

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

Anurag Ravi Nimonkar,

*Department Of Computer Science And Technology,
SRM Institute Of Science And Technology, Kattankulathur,
Chennai, India an5450@srmist.edu.in*

Dr. Priyadarsini K.,

*Department Of Data Science And Business Systems, School of Computing,
SRM Institute Of Science And Technology, Kattankulathur,
Chennai, India priyadak@srmist.edu.in*

Abstract-

Machine Learning Approach for Identifying Disease Prediction Using Machine Learning is based totally on prediction modelling that predicts disorder of the sufferers according to the symptoms provided by using the customers as an i/p to the system. This paper offers an idea of predicting more than one disease using Machine Learning algorithms. Here we will use the notion of supervised Machine Learning in which implementation will be executed by way of making use of Decision Tree, Random Forest, Naïve Bayes and KNN algorithms which will assist in early prediction of illnesses accurately and better sufferers care. The consequences ensure that the machine would be practical and user oriented for sufferers for timely diagnosis of diseases in a patient.

Index Terms - Machine Learning , Disease Prediction , K-Nearest Neighbour, Random Forest

1.INTRODUCTION.

The Earth is going through a purplish patch of science where the demand of intelligence and accuracy is increasing at the back of it [11]. Today's people are probably addicted to the internet however they are not worried about their physical health. People pass by the small trouble and don't go to the health centre which turns into serious disease with time [11]. Taking the gain of this developing technology, our basic intention is to develop such a machine that will predict the multiple illnesses in accordance with signs put down via the patients except journeying the hospitals / physicians. Machine Learning is a subset of AI that frequently deals with the discovery of algorithms which enhance with the use of information and experience. Machine Learning has two phases i.e. Training and Testing [17]. Machine Learning provides an efficient platform in scientific subjects to clear up a number of healthcare troubles at an awful lot quicker rate. There are two kinds of Machine Learning – Supervised Learning and Unsupervised Learning. In supervised getting to know we frame a model with the assist of data that is properly labelled. On the other hand, unsupervised mastering models learn from unlabeled data. The intent is to deduce a fine Machine Learning algorithm which is efficient and accurate for the prediction of disease. In this paper, the supervised Machine Learning idea is used for predicting diseases. The important function will be Machine Learning in which we will be engaging in the usage of algorithms such as Decision Tree, Random Forest, Naïve Bayes and KNN which will help in early prediction of illnesses accurately and better patient care [11].

2.MODEL OBJECTIVE AND DRAWBACKS OF EXISTING SYSTEM

There is a demand to make such a system that will assist end users to predict ailments on the basis of symptoms given in it without travelling to hospitals. By doing so, it will minimise the rush at OPD's of hospitals and carry down the workload on medical staff. Not solely this, this device will reduce the high-priced treatment and panic moment at the quit tiers so that ideal medicinal drug can be furnished at the proper time and we can decrease down the demise rate as well. This gadget also consists of a feature of Database which stores the data entered by the quit users and the name of the sickness the patient is struggling from that can be used as a previous report and will help in further treatment in future. The analysis accuracy is improved via the usage of Machine Learning algorithms. Altogether this the system will help in easier wellness management.

3.PROPOSED SYSTEM

We are proposing such a machine that will flaunt a simple, fee wonderful, elegant User Interface and also be time efficient. Our proposed gadget bridges the hole between medical doctors and sufferers which will assist each training of users to achieve their goal. This machine is used to predict ailments according to symptoms. In this proposed system we are going to take down five symptoms from the users and consider them through applying algorithms such as Decision

Tree, Random Forest, Naïve bayes and KNN which will assist in getting accurate prediction. Our system will discover and merge more datasets which consists of large range of populace to get greater effective results and for that reason our system will improve and enhance the accuracy of the results. Along with the increased accuracy rate, we will proliferate the reliability of our system for this job and can achieve the trust of affected person in this system. Apart from all these, our model will incorporate a Database for storing the statistics entered by way of the users and the name of the sickness the affected person is suffering from which can be used as a reference in future for further treatment. Hence this system will make contributions in less complicated health administration with better satisfaction to the users.

ADVANTAGES

It boost accuracy score by comparison well-liked machine learning algorithms. These reports square measure to the examination of congruity of AI procedures for sleuthing disease in operational circumstances by quality forecast.

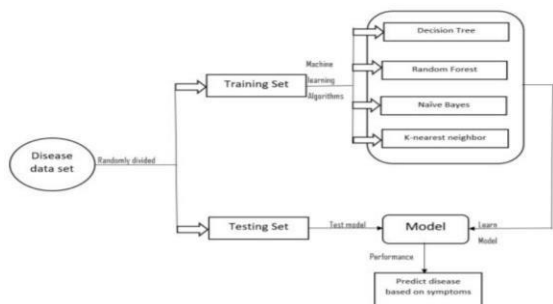


Figure No.1:- System Architecture

4.METHODOLOGY

DATA VALIDATION PROCESS(EDA)

Data validation process:-

Endorsement strategies in ML are utilized to get the slip-up speed of the ML model, which is treated as near the bonafide goof speed of the dataset. On the off chance that the information volume is acceptably colossal to be illustrative of everybody, you may not need the support procedures. Not withstanding, in bona fide conditions, to work with preliminary information that may not be a guaranteed expert of the number of tenants in a given dataset. To tracking down the missing worth, copy worth, and portrayal of information type whether it is coast variable or whole number. The instance of information used to give a fair assessment of a model fit on the plan dataset while tuning model hyper cutoff points. The evaluation winds up being more lopsided as expertise on the underwriting dataset is entwined into the model arrangement. The underwriting set is utilized to overview a given model, at any rate, this is for standard evaluation. It as ML engineers utilizes this information to change the model hyper cutoff points. Information gathering, information assessment, and the way toward looking out for information substance, quality, and design can amount to a drawn-out game plan for the day. The example of information prominent confirmation, it assists with understanding your information and its properties; this information will assist you with picking which assessment to use to gather your model. For instance, time approach information can be bankrupt somewhere near lose the faith calculations; depiction assessments can be utilized to investigate discrete information. (For example

to depict the information type game plan of the given dataset)

```
In [1]: df = pd.read_csv('training.csv')
df.head()

Out[1]:
```

	hunger	skin_rash	head_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	itchers_on_tongue	blackheads	scurfs
0	1	1	1	0	0	0	0	0	0	0	0	0
1	0	1	1	0	0	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0	0	0	0
3	1	1	0	0	0	0	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0	0	0	0	0

```
5 rows * 13 columns

In [3]: X = df.iloc[:, 1:12]
y = df['prognosis']
```

Figure No.2:- After Pre-Processing Data

Data Cleaning/Preparing Process:-

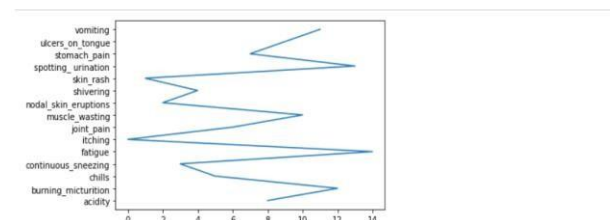
Bringing in the library groups with stacking given dataset. To examining the variable unmistakable verification by data shape, data type and surveying the missing characteristics, duplicate regards. A validation dataset is a model of data held back from setting up your model that is used to give a check of model ability while tuning model's and frameworks that you can use to make the best use of validation and test datasets while surveying your models. Data cleaning /arranging by rename the given dataset and drop the section n, etc to separate the uni-variate, bi-variate and multi-variate measure. The methods and methodologies for data cleaning will move from dataset to dataset. The fundamental target of data cleaning is to perceive and kill bumbles and peculiarities to fabricate the value of data in examination and dynamic.

Data Pre-processing:-

Importing the library packages with loading given dataset. To analyzing the variable identification by data shape, data type and evaluating the missing values, duplicate values. A validation dataset is a sample of data held back from training your model that is used to give an estimate of model skill while tuning model's and procedures that you can use to make the best use of validation and test datasets when evaluating your models. Data cleaning / preparing by rename the given dataset and drop the column etc. to analyze the univariate, bi-variate and multi-variate process. The steps and techniques for data cleaning will vary from dataset to dataset. The primary goal of data cleaning is to detect and remove errors and anomalies to increase the value of data in analytics and decision making.

5.DATA VISUALIZATION

Exploration data analysis of visualization:-Information portrayal is a basic expertise in applied assessments and AI. Encounters does to be sure zero in on quantitative portrayals and assessments of information. Information depiction gives up a basic course of action of instruments for getting a theoretical arrangement. This can be significant while inspecting and getting comfortable with a dataset and can assist with perceiving plans, degenerate, anomalies, and in a general sense more. With a little district information, information depictions can be utilized to examine and show key relationship in plots and diagrams that are more instinctive and assistants than degrees of association or importance. Data insight and exploratory information evaluation are entire fields themselves and it will suggest a more huge jump into some the books alluded to toward the end. Once in a while information doesn't look great until it can take a gander at in a visual development, for example, with diagrams and plots. Having the decision to rapidly imagine of information tests and others is a colossal limit both in applied assessments and in applied AI. It will find the different sorts of plots that you should recognize while envisioning information in Python and how to utilize them to even more plausible understand your own information.



TO TRAIN A MODEL BY GIVEN DATASET USING SKLEARN PACKAGE COMPARISON OF ALGORITHM MS ACCURACY RESULTS OF ALGORITHMS

Random Forest :-

Random Forest comes under the category of supervised Machine Learning algorithm. It is used for classification and regression but mainly deals 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 trees at training time [1]. It selects the best solution by means of voting [2]. Random Forest is composed of multiple decision trees [2]. It creates a forest of trees [2]. The number of trees 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 being 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 uses bootstrap aggregation or bagging[13].

- Loading the dataset.
- Summarizing the dataset.
- Visualizing the dataset.
- Evaluating some algorithms.

Figure No.4:- Random Forest Classification and Report

K-Nearest Neighbor (KNN):- K-Nearest Neighbor is a planned AI calculation which saves all models relate to preparing server farms in n-dimensional space. Precisely when a dim discrete information is gotten, it explores the nearest k number of occasions saved (closest neighbors) and returns the most by and large saw class as the figure and for genuine respected information it returns the mean of k closest neighbors. Somewhere far off weighted closest neighbor assessment, it stacks the duty of the entirety of the k neighbors as per their distance utilizing the going with question giving more observable weight to the nearest neighb ors.

K-Normally KNN is liberal to riotous information since it is aver aging the k-Nearest neighbors. The k-Nearest neighbors assessment is a depiction calculation, and it is guided: it takes a huge load of stepped focuses and utilizes them to figure out some approach to name assorted center interests. To stamp another point, it takes a gander at the named thinks nearest to that new point (those are its closest neighbors), and has those neighbors vote, so whichever name the an immense fragment of the neighbors have is the name for the ne w point (the "k" is the measure of neighbors it checks). Make s measures about the underwriting set utilizing the whole arranging set. KNN makes a presumption with respect to another occasion through looking through the whole set to track down the k "nearest" cases. "Closeness" is settled utilizing a region appraisal (Euclidean) across all highlights.

```
In [8]: from sklearn.neighbors import KNeighborsClassifier
        classifier = KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2)
        classifier.fit(X_train, y_train)

Out[8]: KNeighborsClassifier()

In [9]: y_pred = classifier.predict(X_test)

In [10]: from sklearn.metrics import accuracy_score
         accuracy_score(y_test, y_pred)
         print("Accuracy of SVM is {}".format(a))

Accuracy of KNN: 92.24 %
```

Figure No.5:- KNN classification and Report

- Making some predictions.

7. RESULT

The proposed system can predict the disease faster, more accurately and with higher reliability than the existing systems in the market today. The results are obtained by implementing various Machine Learning algorithms. The Machine Learning classification techniques namely Decision Tree, Random forest and KNN are implemented using Python programming.

```
In [16]: input_vector = np.zeros(len(symptoms_dict))

In [17]: input_vector[[symptoms_dict['itching'], symptoms_dict['skin_rash'], symptoms_dict['nodal_skin_eruptions']] = 1

In [18]: rf_clf.predict_proba(input_vector)

Out[18]: array([[0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.]])

In [19]: rf_clf.predict([input_vector])

Out[19]: array(['Fungal infection'], dtype=object)
```

Figure No.6:- Output

8.CONCLUSION

The major goal of the model is to predict the ailment in accordance with symptoms put down through the patients with proper implementation of Machine Learning algorithm. In this paper we have used four Machine Learning algorithm for prediction and performed the imply accuracy of more than 95% which indicates remarkable rectification and excessive accuracy than preceding work and additionally makes this system greater dependable than the existing one for this job and subsequently offers better pleasure to the consumer in evaluation with the different one. It also stores the facts entered by using the consumer and the name of the sickness the patient is struggling from in the Database which can be used as past report and will assist in future for future treatment and as a consequence contributing in less complicated health management .We have also created a GUI for higher interaction with the system by means of customers which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease without problems with different parameters and models. In the give up we can say that our device has no threshold of the users because anyone can use this system.

9.FUTURE ENHANCEMENT

There are many viable enhancements that ought to be discover to diversify the research through discovering and considering more features. Due to time boundation , the following work required to be performed in future. There is graph to use extra classification strategies \ methods, distinct discretization techniques, multiple classifier voting methods. Would like to use extraordinary rules such as affiliation rule and quite a number algorithms like logistic regression and clustering algorithms. In future, inclined to make use of filter based totally function selection methods in order to achieve extra gorgeous as nicely as functional result. Also, the sequence to the model will also be incorporating a user friendly GUI interface for client interactivity and overall ease of use for the end user.

10. REFERENCE

- [1]Khurana, Sarthak . , Jain, Atishay ., Kataria ,Shikhar .,Bhasin ,Kunal . , Arora ,Sunny . .& Gupta , Dr.Akhilesh . Da s. (2019). Disease Prediction System.International Research Journal Of Engineering and Technology , 6(5) , 51 78-5184.
- [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. ,Harish,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 Personalised Medicine , 10(21), 1- 11.

[7] Ardabili ,Sina. F.,Mosavi ,Amir.,Khamosi, Pedram. , Ferdin and ,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.