

A MINOR PROJECT REPORT
ON
DISEASE PREDICTION SYSTEM
SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF
DEGREE OF
BACHELOR OF TECHNOLOGY
IN
ELECTRONICS AND COMMUNICATION
ENGINEERING



Submitted By:

Under the Guidance of:

Yashi Agarwal (9919102010)

Dr. Banjrang Bansal

Arshita Tyagi (9919102039)

Shivangi Tripathi (9919102038)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING JAYPEE INSTITUTE OF INFORMATION
TECHNOLOGY, NOIDA (U.P.) May, 2022

CERTIFICATE

This is to certify that the minor project report entitled, “DISEASE PREDICTION SYSTEM” submitted by Yashi Agarwal, Shivangi Tripathi and Arshita Tyagi in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in Electronics and Communication Engineering of the Jaypee Institute of Information Technology, Noida is an authentic work carried out by them under my supervision and guidance. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Signature of Supervisor:

Name of the Supervisor: Dr. Bajrang Bansal

ECE Department, JIIT, Sec-128,

Noida-201304

Dated: May 25, 2022

DECLARATION

We hereby declare that this written submission represents our own ideas in our own words and where other's ideas or words have been included, have been adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission.

Place: Noida

Date: May 24, 2022

Name: Yashi Agarwal

Enrollment: 9919102010

Name: Shivangi Tripathi

Enrollment: 9919102038

Name: Arshita Tyagi

Enrollment: 9919102039

Table of Contents

CERTIFICATE	ii
DECLARATION	iii
ABSTRACT	v
LIST OF FIGURES	vi
LIST OF TABLES	vii
Chapter 1 INTRODUCTION	1
1.1 Concluding Remarks	8
Chapter 2 LITERATURE SURVE	9
2.1 Effective Heart Diseases Prediction Using Hybrid Machine Learning Techniques.	
2.2 Designing Disease Prediction Model Using Machine Learning Approach	
2.3 Comparing different supervised machine learning algorithms for disease prediction	
2.4 Heart Disease Prediction using Machine Learning Techniques	
2.5 Feature selection and classification systems for chronic disease prediction: A review	
2.6 Disease Prediction by Machine Learning Over Big Data From Healthcare Communities	
2.7 Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis	
2.8 GDPS - General Disease Prediction System	
2.9 Identification and Prediction of Chronic Diseases Using Machine Learning Approach	
2.10 Implementation of Machine Learning Models for the Prevention of Kidney Diseases (CKD) or Their Derivatives	
2.11 Symptoms Based Disease Prediction Using Machine Learning Techniques	
Chapter 3 METHODOLOGY	17
3.1 Algorithms and Techniques	18
3.1.1 Logistic Regression	18

	3.1.2 Random Forest	19
	3.1.3 Decision Tree	20
	3.1.4 Naive Baye's	22
	3.2 Conclusion	23
Chapter 4	SIMULATION RESULTS	24
	4.1 Performance and Accuracy	24
	4.2 Outputs	25
Chapter 5	CONCLUSION AND FUTURE SCOPE	34
	5.1 Conclusion	34
	5.2 Future Scope	36
	REFERENCES	37

ABSTRACT

Sometimes users neglect to go personally to the hospital for their regular check-ups which they thought is more time consuming. Such a problem can be solved by using the Health App application by giving proper guidance regarding healthy living. Over the past years, in the healthcare sector the use of the prediction tools along with the concerning health has been increased. Thus, this system is concentrating on providing immediate and accurate disease prediction to the users about the symptoms they enter along with the severity of disease predicted. Normally users are not aware about how a particular disease spreads; this project also looks forward to providing proper ways to reduce the spreading of a particular disease. Therefore, this arrangement helps in easier health management. A web application is deployed for users for easy portability, configuring and being able to access remotely where doctors cannot reach easily.

Disease Detection system based on predictive modeling predicts the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and gives the probability of the disease as an output. The main objective is to predict disease accurately with few tests and attribute the presence of the disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having diseases (like diabetes, kidney, heart, breast and many more). Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients.

LIST OF FIGURES

Fig. 2.1	Experiment workflow with UCI dataset.	10
Fig. 2.2	Prediction of heart disease with HRFLM.	17
Fig. 2.3	System Architecture.	20
Fig. 2.4	Feature Selection Process.	21
Fig. 2.5	Adaptive and Parallel classification process.	25
Fig. 2.6	Three ML Algorithms used in the our prediction system.	26
Fig. 2.7	Block diagram for general disease prediction system.	26
Fig. 2.8	Application of data mining.	27
Fig. 3.1	Block Diagram for disease Prediction System.	27
Fig. 3.2	Random Forest Classifier.	28
Fig. 3.3	Decision Forest Classifier.	28
Fig. 3.4	Naive Baeyes algorithm code snippet.	29
Fig. 3.5	Random Forest algorithm code snippet.	30
Fig. 3.6	Decision Tree algorithm code snippet.	30
Fig. 4.1	Selecting the symptoms.	31
Fig. 4.2	Selecting the algorithms to predict the disease.	31
Fig. 4.3	Predicting the disease using all the algorithms.	32
Fig. 4.4	Showing the accuracy of the algorithms.	32

LIST OF TABLES

Table 2.1	The Comparative study of breast cancer.	11
Table 4.1	Accuracy of different supervised learning algorithms.	12
Table 5.1	Conclusion on importance and limitations of supervised learning.	12

Chapter 1

INTRODUCTION

Artificial Intelligence is the study and development of approaches that imitate human intelligence. The technique has been successful in a variety of fields including fraud detection, computer vision, online advertising, robotics, automatic drivers, etc. With its success in areas like disease diagnosis, treatment, patient monitoring, drug discovery, epidemiology, etc, there is a great hope that Artificial Intelligence can be a vibrant area of research to tackle the challenges human faces currently. It is argued that AI will be the key to supporting clinical and academic studies of covid-19 and future crises. For example, at the beginning of outbreak, China initiated a set of actions against the spread of the virus, by adopting a set of AI-based technologies. In this effort, they explored implementation of ideas like the use of facial recognition cameras to track infected people, drones to disinfect places, robots to deliver food and medications etc .There are different fields of applications for which AI approaches are adopted to manage the effects of the disease. This report try to organize the research based on the applications. The applications include clinical applications, disease detection, processing covid-19 related images, pharmaceutical studies and epidemiology. The main categorization is based on applications; however, for the same application, researches are subdivided based on the AI approaches they have employed. Examples of AI approaches include Deep learning, machine learning, Artificial Neural Networks and evolutionary algorithms. AI approaches have shown to be very efficient in modelling complex systems. AI approaches have long been employed for the development of diagnosis and treatment system. Now this pandemic has created a new challenge for this field of science. Developing intelligent systems that can help practitioners in terms of diagnosis, monitoring, prediction of patient's conditions and offering treatment measures can be very helpful to help the already under pressure health systems where there is lack of beds and doctors in the hospital.

variability (HRV) signals (derived from electrocardiogram (ECG) signals) can be effectively used for the non-invasive detection of diabetes. This research paper presents a methodology for classification of diabetic and normal HRV signals using deep learning architecture.

Chapter 2

LITERATURE SURVEY

AI approaches have long been employed for the development of diagnosis and treatment system. Now this pandemic has created a new challenge for this field of science. Developing intelligent systems that can help practitioners in terms of diagnosis, monitoring, prediction of patient's conditions and offering treatment measures can be very helpful to help the already under pressure health systems. AI applications can potentially aid to solve the problem of the “iron triangle” in the healthcare sector. It involves three interlocking factors which are, namely, access, affordability, and effectiveness. AI is usually used for diagnosis and treatment recommendations, patient engagement and adherence or administration activities. The aim of this chapter is to perform a comprehensive survey on the applications of AI in battling against the difficulties the outbreak has caused and to explain the most common techniques and the biggest challenges in disease detection and to summarize the various results from the newest papers. In this sense, we tried to cover every way that AI approaches have been employed and to cover all the research until the writing of this paper. Such a picture, although full of details, is very helpful in understanding where AI sits in current pandemonium. Since the pandemic is a new and developing problem, much of the research has not yet been peer reviewed. Therefore, this paper also covers pre-print works. This report tried to conclude the paper with ideas on how the problems can be tackled in a better way and provide some suggestions for the future.

Therefore, it is very important to compare results of a newly created model produced with the newest information and not only with the state of the art methods. This chapter is a source of such information for researchers in order for them to be precisely correct on result comparison before publishing new achievements in this field. In this literature survey three digital libraries are chosen due to limited resources and the huge number of articles under this topic. However, it can be clearly seen that these libraries cover a significant amount of the related literature sources for our study. Two different digital libraries were used to execute a research:

IEEE Xplore

Web of Science—WOS (previously known as Web of Knowledge)

2.1 Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques

[1] Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a glimpse into predicting heart disease with ML techniques. In this paper, we propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques. We produce an enhanced performance level with an accuracy level of 88.7% through the prediction model for heart disease with the hybrid random forest with a linear model (HRFLM).

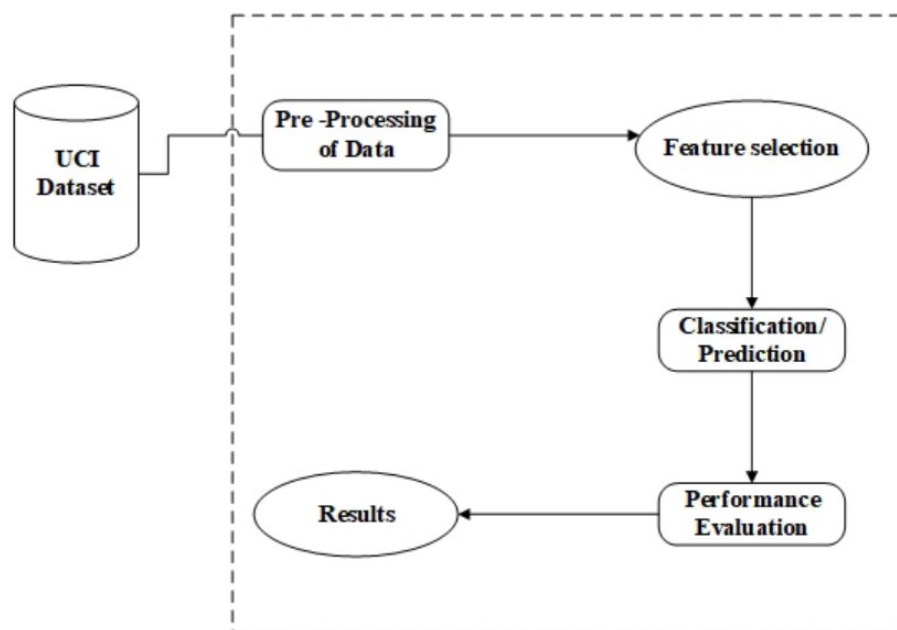


Fig 2.1 Experiment workflow with UCI dataset

Identifying the processing of raw healthcare data of heart information will help in the long term saving of human lives and early detection of abnormalities in heart conditions. Machine learning techniques were used in this work to process raw data and provide a new and novel discernment towards heart disease. Heart disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are

adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed hybrid HRFLM approach combines the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease.

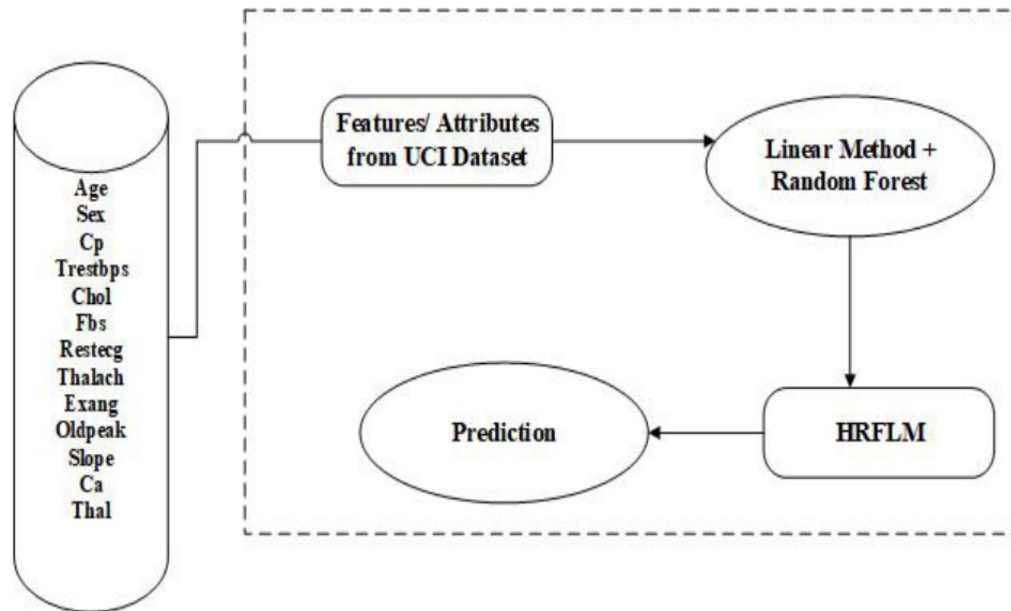


Fig 2.2. Prediction of heart disease with HRFLM.

2.2 Designing Disease Prediction Model Using Machine Learning Approach

[2]Now-a-days, people face various diseases due to environmental conditions and their living habits. So the prediction of disease at an earlier stage becomes an important task. But the accurate prediction on the basis of symptoms becomes too difficult for the doctor. The correct prediction of disease is the most challenging task. To overcome this problem data mining plays an important role in predicting disease. Medical science has a large amount of data growth per year. Due to increased data growth in the medical and healthcare field, the accurate analysis of medical data has benefited from early patient care. With the help of disease data, data mining finds hidden pattern information in the huge amount of medical data. We proposed general disease prediction based on symptoms of the patient. For disease prediction, we use K-Nearest Neighbor (KNN) and Convolutional neural network (CNN) machine learning algorithms for accurate prediction of disease.

For disease prediction, a required disease symptoms dataset. In this general disease prediction, the living habits of a person and checkup information are considered for the accurate prediction. The accuracy of general disease prediction by using CNN is 84.5% which is more than the KNN algorithm. And the time and the memory requirement is also more in KNN than CNN. This system is able to give the risk associated with general disease which is lower risk of general disease or higher.

Today machine learning is present everywhere so that without knowing it, one can possibly use it many times a day. CNN uses both the structured and unstructured data of a hospital to do classification. While other machine learning algorithms work on structured data and time required for computation is high, they are lazy because they store entire data as a training dataset and use complex methods for calculation.

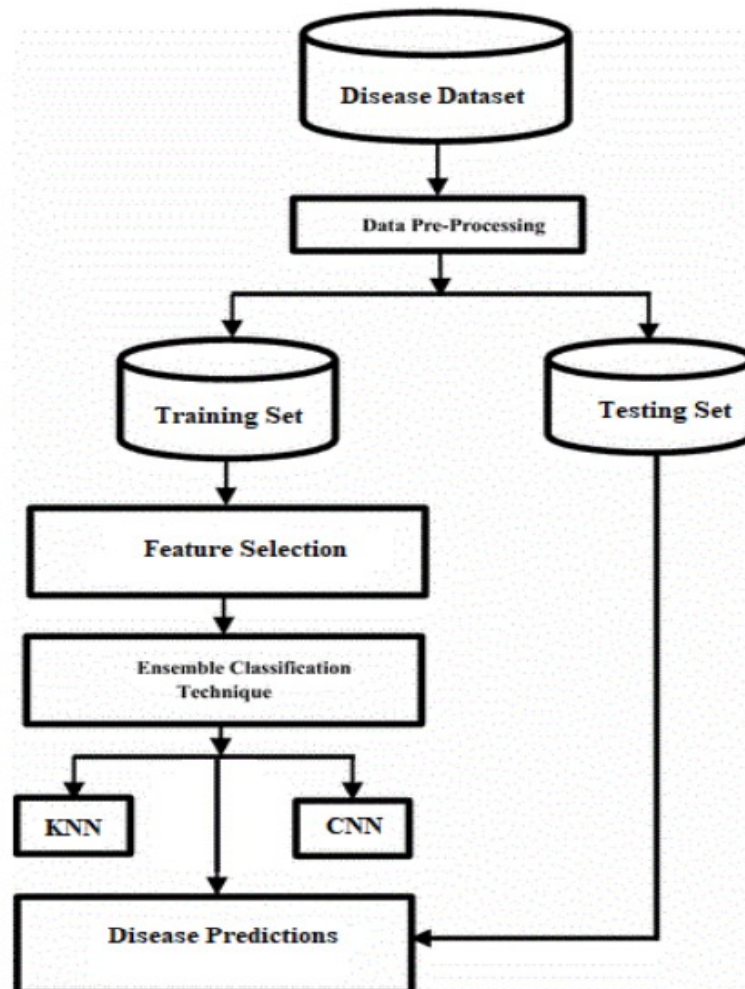


Fig 2.3 . System Architecture

Initially they took a disease dataset from the UCI machine learning website and that is in the form of a disease list with its symptoms. After that, preprocessing is

performed on that dataset for cleaning, that is, removing commas, punctuations and white places. And that is used as a training dataset. After that feature is extracted and selected. Then we classify that data using classification techniques such as KNN and CNN. Based on machine learning we can predict accurate disease

2.3 Comparing different supervised machine learning algorithms for disease prediction

[3]Supervised machine learning algorithms have been a dominant method in the data mining field. Disease prediction using health data has recently shown a potential application area for these methods. This study aims to identify the key trends among different types of supervised machine learning algorithms, and their performance and usage for disease risk prediction.

Methods

Supervised machine learning algorithm

At its most basic sense, machine learning uses programmed algorithms that learn and optimize their operations by analyzing input data to make predictions within an acceptable range. With the feeding of new data, these algorithms tend to make more accurate predictions. Although there are some variations of how to group machine learning algorithms they can be divided into three broad categories according to their purposes and the way the underlying machine is being taught. These three categories are: supervised, unsupervised and semi-supervised.

Logistic regression

Logistic regression (LR) is a powerful and well established method for supervised classification . It can be considered as an extension of ordinary regression and can model only a dichotomous variable which usually represents the occurrence or nonoccurrence of an event. LR helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier, a threshold needs to be assigned to differentiate two classes.This generalized version of LR is known as the multinomial logistic regression.

Support vector machine

Support vector machine (SVM) algorithm can classify both linear and non-linear

data. It first maps each data item into an n -dimensional feature space where n is the number of features. It then identifies the hyperplane that separates the data items into two classes while maximizing the marginal distance for both classes and minimizing the classification errors.

Decision tree

Decision tree (DT) is one of the earliest and most prominent machine learning algorithms. A decision tree models the decision logic i.e., tests and corresponding outcomes for classifying data items into a tree-like structure. The nodes of a DT tree normally have multiple levels where the first or top-most node is called the root node. All internal nodes (i.e., nodes having at least one child) represent tests on input variables or attributes.

Random forest

A random forest (RF) is an ensemble classifier consisting of many DTs similar to the way a forest is a collection of many trees .

Naïve Bayes

Naïve Bayes (NB) is a classification technique based on Bayes' theorem . This theorem can describe the probability of an event based on the prior knowledge of conditions related to that event. This classifier assumes that a particular feature in a class is not directly related to any other feature although features for that class could have interdependence among themselves.

2.4 Heart Disease Prediction using Machine Learning Techniques

[4]Numerous works have been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centers. CKD increases the risk factors of Cardiovascular Disease (CVD) like hypertension, diabetes mellitus, dyslipidemia, and metabolic syndrome. CKD also leads to End Stage Renal Disease (ESRD) which has no cure. U. N. Dulhare.et al [3] extracted action rules based on stages but also predicted CKD by using naïve bayes with OneR attribute selector which helps to prevent the advancing of chronic renal disease to further stages. It is said that the median survival time of past due-stage patients is approximately three years. Evaluating exactly the condition of sufferers is of incredible importance as it might substantially assist to decide appropriate care, medications or medical interventions which amongst them have a complicated interrelationship and have an impact on the final results of the person.

2.5 Feature selection and classification systems for chronic disease prediction: A review

[5]Chronic Disease Prediction plays a pivotal role in healthcare informatics. It is crucial to diagnose the disease at an early stage. This paper presents a survey on the utilization of feature selection and classification techniques for the diagnosis and prediction of chronic diseases. Adequate selection of features plays a significant role for enhancing the accuracy of classification systems. Dimensionality reduction helps in improving overall performance of machine learning algorithms. The application of classification algorithms on disease datasets yields promising results by developing adaptive, automated and intelligent diagnostic systems for chronic diseases. Parallel classification systems can be used to expedite the process and to enhance the computational efficiency of results. This work presents a comprehensive overview of various feature selection methods and their inherent pros and cons. We then analyze adaptive classification systems and parallel classification systems for chronic disease prediction.

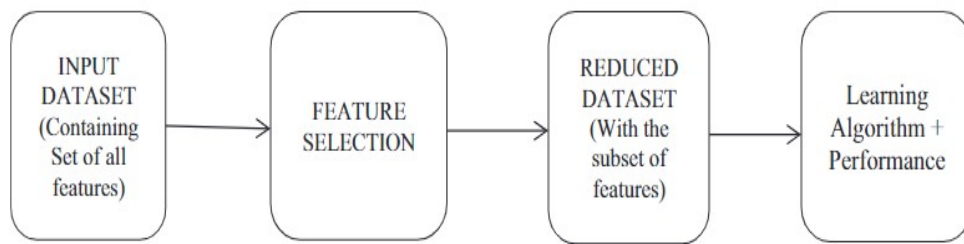


Fig 2.4. Feature selection process.

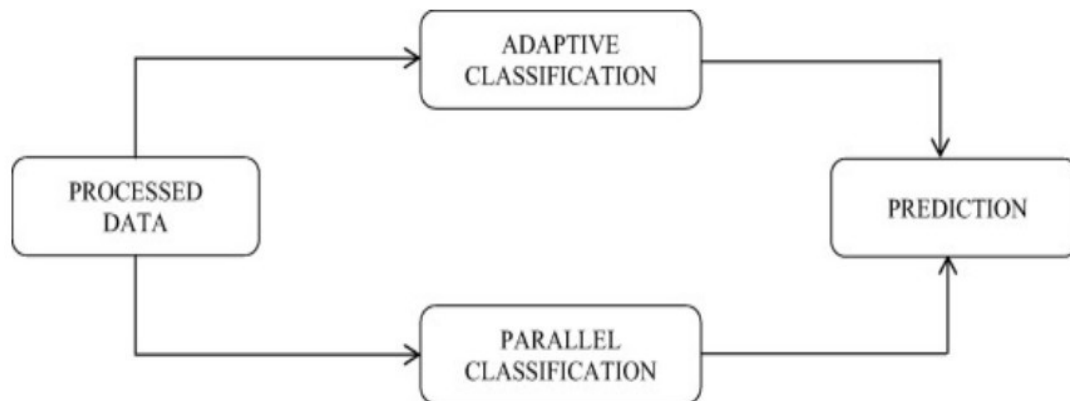


Fig 2.5. Depicts how adaptive classification process and parallel classification process can be applied on processed data to get predictive results.

2.6 Disease Prediction by Machine Learning Over Big Data From Healthcare Communities

[6] With big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. However, the analysis accuracy is reduced when the quality of medical data is incomplete. 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 outbreaks in disease-frequent communities. We experimented with modified prediction models over real-life hospital data collected from central China in 2013–2015. To overcome the difficulty of incomplete data, we use a latent factor model to reconstruct the missing data. We experiment on a regional chronic disease of cerebral infarction. We propose a new convolutional neural network (CNN)-based multimodal disease risk prediction algorithm using structured and unstructured data from hospitals. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared with several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed, which is faster than that of

the CNN-based unimodal disease risk prediction algorithm.

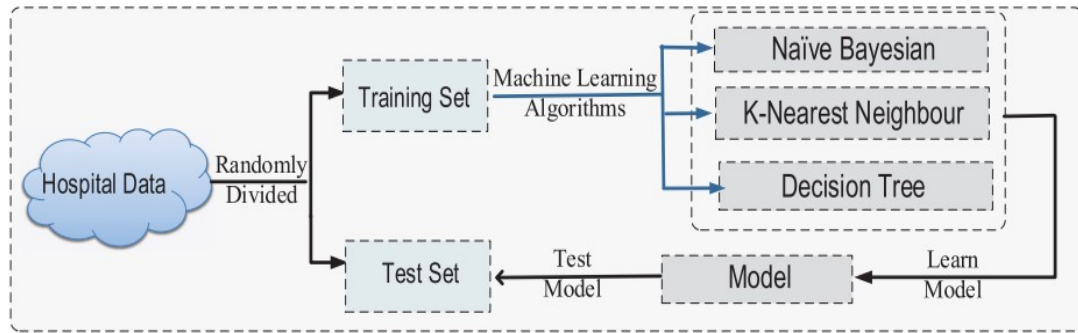


Fig 2.6. The three machine learning algorithms used in our disease prediction experiments.

In this paper, they propose a new convolutional neural network based multi modal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospitals. To the best of our knowledge, none of the existing work focused on both data types in the area of medical big data analytics. Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction (CNN-UDRP) algorithm

2.7 Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis

[7]This paper reviews applications of machine learning (ML) predictive models in the diagnosis of chronic diseases. Chronic diseases (CDs) are responsible for a major portion of global health costs. Patients who suffer from these diseases need lifelong treatment. Nowadays, predictive models are frequently applied in the diagnosis and forecasting of these diseases. In this study, we reviewed the state-of-the-art approaches that encompass ML models in the primary diagnosis of CD. Our outcomes suggest that there are no standard methods to determine the best approach in real-time clinical practice since each method has its advantages and disadvantages. Among the methods considered, support vector machines (SVM), logistic regression (LR), clustering were the most commonly used. These models are highly applicable in classification, and diagnosis of CD and are expected to become more important in medical practice in the near future.

The present study evaluated the studies associated with the diagnosis of chronic diseases. However, the implementation of correct methods or selection of the right models is a

prerequisite to perform ideal decisions, as modern researchers are claiming that some ML models are compromised by enlarging contained datasets with malicious data that can have severe consequences. On the other hand, diagnosis limitations may lead to life-threatening attacks, and sometimes it might be a driving factor of fatality. In contrast, the wrong diagnosis prompts skepticism in machine learning use, which can lead policy makers to avoid predictive model usage. Therefore, reviews on predictive models can provide evidence to propose excellent methods for the CDs diagnosis. In the future, AI techniques like ML, cognitive computing and deep learning may play a critical role in the interpretation of chronic diseases. However, researchers are progressively attracted by predictive model techniques in the advancement of health care. As new advancements in medical care are being established and are expanding the access to electronic data, this opens new doors to decision support and productivity improvements. These models are designed to emphasize the responsibility of patient care quality and cut down medical costs.

2.8 GDPS - General Disease Prediction System

[8]The successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. The medical environment is still information rich but knowledge weak. There is a wealth of data possible within the medical systems. However, there is a lack of powerful analysis tools to identify hidden relationships and trends in data. Disease is a term that assigns to a large number of health care conditions related to the body. These medical conditions describe the unexpected health conditions that directly control all the body parts. Medical data mining techniques like association rule mining, classification, and clustering are implemented to analyze the different kinds of general body-based problems. Classification is an important problem in data mining. A number of popular classifiers construct decision trees to generate class models. The data classification is based on the ID3 Decision Tree algorithm which results in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared.

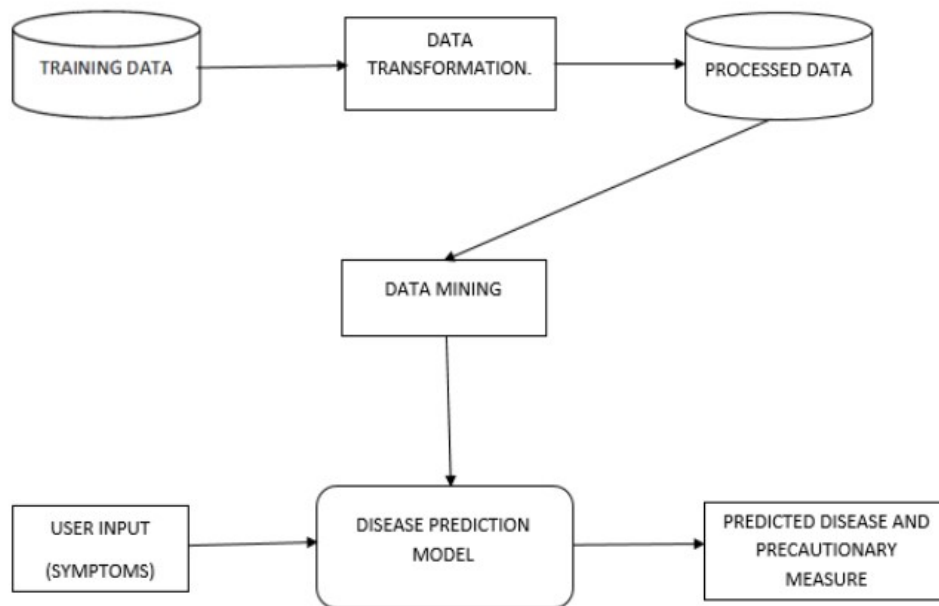


Fig 2.7. Block Diagram for General disease prediction system

The system has been implemented with an accuracy of 86.67% on the dataset of 120 patient data. The current system covers only the general diseases or the more commonly occurring disease, the plan is to include disease of higher fatality, like various cancers in future, so that early prediction and treatment could be done, and the fatality rate of deadly diseases like cancer decreases, with the economic benefit in long sight as well.

2.9 A Survey on technique for prediction of disease in Medical Data

[9]In today's era data mining plays an important role for prediction of diseases in the medical field. With the growing research on disease prediction systems, it has become important to discover hidden patterns and relationships from medical databases. In classical clinical diagnosis, it requires lots of tests which could complicate the disease prediction. Hence the data mining techniques can help medical expertise to take the decision about the disease using computer aided decision support systems. In this paper a comprehensive survey on various data mining techniques used for disease prediction is presented.

Table 2.1. The comparative study of breast cancer

Authors	Year	Accuracy
M. Yaghoobi et al.	(2014)	More than 90%
Burke B.H et al.	(1999)	5yrs-0.909, 10yrs-0.086 &for 15yrs0.883
G.Walker et al.	(2005)	93.6%
E. Gauven et al.	(2006)	87%
R. Ceylan et al.	(2013)	98.05%
Behnam H et al.	(2005)	ROC-0.96
S.Nahavandi et al.	(2015)	97.40%
Sulong et al.	(2012)	Test at 45 years
P. Lim et al.	(2014)	98.84%
D. Chen et al.	(2001)	96.67%
Choi Chul Sang et al.	(2001)	

The main focus of this paper is to discuss decision parameters, attributes, and features used for predicting the disease. The method also throws light on the importance of different classification methods for prediction of disease in medical datasets. The dataset considered in so many existing techniques that we have discussed are related to heart and breast cancer. The various data mining techniques are used as classifier, to build a cost effective model for disease prediction. Hence it is well understood by the exhaustive survey that mining the required information from the medical data help us to support well informed diagnosis and decisions.

2.10 Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis

[10]The worldwide study on causes of death due to heart disease/syndrome has observed that it is the major cause of death. If recent trends are allowed to continue, 23.6 million people will die from heart disease in the coming 2030. The healthcare industry collects large amounts of heart disease data which unfortunately are not “mined” to discover hidden information for effective decision making. In this paper, a study of PCA has been done which finds the minimum number of attributes required to enhance the precision of

various supervised machine learning algorithms. The purpose of this research is to study supervised machine learning algorithms to predict heart disease. Data mining has a number of important techniques like categorization, preprocessing. Diabetic is a life-threatening disease which is prevalent in several urbanized as well as emergent countries like India. The data categorization is diabetic patients datasets which is developed by collecting data from hospital repositories consisting of 1865 instances with dissimilar attributes. The examples in the dataset are two categories of blood tests, urine tests. In this research paper we discuss a variety of algorithm approaches of data mining that have been utilized for diabetic disease prediction. Data mining is a well known practice used by health organizations for classification of diseases such as diabetes and cancer in bio informatics research.

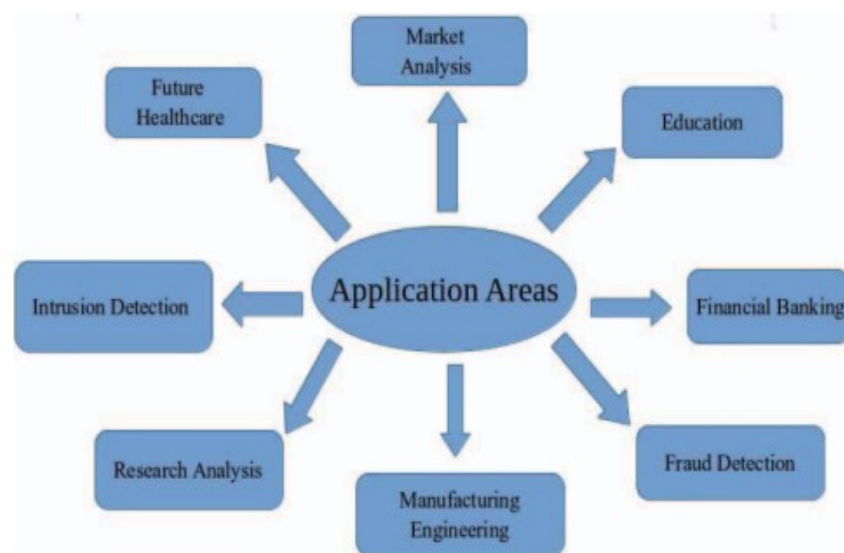


Fig 2.8. Applications of Data Mining

Types of Machine Learning

Some of the main types of machine learning are:

- a) Supervised Learning:** In this learning the training data is labeled with the correct answers for example spam or ham the two most common types of supervised learning are classification and regression.
- b) Unsupervised learning:** In this we are given a collection of unlabeled data, which we wish to analyze & discover patterns within. The two most important examples are dimension reduction and clustering.
- c) Reinforcement learning:** In this an agent for example, a robot or controller seeks to learn the optimal actions to take based on the results of past actions.

In this paper, for heart disease prediction SVM, Naive Bayes and Decision tree has been applied with and without using PCA on the dataset. We used PCA to reduce the number

of attributes. After reducing the size of the dataset, SVM outperforms Naive Bayes and Decision trees. SVM can further be used to predict heart disease. A GUI desktop application can be built using SVM and this dataset to predict the possibility of cardiovascular disease in a patient and for diabetes data prediction, the main aim of this paper is to predict diabetes disease using WEKA data mining tool. Our algorithms were implemented using WEKA data mining technique to analyze algorithm accuracy which was obtained after running these algorithms in the output window. These algorithms compare classifier accuracy to each other on the basis of correctly classified instances, time taken to build the model, mean absolute error and ROC Area. So, using above all observations, we can conclude that Maximum ROC Area means excellent prediction performance as compared to other algorithms.

2.11 Symptoms Based Disease Prediction Using Machine Learning Techniques

[11]Computer Aided Diagnosis (CAD) is a quickly evolving, diverse field of study in medical analysis. Significant efforts have been made in recent years to develop computer-aided diagnostic applications, as failures in medical diagnosing processes can result in medical therapies that are severely deceptive. Machine learning (ML) is important in the Computer Aided Diagnostic test. Objects such as body-organs cannot be identified correctly after using an easy equation. Therefore, pattern recognition essentially requires training from instances. In the biomedical area, pattern detection and ML promises to improve the reliability of disease approach and detection. They also respect the dispensation of the method of decision making. ML provides a respectable approach to make superior and automated algorithms for the study of high dimension and multi - modal biomedical data. The relative study of various ML algorithms for the detection of various diseases such as heart disease, diabetes disease is given in this survey paper. It focuses on the collection of algorithms and techniques for ML used for disease detection and decision making processes.

The evaluation field has been flooded by statistical prediction models that are incapable of generating good quality outcomes. In maintaining generalized knowledge, statistical models are not efficient, coping with missing values and broad data points. The value of MLT stems from all of these causes. In many applications, ML plays a vital role, such as image recognition, data mining, processing of natural languages and diagnosis of diseases. ML provides potential solutions in all these fields. This paper discusses various techniques of ML for the diagnosis of various diseases such as heart, diabetes diseases.

Most models have shown excellent results because they specifically describe the characteristic. It is noted from previous studies that SVM provides 94.60 percent improved performance for heart disease identification. Naive Bayes is a correctly diagnosed diabetes condition. It provides 95 percent of the highest classification precision. The survey shows the benefits and drawbacks of such algorithms. A suite of tools built in the AI community is also presented in this survey paper. These approaches are very useful for the analysis of certain problems and also provide opportunities for an improved decision making process.

Conclusion

In this chapter, a systematic review of the intelligent data analysis tools in the medical field is provided. Some examples of some algorithms used in these areas of the medical field, analyzing some possible trends focused on the goal searched, the technique used, and the application area are also provided. Additionally, the advantages and disadvantages of each technique described to help in a future establishment about which the technique is most suitable for each real-life situation addressed by authors are provided. A systematic review such as the one that we have just presented may become outdated in a short time, given the speed with which new works appear in this emerging area. For this reason, we consider that Table 2.1 to Table 2.5 should be mainly updated, after a careful search for new scientific literature, given that it is more likely that more studies will appear in the short term on the application of existing techniques in this area than on the proposal of new techniques that really constitute a novelty, and not a mere improvement or modification of existing ones.

In the last decades, this detection has been performed through the process of discovering interesting patterns in databases. This knowledge in databases is called Data Mining. However, discovering these patterns is not an easy task. Hence, many techniques were developed within Artificial Intelligence, where Machine Learning appears as a method for providing tools for intelligent data analysis.

Chapter 3

METHODOLOGY

This project is developed using machine learning to overcome general diseases in earlier stages as all know in competitive environment of economic development the mankind has involved so much that he/she is not concerned about health according to research there are 40% peoples ignores about general disease which leads to harmful disease later. This project is implemented completely using python. Even the interface of this project is done using Python's library called flask. This system predicts the chance of presence of a disease present in a patient on the basis of their symptoms. It will also recommend necessary precautionary measures required to treat the predicted disease. The system will initially be fed with data from different sources i.e. available datasets on the internet, the data then pre-processed before further processing is carried out. This is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. Model has been created using different classification algorithms. Fig 3.1 shows the block diagram of the system.

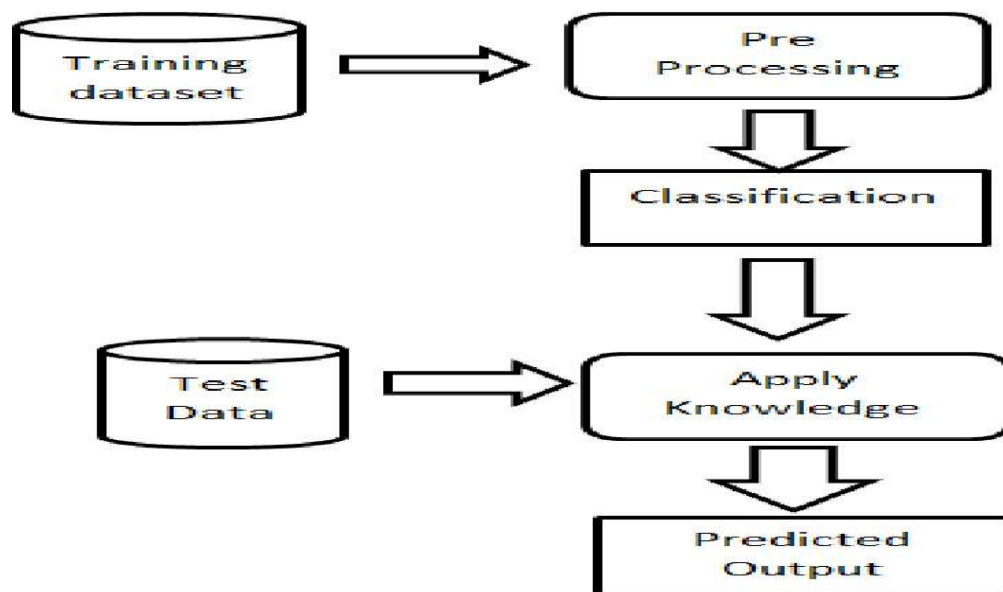


Fig. 3.1 Block Diagram for Disease prediction System.

[12]The system is implemented into two parts, the admin part and the user part. The duty of the admin is training the system for creation of the disease prediction model. The user uses the services provided by the model after logging in as the user, entering the symptoms into the model, which in turn returns the predicted results and necessary precautionary measures.

3.1 Algorithms and Techniques

Classification can be performed on structured or unstructured data. Classification is a technique where data can be categorized into a given number of classes. The main goal of a classification problem is to identify the category/class to which a new data will fall under. Few of the terminologies encountered in machine learning – classification:

- Classifier: An algorithm that maps the input data to a specific category.
- Classification model: A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.
- Feature: A feature is an individual measurable property of a phenomenon being observed.
- Binary Classification: Classification task with two possible outcomes. Eg: Gender classification (Male / Female)
- Multi-class classification: Classification with more than two classes. In multi-class classification each sample is assigned to one and only one target label. Eg: An animal can be a cat or dog but not both at the same time.
- Multi-label classification: Classification task where each sample is mapped to a set of target labels (more than one class). Eg: A news article can be about sports, a person, and location at the same time.

The following are the steps involved in building a classification model:

- Initialize the classifier to be used.
- Train the classifier: All classifiers in scikit-learn uses a $\text{fit}(X, y)$ method to fit the model(training) for the given train data X and train label y .
- Predict the target: Given an unlabeled observation X , $\text{predict}(X)$ returns the predicted label y .
- Evaluate the classifier model.

3.1.1 Logistic regression

[13] Logistic Regression is a classification and not a regression algorithm. It estimates discrete values (Binary values like 0/1, yes/no, true/false) based on a given set of independent variable(s). Simply put, it basically predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression. The values obtained would always lie within 0 and 1 since it predicts the probability. Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS),

which is widely used to predict mortality in injured patients. Many other medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).

3.1.2 Random Forest

[14]Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below Fig 3.2 shows how a random forest would look like with two trees:

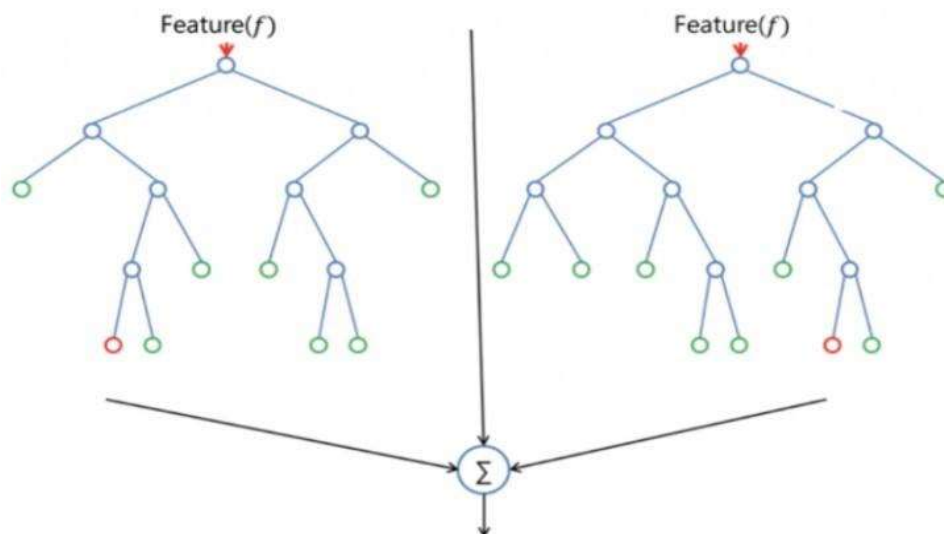


Fig. 3.2 Random Forest Classifier.

Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in a random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

The random forest algorithm is used in a lot of different fields, like banking, the stock market, medicine and e-commerce. In finance, for example, it is used to detect customers more likely to repay their debt on time, or use a bank's services more frequently. In this domain it is also used to detect fraudsters out to scam the bank. In trading, the algorithm can be used to determine a stock's future behavior. In the healthcare domain it is used to identify the correct combination of components in medicine and to analyze a patient's medical history to identify diseases.

3.1.3 Decision Tree

[15]A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursive manner called recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret. Fig 3.3 shows how decision tree looks like.

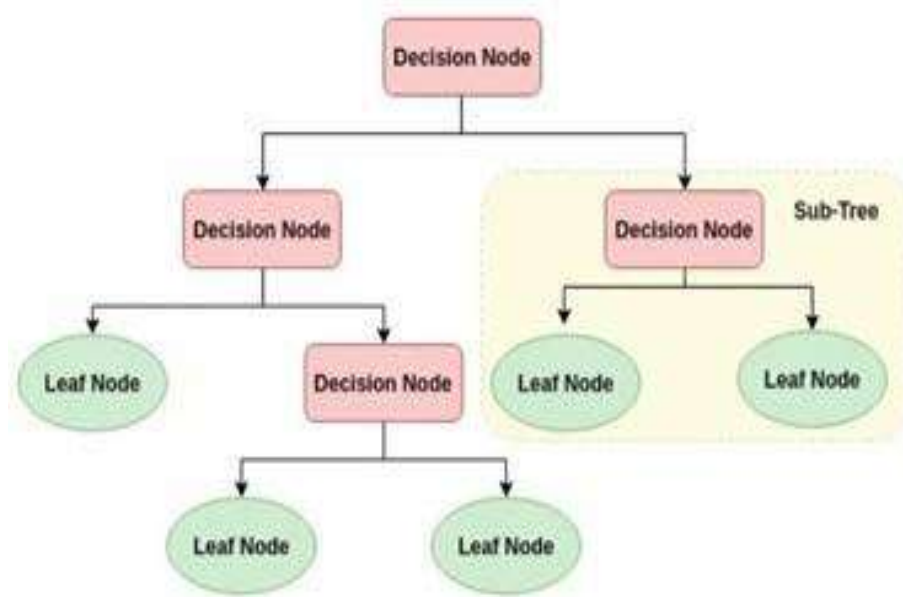


Fig. 3.3 Decision tree Classifier.

Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy. The basic idea behind any decision tree algorithm is as follows:

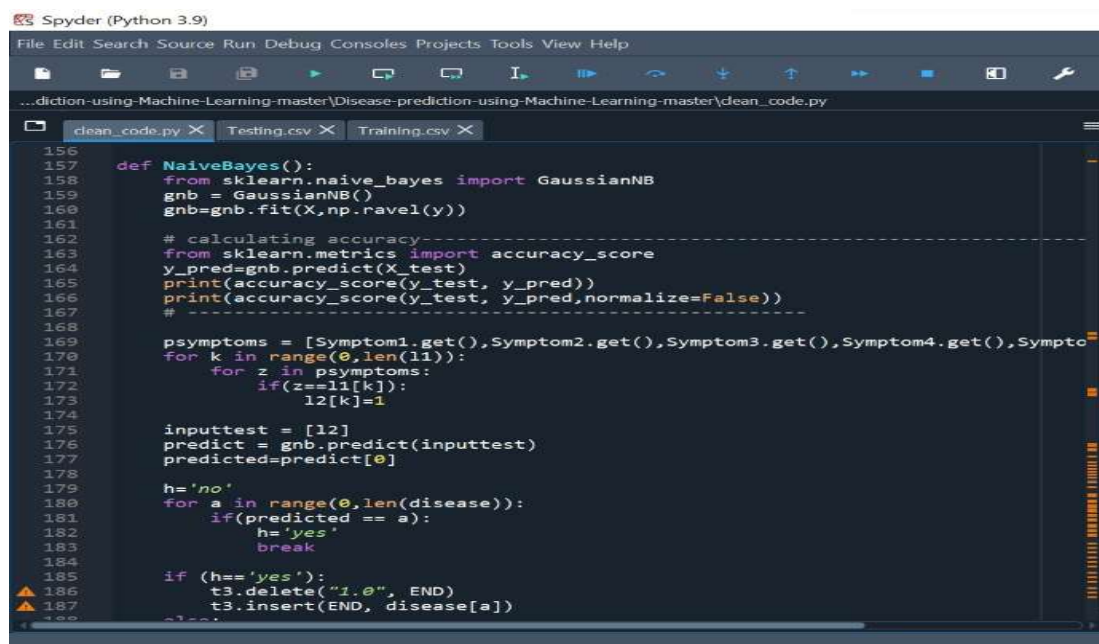
- Select the best attribute using Attribute Selection Measures(ASM) to split the records.
- Make that attribute a decision node and break the dataset into smaller subsets.
- Starts tree building by repeating this process recursively for each child until one of the conditions will match:
 - All the tuples belong to the same attribute value.
 - There are no more remaining attributes.
 - There are no more instances.

3.1.4 Naïve Bayes

The simplest solutions are usually the most powerful ones, and Naïve Bayes is a good example of that. Despite the advances in Machine Learning in the last years, it has proven to not only be simple but also fast, accurate, and reliable. It has been successfully used for many purposes, but it works particularly well with natural language processing (NLP) problems. Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks.

Which tells us: how often A happens given that B happens, written $P(A|B)$ also called posterior probability, When we know: how often B happens given that A happens, written $P(B|A)$ and how likely A is on its own, written $P(A)$ and how likely B is on its own, written $P(B)$.

Simulation:



```
156
157 def NaiveBayes():
158     from sklearn.naive_bayes import GaussianNB
159     gnb = GaussianNB()
160     gnb=gnb.fit(X,np.ravel(y))
161
162     # calculating accuracy-----
163     from sklearn.metrics import accuracy_score
164     y_pred=gnb.predict(X_test)
165     print(accuracy_score(y_test, y_pred))
166     print(accuracy_score(y_test, y_pred,normalize=False))
167     # -----
168
169     psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Sympto
170     for k in range(0,len(l1)):
171         for z in psymptoms:
172             if(z==l1[k]):
173                 l2[k]=1
174
175     inputtest = [l2]
176     predict = gnb.predict(inputtest)
177     predicted=predict[0]
178
179     h='no'
180     for a in range(0,len(disease)):
181         if(predicted == a):
182             h='yes'
183             break
184
185     if (h=='yes'):
186         t3.delete("1.0", END)
187         t3.insert(END, disease[a])
188
189     #1.0
```

Fig 3.4 Naive Bayes algorithm code snippet

```

120 def randomforest():
121     from sklearn.ensemble import RandomForestClassifier
122     clf4 = RandomForestClassifier()
123     clf4 = clf4.fit(X,np.ravel(y))
124
125     # calculating accuracy-----
126     from sklearn.metrics import accuracy_score
127     y_pred=clf4.predict(X_test)
128     print(accuracy_score(y_test, y_pred))
129     print(accuracy_score(y_test, y_pred,normalize=False))
130     # -----
131
132     psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Sympto
133
134     for k in range(0,len(l1)):
135         for z in psymptoms:
136             if(z==l1[k]):
137                 l2[k]=1
138
139     inputtest = [l2]
140     predict = clf4.predict(inputtest)
141     predicted=predict[0]
142
143     h='no'
144     for a in range(0,len(disease)):
145         if(predicted == a):
146             h='yes'
147             break
148
149     if (h=='yes'):
150         t2.delete("1.0", END)
151         t2.insert(END, disease[a])

```

Fig 3.5 Random Forest algorithm code snippet

```

79 def DecisionTree():
80
81     from sklearn import tree
82
83     clf3 = tree.DecisionTreeClassifier() # empty model of the decision tree
84     clf3 = clf3.fit(X,y)
85
86     # calculating accuracy-----
87     from sklearn.metrics import accuracy_score
88     y_pred=clf3.predict(X_test)
89     print(accuracy_score(y_test, y_pred))
90     print(accuracy_score(y_test, y_pred,normalize=False))
91     # -----
92
93     psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Sympto
94
95     for k in range(0,len(l1)):
96         # print (k,)
97         for z in psymptoms:
98             if(z==l1[k]):
99                 l2[k]=1
100
101     inputtest = [l2]
102     predict = clf3.predict(inputtest)
103     predicted=predict[0]
104
105     h='no'
106     for a in range(0,len(disease)):
107         if(predicted == a):
108             h='yes'
109             break
110
111

```

Fig 3.6. Decision Tree algorithm code snippet

3.2 Conclusion

In this chapter, various machine learning algorithms have been discussed. These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. Machine learning is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the pattern or extract information

from the data. In that case, we apply machine learning. With the abundance of datasets available, the demand for machine learning is on the rise. Many industries from medicine to military apply machine learning to extract relevant information. The purpose of machine learning is to learn from the data. Today each and every person is using machine learning knowingly or unknowingly. From getting a recommended product in online shopping to updating photos in social networking sites. This chapter gives an introduction to most of the popular machine learning algorithms.

Chapter 4

SIMULATION RESULTS

Having stated the fundamentals of prediction models, this section presents the experiments conducted to verify hypotheses regarding their performance. It is important to underline the fact that improvement is traceable as long as there is a sufficiently defined evaluation function. In the problem of disease prediction, there are several aspects to be measured. Namely, prediction shall be accurate.

Although immense progress in the field has been seen lately, there is still notable dependency on computational power when applying changes to a model. Thus, when drawing up this thesis it was right away determined that all possible experiments triggering with weights in the model are, unfortunately, infeasible if one seeks to enhance model's performance significantly using an average computer.

4.1 Performance and Accuracy

The comparison of the usage frequency and accuracy of different supervised learning algorithms are shown in Table 4.1. It is observed that Random Forest has been used most frequently. This is followed by Naïve Bayes. Decision tree has better accuracy for kidney and cancer disease.

Table 4.1 Accuracy of different supervised learning algorithms.

Algorithm	Disease	Accuracy (%)
Decision Tree	Diabetes	71
	Heart	62
	Liver	62
	Kidney	98
	Cancer	92
Random Forest	Diabetes	77
	Heart	77
	Liver	71
	Kidney	98
Naive Bayes	Diabetes	74
	Heart	76
	Liver	59
	Kidney	98
	Cancer	94

4.2 Outputs

This portion shows the output of the Web App designed using python library Flask and the ML model which has higher accuracy for each disease. This app takes symptoms as

user input for each disease and predicts whether these symptoms have chances of having the disease or not. This app also shows the precaution must be taken to prevent from having the disease.

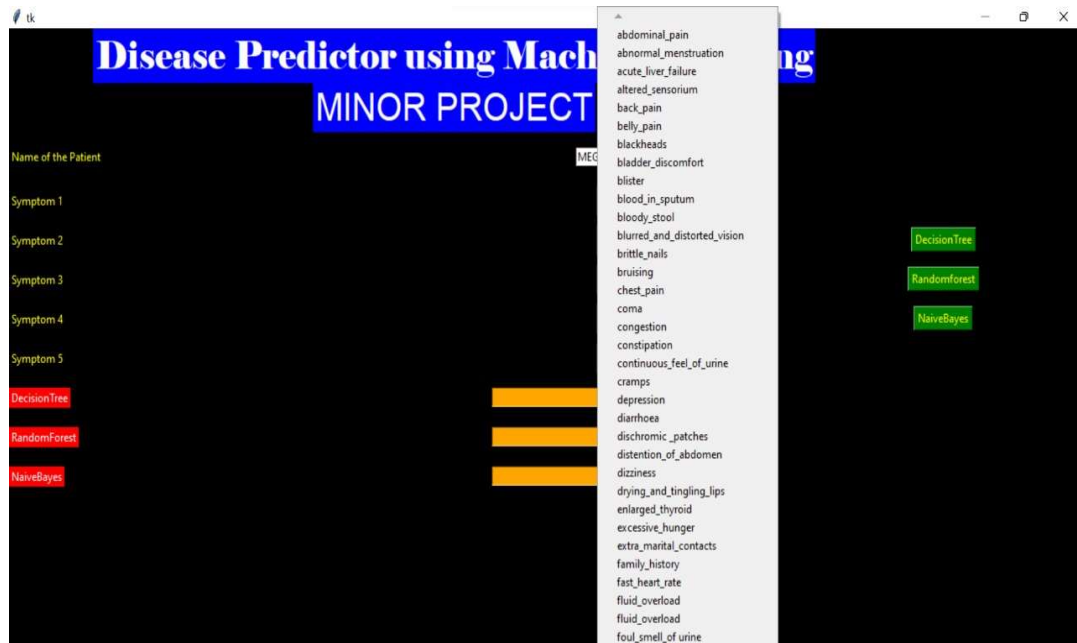


Fig 4.1. Selecting the symptoms

In the above screenshot, user is given a list of symptoms from which they can select their five symptoms.

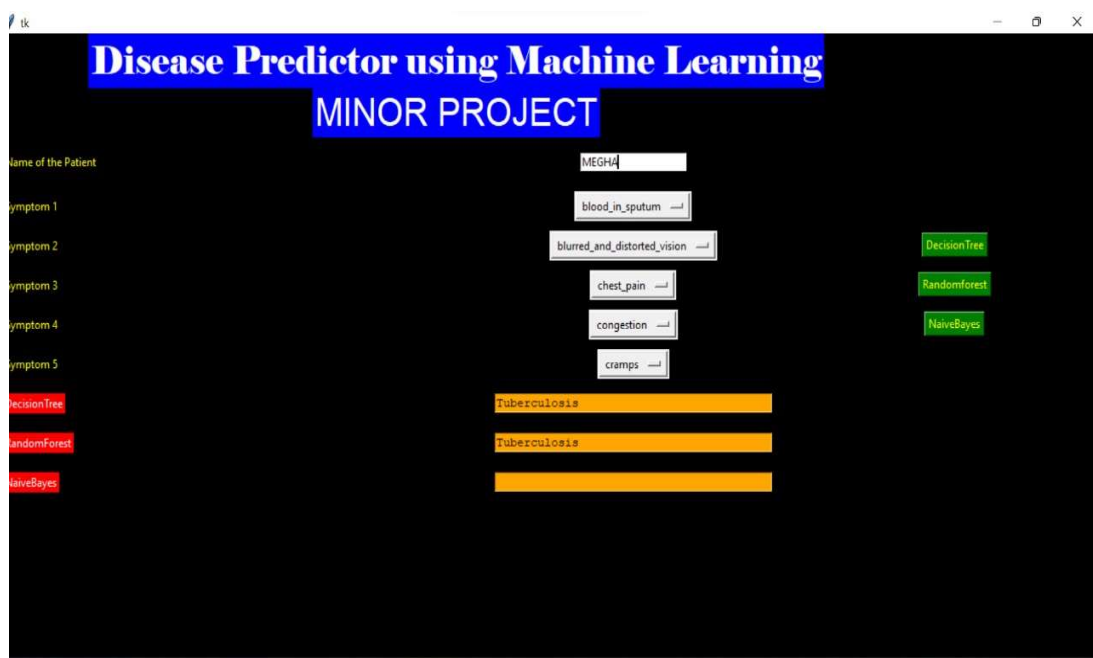


Fig 4.2. Selecting the algorithms to predict the disease

In the above screenshot, the user is selecting the algorithms to predict his/her disease.

The system can predict disease using any of the three algorithms.

Disease Predictor using Machine Learning
MINOR PROJECT

Name of the Patient: SHIVANG

Symptom 1: back_pain

Symptom 2: blister

Symptom 3: blurred_and_distorted_vision

Symptom 4: brittle_nails

Symptom 5: constipation

Algorithms: DecisionTree, RandomForest, NaiveBayes

Predicted Disease: Impetigo

Fig. 3. Predicting the disease using all the algorithms

Disease Predictor using Machine Learning
MINOR PROJECT

Name of the Patient: YASHI

Symptom 1: blurred_and_distorted_vision

Symptom 2: back_pain

Symptom 3: cramps

Symptom 4: dizziness

Symptom 5: excessive_hunger

Algorithms: DecisionTree, RandomForest, NaiveBayes

Predicted Disease: Cervical spondylosis

Fig.4. Showing the accuracy of the algorithms (all the algorithms prediction of the disease)

Chapter 5

CONCLUSION AND FUTURE SCOPE

This project provides a valuable learning experience. First, the data analysis of all the disease datasets and cleaning of it is more complicated than we expected. We learned how to use one of the most popular deep learning frameworks Keras through the project. Second, mathematics and the knowledge of particular software architecture are equally important for the success of the project. Although the web version of the model is implemented very early before the deadline of the project, plenty of time is invested in implementing the model.

Third, working in a team, we could discuss and refine a lot of initial ideas. We could also anticipate problems that could become critical if we were working alone.

5.1 Conclusion

Machine learning can be applied to health care data to develop robust risk models. Healthcare industry is already burdened with the exploding population and lack of trained doctors. Normally the doctor to patient ratio in India is 1:1700 which is far higher than the recommended ratio of 1 in every 100 patients by WHO. Especially during Covid times when people cannot go to hospitals and India has been in shortage of doctors, nurses and beds. The spontaneous increase of efficient healthcare workers is not possible, therefore use of machine learning and artificial intelligence technologies can enhance the productivity and precision of the existing ones and also reduce the healthcare expense.

COVID-19 is a contagious viral infectious disease that has affected a broad spectrum of the global population. The high transmissibility and pathogenic nature make the early detection of infected individuals vital for commendable fighting COVID-19.

Diseases like heart, liver, cancer, kidney are one of the most devastating and fatal chronic diseases that rapidly increase in both economically developed and undeveloped countries and causes death. This damage can be reduced considerably if the patient is diagnosed in the early stages and proper treatment is provided to her. In this report, an intelligent predictive system based on contemporary machine learning algorithms for the prediction and diagnosis of these diseases is developed. Diseases were predicted by using different types of data mining and machine learning techniques (decision tree, KNN, neural network, SVM, Naïve Bayes) to predict the occurrence of disease.

Determine the prediction performance of each algorithm and apply the proposed system for the area it is needed. Use more relevant feature selection methods to improve the accurate performance of algorithms. Table 5.1 shows the conclusion on the importance and limitations of these algorithms.

Machine Learning Algorithm	Benefits	Assumptions and/or Limitations
Decision Tree	<ul style="list-style-type: none"> • Easy to understand and efficient training algorithm • Order of training instances has no effect on training • Pruning can deal with the problem of overfitting 	<ul style="list-style-type: none"> • Classes must be mutually exclusive • Final decision tree dependent upon order of attribute selection • Errors in training set can result in overly complex decision trees • Missing values for an attribute make it unclear about which branch to take when that attribute is tested.
Naive Bayes	<ul style="list-style-type: none"> • Foundation based on statistical modeling • Easy to understand and efficient training algorithm • Order of training instances has no effect on training • Useful across multiple domains 	<ul style="list-style-type: none"> • Assumes attributes are statistically independent • Assumes normal distribution on numeric attributes • Classes must be mutually exclusive • Attribute can class frequency affect accuracy
	<ul style="list-style-type: none"> • Tolerant of noisy inputs • Instances can be classified by more than one output 	<ul style="list-style-type: none"> • overfitting • Optimal network structure can only be determined by experimentation
Support Vector Machine	<ul style="list-style-type: none"> • Models nonlinear class boundaries • Overfitting is unlikely to occur • Computational complexity reduced to quadratic optimization problem • Easy to control complexity of decision rule and frequency of error 	<ul style="list-style-type: none"> • Training is slow compared to Bayes and Decision trees • Difficult to determine optimal parameters when training data is not linearly separable • Difficult to understand structure of algorithm

Table 5.1 Conclusion on importance and limitations of supervised algorithms.

5.2 Future Scope

Of course, this is just a first cut solution and a lot of modifications can be made to

improve this health app especially many more features can be added. The high accuracy obtained may be a cause of concern since it may be a result of overfitting. This can be verified by testing with new data in the near future. The potential of deep learning is so tremendous that it can take a big stride in a future to the so far challengingly difficult area of anomaly prediction from the anomaly detection. This can be achieved if large amount of data is made available to research community from doctors and hospitals. There is lot of research in use of more optimization techniques, feature selection algorithms and classification algorithms to improve the performance of the predictive system for diagnosis of the diseases. The Covid-19 outbreak has had a profound effect on the well being of the people worldwide. In the number disease related to death continuous to grow globally. The Data learning is just one possible successful way to provide promising data driving solution to help the humanity to handle with Covid-19. Let us look the next possible direction of our health app.

Our system will be having it's website which would contain various other functionality like health tracker, availability of the hospitals, etc.

- ❖ Electronic health records consist of entire medical and health data in a single system to ensure data availability and accessibility. Our models can be trained using data set from 1 EHR and can be utilized to predict an outcome from another system.
- ❖ Our system will be able to track in real time the availability of hospital beds, list of oxygen suppliers and will have an updated list of state help line numbers and resources by collecting the data from the official website of the Government.

REFERENCES

- [1] S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, 2019.
- [2] D. Dahiwade, G. Patle and E. Meshram, "Designing Disease Prediction Model Using Machine Learning Approach," 2019 3rd International Conference on Computing Methodologies and IRJET- Disease Prediction using Machine Learning Communication (ICCMC), 2019.
- [3] Uddin, S., Khan, A., Hossain, M. *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak* 19, 281 (2019).
- [4] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. *SN COMPUT. SCI.* 1, 345 (2020).
- [5] Divya Jain, Vijendra Singh, Feature selection and classification systems for chronic disease prediction: A review ,Egyptian Informatics Journal, Volume 19, Issue 3, 2018, Pages 179-189
- [6] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, 2019.
- [7] Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis by Gopi Battineni ,Getu Gamo Sagaro ,Nalini Chinatalapudi and Francesco Amenta Center for Telemedicine and Tele pharmacy, School of Medicinal and Health Sciences Products, University of Camerino, Via Madonna Della carceri 9, 62032 Camerino, Italy.
- [8] GDPS - General Disease Prediction System, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 03, Mar-2018.
- [9] Rayan Alanazi, College of Science and Arts in Qurayyat, Jouf University, Sakakah, Saudi Arabia . Identification and Prediction of Chronic Diseases Using Machine Learning Approach. Published 25 February 2022.
- [10] Khalid Twarish Alhamazani et al., Implementation of Machine Learning Models for the Prevention of Kidney Diseases (CKD) or Their Derivatives. Published 30 December 2021.
- [11] Dr. P. Hamsagayathri, Assistant Professor, Dept. of ECE, Bannari Amman Institute of Technology, Sathyamangalam, Erode, Tamilnadu, India. Symptoms Based Disease Prediction Using Machine Learning Techniques (ICICV 2021)
- [12] <https://www.semanticscholar.org>
- [13] <https://www.geeksforgeeks.org/understanding-logistic-regression>
- [14] <https://www.geeksforgeeks.org/random-forest-regression-in-python>

[15] <https://www.geeksforgeeks.org/decision-tree>