A Critical Analysis of Machine Learning Techniques for Human Disease Prediction Model

Dr. Hemamalini G E[1] Dr. J Prakash [2]

Vemana Institute of Technology,Koramangala,Bangalore-34,9886645402
Bangalore Institute of Technology, V.VPura,Bangalore-1,9845070991

ABSTRACT

Various diseases affect people today because of environmental factors and their lifestyles. Therefore, predicting disease at an early stage becomes crucial. It is difficult for doctors to make an accurate diagnosis based on symptoms alone. It is very challenging to accurately predict disease. Data mining plays a critical role in disease prediction in order to solve this challenge. Each year, there is a tremendous amount of data expansion in the world of medicine. A growing amount of medical and healthcare data has resulted in the ability to analyze medical data in a way that benefits early patient care. A data mining technique uses disease data to find hidden patterns in vast amounts of medical information. The purpose of our proposal was to provide general disease predictions based on symptoms of the patient. In order to predict disease accurately, (DECISION TREE AND NAIVE BAYES) machine learning algorithm is used. It is necessary to collect data on disease symptoms for disease prediction. For the reliability of the general illness diagnosis, a person's lifestyle and check-up facts are taken into account. It is possible to forecast general disease with an accuracy of 84.5% using Naive Bayes and Decision Trees.

Index Terms: Disease Prediction, Machine Learning, Decision Tree, Naive Bayes, Data Mining.

1. INTRODUCTION

For decades, using machine learning and data mining methods to forecast disease has been a difficult task. Machine learning models are more likely to be able to learn hierarchical representations of raw data without any pre-processing, resulting in more accurate outputs, thanks to the recent development of deep learning. More and more research has been conducted on disease prediction using big data technology; various studies have utilized automatic selection of risk factors from large amounts of data rather than conventional selection methods. In healthcare, machine learning is used primarily to supplement patient care for better results. A predictive analysis that employs various machine learning methods improves disease accuracy rate and can aid in more effective patient treatment. The healthcare business generates vast amounts of medical information on a daily basis based on a patient's treatment history and health data, which can be utilized to anticipate future illness that may affect a person.

In the healthcare field, there is much data regarding patient evaluation, cure, follow-ups and medication. It's complicated to orchestrate appropriately. An improper management of information affected the quality of data association. To construct such classifiers, machine learning applications are used to analyse the data and sort it based on its characteristics. DT classifies hospitals using data from both structured and unstructured sources. Other machine learning techniques work solely with structured information and require a high level of processing, but they are inefficient because the full dataset is utilised as a data set, and intricate calculation techniques are used.

For the current prediction system, training data are used with labels for supervised and machine-learning algorithms. High-risk and Low-risk patient classification is done in groups test sets. This approach is only useful in clinical situations and has been extensively studied. (Chen et al. 2017) created a system that employs smart clothes to collect data on patient statistics, outcomes, and medical history, as well as to find available information approaches to lower the cost of healthcare research papers. Six implications of big information in the healthcare industry were proposed by (Bates et al. 2014). In existing systems, diseases can be predicted, but not their subtype. This model is incapable of predicting a person's condition. There has been no specificity or certainty in disease predictions.

In order to assess risk of disease, the proposed concept combines structured and unstructured data in healthcare fields. A data mining technique uses disease data to find hidden patterns in vast amounts of medical information. A method based on patient symptoms was proposed to predict general diseases. In order to predict disease accurately, (DECISION TREE AND NAIVE BAYES) machine learning algorithm is used. To predict disease symptoms, disease symptoms datasets are needed. A person's lifestyle and check-up information are taken into consideration for the accuracy of the general disease prediction. Reconstruction of missing data in medical records collected from hospitals using a latent factor model. Statistical knowledge can also be used to identify the

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

major chronic diseases affecting a particular region or community. For dealing with structured data, hospital experts provide useful insights.

2. LITERATURE SURVEY

The study (Penikalapati et al., 2020) looked at how to use feature choices and Machine Learning techniques like Categorization and Grouping in the domain of disease early prediction and identification. This study analyses the best machine learning models to support their performance measures, utility features, restrictions, and crucial problems in the perspective of effective and efficient use in medical analytics. Healthcare EHRs are increasingly being employed in a seamless manner by combining supervised and unsupervised learning models for efficient and effective use of Machine Learning analytics. The methodologies of classification and clustering are used to improve the quality of healthcare in the specific setting of early prediction and diagnosis of chronic diseases. Patients can save time and money on chronic conditions including diabetes, stress, cancer, and cardiovascular ailments by doing so.

To deal with the issue of missing medical data, (Ambekar et al. 2018) used data cleaning and imputation to convert incomplete information into full information. Data analysis is critical in the healthcare industry when dealing with enormous amounts of data. Rather than creating predictions, previous medical studies focused on managing and absorbing a significant amount of medical records. The biomedical and healthcare industries presently create massive amounts of data, necessitating effective data analysis in order to diagnose diseases early and improve patient care. When medical information is absent in part, however, accuracy suffers. By employing Naive Bayes and KNN algorithms, the researchers developed a prediction method for heart disease based on the dataset. They proposed incorporating structured data into this work to predict disease risk. An algorithm based on convolutional neural networks is used to predict disease risk. Over 65% accuracy is achieved by CNN-UDRP. Furthermore, this system is capable of answering questions related to diseases that people experience during their lifetime.

(Jadhav et al. 2019) analyzed disease prognosis and prediction in healthcare. The progress of biomedical and healthcare records allows accurate review of clinical information to assist early disorder identification, patient care, and network services. Because medical information is incomplete, analysis is inexact. Also, certain regions exhibit specific symptoms of specific diseases, making it difficult to predict epidemic outbreaks in such areas. This system provides knowledge acquisition algorithms to effectively find out disorders occurrences in disorder-prone societies and predict the waiting time for each patient's treatment. A hospital Queueing Suggestion (HQR) system is used to offer treatment mission sequences based on expected wait times. The study uses both structured and unstructured data to examine facts from a health centre regarding a nearby chronic ailment of cerebral infarction. It investigates the disease using both Tree and KNN methods. Currently, none of the existing research in the field of medical big data analytics has focused on each data type. The suggested set of regulations has a computation accuracy of 94.8%, and its convergence rate is quicker than the CNN-based method for completely unimodal sickness danger detection (CNN-UDRP).

(Shivaganga et al., 2019) examined the impact of data mining on diagnosing disease using hospital data. By analyzing the medical data in a variety of ways, such as multidimensional ways and view based, it is able to escape the hard risks and then predict accurately. In hospitals, the data is categorized in two ways: (I) structured data and (II) unstructured data. Based on both big data analytics and predictions in the medical field, the concept fulfils the existing systems. Researchers from diverse fields are continuously working toward developing Achieving Disease Prediction. The purpose of this survey was to summarize the current research and their advantages and disadvantages in relation to disease prediction. In this paper, the merits and disadvantages of the recent techniques are discussed. It is concluded that there have not been any effective methods discovered to achieve Disease Prediction.

Increasing numbers of biomedical and healthcare organizations are employing big data to detect diseases, treat patients, and improve community services. An incomplete set of medical data, however, reduces analysis accuracy. Furthermore, certain regional diseases exhibit unique characteristics in different regions, weakening predictions about outbreaks. Chen et al. (2017) created simplified computational methods for forecasting chronic illness epidemics in disease-prone areas. The researchers tested the updated prediction models using real-world hospital data obtained between 2013 and 2015. In order to overcome the problem of incomplete data, a latent variable model is utilised to recreate incomplete information. A regional chronic infarction of the cerebral cortex was the subject of the study. A CNN-based heterogeneous illness risk prediction system is developed using unstructured and structured data from hospitals. There are currently no studies in the field of medical big data analytics that include both forms of data. Our suggested approach outperforms the CNN-based unimodal illness risk predictive model in terms of convergence efficiency and reliability when compared to many standard predictive models.

According to (Karthika et al. 2017), reliable health information analysis benefits patients, healthcare professionals, and the community by detecting early ailments and resolving problems before they develop.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

However, incomplete medical data may lead to reduced analysis accuracy. In addition, certain diseases have unique characteristics between regions, which can make predicting outbreaks of these diseases more difficult. Using machine learning algorithms, this article rationalizes the onset of chronic diseases in communities with high prevalence of disease. A modified predictive model is being tested with real hospital data. In order to solve the issue of incomplete data, a latent variable model is utilised to recreate missing data. The disease of localised persistent cerebral infarction is being investigated. The use of hospital unstructured and structured data in a neuronal network-based multimodal disease risk prediction system (CNN-MDRP) is presented. None of the previous research in the field of advanced processing for medical applications has addressed both types of data. When compared to the CNN-based unimodal illness risk prediction model, our suggested approach has a predictive performance of over 95%.

(Shinde et al., 2017) analyzed available big data that are connected to health care and organized them in a way that can be used to derive significant patterns and facts, and proposed a health informatics system. The authors used Hadoop and MapReduce as an environment for big data analysis to predict the symptoms in early stages, provide research-oriented treatments, and develop schemes of care in order to minimize the risk of different diseases in our society. A variety of predictive models are discussed in this paper. Data related to healthcare are the focus of the proposed work. In terms of treatment plans, health policies, reducing treatment costs, etc., implementing the proposed system can provide several important patterns. By using mobile phones, emails, social media, and other digital mediums, all health information will be provided to society on a daily basis.

3. PROPOSED METHODOLOGY:

3.1. DECISION TREE

Each internal node in a decision tree indicates a test (for example, whether a coin flip will result in tail or head), each component reflects the test's value, and each leaf node represents the result . A path from root to leaf can be viewed as a categorization rule. For illustrating and analysing the expected utility (or value) of competing choices, decision trees and the strongly linked impact diagram have traditionally been utilised. In a decision tree, there are 3 sorts of nodes: Squares are commonly used to symbolise decision nodes., Chance nodes are usually depicted as circles and End nodes - triangles are commonly used to depict them.

Operational research and management frequently use decision trees. The probabilistic model as an online selection model algorithm should be used, in practice, if decisions must be taken online without recalling results under incomplete knowledge. Determining conditional probabilities is another use for decision trees.

3.1.1 DECISION TREE ELEMENTS

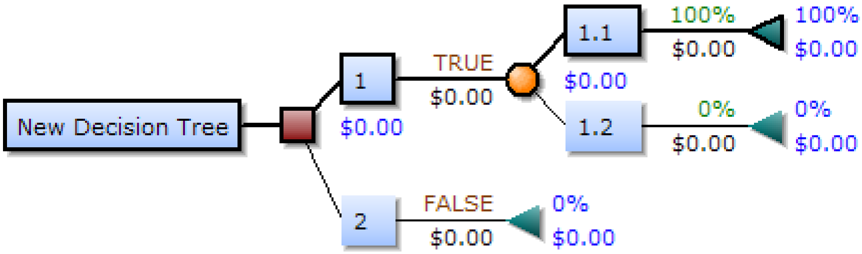


Figure 1: Elements of New Decision Tree

When looking at a decision tree from the left, only outburst nodes and no sink nodes are shown. It can also become exceedingly enormous when sketched by hand, making them difficult to complete. Traditional decision tree techniques were made by hand - as seen in Figure 1 - but specialised software is now being used more frequently.

3.1.2 DECISION RULES

It is possible to linearize the decision tree into decision rules, with the leaf node representing the outcome, and the conditions along the path forming the 'if clause'. Generally, the rules look like this:

if condition1 and condition2 and condition3 then outcome.

The decision rules are generated using the target variable on the right to construct association rules. Additionally, they can indicate a causal or temporal relationship. Decision trees are decision-making aids that employ a tree-like framework of decision and its outcomes, such as random event results and resource constraints. It's a way of illustrating a method made up entirely of dependent control declarations. Using decision trees to determine an approach that is able to obtain an objective is also a common machine learning approach, especially in operations research.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

3.1.3 ADVANTAGES OF DECISION TREES:

Decision trees (and influence diagrams) have a number of benefits over conventional decision-making aids. Decision trees:

- Are easy to follow and understand. The decision tree model can be easily understood after a brief explanation.
- They are valuable even in the absence of hard evidence. Analyzing how experts define a situation (its possibilities, probabilities, and costs) as well as their predictions of the result can provide significant insights.
- For a number of scenarios, provide the best, worst, and expected outcomes.
- When a model provides a result, use a white box design; • Combine with other decision-making strategies.

3.1.4 DISADVANTAGES OF DECISION TREES:

- Even a minor change in the data might cause significant structural changes of an ideal decision tree.
- They have a proclivity for being somewhat imprecise. Other predictors can be tested using similar data. The difficulty can be solved by replacing one decision tree with a Naive Bayes of decision trees; however, random forests are more difficult to read than single decision trees.
- For data containing categorical factors with several levels, data gain is biased in favour of qualities with more stages in decision trees.
- Calculations can become very complex, particularly when a great deal of data is uncertain or if great deals of outputs are linked.

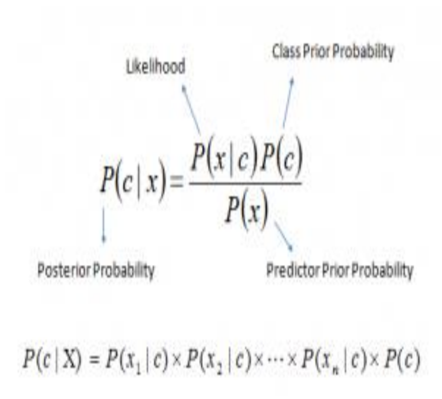
3.2. NAIVE BAYES ALGORITHM:

- The Bayesian theory that underpins it is dependent on the premise that predictors are independent. The assumption of Naive Bayes is that the presence of one characteristic in a class has no bearing on the existence of any other characteristic.
- An apple is a red, round fruit that is smaller than three inches in diameter and takes up less than three inches of area. Regardless of whether these qualities are interdependent or independent of one another, they all contribute to the likelihood that this fruit is an apple, which is why it is dubbed 'Naive.'
- Especially useful for very large datasets, the Naive Bayes model is easy to construct.

  The simplicity of Naive Bayes makes it one of the best classification methods, along with having superior performance.

3.2.1 NAIVE BAYES THEOREM:

The Bayes theorem allows you to calculate posterior possibility P(c|x) from P(x|c), P(c), and P(x). Consider the following equation:



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Above,

• The posterior probability of class (c, target) given predictor is P(c|x) (x, attributes).

• P(c) is the class prior probability.

• The likelihood is P(x|c), which is the probability of a predictor given a class.

• P(x) is the predictor's prior probability.

3.2.2 WORKING OF NAIVE BAYES ALGORITHM:

Consider the following scenario. The weather training set and accompanying goal factor 'Play' are shown below. Players must now be classified based on the weather conditions. Follow these procedures to accomplish this.

Step 1: From the data collection, make a frequency distribution table.

Step 2: Create a likelihood table by calculating probabilities such as Overcast probability of 0.30 and Playing Probability of 0.56.

Step 3: Using the Naive Bayesian model, calculate the bayesian probability of each group. The class with the highest probability distribution is used to make predictions.

Problem: If the climate is sunny, players will play. Is this a true statement?

The probability distribution approach, as discussed above, can be used to address this problem.

P(Yes | Sunny) = P(Sunny | Yes) * P(Yes) / P (Sunny)

Here we have P (Sunny |Yes) = 3/9 = 0.33, P(Sunny) = 5/14 = 0.36, P(Yes)= 9/14 = 0.64

Now, P (Yes | Sunny) = 0.33 * 0.64 / 0.36 = 0.60, which has higher probability.

Naive Bayes calculates the likelihood of classifying things into different groups based on a range of attributes. This approach is most commonly used to classify text and problems with numerous classes.

3.2.3 ADVANTAGES OF NAIVE BAYES ALGORITHM:

- Testing data sets can be predicted easily and quickly. Multi class prediction is also performed well by it
- Naive Bayes classifiers perform better than logistic regression models whenever the assumption of independence holds, and they use fewer training data.
- When using categorical input variables, it performs better than using numerical variables(s). It is assumed that numerical variables have a normal distribution.

3.2.4 DISADVANTAGES OF NAIVE BAYES ALGORITHM:

- When categorical factors in the test information set have a category that does not exist in the training data set, the system will allocate 0 (zero) value, indicating that it is unable to make predictions. The term "Zero Frequency" refers to this phenomenon. This problem can be solved by using smoothing. Laplace estimation is among the simplest smoothing methods.
- Furthermore, because naive Bayes is a lousy estimator, the outputs of predict_proba should not be taken too seriously.
- Another disadvantage of Naive Bayes is the premise of separate predictors. In practise, obtaining a group of independent predictors is nearly impossible.

4. COMPARISON OF NAIVE BAYES AND DECISION TREE:

Table 1 illustrates the comparison of Naive Bayes and Decision Tree Techniques:

Table 1: Evaluation of Naive Bayes and Decision Tree (Jadhav, et al.) (Patankar, et al.)

Sl.No	Parameter	Decision Tree	Naïve Bayesian Algorithm
1	**Effectiveness**	Large data	Huge data
2	**Determination**	Deterministic	Non-Deterministic
3	**Dataset**	Deal with noisy data	Deal with noisy data
4	**Speed**	Faster	Fast
5	**Application**	Pattern, Sequence, and Financial Recognition	Text Classification, Spam Filtering
6	**Accuracy**	High accuracy	Requires a very large amount of records to obtain good results
7	**Understandable**	Simple to understand and generate	Simple to understand and build
8	**Data Types**	Numerical and categorical	Numerical and categorical

5. REQUIREMENT ANALYSIS:

5.1 SYSTEM ARCHITECTURE:

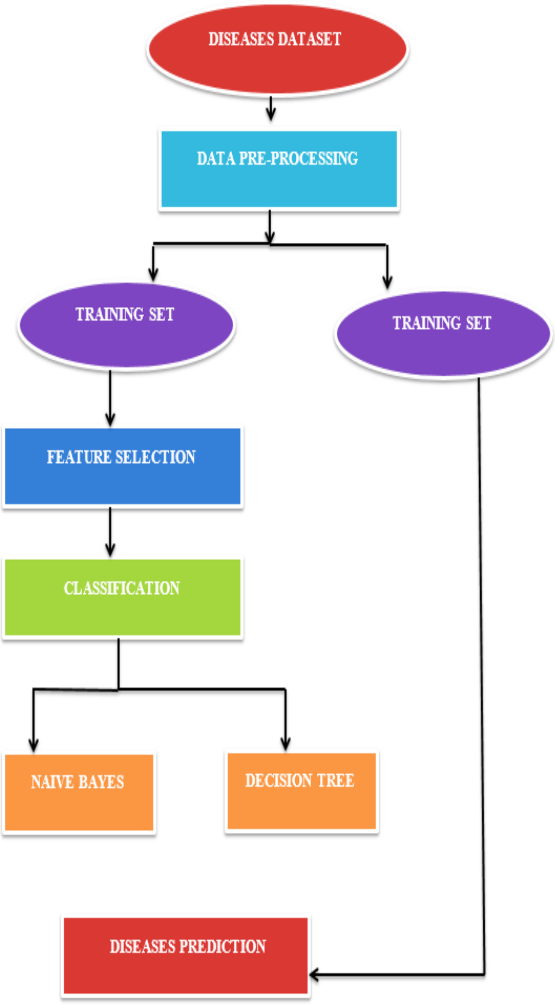


Figure 2. Illustration of Design Architecture (Dahiwade et al.)

The data sets are a collection of diseases and associated signs retrieved from the UCI machine learning website. After that, pre-processing was done on the data to remove commas, punctuations, and whitespace. This is used as a training dataset. This is followed by extracting and selecting the feature. In the next step, this data is classified using KNN and CNN. Machine learning enables accurate disease prediction as shown in figure 2.

5.2 DATA FLOW DIAGRAM:

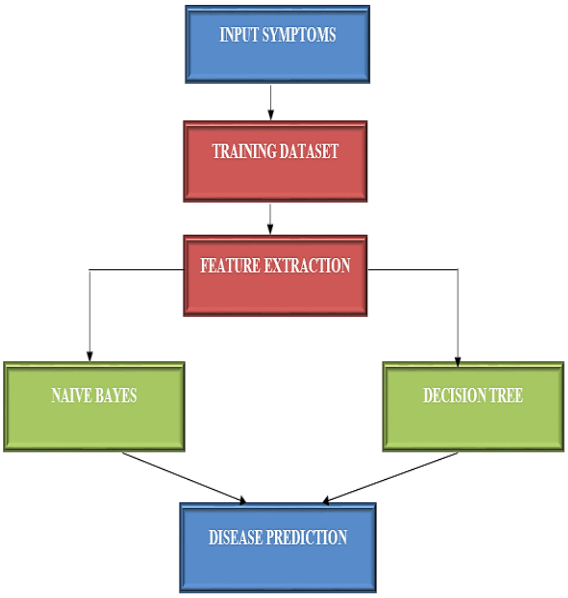


Figure 3. Illustration of Data Flow Diagram in disease prediction

Data flow diagram of disease prediction using machine learning contains all the elements of a regular flow chart. This data flow diagram shows how from the start the model flows from one step to another like symptoms of a patient goes into the system, and compares with the prediction model, Naive Bayes and Decision Tree. At last, it predicts the appropriate disease as in figure 3.

5.3 ACTIVITY DIAGRAM:

A dynamic aspect of a system is well illustrated by an activity diagram. Basically, it's a flowchart that represents the flow from one activity to another. As an operation, this activity constitutes an activity of the system. An operation follows another through the control flow. Here in the diagram, the activity starts with entering the user's symptoms, and then the user proceeds to predict the outcome. After processing the datasets, the analysis will be carried out using Naive Bayes and Decision Trees. A prediction is finally made for the correct disease as in figure 4.

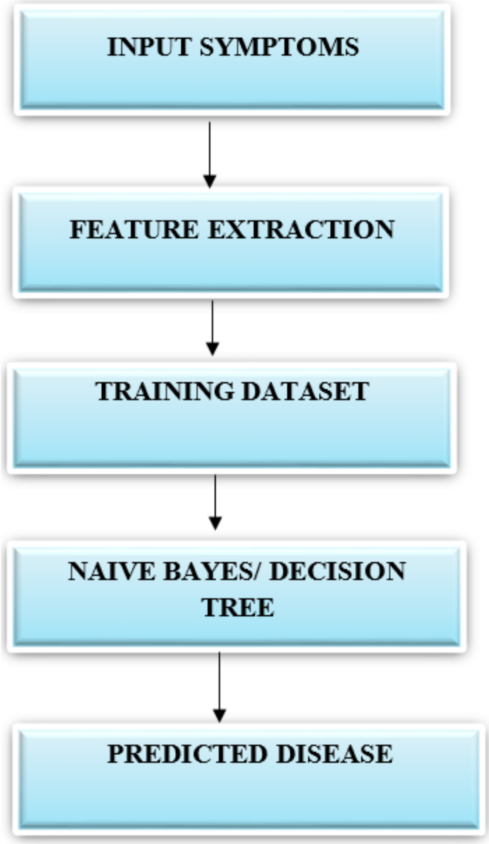


Figure 4: Activity Diagram to Predict Disease

5.4 USE CASE DIAGRAM:

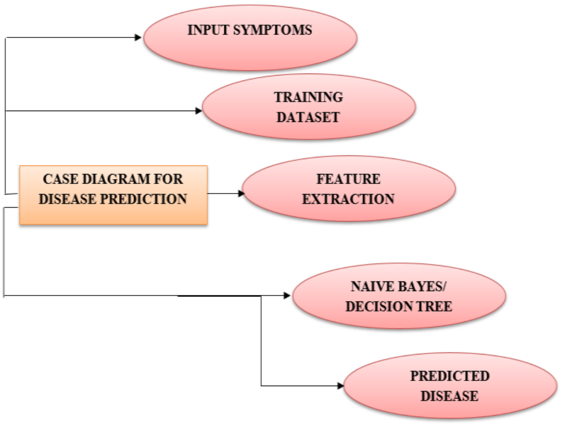


Figure 5: Use case diagram for disease prediction

A typical use case diagram contains all the various elements of a disease prediction project using machine learning. A patient enters all the symptoms and then compares the prediction model, which shows how from the start the model flows. The model predicts the right results. The application shows the user a precautionary measure to take if something goes wrong. Each entity is linked to one another in the use case diagram where the user is introduced to the system as in figure 5.

5.5 ENTITY - RELATIONSHIP DIAGRAM:

An entity relationship diagram is a basic tool for representing collections of data points and their relationships. In addition to being easy to understand, ER is a powerful tool to model real-world problems and is easily applied to database schema. Figure 6 illustrates the features of the ER model based on its typical semantic constructs:

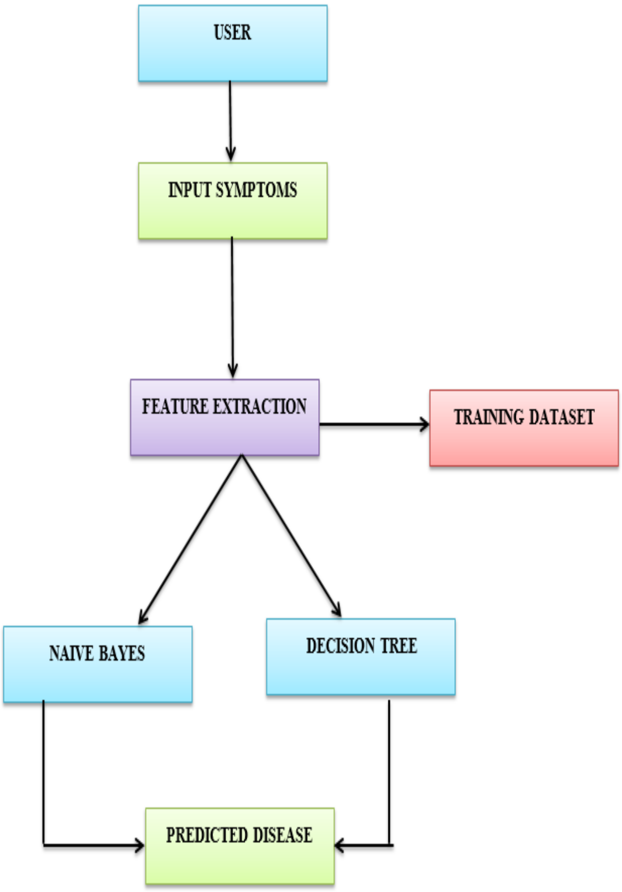


Figure 6: ER diagram for disease prediction

6. SOFTWARE SPECIFICATION:

6.1 SYSTEM SPECIFICATION

The proposed model is done with the following software requirements using the Windows operating system, with a python coding language. For the front end purpose Jupyter Notebook is used and for the back end Anaconda version 3.6 software have been used. The needed hardware requirements are the Intel Pentium Dual Core processor, with the speed of 2.2 GHz, the Ram used for this purpose was a 4 GB RAM. The hard disk was of 200 GB space.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

6.2 PYTHON:

Python has often been described as a "batteries included" language. As a programming language, it has consistently been ranked as one of the most popular. Python provides developers with a choice in coding methodology while pursuing a simpler syntax and grammar. Among Python programmers, there is a term called pythonic that refers to a wide range of styles of code. Pythonic code will use Python idioms well, will be naturally written or show fluency in the language, will be minimalist and will be easy to read. Unpythonic code, on the other hand, is incomprehensible or reads like a scribble in another computing paradigm.

6.3 ANACONDA 3.6:

Anaconda is a package manager and deployment system for programming languages Python and R are two popular programming languages for scientific computing. It includes packages for Linux, Windows, and macOS. By utilizing Anaconda Cloud, users can manage packages, notebooks, environments, and conda and PyPI packages. There are many environments, Python packages and notebooks on Cloud that can be used for many applications. The public packages can be downloaded and installed without requiring a Cloud account or logging in.

7. SYSTEM TESTING AND MAINTENANCE:

Tests are primarily aimed at identifying system errors. Data input is essential to the uncovering process. Our input data should therefore be more conscientious. The correct inputs are crucial to efficient testing. Modules are tested individually as in figure 7. The test data is specifically developed to verify that the system will perform properly in all dimensions after all modules have been tested. The modules are then combined and the finished system is tested. Therefore, system testing is a means of making sure all is in order and demonstrating to the customer that the system works. It is possible to run into errors months after inadequate or non-testing.

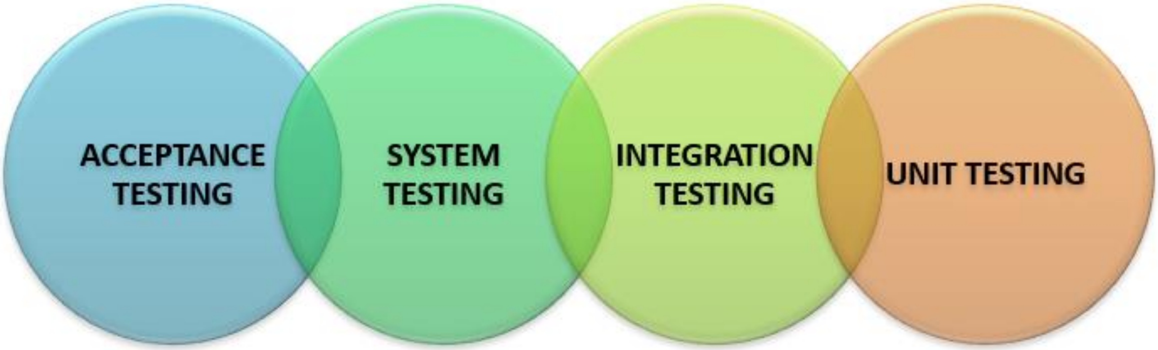


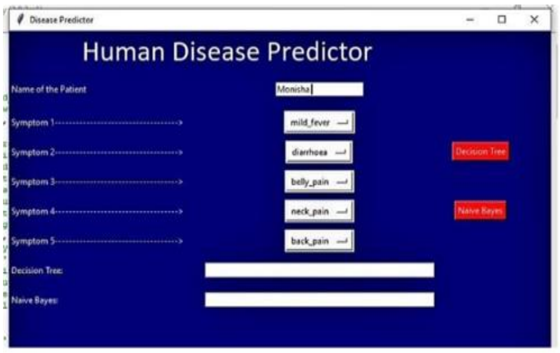
Figure 7: Illustration of System Testing

As part of software testing, integration testing combines individual units and tests them as a whole. During this level, faults that occur during the interaction between integrated units will be exposed. To assist with integration testing, test drivers and test stubs are used. Unit tests are conducted on the smallest unit of design, the module. This process is called "Module Testing". Separate tests are run on each module. Tests are conducted as part of the programming phase. These testing steps verify that each module produces the expected output in accordance with its design. Any project, regardless of size, requires extensive user participation during user acceptance testing. The system must also meet the functional requirements.

7.1 VALIDATION OF DATA SYNCHRONIZATION:

- Upon receiving the packets from the Destination Node, the Sender Node will receive the Acknowledgments.
- Routes are only added when a Route request is needed.
- In the process of updating the cache, information about the status of nodes is automatically collected.

8. RESULTS:



Human Disease Predictor

Name of the Patient Monisha

Symptom 1-----------------------------------> mild_fever

Symptom 2-----------------------------------> diarrhoea

Symptom 3-----------------------------------> belly_pain

Symptom 4-----------------------------------> neck_pain

Symptom 5-----------------------------------> back_pain

Decision Tree:

Naive Bayes:

Figure 8. Result of a user who entered 5 symptoms (Grampurohit et al.)

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

As in figure 8, users can list up to five symptoms they are experiencing. It is necessary to choose algorithms based on the symptoms. Following the selection of the algorithms, the symptoms will be processed, and the disease will be determined by the set of rules defined in the Methodology section.

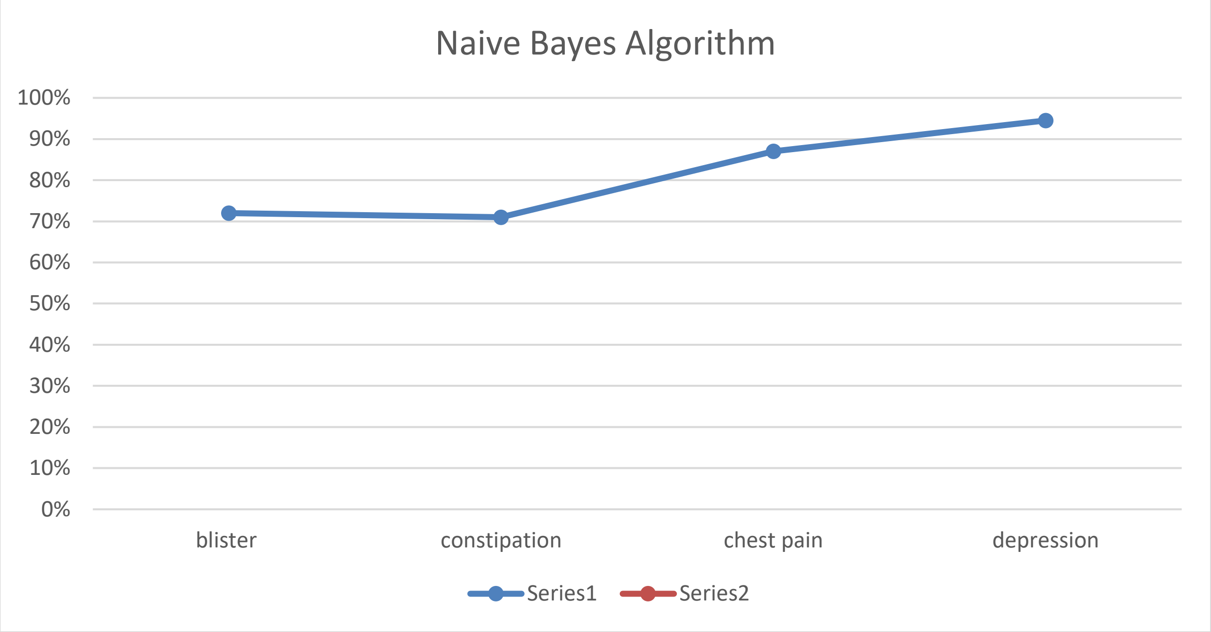


Figure.9 Naïve Bayes for different disease prediction.

The above figure 9 shows the Naïve Bayes for different disease prediction. For Patients XYZ reported the following symptoms: blister, bruising, constipation, chest pain, depression.

The predictions will be Decision Tree – Heart attack and Naïve Bayes - Heart attack. Therefore, the doctor considers that the patient suffers from Heart attack.

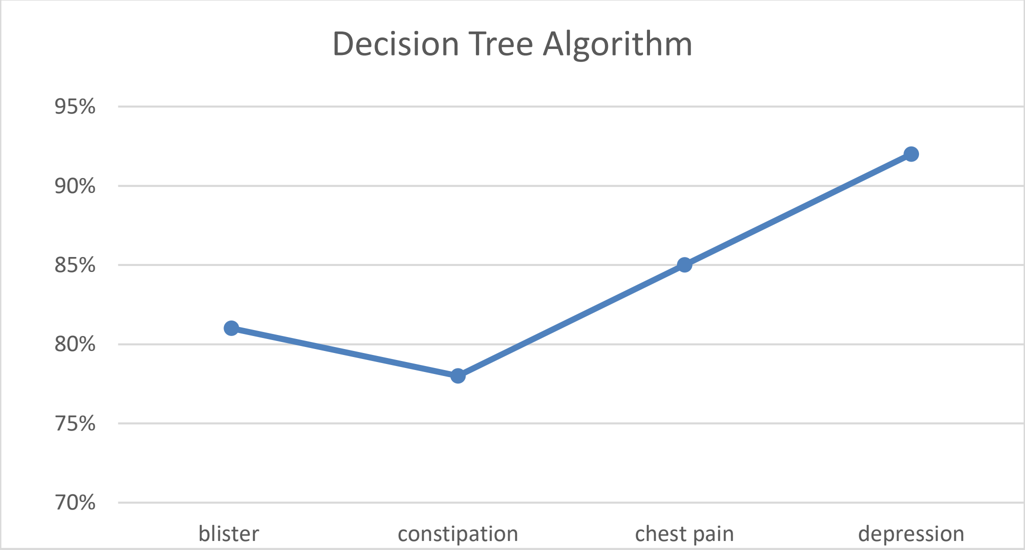


Figure.10 DT for different disease prediction

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

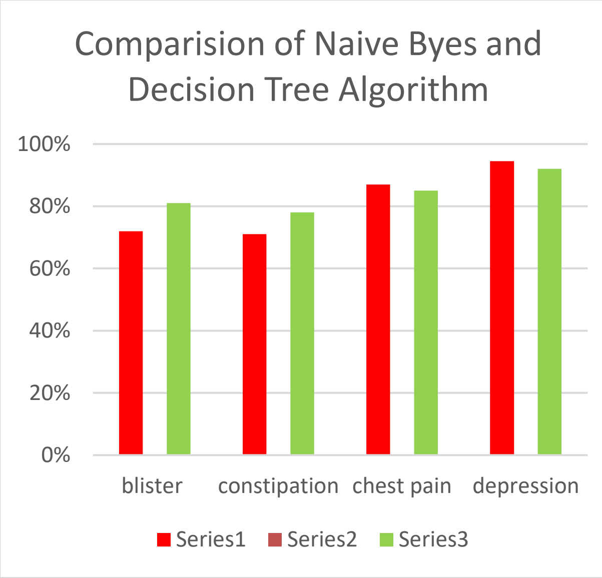


Figure.11 Naïve Bayes and Decision Tree for different disease prediction

Likewise Naïve Bayes DT is also capable of predicting many types of disease. The below Table 2 illustrates the accuracy rate by the proposed algorithm with a few selected disease.

Table 2. Accuracy rate of Decision Tree and Naïve Bayes for certain diseases

DISEASE	ALGORITHM	ACCURACY
Cardiac related	Decision Tree (Purusothaman, et al.)	76%
	Naïve Bayes (Purusothaman, et al.)	69%
Diabetic related	Decision Tree (Sisodia, et al.)	73.82%
	Naïve Bayes (Sisodia, et al.)	76.30%
Cancer related	Decision Tree (Sumbaly, et al.)	94.5%
	Naïve Bayes (Chaurasia, et al.)	97.36%

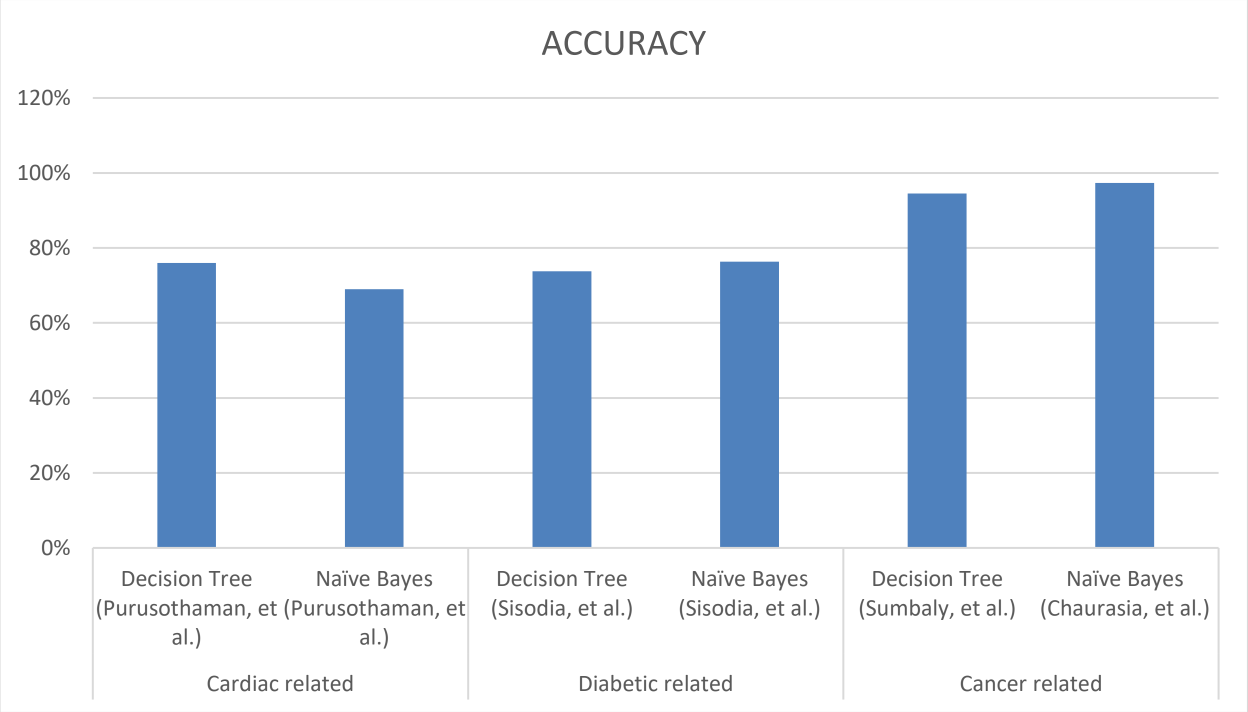


Figure.11. Accuracy rate of Decision Tree and Naïve Bayes for certain diseases

9. CONCLUSION:

It is possible to achieve higher accuracy through the proposed system. The k-mean technique he proposed takes into account both organised and unstructured data provided by the patient. This can be found out by combining both data sets, and the accuracy rate can be as high as 95%. In the realm of medical big data analytics, there is no existing system or work that uses both forms of data. In order to create a disease risk model, structured and unstructured features are combined.

REFERENCES:

1. Ambekar, Sayali, and Rashmi Phalnikar, 2018, "Disease risk prediction by using convolutional neural network." In 2018 Fourth international conference on computing communication control and automation (ICCUBEA), pp. 1-5. IEEE.
2. Amin, S.U.; Agarwal, K.; Beg, R., 2013, “Genetic neural network based data mining in prediction of heart disease using risk factors”, IEEE Conference on Information & Communication Technologies (ICT), vol., no.,pp.1227-31.
3. Bates, David W., Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar, 2014, ‘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. doi: 10.1377/hlthaff.2014.0041
4. Chaurasia, Vikas, Saurabh Pal, and B. B. Tiwari. "Prediction of benign and malignant breast cancer using data mining techniques." Journal of Algorithms & Computational Technology vol. 12, no. 2 (2018): 119-126.
5. Chen, Min, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang, 2017, "Disease prediction by machine learning over big data from healthcare communities." IEEE Access, vol.5, pp. 8869-8879.
6. Dahiwade, Dhiraj, Gajanan Patle, and Ektaa Meshram, 2019, "Designing disease prediction model using machine learning approach." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215. IEEE, DOI: 10.1109/ICCMC.2019.8819782
7. Disease and symptoms Dataset – www.github.com.
8. G. Purusothaman and P. Krishnakumari, “A Survey of Data Mining Techniques on Risk Prediction: Heart Disease”,Indian Journal of Science and Technology, vol. 8, no. 12, (2015).
9. Grampurohit, S., & Sagarnal, C. 2020. ‘Disease Prediction using Machine Learning Algorithms.’ 2020 International Conference for Emerging Technology (INCET). doi:10.1109/incet49848.2020.9154130
10. Heart disease Dataset-WWW.UCI Repository. com
11. IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, 2017, “ Wearable 2.0: Enable human-cloud integration in next generation healthcare system,” IEEE Commun. , vol. 55, no. 1, pp. 54–61.
12. Jadhav, Sayali D., and H. P. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques." International Journal of Science and Research (IJSR) vol 5, no. 1 (2016): 1842-1845.
13. Kunjir, Ajinkya, Harshal Sawant, Nuzhat F.Shaikh, 2017, “Data Mining and Visualization for prediction of Multiple Diseases in Healthcare,” in IEEE big data analytics and computational intelligence, pp. 2325.
14. L. Qiu, K. Gai, and M. Qiu, 2016, “Optimal big data sharing approach for telehealth in cloud computing,” in Proc. IEEE Int. Conf. Smart Cloud (Smart Cloud), pp. 184– 189.
15. Mendis, Shanthi, Pekka Puska, Bo Norrving, World Health Organization, 2011, Global Atlas on Cardiovascular Disease Prevention and Control, PP. 3– 18. ISBN 978-92-4-156437-3.
16. Nithya, B., Dr. V. Ilango Professor, 2017, “Predictive Analytics in Health Care Using Machine Learning Tools and Techniques,” International Conference on Intelligent Computing and Control Systems.
17. Patankar, Bhavesh, and Vijay Chavda. "A comparative study of decision tree, naive Bayesian and k-nn classifiers in data mining." International Journal of Advanced Research in Computer Science and Software Engineering vol 4, no. 12 (2014): 776-779.
18. Penikalapati, Pragathi, and A. Nagaraja Rao. 2020, "Healthcare analytics by engaging machine learning." Science in Information Technology Letters vol.1, no. 1, pp 24-39.
19. Qian, B, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, 2015, “A relative similarity based method for interactive patient risk prediction,” Springer Data Mining Knowl. Discovery, vol. 29, no. 4, pp. 1070–1093.
20. Sharmila, S.Leoni, C.Dharuman and P.Venkatesan “Disease Classification Using Machine Learning Algorithms - A Comparative Study”, International Journal of Pure and Applied Mathematics Volume 114 No. 6 2017, 1-10
21. Shinde, Priyanka P., Kavita S. Oza, Rajanish K. Kamat, and S. V. Katkar. 2017, "Big Data Predictive Model: Towards Digital health." 4th International Conference on Emerging Trends in Engineering and Management Research,

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Institution of Electronics and Telecommunication Engineers, ISBN:9789386171634. pp. 184-189.

22. Shraddha Subhash Shirsath, 2018, “Disease Prediction Using Machine Learning Over Big Data” International Journal of Innovative Research in Science, Vol. 7, Issue 6.
23. Sisodia, Deepti, and Dilip Singh Sisodia. "Prediction of diabetes using classification algorithms." Procedia computer science vol 132 (2018): 1578-1585.
24. Sumbaly, Ronak, N. Vishnusri, and S. Jeyalatha. "Diagnosis of breast cancer using decision tree data mining technique." International Journal of Computer Applications, Vol. 98, no. 10 (2014).
25. Sunny, Allen Daniel, Sajal Kulshreshtha, Satyam Singh, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H, 2018, “ Disease Diagnosis System By Exploring Machine Learning Algorithms”, International Journal of Innovations in Engineering and Technology (IJIET) Volume 10 Issue 2.
26. Vijayashree, J., and N. Ch SrimanNarayanaIyengar. "Heart disease prediction system using data mining and hybrid intelligent techniques: A review." International Journal of Bio-Science and Bio-Technology 8, no. 4 (2016): 139-148.
27. Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, 2017, “HealthCPS: Healthcare cyberphysical system assisted by cloud and big data,” IEEE Syst. J., vol. 11, no. 1, pp. 88–95.