A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives

Anuj Kumar [a], Mr.Analp Pathak [b]

[a] Student, Department of CSE & SRM Institute Of Science And Technology, NCR Campus , India
[b]Assistant Professor , Department of CSE & SRM Institute Of Science And Technology, NCR Campus , India

Abstract: Machine Learning Approach for Identifying Disease Prediction Using Machine Learning is based on prediction modelling that predicts disease of the patients according to the symptoms provided by the users as an i/p to the system. This paper gives an idea of predicting multiple diseases using Machine Learning algorithms. Here we will use the concept of supervised Machine Learning in which implementation will be done by applying Decision Tree, Random Forest, Naïve Bayes and KNN algorithms which will help in early prediction of diseases accurately and better patients care. The results ensured that the system would be functional and user oriented for patients for timely diagnoses of diseases in a patient.

Keywords: Machine Learning , Disease Prediction , Decision Tree, Random Forest.

1. Introduction

The Earth is going through a purplish patch of technology where the demand of intelligence and accuracy is increasing behind it [11]. Today's people are likely addicted to internet but they are not concerned about their physical health. People ignore the small problem and don't visit to visit hospital which turn into serious disease with time [11] . Taking the advantage of this growing technology, our basis aim is to develop such a system that will predict the multiple diseases in accordance with symptoms put down by the patients without visiting the hospitals / physicians.

Machine Learning is a subset of AI that is mainly deal with the study of algorithms which improve with the use of data and experience. Machine Learning has two phases i.e. Training and Testing [17]. Machine Learning provides an efficient platform in medical field to solve various healthcare issues at a much faster rate. There are two kinds of Machine Learning – Supervised Learning and Unsupervised Learning. In supervised learning we frame a model with the help of data that is well labelled. On the other hand, unsupervised learning model learn from unlabeled data.

The intent is to deduce a satisfactory Machine Learning algorithm which is efficient and accurate for the prediction of disease. In this paper, the supervised Machine Learning concept is used for predicting the diseases. The main feature will be Machine Learning in which we will be using algorithms such as Decision Tree, Random Forest, Naïve Bayes and KNN which will help in early prediction of diseases accurately and better patient care [11] .

2. Objective

There is a demand to make such a system that will help end users to predict diseases on the basis of symptoms given in it without visiting hospitals. By doing so, it will decrease the rush at OPD's of hospitals and bring down the workload on medical staff. Not only this, this system will reduce the costly treatment and panic moment at the end stages so that proper medication can be provided at the right time and we can lower down the death rate as well. This system also consists of a feature of Database which stores the data entered by the end users and the name of the disease the patient is suffering from that can be used as a past record and will help in further treatment in future. The analysis accuracy is increased by using Machine Learning algorithms. Altogether this system will help in easier health management.

3.Related Work

There are numerous work that has been done related to disease prediction system using different Machine Learning algorithms and achieved different results for different methods in medical field.

The paper [1] “Disease Prediction System” used Decision tree, Random forest and Naïve Bayes algorithms to predict a disease on the basis of systems and to enable synchronized and well versed medical systems ensuring maximum patient satisfaction

The paper [2]” Heart Disease Prediction with Machine Learning Approaches" made use of LR,NB,KNN,SVM,DT and RF algorithms for prediction of heart disease with proper data processing and implementation of ML algorithm with different parameters and among all Machine Learning algorithms, the highest accuracy is achieved by KNN with 87%.

The paper [3] ”Heart Attack Prediction By Using Machine Learning Techniques" has compared various Machine Learning models with the help of performance metrics and to detect heart related problems with highest accuracy of 89.34% by SVM.

The paper [4] ”Disease Prediction Using Machine Learning over Big Data” has proposed a CNN-MDRP algorithm which combines structured and unstructured data and proved that CNN-MDRP is more accurate than previous prediction algorithm.

The paper [5]” A Review of Heart Disease Prediction Using Machine Learning and Data Analytics Approach” used different types of DM and ML methods to predict the happening of heart disease and apply the proposed system for the area it needed.

The paper [6] “Application of Machine Learning Predictive Models in the Chronic Disease " focused on SVM and LR algorithms andevaluate the study models associated with diagnosis of chronic disease. These models are highly applicable in classification and diagnosis of CD.

The paper [7] “COVID -19 Outbreak Prediction with Machine Learning" made use of MLP and ANBEIS andpresented a comparative analysis of ML and soft models to predict covid-19 outbreak and provides initial benchmarking to demonstrate the potential of ML for future.

The paper [8] "Heart Disease Prediction System Using Machine Learning" has built a heart disease prediction system using NB algorithm that provides 88.163% accuracy among others.

The paper [9] "Heart Disease Prediction Using Machine Learning" suggested a robust model to predict a heart disease and found that Logistic Regression algorithm has the most efficient with an accuracy of 82.89% followed by DT and NB with 80.40% each and SVM was having 81.75%.

The paper [10] " Implementation of Machine Learning Model to Predict Heart Failure Disease” has explored, recommended and applied a Machine Learning model in which Rapid Miner tool is used that calculated the high degree of correctness than Matlab and Weka tool.

The paper [11] “Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively” has proposed a multiple disease prediction system which provides medicine and drug consultation of disease predicted.

The paper [12] “Heart Disease Prediction Using Data Mining Techniques” used KNN, Naïve Bayes and SVM algorithms and collated with respect to the accuracy using heart disease dataset and achieved the highest accuracy of 86.6% using Naïve Bayes.

The paper [13] “Heart Disease Prediction Using Machine Learning Techniques” has proposed a method for heart disease prediction using Machin Learning and the results showed a great accuracy standards for producing a better estimated results.

The paper [14] “Disease Prediction using Machine Learning” used KNN, Naïve Bayes, Logistic Regression and Decision Tree algorithms to make a disease prediction system which can predict the disease on the basis of symptoms and implemented using grails framework.

The paper [15] “Disease Prediction using Machine Learning” used Naïve Bayes, Decision Tree and Random Forest algorithms to create a disease prediction system with better accuracy and it also provides motivational thoughts and images..

4. Proposed System

We are proposing such a system that will flaunt a simple, cost effective , elegant User Interface and also be time efficient . Our proposed system bridges the gap between doctors and patients which will help both classes of users to achieve their goal. This system is used to predict diseases according to symptoms. In this proposed system we are going to take down five symptoms from the users and evaluate them by applying algorithms such as Decision Tree, Random Forest , Naïve bayes and KNN which will help in getting accurate prediction .Our system will explore and merge more datasets which includes large diversity of population to get more effective results and thus our system will improve and enhances the accuracy of the results. Along with the increased accuracy rate, we will proliferate the reliability of our system for this job and can gain the trust of patient in this system. Apart from all these, our system will comprise of a Database for storing the data entered by the users and the name of the disease the patient is suffering from which can be used as a reference in future for further treatment. Hence this system will contribute in easier health management with better satisfaction to the users.

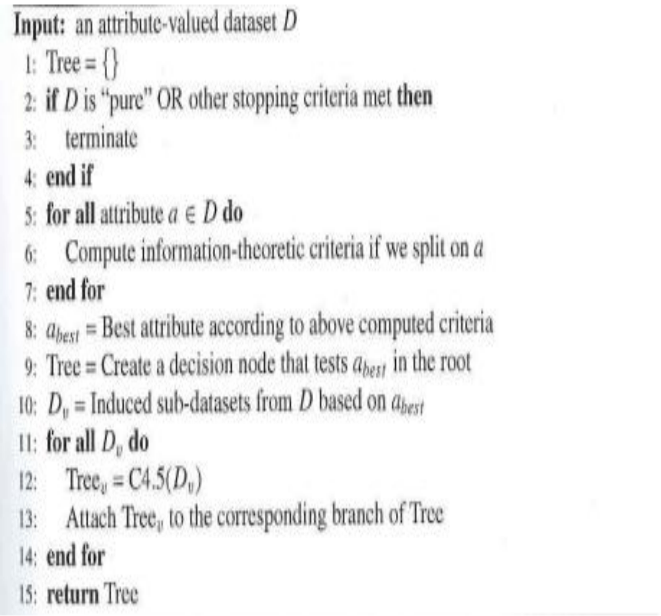
4.1. Methodology

Our project is stand on multiple disease prediction in accordance with symptoms entered by patient. The first task is to determine the problem statement. Then making the dataset ready to work on. After that we conceptualize our data using scatter plot, distribution graph, etc. by doing so we can find out anomalies, missing values, etc. on our data and make our dataset perfect for prediction. And finally the main feature will be Machine Learning in which we will be using algorithms like Decision Tree, Random Forest, Naive Bayes and KNN which will predict accurate disease for early prediction and better patient care. For this model, we have used python as a platform to execute our Machine Learning algorithms. We have also developed an elegant GUI to provide interaction with system.

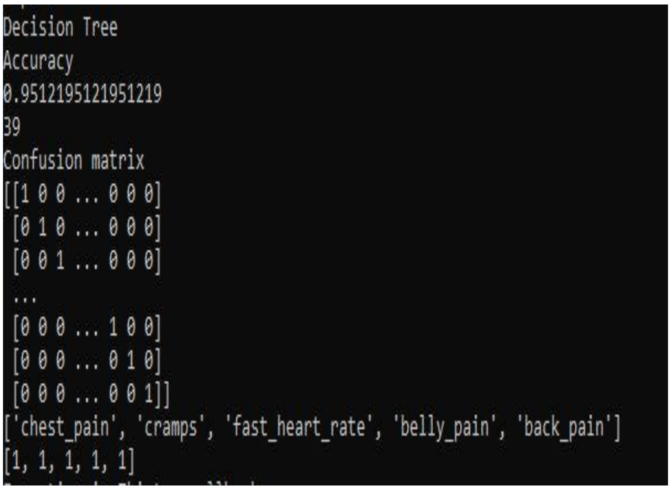
4.1.1. Decision Tree

It is a type of supervised Machine Learning algorithm that mainly deal with classification problem. The main objective of using decision tree is to make a training model that can be used to predict the class or values of the desired value by learning elementary decision procedure surmise from existing data (training data) [9]. In Decision Tree algorithm, we start from root of the tree to predict the class. We collate the values of the root trait with data's trait. On the basis of differentiation we go ground with the branch parallel to that value and move to next node [9] . In this system Decision Tree splits the symptoms as per its classification and lowers down the dataset complexity. It is most effective Machine Learning algorithm to describe Decision tree in graphical manner [2]. It deals with huge and complicated datasets without involvement of multiple parametric structure. With the help of training datasets , decision tree model is decided and a validation dataset decides appropriate tree size to achieve the optimal final model.

(1) Pseudo Code



```
Input: an attribute-valued dataset $D$
Tree = {}
if $D$ is "pure" OR other stopping criteria met then
    terminate
end if
for all attribute $a \in D$ do
    Compute information-theoretic criteria if we split on $a$
end for
$a_{best}$ = Best attribute according to above computed criteria
Tree = Create a decision node that tests $a_{best}$ in the root
$D_v$ = Induced sub-datasets from $D$ based on $a_{best}$
for all $D_v$ do
    $Tree_v$ = C4.5($D_v$)
    Attach $Tree_v$ to the corresponding branch of Tree
end for
return Tree
```



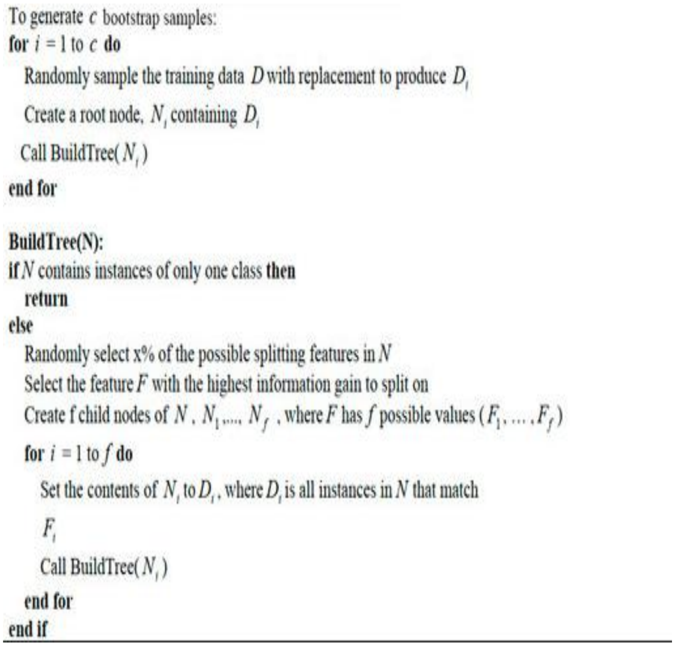
```
Decision Tree
Accuracy
0.9512195121951219
39
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
['chest_pain', 'cramps', 'fast_heart_rate', 'belly_pain', 'back_pain']
[1, 1, 1, 1, 1]
```

4.1.2. Random Forest

Random Forest comes under category of supervised Machine Learning algorithm. It is used for classification and regression but mainly deal with classification problems. The implementation of Random Forest is very easy and easy in use as well [2]. Random Forest is a perfect substitute if we would like to develop a model in short notice [1]. Random Forest is an ensemble learning method that works by creating a horde of decision tree at training time [1]. It selects the best solution by means of voting [2]. Simply Random Forest is composed of multiple decision tree [2]. It creates forest of trees [2]. The number of tress in the forest is directly proportional to the accuracy rate and it prevents the problem of overfitting. Random Forest produces good results over actual problems mainly due to insensitive to noise in the dataset and is not based on overfitting. It works greatly and shows an excellent execution over other tree based algorithms. For tree learning, it mainly use bootstrap aggregation or bagging [13].

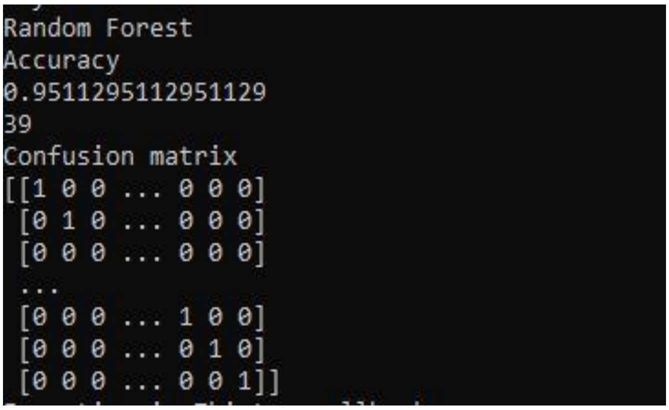
For a given data, X = {m1,m2,m3,…mn} with responses Y = {m1,m2,m3,….mn} which repeats the bagging from b=1 to B [13].

(2) Pseudo Code



```
To generate $c$ bootstrap samples:
for $i = 1$ to $c$ do
    Randomly sample the training data $D$ with replacement to produce $D_i$
    Create a root node, $N_i$ containing $D_i$
    Call BuildTree( $N_i$ )
end for

BuildTree(N):
if $N$ contains instances of only one class then
    return
else
    Randomly select x% of the possible splitting features in $N$
    Select the feature $F$ with the highest information gain to split on
    Create f child nodes of $N$ , $N_1$,..., $N_f$ , where $F$ has $f$ possible values ( $F_1$, ... , $F_f$ )
    for $i = 1$ to $f$ do
        Set the contents of $N_i$ to $D_i$ , where $D_i$ is all instances in $N$ that match
        $F_i$
        Call BuildTree( $N_i$ )
    end for
end if
```



```
Random Forest
Accuracy
0.9511295112951129
39
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
```

4.1.3. Naïve Bayes

Naive Bayes is a type of probabilistic algorithm which is based on probability theory and Bayes Theorem to calculate the probability of diseases [9]. A Naïve Bayes algorithm has a parallel performance with decision tree and other selected classifiers [2]. The computation cost can be brought down significantly [2]. It is very simple to build and useful for large dataset. Naive Bayes classify the data by calculating the probability of independent variable [10]. After the probability of each class is computed , complete transaction is assigned to high probability class [10]. Naïve Bayes works excellent in various complex real world problem . the benefit of using Naïve Bayes is that it needs very less amount of training dataset to evaluate the parameters necessary for classification.

Bayesian rational is functional to decision maker. The portrayal for Naive Bayes is probabilities. It is based on probability theory and Bayes Theorem to forecast the class value of unexplored dataset [12]. A record of likelihood is kept in a report for a learned Naive Bayes model [12].

(3) Pseudo Code

Learning Phase: It is quite swift to learn a Naive Bayes classifier using the training data. A training set D with M traits and N categories is given for each desired value of ao (ao=a1,……aN)

```
P(ao)← evaluate P(ao) with sample in D;
For every trait value djkof each trait dj(j=1,...,M;k=1,.,Zj)
P(dj=djk | ao)← evaluate P(djk|ao)with samples in D;
Result :M*N relative probabilistic models.
```

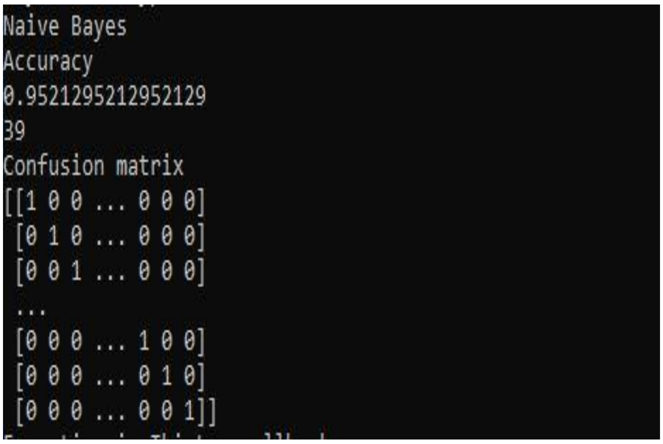
Testing Phase: Since only the likelihood of each group and the likelihood of each group given distinct insert (b) values required to be evaluated so the training is fast. Consider an unknown example b'=(k'1,...,k'n)

Look up tables to allocate the tag a* to B'' if

[$\bar{P}$(k'1|a*).....$\bar{P}$(k'n|a*)]

$\bar{P}$(a*)>[$\bar{P}$(k'1|ai)…...$\bar{P}$(kn|ai)]

$\bar{P}$(ai),ai ≠a*,ai=a1,….,aN



```
Naive Bayes
Accuracy
0.9521295212952129
39
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
```

4.1.4. KNN

KNN is a supervised learning algorithm that use data and classify new data points on similarity measures . It is a non parametric method that implies it does not consider any assumption on underlying data. KNN is also a lazy learner method because it does not perform training at all and it does not pursue any discriminatory function from training data instead it retains the training dataset. KNN based on the idea of feature similarity approach i.e. it considers that the homogeneous things exist in a close proximity. It is also an instance based learning algorithm where the function is approximately locally [2]. This algorithm handles large of amount of data and used where there are non linear decision division between classes .KNN not only handles function approximation problem but also robust to noisy training data. KNN is used to calculate distance between new data point and each training point using distance function.

Suppose (yj, dj) where i = 1, 2……., m be data points.

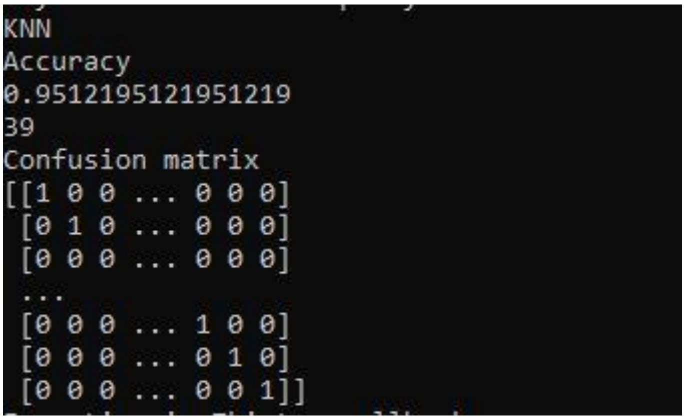
yj represents characteristic values & dj represents tag for each j. Considerig the number of classes as 'd' [13].

dj ∈ {1, 2, 3, ……, d} for all values of j

Consider y be a fact for which label is not identified, and will determine the tag class using KNN algorithm[13].

(4) Pseudo Code

- Compute “d(x, x_i)” i =1, 2, ….., m; in which d represents the Euclidean distance between the points is determined asgiven below.
- Distance= $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
- Arrange the computed m Euclidean distances in upgraded manner .
- Consider z be a positive number, select prime z distances from this orderedlist.
- Locate those z-points parallel to thesez-distances.
- Suppose $z_i$represents the amount of facts fitting to the ith class among z points i.e. z $\geq$0
- Whether $z_i > z_j$ $\forall i \neq j$ then place x in classi.



```
KNN
Accuracy
0.9512195121951219
39
Confusion matrix
[[1 0 0 ... 0 0 0]
 [0 1 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 1 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]
```

5.Dataset And Model Description

In this section we are going to elaborate the dataset which is used in this project to train the Machine Learning model. The dataset we have used for this project is in the structured format [16]. The dataset which is being used contains all the names of diseases with its respective symptoms. Since this system is based on supervised Machine Learning algorithm, the dataset is labelled with 0 or 1. After this, we have divided the dataset into two phases i.e. training dataset and testing dataset . we trained our models using training dataset and then we applied our all Machine Learning algorithms to this training dataset to get trained Machine Learning model [17]. At last we provided the testing dataset to this trained model to test the accuracy of model.

6. Dataset Of Hospital

The dataset of hospital will be in the form of structured format .The dataset that we used is real life hospital data and data stored in data center. The data provided by hospital contains symptoms of patients. The dataset is connected into either 0 or 1 where the value 0 represents feature/ symptoms impacts on disease and the value 1 represents that it does not have any impact on disease [16]. This dataset is a disease symptoms produced by a mechanized system on the basis of information in textual discharge summaries of patients at New York Presbyterian hospital [17].

7. System Architecture

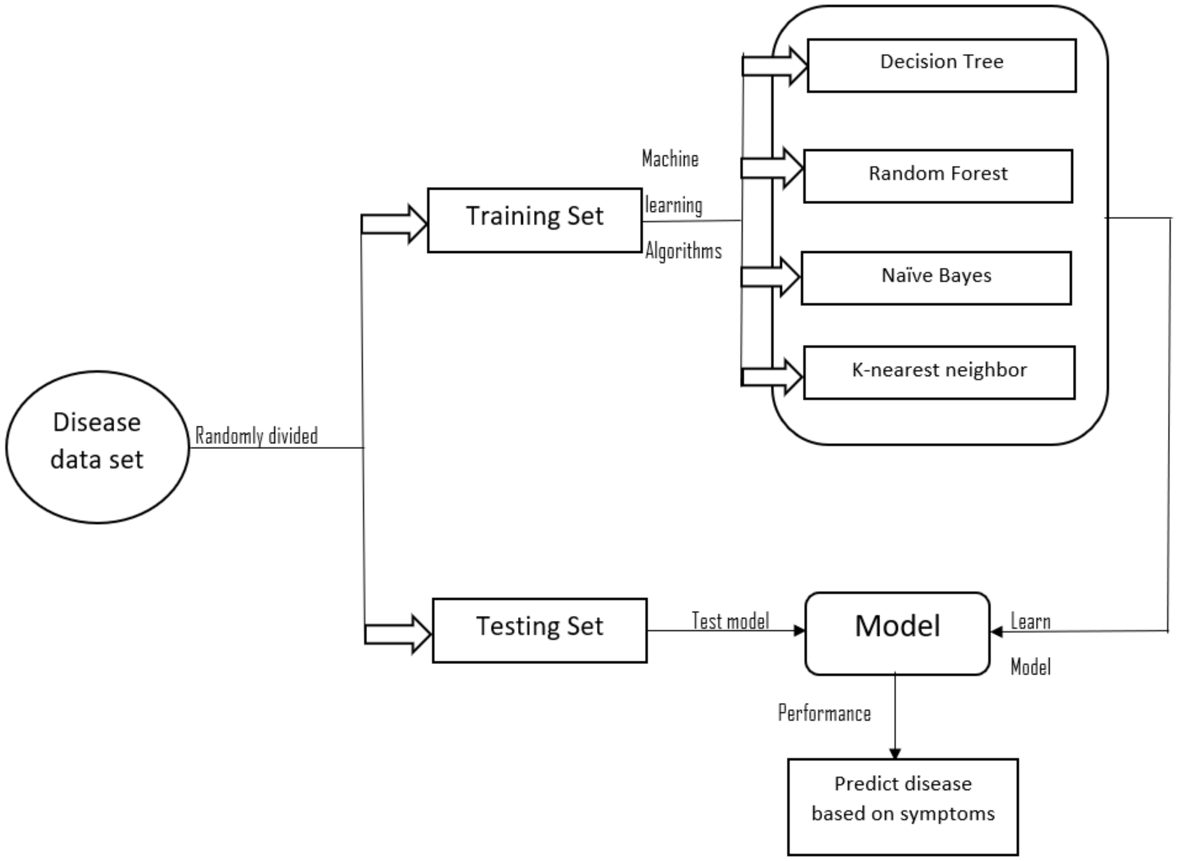


Fig -1: System Architecture

8. Results

• This section represents the proposed system results which can predict the disease faster, more accurate and with high reliability than the existing system. The results are obtained by implementing various Machine Learning algorithms. The Machine Learning classification techniques namely decision Tree, Random forest, Naïve Bayes and KNN are implemented using Python programming.

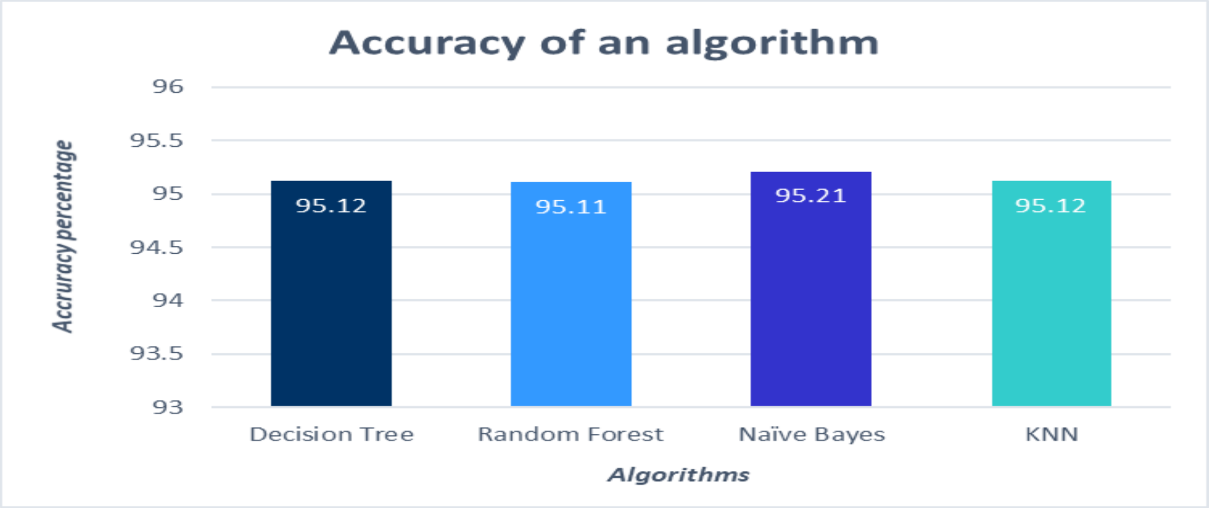
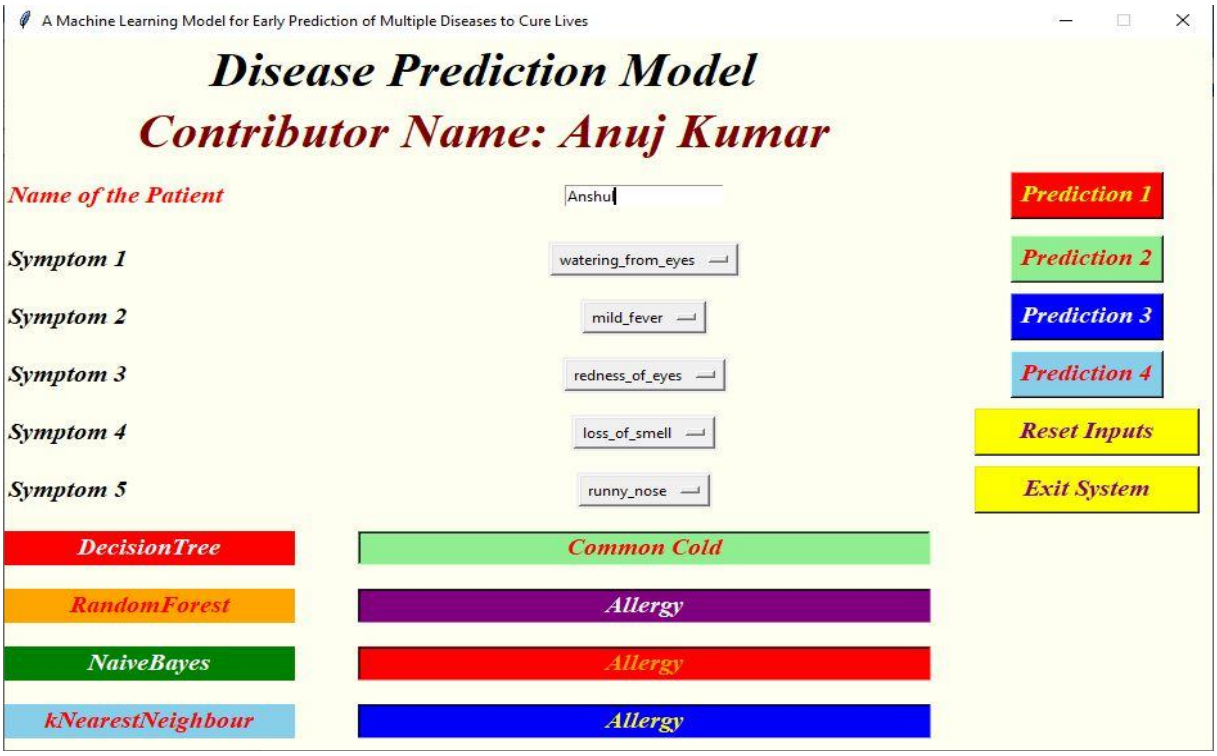


Fig -2: Analysis Of ML Algorithm

• This system also has an elegant interface which takes all the necessary inputs for the evaluation and to facilitate with the system which is very easy to use. The final result of our proposed system can be viewed from this GUI.


A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives

Disease Prediction Model

Contributor Name: Anuj Kumar

Name of the Patient: Anshu

Symptom 1: watering_from_eyes

Symptom 2: mild_fever

Symptom 3: redness_of_eyes

Symptom 4: loss_of_smell

Symptom 5: runny_nose

DecisionTree: Common Cold

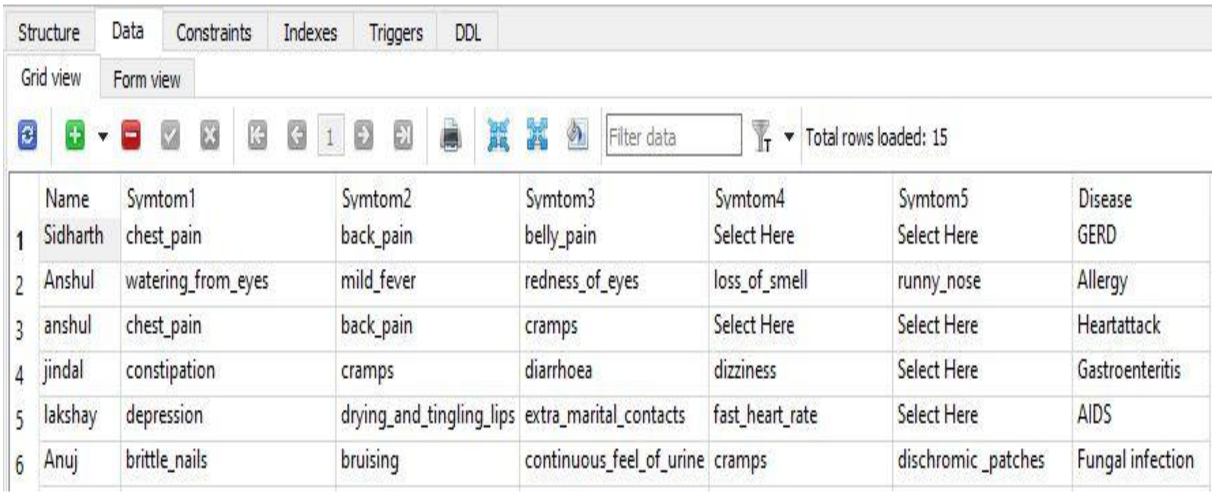
RandomForest: Allergy

NaiveBayes: Allergy

kNearestNeighbour: Allergy

Fig -3: The final / output page

• The proposed system also has a facility of Database for storing the data entered by the end users along with the name of the disease the patient is suffering from that may be used for future references thereby making this system more helpful and easier than the other system for this job.



Total rows loaded: 15

	Name	Symtom1	Symtom2	Symtom3	Symtom4	Symtom5	Disease
1	Sidharth	chest_pain	back_pain	belly_pain	Select Here	Select Here	GERD
2	Anshul	watering_from_eyes	mild_fever	redness_of_eyes	loss_of_smell	runny_nose	Allergy
3	anshul	chest_pain	back_pain	cramps	Select Here	Select Here	Heartattack
4	jindal	constipation	cramps	diarrhoea	dizziness	Select Here	Gastroenteritis
5	lakshay	depression	drying_and_tingling_lips	extra_marital_contacts	fast_heart_rate	Select Here	AIDS
6	Anuj	brittle_nails	bruising	continuous_feel_of_urine	cramps	dischromic _patches	Fungal infection

Fig -4: The Database created using Sqlite3

9. Conclusion

The main aim of this paper is to predict the disease in accordance with symptoms put down by the patients with proper implementation of Machine Learning algorithm. In this paper we have used four Machine Learning algorithm for prediction and achieved the mean accuracy of more than 95% which shows remarkable rectification and high accuracy than previous work and also makes this system more reliable than the existing one for this job and hence provides better satisfaction to the user in comparison with the other one. It also stores the data entered by the user and the name of the disease the patient is suffering from in the Database which can be used as past record and will help in future for future treatment and thus contributing in easier health management .We have also created a GUI for better interaction with the system by users which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease easily with different parameters and models. In the end we can say that our system has no threshold of the users because everyone can use this system.

10. Future Scope

There are many possible improvements that could be explore to diversify the research by discovering and considering extra features. Due to time boundation , the following work required to be performed in future. There is plan to use more classification techniques \ methods, different discretization techniques, multiple classifier voting methods. Would like to use different rules such as association rule and various algorithms like logistic regression and clustering algorithms. In future, willing to make use of filter based feature selection methods in order to achieve more appropriate as well as functional result.

References

[1] Khurana, Sarthak . , Jain, Atishay ., Kataria ,Shikhar. ,Bhasin ,Kunal . , Arora ,Sunny . ,& Gupta , Dr.Akhilesh . Das. (2019). Disease Prediction System.*International Research Journal Of Engineering and Technology* , *6*(5) , 5178-5184.

[2] Kamboj ,Mgha. (2020).Heart Disease Prediction with Machine Learning Approaches.*International Journal Of Science and Research* , *9*(7) , 1454-1458.

[3]Ware,Miss.Sangya . , Rakesh,Mrs.Shanu. K.,&Choudhary,Mr.Bharat . (2020). Heart Attack Prediction By Using Machine Learning Techniques. *International Journal Of Recent Technology and Engineering* , *8*(5), 1577-1580.

[4] Shirsath ,Shraddha.Subhash .,& Patil , Prof. Shubhangi . (2018).Disease Prediction Using Machine Learning over Big Data .*International Journal Of Innovative Research in Science and Technology , 7*(6), 6752-6757

[5]Marimuthu , M. , Abinaya, M. ,Hariesh,K.S., Madhan,K.,& Pavithra, Kumar. V.(2018).A Review of Heart Disease Prediction Using Machine Learning and Data Analytics Approach .*International Journal of Computer Application , 181*(18), 20-25.

[6] Battineni ,Gopi. , Sagaro,Getu.Gamo. ,Chinatalapudi, Nalini. ,&Amenta,Francesco. (2020). Application Of Machine Learning Predictive Models in the Chronic Disease .*International of PersonalisedMedicine , 10*(21), 1-11.

[7]Ardabili ,Sina. F.,Mosavi ,Amir.,Khamosi, Pedram. , Ferdinand ,Filip. ,Varkonyi-Koczy. Annamaria.R. Reuter, Uwe. ,Rabczuk ,Timon. , & Atkinson,Peter M. (2020). COVID -19 Outbreak Prediction with Machine Learning.*Journal of Algorithms,13*(249) , 1-36.

[8]Shrestha,Ranjit.,& Chatterjee,Jyotir. Moy. (2019).Heart Disease Prediction System Using Machine Learning . *LBEF Research Journal of Science Technology and Management , 1*(2), 115-132.

[9] Magar ,Rishabh. , Memane,Rohan.,Raut ,Sura. , &Rupnar,Prof. V.S. (2020) . Heart Disease Prediction Using Machine Learning*. Journal of Emerging Technologies and Innovative Research , 7*(6) , 2081-2085.

[10] Alotaibi, Fahd. Saleh. (2019). Implementation of Machine Learning Model to Predict Heart Failure Disease*. International Journal of Advanced Computer Science and Application , 10*(6) , 261-268.

[11] Godse, Rudra A.,Gunjal,Smita S., JagtapKaran A .,Mahamuni ,Neha S., &Wankhade, Prof. Suchita. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.*International Journal of Advance Research in Computer and Communication Engineering*, *8*(12), 50-52

[12] Anitha ,Dr.S.,& Sridevi,Dr.N. (2019). Heart Disease Prediction Using Data Mining Techniques.*Journal of analysis and Computation ,13*(2) , 48-55.

[13] Bindhika,Galla Siva Sai., Meghana,Munaga., ReddyManchuriSathvika. , &Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. *International Research Journal of Engineering and Technology*, *7*(4) , 5272-5276.

[14] Pingale,Kedar., Surwase, Sushant., Kulkarni,Vaibhav.,Sarage ,Saurabh., &Karve, Prof. Abhijeet .(2019). Disease Prediction using Machine Learning.*International Research Journal of Engineering and Technology, 6*(12) , 2810-2813.

[15]Chauhan Raj H., NaikDaksh N. ,Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction using Machine Learning.*International Research Journal of Engineering and Technology, 7*(5) , 2000-2002.