IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 8, August 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

Human Disease Predictor Using Machine Learning with Claims Data

Yeshaswini N [1], Thejaswini N [2], Soundarya N [3], Saraswathi G [4], Dr. S Usha5

U.G. Student, Department of Computer Engineering, Rajarajeswari College of Engineering, Bangalore, Karnataka,India[1,2,3,4]

Professor & HOD, Department of Computer Engineering, Dean (Research), Rajarajeswari College of Engineering, Bangalore, Karnataka, India[5]

ABSTRACT: "Disease Prediction" system based on prognostic modelling predicts the disease of the client based on the symptoms that user gives as an input to the predictive system. The system analyses the symptoms provided by the user as input and gives the probability of the disease as an output Disease Prediction is done by implementing the information mining calculations like Decision Tree, Random Forest and Naïve Bayes calculations can give a solution for the present circumstance. Thus, we have fostered a mechanized framework that can find and concentrate covered up information related with the infections from a verifiable (illnesses manifestations) data set by the standard arrangement of the particular calculations. This paper presents a thorough relative investigation of three calculations execution on a clinical record each yielding an exactness up to 95 percent. The presentation is dissected through disarray grid and precision score.

KEYWORDS: Dataset, Machine learning-Classification method, Python IDE, Decision Tree, Random Forest, Naive-Bayes Classifier, Support Vector Machines, Human Symptoms, Accuracy Score.

I. INTRODUCTION

At this point of time, when patient suffering from any disease, then it is necessary for a people to visit a doctor which very risky and have to spend money more. In adding, if the user is not reachable to doctor and hospitals are far, it may be difficult for the client as the diseasecannot be acknowledged. Hence, if the above procedure could be accomplished using an computerized program which can save lot of timeas well as expenditure, it is easier to the patient or the client, which makes the procedure much more easier. There are many more Heart related Disease Calculation System using data mining procedures that analyzes the risk level of the client or the patient. Disease Predictor is a web based submission that forecasts the disease of the client with respect to the symptoms which is provided by the client. Disease Estimate system which collects the diseases from many health disease related website. Hence , we can say that with the help of the disease estimator we can analyse the diseases based on the input given by the client or the patient.

Artificial Intelligence completed computer made more intelligent and can allow the computer to analyze and decide. AI by studying determines the machine learning as the sub group of research work. Different analysts makes us to realize without learning we can't decide it. There are numerous kinds of Machine Learning Procedures like Unsupervised, Semi Supervised, Reinforcement, Supervised Evolutionary Learning and Deep Learning. These learnings classify the gaint data efferently and quickly. The medical services and clinical area are more needing datamining today. At the point when certain information mining techniques are utilized in a correct manner, significant data can be extricated from enormous data set and that can assist the clinical expert with taking early choice and further develop wellbeing administrations. The soul is to utilize the characterization to help the doctor.

Much more tools are available which relates to disease forecast. But particularly every diseases have been analyzed and generates the risk level. But generally there no tools that are used for forecast of general diseases. Therefore, Disease Forecaster helps for the prediction of the diseases in general. It is dangerous to know the accurate analysis of clients by Medical examination and assessment. For convincing willpower, decision support systems that depends on computer that may accept an indispensable task. Health care ground creates huge information about Medical estimation, report in esteems to client, cure, succeeding meet-ups, remedy and so forth. It is complex to compose appropriately. Quality of

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

the data suggestion has been prejudiced due to improper organization of the data. Elevation in the amount of data needs some sincere way to essence and process data viably and professionally.

The disadvantages of the existing system is as follows :

- Although many other machine learning algorithms only effort on structured data and time required for calculation is more high also they are idle because they store whole data as a training dataset and customs complex method for calculation
- Accuracy is very low to make enough good decision on prediction.

We future general disease forecast system based on machine learning algorithm. It is based on prognostic modelling forecasts the disease of the client on the base of the Symptoms that client delivers as an input to the system. The system examines the symptoms providing by the client as input and gives the likelihood of the disease as an output Disease Forecast is done by realizing the information mining calculations like Decision Tree, Random Forest and Naïve Bayes, Svm vector algorithm ,calculations can give a solution for the present circumstance. Henceforth, we have fostered a computerized framework that can find and concentrate covered up information related with the illnesses from a chronicled (infections side effects) data set by the standard arrangement of the separate calculations.

This paper presents an extensive similar investigation of three calculations execution on a clinical record yielding an 95 % accuracy.

This paper has the solution where the advantages over the present technologies are at a good position. Some of themare as follows:

- Predominantly every diseases have been examined and risk level is Pretested. But generally there are no tools that are cast-off for forecast of overall diseases.
- Hence Disease Forecaster helps for the forecast of the overall diseases.
- Due to comparing the algorithms leads to high accuracy.

II. RELATED WORK

Human disease Forecaster using Machine learning which claims data is very much helpful in forecasting the disease based upon the client symptoms. Here there are four algorithms i.e., Naive Bayes, Random Forest and Decision Tree ,Support Vector Machine to forecast the correctness of the disease present in the client based on the symptoms, which is provided as an effort by the client.

In this paper, we have proposed 4 layers of modules on working of the solution i.e. Data Collection & Evaluation, Data processing & training, Data Analysis and Testing of Data.

Data Collection and Evaluation

Data collection has been done from the cyberspace to recognize the disease and the real symptoms of the disease are composed that is current. No imitation values are arrived. The symptoms of the disease are composed from diverse many health related websites (Kaggle). we have fostered a robotized framework that can find and concentrate covered up information related with the sicknesses from an illnesses side effects data set by the standard arrangement of the individual calculations.

Data Processing and Training

The information mining strategy that changes the crude information or encodes the information to a structure, which can be effectively deciphered by the calculation, is called data mining or pre-processing. Data mining process cast-off to convert the data hooked on appropriate forms. It is frequently done in instruction to adapt the data into compulsory format besides on the additional courses of analyzing.

The role of this module is to maintain good information consistency, which means classifying and rectifying lost values, identifying and reducing outliers. With respect to switch this condition, there are two ways:

Evade tuples: This system is appropriate when the datasets are rather large and many morals are missing in tuple. For example, in this examination, features like fever and precipitation were removed in feature extraction.

Missing values: This way can be occupied manually by quality mean or most credible worth.

Data Analysis

This a technique for the reviewing data sets, often using visual tools, to outline their key features. This is step

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

where data sets are explored and understood.
It is necessary step before the data modelling or machine learning. By doing this analysis, it is easier to find out if the selected features are good enough to model, if all the features are needed, whether there are any correlations depending on which we can rather go back to the phase of Data Pre-processing or continue with modelling.

The principle goals of the exploratory period of information investigation is to comprehend the fundamental attributes of the information by utilizing spellbinding insights, connection examination, visual assessment and other basic displaying.
The following steps are carried in this analysis process:

- Overview of data
- Dealing with missing data.
- Dealing outliers.

Overview of Data:
Before we can go on to the other measures, we need to learn the various kinds of data and other information regarding our data. example: describe ()
This helps to generate descriptive statistics which describe dispersion, shape of distribution of datasets and central tendency.

Dealing with the missing data:
In real world, data are rarely homogenous and clean. It can be missing either in data extraction or in data collection. It can be carefully handled because quality is reduced of any performance, which would lead to wrong classification or prediction. There are many options for dealing with missing values:

- Drop-missing values or NULL: in this method, all the NULL values or missing values are dropped. This methodreduces the quality of data model because the data sample is reduced.
- Fill missing values: in this method, all the missing values are replaced with statics value like median, mean of that feature or attribute where the value belongs.
- Predict values using algorithms: in this method, we use algorithms of regression or classification algorithms to deal with missing values.

Dealing with Outliers:
Outlier is something different or distinct from the crowd. It can be product of data collection mistake or indicator of inconsistency in data. There are many options to identify and handle outliers:
The training data must include the appropriate answer, which is recognized as a goal attribute. The learning algorithm finds patterns that map the attributes of input data to the target and it produces an ML model that predicts the diseases with use of user input as symptoms.

Data Testing
A methodology where assortments of information are characterized into classes. This should be possible for both unstructured and organized information. This methodology starts with expectation of class with information focuses. Classes can likewise allude as classifications. In this scenario, the classifier needs training data to determine how the key variables are associated to the group And when the classifier is properly trained, it can be used to determine whether the quality of air is good or not.
Since characterization is a type of administered learning, input information is likewise introduced to the objectives. Prescient displaying of arrangement is the way toward approximating the sifting capacity from the factors of contribution to the factors of particular yield. The fundamental objective is to arrange which class/classification the most recent information falls under. The accompanying advances are utilized to utilize the model's suitable calculation

- Evaluating the data.
- Based on features produce reliant and independent data sets.
- Separating datasets into training and testing.
- Prepare the data model with classifiers like support vector machines, decision tree and random forest.
- Analyzing the classifier.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

- Selecting the classifier with best precision

Preparing information alludes to assortment of information to help a program and produce complex results by comprehension and carrying out different AI methods.

Testing information comprises of a progression of results used to decide model effectiveness utilizing certain yield measure. It is fundamental that the test set doesn't contain any outcomes from the preparation set.
This phase includes providing an ML algorithm (i.e., the learning algorithm) with learning data from which to learn. The term ML model referred as artefact of that model where the training method generated. The algorithms such as Naive Bayes, Random Forest, Decision tree and Support Vector algorithms are used for analysis and prediction purpose.

III. METHODOLOGY

There are so many algorithms to predict data such as Naive Bayes, Decision tree, Random forest, Support Vector Machines etc. We have combined all the following discussed algorithms to obtain the maximum accuracy of Disease Prediction. The following Figure shows the general working of Machine Learning Algorithms..

A. Random Forest Algorithm

Random Forest grows multiple decision trees which are merged together for a more accurate prediction.The logic behind the Random Forest model is that multiple uncorrelated models (the individual decision trees) perform much better as a group than they do alone. When using Random Forest for classification, each tree gives a classification or a "vote." The forest chooses the classification with the majority of the "votes." When using Random Forest for regression, the forest picks the average of the outputs of all trees.

The key here lies in the fact that there is low (or no) correlation between the individual models— that is, between the decision trees that make up the larger Random Forest model. While individual decision trees may produce errors, the majority of the group will be correct, thus moving the overall outcome in the right direction of predictingthe disease based on the user input provided, that is symptoms of the user.

B. Naive Bayes Algorithm

Naive Bayes classifier is used to predict the disease on the user input accordingly as a symptoms. This simplified Bayesian classifier is called as naive Bayes. The Naive Bayes classifier assumes the presence of a particular featurein a class is unrelated to the presence of any other feature. It is very easy to build and useful for large datasets.
Naive Bayes is a supervised learning model. Bayes theorem provides some way of calculative posterior chance P(b|a) from P(b), P(a) and P(a|b). Look at the equation below:

- P(b|a) is that the posterior chance of class (b,target) given predictor (a, attributes).
- P(b) is the prior probability of class.
- P(a|c) is that chance that is that the chance of predictor given class.
- P(a) is the priori probability of predictor. In our system, Naïve Bayes decides which symptom is to put in classifier and which is not.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

Naive Bayes Theorem depends on Bayes theorem,

$$P(C_i|\, x_1, x_2 \ldots, x_n) = \frac{P(x_1, x_2 \ldots, x_n|C_i).P(C_i)}{P(x_1, x_2 \ldots, x_n)} \; for\; 1 < i < k$$

Where,

Ci == Class variable;

{x1, x2, …..,xn} == Dependent Features;

C. Decision Tree

Hierarchy structure is the form in which Decision tree builds regression or classification models. It separates a dataset into more modest and more modest subsets while simultaneously a related choice tree is gradually evolved. The outcome is a tree with choice hubs and leaf hubs. The center calculation for building choice trees called ID3 by J. R. Quinlan which utilizes a hierarchical, voracious pursuit through the space of potential branches with no backtracking.

The ID3 calculation can be utilized to build a choice tree for relapse by supplanting Information Gain with Standard Deviation Reduction.

D. Support Vector Machine

In AI, support-vector machines (SVMs, additionally support-vector organizations) are managed AI models ,with which are related learning calculations that investigations the utilization of information for arrangement and relapse examination. The Support Vector Machine (SVM) calculation is a mainstream AI device that offers answers for both order and relapse issues. On account of relapse, an edge of resistance (epsilon) is set in estimate to the SVM which would have effectively mentioned from the issue. Yet, other than this reality, there is likewise a more muddled explanation; the calculation is more confounded hence to be taken in thought.

In any case, the primary thought is consistently something very similar: to limit blunder, individualizing the hyper plane which amplifies the edge, remembering that piece of the mistake is endured.

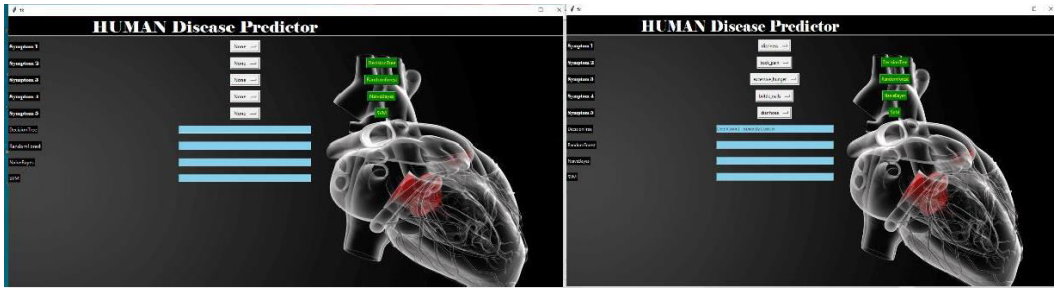
IV. EXPERIMENTAL RESULTS

After all the successful feeding of all the four machine learning algorithms the solution that we have proposedhas the accuracy of 95% data analysis based on the user conditions and symptoms. Our solution was capableof determining the type and the condition of the user as per the common inputs by comparing throughout the whole datasets that we had loaded. That it not only is also able to distinguish between diseases and alerts the user to undergo certain test to confirm this.
In this paper, we also give an option for the user based on what algorithm should be used for his analysis of the inputs given by them and give them the best accuracy by their comparison of results. The user can be aware of the disease, which might affect them and be prepared by their precautions to avoid them.
Fig A shows login page of our website where Fig B shows predicting data in Decision Tree, Fig C shows predicting data in Random Forest, Fig D shows predicting data in Naïve Bayes, Fig E shows predicting data in Support Vector Machine.

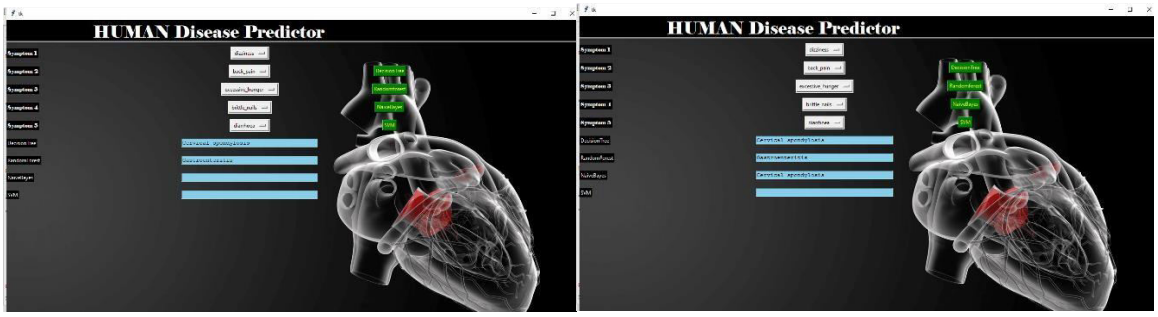
| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

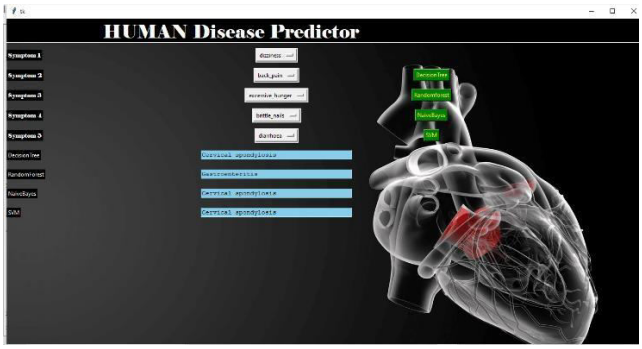
| DOI:10.15680/IJIRSET.2021.1008026 |



(A) (B)



(C) (D)



(E)

IV. CONCLUSION

This venture intends to anticipate the infection based on the side effects. The venture is planned so that the framework accepts side effects from the client as information and produces yield for example foresee illness. Normal forecast exactness likelihood of 95% is acquired. Illness Predictor was effectively executed utilizing chalices structure.

The efficient strategy began from information cleaning and preparing, missing worth, exploratory examination and last model structure and assessment. The best exactness on open test set is higher precision score choice tree calculation for foreseeing the air quality by given ascribes. This application can help India Medical division in anticipating the infections and its status and relies upon that they can make a move.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|

|| Volume 10, Issue 8, August 2021 ||

| DOI:10.15680/IJIRSET.2021.1008026 |

REFERENCES

1. A.Davis, D., V.Chawla, N., Blumm, N., Christakis, N., & Barbasi, A. L. (2008).Predicting Individual Disease Risk Based On Medical History.
2. Adam, S., & Parveen, A. (2012). Prediction System For Heart DiseaseUsing Naive Bayes.
3. Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification With Naive Bayes Approach. *Information Technology Journal* .
4. Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based OnMedical History.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details