International Conference on Recent Advances in Computer Science, June 2021
Dept. of Computer Science & Engineering
Saveetha Engineering College ,Chennai, India.

DISEASE PREDICTION USING MACHINE LEARNING

[1] Mrs Swathi Mandava, [2] Mr.Prabhakar Marry , [3] Ms. Wuriti Kavya
[1] [2] [3] Assiastant professor ,Department of IT, Vignan Institution of Technology and science,Hyderabad, India,
[1] swathi.mandava599@gmail.com [2] marryprabhakar@gmail.com [3] kavyawuriti@gmail.com

ABSTRACT

Big data has a major impact on healthcare analytics and it have the capacity to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life. Accurate analysis of medical data benefits in early disease detection and well patient care in big data. The analysis accuracy is reduced when we have incomplete data. In this paper, machine learning algorithms is used for effective prediction of diseases. Latent factor model is used to overcome the difficulty of missing data.

Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. In this paper, we streamline machine learning algorithms for effective prediction of chronic disease outbreak in disease-frequent communities.

One such implementation of machine learning algorithms is in the field of healthcare. Medical facilities need to be advanced so that better decisions for patient diagnosis and treatment options can be made. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyze them into clinical insights. This then can further be used by physicians in providing medical care.

Hence machine learning when implemented in healthcare can leads to increased patient satisfaction. In this paper, we try to implement functionalities of machine learning in healthcare in a single system. Instead of diagnosis, when a disease prediction is implemented using certain machine learning predictive algorithms then healthcare can be made smart. Some cases can occur when early diagnosis of a disease is not within reach.

Hence disease prediction can be effectively implemented. As widely said “Prevention is better than cure”, prediction of diseases and epidemic outbreak would lead to an early prevention of an occurrence of a disease. This paper mainly focuses on the development of a system or we could say an immediate medical provision which would incorporate the symptoms collected from multisensory devices and other medical data and store them into a healthcare dataset. The Machine Learning algorithms that we use on this dataset are Decision Tree algorithm, Random Forest algorithm, Naïve Bayes algorithm.

KEYWORDS: Big Data, healthcare, Machine learning, Decision Tree algorithm, Random Forest algorithm, Naïve Bayes algorithm etc

INTRODUCTION

The 21st century Industries are generating more data that will grow faster. The organizations utilize this data to make important decisions. The industries that generate huge data are Hospitals, Educational Institutions and many other companies. Healthcare is one among the top that generates large amount of data. Here we apply Machine learning algorithms to maintain complete hospital data. Machine learning allows building models to quickly analyse data and deliver results, leveraging both past and real-time data. With machine learning techniques, doctors can make better decisions on patient’s diagnoses and treatment options, which leads to the improvement of healthcare services.

Healthcare is a prime example of how the three dimensions of Big data is used. First is

velocity, second is variety and third one is volume.

MOTIVATION

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data.

These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this research.

PROBLEM DEFINITION

Beforehand, medical team was trying for human services experts to gather and examine the enormous volume of information for powerful expectations and medicines. Since around then there were no advancements or apparatuses are accessible for them. Presently, with machine learning, we make it moderately simple. Huge information advancements, for example, Hadoop are all the sufficiently more for wide-scale selection. Indeed, 54% of associations are utilizing Hadoop as large information handling instrument to get data in human services. 94% of Hadoop clients perform investigation on voluminous information. Machine learning calculations can likewise be useful in giving essential insights, constant information and progressed examination as far as the patient's malady, lab test comes about, circulatory strain, family history, clinical trial information and more to specialists. Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited.

They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." 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. Integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome.

OBJECTIVE OF THE PRODUCT

The main objective of this research is to develop a prototype named Diseases Prediction using Machine Learning. This can discover and extract hidden knowledge

(patterns and relationships) associated with various disease from a historical disease and symptoms database. It can answer complex queries for diagnosing particular disease with comparing symptoms and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms. To tackle these issues, the organized and unstructured information is joined in social insurance field for successfully anticipating the sicknesses.

LITERATURE SURVEY

INTRODUCTION

Beforehand, medical team was trying for human services experts to gather and examine the enormous volume of information for powerful expectations and medicines. Since around then there were no advancements or apparatuses are accessible for them. Presently, with machine learning, we make it moderately simple. Huge information advancements, for example, Hadoop are all the sufficiently more for wide-scale selection. Indeed, 54% of associations are utilizing Hadoop as large information handling instrument to get data in human services. 94% of Hadoop clients perform investigation on voluminous information. Machine learning calculations can likewise be useful in giving essential insights, constant information and progressed examination as far as the patient's malady, lab test comes about, circulatory strain, family history, clinical trial information and more to specialists.

Human services framework creates expansive measure of information, the test is to gather this information and successfully utilize it for investigation, forecast and treatment. The principle way to deal with human services framework is to keep the sickness with early location instead of go for a treatment after conclusion.

Customarily, specialists utilize a hazard number cruncher to survey the likelihood of sickness advancement. These adding machines utilize central data like socioeconomics, medicinal conditions, life schedules and more to figure the likelihood of advancement of a specific malady. Such counts are finished utilizing condition based scientific techniques and devices. The issue with this technique is the low exactness rate with a comparable conditional based approach. Be that as it may, with late improvement in advancements, for example, huge information and machine taking in, it's conceivable to get more precise outcomes for illness expectation.

Doctors are collaborating with analysts and PC researchers to grow better instruments to foresee the sicknesses. Specialists in this field are chipping away at the procedures to recognize, create, and tweak machine learning calculations and models that can convey exact forecasts. To build up a solid and more precise machine learning model, we can utilize information gathered from considers, quiet socioeconomics, restorative wellbeing records, and different sources. The distinction between conventional approach and the machine learning approach for sickness forecast is the quantity of ward factors to consider. In a customary approach, not very many factors are viewed as, for example, age, weight, stature, sex, and then some. Then again, machine learning can think about countless, which brings about a superior precision of social insurance information. As indicated by a current report, the analyst got better symptomatic precision, utilizing whole medicinal records by considering around 200 factors.

Aside from illness expectation, there are the couple of more potential zones like medication revelation or electronic wellbeing records where machine learning can enhance social insurance industry. With machine learning applications, the human services and solution

fragment can progress into another domain and totally change social insurance tasks.

Subsequently the grouping of danger of infections in light of enormous information investigation has the accompanying difficulties: How should the missing information be tended to? In what capacity should the fundamental incessant maladies in a specific locale decided? By what means can huge information examination innovation be utilized to break down the infection and make a superior model? To tackle these issues, the organized and unstructured information is joined in social insurance field for successfully anticipating the sicknesses.

ALGORITHMS AND FLOWCHARTS

DECISION TREE ALGORITHM

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

In medicinal choice like arrangement, diagnosing there are numerous circumstances where choice must be made successfully and dependably. Reasonable straightforward basic leadership models with the likelihood of programmed learning are the most fitting for performing such undertakings. Choice trees are a solid and successful basic leadership procedure that furnish high grouping exactness with a straightforward portrayal of accumulated learning and they have been utilized as a part of various zones of restorative basic leadership.

The below are the some of the assumptions we make while using Decision tree:

- At the beginning, the whole training set is considered as the **root.**
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

DECISION TREE ALGORITHM PSEUDOCODE

1. Place the best attribute of the dataset at the **root** of the tree.
2. Split the training set into **subsets**. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find **leaf nodes** in all the branches of the tree.

ADVANTAGES

COMPREHENSIVE

A significant advantage of a decision tree is that it forces the consideration of all possible outcomes of a decision and traces each path to a conclusion. It creates a comprehensive analysis of the consequences along each branch and identifies decision nodes that need further analysis.

SPECIFIC

Decision trees assign specific values to each problem, decision path and outcome. Using monetary values makes costs and benefits explicit. This approach identifies the relevant decision paths, reduces uncertainty, clears up ambiguity and clarifies the financial consequences of various courses of action.

When factual information is not available, decision trees use probabilities for conditions to keep choices in perspective with each other for easy comparisons.

EASY TO USE

Decision trees are easy to use and explain with simple math, no complex formulas. They present visually all of the decision alternatives for quick comparisons in a format that is easy to understand with only brief explanations.

They are intuitive and follow the same pattern of thinking that humans use when making decisions.

VERSATILE

A multitude of business problems can be analysed and solved by decision trees. They are useful tools for business managers, technicians, engineers, medical staff and anyone else who has to make decisions under uncertain conditions. The algorithm of a decision tree can be integrated with other management analysis tools such as Net Present.

Value and Project Evaluation Review Technique (PERT). Simple decision trees can be manually constructed or used with computer programs for more complicated diagrams.

Decision trees are a common-sense technique to find the best solutions to problems with uncertainty. Should you take an umbrella to work today? To find out, construct a simple decision-tree diagram.

DISADVANTAGES

- They are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree.
- They are often relatively inaccurate.
- Many other predictors perform better with similar data.
- This can be remedied by replacing a single decision tree with a random forest of decision trees, but a random forest is not as easy to interpret as a single decision tree.
- For data including categorical variables with different number of levels, information gain in decision trees is biased in 11amour of those attributes with more levels.
- Calculations can get very complex, particularly if many values are uncertain and/or if many outcomes are linked.

COMMON TERMS USED WITH DECISION TREES

1. **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.

4. **Leaf/ Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
6. **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

EXAMPLE

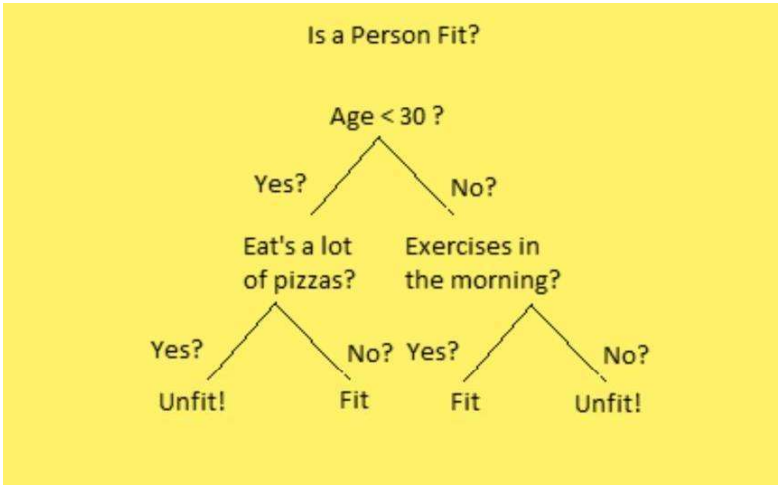

FIG 1

RANDOM FOREST ALGORITHM

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

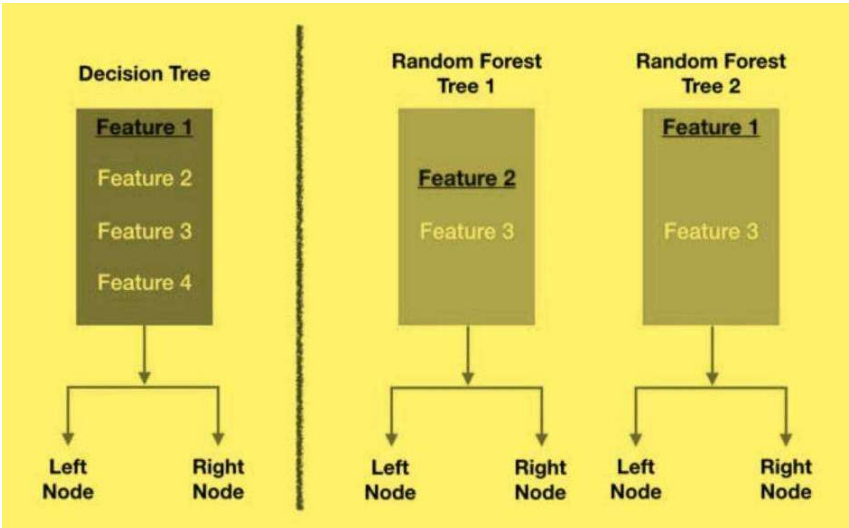

FIG 2

You start with a node which then branches to another node, repeating this process until you reach a leaf. A node asks a question in order to help classify the data. A branch represents the different possibilities that this node could lead to. A leaf is the end of a decision tree, or a node that no longer has any branches.

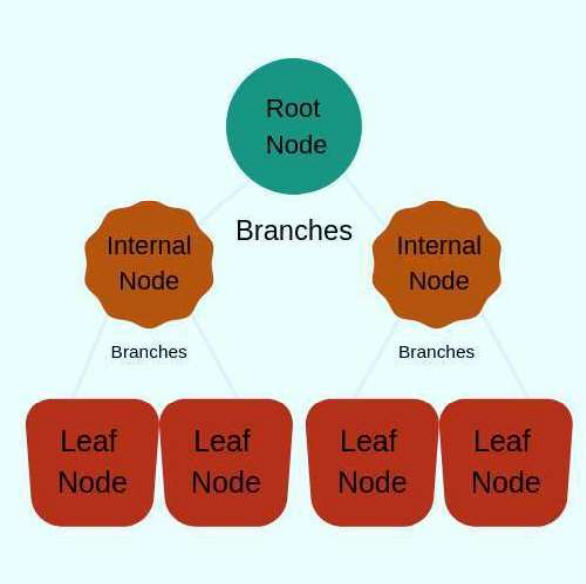

FIG 3

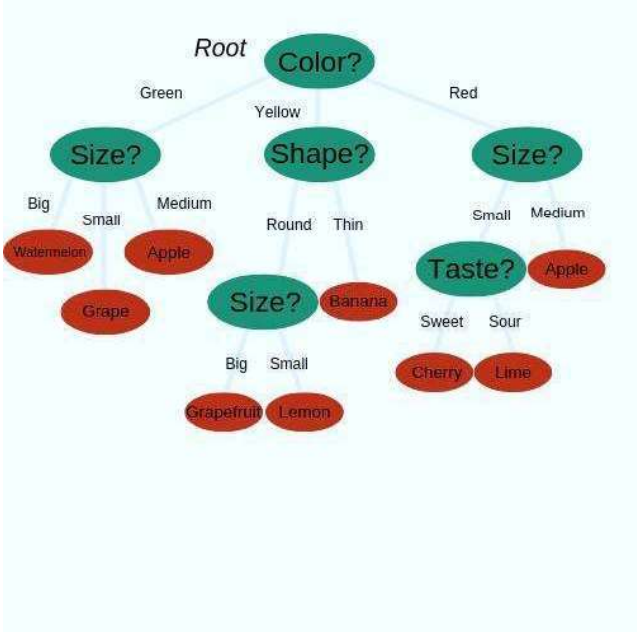

FIG.4.

The Random Forest Algorithm is composed of different decision trees, each with the same nodes, but using different data that leads to different leaves. It merges the decisions

of multiple decision trees in order to find an answer, which represents the average of all these decision trees.

The random forest algorithm is a supervised learning model; it uses labeled data to "learn" how to classify unlabeled data. This is the opposite of the K-means Cluster algorithm, which we learned in a past article was an unsupervised learning model. The Random Forest Algorithm is used to solve both regression and classification problems, making it a diverse model that is widely used by engineers.

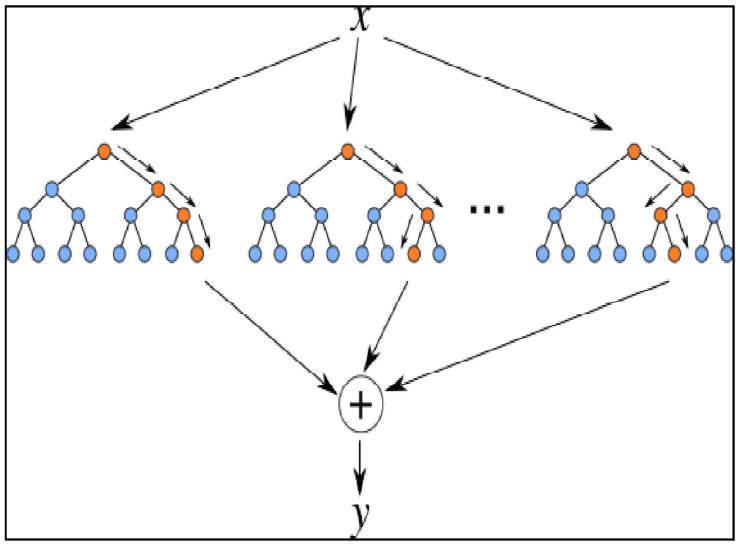

FIG.5

In the above example, we have three individual decision trees which together make up a Random Forest. Random Forest is considered ensemble learning, meaning it helps to create more accurate results by using multiple models to come to its conclusion. The algorithm uses the leaves, or final decisions, of each node to come to a conclusion of its own. This increases the accuracy of the model since it's looking at the results of many different decision trees and finding an average.

NAÏVE BAYES ALGORİTHM

The finding of a medicinal condition depends on the side effects, physical examination and restorative history of a patient. There have been numerous situations where treatment for an infection has not been done precisely because of absence of legitimate investigation and obliviousness of various imperative components.

To make the fundamentals of good wellbeing hones open to everybody, a Naïve Bayes approach for Disease Diagnosis is proposed here. This administration plans to help patients and specialists by directing them to anticipate conceivable illnesses utilizing the side effects gave by the client.

Guileless Bayes classifier is utilized for characterization of information. The framework contains a manifestation informational collection which is additionally sorted as Name, Attributes, and Record information. In light of the positioning of every side effect gave, the analyzed outcome recommends the likelihood weight of every illness. It additionally gives the detail of recognized sicknesses including the cure, counteractive action and the conceivable treatment. Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called *naive Bayes* or *idiot Bayes* because the calculation of the probabilities for

each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on. This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

DESIGN

INTRODUCTION

The analysis accuracy is reduced when the quality of meical data in incomplete. Moreover, different regions exhibit unique characteristics of certain regional diseases, which may weaken the prediction of disease outbreaks. However, those existing work mostly considered structured data. There are no proper methods to handle semi structured and unstructured. The proposed system will consider both structured and unstructured data. The analysis accuracy is increased by using Machine Learning algorithm and Map Reduce algorithm.

MODULE DESIGN AND DESCRIPTION

ARCHITECTURE DIAGRAM

An architecture is constructed with input data coming from the user/patient giving out his/her symptoms. Later the data sent is sent to a database and stored. Later the data is processed with the three algorithms named as Decision Tree Algorithm, Random Forest algorithm and Naïve Bayes Algorithm which are the machine learning algorithms that help in processing of data and helps in getting the output.

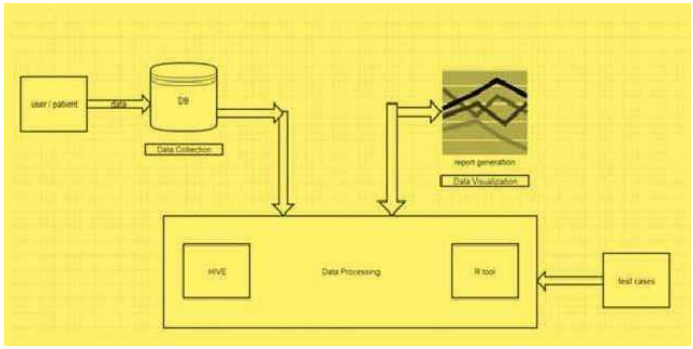

FIG 6

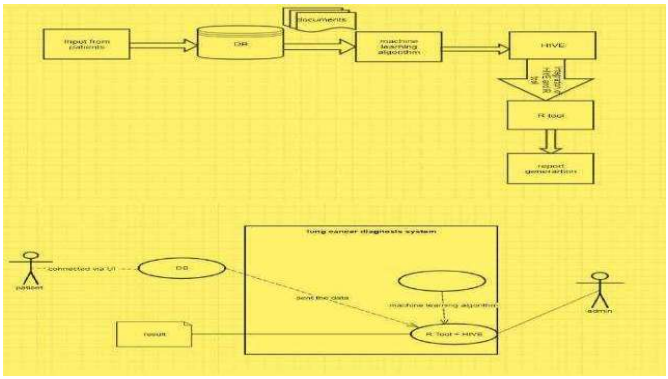

FIG.7

IMPLEMENTATION AND RESULTS

INTRODUCTION

Machine Learning (ML) delivers methodologies, approaches, and apparatuses that can help resolving analytic and predictive hitches in a miscellany of medicinal areas. ML is being used for the inquiry of the wild of controlled edges and their mixtures for forecast, e.g. forecast of illness development, removal of medicinal information for consequence investigation, treatment guidance and provision, and for the overall enduring organization. ML is also being used for statistics examination, such as discovery of proportions in the data by rightly commerce with flawed data, clarification of incessant data used in the Strenuous Care Unit, and brainy troubling subsequent in real and ordered nursing. It is contended that the successful presentation of ML attitudes can help the tally of computer-based structures in the healthcare setting providing chances to ease and enhance the exertion of medical boffins and eventually to recover the competence and excellence of medicinal repair. Below, it précises some main ML requests in medicine. Machine Learning learns the data and produces the result.

EXPLANATION OF KEY FUNCTIONS

The proposed system mainly contains one module i.e. User module

USER MODULE

In this module, user opens the web page and enters the required details such as values of 14 attributes and click the submit button. If any of the values are missing, a message appears as “missing values”. Once every value is entered, the prediction system in the back end predicts the result and produces the output to the user. There is a refresh button where the values can be refreshed by the user if required.

METHOD OF IMPLEMENTATION

DECISION TREE

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

```
def DecisionTree():

    from sklearn import tree

    clf3 = tree.DecisionTreeClassifier()   # empty model of the decision tree
    clf3 = clf3.fit(X,y)

    # calculating accuracy-------------------------------------------------------------------
    from sklearn.metrics import accuracy_score
    y_pred=clf3.predict(X_test)
    print(accuracy_score(y_test, y_pred))
    print(accuracy_score(y_test, y_pred,normalize=False))
    # -------------------------------------------------------

    psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Symptom5.get()]
```

FIG 8

RANDOM FOREST ALGORITHM

Random forests or random decision forests are an ensemble learning method for classification, regression **and** other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

```
def randomforest():
    from sklearn.ensemble import RandomForestClassifier
    clf4 = RandomForestClassifier()
    clf4 = clf4.fit(X,np.ravel(y))

    # calculating accuracy-------------------------------------------------------------------
    from sklearn.metrics import accuracy_score
    y_pred=clf4.predict(X_test)
    print(accuracy_score(y_test, y_pred))
    print(accuracy_score(y_test, y_pred,normalize=False))
    # ---------------------------------------------------

    psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Symptom5.get()]
```

FIG 9

NAÏVE BAYES ALGORITHM

The finding of a medicinal condition depends on the side effects, physical examination and restorative history of a patient. There have been numerous situations where treatment for an infection has not been done precisely because of absence of legitimate investigation and obliviousness of various imperative components. To make the fundamentals of good wellbeing hones open to everybody, a Naïve Bayes approach for Disease Diagnosis is proposed here. This administration plans to help patients and specialists by directing them to anticipate conceivable illnesses utilizing the side effects gave by the client. Guileless Bayes classifier is utilized for characterization of information. The framework contains a manifestation informational collection which is additionally sorted as Name, Attributes, and Record information. In light of the positioning of every side effect gave, the analyzed outcome recommends the likelihood weight of every illness. It additionally gives the detail of recognized sicknesses including the cure, counteractive action and the conceivable treatment.

CONCLUSION

In this paper, we propose a machine learning Decision Tree, Random Forest & Naïve Bayes algorithms using structured and unstructured data from hospital for effective prediction of diseases. Existing work is not focused on both data types in the area of healthcare. Compared to several typical prediction algorithms, the proposed algorithm accuracy prediction reaches 94.8% In this paper, it bid a Machine learning Decision tree map algorithm by using structured and unstructured data from hospital. It also uses Map Reduce algorithm for partitioning the data. To the highest of gen, none of the current work attentive on together data types in the zone of remedial big data analytics. The report consists of possibility of occurrences of diseases.

FUTURE SCOPE

Medical related information's are huge in nature and it can be derived from different birthplaces which are not entirely applicable in feature. In this work, heart disease prediction system was developed using clustering and classification algorithms to predict the effective risk level and accuracy of the patients. In future work, we have planned to propose an effective disease prediction system to predict the disease with better accuracy using different data mining techniques and compare the performance of algorithm with other related data mining algorithms.

REFERENCES

1] Anand Borad, Healthcare and Machine Learning: The Future with Possibilities Jan. 18.
2] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system, "IEEE Commun., vol. 55, no. 1, pp. 54_61, Jan. 2017.
3] P. Groves, B. Kayyali, D. Knott, and S. van Kuiken, The Big Data Revolution in Healthcare: Accelerating Value and Innovation. USA: Center for US Health System Reform Business Technology Office, 2016.
4] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction, "Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070_1093, 2015.
5] S. Bandyopadhyayet al., "Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data", Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1033_1069, 2015
6] Prediction of probability of disease based on Symptoms using machine Learning Algorithm by Harini DK and Natesh M