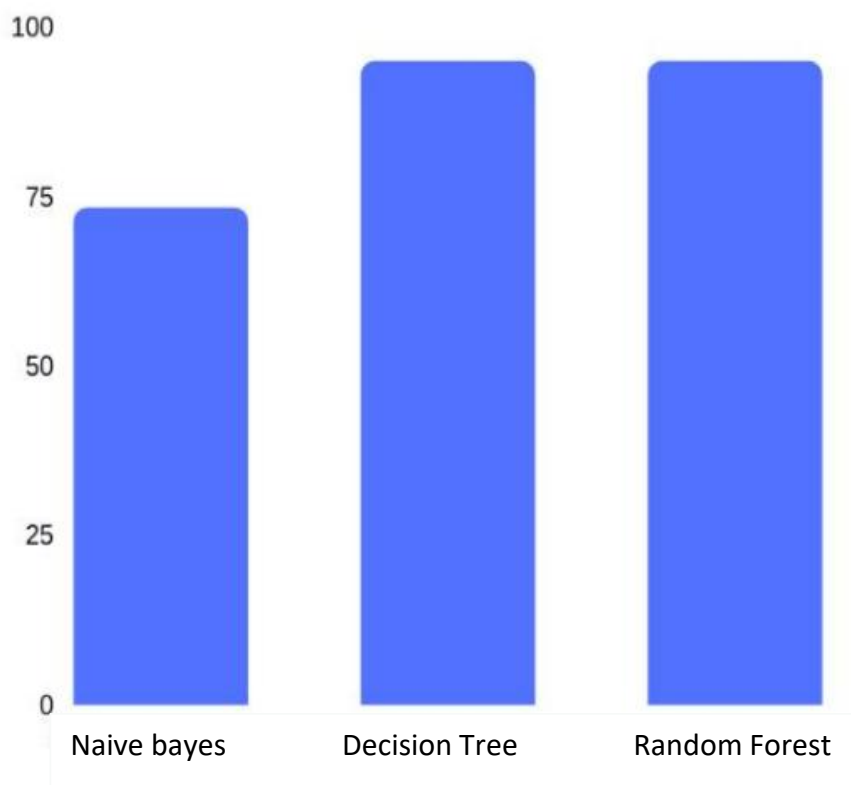
Fig 4.3 Graph of the training accuracies
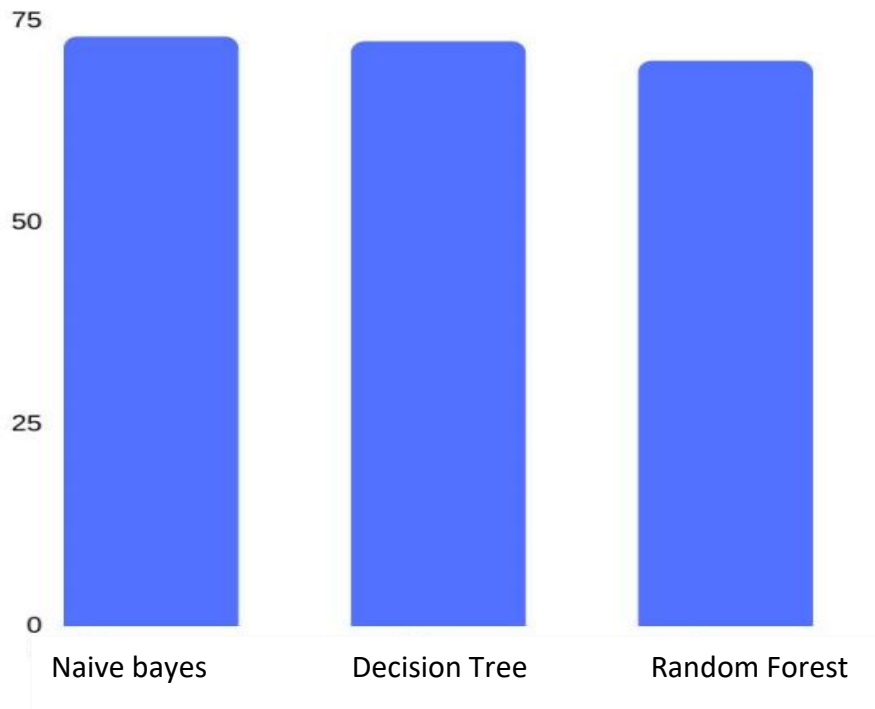
4.3.3 TESTING ACCURACIES:

Fig 4.4   Graph of the testing accuracies

4.3.4 Decision Tree Training Accuracies:

Cost/Depth	1	2	3	4	5	6	None
Entropy	69%	69%	69.7%	71%	71.9%	72.6%	95%
Gini	69%	69.6%	70.7%	71.1%	71.8%	72.8%	95%

4.3.5 Random Forest Training Accuracies:

Cost/Depth	1	2	3	4	5	6	None
Entropy	84.4%	84.6%	90%	90.2%	91.8%	92.2%	95%
Gini	84.6%	84.8%	89.8%	90.2%	91.9%	92.2%	95%

4.3.6   Decision Tree Testing Accuracies:

Cost/Depth	1-2	3	4	5-7	8-9	10	None
Entropy	68.7%	69%	70.7%	71.9%	72.4%	71.8%	64.4%
Gini	69.3%	70%	70.8%	71.9%	72.4%	71.8%	64.4%

4.3.7   Random Forest Testing Accuracies:

Cost/Depth	1	2	3	4	5	6	None
Entropy	64.7%	63.3%	67.2%	66.4%	68.3%	67.4%	70.2%
Gini	64.4%	62.9%	67.4%	65.8%	67.7%	67.4%	70.4%

4.4   APPLICATION:

CHAPTER – 5

CONCLUSION AND FUTURE SCOPE

5.1 CONCLUSION

Now-a-days, diseases are the major killers in our country-India as well as across the globe. The number of people suffering with various kinds of diseases are being increased on a large scale every year. Many people in present days, completely dependent upon the different kind of technologies to find out the solutions for their problems which they cannot solve on their own. In the same way Health care related issues are the one of those things which people can't solve their own problems on their own. Early checkups can help in making the right decisions which can lead to reduction oh high complication risks.

Having a technology like the machine learning will show a great impact on our current society. A full suitable technology can be so beneficial to the health care industries.

But in our project there is an automated model help will be provided which can benefit the people who are really suffering from the various kinds of diseases by providing a few details and information to the model just by only few steps.

Disease Prediction System will play a major role in every individual's life. It will help those people who are suffering with various kinds of diseases knowingly or unknowingly. It helps people in understanding their condition by taking a small test. It is not at all a time -consuming process. They need not to go far places when hospitals are not nearby. They can simply use our model. It gets better over the period of time and the accuracy of disease prediction will also be on rising.

To Conclude in total, we have developed a machine learning model with better accuracy provider at which it predicts whether any person is suffering with any kind of diseases are not.

5.2 FUTURE SCOPE

Further, this system can be used for detecting and predicting the diseases that people are suffering by analyzing the various sources of data sets. Many other different types of symptoms and diseases can be added into this system. Data mining will be more helpful in deciding the procedure of treatment to be followed by obtaining complete knowledge from such databases.

BIBLIOGRAPHY

A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of Heart disease using machine learning," in 2018 Second International Conference on Electronics, Communication and Aerospace Technology(ICECA), 2018, pp. 1275–1278.

Y. Hasija, N. Garg, and S. Sourav, "Automated detection of dermatological disorders through image-processing and machine learning," in 2017International Conference on Intelligent Sustainable Systems (ICISS),2017, pp. 1047–1051.

S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction,"BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1–16, 2019.

R. Katarya and P. Srinivas, "Predicting heart disease at early stages using machine learning: A survey," in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020,pp. 302–305.

P. S. Kohli and S. Arora, "Application of machine learning in disease prediction," in 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018, pp. 1–4.

M. Patil, V. B. Lobo, P. Puranik, A. Pawaskar, A. Pai, and R. Mishra ,"A proposed model for lifestyle disease prediction using support vector machine," in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1–6.

F. Q. Yuan, "Critical issues of applying machine learning to condition monitoring for failure diagnosis," in 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM),2016, pp. 1903–1907.

S. Ismaeel, A. Miri, and D. Chourishi, "Using the extreme learning ma-chine (elm) technique for heart disease diagnosis," in 2015 IEEE CanadaInternational Humanitarian Technology Conference (IHTC2015), 2015,pp. 1–3.

D. Dahiwade, G. Patle, and E. Meshram, "Designing disease predictionmodel using machine learning approach," Proceedings of the 3rd Inter-national Conference on Computing Methodologies and Communication, ICCMC 2019, no. Iccmc , pp. 1211–1215, 2019.

S. Jadhav, R. Kasar, N. Lade, M. Patil, and S. Kolte, "Disease Predictionby Machine Learning from Healthcare Communities," International Journal of Scientific Research in Science and Technology, pp. 29–35,2019.

APPENDIX -I

PYTHON CODE:

```python
from tkinter import *

import numpy as np

import pandas as pd

# from gui_stuff import *

l1['back_pain','constipation','abdominal_pain',

'diarrhoea','mild_fever','yellow_urine',

'yellowing_of_eyes','acute_liver_failure'

'fluid_overload','swelling_of_stomach',

'swelled_lymph_nodes','malaise','blurred_and_distorted_vision','phleg
m','throat_irritatio',

'redness_of_eyes','sinus_pressure','runny_nose',

'congestion','chest_pain',
```

'weakness_in_limbs','fast_heart_rate','pain_during_bowel_

movement',

'pain_in_anal_region','bloody_stool'

,'irritation_in_anus'

,'neck_pain','dizziness',

'cramps','bruising','obesity','swollen_legs',

'swollen_blood_vessels',

'puffy_face_and_eyes','enlarged_thyroid',

'brittle_nails','swollen_extremeties'

,'excessive_hunger','extra_marital_contacts',

'drying_and_tingling_lips','slurred_speech',

'knee_pain','hip_joint_pain','muscle_weakness',

'stiff_neck','swelling_joints',

'movement_stiffness','spinning_movements',

'loss_of_balance','unsteadiness',

'weakness_of_one_body_side'

,'loss_of_smell','bladder_discomfort','foul_smell_of_urine',

'continuous_feel_of_urine','passage_of_gases',

'internal_itching','toxic_look_(typhos)',

'depression','irritability','muscle_pain',

'altered_sensorium','red_spots_over_body','belly_pain',

'abnormal_menstruation','dischromic_patches'

,'watering_from_eyes','increased_appetite',

'polyuria','family_history','mucoid_sputum',

'rusty_sputum','lack_of_concentration',

'visual_disturbances','receiving_blood_transfusion',

'receiving_unsterile_injections','coma',

'stomach_bleeding','distention_of_abdomen',

'history_of_alcohol_consumption',

'fluid_overload','blood_in_sputum',

'prominent_veins_on_calf','palpitations',

'painful_walking','pus_filled_pimples',

'blackheads','scurring','skin_peeling','silver_like_dusting',

'small_dents_in_nails','inflammatory_nails',

'blister','red_sore_around_nose','yellow_crust_ooze']

disease=['Fungal infection', 'Allergy' ,'GERD' ,

'Chronic cholestasis' ,'Drug Reaction','Peptic ulcer

diseae','AIDS','Diabetes','Gastroenteritis',

'Bronchial Asthma','Hypertension',' Migraine',

'Cervical spondylosis',

'Paralysis (brain hemorrhage)','Jaundice',

'Malaria','Chicken pox','Dengue','Typhoid','hepatitis A',

'Hepatitis B','Hepatitis C','Hepatitis D','Hepatitis E',

'Alcoholic hepatitis','Tuberculosis',

'Common Cold','Pneumonia','Dimorphic
hemmorhoids(piles)','Heartattack','Varicoseveins',

'Hypothyroidism','Hyperthyroidism','Hypoglycemia',

'Osteoarthristis','Arthritis','(vertigo)

 Paroymsal Positional Vertigo',

'Acne','Urinary tract infection','Psoriasis','Impetigo']

```python
l2=[]

for x in range(0,len(l1)):

    l2.append(0)

# TESTING DATA

df=pd.read_csv("Training.csv")

df.replace({'prognosis':{'Fungal infection':0,

'Allergy':1,'GERD':2,'Chronic cholestasis':3,

'Drug Reaction':4,
'Peptic ulcer diseae':5,'AIDS':6,'Diabetes ':7,

'Gastroenteritis':8,'Bronchial Asthma':9,

'Hypertension ':10,

'Migraine':11,'Cervical spondylosis':12,'Paralysis (brain

hemorrhage)':13,'Jaundice':14,'Malaria':15,'Chicken pox':16,
```

'Dengue':17,'Typhoid':18,

'hepatitis A':19,'Hepatitis B':20,

'Hepatitis C':21,'Hepatitis D':22,'Hepatitis E':23,'Alcoholic

hepatitis':24,'Tuberculosis':25,'Common Cold':26,

'Pneumonia':27,'Dimorphic hemmorhoids(piles)':28,

'Heart attack':29,

'Varicoseveins':30,'Hypothyroidism':31,'Hyperthyroidism':32,

'Hypoglycemia':33,'Osteoarthristis':34,

'Arthritis':35,'(vertigo) Paroymsal Positional Vertigo':36,

'Acne':37,'Urinary tract infection':38,'Psoriasis':39,

'Impetigo':40}},inplace=True)

X= df[l1]

y = df[["prognosis"]]

```
np.ravel(y)


# TRAINING DATA

tr=pd.read_csv("Testing.csv")


tr.replace({'prognosis':{'Fungal infection':0,'Allergy':1,


'GERD':2,'Chronic cholestasis':3,


'Drug Reaction':4,'Peptic ulcer diseae':5,


'AIDS':6,'Diabetes ':7,'Gastroenteritis':8,


'Bronchial Asthma':9,


'Hypertension ':10,'Migraine':11,'Cervical spondylosis':12,


'Paralysis (brain hemorrhage)':13,'Jaundice':14,


'Malaria':15,'Chicken pox':16,'Dengue':17,


'Typhoid':18,'hepatitis A':19,


'Hepatitis B':20,'Hepatitis C':21,'Hepatitis D':22,
```

'Hepatitis E':23,'Alcoholic hepatitis':24,'Tuberculosis':25,

'Common Cold':26,'Pneumonia':27

,'Dimorphic hemmorhoids(piles)':28,'Heart

attack':29,'Varicose

veins':30,'Hypothyroidism':31,'Hyperthyroidism':32,

'Hypoglycemia':33,'Osteoarthristis':34,'Arthritis':35,

'(vertigo) Paroymsal Positional Vertigo':36,

'Acne':37,'Urinary tract infection':38,'Psoriasis':39,

'Impetigo':40}},inplace=True)

X_test= tr[l1]

y_test = tr[["prognosis"]]

np.ravel(y_test)

def DecisionTree():

```python
from sklearn import tree

clf3 = tree.DecisionTreeClassifier()

clf3 = clf3.fit(X,y)




# calculating accuracy




from sklearn.metrics import accuracy_score  y_pred=clf3.

predict(X_test)

print(accuracy_score(y_test, y_pred,normalize=True))

psymptoms =[Symptom1.get(),Symptom2.get(),Symptom3.

get(),Symptom4.get(),Symptom5.get()]

for k in range(0,len(l1)):

for z in psymptoms:
```

```python
        if(z==l1[k])

            l2[k]=1

    inputtest = [l2]

    predict = clf3.predict(inputtest)

    predicted=predict[0]

    h='no'

    for a in range(0,len(disease)):

        if(predicted == a):

            h='yes'

            break

    if (h=='yes'):

        t1.delete("1.0", END)

        t1.insert(END, disease[a])
```

```python
    else:

        t1.delete("1.0", END)

        t1.insert(END, "Not Found")

def randomforest():

    from sklearn.ensemble import RandomForestClassifier

    clf4 = RandomForestClassifier()

    clf4 = clf4.fit(X,np.ravel(y))

        # calculating accuracy


from sklearn.metrics import accuracy_score

y_pred=clf4.predict(X_test)

print(accuracy_score(y_test, y_pred,normalize=True))
```

```python
psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.
get(),Symptom4.get(),Symptom5.get()]

for k in range(0,len(l1)):

    for z in psymptoms:

        if(z==l1[k]):

            l2[k]=1

inputtest = [l2]
predict = clf4.predict(inputtest)

predicted=predict[0]

h='no'

for a in range(0,len(disease)):

    if(predicted == a):

        h='yes'

Break
```

```python
    if (h=='yes'):

    t2.delete("1.0", END)

    t2.insert(END, disease[a]

    else:

    t2.delete("1.0", END)

    t2.insert(END, "Not Found")

        def NaiveBayes():

    from sklearn.naive_bayes import GaussianNB

    gnb = GaussianNB()

    gnb=gnb.fit(X,np.ravel(y))


    # calculating accuracy

    from sklearn.metrics import accuracy_score
```

```python
y_pred=gnb.predict(X_test)

print(accuracy_score(y_test, y_pred,normalize=True))

psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.

get(),Symptom4.get(),Symptom5.    get()]

for k in range(0,len(l1)):

for z in psymptoms:

if(z==l1[k]):

l2[k]=1


inputtest = [l2]

predict = gnb.predict(inputtest)

predicted=predict[0]

h='no'
```

```python
    for a in range(0,len(disease)):

if(predicted == a):

        h='yes'
        break



    if (h=='yes'):

 t3.delete("1.0", END)

t3.insert(END, disease[a])

else:

        t3.delete("1.0", END)

        t3.insert(END, "Not Found")

# gui_stuff

root = Tk()
```

```python
root.configure(background='white')

# entry variables


Symptom1 = StringVar()

Symptom1.set(None)

Symptom2 = StringVar()

Symptom2.set(None)

Symptom3 = StringVar()

Symptom3.set(None)

Symptom4 = StringVar()

Symptom4.set(None)

Symptom5 = StringVar()

Symptom5.set(None)
```

```python
Name = StringVar()


# Heading


w2 = Label(root, justify=CENTER, text="Disease Prediction",


fg="white", bg="blue")


w2.config(font=("Elephant", 30))


w2.grid(row=1, column=0, columnspan=2, padx=100)


# labels


S1Lb = Label(root, text="Symptom 1", fg="yellow", bg="black")
S1Lb.grid(row=7, column=0, pady=10, sticky=W)


S2Lb = Label(root, text="Symptom 2", fg="yellow", bg="black")
S2Lb.grid(row=8, column=0, pady=10, sticky=W)
```

```python
S3Lb = Label(root, text="Symptom 3", fg="yellow", bg="black")
S3Lb.grid(row=9, column=0, pady=10, sticky=W)




S4Lb = Label(root, text="Symptom 4", fg="yellow", bg="black")
S4Lb.grid(row=10, column=0, pady=10, sticky=W)




S5Lb = Label(root, text="Symptom 5", fg="yellow", bg="black")
S5Lb.grid(row=11, column=0, pady=10, sticky=W)




lrLb = Label(root, text="DecisionTree", fg="white", bg="red")
lrLb.grid(row=15, column=0, pady=10,sticky=W)




destreeLb = Label(root, text="RandomForest", fg="white", bg="red")
destreeLb.grid(row=17, column=0, pady=10, sticky=W)




ranfLb = Label(root, text="NaiveBayes", fg="white", bg="red")
ranfLb.grid(row=19, column=0, pady=10, sticky=W)




# entries
```

```python
OPTIONS = sorted(l1)


S1En = OptionMenu(root, Symptom1,*OPTIONS)

S1En.grid(row=7, column=1)




S2En = OptionMenu(root, Symptom2,*OPTIONS)

S2En.grid(row=8, column=1)




S3En = OptionMenu(root, Symptom3,*OPTIONS)

S3En.grid(row=9, column=1)




S4En = OptionMenu(root, Symptom4,*OPTIONS)

S4En.grid(row=10, column=1)




S5En = OptionMenu(root, Symptom5,*OPTIONS)

S5En.grid(row=11, column=1)




dst = Button(root, text="DecisionTree", command=DecisionTree,bg="green",fg="yellow")

dst.grid(row=8, column=3)
```

```python
rnf = Button(root, text="Randomforest", command=randomforest,bg="green",fg="yellow")
rnf.grid(row=9, column=3,padx=10)




lr = Button(root, text="NaiveBayes", command=NaiveBayes,
bg="green",fg="yellow")
lr.grid(row=10, column=3,padx=10)




#textfileds

t1 = Text(root, height=1, width=40,bg="black",fg="white")
t1.grid(row=15, column=1, padx=10)




t2 = Text(root, height=1, width=40,bg="black",fg="white")
t2.grid(row=17, column=1 , padx=10)




t3 = Text(root, height=1, width=40,bg="black",fg="white")
t3.grid(row=19, column=1 , padx=10)
    root.mainloop()
```

Paper publication letter

2 messages

Microsoft CMT <email@msr-cmt.org>
Apr 12, 2023 at 10:06 PM

Reply-to: Microsoft CMT - Do Not Reply <noreply@msr-cmt.org> To: naraganib@gmail.com

Hello,

The following submission has been created. Track Name: ASIANCON2023

Paper ID: 114

Paper Title: Smart Health Prediction using Machine Learning Abstract:

Now-a-days, Health care industries are literally playing a major role in curing the diseases that are suffering the people. And this will be one kind of help to health care industries. In present days, people are facing lot of issues related to their health due to their life style and their livelihood. Due to their busy schedule in their lives, people are not at all taking care of their health. They are not having the time to consult doctors and know what they are going through and it may lead to severe risk for them. So, People should be aware of what they are going through at early stage will reduce the high risk. In our proposed system we used Naïve Bayes Classifier, Random Forest Classifier and Decision tree classifier in prediction of the disease. Disease Prediction is a supervised model that is used for prediction of diseases from the symptoms or the information provides by the user. This proposed system will process the symptoms entered by the

user and provide the predicted disease as an output.

Created on: Wed, 12 Apr 2023 16:36:03 GMT

Last Modified: Wed,

12 Apr 2023

16:36:03 GMT

Authors:

- **chrprasad20@gmail.com** **(Primary)**
- **pillalamarrishivapriya@gmail.com**
- **naraganib@gmail.com**
- **nagarajpatel999@gmail.com**
- **suprajalakshmidevi366@gmail.com**
- **ksreedhar829@gmail.com**

Secondary Subject Areas: Not Entered

Submission Files: Manuscipt11.docx (483 Kb, Wed, 12 Apr 2023 16:35:57 GMT)

Submission Questions Response: Not Entered Thanks, CMT team.

Download the CMT app to access submissions and reviews on the move and receive notifications:
https://apps.apple.com/us/app/conference-management-toolkit/id1532488001
https://play.google.com/store/apps/details?id=com.microsoft.research.cmt

bhargavi naragani <naraganib@gmail.com>

at, Apr 22, 2023 at 7:09 PM