Mini Project Report on

DISEASE PREDICTION SYSTEM USING MACHINE LEARNING

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING (AI & ML)

Submitted by:

Ashu Ishan **University Roll No.-2019468**

Under the Mentorship of

Dr. Sharon Christa
Assistant Professor, Department of Computer Science

Department of Computer Science and Engineering
Graphic Era (Deemed to be University)
Dehradun, Uttarakhand
January 2023

Graphic Era Deemed to be University DEHRADUN

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled **"Disease Prediction System Using Machine Learning"** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Sharon Christa, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name- Ashu Ishan University Roll no.- 2019468

signature

Table of Contents

Chapter 1

INTRODUCTION

Abstract

The Disease Prediction System based on various prediction models that help to predict the disease of the user on the basis of the symptoms that user enters as an input to the system. Predictive models with the help of machine learning classification algorithms analyzes the symptoms provided by the user as input and gives the name and probability of the disease as an output. Disease Prediction is done by implementing the Naive Bayes Classifier, Decision tree and Random Forest Algorithm. The Naive Bayes helps to calculate the probability of the dis-ease which is predicted. Average prediction accuracy probability 87% is obtained. The model uses a dataset with the count of 132 symptoms from which the user can select their symptoms. The user does not need to have a medical report to use this system as the prediction is based on the symptoms which will save the money. The system also has a very easy to use user interface so all the users can use it to predict the generic diseases

1.1 About

There are times when we need a doctor all of a sudden but sometimes they are not available due to some reason and we are left in trouble. The system we have proposed is user friendly to get help and advice on health issues immediately through the online healthcare system. Now adays, with the help of the statistics and posterior distribution the problems are swiftly and easily. As the Bayesian statistics has a great success rate in the field of economic, social science and a few other fields just like that, in medical fields, people have solved various medical problems that are tiresome to be settled in classic statistics by classification and can be solved easily. Naive Bayes is among the basic common classification techniques introduced by Reverend Thomas Bayes. The classification rules which help in solving the prediction of disease are generated by the samples trained by themselves and help in solving the problem easily. It is approximated that greater than 70% of people in India are prone to various body dis-eases like viral, flu, cough, cold etc. in intervals of 2 months. As many people don’t understand that the general body diseases could

be symptoms of something more harmful, 25% of this population dies or gets some serious medical problem because of ignoring the early general body symptoms and this is a very serious condition that we are facing and the problem can be proven to be a very dangerous situation for the population and can be alarming if the people will continue ignoring these diseases. Hence identifying or predicting the disease at the very basic stage is very important to avoid any unwanted problems and deaths. The systems which are available now a days are the systems that are either dedicated to a particular disease or are in development or the research for solving the algorithms related to the problem when it comes to generalized disease. The main motive of the proposed system is the prediction of the commonly occurring dis-eases in the early phase as when they are not checked or examined they can turn into a disease more dangerous disease and can even cause death. The system applies data mining techniques, decision tree algorithms, Naive Bayes algorithm and Random Forest algorithm. This system will predict the most possible disease based on the given symptoms by the user and precautionary measures required to avoid the aggression of disease, it will also help doctors to analyze the patterns of diseases in the society. This project is dedicated to the Disease prediction System that will have data mining techniques for the basic stages of the dataset and the main model will be trained using the Machine Learning (ML) algorithms and will help in the prediction of general diseases.

1.2 Data Mining and Machine Learning Algorithm

The Data Mining and the Machine Learning Algorithms are used for the prediction of Disease in the Project. There are different Data Mining and Machine Learning used for the purpose of correcting and evaluating the dataset and then testing the dataset on the basis of train score and the test score of the ML model.

1.2.1 Data Analysis and Data Mining

The Data Mining is a process in which raw data is prepared and structured from the unstructured data as to take meaningful information from the data which can be used in the project. Task of making data organized and reflective about data is to way to get what this information does the data contains in it and what it does not have in it. There are so many different types of methods in which the people can make use of data analysis. It is simply very easy to use data during the analysis phase and get to some certain conclusions or some agendas. The analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the objective of highlighting useful information, suggesting conclusions, and supporting decision

making which are helpful to the user. Data analysis has multiple facets and approaches, encompassing diverse techniques under an array of names, in different business, science, and social science domains. Data Mining is the discovery of unknown information found in databases, data mining functions has some different methods for clustering, classification, prediction, and associations. In the data mining important application is that of mining association rules, association rules was first introduced in 1993 and are used to identify relationships among a set of items in databases these different properties are not based on the properties of the data, but rather based on co-occurrence of the data items. The Data mining helps in giving new and different perspectives for data analysis the main role of data mining is to extract and discover new knowledge from data. In the past few years, different methods have been coined and developed about the capabilities of data collection and data generation, data collection tools have provided us with a huge amount of data, data mining processes have integrated techniques from multiple disciplines such as, statistics, machine learning, database technology, pattern recognition, neural networks, information retrieval and spatial data analysis. The data mining techniques have been used in many different fields such as, business management, science, engineering, banking, data management, administration, and many other applications.

1.2.2 Machine Learning Algorithms

The ML is a small part of Artificial Intelligence (AI) which is used in the computation work and the analysis work in the AI. The ML algorithms are used to find different patters and different structures in the dataset which is provided to the dataset, the ML algorithms are used to give a large computation capabilities to the system by which a large amount of data is given to the model for the purpose of training and testing the data, the ML algorithms are used in decision making process the model which is prepared by using the ML has a large amount of data in it which makes it a very good for the process of decision making. ML algorithms have very high computational power and are proven to be very helpful in today's world. Different types of ML algorithms are organized into different ways, based on the desired outcome of the algorithm. Common algorithm types include

- Supervised learning — The supervised learning algorithm can apply what has been learned in the past to new data using labelled examples to predict the future events. Starting from analysis of a known training dataset. This algorithm is used to provide targetsfor any new values after sufficient amount of training of the model.

- Unsupervised learning — Unsupervised machine learning algorithms are used when theinformation used to train is neither classified nor labeled. This algorithm shows how thesystem can infer a function to describe a hidden structure from unlabeled data.

- Semi-supervised learning — This category of the ML algorithms falls somewhere be-tween the supervised learning and the unsupervised learning algorithm which combinesboth labeled and unlabeled examples to generate an appropriate function or classifierwhich is used to make a model for the purpose of prediction or classification.

- Reinforcement learning — This is the algorithm where the algorithm learns a policyof how to act given an observation of the world. Every action has some impact in theenvironment, and the environment provides feedback that guides the learning algorithm.

- Transduction — This algorithm is similar to supervised learning, but does not explic-itly construct a function: instead, tries to predict new outputs based on training inputs,training outputs, and new inputs

Chapter 2

Literature Survey

2.1 Introduction to Old Models

• In the model proposed by [1] showed important ML approaches to predict the disease but this model which was proposed by [1] works on the K-Nearest Neighbour (KNN)and Convolution Neural Network (CNN) approach of the machine learning algorithm. Both the KNN and CNN approaches are used in this system which is different from the approach which is used in our project. The CNN uses both the structures as well as the unstructured data for the prediction of the disease which makes it more time consuming. The accuracy of the system proposed by [1] comes out to be very high i.e. above 95% for the KNN algorithm and 100% for the CNN algorithm that is very high for a ML model, In such cases the model is said to be overfitting.

Figure 2.1: Accuracy Of The Model

• The model proposed by [2] is used for Disease Prediction and uses different ML algorithms like I forest for correcting the dataset problems and SMOTET for balancing the dataset and then it uses the Ensemble learning technique. The Input the ML model is taken only by the electronic reports which are produced by the blood examination of the patient or the user. Some of the input taken in this model

are glucose level, cholesterol, lipoprotein, blood pressure and other inputs which are only be possible by the physical examination the user or the patient.

2.2 Identification of Research Gap and Problems

• The model given by [1] uses KNN and CNN algorithms which is more time consuming gas it involves both the structured and the unstructured data so the time taken to process the data is more as compared to the dataset which contains only the structured data as in the proposed project which contains only the structured data and the classification algorithms used in the proposed project are decision tree, Naive Bayes and Random forest. The accuracy of the model given by [1] is above 90% which is not good for a ML model as it is said to be in an over fitting situation whereas the proposed model has accuracy of about 86% which is good enough for a model of disease prediction.• The model given by [2] has a very limited scope as it is only meant for the prediction of the diabetes and hypertension whereas the proposed model is used for the prediction of the basic general disease. The model given by [2] needs the blood report of the patient or the user for the prediction of the diabetes or the hypertension and the algorithms used in this model are ensemble learning techniques whereas the predicted model does not need any blood report or physical presence of the user or the patient. The system contains a list of symptoms from which the user can select the symptoms which the user is facing and can predict the disease very easily and the algorithms used are different from the given model. The input required in the given model are based on the medical report of the user like cholesterol, blood glucose etc whereas the proposed system does not require any type of blood report for the prediction of the disease.

• The model given by [3] uses a very big data set and to manage that dataset the big data analytics are used which makes this system slow as needs a lot of system requirements to run this project and the deep learning algorithms are used in this project are FISM,NAIS, Deep ICF which is different from the proposed model which uses the classification algorithms which are light weight for a PC and run faster as compared as compared to the big deep learning techniques and big data analytics which takes more time and space.

• The model given by [4] is best for the prediction of disease related to breast cancer, diabetes and heart related problems and has different dataset for all the three different kind of disease which is different from the proposed model as the proposed model helps in the prediction of the general diseases with the help of the symptoms and has

a single dataset for all the diseases. The accuracy in some case of the given model by [4] is very high.

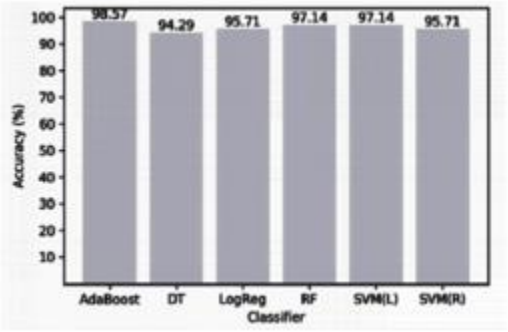
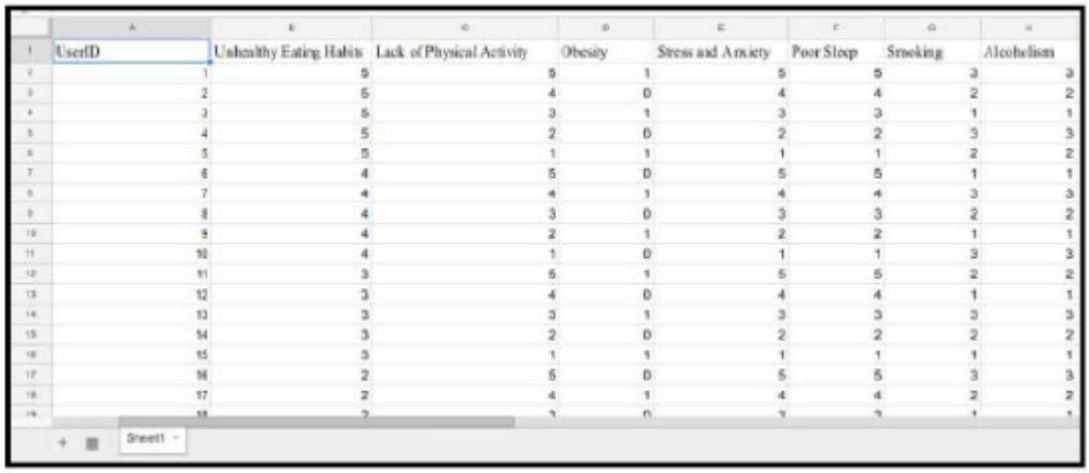


Figure 2.2: Accuracy of the breast cancer dataset

so the model given by [4] can be said to be a over fitting model as the accuracy is too high where as the accuracy in the proposed model is about 86% which is good for a model for the disease prediction.

• The model given by [5] uses KNN algorithm for the prediction of heart related diseases and uses parameters like high cholesterol, high blood sugar, diabetes, smoking habits, consuming too much alcohol as the input for the prediction of heart related diseases this model also gives information about the cardio vascular diseases and the cardiac arrest and many more heart related problems, the efficiency of the system is high for the decision tree algorithm i.e. about 91% whereas the proposed system has the capability to predict the general diseases and is more helpful as compared to a simple heart disease prediction system which is only helpful for the heart diseases but the simple disease prediction system is helpful for the prediction of more diseases.

• The model given by [6] uses SVM algorithm for the prediction. This model is used to predict the lifestyle and weather a person is suffering from any disease or not. The input in the model is given as per the rating i.e. from 1-5 where 1 is for excellent and 5 is for very bad and the symptoms which are rated are lack of physical activity, obesity, stress and activity, smoking etc.



UserID	Unhealthy Eating Habits	Lack of Physical Activity	Obesity	Stress and Anxiety	Poor Sleep	Smoking	Alcoholism
1	5	5	1	5	5	3	3
2	5	4	0	4	4	2	2
3	5	3	1	3	3	1	1
4	5	2	0	2	2	3	3
5	5	1	1	1	1	2	2
6	4	5	0	5	5	1	1
7	4	4	1	4	4	3	3
8	4	3	0	3	3	2	2
9	4	2	1	2	2	1	1
10	4	1	0	1	1	3	3
11	3	5	1	5	5	2	2
12	3	4	0	4	4	1	1
13	3	3	1	3	3	3	3
14	3	2	0	2	2	2	2
15	3	1	1	1	1	1	1
16	2	5	0	5	5	3	3
17	2	4	1	4	4	2	2

Figure 2.3: Dataset for [6]

whereas the proposed model dataset has the data on the basis of 0 and 1 that weather a symptom is present or not and helps in the prediction of disease in a better way as the given model is only to predict the lifestyle of a person that he or she is physically active or not and many other things and what are the chances that a person is prone to a disease whereas the proposed system predicts the disease.

• The model given by [7] uses very big data set and uses big data analytics for the prediction of disorders. this model uses mahout of the had loop file system for the prediction as the mahout contains all the data mining and analysis techniques for the prediction of the disorders but as we can see that there is a huge amount of data associated with this model so it's a bit hard to process all the data for the predictions of the disorders and the overall speed of the system becomes a bit slow as there is a huge amount of data to be processed. This model helps in the prediction of the chronic disorders like thyroid and needs a medical examination of the user as well whereas the proposed system is fast as it has light weight algorithms and has higher efficiency and helps in the prediction of the more commonly occurring diseases which can later result in big problems later to the user or the patient. The system also helps in getting medical advice from the doctor as the doctors are also registered with the system which helps in better diagnosis of the disease and getting medical treatment

Chapter 3

Methodology

3.1 Major Inputs Required

The inputs required for the project are:

• Software Inputs:

–Jupyter Notebook

–Python version 3

–Pip version 3

–Pip virtual environment

–Django version 2

–Postgresql version 10

–Pgadmin version 3

• Hardware Inputs:

–Windows/Linux/Mac OS

–At least of 2 GB RAM

–At least 512 GB ROM

–At least a Integrated Graphic card

• User Inputs:

–Basic Details–Symptoms

3.2 Feasibility Analysis

3.2.1 Technical Feasibility

The project is technically feasible as it can be built using the existing available technologies. It is a web based applications that uses Django Framework. The technology required by Disease Predictor is available and hence it is technically feasible.

3.2.2 Economic Feasibility

The project is economically feasible as the cost of the project is involved only in the hosting of the project. As the data samples increases, which consume more time and processing power. In that case better processor might be needed.

3.2.3 Operational Feasibility

The project is operationally feasible as the user having basic knowledge about computer and Internet. Disease Predictor is based on client-server architecture where client is users and server is the machine where dataset and project are stored

3.3 System Design

The whole project can be divided into two parts i.e. The Machine Learning Model and The User Interface and they can be elaborated as

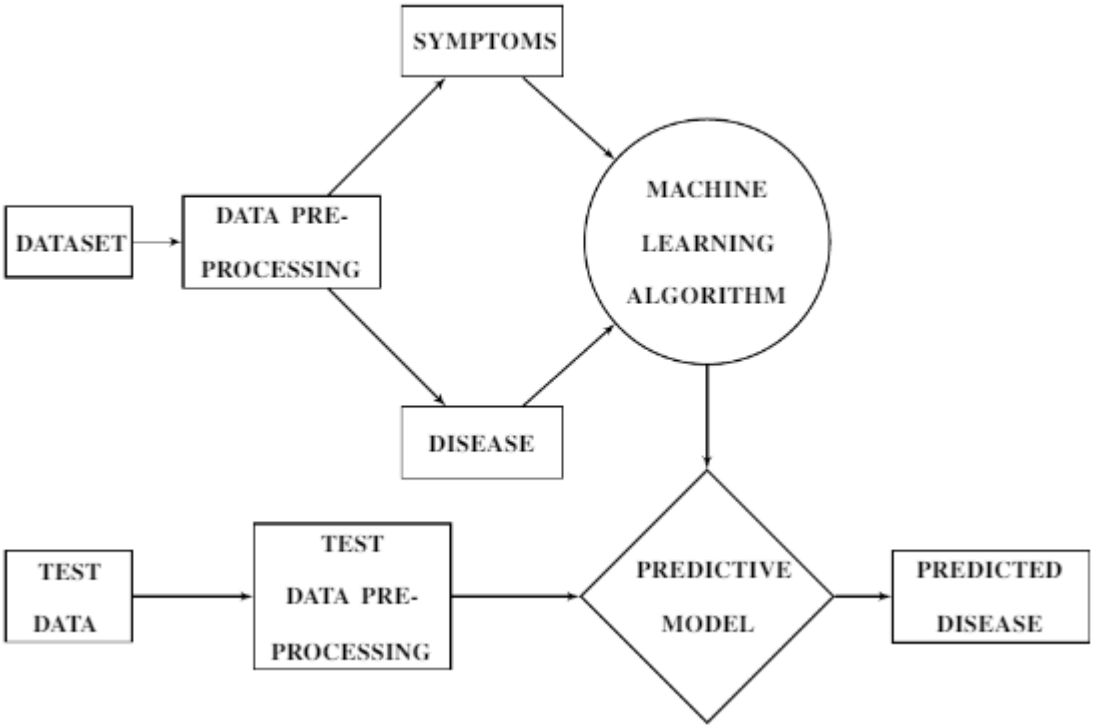


Figure 3.1: Detailed Design of Model

Data mining techniques are used in the project to see weather the dataset is good for prediction or not. Various data mining libraries used in the project are:

1. Scipy: This is used for implementing scientific computing in Python programming language. It is a collection of mathematical algorithms and convenience functions built on Numpy. Following are some of the functionalities it provides Special Functions (special), Integration (integrate), Optimization (optimize), Fourier Transforms (interpolate), Signal Processing (signal), Linear Algebra (linalg), Statistics (stats), File IO (io) etc. In this project stats (Statistics) library of this package is primarily used.

2. Sklearn : This stands for Scikit learn and is built on the Scipy package. It is the primary package being used in this project. It is used for providing interface for

supervised and unsupervised learning algorithms. Following groups of models are provided by sklearn Clustering, Cross Validation, Datasets, Dimensionality Reduction, Ensemble methods, Feature extraction, Feature selection, Parameter Tuning, Manifold Learning, Supervised Models.

3. Numpy : It is a library for the Python programming language, adding support for multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It provides functions for Array Objects, Routines, Constants, Universal Functions, Packaging etc. In this project it is used for performing multi-dimensional array operations.

4. Pandas : This library is used to provide high-performance, easy-to-use data structures and data analysis tools for the Python programming language. It provides functionalities like table manipulations, creating plots, calculate summary statistics, reshape tables, combine data from tables, handle times series data, manipulate textual data etc. In this project it is used for reading csv files, comparing null and alternate hypothesis etc.

5. Matplotlib : It is a library for creating static, animated, and interactive visualizations in Python programming language. In this project it is used for creating simple plots, sub-plots and its object is used alongside with the seaborn object to employ certain functions such as show, grid etc. A %matplotlib inline function is also used for providing more concise plots right below the cells that create that plot.

6. Seaborn : It provides a interface for making graphs that are more attractive and interactive in nature. It is based on the matplotlib module. These graphs can be dynamic and are much more informative and easier to interpret. It provides different presentation formats for data such as Relational, Categorical, Distribution, Regression, Multiples and style and color of all these types. In this project they are used for creating complex plots that use various attributes.

7. Warning : It is used for handling any warnings that may arise when the program is running. It is a subclass of Exception.8. Stats : This library is used to incorporate statistics functionality in Python programming language. This library is included in the scipy package. This library is not directly used rather the required functions are directly imported as and when required i.e. for Measures of Central Tendency, Measures of Variability . The functions used can be for simple concepts like mean, median, mode, variance, percentiles, skew, kurtosis, range, Cumulative25

Chapter 4

Result and Conclusion

• The outcome which is expected from the project is given with the help of machine leaning models that are used in the project to predict the disease on the basis of input provided by the user as a content of symptoms which are selected from a given list of symptoms provided to the user as in fig:4.1.

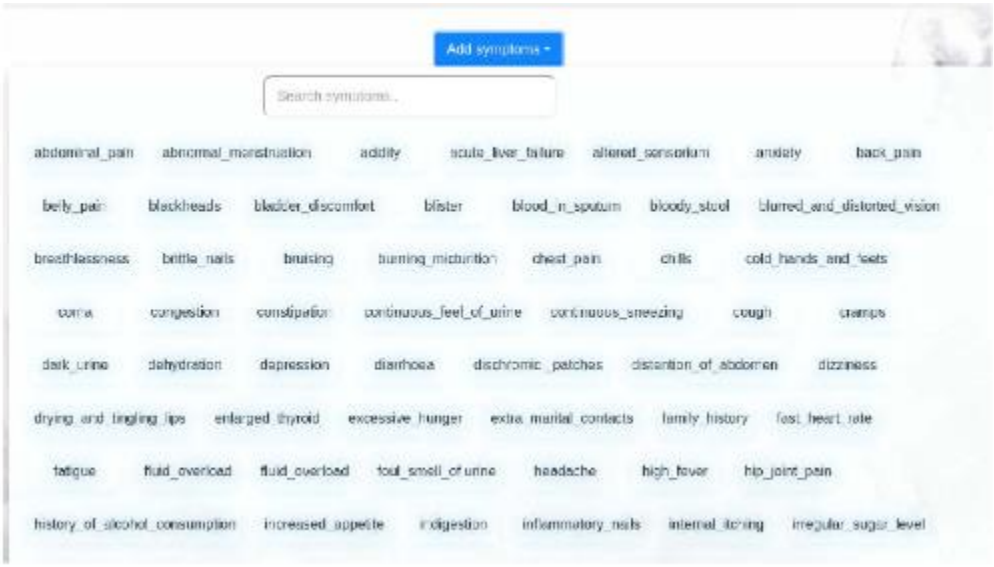


Figure 4.1: Symptoms Selection List

• The expected outcome observed will also have an UI , so that it will be more easier ,understandable and interactive to help a user to operate and predict the disease on the basis of input provided and becomes an easier task to perform.

• The UI is made in such a manner that when the project is accessed on a mobile web browser then the functions and tabs in the page adapt the size of the screen and does not create any problem while being accessed on a mobile web browser.

• This model also helps user to interact or concern a doctor at the end of showing the predicted result as in fig:6.2 which is purely based upon input which is provided by the user, in the end the user can take help of a specific doctor who would be a

specialist of the respective disease and is registered with the disease prediction system.

Figure 4.2: Doctor Consultation

• The doctors will be registered with the system on the basis of their working area and will be addressing the patients which will face the problem which are related to their department.

• The doctors upon signing in to the account will see how many patients approached them and the details of the patients, the doctors can interact with the users and give response specific to the user queries and doctors can also suggest them various medications that can be helpful to the user.• The project designed is absolutely user friendly and essential machine learning models like Naive Bayes, Random Forest and Decision Tree are applied which will help in predicting the disease in a better and easy way.

• The dataset used in the project contains 132 symptoms which are related to almost every kind of general disease and contains data in the form of 0 and 1 where 1 means that the symptom is present and 0 means that the symptom is not present.• The system also shows the probability of the disease that how much chances are there that the user is suffering from the disease which is predicted by the system which helps in better identification of the disease and a better diagnosis also as in figure:4.3

Figure 4.3: Predicted Disease

• As the result obtained is specific and depends upon the input provided by the users as symptoms should be accurate with prior knowledge , so there is no misinterpretation of disease.

• The accuracy of the disease prediction is about 86% so we can say that the dataset is not in overfitting situation and predicts the correct output.

Chapter 5

Future Work

Today's, world most of the data is computerized, the data is distributed, and it is not utilizing properly. With the help of the already present data and analyzing it, we can also use for un-known patterns. The primary motive of this project is the prediction of diseases with high rate of accuracy. For predicting the disease, we can use logistic regression algorithm, naive Bayes, sklearn in machine learning. The future scope of the paper is the prediction of diseases by using advanced techniques and algorithms in less time complexity.

A technology called CAD is more beneficial as sometimes systems are better diagnostics than Doctors. Machine Learning and its different branches are used in Cancer detection as well. It helps or can say assist in making decisions on critical cases or on therapies. Artificial intelligence plays an important role in development of many health related procedure or methods. Artificial intelligence is very common now a days in surgeries, like Robotics surgery. Since we are in the circumstances of growing population, we must need technology which can help us to meet the expectations of the patients, their flawless cure, their better health and their smooth and easy approachable access to health care industries to heal and get well soon!!

References

[1] Dahiwade, D., Patle, G., and Meshram, E. (2019). “Designing disease prediction model using machine learning approach.” 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), IEEE. 1211–1215.

[2] Fitriyani, N. L., Syafrudin, M., Alfian, G., and Rhee, J. (2019). “Development of diseaseprediction model based on ensemble learning approach for diabetes and hypertension.” IEEEAccess, 7, 144777–144789.

[3] Hong, W., Xiong, Z., Zheng, N., and Weng, Y. (2019). “A medical-history-based potentialdisease prediction algorithm.” IEEE Access, 7, 131094–131101.

[4] Kohli, P. S. and Arora, S. (2018). “Application of machine learning in disease predic-tion.” 2018 4th International Conference on Computing Communication and Automation(ICCCA), IEEE. 1–4.

[5] Krishnan.J, M. “Prediction of heart disease using machine learning algorithms.

[6] Patil, M., Lobo, V. B., Puranik, P., Pawaskar, A., Pai, A., and Mishra, R. (2018). “Aproposed model for lifestyle disease prediction using support vector machine.” 2018 9th In-ternational Conference on Computing, Communication and Networking Technologies (IC-CCNT), IEEE. 1–6.

[7] Shobana, V. and Kumar, N. (2017). “A personalized recommendation engine for predic-tion of disorders using big data analytics.” 2017 International Conference on Innovations inGreen Energy and Healthcare Technologies (IGEHT), IEEE. 1–4.