

# **DESIGN AND DEVELOPMENT OF DISEASE DETECTION USING M.L**

by

Akshita Agarwal (1601410012)

Lavi Thukral (1601410060)

Priyanshi Kotnala (1601410081)

Submitted to the Department of Computer Science & Engineering

in partial fulfillment of the requirements

for the degree of

Bachelor of Technology

in

Computer Science & Engineering



Department of Computer Science & Engineering

**Shri Ram Murli Smarak College of Engineering & Technology, Bareilly**

**Dr.A.P.J. Abdul Kalam Technical University, Lucknow**

July, 2020

# TABLE OF CONTENTS

Page

DECLARATION.....	4
CERTIFICATION.....	5
ACKNOWLEDGEMENTS.....	6
ABSTRACT.....	7
LIST OF SYMBOLS.....	8
CHAPTER 1.....	9
DESCRIPTION.....	9
1.1    INTRODUCTION.....	9-11
1.2    BACKGROUND OF PROBLEM.....	12
1.3    PROBLEM DEFINITION.....	13
1.4    CURRENT SYSTEM.....	13
1.5    ADVANTAGES.....	13
1.6    APPLICATIONS.....	13-14
1.7    LIMITATIONS.....	14
CHAPTER 2.....	15
LITERATURE REVIEW.....	15
2.1    LITERATURE SURVEY.....	15-16
CHAPTER 3.....	17
REQUIREMENT ANALYSIS.....	17
3.1    SUPPORTING OPERATING SYSTEMS.....	17
3.2    SOFTWARE REQUIREMENTS.....	17
3.3    HARDWARE REQUIREMENTS.....	17
CHAPTER 4.....	18
METHODOLOGY.....	18
4.1    MODULAR DESCRIPTION.....	18-20
4.1.1    DATA.....	21

4.1.2	CLASSIFICATION.....	21
4.1.3	DECISION TREE.....	22
4.1.4	RANDOM FOREST.....	23
4.1.5	NAÏVE BAYES.....	23-24
4.1.6	KNN ALGORITHM.....	24
4.1.7	FEATURE SELECTION.....	24
4.1.8	PRINCIPAL COMPONENT ANALYSIS.....	24-25
4.2	MEASUREMENT.....	25-26
CHAPTER 5.....		27
IMPLEMENTATION METHODOLOGIES.....		27
5.1	MACHINE LEARNING.....	27
5.2	DJANGO.....	28
CHAPTER 6.....		29
PROJECT SNAPSHOTS.....		29
6.1	SNAPSHOTS.....	29-31
CHAPTER 7.....		32
CONCLUSION AND FUTURE SCOPE.....		32
7.1	CONCLUSION .....	32
7.2	LIMITATION.....	32
7.3	FUTURE ENCHANCEMENT.....	32
REFERENCES.....		33

## DECLARATION

I hereby declare that the submission of Design and Development of Disease Detection using Machine Learning is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature.....

Name Akshita Agarwal

Roll No 1601410012

Date.....

Signature.....

Name Lavi Thukral

Roll No 1601410060

Date.....

Signature.....

Name Priyanshi Kotnala

Roll No 1601410081

Date.....

# **CERTIFICATE**

This is to certify that the Project Report entitled DESIGN AND DEVELOPMENT OF DISEASE DETECTION USING MACHINE LEARNING which is submitted by Akshita Agarwal (1601410012), Lavi Thukral (1601410060) and Priyanshi Kotnala (1601410081) is a record of the candidates own work carried out by them under my supervision. The matter embodied in this work is original and has not been submitted for the award of any other work or degree.

Dr. L. S. Maurya

**HOD (CSE/IT)**

Ms. Neha Sharma

**Project Incharge (CS1)**

Dr. L. S. Maurya

**Supervisor**

## ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Dr. L. S. Maurya, Computer Science & Technology, S.R.M.S.C.E.T, Bareilly for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Dr. L. S. Maurya, Head, Department of Computer Science & Engineering/Information Technology, S.R.M.S.C.E.T, Bareilly for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature.....

Name Akshita Agarwal

Roll No 1601410012

Date.....

Signature.....

Name Lavi Thukral

Roll No 1601410060

Date.....

Signature.....

Name Priyanshi Kotnala

Roll No 1601410081

Date.....

## **ABSTRACT**

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.

But with the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not.

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million.

Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc.

Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques.

## LIST OF SYMBOLS

Attribute  $A$  information gain:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Pre-segmentation information entropy:

$$\text{Info}(D) = \text{Entropy}(D) = -\sum_j p(j|D) \log p(j|D)$$

Distributed information entropy:

$$\text{Info}_A(D) = \sum_{i=1}^n \text{Info}(D_i)$$

We suppose the input vector is  $\vec{x}$ , the weight vector is  $\vec{w}$ , and the activation function is a sigmoid function, then the output is:

$$y = \text{sigmoid}(\vec{w}^T \cdot \vec{x})$$

and the sigmoid is:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad \text{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$



# CHAPTER-1

## DESCRIPTION

### 1.1 INTRODUCTION

A Diabetes is a common chronic disease and poses a great threat to human health. The characteristic of diabetes is that the blood glucose is higher than the normal level, which is caused by defective insulin secretion or its impaired biological effects, or both (Lonappan et al., 2007). Diabetes can lead to chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves (Krasteva et al., 2011).

Diabetes can be divided into two categories:

- Type 1 diabetes (T1D)
- Type 2 diabetes (T2D)

Patients with Type 1 diabetes are normally younger, mostly less than 30 years old. The typical clinical symptoms are increased thirst and frequent urination, high blood glucose levels. This type of diabetes cannot be cured effectively with oral medications alone and the patients are required insulin therapy.

Type 2 diabetes occurs more commonly in middle-aged and elderly people, which is often associated with the occurrence of obesity, hypertension, dyslipidemia, arteriosclerosis, and other diseases.

There is no doubt that this alarming figure needs great attention. With the rapid development of machine learning, machine learning has been applied to many aspects of medical health. The results showed that prediction with random forest could reach the highest accuracy ( $ACC = 0.8084$ ) when all the attributes were used. Machine learning methods are widely used in predicting diabetes, and they get preferable results.

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.

With the development of living standards, diabetes is increasingly common in people's daily life. Therefore, how to quickly and accurately diagnose and analyze diabetes is a topic worthy studying. In medicine, the diagnosis of diabetes is according to fasting blood glucose, glucose tolerance, and random blood glucose levels. The earlier diagnosis is obtained, the much easier we can control it.

Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors. For machine learning method, how to select the valid features and the correct classifier are the most important problems.

Recently, numerous algorithms are used to predict diabetes, including the traditional machine learning method such as:

- Support Vector Machine (SVM)
- Decision Tree (DT)
- Logistic Regression

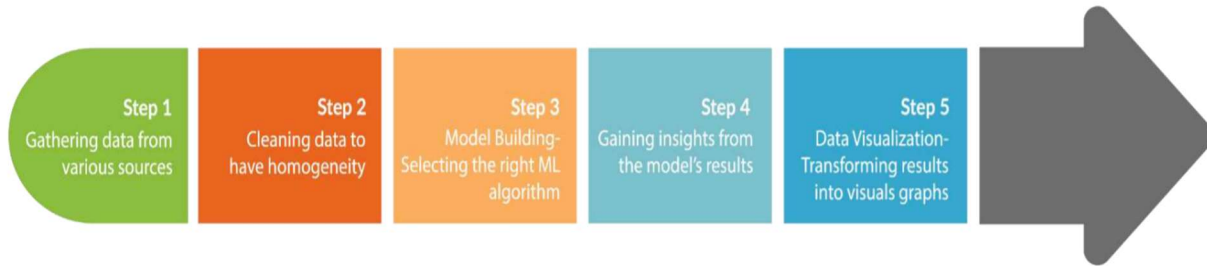
and so on...

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable.

Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" Although data mining has been around for more than two decades, its potential is only being realized now.

Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms.

# The Machine Learning Process



## Types of Diabetes



## 1.2 BACKGROUND OF THE PROBLEM

The term diabetes is the shortened version of the full name diabetes mellitus. Diabetes mellitus is derived from the Greek word *diabetes* meaning siphon - to pass through and the Latin word *mellitus* meaning honeyed or sweet. This is because in diabetes excess sugar is found in blood as well as the urine. It was known in the 17th century as the “pissing evil”.

The term diabetes was probably coined by Apollonius of Memphis around 250 BC. Diabetes is first recorded in English, in the form diabetes, in a medical text written around 1425. It was in 1675 that Thomas Willis added the word “mellitus” to the word diabetes. This was because of the sweet taste of the urine. This sweet taste had been noticed in urine by the ancient Greeks, Chinese, Egyptians, Indians, and Persians as is evident from their literature.

Sushruta, Arataeus, and Thomas Willis were the early pioneers of the treatment of diabetes. Greek physicians prescribed exercise - preferably on horseback to alleviate excess urination. Some other forms of therapy applied to diabetes include wine, overfeeding to compensate for loss of fluid weight, starvation diet, etc.

In 1776, Matthew Dobson confirmed that the sweet taste of urine of diabetics was due to excess of a kind of sugar in the urine and blood of people with diabetes.

In ancient times and medieval ages diabetes was usually a death sentence. Aretaeus did attempt to treat it but could not give a good outcome. Sushruta an Indian healer identified diabetes and classified it as “Madhumeha”. Here the word “madhu” means honey and combined the term means sweet urine. The ancient Indians tested for diabetes by looking at whether ants were attracted to a person's urine. The Korean, Chinese, and Japanese words for diabetes are based on the same ideographs which mean “sugar urine disease”.

In Persia Avicenna provided a detailed account on diabetes mellitus in “The Canon of Medicine”. He described abnormal appetite and the decline of sexual functions along with sweet urine. He also identified diabetic gangrene. Avicenna was the first to describe diabetes insipidus very precisely. It was much later in the 18th and 19th century that Johann Peter Frank differentiated between diabetes mellitus and diabetes insipidus.

## **1.3 PROBLEM DEFINITION**

The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on following classification methods namely, Logistic Regression, Naïve Bayes and Random Forest. we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not.

## **1.4 CURRENT SYSTEMS**

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain.

## **1.5 ADVANTAGES**

- i. No waste of Money.
- ii. Save time
- iii. The usage of this application greatly reduces the time required for your report
- iv. The application also leads to quicker decision making to know whether person is diabetic or not.

## **1.6 APPLICATIONS**

This system can be used to by any Hospital related to diabetes prediction which may help their patient or agencies to help the users in gaining an advantage over a Paper Report.

Machine Learning is concerned with the development of algorithms and techniques that allows

the computers to learn and gain intelligence based on the past experience. It is a branch of Artificial Intelligence (AI) and is closely related to statistics. By learning it means that the system is able to identify and understand the input data, so that it can make decisions and predictions based on it.

- Now days, machine learning algorithms are used for automatic analysis of high dimensional biomedical data.

- Diagnosis of liver disease, skin lesions, cancer classification, risk assessment for cardiovascular disease and analysis of genetic and genomic data are some of the examples of biomedical application of ML.

- For liver disease diagnosis, Hashemi et al. (2012) has successfully implemented SVM algorithm.

- In order to diagnose major depressive disorder (MDD) based on EEG dataset, Mumtaz et al. (2017) have used classification models such as support vector machine (SVM), logistic regression(LR) and Naïve Bayesian (NB)

Diabetes is a very common metabolic disease. Usually onset of type 2 diabetes happens in middle age and sometimes in old age. But nowadays incidences of this disease are reported in children as well. There are several factors for developing diabetes like genetic susceptibility, body weight, food habit and sedentary lifestyle.

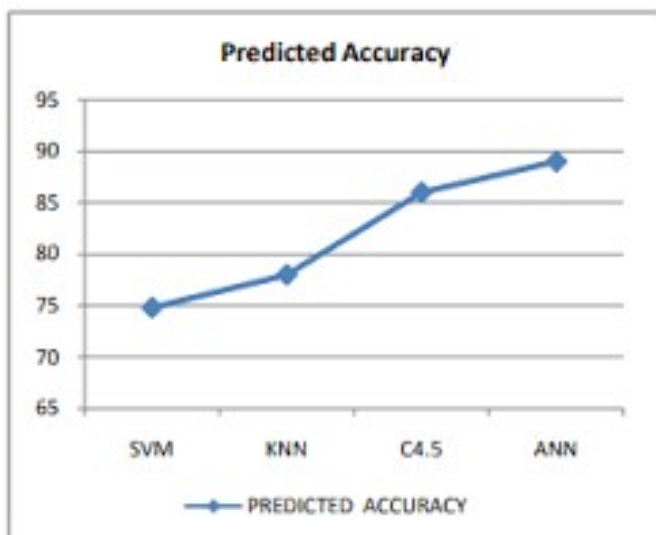
Undiagnosed diabetes may result in very high blood sugar level referred as hyperglycemia which can lead to complication like diabetic retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer.

## CHAPTER -2

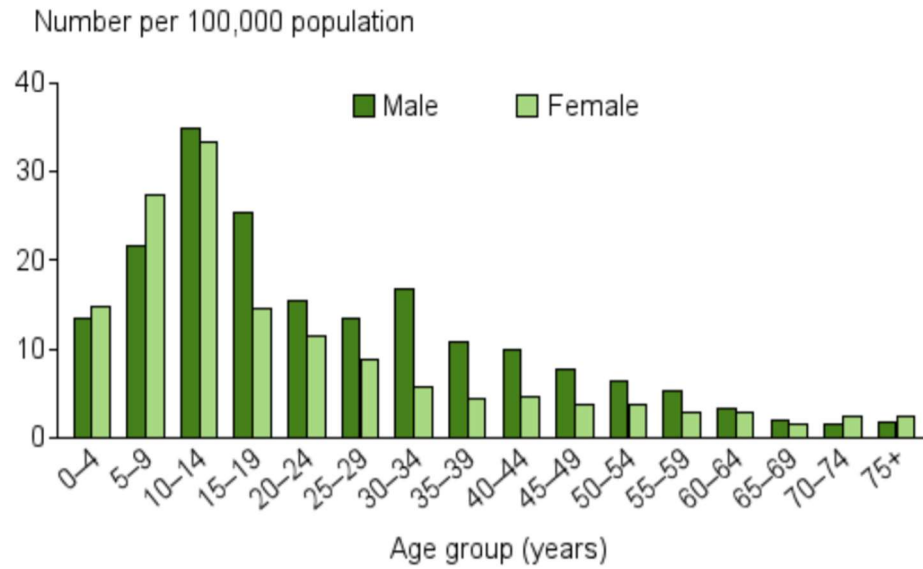
### LITERATURE REVIEW

#### 2.1 LITERATURE SURVEY

Predictive analytics which is help to healthcare organizations to evaluate data on the past behavior and predict likelihood of future behavior to enable better decisions and outcomes of their patient. Predictive models can make human decisions more effective and highly automate an entire decision-making process. It increasingly, predictive analytics uses data to improve safety and performance of patient outcomes. Here, the Modernizing healthcare industry's move towards processing massive health records, and to access those for analysis and this will greatly increases the complexities. Due to the unstructured nature of Machine Learning form health industry, it is necessary to structure and emphasis their size into nominal value with possible solution. Healthcare industry faces many challenges that make us to know the importance to develop the data analytics of the diabetes mellitus.



## Diabetes compendium





## **CHAPTER-3**

### **REQUIREMENT ANALYSIS**

#### **3.1 SUPPORTING OPERATING SYSTEMS**

The supported Operating Systems for client include:

- i. Windows 7(ultimate, enterprise) or higher

Windows is two of the operating systems that will support comparative applications.

Windows is a Meta family of graphical operating systems developed consists of several families of operating systems. The project is developed on Windows.

#### **3.2 SOFTWARE REQUIREMENTS**

- i. A JSON parsing library such as Json pickle
- ii. Machine Learning/Data Science libraries such as numpy and sci-kit learn
- iii. Jupiter Notebook
- iv. Pima Dataset API
- v. VMware

#### **3.3 HARDWARE REQUIREMENTS**

- i. i3 Processor Based Computer or higher
- ii. Memory: 1 GB RAM
- iii. Hard Drive: 50 GB
- iv. Monitor
- v. Internet Connection

# CHAPTER-4

## METHODOLOGY

### 4.1 MODULAR DESCRIPTION

The system comprises of 3 major modules as follows:

- a. Check Diabetes (By providing Details like)
  - Pregnancies
  - Glucose
  - Blood Pressure
  - Skin Thickness
  - Insulin
  - BMI (Body Mass Index)
  - Diabetes Pedigree Function
  - Age

Diabetes Disease Prediction

You don't have diabetes disease.

Pregnancies:

Glucose:

Blood Pressure:

Skin Thickness:

Insulin:

BMI:

Diabetes Pedigree Function:

Age:

Activate Windows  
Go to Settings to activate Windows.

b. Check Heart Disease (By providing Details like)

- Age
- Sex
- Chest Pain
- Trestbps (resting blood pressure)
- Cholestrol
- Fbs (fasting blood sugar)
- Restecg (resting electrocardiographic results)
- Thalach (maximum heart rate achieved)
- Exang (exercise induced angina)
- Oldpeak (depression induced by exercise relative to rest)
- Slope (the slope of the peak exercise ST segment)
- CA ( number of major vessels colored by flourosopy)
- THAL

Heart Disease Prediction

Age	FBS	SLOPE
<input type="text" value="63"/>	<input type="text" value="1"/>	<input type="text" value="3"/>
Sex	RESTECG	CA
<input type="text" value="1"/>	<input type="text" value="0"/>	<input type="text" value="0"/>
CP	THALACH	THAL
<input type="text" value="3"/>	<input type="text" value="150"/>	<input type="text" value="0"/>
TRESTBPS	EXANG	<input type="button" value="Submit"/>
<input type="text" value="145"/>	<input type="text" value="0"/>	
CHOL	OLDPEAK	
<input type="text" value="233"/>	<input type="text" value="2"/>	

Activate Windows  
Go to Settings to activate Windows.

c. Check Breast Cancer (By providing Details like)

- Mean Radius
- Mean Texture
- Mean Perimeter
- Mean Area
- Mean Smoothness

The screenshot shows a web browser window with the address bar displaying "127.0.0.1:8000/breast". The page title is "Disease Predictor". The main content area has a blue background and is titled "Breast Cancer Prediction". It contains six input fields, each with a label and a value of "0":

- Mean Radius
- Mean Texture
- Mean Perimeter
- Mean Area
- Mean Smoothness

Below the input fields is a red "Submit" button. In the bottom right corner, there is a "Activate Windows" watermark with the text "Go to Settings to activate Windows." The Windows taskbar is visible at the bottom, showing the Start button, search icon, and several application icons. The system tray shows the language as "ENG US", the time as "02:03 PM", and the date as "30-07-2020".

### 4.1.1 DATA

The dataset was obtained from hospital physical examination data in Luzhou, China. This dataset is divided into two parts: the healthy people and the diabetes. There are two healthy people physical examination data. We used one of healthy people physical examination data that contains 164431 instances as the training set. In the other data set, 13700 samples were randomly selected as an independent test set. The physical data include 14 physical examination indexes: age, pulse rate, breathe, left systolic pressure (LSP), right systolic pressure (RSP), left diastolic pressure (LDP), right diastolic pressure (RDP), height, weight, physique index, fasting glucose, waistline, low density lipoprotein (LDL), and high density lipoprotein (HDL). In the training dataset, there are many missing data. We deleted the abnormal and missing samples to reduce the impact of data processing on result. Consequently, we got 151598 diabetic physical data and 69082 healthy people physical data. So, we randomly selected 68994 healthy people and diabetic patients' data, respectively as training set. Due to the data unbalance, we randomly extracted 5 times. The final result was the mean value of 5 experiments. The 13,700 patients physical examination data, which were randomly selected as the independent test set, were different from the previous five sets which were used as training set.

Another dataset is Pima Indians diabetics data ([Jegan, 2014](#)). In particular, all patients are females at least 21 years old of Pima Indian heritage. The dataset contains 8 attributes which are times of pregnancy, plasma glucose concentration after an 2-h oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-h serum insulin, body mass index, diabetes pedigree function and age. In this dataset, the original 786 diabetics data reduces to 392 after deleted the missing data.

### 4.1.2 CLASSIFICATION

In this section, we used decision tree, RF and neural network as the classifiers. Decision tree and RF can implement in WEKA, which is a free, non-commercial, open source machine learning and data mining software based on JAVA environment. Neural network can be implemented in MATLAB, which is a commercial mathematics software exploited by MathWorks, Inc. It is used for algorithmic development, data visualization, data analysis and provides advanced computational language, and interactive environment for numerical calculation.

### 4.1.3 DECISION TREE

Decision tree is a basic classification and regression method. Decision tree model has a tree structure, which can describe the process of classification instances based on features). It can be considered as a set of if-then rules, which also can be thought of as conditional probability distributions defined in feature space and class space.

Decision tree uses tree structure and the tree begins with a single node representing the training samples. If the samples are all in the same class, the node becomes the leaf and the class marks it. Otherwise, the algorithm chooses the discriminatory attribute as the current node of the decision tree. According to the value of the current decision node attribute, the training samples are divided into several subsets, each of which forms a branch, and there are several values that form several branches. For each subset or branch obtained in the previous step, the previous steps are repeated, recursively forming a decision tree on each of the partitioned samples.

The typical algorithms of decision tree are ID3, C4.5, CART and so on. In this study, we used the J48 decision tree in WEKA. J48 another name is C4.8, which is an upgrade of C4.5. J48 is a top-down, recursive divide and conquer strategy. This method selects an attribute to be root node, generates a branch for each possible attribute value, divides the instance into multiple subsets, and each subset corresponds to a branch of the root node, and then repeats the process recursively on each branch. When all instances have the same classification, the algorithm stops. In J48, the nodes are decided by information gain. According to the following formulas, in each iteration, J48 calculates the information gain of each attribute, and selects the attribute with the largest value of information gain as the node of this.

Attribute  $A$  information gain:

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

Pre-segmentation information entropy:

$$\text{Info}(D) = \text{Entropy}(D) = -\sum_j p(j|D) \log_2 p(j|D)$$

Distributed information entropy:

$$\text{Info}_A(D) = \sum_{i=1}^n \frac{n_i}{n} \text{Info}(D_i)$$

## 4.1.4 RANDOM FOREST

RF is a classification by using many decision trees. This algorithm proposed by Breiman. RF is a multifunctional machine learning method. It can perform the tasks of prediction and regression. In addition, RF is based on Bagging and it plays an important role in ensemble machine learning. RF has been employed in several biomedicine research.

RF generates many decision trees, which is very different from decision tree. When the RF is predicting a new object based on some attributes, each tree in RF will give its own classification result and 'vote,' and then the overall output of the forest will be the largest number of taxonomy. In the regression problem, the RF output is the average value of output of all decision trees.

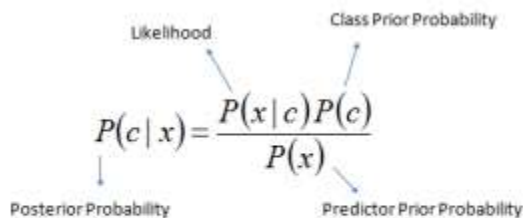
## 4.1.5 NAÏVE BAYES

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Look at the equation below:



The diagram shows the equation  $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$  with labels and arrows: 'Likelihood' points to  $P(x|c)$ , 'Class Prior Probability' points to  $P(c)$ , 'Posterior Probability' points to  $P(c|x)$ , and 'Predictor Prior Probability' points to  $P(x)$ .

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Above,

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

- $P(x|c)$  is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$  is the prior probability of *predictor*.

### 4.1.6 KNN ALGORITHM

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2. Calculation time

3. Predictive Power

- **Lazy learning algorithm** – KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
- **Non-parametric learning algorithm** – KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

### 4.1.7 FEATURE SELECTION

Feature selection methods can reduce the number of attributes, which can avoid the redundant features. There are many feature selection methods. In this study, we used PCA to reduce the dimensionality.

### 4.1.8 PRINCIPAL COMPONENT ANALYSIS

PCA obtains the  $K$  vectors and unit eigenvectors by solving the characteristic equation of the correlation matrix of the observed variables. The eigenvalues are sorted from large to small, representing the variance of the observed variables explained by  $K$  principal components, respectively.

The model for extracting principal component factors is:



$$F_i = T_{i1}X_1 + T_{i2}X_2 + \dots + T_{ik}X_k \quad (i=1, 2, \dots, m)$$

where,  $F_i$  is the  $i$  principal component factor;  $T_{ij}$  is the load of the  $i$  principal component factor on the  $j$  index;  $m$  is the number of principal component factors;  $k$  is the number of indicators.

The PCA method can reduce the original multiple indicators to one or more comprehensive indicators. This small number of comprehensive indicators can reflect the vast majority of the information reflected by the original indicators, and they are not related to each other, and they can avoid the repeated information (Jackson, 1993; Jolliffe, 1998). At the same time, the reduction of indicators facilitates further calculation, analysis and evaluation.

We used Statistical Product and Service Solutions (SPSS) to implement the PCA algorithm. SPSS is a general term for a series of software products and related services launched by IBM. It is mainly used for statistical analysis, data mining, predictive analysis and other tasks. SPSS has a friendly visual interface and is easy to operate.

## 4.2 MEASUREMENT

In this study, we used sensitivity (SN), specificity (SP), accuracy (ACC), and Matthews correlation coefficient (MCC) to measure the classified effectiveness. And the formulas are as follow:

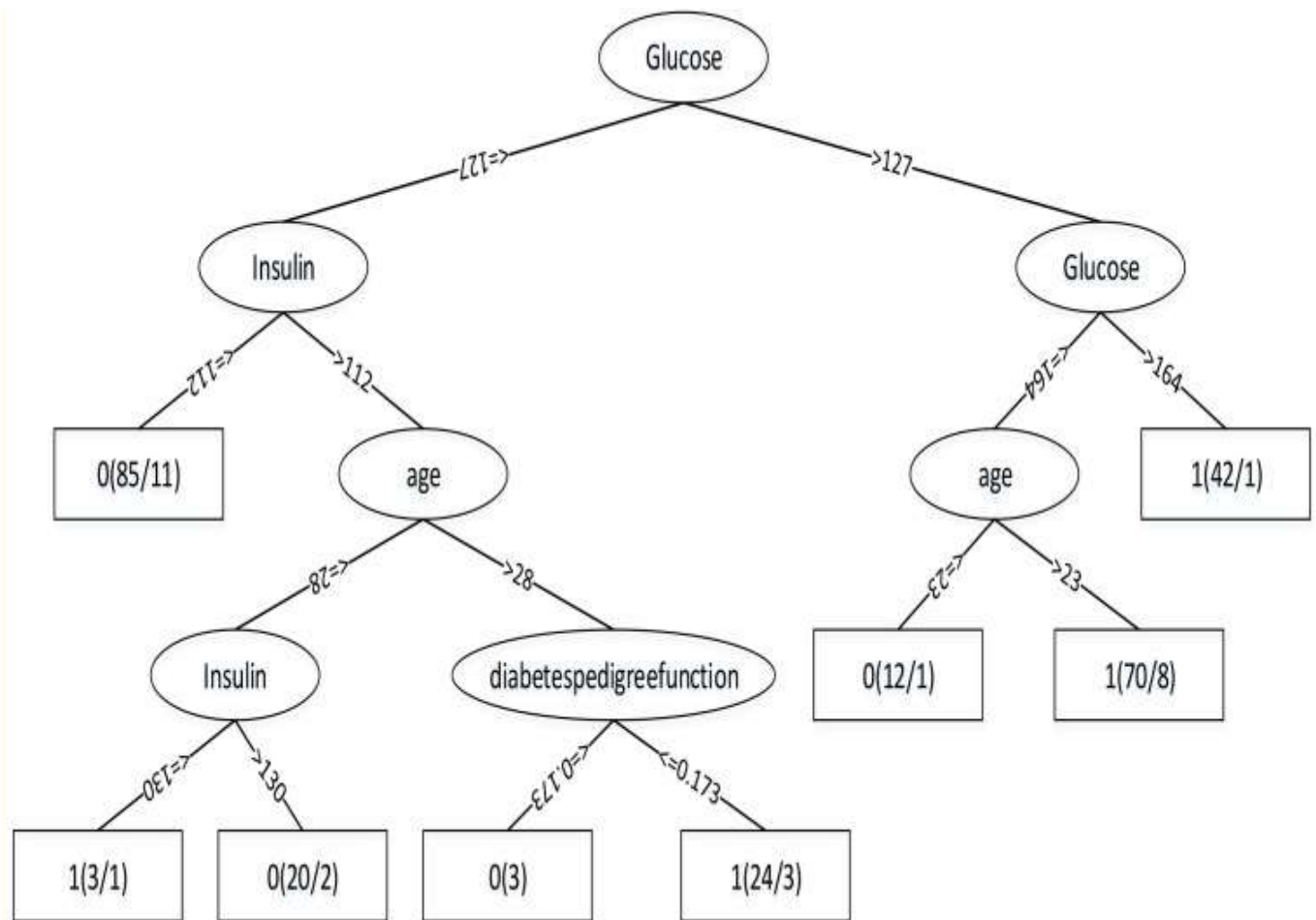
$$SN = TP / (TP + FN)$$

$$SP = TN / (TN + FP)$$

$$ACC = (TN + TP) / (TN + FP + TP + FN)$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

where true positive represents (TP) the number of identified positive samples in the positive set. True negative (TN) means the number of classification negative samples in the negative set. False positive (FP) is the number of the number of identified positive samples in the negative set. And false negative (FN) represents the number of identified negative samples in the positive set. It is often used to evaluate the quality of classification models. The accuracy is defined as the ratio of the number of samples correctly classified by the classifier to the total number of samples. In medical statistics, there are two basic characteristics, sensitivity (SN) and specificity (SP). Sensitivity is the true positive rate, and specificity is the true negative rate. The MCC is a correlation coefficient between the actual classification and the predicted classification. Its value range is [-1, 1]. When the MCC equals one, it indicates a perfect prediction for the subject. When the MCC value is 0, it indicates the predicted result is not as good as the result of random prediction, and -1 means that the predicted classification is completely inconsistent with the actual classification.



Decision tree structure by using all features and Pima Indians dataset. From this figure, we can find in this method glucose as the root node, which can indicate the index has the highest information gain and insulin and age play important roles in this method.

# **CHAPTER-5**

## **IMPLEMENTATION METHODOLOGIES**

### **5.1 MACHINE LEARNING**

ML algorithms are very well-known in the medical field for predicting diseases. Many researchers have used ML techniques to predict diabetes in an effort to obtain the best and most accurate results .

#### **INDIVIDUAL ELEMENTS OF Machine learning**

The machine learning algorithms can be roughly categorized into three types namely supervised learning, unsupervised learning and semi-supervised learning.

The supervised learning algorithms are used when human expertise does not exist (navigating on Mars), humans are unable to explain their expertise (speech recognition). Solution changes in time series (routing on a computer function) and to solution needs to be adapted to particular cases (user biometrics). The supervised learning algorithms are classified into different types such as probability-based, function-based, rule-based, tree-based, instance-based, etc.

The unsupervised learning is the descriptive type learning. This learning is used to describe or summarize the data. The examples of the unsupervised learning algorithms are clustering, association rule mining, etc.

The semi-supervised learning is the combination of supervised and unsupervised. This presents a diabetes prediction system to diagnosis the diabetics.

Moreover, the supervised learning algorithm is used to learn the diabetes data and to develop diabetes predication system for diagnosing diabetes. The accuracy of this prediction system is improved using pre-processing technique.

## 5.2 DJANGO

**Django** is a Python-based free and open-source web framework that follows the model-template-view (MVC) architectural pattern. It is maintained by the Django Software Foundation (DSF), an American independent organization established as a 501(c)(3) non-profit.

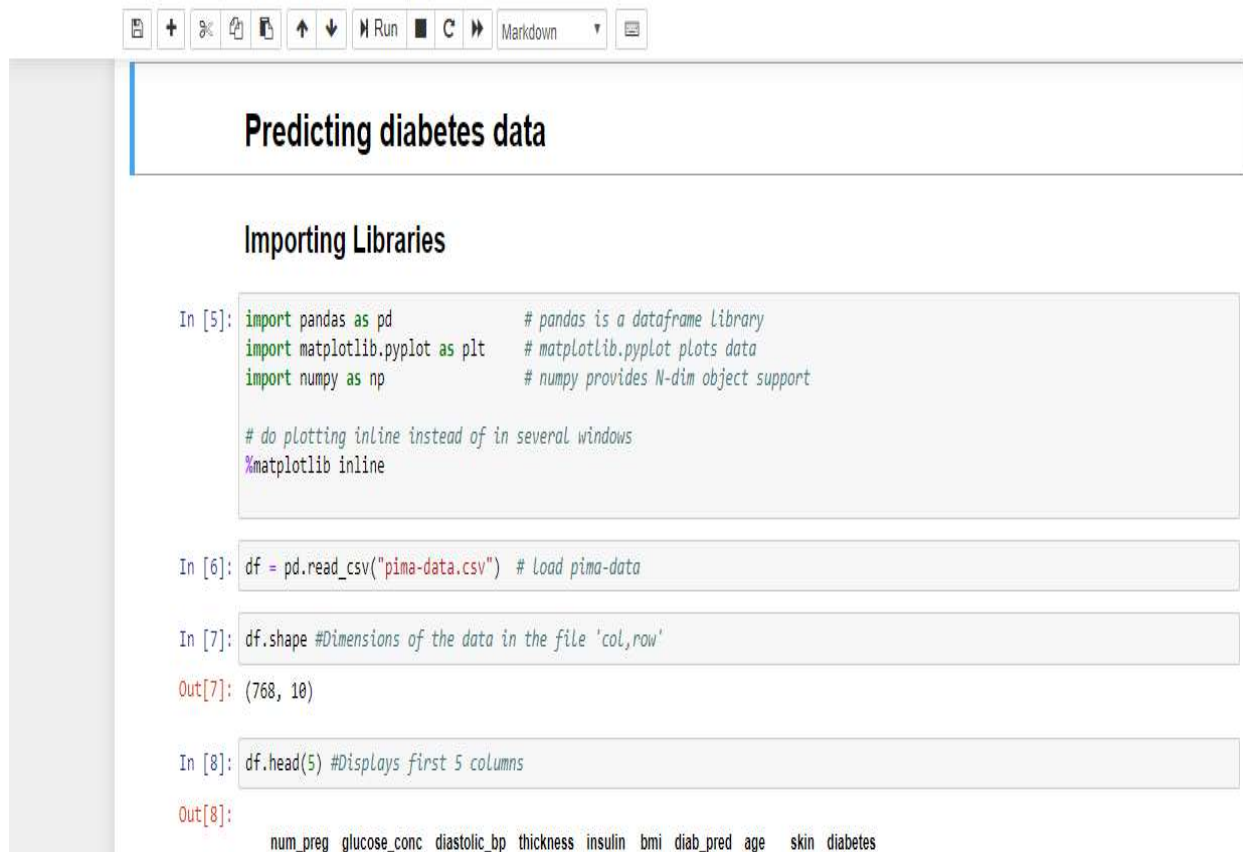
Django's primary goal is to ease the creation of complex, database-driven websites. The framework emphasizes reusability and "pluggability" of components, less code, low coupling, rapid development, and the principle of don't repeat yourself. Python is used throughout, even for settings files and data models. Django also provides an optional administrative create, read, update and delete interface that is generated dynamically through introspection and configured via admin models.

Some well known sites that use Django include PBS, Instagram, Mozilla, *The Washington Times*, Disqus, Bitbucket and Nextdoor.

# CHAPTER-6

## PROJECT SNAPSHOTS

### 6.1 SNAPSHOTS



The image shows a Jupyter Notebook interface with a toolbar at the top containing icons for file operations, execution, and markdown. The notebook has a title bar that says "Predicting diabetes data". The content is organized into sections and code cells. The first section is "Importing Libraries", followed by a code cell for importing pandas, matplotlib, and numpy. The next code cell loads the "pima-data.csv" file. The following code cell prints the shape of the data. The final code cell prints the first 5 columns of the data.

Predicting diabetes data

#### Importing Libraries

```
In [5]: import pandas as pd          # pandas is a dataframe library
import matplotlib.pyplot as plt    # matplotlib.pyplot plots data
import numpy as np                # numpy provides N-dim object support

# do plotting inline instead of in several windows
%matplotlib inline
```

```
In [6]: df = pd.read_csv("pima-data.csv") # Load pima-data
```

```
In [7]: df.shape #Dimensions of the data in the file 'col,row'
```

```
Out[7]: (768, 10)
```

```
In [8]: df.head(5) #Displays first 5 columns
```

```
Out[8]:
```

num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
----------	--------------	--------------	-----------	---------	-----	-----------	-----	------	----------

jupyter Python-Pima-data Last Checkpoint: 24 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

In [8]: `df.head(5)` #Displays first 5 columns

Out[8]:

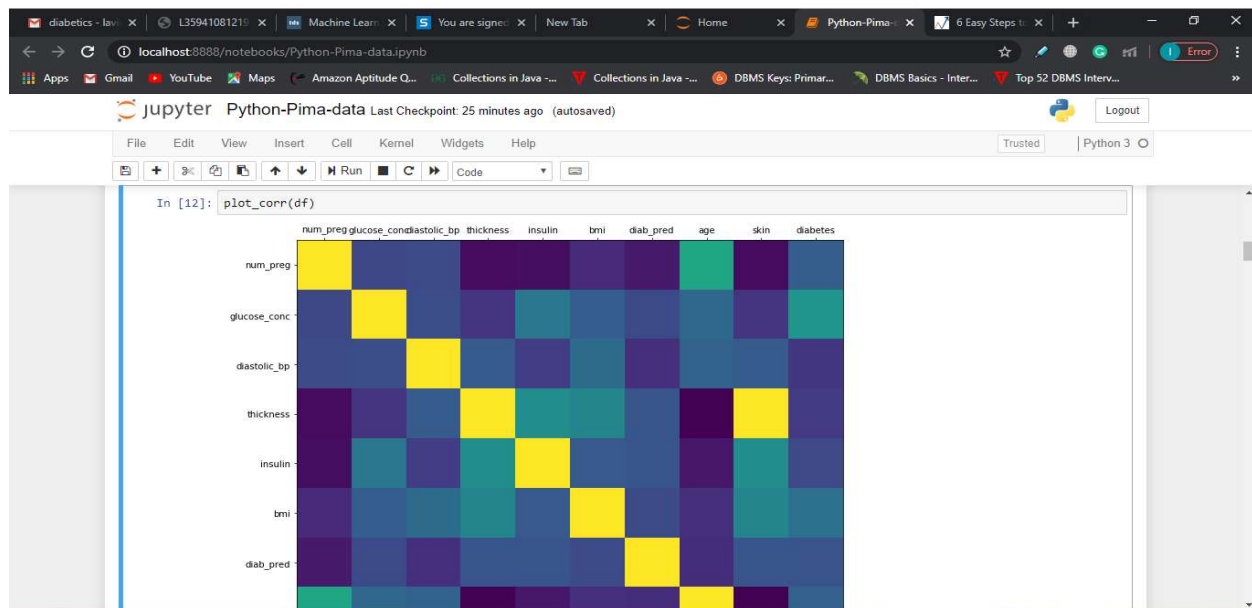
	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
0	6	148	72	35	0	33.6	0.627	50	1.3790	True
1	1	85	66	29	0	26.6	0.351	31	1.1426	False
2	8	183	64	0	0	23.3	0.672	32	0.0000	True
3	1	89	66	23	94	28.1	0.167	21	0.9062	False
4	0	137	40	35	168	43.1	2.288	33	1.3790	True

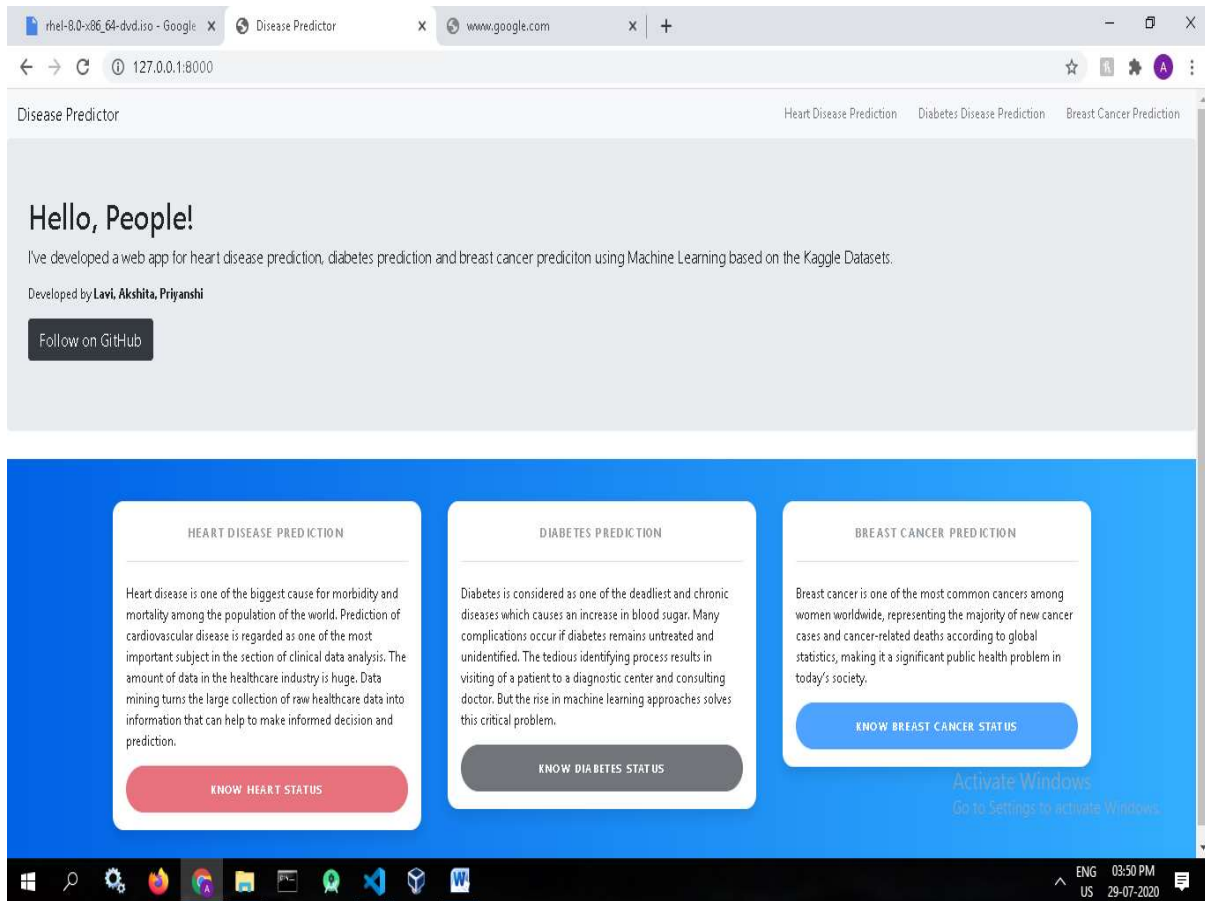
In [9]: `df.tail(5)` #Displays last 5 columns

Out[9]:

	num_preg	glucose_conc	diastolic_bp	thickness	insulin	bmi	diab_pred	age	skin	diabetes
763	10	101	76	48	180	32.9	0.171	63	1.8912	False
764	2	122	70	27	0	36.8	0.340	27	1.0638	False
765	5	121	72	23	112	26.2	0.245	30	0.9062	False
766	1	126	60	0	0	30.1	0.349	47	0.0000	True
767	1	93	70	31	0	30.4	0.315	23	1.2214	False

In [10]: `df.isnull().values.any()` #Checking to see if there are any empty spaces in the table





# **CHAPTER-7**

## **CONCLUSION AND FUTURE SCOPE**

### **7.1 CONCLUSION**

Since Diabetes mellitus is a chronic disease characterized by hyperglycemia. It may cause many complications. According to the growing morbidity in recent years, in 2040, the world's diabetic patients will reach 642 million, which means that one of the ten adults in the future is suffering from diabetes.

There is no doubt that this alarming figure needs great attention.. Most people

without using the latest technology waste a lot of time and money in Paper report. So, an

application like DESIGN AND DEVELOPMENT OF DIABTIES PREDICTION USING DATA MINING really helps people to utilize their precious time to the fullest and get their report at the same time.

### **7.2 LIMITATIONS**

- i. It requires active internet connection else error may occur.
- ii. Since for the data the system is dependent so if anything goes wrong with the foursquare, the system is liable to give wrong data.

### **7.3 FUTURE ENHANCEMENTS**

As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment. Because it is based on object oriented design, any further changes can be easily adaptable. For future upgrades we will be linking our application with the databases of various different diseases.



## REFERENCES

1. Alghamdi M., Al-Mallah M., Keteyian S., Brawner C., Ehrman J., Sakr S. . Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the henry ford exercise testing (FIT) project. *PLoS One* 12:e0179805. 10.1371/journal.pone.0179805 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
2. American Diabetes Association (2012). Diagnosis and classification of diabetes mellitus. *Diabetes Care* 35(Suppl. 1), S64–S71. 10.2337/dc12-s064 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
3. Bengio Y., Grandvalet Y. (2005). *Bias in Estimating the Variance of K -Fold Cross-Validation*. New York, NY: Springer, 75–95. 10.1007/0-387-24555-3\_5 [CrossRef] [Google Scholar]
4. Breiman L. (2001). Random forest. *Mach. Learn.* 45 5–32. 10.1023/A:1010933404324 [CrossRef] [Google Scholar]
5. Chen X. X., Tang H., Li W. C., Wu H., Chen W., Ding H., et al. (2016). Identification of bacterial cell wall lyases via pseudo amino acid composition. *Biomed. Res. Int.* 2016:1654623. 10.1155/2016/1654623 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
6. Cox M. E., Edelman D. (2009). Tests for screening and diagnosis of type 2 diabetes. *Clin. Diabetes* 27 132–138. 10.2337/diaclin.27.4.132 [CrossRef] [Google Scholar]
7. Duygu ç., Esin D. (2011). An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. *Expert Syst. Appl.* 38 8311–8315. [Google Scholar]
8. Friedl M. A., Brodley C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sens. Environ.* 61 399–409. [Google Scholar]
9. Georga E. I., Protopappas V. C., Ardigo D., Marina M., Zavaroni I., Polyzos D., et al. (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE J. Biomed. Health Inform.* 17 71–81. 10.1109/TITB.2012.2219876 [PubMed] [CrossRef] [Google Scholar]
10. Habibi S., Ahmadi M., Alizadeh S. (2015). Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining. *Glob. J. Health Sci.* 7 304–310. 10.5539/gjhs.v7n5p304 [PMC free article] [PubMed] [CrossRef] [Google Scholar]