HUMAN DISEASE PREDICTION

A.Mahendar[1], P.Krishna tulasi[2], B.Harika[3], N.Sindhu[4]

[1]Associate Professor, CSE, CMR Technical Campus, Hyderabad, India.

[2, 3, 4] Student, CSE, CMR Technical Campus, Hyderabad, India.

***ABSTRACT*— Disease prediction using Machine Learning is a system that predicts the disease based on the information or the symptoms provided by the user based on that accurate result shown. Currently, the health industry plays a major role in curing the illness of the patients, just in case he/she doesn't want to travel to the hospital, therefore just by giving the symptoms and all other validated information the user can get to know the illness they are suffering from. Machine Learning algorithms such as Naive Bayes, Decision Tree, KNN and Random forest are used on the provided dataset and predict the disease of the patient. The implementation is done through the Python Programming Language.**

***Keywords*— Machine Learning, Naive Bayes algorithm, KNN algorithm, Decision Tree algorithm, Random forest algorithm, Python.**

I. INTRODUCTION

Machine learning[1] is a type of artificial intelligence application that enables self-learning from data and then applies the learning without the need for human intervention and improves from experience without being explicitly programmed. In actuality, there are many different types of machine learning, as well as many strategies and algorithms so how to best employ them. It is totally different from traditional programming, here data and the output is given to the computer and in return, it gives us the program which provides a solution to various problems. Machine learning is the combination of algorithms, datasets and programs. Now-a day's. Machine learning is playing a major role in health industries, Stock market ,IT industry etc.,

Machine Learning [2] is complex in itself that is why it has been divided into two main areas, supervised learning [3] and unsupervised learning. Each one has a reason and action with Machine Learning, to yield particular results and utilizing various forms of data. Approximately 70 percent of Machine Learning is supervised learning such as Decision tree, Random forest, KNN etc. while unsupervised learning ranges from 10-20 percent such as K-means clustering, Apriority etc. Other method that is used less often is reinforcement learning.

II. LITERATURE REVIEW

Health care center's purpose is to improve the current health of the community. A health care institution such as hospitals or medical centers would essentially consist of many doctors who were qualified and specialized in treating various patients of their current disease that they are suffering from and trying to cure them and give proper health.

Currently, new technologies have developed and came into existence to improve people's daily life and routine especially in health care centers. The project disease prediction using machine learning is developed to overcome general disease or illness in early stages as we all know in the competitive environment of economic development the people has involved so much that he/she is not worried about health according to research there are 40% people who ignore the general disease or illness which leads to harmful disease later. The main reason for ignorance is laziness to consult a doctor or hospital and time concerns the people have involved themselves so much that they have no time to make an appointment and consult the doctor which later results in fatal dangerous disease. According to research there are 70% of people in India suffering from general disease and 25% of people are facing death due to early ignorance the main motive to develop this project Is that a user can sit at their place and have a check-up of their health the UI designed in such a simple way that everyone can easily operate it and can have a check-up.

III. ARCHITECTURE

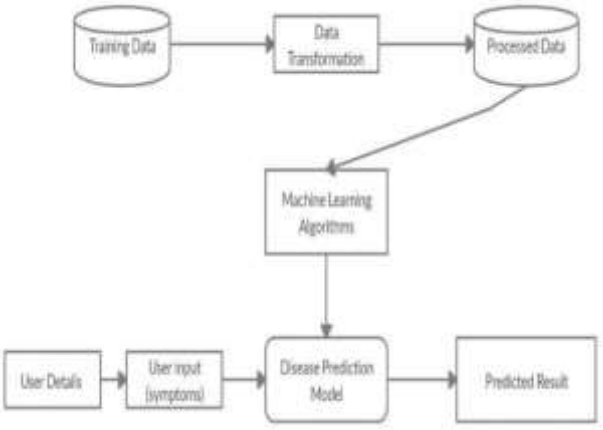


Fig 1: Architecture of Human Disease Prediction

From the above architectural diagram, first the training data from the dataset is transformed by removing the raw and the noisy data and then the data is processed further. Now the machine learning algorithms will be applied or predict the disease. When the user enters the basic details and the symptoms from which the patient is suffering. After entering the symptoms the Disease prediction model will predict the accurate disease as the output.

IV. EXISTING SYSTEM

Prediction using traditional methods [4] and models involves various risk factors and it contains various measures of algorithms such as datasets, programs and much more. In this system, there are many models which may not provide accurate results and use only one or two algorithms for predicting the resultant output. High-risk and low-risk patient classification is done based on the tests that are done in a group. But these models are only valuable in clinical situations and not in the big industry sector. So, to include the disease predictions in various health-related industries, we have used the concepts of machine learning and supervised learning methods to build the prediction system. This system can predict the disease but not the sub-type of the disease and it fails to predict the condition of the people, the predictions of the disease have been indefinite and non-specific. And the most difficult thing in the current system is it makes the patient undergo lengthy procedures and questionnaires which is time-consuming process.

Disadvantages:

- Only a few diseases can be detected.
- Less accuracy.
- Weak generalization.
- Time-consuming.

V. PROPOSED SYSTEM

The proposed system [5] of disease prediction using machine learning is that we have used many techniques and algorithms and all other various tools to build a system that predicts the disease of the patient using the symptoms and by taking those symptoms we are comparing with the system's dataset that is previously available. By taking those datasets and taking those symptoms we are comparing with the system's previously available dataset patient's disease.we will predict the accurate percentage of disease of the patient. In proposed system of disease prediction using machine learning is that we have used many different algorithms such as Decision tree, KNN, Naive Bayes, and Random Forest[6] which predicts the disease of the patient by entering the basic details of the patient like name and the symptoms from which the patient is suffering from. Now, by taking those symptoms we are comparing with the system dataset that is previously available which will predict accurate disease. Our main motive with this system is to be connecting the bridge between doctors and patients.

Advantages:

- Data can be stored and used for doctor reference.
- Analysis based on multiple algorithms.
- More accurate.

VI. ALGORITHMS

1. DECISION TREE ALGORITHM:

Decision Tree [8] algorithm belongs to supervised learning algorithms. Instead of other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

The reason for using a Decision Tree is to create a training model that can use to predict the class or value of the variable by learning decision rules which are used and inferred from the training data.

In Decision Trees, for predicting a class label we start from the root point of the tree. Next, we will compare the values of the root point with the record's attribute values. At this point of Comparison, we follow the branch value corresponding to that value and shift to the next node.

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition, *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_attribute* and, possibly, either a *split-point* or *splitting subset.*

Output: A decision tree.

Method:

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C, then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let Dj be the set of data tuples in D satisfying outcome j; // a partition
(12)     if Dj is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(Dj, attribute_list) to node N;
     endfor
(15) return N;
```

Fig 2: Decision tree Algorithm

EXAMPLE:

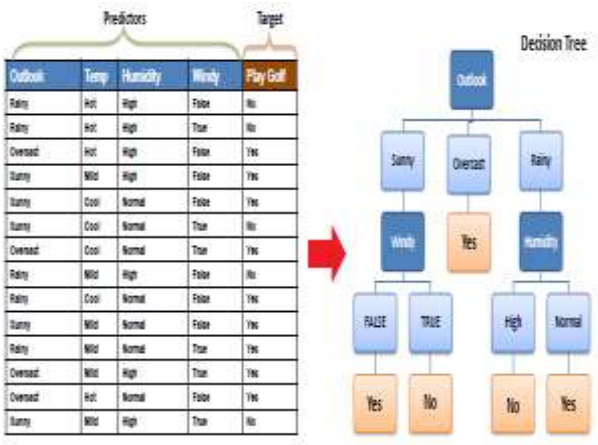


Fig 3: Decision tree Example

2. RANDOM FOREST ALGORITHM:

Random forest [9] is an algorithm. The "forest" it builds, is an ensemble of decision trees, typically trained with the "bagging" technique. The main idea of the bagging technique is that a mixture of learning models wills increases the final result.

In simple words, random forest builds multiple decision trees and combines them to get a more accurate and stable prediction and prevents the drawback of overfitting.

One massive advantage of random forest is that it may be used for both classification and regression issues, which form the bulk of current machine learning systems.

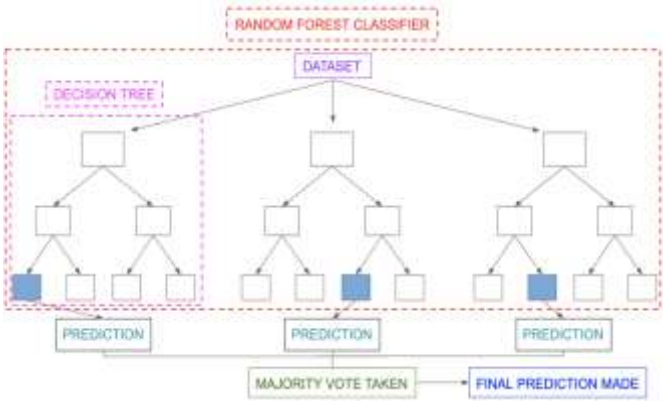


Fig 3: Working of Random Forest

The following are the steps for the Random Forest algorithm:

1: procedure RANDOM FOREST ALGORITHM
2: for $i = 1$ to T trees do
3: Pick n data points ($D_{i=1\_n}$) with replacement from training dataset (D)
4: Build full decision tree on $D_{i=1\_n}$
5: for each split, consider only k features that are picked uniformly at random new features for every split
6: Prune tree to minimize out-of-bag erro
7: end for
8: Average all T trees
9: end procedure

Fig 4: Random Forest Algorithm

3. K-NEAREST NEIGHBOR ALGORITHM:

K-Nearest Neighbour [10] is one of the easiest Machine Learning algorithms based on the Supervised Learning technique. KNN algorithm detects the similarity between the new data and available data cases and puts the new case data into the file that is most similar to the available one.

KNN algorithm stores all the data which is available and classifies a new case data point based on the similarity. This means when new data appears then it can be easily classified into a well-suited category by using the KNN algorithm technique. It can be used for Regression issues and also for Classification but mostly it is used for Classification issues.KNN is a non-parametric algorithm technique, which means it can not make any assumption or idea on underlying data.

KNN is also called a **lazy** learner algorithm because it does not learn the issues generated from the training set immediately instead it stores the issues in the dataset and at the time of classification, it performs on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new case data, and then it classifies that case data into a category based that is much familiar to the new data. The following are the steps for KNN algorithm:

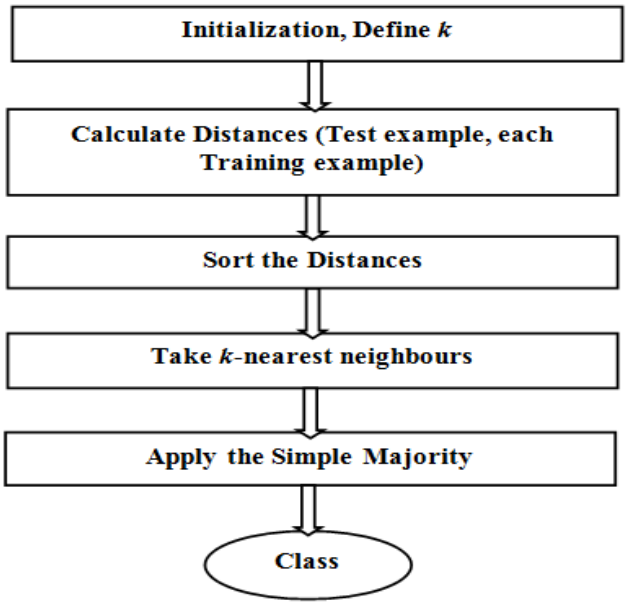


Fig 5: KNN Algorithm

EXAMPLE:

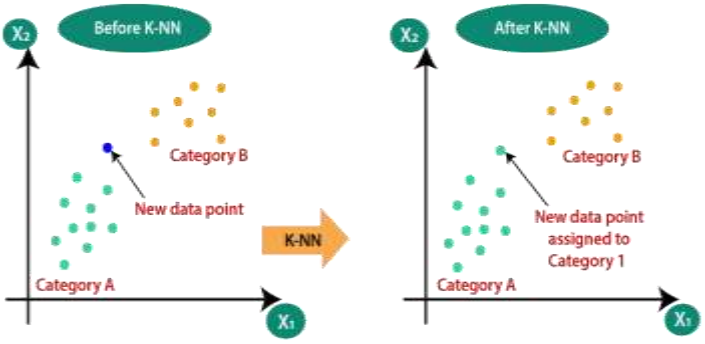


Fig 6: KNN Example

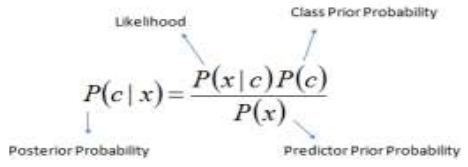
4. NAIVE BAYES ALGORITHM:

It is a supervised learning technique based on Bayes' Theorem [7]. In simple words, a Naive Bayes classifier thinks that the presence of the main feature in a class is not related to any of the other class features. For example, a fruit may be said to be an apple if it is red in color, round in shape, and about 4 inches in diameter.

Even if these common features depend upon each other, all of these property methods separately say that this fruit is an apple and that is why it is called 'Naive'. The Navie Bayes model is very easy to build and mainly useful for very large data sets problems/issues. Along with simplicity, Naive Bayes is called to perform even highly sophisticated classification techniques.

The following are the steps for the Naive Bayes algorithm:

Bayes theorem gives a way for calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). See the equation given below:



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is a posterior probability of the *class* (c, *target*) and given *predictor* (x, *attributes*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of the *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

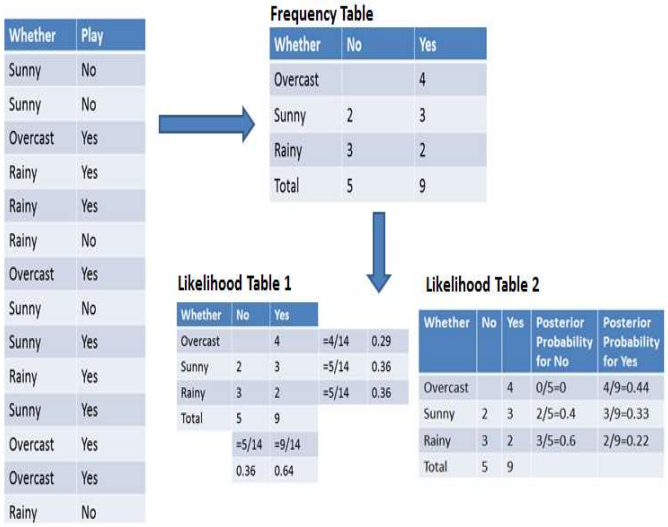
Algorithm Steps:

Step 1: Convert the data set into a frequency table model.

Step 2: Create Likelihood tables by finding the probabilities of each dataset.

Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each of the class separately. The data classes with the high posterior probability is the result of the prediction method.

EXAMPLE:



Whether	Play
Sunny	No
Sunny	No
Overcast	Yes
Rainy	Yes
Rainy	Yes
Rainy	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rainy	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rainy	No

Whether	No	Yes
Overcast		4
Sunny	2	3
Rainy	3	2
Total	5	9

Whether	No	Yes		
Overcast		4	=4/14	0.29
Sunny	2	3	=5/14	0.36
Rainy	3	2	=5/14	0.36
Total	5	9		
	=5/14	=9/14		
	0.36	0.64		

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

Fig 7: Naive Bayes Example

VII. RESULT

- First, the user needs to register into the system with the username and the password. If the user is already registered then he/she need to log in into the system with their credentials.
- Now the user needs to select minimum of two symptoms as compulsory from the given dropdown menu, for getting more accurate results.
- After entering all the symptoms which they are suffering from, the user needs to press the buttons of the respective algorithm to predict the resultant disease the disease will be displayed that they are suffering from and here we are going to use four different algorithms to provide a clearer picture and accurate result to the user.

Based on our statistics,the Decision tree algorithm is giving more accurate results than the other algorithms.

This is the resultant output of the particular patient which will be stored in our database system.

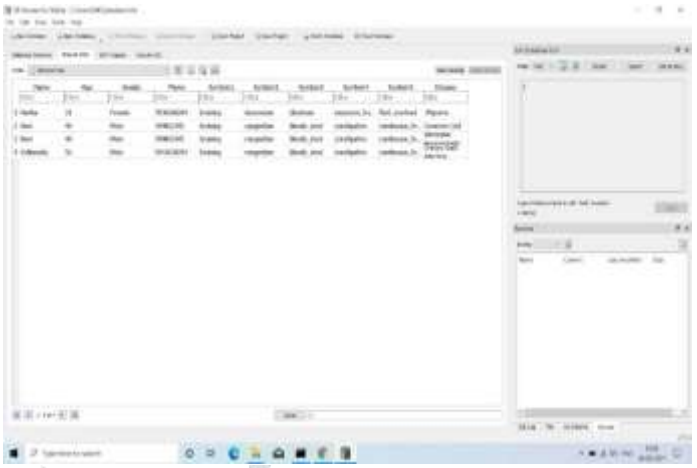


Fig 8: Resultant output of database stored by Human disease Prediction project

VIII. CONCLUSION

Nowadays machine learning is playing important role in the medical prediction field of the health sector. Health is a very important aspect of everyone's life.

So, after doing the research and comparison of all the algorithms and theorems of machine learning we have come to the conclusion that all those algorithms such as Decision Tree, KNN, Naive ayes, and Random Forest Algorithms all are used in building a disease prediction system which predicts the disease of the patients from which he/she is suffering from. But, according to accuracy percentage Decision tree is predicting the most accurate result.

IX. REFERENCES

[1] https://link.springer.com/chapter/10.1007/978-981-15-50898_55#:~:text=Machine%20Learning%20is%20an%20emerging,dataset%20ana%20predict%20the%20disease.

[2] https://www.expert.ai/blog/machine-learning-definition/

[3] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8

[4] https://www.youtube.com/watch?v=3YmAbta16yk&t=1s

[5] https://www.youtube.com/watch?v=NUw2cPNwDZs

[6] https://www.youtube.com/watch?v=I7NrVwm3apg

[7] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[8] https://www.geeksforgeeks.org/decision-tree-introduction-example/#:~:text=Decision%20tree%20algorithm%20falls%20under%20the%20category%20of%20supervised%20learning.&text=Decision%20tree%20uses%20the%20tree,internal%20node%20of%20the%20tree.

[9] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm#:~:text=Random%20forest%20is%20a%20supervised,classification%20as%20well%20as%20regression.&text=Similarly%2C%20random%20forest%20algorithm%20creates,solution%20by%20means%20of%20voting.

[10] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm#:~:text=With%20the%20help%20of%20KNN,Image%20Recognition%20and%20Video%20Recognition.