A

REPORT ON PROJECT STAGE I

DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

SUBMITTED TO SAVITRIBAI PHULE PUNE UNIVERSITY

FOR PARTIAL FULFILLMENT

OF THE REQUIREMENTS

FOR THE AWARD OF THE DEGREE OF

BACHELOR OF ENGINEERING

In

Electronics and Telecommunication Engineering

By

ISHAN DHAWANE PRN 71828686D

HIMANSHU GAWAI PRN 71923926F

KALYANI GAWARE PRN 71923927D

GUIDE

Dr. S. S. NARKHEDE

DEPARTMENT OF

ELECTRONICS AND TELECOMMUNICATION ENGINEERING

PUNE INSTITUTE OF COMPUTER TECHNOLOGY

PUNE – 43

DECEMBER 2020

Department of Electronics and Telecommunication Engineering

Pune Institute of Computer Technology, Pune – 43

CERTIFICATE

This is to certify that the Project Stage I Report entitled

DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

has been successfully completed by

ISHAN DHAWANE PRN 71828686D

HIMANSHU GAWAI PRN 71923926F

KALYANI GAWARE PRN 71923927D

Is a bona fide work carried out by them towards the partial fulfillment of the requirement of the Savitribai Phule Pune University, Pune for the award of the degree of the Bachelor of Engineering in Electronics and Telecommunication Engineering.

Dr. S. S.Narkhede	Dr. S.V. Gaikwad	Dr. R. Sreemathy
Project Guide	HOD, E&TC Dept	Principal, PICT

Place: Pune

Date :

ACKNOWLEDGEMENTS

The satisfaction and euphoria that accompany the successful completion of any task would be impossible without the mention of the people who made it possible, whose constant guidance and encouragement crowned our efforts with success.

I take this opportunity to express my profound gratitude to **Dr. R. Sreemathy** Principal of PICT , for her constant support and encouragement.

I would also like to thank **Dr. S. V. Gaikwad** Professor and Head , Department of Electronics and Telecommunication Engineering, for his constant support.

High appreciation and deepest gratitude are given to my Guide **Dr. S .S. Narkhede** for his invaluable guidance, encouragement, open assistance, patience and strong support throughout this learning project, as well as his confidence in my ability to produce exemplary research for this project. I have great pleasure in expressing my deep sense of gratitude to Professor **Mr N. B. Patil** and **Mr A. A. Bhandekar**, for providing necessary infrastructure and creating good environment.

Finally, a note of thanks to the teaching and non-teaching staff of Dept of Electronics and Telecommunication Engineering, for their cooperation extended to me, and my friends, who helped me directly or indirectly in the course of the project work.

Thanking You

Ishan Dhawane

Himanshu Gawai

Kalyani Gaware

CONTENTS

ABSTRACT

Disease Prediction system is based on predictive modeling which is used to predict the disease of the user on the basis of the symptoms that user provides as an input to the system. The system analyzes the symptoms provided by the user as input and provides only the probability of a disease such as an Outbreak. With the advancement of machine learning technology in the biomedical and healthcare communities, accurate analysis of medical data facilitates early diagnosis and patient care. Nowadays the use of learning materials is increasing in the areas of natural and human resources services, accurate medical research to facilitate early diagnosis, peaceful care and group management. It is a system which provides the user a way to find out the disease using this disease prediction system. It is useful for the user in case he/she doesn't want to go to the hospital or any other clinics, so they can do simple thing just by entering the symptoms and all other useful information the user can get to know the disease he/she is suffering from and also the health industry can get benefit from this system and in just few seconds they can tell the exact and up to some extent the accurate diseases. This Disease Prediction Using Machine Learning is completely done with the help of Machine Learning and Python Programming language with useful libraries such as Pandas , Numpy , Sklearn for it and also using the dataset that is available previously by the hospitals using that we will predict the disease.

Abbreviations and Acronyms	
HTML	Hypertext Markup Language
CSS	Cascading Style Sheets
KNN	K – Nearest Neighbors

List of Symbols	
	Not used any Symbol

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Background :

The worldwide study ensures that the causes of death is due to disease or syndrome including heart issue , cancer , diabetes, communicable and non-communicable diseases as indicated by new information discharged by World Health Organization (WHO) . In India, an aggregate of 58, 17,000 deaths were evaluated from illnesses. Over the last decade heart disease and diabetes is the main reason for death in the world. Almost one person dies of Heart disease and diabetes about every minute in India alone. In order to lower the number of deaths from diseases, there has to be a fast and efficient detection technique. Predicting disease using patient medical history and health data using machine learning strategies has been an ongoing struggle for decades. Also, some methods try to predict the control and progression of the disease.

The healthcare industry has become a major business in the world .This healthcare sector generates a lot of daily health data that can be used to generate predictable patient morbidity data while using medical history and health data. The rapid adoption of the electronic health records system has created an enormous wealth of new information about patients, which is the gold standard for improving people's understanding of human health. Many research industries have used a variety of techniques in the details of disease or medical profiles to predict specific diseases. These methods attempted to predict the recurrence of diseases. Hence by considering the above areas of improvements we need to enhance the disease prediction standards by implementing such system through which people can come to know about the syndrome or disease that they are facing in their daily life.

1.2 Relevance :

Traditionally, standardized methods of mathematical statistics and sentiments, knowledge and experience were used to predict and predict disease risk. However, this practice often leads to errors, unwanted discrimination, and high costs, and negatively affects the quality of service provided to patients. Disease forecasts and in the broader context, medical informatics, have recently received much attention in the field of data science research in recent years. This is primarily helping to the students like us to the wide adaptation and learning as well as working sector of computer-based technology into the health sector in different forms (e.g., electronic health records and administrative data) and subsequent availability of large health databases for researchers.

1.3 Literature Survey

We have undergone into the deep study of software solution of smart health prediction using Machine Learning techniques . This project aims in developing a computerized system to check and maintain your health by knowing the symptoms. It has a symptom checker module which actually defines our body related problem and gives us liability to select the affected area and checkout the symptoms. Technologies implemented in this paper are: The front end is designed with the help of HTML, Bootstrap and CSS. The back end is designed using Flask. It also comprises of different classifiers techniques which resides in Machine Learning which is used to to improve the prediction performance. This paper also contains the information of testing which is done at server side or we can say at the developer's end, this is an actual testing done with potential users or as an independent testing process at server end .

1.4 Motivation

This disease attacks a person so instantly that it hardly gets any time to get treated . In today's modern world disease is the most lethal one reason for the death of humans. So diagnosing patients correctly on timely basis is the most challenging task for the medical fraternity. A wrong diagnosis by the hospital leads to earn a bad name and loosing reputation. Almost all the hospitals use some hospital management system to manage healthcare in patients. At the same time treatment of the said disease is quite high and not affordable by most of the patients particularly in India. So, in this direction lots of efforts are required to make intelligent decisions in this sector .After reviewing many literatures related to this we were motivated to work on classfication model which will do smart prediction of the diseases .Hence we have defined the problem and determined its medical goals in this project .

1.5 Aim of the project

Now a days health industries are facing many problems related to devices or machines which will give unaccepted and wrong result due to less calibration of the instruments . So to avoid those drawbacks and get the desired and correct results we are building a project which will give the accurate prediction based on information provided by the user and also by dataset that are available in that machine . Furthermore the problem is that many people goes to hospitals or clinic to know how is their health and how much they are improving in those sick days but for this they have to travel to get to know their answers and sometimes it may happens that doctor wont be available or he/she may be on leave and many more reasons will be there so to avoid all these confusion and reasons we are making this project which will help all those person's and all the patients who are in need to know the condition of their health. This project will help to the various people's ranging from children to teenagers to adults and also the senior citizens .

1.6 Scope and Objectives

The purpose of this system is to provide prediction for the general and more commonly occurring disease in daily life and when such diseases are unchecked that can turn into fatal disease. This system will predict the most possible disease based on the given symptoms to avoid the aggression of diseases. The analysis accuracy is reduced when the quality of medical data in incomplete. The proposed system will consider both such problems. The analysis accuracy is increased by using Machine Learning algorithm .

1.7 Technical Approach

Machine Learning (ML) delivers methodologies, approaches, and apparatuses that can help resolving analytic and predictive hitches in a miscellany of medicinal areas. Machine Learning learns the data and produces the result. Decision Tree Machine Learning Algorithm predicts Diseases as well as all sub diseases. This algorithm is implemented to increase operational efficiency. It reduces Query retrieval time. Accuracy is improved using Machine Learning algorithm. This system is used by end users. In this we describe dataset which is being use to train the machine learning model. The dataset will contain symptoms of various diseases. The hospital data or symptoms of the patients will be in the form of structural format. The dataset used in this project is real life data. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e. predict disease.

CHAPTER 2

Theoretical Description of Project

2.1 Theoretical background

There is a need to study and make a system which will make it easy for an end users to predict the communicable and non-communicable diseases without visiting physician or doctor for diagnosis. To detect the Various Diseases through the examining Symptoms of patient's using different techniques of Machine Learning Models. The Predictions Accuracy will Increase using Machine Learning. We are applying machine learning to maintained complete solution through 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 good decision for patient diagnoses and treatment options, which leads to improvement of patient healthcare services. Healthcare is the most prime example of how machine learning is use in medical field. When the quality of medical data is incomplete the exactness of study is reduced. Moreover, different regions exhibit unique appearances of certain regional diseases, which may results in weakening the prediction of disease outbreaks. In the proposed system, it provides machine learning algorithms for effective prediction of various disease occurrences in disease-frequent societies. To the best of knowledge in the area of medical analytics none of the existing work focused on easiness and reliable patient's services at their end to check their diseases from which they are suffering through a long time . Hence this project is short approach to solve these problems in healthcare field to assess the risk of disease.

2.2 Technical specification of the project, resources required

2.2.1 Technical specification

Existing method fails to predict all possible diseases conditions of the people. Existing system handles only structured data. The prediction system are ambiguous and broad .The standing organizations arrange a blend of machine learning algorithms which are judiciously exact in envisaging diseases. However the restraint with the prevailing systems are speckled. First, the prevailing systems are dearer only rich people could pay for to such calculation systems. And also, when it comes to folks, it becomes even higher .

Our application will be at affordable in cost. Decision Tree Machine Learning Algorithm and K – Nearest Neighbors predicts Diseases as well as all sub diseases which increases operational efficiency. It reduces Query retrieval time. Accuracy is improved using Machine Learning algorithm. Thus making our presentation broadly open by all at cheap cost. This disease prediction using machine learning is completely done with the help of machine learning and python programming with important libraries such as Pandas , Numpy , Sklearn and also using the dataset that is precisely checked by hospitals using that we will predict the disease . So just by entering the symptoms into the system and all other useful information the user can get to know the disease he/she is suffering from and the health industry can also get the benefits from this system .

2.2.2 Resources required

- Hardware Requirements :

- System = Intel core i3 ,i5, i7 and 2 GHz minimum
- Ram = 512 mb or above
- Hard Disk = 10GB or above
- Input Device = Keyboard and Mouse
- Output Device = Monitor or PC

- Software Requirements :

- Operating System = Windows 7 ,10 , or higher versions
- Platform = Jupyter Notebook
- Programming Language = Python
- Front End = HTML , CSS , Bootstrap
- Back End = Flask , Python 3.8
- Database = Sqlite3
- Libraries = Pandas , Numpy , Sklearn

2.3 Block diagram

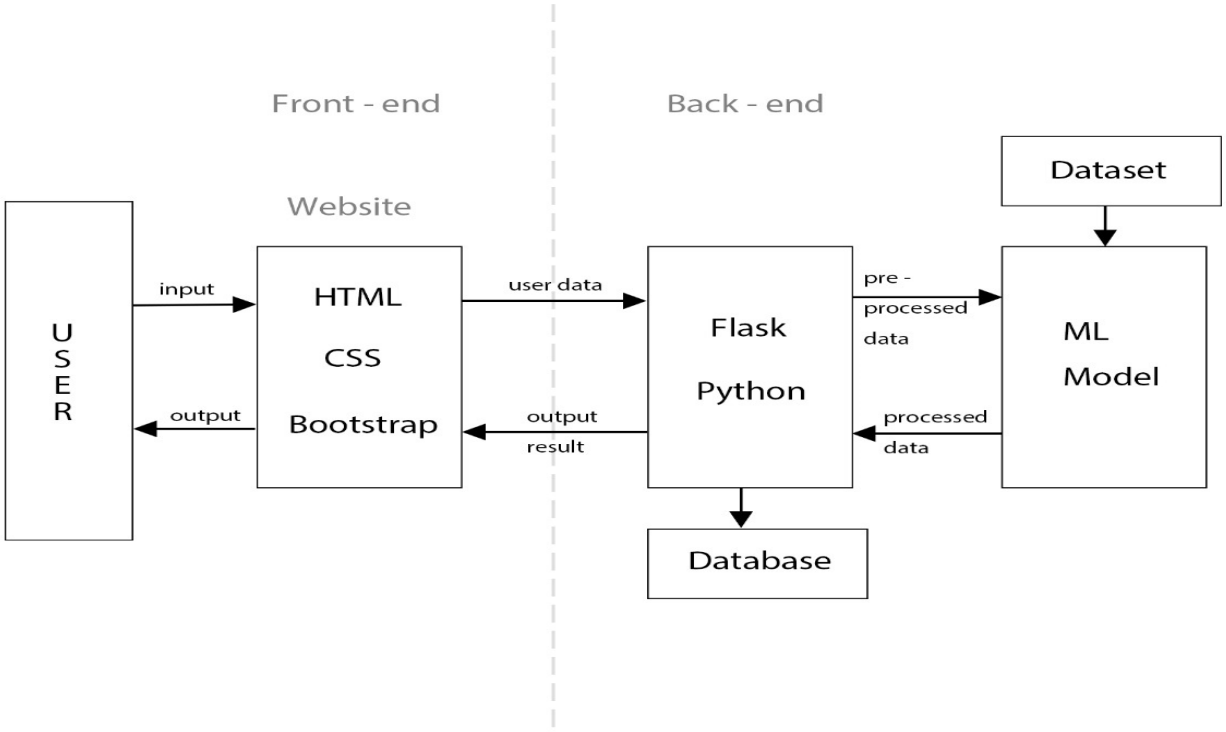

Fig 1.1 : Block Diagram

- User

Here the user has to enter the symptoms of the disease that they are facing, in order to use the prediction .Once the user enters the appropriate symptoms then system starts predicting the disease through its control flow by machine learning model which resides in it.

- Frontend of Website :
- HTML

HTML Hypertext Markup Language is used to create Web pages and tells the browser how to display them. It designs the basic layout and formatting of Web pages.HTML is made up of elements or tags and attributes which work together to identify document parts and tell the browser how to display them. An HTML [Hypertext Markup Language] elements are a set of tags and attributes that define different parts of web documents and inform web browsers how to display them. An HTML document is a text file that contains the tag-based information to publish. It also contains embedded instructions called “Elements”. For generally understood, the position of an element is indicated as spanning from a start tag, possibly including some child content, and is terminated by an end tag. An HTML element usually consists of a start tag and an end tag, with the content inserted in between start tag and end tag, syntax -

<start tag>Content Goes Here</end tag>

- CSS

CSS is the language for describing the presentation of Web pages, including colors, layout, and fonts. It allows one to adapt the presentation to different types of devices, such as large screens, small screens, or printers. CSS is independent of HTML and can be used with any XML-based markup language. The purpose of CSS is to provide Web developers with a standard way to define, apply, and manage sets of style characteristics. The "cascading" in Cascading Style Sheets refers to how property values are applied in the context of the

parent/child hierarchy of the Web document. Child elements either inherit or override property values bound to their parent elements. A style sheet is the encapsulation of style rules in a centralized location, either in the head section of the HTML document or in a separate linked file. The Web browser reads these styles and applies the specified formatting rules before displaying the content.

- Bootstrap

Bootstrap is a web framework that focuses on simplifying the development of informative web pages (as opposed to web apps). The primary purpose of adding it to a web project is to apply Bootstrap's choices of color, size, font and layout to that project. Bootstrap is a framework to help you design websites faster and easier. It includes HTML and CSS based design templates for typography, forms, buttons, tables, navigation, modals, image carousels, etc. It also gives you support for JavaScript plugins. Bootstrap's responsive CSS adjusts to phones, tablets, and desktops . Bootstrap is compatible with all modern browsers (Chrome, Firefox, Internet Explorer, Safari, and Opera). is a powerful toolkit - a collection of HTML, CSS, and JavaScript tools for creating and building web pages and web applications. It is a free and open source project .

- Flask

Flask is a web framework. This means that flask provides you with the tools, libraries and technologies that allow you to build a web system. This web application can be other web pages, blog, wiki or more or more like a web-based calendar or commercial website. Flask is a web framework, a Python module that allows you to easily develop web applications. Flask is based on WSGI (Web Server Gateway Interface) tools and Jinja2 template engine.

Flask offers a variety of developer options when developing web applications, providing you with tools, libraries, and equipment that allow you to create a web application but will not force any reliance or tell you what the project should look like.

- Python 3.8

The Python 3.8 series is the newest major release of the Python programming language, and it contains many new features and optimizations. Python is an interpreted , object-oriented, high-level programming language with dynamic semantics. Python's simple, easy to learn syntax emphasizes readability and it reduces the cost of maintenance .Python supports modules and packages, which encourages program modularity and code reuse. It is used in web development, data science, creating software prototypes, and so on. Python can be used to build server-side web applications

- Machine Learning Model
- Decision Tree

Decision trees are constructed via an algorithmic approach that identifies ways to split a data set based on different conditions. It is one of the most widely used and practical methods for supervised learning. Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. Algorithm for constructing decision tree usually works top-down, by choosing a variable at each step that best splits the set of items. Decision tree is one of the predictive modelling approaches used in machine learning.A decision tree is a flowchart-like structure in which each internal node represents a test on a feature (e.g. whether a coin flip comes up heads or tails) , each leaf node represents a class label (decision taken after computing all features) and branches represent conjunctions of features that lead to those class labels. The paths from root to leaf represent classification rules .

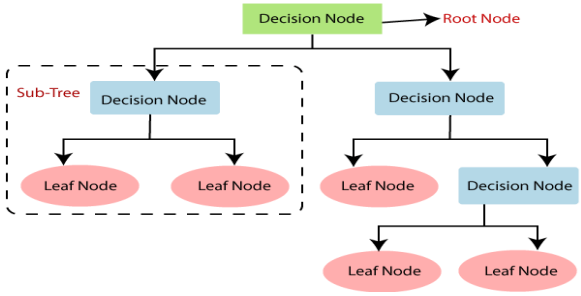

Fig 1.2: Decision Tree Flowdiagram

- KNN

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The K-NN algorithm matches the similarity between a new case / data and available cases and puts a new case in the same category as the available categories. The K-NN algorithm stores all available data and separates the new data point accordingly. This means that where new information appears then it can be easily separated into a suite section well using the K-NN algorithm.

can be used for reconstruction and segregation but is mainly used for segregation issues. K-NN is an a non-parametric algorithm, which means it makes no sense in data thinking. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

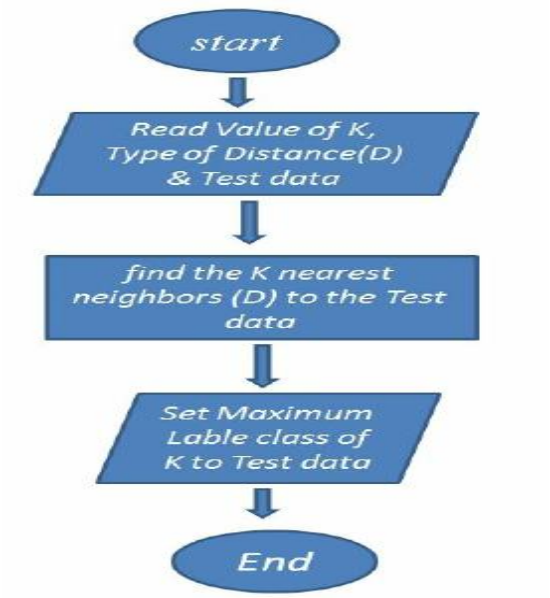

Fig 1.3: KNN flowdiagram

2.4 Flowchart

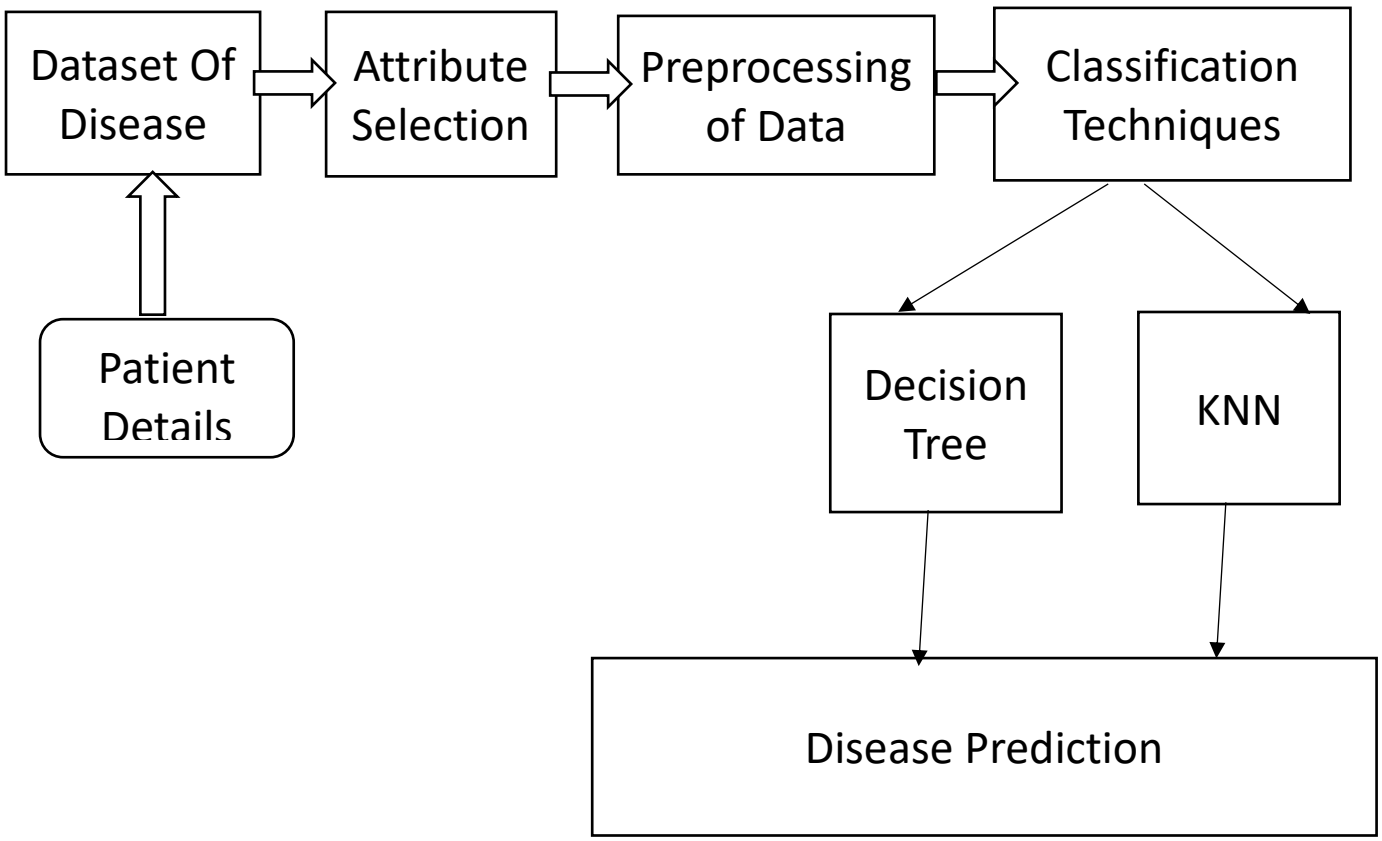

Fig 1.4: Flowchart

2.4 Algorithm

1] Once user open the system user needs to provide the symptoms which he/she is going through based on which we have several algorithms which predict the disease .

2] The user needs to enter at least two symptoms and maximum five symptoms to get the accurate result .

3] Data collection and dataset preparation this will involve collection of medical information from various sources like hospitals , then pre-processing is applied on dataset which will remove all the unnecessary data and extract the essential features from data .

4] Machine learning model consisting of Decision tree and KNN algorithm is to be developed, it will run effectively on extensive database of healthcare and also it can deal with a huge number of information variables without variable deletion .

5] In this project two different algorithms are used :

- Decision Tree Algorithm
- K-Nearest Neighbors Algorithm

6] Deployment and analysis on real life scenario will be done hence the trained and tested prediction model will be deployed in a real life scenario and will be leveraged further .

CHAPTER 3

System Design

3.1 Block wise design

The system design of the project " Disease prediction using Machine learning " consist of all the various aspects a normal application requires . This snapshots of the system design shows how from starting the model flows from one step to another , like user enters into the system then enters all the information's and all other general information along with symptoms that goes into the system compares with the prdiction model and if it is true then predicts the appropriate result otherwise it shows the details where the user went wrong while entering the information .

The control flow and the operation of the system is described as follows through snapshots .

1] Main Page of the interface :

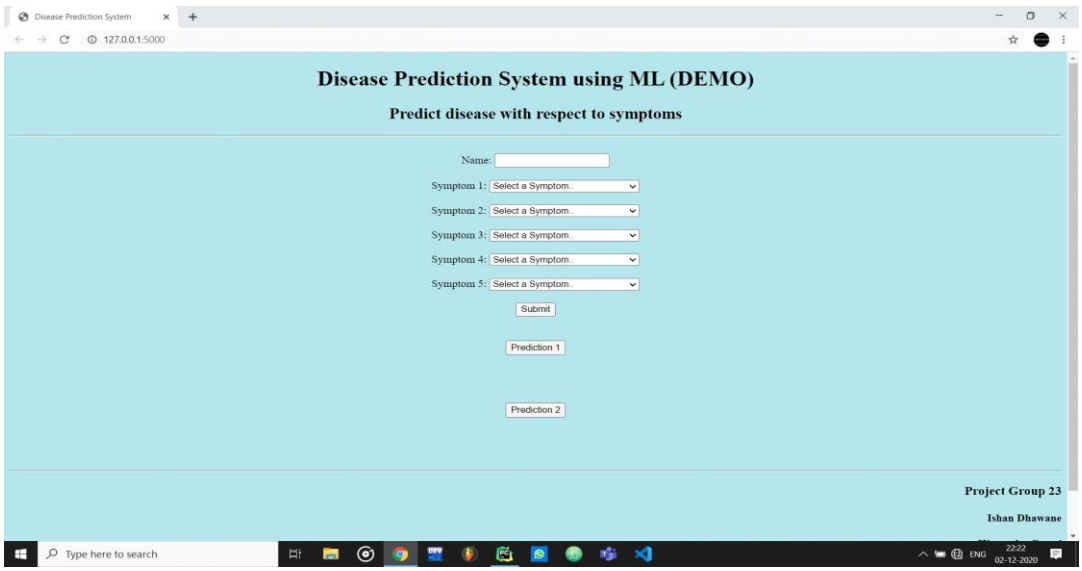

Fig 1.5 : Main Page

2] Enter the appropriate symptoms :

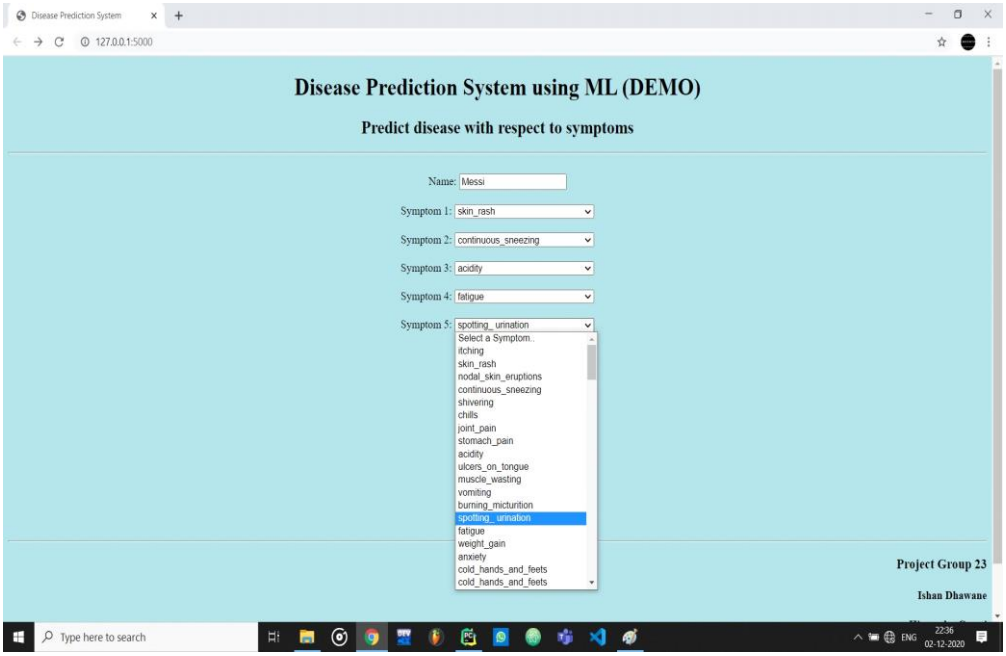

Fig 1.6 : Selection of Symptoms

3] Invalid selection of symptoms :

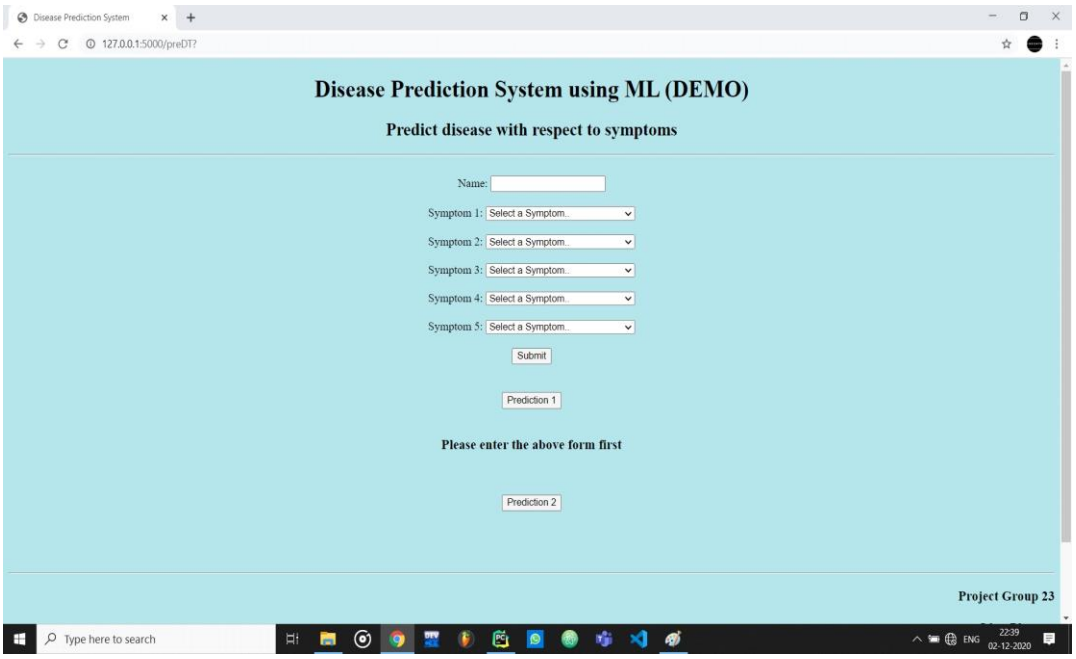

Fig 1.7 : Warning if data not entered correctly

4] Warning given through required field if the data is not entered properly :

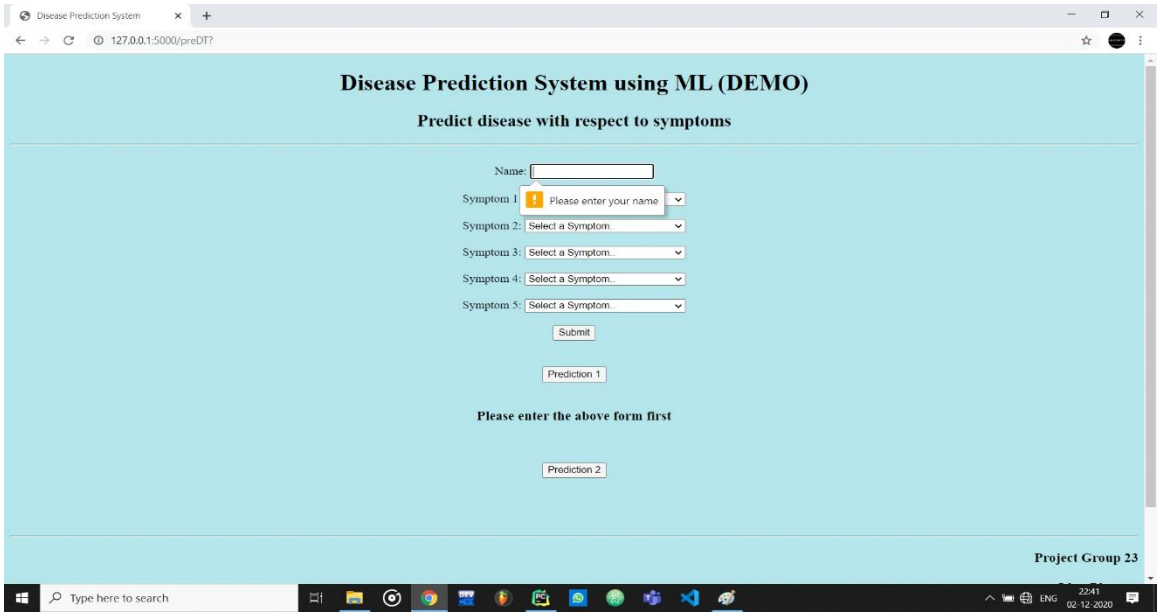

Fig 1.8 : Required field warning

5] Warning given through required field if the data is not entered sufficiently upto 2 symptoms limit :

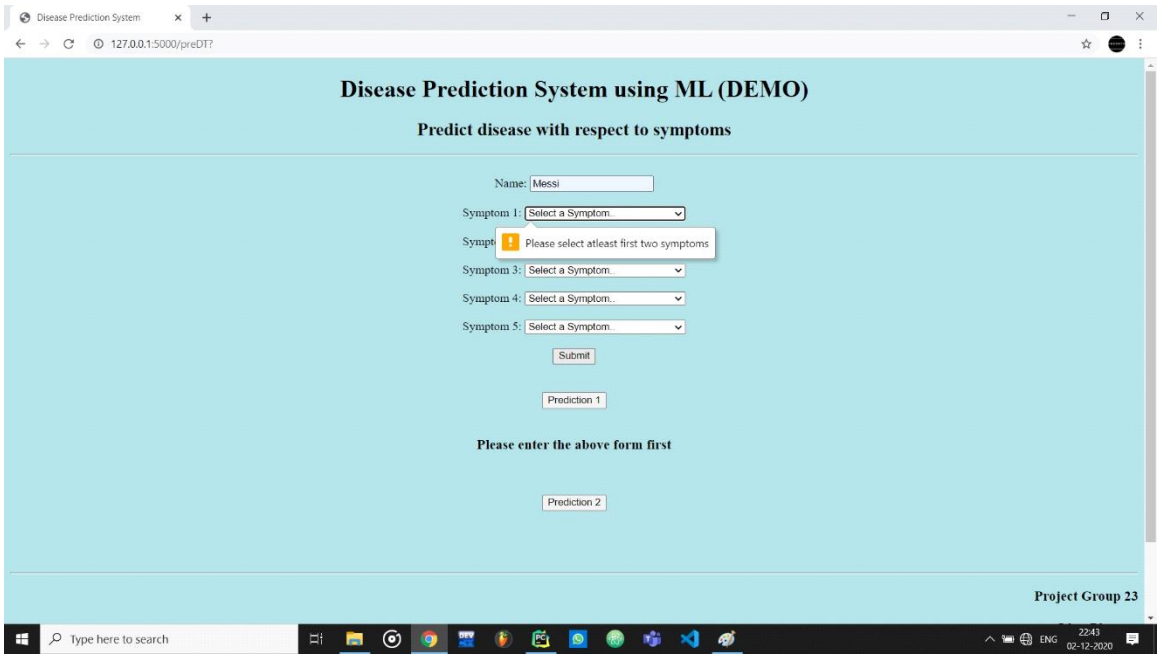

Fig 1.9 : Requires at least 2 symptoms

6] Data successfully submitted :

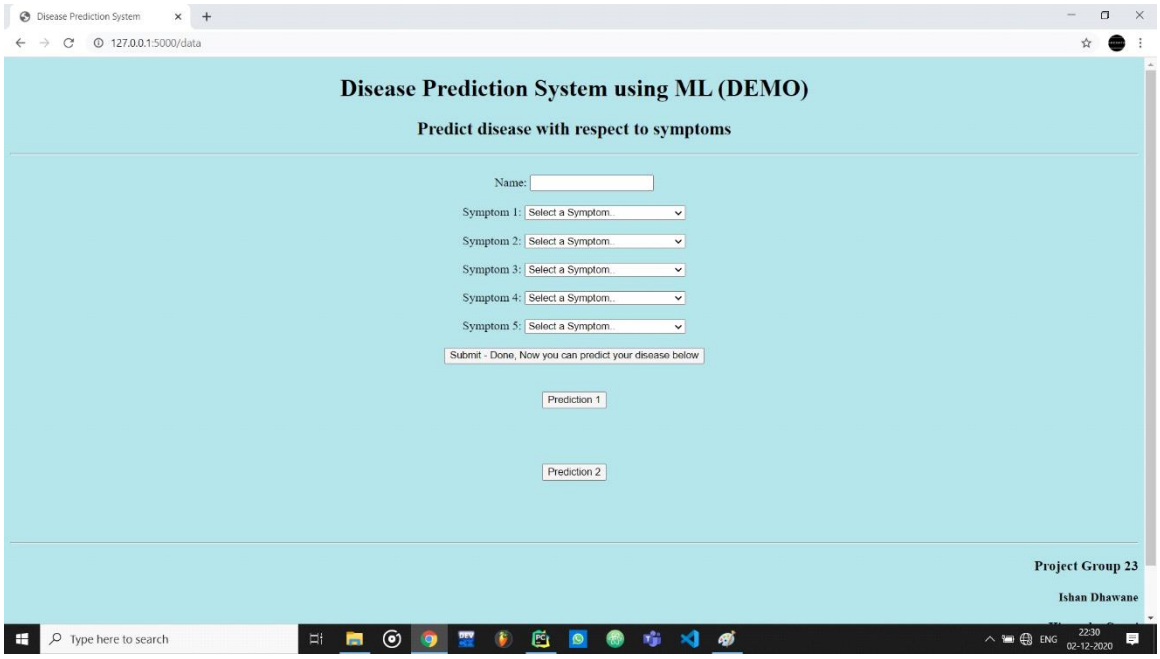

Fig 2.1 : Data entered succesfully

7] Predicted Result Stage 1 :

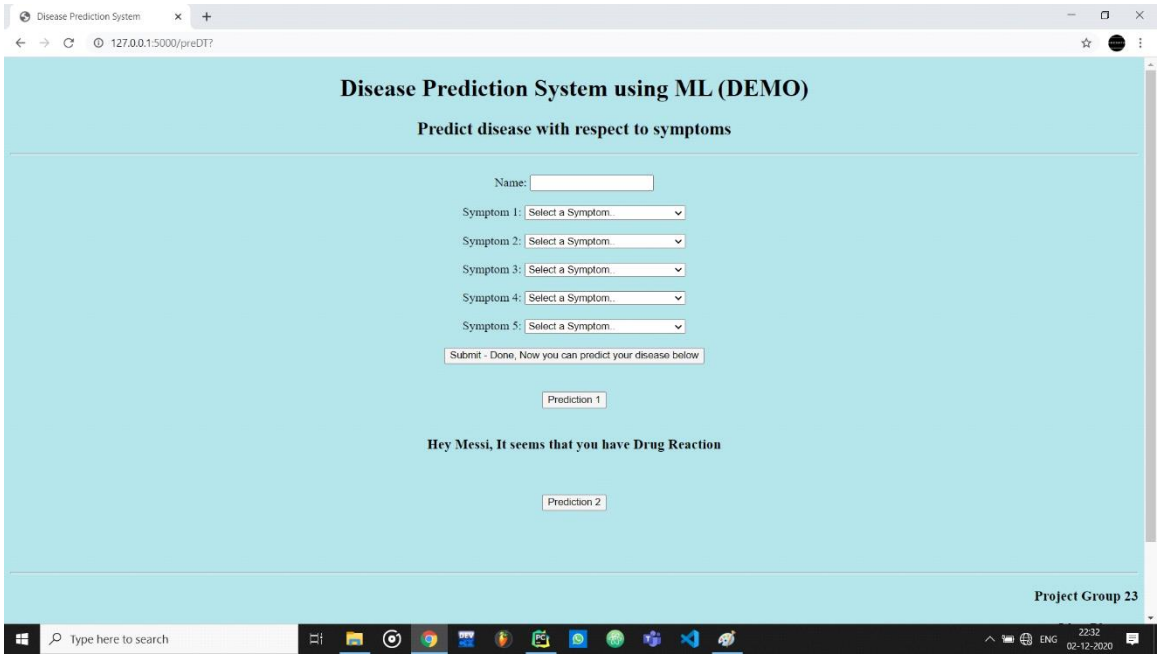

Fig 2.2 : Prediction using Decision Tree model

8] Predicted Result Stage 2 :

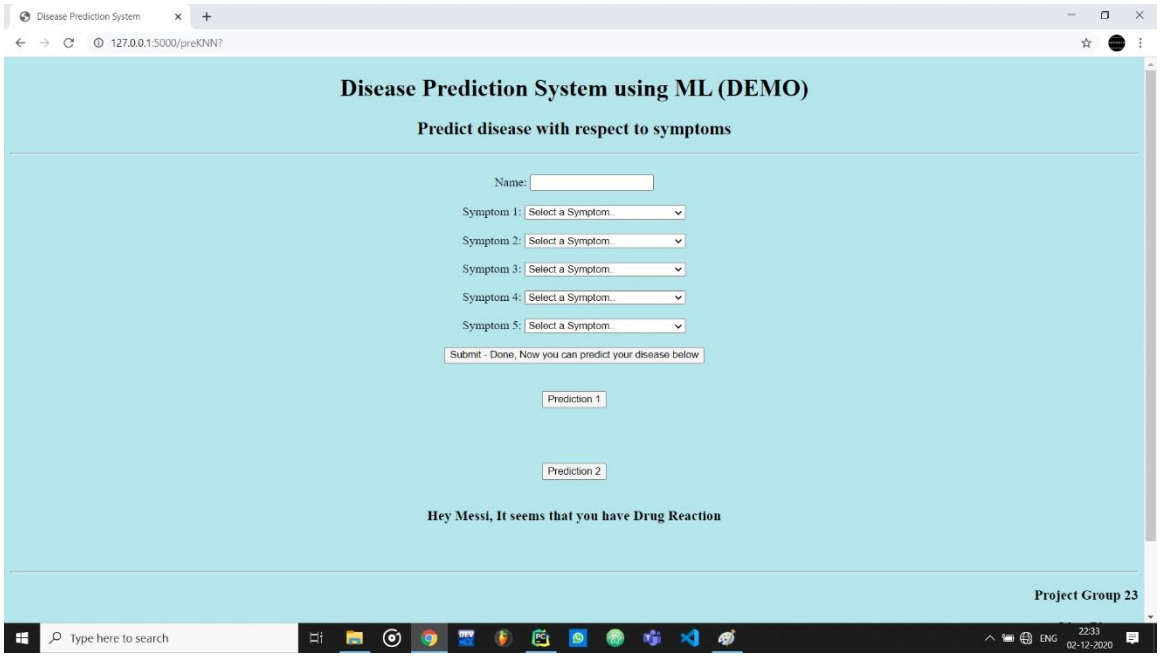

Fig 2.3 : Prediction using KNN model

9] Testing Dataset added into the project :

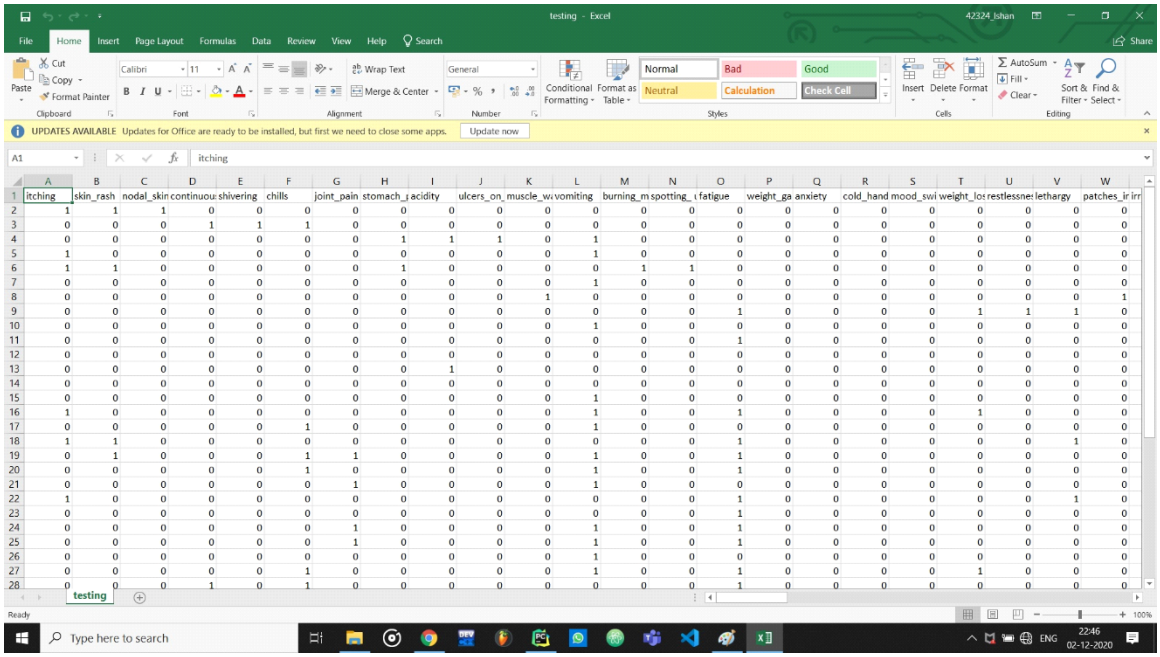

Fig 2.4 : Testing data

10] Training Dataset added into the project :

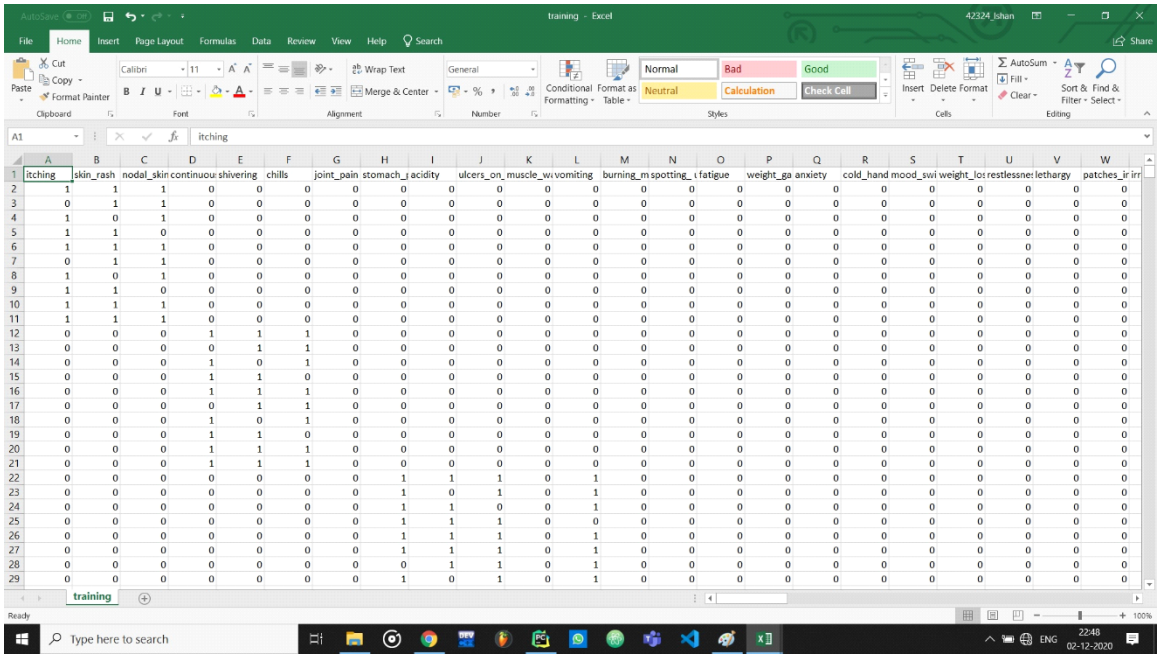

Fig 2.5: Training data

11] Test Case for Prediction Result :

Serial Number of Test Case	TC 03
Module Under Test	Prediction Result
Description	User needs to enter the name and symptoms to get the prediction result.
Input	Name and Symptoms
Output	If user enters all 5 correct symptoms then the accuracy will be high. If user enters only few symptoms then accuracy will be low.
Remarks	Test Successful.

Table 1.1 : Test Case

CHAPTER 4

Progress of Project and Discussion

In order to ensure effective implementation of our Project "Disease prediction system using Machine Learning " we have tried to accomplish the following things indivisually with desired task domain .

It includes :

1] Front End Development :

We have learned web development through HTML , CSS , Bootstrap for creation of Effective user friendly website .

2] Back End Development :

It is carried out by Flask framework and Python 3.8 .

3] Real-time prediction of the disease :

This has been done through Machine Learning algorithm and by using dataset comprises of various symptoms data .

- Succesfully Completion of Following things in this Project :

1] Accurate prediction of the diseases

2] Created Website for driving this "disease prediction System" and hence user will enter all the data through this website interface .

3] Machine learning model is ready with Decision tree and KNN algorithm by checking and testing of it .

- Additional Thing yet to do :

1] Need to add database through which user can store their login details inside the Disese prediction system .

CHAPTER 5

Conclusions

In this paper, we have presented a system which is suitable for real-time diseases prediction and can be used by the users who have communicable or non-communicable disease. The diagnosis system of the system is able to predict the any disease by using Machine Learning algorithms and the prediction results are based on the disease dataset instance. On the other hand, the system is very inexpensive. To prove the effectiveness of the system we have carried out experiments for diagnosis system . We ran experiments with some popular algorithms like KNN, Decision Tree. The experiment was carried out with the holdout test .Such a system can decrease the rush at OPDs of hospitals and reduce the workload on medical staff. Such a system can be largely reliable to do the job creating this system we also added a way to store the data entered by the user in the database which can be used in future to help in creating better version of such system. Our system also has an easy to use interface. It also has various visual representation of data collected and results.

CHAPTER 6

Future Scope

1] To propose Community question answering system

(CQA) which is used to help user to post the questions and get back the answers related to the disease.

2] To append precautionary measures with respective Disease .

3] More interactive user interface

4] It can be done as Mobile Application

5] Update more latest Global Diseases .

6] To depict latest disease outbreaks on webpage

References

[1] "P. B. Jensen, L. J. Jensen and S. Brunak. Mining electronic health records: Towards better research applications and clinical care".

[2] "Dmitry Ignatov and Andrey Ignatov. Decision Stream: Cultivating Deep Decision Trees", 3 Sep 2017 IEEE".

[3] "Disease Prediction Using Machine Learning Over Big Data"Vinitha S. Sweetlin S. Vinusha H and Sajini S (2018).

[4] "Multi Disease Prediction Using Data Mining Techniques". Gomathi, Dr. D. Shanmuga Priyaa (2017).

[5] Seema sharma, Jitendra Agarwal, Shikha Agarwal, Sanjeev Sharma, "Machine Learning Techniques for Data Mining:A Survey, in Computational Intelligence and Computing Research", IEEE International Conference on IEEE, 2013, pp.1-6.

[6] N. Nori, H. Kashima, K. Yamashita, H. Ikai, and Y. Imanaka, "Simultaneous modeling of multiple diseases for mortality prediction in acute hospital care", in Proceedings of the 21th ACM, 2015.

[7] Algorithms Details from www.dataspirant.com

Appendix

- Bill of material :

This system of disease prediction system is completely software project and it did not cost anything. It is totally inexpensive .

- Testing Photos :

1] Disease Prediction done by using Decision Tree algorithm :

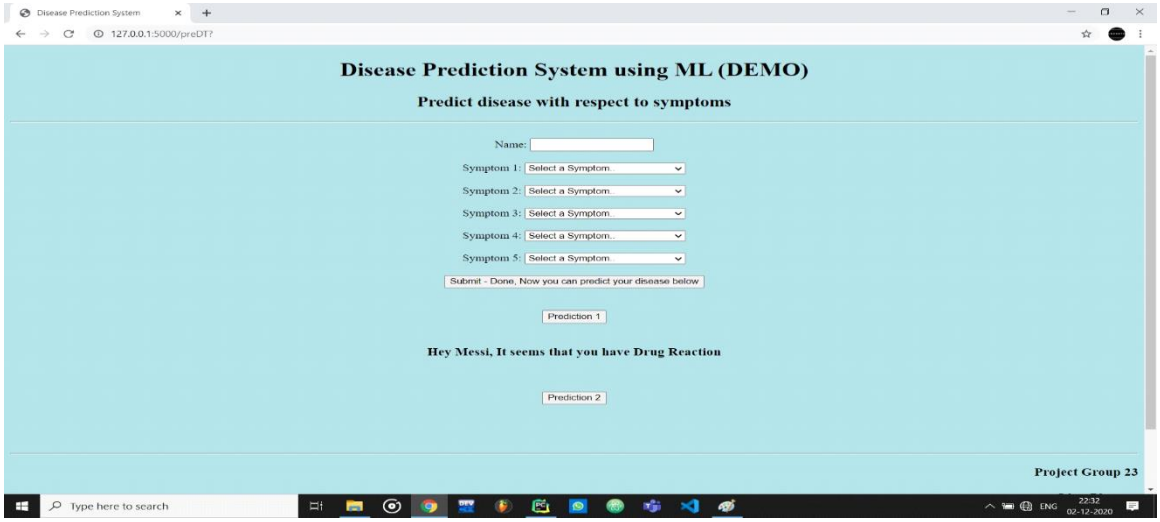

Fig 2.6 : Testing with Decision Tree Algorithm

2] Disease Prediction done by using KNN algorithm :

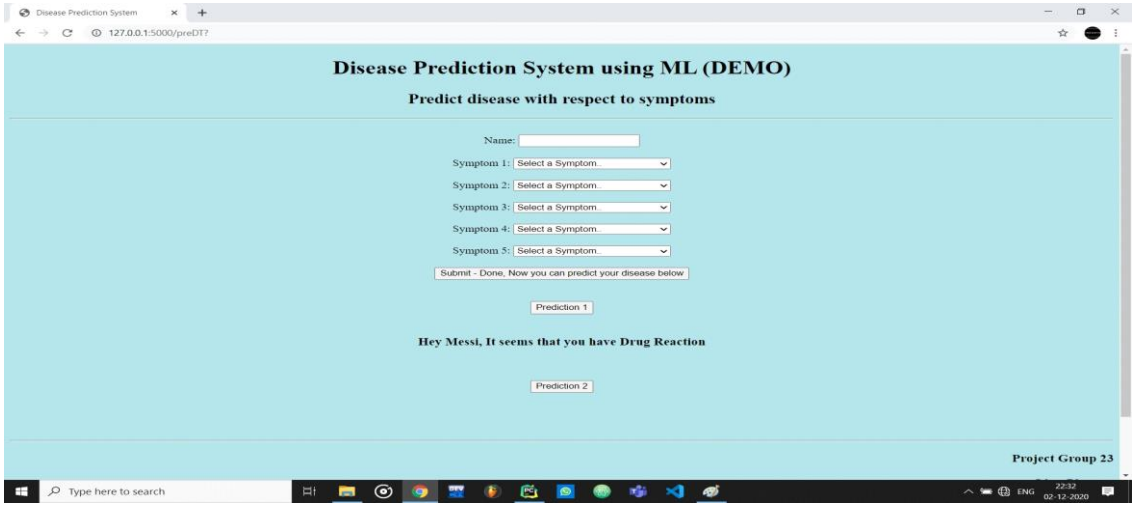

Fig 2.7 : Testing with KNN Algorithm

Output

Here we are succesfully getting the same result through both the approaches of Machine Learning algorithm. In both the testing photos user is getting the same disease incurred with repective of both different algorithms (Decision Tree, KNN) prediction .

Datasheets

We have used two datasets here . They are as follows

1] Testing Dataset

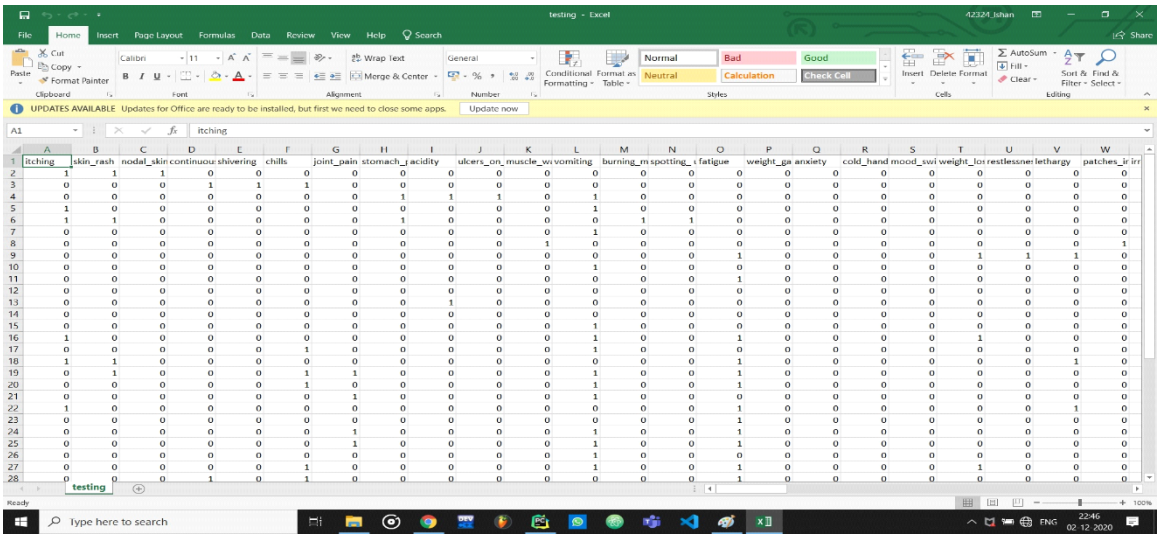

Fig 2.8 : Testing Dataset

2] Training Dataset

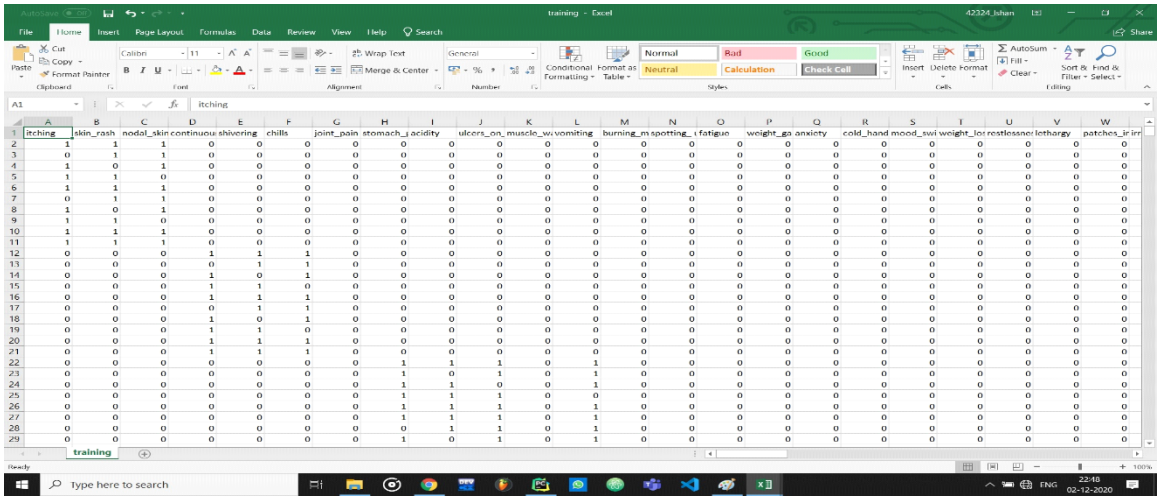

Fig 2.9 : Training Dataset