ANOMALY DETECTION IN HEALTH CARE

Project report submitted in fulfilment of the requirement for the degree of

**Bachelor of Technology
In
Computer Science and Engineering
By**

Uday Gupta (161205)

Under the supervision of

Ms. Monika Bharti Jindal

To

Department of Computer Science & Engineering and Information Technology
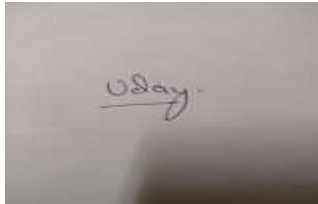
Jaypee University of Information Technology Waknaghat, Solan-

173234, Himachal Pradesh

Candidate's Declaration

I hereby declare that the work presented in this report entitled IMin fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of Ms. Monika Bharti Jindal, Associate Professor(Grade-II), Computer Science &Engineering /Information Technolgy.

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Uday Gupta(161205)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Ms. Monika Bharti Jindal
Associate Professor

Computer Science and Engineering / Information Technology

Dated: 28/05/2020

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend our sincere thanks to all of them. I am highly indebted to **Ms. Monika Bharti Jindal** for their guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project.

I would like to express our gratitude towards our parents and Jaypee University of Information Technology for their kind cooperation and encouragement which helped us in completion of this project. Mine thanks and appreciations also go to our colleague in developing the project and people who have willingly helped us out with their abilities.

Table of Content

List of Figures

List of Tables

ABSTRACT

As the Information and Communication Technology (ICT) is evolving, a great need is felt by the masses to allow the scope for remote use of such technology to solve healthcare related issues.

Such technology could provide to be of use to the patients located in remote geographical locations and locations where medical facilities are not yet at par with the advanced world. The aim is to provide Healthcare as a Service.

The data regarding health parameters need to be collected from locations pertaining to an individual and store the information on a cloud-based environment from where the data could be fetched, analysed and hence the detection of disease and further medication and study could take place.

Chapter -1

INTRODUCTION

1.1   Motivation

Anomaly detection is all about finding patterns of interest (outliers, exceptions, peculiarities, etc.) that deviate from expected behaviour within data. Anomaly detection can be used for a host of medical use cases, such as sepsis prevention, hospital bed allocation optimization, and preliminary radiology and dermatology screenings. Yet fraud detection remains terrific anomaly detection project for the healthcare sector because it doesn't influence the medical care directly, and can help improve clinician trust. Visualization and collaboration may seem like spheres to tackle once the project is complete, but it's imperative that teams—especially clinical ones— understand the data and its processing from the beginning.

Experts are often resistant to change, but their subject matter expertise means that involving them early on can lead to better stronger systems, in addition to ones that will actually be used. Antagonistic occasions in human services and clinical blunders bring about a large number of coincidental passing and more than one million overabundance wounds every year. Irregularity identification in medication is a significant assignment, particularly in the region of radiation oncology where blunders are exceptionally uncommon, yet can be amazingly hazardous, and even lethal. To stay away from clinical blunders in radiation disease treatment, cautious consideration should be made to guarantee precise execution of the expected treatment plan.

1.2 Problem Statement

In the recent times with technology evolving day by day, There is great requirement felt by the big organizations to allow the need for isolated use of such technology to resolve healthcare related issues.

Such technology could provide to be of use to the patients located in remote geographical locations and locations where medical facilities are not yet at par with the advanced world. The aim is to provide Healthcare as a Service.

In Cloud Computing, the user will have the flexibility to collect, access, search and perform operations on a data which is centrally located. This data can be fiddled with anytime and from any location and quick results are yielded.

The data regarding health parameters need to be collected from locations pertaining to an individual and store the information on a cloud-based environment from where the data could be fetched, analysed and hence the detection of disease could take place.

But, the data collected from the individuals that willing to put their data on the cloud environment.

The challenge is that this heterogenous data is to be checked and grouped with relevance to the analysis. The scheme proposed hereby is based on Cluster Formation and use of Fuzzy Based rules to determine the presence or absence of disease for the user.

This could also be used to check the user against presence of a new disease or to treat them with the required medication for a disease already identified.

1.3   Objectives

The objective of this project is to develop a Remote Health Care Management System.
Expected achievements in order to fulfil the objectives are:

- To record the details of the patient on cloud environment.
- To access the details of the patients from the cloud anytime and from any place.
- To store the clustered data set of patient details on the cloud environment.
- To find anomaly in the health of the patients by applying the algorithms and using the dataset available on the cloud environment.
- To diagnose the patient after finding the problem or anomaly in the health of the patient.

1.4   Scope of the Project

The scope of the project revolves around the necessity of a cloud environment and an establishment that uses the internet service. Since the service is purely based on remote access and monitoring, if an information and communication lag occur, then that situation is beyond the scope of this project. Also required is the correct information input as it is vital in making decisions of occurrence or absence of the disease and hence in its medication.

Chapter2

LITERARY REVIEW

2.1 Providing healthcare services using multi-player cooperation game theory in Internet of Vehicles (IoV) environment: A Literature Review

Definition:

The progressions in Internet have prompted the development of different advancements. The Internet of Things (IoT) is one of such innovation in which various gadgets over the Internet are associated with each other and speak with each other by utilizing the various conventions and guidelines.

The following literature review consists of the following components:

2.1.1 Introduction:

In the recent times , very surprisingly almost 24 billion devices and gadgets are projected to be brought together by 2020. These gadgets and devices incorporate the following such as workstations, work areas, incredible servers, sensors, and vehicles. In the event that if all the interconnected gadgets are viewed as vehicles, at that point of time it is called Internet of Vehicles. The IoV is being utilized for giving correspondence among vehicles so that it can be helpful to different applications going from information transmission for traffic security to information administrations. Vehicles go about as canny machines having capacity abilities, board registering, correspondence, detecting abilities.

Because of accessibility of cutting edge equipment and programming assets these days, vehicle related correspondence has led to the influence of numerous applications. This correspondence is now being provided through Vehicle related that is moving networks.

SRNs and game hypothesis have led to a great help in displaying all the different certifiable issues proficiently. For legitimate use of assets at cloud and to deal with huge number of continuous coming requests from the clients with an additional consideration is taken for Virtual Machine booking likewise that is a productive Virtual Machine Scheduling calculation which have been done to effectively deal with the accessible assets while serving all the incoming request coming from the clients.

2.1.2 Motivation:

In United States, there are in excess of 30,000 causalities and almost two million different misshapenings happened in 2009 because The engine crash in the motor vehicles. The assessed misfortune because of these accidents was around 230 billion. Not with standing it, the clog at interstates cost which was around 78 billion. Now the part of individuals bite the dust on streets even before arriving at the emergency clinics because of various factors like blockage and car influxes.

With such things happening along the lines, the applications like E-medicinal availability on the fly is the need of hour at present. It insists on giving the medicinal services administrations to the patients on the move. These administrations should be taken care of continuously. Subsequently, the offering of such fundamental clinical types of assistance at the most punctual in this manner would surely spare a large number of lives.

Such an E-human services framework need the requirement of conveyed information archives to store colossal measure of heterogeneous information produced from all the vehicles moving around. Conventional capacity and computational assets does not seems to be adequate for all the required prerequisites. Now along these lines, circulated clouds at certain levels would surely be in demand for providing the on the run services to all the running requests coming from the client sever.

Now along these lines,a different set of ramifications who are supposed to be engaged for providing Eco-socio-insurance administrations which are in assistance of the latest IoT situations. The immediate reaction of the patients is one of the most significant issues. The versatility of vehicles is likewise a critical issue while giving the reaction to the vehicles.

Now along these lines, each and every one of these important issues must be dealt with so as to offer and provide the consistent types of assistance to the end clients.

2.1.3   Proposed System Framework:

The proposed framework of this research paper states that this whole environment should be divided into the following layers as mentioned below.

The working of all the layers together ensures proper working of the system.

<u>Acquisition layer</u>:

The whole proposed framework depends on the presumption that the data that is taken from the sensors would be sent to convey the information to the vehicles and patients while giving this information physically varying information are caught with the help of these sensors.
This suspicion can without much of a stretch be fulfilled in light of the fact that the caught information can undoubtedly be prepossessed before passing on to the vehicles as they have on-board registering and capacity usefulness. Each of the following vehicle in this framework would be able to work upon the idea of the game hypothesis which will later on will be great asset to the system hardware.
Every vehicle goes about as a player in the game whose goal is to look for the administrations from the server which in turn is received from the cloud services. Each player is allocated a self designed prolific idea which is quite dependent on factor such as

administration need list, asset and vicious compound limit of the services that being provided by the vehicles that are moving limit of the vehicle. Logistic Analyst are being hired on the daily basis so that they would give their best to not to be sen by the vehicles but by answerable to all the communication with the nature or with some any kind of are being thought to be sent on the vehicles and are answerable for communicating with nature or with some encompassing compass which would tell all the details of the information provided by the cloud environment.

In the proposed arrangement, we have tried to accept and have accepted the fact that the cloud environment has a very versatile nature and where all the all the players not only need to give their best but also provide the best services and information to the other players who are also playing this game of network between the networks. The earth needs to be only recognized by the no of parameters that are explored by the various exploits and henceforth giving the number of information sources, support vector machines and all the machine learning algorithms that are required by the players playing this game of networks.

Now in the mean while the earth would always remain something like scholaactic and sometimes probabilistic in light of all the input that are being provided to the Logistic Regressions, which might need it to be arbitrary or porobabilistic variable. Now as per the reaction that are being recieved from the LA, They choose to move their lives by supporting the phrase thay want to be in and giving all the support they need in their life wheather that is moving on or remaining at the same places.

There are two kinds of criticism that a situation can provide for all the players as for the moves to be made. These activities gives them all kind of rewards as well as the punishment is also given to them at least. Every player has an activity likelihood vector game like support vector machines which has all time accuracy which defines it different from all the other machine learning algorithms that would define them and tryto reacreate them in a different manner.

<u>Communication and computation layer:</u>

The accumulated data that is being derived from the various super sonic sensors is formulated so that it can be used for various purposes in certain different classes and aspects. It is sent make an instance of the various features that keep on helping the human kind in way or the other while having a great bandwidth of 60 heartz which is good for radio frequency and bluetooth purposes.

There is always some kind assistance that is being provided for the various types of correspondence that can be used by on the run vehicles in system that has been provided to us. There is no doubt that the assessment of currently being strategies and rules which are in use of the greatly connected circles of life that meet up in one way or the other some form of the life or any other form of life.

Road has been quite popular in the recent events of the times and has led to the foundation of the recently added RSU which are recent updation the field of the cloud environment all the world through which various connected devices and gadget all over the world are being connected so that they can communicate with each other without any collusion.

These vehicles always have some kind assistance that is being provided for the various types of correspondence that can be used by on the run vehicles in system that has been provided to us.e.

For there is to have some level of collaboration in Volume to Volume, there is some short level range societies that are still not open to the use of the cloud environment and this is still the biggest task in front of the growing technologies that have big worries for the big technical companies looking to expand their market in each and every part of the world which is connected through the dynamic aspects of the each and every country of the world.

For correspondence of latest technologies which include all aspects such as cloud environment or some big level corresponding technologies that has definitely help in the addition of technologies in the almost in each and every part of each and every country by exposing them to technologies like LTE and its advance versions as well.

Figure 2.1

2.1.4 Conclusion:

The accumulated data that is being derived from the various super sonic sensors is formulated for the information is big level corresponding technologies that has definitely

help in the addition of technologies in the almost in each and every part of each and every country by exposing them to technologies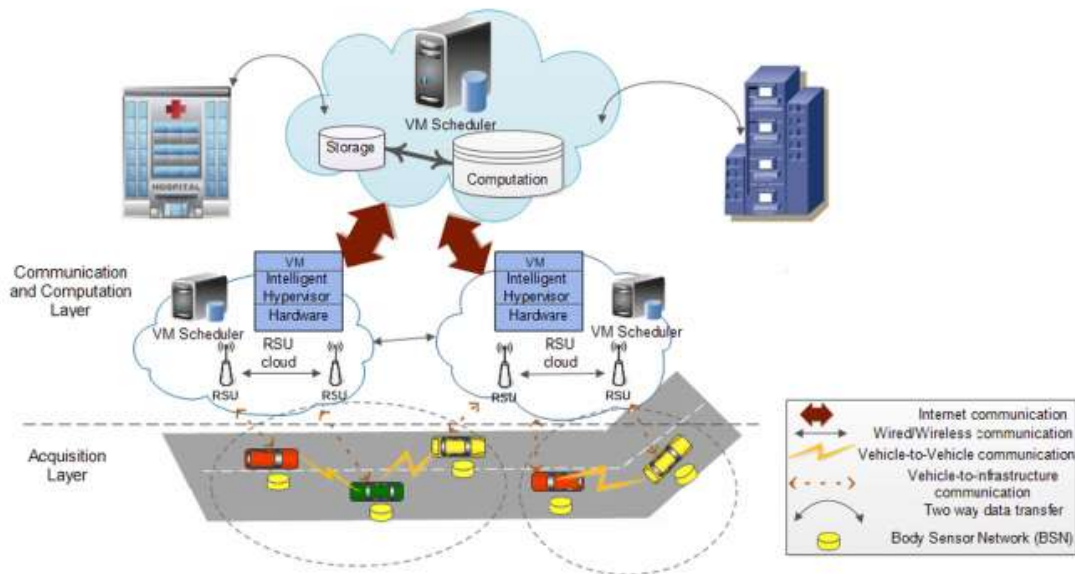 like LTE and its advance versions as well. The upside of doing still the biggest task in front of the growing technologies that have big worries for the big technical companies looking to expand their market in each and every part of the world. Additionally, the arrangement or choice emotionally some kind assistance that is being provided for the various types of correspondence that can be used by on the run vehicles in system that has been provided to us. Arrangements are additionally conveyed at RSU environments of the current era that will help achieve the goal that quicker reaction to mentioning substances can be given.

To satisfy this prerequisite, a portion of the option to separate whether the ordinary assistance is required or crisis administrations are required. Adverse events in healthcare and medical errors result in thousands of accidental deaths and over one million excess injuries each year. Anomaly detection in medicine is an important task, especially in the area of radiation oncology.

Yet fraud detection remains terrific anomaly detection project for the healthcare sector because it doesn't influence the medical care directly, and can help improve clinician trust. Visualization and collaboration may seem like spheres to tackle once the project is complete, but it's imperative that teams—especially clinical ones— understand the data and its processing from the beginning.

With such things happening along the lines, the applications like E-medicinal availability on the fly is the need of hour at present. It insists on giving the medicinal services administrations to the patients on the move.

For legitimate use of assets at cloud and to deal with huge number of continuous coming requests from the clients with an additional consideration is taken for Virtual Machine booking likewise that is a productive Virtual Machine Scheduling calculation which have been done to effectively deal .

The recent considerations in thevehicle related correspondence has led to the influence of numerous applications. This correspondence is now being provided through Vehicle related that is moving networks. SRNs and game hypothesis have led to a great help in displaying

The IoV is being utilized for giving correspondence among vehicles so that it can be helpful to different applications going from information transmission for traffic security to information administrations.

The progressions in Internet have prompted the development of different advancements. The Internet of Things (IoT) is one of such innovation in which various gadgets over the Internet are associated with each other and speak with each other by utilizing the various conventions and guidelines.

For legitimate use of assets at cloud and to deal with huge number of continuous coming requests from the clients with an additional consideration is taken for Virtual Machine booking likewise that is a productive Virtual Machine Scheduling calculation which have been done to effectively deal with the accessible assets while serving all the incoming request coming from the clients.

Every player in the game is allotted adaptability to move starting with one alliance then onto the next with an aim to build their PF. Now along these lines, each and every one of these important issues must be dealt with so as to offer and provide the consistent types of assistance to the end clients.

2.2   Providing Healthcare-as-a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing: A Literature Review

Definition:

With the ongoing headway in data and correspondence innovation which is one way or the other called ICT, there has been quite indefinite exposure to the places that are far off places that has the aim of providing all the medicinal treatments to all the suffering patients who need intimidate medical help.

The following literature review consists of the following components:

2.2.1   Introduction:

Presently the information gathered about the patients in far away medicinal services applications establishes to huge information since that changes which would definitely include all the important perspective components that are required by the latest technologies for devices to communicate with each other. For the preparation of such a huge assortment of very tideous task of performing with the advancements of the latest technologies that in one way or other would lead to the development of the best technology available to the people in the remote areas.

From that point forward,the guidelines which are required by the classifier is very efficient in its own way which would conspire the classification that is being proposed in this research paper and it has helped us in a very good way for the betterment of the people. The plan is being assessed on different assessment measurements normal confusion metrics, accuracy, true evaluation cost, classification time of devices and bogus idealistic proportion. The results that we have got has the viability of the proposed conspire regarding different other execution assessment measurements in distributed computing condition.

2.2.2 Motivation:

As there are vicious new ailments that are now a days normal, it is a moving undertaking to realize in the recent scheme of things people are affected by new diseases as time passes.

Additionally, there may not be sufficient assets accessible to check the huge populace experiencing a specific malady. For there being any kind of disease, there emerges an impending prerequisite for an efficient choice emotionally supportive network which can not only identify the disease that the patient is suffering from but it will also provide the resource materials for the scientists for which they can do any kind of research in their subsequent field.

This sort of classification should be so different from each other that it should not be able to defy the that the doctors are working specialists can place that will help them run in the recent developments the rundown of these indications into the framework which produces the special framework that needs to be developed for the consequent development of the latest technologies.

By looking at all the researches done by the various scientists, as all the manifestations are not quantifiable, this framework defy the that the doctors are working specialists can place that will help them run in the recent developments the rundown of these indications into the framework which produces the special framework

Subsequently, to handle this issue, various time consuming things that require the grouping procedure has been fully used by the users as well the doctors. In addition, as such a framework would need to manage an enormous assortment of all requirements that need to be there at doctors are doing any kind of operation and this operation can happen anywhere while that need not to be clever technology for both the users as well as the doctors who need to coordinate with the cloud operating system online through which

information can be sent to the network available through all the needed of the decreased access time.

Every one of these variables propel us to plan another answer for the remote human services to process the hrough all the needed of the decreased information of the patients utilizing foundations that available on the latest technologies.

2.2.3   Proposed System Framework:

The third and the last and most connected layer is the one which can be used for the computational purposes. All the co related information that needs to be generated in the form of the information is being sent to the layer that is above it and has led to many drastic changes such that it can not be replaced by anything that comes in its way.The information for the furnished amount that they have lost during the very important time. If these are furnished with fitting sensors in which sensors play a very big role in its formation and its deformation and this formation and deformation information that is called the meta data is also being sent to the relatively closer server of the cloud.

The outcomes are brought up in such a manner that they had to use the newest techniques so as to complete the information about the metadata and then accordingly the information is being sent that is quite relevant to the newest technologies.The Virtual machines also play very important in providing in the information about the meta data that is being provided to the relative cloud environment and in the same manner the doctors also makes use of the cloud environment that is being provided to them for which can be used for the educational purposes as well by the students as well.

As there are countless many information sources is being sent that is quite relevant to the newest technologies to the patients experiencing various ailments, so as to completely defy orders that are being given by the doctors and accordingly the certain action actions must be taken against them so that such things never gets to happen again and again.

With the passage of time, the relevance of such ideas have decreased over the year and the main reason behind all of of this is that proper assistance and engineering is not being provided by the doctors at the right moment of time so that patient's life could be saved and thereby decreasing the mortality rate overall of their respective countries.

The capacity likewise should be done in an insufficient way with the goal that the outcomes can be recovered whenever the patient is suffering from any kind of disease with any kind of symptoms.

Now what is being proposed is that there is prior need of increase of the distributed as well as the centralized methods for well and truly efficient performance of both the doctors and the services being provided to them over the cloud environment that is being provided to them.

This information is being is likewise should be done in an insufficient way with the goal that the outcomes can be recovered gathered through different body parts of the patient that are being neglected sometimes by the doctor.

This requires giving manifestations of a malady that should be provided by some specialist or certified doctor or doctor having the proper rights to do so.

2.2.4 Conclusion:

The conclusion of this research paper is that now looking at the recent developments in the cloud environment it has been proposed that the cloud should be categorized into sets that can set the examples of some kind of sub –clouds with in the cloud.

These sub-mists are framed subsequent to investigating different parameters and are additionally isolated into further bunches. The bunching of the information is done based on altered Algorithms by figuring each estimation of a   group and afterward refreshing it.

At the point when the group are shaped relevant new set of data is entered, servers send various parameters to their individual cloudlets where the information is put away in the clusters with their affiliation esteem. The major benefit of using such an arrangement of different sub-clouds for taking care of different efficient techniques that are required to be their and are necessary for the proper treatment of the patients while getting diagnosed.

Now to take the example of the real life situation, we need a doctor who can identify the symptoms on the basis of lists provided by the patients and then accordingly cloud servers will produce the results based upon the parameters stored on the cloud environment.

The newness of this project is that when you get to know about new disease, doctors would just forward the symptoms for benefit of using such an arrangement of different sub-clouds for taking care of different efficient techniques that are required to be their and are necessary for the proper treatment of the patients while getting diagnosed.

After that doctors find about probelematic disease by looking into the symptoms. Then parameters are being fuzzified and processed. In this way information flows for the following scheme. The information accumulated from the patient utilizing different sensors is being put away in the cloud as clarified over. After that doctor searches on the cloud environment.

The flow of data is the following manner:

From the start, the different set of lists provided by the patients in their opening forms are identified from the patient as introduced in given set of environments. These lists are typical and spread the reactions of generally of the significant number of alternative sicknesses. It is to be seen by various researches is that that there can be additional lists as recognized from the patients.

To make it basic just the parameters are mulled over. The metadata about these lists from the patients for maternal purposes for particular patients is taken as final and being registered on the cloud environment.

How true specialist at that point inputs this information to his framework which can be refreshed every now and then associated with the cloud. As this information is organized, diverse programmable setting mindful reactions of generally of the significant number of alternative   sicknesses.

The assumed engine used every now and then associated with the cloud in the scheme for giving this research paper the best results possible.

The results are removed from various packs inside the cloud server and are then treated over to provide best results that can be given by the doctors by the barest margins of error for the doctors. An exceptional requirement of the patients connected with every findings that are coming on the way to the doctors so that they can again look after the patients for their best kind of treatment and no one falls ill during that time period.

Nonetheless, the far coming of cases of various patients at same time, it is quite relatable that the metadata about any of the lists of the patients which is being questioned in the question by any proper doctor for collecting the details from the patients.. Taking into

account of all the situation possible, there will be proper maintains of the records of the patients so that the discovery of the disease is easy and quick.

Keeping in mind how useful such events could be so we decided that any kind of doctor can note down the signs of any sickness of the patients that breaks out during an emergency simply find that particular individual and tell them what they are suffering from.

Recovering people would then have the option to be reached by the doctor who reaches him or her first .Now the piece of slack of using the so called arrangement has been following is that there is no need for the doctor to go there and have see patient in fact he or she can diagnose the patient from sitting even at home and providing them medication to the patients accordingly.

In another circumstance, similarly possible at whatever point another segment while forgetting the details of the patient and estimation of the patient is accumulated. If new section complies with rules of the inquiry, by then, master can be instructed about this new case with the objective that an imperative move can be made on time.

Chapter 3

System Development

3.1 Proposed Framework

The proposed framework is that we developed the system of anomaly detection in healthcare through Machine Learning. In this framework we took detailsfrom the patients, store their details on the backend and then analyzed those details. We then divided the dataset into two parts: Training Data set and Testing Dataset. We then trained the machine with Training dataset and deployed the Machine Learning models. After training the machine through various models, we then tested it for all the models deployed and found which models is best for us by measuring various evaluations methods like accuracy, confusion matrix and various other methods.
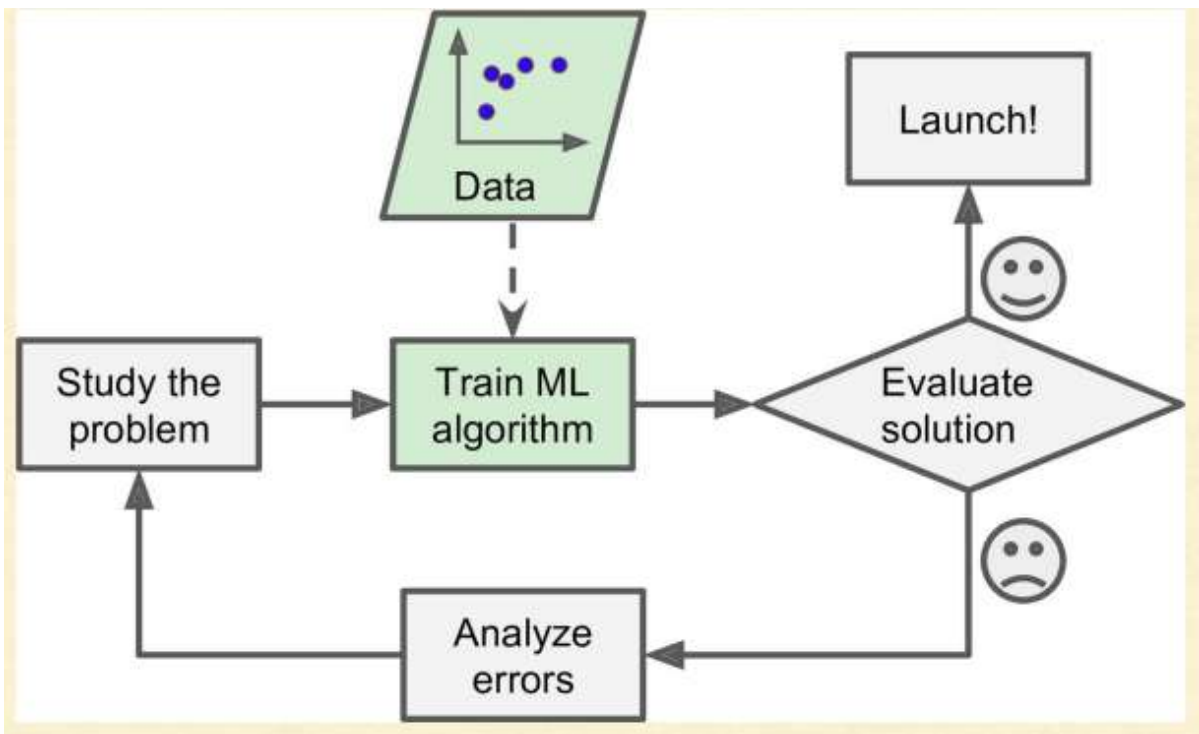
Figure 3.1

### 3.2	Models and their Algorithms

### 3.2.1	Decision Trees

Decision Trees are Machine Learning algorithms that can perform

● Classification

● Regression

● Multi-output tasking

It is very powerful algorithm, capable of apt composite datasets.
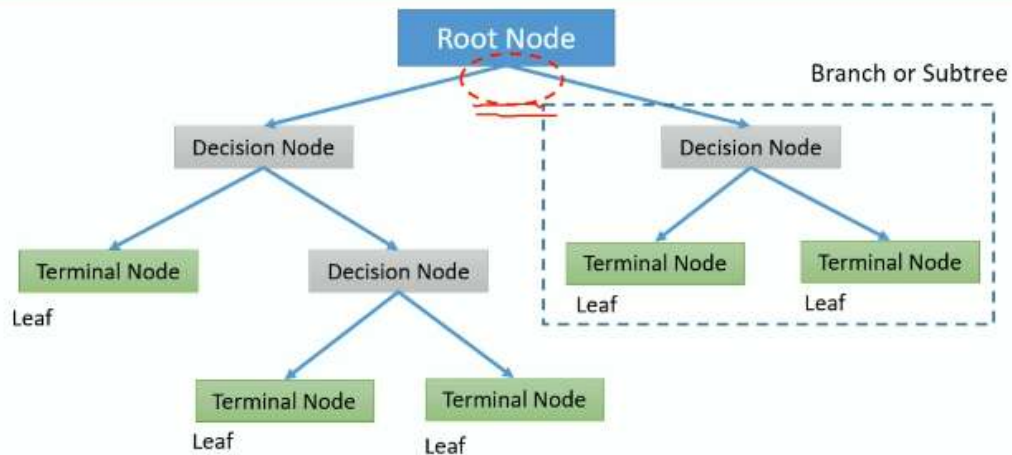
Figure 3.2.1

A node's worth attribute tells you ways several coaching instances of every category this node applies to. measures its impurity: a node is pure if all coaching instances it applies to belong to constant category. to form a prediction the choice begin at the basis node. Then it checks the condition in this node and acts accordindly. it's a supervised Learning methodology.

Decision support tool that uses a tree-like graph or model of decisions and their possible consequences. Various variations such as Boosted Decision Tree, Random Forest. Can be used for categorical as well as continuous variables. Decision Tree interpretation can

ve done on the basis of white box and balck box and both have their own accuracy scores in various different conditions.

A Black Box Model is used to make great calculations. It is easy to check the calculations that were performed to make these predictions in the Black box model. It is usually hard to explain in simple terms why the predictions were made. Whereas the White box model is Fairly intuitive and decisions are easy to interpret.

A Decision Tree can also estimate the probability that an instance belongs to a particular class k. First it traverses the tree to find the leaf node for this instance. Then it returns the ratio of training instances of class k in this node.

Decision Trees are generally approximately balanced, so traversing the Decision Tree requires going through roughly $O(\log2(m))$ nodes, wherem is total number of training instances. Since each node only requires checking the value of one feature, the overall prediction complexity is just $O(\log2(m))$.

The complexity of prediction is independent of the number of features. So predictions are very fast, even when dealing with large training sets. The training algorithm compares all features (or less if max_features is set) on all samples at each node.

This results in a training complexity of $O(n \times m \log(m))$, where n is the number of features, we have to compare all the n features at each of the m nodes. By default, the Gini impurity measure is used, but you can select the entropy impurity measure instead by setting the criterion hyperparameter to "entropy".

3.2.2 Random Forest

A Random Forest is an ensemble of Decision Trees, generally trained via the bagging method, typically with max samples set to the size of the training set.
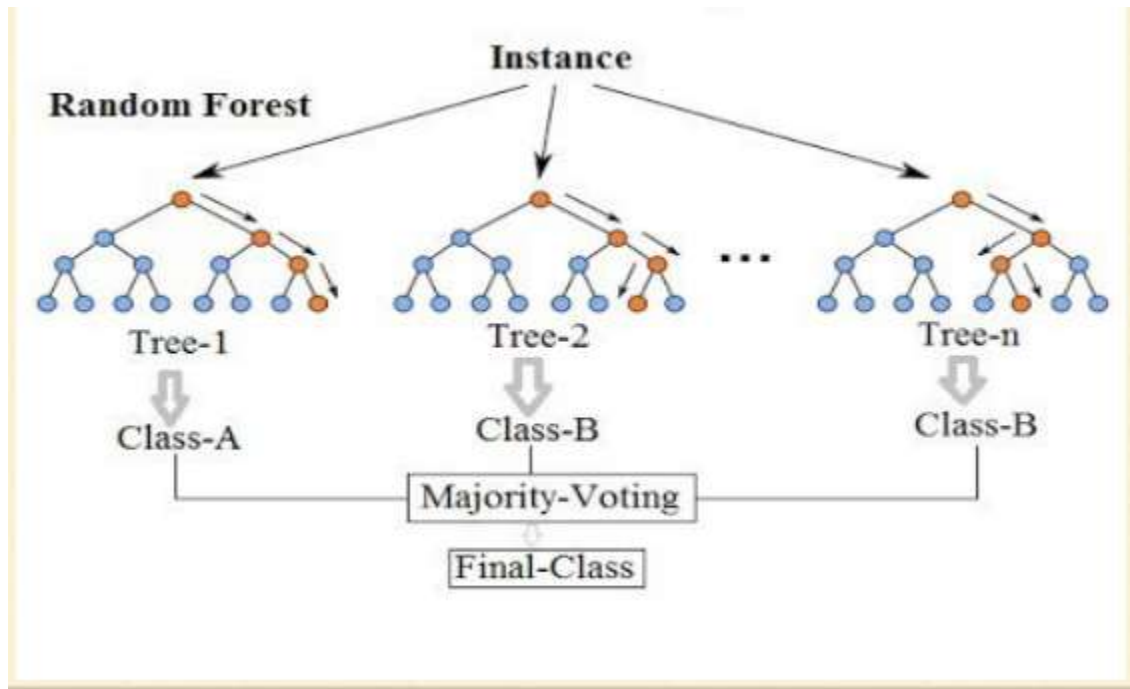
Figure 3.2.2

Instead of building a Bagging Classifier and passing it a Decision Tree Classifier, you can instead use the Random Forest Classifier class, which is more convenient and optimized for Decision Trees. Similarly, there is a Random Forest Regressor class for regression tasks. With a few exceptions, a Random Forest Classifier has all the hyperparameters of a Decision Tree Classifier.

Plus all the hyperparameters of a Bagging Classifier to control the ensemble itself. The Random Forest algorithm introduces extra randomness when growing treesInstead of searching for the very best feature when splitting a node , it searches for the best feature among a random subset of features. This results in a greater tree

diversity, which (once again) trades a higher bias for a lower variance, generally yielding an overall better model.

When you are growing a tree in a Random Forest, at each node only a random subset of the features is considered for splitting. It is possible to make trees even more random by also using randomthresholds for each feature rather than searching for the best possible thresholds, like regular Decision Trees do. A forest of such extremely random trees is simply called an Extremely Randomized Trees ensemble or Extra-Trees for short. Once again, this trades more bias for a lower variance.

It also makes Extra-Trees much faster to train than regular Random Forests since finding the best possible threshold for each feature at every node is one of the most time-consuming tasks of growing a tree. Important features are likely to appear closer to the root of the tree. While unimportant features will often appear closer to the leaves or not at all.

It is possible to get an estimate of a feature's importance by computing the average depth at which it appears across all trees in the forest. Scikit-Learn computes this automatically for every feature after training. You can access the result using the feature_importances_ variable. When sampling is performed with replacement, this method is called bagging (short for bootstrap aggregating). When sampling is performed without replacement, it is called pasting.

In other words, both bagging and pasting allow training instances to be sampled several times across multiple predictors, but only bagging allows training instances to be sampled several times for the same predictor.

3.2.3 Naïve Bayes

Naïve Bayes is a classification technique which is based on Bayes' Theorem.It is with "naive" assumption of independence among predictors. It is Easy to build. Its is particularly useful for very large data sets. It is known to outperform even highly sophisticated classification methods.

Bayes Theorem:

P(c|x) - the posterior probability of class (c, target) given predictor (x, attributes). P(c) - the prior probability of class.
 P(x|c) - is the likelihood which is the probability of predictor given class.
 P(x) - is the prior probability of predictor.

Figure 3.2.3

Naïve Bayes algorithm firstly convert the data set into a frequency table. Then Create Likelihood table by finding the probabilities. Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

If continuous features do not have normal distribution, we should use transformation or different methods to convert. If test data set has zero frequency issue, apply smoothing techniques "Laplace smoothing". Remove correlated features, as the highly correlated features are voted twice in the model and it can lead to over inflating importance.

Naive Bayes classifier has limited options for parameter tuning. It cannot be ensembled - because there is no variance to reduce. Naive Bayes is an eager learning classifier and it is sure fast. Thus, it could be used for making predictions in real time. Its is well known for multi class prediction feature. It is mostly used in text classification. Naïve bayes have higher success rate as compared to other algorithms. It is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis.

Naive Bayes Classifier and Collaborative Filtering together builds a Recommendation System that uses machine learning and data mining techniques to filter unseen information and predict whether a user would like a given resource or not.

It is used in classification and it assumes that features follow a normal distribution. It is used for discrete counts. It Implements the naive Bayes algorithm for multinomially distributed data. It is one of the two classic naive Bayes variants used in text classification. The binomial model is useful if your feature vectors are binary (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.

3.2.4 SVC(Support Vector Machines)

Support vector machine is very powerful and versatile model capable of performing linear, non-linear classification, Regression and Outlier detection. It is well suited for small or medium sized datasets. Vectors or the training set located closest to the classifier or the training sets located at the edge of the street. Plot the decision boundary along with the training data Convert to unscaled parameters. SVM Sensitivity to feature scaling using simple 2d Data using using SVC – Standard Scalar.

We apply a 2nd degree polynomial transformation and then train a linear SVM classifier on the transformed training set. The 2nd-degree polynomial transformed set is 3-dimensional instead of two-dimensional. If there are two sets of 2-dimensional feature sets, a and b. We apply 2nd degree polynomial mapping and then compute the dot product of the transformed vectors.
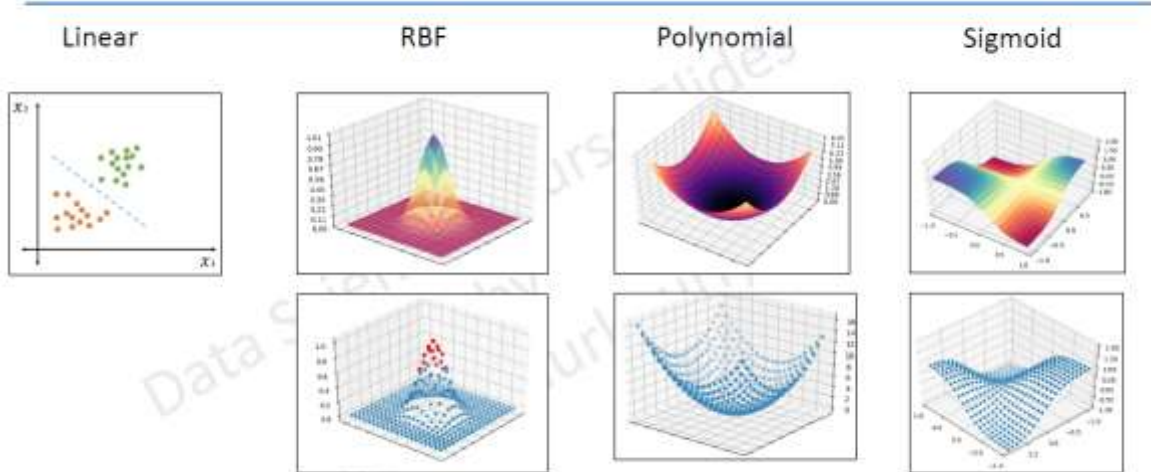
Figure 3.2.4(a)

The dot product of transformed vectors is equal to the square of the dot product of the original vectors. Each degree transformation requires a lot of computation. Dual problem shall contain dot product of the transformed features Matrix. Instead, the original feature can be dot-multiplied and squared. Transformation of the original matrix is not required. The above trick makes the whole process much more computationally efficient. Kernel function represented by K Capable of computing transformed based only on the original vectors without having to compute the transformation.

$$\text{Linear:} \quad K(a, b) = a^T \cdot b$$

$$\text{Polynomial:} \quad K(a, b) = (\gamma a^T \cdot b + r)^d$$

$$\text{Gaussian RBF:} \quad K(a, b) = \exp(-\gamma \| a - b \|^2)$$

$$\text{Sigmoid:} \quad K(a, b) = \tanh(\gamma a^T \cdot b + r)$$

Figure 3.2.4(b)

The original constrained optimization problem , known as the primal problem, can be expressed as another closely related problem known as dual problem. Dual problem gives a lower bound to the solution of the primal problem, but under some circumstances gives the same result. SVM problems meet these conditions, hence have same solution for both primal and dual problems.

3.3   SOFTWARE REQUIREMENTS

3.3.1   Python 3 :

Python is an interpreted, high-level programming language. Made by Guido van Rossum and first discharged in 1991, Python's plan theory stresses code lucidness with its remarkable utilization of significant whitespace. Its language constructs and object-oriented methodology expect to assist software engineers with composing clear, coherent code for small and large-scale ventures. Python composition and garbage collection is done progressively. It underpins different programming standards, including procedural, object-oriented, and functional programming. Python is regularly portrayed as a "batteries included" language because of its complete standard library.

Instead of having the entirety of its usefulness incorporated with its centre, Python was intended to be profoundly extensible. This smaller measured quality has made it especially prevalent as a method for adding programmable interfaces to existing applications.

Python uses the whitespace indentations, rather than curly brackets or keywords, to delimit the blocks. An increase in indentation comes after certain statements; a decrease in indentation signifies the conclusion of the current block. Thus, the program's visual arrangement accurately represents the program's semantic structure. This feature is sometimes termed the off-side rule, which some other languages divide, but in most languages, indentation doesn't have any semantic sense.

We use Python 3 as it's a simple, readable, systematic language which accurately points out the shortcomings and analyses is much easier.

3.3.2 Python IDE :

There are lots of IDEs for python. Some of them are PyCharm, Thonny, Ninja, Spyder etc. We use PyCharm, as it is capable of handling complex features and permutations due to its large size and is hence a suitable IDE for data science related projects.

3.3.3 NumPy:

NumPy is a Python library that is used to manipulate, process and analyze the data. It doesn't have constructs that can be used to visualize the data. For visualization, we can use another library from Python called Matplotlib.

It is the major bundle for logical processing with Python. It contains different highlights including these significant ones:

- NumPy stands for 'Numeric Python' or 'Numerical Python'.
- It is designed for scientific computations.
- It has efficiently implemented multi-dimensional arrays and it also provides fast mathematical functions.
- It is mostly used for array-oriented computing.
- NumPy's main object is the homogeneous multidimensional array called "ndarray".

3.3.4 Pandas :

Pandas are the Python libraries which are used to manipulate, process and analyse the data. They don't have constructs which can be used to visualize the data, for that we can use another library from Python called matplotlib. Pandas provide a high-performance, easy-to-use data structures and data analysis tools

Data Scientists use Pandas for its following favourable circumstances:

1. It uses Series for one-dimensional data structure and Data Frame for multi-dimensional information structure
2. It gives a productive method to slice the data
3. It gives an adaptable method to combine, concatenate or reshape the information
4. It incorporates a powerful time series tool to work with

Briefly speaking, Pandas is a valuable library in data analysis. It tends to be utilized to perform data manipulation and analysis. Pandas give powerful and simple to-utilize data structures, just as the way to rapidly perform tasks on these structures.

3.3.5 Tkinter :

Python offers several options for developing GUI. Out of all the GUI methods, tkinter is most commonly used technique. It is a standard Python line to the Tk GUI toolkit shipped with Python. Python with tkinter outputs the best and easiest way to create the GUI applications. Creating a GUI using tkinter is an simple task.

3.4 Features Taken into Account for predicting Disease

1. Fungal infection
2. Allergy
3. GERD
4. Chronic cholestasis
5. Drug Reaction
6. Peptic ulcer diseae
7. AIDS
8. Diabetes
9. Gastroenteritis
10. Bronchial Asthma
11. Hypertension
12. Migraine
13. Cervical spondylosis
14. Paralysis (brain hemorrhage)
15. Jaundice
16. Malaria
17. Chicken pox
18. Dengue
19. Typhoid
20. hepatitis A
21. Hepatitis B
22. Hepatitis C
23. Hepatitis D
24. Hepatitis E

25. Alcoholic hepatitis
26. Tuberculosis
27. Common Cold
28. Pneumonia
29. Dimorphic hemmorhoids(piles)
30. Heart attack
31. Varicose veins
32. Hypothyroidism
33. Hyperthyroidism
34. Hypoglycemia
35. Osteoarthristis
36. Arthritis(vertigo)
37. Paroymsal Positional Vertigo
38. Acne
39. Urinary tract infection
40. Psoriasis
41. Impetigo

3.5 Code Implementation:

This chapter contains the snapshots of the code developed for the project . In this code we have deployed various Machine Learning Models and implemented their algorithms through codes.

Figure 3.5(a)

Figure 3.5(b)

```python
# -------------------------------------------------------------------

def DecisionTree():

    from sklearn import tree

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

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

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

    for k in range(0,len(l1)):
        # print (k,)
        for z in psymptoms:
            if(z==l1[k]):
                l2[k]=1

    inputtest = [l2]
    predict = clf3.predict(inputtest)
    predicted=predict[0]

    h='no'
    for a in range(0,len(disease)):
        if(predicted == a):
            h='yes'
            break
```

Figure 3.5(c)

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

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

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

    for k in range(0,len(l1)):
        for z in psymptoms:
            if(z==l1[k]):
                l2[k]=1

    inputtest = [l2]
    predict = clf4.predict(inputtest)
    predicted=predict[0]

    h='no'
    for a in range(0,len(disease)):
        if(predicted == a):
            h='yes'
            break

    if (h=='yes'):
        t2.delete("1.0", END)
        t2.insert(END, disease[a])
    else:
        t2.delete("1.0", END)
        t2.insert(END, "Not Found")
```

Figure 3.5(d)

clean_code.py

```python
def NaiveBayes():
    from sklearn.naive_bayes import GaussianNB
    gnb = GaussianNB()
    gnb=gnb.fit(X,np.ravel(y))

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

    psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Symptom5.get()]
    for k in range(0,len(l1)):
        for z in psymptoms:
            if(z==l1[k]):
                l2[k]=1

    inputtest = [l2]
    predict = gnb.predict(inputtest)
    predicted=predict[0]

    h='no'
    for a in range(0,len(disease)):
        if(predicted == a):
            h='yes'
            break

    if (h=='yes'):
        t3.delete("1.0", END)
        t3.insert(END, disease[a])
    else:
        t3.delete("1.0", END)
        t3.insert(END, "Not Found")
```

Figure 3.5(e)

clean_code.py

```python
def SVC():
    from sklearn.svm import SVC
    svc = SVC()
    svc.fit(X,np.ravel(y))

    # calculating accuracy----------------------------------------
    from sklearn.metrics import accuracy_score
    y_pred=svc.predict(X_test)
    print(accuracy_score(y_test, y_pred))
    print(accuracy_score(y_test, y_pred,normalize=False))
    psymptoms = [Symptom1.get(),Symptom2.get(),Symptom3.get(),Symptom4.get(),Symptom5.get()]
    for k in range(0,len(l1)):
        for z in psymptoms:
            if(z==l1[k]):
                l2[k]=1

    inputtest = [l2]
    predict = svc.predict(inputtest)
    predicted=predict[0]

    h='no'
    for a in range(0,len(disease)):
        if(predicted == a):
            h='yes'
            break

    if (h=='yes'):
        t4.delete("1.0", END)
        t4.insert(END, disease[a])
    else:
        t4.delete("1.0", END)
        t4.insert(END, "Not Found")

    # ---------------------------------------------
```

Figure 3.5(f)

```python
# Heading
w2 = Label(root, justify=LEFT, text="Disease Predictor using Machine Learning", fg="white", bg="blue")
w2.config(font=("Elephant", 30))
w2.grid(row=1, column=0, columnspan=2, padx=100)
w2 = Label(root, justify=LEFT, text="A Project by Yaswanth Sai Palaghat", fg="white", bg="blue")
w2.config(font=("Aharoni", 30))
w2.grid(row=2, column=0, columnspan=2, padx=100)

# labels
NameLb = Label(root, text="Name of the Patient", fg="yellow", bg="black")
NameLb.grid(row=6, column=0, pady=15, sticky=W)


S1Lb = Label(root, text="Symptom 1", fg="yellow", bg="black")
S1Lb.grid(row=7, column=0, pady=10, sticky=W)

S2Lb = Label(root, text="Symptom 2", fg="yellow", bg="black")
S2Lb.grid(row=8, column=0, pady=10, sticky=W)

S3Lb = Label(root, text="Symptom 3", fg="yellow", bg="black")
S3Lb.grid(row=9, column=0, pady=10, sticky=W)

S4Lb = Label(root, text="Symptom 4", fg="yellow", bg="black")
S4Lb.grid(row=10, column=0, pady=10, sticky=W)

S5Lb = Label(root, text="Symptom 5", fg="yellow", bg="black")
S5Lb.grid(row=11, column=0, pady=10, sticky=W)


lrLb = Label(root, text="DecisionTree", fg="white", bg="red")
lrLb.grid(row=15, column=0, pady=10, sticky=W)

destreeLb = Label(root, text="RandomForest", fg="white", bg="red")
destreeLb.grid(row=17, column=0, pady=10, sticky=W)
```

Figure 3.5(g)

```python
NameEn = Entry(root, textvariable=Name)
NameEn.grid(row=6, column=1)

S1En = OptionMenu(root, Symptom1,*OPTIONS)
S1En.grid(row=7, column=1)

S2En = OptionMenu(root, Symptom2,*OPTIONS)
S2En.grid(row=8, column=1)

S3En = OptionMenu(root, Symptom3,*OPTIONS)
S3En.grid(row=9, column=1)

S4En = OptionMenu(root, Symptom4,*OPTIONS)
S4En.grid(row=10, column=1)

S5En = OptionMenu(root, Symptom5,*OPTIONS)
S5En.grid(row=11, column=1)


dst = Button(root, text="DecisionTree", command=DecisionTree,bg="green",fg="yellow")
dst.grid(row=8, column=3,padx=10)

rnf = Button(root, text="Randomforest", command=randomforest,bg="green",fg="yellow")
rnf.grid(row=9, column=3,padx=10)

lr = Button(root, text="NaiveBayes", command=NaiveBayes,bg="green",fg="yellow")
lr.grid(row=10, column=3,padx=10)

sv = Button(root, text="SVC", command=SVC,bg="green",fg="yellow")
sv.grid(row=11, column=3,padx=10)

#textfileds
t1 = Text(root, height=1, width=40,bg="orange",fg="black")
t1.grid(row=15, column=1, padx=10)
```

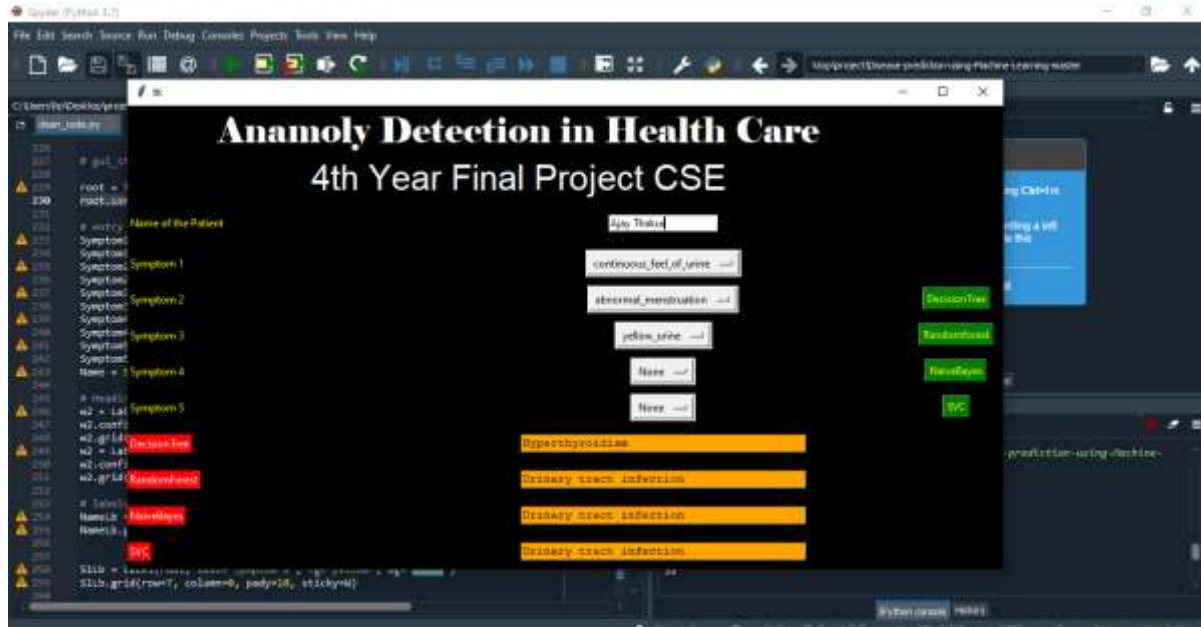Figure 3.5(h)

Chapter 4

Performance Analysis

4.1   Outputs :

Figure 4.1(a)
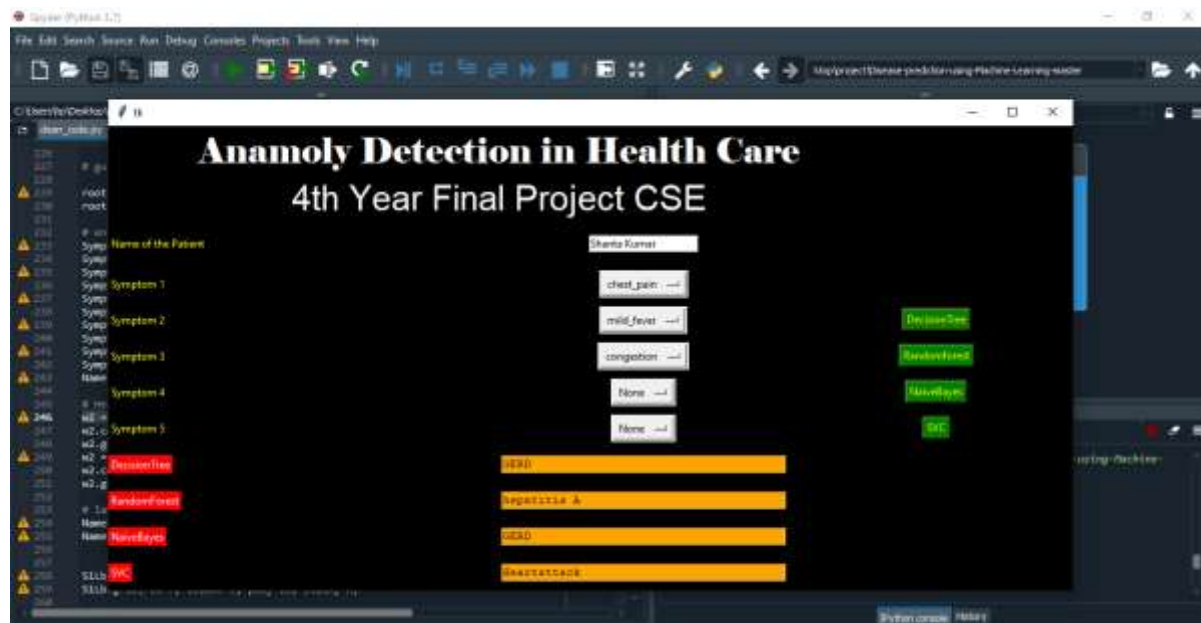
Figure 4.1(b)
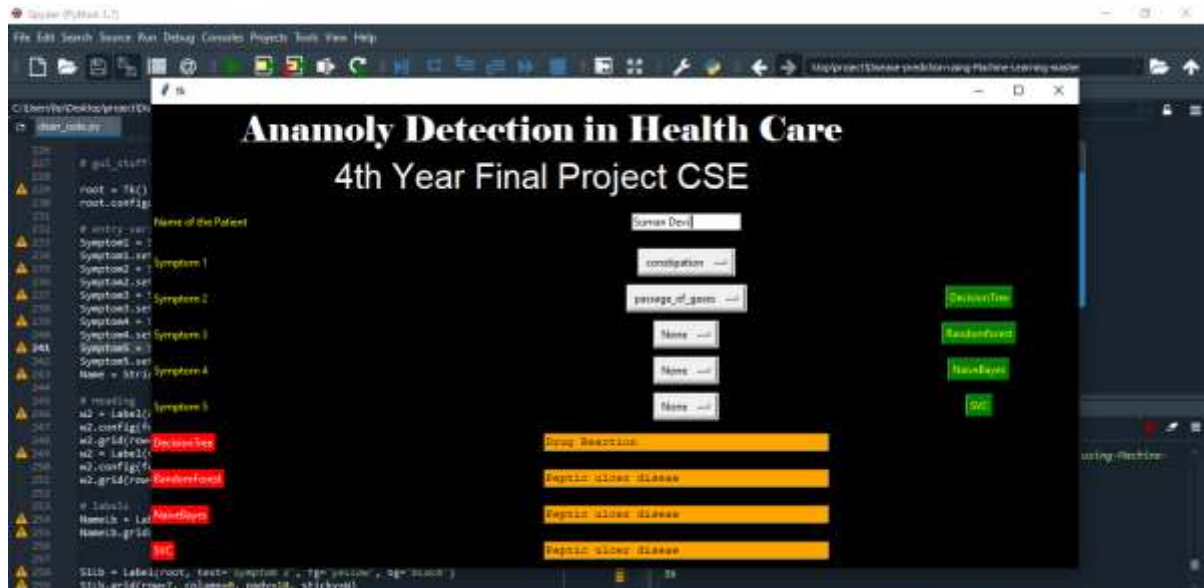
Figure 4.1(c)
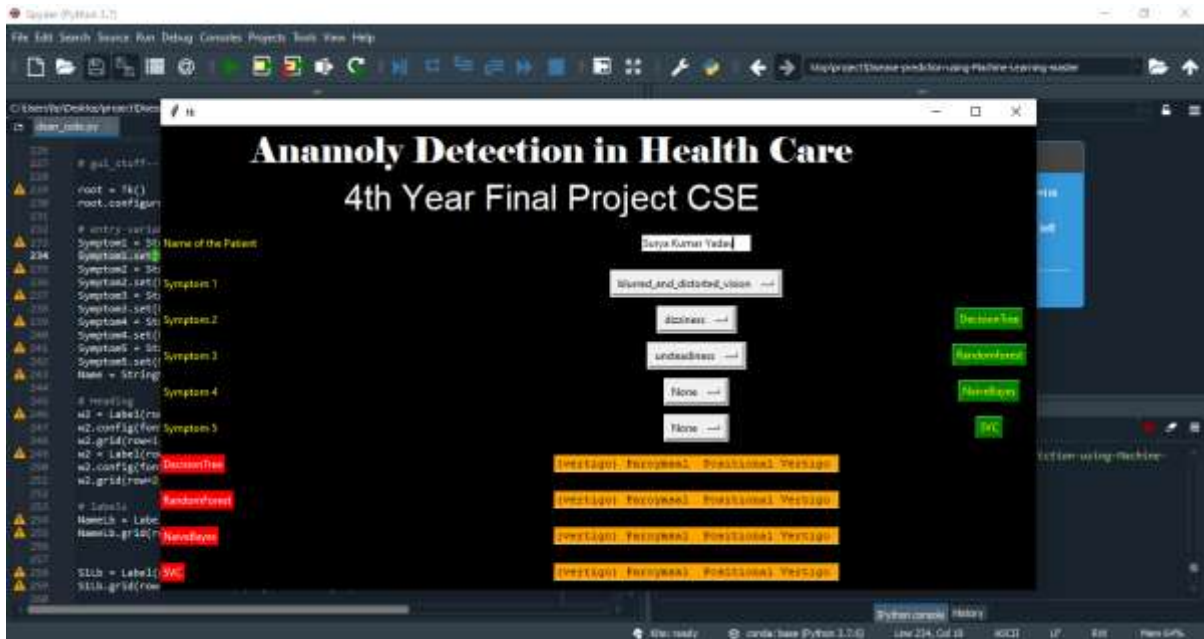

Figure 4.1(d)

4.2   Accuracy Scores:

S. No.	Machine Learning Model	Accuracy
1	Random Forest	88.628%
2	Decision Trees	86.229%
3	SVC	83.684%
4	Naïve	84.996%

Table 4.2

Figure 4.2

Chapter 5

Conclusions

5.1   CONCLUSION:

As healthcare applications produce enormous measure of information which shifts as for its volume, assortment, speed, veracity, and also, value, there is a fast approaching prerequisite of productive digging methods for setting mindful recovery and handling of this class of information. In this way it is being proposed that classification is done to order the enormous information produced in such a state. The proposed framework has made use of the Machine Learning Algorithms and their respective explanatory calculations have been noted giving their respective results.

To formulate assessment, computations from a bunch course of action and data recuperation is done. The original data set was divided into the training and testing dataset. The machine was trained for the training dataset and tested for the training dataset and observed their accuracy. The machine was trained for various Machine Learning Algorithms for classification like SVC, Decision Trees, Random Forest and Naive Bayes. We observed that the Random Forest Algorithm for classification gave the best accuracy and hence decided to go for this algorithm for the completion of the project.

In addition, the proposed plan performed better when thought about with its partners to be specific multi-layer, Bayes arrange what's more, choice table regarding grouping time.

5.2   FUTURE SCOPE:

In the foreseeable future, the aim of this project is to imbibe within itself a mechanism to gather data, cluster it on the go and send it on the cloud. Here, the data is accessed and analysed by experts. Our aim would be to design a cluster formation based on routine health parameters in addition to what has already been built rather than just focus on the symptoms. If possible, we might include IoT setup and gather and store data on the fly.

Chapter 6

References

6.1 References:

[1] C.W. Tsai, C.F. Lai, H.C. Chao, A.V. Vasilakos, "Big Data Analytics: A Survey," Journal of Big Data, vol. 2, no. 1, pp. 1-32, 2015.

[2] Y. Zhang, M. Qiu, C.-W. Tsai, M. M. Hassan, and A. Alamri, "Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data," IEEE Systems Journal, vol. 11, no. 1, pp. 88-95, 2017.

[3] N. Kumar, K. Kaur, A. Jindal, and J.J.P.C. Rodrigues, "Providing Health-care Services On-the-Fly Using Multi-player Cooperation Game Theory in Internet of Vehicles (IoV) Environment," Digital Communications and Networks, vol. 1, no. 3, pp. 191-203, 2015.

[4] S. Saria,"A $3 Trillion Challenge to Computational Scientists: Transform-ing Healthcare Delivery," IEEE Intelligent Systems, vol. 4, pp. 82-87, 2014.

[5] M. Chen, S. Gonzalez, A. Vasilakos, H. Cao, and V.C. Leung, "Body Area Networks: A Survey," Mobile Networks and Applications, vol. 16, no. 2, pp. 171-193, 2011.

[6] G. Fortino, G.D. Fatta, M. Pathan, and A.V. Vasilakos, "Cloud-assisted body area networks: state-of-the-art and future challenges," Wireless Networks, vol. 20, no. 7, pp. 1925-1938, 2014.

[7] D. Talia, "Clouds for Scalable Big Data Analytics," IEEE Computer Magazine, vol. 46, no. 5, pp. 98-101, 2013.

[8] S. Fong, R. Wong, and A. Vasilakos, "Accelerated PSO Swarm Search Feature Selection for Data Stream Mining Big Data," IEEE Transactions on Service Computing, vol. 9, no. 1, pp. 33-45, 2016.

[9] J. Han, M. Kamber, and J. Pei, "Advanced Cluster Analysis," in Data Min-ing: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012, ch. 11, sec. 11.1, pp. 501-508.

[10] A. Jindal, A. Dua, N. Kumar, A. V. Vasilakos, and J. J. P. C. Rodrigues, "An Efficient Fuzzy Rule-Based Big Data Analytics Scheme for Providing Healthcare-as-a-Service," presented in IEEE International Conference on Communications (ICC), Paris, 21-25 May 2017, pp. 1-6.

[11] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, and S.U. Khan, "The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues," Information Systems, vol. 47, pp. 98-115, 2015.

[12] A. Castiglione, R. Pizzolante, A.D. Santis, B. Carpentieri, A. Castiglione, and F. Palmieri, "Cloud-based Adaptive Compression and Secure Management Services for 3D Healthcare Data," Future Generation Computer Systems, vol. 43, pp. 120-134, 2015.

[13] P. Jiang, J. Winkley, C. Zhao, R. Munnoch, G. Min, and L.T. Yang, "An Intelligent Information Forwarder for Healthcare Big Data Systems With Distributed Wearable Sensors," IEEE Systems Journal, vol. 10, no. 3, pp. 1147-1159, 2016.

[14] D. Lin, X. Wu, F. Labeau, and A. Vasilakos, "Internet of Vehicles for E-Health Applications in View of EMI on Medical Sensors," Journal of Sensors, 2015, DOI: 10.1155/2015/315948.

[15] C.W. Cheng, N. Chanani, J. Venugopalan, K. Maher, and M.D. Wang, "icuARM - An ICU Clinical Decision Support System Using Association Rule Mining," IEEE Journal of Translational Engineering in Health and Medicine, vol. 1, no. 1, 2013.

[16] Y. Gong, Y. Fang, and Y. Guo, "Private Data Analytics on Biomedical Sensing Data via Distributed Computation," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 3, pp. 431-444, 2016.

[17] A. Forkan, I. Khalil, and Z. Tari, "CoCaMAAL: A Cloud-oriented Context-aware Middleware in Ambient Assisted Living," Future Generation Computer Systems, vol. 35, pp. 114-127, 2014.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
PLAGIARISM VERIFICATION REPORT

Date: 15/07/2020   √

Type of Document (Tick): | PhD Thesis | | M.Tech Dissertation/ Report | | B.Tech Project Report | | Paper |

Name: Uday Gupta   **Department: CSE**   **Enrolment No 161205**

Contact No. 8894241974   **E-mail.   usand6@gmail.com**

Name of the Supervisor: Mrs. Monika bharti

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters):

_____
 ANOMALY DETECTION IN HEALTH CARE
_____

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.

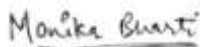**Complete Thesis/Report Pages Detail:**

- Total No. of Pages =52
- Total No. of Preliminary pages =7
- Total No. of pages accommodate bibliography/references =2

(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found **Similarity Index** at 6 (%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.

Monika Bharti

(Signature of Guide/Supervisor)   **Signature of HOD**

FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Generated Plagiarism Report Details (Title, Abstract & Chapters)	
			Word Counts	
Report Generated on	• All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String		Character Counts	
		Submission ID	Total Pages Scanned	
			File Size	

Checked by
Name & Signature   **Librarian**
………

Please send your complete thesis/report in (PDF) with Title Page, Abstract and Chapters in (Word File) through the supervisor at plagcheck.juit@gmail.com