DISEASE PREDICTION USING MACHINE LEARNING OVER BIGDATA

Afroz Fatima

M.Tech(CSE), Final Year,

PDA College of Engineering, kalaburagi, India.

Abstract : The Advanced Healthcare System applies Machine Learning Algorithms for the implementation and development of medical facilities, enabling precise decisions about patient treatment and diagnosis. This System helps people to process and analyze, generate deep insights of large datasets of medical procedures. Physicians could use these advanced application, further in the provision of medical care. Therefore the significant contribution towards Machine Learning applications in healthcare could improve patient satisfaction. In this paper, we are attempting to incorporate Healthcare- Machine learning functionalities inside a single program. This sophisticated technology of machine leaning has made the healthcare system intelligent such that instead of diagnosis, the disease could be predicted using predictive algorithms of machine learning. Some cases can occur when diagnosing a disease early is not within reach. It is also possible to successfully apply the prediction of disease. As widely stated "Prevention is better than cure," disease prediction and epidemic outbreak would lead to an early prevention of a disease occurrence. This paper focuses mainly on the implementation of a program or we might claim an immediate medical service that would integrate the symptoms gathered from multisensory tools and other medical data and store them in a healthcare dataset. This dataset would then be analyzed using Decision tree, Random Forest, NaïveBayes machine learning algorithms to ensure maximum accuracy of results.

Keywords - Decision tree, Random Forest, NaïveBayes, Machine Learning.

I. INTRODUCTION

Predicting disease through the use of data mining and machine learning techniques using records of patient care and health data is a constant challenge over the past decades. Some methods often seek to forecast disease prevention and progression. The recent advancement of deep learning in various machine learning applications has led to a change to machine learning models that can learn complex, hierarchical representations of raw data with little pre-processing and produce more precise results. The recent development in big data technology, more attention was given for disease from the perspective of big data analysis; various research was carried out automatically that could select the characteristics from the huge data and could improve the accuracy of risk classification rather than the characteristics previously selected. The main focus of this Healthcare -Machine learning application is to supplement patient care for better outcomes. The Machine learning algorithms made easier to correctly identify and diagnose the various diseases. With the aid of effective multiple machine learning algorithms, predictive modeling helps to predict the disease more accurately and helps to treat patients. The huge amount of healthcare data is produced from healthcare industry daily and that could be used to extract information for disease prediction that could help patient in the future while using history Machine learning in healthcare helps people process and analyze large and complex medical datasets into clinical insights. Physicians can then use this further in the provision of medical care. Therefore machine learning can contribute to improved patient satisfaction when applied in healthcare. The Decision tree, Random Forest, NaïveBayes algorithm is used to predict illnesses using the history and health details of patient care.

ISSUES:-

1. It can predict the diseases but not the subtype of diseases
2. It fails to predict the condition of people. The prediction **of** diseases have been nonspecific and indefinite

II. RELATED WORK

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: using analytics to identify and manage high-risk and high-cost patients," Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014. : This paper is suggested that the Data Mining technique is best which is able to estimate dangerousness of the running seismic events quickly and correctly. Many seismic recordings of Japanese earthquakes were analyzed and a model was obtained by means of ten data mining techniques, namely, Binomial distribution, Earlang distribution, Exponential smoothing, Poisson distribution, Same birthday paradox, Linear regression, Artificial Neural Networks (ANN), Decision tree (C5), M8 algorithm and MSc algorithm, which were tested at the time of first recordings of seismic events to reduce the decision time and the test results were very satisfactory.

[2] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, "Performance the comparison of three data mining techniques for forecast kidney disease survivability", International Journal of Advances in Engineering & Technology, Mar. 2014.: We discuss the types of insights that are probable to appear from clinical analytics, the types of data needed to obtain such insights, and the infrastructure—analytics, algorithms, registries, assessment scores, monitoring devices, and so forth—that organizations will need to perform the necessary examine and to implement changes that will improve care while reducing costs. Our findings have policy implications for regulatory surveillance, ways to address privacy concerns, and the support of research on analytics.

[3] Mr. Chala Beyene, Prof. Pooja Kamat, "Survey on projection and examination the Occurrence of Heart Disease Using Data Mining Techniques", International Journal of Pure and Applied Mathematics, 2018. : One of the most common problems in medical centers is that all experts have not an equal knowledge and skill to treat their patients, they give their own decision that may give poor results and cause the patients to death. To overcome such problems projection the occurrence of heart diseases using data mining techniques and machine learning algorithms are playing important Some machine algorithms used for predicting the occurrence of heart diseases are Support Vector Machine, Decision Tree, Naïve Bayes, K-Nearest Neighbor, and Artificial Neural Network.

PROPOSED SYSTEM

The proposed framework combined the structure with unstructured data in healthcare fields that allow us to assess disease risk. The latent factor model method to recreate the missing data in medical records that are obtained from the hospital. And we could evaluate the major chronic diseases in a specific area and in particular society by using statistical information. We work with hospital specialists to learn valuable features for managing structured data. With the support of k-mean algorithm we select the features automatically in the case of unstructured text data. For structured as well as unstructured data we suggest a decision tree, Random forest, Naïve Bays algorithm.

III. METHODOLOGY

Decision Tree: A decision tree is a structure comprising a node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the test result, and each leaf node is labeled as a class. The root node is the topmost node in the tree.

Algorithm Steps:

1. Pick the best attribute/feature. The best attribute is one which best splits or separates the data.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive to the answer.

Naïve Bayes: Naïve Bayes algorithms is a classification technique based on the implementation of the theorem of Bayes, with a clear assumption that all the predictors are independent. The presumption, in plain words, is that the existence of a feature in a class is independent of any other feature in the same class.

Algorithm Steps:

Naive Bayes classifier calculates the probability of an event in the following steps:

1. **Step** 1: Calculate the prior probability for given class labels.

2. **Step** 2: Find Likelihood probability with each attribute for each class.
3. **Step** 3: Put these value in **Bayes** Formula and calculate posterior probability.

Random Forest: Random Forest is a popular learning machine algorithm within the supervised learning technique. This can be used in ML for Classification and Regression problems. It is based on the principle of learning the ensemble, which is a process of combining multiple classifiers in order to solve a complex problem and to improve model efficiency.

Algorithm Steps:

The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

MODULES

1. Load dataset: A data set is a collection of data in Machine Learning , we need a training data set. It is the actual data set used to train the model for performing various actions. here in this module loading the patient dataset
2. Preprocessing: In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm. Here it preprocesses the data as per our dataset.
3. Classification: Classification is a type of supervised learning. It specifies the class to which data elements belong to and is best used when the output has finite and discrete values. It predicts a class for an input variable as well. here it predicts the disease using naive bayes, random forest and decision tree.

Experimental results

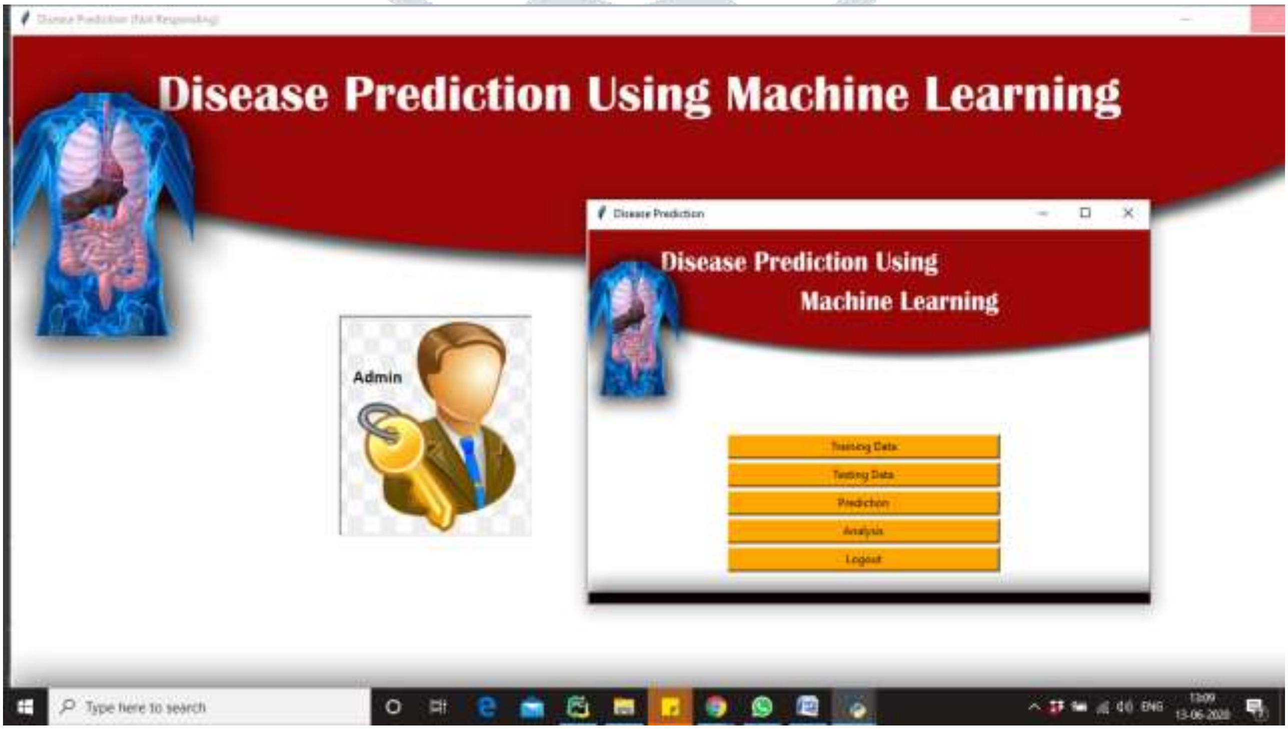
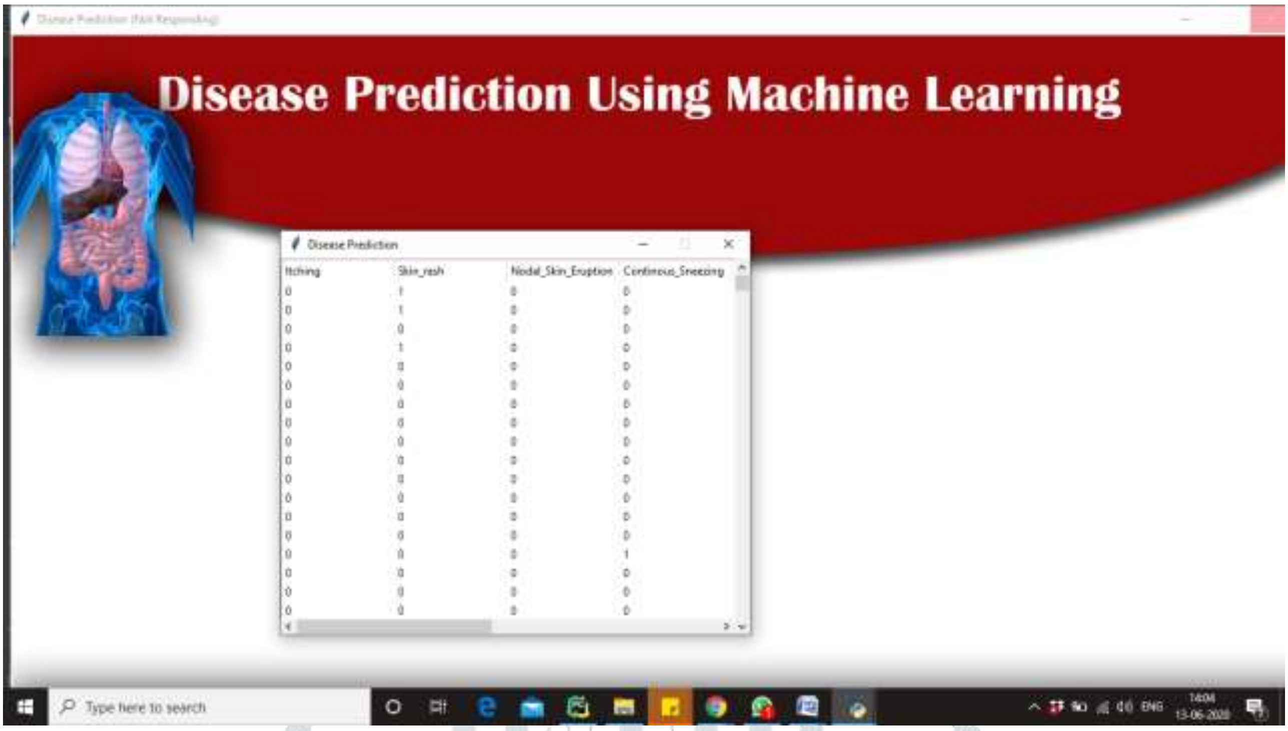
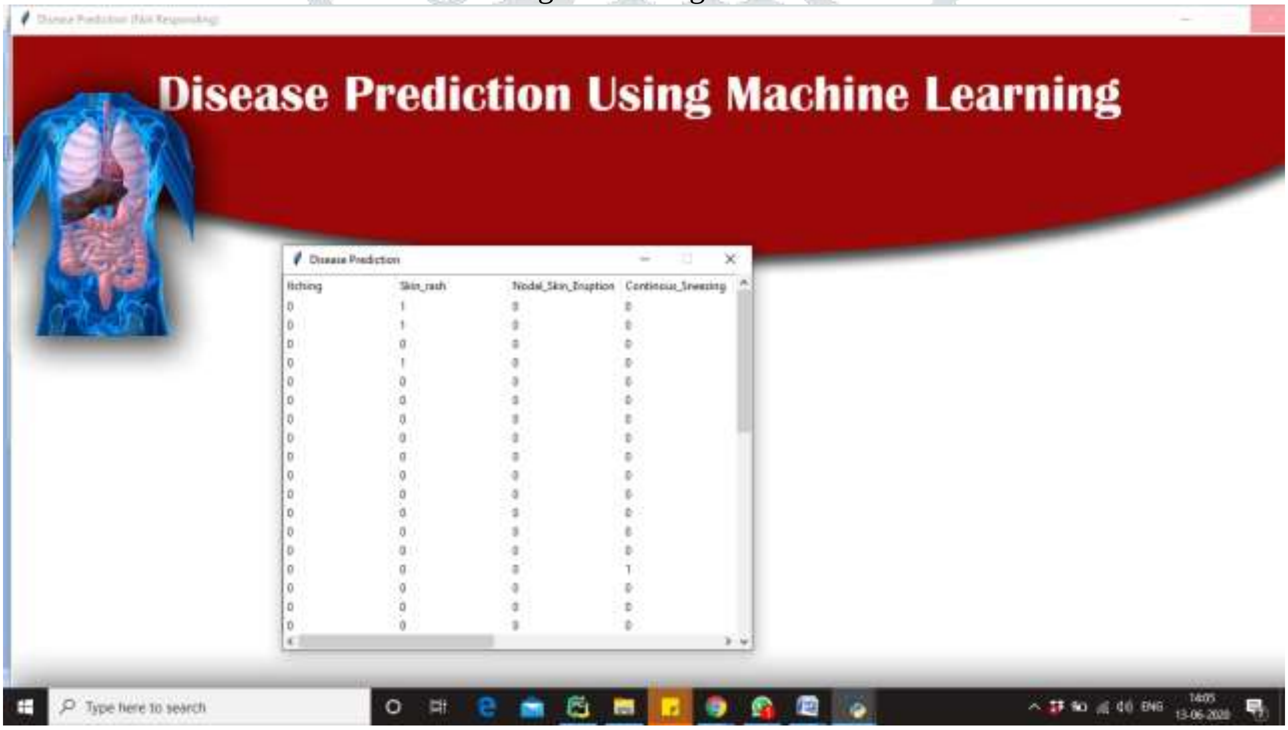


Fig 1:Menu


Disease Prediction (Not Responding)
Disease Prediction Using Machine Learning
Disease Prediction
Type here to search
ENG
14:04
13-06-2020

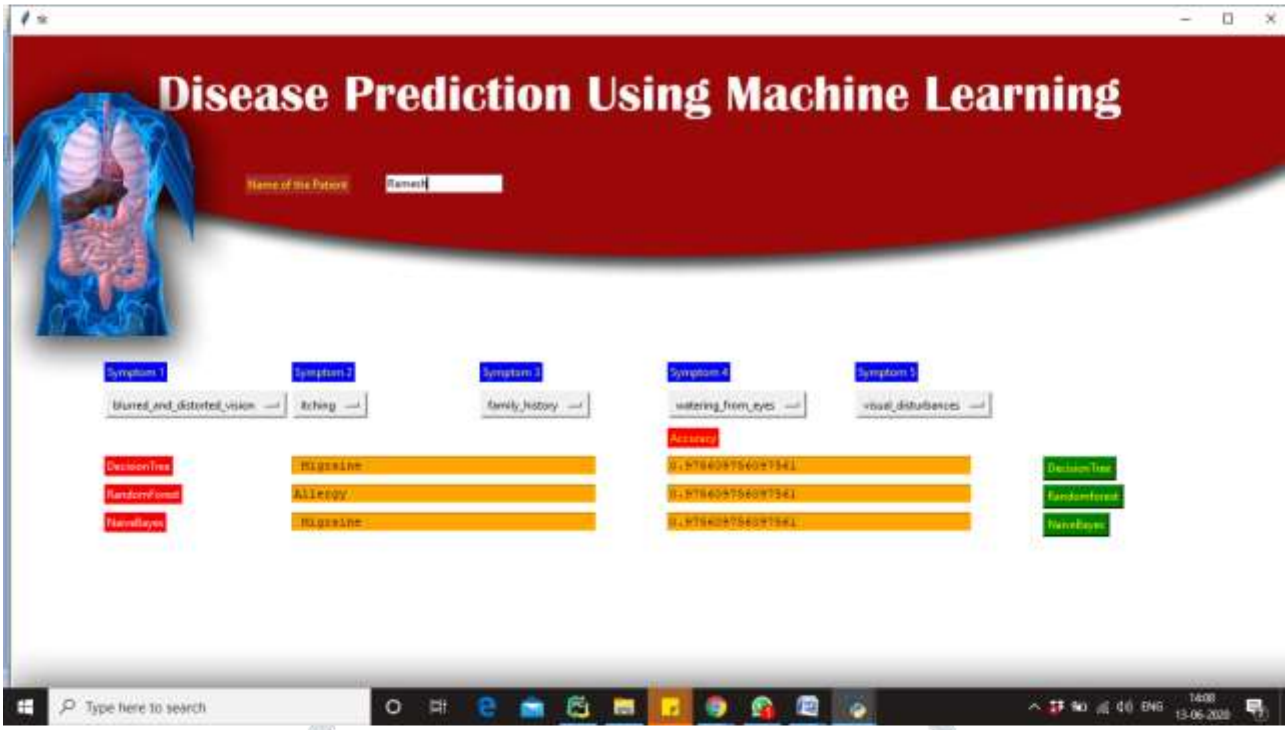
Itching	Skin_rash	Nodal_Skin_Eruption	Continous_Sneezing
0	1	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0

Fig 2: Training Data


Disease Prediction (Not Responding)
Disease Prediction Using Machine Learning
Disease Prediction
Type here to search
ENG
14:05
13-06-2020

Itching	Skin_rash	Nodal_Skin_Eruption	Continous_Sneezing
0	1	0	0
0	1	0	0
0	0	0	0
0	1	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	1
0	0	0	0
0	0	0	0
0	0	0	0

Fig 3: Testing Data



Name of the Patient: Ramesh

Symptom 1: blurred_and_distorted_vision
Symptom 2: itching
Symptom 3: family_history
Symptom 4: watering_from_eyes
Symptom 5: visual_disturbances

		Accuracy
DecisionTree	Migraine	0.975609756097561
RandomForest	Allergy	0.975609756097561
NaiveBayes	Migraine	0.975609756097561

Fig 4: Classification Algorithm

IV. CONCLUSION

With the proposed system, higher accuracy can be achieved. We not only use structured data, but also the text data of the patient based on the proposed k-mean algorithm. To find that out, we combine both data, and the accuracy rate can be reached up to 95%. None of the existing system and work is focused on using both the data types in the field of medical big data analytics. We propose a K-Mean clustering algorithm for both structured and unstructured data. The disease risk model is obtained by combining both structured and unstructured features.

REFERENCES

[1] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, “Big data in health care: using analytics to identify and manage high-risk and high-cost patients,” Health Affairs, vol. 33, no. 7, pp. 1123–1131, 2014.

[2] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna, ”Performance comparison of three data mining techniques for predicting kidney disease survivability”, International Journal of Advances in Engineering & Technology, Mar. 2014. [3] Mr. Chala Boyne, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics, 2018.

[4] Boshra Brahmi, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164168. [5] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, “Incorporating temporal ehr data in predictive models for risk stratification of renal function deterioration,” Journal of biomedical informatics, vol. 53, pp. 220–228, 2015.

[6] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” Int. J. of Inform. Sci. and Tech., Vol. 6, pp. 53-60,March 2016.