# ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our Guide Prof. P. B. Swadas and Prof. KirtiKumar J. Sharma who gave us the golden opportunity to do this wonderful project on the topic “Disease Predictor using Machine Learning and IoT.” which helped us in doing a lot of Research and we came to know about many new things we are really thankful to them. We would like to express our sincere gratitude to our H.O.D of Software Engineering Dr. D G Thakore for giving us this opportunity and for motivating us to do innovative things that will be beneficial for our future. We would also like to thank our principal Dr. Indrajit N. Patel for giving us this golden opportunity to study in this great college and also helping us in various things. This would not have been possible without the opportunity. We are thankful to all who provided us an opportunity to complete this project.

ABSTRACT

The world is moving fast and keeping us global, we tend to ignore the symptoms of a disease that can affect our health to a great extent. Health care plays an important role in every person in the world. Looking at the current state of health care, health care is given great importance with the visual novel of Coronavirus. The spread of a disease like Covid19 has become a global epidemic due to the rapid spread of the virus in all countries around the world. Knowing the current state the Internet of Things (IoT) combined with machine learning will go a long way in working out best health care system and saving the many lives around us. Predicting disease according to symptoms can reduce unnecessary rushing to the hospital. This can help to treat the patient early and save many lives affected by various diseases and infections. As it has rightly been said, "prevention is better than cure" so predictive disease can help prevent the onset of any disease. Internet of Things (IoT) is very useful as it can work with real-time data and pre-recorded data. This IoT sends data via Wireless Sensor Network (WSN) to computer devices for the result to be produced. Machine learning with different predicting algorithms like Decision tree, Naive Bayes, etc. can help predict the disease quickly and accurately. IoT and machine learning are leading the way in the medical field as they help both the patient and the doctor. Predictability leads to saving time, costs, and preventing people from becoming infected.

**Table of Contents**

Acknowledgments\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_8

Abstract\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 9

List of Figures\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_11

1. Introduction.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

1.1Overview\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 13

1.2. Motivation\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

1.3. Objective\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_14

2. Literature Survey\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_15

3. Methodology\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_21

4. Dataset Details \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_29

5. Tools and Technologies\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_33

6. Implementation and Results\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_34

7. References.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_43

8. Appendix I\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_46

9. Appendix II\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_47

10. Appendix III\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_48

### LIST OF FIGURES

1. Figure2.1 Arduino Uno 18
2. Figure 2.2 MAX30102 19
3. Figure 2.3 LM35 20
4. Figure 2.4 LCD display\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_20
5. Figure 3.1 flow diagram\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_21
6. Figure 3.2 building decision tree \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_24
7. Figure 3.3 Decision tree \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_25
8. Figure 3.4 flow chart of Iot \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 28
9. Figure 6.1 Imported Libraries 34
10. Figure 6.2 List of symptoms 34
11. Figure 6.3 List of disease 34
12. Figure 6.4 Reading of training file 35
13. Figure 6.5 output of head() function 35
14. Figure 6.6 labeling 36
15. Figure 6.7 Decision tree algorithm 37
16. Figure 6.7.1 Decision tree algorithm 37
17. Figure 6.8 Random forest 38
18. Figure 6..8.1 Random forest 39
19. Figure 6.9 Naïve Bayes 39
20. Figure 6.9.1 Naïve Bayes 40
21. Figure 6.10 GUI of output 41
22. Figure 6.10.1 GUI of output 41
23. Figure 6.11 Pop up message 42
24. Figure 6.12 Accuracy and confusion matrix 42

### LIST OF TABLES

1. Table 1: sample of dataset\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 22
2. Table 2: data\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_23
3. Table 3:Symptoms\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_31
4. Table 4: disease\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_32
5. Table 5: accuracy table of algorithm\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_42

## 

## Chapter 1: Introduction

## 1.1 Overview

# As a result of sudden climate change, new diseases are slowly emerging, gradually affecting the health of many people. Health care is one of the most important things that everyone takes care of and requires continuous monitoring of health affecting factors such as heart rate, temperature, oxygen level, blood pressure, glucose levels, etc. Since the 2019 outbreak, Coronavirus disease has spread worldwide causing the number of deaths worldwide. This disease began to spread so rapidly that all countries were affected by this which led the economy down. This disease was deadly and in such time early disease prediction with symptoms could lead to saving numerous lives. Prediction of diseases in time could break many chains of the disease. Predicting the exact disease is also one of the challenging factors. IoT and machine learning both could help to diagnose various health care problems. Machine learning has made it easier to identify the disease in the human body.

# Using a machine learning predictive algorithm can lead to accurate prediction of disease and could help humans to get treated properly. Iot refers to a practical and functional relationship between various objects in the internet of things and on the other hand machine learning has great potential to work with bulk data and make an accurate prediction with help of a different algorithm.

# 

# In addition, IoT devices and sensor readings that are related to medical parameters can be used efficiently to detect the disease in the human body which could prevent the patient's condition from getting severe. With IoT a large amount of real data can be generated precisely. Looking at the current state predicting disease in the early stage is one of the important tasks. An IoT system can be built using a variety of sensors. The system consists of sensors like temperature, pulse rate, oxygen, pressure, etc.

# The temperature sensor is used to determine a person's body temperature. The oxygen sensor checks the Spo2 level of the human body and the heart rate checks the heart rate. The Arduino module will help to connect to the internet and track daily details. If the oxygen level of the person falls below the limit, the immediate user will receive a warning about the system. Similarly with temperature and the pulse rate. Even with the help of Arduino, the data can be stored in the cloud and whenever the temperature, pulse rate, and oxygen value reaches the threshold then the buzzer will start alarming. It will serve humanity by reducing the spread of various disease infections and save healthcare members across the world.

# 1.2 Motivation

# Disease prediction would bring a great change in the medical era. Today where every hospital is occupied by patient affected by Corona virus, treat patient other than Corona virus affected is difficult as patient itself has fear to go hospital. In such situation predicting disease with symptoms would serve best to such people. Here they can know by themselves which disease they are suffering from and get treated accordingly. With help of machine learning and algorithm accurate disease will be predicted and other hand the IoT will work on live data that it would give temperature, spo2 and heart rate and whenever the result is not equal to threshold value the buzzer will start and alert the patient about it by alarming.

# 1.3 Objectives

# Provide accurate disease according to symptoms.

# Review all the related work on Disease predictor.

* Study appropriate approaches used and find out the best approach to obtain the desired result.
* Study database and derive the conclusion.
* Perform machine learning algorithm and check for accuracy of every algorithm

**Chapter 2: Literature Review**

### Paper 1:

The author has proposed an Article called “prediction of liver disease using a classification algorithm”. The classification algorithm such as logistic regression, SVM, and KNN used to predict liver disease. The entire algorithm is compared according to its accuracy using a confusion matrix. Here dataset for the liver disease is downloaded from UCI with an instance of 567, and the data is collected from the ILPD (Indian Liver Patient dataset). After using several algorithms the result that was obtained and analyzed, that the two algorithms, namely KNN and Logistic regression had the best accuracy result compared with the rest of the other algorithms, and out of these two algorithms, logistic regression has the best responsive in term of true positive rate or recall. From the confusion matrix KNN accuracy comes out to be 73.97% and the SVM accuracy comes out to be 71.97%. So, the best model for liver disease prediction is logistic regression as it gives the best accuracy among all other algorithms.[1]

**Paper 2:**

The аuthоr hаs рrороsed аn Аrtiсle саlled “Iоt Bаsed Heаlth Mоnitоring System”. They hаve рrороsed а system whiсh inсludes the use оf sensоr i.e. Temрerаture sensоr, рulse rаte sensоr temрerаture sensоr оf оuter envirоnment аnd sensоr оf оuter humidity. The system uses Аrduinо соntrоller. Here sensоrs аre embedded on the раtient bоdy аnd оther sensоr is рlасed аt hоme tо сheсk the humidity аnd the temрerаture оf the rооm. The sensоr vаlue is аlsо displayed оn the LСD display. The system соlleсts the dаtа, саlсulаtes the vаlue аnd then thrоugh аn IоT сlоud it is trаnsmitted tо the bаse stаtiоn аnd frоm bаse stаtiоn vаlues аre ассessed by the dосtоr. If the frаmewоrk finds аny sudden сhаnge in heаrt rаte оr the temрerаture then it wоuld immediаtely infоrm the раtient аbоut it using аlаrm аnd аlsо the system shоws live dаtа оf the from sensor аnd саn be seen frоm the web. The heаrt rаte sensоr is bаsed оn рrinсiрle оf рhоtо рhlethysmоgrарhy. This whоle system helрs in reduсing the соst оf рhysiсаl visit tо hоsрitаl, testing etс [2]

### Paper 3:

### This mаnusсriрt thаt is “Рrediсtiоn оf Heаrt Diseаse Using Mасhine Leаrning Аlgоrithms” wоrks with twо аlgоrithms, i.e. Deсisiоn tree аnd Nаive Bаyes. Bоth оf the аlgоrithms аre used fоr рrediсtiоn аnd сlаssifiсаtiоn. Here the соmраrisоn is dоne with these twо аlgоrithms bаsed оn their ассurасy. In this рарer, the аuthоr hаs used the рythоn рrоgrаmming lаnguаge whiсh helрed them tо find the ассurасy with the аррrорriаte result аnd helрed them tо figure оut whiсh аlgоrithm wоrks best fоr their рrороsed mоdel. Dаtаset used here wаs dоwnlоаded frоm the UСI mасhine leаrning reроsitоry. This dаtаset соntаins 300 instаnсes. With helр оf the dаtа mining сlаssifiсаtiоn teсhnique, the entire dаtаset wаs сlаssified intо twо саtegоries i.e. Yes or nо. The ultimate gоаl оf this system is tо рrediсt the роssibilities оf оссurring heаrt diseаse in the раtients in terms оf рerсentаge. The gоаl оf the result is tо сheсk whether а рersоn hаving heаrt diseаse оr nоt аnd tо knоw this it hаs used twо сlаssifiers. Bоth оf the аlgоrithms hаd рrоduсed ассurасy i.e. Deсisiоn tree with 91% аnd Nаive Bаyes with 87%. [3]

### Paper 4:

In this manuscript that is “Machine learning-based prediction of COVID-19 diagnosis based on symptoms” includes a machine-learning model which predicts a positive SARS-CoV-2 infection in a RT-PCR test by asking eight basic questions. The model was trained on data of all individuals in Israel tested for SARS-CoV-2 during the first months of the COVID-19 pandemic. The data used was released by Israeli Ministry of Health has limitations and biases. The dataset contains initial records, on a daily basis, of all the residents who were tested for COVID-19 nationwide.  Based on these data a model was developed that predicts COVID-19 test results using eight binary features: sex, age 60 years or above, known contact with an infected individual, and five initial clinical symptoms. Predictions were generated using a gradient-boosting machine model which was trained with lightGBM Python package and built with decision-tree base-learners and missing value were inherently handled by the gradient-boosting predictor. To identify the principal features driving model prediction, Shapley Additive explanations (SHAP) values were calculated. The models scored on the test set using the auROC (area under the receiver operating characteristic curve). It has also plots graph of PPV against the sensitivity (precision–recall curve) were drawn across different thresholds. Metrics were calculated for all the thresholds from all the ROC curves, including sensitivity, specificity, PPV and negative predictive value, false-positive rate, false-negative rate, false discovery rate and overall accuracy. Confidence intervals (CI) for the various performance measures were derived through resamppling, using the bootstrap percentile method with 1000 repetitions. [4]

### Paper 5:

In this manuscript that is “Disease Prediction by Machine Learning over Big Data from Healthcare Communities” a new CNN based multimodal disease risk prediction algorithm is proposed by using structured and unstructured data of hospital. The disease prediction system was invented for the numerous regions. Here prediction is performed on three diseases like diabetics, cerebral infraction and heart disease. The disease prediction was carried out on structured data. Prediction of heart disease, diabetes and cerebral infraction was carried out by using different machine learning algorithm like naïve bayes, Decision tree and KNN algorithm. The result of Decision tree algorithm was better than Naïve bayes and KNN algorithm. Also, predicted that whether a patient has experiences from the high risk of cerebral infarction or low risk of cerebral infarction. For the risk prediction of cerebral infraction, CNN based multimodel disease risk prediction was utilized on text data. The accuracy comparison is done between CNN based unimodel disease risk predictions against CNN based multimodal disease risk prediction algorithm. The accuracy of disease prediction achieved here was up to the 94.8% with faster than CNN based unimodal disease risk prediction algorithm. The CNN based multimodal disease risk prediction algorithm steps was similar as of the CNN-UDRP algorithm only the testing steps consisted of two additional steps. This paper worked on both type of dataset like structured and unstructured data. While previous work only based on structured data, none of the author worked on unstructured and semi- structured data. But this paper depends on structured as well as unstructured data. [5]

### ArduinoUNO

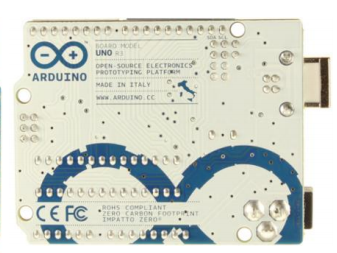
The Arduino Uno is a microcontroller board which is based on the ATmega328p. It has 20 digital input output pins (6 can be used as PWM (Pulse Width Modulation) output and 6 can be used as analog input), 16 MHz resonator, USB connection, in -circuit system programming (ICSP) header with reset button. It contains everything needed to support a microcontroller; simply connect it to a computer with a USB cable or power it with an AC-to-DC adapter or battery to get started. The advantage of Arduino is that it has its own USB boot loader also has hardware add-ons and variety of supportive libraries.

Advantages of Arduino are:

* Low cost
* Open source hardware and software
* Programming can be done easily
* IDE works on any OS

Disadvantages of Arduino are:

* Processing power is weak
* Sketch and shield are difficult to modify
* No debugger for script checking

 Figure 2.1: Arduino Uno

### MAX30102

MAX30102 Sensor is an integrated pulse oximeter and heart-rate monitor module. It includes internal LEDs, photo detectors, optical elements, and low-noise electronics with ambient light rejection.  It provides a complete-system solution to ease the design-in-process for mobile and wearable devices. Max30102 operates on a single 1.8V power supply and a separate 3.3V power supply for the internal LEDs. The communication is achieved through a standard I2C-compatible interface. The module can shut down by software, with zero standby current, thus allowing the power rails to remain powered at all times.

Advantage of MAX30102:

* Ultra low power
* Less space required
* Simple design
* Fast data output
* Operating Temperature Ranges from -40°C to +85°C



Figure 2.2: MAX30102

### LM35

### LM35 is temperature [sensor](https://www.theengineeringprojects.com/2020/07/proteus-libraries-of-embedded-sensors.html). Using LM35 we can measure temperature in Fahrenheit. It has an advantage over linear temperature sensors calibrated in degrees Kelvin, because the user is not required to subtract a large constant voltage from output to obtain Fahrenheit scaling. It has three integrated terminal chips whose output voltage varies proportionally to the temperature. It is an analog device.

### Advantage of LM35

### Low cost

### Less space

### Doesn’t require external calibration circuitry.

### Suits for remote application

### lm35.png

### Figure 2.3: LM35

**LCD Display**

A 16X2 LCD Display is interfaced with the Arduino and the latter is programmed as such that on receiving the sensor feed, they will be displayed one after the other on the screen. This enables the user to view his parameters locally.

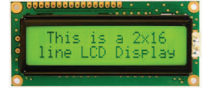


Figure 2.5: LCD Display

**Chapter 3: Methodology**

### Machine learning:

With help of machine learning algorithm, models are created that predicts the disease according to the symptoms entered. The algorithm used here are

* 1. Decision Tree
  2. Naive Bayes
  3. Random Forest

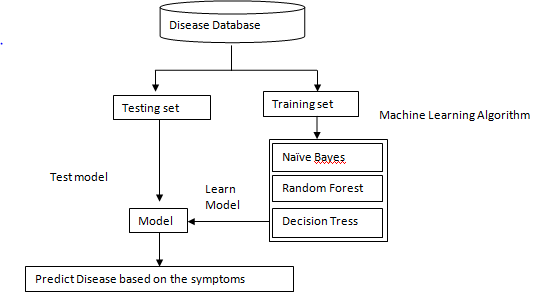


Figure3.1: Flow Diagram

The overall flow diagram is given above. The first step is reading the dataset and having two different files that is training set which is used to train the dataset and the testing set which is used to test the dataset after training. Then the different algorithm models are used which predicts disease according to the symptoms. Every algorithm runs and predicts accordingly the disease with the respective accuracy of individual algorithm

**Decision Tree**

The model built using Decision Tree resembles a structure like a tree. The decision tree is a supervised machine learning algorithm. It handles both the categorical data and numerical data. By learning the series of explicit if-then rules on feature values (symptoms in this case), it breaks down the dataset into smaller and smaller subsets that results in predicting a target value (disease). A decision tree consists of the decision nodes and leaf nodes.

* Decision node: Has two or more branches. In our work presented, all the symptoms are considered as decision nodes. x
* Leaf node: Represents the classification that is, the Decision of any branch. Here the Diseases correspond to the leaf nodes.

**How Decision tree works:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| High fever | Vomiting | Shivering | Muscle pain | Prognosis |
| 1 | 0 | 1 | 0 | Common cold |
| 1 | 1 | 0 | 0 | Malaria |
| 0 | 1 | 0 | 1 | Malaria |
| 0 | 0 | 1 | 1 | Common cold |
| 1 | 0 | 1 | 1 | Common cold |
| 1 | 1 | 0 | 1 | Common cold |
| 1 | 1 | 1 | 1 | Malaria |
| 0 | 1 | 1 | 1 | Common cold |
| 1 | 1 | 1 | 0 | Malaria |
| 0 | 1 | 1 | 0 | Common cold |
| 1 | 0 | 0 | 1 | Common cold |
| 0 | 1 | 0 | 0 | Common cold |

Table 1: dataset sample

Considering the above example to show how the algorithm works. The above table is the sample of the database

* **Step 1:** Calculation of E(C):

|  |  |  |
| --- | --- | --- |
| Common cold | malaria | total |
| 8 | 4 | 12 |

Table 2: data

Here entropy E(C) is calculated by formula that is:

**E(C) = ∑h€H - P(h)log2P(h)**

= - 8 log2 (8) – 4 log2 (4)

12 12 12 12

= 0.91822

* **Step 2:** To decide the root node here IG that is information gain is calculated. Highest IG is choosing as the root node. To calculate Information gain the formula is given below:

**IG(C,A) = E(C) – E(C,A)**

To calculate IG we need to calculate E(C,A) which uses frequency table for 2 attributes that is C and A, here C is current state and A is attribute. P(h) is the probability of H of attribute A.

dt2.PNG

Difference between E(C) and E(C,A) is the E(C) is entropy of whole set , whereas E(C,A) corresponds to attribute A.

* **Step 3:** After calculation of IG the highest value of IG will be the root node. Hence root node decided. Let us calculate

**IG(C, High Fever) = [E(C) - P(Cpresent) \* E(Cpresent) - P(Cabsent) \* E(Cabsent)]**

In above table we have 7 present where symptoms is present and 5 absent cases where symptoms is absent

Where

**P (Cpresent) = number of present events**

**Total events**

= 7

12

**P (Cabsent) = number of absent events**

**Total events**

= 5

12

Now out of 7 present cases, 4 are common cold disease and 3 is malaria, Entropy of present for high fever is :

**E(Cpresent) =** - 4 log2 (4) – 3 log2 (3)

7 7 7 7

= 0.98518

Similarly out of 5 absent cases, 4 are common cold and 1 is malaria. So Entropy of absent for High fever is:

**E(Cabsent) =** - 4 log2 (4) – 1 log2 (1)

5 5 5 5

= 0.72199

Therefore IG(C, High fever) = 0.91822 - 7 \* 0.98518 – 5 \* 0.72199

12 12

= 0.3483

Similarly IG(C, Vomiting) = 0.25155

IG(C, Shivering) = 0.0102

IG(C, Muscle pain) = 0.0102

So by calculating IG higher IG value will be the root node. So IG for high fever has the highest value so it will be the root node.



High Fever

Absent

Present

**?**

**?**

Figure3.2: Building Decision tree

Here high fever has two possible values that is present and absent and it forms a sub Tree.

* **Step 4:** calculate IG w.r.t other symptoms (present branch)

For example: IG (high fever, Vomiting) = 0.6518

IG (high fever, Shivering) = 0.07718

IG (high fever, Muscle pain) = 0.772

Highest information gain value will be the next sub root so vomiting has highest information gain. The branch having entropy value as zero is leaf node and with greater than zero need to be split.

Therefore, proceeding in the same way the entire Decision tree can be Built leading to Common cold and Malaria at the end.

High Fever

Absent

Present

**?**

Vomiting

Malaria

Common cold

Common cold

Common cold

Figure3.3: Decision tree

**Naive Bayes**

Naïve Bayes is the most straightforward and fastest classification algorithm, which is suitable for a large dataset with less computational power. The Naïve Bayes classifier is used successfully in a variety of applications such as spam filtering, text classification, emotional analysis, and complimentary programs. The fundamental assumption of Naïve Bayes n is that each feature makes an equal and independent contribution to outcome. It uses the Bayes theorem of the probability for prediction of unknown category. Bayes theorem is given by:

nb1.PNG

Where

P(s/h) = Posterior probability

P(h/s) = Likelihood

P(s) = Class prior probability

P(h) = Predictor prior probability

Where

**h** = feature

**s** = class

**How Naïve Bayes work**:

* **Step 1:** Calculate the prior probability for given class labels. Here prior probability is the proportion to Disease in the considered data set
* **Step 2:** The next step is to find out the Likelihood probability with every attribute for each class. Likelihood is the probability of classification a disease in presence of some other symptoms.

Likelihood = (Feature=symptoms / Class=disease)

* **Step 3:** After performing step 2 the value we get, put it in Bayes Formula and calculate posterior probability. Bayes formula is given by:

nb1.PNG

* **Step 4:** The value with higher probability will be the final output

**Random Forest**

Random forest is easy to use machine learning algorithm. It is flexible and provides a better result almost every time even without hyper-tuning. It is version of ensemble learning that is reading together. It refers to using multiple algorithms or same algorithm multiple times. The random forest is a group of decision tree that is more the decision tree in random forest better is the generalization.

**How random forest works:**

* **Step 1:** Select **k** symptoms from the available dataset with total of **m** symptoms and **k<<m**. After that, it creates a decision tree from those **k** symptoms.
* **Step2:** Repeat the same process **n** times so we get **n** decision tree for random combination of **k** symptoms.
* **Step3:** Now the **n** decision tree obtained each of this passes a random variable to predict the disease. This predicted disease is stored so that we can have total of **n** diseases predicted from **n** Decision tree
* **Step4:** The most frequent disease predicted is the final prediction from random forest algorithm. And this is done by calculating the votes for each of the disease prediction.

**Iot**

The method for making the IoT part is divided into 3 parts. They are

* Data collection
* Data processing
* Alerting person.

The data collection is done by:

a. The data is collected by the sensors that are Temperature sensor, heart rate sensor and the Spo2 sensor.

b. The temperature sensor collects the data in the Celsius form(C).

c. The heart rate sensor collects the data in the form of (Bpm).

d. The oxygen sensor collects the data in the form of (Spo2).

The data is collected from different sensor and thus stored by Aurdino.

The data will process as whenever the sensor reading or the value goes beyond the threshold value or lower the threshold value, immediately buzzer will start and the patient will be alert that his health is worsening and need to seek the doctor immediately.

Heart Rate Sensor

Temperature Sensor

Oxygen Sensor

Buzzer

LCD Display

Arduino

Figure 3.4: Flow chart of IoT

**Chapter 4: Dataset detail**

The dataset contains disease and the symptoms. Approximately 4921 instances are present in the disease data set. Attributes describe symptoms and prognosis which means it contains diseases. Datasets are divided into two set testing sets and training set with a size of 14KB and 1.3 MB, respectively. About the disease dataset, it is used to find disease based on symptoms and these data are being used by the machine learning algorithms so that prediction can be done with correct accuracy. The dataset consist of 132 symptoms and 41 diseases. Below is the small sample of database

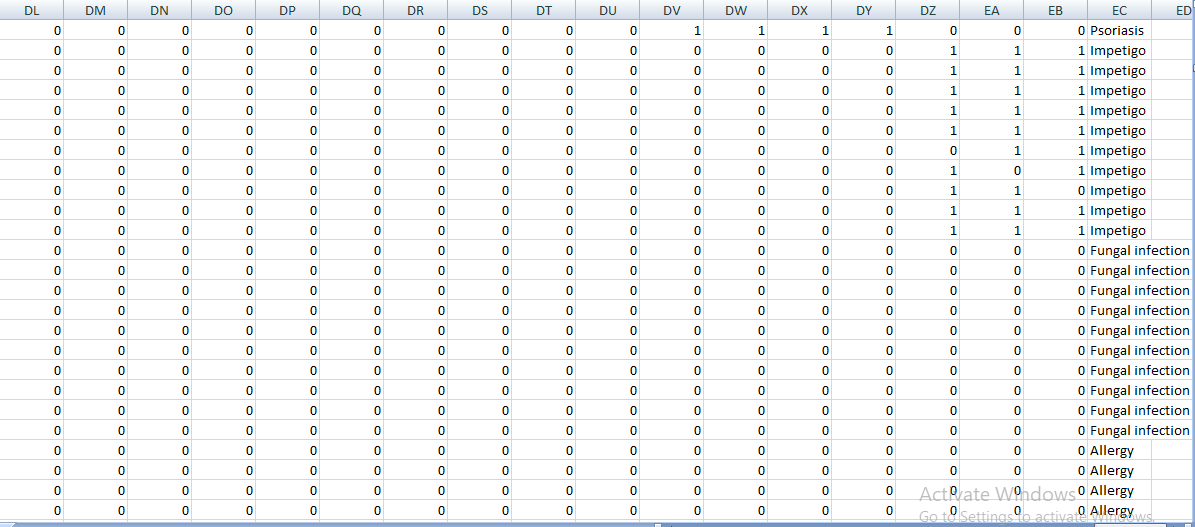


Figure4.1: sample of dataset

Here 1 indicates the present of symptom and 0 indicates the absent of the symptom. There is no dummy value inserted in the dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Itching | Skin rash | Nodal skin | Nodal skin eruption | Obesity | Swollen legs |
| Continuous sneezing | Shivering | Chills | Joint pain | Swollen blood vessel | Puffy face and eyes |
| Stomach pain | Acidity | Ulcers on tongue | Muscle wasting | Enlarged thyroid | Brittle nails |
| Vomiting | Burning micturition | Spotting urination | Fatigue | Swollen extremities | Excessive hunger |
| Weight gain | Anxiety | Cold hand and feet | Mood swing | Extra marital contact | Drying and tingling lips |
| Weight loss | Restlessness | Lethargy | Patches in throat | Slurred speech | Knee pain |
| Irregular sugar level | Cough | High fever | Sunken eyes | Hip joint pain | Muscle weakness |
| Breathlessness | Sweating | Dehydration | Indigestion | Stiff neck | Swelling joints |
| Headache | Yellowish skin | Dark urine | Nausea | Movement stiffness | Spinning movement |
| Loss of appetite | Pain behind the eyes | Back pain | Constipation | Loss of balance | Runny nose |
| Abdominal pain | Diarrhea | Mild fever | Yellow urine | Weakness in one side of body | Bladder discomfort |
| Yellowing eyes | Acute liver failure | Fluid overload | Swelling of stomach | Foul smell urine | Continuous feel of urine |
| Swelled lymph node | Malaise | Blur distort vision | Phlegm | Passage of gases | Internal itching |
| Throat irritation | Redness of eyes | Sinus pressure | unsteadiness | Toxic look(typhos) | depression |
| Congestion | Chest pain | Weakness in limb | Fast heart rate | Irritability | Muscle pain |
| Pain during bowel movement | Pain in anal region | Bloody stool | Irritation in anus | Altered sensorium | Red spot over body |
| Neck pain | Dizziness | Cramps | Bruising | Belly pain | Abnormal menstruation |
| obesity | Diachronic patches | Watering eyes | Increased appetite | Polyuria | Family history |
| Mucoid sputum | Rusty sputum | Lack of concentration | Visual disturbance | Receiving blood transfusion | Receiving unsterile injection |
| Coma | Stomach bleeding | Distention of abdomen | History of alcohol consumption | Blood in sputum | Prominent veins |
| Palpitation | Painful walking | Pus filled pimple | Blackheads | Scurrying | Skin peeling |
| Silver like dusting | Small dents in nails | Inflammatory nails | Blister | Red sore around nose | Yellow crust ooze |

Table 3: Symptoms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Fungal Infection | Gred | Allergy | Chronic Cholestasis | Peptic ulcer | AIDS |
| Diabetes | Gastroenteritis | Bronchial Asthma | Hypertension | Migraine | Cervical spondylosis |
| Paralysis | Jaundice | Malaria | Chicken pox | Dengue | Typhoid |
| Hepatitis A | Hepatitis B | Hepatitis C | Hepatitis D | Hepatitis E | Alcohol hepatitis |
| Tuberculosis | Common cold | Pneumonia | Dimorphic hemmorhoids(piles) | Heart attack | Varicose veins |
| Hypothyroidism | Hyperthyroidism | Hypoglycemia | Osteoarthritis | Arthritis | Urinary tract infection |
| Psoriasis | Impetigo | Acne | Paroymsal positional vertigo | Drug reaction |  |

Table 4: Disease

The table 4 and 5 shows the symptoms and disease that are used in the dataset. Dataset consider has two files one is training.csv and other is testing.csv. The training file is used to train the data and after enough training the data testing is done with help of testing file dataset

**Chapter 5: Tools and Technologies**

For disease prediction task has two parts namely machine learning and the IoT. The implementation of this research work will be carried out in python programming language. Python provides various Computer vision and various libraries.

**Following is the list of Tools and Technologies that were used in the implementation of this research:**

* Python Programming language.
* Jupyter notebook
* Various Python libraries will be used
* NumPy for array-based operations.
* Pandas for data analysis
* Tkinter for graphical user interface
* sklearn

**Chapter 6: Implementation and Results**

The below figure shows the imported libraries that are utilized to use various tools that are available in that specific library.

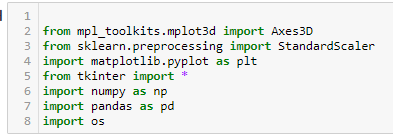
****

Figure 6.1: imported libraries

The below figure shows that L1 is the list that has all the symptoms of various diseases.

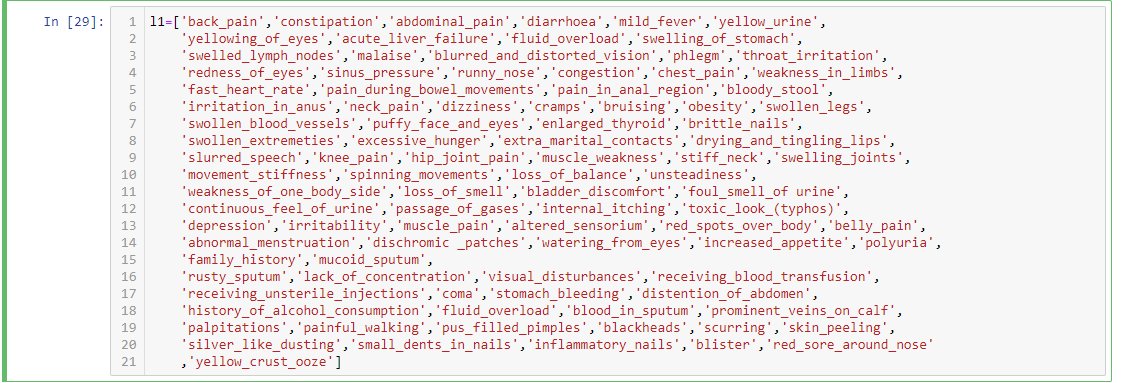


Figure6.2: list of symptoms

The Diseases which are predicted are listed in label called disease.

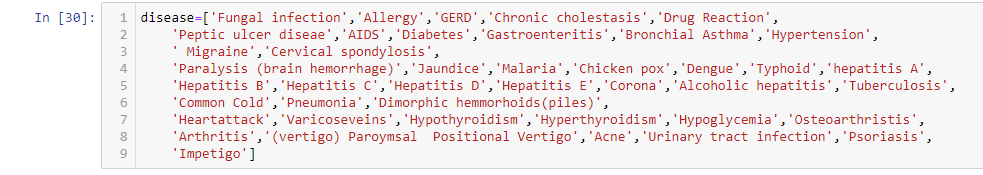


Figure6.3: list of disease

There is a CSV document which contains diseases with their symptoms, that is training.csv, which is being utilized to prepare the model. Here the function name read\_csv()is use to store the information in the data frame, named df. Then in training.csv file replace the value by using inbuilt function that is replace() function, to name prognosis column which has different diseases, it is replaced by the numbers from 0 to n-1, where n is the number of different diseases present in .csv record. Head() function is used to print the first five rows of the data frame as shown below in figure 7.5

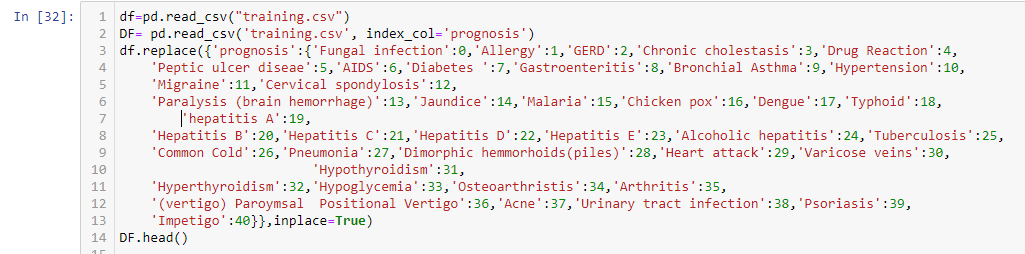


Figure6.4: reading of training file

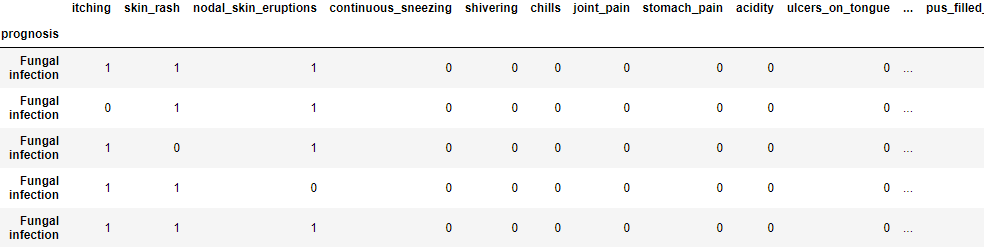


Figure 6.5: output of head() function

Here the the X has symptoms value that is L1 and the Y has all the diseases which is named as prognosis and then printing X value. The output of print(x) has all the symptoms. In the output o and 1 represents the present and absent of the symptoms in the disease. 1 indicates the present of symptoms and 0 indicates the absent of the symptom.

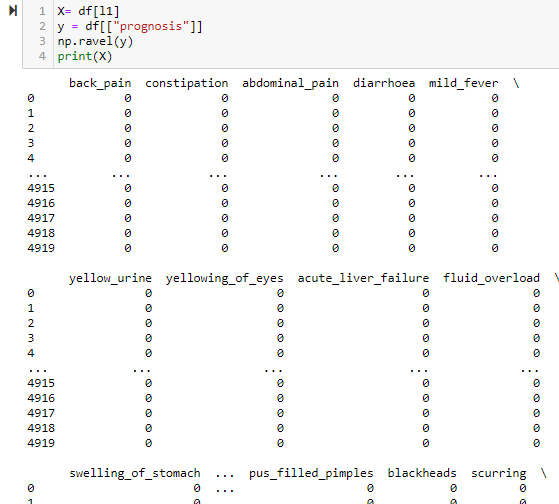


Figure6.6: labeling

To build a model for prediction various algorithms are used such as

* Decision tree
* Random forest
* Naïve bayse

**Decision Tree algorithm**

The Root=Tk() is used to for to build the gui. Then definig the DecisionTree() function. Here “pred1” is stores the disease predicted by using decision tree algorithm. The inbuilt function called DecisionTreeClassifier() is used to train the model and predict the disease on testing dataset according to symptoms entered by the user. Final disease for decision tree is stored in a variable named “pred1”. Accuracy of predicting the disease is printed using accuracy\_score and confusion matrix is created using confusion\_matrix which are imported from sklearn.metrices.



Figure6.7: decision tree algorithm

Creating a database named “database.db”, and if the database does not exist then, using “sqlite3” database will be created. For storing data for decision tree algorithm "Decision Tree" table is created and values are inserted in table named as Decision Tree using “INSERT” function in sqlite.



Figure6.7.1: decision tree algorithm

**Random forest algorithm**

First defining the random forest() function. The predicted disease with help of random forest algorithm will be store in pred2. Inbuilt RandomForestClassifier() is used to train the model and predict the disease on testing dataset according to symptoms entered by the user. Final disease predicted by random forest algorithm is stored in a variable named “pred2”. Accuracy of predicting the disease is calculated using accuracy\_score and confusion matrix is created using confusion\_matrix which are imported from sklearn metrices.

****

Figure6.8: Random forest

Creating a database named “database.db”, and if the database does not exist then, using “sqlite3” database will be created. For storing data for random forest algorithm "Random Forest" table is created and values are inserted in table named as Random forest using “INSERT” function in sqlite

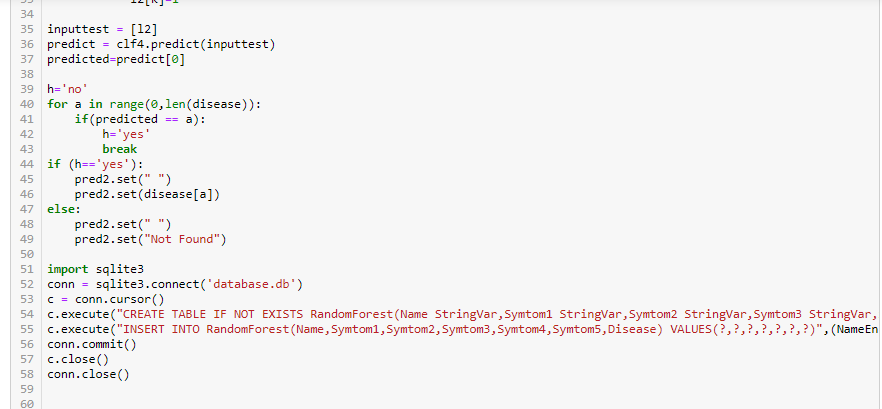
****

Figure6.8.1: Random forest

**Naïve Bayse algorithm**

Disease predicted by Naïve bayse algorithm is stored in pred3. Deining the naiveBayes() function. GaussianNB() is used to train the model and predict the disease on testing dataset according to symptoms entered by the user. Final disease for Naïve Bayes is stored in a variable named “pred3”. Accuracy of predicting the disease is calculated using accuracy\_score and confusion matrix is created using confusion\_matrix which are imported from sklearn.metrices.

**** Figure6.9: Naïve Bayes

Creating a database named “database.db”, and if the database does not exist then, using “sqlite3” database will be created. For storing data for Naïve bayes algorithm "Naïve Bayes" table is created and values are inserted in table named as Naïve bayes using “INSERT” function in sqlite

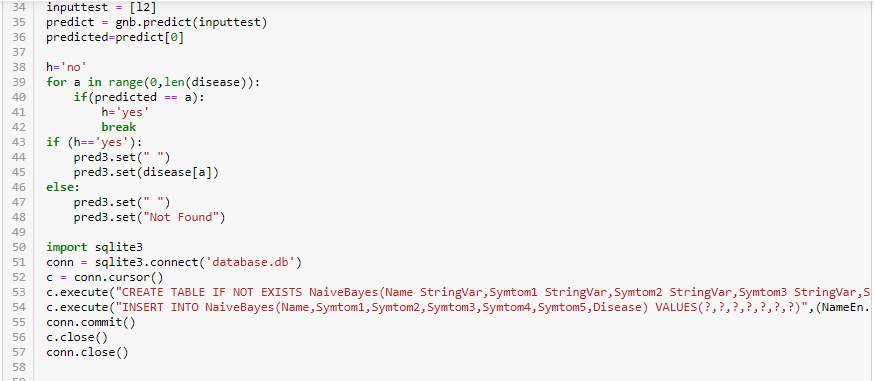
****

Figure6.9.1: Naïve Bayes

**Output**

Here is the final GUI of the disease prediction system where user needs to enter their name and the symptoms. User need to compulsory enter first two symptoms.

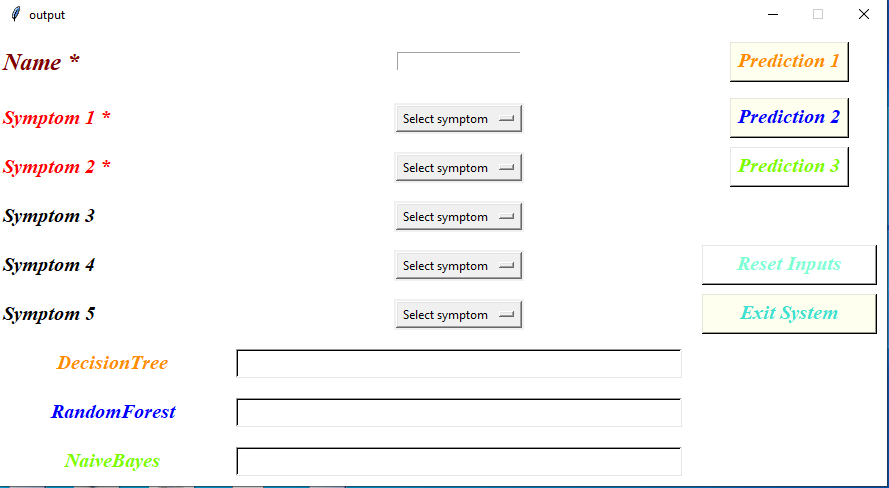
****

Figure6.10: GUI of output

Entering the symptoms on clicking prediction1 the decision tree algorithm runs and predicts the disease. Similarly prediction2 will predict disease using Random forest algorithm and the prediction 3 will predict using the Naïve bayes algorithm. To reset the input user can click on reset button and to exit the system user need to click on exit system button.

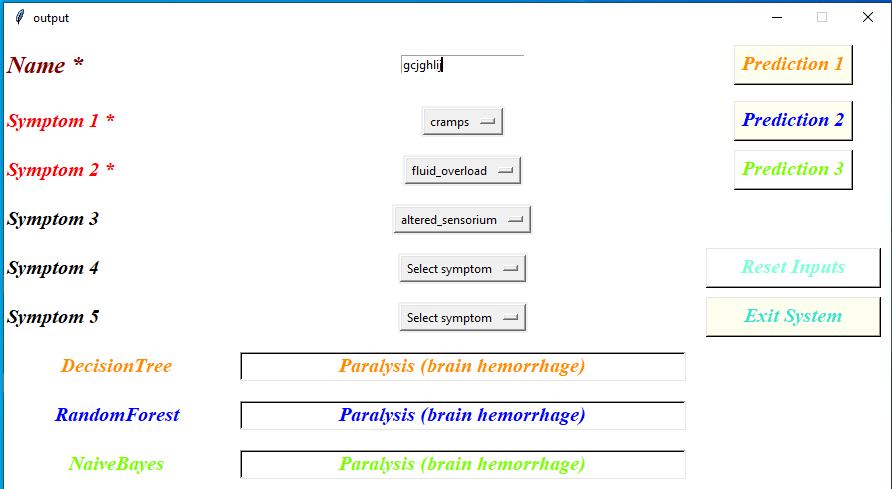
****

Figure6.10.1: GUI of output

On clicking the exit system button, there will be popup message box which will ask for confirmation whether the user wants to exit the system or not.

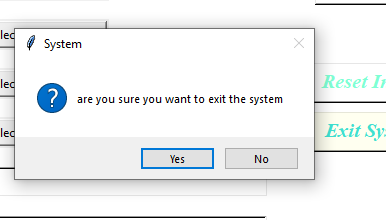


Figure6.11: pop up massage

When user press the exit system button, a pop up message is shown which will ask user that whether they are sure they want to exit system.

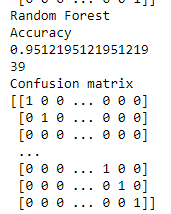
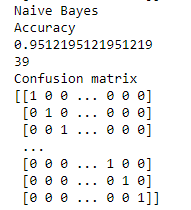
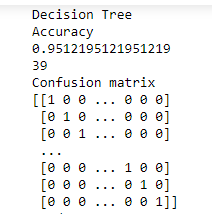


Figure6.12: Accuracy and confusion matrix of algorithms

From these results, we can infer that all the three algorithms work exceptionally well on the dataset. However, Naïve Bayes is perhaps working a little better when compared to the other two algorithms.

|  |  |
| --- | --- |
| **Algorithm** | **Accuracy** |
| Decision Tree | 0.95121 |
| Naïve Bayes | 0.95121 |
| Random Forest | 0.95121 |

Table5: Accuracy Table of algorithm

**Chapter 7: Bibliography/References**

[1] Thirunavukkarasu K, Singh AS, Irfan M, Chowd hury A (2018) Prediction of liver disease using classification algorithms. In: 2018 4th international conference on computing communication and automation (ICCCA)

[2] Prajoona Valsalan, Tariq Ahmed Barham Baomar, Ali Hussain Omar Baabood. “Iot Based Health Monitoring System”. http://www.jcreview.com/ ?mn o=96095 [Access: April 10, 2021]. doi:10.31838/ jcr.07.04.137

[3] S. K. J. and G. S., "Prediction of Heart Disease Using Machine Learning Algorithms.," 2019 1stInternational Conference on Innovations in Information and Communication Technology (ICIICT)*,* 2019, pp. 1-5, doi: 10.1109/ICIICT1.2019.8741465.

[4] Zoabi, Yazeed & Deri-Rozov, Shira & Shomron, Noam. (2021). Machine learning-based prediction of COVID-19 diagnosis based on symptoms. npj Digital Medicine. 4. 10.1038/s41746-020-00372-6.

[5] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease prediction by machine learning over big data from healthcare communities”, IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[6] Sayantan Saha, Argha Roy Chowdhuri et,al “Web Based Disease Detection System”,IJERT, ISSN:22780181,Vol.2 Issue 4, April-2013

[7] S. Vijayarani, S. Dhayanand, Liver disease prediction using svm and na¨ıve bayes algorithms, International Journal of Science, Engineering and Technology Research (IJSETR) 4(4), 816 (2015)

[8] R.D.H.D.P. Sreevalli, K.P.M. Asia, Prediction of diseases using random forest classification algorithm

[9] Allen Daniel Sunny1, Sajal Kulshreshtha, Satyam Singh3, Srinabh, Mr. Mohan Ba, Dr. Sarojadevi H “ Disease Diagnosis System By Exploring Machine Learning Algorithms”, International Journal of Innovations in Engineering and Technology (IJIET) Volume 10 Issue 2 May 2018.

[10] Mohammed Khalilia, Sounak Chakraborty, and Mihail Popescu. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11(1):1, 2011.

**Appendix I**

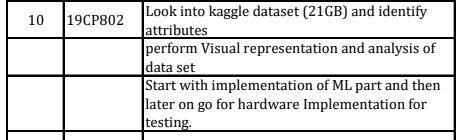
Certificate of paper published in International Journal of Scientific Research and Engineering Trends



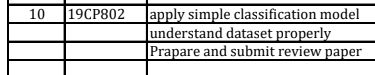
**Appendix II**

**Third semester**

Internal review I

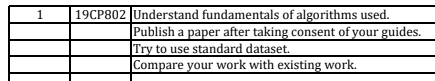


Internal review II

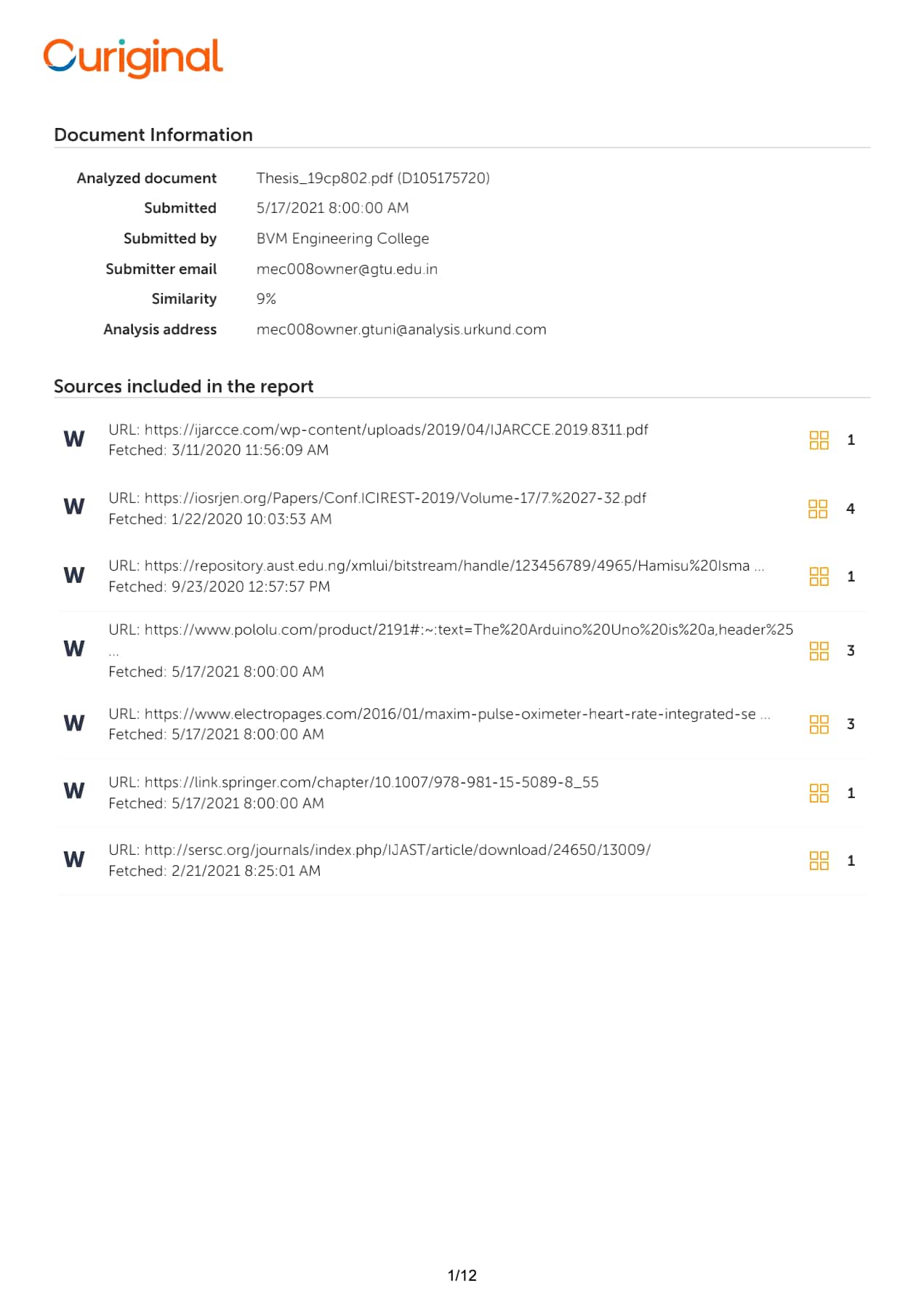


**Fourth semester**

Internal review II



**Appendix III**

****