IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS

[1] A.Vijetha, [2] Madharam Ravalika, [3] Aijaz Hussain, [4] Thulla Abhinay Goud

[1] Assistant Professor, Department of Information Technology, Teegala Krishna Reddy Engineering College Hyderabad, Telangana, India.

[1] vullojuvijetha@gmail.com

[2,3,4] UG Scholars Department of Information Technology, Teegala Krishna Reddy Engineering College , Hyderabad, Telangana, India.

[2]madaramravalika258@gmail.com , [3] aijazhussain683@gmail.com ,[4] abhinaytulla1234@gmail.com

Abstract

The Disease Prediction System based on various prediction models that help to predict the disease of the user on the basis of the symptoms that user enters as an input to the system. Predictive models with the help of machine learning classification algorithms analyzes the symptoms provided by the user as input and gives the name and probability of the disease as an output. Disease Prediction is done by implementing the Naive Bayes Classifier, Decision tree and Random Forest Algorithm. The Naive Bayes helps to calculate the probability of the disease which is predicted. Average prediction accuracy probability 87% is obtained. The model uses a dataset with the count of 132 symptoms from which the user can select their symptoms. The user does not need to have a medical report to use this system as the prediction is based on the symptoms which will save the money. The system also has a very easy to use user interface so all the users can use it to predict the generic diseases. People are currently suffering from a variety of diseases. Many people are unsure if the symptoms they are experiencing are indicative of a certain disease, and hence they are unable to take the required safeguards.People will not be able to visit a doctor every time they experience a symptom. It may sometimes become a serious ailment if not treated. A model is suggested that uses a variety of symptoms as input to predict the illness. For disease prediction, the suggested method utilizes Decision trees, Naive Bayes, and Random forest classifiers. The ultimate result will be the mode of all these machine learning models. Users will be given a graphical user interface (GUI) to choose their symptoms. The final result will be shown on the interface using all three machine learning techniques, and feature extraction will be done depending on their symptoms. Four modules make up the proposed methodology. Preprocessing will be done on the dataset in the first module. The decision tree classifier is used to generate a prediction model in the second module. The Random forest method is used for forecast the illness in the third module, and the Naive Bayes technique is utilized in the fourth model, with the mode of the outputs from all the three models taken into account.

1. INTRODUCTION

There are times when we need a doctor all of a sudden but sometimes they are not available due to some reason and we are left in trouble. The system we have proposed is user friendly to get help and advice on health issues immediately through the online healthcare system. Now a days, with the help of the statistics and posterior distribution the problems are swiftly and easily. As the Bayesian statistics has a great success rate in the field of economic, social science and a few other fields just like that, in medical fields, people have solved various medical problems that are tiresome to be settled in classic statistics by classification and can be solved easily. Naive Bayes is among the basic common classification techniques introduced by Reverend Thomas Bayes.

The classification rules which help in solving the prediction of disease are generated by the samples trained by themselves and help in solving the problem easily. It is approximated that greater than 70% of people in India are prone to various body diseases like viral, flu, cough, cold etc. in intervals of 2 months. As many people don't understand that the general body diseases could be symptoms of something more harmful, 25% of this population dies or gets some serious medical problem because of ignoring the early general body symptoms and this is a very serious condition that we are facing and the problem can be proven to be a very dangerous situation for the population and can be alarming if the people will continue ignoring these diseases. Hence identifying or predicting the disease at the very basic stage is very important to avoid any unwanted problems and deaths.

The systems which are available now a days are the systems that are either dedicated to a particular disease or are in development or the research for solving the algorithms related to the problem when it comes to generalized disease. The main motive of the proposed system is the prediction of the commonly occurring diseases in the early phase as when they are not checked or examined they can turn into a disease more dangerous disease and can even cause death. The system applies data mining techniques,decision tree algorithms, Naive Bayes algorithm and Random Forest algorithm.

This system will predict the most possible disease based on the given symptoms by the user and precautionary measures required to avoid the aggression of disease, it will also help doctors to analyze the patterns of diseases in the society.

This project is dedicated to the Disease prediction System that will have data mining techniques for the basic stages of the dataset and the main model will be trained using the Machine Learning (ML) algorithms and will help in the prediction of general diseases.

1.1 Data Mining and Machine Learning Algorithms The Data Mining and the Machine Learning Algorithms are used for the prediction of Disease in the Project. There are different Data Mining and Machine Learning used for the purpose of correcting and evaluating the dataset and then testing the dataset on the basis of train score and the test score of the ML model.

1.1.1 Data Analysis and Data Mining

The Data Mining is a process in which raw data is prepared and structured from the unstructured data as to take meaningful information from the data which can be used in the project. Task of making data organized and reflective about data is to way to get what this information does the data contains in it and what it does not have in it. There are so many different types of methods in which the people can make use of data analysis. It is simply very easy to use data during the analysis phase and get to some certain conclusions or some agendas. The analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the objective of highlighting useful information, suggesting conclusions, and supporting decision making which are helpful to the user. Data analysis has multiple facets and approaches, encompassing diverse techniques under an array of names, in different business, science, and social science domains.

Data Mining is the discovery of unknown information found in databases, data mining functions has some different methods for clustering, classification, prediction, and associations. In the data mining important application is that of mining association rules, association rules was first introduced in 1993 and are used to identify relationships among a set of items in databases these different properties are not based on the properties of the data, but rather based on cooccurrence of the data items.

The Data mining helps in giving new and different perspectives for data analysis the main role of data mining is to extract and discover new knowledge from data. In the past few years, different methods have been coined and developed about the capabilities of data collection and data generation, data collection tools have provided us with a huge amount of data, data mining processes have integrated techniques from multiple disciplines such as, statistics, machine learning, database technology, pattern recognition, neuralnetworks, information retrieval and spatial data analysis. The data mining techniques have been used in many different fields such as, business management, science, engineering, banking, data management, administration, and many other applications.

1.1.2 Machine Learning Algorithms

The ML is a small part of Artificial Intelligence (AI) which is used in the computation work and the analysis work in the AI. The ML algorithms are used to find different patters and different structures in the dataset which is provided to the dataset, the ML algorithms are used to give a large computation capabilities to the system by which a large amount of data is given to the model for the purpose of training and testing the data, the ML algorithms are used in decision making process the model which is prepared by using the ML has a large amount of data in it which makes it a very good for the process of decision making. ML algorithms have very high computational power and are proven to be very helpful in today’s world. Different types of ML algorithms are organized into different ways, based on the desired outcome of the algorithm. Common algorithm types include:

 Supervised learning — The supervised learning algorithm can apply what has been learned in the past to new data using labelled examples to predict the future events. Starting from analysis of a known training dataset. This algorithm is used to provide targets for any new values after sufficient amount of training of the model.

 Unsupervised learning — Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. This algorithm shows how the system can infer a function to describe a hidden structure from unlabeled data.

 Semi-supervised learning — This category of the ML algorithms falls some where between the supervised learning and the unsupervised learning algorithm which combines both labeled and unlabeled examples to generate an appropriate function or classifier which is used to make a model for the purpose of prediction or classification.

 Reinforcement learning — This is the algorithm where the algorithm learns a policy of how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback that guides the learning algorithm.

 Transduction — This algorithm is similar to supervised learning, but does not explicitly construct a function: instead, tries to predict new outputs based on training inputs, training outputs, and new inputs.

 Learning to learn — This method is where the algorithm learns its own inductive bias based on previous experience. The performance and computational analysis of ML algorithms is a branch of statistics known as computational learning theory.

Machine learning is about designing algorithms that allow a computer to learn. Learning is not necessarily involves consciousness but learning is a matter of finding statistical regularities or other patterns in the data. Thus, many machine learning algorithms will barely resemble how human might approach a learning task. However, learning algorithms can give insight into the relative difficulty of learning in different environments Machine learning is made up of three parts:

 The computational algorithm at the core of making determinations.

 Variables and features that make up the decision.

 Base knowledge for which the answer is known that enables (trains) the system to learn. Initially, the model is fed parameter data for which the answer is known. The algorithm is then run, and adjustments are made until the algorithm’s output (learning) agrees with the known answer. At this point, increasing amounts of data are input to help the system learn and process higher computational decisions.

2.LITERATURE SURVEY

In the paper “Disease Prediction System using data mining techniques”[1] the author has discussed about the data mining techniques like association rule mining, classification, clustering to analyse the different kinds of heart based problems. The database used contain collection of records, each with a single class label, a classifier performs a brief and clear definition for each class that can be used to classify successive records.

The data classification depends on MAFIA algorithms that cause accuracy, the info is calculable exploitation entropy primarily based cross validations and partition techniques and also the results are compared. C4.5 algorithmic rule is employed because the coaching algorithmic rule to indicate rank of attack with the choice tree.

The heart unwellness information is clustered mistreatment the K-means clump algorithmic rule, which will remove the data applicable to heart attack from the database. Some limitations square measure faced by the system like, time complexity is more due to DFS traversal, C4.5-Time complexity increases while searching for insignificant branches and lastly no precautions are defined. In the paper “A study on data mining prediction techniques in healthcare sector” [2] the fields that mentioned are, information Discovery method (KDD) is that the method of adjusting the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial removal of implicit, antecedently unknown and doubtless helpful data from information in databases.

The repetitious method consists of the subsequent steps: information cleansing, information integration, information choice, information transformation, data processing, Pattern analysis, Knowledge. Healthcare data processing prediction supported data processing techniques are as follows: Neural network, Bayesian Classifiers, call tree, Support Vector Machine. The paper states the comparative study of various aid predictions, Study of information mining techniques and tools for prediction of cardiovascular disease, numerous cancers, and diabetes, disease and medicine conditions.

Few limitations are that if attributes are not related then Decision trees prediction is less accurate and ANN is computationally intensive to train also it does not lead to specific conclusion. The paper “Predicting Disease by Using Data Mining Based on Healthcare Information System” [4] applies the information mining process to predict high blood pressure from patient medical records with eight alternative diseases. The data was extracted from a true world health care system info containing medical records. Under- sampling technique has been applied to come up with coaching knowledge sets, and data processing tool wood hen has been wont to generate the Naive Bayesian and J48 classifiers created to improve the prediction performance, and rough set tools were wont to scale back the ensemble supported the concept of second- order approximation. Experimental results showed a bit improvement of the ensemble approach over pure Naive Bayesian and J-48 in accuracy, sensitivity and F-measure. Initially they'd a classification and so ensemble the classifiers and so the reduction of Ensemble Classifiers is employed.

But the choice trees generated by J-48 is typically lacking within the leveling therefore the overall improvement of victimization ensemble approach is a smaller amount. The paper “An approach to devise an Interactive software solution for smart health prediction using data mining” [5] aims in developing a computerized system to check and maintain your health by knowing the symptoms.

It has a symptom checker module which actually defines our body structure and gives us liability to select the affected area and checkout the symptoms. Technologies implemented in this paper are: The front end is designed with help of HTML, Java Script and CSS. The back end is designed using MySQL which is used to design the databases.

This paper also contains the information of testing like Alpha testing which is done at server side or we can say at the developer's end, this is an actual testing done with potential users or as an independent testing process at server end. And Beta testing is done after performing alpha testing, versions of a system or software known as beta versions are given to a specific audience outside the programming team. Only the limitation of this paper is it suggests only the award winning doctors and not the nearby doctors to the patient.

3. PROBLEM STATEMENT

Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects. The National Patient Safety Foundation cites that 42% of medical patients feel they have had experienced a medical error or missed diagnosis. Patient safety is sometimes negligently given the back seat for other concerns, such as the cost of medical testsdrugs, and operations. Medical Misdiagnoses are a serious risk to our healthcare profession. If they continue, then people will fear going to the hospital for treatment. We can put an end to medical misdiagnosis by informing the public and filing claims and suits against the medical practitioners at fault.

4. PROPOSED SYSTEM

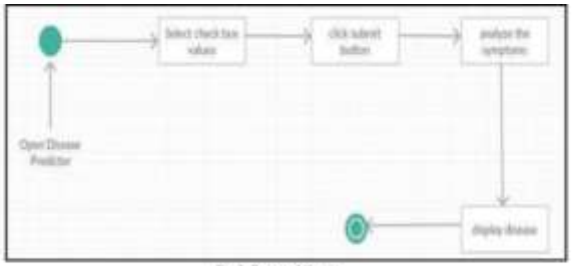


Fig.3: Proposed System

We are predicting a disease which a person is suffering from depending upon the symptoms he or she is suffering. Here we take five symptoms from the patient and evaluate them by using algorithms such as Random Forest , Decision Tree, Naïve Bayes. Steps of model building:

i. Objective

We want to predict the disease suffered by a patient depending upon the symptoms.

ii. Collecting data

Be it the raw data from excel, access, text files etc., this step (gathering past data) forms the foundation of the future learning. The better the variety, density and volume of relevant data, better the learning prospects for the machine becomes.

iii. Preparing the data

Any analytical process thrives on the quality of the data used. One needs to spend time determining the quality of data and then taking steps for fixing issues such as missing data and treatment of outliers. Exploratory analysis is perhaps one method to study the nuances of the data in details thereby burgeoning the nutritional content.

iv. Training a model

This step involves choosing the appropriate algorithm and representation of data in the form of the model. The cleaned data is split into two parts – train and test (proportion depending on the prerequisites); the first part (training data) is used for developing the model.

v.Evaluating the model

To test the accuracy, the second part of the data (holdout / test data) is used. This step determines the precision in the choice of the algorithm based on the outcome. A better test to check accuracy of model is to see its performance on data which was not used at all during model build.

vi. Improving the performance

This step might involve choosing a different model altogether or introducing more variables to augment the efficiency. That’s why significant amount of time needs to be spent in data collection and preparation.

5. SYSTEM ARCHITECTURE

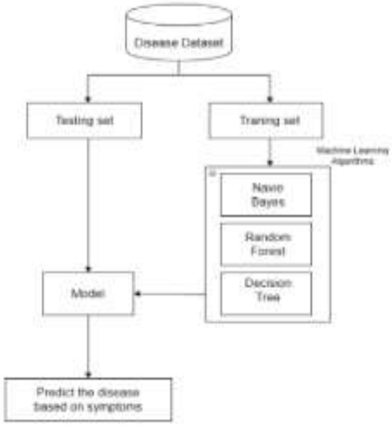


Fig.4:System Architecture

6. IMPLEMENTATION

There are three different kind of models present in our project to predict the disease these are

Decision tree, Random forest tree AND Gaussian Naïve Bayes.

6.1 Decision Tree

Decision tree is classified as a very effective and versatile classification technique. It is used in pattern recognition and classification for image. It is used for classification in very complex problems dew to its high adaptability. It is also capable of engaging problems of higher dimensionality. It mainly consists of three parts root, nodes and leaf. Roots consists of attribute which has most effect on the outcome, leaf tests for value of certain attribute and leaf gives out the output of tree.

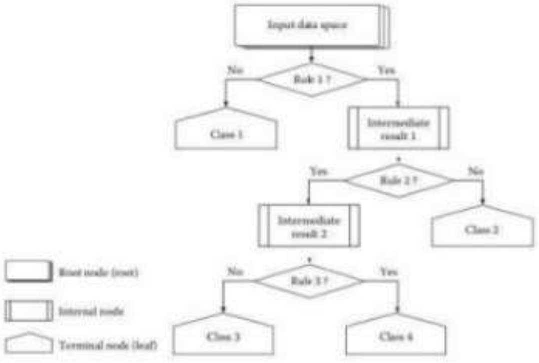


Fig.1: Decision Tree

6.2 Random Forest

Random Forest is a great algorithm to train early in the model development process, to see how it performs and it’s hard to build a “bad” Random Forest, because of its simplicity. This rule is additionally an excellent alternative, if you would like to develop a model during a short amount of your time. On prime of that, it provides a fairly sensible indicator of the importance it assigns to your options. Random Forests are terribly onerous to ram down terms of performance. And on prime of that, they'll handle tons of various feature varieties, like binary, categorical and numerical. Overall, Random Forest may be a (mostly) quick, easy and versatile tool, though it's its limitations. Random forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the categories (classification) or mean prediction (regression) of the individual trees Random call forests correct for call trees' habit of over fitting to their training set.

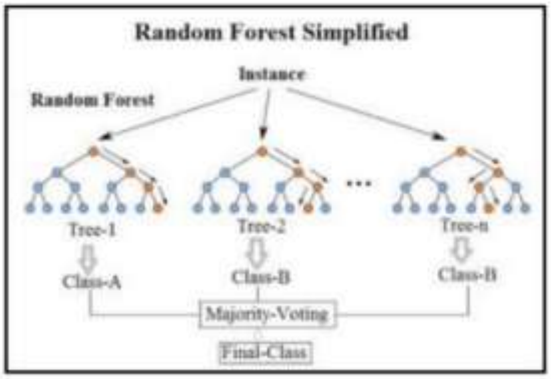


Fig.2: Random Forest

6.3 Naive Bayes Algorithm

Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. For instance, if you are trying to identify a fruit based on its color, shape and taste, then an orange colored, spherical, and tangy fruit would most likely be an orange. All these properties individually contribute to the probability that this fruit is an orange and that is why it is known as “naive”. As for the “Bayes” part ,it refers to statistician and philosopher, Thomas Bayes and the theorem named after him , Bayes’ theorem , which is the base for Naïve Bayes Algorithm. More formally, Bayes ‘Theorem is stated as the following equation:

P(A/B) = (P(B/A)*P(A)) / P(B)

7. LIBRARY USED

In this project standard libraries for database analysis and model creation are used. The following are the libraries used in this project

1. tkinter: It’s a standard GUI library of python. Python when combined with tkinter provides fast and easy way to create GUI. It provides powerful object-oriented tool for creating GUI. It provides various widgets to create GUI some of the prominent ones being:

□ Button
□ Canvas
□ Label
□ Entry
□ Check Button
□ List box
□ Message
□ Text
□ Messagebox

Some of these were used in this project to create our GUI namely messagebox, button, label, Option Menu, text and title. Using tkinter we were able to create an interactive GUI for our model.

2. Numpy: Numpy is core library of scientific computing in python. It provides powerful tools to deal with various multi-dimensional arrays in python. It is a general purpose array processing package.

Numpy's main purpose is to deal with multidimensional homogeneous array. It has tools ranging from array creation to its handling. It makes it easier to create a n dimensional array just by using np.zeros() or handle its contents using various other methods such as replace, arrange, random, save, load it also helps I array processing using methods like sum, mean, std, max, min, all, etc Array created with numpy also behave differently then arrays created normally when they are perated upon using operators such as +,-,*,/. All the above qualities and services offered by numpy array makes it highly suitable for our purpose of handling data. Data manipulation occurring in arrays while performing various operations need to give the desired results while predicting outputs require such high operational capabilities.

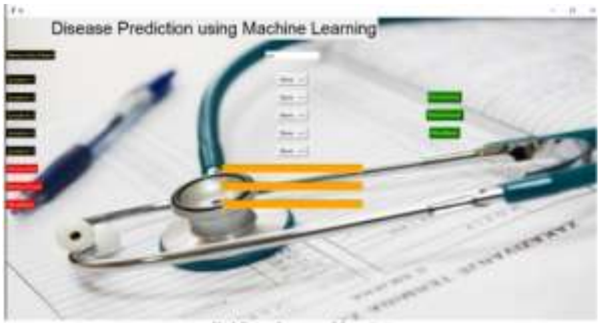
3. pandas : it is the most popular python library used for data analysis. It provides highly optimized performance with back-end source code purely written in C or python. Data in python can be analysed with 2 ways

- Series
- Dataframes

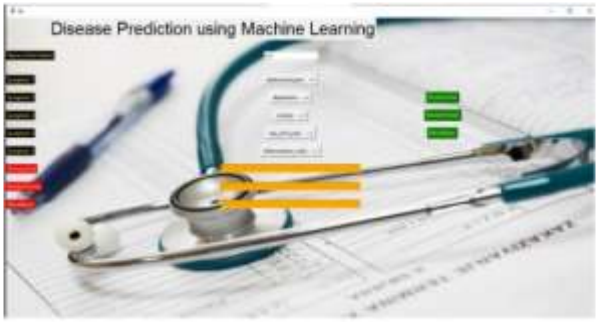
Series is one dimensional array defined in pandas used to store any data type. Dataframes are two-dimensional data structure used in python to store data consisting of rows and columns. Pandas dataframe is used extensively in this project to use datasets required for training and testing the algorithms. Dataframes makes it easier to work with attributes and results. Several of its inbuilt functions such as replace were used in our project for data manipulation and preprocessing.

4. sklearn: Sklearn is an open source python library with implements a huge range of machine- learning, pre-processing, cross-validation and visualization algorithms. It features various simple and efficient tools for data mining and data processing. It features various classification, regression and clustering algorithm such as support vector machine, random forest classifier, decision tree, gaussian naïve-Bayes, KNN to name a few. In this project we have used sklearn to get advantage of inbuilt classification algorithms like decision tree, random forest classifier, KNN and naïve Bayes. We have also used inbuilt cross validation and visualization features such as classification report, confusion matrix and accuracy score.

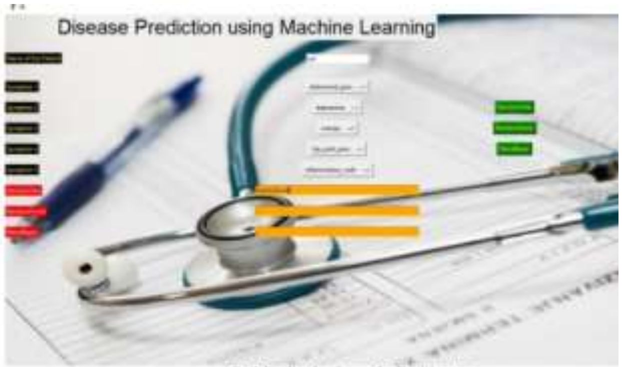
8. RESULTS



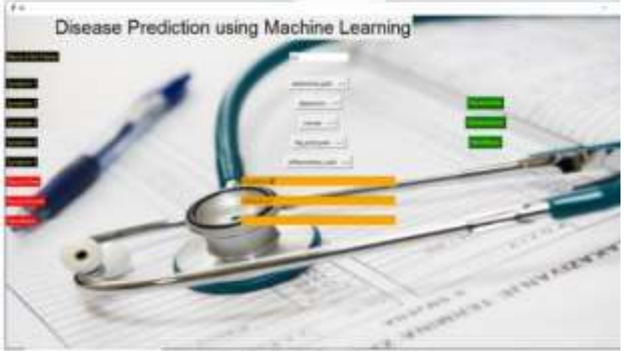
Op 1 Enter the name of the patiant



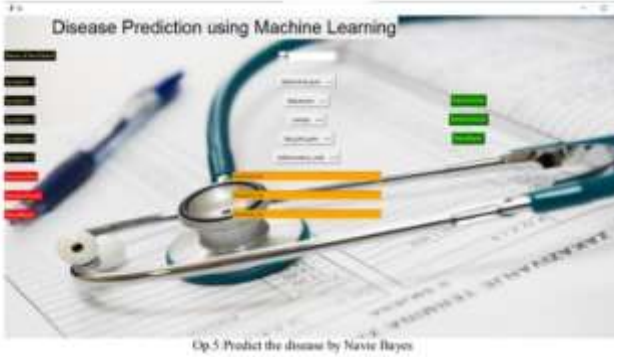
Op 2 Input the symptoms



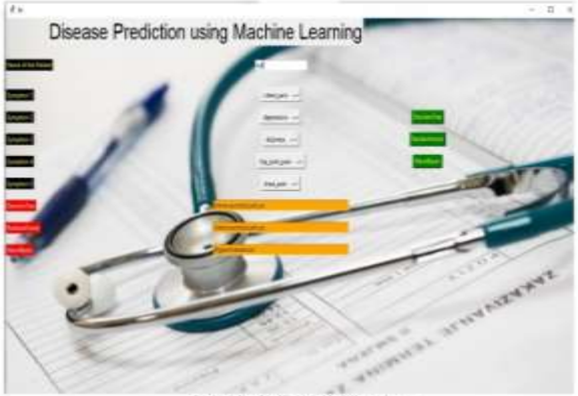
Op.3 Predict the disease by Decision tree



Op.4 Predict the disease by Random forest



Op.5 Predict the disease by Navie Bayes



Op.6 Predict the disease with different inputs

9. CONCLUSION

The ultimate goal is to facilitate coordinated and well-informed health care systems capable of ensuring maximum patient satisfaction. In developing nations, predictive analytics are the next big idea in medicine –the next evolution in statistics – and roles will change as a result. Patients can get to become higher knowing and can get to assume a lot of responsibility for his or her own care, if they are to make use of the information derived. Physician roles can probably modification to a lot of an advisor than head, who will advise, warn and help individual patients. Physicians might notice a lot of joy in apply as positive outcomes increase and negative outcomes decrease. Perhaps time with individual patients can increase and physicians will another time have the time to create positive and lasting relationships with their patients. Time to assume, to interact, and to really help people; relationship formation is one of the reasons physicians say they went into medicine, and when these diminish, so does their satisfaction with their profession. Hospitals, pharmaceutical corporations and insurance suppliers can see changes furthermore. .

10. FUTURE ENHANCEMENT

Every one of us would like to have a good medical care system and physicians are expected to be medical experts and take good decisions all the time. But it's highly unlikely to memorize all the knowledge, patient history, records needed for every situation. Although they have all the massive amount of data and information; it's difficult to compare and analyse the symptoms of all the diseases and predict the outcome. So, integrating information into patient's personalized profile and performing an in-depth research is beyond the scope a physician. So the solution is ever heard of a personalized healthcare plan – exclusively crafted for an individual. Predictive analytics is the process to make predictions about the future by analyzing historical data. For health care, it would be convenient to make best decisions in case of every individual. Predictive modeling uses artificial intelligence to create a prediction from past records, trends, individuals, diseases and the model is deployed so that a new individual can get a prediction instantly.

11. REFERENCES

[1] Aditya Tomar, "Disease Prediction System using data mining techniques", in International Journal of Advanced Research in computer and Communication Engineering, ISO 3297, July 2016.

[2] Dr. B.Srinivasan, K.Pavya, "A study on data mining prediction techniques in healthcare sector", in International Research Journal of Engineering and Technology (IRJET), March-2016.

[3] Megha Rathi, Vikas Pareek, "An integrated hybrid data mining approach for healthcare" , in IRACST -International Journal of Computer Science and Information Technology Security (IJCSITS), ISSN: 2249-9555 , Vol.6, No.6,Nov-Dec 2016.

[4] Feixiang Huang, Shengyong Wang, and Chien-Chung Chan, "Predicting Disease By Using Data Mining Based on Healthcare Information System" , in IEEE 2012.

[5] M.A. Nishara Banu,B Gomathy, "An approach to devise an Interactive software solution for smart health prediction using data mining, in International Journal of Technical Research and Applications , eISSN, Nov-Dec 2013.

[6] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2012). Medical Data Classification with Naïve Bayes Approach. Information Technology Journal.

[7] Darcy A. Davis, N. V.-L. (2008). Predicting Individual Disease Risk Based On Medical History.