A PROPOSED IMPLEMENTATION OF END TO END MACHINE LEARNING PROJECT FOR DISEASE PREDICTION

**A PROJECT REPORT**

###### ***Submitted by***

**Bipin Kumar Chaudhary (20BCE10325)**

**Mayank Malpani (20BCE10345)**

**Smitesh Somkuwar (20BCE10355)**

**Areeb Ahmad Siddiqui (20BCE11049)**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

# **COMPUTER SCIENCE AND ENGINEERING**

****

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

**KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

APR 2022

**VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE**

**MADHYA PRADESH – 466114**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“Disease Prediction using Machine Learning”** is the bonafide work of “**Bipin Kumar Chaudhary (20BCE10325), Mayank Malpani (20BCE10345), Smitesh Somkuwar (20BCE10355) and Areeb Ahmad Siddiqui (20BCE11049)”** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported at this time does not form part of any other project/research work based on which a degree or award was conferred on an earlier occasion on this or any other candidate.

**PROGRAM CHAIR PROJECT GUIDE**

Prof Anand Motwani

Senior Teaching Fellowship

School of Computer Science and Engineering

VIT BHOPAL UNIVERSITY

Dr Sandip Mal

Assistant Professor Senior Grade I

School of Computer Science and Engineering

VIT BHOPAL UNIVERSITY

The Project Exhibition II Examination is held on 23/04/2022

**ACKNOWLEDGEMENT**

First and foremost I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to Dr Sandip Mal, Head of the Department, School of Computer Science and Engineering(SCSE) for much of his valuable support encouragement in carrying out this work.

I would like to thank my internal guide Dr. Anand Motwani Sir for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering(SCSE), who extended directly or indirectly all support.

Last, but not least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

**ABSTRACT**

Machine Learning and its approaches are generally helpful for healthcare and biomedical sectors for predicting the disease. For trivial symptoms, the difficulty is to meet the doctors at any time in the hospital. Big Data provides essential data regarding the diseases on the basis of the patient's symptoms. For several medical organizations, disease prediction is important for making the best feasible health care decisions. Conversely, the conventional medical care model offers input as structured that requires more accurate and consistent prediction. This project is planned to develop the multi-disease prediction using the various machine learning concept.

Machine Learning Approach for Identifying Disease Prediction Using Machine Learning is based on prediction modelling that predicts disease of the patients according to the symptoms provided by the users as an input / output to the system.

Here, the different datasets pertain to "Diabetes, lung cancer, liver and heart disease ", from the benchmark UCI repository is gathered for conducting the experiment also few datasets are taken from Kaggle.

Here we will use the concept of supervised Machine Learning in which implementation will be done by applying Decision Tree, Random Forest and few other algorithms which will help in early prediction of diseases accurately and better patients care. The results ensured that the system would be functional and user oriented for patients for timely diagnoses of diseases in a patient.

**Keywords:**Machine Learning; UCI repository; Decision Tree; Random Forest.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO.** |
| 1 | **CHAPTER-1:**  **PROJECT DESCRIPTION AND OUTLINE** Introduction 1.2 Motivation for the work  1.3 Problem Statement  1.4 Objective of the work | 7 |
| 2 | **CHAPTER-2:**  **RELATED WORK INVESTIGATION**  2.1 Introduction  2.2 Existing works with Limitations  2.3 Solution provided by the project | 8 |
| 3 | **CHAPTER-3:**  **REQUIREMENT ARTIFACTS**  3.1 Introduction  3.2 Hardware and Software requirements | 14 |
| 4 | **CHAPTER-4:**  **DESIGN METHODOLOGY AND ITS NOVELTY**  4.1 Methodology and goal  4.2 Functional modules design and analysis  4.3 Software Architectural designs  4.4 User Interface designs | 15 |
| 5 | **CHAPTER-5:**  **TECHNICAL IMPLEMENTATION & ANALYSIS**    5.1Outline  5.2 Technical coding and code solutions | 24 |
| 6 | **CHAPTER-6:**  **PROJECT OUTCOME AND APPLICABILITY**  6.1 Key implementations outlines of the System  6.2 Significant Outcomes  6.3 Project applicability on Real-world applications | 31 |
| 7 | **CHAPTER-7:**  **CONCLUSIONS AND RECOMMENDATION**  7.1 Conclusion  7.2 Limitation/Constraints of the System  7.3 Future Enhancements | 32 |
| 8 | **REFERENCES** | 34 |

**CHAPTER 1**

**PROJECT DESCRIPTION AND OUTLINE**

* 1. **INTRODUCTION**

Now-a-days, people face various diseases due to the environmental condition and their living habits. Especially due to Covid, one never knows what problem he can encounter. So, the prediction of disease at earlier stage becomes important task. Disease prediction using patient treatment history and health data by applying data mining and machine learning techniques is ongoing struggle for the past decades. The correct prediction of disease is the most challenging task. The recent success of deep learning in disparate areas of machine learning has driven a shift towards machine learning models that can learn rich, hierarchical representations of raw data with little pre-processing and produce more accurate results.

With the help of disease data, Machine Learning finds hidden pattern information in the huge amount of data, which in turn helps the patient to identify the problem. The main focus is on to use machine learning in healthcare to supplement patient care for better results. Machine learning has made easier to identify different diseases and diagnosis correctly. Predictive analysis with the help of efficient multiple machine learning algorithms helps to predict the disease more correctly and help treat patients. Machine learning in healthcare aids the humans to process huge and complex medical datasets and then analyse them into clinical insights. This then can further be used by physicians in providing medical care. Hence machine learning when implemented in healthcare can leads to increased patient satisfaction.

* 1. **MOTIVATION FOR THE WORK**

There is a demand to make such a system that will help end users to predict diseases on the basis of symptoms given in it without visiting hospitals. By doing so, it will decrease the rush at OPD‟s of hospitals and bring down the workload on medical staff. Not only this, this system will reduce the costly treatment and panic moment at the end stages so that proper medication can be provided at the right time and we can lower down the death rate as well. This system also consists of a feature of Database which stores the data entered by the end users and the name of the disease the patient is suffering from that can be used as a past record and will help in further treatment in future. The analysis accuracy is increased by using Machine Learning algorithms. Altogether this system will help in easier health management.

**1.3 PROBLEM STATEMENT**

In day to day life, it is difficult for one person to go to a doctor and get a check of diseases as it takes quite a lot amount of money as well as time. In our developing and technology dependent life we totally rely on gadgets. So, there should be a way, with whose help a person can at least check whether he has a particular disease or not. Using tech like machine learning in predicting the diseases using symptoms or concerned medical data is a need of the hour.

**1.4 OBJECTIVE OF THE WORK**

The aim of this study is to test the proposed hypothesis that supervised ML algorithms can improve health care bythe accurate and early detection of diseases. In this study, weinvestigate studies that utilize more than one supervised MLmodel for each disease recognition problem. This approachrenders more comprehensiveness and precision because theevaluation of the performance of a single algorithm overvarious study settings induces bias which generates imprecise results. The analysis of ML models will be conducted on few diseases located at heart, kidney, breast, and brain. For the detection of the disease, numerous methodologies will be evaluated such as RF, DT, SVM, and LR. At the end of this literature, the best performing ML models inrespect of each disease will be concluded.

**CHAPTER 2**

**RELATED WORK INVESTIGATION**

**2.1 INTRODUCTION**

Since the arrival of advanced computing, the doctors' still requires the technology in various possible ways like surgical representation process and x-ray photography, but the technology perceptually stayed behind. The method still requires the doctor's knowledge and experience due to alternative factors starting from medical records to weather conditions, atmosphere, blood pressure and numerous alternative factors. The huge numbers of variables are granted as entire variables that are required to understand the complete working process itself, nevertheless, no model has analyzed successfully. To tackle this drawback, Medical decision support systems must be used. This system can assist the doctors to make the correct decision. We are applying machine learning to maintained complete hospital data Machine learning technology which allows building models to get quickly analyze data and deliver results faster, with the use of machine learning technology doctors can make a big decision for patient diagnoses and treatment choices, which leads to enhancement of patient healthcare services. Healthcare is the most prime example of how machine learning is used in the medical field.

**2.2 EXISTING WORKS**

1. **Heart Disease**

Marimuthu et al. [16] aimed to predict heart diseases using supervised ML techniques. The authors structured the attributes of data as gender, age, chest pain, gender, target and slope [16]. The applied ML algorithms that were deployed are DT, KNN, LR and NB. As per analysis, the LR algorithm gave a high accuracy of 86.89%, which deemed to be the most effective compared to the other mentioned algorithms.

In 2018, Dwivedi [17] attempted to add more precision to the prediction of heart diseases by accounting for additional parameters such as Resting blood pressure, Serum Cholesterol in mg/dl, and Maximum Heart Rate achieved. The used dataset was imported from the UCI ML laboratory; it was comprised with 120 samples that were heart disease positive, and 150 samples that were heart disease negative. Dwivedi attempted to evaluate the performance of Artificial Neural Networks (ANN), SVM, KNN, NB, LR and Classification Tree.

At the appliance of tenfold cross validation, the results showed that LR has the highest classification accuracy and sensitivity, which shows high dependability at detecting heart diseases [17]. This conclusion is strengthened by the findings of Polaraju [18] and Vahid et al. [19], where the Logistic Regression outperformed other techniques such as ANN, SVM, and Adaboost. The studies excelled in conducting an extensive analysis on the ML models. For instance, various hyper-parameters were tested at each ML algorithm to converge to the best possible accuracy and precision values. Despite that advantage, the small size of the imported datasets constraints the learning models from targeting diseases with higher accuracy and precision.

1. **Liver Disease**

P. Kuppan et al. [10] in their research work authors have worked on doing an analysis of the data related to Liver Disorder with the help of Naive Bayes, Decision Table, and J48. However, attributes like case history of the patient, diabetes, smoking, obesity, alcohol intake, smoking etc were used. Based upon the given database it has concluded Jagdeep Singh et al. / Procedia Computer Science 167 (2020) 1970–1980 1973 4 Jagdeep Singh et al. / Procedia Computer Science 00 (2019) 000–000 that male people are having more liver disorder than the females. Age group of 35-65 is mostly affected and out of these 26% people are having the disorder because of alcohol, smoking contributed to 22% of people, obesity, and diabetic of 4 & 5 percent respectively.

A. Gulia et al. [11] in their proposed work researchers have done classification of the liver patient data using the algorithms like Bayesian Network, Support Vector Machine, J48, Multi-Layer Perceptron and Random Forest. The data from the UCI repository which is afforded by Center of Machine Learning and Intelligent Systems has used. After completion of their three-phase analysis, the Random Forest Algorithm is the best one with an accuracy of 71.87% has been concluded.

Y. Kumar et al. [12] in their proposed work researchers have used Rule-Based Classification Model (RBCM) for the prediction of liver diseases. Without the rule-based classification the efficiency of all the common algorithm decreases was analyzed. In their proposed work 20 rules were used for the classification of liver diseases. The decision tree-based algorithm gives the best performance using rule-based classification and accordingly its accuracy decreases when rule-based is not used. M. Pasha et al. [13] work on the dataset from the UCI repository is used which is having 583 instances, the metalearning algorithms like Grading, logit boost, Adaboost, and Bagging were used. The comparisons of the algorithms based upon the amount of correct and incorrect classifications and time of execution have done. After doing detailed analysis the grading is the best algorithm in terms of accuracy and execution time have been concluded.

M. Abdar et al. [14] in their research work focuses on the early prediction of liver disease using Multilayer Perceptron Algorithm (MLPNN) which uses (CART) (classification and regression tree, (CHAID) Chi-square Automatic interaction detector, See5(C5.0). Their dataset is from UCI repository of the University of California, Irvine relevant to Indian Liver Patient Dataset (ILPD). From their results, it can be concluded that MLPNNBCHAID is the best algorithm with an innovative accuracy of 14.57%. The 70% of the data as a training data and rest of the 30% for the testing stage were used.

1. **Diabetes**

Yasodhaet al.[1] uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient’s data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Aiswaryaet al. [2] aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. Using PIMA dataset and cross validation approach the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

Gupta et al. [3] aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of same classifiers when implemented on some other tools which includes Rapidminer and Matlabusing the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and BayesNet algorithms. The result shows that Jgraft shows highest accuracy i.e 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner.

**d**. **Lung Cancer**

Roy et al[34]. They use a combination of image processing biomedical techniques and information discovery in data to improve accuracy and assess precise significance for early detection of lung carcinoma. The representation of the lungs acquired from CT (Computer Tomography) The scan images are pre-processed, and the Region of Interest is segmented (ROI) is performed. The Random Forest procedure is used to distinguish the distinct features. Using an SVM Classifier, the SURF (Speeded Up Robust Functionality) algorithm was used to extract features like entropy, co-relation, power, and variance from Saliency Enhanced images. The image's classification determines if it is safe or toxic (carcinomic). CT scan images were used as the dataset. The SVM classification and random forest algorithm were used to carry out the whole operation. Using SVM classification, the best outcome is achieved. This technique is 94.5 percent effective in general, 74.2 percent sensitive, 66.3 percent recall, and 77.6% specific.

For lung cancer diagnosis, Faisal et al [12] recommend evaluating machine learning classifiers as well as, classifiers such as Multilayer perceptron (MLP), Nave Bayes, Decision 143 Tree, Neural Network, Gradient Boosted Tree, and SVM are evaluated. The dataset was downloaded from the UCI registry and is used to analyze random forest and plurality voting-based ensembles for predict lung cancer. Gradient Boosted Tree was found to outperform all other person and ensemble classifiers. Gradient-boosted Tree outperformed all others as well as ensemble classifiers, achieving 90% precision, according to performance assessments.

Delta Radiomics uses the machine learning methods proposed by Baskar et al [35] to extract the characteristics of the cancer nodules. Lung cancer nodule malignancy is predicted by using the Support Vector Machine (SVM). The SVM can examine compact features in a lung cancer nodule photograph, and image classification is useful in distinguishing between the multiple nodules. As a result, SVM is recommended as the best tool for diagnosing.

A technique k-Nearest-Neighbors was developed by Maleki et al [39], for which a genetic algorithm was used to efficiently pick features, to reduce the dimensions of the dataset and to improve the speed of the classifier. The experimental approach is used to determine the best value for k to increase the precision of the proposed algorithm. Use of the proposed solution to the database for lung cancer shows 100% accuracy.

**2.3 SOLUTION PROVIDED BY THE PROJECT**

While there are many, disease prediction model available, with new research papers regularly publishing most of them, detail their experiments and findings limited to only one disease such as heart, liver etc. There are very few works which use multi disease classification models, i.e predicting multiple disease at the same time also many of the works lacks an user interface.

To overcome the problems posed by the existing models, we are proposing such a system that will flaunt a simple, cost effective , elegant User Interface and also be time efficient . Our proposed system bridges the gap between doctors and patients which will help both classes of users to achieve their goal. This system is used to predict diseases according to symptoms. This system is used by end-users. We have named this system as 'MedBuddy'. This system is for those people who are always fretting about their health.

In this proposed system we are going to take down required symptoms from the users and evaluate them by applying algorithms such as Decision Tree, Random Forest , and Logistic Regression which will help in getting accurate prediction. Our system will explore and merge more datasets which includes large diversity of population to get more effective results and thus our system will improve and enhances the accuracy of the results. Along with the increased accuracy rate, we will proliferate the reliability of our system for this job and can gain the trust of patient in this system. Apart from all these, our system will comprise of a Database for storing the data entered by the users and the name of the disease the patient is suffering from which can be used as a reference in future for further treatment. Hence this system will contribute in easier health management with better satisfaction to the users.

**CHAPTER 3**

**REQUIREMENTS AND ARTIFACTS**

**3.1 INTRODUCTION**

Our project “Med Buddy” is a simple user-friendly web application for disease prediction purposes. It uses HTML (hypertext markup language) as its basic frame work for website, CSS (cascading style sheet) to organize and neatly tie up the containers, Java script so that it could be dynamic and add functionalities to our website such as the user input. For the backend we have used Flask as it provides an easy-to-use functional interface and helps in easily manipulating the form values from the ‘html forms’.

**3.2 HARDWARE AND SOFTWARE REQUIREMENTS**

Note that our website is not been hosted so it requires the following to execute our project.

Hardware requirement:

* 1.6 GHz or faster processor
* 1 GB of RAM

Software requirement:

* Any code editor (we have used VS code) for flask.
* Any browsing engine
* Google Collab
* Python version (>3.0)

For Windows-

* Windows 7 (with .NET Framework 4.5.2), 8.0, 8.1 and 10, 11 (32-bit and 64-bit)
* Microsoft .NET Framework 4.5.2

For Linux-

* Linux (Red Hat): Red Hat Enterprise Linux 7, CentOS 8, Fedora 24
* GLIBCXX version 3.4.21 or later
* GLIBC version 2.15 or later

**CHAPTER 4**

**DESIGN METHODOLOGY AND ITS NOVELTY**

* 1. **METHODOLOGY AND GOAL**
* From the starting human health has been the one of the most important topic to the human society.
* Over the period of time it has been continuously developing at an exponential rate.
* The identification and prediction of such diseases at their earlier stages are much important, so as to prevent the extremity of it. It is difficult for doctors to manually identify the diseases accurately most of the time.
* Goal of our project is to predict the possibility of having such diseases by using machine learning.
* Machine learning is the understanding of computer system under which the machine learning model learn from data and experience.
* The machine learning algorithm has two phase:  
   1)Training and 2)Testing.

**IDENTIFICATION OF THE PROBLEM:**

This Module deals with the identification of the type of the problem, whether it is a classification problem or a regression problem etc

**DATASET PREPARATION AND PREPROCESSING**

This module deals with the filtering of data. Separating the data. Identifying the feature variables (input) and the Labels(output). Feature imputation: filling missing values .Feature encoding (turning values into numbers). One hot encoding: Turn all unique values into lists of 0’s and 1’. Label Encoding: Turn labels into distinct numerical values.

**DATASET SPLITTING**

This module deals with the Splitting the data into the Training set (usually 70-80% of data): Model learns on this. Validation set (usually 10-15% of data): Model hyperparameters are tuned on this. Test set (usually 10-15% of data). The Final test is done on this set.

**MODELLING**

Choosing an algorithm, whether to choose Supervised algorithms – Linear Regression, Logistic Regression, Random forests, etc. Overfit the model, reducing the overfitting with regularization. Ensemblingand Hyper parameter Tuning – run a bunch of experiments with different settings and see which works best.

**ANLYSIS/EVALUATION**

Comparing the model on various evaluation metrics such as Classification- Accuracy, Precision, Recall, F1,Confusion matrix, Mean average precision (object detection), Regression – Mean Squared Error(MSE), Mean Absolute Error (MAE) and Coefficient of Determination or sometimes also known as Goodness of fit( R2)

**MODEL DEPLOYMENT**

Deploying the model for the use of Public, In our case deploying, the model on a website with Flask as  back-end framework.

* 1. **FUNCTIONAL MODULES DESIGN AND ANALYSIS**We will be using our lung disease prediction model for explanations and examples and it would be similar for rest of the models.
* First we checked for any missing files in the csv file.
* We encoded the ‘yes’ and ‘no’ of **LUNG\_CANCER** feature as 1 and 0 where 1 represent ‘YES’ and 0 represents ‘NO’ and similarly for ‘GENDER’ feature we have encoded ‘MALE’ as 1 and ‘FEMALE’ as 0.
* Normalization of data

=

Here Xmax and Xmin are the maximum and the minimum values of the feature respectively.

* The data set is imported and it is split into the Train set and Test set with the test\_size = 0.20.
* We have used logistic regression, decision tree classifier, random forest classifier.

**Logistic regression**

LR is considered as an extension of ordinary regression and can model only a dichotomous variable which usually represents the occurrence or non-occurrence of an event. LR helps in finding the probability that a new instance belongs to a certain class. Since it is a probability, the outcome lies between 0 and 1. Therefore, to use the LR as a binary classifier.

**Decision tree (DT)**

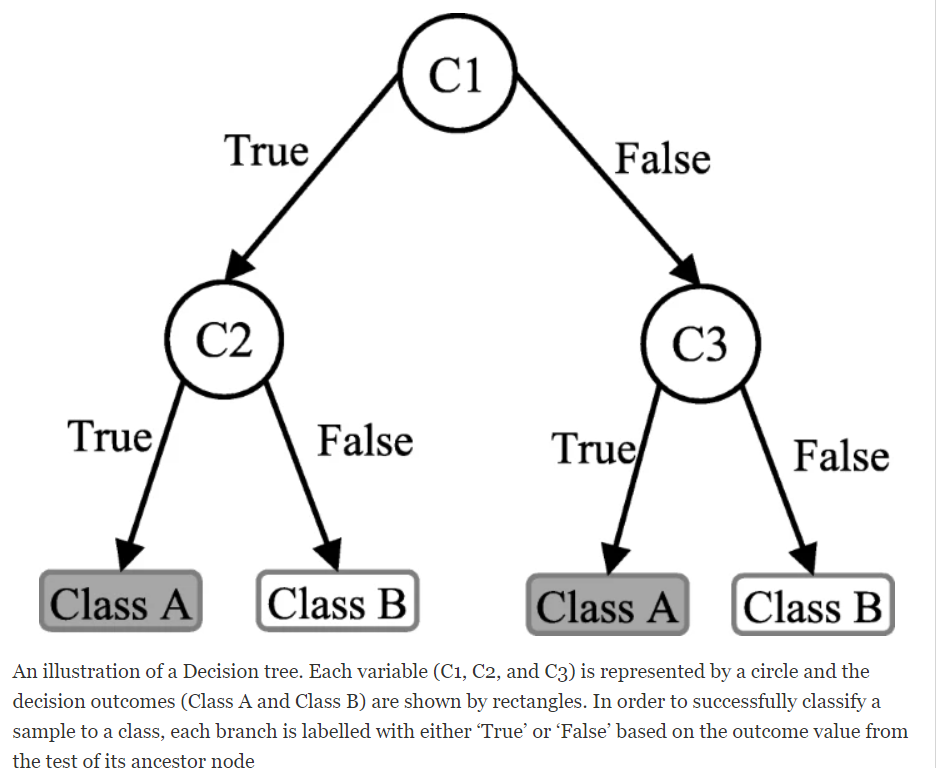
DT is one of the earliest and prominent machine learning algorithms. A decision tree models the decision logics i.e., tests and corresponds outcomes for classifying data items into a tree-like structure. The nodes of a DT tree normally have multiple levels where the first or top-most node is called the root node. All internal nodes represent tests on input variables or attributes. Depending on the test outcome, the classification algorithm branches towards the appropriate child node where the process of test and branching repeats until it reaches the leaf node. The leaf or terminal nodes correspond to the decision outcomes. DTs have been found easy to interpret and quick to learn, and are a common component to many medical diagnostic protocols . When traversing the tree for the classification of a sample, the outcomes of all tests at each node along the path will provide sufficient information to conjecture about its class.

Fig 1. An illustration of Decision tree

**Random Forest (RF)**

RF is an ensemble classifier and consisting of many DTs similar to the way a forest is a collection of many trees. DTs that are grown very deep often cause overfitting of the training data, resulting a high variation in classification outcome for a small change in the input data. They are very sensitive to their training data, which makes them error-prone to the test dataset. The different DTs of an RF are trained using the different parts of the training dataset. To classify a new sample, the input vector of that sample is required to pass down with each DT of the forest. Each DT then considers a different part of that input vector and gives a classification outcome. The forest then chooses the classification of having the most ‘votes’ (for discrete classification outcome) or the average of all trees in the forest.

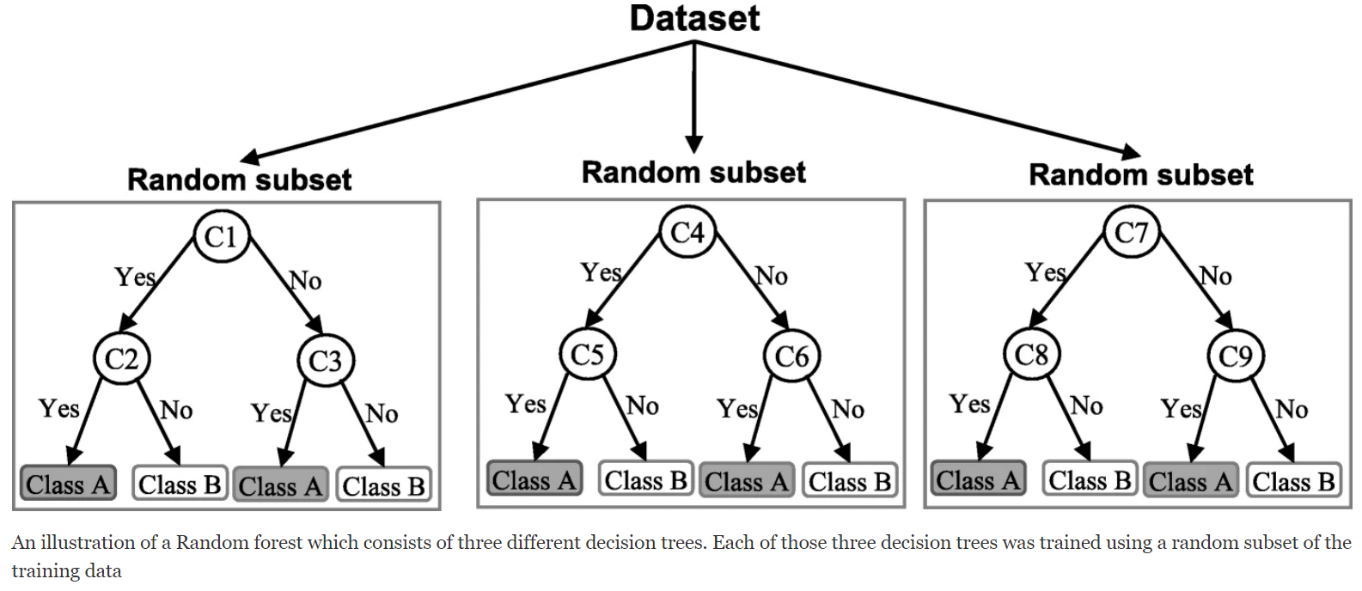


Fig 2. An illustration of Random Forest Classifier

* 1. **SOFTWARE ARCHITECTURAL DESIGNS**
* Now we will discuss the system architecture of this machine learning project and this includes all the steps required to build the proper machine learning project from scratch.
* We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine learning Model performance.

We can define the machine learning workflow in 5 main stages.

* Gathering data
* Data pre-processing
* Researching the model that will be best for the type of data
* Training and testing the model
* Evaluation

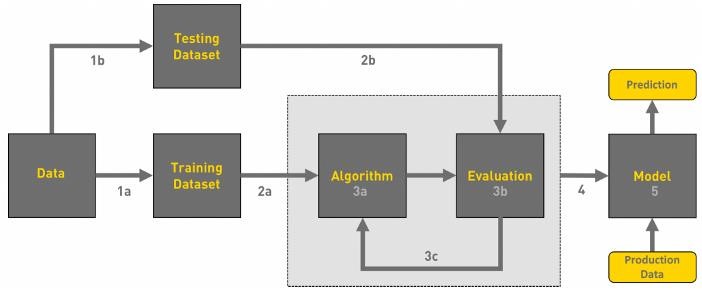


Fig 3. Workflow of a Machine Learning Model

**1. GATHERING DATA**

The process of gathering data depends on the type of project we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data.

The data set can be collected from various sources such as a file, database, sensor and many other such sources but the collected data cannot be used directly for performing the analysis process.

**2. DATA PRE-PROCESSING**

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately.

In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data pre-processing and 20% time to actually perform the analysis.

**3. RESEARCHING THE MODEL THAT WILL BE BEST FOR THE TYPE OF DATA**

Our main goal is to train the best performing model possible, using the pre-processed data.

#### 

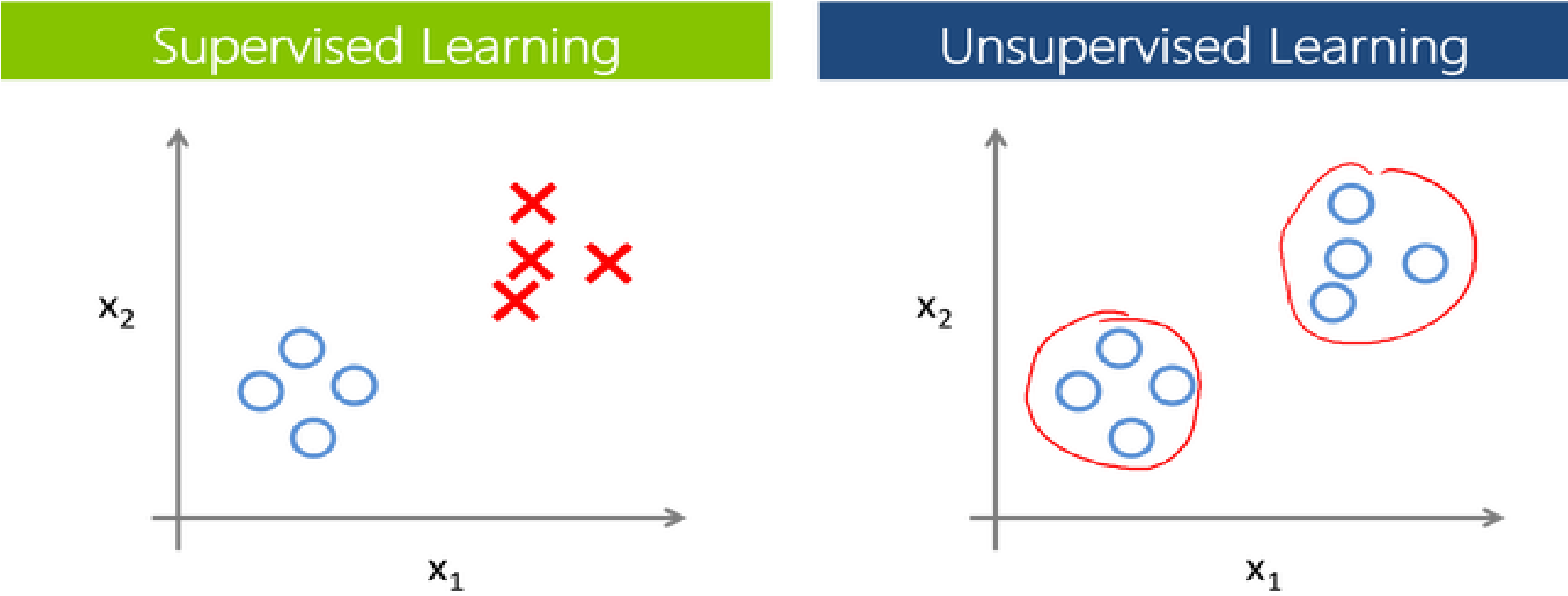
Fig 4. Different types of Learning Supervised and Unsupervised

#### **4. TRAINING AND TESTING THE MODEL ON DATA**

You train the classifier using ‘training data set’, tune the parameters using ‘validation set’ and then test the performance of your classifier on unseen ‘test data set’. For training a model we initially split the model into 3 three sections which are ‘Training data’ ,‘Validation data’ and ‘Testing data’.

**5.EVALUATION**

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To improve the model we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives



**4.4 USER INTERFACE AND DESIGNS**

Flask is a Python-based micro framework used for developing small scale websites. Flask is very easy to make Restful API’s using python.

Home page:

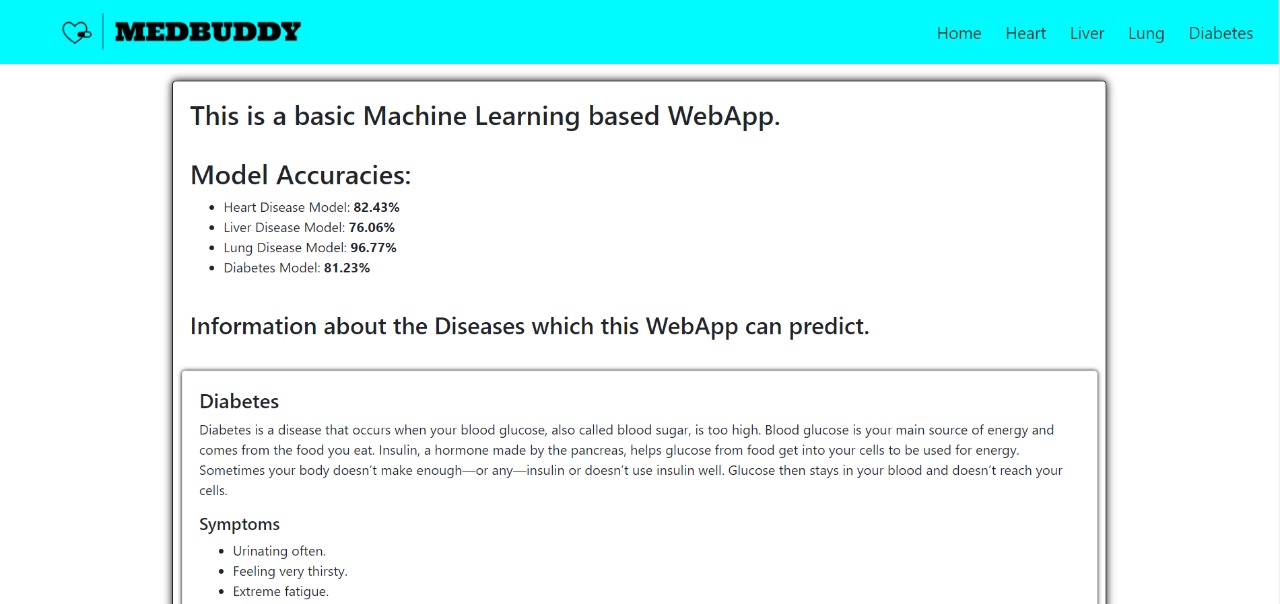
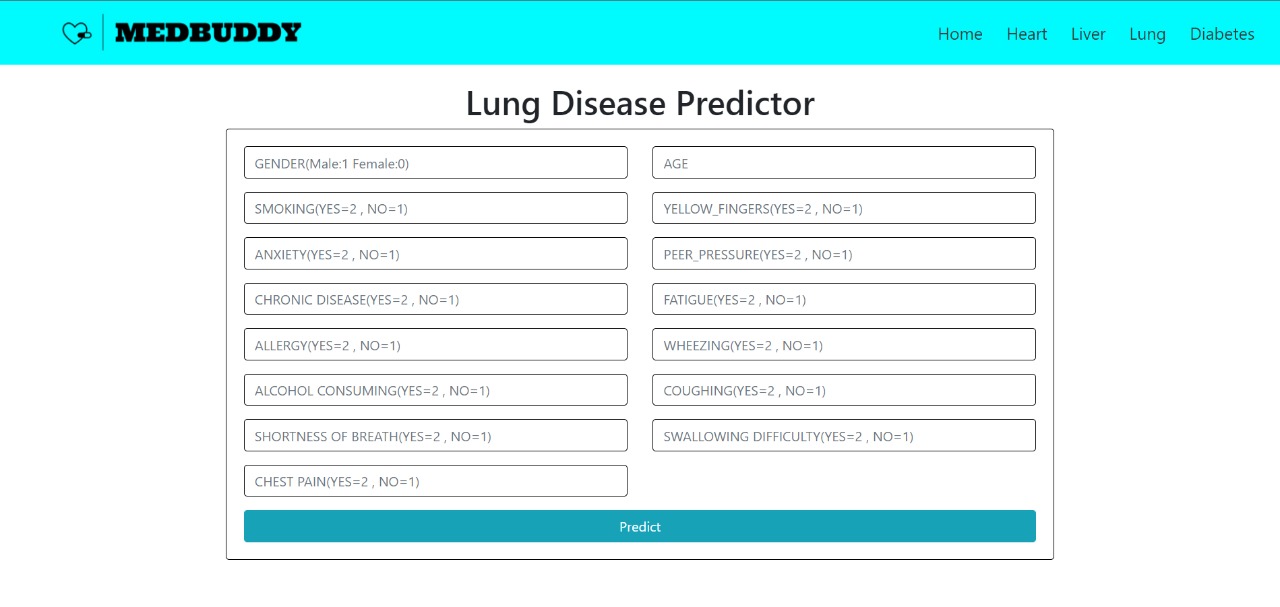


Fig 5. UI of the Homepage of the website

Lung disease predictor:

Fig 6. UI of the Lung Disease Predictor 

Result for unhealthy inputs:

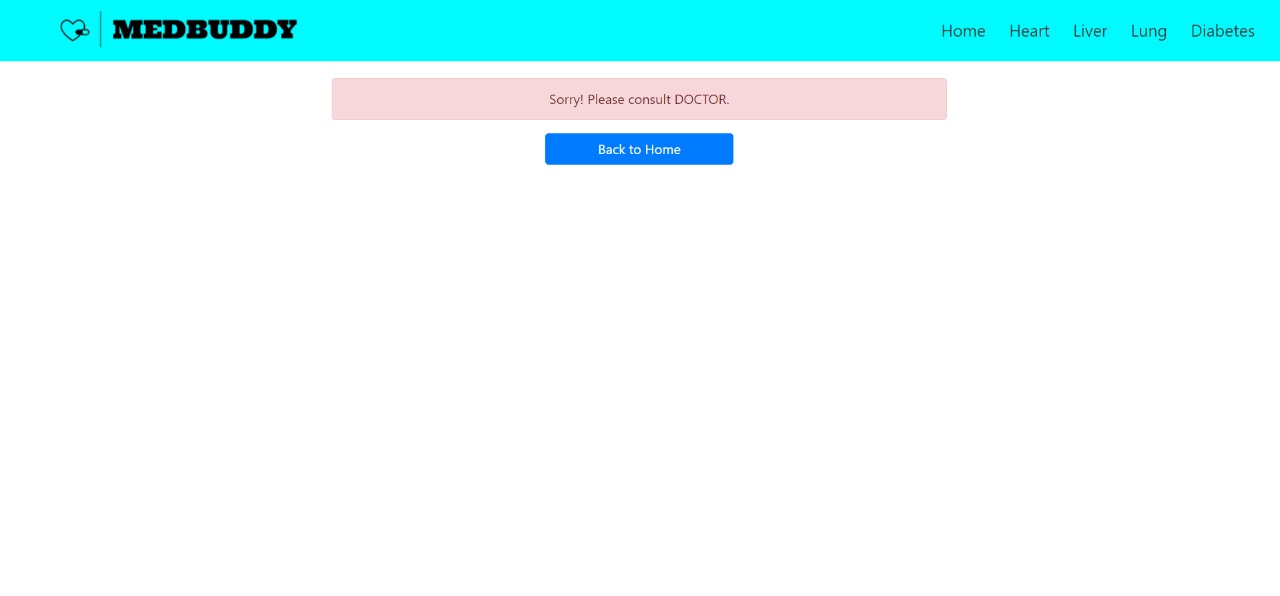
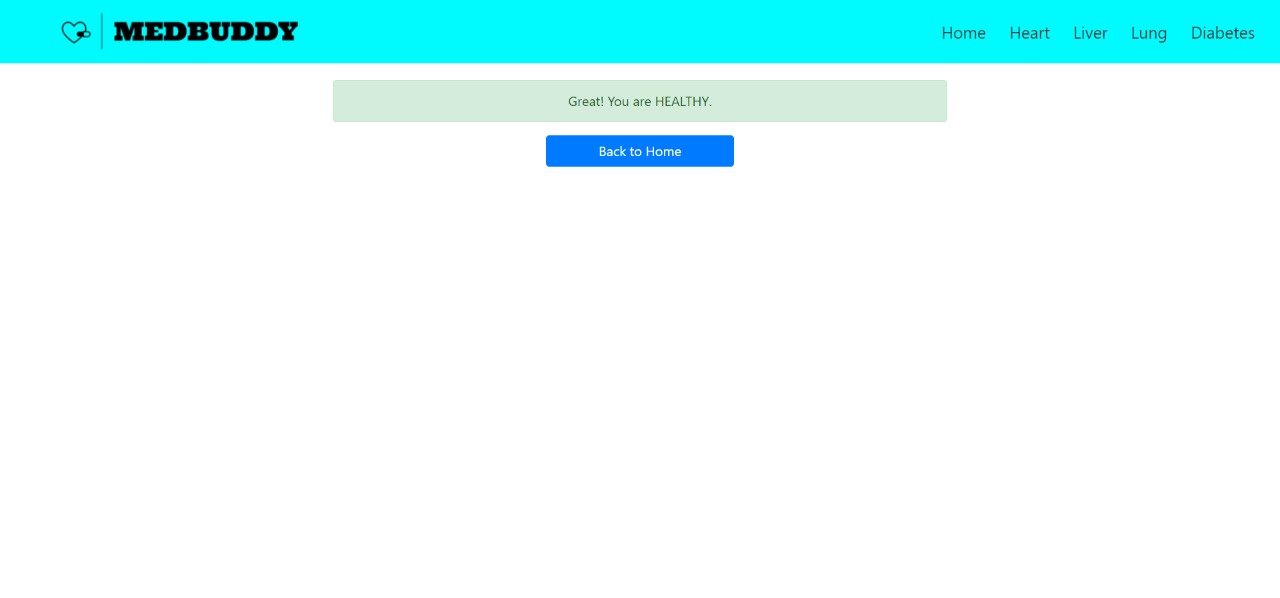


Fig 7. Message for Unhealthy Condition

Result for healthy inputs:

Fig 8. Message for Healthy Condition

**4.5 SUMMARY**We will save our trained model to the disk using the pickle library. Pickle is used to serializing and de-serializing a Python object structure. In which python object is converted into the byte stream. dump() method dumps the object into the file specified in the arguments.  
In our case, we want to save our model so that it can be used by the server.

**CHAPTER 5**

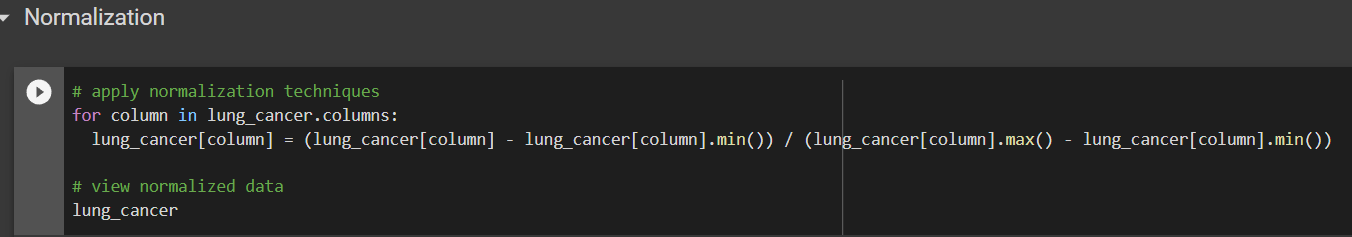
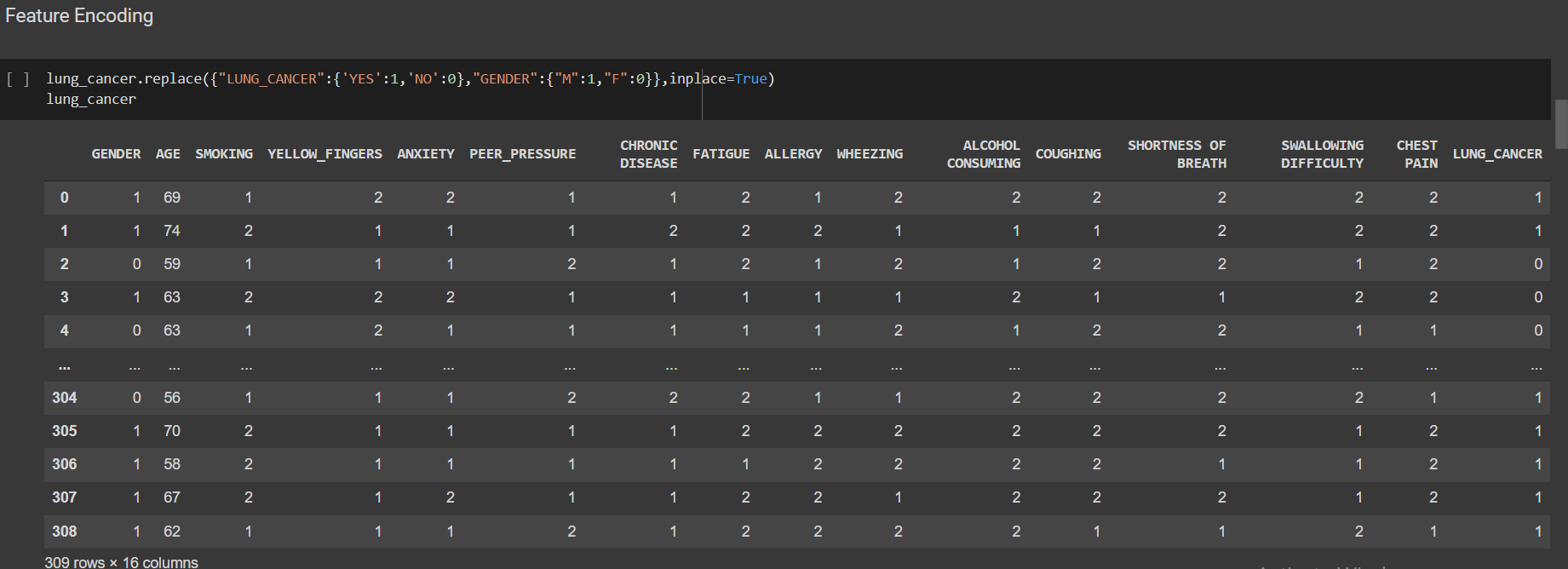
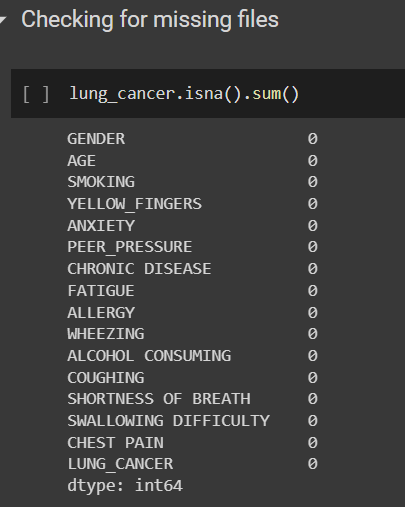
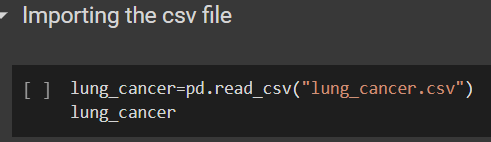
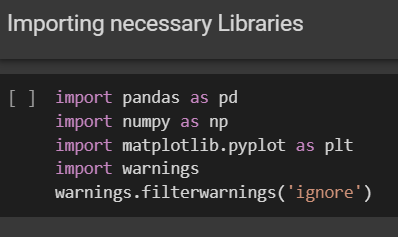
**TECHNICAL IMPLEMENTATION & ANALYSIS**

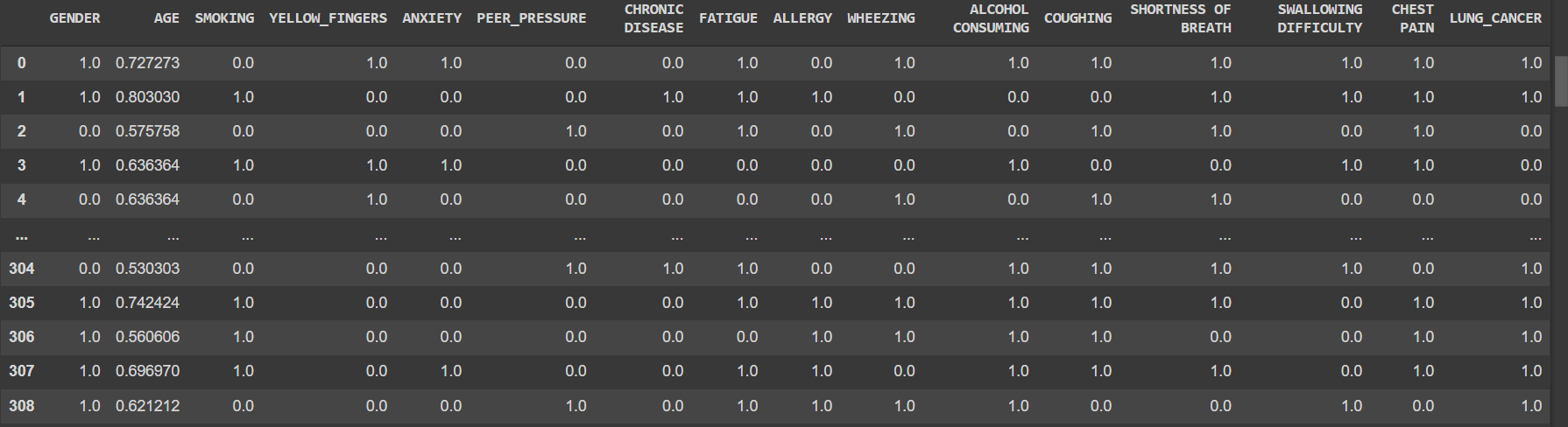
5**.1 OUTLINE**

In supervised machine learning algorithms, a labelled training dataset is used first to train the underlying algorithm. This trained algorithm is then fed on the unlabeled test dataset to categories them into similar groups.

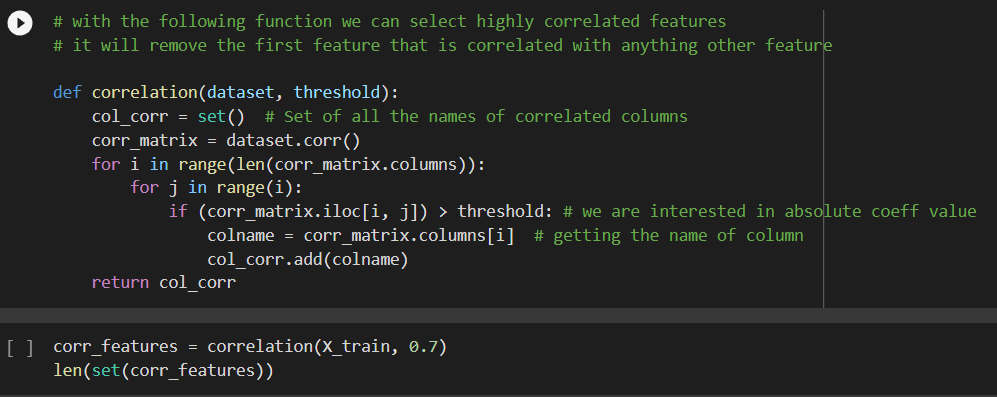
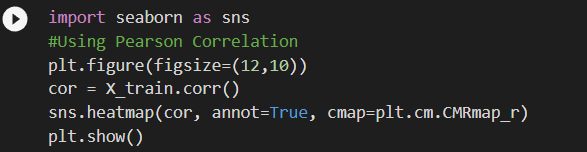
In classification problems, the underlying output variable is discrete

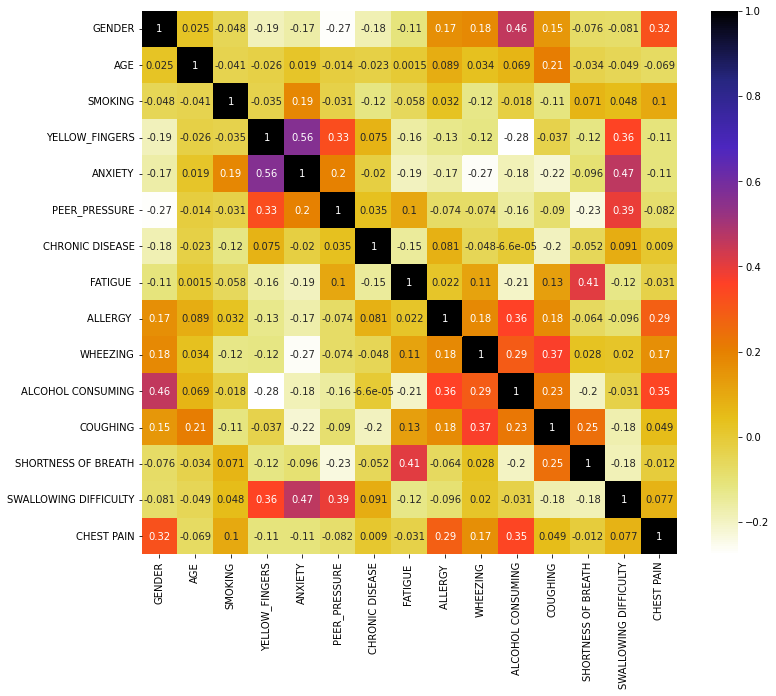
**5.2 TECHNICAL CODING AND CODE SOLUTION**



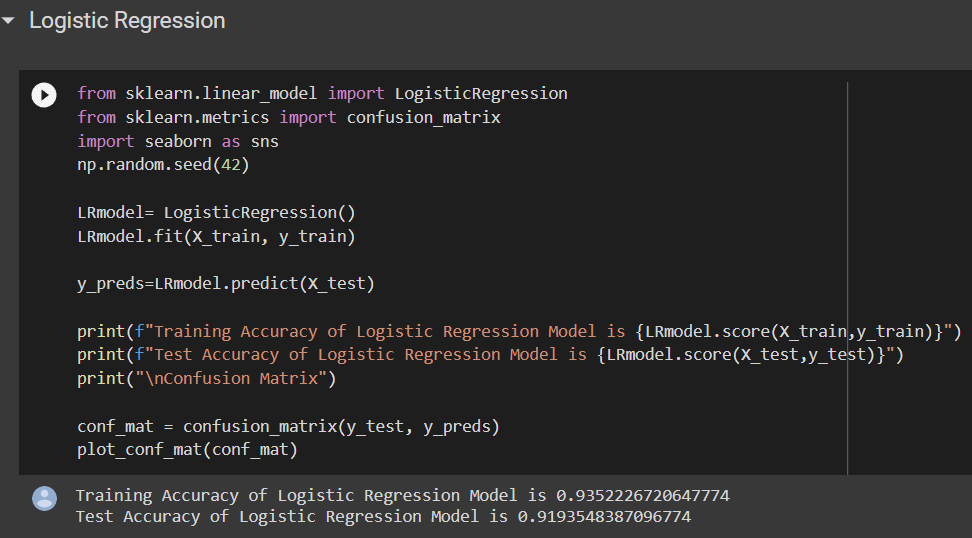


**FEATURE SELECTION**

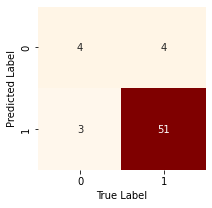


**Correlation heatmap**

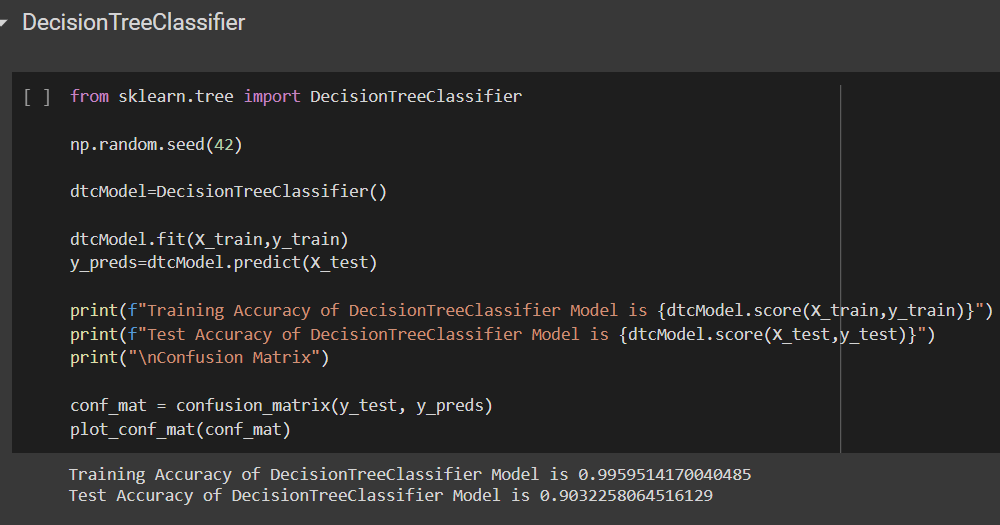
**Logistic Regression Model**



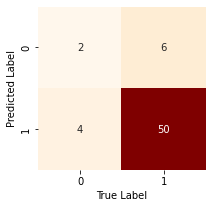
**Confusion Matrix**

****

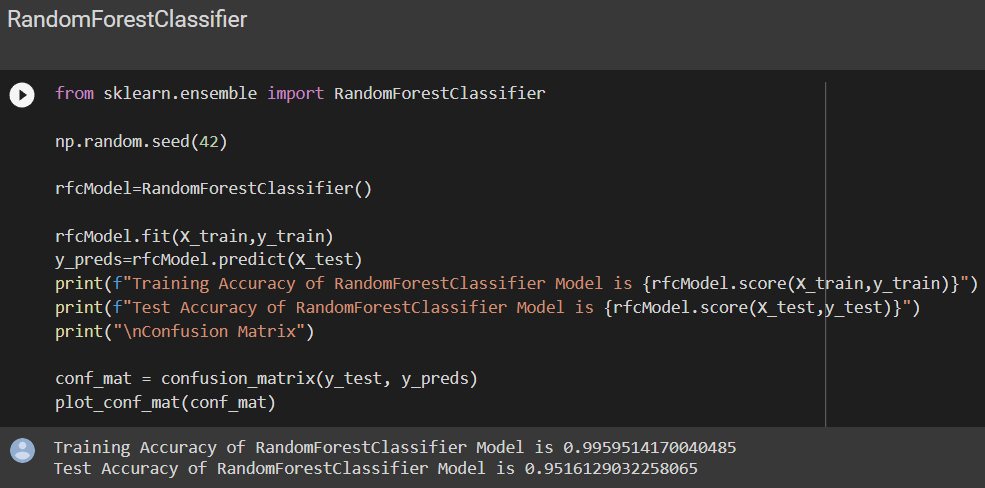
**Decision Tree Classifier Model**



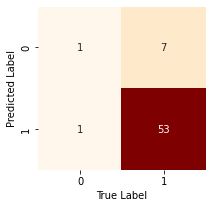
**Confusion Matrix**

****

**Random Forest Classifier Model**



**Confusion Matrix**



**5.5 SUMMARY**

* Out of all machine learning algorithms Random Forest gave the highest accuracy of 95.16%
* The random forest model combines the predictions of the estimators to produce a more accurate prediction
* It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions.

**CHAPTER 6**

**PROJECT OUTCOME AND APPLICABILITY**

**6.1 KEY IMPLEMENTATION AND OUTLINE OF THE SYSTEM**

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the

symptoms which are given by the patients or any user. The system processes the symptoms provided by the

user as input and gives the output as the probability of the disease. Naïve Bayes classifier is used in the

prediction of the disease which is a supervised machine learning algorithm. The probability of the disease is

calculated by the Naïve Bayes algorithm. With an increase in biomedical and healthcare data, accurate

analysis of medical data benefits early disease detection and patient care

Disease Prediction using Machine Learning is the system that is used to predict the diseases from the symptoms which are given by the patients or any user. The system processes the symptoms provided by the user as input and gives the output as the probability of the disease. Random Forest, Decision Tree, Logistic Regression etc. are used in the prediction of the disease which are supervised machine learning algorithms. The probability of the disease is calculated by the many different algorithms. With an increase in biomedical and healthcare data, accurate analysis of medical data benefits early disease detection and patient care.

* 1. **SIGNIFICANT OUTCOMES OF THE PROJECT**

After testing out different algorithms for different diseases with their supervised data we got the following as the best accuracy for the specific algorithms which are as the following:

For lung disease-95.16% accuracy given by Random Forest Classifier

Similarly for Heart disease-98.53% accuracy given by Decision Tree

And for Daibetes-82.29% accuracy given by Random Forest Classifier

For Liver Disease-67.52% accuracy given by Random Forest Classifier

* 1. **PROJECTS APPLICABILITY AND OTHER APPLICATIONS**

1. The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This
2. system takes the symptoms of the user from which he or she suffers as input and generates final output as a
3. prediction of disease

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease.

This system gives a user-friendly environment and easy to use.

As the system is based on the web application, the user can use this system from anywhere and at any time. In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the supervised data provided.

As the system is based on the web application, the user can use this system from anywhere and at any time. In

conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the

hospital dat

As the system is based on the web application, the user can use this system from anywhere and at any time. In

conclusion, for disease risk modeling, the accuracy of risk prediction depends on the diversity feature of the

hospital

**CHAPTER 7**

**CONCLUSION AND RECOMMENDATION**

**7.1 CONCLUSION**

The project presented the technique of predicting the disease based on the symptoms, age, and gender of an individual patient. Different Machine learning algorithms were used to carry out the project such as the RandomForest, Decision Tree and Logistic Regression. Almost all the ML models gave good accuracy values. As some models were dependent on the parameters, they couldn’t predict the disease and the accuracy percentage was quite low.

Once the disease is predicted, we could easily manage the medicine resources required for the

treatment. This model would help in lowering the cost required in dealing with the disease and would also improve the recovery process.

We have also created a GUI for better interaction with the system by users which is very easy to operate .This paper shows that Machine Learning algorithm can be used to predict the disease easily with different parameters and models. To conclude, our system is helpful to those people who are always worrying about their health and they need to know what happens with their body. Our main motto to develop this system is to know them for their health. Especially, people who are suffering from mental illness like depression, anxiety.

They can come out of these problems and can live their daily lives easily. Besides, our system provides better accuracy of disease prediction according to symptoms of the user, and also it will provide motivational thoughts and images. In the end, we can say that our system has no boundary of the user because everyone can use this system.

**7.2 LIMITATIONS OF THE PROJECT**

Though our project is able to predict the chances of a person suffering from a disease with a remarkably good accuracy scores, but in fact it is not a doctor and predictive result from the app should not be taken as the final result as for many diseases the dataset was not very large and as long as we don’t have a large dataset for training out model it is very risky to use for the community use as it creates a doubt whether the prediction is correct or not. Also, the web app corresponding to our project requires the user to input the values of several reading which he might not know.

**7.3 FUTURE ENHANCEMENTS**

The main enhancement which we plan to do is to collect more and more data so that the training of our model can be done in the best possible way. Also, the use several other Machine Learning algorithms could also help in taking our accuracy a level up. We would work on the UI a bit more wherein the user need not input every thing only the important features( features selected after feature selection). This will not only make our model light weight but also provide the user a better experience.

Also use of different techniques such as hyperparameters tuning on each model to get the best out of the models, techniques such as boosting and bagging etc, are in our todo for the future version of the project.

**REFERENCES**

1. A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” in 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 1275–1278.
2. Y. Hasija, N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” in 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 1047–1051
3. S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, vol. 19, no. 1, pp. 1– 16, 2019
4. Shrestha,Ranjit.,& Chatterjee,Jyotir. Moy. (2019).Heart Disease Prediction System Using Machine Learning . LBEF Research Journal of Science Technology and Management , 1(2), 115-132
5. Godse, Rudra A.,Gunjal,Smita S., JagtapKaran A .,Mahamuni ,Neha S., &Wankhade, Prof. Suchita. (2019). Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively.International Journal of Advance Research in Computer and Communication Engineering, 8(12), 50-52
6. Anitha ,Dr.S.,& Sridevi,Dr.N. (2019). Heart Disease Prediction Using Data Mining Techniques.Journal of analysis and Computation ,13(2) , 48-55.
7. Bindhika,Galla Siva Sai., Meghana,Munaga., ReddyManchuriSathvika. , &Rajalakshmi. (2020). Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology, 7(4) , 5272-5276.
8. Pingale,Kedar., Surwase, Sushant., Kulkarni,Vaibhav.,Sarage ,Saurabh., &Karve, Prof. Abhijeet .(2019). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 6(12) , 2810-2813.
9. Chauhan Raj H., NaikDaksh N. ,Halpati,Rinal A., Patel,Sagarkumar J. , &PrajapatiMr. A.D. (2020). Disease Prediction using Machine Learning.International Research Journal of Engineering and Technology, 7(5) , 2000-2002.